STATISTICAL TECHNIQUES B

Probability

1. Introduction

There is an experiment whose *outcome* is random and which can be repeated. Define Ω as the *sample space*, which is the set of all possible outcomes of the experiment and the basic (or elementary) outcomes are defined as C_i, i=1,…k, and these are the list of all possible outcomes and only one of the list can be the outcome in any particular experiment.

Example:

(i) If the experiment is rolling a dice then $\Omega = \{1,2,3,4,5,6\}$ and C_i =the number on the face of the die; $C_1 = \{1\}$, $C_2 = \{2\} \ldots C_6 = \{6\}$

(ii) If the experiment was rolling two die then $\Omega = \{(1,1),(1,2),\ldots(1,6)\ldots(6,6)\}$ and C_i is the number on the faces combined.

The *events*, A, are a collection of one or more outcomes of the basic outcomes (C_i) and are a subset of Ω.

Example:

(i) Let A = the event that an odd face is rolled = {1,3,5} and $A \subset \Omega$.

(ii) Let A = Ordered pair of Ω which sum to 7 = {(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)}.

2. Set Theory and Venn Diagrams

A set is a selection of objects. Members of a set are referred to as elements and are written inside brackets, $\{\}$. Some notation:

(1) $\in$ ($\notin$) means belongs (does not belong) to

(a) The set $A = \{1, 2, 3, 4\}$ and $1 \in A$

(b) The set $B = \{x \mid 0 \leq x \leq 1\}$ and $0 \in$ B

(c) The set $C = \{\text{Heads,Tails}\}$ and Heads $\in$ C

(d) The set $D = \{\text{real numbers}|\text{x}^2 = -1\}$ and $D = \varnothing$

(2) $E \subset A$, meaning the set *E* is a subset of the set *A*, implying that all elements that belong to *E* also belong to *A*.

(a) $A = \{1, 2, 3, 4\}$ and $E = \{1, 2\}$ then $E \subset A$

(3) $A \cup E$ is the set of all elements which are in *A* or *E* or both

(a) $A = \{0, 0.5, 1\}$, $E = \{1, 2, 3\} \Rightarrow A \cup E = \{0, 0.5, 1, 2, 3\}$

(4) $A \cap E$ is the set of all elements which belong to both A and E

(a) $A = \{0, 0.5, 1\}$, $E = \{1, 2, 3\} \Rightarrow A \cap E = \{1\}$

(5) $\overline{A}$ is the complementary event to *A*, and $\overline{A} = \left\{a \in \Omega; a \notin A\right\}$

(a) $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, $A = \{1, 2, 3, 4\} \Rightarrow \overline{A} = \{5, 6, 7, 8, 9, 10\}$

(6) The elementary events $C_1, C_2, \ldots C_k$ are said to mutually exclusive, meaning $C_i \cap C_j = \varnothing$ for $i \neq j$ and exhaustive, meaning $C_1 \cup C_2 \cup C_3 \cup ... \cup C_k = \Omega$. If *A* is some other event then:

$$A = (C_1 \cap A) \cup (C_2 \cap A) \cup \ldots (C_k \cap A)$$

This is the "union-intersection" rule.

2.1 Rules

(1) $A \cup B = B \cup A$

(2) $A \cap B = B \cap A$

(3) $A \cup (B \cup C) = (A \cup B) \cup C$

(4) $A \cap (B \cap C) = (A \cap B) \cap C$

(5) $A \cap \varnothing = \varnothing$

(6) $A \cup \varnothing = A$

(7) $\overline{\overline{A}} = A$

(8) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(9) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(10) $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$

(11) $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$

3. Probability

The probability of an event *A*, *P(A)*, is the probability that an outcome of the experiment is *A*.

Example:

(i) *A* is made up of 3 outcomes and Ω of 6, then *P(A*)=3/6=1/2.

(ii) *A* is made up of 6 outcomes, Ω of 36, then *P(A)*=6/36=1/6.

Let *P* be a function which assigns a real number, *P(A)*, to *A*, $\forall A \subset \Omega$. Then *P* is a probability measure if:

(i) *P(A)*≥0

(ii) If $C_i \cap C_j = \varnothing$, for *i≠j*, then $P(C_1 \cup C_2 \cup C_3 \ldots) = P(C_1) + P(C_2) + \ldots$

(iii) *P(Ω)*=1

3.1 Useful properties of probability

(i) Let $\overline{A}$ be the complementary event to *A*, then $\overline{A} = \{a \in \Omega; a \notin A\}$ (where $\in(\notin)$ belongs to (does not belong to)), $P(\overline{A}) = 1 - P(A)$

Proof

$A \cap \overline{A} = 0 \Rightarrow P(A \cup \overline{A}) = P(A) + P(\overline{A})$

but, $A \cup \overline{A} = \Omega \Rightarrow P(A \cup \overline{A}) = P(A) + P(\overline{A}) = P(\Omega) = 1 \Rightarrow P(\overline{A}) = 1 - P(A)$.

(ii) $P(0) = 0$

(iii) If $A \subset B$ then $P(A) \le P(B)$

Proof

Let $D = B \cap \overline{A} \Rightarrow D \cap A = 0$ Therefore, $P(D \cup A) = P(B) = P(D) + P(A)$ as $P(D) \ge 0$ this implies $P(B) \ge P(A)$

(iv) For each event $C_i \in \Omega$, $0 \le P(C_i) \le 1$

(v) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof

$P(A) = P(A \cap B) + P(A \cap \overline{B})$ and $P(B) = P(A \cap B) + P(\overline{A} \cap B)$

$$P(A \cup B) = P(A) + P(D) = P(A) + \underbrace{P(\overline{A} \cap B)}_{P(B) - P(A \cap B)} = P(A) + P(B) - P(A \cap B)$$

If *A* and *B* are mutually exclusive, such that $A \cap B = \varnothing$ then,

$P(A \cup B) = P(A) + P(B)$.

4. Bivariate probabilities

Consider the two events *A* and *B*, which have elementary events $C_1^A, C_2^A, \ldots C_h^A$ and $C_1^B, C_2^B, \ldots C_k^B$, such that:

Each event C_i^A can occur jointly with any C_j^B and these joint outcomes can be thought of as the basic outcomes.

Under this scenario we can observe the following sets of possible outcomes as shown in Table 1 with associated probabilities.

Table 1 Probabilities table

	C_1^B	C_2^B	$\ldots$	C_k^B	Total
C_1^A	$P(C_1^A \cap C_1^B)$	$P(C_1^A \cap C_2^B)$		$P(C_1^A \cap C_k^B)$	$P(C_1^A) = \sum_{j=1}^{k} P(C_1^A \cap C_j^B)$
C_2^A	$P(C_2^A \cap C_1^B)$	$P(C_2^A \cap C_2^B)$	$\ldots$	$P(C_2^A \cap C_k^B)$	$P(C_2^A) = \sum_{j=1}^{k} P(C_2^A \cap C_j^B)$
$\ldots$	$\ldots$				$\ldots$
C_h^A	$P(C_h^A \cap C_1^B)$	$P(C_h^A \cap C_2^B)$	$\ldots$	$P(C_h^A \cap C_k^B)$	$P(C_h^A) = \sum_{j=1}^{k} P(C_h^A \cap C_j^B)$
	$P(C_1^B) =$ $\sum_{i=1}^{h} P(C_i^A \cap C_1^B)$	$P(C_2^B) =$ $\sum_{i=1}^{h} P(C_i^A \cap C_2^B)$		$P(C_k^B) =$ $\sum_{i=1}^{h} P(C_i^A \cap C_k^B)$	

where $P(C_i^A \cap C_j^B)$ are the joint probabilities of the event (C_i^A and C_j^B) and $P(C_i^A)$ is the marginal probability of event C_i^A occurring irrespective of outcome of *B*.

5. Conditional Distributions

Consider now the probability that any one of the outcomes associated with experiment A occurs given that the outcome from experiment *B* was B_k, this is written as $P(C_i^A \mid C_k^B),\ i=1,\dots h$.

Correspondingly we only need the penultimate column of Table 2. However, these probabilities (which define all of the possible outcomes of the experiment C_i^A given that C_k^B occurred) sum to $P(C_k^B)$, rather than unity, scaling each probability in the table by $P(C_k^B)$, gives

$$P(C_i^A \mid C_k^B) = \frac{P(C_i^A \cap C_k^B)}{P(C_k^B)}.$$

Similarly we can find

$$P(C_j^B \mid C_h^A) = \frac{P(C_h^A \cap C_j^B)}{P(C_h^A)}.$$

From the above we therefore have that:

$$P(C_i^A \mid C_k^B)P(C_k^B) = P(C_i^A \cap C_k^B)$$

$$P(C_j^B \mid C_h^A)P(C_h^A) = P(C_h^A \cap C_j^B)$$

and

$$P(C_i^A \mid C_k^B) = \frac{P(C_k^B \mid C_i^A)P(C_i^A)}{P(C_k^B)}$$

Imagine there are two groups (*Group1* and *Group2*) and we are interested in the probability of some event *A*, then:

Relative Risk is calculated as: $\dfrac{P(A \mid Group1)}{P(A \mid Group2)}$ and is the ratio of the probability of some event for two different groups.

Odds Ratio is calculated as: $\dfrac{P(A \mid Group1)/P(\overline{A} \mid Group1)}{P(A \mid Group2)/P(\overline{A} \mid Group2)}$ and is the ratio of undertaking an activity (compared to not) for *Group1* compared to *Group2*. Note this would be the same if $P(\overline{A} \mid Group1) \simeq 1$ and $P(\overline{A} \mid Group2) \simeq 1$.

6. Statistical Independence

The events C_i^A and C_j^B are said to be statistically, if

$$P(C_i^A \mid C_j^B) = \frac{P(C_i^A \cap C_j^B)}{P(C_j^B)} = P(C_i^A)$$

that is, the events are independent if probability of C_i^A occurring, conditional on C_j^B having occurred is simply the marginal probability of C_i^A (the conditioning has no effect).

Independence implies:

$$P(C_i^A \cap C_j^B) = P(C_i^A).P(C_j^B)$$

7. Bayes Theorem

Define an event C_i^A and some mutually exclusive and exhaustive basic events $C_1^B, C_2^B, \dots C_k^B$. Then we know

$$P(C_i^A) = P(C_i^A \cap C_1^B) + P(C_i^A \cap C_2^B) \cdots + P(C_i^A \cap C_k^B)$$

$$P(C_i^A) = P(C_i^A \mid C_1^B)P(C_1^B) + P(C_i^A \mid C_2^B)P(C_2^B) \cdots + P(C_i^A \mid C_k^B)P(C_k^B)$$

$$P(C_i^A) = \sum_{j=1}^{k} P(C_i^A \mid C_j^B)P(C_j^B)$$

and therefore we get Bayes Theorem:

$$P(C_j^B \mid C_i^A) = \frac{P(C_i^A \mid C_j^B).P(C_j^B)}{P(C_i^A)} = \frac{P(C_i^A \mid C_j^B).P(C_j^B)}{\sum_{j=1}^{k} P(C_i^A \mid C_j^B)P(C_j^B)}$$

8. Combinations and Permutations

There are *n* objects to be arranged in order: how many different ways are there of doing this?

${}_nP_n = n(n-1)(n-2)\ldots 1 = n!$ - permutations

There are *n* different objects and you choose *r* of them, how many ways can you order these r objects?

$${}_nP_r = n! = n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!}$$

There are *n* different objects and you choose *r* of them, how many ways can you choose *r* (without ordering) ${}_nC_r = \dfrac{n!}{r!(n-r)!} = \dfrac{{}_nP_r}{r!}$

Consider 5 people A, B, C, D, E entering a room and we are interested the order in which they enter then we can see this as:

		1st person				
		A	B	C	D	E
	A	-	1	1	1	1
	B	1	-	2	2	2
	C	2	2	-	3	3
	D	3	3	3	-	4
	E	4	4	4	4	-

And so if A enters 1st there are 4 ways other people can enter 2nd. Similarly if B enters 1st there are 4 ways people can enter 2nd. So the number of ways 2 people can enter the room is $5 \times 4 = 20 = {}^5P_2 = \dfrac{5!}{(5-2)!}$. If we are looking at getting a 3rd person into the room and A followed by B are the 1st two, then how many ways can we do this and the answer is 3 (C or D or E) and the answer would therefore be

$$20 \times 3 = {}^5P_3 = \frac{5!}{(5-3)!}\ .$$

If we do not care about the order in how many ways can two people enter the room the answer is $\dfrac{{}^5P_2}{2} = \dfrac{5!}{(5-2)!2!} = {}^5C_2 = 10$. If 3 people enter the room in how many different ways can that happen: $\frac{{}^{5}P_{3}}{6} = \frac{5!}{(5-3)!3!} = {}^{5}C_{3} = 10$ as there are 6 ways of rearranging any combination of three letters (e.g. ABC, ACB, BAC, BCA, CAB, CBA).

Sample Questions

Question 1

Let $A=\left(2,4,6,8\right)$, $B=\left(1,3,5,9\right)$ and $\Omega=\left(1,2,3,4,5,6,7,8,9\right)$

Evaluate the sets:

(a) $\overline{A}$, (b) $\overline{B}$, (c) $A\cup B$, (d) $\overline{A\cup B}$, (e) $A\cap B$, (f) $\overline{A}\cup\overline{B}$, (g) $\overline{A}\cap\overline{B}$

Question 2

If $P(A)=1/3$, $P(B)=1/2$ and $P(A\cup B)=3/4$.

Find (a) $P(A\cap B)$, (b) $P(\overline{A\cap B})$, (c) $P(\overline{A\cup B})$, (d) $P(A\cap\overline{B})$ (e) $P(\overline{A}\cap B)$, (f) $P(\overline{A}\cap\overline{B})$, (g) $P(\overline{A}\cup\overline{B})$

Question 3

A fair octagonal (eight sided) die, with faces marked 1 to 8, is thrown as an experiment, the result being the number on the face of the die. Define the following events: $E_1=(1,2,3,4,5)$, $E_2=(2,4,6,8)$, $E_3=(1,3,5,7)$. Find the following:

(a) $\Pr(E_1\,|\,E_2)$, (b) $\Pr(E_1\,|\,E_3)$, (c) $\Pr(\overline{E}_1\,|\,E_2)$, (d) $\Pr(E_2\,|\,E_1)$,

(e) $\Pr(E_3\,|\,E_1\cup E_2)$.

Question 4

A town has three bus routes A, B and C. In the “rush hour”, route A has twice as many buses on its route as both B and C. Over a period of time it has been found that, along a certain stretch of road, where the three bus routes converge, the buses run more than five minutes late depending on their route with probabilities: 0.5, 0.2 and 0.1, respectively. If an inspector finds that a bus is more than five minutes late, find the probability that it is a route B bus.

Question 5

A child uses a home-made metal detector to look for valuable metallic objects on a beach. There is fault in the machine which causes it to signal the presence of only 95% of metallic objects over which it passes and to signal the presence of 6% of non-metallic objects. Of the objects over which the machine passes, 20% are metallic.

(a) Find the probability that a given object is metallic and the machine gives a signal.

(b) Find the probability of a signal being received by the child for any given object.

(c) Find the probability that the child has found a metallic object when they receive a signal.

(d) Given that 10% of metallic objects found on the beach are valuable, find the proportion of objects, discovered by a signal from the detector, that are valuable.

Question 6

A passenger compartment on a train has six seats, three facing forwards and three facing backwards. Three men and two women enter the compartment and seat themselves randomly.

(a) In how many ways can they be seated?

(b) In how many ways will the women be seated opposite each other?

(c) In how many ways can two men be seated opposite each other?

Sample Questions (with Answers)

Question 1

Let $A=\left(2,4,6,8\right)$, $B=\left(1,3,5,9\right)$ and $\Omega=\left(1,2,3,4,5,6,7,8,9\right)$

Evaluate the sets:

(a) $\overline{A}$, (b) $\overline{B}$, (c) $A\cup B$, (d) $\overline{A\cup B}$, (e) $A\cap B$, (f) $\overline{A}\cup\overline{B}$, (g) $\overline{A}\cap\overline{B}$

Answer

(a) $\overline{A}=(1,3,5,7,9)$

(b) $\overline{B}=(2,4,6,7,8)$

(c) $A\cup B=(1,2,3,4,5,6,8,9)$

(d) $\overline{A\cup B}=(7)$

(e) $A\cap B=(\varnothing)$

(f) $\overline{A}\cup\overline{B}=(1,2,3,4,5,6,7,8,9)=\overline{A\cap B}$

(g) $\overline{A}\cap\overline{B}=(7)=\overline{A\cup B}$

Question 2

If $P(A)=1/3$, $P(B)=1/2$ and $P(A\cup B)=3/4$.

Find (a) $P(A\cap B)$, (b) $P(\overline{A\cap B})$, (c) $P(\overline{A\cup B})$, (d) $P(A\cap \overline{B})$ (e) $P(\overline{A}\cap B)$, (f) $P(\overline{A}\cap \overline{B})$, (g) $P(\overline{A}\cup \overline{B})$

Answer

(a) $P(A\cup B)=3/4=1/3+1/2-P(A\cap B)\Rightarrow P(A\cap B)=1/12$

(b) $P(\overline{A\cap B})=1-1/12=11/12$

(c) $P(\overline{A\cup B})=1-3/4=1/4$

(d) $P(A)=P(A\cap B)+P(A\cap \overline{B})\Rightarrow P(A\cap \overline{B})=1/3-1/12=1/4$

(e) $P(B)=P(A\cap B)+P(\overline{A}\cap B)\Rightarrow P(\overline{A}\cap B)=1/2-1/12=5/12$

(f) $P(\overline{A})=P(\overline{A}\cap B)+P(\overline{A}\cap \overline{B})\Rightarrow P(\overline{A}\cap \overline{B})=2/3-5/12=1/4 \; =P(\overline{A\cup B})$

(g) $P(\overline{A}\cup \overline{B})=P(\overline{A})+P(\overline{B})-P(\overline{A}\cap \overline{B})=2/3+1/2-1/4=11/12 \; =P(\overline{A\cap B})$

Question 3

A fair octagonal (eight sided) die, with faces marked 1 to 8, is thrown as an experiment, the result being the number on the face of the die. Define the following events: $E_1 = (1,2,3,4,5)$, $E_2 = (2,4,6,8)$, $E_3 = (1,3,5,7)$. Find the following:

(a) $\Pr(E_1 \mid E_2)$, (b) $\Pr(E_1 \mid E_3)$, (c) $\Pr(\overline{E}_1 \mid E_2)$, (d) $\Pr(E_2 \mid E_1)$,

(e) $\Pr(E_3 \mid E_1 \cup E_2)$.

Answer

(a) $\Pr(E_1 \mid E_2) = \dfrac{\Pr(E_1 \cap E_2)}{\Pr(E_2)} = \dfrac{\Pr(2,4)}{\Pr(2,4,6,8)} = \dfrac{0.25}{0.5} = 0.5$

(b) $\Pr(E_1 \mid E_3) = \dfrac{\Pr(E_1 \cap E_3)}{\Pr(E_3)} = \dfrac{\Pr(1,3,5)}{\Pr(1,3,5,7)} = \dfrac{0.375}{0.5} = 0.75$

(c) $\Pr(\overline{E}_1 \mid E_2) = \dfrac{\Pr(\overline{E}_1 \cap E_2)}{\Pr(E_2)} = \dfrac{\Pr(6,8)}{\Pr(2,4,6,8)} = \dfrac{0.25}{0.5} = 0.5$

(d) $\Pr(E_2 \mid E_1) = \dfrac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} = \dfrac{\Pr(2,4)}{\Pr(1,2,3,4,5)} = \dfrac{0.25}{0.625} = 0.4$

(e) $\Pr(E_3 \mid E_1 \cup E_2) = \dfrac{\Pr(E_3 \cap (E_1 \cup E_2))}{\Pr(E_1 \cup E_2)} = \dfrac{\Pr(1,3,5)}{\Pr(1,2,3,4,5,6,8)} = \dfrac{0.375}{0.875} = \dfrac{3}{7}$

Question 4

A town has three buses A, B and C. In the "rush hour", A has twice as many buses on its route as both B and C. Over a period of time it has been found that, along a certain stretch of road, where the three buses converge, the probability of a bus being at least 5 minutes late is 0.5, 0.2 and 0.1 for each given bus, respectively. If an inspector (standing near this stretch of road) finds that the first bus is more than five minutes late, find the probability that it is route B bus.

Answer

$\Pr(A) = 0.5$, $\Pr(B) = \Pr(C) = 0.25$

In addition, we know

$\Pr(L \mid A) = 0.5$, $\Pr(L \mid B) = 0.2$, $\Pr(L \mid C) = 0.1$

we want to know:

$$\Pr(B \mid L) = \frac{\Pr(L \mid B).\Pr(B)}{\Pr(L)} = \frac{0.2(0.25)}{0.325} = \frac{0.05}{0.325} = 0.154$$

and from the union-intersection rule:

$$\Pr(L) = \Pr(L \mid A).\Pr(A) + \Pr(L \mid B).\Pr(B) + \Pr(L \mid C).\Pr(C)$$

$$\Pr(L) = 0.5(0.5) + 0.2(0.25) + 0.1(0.25) = 0.325$$

Question 5

A child uses a home-made metal detector to look for valuable metallic objects on a beach. There is fault in the machine which causes it to signal the presence of only 95% of metallic objects over which it passes and to signal the presence of 6% of non-metallic objects. Of the objects over which the machine passes, 20% are metallic.

(a) Find the probability that a given object is metallic and the machine gives a signal.

(b) Find the probability of a signal being received by the child for any given object.

(c) Find the probability that the child has found a metallic object when they receive a signal.

(d) Given that 10% of metallic objects found on the beach are valuable (and non-metal objects are not valuable), find the proportion of objects, discovered by a signal from the detector, that are valuable.

Answer

$\Pr(S \mid M) = 0.95$, $\Pr(S \mid NM) = 0.06$, $\Pr(M) = 0.2$

(a) $\Pr(M \cap S) = \Pr(S \mid M).\Pr(M) = 0.95(0.2) = 0.19$.

(b) $\Pr(S) = \Pr(S \cap M) + \Pr(S \cap NM) = 0.19 + 0.06(0.8) = 0.238$

(c) $\Pr(M \mid S) = \dfrac{\Pr(M \cap S)}{\Pr(S)} = \dfrac{0.19}{0.238} = 0.798$

(d) $P(V \mid M \cap S) = 0.1 \Rightarrow P(V \cap M \cap S) = 0.1(0.19) = P(V \cap S)$

as $P(V \cap S \cap NM) = 0$, therefore $\Pr(V \mid S) = \dfrac{\Pr(V \cap S)}{P(S)} = \dfrac{0.1(0.19)}{0.238} = 0.08$

Question 6

A passenger compartment on a train has six seats, three facing forwards and three facing backwards. Three men and two women enter the compartment and seat themselves randomly.

(a) In how many ways can they be seated?

(b) In how many ways will the women be seated opposite each other?

(c) In how many ways can two men be seated opposite each other?

Answer

(a) In how many ways can we arrange 5 people in 6 seats - ${}_6P_6 = 720$.

(b) If the two women sit opposite one another (next to the window) then how many ways can the three men occupy the remaining 4 seats - ${}_4P_4 = 24$. In total the women can sit opposite each other in 3 ways (window, aisle or middle seats and can swap places) – therefore we have 6*24=144.

(c) Clearly if there we only 2 men then the answer would be the same as (b), but there are three men and these can sit opposite each other as man 1 and man 2, man1 and man 3, or man 2 and man 3 (and can swap around). Therefore we have 432.

STATISTICAL TECHNIQUES B

Univariate and Bivariate Distributions

1. Introduction to Univariate Distributions

For a random experiment, with a sample space, Ω, a function *X*, which assigns to each element of *C*, a real number *X*(*C*)=*x*, is called a random variable. We must distinguish between the random variable, *X*, and the possible outcomes, $x \in \Omega$.

Example 1:

Consider rolling a dice then $\Omega = \{1,2,3,4,5,6\}$, define a random variable, which considers only odd or even numbers and $X(C) = \begin{cases} 1 \text{ if even number} \\ 0 \text{ if odd number} \end{cases}$, then

x	0	1
P(*X*=*x*)	½	½

Example 2:

Toss two coins $\Omega = \{HH, HT, TH, TT\}$. Let X(C)=Number of tails

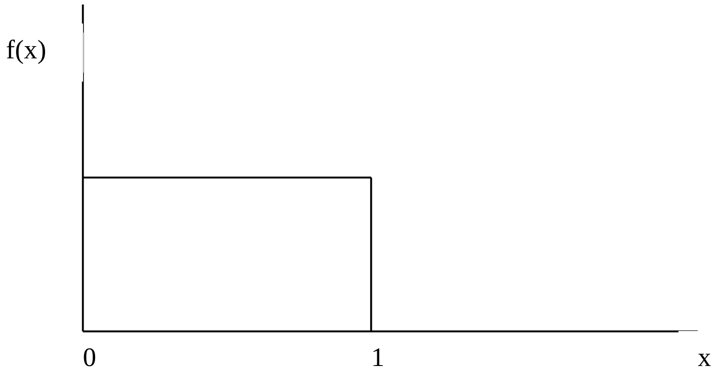
$X(HH) = 0, X(HT) = 1, X(TH) = 1, X(TT) = 2$ then $\begin{cases} \Omega = \{HH, HT, TH, TT\} \\ \text{A} = \{0,1,2\} \end{cases}$ and

P(*X*=1)=P(HT,TH)=1/2

x	0	1	2
P(*X*=*x*)	¼	½	¼

Example 3:

$f(x) = 1 \quad 0 \le x \le 1 \quad \Omega = \{x; 0 \le x \le 1\}$

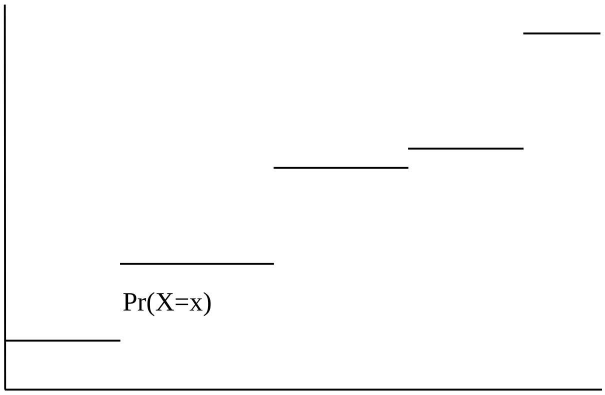


2. Discrete Univariate Distributions

Suppose X is a scalar random variable with a finite number of values - this is a discrete set of points. Let $p_X(x)$ be a function such that, (i) $p_X(x) \geq 0$, and (ii) $\sum_x p_X(x) = 1$, then X is a discrete random variable with probability density (mass) function, $p_X(x)$ and $P(a \leq X \leq b) = \sum_{x=a}^{b} f(x)$. (for rules on summations see Appendix 2)

The cumulative distribution (mass) function of X is such that for each $F_X(x_0) = P(X \leq x_0) = \sum_{x \leq x_0} p_X(x)$.

(a) cdf of a discrete random variable



(iii) $\lim_{x \to -\infty} F(x) = 0$

(iv) $\lim_{x \to \infty} F(x) = 1$

(v) $\Pr\left(a < X \leq b\right) = F(b) - F(a)$

2.1 Measures of central tendency and dispersion (spread)

2.1.1 Median

For a discrete random variable *X*, if $p\left(X \leq x\right) \geq 1/2$, and $p\left(X \geq x\right) \geq 1/2$ then the median if the variable *X* is *x*. Note there might be a case where the median is not defined.

2.1.2 Mode

That value of *x* such that $p(x)$ is maximised.

2.1.3 Expectation (simple mean)

X takes on a finite number of outcomes $x = x_1, x_2, \ldots, x_n$ and each has an associated probability:

X	x_1	x_2	…	x_n
$P(X=x)$	p_1	p_2	…	p_n

$E(X) = \sum_x p_X(x)x = \mu_X$ (for rules on expectations see Appendix 3)

2.1.4 Variance

One is also interested in the spread or dispersion in the data and a common statistic to measure this is the variance, defined as:

$V(X) = E\left[(X-\mu_X)^2\right] = \sum_x p_X(x)(x-\mu_X)^2$ (for rules on variances see Appendix 3)

$$= \sum_x p_X(x)x^2 - 2\mu_X \underbrace{\sum p_X(x)x}_{\mu_X} + \mu_X^2 \underbrace{\sum_x p_X(x)}_{1} = \sum_x p_X(x)x^2 - \mu_X^2 = E(X^2) - \mu_X^2$$

$$V(X) = E\left[(X-\mu_X)^2\right] = E(X^2) - E(X)^2$$

The variance looks at how far each point is the mean of the variable (deviations from the mean) and then squares that measure (to ensure all values are positive) and then looks at the expected value of this transformed series. By construction series which are more dispersed around the mean have a higher variance and at the limit if a variable only takes 1 value it has a zero variance.

2.1.5 Higher moments

In general,

$$E\left[g(X)\right] = \sum_x p_X(x)g(x)$$

Example

X=*x*	1	2	3	4
Pr(X=*x*)	0.2	0.3	0.4	0.1

$$E(X) = 0.2(1) + 0.3(2) + 0.4(3) + 0.1(4) = 2.4$$

$$E(X^2) = 0.2(1^2) + 0.3(2^2) + 0.4(3^2) + 0.1(4^2) = 6.6$$

$$V(X) = E(X^2) - \left[E(X)\right]^2 = 6.6 - 2.4^2 = 0.84$$

3. Introduction to Discrete Bivariate Distributions

Suppose that the sample space of X_1 is $\Omega_1 = \{x_1^1, x_2^1, \dots x_h^1\}$ and the sample space of X_2 is $\Omega_2 = \{x_1^2, x_2^2, \dots x_h^2\}$. Then we can define all possible outcomes and the joint probability density function using the Table 1 below (similar to Table 1 from Handout 1 - Probability):

Table 1: Joint Probability Table

	x_1^2	x_2^2	$\cdots$	x_k^2	Total
x_1^1	$p(x_1^1 \cap x_1^2)$	$p(x_1^1 \cap x_2^2)$		$p(x_1^1 \cap x_k^2)$	$p(x_1^1) = \sum_{j=1}^{k} p(x_1^1 \cap x_j^2)$
x_2^1	$P(x_2^1 \cap x_1^2)$	$P(x_2^1 \cap x_2^2)$	$\cdots$	$P(x_2^1 \cap x_k^2)$	$p(x_2^1) = \sum_{j=1}^{k} p(x_2^1 \cap x_j^2)$
$\cdots$	$\cdots$				$\cdots$
x_h^1	$P(x_h^1 \cap x_1^2)$	$P(x_h^1 \cap x_2^2)$	$\cdots$	$P(x_h^1 \cap x_k^2)$	$p(x_h^1) = \sum_{j=1}^{k} p(x_h^1 \cap x_j^2)$
	$p(x_1^2) =$ $\sum_{i=1}^{h} p(x_i^1 \cap x_1^2)$	$p(x_2^2) =$ $\sum_{i=1}^{h} p(x_i^1 \cap x_2^2)$		$p(x_k^2) =$ $\sum_{i=1}^{h} p(x_i^1 \cap x_k^2)$	

The table defining a probability for each pair of events, x_i^1, x_j^2 for i=1,…h and j=1, …k, such that $p(x_i^1 \cap x_j^2) \geq 0$ and $\sum_i \sum_j p(x_i^1 \cap x_j^2) = 1$ then it is a valid probability density function.

3.1 Marginal Distributions

For the bivariate case above, consider the event, $X_1 = x_i^1$. This event occurs when $X_1 = x_i^1$ and X_2 takes any possible value. This probability is

$$P(X_1 = x_i^1, x_1^2 \leq X_2 \leq x_k^2) = \sum_j p(x_i^1, x_j^2) = p_1(x_i^1)$$

and this is the MARGINAL PROBABILITY DENSITY (MASS) FUNCTION of X_1. From these marginal distributions we can calculate the moments of the random variables X_1 and X_2, that is, $E(X_1)$, $E(X_2)$, $V(X_1)$, $V(X_2)$ and $\text{cov}(X_1, X_2)$ as we did before. For the discrete random variable X_1, with marginal probability density function $p_1(x_i^1)$:

$$E(X_1)=\sum_i x_i^1 p_1(x_i^1)$$

$$V(X_1)=\sum_i (x_i^1)^2 p_1(x_i^1)-E(X_1)^2$$

similarly for X_2,

$$E(X_2)=\sum_j x_j^2 p_2(x_j^2)$$

$$V(X_2)=\sum_j (x_i^2)^2 p_2(x_j^2)-E(X_2)^2 .$$

Between two random variables one can also have measures of association and one common measure of association is the covariance, defined as:

$$\text{cov}(X_1,X_2)=E(X_1-E(X_1))(X_2-E(X_2))=E(X_1X_2)-E(X_1)E(X_2)$$

$$\text{cov}(X_1,X_2)=\sum_i\sum_j x_i^1x_j^2 p(x_i^1,x_j^2)-E(X_1)E(X_2)$$

The covariance measures both series in terms of deviations from the mean and then measures what is the expected value of the cross-product of the two terms. This measure has two elements contained within it:

(i) A measure of sign are positive (negative) deviations of X_1 associated with positive (negative) deviations of X_2.

(ii) Size how big are these deviations and the bigger they are the bigger (in an absolute sense) the covariance.

Appendix 4 contains example scatter plots of values of the variables of X_1 and X_2 and the associated covariance sign.

Rules on expectations and variance for combinations of random variables are in Appendix 5.

3.2 Conditional Distributions

The conditional probability density function for random variables, X_1 and X_2 with a joint probability density (mass) function $p(x_1,x_2)$ and marginals $p_1(x_i^1)$ and $p_2(x_i^2)$, is written as:

$$p(x_i^1\mid x_j^2)=\frac{p(x_i^1,x_j^2)}{p_2(x_j^2)} .$$

This is a valid pdf as

$$\sum_i p(x_i^1\mid x_j^2)=\sum_i\frac{p(x_i^1,x_j^2)}{p_2(x_j^2)}=\frac{1}{p_2(x_j^2)}\sum_i p(x_i^1,x_j^2)=\frac{p_2(x_j^2)}{p_2(x_j^2)}=1$$

and $p(x_i^1 \mid x_j^2) = \dfrac{p(x_i^1, x_j^2)}{p_2(x_j^2)} \geq 0$

Rearranging the above expression we also have that

$p(x_i^1 \mid x_j^2)\, p_2(x_j^2) = p(x_i^1, x_j^2)$.

This idea can be extended to more than two events, in which case we have

$$p(x_i^1 \mid x_j^2, x_m^3) = \frac{p(x_i^1, x_j^2, x_m^3)}{p_{2,3}(x_j^2, x_m^3)}$$

Rearranging and using the rule that $p(x_i^1, x_j^2, x_m^3) = p(x_i^1 \mid x_j^2, x_m^3)\, p_{2,3}(x_j^2, x_m^3)$, we have

$$p(x_i^1, x_j^2, x_m^3) = p(x_i^1 \mid x_j^2, x_m^3)\, p_{2,3}(x_j^2, x_m^3) = p(x_i^1 \mid x_j^2, x_m^3).p(x_j^2 \mid x_m^3).p_3(x_m^3)$$

NOTE:

$E(X) = E(E(X \mid Y))$, i..e.:

$$E(X) = E(X \mid Y = 1).P(Y = 1) + E(X \mid Y = 2).P(Y = 2) = 1.8(0.5) + 3.0(0.5) = 2.4$$

$$V(X) = E(V(X \mid Y)) + V(E(X \mid Y))$$

4. Continuous Univariate Distributions

Suppose *X* is a scalar random variable along the real line and (i) $f_X(x) \geq 0$ and (ii) $\int_x f_X(x)dx = 1$, then *X* is a continuous random variable with probability density function, $f_X(x)$ and $P(a \leq X \leq b) = \int_a^b f_X(x)dx$. Moreover $P(X = a) = \int_a^a f_X(x)dx = 0$, therefore $P(a < X < b) = P(a \leq X \leq b)$ (see Appendix 7 for the basics of integration).

The cumulative distribution function of *X* is such that for each

$$F_X(x_0) = P(X \leq x_0) = \int_{x \leq x_0} f_X(x)dx,$$

The cdf of a random variable *X* is:

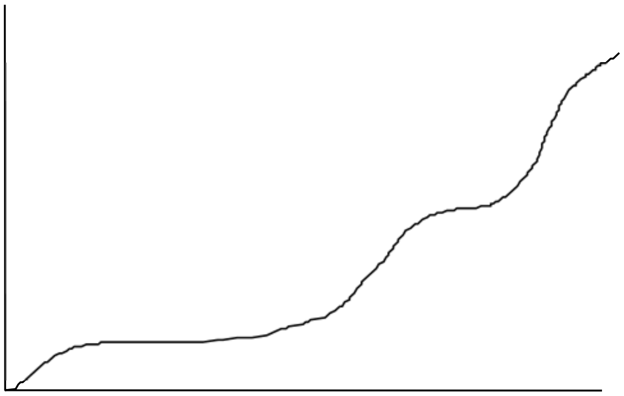
$$F(x) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(y)dy \quad \text{or} \quad \sum_{-\infty}^{x_0} f(y)$$

From cdf can determine all the relevant probability statements.

(i) F is non-decreasing, that is, if $y \leq x$ then $F(y) \leq F(x)$

(ii) cdfs are everywhere continuous from the right, that is, $\lim_{h \to 0} F(x+h) = F(x)$

(b) cdf of a continuous random variable



4.1 Measures of central tendency and dispersion

4.1.1 Median

For a continuous random variable *X*, with cdf *F*, the median is the point *x*, such that *F*(*x*)=1/2.

4.1.2 Mode

That value of *x* such that $f(x)$ is maximised.

4.1.3 Expectations and variances

For continuous random variables

$$E(X) = \int_x x f_X(x) dx = \mu_X$$

$$V(X) = E(X^2) - E(X)^2 = \int_x x^2 f_X(x) dx - \mu_x^2$$

5. Introduction to Continuous Bivariate Distributions

Let $f_{X_1,X_2}(x_1,x_2)$ be the joint probability density function for the continuous random variables X_1 and X_2. This function is a valid probability density function, if,

(i) $f_{X_1,X_2}(x_1,x_2) \geq 0$

(ii) $\int_{x_1}\int_{x_2} f_{X_1,X_2}(x_1,x_2)\partial x_2 \partial x_1 = 1$

5.1 Marginal Distributions

For the continuous random variables X_1 and X_2, with $f_{X_1,X_2}(x_1,x_2)$ as the joint probability density function, the marginal probability of the marginal probability of X_1 is:

$$P(X_1, -\infty < X_2 < \infty) = \int_{-\infty}^{\infty} f_{X_1X_2}(x_1,x_2)\,dx_2 = f_1(x_1)$$

and this removes the variable X_2 out of the formula, leaving the marginal probability density function of X_1, $f_1(x_1)$, a function of x_1 alone. Similarly to above the marginal probability of X_2 is:

$$P(X_2, -\infty < X_2 < \infty) = \int_{-\infty}^{\infty} f_{X_1X_2}(x_1,x_2)\,dx_1 = f_2(x_2).$$

$$E(X_1) = \int_{x_1} x_1 f_{X_1}(x_1)dx_1 = \mu_{X_1}$$

$$V(X_1) = E(X_1^2) - E(X_1)^2 = \int_{x_1} x_1^2 f_{X_1}(x_1)dx_1 - \mu_{x_1}^2$$

similarly for X_2,

and

$$\operatorname{cov}(X_1,X_2) = E(X_1 - E(X_1))(X_2 - E(X_2)) = E(X_1X_2) - E(X_1)E(X_2)$$

$$\operatorname{cov}(X_1,X_2) = \int_{x_1}\int_{x_2} x_1 x_x f(x_1,x_2)\partial x_2 \partial x_1 - E(X_1)E(X_2)$$

Sample Questions

Question 1

For the following discrete distribution:

x	1	3	5	8	9
Pr(X=x)	0.1	0.4	0.3	0.15	0.05

Find (a) $E\left(X\right)$, (b) $E\left(X-1\right)$, (c) $E\left[3\left(X-1\right)\right]$, (d) $V\left(X\right)$, (e) $V\left(X-1\right)$,

(f) $V\left[3\left(X-1\right)\right]$.

Question 2

For the discrete random variable defined by the pdf.

x	-3	2	4
Pr(X=x)	0.4	0.3	0.3

Find $E(Y)$ if $Y=2(X-1)^2+3(X-1)-5$

Question 3

Consider the following bivariate distribution for X_1 and X_2.

		x_2			
		1	2	3	4
	1	0.10	0.05	0.00	0.10
x_1	2	0.10	0.00	0.20	0.00
	3	0.05	0.10	0.25	0.05

(a) Write out the marginal distributions for X_1 and X_2.

(b) Calculate $E\left(X_1\right)$ and $E\left(X_2\right)$, $V\left(X_1\right)$ and $V\left(X_2\right)$.

(c) Calculate $cov\left(X_1,X_2\right)$.

(d) Write out the distribution of $\left(X_1 \mid X_2=2\right)$ and calculate $E\left(X_1 \mid X_2=2\right)$.

Question 4

The random variables X_1 is distributed with a mean of 50 and variance of 10, while is independently X_2 is distributed with mean of 50 and variance 5. Find the mean and variance of (a) $X_1 + X_2$, (b) $X_1 - 2X_2$, (c) $X_2 - 0.4X_1$.

Question 5

The random variables X_1, X_2 and X_3 are a random sample from a population with a mean of μ and variance σ^2. Find the mean and variance of

(a) $X_1 + X_2 + X_3$,

(b) $X_1 - X_2 + X_3$,

(c) $\left(X_1 + X_2 + X_3\right)/3$.

Question 6

The continuous random variable X has pdf $f(x)$, where

$$f(x) = \begin{cases} 0 & x < 2 \\ k(3-x) & 2 \le x \le 3 \\ 0 & x > 3 \end{cases}$$

Calculate (a) the constant, k, (b) the median of X.

Question 7

The continuous random variables X and Y have a joint pdf, $f(x,y)$, given by:

$$f(x,y) = \frac{x(1+3y^2)}{4} \quad 0 \le x \le 2, 0 \le y \le 1$$

Find

(a) the marginal distribution of X,

(b) the marginal distribution of Y,

(c) the conditional distribution of Y given $X=x$,

(d) the conditional distribution of X given $Y=y$.

Sample Questions (with Answers)

Question 1

For the following discrete distribution:

x	1	3	5	8	9
Pr(X=x)	0.1	0.4	0.3	0.15	0.05

Find (a) $E\left(X\right)$, (b) $E\left(X-1\right)$, (c) $E\left[3\left(X-1\right)\right]$, (d) $V\left(X\right)$, (e) $V\left(X-1\right)$,

(f) $V\left[3\left(X-1\right)\right]$.

Answer

(a) $E(X)=1(0.1)+3(0.4)+5(0.3)+8(0.15)+9(0.05)=4.45$

(b)

x-1	0	2	4	7	8
Pr[(X-*1*=x-1]	0.1	0.4	0.3	0.15	0.05

$E(X-1)=0(0.1)+2(0.4)+4(0.3)+7(0.15)+8(0.05)=3.45=E(X)-1$

(c)

3(x-1)	0	6	12	21	24
Pr[3(X-*1)*=3(x-1)]	0.1	0.4	0.3	0.15	0.05

$E[3(X-1)]=0(0.1)+6(0.4)+12(0.3)+21(0.15)+24(0.05)=10.35=3E(X)-3$

(d) $$\begin{aligned}V(X)&=E(X^2)-E(X)^2\\&=1^2(0.1)+3^2(0.4)+5^2(0.3)+8^2(0.15)+9^2(0.05)-4.45^2=5.05\end{aligned}$$

(e) $V(X-1)=0^2(0.1)+2^2(0.4)+4^2(0.3)+7^2(0.15)+8^2(0.05)-3.45^2=5.05$

$V(X-1)=E[(X-1)^2]-E[(X-1)]^2=E[X^2-2X+1]-[E(X)-1]^2$

$V(X-1)=E(X^2)-2E(X)+1-E(X)^2+2E(X)-1=E(X^2)-E(X)^2=V(X)$

(f) $V[3(X-1)]=0^2(0.1)+6^2(0.4)+12^2(0.3)+21^2(0.15)+24^2(0.05)-10.35^2=45.45$

$V[3(X-1)]=E[\{3(X-1)\}^2]-E[3(X-1)]^2=E[9X^2-18X+9]-[3E(X)-3]^2$

$V[3(X-1)]=9E(X^2)-18E(X)+9-9E(X)^2+18E(X)-9=9[E(X^2)-E(X)^2]=9V(X)$

Question 2

For the discrete random variable defined by the pdf.

x	-3	2	4
Pr(X=x)	0.4	0.3	0.3

Find $E(Y)$ if $Y = 2(X-1)^2 + 3(X-1) - 5$

Answer

$E(X) = -3(0.4) + 2(0.3) + 4(0.3) = 0.6$

y	15	0	22
Pr(Y=y)	0.4	0.3	0.3

$E(Y) = 15(0.4) + 0(0.3) + 22(0.3) = 12.6$

$E(Y) = 2E(X^2) - 4E(X) + 2 + 3E(X) - 3 - 5 = 2E(X^2) - E(X) - 6$

$E(Y) = 2E(X^2) - E(X) - 6 = 19.2 - 0.6 - 6 = 12.6$

Question 3

Consider the following bivariate distribution for X_1 and X_2.

		x_2			
		1	2	3	4
	1	0.10	0.05	0.00	0.10
x_1	2	0.10	0.00	0.20	0.00
	3	0.05	0.10	0.25	0.05

(a) Write out the marginal distributions for X_1 and X_2.

(b) Calculate $E\left(X_1\right)$ and $E\left(X_2\right)$, $V\left(X_1\right)$ and $V\left(X_2\right)$.

(c) Calculate $cov\left(X_1, X_2\right)$.

(d) Write out the distribution of $\left(X_1 \mid X_2 = 2\right)$ and calculate $E\left(X_1 \mid X_2 = 2\right)$.

Answer

(a)

x_1	1	2	3
$P(X_1 = x_1)$	0.25	0.30	0.45

x_2	1	2	3	4
$P(X_2 = x_2)$	0.25	0.15	0.45	0.15

(b) $E(X_1) = 1(0.25) + 2(0.3) + 3(0.45) = 2.2$

$E(X_2) = 1(0.25) + 2(0.15) + 3(0.45) + 4(0.15) = 2.5$

$V(X_1) = E(X_1^2) - E(X_1)^2 = 1^2(0.25) + 2^2(0.3) + 3^2(0.45) - (2.2)^2 = 0.66$

$V(X_2) = E(X_2^2) - E(X_2)^2 = 1^2(0.25) + 2^2(0.15) + 3^2(0.45) + 4^2(0.15) - (2.5)^2 = 1.05$

(c) $\text{cov}(X_1, X_2) = E(X_1X_2) - E(X_2)E(X_2)$

$$\begin{aligned}\text{cov}(x_1, x_2) &= 1(1)(0.1) + 1(2)(0.05) + 1(3)(0.0) + 1(4)(0.1)\\ &+2(1)(0.1) + 2(2)(0.0) + 2(3)(0.2) + 2(4)(0.0) + 3(1)(0.05) + 3(2)(0.1)\\ &+3(3)(0.25) + 3(4)(0.05) - (2.2)(2.5) = 5.6 - 5.5 = 0.1\end{aligned}$$

(d)

$x_1 \mid X_2 = 2$	1	2	3
$P(X_1 = x_1 \mid X_2 = 2)$	0.05/0.15	0.00	0.1/0.15

$x_1 \mid X_2 = 2$	1	2	3
$P(X_1 = x_1 \mid X_2 = 2)$	0.3333	0.00	0.6666

(e) $E(X_1 \mid X_2 = 2) = 1(0.3333) + 2(0.0) + 3(0.66666) = 2.3333$

Question 4

The random variables X_1 is distributed with a mean of 50 and variance of 10, while is independently X_2 is distributed with mean of 50 and variance 5. Find the mean and variance of (a) $X_1 + X_2$, (b) $X_1 - 2X_2$, (c) $X_2 - 0.4X_1$.

Answer

(a) X_1 and X_2 are assumed to be independent.

$E(X_1 + X_2) = E(X_1) + E(X_2) = 50 + 50 = 100$

$V(X_1 + X_2) = V(X_1) + V(X_2) + 2\operatorname{cov}(X_1, X_2) = 10 + 5 = 15$

(b) $E(X_1 - 2X_2) = E(X_1) - 2E(X_2) = 50 - 2(50) = -50$

$V(X_1 - 2X_2) = V(X_1) + V(-2X_2) + 2\operatorname{cov}(X_1, -2X_2) = V(X_1) + 4V(X_2) - 4\operatorname{cov}(X_1, X_2)$

$V(X_1 - 2X_2) = 10 + 4(5) = 30$

(c) $E(X_2 - 0.4X_1) = E(X_2) - 0.4E(X_2) = 50 - 0.4(50) = 30$

$V(X_2 - 0.4X_1) = V(X_2) + V(-0.4X_1) + 2\operatorname{cov}(X_2, -0.4X_1) = V(X_2) + 0.16V(X_1) - 0.8\operatorname{cov}(X_1, X_2)$

$V(X_2 - 0.4X_1) = 5 + 0.16(10) = 6.6$

Question 5

The random variables X_1, X_2 and X_3 are a random sample from a population with a mean of μ and variance σ^2. Find the mean and variance of

(a) $X_1 + X_2 + X_3$,

(b) $X_1 - X_2 + X_3$,

(c) $\left(X_1 + X_2 + X_3\right)/3$.

Answer

(a) Assuming independence

$E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = \mu + \mu + \mu = 3\mu$

$V(X_1 + X_2 + X_3) = V(X_1) + V(X_2) + V(X_3) + 2\operatorname{cov}(X_1, X_2) + 2\operatorname{cov}(X_1, X_3) + 2\operatorname{cov}(X_2, X_3)$

$V(X_1 + X_2 + X_3) = \sigma^2 + \sigma^2 + \sigma^2 = 3\sigma^2$

(b) $E(X_1 - X_2 + X_3) = E(X_1) - E(X_2) + E(X_3) = \mu - \mu + \mu = \mu$

$V(X_1 - X_2 + X_3) = V(X_1) + V(X_2) + V(X_3) - 2\operatorname{cov}(X_1, X_2) + 2\operatorname{cov}(X_1, X_3) - 2\operatorname{cov}(X_2, X_3)$

$V(X_1 + X_2 + X_3) = \sigma^2 + \sigma^2 + \sigma^2 = 3\sigma^2$

(c) $E[(X_1 + X_2 + X_3)/3] = 1/3[E(X_1) + E(X_2) + E(X_3)] = 1/3[\mu + \mu + \mu] = \mu$

$V[(X_1 + X_2 + X_3)/3] = 1/9[V(X_1) + V(X_2) + V(X_3)]$

$V[(X_1 + X_2 + X_3)/3] = 1/9[\sigma^2 + \sigma^2 + \sigma^2] = \sigma^2/3$

Question 6

The continuous random variable *X* has pdf p*f*(*x*), where

$$f(x)=\begin{cases}0 & x<2\\ k(3-x) & 2\le x\le 3\\ 0 & x>3\end{cases}$$

Calculate (a) the constant, *k*, (b) the median of *X*.

Answer

(a) $\int_2^3 k(3-x)\,dx = 1 \Rightarrow k[3x-x^2/2]_2^3 = 1 \Rightarrow k\{4.5-4\} = 1 \Rightarrow k=2$

(b) $\int_2^d 2(3-x)\,dx = 0.5 \Rightarrow 2[3x-x^2/2]_2^d = 0.5 \Rightarrow 2\{3d-d^2/2-4\} = 0.5$

$d^2-6d+8.5=0 \Rightarrow d=2.29$

Question 7

The continuous random variables *X* and *Y* have a joint pdf, *f*(*x*,*y*), given by:

$$f(x,y) = \frac{x(1+3y^2)}{4} \quad 0 \le x \le 2, 0 \le y \le 1$$

Find

(a) the marginal distribution of *X*,

(b) the marginal distribution of *Y*,

(c) the conditional distribution of *Y* given *X*=*x*,

(d) the conditional distribution of *X* given *Y*=*y*.

Answer

(a) $f(x) = \int_0^1 \frac{x(1+3y^2)}{4} dy = \frac{x}{4}[y + 3y^3/3]_0^1 = \frac{x}{2} \quad 0 \le x \le 2$

(b) $f(y) = \int_0^2 \frac{x(1+3y^2)}{4} dx = \frac{(1+3y^2)}{4}[x^2/2]_0^2 = \frac{(1+3y^2)}{2} \quad 0 \le y \le 1$

(c) $f(y,|X = x) = \frac{\frac{x(1+3y^2)}{4}}{\frac{x}{2}} = \frac{(1+3y^2)}{2} = f(y) \quad 0 \le y \le 1$

(d) $f(y,|X = x) = \frac{\frac{x(1+3y^2)}{4}}{\frac{(1+3y^2)}{2}} = \frac{x}{2} = f(x) \quad 0 \le x \le 2$

Note: *X* and *Y* are independent

Rules of summation

1. $$\sum_{i=1}^{n} X_i = (X_1 + X_2 + X_3 + \ldots + X_n)$$

2. $$\sum_{i=1}^{n} c = (c + c + c + \ldots + c) = nc$$

3. $$\sum_{i=1}^{n} cX_i = (cX_1 + cX_2 + cX_3 + \ldots + cX_n) = c(X_1 + X_2 + X_3 + \ldots + X_n) = c\sum_{i=1}^{n} X_i$$

4. $$\begin{aligned}&\sum_{i=1}^{n} (X_i + Y_i) = (X_1 + Y_1) + (X_2 + Y_2) + (X_3 + Y_3) \ldots + (X_n + Y_n)\\ &= (X_1 + X_2 + X_3 \ldots + X_n) + (Y_1 + Y_2 + Y_3 + \ldots + Y_n) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i\end{aligned}$$

5. $$\begin{aligned}&\sum_{i=1}^{n} (X_i + Y_i)^2 = (X_1 + Y_1)^2 + (X_2 + Y_2)^2 + (X_3 + Y_3)^2 \ldots + (X_n + Y_n)^2\\ &= (x_1^2 + Y_1^2 + 2X_1Y_1) + (X_2^2 + Y_2^2 + 2X_2Y_2) + \ldots (X_n^2 + Y_n^2 + 2X_nY_n)\\ &= \sum_{i=1}^{n} (X_i^2 + Y_i^2 + 2X_iY_i)\end{aligned}$$

6. $$\begin{aligned}&\sum_{i=1}^{n} (cX_i + cY_i)^2 = \sum_{i=1}^{n} (c^2X_i^2 + c^2Y_i^2 + 2c^2X_iY_i) = c^2\sum_{i=1}^{n} (X_i^2 + Y_i^2 + 2X_iY_i)\\ &= c^2\sum_{i=1}^{n} (X_i + Y_i)^2\end{aligned}$$

Rules on expectations variances

Define

$E(X)=\sum_{i=1}^{k} p_i x_i$ as the expected value of the random variable *X* and

$$V(X)=E[X-E(X)]^2=E(X^2)-E(X)^2=\sum_{i=1}^{k} p_i (x_i-E(X))^2 .$$

1. $E(a+X)=\sum_{i=1}^{k} p_i (a+x_i)=a+\sum_{i=1}^{k} p_i x_i = a+E(X)$

2. $E(aX)=\sum_{i=1}^{k} p_i (ax_i)=a\sum_{i=1}^{k} p_i x_i = aE(X)$

3. $$\begin{aligned} V(a+X)&=\sum_{i=1}^{k} p_i \left[(a+x_i)-E(a+X)\right]^2=\sum_{i=1}^{k} p_i \left[(a+x_i)-a-E(X)\right]^2 \\ &=\sum_{i=1}^{k} p_i \left[x_i-E(X)\right]^2=V(X)\end{aligned}$$

4. $$\begin{aligned} V(aX)&=\sum_{i=1}^{k} p_i \left[ax_i-E(aX)\right]^2=\sum_{i=1}^{k} p_i \left[ax_i-aE(X)\right]^2 \\ &=a^2\sum_{i=1}^{k} p_i \left[x_i-E(X)\right]^2=a^2V(X)\end{aligned}$$

Suppose $E(X)=\mu$ and $V(X)=\sigma^2$ and define $Z=\frac{X-\mu}{\sigma}$, then

$$E(Z)=E\left[\frac{X-\mu}{\sigma}\right]=\frac{1}{\sigma}E\left[X-\mu\right]=\frac{1}{\sigma}\left[\underbrace{E(X)}_{\mu}-\mu\right]=0$$

(using rules (2) and (1))

$$V(Z)=V\left[\frac{X-\mu}{\sigma}\right]=\frac{1}{\sigma^2}V(X-\mu)=\frac{1}{\sigma^2}\underbrace{V(X)}_{\sigma^2}=1$$

(using rules (4) and (3)) therefore *E*(*Z*)=0 and *V*(*Z*)=1, this is a standardised variable. While these rules have been derived for a discrete distribution on the random variable, *X*, similar arguments would hold if *X* was a continuous random variable (although the summation sign would be replaced by an integral).

Figure 1a: cov(x,y)>0

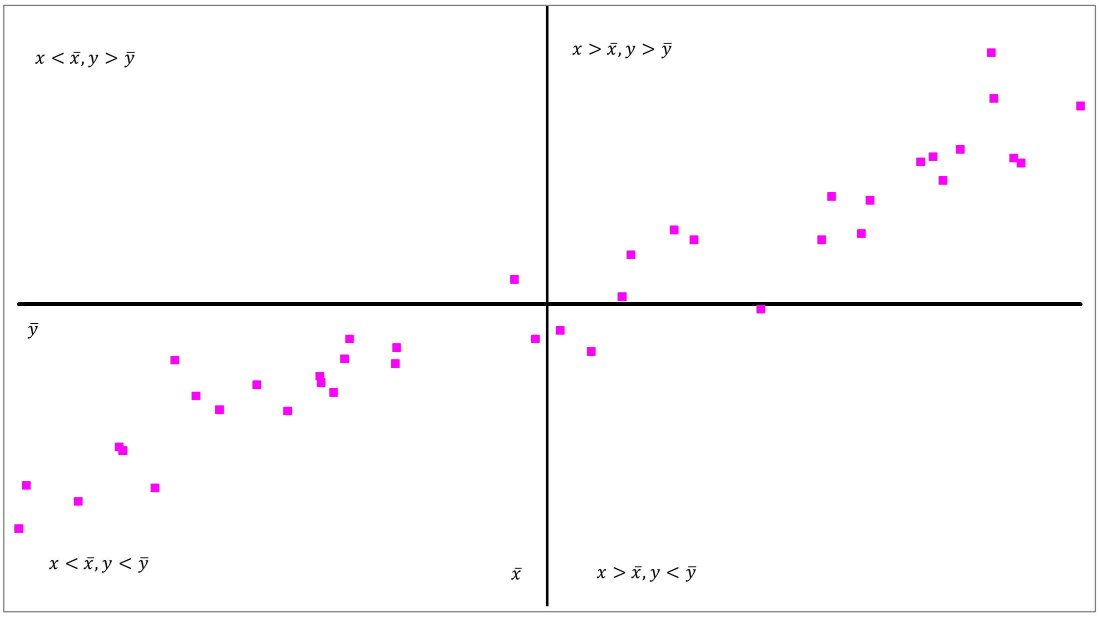


Figure 1b: cov(x,y)<0

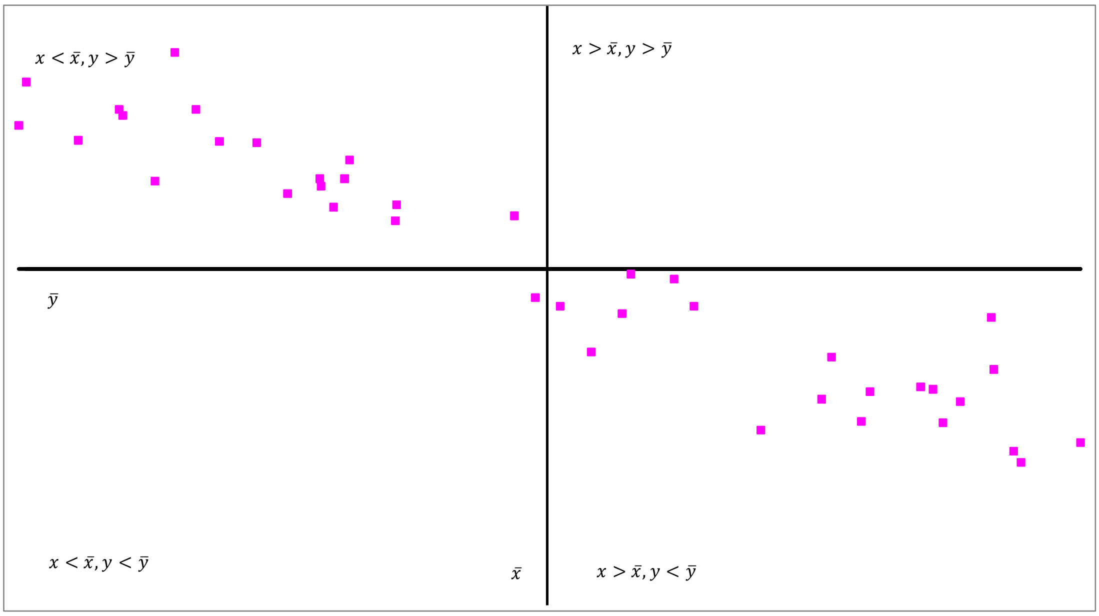


Figure 1c: cov(x,y)=0

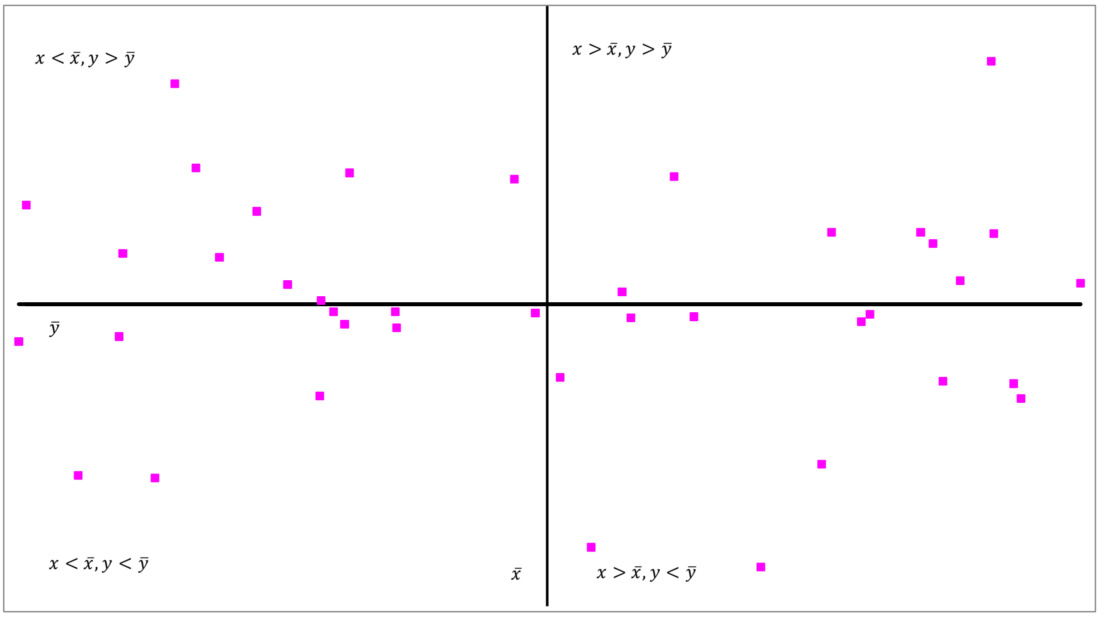
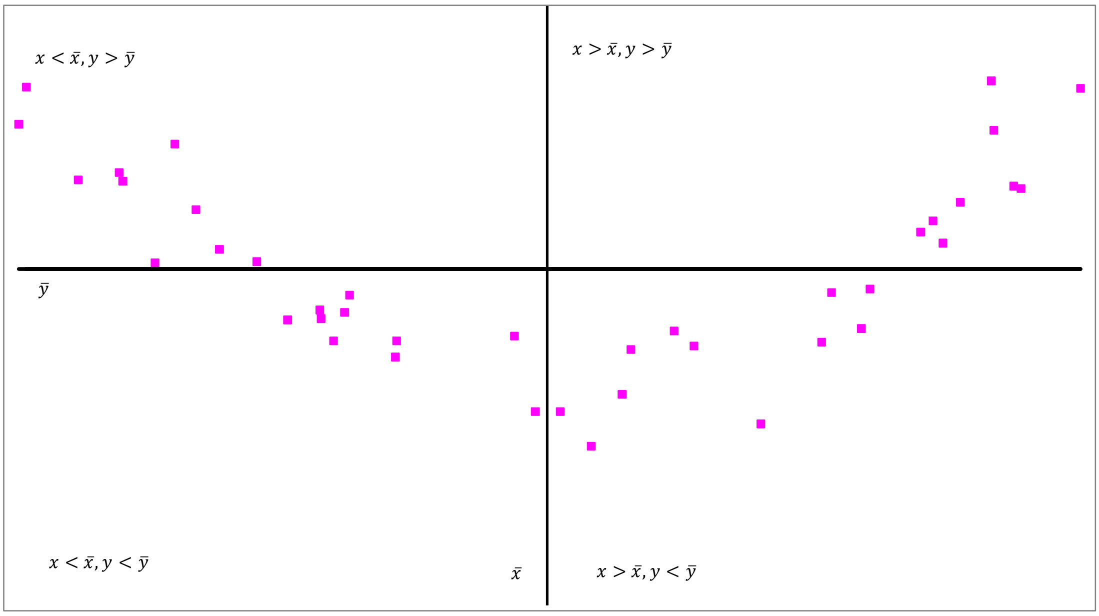


Figure 1d: cov(x,y)=0



More rules on expectations variances

Define the probability density function for the random variables *X*, *Y* as $p(X,Y)$, such that, $\sum_x \sum_y p(x,y) = 1$. Define the marginal density of *X* as *p*(*x*) and the marginal probability density for *Y* as *p*(*y*), such that $p(X) = \sum_y p(x,y)$ and $p(Y) = \sum_x p(x,y)$.

Then define $E(X) = \sum_y xp(y)$, $E(Y) = \sum_y yp(y)$, $V(X) = \sum_x (x - E(X))^2 p(x)$, $V(Y) = \sum_y (y - E(Y))^2 p(y)$ and $\mathrm{cov}(X,Y) = \sum_x \sum_y (x - E(X))(y - E(Y))\, p(x,y)$.

Then we can show that

1. $E(X+Y) = \sum_x \sum_y (x+y)\, p(x,y) = \sum_x \sum_y xp(x,y) + \sum_x \sum_y yp(x,y)$

$$\sum_x x \underbrace{\sum_y p(x,y)}_{p(x)} + \sum_y y \underbrace{\sum_x p(x,y)}_{p(y)} = \sum_x xp(x) + \sum_y yp(y) = E(X) + E(Y)$$

2. $\mathrm{cov}(a + X, Y) = \sum_x \sum_y ((a+x) - E(a+X))(y - E(Y))\, p(x,y)$

$$= \sum_x \sum_y (a + x - a - E(X))(y - E(Y))\, p(x,y) = \mathrm{cov}(X,Y).$$

3. $\mathrm{cov}(aX,Y) = \sum_x \sum_y (ax - E(aX))(y - E(Y))\, p(x,y)$

$$\mathrm{cov}(aX,Y) = \sum_x \sum_y (ax - aE(X))(y - E(Y))\, p(x,y)$$

$$\mathrm{cov}(aX,Y) = a \sum_x \sum_y (x - E(X))(y - E(Y))\, p(x,y) = a\,\mathrm{cov}(X,Y)$$

4. $V(X+Y) = \sum_x \sum_y \left[(x+y) - E(X+Y)\right]^2 p(x,y) = \sum_x \sum_y \left[(x - E(X)) + (y - E(Y))\right]^2 p(x,y)$

$$\sum_x \sum_y \left[(x - E(X)) + (y - E(Y))\right]^2 p(x,y)$$

$$\sum_x \sum_y (x - E(X))^2 p(x,y) + \sum_x \sum_y (y - E(Y))^2 p(x,y) + 2 \sum_x \sum_y (x - E(X)(y - E(Y))\, p(x,y)$$

$$= \sum_x (x - E(X))^2 \underbrace{\sum_y p(x,y)}_{p(x)} + \sum_x (y - E(Y))^2 \underbrace{\sum_y p(x,y)}_{p(y)} + 2\,\mathrm{cov}(X,Y)$$

$$= V(X) + V(Y) + 2\,\mathrm{cov}(X,Y)$$

5. $E(X - Y) = E(X) - E(Y)$

6. $V(X - Y) = V(X) + V(Y) - 2\,\mathrm{cov}(X,Y)$

7. $E(aX - bY) = aE(X) - bE(Y)$

8. $V(aX - bY) = a^2V(X) + b^2V(Y) - 2ab\,\mathrm{cov}(X,Y)$

9. $E(X + Y + Z) = E(X) + E(Y) + E(Z)$

10. $V(X + Y + Z) = V(X) + V(Y) + V(Z) + 2\,\mathrm{cov}(X,Y) + 2\,\mathrm{cov}(X,Z) + 2\,\mathrm{cov}(Y,Z)$

11. $E(X - Y - Z) = E(X) - E(Y) - E(Z)$

12. $V(X - Y - Z) = V(X) + V(Y) + V(Z) - 2\,\mathrm{cov}(X,Y) - 2\,\mathrm{cov}(X,Z) + 2\,\mathrm{cov}(Y,Z)$

Suppose $E(X_i) = \mu$, $V(X_i) = \sigma^2$ for all *i* and $\mathrm{cov}(X_i, X_j) = \gamma_{|i-j|}$ for all *i* and *j*, $i \neq j$.

Now define $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n} = \dfrac{(X_1 + X_2 + \ldots + X_n)}{n}$

$$E(\bar{X}) = E\left[\frac{(X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{1}{n}E(X_1 + X_2 + \ldots + X_n)$$

$$= \frac{1}{n}\left[E(X_1) + \ldots + E(X_n)\right] = \frac{1}{n}\left[\mu + \mu + \ldots + \mu\right] = \mu$$

$$V(\bar{X}) = V\left[\frac{(X_1 + X_2 + \ldots + X_n)}{n}\right] = \frac{1}{n^2}[V(X_1) + V(X_2) + \ldots + V(X_n)$$

$$+2\,\mathrm{cov}(X_1, X_2) + 2\,\mathrm{cov}(X_1, X_3) + \ldots + 2\,\mathrm{cov}(X_1, X_n)$$

$$+2\,\mathrm{cov}(X_2, X_3) + \ldots + 2\,\mathrm{cov}(X_2, X_n)$$

$$+ \ldots +$$

$$\ldots + 2\,\mathrm{cov}(X_{n-1}, X_n)]$$

assuming $X_1, \ldots, X_n$ are a random sample (with or without replacement, providing *n* is sufficiently large) then $\mathrm{cov}(X_{,}, X_j) = 0$ for $i \neq j$ and

$V(\bar{X}) = \dfrac{1}{n^2}[\sigma^2 + \sigma^2 + \ldots + \sigma^2] = \dfrac{\sigma^2}{n}$ (therefore the variance of the sample mean falls as the number of points in the sample increases – this is the effect of smoothing see Appendix 6: Figures 1a-1d).

Figure 2: Effects of averaging of the standard deviation

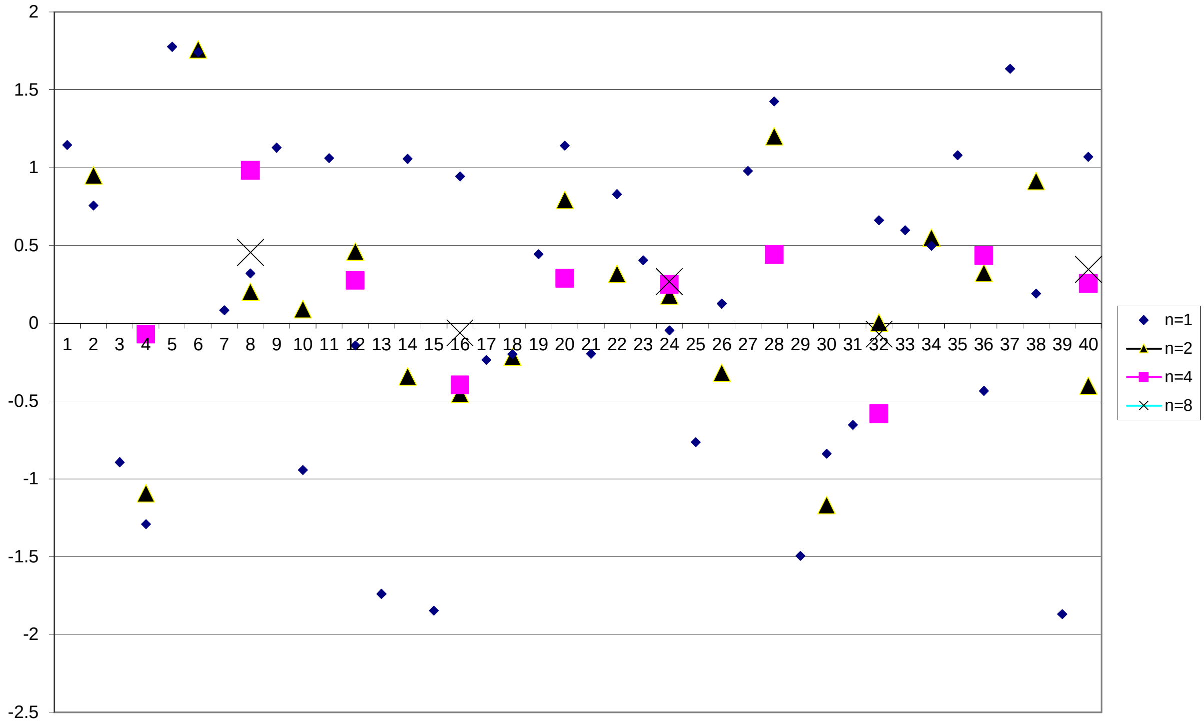
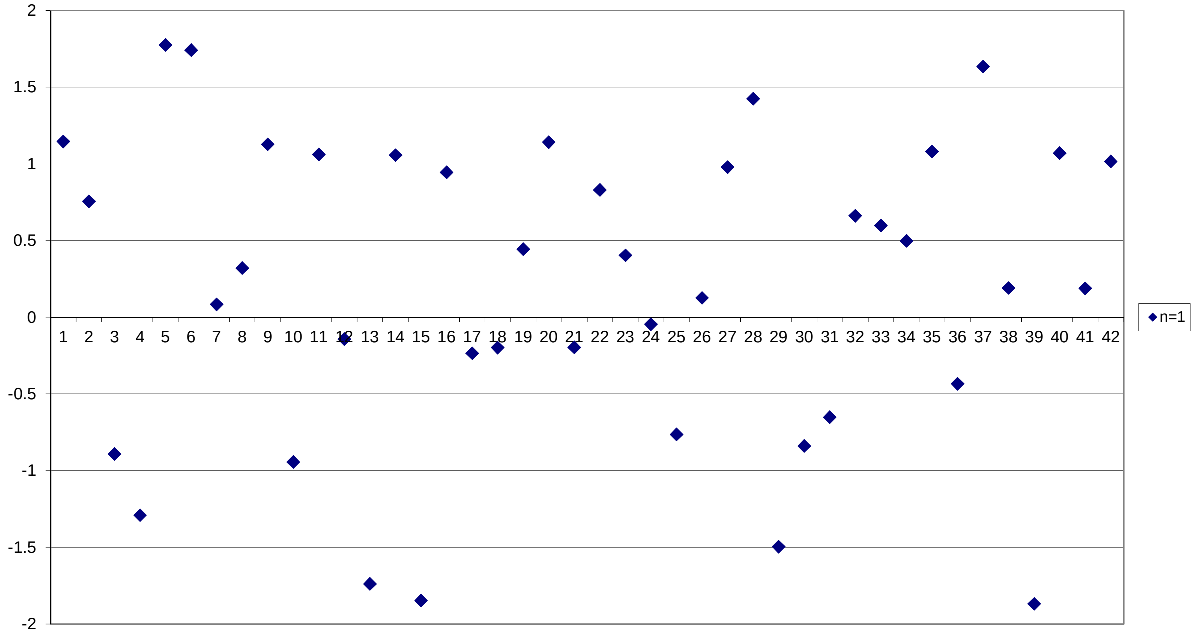


Figure 2a: Dispersion with an average of n=1



Appendix 2: Handout 3

Figure 2b: Dispersion with an average of n=2

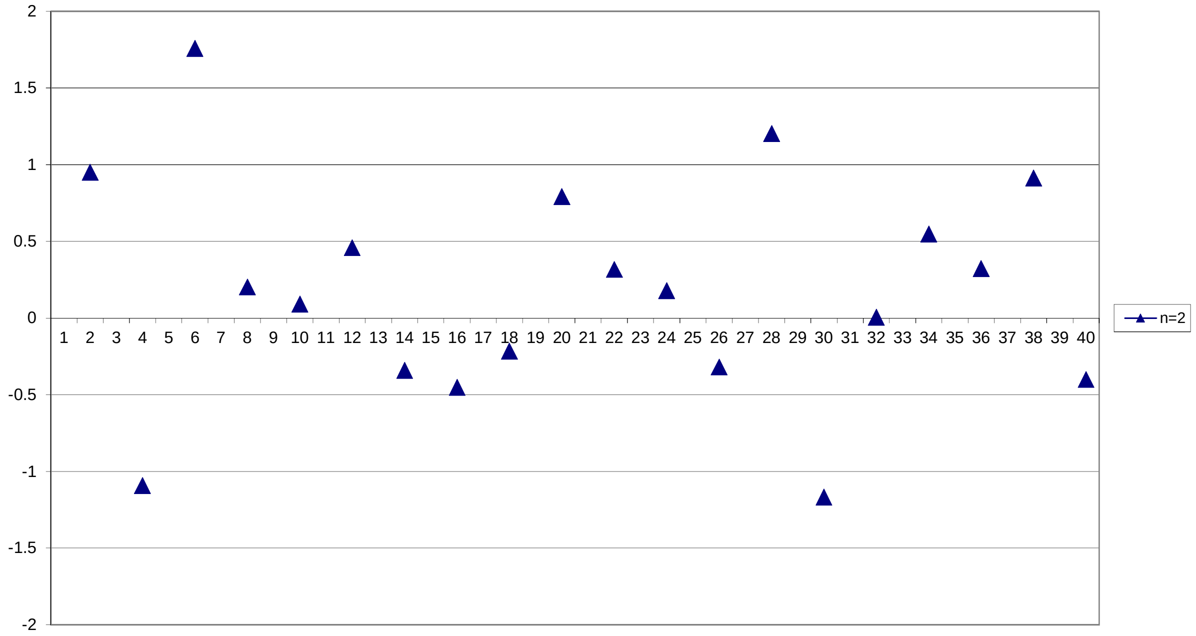


Figure 2c: Dispersion with an average of n=4

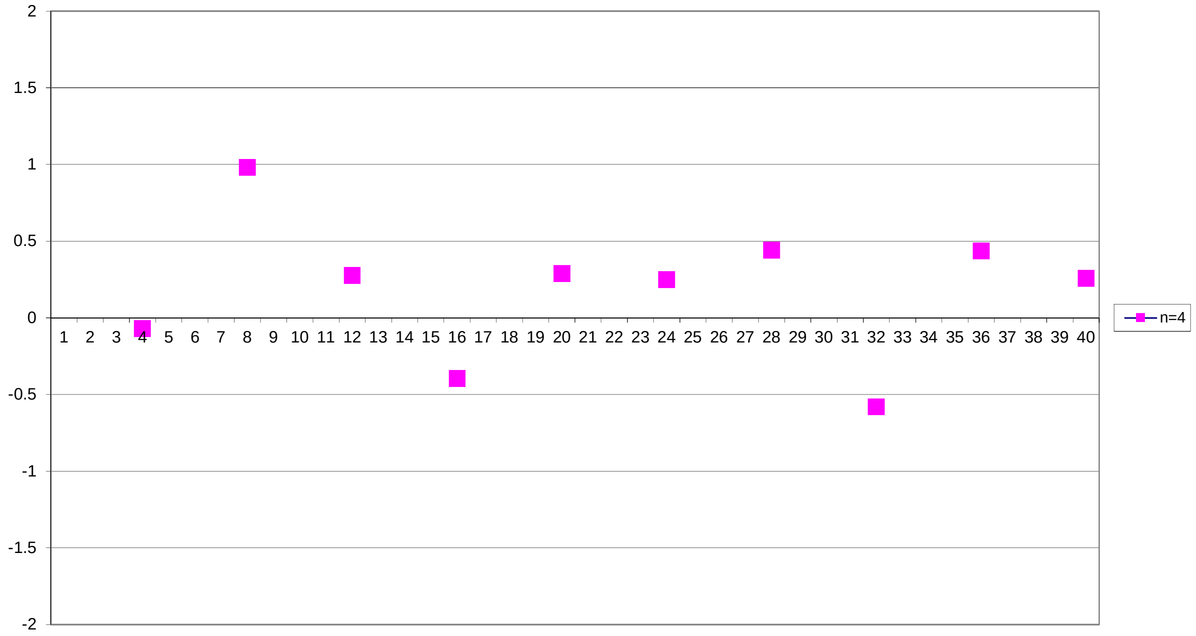
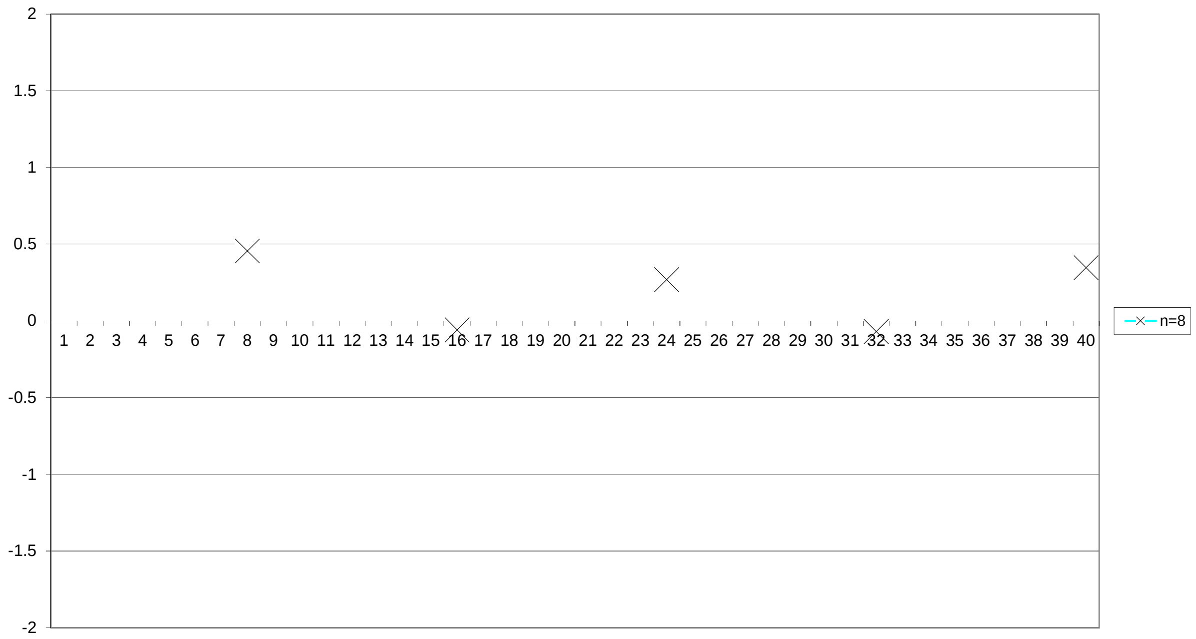


Figure 2d: Dispersion with an average of n=8



Basic Integration

Integration can be approximated as a summation of the function over small changes in *x*, that is,

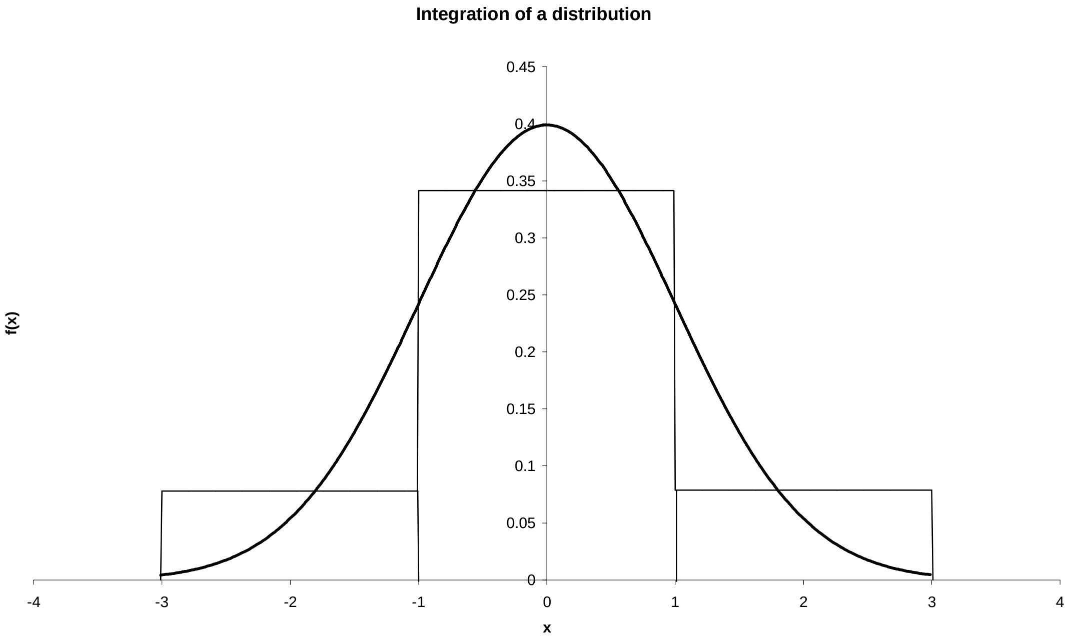
$\int_x f(x)\partial x \approx \sum_{i=1} f(x_i)(x_i - x_{i-1})$, the approximation is best as the changes in *x* become infinitesimal (very small) – see figures overleaf – where we can see that as the changes in *x*, ∂x, become smaller so the approximation of the areas of the rectangles become a better approximation to the area under the continuous line.

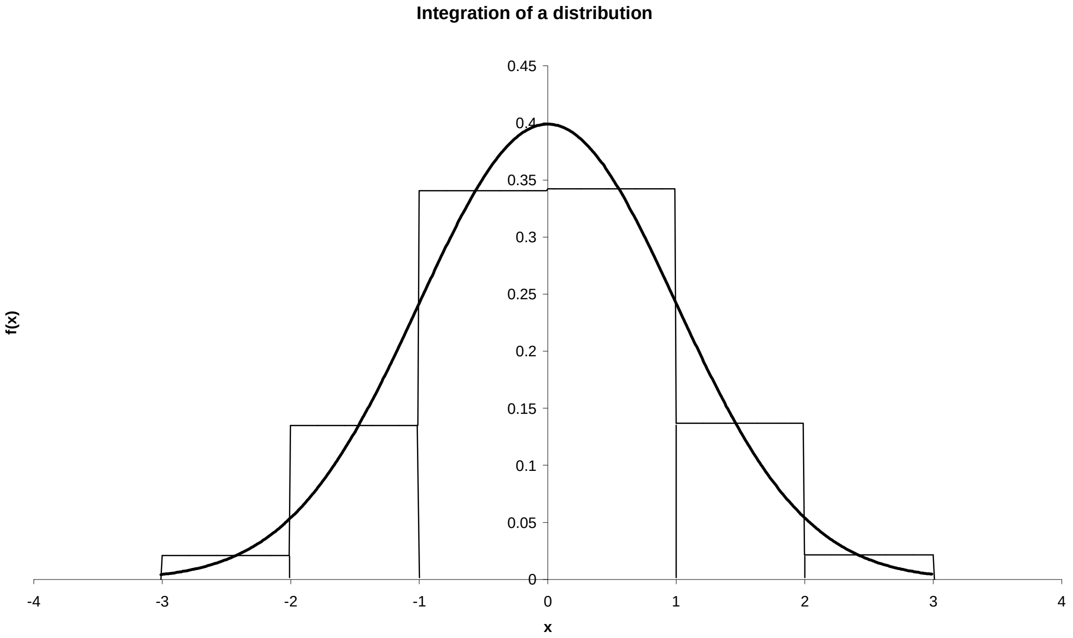
Some basic rules of integration:

1. $\int_c^d x^n \partial x = \left.\frac{x^{n+1}}{n+1}\right|_c^d = \frac{d^{n+1}}{n+1} - \frac{c^{n+1}}{n+1}$ for all $n \neq -1$
2. $\int_c^d x^{-1} \partial x = \ln(x)\big|_c^d = \ln(d) - \ln(c)$
3. $\int_c^d e^x \partial x = e^x\big|_c^d = e^d - e^c$

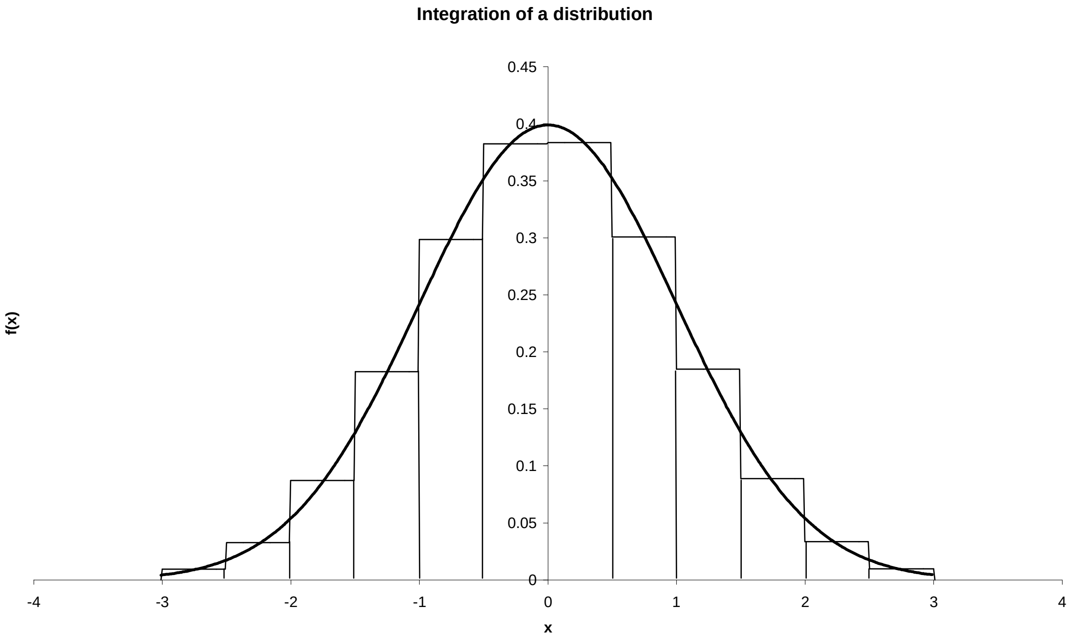
Remember integration is simply the inverse function of differentiation, so

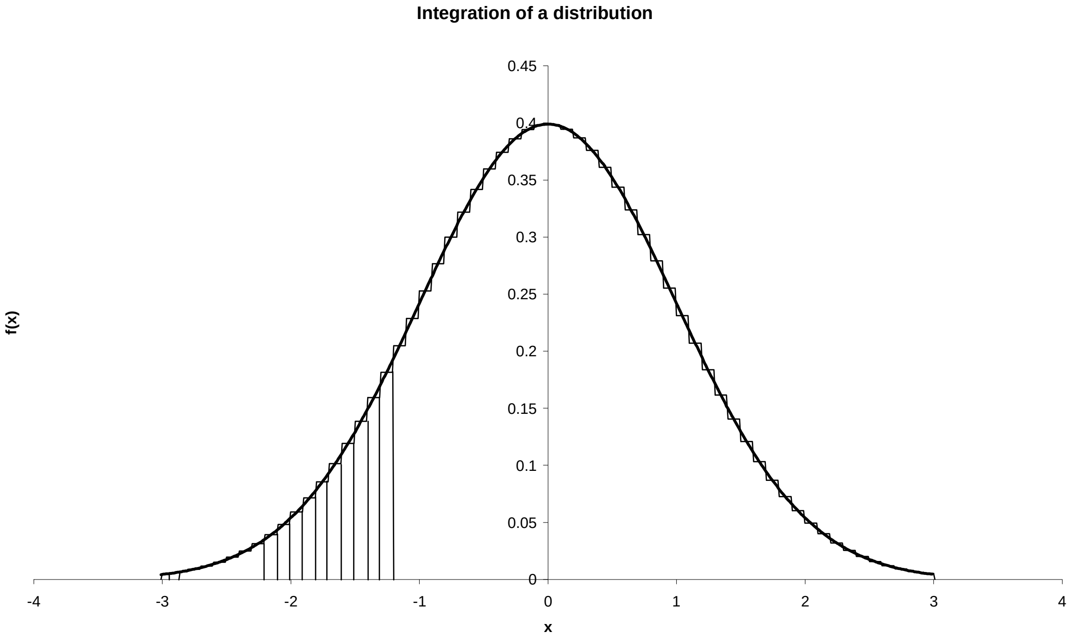
$$\frac{\partial x^n}{\partial x} = nx^{n-1} \Rightarrow \int nx^{n-1}\partial x = \frac{nx^n}{n} = x^n$$


Integration of a distribution
f(x)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
x


Integration of a distribution
f(x)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
x

Handout 2: Appendix7





STATISTICAL TECHNIQUES B

Special Distributions

1. Bernoulli distribution

An experiment leading to only two outcomes – a 'success' and a 'failure' – called a Bernoulli trial *x*=0 (=failure) with probability, 1-*p*, and 1 (=success) with probability, *p*.

x	0	1
P(*X*=*x*)	1-*p*	*p*

This is a valid pdf as:

$$\sum_{x=0} p_X(x) = (1-p) + p = 1$$

1.1 Mean

$$E(X) = 0(1-p) + 1p = p$$

$$E(X^2) = 0^2(1-p) + 1^2 p = p$$

1.2 Variance

$$V(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$$

2. Binomial distribution

This consists of having n independent Bernoulli trials, where we define *X*=number of `successes’ in *n* trials and *X*=0,1,2,…*n*.

We can then calculate the probability of a specific outcome such as:

$$P(S_1 \cap S_2 \ldots \cap S_x \cap F_{x+1} \cap F_{x+2} \ldots \cap F_n) = p^x (1-p)^{n-x}$$

However, as there are ${}_nC_x = \dfrac{n!}{(n-x)!x!}$ in which we can get *x* successes in *n* trials, given that the order is unimportant, we have:

$$p_X(x) = {}_nC_x p^x (1-p)^{n-x}$$

This is a valid probability density function as:

$$\sum_{x=0}^{n} p_X(x) = \sum_{x=0}^{n} {}_nC_x p^x (1-p)^{n-x} = (p + (1-p))^n = 1$$

For, *n*=3

$$p_X(0) + p_X(1) + p_X(2) + p_X(3) = p^0 (1-p)^3 + 3p(1-p)^2 + 3p^2(1-p) + p^3 (1-p)^0$$

2.1 Mean

$$E(X) = \sum_{x=0}^{n} x p_X(x) = \sum_{x=0}^{n} x \ {}_nC_x p^x (1-p)^{n-x} = np \sum_{x=1}^{n} {}_{n-1}C_{x-1} p^{x-1} (1-p)^{n-x}$$

defining *m*=*n*-1 and *y*=*x*-1 then,

$$E(X) = np \underbrace{\sum_{y=0}^{m} {}_mC_y p^y (1-p)^{m-y}}_{1} = np$$

2.2 Variance

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = E[X(X-1)] + E(X)$$

$$E[X(X-1)] =$$

$$\sum_{x=0}^{n} x(x-1) p_X(x) = \sum_{x=0}^{n} x(x-1) \ {}_nC_x p^x (1-p)^{n-x} = n(n-1)p^2 \sum_{x=2}^{n} {}_{n-2}C_{x-2} p^{x-2} (1-p)^{n-x}$$

$$= n(n-1)p^2 \sum_{y=0}^{m} {}_mC_y p^y (1-p)^{m-y} = n(n-1)p^2$$

$$E(X^2) = n(n-1)p^2 + np$$

$$V(X) = n(n-1)p^2 + np - n^2p^2 = np\left[np - p + 1 - np\right] = np(1-p)$$

2.3 Probability values (see Appendix 1: Table 7)

$$P(a \le X \le b) = P(a) + P(a+1) + \dots + P(b)$$

$$P(a \le X \le b) = {}_nC_a p^a (1-p)^{n-a} + {}_nC_{a+1} p^{a+1} (1-p)^{n-a-1} + \dots + {}_nC_b p^b (1-p)^{n-b}$$

(see Appendix 3 for a binomial expansion)

3. Poisson

We are interested in the number of occurrences of an event during a period of time, where the period of time, *T*, is divided into *n* unit time intervals (and n is very large). The probability of an event in an interval of time is *p* (which is small) and we assume that the events are independent occurrences.

$\lambda =$ mean rate of occurrence

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0,1,2\ldots$$

This is a valid probability density function as:

$$\sum_{x=0} p_X(x) = \sum_{x=0} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \underbrace{\sum_{x=0} \frac{\lambda^x}{x!}}_{e^{\lambda}} = 1$$

as,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

3.1 Mean

$$E(X) = \sum_{x=0} x p_X(x) = \sum_{x=0} x \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\left[\frac{1\lambda}{1!} + \frac{2\lambda^2}{2!} + \frac{3\lambda^3}{3!} + \frac{4\lambda^4}{4!} + \ldots\right]$$

$$= e^{-\lambda}\lambda\left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right] = \lambda e^{-\lambda} \underbrace{\sum_{y=0} \frac{\lambda^y}{y!}}_{e^{\lambda}} = \lambda$$

3.2 Variance

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = E[X(X-1)] + E(X)$$

$$E(X(X-1)) = \sum_{x=0} x(x-1) p_X(x) = \sum_{x=0} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\left[\frac{2\lambda^2}{2!} + \frac{3(2)\lambda^3}{3!} + \frac{4(3)\lambda^4}{4!} + \ldots\right]$$

$$= e^{-\lambda}\lambda^2\left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right] = \lambda^2$$

$$E(X^2) = \lambda^2 + \lambda$$

$$V(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

3.3 Probability values (see Appendix 1: Table 8)

$$P(a \le X \le b) = P(a) + P(a+1) + \ldots + P(b)$$

$$P(a \le X \le b) = e^{-\lambda}\left[\frac{\lambda^{a}}{a!} + \frac{\lambda^{a+1}}{(a+1)!} + \ldots \frac{\lambda^{b}}{b!}\right]$$

NOTE:

The Poisson distribution can be used as an approximation to the Binomial distribution having the same mean. If a binomial distribution has a large, *n* (*n*>50) and a small *p* (*p*<0.1), then the probabilities of 0, 1, 2, … successes given by a Poison distribution with parameter $\lambda = np$ approximates well to the true probabilities given by the defined binomial distribution.

4. Uniform

Define the probability density function, such that all points in the interval (*a*,*b*) have an equal likelihood of occurring,

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{elsewhere} \end{cases}$$

This is a valid probability density function as:

$$\int_a^b f(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b 1\,dx = \frac{1}{b-a}\Big[x\Big|_a^b = 1$$

4.1 Mean

$$E(X) = \int_a^b xf(x)\,dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\Big[x^2/2\Big|_a^b = \frac{1}{b-a}\Big[b^2/2 - a^2/2\Big]$$

$$= \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2}$$

4.2 Variance

$$V(X) = E(X^2) - E(X)^2$$

$$E(X^2) = \int_a^b x^2 f(x)\,dx = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\Big[x^3/3\Big|_a^b = \frac{1}{b-a}\Big[b^3/3 - a^3/3\Big]$$

$$= \frac{(b^3-a^3)}{3(b-a)} = \frac{(b^2+a^2)(b-a)+ab(b-a)}{3(b-a)} = \frac{(b^2+a^2)+ab}{3}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{(b^2+a^2)+ab}{3} - \frac{(b+a)^2}{4}$$

$$V(X) = \frac{4(b^2+a^2)+4ab-3(b^2+a^2+2ab)}{12} = \frac{b^2+a^2-2ab}{12} = \frac{(b-a)^2}{12}$$

4.3 Probability values

$$P(c \le X \le d) = \int_c^d \frac{1}{b-a}\,dx = \frac{x}{b-a}\Bigg|_c^d = \frac{d-c}{b-a}$$

5. Normal Distribution

Define the probability density function

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty$$

This is a valid probability density function as:

$$\int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx = 1$$

although, showing this is non-trivial.

5.1 Mean

$E(X) = \mu$

although, showing this is non-trivial.

5.2 Variance

$V(X) = \sigma^2$

although, showing this is non-trivial.

5.3 Probability values

$$P(a \le X \le b) = \int_{a}^{b} (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$

and this is non-trivial. However, statistical tables are available for the standard normal distribution, *Z*, where $E(Z) = 0$ and $V(Z) = 1$, such that:

$$P(Z \le c) = (2\pi)^{-1/2} \int_{-\infty}^{c} \exp\left(\frac{-z^2}{2}\right) dz$$

As we know that

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right)$$

$$\Rightarrow P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$

and this can be calculated from the standard normal statistical tables as:

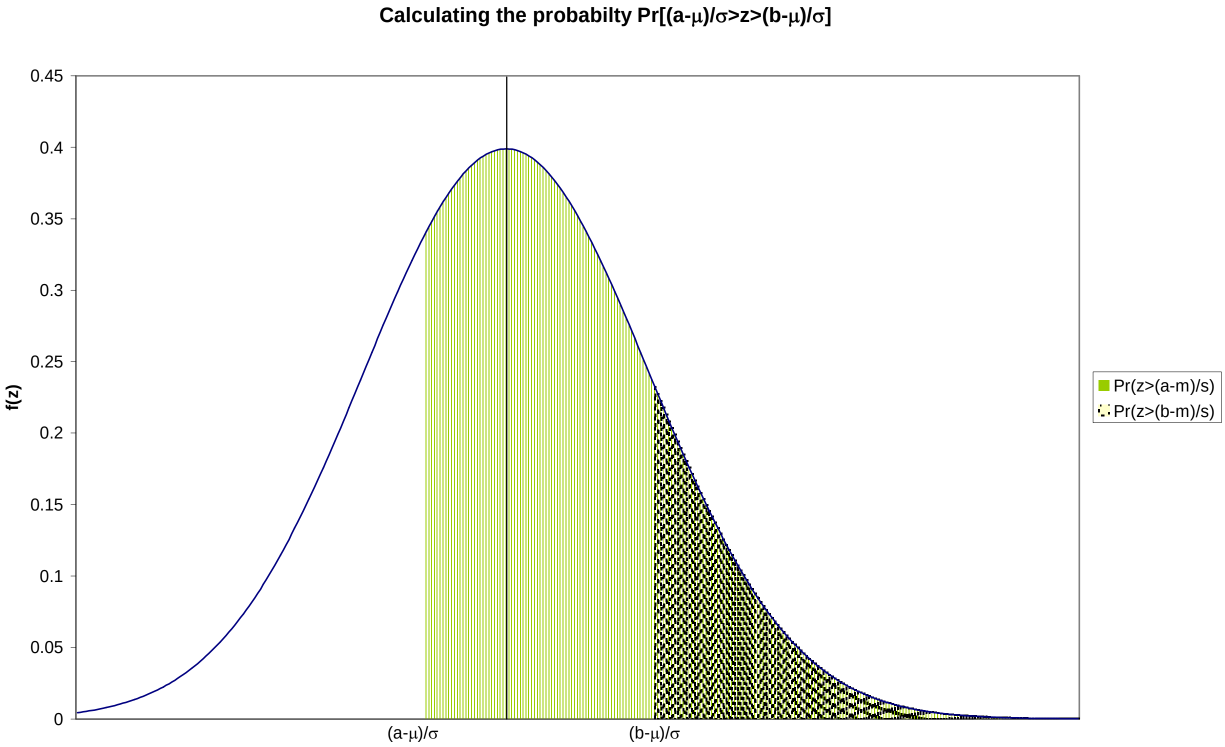
$$P(a \le X \le b) = P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right)$$

(see Figure 1 and **Appendix 1: Table 1**).

NOTE: As the area under the pdf is unity, $P(Z > c) = 1 - P(Z < c)$. In addition due to symmetry we have that $P(Z < -c) = P(Z > c) = 1 - P(Z < c)$ where $c > 0$.

<u>NOTE:</u> If X_1 and X_2 are both normally distributed then any linear combination of them is also normally distributed. Suppose we take a random sample from some population such that $\underline{X_i \sim N(\mu, \sigma^2)}$ then $\underline{\sum_{i=1}^{k} a_i X_i \sim N(\mu \sum_{i=1}^{k} a_i, \sigma^2 \sum_{i=1}^{k} a_i^2)}$.

Figure 1



6. Chi-squared distribution

Define the pdf as:

$$f(x) = \frac{1}{\Gamma(\upsilon/2)}\left(\frac{1}{2}\right)^{\frac{\upsilon}{2}} x^{\upsilon/2-1}e^{-x/2} \quad x > 0$$

This is a valid probability density function and is denoted as χ^2_υ, where υ are the degrees of freedom.

NOTE: $N(0,1)^2 = \chi^2_1$ and if $W_i \sim \chi^2_1$ and these are independent, then $\sum_{i=1}^{n} W_i = \chi^2_n$

6.1 Mean

$E(X) = \upsilon$

6.2 Variance

$V(X) = E(X^2) - E(X)^2 = 2\upsilon$

6.3 Probability values (see Appendix 1: Table 3)

These are tabulated in the statistical tables.

7. F-distribution

Define the pdf as:

$$f(x) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2} \frac{x^{(m\text{-}2)/2}}{[1+(m/n)x]^{(m+n)/2}} \quad x > 0$$

This is a valid probability density function and is denoted as an F-distribution with degrees of freedom *m* and *n*.

NOTE: An F distribution is formed as the ratio of 2 independent chi-squared distributions, $\frac{\chi_m^2/m}{\chi_n^2/n} \sim F_{m,n}$. As $n \to \infty$ so $F_{m,n} = \frac{\chi_m^2/m}{\chi_\infty^2/\infty} \sim \chi_m^2/m$.

7.1 Mean (for a F$_{m,n}$)

$$E(X) = \frac{n}{n-2} \qquad \text{for } n > 2$$

7.2 Variance (for a F$_{m,n}$)

$$V(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad \text{for } n > 4$$

7.3 Probability values (see Appendix 1: Table 5)

These are tabulated in the statistical tables.

8. Student t-distribution

Define the pdf as:

$$f(x) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)} \frac{1}{\sqrt{n\pi}} \frac{1}{(1+x^2/n)^{(n+1)/2}}$$

This is a valid probability density function and is denoted as a t-distribution with degrees of freedom n.

NOTE: A t-distribution is formed as the ratio of a N(0,1) to a chi-square distribution, $\frac{N(0,1)}{\sqrt{\chi_n^2/n}} \sim t_n$. As $n \to \infty$ so $t_n \sim N(0,1)$.

8.1 Mean (for a t*n*)

$E(X) = 0$

8.2 Variance (for a t*n*)

$$V(X) = \frac{n}{(n-2)} \quad \text{for } n > 2$$

8.3 Probability values (see Appendix 1: Table 2)

These are tabulated in the statistical tables.

Table 7 (from Statistical tables): Binomial Distribution (cont'd)

$\Pr(X \leq k)$

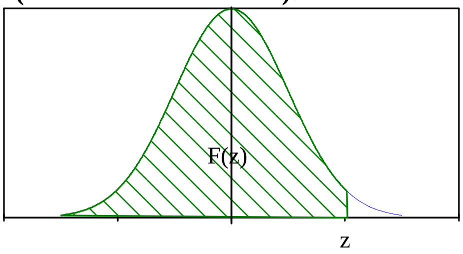
	n=8								n=9					
p	k=0	1	2	3	4	5	6	7	k=0	1	2	3	4	5
0.01	0.923	0.997	1.000						0.914	0.997	1.000			
0.02	0.851	0.990	1.000						0.834	0.987	0.999	1.000		
0.03	0.784	0.978	0.999	1.000					0.760	0.972	0.998	1.000		
0.04	0.721	0.962	0.997	1.000					0.693	0.952	0.996	1.000		
0.05	0.663	0.943	0.994	1.000					0.630	0.929	0.992	0.999	1.000	
0.06	0.610	0.921	0.990	0.999	1.000				0.573	0.902	0.986	0.999	1.000	
0.07	0.560	0.897	0.985	0.999	1.000				0.520	0.873	0.979	0.998	1.000	
0.08	0.513	0.870	0.979	0.998	1.000				0.472	0.842	0.970	0.996	1.000	
0.09	0.470	0.842	0.971	0.997	1.000				0.428	0.809	0.960	0.994	0.999	1.000
0.10	0.430	0.813	0.962	0.995	1.000				0.387	0.775	0.947	0.992	0.999	1.000
0.11	0.394	0.783	0.951	0.993	0.999	1.000			0.350	0.740	0.933	0.988	0.999	1.000
0.12	0.360	0.752	0.939	0.990	0.999	1.000			0.316	0.705	0.917	0.984	0.998	1.000
0.13	0.328	0.721	0.926	0.987	0.999	1.000			0.286	0.670	0.899	0.979	0.997	1.000
0.14	0.299	0.689	0.911	0.983	0.998	1.000			0.257	0.634	0.880	0.973	0.996	1.000
0.15	0.272	0.657	0.895	0.979	0.997	1.000			0.232	0.599	0.859	0.966	0.994	0.999
0.16	0.248	0.626	0.877	0.973	0.996	1.000			0.208	0.565	0.837	0.958	0.993	0.999
0.17	0.225	0.594	0.859	0.967	0.995	1.000			0.187	0.532	0.814	0.949	0.990	0.999
0.18	0.204	0.563	0.839	0.960	0.993	0.999	1.000		0.168	0.499	0.790	0.938	0.988	0.998
0.19	0.185	0.533	0.819	0.952	0.992	0.999	1.000		0.150	0.467	0.764	0.927	0.984	0.998
0.20	0.168	0.503	0.797	0.944	0.990	0.999	1.000		0.134	0.436	0.738	0.914	0.980	0.997
0.21	0.152	0.474	0.775	0.934	0.987	0.998	1.000		0.120	0.407	0.711	0.901	0.976	0.996
0.22	0.137	0.446	0.751	0.924	0.984	0.998	1.000		0.107	0.378	0.684	0.886	0.971	0.995
0.23	0.124	0.419	0.728	0.912	0.981	0.997	1.000		0.095	0.351	0.657	0.870	0.965	0.994
0.24	0.111	0.392	0.703	0.900	0.977	0.997	1.000		0.085	0.325	0.629	0.852	0.958	0.992
0.25	0.100	0.367	0.679	0.886	0.973	0.996	1.000		0.075	0.300	0.601	0.834	0.951	0.990
0.26	0.090	0.343	0.653	0.872	0.968	0.995	1.000		0.067	0.277	0.573	0.815	0.943	0.988
0.27	0.081	0.319	0.628	0.857	0.962	0.994	0.999	1.000	0.059	0.255	0.545	0.795	0.934	0.985
0.28	0.072	0.297	0.603	0.841	0.956	0.992	0.999	1.000	0.052	0.234	0.517	0.774	0.924	0.982
0.29	0.065	0.276	0.577	0.824	0.949	0.991	0.999	1.000	0.046	0.214	0.490	0.752	0.913	0.979
0.30	0.058	0.255	0.552	0.806	0.942	0.989	0.999	1.000	0.040	0.196	0.463	0.730	0.901	0.975
0.31	0.051	0.236	0.526	0.787	0.934	0.987	0.998	1.000	0.035	0.179	0.436	0.706	0.888	0.970
0.32	0.046	0.218	0.501	0.768	0.925	0.984	0.998	1.000	0.031	0.163	0.411	0.683	0.875	0.965
0.33	0.041	0.201	0.476	0.748	0.915	0.981	0.998	1.000	0.027	0.148	0.385	0.658	0.860	0.960
0.34	0.036	0.184	0.452	0.728	0.905	0.978	0.997	1.000	0.024	0.134	0.361	0.634	0.845	0.953
0.35	0.032	0.169	0.428	0.706	0.894	0.975	0.996	1.000	0.021	0.121	0.337	0.609	0.828	0.946
0.36	0.028	0.155	0.404	0.685	0.882	0.971	0.996	1.000	0.018	0.109	0.314	0.584	0.811	0.939
0.37	0.025	0.141	0.381	0.663	0.869	0.966	0.995	1.000	0.016	0.098	0.292	0.558	0.793	0.930
0.38	0.022	0.129	0.359	0.640	0.856	0.961	0.994	1.000	0.014	0.088	0.271	0.533	0.774	0.921
0.39	0.019	0.117	0.337	0.617	0.841	0.956	0.993	0.999	0.012	0.079	0.251	0.508	0.754	0.911
0.40	0.017	0.106	0.315	0.594	0.826	0.950	0.991	0.999	0.010	0.071	0.232	0.483	0.733	0.901
0.41	0.015	0.096	0.295	0.571	0.810	0.944	0.990	0.999	0.009	0.063	0.213	0.458	0.712	0.889
0.42	0.013	0.087	0.275	0.547	0.794	0.937	0.988	0.999	0.007	0.056	0.196	0.433	0.690	0.877
0.43	0.011	0.078	0.256	0.524	0.776	0.929	0.986	0.999	0.006	0.049	0.180	0.409	0.668	0.863
0.44	0.010	0.070	0.238	0.500	0.758	0.921	0.984	0.999	0.005	0.044	0.164	0.385	0.645	0.849
0.45	0.008	0.063	0.220	0.477	0.740	0.912	0.982	0.998	0.005	0.039	0.150	0.361	0.621	0.834
0.46	0.007	0.057	0.203	0.454	0.720	0.902	0.979	0.998	0.004	0.034	0.136	0.339	0.598	0.818
0.47	0.006	0.050	0.187	0.431	0.700	0.891	0.976	0.998	0.003	0.030	0.123	0.316	0.573	0.801
0.48	0.005	0.045	0.172	0.408	0.680	0.880	0.973	0.997	0.003	0.026	0.111	0.295	0.549	0.784
0.49	0.005	0.040	0.158	0.385	0.658	0.868	0.969	0.997	0.002	0.023	0.100	0.274	0.525	0.765
0.50	0.004	0.035	0.145	0.363	0.637	0.855	0.965	0.996	0.002	0.020	0.090	0.254	0.500	0.746

Table 8 (from Statistical Tables): Poisson Distribution

$\Pr(X \le k)$

λ	k=0	1	2	3	4	5	6	7	8	9	10
0.1	0.905	0.995	1.000	1.000							
0.2	0.819	0.982	0.999	1.000							
0.3	0.741	0.963	0.996	1.000							
0.4	0.670	0.938	0.992	0.999	1.000						
0.5	0.607	0.910	0.986	0.998	1.000						
0.6	0.549	0.878	0.977	0.997	1.000						
0.7	0.497	0.844	0.966	0.994	0.999	1.000					
0.8	0.449	0.809	0.953	0.991	0.999	1.000					
0.9	0.407	0.772	0.937	0.987	0.998	1.000					
1.0	0.368	0.736	0.920	0.981	0.996	0.999	1.000				
1.1	0.333	0.699	0.900	0.974	0.995	0.999	1.000				
1.2	0.301	0.663	0.879	0.966	0.992	0.998	1.000				
1.3	0.273	0.627	0.857	0.957	0.989	0.998	1.000				
1.4	0.247	0.592	0.833	0.946	0.986	0.997	0.999	1.000			
1.5	0.223	0.558	0.809	0.934	0.981	0.996	0.999	1.000			
1.6	0.202	0.525	0.783	0.921	0.976	0.994	0.999	1.000			
1.7	0.183	0.493	0.757	0.907	0.970	0.992	0.998	1.000			
1.8	0.165	0.463	0.731	0.891	0.964	0.990	0.997	0.999	1.000		
1.9	0.150	0.434	0.704	0.875	0.956	0.987	0.997	0.999	1.000		
2.0	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000		
2.1	0.122	0.380	0.650	0.839	0.938	0.980	0.994	0.999	1.000		
2.2	0.111	0.355	0.623	0.819	0.928	0.975	0.993	0.998	1.000		
2.3	0.100	0.331	0.596	0.799	0.916	0.970	0.991	0.997	0.999	1.000	
2.4	0.091	0.308	0.570	0.779	0.904	0.964	0.988	0.997	0.999	1.000	
2.5	0.082	0.287	0.544	0.758	0.891	0.958	0.986	0.996	0.999	1.000	
2.6	0.074	0.267	0.518	0.736	0.877	0.951	0.983	0.995	0.999	1.000	
2.7	0.067	0.249	0.494	0.714	0.863	0.943	0.979	0.993	0.998	0.999	1.000
2.8	0.061	0.231	0.469	0.692	0.848	0.935	0.976	0.992	0.998	0.999	1.000
2.9	0.055	0.215	0.446	0.670	0.832	0.926	0.971	0.990	0.997	0.999	1.000
3.0	0.050	0.199	0.423	0.647	0.815	0.916	0.966	0.988	0.996	0.999	1.000
3.1	0.045	0.185	0.401	0.625	0.798	0.906	0.961	0.986	0.995	0.999	1.000
3.2	0.041	0.171	0.380	0.603	0.781	0.895	0.955	0.983	0.994	0.998	1.000
3.3	0.037	0.159	0.359	0.580	0.763	0.883	0.949	0.980	0.993	0.998	0.999
3.4	0.033	0.147	0.340	0.558	0.744	0.871	0.942	0.977	0.992	0.997	0.999
3.5	0.030	0.136	0.321	0.537	0.725	0.858	0.935	0.973	0.990	0.997	0.999
3.6	0.027	0.126	0.303	0.515	0.706	0.844	0.927	0.969	0.988	0.996	0.999
3.7	0.025	0.116	0.285	0.494	0.687	0.830	0.918	0.965	0.986	0.995	0.998
3.8	0.022	0.107	0.269	0.473	0.668	0.816	0.909	0.960	0.984	0.994	0.998
3.9	0.020	0.099	0.253	0.453	0.648	0.801	0.899	0.955	0.981	0.993	0.998
4.0	0.018	0.092	0.238	0.433	0.629	0.785	0.889	0.949	0.979	0.992	0.997
4.1	0.017	0.085	0.224	0.414	0.609	0.769	0.879	0.943	0.976	0.990	0.997
4.2	0.015	0.078	0.210	0.395	0.590	0.753	0.867	0.936	0.972	0.989	0.996
4.3	0.014	0.072	0.197	0.377	0.570	0.737	0.856	0.929	0.968	0.987	0.995
4.4	0.012	0.066	0.185	0.359	0.551	0.720	0.844	0.921	0.964	0.985	0.994
4.5	0.011	0.061	0.174	0.342	0.532	0.703	0.831	0.913	0.960	0.983	0.993
4.6	0.010	0.056	0.163	0.326	0.513	0.686	0.818	0.905	0.955	0.980	0.992
4.7	0.009	0.052	0.152	0.310	0.495	0.668	0.805	0.896	0.950	0.978	0.991
4.8	0.008	0.048	0.143	0.294	0.476	0.651	0.791	0.887	0.944	0.975	0.990
4.9	0.007	0.044	0.133	0.279	0.458	0.634	0.777	0.877	0.938	0.972	0.988
5.0	0.007	0.040	0.125	0.265	0.440	0.616	0.762	0.867	0.932	0.968	0.986

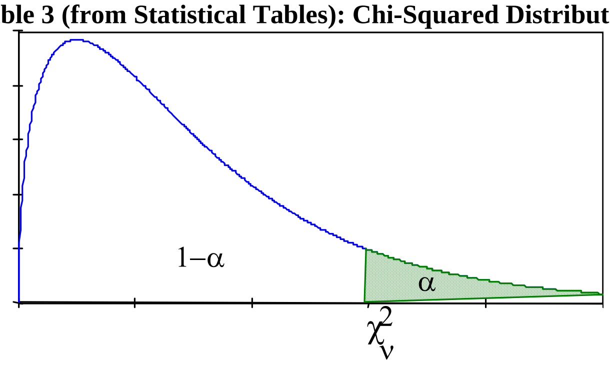
Table 1 (from Statistical Tables): Normal Distribution



$$\Pr(Z \le z) = F(z)$$

z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)
0.00	0.500	0.50	0.691	1.00	0.841	1.50	0.933	2.00	0.977	2.50	0.994	3.00	0.999
0.01	0.504	0.51	0.695	1.01	0.844	1.51	0.934	2.01	0.978	2.51	0.994	3.01	0.999
0.02	0.508	0.52	0.698	1.02	0.846	1.52	0.936	2.02	0.978	2.52	0.994	3.02	0.999
0.03	0.512	0.53	0.702	1.03	0.848	1.53	0.937	2.03	0.979	2.53	0.994	3.03	0.999
0.04	0.516	0.54	0.705	1.04	0.851	1.54	0.938	2.04	0.979	2.54	0.994	3.04	0.999
0.05	0.520	0.55	0.709	1.05	0.853	1.55	0.939	2.05	0.980	2.55	0.995	3.05	0.999
0.06	0.524	0.56	0.712	1.06	0.855	1.56	0.941	2.06	0.980	2.56	0.995	3.06	0.999
0.07	0.528	0.57	0.716	1.07	0.858	1.57	0.942	2.07	0.981	2.57	0.995	3.07	0.999
0.08	0.532	0.58	0.719	1.08	0.860	1.58	0.943	2.08	0.981	2.58	0.995	3.08	0.999
0.09	0.536	0.59	0.722	1.09	0.862	1.59	0.944	2.09	0.982	2.59	0.995	3.09	0.999
0.10	0.540	0.60	0.726	1.10	0.864	1.60	0.945	2.10	0.982	2.60	0.995	3.10	0.999
0.11	0.544	0.61	0.729	1.11	0.867	1.61	0.946	2.11	0.983	2.61	0.995	3.11	0.999
0.12	0.548	0.62	0.732	1.12	0.869	1.62	0.947	2.12	0.983	2.62	0.996	3.12	0.999
0.13	0.552	0.63	0.736	1.13	0.871	1.63	0.948	2.13	0.983	2.63	0.996	3.13	0.999
0.14	0.556	0.64	0.739	1.14	0.873	1.64	0.949	2.14	0.984	2.64	0.996	3.14	0.999
0.15	0.560	0.65	0.742	1.15	0.875	1.65	0.951	2.15	0.984	2.65	0.996	3.15	0.999
0.16	0.564	0.66	0.745	1.16	0.877	1.66	0.952	2.16	0.985	2.66	0.996	3.16	0.999
0.17	0.567	0.67	0.749	1.17	0.879	1.67	0.953	2.17	0.985	2.67	0.996	3.17	0.999
0.18	0.571	0.68	0.752	1.18	0.881	1.68	0.954	2.18	0.985	2.68	0.996	3.18	0.999
0.19	0.575	0.69	0.755	1.19	0.883	1.69	0.954	2.19	0.986	2.69	0.996	3.19	0.999
0.20	0.579	0.70	0.758	1.20	0.885	1.70	0.955	2.20	0.986	2.70	0.997	3.20	0.999
0.21	0.583	0.71	0.761	1.21	0.887	1.71	0.956	2.21	0.986	2.71	0.997	3.21	0.999
0.22	0.587	0.72	0.764	1.22	0.889	1.72	0.957	2.22	0.987	2.72	0.997	3.22	0.999
0.23	0.591	0.73	0.767	1.23	0.891	1.73	0.958	2.23	0.987	2.73	0.997	3.23	0.999
0.24	0.595	0.74	0.770	1.24	0.893	1.74	0.959	2.24	0.987	2.74	0.997	3.24	0.999
0.25	0.599	0.75	0.773	1.25	0.894	1.75	0.960	2.25	0.988	2.75	0.997	3.25	0.999
0.26	0.603	0.76	0.776	1.26	0.896	1.76	0.961	2.26	0.988	2.76	0.997	3.26	0.999
0.27	0.606	0.77	0.779	1.27	0.898	1.77	0.962	2.27	0.988	2.77	0.997	3.27	0.999
0.28	0.610	0.78	0.782	1.28	0.900	1.78	0.962	2.28	0.989	2.78	0.997	3.28	0.999
0.29	0.614	0.79	0.785	1.29	0.901	1.79	0.963	2.29	0.989	2.79	0.997	3.29	0.999
0.30	0.618	0.80	0.788	1.30	0.903	1.80	0.964	2.30	0.989	2.80	0.997	3.30	1.000
0.31	0.622	0.81	0.791	1.31	0.905	1.81	0.965	2.31	0.990	2.81	0.998	3.31	1.000
0.32	0.626	0.82	0.794	1.32	0.907	1.82	0.966	2.32	0.990	2.82	0.998	3.32	1.000
0.33	0.629	0.83	0.797	1.33	0.908	1.83	0.966	2.33	0.990	2.83	0.998	3.33	1.000
0.34	0.633	0.84	0.800	1.34	0.910	1.84	0.967	2.34	0.990	2.84	0.998	3.34	1.000
0.35	0.637	0.85	0.802	1.35	0.911	1.85	0.968	2.35	0.991	2.85	0.998	3.35	1.000
0.36	0.641	0.86	0.805	1.36	0.913	1.86	0.969	2.36	0.991	2.86	0.998	3.36	1.000
0.37	0.644	0.87	0.808	1.37	0.915	1.87	0.969	2.37	0.991	2.87	0.998	3.37	1.000
0.38	0.648	0.88	0.811	1.38	0.916	1.88	0.970	2.38	0.991	2.88	0.998	3.38	1.000
0.39	0.652	0.89	0.813	1.39	0.918	1.89	0.971	2.39	0.992	2.89	0.998	3.39	1.000
0.40	0.655	0.90	0.816	1.40	0.919	1.90	0.971	2.40	0.992	2.90	0.998	3.40	1.000
0.41	0.659	0.91	0.819	1.41	0.921	1.91	0.972	2.41	0.992	2.91	0.998	3.41	1.000
0.42	0.663	0.92	0.821	1.42	0.922	1.92	0.973	2.42	0.992	2.92	0.998	3.42	1.000
0.43	0.666	0.93	0.824	1.43	0.924	1.93	0.973	2.43	0.992	2.93	0.998	3.43	1.000
0.44	0.670	0.94	0.826	1.44	0.925	1.94	0.974	2.44	0.993	2.94	0.998	3.44	1.000
0.45	0.674	0.95	0.829	1.45	0.926	1.95	0.974	2.45	0.993	2.95	0.998	3.45	1.000
0.46	0.677	0.96	0.831	1.46	0.928	1.96	0.975	2.46	0.993	2.96	0.998	3.46	1.000
0.47	0.681	0.97	0.834	1.47	0.929	1.97	0.976	2.47	0.993	2.97	0.999	3.47	1.000
0.48	0.684	0.98	0.836	1.48	0.931	1.98	0.976	2.48	0.993	2.98	0.999	3.48	1.000
0.49	0.688	0.99	0.839	1.49	0.932	1.99	0.977	2.49	0.994	2.99	0.999	3.49	1.000

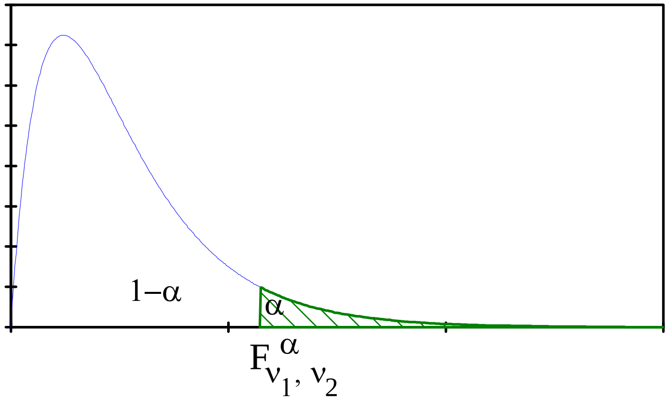
Table 3 (from Statistical Tables): Chi-Squared Distribution



$\Pr(X \geq \chi^2_{\nu,\alpha})$

	α									
ν	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
31	14.46	15.66	17.54	19.28	21.43	41.42	44.99	48.23	52.19	55.00
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49	56.33
33	15.82	17.07	19.05	20.87	23.11	43.75	47.40	50.73	54.78	57.65
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06	58.96
35	17.19	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.21
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17

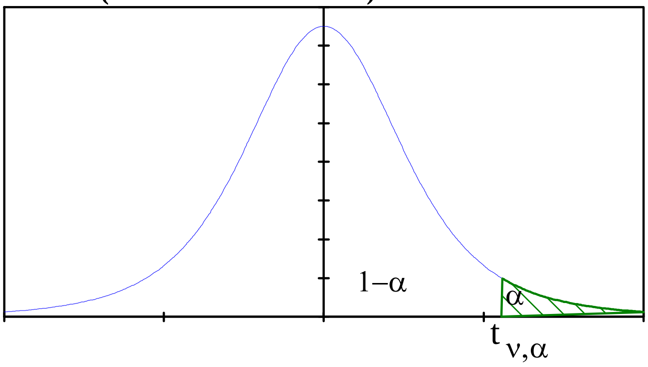
Table 5 (from Statistical Tables): F-distribution ($\alpha = 0.05$)



$$\Pr(X > F^{\alpha}_{v_1,v_2}) = \alpha$$

	v_1									
v_2	1	2	3	4	5	6	7	8	9	10
1	161.55	199.71	215.95	224.84	230.42	234.25	237.04	239.16	240.82	242.16
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.73
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Table 2 (from Statistical Tables): Student t-distribution



$$\Pr(X > t_{\nu,\alpha}) = \alpha$$

	α				
ν	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.327	2.576

Sample Questions

Question 1

A large batch of clay pots is moulded and fired. After firing, a random sample of 10 pots is inspected for flaws before glazing, decoration and final firing. If 20% of pots in the batch have flaws, calculate the probability that the random sample contains:

(a) no pots with flaws,

(b) exactly one pot with a flaw,

(c) exactly two pots with flaws,

(d) less than three pots with flaws.

Question 2

A manufacturer sets up a ‘double sampling’ scheme as follows. A sample of 8 items is taken from a large lot ready for dispatch to customers. If there are no defectives, the lot is accepted and if there are 3 or more defectives, the lot is rejected. If there is either 1 or 2 defectives in the sample, a second sample is taken from the same lot, and the lot is rejected only if there are 3 or more defectives in the 2 samples combined. 12% of items are defective.

(a) What proportion of lots will be accepted using only a single sampling scheme (with 3 or more defectives per sample causing a lot rejection,

(b) What proportion of lots will be accepted using the ‘double sampling’ scheme.

Question 3

A large batch of items is known to have a proportion 0.03 defective. If a sample of 200 is taken, what is the probability that the sample will contain:

(a) no defectives,

(b) 4 defectives or less,

(c) more than 5 defectives.

Question 4

A random variable, Y is N(3,16). Find the probability that a value of Y taken at random will be negative. If 20 values are taken randomly, what is the probability that at least 3 have negative values?

Question 5

As a result of tests on electric light bulbs, it was found that the lifetime of a particular make of bulb was distributed normally, with a mean of 2040 hours and standard deviation of 60 hours. What proportion of bulbs can be expected to burn:

(a) For more than 2150 hours,

(b) for more than 1960 hours?

Question 6

If the random variables X_1, X_2, and X_3 are distributed as χ_1^2, χ_5^2 and χ_{10}^2, respectively, find the distribution of:

(a) $X_1 + X_2$,

(b) $X_1 + X_3$.

Question 7

Use the chi-squared tables such that:

(a) $\Pr(\chi_9^2 > 19.02) = p$,

(b) $\Pr(\chi_{40}^2 > 24.43) = p$,

(c) $\Pr(\chi_{29}^2 > x) = 0.005$,

(d) $\Pr(\chi_4^2 > x) = 0.99$

Question 8

Use the F tables such that:

(a) $\Pr(F_{5,7} > 7.46) = p$,

(b) $\Pr(F_{1,60} > 2.79) = p$,

(c) $\Pr(F_{10,1} > x) = 0.10$,

(d) $\Pr(F_{15,20} > x) = 0.05$

Question 9

If $X \sim B(3, 0.667)$ and $Y \sim P(1)$, find:

(a) $\Pr(X + Y = 4)$,

(b) $\Pr(X + Y \le 2)$

Sample Questions (with Answers)

Question 1

A large batch of clay pots is moulded and fired. After firing, a random sample of 10 pots is inspected for flaws before glazing, decoration and final firing. If 20% of pots in the batch have flaws, calculate the probability that the random sample contains:

(a) no pots with flaws,

(b) exactly one pot with a flaw,

(c) exactly two pots with flaws,

(d) less than three pots with flaws.

Answer

$X \sim B(10, 0.2)$

(a) $\Pr(X = 0) = 0.2^0 0.8^{10} = 0.107$,

(b) $\Pr(X = 1) = 10(0.2^1)0.8^9 = 0.268$

(c) $\Pr(X = 2) = 45(0.2^2)0.8^8 = 0.302$

(d) $\Pr(X \le 2) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) = 0.678$ (from Statistical Tables)

Question 2

A manufacturer sets up a ‘double sampling’ scheme as follows. A sample of 8 items is taken from a large lot ready for dispatch to customers. If there are no defectives, the lot is accepted and if there are 3 or more defectives, the lot is rejected. If there is either 1 or 2 defectives in the sample, a second sample is taken from the same lot, and the lot is rejected only if there are 3 or more defectives in the 2 samples combined. 12% of items are defective.

(a) What proportion of lots will be accepted using only a single sampling scheme (with 3 or more defectives per sample causing a lot rejection,

(b) What proportion of lots will be accepted using the ‘double sampling’ scheme.

Answer

(a) $\Pr(X < 2) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2)$

As $X \sim B(8, 0.12)$, then $\Pr(X \leq 2) = 0.939$.

(b) Proportion of rejections is

$$\Pr(X_1 \geq 3) + \Pr(X_1 = 1).\Pr(X_2 \geq 2) + \Pr(X_1 = 2).\Pr(X_2 \geq 1)$$

$$\Pr(X_1 \geq 3) + \Pr(X_1 = 1).[1 - \Pr(X_2 \leq 1)] + \Pr(X_1 = 2).[1 - \Pr(X_2 = 0)]$$

$$0.061 + 0.392(0.248) + 0.187(0.64) = 0.278.$$

Therefore proportion of acceptances=0.722.

Question 3

A large batch of items is known to have a proportion 0.03 defective. If a sample of 200 is taken, what is the probability that the sample will contain:

(a) no defectives,

(b) 4 defectives or less,

(c) more than 5 defectives.

Answer

Poisson approximation to binomial distribution.

$X \sim B(200, 0.03) \approx X \sim P(200(0.03))$

(a) $\Pr(X = 0) = \dfrac{6^0}{0!} e^6 = 0.0025$

(b) $\Pr(X \leq 4) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) + \Pr(X = 3) + \Pr(X = 4)$

$$e^{-6} + \frac{6^1}{1!} e^{-6} + \frac{6^2}{2!} e^{-6} + \frac{6^3}{3!} e^{-6} + \frac{6^4}{4!} e^{-6} = 0.0025 + 0.0149 + 0.0446 + 0.0892 + 0.1339 = 0.285$$

Alternatively, from Statistical Tables $\lambda = 6$ and *k*=4.

(c)

$$\Pr(X > 5) = 1 - \Pr(X \leq 5) = 1 - [\Pr(X \leq 4) + \Pr(X = 5)] = 1 - [0.285 + 0.161] = 1 - 0.446 = 0.554$$

(also from Statistical Tables).

Question 4

A random variable, Y is N(3,16):

(a)Find the probability that a value of Y taken at random will be negative.

(b)If 20 values are taken randomly, what is the probability that at least 3 have negative values?

Answer

(a) $\Pr(Y<0)=\Pr\left(\frac{Y-3}{4}<\frac{0-3}{4}\right)=\Pr(z<-0.75)=0.227$

(b) $X\sim B(20,0.227)\Rightarrow\Pr(X\geq 3)=1-\Pr(X\leq 2)$

$1-(0.0058+0.0341+0.0951)=0.865$

Question 5

As a result of tests on electric light bulbs, it was found that the lifetime of a particular make of bulb was distributed normally, with a mean of 2040 hours and standard deviation of 60 hours. What proportion of bulbs can be expected to burn:

(a) For more than 2150 hours,

(b) for more than 1960 hours?

Answer

(a) $\Pr(X > 2150) = \Pr\left(\frac{X - 2040}{60} > \frac{2150 - 2040}{60}\right) = \Pr(z > 1.83) = 0.034$

(b) $\Pr(X > 1960) = \Pr\left(\frac{X - 2040}{60} > \frac{1960 - 2040}{60}\right) = \Pr(z > -1.33) = 0.908$

Question 6

Use the chi-squared tables such that:

(a) $\Pr(\chi_9^2 > 19.02) = p$,

(b) $\Pr(\chi_{40}^2 > 24.43) = p$,

(c) $\Pr(\chi_{29}^2 > x) = 0.005$,

(d) $\Pr(\chi_4^2 > x) = 0.99$

Answer

(a) 0.025,

(b) 0.975,

(c) 52.34,

(d) 0.30.

Question 7

Use the F tables such that:

(a) $\Pr(F_{5,7} > 7.46) = p$,

(b) $\Pr(F_{1,60} > 2.79) = p$,

(c) $\Pr(F_{10,1} > x) = 0.10$,

(d) $\Pr(F_{15,20} > x) = 0.05$

Answer

(a) 0.01,

(b) 0.10,

(c) 60.24,

(d) 2.20

Question 8

If $X \sim B(3, 0.667)$ and $Y \sim P(1)$, find:

(a) $\Pr(X+Y=4)$,

(b) $\Pr(X+Y \le 2)$

Answer

X and *Y* are independent

(a) $$\begin{aligned}\Pr(X+Y=4) &= \Pr(X=0).\Pr(Y=4)+\Pr(X=1).\Pr(Y=3)+\Pr(X=2).\Pr(Y=4)\\ &\quad +\Pr(X=3).\Pr(Y=1)\end{aligned}$$

(as Pr(*X*=4)=0).

$$\Pr(X+Y=4)=0.0006+0.0136+0.0818+0.1090=0.2049$$

(b) $$\begin{aligned}\Pr(X+Y\le 2) &= \Pr(X=0).\Pr(Y=0)+\Pr(X=0).\Pr(Y=1)+\Pr(X=0).\Pr(Y=2)\\ &\quad +\Pr(X=1).\Pr(Y=0)+\Pr(X=1).\Pr(Y=1)+\Pr(X=2).\Pr(Y=0)\end{aligned}$$

$$\Pr(X+Y\le 2)=0.0136+0.0136+0.0068+0.0817+0.0817+0.1635=0.3611$$

Binomial Distribution

Binomial expansion of $(x+y)^d$

d											
1										x	y
2									x^2	$2xy$	y^2
3								x^3	$3x^2y$	$3xy^2$	y^3
4							x^4	$4x^3y$	$6x^2y^2$	$4xy^3$	y^4
5						x^5	$5x^4y$	$10x^3y^2$	$10x^2y^3$	$5xy^4$	y^5
6					x^6	$6x^5y$	$15x^4y^2$	$20x^3y^3$	$15x^2y^4$	$6xy^5$	y^6
7				x^7	$7x^6y$	$21x^5y^2$	$35x^4y^3$	$35x^3y^4$	$21x^2y^5$	$7xy^6$	y^7
8			x^8	$8x^7y$	$28x^6y^2$	$56x^5y^3$	$70x^4y^4$	$56x^3y^5$	$28x^2y^6$	$8xy^7$	y^8
9		x^9	$9x^8y$	$36x^7y^2$	$84x^6y^3$	$126x^5y^4$	$126x^4y^5$	$84x^3y^6$	$36x^2y^7$	$9xy^8$	y^9
10	x^{10}	$10x^9y$	$45x^8y^2$	$120x^7y^3$	$210x^6y^4$	$252x^5y^5$	$210x^4y^6$	$120x^3y^7$	$45x^2y^8$	$10xy^9$	y^{10}

Coefficients on the binomial expansion of $(x+y)^d$

d											
1										1	1
2									1	2	1
3								1	3	3	1
4							1	4	6	4	1
5						1	5	10	10	5	1
6					1	6	15	20	15	6	1
7				1	7	21	35	35	21	7	1
8			1	8	28	56	70	56	28	8	1
9		1	9	36	84	126	126	84	36	9	1
10	1	10	45	120	210	252	210	120	45	10	1

Coefficients on the binomial expansion of $(x+y)^d$

d											
1										${}_{1}C_{1}$	${}_{1}C_{0}$
2									${}_{2}C_{2}$	${}_{2}C_{1}$	${}_{2}C_{0}$
3								${}_{3}C_{3}$	${}_{3}C_{2}$	${}_{3}C_{1}$	${}_{3}C_{0}$
4							${}_{4}C_{4}$	${}_{4}C_{3}$	${}_{4}C_{2}$	${}_{4}C_{1}$	${}_{4}C_{0}$
5						${}_{5}C_{5}$	${}_{5}C_{4}$	${}_{5}C_{3}$	${}_{5}C_{2}$	${}_{5}C_{1}$	${}_{5}C_{0}$
6					${}_{6}C_{6}$	${}_{6}C_{5}$	${}_{6}C_{4}$	${}_{6}C_{3}$	${}_{6}C_{2}$	${}_{6}C_{1}$	${}_{6}C_{0}$
7				${}_{7}C_{7}$	${}_{7}C_{6}$	${}_{7}C_{5}$	${}_{7}C_{4}$	${}_{7}C_{3}$	${}_{7}C_{2}$	${}_{7}C_{1}$	${}_{7}C_{0}$
8			${}_{8}C_{8}$	${}_{8}C_{7}$	${}_{8}C_{6}$	${}_{8}C_{5}$	${}_{8}C_{4}$	${}_{8}C_{3}$	${}_{8}C_{2}$	${}_{8}C_{1}$	${}_{8}C_{0}$
9		${}_{9}C_{9}$	${}_{9}C_{8}$	${}_{9}C_{7}$	${}_{9}C_{6}$	${}_{9}C_{5}$	${}_{9}C_{4}$	${}_{9}C_{3}$	${}_{9}C_{2}$	${}_{9}C_{1}$	${}_{9}C_{0}$
10	${}_{10}C_{10}$	${}_{10}C_{9}$	${}_{10}C_{8}$	${}_{10}C_{7}$	${}_{10}C_{6}$	${}_{10}C_{5}$	${}_{10}C_{4}$	${}_{10}C_{3}$	${}_{10}C_{2}$	${}_{10}C_{1}$	${}_{10}C_{0}$

STATISTICAL TECHNIQUES B

Statistics and Properties of Statistics

1. Sample statistics

1.1 Measures of central tendency

Arithmetic (simple) mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Median: Middle value of an ordered set of observations, for an odd number of observations[1] this is: $x_1, x_2, \ldots, x_n$: $x_{0.5(n+1)}$

Mode: Value which occurs most frequently from the set of observations the largest.

1.2 Measures of spread (dispersion)

Variance: expresses how spread out are a set of numbers and is constructed as the average squared deviation around the mean: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

where n-1 are the degrees of freedom i.e. the number of observations of x_i you can freely choose. The square root of the variance is called the **standard deviation**.[2]

Range: is the difference between the largest and smallest observations. Assuming the data has been sorted by size, from smallest to largest: $x_n - x_1$

Interquartile range: measures the difference between the 25th and 75th percentile points: $x_{0.75(n+1)} - x_{0.25(n+1)}$ in ordered data.

Mean absolute deviation: $MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$

[1] For an even number of observations it is $\frac{x_{0.5n} + x_{0.5n+1}}{2}$.

[2] Tchebychev's rule states that for any population with mean μ and standard deviation σ, at least *100(1-1/m²)*% of the population lie within *m* standard deviations around the mean, for *m*>1.

1.3 Skewness

Skewness gives a numerical measure of how asymmetric is a distribution:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{ns^3}$$

A zero value implies a symmetric distribution, a positive number implies a distribution skewed to the right (positively skewed) and a negative number implies a distribution skewed to the left (negatively skewed).

1.4 Kurtosis

Kurtosis gives a measure of how many observations lie in the tails of the distribution:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{ns^4}$$

a value of three implies the distribution has the same proportion of observations in the tails as a normal distribution. A value less than three implies the distribution is platykurtic – meaning the distribution is flat-topped. A value greater than three implies the distribution is leptokurtic – meaning the distribution is more peaked.

1.5. Measures of linear association

In Economics we are interested in the relation between 2 or more random variables, for example:

Personal consumption and disposable income

Investment and interest rates

Earnings and schooling

while there are many ways in which these pairs of random variables might be related – a linear relationship is often a useful first approximation.

1.5.1 Covariance

The association might be STRONG, when a scatter plot, of *Y* against *X*, will be tightly clustered around a straight line, or weak with a scatter plot more widely dispersed about a line. A plot of the data is a necessary preliminary to data analysis, but more sophisticated techniques than a graphical inspection are often required.

The sample covariance between two random variables *X* and *Y* is defined as:

$$s_{XY} = \mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}.\overline{y}}{n-1}$$

the degrees of freedom are only *n*-1 as we actually only need to know the mean of *x* or *y*.

1.5.2 Correlation

The covariance measure is not scale free and multiplying the *x* variable by 100 multiplies the covariance by 100. A scale free measure is a correlation:

$r_{XY} = \frac{s_{XY}}{s_X s_Y}$ where,

$s_X^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 / (n-1)$, $s_Y^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 / (n-1)$, $\overline{x} = \sum_{i=1}^{n} x_i / n$ and $\overline{y} = \sum_{i=1}^{n} y_i / n$.

Where the correlation has the following properties

1. $-1 \le r_{XY} \le 1$
2. $r_{XY} = -1 \Rightarrow$ perfect negative association
3. $r_{XY} = 1 \Rightarrow$ perfect positive linear association
4. $r_{XY} = 0 \Rightarrow$ no linear association
5. As $|r_{XY}|$ increases $\Rightarrow$ stronger association.

2. Estimators and estimates

An *estimator* of a population parameter is a random variable and is a function of the data. Whereas an *estimate* is a particular realisation, or an actual value, based on a specific sample of data points.

3. Unbiasedness

An estimator, θ, is said to be UNBIASED if:

$E(\hat{\theta}) = \theta$, that is, if the mean of the sampling distribution of $\hat{\theta}$ is centred on θ. The ESTIMATORS, $\bar{X}$, s_X^2 and $\hat{p}_X$ are all unbiased:

3.1 Unbiasedness of $\bar{X}$

$$E(\bar{X}) = E(X_1 + X_2 + \ldots + X_n)/n = \frac{1}{n}\left[E(X_1) + E(X_2) + \ldots + E(X_n)\right]$$

$$= \frac{1}{n}\left[\mu + \mu + \ldots + \mu\right] = \mu$$

3.2 Unbiasedness of s_X^2

$$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n}\left[(X_i - \mu) - (\mu - \bar{X})\right]^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2}{n-1}$$

$$E(s_X^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2}{n-1}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right]$$

$$E(s_X^2) = \frac{1}{n-1}\left[n\sigma_X^2 - n\left(\frac{\sigma_X^2}{n}\right)\right] = \sigma_X^2$$

4. Efficiency

Consider two alternative estimators of θ, $\hat{\theta}_1$ and $\hat{\theta}_2$, based on the same information, we say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$, if $V(\hat{\theta}_1) < V(\hat{\theta}_2)$. One possible measure of this is relative efficiency constructed as: $\dfrac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)}$.

In general if we are choosing between two unbiased estimators then we choose the estimator with the smaller variance.

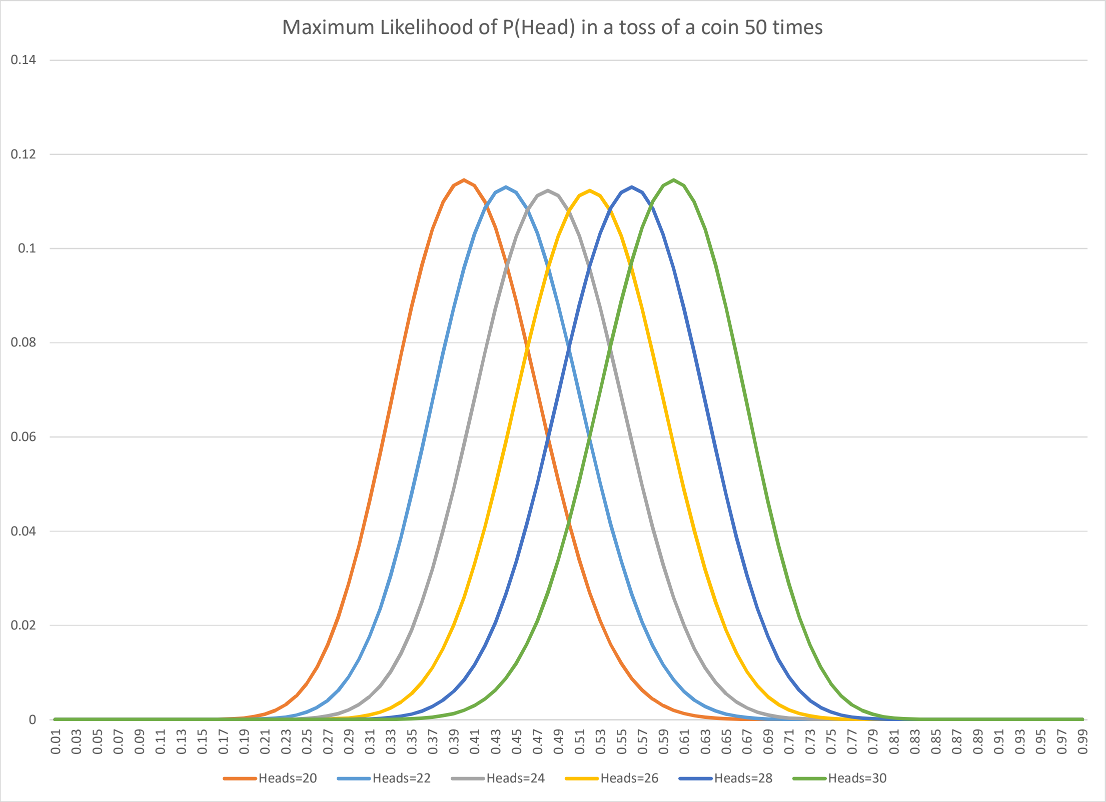
For example, consider 3 alternative estimators for the population parameter, μ,

Estimator	E(.)	V(.)
$\bar{X}_1 = X_1$	μ	σ^2
$\bar{X}_2 = \dfrac{(X_1 + X_2)}{2}$	μ	$\sigma^2 / 2$
$\bar{X}_3 = \dfrac{(X_1 + X_2 + X_3 + \ldots X_n)}{n}$	μ	σ^2 / n
$\bar{X}_4 = 3$	3	0

Last estimator is biased, but with very small variance.

5. Maximum Likelihood Estimation

Consider the toss of a coin with the outcomes of Heads and Tails, with probabilities *p* and (1-*p*). You toss a coin 50 times and get 20 Heads and 30 Tails. Then we know the probability of this happening is $\Pr(Head = 20, Tails = 30) = {}^{50}C_{20}\, p^{20}(1-p)^{30}$. Now think about choosing the value of *p* which maximizes this probability. This is plotted in the figure below as the orange line for all values of *p*. If you choose *p*=0.2 this gives a probability of 0.0006, if you choose a value of *p*=0.3 this gives a probability of 0.0370, a value of p=0.4 gives a joint probability of 0.1146 and *p*=0.5 gives probability of 0.0419 and so p=0.4 maximises the probability and would therefore be the maximum likelihood estimate. If alternatively you toss a coin 50 times and get 28 Heads and 22 Tails. Then we know the probability of this happening is $\Pr(Head = 28, Tails = 22) = {}^{50}C_{28}\, p^{28}(1-p)^{22}$. If you choose *p*=0.3 this gives a probability of 0.0001, a value of *p*=0.5 gives probability of 0.0788 and *p*=0.56 gives a probability of 0.1131 and p=0.6 gives a probability of 0.0959 and so p=0.56 maximises the probability and would therefore be the maximum likelihood estimate.



Algebraically, consider the probability of getting *k* heads out of *n* tosses of a coin. This can be written as:

$$P(k \mid n,p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

So which value of *p* makes this outcome most likely? Consider the maximization problem:

$$\max_p P(k \mid n,p) = \max_p \ln[P(k \mid n,p)]$$

$$\max_p \ln[P(k \mid n,p)] = \max_p [\ln(\frac{n!}{k!(n-k)!}) + k\ln(p) + (n-k)\ln(1-p)]$$

$$\max_p P(k \mid n,p) = \max_p \ln[P(k \mid n,p)]$$

$$\max_p \ln[P(k \mid n,p)] = \max_p [l(p)],\ \text{where } l(p) = c + k\ln(p) + (n-k)\ln(1-p)]$$

$$\frac{\partial l(p)}{p} = \frac{k}{p} - \frac{(n-k)}{(1-p)} = 0 \Rightarrow (n-k)p = k(1-p)$$

$$np - kp = k - kp \Rightarrow \hat{p}_n = k/n$$

and $\operatorname{plim}(\hat{p}_n) = p$ (see below).

6. Consistency

Suppose $\hat{\theta}_n$ is an estimator of θ on a sample of $X_1, X_2, \ldots, X_n$. Then, $\hat{\theta}_n$ is a consistent estimator of θ if for every $\varepsilon > 0$, $P\left(\left|\hat{\theta}_n - \theta\right| \; > \varepsilon\right) \to 0$ as $n \to \infty$. This says that the probability that the absolute difference between $\hat{\theta}_n$ and θ being larger than ε goes to zero as *n* gets bigger. In other words, we say that θ is the probability limit of $\hat{\theta}_n$:

$\text{plim}(\hat{\theta}_n) = \theta$.

7. Central Limit Theorem (CLT)

Many random variables can be characterised as either the sum or the average of a large number of independent random variables. Let, $X_1, X_2, \ldots X_n$, be *n* independent random variables having identical distributions with mean, μ, and variance, σ^2. Denote their sum by $X = X_1 + X_2 + \ldots + X_n$. Now we know that:

$$E(X) = E(X_1 + X_2 + \ldots + X_n) = \underbrace{E(X_1)}_{\mu} + \underbrace{E(X_2)}_{\mu} + \ldots + \underbrace{E(X_n)}_{\mu} = n\mu$$

$$V(X) = V(X_1 + X_2 + \ldots + X_n) = V(X_1) + V(X_2) + \ldots + V(X_n)$$
$$+2\underbrace{\text{cov}(X_1, X_2)}_{0} + 2\underbrace{\text{cov}(X_1, X_3)}_{0} + \ldots + 2\underbrace{\text{cov}(X_{n-1} X_n)}_{0} = n\sigma^2$$

The CLT states, that whatever the distribution of *Xi* (provided that σ^2 is finite) as the number of terms in the sum become large, the distribution of *X* tends to a normal distribution, that is,

$$X \overset{a}{\rightarrow} N(n\mu, n\sigma^2).$$

Therefore the normal distribution will provide a satisfactory approximation to the true distribution for many statistical problems as these involve either sums or averages. This results applies regardless of whether the underlying distribution is continuous and symmetric like the uniform distribution, continuous and asymmetric like the chi-squared distribution, or even discrete such as the Binomial distribution. In fact, Appendix 2: Figures 1-6 show the shape of Binomial distribution when *n*=1, *n*=3, *n*=10, *n*=30 and *n*=100 – it is clear by the last graph (*n*=100) the distribution of the trial is normal. Additionally in Appendix 2: Figure 7 we show a χ^2_d for *d* =8, 12, 16, 20, 30 and one can see the distribution becomes more "normal" as *d* increases. Finally Appendix 2: Figure 8 plots the sample mean ($\bar{X}_n = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$) from the distribution:

$X = x$	1	2	3
$P(X = x)$	0.7	0.2	0.1

Appendix 1 (Question 4 and 5) shows some probability calculations associated with an exact Binomial/Poisson distribution and the approximate probability calculations from a normal distribution.

Sample Questions

Question 1

The data given below is the average weight gain (lbs) between 1986 and 1995 for a sample of 84 males in the US (aged around 30 in 1986). Calculate (a) the sample mean, (b) median, (c) the sample variance

Males

10	0	-18	10	45	10	70	30	33	6	5	25	13	5	30	7	35	25	28	25
30	65	34	20	13	71	10	5	37	25	22	25	40	10	105	65	19	60	29	25
20	28	0	5	15	28	35	40	35	48	13	10	20	45	26	10	2	40	60	58
50	5	43	20	5	46	15	35	-12	25	4	23	0	30	28	21	45	32	20	9
10	12	10	0																

Question 2

The amount of time spent studying on a particular module outside the usual lecture and class hours in an average (typical) week for the sample of 35 students, given in the Table below.

Amount of time spent studying on a particular module

Minutes	No. students
<20	2
20-<40	5
40-<60	4
60-<90	6
90-<120	5
120-<180	7
180-<240	3
240-<360	2
≥360	1
Total	35

Using the data given in the Table above, calculate: (a) sample mean; (b) sample standard deviation; (c) median; (d) the inter-quartile range.

Question 3

Consider the following sample of bivariate data

x	42	24	82	74	70	36	57	29	63	74	80	30
y	136	141	133	135	120	142	130	153	128	112	124	146

(a) Calculate the covariance between *y* and *x*.
(b) Calculate the correlation between *y* and *x*.

Question 4

In each of the following cases work out the exact probability and the probability based on a normal approximation to the exact distribution (using the continuity correction):

(a) For $X \sim B(3, 0.7)$ what is the $\Pr(X \geq 3)$,

(b) For $X \sim B(10, 0.7)$ what is the $\Pr(X \geq 7)$,

(c) For $X \sim B(30, 0.7)$ what is the $\Pr(X \geq 21)$,

(d) For $X \sim B(100, 0.7)$ what is the $\Pr(X \geq 70)$,

(e) For $X \sim B(200, 0.7)$ what is the $\Pr(X \geq 140)$.

Question 5

In each of the following cases work out the exact probability and the probability based on a normal approximation to the exact distribution (using the continuity correction):

(a) For $X \sim P(1)$ what is the $\Pr(X \geq 2)$,
(b) For $X \sim P(4)$ what is the $\Pr(X \geq 5)$,
(c) For $X \sim P(10)$ what is the $\Pr(X \geq 11)$,
(d) For $X \sim P(30)$ what is the $\Pr(X \geq 31)$,
(e) For $X \sim P(100)$ what is the $\Pr(X \geq 101)$.

Sample Questions (with Answers)

Question 1

The data given below is the average weight gain (lbs) between 1986 and 1995 for a sample of 84 males in the US (aged around 30 in 1986). Calculate (i) the sample mean, (ii) median, (iii) the sample variance

Males

10	0	-18	10	45	10	70	30	33	6	5	25	13	5	30	7	35	25	28	25
30	65	34	20	13	71	10	5	37	25	22	25	40	10	105	65	19	60	29	25
20	28	0	5	15	28	35	40	35	48	13	10	20	45	26	10	2	40	60	58
50	5	43	20	5	46	15	35	-12	25	4	23	0	30	28	21	45	32	20	9
10	12	10	0																

Answer

(a) $\bar{x}^M = \frac{10+0-18+10+45+10...+0}{84} = \frac{2118}{84} = 25.21$

(b) median^M=25

(c) $s_M^2 = \frac{10^2+0^2+(-18)^2+10^2+45^2+10^2...+0^2-84\times 25.21^2}{83} = 421.76$

Question 2

The amount of time spent studying on a particular module outside the usual lecture and class hours in an average (typical) week for the sample of 35 students, given in the Table below.

Amount of time spent studying on a particular module

Minutes	No. students
<20	2
20-<40	5
40-<60	4
60-<90	6
90-<120	5
120-<180	7
180-<240	3
240-<360	2
≥360	1
Total	35

Using the data given in the Table, calculate: (a) sample mean, (b) sample standard deviation, (c) median, (d) the inter-quartile range.

Answer

Amount of time spent studying on a particular module

Range	Mid point	f	F	xf	$x^2 f$
<20	10	2	2	20	200
20-40	30	5	7	150	4500
40-60	50	4	11	200	10000
60-90	75	6	17	450	33750
90-120	105	5	22	525	55125
120-180	150	7	29	1050	157500
180-240	210	3	32	630	132300
240-360	300	2	34	600	180000
>360	420	1	35	420	176400
	Total	35		4045	749775

(a) $\bar{x} = \frac{4045}{35} = 115.6$,

(b) $s^2 = \frac{749775 - 35(115.6^2)}{34} = 8302.6 \Rightarrow s = 91.1$,

(c) For the median sometimes people actual calculate the median point in the interval as: $F_{50} = 90 + \left[\frac{18-17}{5}\right]30 = 95$ minutes.

(d) $F_{25} = 40 + \left[\frac{9-7}{4}\right]20 = 50$, $F_{75} = 120 + \left[\frac{27-22}{7}\right]60 = 162.86$

Question 3

Consider the following sample of bivariate data

x	42	24	82	74	70	36	57	29	63	74	80	30
y	136	141	133	135	120	142	130	153	128	112	124	146

(a) Calculate the covariance between *y* and *x*.

(b) Calculate the correlation between *y* and *x*.

Answer

(a) $\operatorname{cov}(x, y) = \dfrac{\sum_{i=1}^{12} x_i y_i - n\overline{xy}}{n-1} = \dfrac{86003 - 12 \times 55.08 \times 133.33}{11} = -193.67$

(b) $corr(x, y) = \dfrac{\operatorname{cov}(x, y)}{s_x s_y} = \dfrac{-193.67}{21.66 \times 11.48} = -0.779$.

Question 4

In each of the following cases work out the exact probability and the probability based on a normal approximation to the exact distribution (using the continuity correction):

(a) For $X \sim B(10, 0.7)$ what is the $\Pr(X \geq 7)$,

(b) For $X \sim B(30, 0.7)$ what is the $\Pr(X \geq 21)$,

(c) For $X \sim B(100, 0.7)$ what is the $\Pr(X \geq 70)$,

(d) For $X \sim B(200, 0.7)$ what is the $\Pr(X \geq 140)$.

Answer

(a) $\Pr(X \geq 7) = 0.65$

$X \overset{a}{\sim} N(7, 2.1)$

$$\Pr(X \geq 7) = \Pr(X \geq 6.5) = \Pr(Z \geq \frac{6.5-7}{1.449}) = \Pr(Z \geq -0.345) = 0.633$$

(b) $\Pr(X \geq 21) = 0.589$

$X \overset{a}{\sim} N(21, 6.3)$

$$\Pr(X \geq 21) = \Pr(X \geq 20.5) = \Pr(Z \geq \frac{20.5-21}{2.510}) = \Pr(Z \geq -0.199) = 0.579$$

(c) $\Pr(X \geq 70) = 0.549$

$X \overset{a}{\sim} N(70, 21)$

$$\Pr(X \geq 70) = \Pr(X \geq 69.5) = \Pr(Z \geq \frac{69.5-70}{4.5826}) = \Pr(Z \geq -0.109) = 0.544$$

(d) $\Pr(X \geq 140) = 0.534$

$X \overset{a}{\sim} N(140, 42)$

$$\Pr(X \geq 140) = \Pr(X \geq 139.5) = \Pr(Z \geq \frac{139.5-140}{6.4807}) = \Pr(Z \geq -0.077) = 0.532$$

Question 5

In each of the following cases work out the exact probability and the probability based on a normal approximation to the exact distribution (using the continuity correction):

(a) For $X \sim P(4)$ what is the $\Pr(X \geq 5)$,

(b) For $X \sim P(10)$ what is the $\Pr(X \geq 11)$,

(c) For $X \sim P(30)$ what is the $\Pr(X \geq 31)$,

(d) For $X \sim P(100)$ what is the $\Pr(X \geq 101)$.

Answer

(a) $\Pr(X \geq 5) = 0.371$

$X \overset{a}{\sim} N(4,4)$

$\Pr(X \geq 5) = \Pr(X \geq 4.5) = \Pr(Z \geq \frac{4.5-4}{2}) = \Pr(Z \geq 0.25) = 0.401$

(b) $\Pr(X \geq 11) = 0.417$

$X \overset{a}{\sim} N(10,10)$

$\Pr(X \geq 11) = \Pr(X \geq 10.5) = \Pr(Z \geq \frac{10.5-10}{3.1623}) = \Pr(Z \geq 0.158) = 0.436$

(c) $\Pr(X \geq 31) = 0.452$

$X \overset{a}{\sim} N(30,30)$

$\Pr(X \geq 31) = \Pr(X \geq 30.5) = \Pr(Z \geq \frac{30.5-30}{5.477}) = \Pr(Z \geq 0.091) = 0.464$

(d) $\Pr(X \geq 101) = 0.473$

$X \overset{a}{\sim} N(100,100)$

$\Pr(X \geq 101) = \Pr(X \geq 100.5) = \Pr(Z \geq \frac{100.5-100}{10}) = \Pr(Z \geq 0.050) = 0.480$

Appendix 2: Binomial versus Normal Approximation

Figure 1: Distribution for a binomial with p=0.7 and n=1

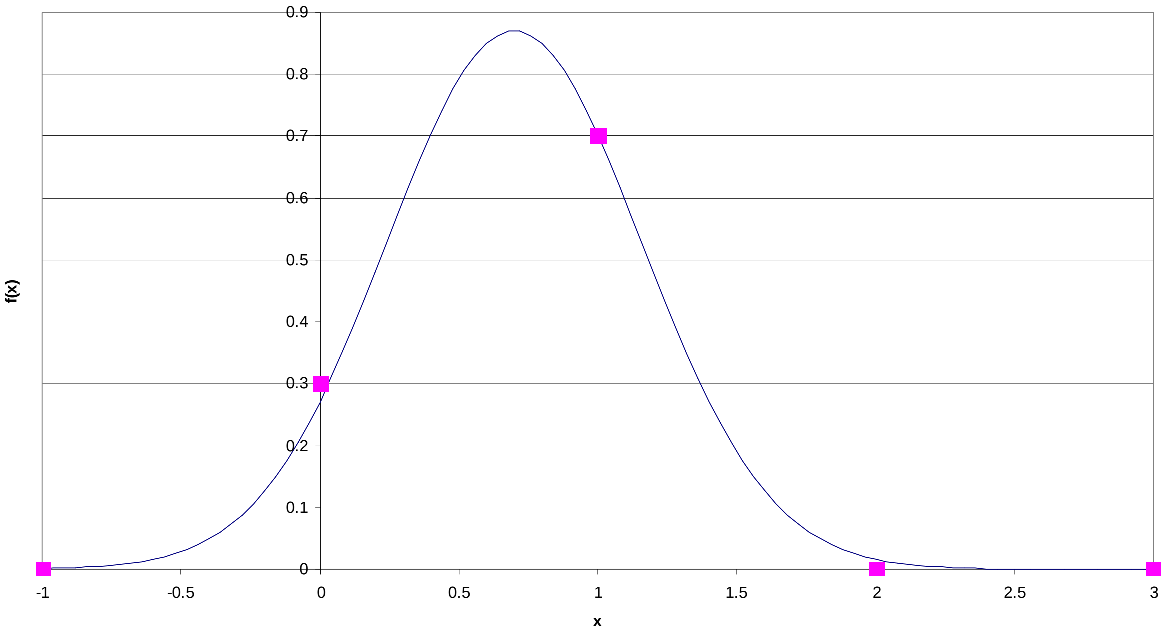


Figure 2: Distribution for a binomial with p=0.7 and n=3

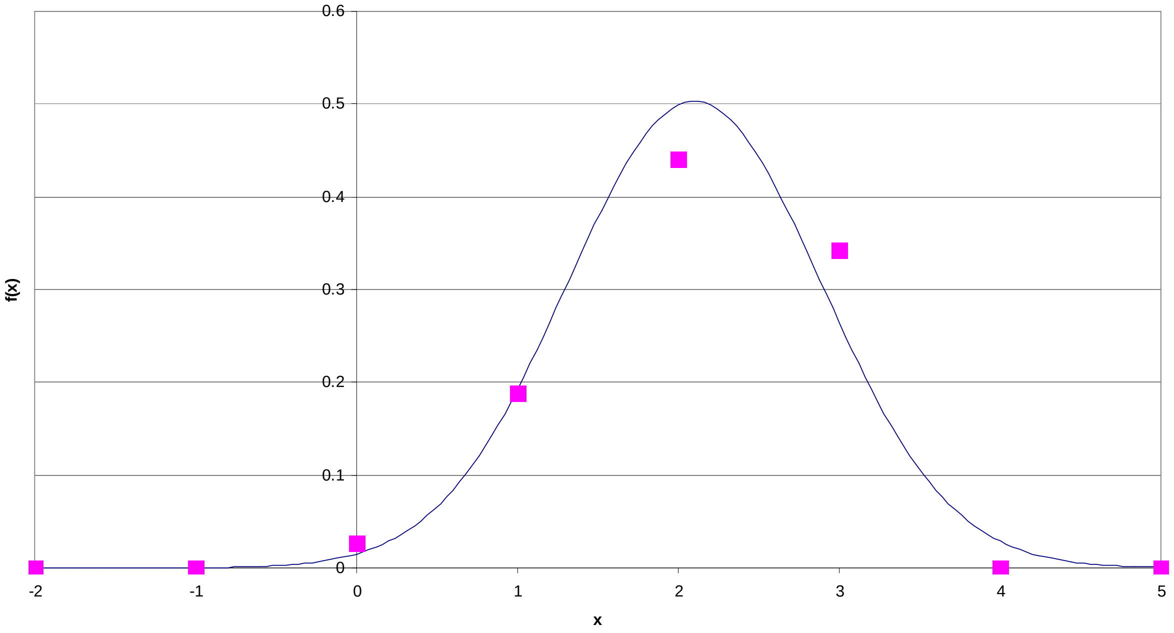


Figure 3: Distribution for a binomial with p=0.7 and n=10

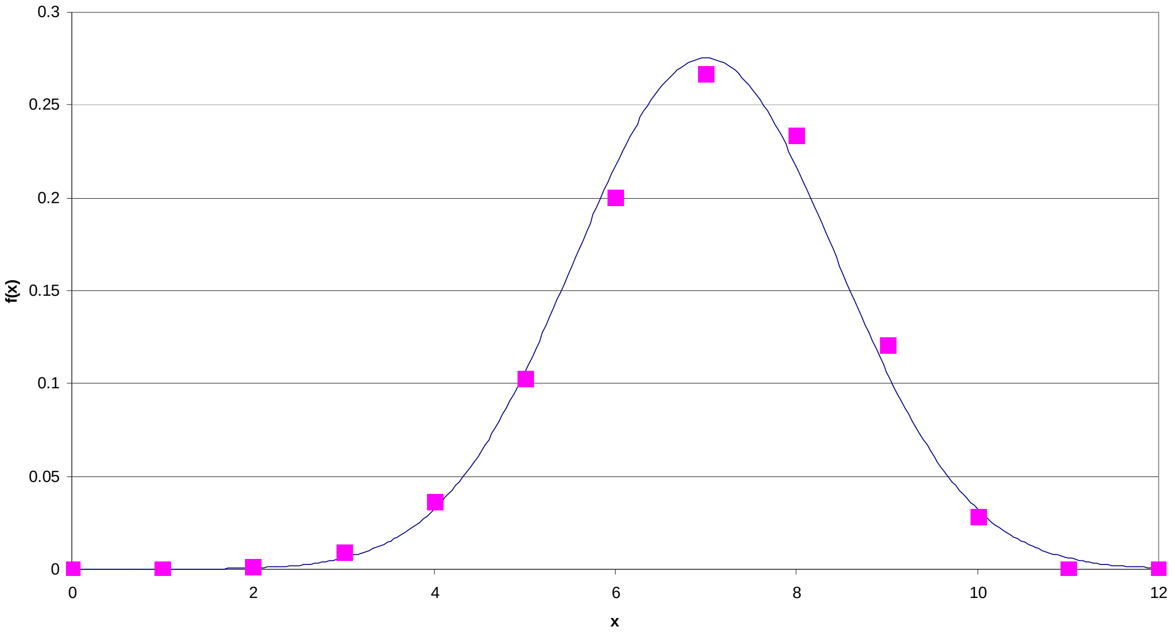


Figure 4: Distribution for a binomial with p=0.7 and n=20

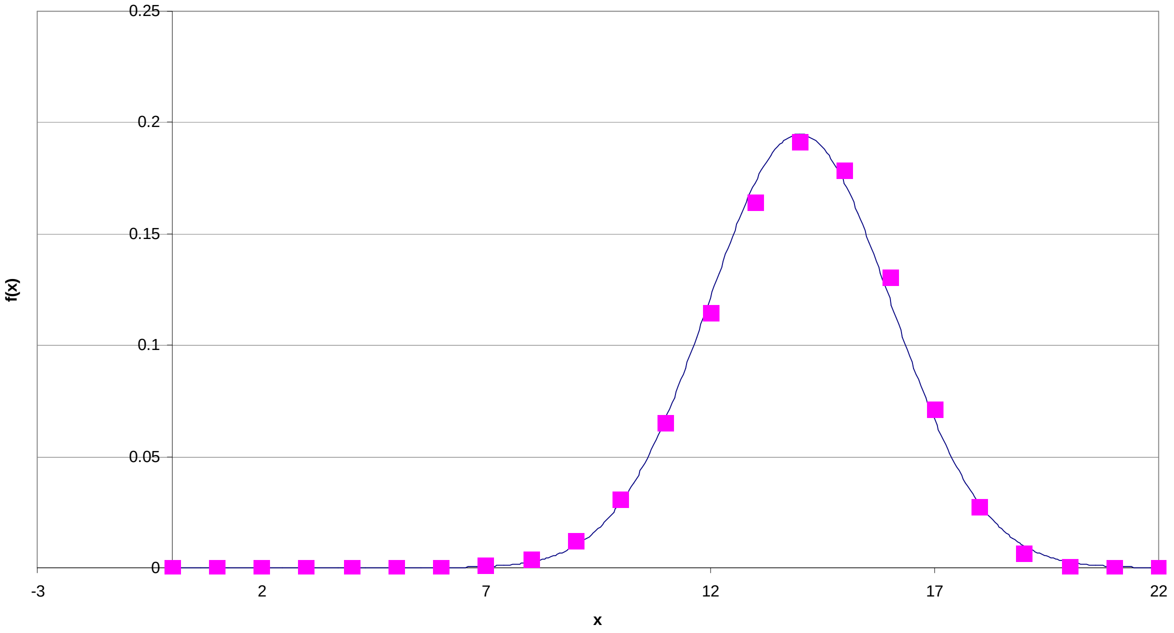


Figure 5: Distribution for a binomial with p=0.7 and n=30

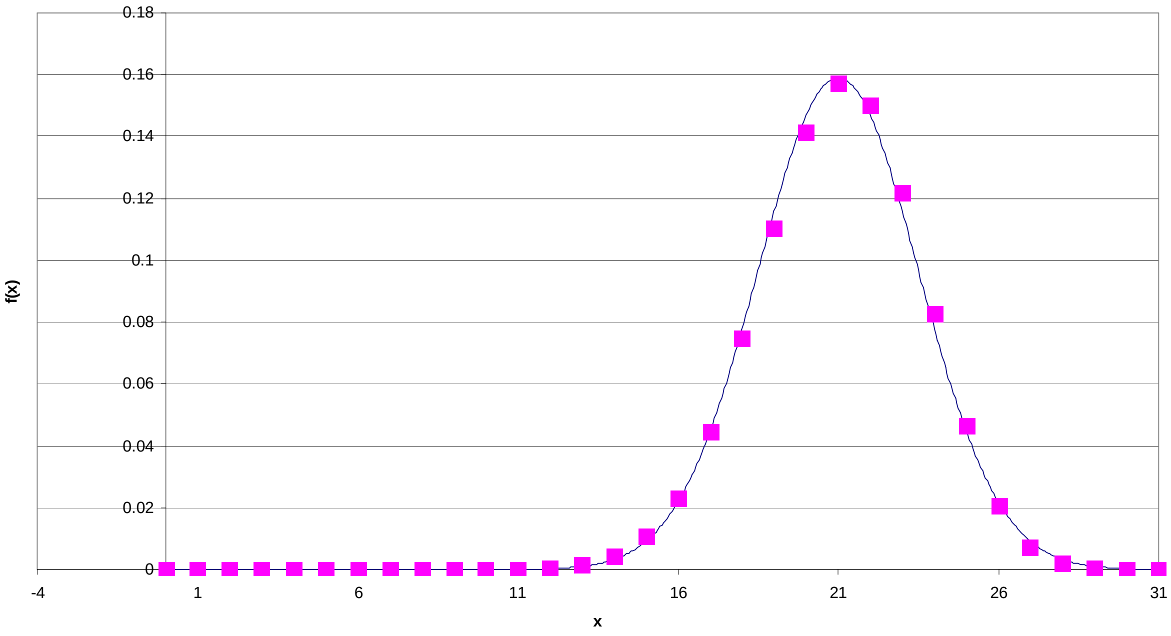
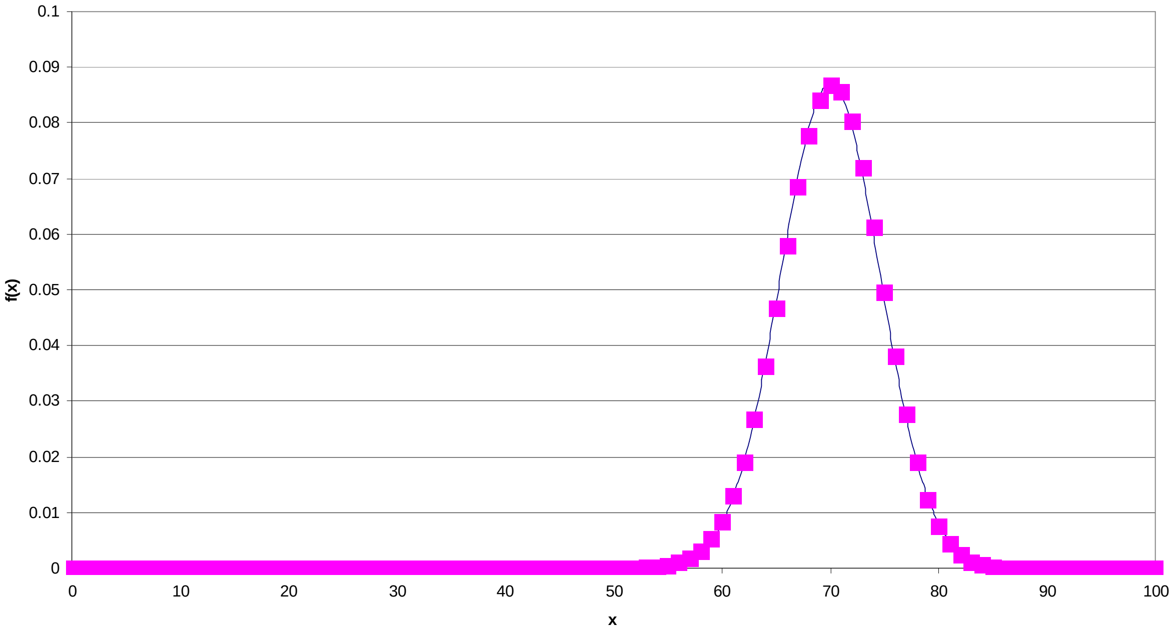


Figure 6: Distribution for a binomial with p=0.7 and n=100



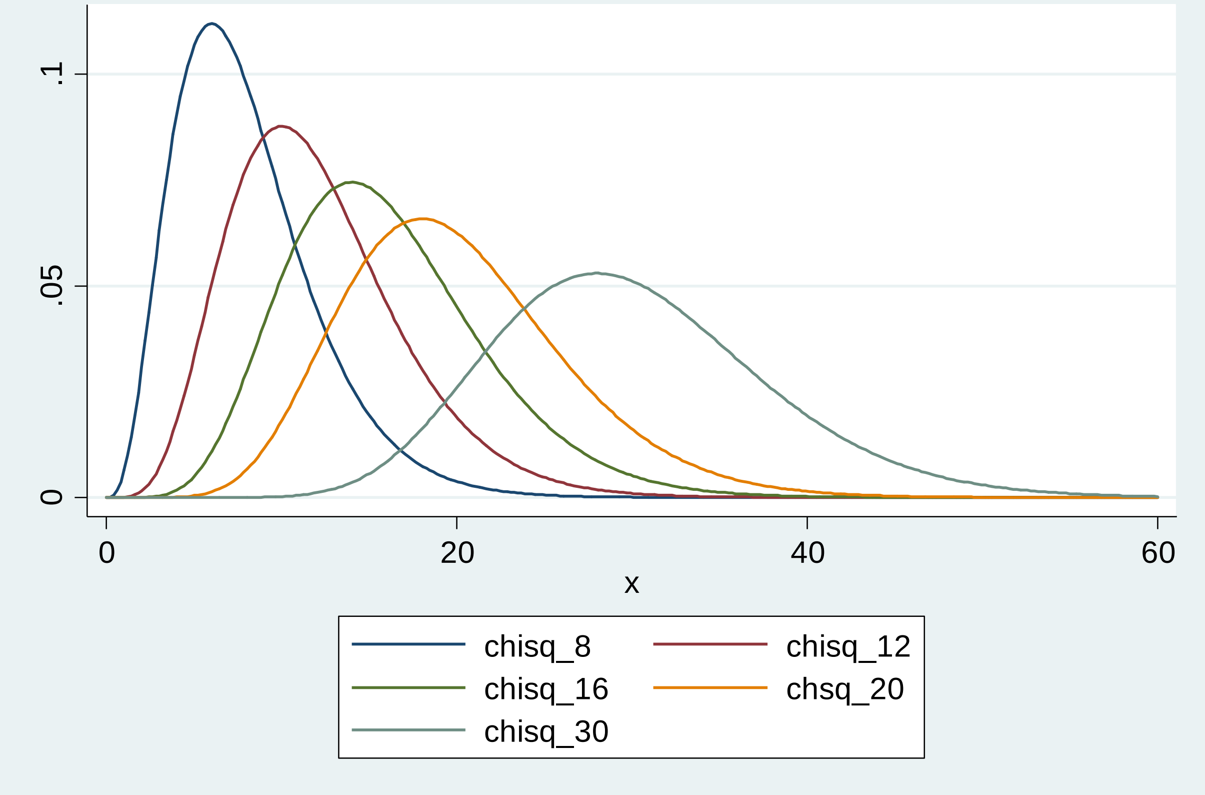


Figure 7: Plot of a series of chi-squared distributions

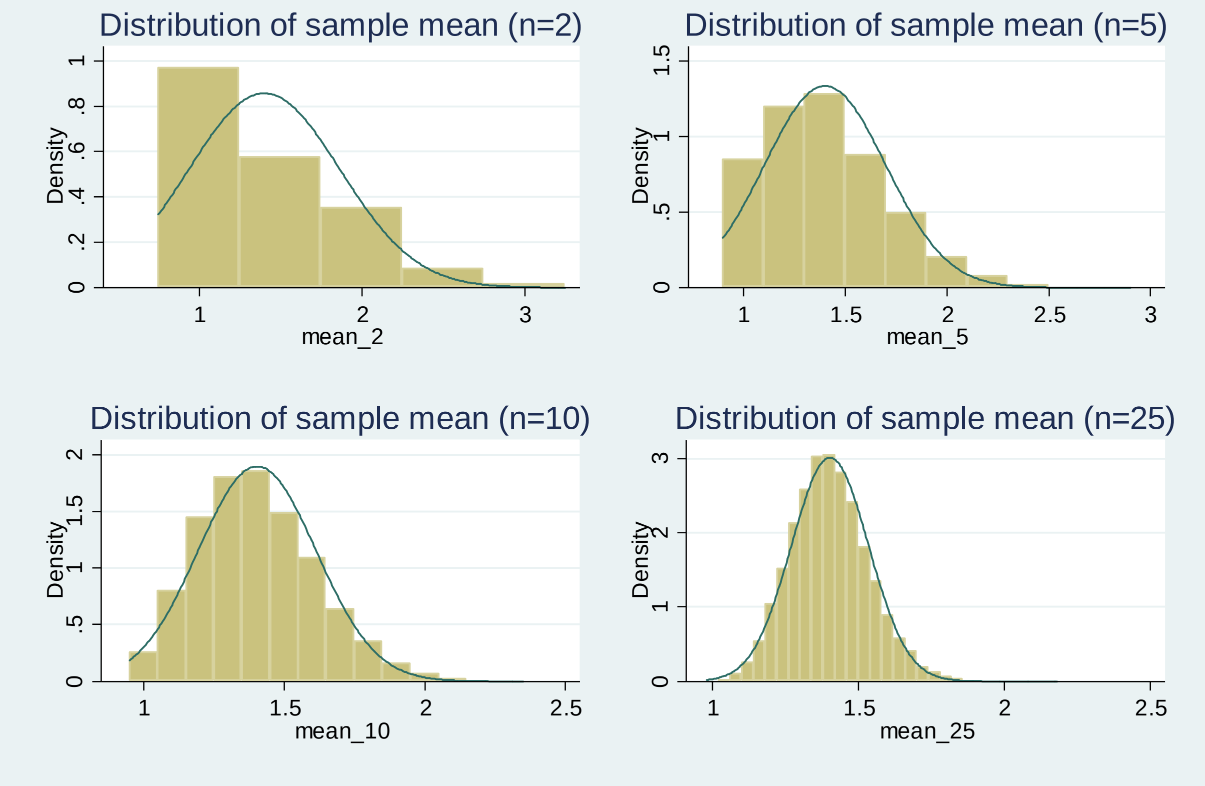


Figure 8

Distribution of sample mean (n=2)

Distribution of sample mean (n=5)

Distribution of sample mean (n=10)

Distribution of sample mean (n=25)

STATISTICAL TECHNIQUES B

Hypothesis Testing

1. Introduction

Hypothesis testing involves assessing the validity of some conjecture or hypothesis. Within statistics some hypothesis is made about some unknown population parameter, θ, This is referred to as the maintained or NULL HYPOTHESIS and is denoted as H_0. If the null hypothesis is NOT true, then some alternative is TRUE. The investigator then formulates as ALTERNATIVE HYPOTHESIS (H_1) against which to test the null hypothesis. This alternative hypothesis is invariably a composite hypothesis (encompassing many values of θ). The null hypothesis is always assumed to be true until counter evidence forces us to reject this working hypothesis.

For example,

$H_0 : \theta = \theta_0$ simple null
$H_1 : \theta \neq \theta_0$ composite 2-sided alternative

$H_0 : \theta \leq \theta_0$ composite null
$H_1 : \theta > \theta_0$ composite 1-sided alternative

$H_0 : \theta \geq \theta_0$ composite null
$H_1 : \theta < \theta_0$ composite 1-sided alternative

Now any null hypothesis can be TRUE or FALSE (as the population parameter, θ, is unknown). Based on the sample evidence we are going to draw conclusions about the population parameters.

2. Types of errors

Needless to say one can clearly make errors when testing a particular hypothesis. In particular, there are two types of errors one can make:

Type I error – Rejecting a TRUE H_0

$\Pr(\text{Type I error}) = \alpha = \text{significance level}$

Type II error – Accept a FALSE H_0

$\Pr(\text{Type II error}) = \beta$

$Power = 1 - \beta = \Pr(\text{Correctly rejecting a FALSE } H_0)$ (see Appendix 1: Figure 1).

Suppose, we believe that that a random variable X is normally distributed with a mean of zero and a variance of unity, that is, $X \sim N(0,1)$ and we wish to test the hypothesis

$$H_0 : \mu = 0$$
$$H_1 : \mu > 0$$

Now if a randomly selected individual had a value of x=1.2. Test the hypothesis that this came from a distribution with a mean of zero.

You proceed by asking the question: What is the probability of observing a number as big as (as small as, for a negative number) the one observed, given $\mu = 0$?

$$\Pr(X \geq 1.2) = \Pr\left(Z \geq \frac{(1.2-0)}{1}\right) = \Pr(Z \geq 1.2) = 0.115$$

so there is an 11.51% chance of observing x>1.2. This probability is known as the p-value, the probability of observing a sample mean as big (or as small) as the one actually observed). However, the question remains:

At what point would you start to question H_0?

The answer depends on the significance level. If you are prepared to only reject H_0 for a p-value of say 0.001 (0.1%), then you really have a low Pr(Type I error) – you must strongly believe in H_0 (naturally conservative). If you are prepared to reject H_0 at say 0.20, then you are prepared to have a high Pr(Type I error) – naturally prepared to overthrow prior beliefs. In statistics the significance level (the probability at which you are prepared to reject H_0) are generally set at $\alpha = 0.01, 0.05, 0.10$, that is, 1%, 5% or 10% and this should be determined before undertaking the test. If we choose to use a significance level of 5%, this implies that you are accepting that 1 time in 20 will incorrectly reject H_0:

Why do we not make Pr(Type I error)≈0.000? Because there is a trade-off between type I and type II errors. So that by choosing a very low type I error probability – that is, minimising the probability of rejecting a true null, you increase the probability of accepting a false null, compare Appendix 1: Figures 1 and 2. Do we regard this as sufficiently rare to reject H_0?

In hypothesis testing, for a given significance level, as we are only interested in the dichotomous decision of either rejecting, or not rejecting, H_0 we do not need to calculate the p-value, but simply calculate test statistic (z=1.2, in the example above) and compare this to a critical value – where the critical value is that value associated with a probability of α (significance level), under the hypothesised distribution.

Table 1: Critical values from a standard normal distribution

a	Pr(X>a)
1.280	0.100
1.645	0.050
1.960	0.025
2.320	0.010
2.575	0.005

In our example as the test statistics of1.2 is less than the critical value of 1.645 (at the 5% significance level) we would not reject H_0.

3. Testing the mean of a normal distribution

Procedure

1. Specify a null hypothesis.
2. Specify an alternative hypothesis.
3. Choose a significance level and corresponding critical region.
4. Calculate the test under the null hypothesis, by calculating how far the sample statistic is from the hypothesised value.
5. Compare the test statistic with the critical value and formulate a decision.

Suppose $X_1, X_2, \ldots, X_n$ denote a random sample of *n* observations from a normal distribution with unknown mean, μ, and known variance, σ^2. Then, $\mathrm{E}(\bar{X}) = \mu$, $V(\bar{X}) = \sigma^2 / n$ and $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. By standardising we have, $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$.

We are interested in testing the hypothesis that the population mean equals μ_0 against an alternative, e.g. $\mu > \mu_0$, the 5-step procedure is:

1. $H_0 : \mu = \mu_0$
2. $H_1 : \mu > \mu_0$. This alternative hypothesis is a 1-sided alternative, implying we reject H_0 only when we observe a sample mean a long way above the hypothesised value, μ_0.
3. We choose some appropriate significance level of α, and find the corresponding critical value from a NORMAL distribution (as distribution of sample mean is normal), denoted z_α - this is the value which occurs with exactly $100\alpha\%$ probability.
4. Under $H_0 : \mu = \mu_0$ then $\bar{X} \sim N(\mu_0, \sigma^2 / n)$, hence $\Pr(\bar{X} > \bar{x}) = \Pr\left(Z > \frac{(\bar{x} - \mu_0)}{\sigma / \sqrt{n}}\right)$.

   In which case we calculate test statistic as $Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$, this essentially measures how far the sample mean is from the hypothesised population mean, μ_0 and scales this distance by the standard error of the sample mean.
5. Then if *Z* is greater than z_α then we have observed an event which occurs with a probability of less than α and should therefore reject H_0. The decision rule is Reject

H0 if $Z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha$ Do not reject H0 if $Z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} < z_\alpha$. (Appendix 1: Figure 3 shows the appropriate acceptance and rejection regions)

If the alternative hypothesis had been $H_1 : \mu < \mu_0$ ($H_1 : \mu \neq \mu_0$), then the corresponding critical-value would have been $-z_\alpha$ ($-z_{\alpha/2}$ and $z_{\alpha/2}$) and the decision rule is: Reject *H*0 if $Z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} < -z_\alpha$ $\left(Z = \left| \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} \right| > z_{\alpha/2} \right)$; Do not reject H0 if $Z = \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} > -z_\alpha$ $\left(Z = \left| \dfrac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} \right| < z_{\alpha/2} \right)$ (Appendix 1: Figures 4 and 5 show the appropriate rejection regions).

3.1 Variants of this basic hypothesis test case:

(1) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of n observations from a distribution which is NOT NORMAL, with an unknown mean, μ, but σ^2 known. Then $\mathrm{E}(\overline{X}) = \mu$, $V(\overline{X}) = \sigma^2 / n$ and if *n*>30 then by a CLT we can say that $\overline{X} \overset{a}{\sim} N(\mu, \sigma^2 / n)$ in which case, the 5 step procedure is as above.

(2) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of *n* observations from a distribution which is NOT NORMAL, with an unknown mean, μ, but σ^2 **unknown** (and a sample variance of s_X^2). Then $\mathrm{E}(\overline{X}) = \mu$, $V(\overline{X}) = \sigma^2 / n$ and if *n*>30 then by a CLT we can say that $\overline{X} \overset{a}{\sim} N(\mu, \mathrm{s}_X^2 / n)$ in which case, the 5 step procedure is as above. (Appendix 2: Example 1).

(3) As a specific example of the case above suppose now that $X_1, X_2, \ldots, X_n$ comes from a Bernoulli distribution, that is,

x	0	1
Pr(*X*=*x*)	1-π	π

$E(X) = \mu = \pi$ and $V(X) = \sigma^2 = \pi(1-\pi)$, with a unknown mean, μ (=π), and an unknown population variance, σ^2(=π (1-π)). Then $\mathrm{E}(\overline{X}) = \pi$, $V(\overline{X}) = \pi(1-\pi)/n$

and if *n*>30 then by a CLT and under the null hypothesis $H_0:\mu=\pi_0$, $\bar{X}\overset{a}{\sim}N(\pi_0,\pi_0(1-\pi_0)/n)$ in which case, the 5 step procedure is as above. (Appendix 2: Example 2).

(4) Suppose now that $X_1,X_2,\ldots,X_n$ denote a random sample of *n* observations from a distribution which is NORMAL, with unknown σ^2 (and sample variance of s_X^2). Then $\mathrm{E}(\bar{X})=\mu$, $V(\bar{X})=\sigma^2/n$ and $\bar{X}\sim N(\mu,\sigma^2/n)$, implying. But we know that $\frac{(n-1)s_X^2}{\sigma^2}\sim\chi_{n-1}^2$. In which case

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s_X^2}{\sigma^2(n-1)}}}\equiv\frac{\bar{X}-\mu}{s_X^2/\sqrt{n}}\sim\frac{N(0,1)}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}}\sim t_{n-1}$$

where $t_{n-1}^{\alpha/2}$ is the critical value from a t-distribution with *n*-1 degrees of freedom. A t-distribution looks similar to a normal distribution (symmetric and bell-shaped); however, this distribution has a higher proportion of points in its tails, see Appendix 1: Figure 6, which compares the distributions of a t_8 with a $N\left(0,1\right)$.

Table 3: Normal compared with t-distributions

	a	Pr(*X*>*a*)	*b*	Pr(*X*>*b*)	*c*	Pr(*X*>*c*)
Normal	2.32	0.010	1.96	0.025	1.645	0.050
t-dist(5)	2.32	0.034	1.96	0.054	1.645	0.080
t-dist(10)	2.32	0.021	1.96	0.039	1.645	0.065
t-dist(15)	2.32	0.017	1.96	0.034	1.645	0.060
t-dist(20)	2.32	0.016	1.96	0.032	1.645	0.058
t-dist(30)	2.32	0.014	1.96	0.030	1.645	0.055
t-dist(50)	2.32	0.012	1.96	0.028	1.645	0.053
t-dist(100)	2.32	0.011	1.96	0.026	1.645	0.052

and so $t_\infty\to N(0,1)$. Many people argue for n>30 the t-distribution can be reasonably well approximated by a standard normal distribution, but this is only an approximation. In which case the 5 step procedure for testing the null and alternative below is:

1. $H_0:\mu=\mu_0$.

2. $H_1:\mu\neq\mu_0$.

3. The critical values come from a t-distribution, denoted $-t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$.

4. The test statistic is $t = \dfrac{\overline{x} - \mu_0}{s_x / \sqrt{n}}$.

5. The decision rule is: Reject *H0* if $t = \left| \dfrac{\overline{x} - \mu_0}{s_x / \sqrt{n}} \right| > t_{\alpha/2,n-1}$; Do not reject *H0* if $t = \left| \dfrac{\overline{x} - \mu_0}{s_x / \sqrt{n}} \right| < t_{\alpha/2,n-1}$. (Appendix 2: Example 3).

4. Test for the difference in means

4.1 Independent samples:

Assume we have two samples of size, n_1 and n_2, on the random variables X_1 and X_2, with unknown means μ_1 and μ_2, respectively. The sample means are denoted as $\bar{X}_1$ and $\bar{X}_2$ and we know that: $E(\bar{X}_1-\bar{X}_2)=\mu_1-\mu_2$ and $V(\bar{X}_1-\bar{X}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$.

If the underlying distributions are NORMAL and the population variances are KNOWN then:

$\bar{X}_1-\bar{X}_2 \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$. In which case the 5-step procedure is:

1. $H_0:\mu_1-\mu_2=D_0$ and so under H_0 $\bar{X}_1-\bar{X}_2 \sim N\left(D_0, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$
2. $H_1:\mu_1-\mu_2\neq D_0$
3. The critical values are $-z_{\alpha/2}$ and $z_{\alpha/2}$.
4. The test statistic is $Z=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$.
5. The decision rule is: Reject *H0* if $Z=\left|\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right|>z_{\alpha/2}$; Do not reject *H0* if

$$Z=\left|\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right|<z_{\alpha/2}.$$

4.1.2 Variants of the difference in means hypothesis test

(1) We have two samples of size, n_1 and n_2, on the random variables X_1 and X_2, with unknown means μ_1 and μ_2, respectively. The sample means are denoted as $\bar{X}_1$ and $\bar{X}_2$ and we know that: $E(\bar{X}_1-\bar{X}_2)=\mu_1-\mu_2$ and $V(\bar{X}_1-\bar{X}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$.

If the underlying distribution of X_1 and X_2 is NOT NORMAL, the population variances (σ_1^2 and σ_2^2) are KNOWN, then providing n_1 and n_2>30, then we can apply a CLT and $\bar{X}_1-\bar{X}_2\overset{a}{\sim}N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$, in which case, the 5 step procedure is as above.

(2) We have two samples of size, n_1 and n_2, on the random variables X_1 and X_2, with unknown means μ_1 and μ_2, respectively. The sample means are denoted as $\bar{X}_1$ and $\bar{X}_2$ and we know that: $E(\bar{X}_1-\bar{X}_2)=\mu_1-\mu_2$ and $V(\bar{X}_1-\bar{X}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$. If the underlying distribution of X_1 and X_2 is NOT NORMAL, the population variances (σ_1^2 and σ_2^2) are UNKNOWN, then providing n_1 and n_2>30, then we can apply a CLT and $\bar{X}_1-\bar{X}_2\overset{a}{\sim}N\left(\mu_1-\mu_2,\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)$, in which case, the 5 step procedure is as above. (Appendix 2: Example 4).

(3) As an example of the above if X_1 and X_2 are both Bernoulli distributions then

$$\bar{X}_1-\bar{X}_2\overset{a}{\sim}N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

1. $H_0:\pi_1-\pi_2=0$ and so under H_0 $\bar{X}_1-\bar{X}_2\overset{a}{\sim}N\left(0,\frac{\pi_0(1-\pi_0)}{n_1}+\frac{\pi_0(1-\pi_0)}{n_2}\right)$, where π_0 is the true overall proportion.
2. $H_1:\pi_1-\pi_2\neq 0$
3. The critical values are $-z_{\alpha/2}$ and $z_{\alpha/2}$.
4. The test statistic is $z=\frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{(p_0(1-p_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$. As under H_0 the population proportions are equal the standard error is based on $p_0=\frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$.

5. The decision rule is: Reject *H0* if $z = \left| \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(p_0(1-p_0)\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \right| > z_{\alpha/2}$; Do not reject *H0* if $z = \left| \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(p_0(1-p_0)\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \right| < z_{\alpha/2}$ (Appendix 2: Example 5).

(4) If the underlying distribution of X_1 and X_2 is NORMAL and the population variances are UNKNOWN, but EQUAL, i.e. $\sigma_1^2 = \sigma_2^2$, then

1. $H_0 : \mu_1 - \mu_2 = D_0$
2. $H_1 : \mu_1 - \mu_2 \neq D_0$
3. The critical values are from a t-distribution, denoted $-t_{\alpha/2,(n_1+n_2-2)}$ and $t_{\alpha/2,(n_1+n_2-2)}$.
4. The test statistic is $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_x\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$, where $s_x^2 = \dfrac{(n_1-1)s_{x_1}^2 + (n_2-1)s_{x_2}^2}{(n_1+n_2-2)}$. Note: To use this test it MUST be the case that there is no evidence that the population variances are different.
5. The decision rule is: Reject *H0* if $t = \left| \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_x\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \right| > t_{\alpha/2,(n_1+n_2-2)}$; Do not reject *H0* if $t = \left| \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_x\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \right| < t_{\alpha/2,(n_1+n_2-2)}$. (Appendix 2: Example 6).

(5) We have two samples of size, n_1 and n_2, on the random variables X_1 and X_2, with unknown means μ_1 and μ_2, respectively. The sample means are denoted as $\overline{X}_1$ and $\overline{X}_2$ and we know that: $E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$ and $V(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

If the underlying distribution of X_1 and X_2 is NORMAL, the population variances (σ_1^2 and σ_2^2) are UNKNOWN and NOT equal, then

1. $H_0 : \mu_1 - \mu_2 = D_0$
2. $H_1 : \mu_1 - \mu_2 \neq D_0$
3. The critical values are from a t-distribution, denoted $-t_{\alpha/2,DoF}$ and $t_{\alpha/2,DoF}$.
4. The test statistic is $t = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}}}$, where $DoF = \frac{\left[s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2 \right]^2}{\frac{\left(s_{x_1}^2 / n_1\right)^2}{(n_1 - 1)} + \frac{\left(s_{x_2}^2 / n_2\right)^2}{(n_2 - 1)}}$.
5. The decision rule is: Reject H_0 if $t = \left| t = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}}} \right| > t_{\alpha/2,DoF}$; Do not reject H_0 if $t = \left| t = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}}} \right| < t_{\alpha/2,DoF}$. (Appendix 2: Example 7).

5. Test of the variance of a distribution

To formulate a hypothesis testing on a sample variance, *X*, MUST be normally distributed, $X_i \sim N(\mu, \sigma^2)$. In which case, $w = \frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$. In which case the 5-step procedure is:

1. $H_0 : \sigma^2 = \sigma_0^2$
2. $H_1 : \sigma^2 > \sigma_0^2$
3. The critical value is from a χ^2-distribution, denoted $\chi_{\alpha,n-1}^2$
4. $\chi^2 = \frac{(n-1)s_X^2}{\sigma_0^2}$
5. The decision rule is: Reject *H₀* if $\chi^2 = \frac{(n-1)s_x^2}{\sigma_0^2} > \chi_{\alpha,n-1}^2$; Do not reject *H₀* if $\chi^2 = \frac{(n-1)s_x^2}{\sigma_0^2} < \chi_{\alpha,n-1}^2$ (Appendix 2: Example 8).

6. Testing equality of variances

This must be done before you can use the 4th option from section 4.1.2 (variants of the difference in means hypothesis test)

1. $H_0 : \sigma_1^2 = \sigma_2^2$
2. $H_1 : \sigma_1^2 \neq \sigma_2^2$
3. The critical value is from the F-distribution, denoted $F^{\alpha/2}_{n_1-1,n_2-1}$.
4. The test statistic as $F = \frac{s_{x_1}^2}{s_{x_2}^2}$, when $s_{x_1}^2 > s_{x_2}^2$ (or $F = \frac{s_{x_2}^2}{s_{x_1}^2}$ when $s_{x_2}^2 > s_{x_1}^2$).
5. The decision rule is: Reject *H0* if $F = \frac{s_{x_1}^2}{s_{x_2}^2} > F^{\alpha/2}_{n_1-1,n_2-1}$; Do not reject *H0* if $F = \frac{s_{x_1}^2}{s_{x_2}^2} < F^{\alpha/2}_{n_1-1,n_2-1}$ (Appendix 2: Example 6).

In Appendix 2, we include a reference table for the different hypothesis test formulas for alternative distributions.

7. Matched pairs

We are interested in formulating tests about $\mu_1 - \mu_2$, when the two experiments (X_1 and X_2) are undertaken with the same sample and are not therefore independent. Given the outcomes of the two trials for the same population we form the difference in the outcomes of the two random variables, that is, $D = X_1 - X_2$, where $E(D) = E(X_1 - X_2) = \mu_1 - \mu_2 \equiv \mu_d$, $V(D) = V(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} \equiv \sigma_d^2$ and $\sigma_{12} > 0$.

If the underlying distribution of $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ then $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_d^2)$ then we know that $\overline{X_1 - X_2} \sim N(\mu_1 - \mu_2, \sigma_d^2 / n)$. Normalising this expression we have: $\dfrac{(\overline{X_1 - X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_d^2 / n}} \sim N(0,1)$, but σ_d^2 is unknown and we need to replace this by the sample variance, in which case $\dfrac{(\overline{X_1 - X_2}) - (\mu_1 - \mu_2)}{\sqrt{s_d^2 / n}} \sim t_{n-1}$.

Given sample of data for the random variable X_1 and X_2, define the difference as: $d_1 = x_1^1 - x_{21}^1$, $d_2 = x_1^2 - x_2^2$, $d_3 = x_1^3 - x_2^3$, $\ldots d_n = x_1^n - x_2^n$

And we calculate the sample moments of this series: $\bar{d} = \sum_{i=1}^{n} d_i / n$ and $s_d^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2 / (n-1)$.

The 5 steps are then:

1. $H_0 : \mu_d \equiv \mu_1 - \mu_2 = D_0$
2. $H_1 : \mu_d \equiv \mu_1 - \mu_2 \neq D_0$
3. The critical values from the t-distribution are $-t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$.
4. The test statistic as $t = \dfrac{\bar{d} - D_0}{s_d / \sqrt{n}}$.
5. The decision rule is : Reject *H0* if $t = \dfrac{\bar{d} - D_0}{s_d / \sqrt{n}} > t_{\alpha/2,n-1}$; Do not reject *H0* if $t = \dfrac{\bar{d} - D_0}{s_d / \sqrt{n}} < t_{\alpha/2,n-1}$ (Appendix 2: Example 9)

8. Calculating the power of a test

$\text{Power} = \Pr(\text{Rejecting } H_0 \mid H_0 \text{ false})$ (see Appendix 1: Figure 1). This is calculated as three steps, for

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

given $\mu = \mu_1$.

(1) Define the critical value as the point at which you just reject *H₀*, for example, $\pm z_{\alpha/2}$?

(2) Find the sample mean, $\bar{x}^c$, corresponding to the critical value, that is,

$$\frac{\bar{x}^c - \mu_0}{s/\sqrt{n}} = \pm z_{\alpha/2} \Rightarrow \bar{x}_1^c = z_{\alpha/2}(s/\sqrt{n}) + \mu_0,\ \bar{x}_1^c = z_{\alpha/2}(s/\sqrt{n}) + \mu_0$$

$$\Rightarrow \bar{x}_2^c = -z_{\alpha/2}(s/\sqrt{n}) + \mu_0$$

(3) Calculate $\Pr(\bar{X} > \bar{x}_1^c \mid \mu = \mu_1) + \Pr(\bar{X} < \bar{x}_2^c \mid \mu = \mu_1)$

$$\Rightarrow \Pr(Z > \frac{\bar{x}_1^c - \mu_1}{s/\sqrt{n}}) + \Pr(Z < \frac{\bar{x}_2^c - \mu_1}{s/\sqrt{n}})$$

Appendix 1: Figure 7-9 show the effect on power as the true mean, μ_1, moves increasingly further away from the null hypothesis, μ_0. Appendix 1: Figure 10 shows that as $\mu_1 \to \mu_0$, then power approaches the significance level, α. (Appendix 2: Example 10).

9. ANOVA

This enables you to test for the equality of means among two or more groups. For two groups we might consider a t-test testing for a difference of means but for 3 groups one would have to do a difference in means between group 1 and group 2, then between group 1 and group 3 and then finally between group 2 and group 3 (i.e. 3 tests). If you had 4 groups then you might have to undertake ${}^4C_2 = 6$ separate tests.

To undertake this analysis we assume:

1. Distribution of the random variables comes from a normal distribution.
2. The variance for each group is the same
3. Random sample

That is we assume we have some random variable (*X*) and we observe it across a categorical variable with *k* outcomes, and we denote the corresponding random variables as $X^1, X^2, \ldots, X^k$. We assume $E(X^j) = \mu^j$ and $V(X^j) = \sigma^2$, and $X^j \sim N(\mu^j, \sigma^2)$.

A random sample from each sub-group is then denoted: $X_1^1, X_2^1, \ldots X_{n_1}^1$ $X_1^2, X_2^2, \ldots X_{n_2}^2$ and $X_1^k, X_2^k, \ldots X_{n_k}^k$ (in that we are allowing a different sample size for each sub-group) and where $n_1 + n_2 + \ldots + n_k = n$.

Now we know that: $\dfrac{X_1^j + X_2^j + \ldots + X_{n_j}^j}{n_j} \equiv \bar{X}^j \sim N(\mu^j, \sigma^2 / n_j)$ and we want to test the hypothesis:

$H_0 : \mu^1 = \mu^2 = \ldots = \mu^k$ against the alternative: $H_0 : \mu^j \neq \mu^l$ for $j \neq l$.

Define $\bar{X} = \dfrac{n_1\bar{X}^1 + n_2\bar{X}^2 + \ldots + n_k\bar{X}^k}{n_1 + n_2 + \ldots + n_k}$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_i^j - \bar{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_i^j - \bar{X}^j)^2 + \sum_{j=1}^{k} n_j(\bar{X}^j - \bar{X})^2$$

Which is telling us that

Total Sum of Squares=*Within Sum of Squares* + *Between Sum of Squares*.

The 5 steps are then:

1. $H_0 : \mu^1 = \mu^2 = \ldots = \mu^k$
2. $H_1 : \mu^j \neq \mu^l$ for $j \neq l$
3. The critical values are from an F-distribution, denoted $F_{k-1,n-k}^{\alpha}$.

4. The test statistic is $F = \dfrac{\textit{Between}\ \text{SS}/(k\text{-}1)}{\textit{Within}\ \text{SS}/(\text{n-}k)}$.

5. The decision rule is: Reject *H₀* if $F = \dfrac{\textit{Between}\ \text{SS}/(k\text{-}1)}{\textit{Within}\ \text{SS}/(\text{n-}k)} > F^{\alpha}_{k-1,n-k}$; Do not

   reject otherwise. (Appendix 2: Example 11).

The test works because under H_0 we know:

$$E\left(Within\ SS\right) =$$

$$E\left[\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_i^j - \bar{X}^j)^2\right] = \sum_{j=1}^{k}\sum_{i=1}^{n_j}E(X_i^j - \bar{X}^j)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}V(X_i^j) = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\sigma^2 = n\sigma^2$$

$$E\left(Between\ SS\right) =$$

$$E\left[\sum_{j=1}^{k}n_j(\bar{X}^j - \bar{X})^2\right] = \sum_{j=1}^{k}n_jE(\bar{X}^j - \bar{X})^2 = \sum_{j=1}^{k}n_jV(\bar{X}^j) = \sum_{j=1}^{k}n_j(\sigma^2/n_j) = k\sigma^2$$

in which case we can say something like:

$$E(F) = E\left(\frac{\textit{Between}\ \text{SS}/(k-1)}{\textit{Within}\ \text{SS}/(n-k)}\right) = \frac{k\sigma^2/(k-1)}{n\sigma^2/(n-k)} \approx 1$$

Whereas under $H_1 : \mu^j \neq \mu^l$ for $j \neq l$ and therefore

$$E(Between\ SS) = E\left[\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}^j - \bar{X})^2\right] = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\mathrm{E}(\bar{X}^j - \bar{X})^2 = k\sigma^2 + c$$

where $c > 0$ and gets bigger the greater the variation in $\bar{X}^j$ across *j*, in which case we can say something like:

$$E(F) = E\left(\frac{\textit{Between}\ \text{SS}/(k-1)}{\textit{Within}\ \text{SS}/(n-k)}\right) = \frac{(k\sigma^2 + c)/(k-1)}{n\sigma^2/(n-k)} >> 1$$

Handout 5: Appendix 1

Figure 1: Pr(Type I error), Pr(Type II error) and Power (α=5%)

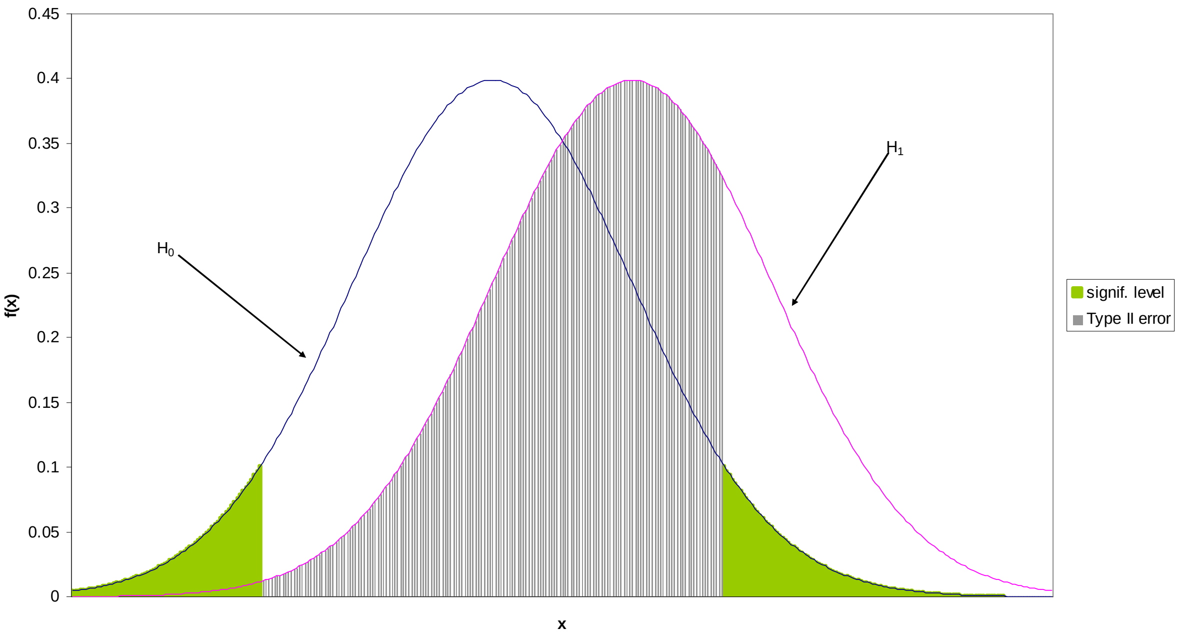


Figure 2: Pr(Type I error), Pr(Type II error) and Power (α=1%)

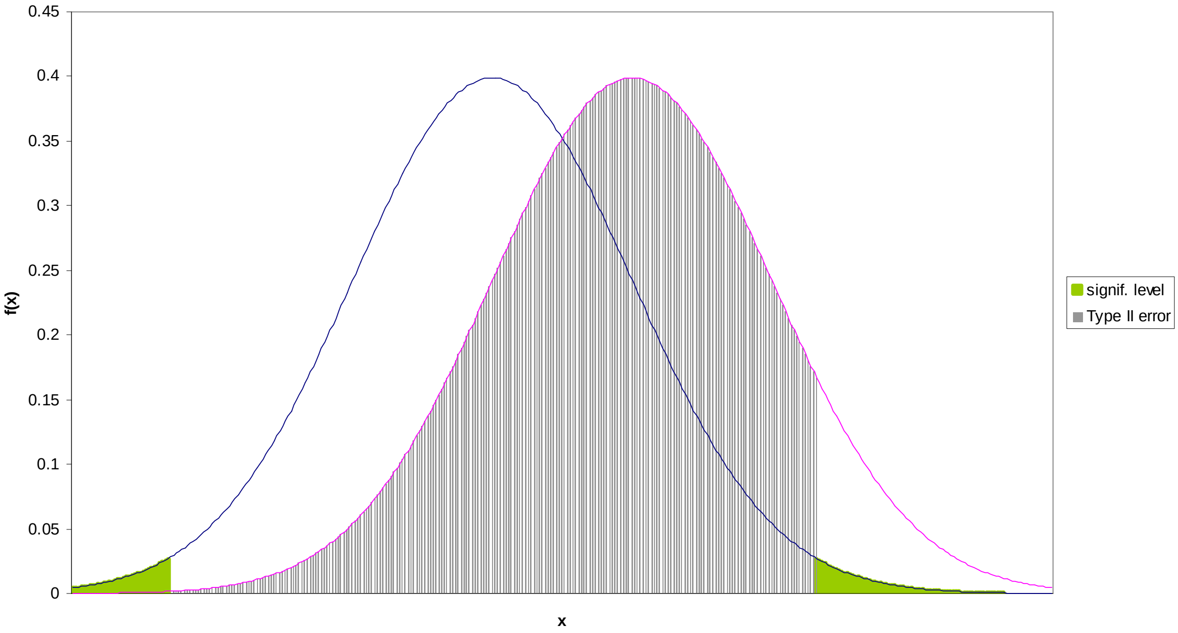


Figure 3: Significance level and critical region for a one-sided alternative H_1:$\mu>\mu_0$

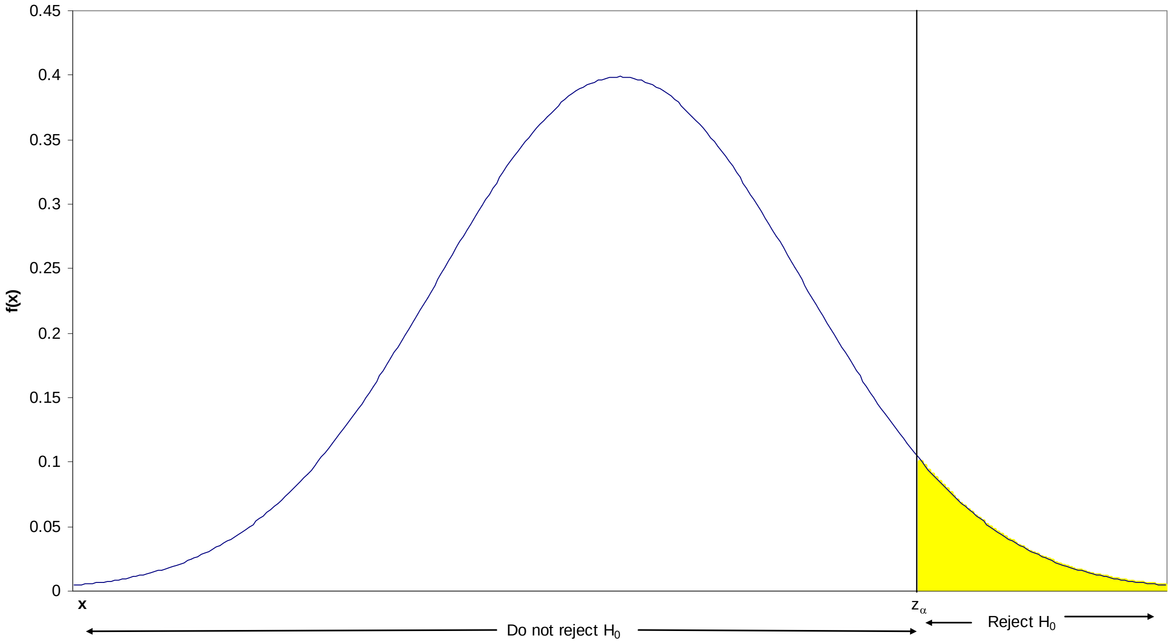
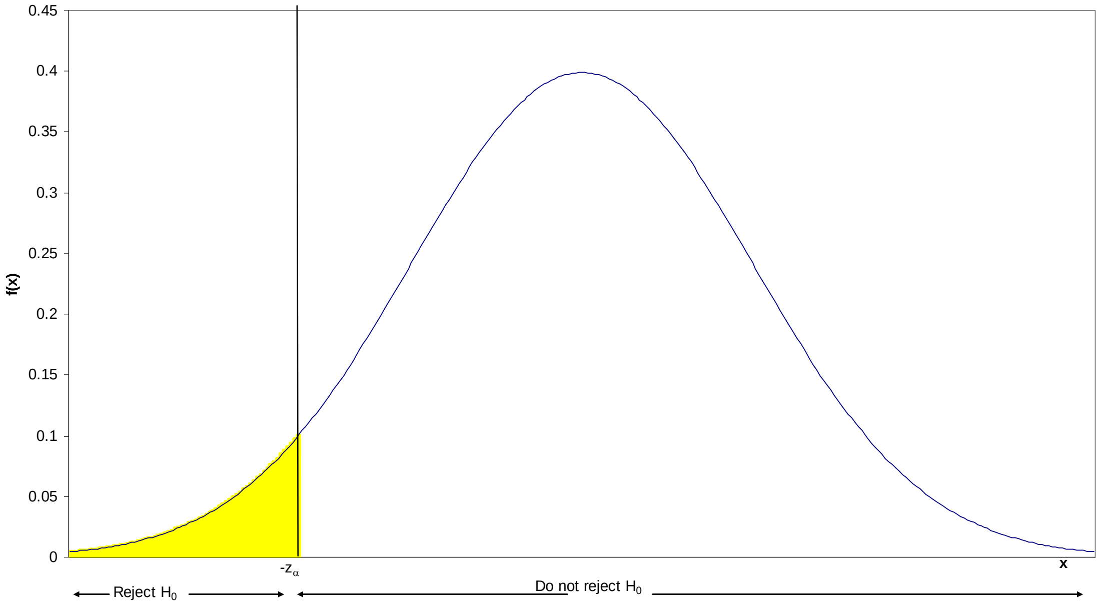


Figure 4: Significance level and critical region for a one-sided alternative H_1 $\mu<\mu_0$



Handout 5: Appendix 1

Figure 5: Significance level and critical regions for a two-sided alternative H_1: $\mu=\mu_0$

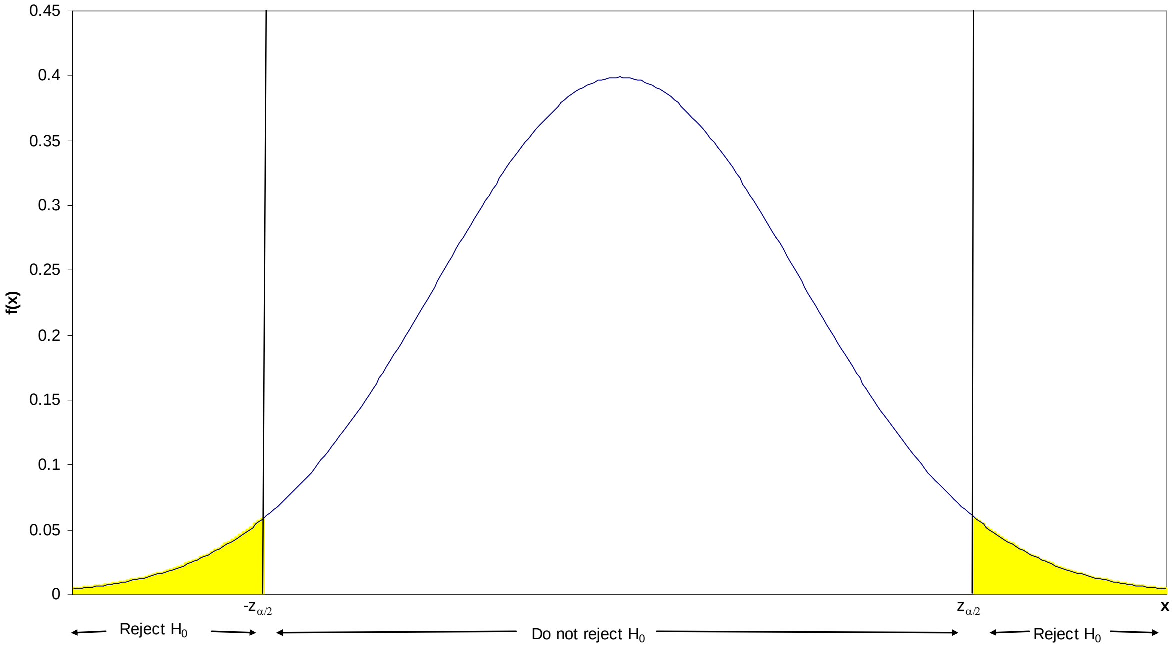


Figure 6: t-distribution vs standard normal distribution

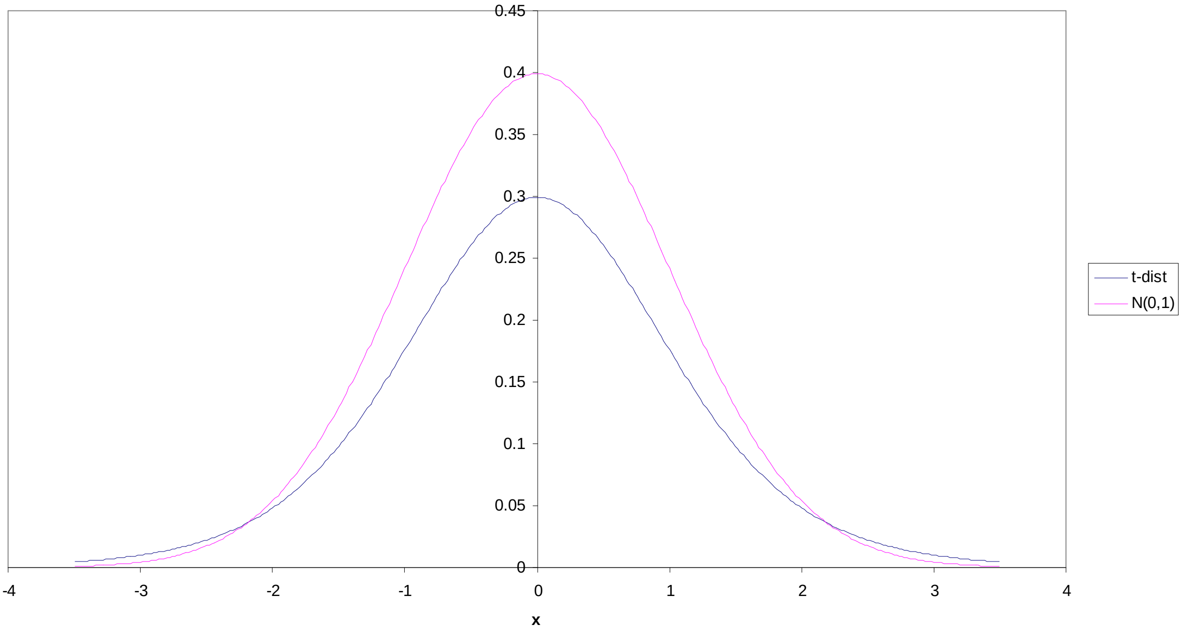


Figure 7: Size and Power: H_0:μ=0 when μ=1.0

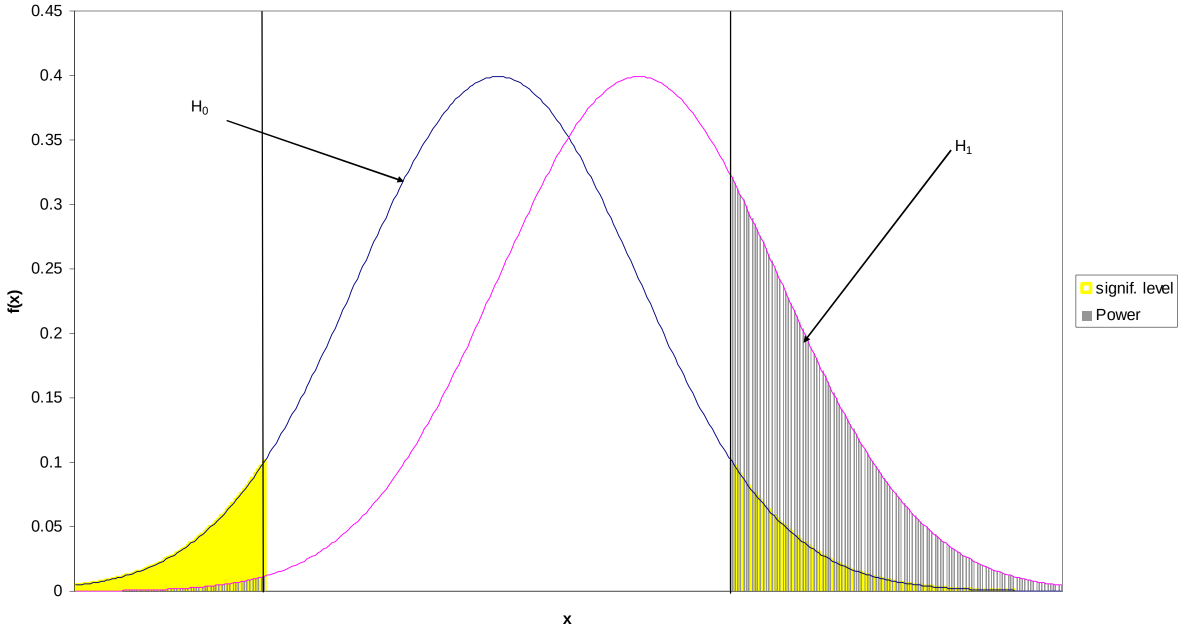


Figure 8: Size and Power: H_0:μ=0 when μ=1.5

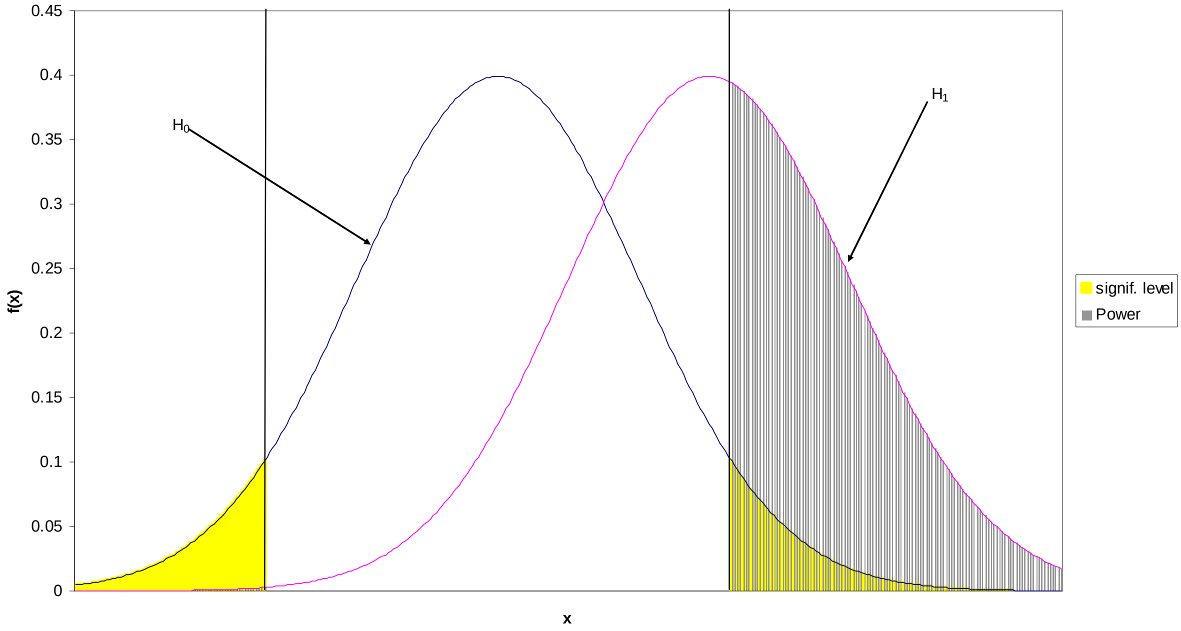


Figure 9: Size and Power: H_0:μ=0 when μ=2.5

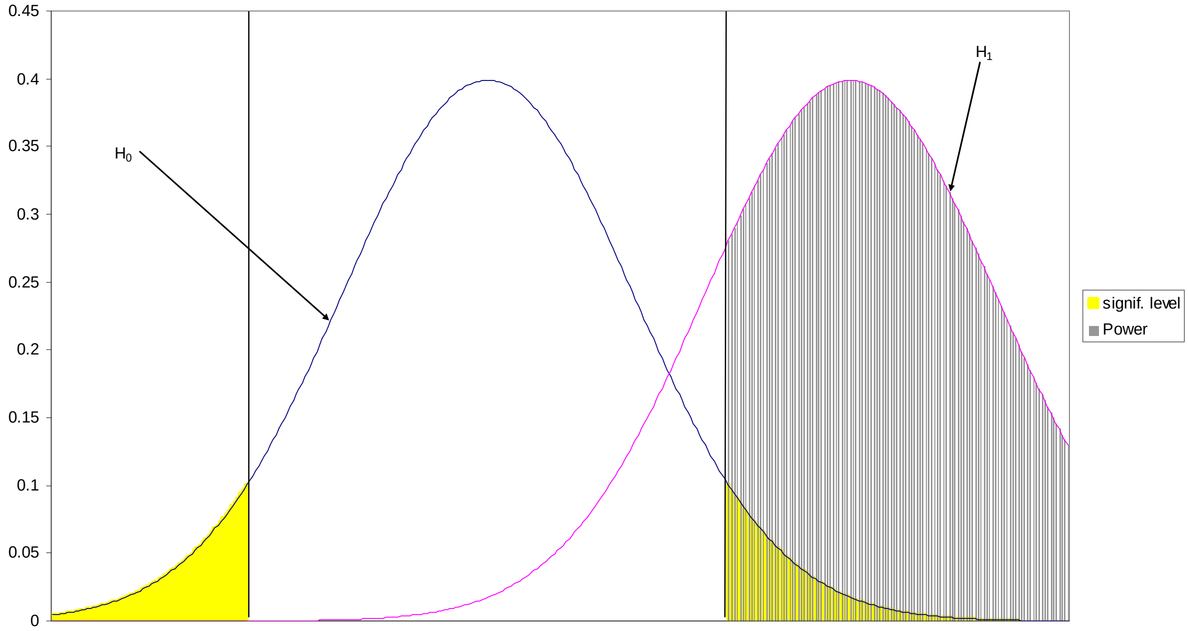
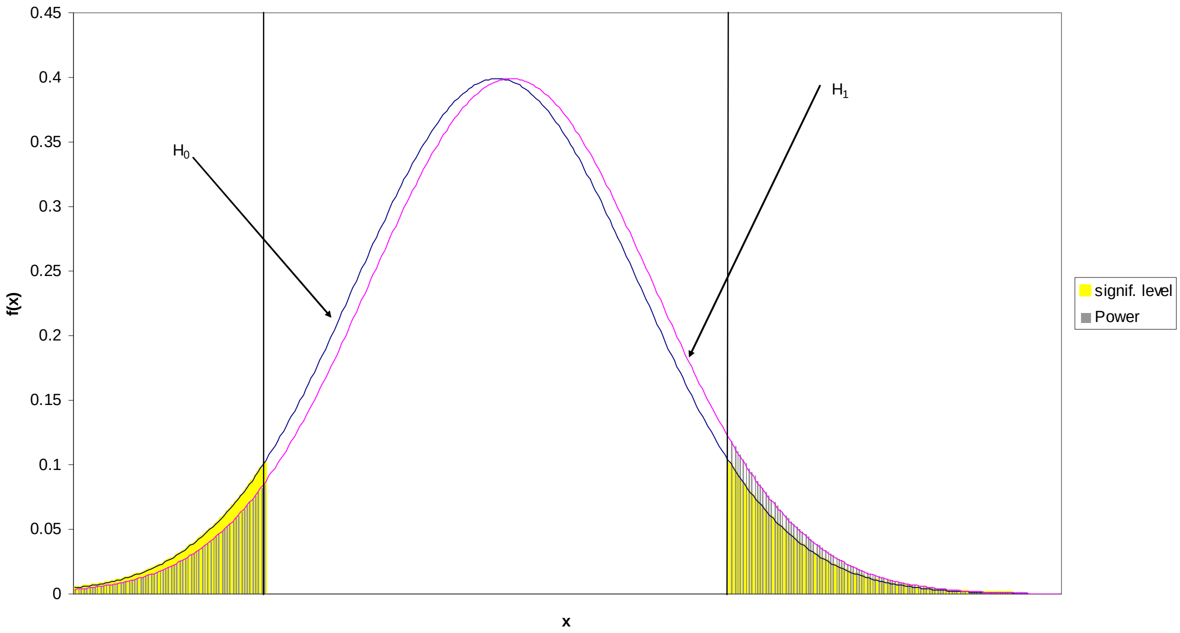


Figure 10: Size and Power: H_0:μ=0 when μ=0.2



Sample Questions

Question 1

A manufacturer claims that through the use of a fuel additive, automobiles should achieve on average an additional 3 miles per gallon of gas. A random sample of 100 automobiles was used to evaluate this product. The sample mean increase in miles per gallon achieved was 2.6 miles and a sample standard deviation was 1.8 miles per gallon. Test the null hypothesis that the population mean is at least 3 miles per gallon. Find the p-value of this test.

Question 2

A mayor in a major city claims that in one particularly depressed neighbourhood, at least 20% of all males between the ages of 18 and 65 are unemployed. A random sample of 120 people from this population contained 20 unemployed people. Test the mayor's claim.

Question 3

A beer manufacturer claims that a new display featuring a life-size picture of a well-known footballer will increase product sales in supermarkets by an average of 50 cases. For a random sample of 20 supermarkets, the average sales increase was 44.3 cases with a sample standard deviation of 12.2 cases. Test at the 5% significance level the null hypothesis that the population mean sales increase is at least 50 cases, stating any assumptions you make.

Question 4

The MATWES procedure was designed to measure attitudes toward women as managers. High scores indicate negative attitudes and low scores indicate positive attitudes. Independent random samples were taken of 151 male MBA students and 108 female MBA students. For the former group, the sample mean and standard deviation MATWES scores were 75.8 and 19.3, while the corresponding figures for the latter group were 71.5 and 12.2. Test the hypothesis that the two population means are equal against the alternative that the true mean MATWES score is higher for male than for female MBA students.

Question 5

Of a random sample of 381 investment grade corporate bonds, 191 had sinking funds. Of an independent random sample of 166 speculative-grade corporate bonds, 98 had sinking funds. Test a 2-sided alternative against the null hypothesis that the two population proportions are equal.

Question 6

A publisher is interested in the effect on sales of university textbooks of cover design. The publisher is planning to bring out 20 texts in the area of business and randomly chooses 10 text to have expensive designs, with the remaining texts having a plain cover. For those with expensive cover designs average sales in the first year were 9254 with a sample standard deviation of 2107. For books with a plain cover average first year sales were 8167, with a standard deviation of 1681. Assuming the population distributions are normal (and the population variances are equal), test the hypothesis that the population means are equal against the alternative that the true mean is higher for books with an expensive cover.

Question 7

A publisher is interested in the effect on sales of university textbooks of cover design. The publisher is planning to bring out 20 texts in the area of business and randomly chooses 10 text to have expensive designs, with the remaining texts having a plain cover. For those with expensive cover designs average sales in the first year were 9254 with a sample standard deviation of 2107. For books with a plain cover average first year sales were 8167, with a standard deviation of 1081. Assuming the population distributions are normal (and the population variances are not equal), test the hypothesis that the population means are equal against the alternative that the true mean is higher for books with an expensive cover.

Question 8

A company produces electric devices operated by a thermostat control. The standard deviation of the temperature at which these controls actually operate should not exceed 2°F. For a random sample of 20 of these controls, the sample standard deviation of operating temperatures was 2.36°F. Stating any assumptions you need to make, test at the 5% significance level the null hypothesis that the population standard deviation is 2°F against the alternative that it is bigger.

Question 9

A doctor is interested in the placebo effect. A random sample of 8 individuals are given a series of tests and are scored out of 100 (Case 1). The same set of individuals are given a sugar coated pill and told that it is designed to increase mental capabilities (even though this is not the case) and are then retested (Case 2). It is known that the performance of individuals in the test follows a normal distribution.

	1	2	3	4	5	6	7	8
Case 1	68	78	45	52	88	56	64	66
Case 2	70	74	48	56	90	59	70	66

At the 5% significance level, test the hypothesis of no difference in the performance of the individuals after being given the placebo.

Question 10

A manufacturer claims that through the use of a fuel additive, automobiles should achieve on average an additional 3 miles per gallon of gas. A random sample of 100 automobiles was used to evaluate this product. The sample mean increase in miles per gallon achieved was 2.6 miles and a sample standard deviation was 1.8 miles per gallon. At the 5% significance level, calculate the power of the test that the population mean is at least 3 miles per gallon, given that the true mean increase in miles per gallon is 2.6

Question 11

The scores in a maths test (out of 30) there collected from a random sample of 16 females. The scores are reported below, according to the females' height split by the lower quartile, the middle two quartiles and the upper quartile:

<25%	25-75%	>75%
18	16	10
19	15	13
15	17	14
27	19	18
21	18	19
		16

Use an ANOVA analysis to test at the 5% significance level whether the three groups of individuals have the same level of performance.

Sample Questions (with Answers)

Question 1

A manufacturer claims that through the use of a fuel additive, automobiles should achieve on average an additional 3 miles per gallon of gas. A random sample of 100 automobiles was used to evaluate this product. The sample mean increase in miles per gallon achieved was 2.6 miles and a sample standard deviation was 1.8 miles per gallon. Test the null hypothesis that the population mean is at least 3 miles per gallon. Find the p-value of this test.

Answer

The distribution of the underlying series is unknown. Nevertheless the distribution of the sample mean will be approximately normal as n→∞.

$$\bar{X} \overset{a}{\sim} N(\mu, s_X^2 / n) \Rightarrow \frac{(\bar{X} - \mu)}{s_X / \sqrt{n}} \sim N(0,1)$$

$$H_0 : \mu \geq 3$$

$$H_1 : \mu < 3$$

$$z = \frac{(2.6 - 3.0)}{1.8 / \sqrt{100}} = -2.22 \Rightarrow P(Z < -2.22) = 0.013$$

and so we reject the null hypothesis at the 1.3% significance level. The probability of observing a value as low as 2.6 miles (assuming null hypothesis is true, that is, $\mu \geq 3$) is 1.3%.

Question 2

A mayor in a major city claims that in one particularly depressed neighbourhood, at least 20% of all males between the ages of 18 and 65 are unemployed. A random sample of 120 people from this population contained 20 unemployed people. Test the mayor's claim.

Answer

The underlying series is a Bernoulli trial. The distribution cannot therefore be normal. Nevertheless the distribution of the sample mean will be approximately normal as $n \to \infty$.

In particular:

$$\overline{X} \overset{a}{\sim} N(\pi, \pi(1-\pi)/n)$$

$$H_0 : \pi \geq 0.2 \qquad\qquad z_{0.05} = -1.645$$

$$H_1 : \pi < 0.2$$

Now under H$_0$ we have: $\Rightarrow \overline{X}_1 \overset{a}{\rightarrow} N(0.2, 0.2(0.8)/120)$ $z = \dfrac{(0.1666 - 0.2)}{\sqrt{\dfrac{(0.2)(0.8)}{120}}} = -0.913$

and so we are unable to reject the null hypothesis at the 5% significance level.

Question 3

A beer manufacturer claims that a new display featuring a life-size picture of a well-known footballer will increase product sales in supermarkets by an average of 50 cases. For a random sample of 20 supermarkets, the average sales increase was 44.3 cases with a sample standard deviation of 12.2 cases. Test at the 5% significance level the null hypothesis that the population mean sales increase is at least 50 cases, stating any assumptions you make.

Answer

This question can only be answered if we are prepared to assume that the increase in product sales will be normally distributed.

$$\bar{X} \sim N(\mu, \sigma^2 / n) \Rightarrow \frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}} \sim N(0,1)$$

we also know that

$$\frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\bar{X} - \mu)}{s_X / \sqrt{n}} \sim t_{n-1}$$

$$H_0 : \mu \geq 50 \qquad t_{19,0.05} = -1.729$$

$$H_1 : \mu < 50$$

$$t = \frac{(44.3 - 50)}{12.2 / \sqrt{20}} = -2.089$$

A t-value of –1.729 occurs with probability of 5%. By using a significance level of 5%, we are saying that an event which occurs with a probability of 5%, or less, is sufficiently rare that we should question the assumption under which the test was undertaken. As we obtained a test statistic of –2.089 this occurs with a probability of less than 5% and we therefore reject H_0.

Question 4

The MATWES procedure was designed to measure attitudes toward women as managers. High scores indicate negative attitudes and low scores indicate positive attitudes. Independent random samples were taken of 151 male MBA students and 108 female MBA students. For the former group, the sample mean and standard deviation MATWES scores were 75.8 and 19.3, while the corresponding figures for the latter group were 71.5 and 12.2. Test the hypothesis that the two population means are equal against the alternative that the true mean MATWES score is higher for male than for female MBA students.

Answer

The underlying series has an unknown distribution, but the distribution of the sample means will both be approximately normally distributed as $n \to \infty$. In particular:

$$\bar{X}_1 \overset{a}{\sim} N(\mu_1, s_{X_1}^2 / n_1) \text{ and } \bar{X}_2 \overset{a}{\sim} N(\mu_2, s_{X_2}^2 / n_2)$$

implying:

$$\bar{X}_1 - \bar{X}_2 \overset{a}{\sim} N\left(\mu_1 - \mu_2, \left[\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2} \right] \right)$$

$$H_0 : \mu_1 - \mu_2 = 0 \qquad z_{0.05} = 1.645$$

$$H_1 : \mu_1 - \mu_2 > 0 \qquad z_{0.01} = 2.323$$

$$z = \frac{(75.8 - 71.5) - 0.0}{\sqrt{\frac{19.3^2}{151} + \frac{12.2^2}{108}}} = 2.193$$

and so we are able to reject the null hypothesis at the 5% significance level, but not at the 1% significance level..

Question 5

Of a random sample of 381 investment grade corporate bonds, 191 had sinking funds. Of an independent random sample of 166 speculative-grade corporate bonds, 98 had sinking funds. Test a 2-sided alternative against the null hypothesis that the two population proportions are equal.

Answer

The underlying series follows a Bernoulli trial and hence the distribution cannot be normal, however, the distribution of the sample mean (sample proportion) will be approximately normally distributed as $n \to \infty$. In particular:

$\bar{X}_1 \overset{a}{\sim} N(\pi_1, \pi_1(1-\pi_1)/n_1)$ and $\bar{X}_2 \overset{a}{\sim} N(\pi_2, \pi_2(1-\pi_2)/n_2)$

implying:

$$\bar{X}_1 - \bar{X}_2 \overset{a}{\sim} N\left(\pi_1 - \pi_2, \left[\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right]\right)$$

$H_0 : \pi_1 - \pi_2 = 0$ $\qquad$ $z_{0.025} = \pm 1.96$

$H_1 : \pi_1 - \pi_2 \neq 0$

Under $H_0 \Rightarrow \bar{X}_1 - \bar{X}_2 \overset{a}{\sim} N\left(0, \left[\frac{p_0(1-p_0)}{n_1} + \frac{p_0(1-p_0)}{n_2}\right]\right)$,

where $p_0 = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

$$z = \frac{(0.5013 - 0.5904) - 0.0}{\sqrt{\frac{0.5283(0.4717)}{381} + \frac{0.5283(0.4717)}{166}}} = -1.919$$

and so we are just unable to reject the null hypothesis at the 5% significance level.

Question 6

A publisher is interested in the effect on sales of university textbooks of cover design. The publisher is planning to bring out 20 texts in the area of business and randomly chooses 10 text to have expensive designs, with the remaining texts having a plain cover. For those with expensive cover designs average sales in the first year were 9254 with a sample standard deviation of 2107. For books with a plain cover average first year sales were 8167, with a standard deviation of 1681. Assuming the population distributions are normal (and the population variances are equal), test the hypothesis that the population means are equal against the alternative that the true mean is higher for books with an expensive cover.

Answer

The underlying distribution is normal, i.e.

$\bar{X}_1 \sim N(\mu_1, \sigma^2 / n_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma^2 / n_2)$

if in addition we assume the variances are equal, we have:

$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right)$ we also know that $\frac{(n_1 + n_2 - 2)s_X^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_X\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$ where $s_x^2 = \frac{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2}{n_1 + n_2 - 2}$.

This implies $s^2 = \frac{(9)2107^2 + (9)1681^2}{18} \Rightarrow s = 1905.94$

$H_0 : \mu_1 - \mu_2 = 0$ $\qquad$ $t_{18}^{0.05} = 1.734$

$H_1 : \mu_1 - \mu_2 > 0$

$$t = \frac{(9254 - 8167) - 0}{1905.94\sqrt{\frac{1}{10} + \frac{1}{10}}} = 1.275$$

and so we are unable to reject the null hypothesis at the 5% significance level.

NOTE:

To test the assumption the variances are equal (at 10% significance level):

$H_0 : \sigma_1^2 = \sigma_2^2$ $\qquad$ $F_{9,9}^{0.05} = 3.18$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{2107^2}{1681^2} = 1.571$$

An F value of 3.18 occurs with a probability of 10%. As we obtained a test statistic of 1.571 this occurs with a probability of more than 10% and we therefore we are unable to reject H_0. In which case our assumption is reasonable:

Question 7

A publisher is interested in the effect on sales of university textbooks of cover design. The publisher is planning to bring out 20 texts in the area of business and randomly chooses 10 text to have expensive designs, with the remaining texts having a plain cover. For those with expensive cover designs average sales in the first year were 9254 with a sample standard deviation of 2707. For books with a plain cover average first year sales were 8167, with a standard deviation of 1062. Assuming the population distributions are normal (and that the population variances are not equal), test the hypothesis that the population means are equal against the alternative that the true mean is higher for books with an expensive cover.

Answer

As the underlying distribution is normal and we have:

$\bar{X}_1 \sim N(\mu_1, \sigma^2 / n_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma^2 / n_2)$

in addition as the population variances are NOT equal, we have:

$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ but as we only have sample variances as opposed to the unknown population variances, means that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim t_{DoF} \text{ where } DoF = \frac{\left(\left(\frac{2707^2}{10}\right) + \left(\frac{1062^2}{10}\right)\right)^2}{\frac{\left(\frac{2707^2}{10}\right)^2}{9} + \frac{\left(\frac{1062^2}{10}\right)^2}{9}} = 11.7$$

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}\right)}} \sim t_{DoF} \text{ where } DoF = \frac{\left[\frac{2707^2}{10} + \frac{1062^2}{10}\right]^2}{\left[\left(\frac{2707^2}{10}\right)^2 / 9 + \left(\frac{1062^2}{10}\right)^2 / 9\right]} = 11.7.$$

$H_0 : \mu_1 - \mu_2 = 0$ $\qquad t_{11}^{0.05} = 1.796$

$H_1 : \mu_1 - \mu_2 > 0$

$$t = \frac{(9254 - 8167) - 0}{\sqrt{\frac{2707^2}{10} + \frac{1062^2}{10}}} = 1.18$$

and so we are unable to reject the null hypothesis at the 5% significance level. NOTE one should test for the equality of the population variances (see Question 6).

Question 8

A company produces electric devices operated by a thermostat control. The standard deviation of the temperature at which these controls actually operate should not exceed 2°F. For a random sample of 20 of these controls, the sample standard deviation of operating temperatures was 2.36°F. Stating any assumptions you need to make, test at the 5% significance level the null hypothesis that the population standard deviation is 2°F against the alternative that it is bigger.

Answer

Assuming the underlying distribution of the temperature at which the controls operate is normal, then we have:

$$\frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$$

$H_0 : \sigma^2 = 4$ $\qquad\qquad$ $\chi_{19,0.05}^2 = 30.14$

$H_1 : \sigma^2 > 4$

$$\chi^2 = \frac{19(2.36)^2}{4} = 26.46$$

A chi-squared value of 30.14 occurs with probability of 5%. As we obtained a test statistic of 26.46 this occurs with a probability of more than 5% and we therefore we are unable to reject H$_0$.

Question 9

A doctor is interested in the placebo effect. A random sample of 8 individuals are given a series of tests and are scored out of 100 (Case 1). The same set of individuals are given a sugar coated pill and told that it is designed to increase mental capabilities (even though this is not the case) and are then retested (Case 2). It is known that the performance of individuals in the test follows a normal distribution.

	1	2	3	4	5	6	7	8
Case 1	68	78	45	52	88	56	64	66
Case 2	70	74	48	56	90	59	70	66

At the 5% significance level, test the hypothesis of no difference in the performance of the individuals after being given the placebo.

Answer

As the underlying distribution is normal and we have:

$X_1 \sim N(\mu_1,\sigma_1^2)$ and $X_2 \sim N(\mu_2,\sigma_2^2)$, then we know that $X_1 - X_2 \sim N(\mu_1 - \mu_2,\sigma_d^2)$. This implies:

$$\overline{X_1 - X_2} \sim N\left(\mu_1 - \mu_2,\sigma_d^2 / n\right) \text{ and } \frac{(\overline{X_1 - X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_d^2 / n}} \sim N\left(0,1\right)$$

we also know that

$$\frac{(n-1)s_d^2}{\sigma_d^2} \sim \chi_{n-1}^2$$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\overline{X_1 - X_2}) - (\mu_1 - \mu_2)}{\sqrt{s_d^2 / n}} \sim t_{n-1}$$

From the data above $\bar{d} = -2.0$, $s_d^2 = 8.857$.

$H_0 : \mu_d \equiv \mu_1 - \mu_2 = 0$ $\qquad t_{7,0.05} = -1.895$

$H_1 : \mu_d \equiv \mu_1 - \mu_2 < 0$

$$t = \frac{(2.0 - 0)}{\sqrt{8.857 / 8}} = -1.90$$

and so we are just able to reject the null hypothesis at the 5% significance level.

Question 10

A manufacturer claims that through the use of a fuel additive, automobiles should achieve on average an additional 3 miles per gallon of gas. A random sample of 100 automobiles was used to evaluate this product. The sample mean increase in miles per gallon achieved was 2.6 miles and a sample standard deviation was 1.8 miles per gallon. At the 5% significance level, calculate the power of the test that the population mean is at least 3 miles per gallon, given that the true mean increase in miles per gallon is 2.6

Answer

$H_0 : \mu \geq 3 \qquad H_1 : \mu < 3 \qquad\qquad z_{0.05} = -1.645$

$$z = \frac{(\overline{x}^c - 3.0)}{1.8 / \sqrt{100}} = -1.645 \Rightarrow \overline{x}^c = -1.645(0.18) + 3 \Rightarrow \overline{x}^c = 2.7039$$

Power=Pr(Reject H$_0$| H$_0$ false)=Pr(Reject $\mu \geq 3 \,|\, \mu = 2.0$ **)**

$$\Pr(\overline{X} < 2.7039 \,|\, \mu = 2.6) = \Pr(Z < \frac{2.7039 - 2.6}{0.18}) = \Pr(Z < 0.58) = 0.718$$

Question 11

The scores in a maths test (out of 30) there collected from a random sample of 16 females. The scores are reported below, according to the females' height split by the lower quartile, the middle two quartiles and the upper quartile:

<25%	25-75%	>75%
18	16	10
19	15	13
15	17	14
27	19	18
21	18	19
		16

Use an ANOVA analysis to test at the 5% significance level whether the three groups of individuals have the same level of performance.

Answer

	<25%	25-75%	>75%	Overall
	18	16	10	
	19	15	13	
	15	17	14	
	27	19	18	
	21	18	19	
			16	
n	5	5	6	16
$\bar{x}$	20	17	15	17.1875
$\sum_{i=1}(x_i - \bar{x})^2$	80	10	56	214.538

$$BSS = 5 \times (20 - 17.1875)^2 + 5 \times (17 - 17.1875)^2 + 6 \times (15 - 17.1875)^2 = 68.4375$$

$$WSS = 80 + 10 + 56 = 146$$

$$F = \frac{68.4375/(3-1)}{146/(16-3)} = 3.05 \qquad F_{2,13}^{0.05} = 3.81$$

Do NOT reject H_0.

Handout 5: Appendix 3

Hypothesis Testing Formulas

Hypothesis Testing: Test Statistics for Tests of Means

One Population

Sample	Hypothesis	Distrib of *X_i*	σ^2 Known	σ^2 Not Known
Large/Small	$H_0 : \mu = \mu_0$	Normal	$z = \frac{(\bar{x} - \mu_0)}{\sigma / \sqrt{n}}$	$t = \frac{(\bar{x} - \mu_0)}{s_x / \sqrt{n}}$
Large	$H_0 : \mu = \mu_0$	Non-Normal	$z = \frac{(\bar{x} - \mu_0)}{\sigma / \sqrt{n}}$	$z = \frac{(\bar{x} - \mu_0)}{s_x / \sqrt{n}}$
Small	$H_0 : \mu = \mu_0$	Non-Normal	?	?

Two Populations

Sample	Hypothesis	Distrib of *X_i*	σ^2 Known	σ^2 Not Known
Large/Small	$H_0 : \mu_1 - \mu_2 = \delta$	Normal	$z = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2\}}}$	---
Large	$H_0 : \mu_1 - \mu_2 = \delta$	Non-Normal	$z = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}}$	$z = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2)}}$
Large/Small	$H_0 : \mu_1 - \mu_2 = \delta$	Normal	---	$t = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{s_0\sqrt{(1/n_1 + 1/n_2)}}$ where $s_0^2 = \{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2\} / (n_1 + n_2 - 2)$
Large/Small	$H_0 : \mu_1 - \mu_2 = \delta$	Normal	---	$t = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2)}}$ where $DoF = \frac{\left[s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2\right]^2}{\left(s_{x_1}^2 / n_1\right)^2 / (n_1 - 1) + \left(s_{x_2}^2 / n_2\right)^2 / (n_2 - 1)}$
Small Sample	$H_0 : \mu_1 - \mu_2 = \delta$	Non-Normal	?	?

Handout 5: Appendix 3

Tests on Proportions

One Population

	Hypothesis	Distrib of X_i	Test
Large sample	$H_0 : \pi = \pi_0$	Non-normal	

$$z = \frac{\bar{x} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Small sample	$H_0 : \pi = \pi_0$	Non-normal	?

Two Populations

Large sample	$H_0 : \pi_1 - \pi_2 = 0$	Non-normal	

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{p_0(1-p_0)(1/n_1 + 1/n_2)}} \text{ where } p_0 = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2}{n_1 + n_2}$$

Small sample	$H_0 : \pi_1 - \pi_2 = 0$	Non-normal	?

Tests on variances

One Population

Large/Small	$H_0 : \sigma^2 = \sigma_o^2$	Normal	

$$u = \frac{(n-1)s_x^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Two Populations

Large/Small	$H_0 : \sigma_1^2 = \sigma_2^2$	Normal	

$$F = s_{x_1}^2 / s_{x_2}^2 \sim F_{(n_1-1, n_2-1)}$$

ANOVA: $F = \dfrac{\textit{Between}\ \text{SS}/(k\text{-}1)}{\textit{Within}\ \text{SS}/(\text{n-}k)} \sim F_{k-1,n-k}^{\alpha}$

7STATISTICAL TECHNIQUES B

Confidence Intervals

1. Introduction

Confidence intervals give a range of likely values for the TRUE (but unknown) population parameter, together with a measure of the confidence (or likelihood) that the range contains the true value. For some unknown population parameter, θ, based on sample data we find two values a and b, such that,

$$\Pr\left\{a < \theta < b\right\} = 1 - \alpha \qquad for \;\; 0 < \alpha < 1$$

then we can say with $100(1-\alpha)\%$ confidence that θ lies in the range *a* to *b*. That means in repeated samples, $100(1-\alpha)\%$ of the time, θ would lie within intervals calculated this way.

Consider a N(0,1) distribution we know that

$$\Pr\left\{-1.645 < Z < 1.645\right\} = 0.90 .$$

We take symmetric points around zero as this minimises the range for the interval (compare Appendix 1: Figures 1 and 2).

Table 1: Range for a N(0,1) for a 90% interval

p_a	a	p_b	b	range
0.05	-1.645	0.05	1.645	3.29
0.04	-1.74	0.06	1.56	3.30
0.01	-2.32	0.09	1.34	3.66

Table 2: Critical values for a N(0,1)

CI(%)	Lower limit	Upper limit
90	-1.645	1.645
95	-1.96	1.96
99	-2.575	2.575

A diagrammatic illustration of this is provided by Appendix 1: Figure 3. To be more confident of a statement or value our degree of uncertainty or range of possible values has to increase.

2. Confidence Interval for mean of a distribution

Let $X_1, X_2, \ldots, X_n$ denote a random sample of n observations from a normal distribution with unknown mean, μ, and known variance, σ^2. Then, we know that

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and this implies (by standardising) that, $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$.

As,

$\Pr(-1.645 < Z < 1.645) = 0.9$.

Therefore,

$$\Pr(-1.645 < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < 1.645) = 0.9$$

$$\Pr(-1.645(\sigma / \sqrt{n}) < \bar{X} - \mu < 1.645(\sigma / \sqrt{n})) = 0.9$$

$$\Pr(-\bar{X} - 1.645(\sigma / \sqrt{n}) < -\mu < -\bar{X} + 1.645(\sigma / \sqrt{n})) = 0.9$$

$$\Pr(\bar{X} - \underbrace{1.645(\sigma / \sqrt{n})}_{D} < \mu < \bar{X} + \underbrace{1.645(\sigma / \sqrt{n})}_{D}) = 0.9\,.$$

such that we expect the interval $\left\{\bar{X} - D, \bar{X} + D\right\}$ to contain μ on 90% of occasions.

However, after taking a sample and calculating the actual sample mean, $\bar{x}$, we can say that we are 90% confident that the interval $\left\{\bar{x} - D, \bar{x} + D\right\}$ contain the (unknown) population mean, μ.

The width of the confidence interval 2D depends three factors:

(i) The level of confidence, that is, 90%, 95% or 99%. The interval for 99% being far wider than that for 90%.

(ii) The variability in the underlying distribution, σ, the greater the variability the wider the interval.

(iii) The number of observation in the sample, *n*, as this effects the standard error of the sample mean, $\sigma / \sqrt{n}$, as *n* increases the standard error of the sample mean falls and hence the interval narrows. (Appendix 2: Example 1)

2.1 Variants of this basic confidence interval case:

(1) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of *n* observations from a distribution which is NOT NORMAL, with an unknown mean, μ, but the population variance is KNOWN, σ^2. If *n* is large enough (*n*>30) then by appealing to the central limit theorem (CLT) we can say that

$$\bar{X} \overset{a}{\sim} N(\mu, \sigma^2 / n)$$

in which case, the $100(1-\alpha)\%$ confidence interval is still written as:

$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n})$, where $z_{\alpha/2}$ is the critical value from a N(0,1).

(2) In the previous case, if the σ^2 is UNKNOWN, then if *n*>30 $\bar{X} \overset{a}{\sim} N(\mu, s^2 / n)$ the confidence interval is written as: $\bar{x} \pm z_{\alpha/2}(s_x / \sqrt{n})$, where $z_{\alpha/2}$ is the critical value from a N(0,1) (Appendix 2: Example 2).

(3) A specific example of the previous case is when $X_1, X_2, \ldots, X_n$ denotes a random sample from a Bernoulli (NOT NORMAL) distribution, that is,

X	0	1
Pr(*X*)	$1-\pi$	π

$E(X) = \pi$ and $V(X) = \pi(1-\pi)$

with a unknown mean, $\mu(=\pi)$, and an unknown population variance, $\sigma^2(=\pi(1-\pi))$.

Then if *n*>30 by a CLT $\bar{X} \overset{a}{\sim} N(\pi, \pi(1-\pi)/n)$ in which case, the $100(1-\alpha)\%$ confidence interval is written as: $p \pm z_{\alpha/2}(\sqrt{p(1-p)/n})$, where p is the sample proportion. (Appendix 2: Example 3).

(4) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of *n* observations from a distribution which is NORMAL, with a unknown mean, μ, and an UNKNOWN σ^2.

Then: $\dfrac{\bar{X} - \mu}{s_X / \sqrt{n}} \sim t_{n-1}$ in which case, the $100(1-\alpha)\%$ confidence interval is written as:

$\bar{x} \pm t_{n-1}^{\alpha/2}(s_x / \sqrt{n})$, where $t_{n-1}^{\alpha/2}$ is the critical value from a t-distribution with *n*-1 degrees of freedom. (Appendix 2: Example 4).

3. Confidence Interval for the difference in means

3.1 Independent samples:

Assume we have two ***independent*** samples of size, n_1 and n_2, on X_1 and X_2, respectively.

The sample means are $\bar{X}_1$ and $\bar{X}_2$: $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ and $V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

If the underlying distribution of X_1 and X_2 are normal and the population variances are KNOWN, then: $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ and the confidence interval for $\mu_1 - \mu_2$ is: $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]}$

3.1.2 Variants on the CI for the difference in means (independent samples)

(1) If the underlying distributions are NOT NORMAL and the population variances are KNOWN, providing $n_1 > 30$ and $n_2 > 30$, then from a CLT the CI is: $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]}$.

(2) If the underlying distributions is NOT NORMAL and the population variances are UNKNOWN, providing $n_1 > 30$ and $n_2 > 30$, then from a CLT the CI is $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\left[\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}\right]}$ (Appendix 2: Example 5).

(3) Difference in sample proportions, applying CLT we get CI:

$(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\left[\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right]}$, where p_1 and p_2 are the two sample proportions from the two samples. (Appendix 2: Example 6)

(4) If the underlying distribution are NORMAL, and the population variances are UNKNOWN (and EQUAL), the CI is: $(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{\left[\frac{s_x^2}{n_1} + \frac{s_x^2}{n_2}\right]}$, where,

$s_x^2 = \frac{(n_1-1)s_{x_1}^2 + (n_2-1)s_{x_2}^2}{n_1+n_2-2}$ (Appendix 2: Example 7).

(5) If the underlying distribution are NORMAL, and the population variances are UNKNOWN (and UNEQUAL), the CI is: $(\bar{x}_1 - \bar{x}_2) \pm t_{DoF,\alpha/2}\sqrt{\left[\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}\right]}$, where,

$$DoF = \frac{\left[s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2\right]^2}{\left(s_{x_1}^2 / n_1\right)^2 / (n_1 - 1) + \left(s_{x_2}^2 / n_2\right)^2 / (n_2 - 1)}$$ (Appendix 2: Example 8).

Appendix 3 is a reference table for the different confidence interval formulas.

<u>3.2 Matched pairs</u>

Again we are interested in formulating the confidence interval for $\mu_1 - \mu_2$, however, in this case, the two experiments are with the same sample of individuals and cannot therefore be independent. Given the outcomes of the two trials are for the same individuals we form the difference in the outcomes of the two random variables:

$D = X_1 - X_2$,

where $E(D) = E(X_1 - X_2) = \mu_1 - \mu_2 \equiv \mu_d$, $V(D) = V(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} \equiv \sigma_d^2$ and $\sigma_{12} > 0$.

If the underlying distribution of $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ then $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_d^2)$ then we know that $\overline{X_1 - X_2} \sim N(\mu_1 - \mu_2, \sigma_d^2 / n)$.

Normalising this expression we have: $\frac{(\overline{X_1 - X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_d^2 / n}} \sim N(0,1)$, but σ_d^2 is unknown and we need to replace this by the sample variance in which case $\frac{(\overline{X_1 - X_2}) - (\mu_1 - \mu_2)}{\sqrt{s_d^2 / n}} \sim t_{n-1}$.

Given sample of data for the random variable X_1 and X_2, define the difference as:

$d_1 = x_1^1 - x_{21}^1$, $d_2 = x_1^2 - x_2^2$, $d_3 = x_1^3 - x_2^3$, $\ldots d_n = x_1^n - x_2^n$

And we calculate the sample moments of this series: $\bar{d} = \sum_{i=1}^{n} d_i / n$ and $s_d^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2 / (n-1)$. The $100(1-\alpha)\%$ confidence interval is written as

$\bar{d} \pm t_{n-1}^{\alpha/2}(s_d / \sqrt{n})$ (Appendix 2: Example 9).

4. Confidence Interval for the variance of a distribution

Similarly to producing confidence intervals for the population mean, we wish to produce an interval estimate of the population variance on the basis of a sample of data. To formulate a confidence interval the random variable, *X*, MUST be normally distributed. In which case, $w = \frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$.

Now, $\Pr\left\{\chi_{n-1,0.95}^2 < w < \chi_{n-1,0.05}^2\right\} = 0.90$. Note that for a χ^2 distribution, a symmetric confidence interval does not necessarily minimise the range – in particular, a smaller range will be obtained by having only 1% in the left tail and 9% in the right tail for a 90% confidence interval (see Appendix 1: Figure 4). Substituting for *w* and rearranging gives:

$\Pr\left\{\chi_{n-1,0.95}^2 < \frac{(n-1)s_X^2}{\sigma^2} < \chi_{n-1,0.05}^2\right\} = 0.90$. Rearranging, we get:

$$\Pr\left\{\frac{\chi_{n-1,0.95}^2}{(n-1)s_X^2} < \frac{1}{\sigma^2} < \frac{\chi_{n-1,0.05}^2}{(n-1)s_X^2}\right\} \Rightarrow \Pr\left\{\frac{(n-1)s_X^2}{\chi_{n-1,0.05}^2} < \sigma^2 < \frac{(n-1)s_X^2}{\chi_{n-1,0.95}^2}\right\} = 0.90$$

Appendix 1: Figure 5 compares the distribution of a χ_4^2 to that of a χ_8^2 (Appendix 2: Example 10).

Figure 1: 90% Confidence interval for N(0,1) (symmetric)

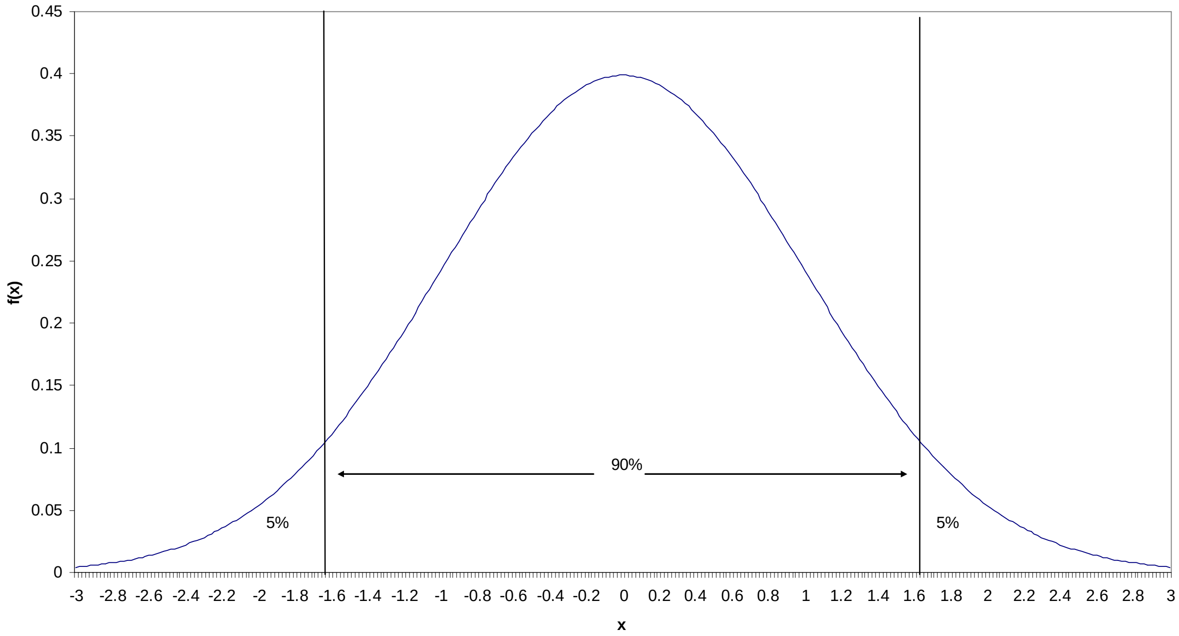
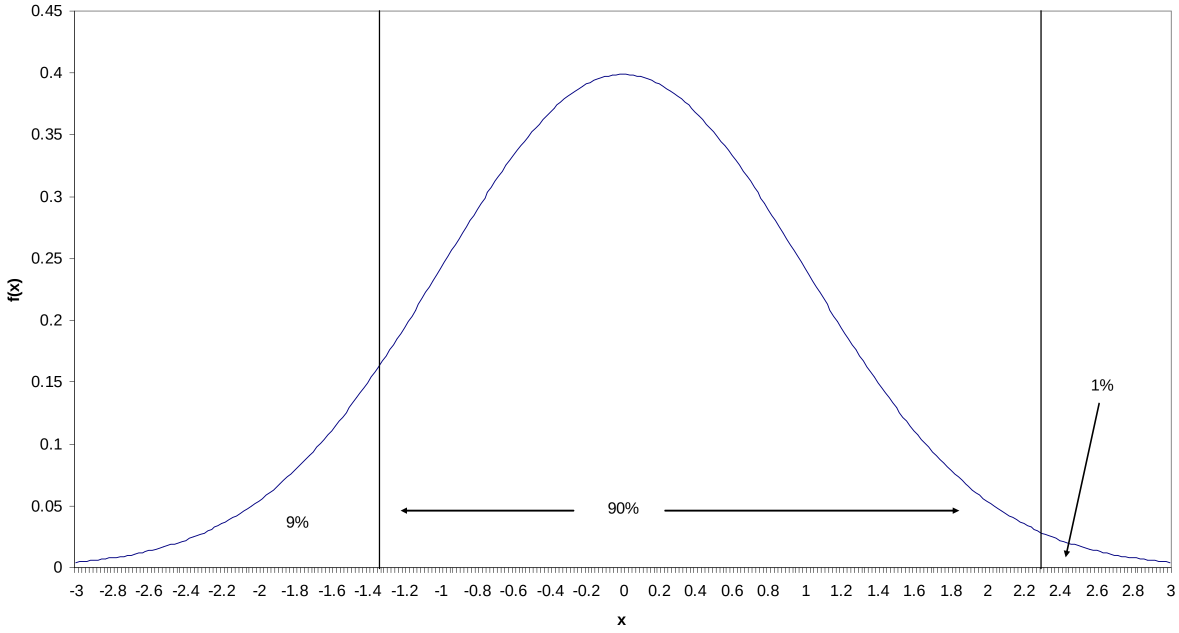


Figure 2: Alternate 90% confidence interval for N(0,1)



Handout 6: Appendix 1

Figure 3: Confidence limits for a N(0,1)

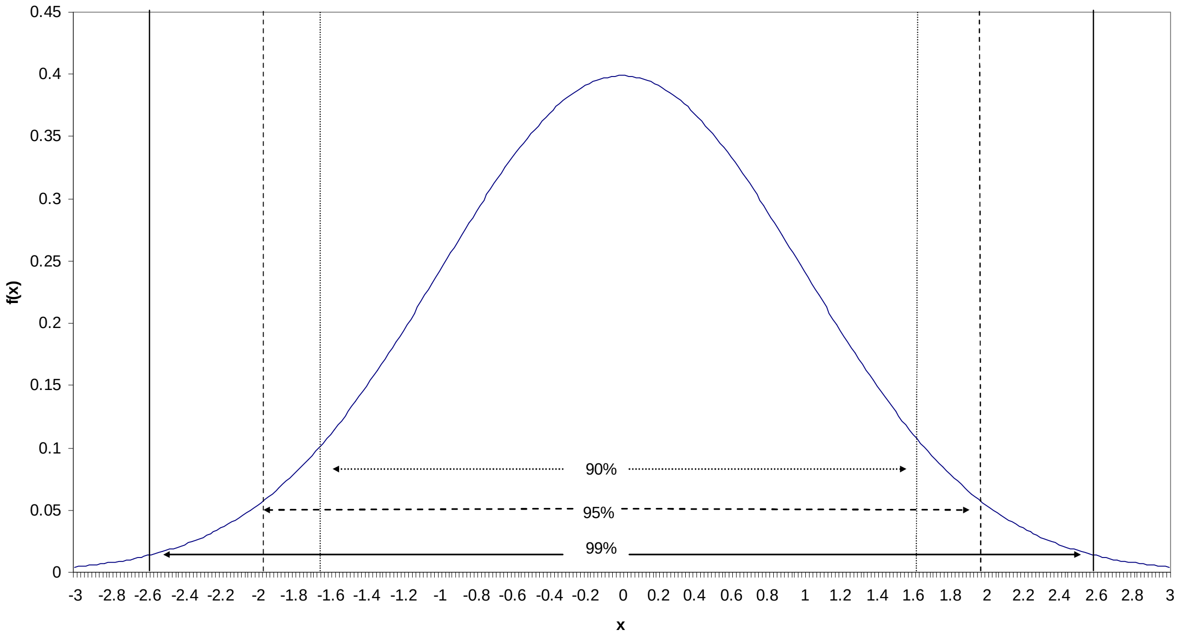


Figure 4: Confidence intervals for Chi-squared

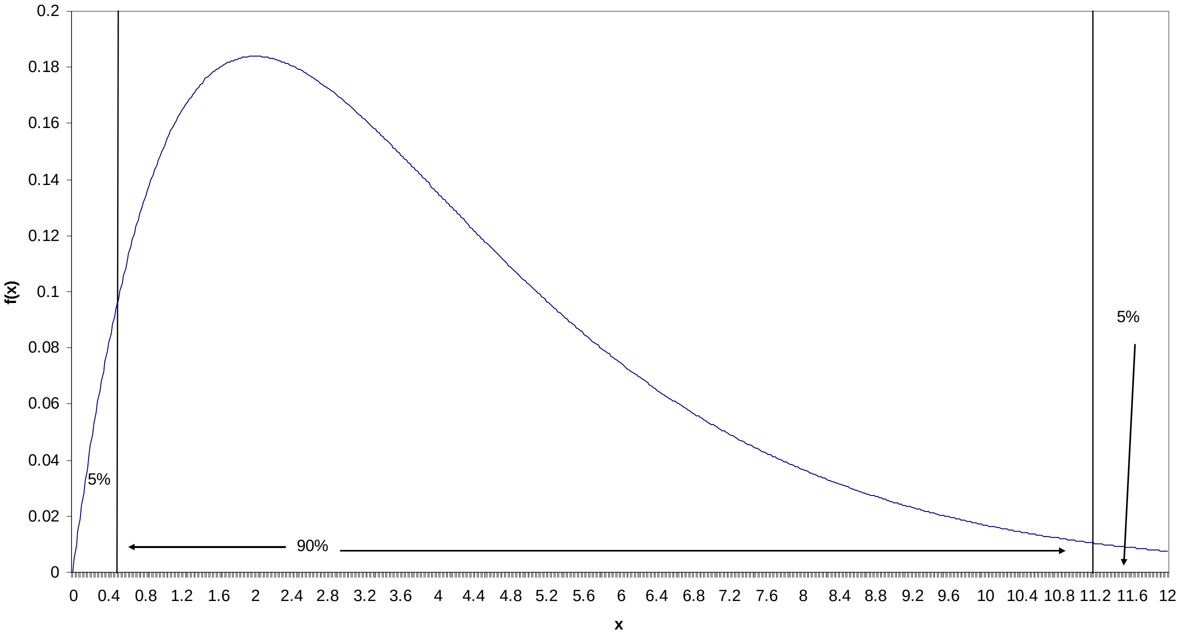
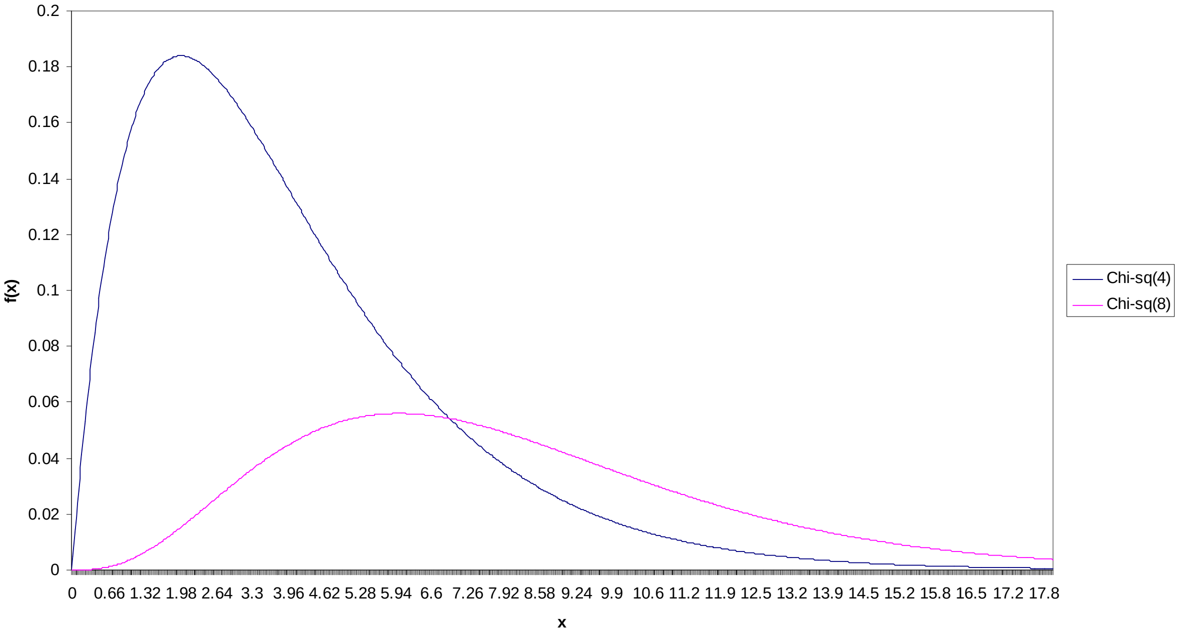


Figure 5: Chi-squared(4) and Chi-squared(8)



Sample Questions

Question 1

A personnel manager found that historically, the scores on aptitude tests given to applicants are normal with a standard deviation of 32.4 points. A random sample of 9 scores produced a mean of 187.9 points. Based on these results a statistician found a population mean confidence interval of 165.8-210.0 points.

(a) Find the probability of this interval.

(b) Find an 80% confidence interval for the population mean score.

Question 2

A random sample of 125 economics students were asked to rate the importance of particular job characteristics on a scale from 1 (not important) to 5 (extremely important). For the question on job security the sample mean rating was 4.18 and the sample standard deviation 0.80. Find a 99% confidence interval for the population mean.

Question 3

A random sample of 850 voters were asked, "If there was a referendum tomorrow on Britain joining the ERM, how would you vote?" 391 voters reported support for Britain joining the ERM. Find a 95% confidence interval for the population proportion of all voters supporting Britain joining the ERM.

Question 4

GAP is interested in the expenditure on clothes of university students in the first month of the academic year. For a random sample of 15 students, the mean expenditure was £89.56 and the sample standard deviation was £20.13. Assuming that the population distribution is normal, find a 95% confidence interval for population mean expenditure.

Question 5

Independent samples of Vice-Chancellors (VCs) and Chief Executive Officers (CEOs) in large private companies were asked the importance of salary to their job satisfaction on a scale of 1 (not important at all) to 10 (the most important aspect). A random sample of 42 VCs had a mean rating of 4.01 and sample standard deviation of 1.2. For an independent random sample of 68 CEOs the mean rating was 5.43 and a sample standard deviation of 1.7. Find a 95% confidence interval for the difference in the population mean responses.

Question 6

Of a random sample of 150 Economics students 105 said that teaching as a career was very unappealing. For an independent sample of 120 English Literature students 72 had the same reaction to teaching as a career. Find a 95% confidence interval for the difference between the population proportions regarding teaching as an unappealing career.

Question 7

A researcher intends to estimate the effect of a drug on scores of human subjects performing a task of psychomotor coordination. The members of a random sample of 9 subjects were given the drug prior to testing, their mean score was 9.78 and the sample standard deviation was 4.2. An independent sample of 10 subjects were given a placebo prior to testing, the mean score for this group was 15.10 and the sample standard deviation was 5.2. Assuming the population distributions are normal, find a 90% confidence interval for the difference in the population means.

Question 8

A researcher intends to estimate the effect of a drug on scores of human subjects performing a task of psychomotor coordination. The members of a random sample of 9 subjects were given the drug prior to testing, their mean score was 9.78 and the sample standard deviation was 3.2. An independent sample of 10 subjects were given a placebo prior to testing, the mean score for this group was 15.10 and the sample standard deviation was 7.2. Assuming the population distributions are normal, find a 90% confidence interval for the difference in the population means.

Question 9

A random sample of 15 financial analysts' forecasts of next years' earnings per share for a large corporation was taken. The sample standard deviation was £0.88. Find a 95% confidence interval for the variance of predicted earnings per share for all analysts.

Sample Questions (with Answers)

Question 1

A personnel manager found that historically, the scores on aptitude tests given to applicants are normal with a standard deviation of 32.4 points. A random sample of 9 scores produced a mean of 187.9 points. Based on these results a statistician found a population mean confidence interval of 165.8-210.0 points.

(c) Find the probability of this interval.

(d) Find an 80% confidence interval for the population mean score.

Answer

The underlying distribution is normal with a known population variance, therefore the distribution of the sample mean will be normal:

$$\bar{X} \sim N(\mu, \sigma^2 / n) \Rightarrow \bar{X} \sim N(\mu, 32.4^2 / 9)$$

(a) Now, $187.9 - z_{\alpha/2}\left(\frac{32.4}{3}\right) \leq \mu \leq 187.9 + z_{\alpha/2}\left(\frac{32.4}{3}\right) \Rightarrow 165.8 \leq \mu \leq 210.0$

$$D = 22.1 \Rightarrow z_{\alpha/2}\left(\frac{32.4}{3}\right) = \pm 22.1 \Rightarrow z_{\alpha/2} = \pm 2.046 \Rightarrow \alpha / 2 = 0.0204$$

$\alpha = 0.0408 \Rightarrow$CI is 95.92%

(b) To construct an 80% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha / 2 = 0.10 \Rightarrow z_{0.10} = 1.28$$

Therefore the 80% confidence interval is:

$$\bar{x} - z_{0.10}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.10}\frac{\sigma}{\sqrt{n}}$$

where $\bar{x} = 187.9$.

The required 80% confidence interval is therefore:

$$187.9 - 1.28\left(\frac{32.4}{3}\right) \leq \mu \leq 187.9 + 1.28\left(\frac{32.4}{3}\right) \Rightarrow 174.05 \leq \mu \leq 201.75$$

and this is our 80% confidence interval for the population mean rating.

Question 2

A random sample of 125 economics students were asked to rate the importance of particular job characteristics on a scale from 1 (not important) to 5 (extremely important). For the question on job security the sample mean rating was 4.18 and the sample standard deviation 0.80. Find a 99% confidence interval for the population mean.

Answer

The underlying series has outcomes taking one of five integer values between 1 and 5. The distribution cannot therefore be normal. Nevertheless the distribution of the sample mean will be approximately normal as $n \rightarrow \infty$. In particular:

$$\overline{X} \overset{a}{\sim} N(\mu, s_X^2 / n) \Rightarrow \overline{X} \overset{a}{\sim} N(\mu, 0.64/125)$$

To construct a 99% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha / 2 = 0.005 \Rightarrow z_{0.005} = 2.575$$

Therefore the 99% confidence interval is:

$$\overline{x} - z_{0.005} \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{0.005} \frac{s}{\sqrt{n}}$$

where $\overline{x} = 4.18$.

The required 99% confidence interval is therefore:

$$4.18 - 2.575\left(\frac{0.80}{\sqrt{125}}\right) \leq \mu \leq 4.18 + 2.575\left(\frac{0.80}{\sqrt{125}}\right) \Rightarrow 3.996 \leq \mu \leq 4.364$$

and this is our 99% confidence interval for the population mean rating.

Question 3

A random sample of 850 voters were asked, “If there was a referendum tomorrow on Britain joining the ERM, how would you vote?” 391 voters reported support for Britain joining the ERM. Find a 95% confidence interval for the population proportion of all voters supporting Britain joining the ERM.

Answer

While the underlying series follows a Bernoulli trial and hence the distribution cannot be normal, the distribution of the sample mean (sample proportion) will be approximately normally distributed as *n*→∞. In particular:

$$\bar{X} \overset{a}{\sim} N(\pi, p(1-p)/n) \Rightarrow \bar{X} \overset{a}{\sim} N(\pi, (0.46)(0.54)/850)$$

To construct a 95% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha / 2 = 0.025 \Rightarrow z_{0.025} = 1.96$$

Therefore the 95% confidence interval is:

$$p - z_{0.025}\sqrt{\frac{p(1-p)}{n}} \le \pi \le p + z_{0.025}\sqrt{\frac{p(1-p)}{n}}$$

where $p = 0.46$.

The required 95% confidence interval is therefore:

$$0.46 - 1.96\sqrt{\left(\frac{0.46(0.54)}{850}\right)} \le \pi \le 0.46 + 1.96\sqrt{\left(\frac{0.46(0.54)}{850}\right)} \Rightarrow 0.426 \le p \le 0.493$$

Question 4

GAP is interested in the expenditure on clothes of university students in the first month of the academic year. For a random sample of 15 students, the mean expenditure was £89.56 and the sample standard deviation was £20.13. Assuming that the population distribution is normal, find a 95% confidence interval for population mean expenditure.

Answer

The underlying series has a normal distribution, but the population standard deviation is unknown.

$$\bar{X} \sim N(\mu, \sigma^2 / n) \Rightarrow \frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}} \sim N(0,1)$$

we also know that

$$\frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\bar{X} - \mu)}{s_X / \sqrt{n}} \sim t_{n-1}$$

To construct a 95% confidence interval we need to find a point $t_{14,\alpha/2}$, such that:

$$P(t_{14} > t_{14,\alpha/2}) = \alpha / 2 = 0.025 \Rightarrow t_{14,0.025} = 2.145$$

Therefore the 95% confidence interval is:

$$\bar{x} - t_{14,0.025} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{14,0.025} \frac{s}{\sqrt{n}}$$

where $\bar{x} = 89.56$ and $s = 20.13$.

The required 95% confidence interval is therefore:

$$89.56 - 2.145\left(\frac{20.13}{\sqrt{15}}\right) \leq \mu \leq 89.56 + 2.145\left(\frac{20.13}{\sqrt{15}}\right) \Rightarrow 78.41 \leq \mu \leq 100.71.$$

Question 5

Independent samples of Vice-Chancellors (VCs) and Chief Executive Officers (CEOs) in large private companies were asked the importance of salary to their job satisfaction on a scale of 1 (not important at all) to 10 (the most important aspect). A random sample of 42 VCs had a mean rating of 4.01 and sample standard deviation of 1.2. For an independent random sample of 68 CEOs the mean rating was 5.43 and a sample standard deviation of 1.7. Find a 95% confidence interval for the difference in the population mean responses.

Answer

The underlying series has outcomes taking values between 1 and 10. The distribution cannot therefore be normal. Nevertheless the distribution of the sample mean will be approximately normal as $n \to \infty$. In particular:

$$\overline{X}_1 \overset{a}{\sim} N(\mu_1, s_{x_1}^2 / n_1) \Rightarrow \overline{X}_1 \overset{a}{\sim} N(\mu_1, 1.44/42)$$

$$\overline{X}_2 \overset{a}{\sim} N(\mu_2, s_{x_2}^2 / n_2) \Rightarrow \overline{X}_2 \overset{a}{\sim} N(\mu_2, 2.89/68)$$

and

$$\overline{X}_1 - \overline{X}_2 \overset{a}{\sim} N\left(\mu_1 - \mu_2, \frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}\right) \Rightarrow \overline{X}_1 - \overline{X}_2 \overset{a}{\sim} N(\mu_1 - \mu_2, 0.0768)$$

To construct a 95% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha / 2 = 0.025 \Rightarrow z_{0.025} = 1.96$$

Therefore the 95% confidence interval is:

$$(\overline{x}_1 - \overline{x}_2) - z_{0.025}\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}} \le \mu_1 - \mu_2 \le (\overline{x}_1 - \overline{x}_2) + z_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\overline{x}_1 - \overline{x}_2 = -1.42$.

The required 95% confidence interval is therefore:

$$-1.42 - 1.96(0.277) \le \mu_1 - \mu_2 \le -1.42 + 1.96(0.277) \Rightarrow -1.963 \le \mu_1 - \mu_2 \le -0.877.$$

Question 6

Of a random sample of 150 Economics students 105 said that teaching as a career was very unappealing. For an independent sample of 120 English Literature students 72 had the same reaction to teaching as a career. Find a 95% confidence interval for the difference between the population proportions regarding teaching as an unappealing career.

Answer

The underlying series follows a Bernoulli trial and hence the distribution cannot be normal, however, the distribution of the sample mean (sample proportion) will be approximately normally distributed as $n \to \infty$. In particular:

$\bar{X}_1 \overset{a}{\sim} N(\pi_1, p_1(1-p_1)/n_1)$ and $\bar{X}_2 \overset{a}{\sim} N(\pi_2, p_2(1-p_2)/n_2)$

implying:

$$\bar{X}_1 - \bar{X}_2 \overset{a}{\sim} N\left(\pi_1 - \pi_2, \left[\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right]\right)$$

To construct a 95% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$P(Z > z_{\alpha/2}) = \alpha/2 = 0.025 \Rightarrow z_{0.025} = 1.96$

Therefore the 95% confidence interval is:

$$(p_1 - p_2) - z_{0.025}\sqrt{\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)} \le \pi_1 - \pi_2 \le (p_1 - p_2) + z_{0.025}\sqrt{\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)}$$

The required 95% confidence interval is therefore:

$0.1 - 1.96(0.0583) \le \pi_1 - \pi_2 \le 0.1 + 1.96(0.0583) \Rightarrow -0.014 \le \pi_1 - \pi_2 \le 0.214$.

Question 7

A researcher intends to estimate the effect of a drug on scores of human subjects performing a task of psychomotor coordination. The members of a random sample of 9 subjects were given the drug prior to testing, their mean score was 9.78 and the sample standard deviation was 4.2. An independent sample of 10 subjects were given a placebo prior to testing, the mean score for this group was 15.10 and the sample standard deviation was 5.2. Assuming the population distributions are normal, find a 90% confidence interval for the difference in the population means.

Answer

The underlying series has a normal distribution, but the population standard deviation is unknown, assuming they are equal:

$\bar{X}_1 \sim N(\mu_1, \sigma^2 / n_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma^2 / n_2)$

implying:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right)$$

we also know that $\frac{(n_1 + n_2 - 2)s_X^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_X\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$ where $s_x^2 = \frac{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2}{n_1 + n_2 - 2}$, $s_1^2 = 4.2^2$ and $s_{21}^2 = 5.2^2$.

To construct a 90% confidence interval we need to find a point $t_{17,\alpha/2}$, such that:

$P(t_{17} > t_{17,\alpha/2}) = \alpha / 2 = 0.05 \Rightarrow t_{17,0.05} = 1.74$ Therefore the 90% confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) - t_{17,0.05}s_x\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_{17,0.05}s_x\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $\bar{x}_1 - \bar{x}_2 = -5.32$ and $s = 4.7557$.

The required 95% confidence interval is therefore:

$$-5.32 - 1.74(4.7557)\sqrt{\frac{1}{9} + \frac{1}{10}} \le \mu_1 - \mu_2 \le -5.32 + 1.74(4.7557)\sqrt{\frac{1}{9} + \frac{1}{10}}$$

$\Rightarrow -9.121 \le \mu_1 - \mu_2 \le -1.519$.

Question 8

A researcher intends to estimate the effect of a drug on scores of human subjects performing a task of psychomotor coordination. The members of a random sample of 9 subjects were given the drug prior to testing, their mean score was 9.78 and the sample standard deviation was 3.2. An independent sample of 10 subjects were given a placebo prior to testing, the mean score for this group was 15.10 and the sample standard deviation was 7.2. Assuming the population distributions are normal, find a 90% confidence interval for the difference in the population means.

Answer

The underlying series has a normal distribution, but the population standard deviation is unknown:

$\bar{X}_1 \sim N(\mu_1, \sigma_{x_1}^2 / n_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma_{x_2}^2 / n_2)$

implying:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \left[\frac{\sigma_{x_1}^2}{n_1} + \frac{\sigma_{x_2}^2}{n_2}\right]\right)$$

The greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}\right)}} \sim t_{DoF} \text{, where } DoF = \frac{\left[3.2^2 / 9 + 7.2^2 / 10\right]^2}{\left(3.2^2 / 9\right)^2 / 8 + \left(7.2^2 / 10\right)^2 / 9} = 12.70.$$

To construct a 90% confidence interval we need to find a point $t_{12,\alpha/2}$, such that:

$$P(t_{12} > t_{12,\alpha/2}) = \alpha / 2 = 0.05 \Rightarrow t_{12,0.05} = 1.782$$

Therefore the 90% confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) - t_{12,0.05}\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_{12,0.05}\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}},$$ where

$\bar{x}_1 - \bar{x}_2 = -5.32$. The required 90% confidence interval is therefore:

$$-5.32 - 1.782\sqrt{\frac{3.2^2}{9} + \frac{7.2^2}{10}} \le \mu_1 - \mu_2 \le -5.32 + 1.782\sqrt{\frac{3.2^2}{9} + \frac{7.2^2}{10}}$$

$$\Rightarrow -9.800 \le \mu_1 - \mu_2 \le -0.839$$

Question 9

A random sample of 15 financial analysts' forecasts of next years' earnings per share for a large corporation was taken. The sample standard deviation was £0.88. Find a 95% confidence interval for the variance of predicted earnings per share for all analysts.

Answer

Assuming the underlying distribution of predicted earnings per share is normal. Then the sample variance of the predicted earnings will follow a chi-squared distribution.

$$\frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$$

To construct a 95% confidence interval we need to find a the points $\chi_{\alpha/2}^2$ and $\chi_{1-\alpha/2}^2$, such that:

$$P(\chi_{1-\alpha/2}^2 > \chi_{n-1}^2 > \chi_{\alpha/2}^2) = \alpha/2 = 0.025 \Rightarrow \chi_{0.025}^2 = 26.12 \Rightarrow \chi_{0.975}^2 = 5.63$$

Therefore the 95% confidence interval is:

$$\frac{(n-1)s_x^2}{\chi_{0.025}^2} \le \sigma^2 \le \frac{(n-1)s_x^2}{\chi_{0.975}^2}$$

where $s^2 = 0.88$.

The required 95% confidence interval is therefore:

$$\frac{14(0.88)^2}{26.12} \le \sigma^2 \le \frac{14(0.88)^2}{5.63} \Rightarrow 0.415 \le \sigma^2 \le 1.925$$

Confidence Intervals Sheet

Confidence Intervals for Means

Parameter	Distrib. of X_i	Sample	Variance	Confidence Interval
One Population				
μ	Normal	Large/Small	Known	$\bar{x} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
μ	Normal	Large/Small	Not Known	$\bar{x} - t_{n-1}^{\alpha/2}\dfrac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1}^{\alpha/2}\dfrac{s}{\sqrt{n}}$
μ	Non-Normal	Large	Known	$\bar{x} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
μ	Non-Normal	Large	Not Known	$\bar{x} - z_{\alpha/2}\dfrac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\dfrac{s}{\sqrt{n}}$
μ	Non-Normal	Small	Known/Not Known	$? \le \mu \le ?$
Two Populations				
$\mu_1 - \mu_2$	Normal	Large/Small	Known	$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
$\mu_1 - \mu_2$	Normal	Large/Small	Not Known (Equal)	$(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2}^{\alpha/2}\sqrt{\dfrac{s_0^2}{n_1} + \dfrac{s_0^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2}^{\alpha/2}\sqrt{\dfrac{s_0^2}{n_1} + \dfrac{s_0^2}{n_2}}$
where, $s_0^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$				
$\mu_1 - \mu_2$	Normal	Large/Small	Not Known (Unequal)	$(\bar{x}_1 - \bar{x}_2) - t_{DoF}^{\alpha/2}\sqrt{\dfrac{s_{x_1}^2}{n_1} + \dfrac{s_{x_2}^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_{DoF}^{\alpha/2}\sqrt{\dfrac{s_{x_1}^2}{n_1} + \dfrac{s_{x_2}^2}{n_2}}$

where, $DoF = \dfrac{\left[s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2 \right]^2}{\left(s_{x_1}^2 / n_1 \right)^2 / (n_1 - 1) + \left(s_{x_2}^2 / n_2 \right)^2 / (n_2 - 1)}$

$\mu_1 - \mu_2$	Non-Normal	Large	Known	$(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$\mu_1 - \mu_2$	Non-Normal	Large	Not Known	$(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$\mu_1 - \mu_2$	Non-Normal	Small	Known/Not Known	$? \le \mu_1 - \mu_2 \le ?$

Confidence Intervals for Proportions

p	Non-Normal	Large	Not Known	$p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le \pi \le p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
p	Non-Normal	Small	Not Known	$? \le \pi \le ?$
$p_1 - p_2$	Non-normal	Large	Not Known	

$$(p_1 - p_2) - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \le \pi_1 - \pi_2 \le (p_1 - p_2) - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$p_1 - p_2$	Non-normal	Small	Not Known	$? \le \pi_1 - \pi_2 \le ?$

Confidence Intervals on variances

σ^2	Normal	Large/Small	Not Known	$\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}$

STATISTICAL TECHNIQUES B

Nonparametric Tests

1. Introduction

In Handout 5 (Hypothesis Testing) a number of our tests were reliant on the assumption of normality of the underlying distribution of X_i. For example, we learnt that if the underlying distribution is normal: (i) and the population variance is known the distribution of the resultant test statistic of the sample mean would be normal; (ii) and the population variance is unknown the distribution of the resultant test statistic of the sample mean would be a t-distribution. By virtue of a Central Limit Theorem, the distribution of the test statistic of the sample mean will be approximately normal, in large samples, even if the population distribution is not normal. So for example, we might have: The Thomas Pink Gold Cup (held in November) and the Cheltenham Gold Cup (in March) are 2-mile National Hunt jumps races for horses at Cheltenham Racecourse. The same nine two-year old horses were timed in each race of these races and we are interested in testing the hypothesis $H_0 : \mu_d = 0$ (mean time difference is zero) against a 2-sided alternative at the 5% significance level. The times taken were as follows (in minutes):

PARAMETRIC TEST

Thomas Pink	8.1	8.2	8.0	8.0	8.4	8.6	8.5	8.4	8.9
Cheltenham	8.3	8.4	8.3	8.5	8.5	8.2	8.9	8.5	9.0
Difference	-0.2	-0.2	-0.3	-0.5	-0.1	+0.4	-0.4	-0.1	-0.1

Matched pairs: $\bar{x}_d = -0.167, s_d = 0.255$ if underlying distribution is normal then $\bar{X}_d \sim N(\mu_d, \sigma_d^2 / n)$ and with σ_d^2 unknown we have: $\frac{\bar{X}_d - \mu_d}{\sqrt{s_d^2 / n}} \sim t_{n-1}$. In which case:

$\Pr\left(t_{n-1} < \frac{(-0.167 - 0)}{\sqrt{0.255^2 / 9}} \right) = \Pr\left(t_{n-1} < -1.960\right) = 0.043$ and for a 2-sided test this probability is 0.086 and we do NOT reject H$_0$.

However, it often is the case that the normality assumption is not reasonable and/or the sample size is not large. In these cases it is desirable to base inference on tests which are valid over a wide range of distributions of X_i (although they do require certain assumptions to be valid, e.g. independent random samples). These tests are often referred to as nonparametric tests.

2. Sign Test (Wilcoxon)

This is the simplest test to undertake and is used for testing hypotheses about the central location of a population distribution. This is most frequently used in analysing matched pairs data and is based on assigning a plus(+) if the value from the 1st sample is greater than that from the 2nd sample and a minus(-) if the value from the 1st sample is less than that from the 2nd sample and dropping those cases where the two values are equal. The null hypothesis is that in the population the two values have the same mean.

Based on the sample of *n* observations, for which there was + or - recorded, under the null hypothesis the number of + and - values should equal, such that Pr(+)=0.5 and the Pr(-)=0.5. Consider only + values and denote *p* as the true proportion of +'s in the population, then $H_0 : p = 0.5$ and the distribution of *W*, the number of + values follows a Binomial distribution, *W*~*B*(*n*,0.5) . For an alternative hypothesis $H_1 : p < 0.5\left(p > 0.5\right)$, we want $\Pr(W \le w)$ $\left(\Pr(W \ge w)\right)$, for a 1-sided test and for an alternative hypothesis $H_1 : p \ne 0.5$, we want $2 \times \Pr(W \le w)$ (Appendix 2: Example 2)

Note that for large *n* (>25) $W / n \overset{a}{\sim} N(0.5, 0.25 / n) \Rightarrow \dfrac{W / n - 0.5}{\sqrt{0.25 / n}} \sim N(0,1)$ (Appendix 2: Example 2).

Thomas Pink	8.1	8.2	8.0	8.0	8.4	8.6	8.5	8.4	8.9
Cheltenham	8.3	8.4	8.3	8.5	8.5	8.2	8.9	8.5	9.0
	-	-	-	-	-	+	-	-	-

$\Pr(W \le 1) = {}_9C_0 (0.5)^9 + {}_9C_1 (0.5)^9 = 0.020$ and for a 2-sided alternative hypothesis the p-value is 0.040 and we reject H_0.

3. Wilcoxon Signed Rank Test

The problem with the Sign Test is that it only uses a very limited amount of information (namely the sign of the difference) and therefore ignores the strength of preference of one value over the other. As a result the test can lack power in small samples. The signed rank test, uses not only the sign of the difference, but also the magnitude of the difference. This test is also applied to matched pairs and is testing the null hypothesis $H_0 : \mu_d = 0$, where μ_d is the population mean difference in scores across the matched pairs and as with the Sign Test differences of 0 are ignored. The nonzero absolute differences are then ranked in ascending order of magnitude (where equal values are assigned the average rank). The ranks of positive and negative differences are then summed separately as $W_+ = \sum_{i=1}^{n} \phi_i^+ R_i$ and $W_- = \sum_{i=1}^{n} \phi_i^- R_i$, where $\phi_i^+ = \begin{cases} 1 & \text{if difference positive} \\ 0 & \text{otherwise} \end{cases}$, $\phi_i^- = \begin{cases} 1 & \text{if difference negative} \\ 0 & \text{otherwise} \end{cases}$ and R_i is the rank of the absolute value of the difference in the scores for the i^{th} value.

We denote the Wilcoxon signed rank test as $T = \min(W_+, W_-)$. For a 1-sided alternative hypotheses $H_1 : \mu_d < 0 \left(H_1 : \mu_d > 0\right)$, you want $\Pr(T \le cv)$, but for a 2-sided alternative hypothesis $H_1 : \mu_d \ne 0$, you want $2 \times \Pr(T \le cv)$ and for small samples these probabilities are based on critical values (*cv*) reported in the Table below (Appendix 2: Example 3).

Under the null hypothesis that the true population difference in scores across the matched pairs is zero, it can be shown that:

$$E(T) = \frac{n(n+1)}{4} \text{ and } V(T) = \frac{n(n+1)(2n+1)}{24}.$$

For *n*>25 then $T \overset{a}{\sim} N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$ and therefore

$$\frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \overset{a}{\sim} N(0,1) \text{ (Appendix 2: Example 4).}$$

Table 1: Critical values of the Wilcoxon Signed Rank Test (*n*<30)

1-tailed	*α*=0.05	*α*=0.025	*α*=0.01	*α*=0.005
2-tailed	*α*=0.10	*α*=0.05	*α*=0.02	*α*=0.01
n				
6	2	0	-	-
7	3	2	0	-
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15
16	35	29	23	19
17	41	34	27	23
18	47	40	32	27
19	53	46	37	32
20	60	52	43	37
21	67	58	49	42
22	75	65	55	48
23	83	73	62	54
24	91	81	69	61
25	100	89	76	68

Thomas Pink	8.1	8.2	8.0	8.0	8.4	8.6	8.5	8.4	8.9
Cheltenham	8.3	8.4	8.3	8.5	8.5	8.2	8.9	8.5	9.0
Difference	-0.2	-0.2	-0.3	-0.5	-0.1	+0.4	-0.4	-0.1	-0.1
Rank	(-)4.5	(-)4.5	(-)6	(-)9	(-)2	(+)7.5	(-)7.5	(-)2	(-)2

$W_- = 37.5$, $W_+ = 7.5$, so $T = 7.5$ and this is greater than the critical value, *cv*=5, and so we do not reject H$_0$.

4. Mann-Whitney Test

This test compares the central location of two populations, but in this case the samples come from independent random samples. Suppose that n_1 observations are available from the first population and n_2 from the second. All observations ($n = n_1 + n_2$) are then ranked in ascending order of magnitude (where equal values are assigned the average rank) and we denote R_1 as the sum of ranks of observations from the first population. The Mann-Whitney test is then defined as:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

For small sample the critical values are reported in Table 2. For this statistics we know:

$$E(U) = \frac{n_1 n_2}{2} \text{ and } V(U) = \frac{n_1 n_2(n_1+n_2+1)}{12}$$

And for samples (n>25) $U \overset{a}{\sim} N\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$ implying

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \overset{a}{\sim} N(0,1)$$. (Appendix 2: Example 5)

8.1	8.2	8.0	8.0	8.4	8.6	8.5	8.4	8.9	8.3	8.4	8.3	8.5	8.5	8.2	8.9	8.5	9.0
8.0	8.0	8.1	8.2	8.2	8.3	8.3	8.4	8.4	8.4	8.5	8.5	8.5	8.5	8.6	8.9	8.9	9.0
1.5	1.5	3	4.5	4.5	6.5	6.5	9	9	9	12.5	12.5	12.5	12.5	15	16.5	16.5	18

where red is Thomas Pink and Green is Cheltenham. Then $R_1 = 72.5$ and $R_{12} = 98.5$, in which case: $U_1 = 81 + 45 - 72.5 = 53.5$ and $U_2 = 81 + 45 - 98.5 = 27.5$ and $U = \min(U_1, U_2) = 27.5$, with a 5% cv=18 we do not reject H_0.

Alternatively, we know, $E(U) = 40.5$ and $V(U) = 128.5$, in which case:

$\Pr\left(z > \frac{(53.5 - 40.5)}{\sqrt{128.5}}\right) = \Pr(z > 1.148) = 0.125$ and for a 2-sided test this probability is 0.250, in which case we do NOT reject H_0.

Table 2: Mann-Whitney 1% Critical Values (2-sided)

									n_2								
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	5	1	2	3	3	4	5	6	7	8	8	10	10	11	12	13	13
	6	2	3	4	5	6	7	8	9	10	12	13	14	16	17	18	19
	7	3	4	5	7	8	9	11	12	14	15	17	18	20	21	23	24
	8	3	5	7	9	10	12	14	15	17	19	21	23	24	26	28	30
	9	4	6	8	10	13	14	16	19	21	23	25	28	30	32	34	36
	10	5	7	9	12	14	17	19	21	25	27	29	32	35	38	40	42
	11	6	8	11	14	16	19	22	24	28	30	32	36	39	42	45	48
n_1	12	7	9	12	15	19	17	24	28	30	34	37	41	44	47	51	55
	13	8	10	14	17	21	21	28	30	35	38	42	46	50	53	57	61
	14	8	12	15	19	23	25	30	34	38	43	47	51	56	59	63	66
	15	10	13	17	21	25	27	32	37	42	47	51	56	61	64	68	72
	16	10	14	18	23	28	29	36	41	46	51	56	61	66	70	74	78
	17	11	16	20	24	30	32	39	44	50	56	61	66	71	76	82	86
	18	12	17	21	26	32	35	42	47	53	59	64	70	76	83	88	93
	19	13	18	23	28	34	40	45	51	57	63	68	74	82	88	94	99
	20	13	19	24	30	36	42	48	55	61	66	72	78	86	93	99	107

Table 2: Mann-Whitney 5% Critical Values (2-sided)

									n_2								
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	5	3	4	5	7	8	9	10	11	13	14	15	16	18	19	20	21
	6	4	6	7	9	10	12	14	15	17	18	20	22	23	25	26	28
	7	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35
	8	7	9	11	14	16	18	21	23	25	28	30	32	35	37	40	42
	9	8	10	13	16	18	21	24	26	29	32	35	37	40	42	46	48
	10	9	12	15	18	21	24	27	30	34	37	39	43	46	49	52	55
	11	10	14	17	21	24	27	30	34	37	41	44	48	51	55	59	62
n_1	12	11	15	19	23	26	24	34	38	41	45	49	53	57	61	65	69
	13	13	17	21	25	29	30	37	41	46	50	54	58	62	67	71	75
	14	14	18	23	28	32	34	41	45	50	55	59	64	68	74	78	84
	15	15	20	25	30	35	37	44	49	54	59	64	69	75	80	86	91
	16	16	22	27	32	37	39	48	53	58	64	69	76	81	87	93	99
	17	18	23	29	35	40	43	51	57	62	68	75	81	87	94	100	106
	18	19	25	31	37	42	46	55	61	67	74	80	87	94	99	106	113
	19	20	26	33	40	46	52	59	65	71	78	86	93	100	106	113	120
	20	21	28	35	42	48	55	62	69	75	84	91	99	106	113	120	128

5. Goodness-of-fit test

Suppose that we are given a random sample of *n* observations, each of which can be classified into exactly one of *K* categories. Denote the observed number of cases in each category as $O_1, O_2, O_3, \ldots, O_K$. If a null hypothesis ($H_0$) specifies probabilities $p_1, p_2, p_3, \ldots, p_K$ for an observation falling into each of these categories, the expected numbers in each category, under H_0, would be $E_i = np_i \qquad (i = 1, 2, \ldots, K)$

We then test whether the actual data is a close fit to the expect data (based on some assumed population distribution for probabilities) – and this is done by looking at the magnitude of the discrepancy between the observed and expected values. Where large (absolute) values ought to make one increasingly suspicious of the null hypothesis. The test is constructed as (see Appendix 1 for a proof of this equivalence):

$$\sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{K} \frac{O_i^2}{E_i} - n \sim \chi^2_{(K-1)}$$

And H_0 is rejected at significance level α, if $\sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{(K-1),\alpha}$. (Appendix 2: Example 6)

In November 2015, Economists were asked about inflation expectations for December 2016, with 10% reporting <1%, 40% reporting 1-2%, 40% reporting 2-3% and 10% reporting >3%. In November 2017, 80 Economists reported their inflation expectations for December 2018:

Range	<1%	1-2%	2-3%	>3%
Frequency	12	15	33	20

Test the hypothesis the distribution has not changed.

Range	<1%	1-2%	2-3%	>3%
Frequency	12	15	33	20
Expected	0.1×80=8	0.4×80=32	0.4×80=32	0.1×80=8

$\frac{12^2}{8} + \frac{15^2}{32} + \frac{33^2}{32} + \frac{20^2}{8} - 80 = 29.1$, $\chi^2_{3,0.05} = 7.81$ and so we reject H0.

6. Contingency Tables

Suppose we have two attributes *A* and *B*. There are *K* categories in *A* and *H* in *B* so that there are *KH* cross-classifications in total. The number of sample observations belonging to the *i*th category of *A* and the *j*th category of B is denoted as O_{ij} and there are *n* observations in total. To test the null hypothesis of no association (independence) between the two attributes, we want to know how many observations were would expect to find in each cross-classification. Under the null hypothesis of independence between the two attributes A and B we know that the joint probability (p_{ij}) is equal to the product of the marginal probabilities ($p_i.p_j$), in other words:

$$p_{ij} \equiv \Pr(A=i, B=j) = \Pr(A=i).\Pr(B=j) \equiv p_i.p_j\,.$$

Now $p_i \equiv \Pr(A=i) = \sum_{j=1}^{H} O_{ij}/n$ and $p_j \equiv \Pr(B=j) = \sum_{i=1}^{K} O_{ij}/n$. In which case under the null hypothesis of independence the expected number of observations is

$$E_{ij} = np_{ij} = \sum_{j=1}^{H} O_{ij} \sum_{i=1}^{K} O_{ij}/n\,.$$

We then test whether the actual data is a close fit to the expect data and this is done by looking at the magnitude of the discrepancy between the observed and expected values. Where large (absolute) values ought to make one increasingly suspicious of the null hypothesis. The test is constructed as:

$$\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{K}\sum_{j=1}^{H}\frac{O_{ij}^{\ 2}}{E_{ij}} - n \sim \chi^2_{(K-1)(H-1)}$$

And H_0 is rejected at some significance level α , if $\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} > \chi^2_{(K-1)(H-1),\alpha}$.

(Appendix 2: Example 7)

150 Economists were asked about inflation expectations for December 2019 in both November 2016 and in November 2018 and the results were reported as follows:

		Nov 2016				
		<1%	1-2%	2-3%	>3%	
Nov 2018	<1%	6	10	8	6	30
	1-2%	8	12	12	8	40
	2-3%	8	12	20	10	50
	>3%	8	6	10	6	30
		30	40	50	30	150

Test the hypothesis that there is no association between these two series.

Under independence

$$P(<1\%,<1\%)=\frac{30}{150}\times\frac{30}{150}=\frac{900}{22500}=0.040,\ P(<1\%,1-2\%)=\frac{30}{150}\times\frac{40}{150}=\frac{1200}{22500}=0.0533$$

$$P(<1\%,2-3\%)=\frac{30}{150}\times\frac{50}{150}=\frac{1500}{22500}=0.0667,\ P(<1\%,>3\%)=\frac{30}{150}\times\frac{30}{150}=\frac{900}{22500}=0.040,$$

$$P(1-2\%,<1\%)=\frac{40}{150}\times\frac{30}{150}=\frac{120}{22500}=0.0533,\ P(1-2\%,1-2\%)=\frac{40}{150}\times\frac{40}{150}=\frac{1600}{22500}=0.0711$$

$$P(1-2\%,2-3\%)=\frac{40}{150}\times\frac{50}{150}=\frac{2000}{22500}=0.0889,\ P(1-2\%,>3\%)=\frac{40}{150}\times\frac{30}{150}=\frac{1200}{22500}=0.0533,$$

$$P(2-3\%,<1\%)=\frac{50}{150}\times\frac{30}{150}=\frac{1500}{22500}=0.0667,\ P(2-3\%,1-2\%)=\frac{50}{150}\times\frac{40}{150}=\frac{2000}{22500}=0.0889,$$

$$P(2-3\%,2-3\%)=\frac{50}{150}\times\frac{50}{150}=\frac{2500}{22500}=0.111,\ P(2-3\%,>3\%)=\frac{50}{150}\times\frac{30}{150}=\frac{1500}{22500}=00667,$$

$$P(>3\%,<1\%)=\frac{30}{150}\times\frac{30}{150}=\frac{900}{22500}=0.040\quad P(>3\%,1-2\%)=\frac{30}{150}\times\frac{40}{150}=\frac{1200}{22500}=0.0533$$

$$P(>3\%,2-3\%)=\frac{30}{150}\times\frac{50}{150}=\frac{1500}{22500}=0.0667,\ P(>3\%,>3\%)=\frac{30}{150}\times\frac{30}{150}=\frac{90}{22500}=0.040$$

Expected numbers

	<1%	1-2%	2-3%	>3%
<1%	0.040×150=6	0.053×150=8	0.066×150=10	0.040×150=6
1-2%	0.053×150=8	0.071×150=10.6	0.088×150=13.3	0.053×150=8
2-3%	0.067×150=10	0.088×150=13.3	0.111×150=16.6	0.067×150=10
>3%	0.040×150=6	0.053×150=8	0.066×150=10	0.040×150=6
	30	40	50	30

$$\frac{6^2}{6}+\frac{10^2}{8}+\frac{8^2}{10}+\frac{6^2}{6}+\frac{8^2}{8}+\frac{12^2}{10.6}+\frac{12^2}{13.3}+\frac{8^2}{8}+\frac{8^2}{10}+\frac{12^2}{13.3}+\frac{20^2}{16.6}+\frac{10^2}{10}+\frac{8^2}{6}+\frac{6^2}{8}+\frac{10^2}{10}+\frac{6^2}{6}-150=3.567$$

$\chi^2_{9,0.05}=16.92$, therefore we DO NOT reject H_0.

Equivalence of Contingency Tests

$$\sum_{i=1}^{K}\frac{(O_i-E_i)^2}{E_i}=\sum_{i=1}^{K}\frac{O_i^2+E_i{}^2-2O_iE_i}{E_i}=\sum_{i=1}^{K}\left(\frac{O_i^2}{E_i}+E_i-2O_i\right)$$

$$=\sum_{i=1}^{K}\frac{O_i^2}{E_i}+\sum_{i=1}^{K}E_i-2\sum_{i=1}^{K}O_i$$

As $\sum_{i=1}^{K}E_i=\sum_{i=1}^{K}O_i=n$

$$=\sum_{i=1}^{K}\frac{O_i^2}{E_i}+n-2n=\sum_{i=1}^{K}\frac{O_i^2}{E_i}-n$$

Sample Questions

Question 1

A random sample of twelve financial analysts was asked to predict the percentage increases in the prices of two common stocks over the next year. The results obtained are shown in the table below. Use the sign test to test the null hypothesis that for the population of analysts, there is no overall preference for one stock over the other:

Analyst	Stock 1	Stock 2	Analyst	Stock 1	Stock 2
A	6.8	7.1	G	9.3	10.1
B	9.8	12.3	H	1.0	2.7
C	2.1	5.3	I	-0.2	1.3
D	6.2	6.8	J	9.6	9.8
E	7.1	7.2	K	12.0	12.0
F	6.5	6.2	L	6.3	8.9

Question 2

In a random sample of 130 voters, 44 favoured tax increases to raise funding for education, 68 opposed the tax increase, and 18 expressed no opinion. Test against a 2-sided alternative the null hypothesis that voters in the state are evenly divided on the issue of a tax increase.

Question 3

Using the data in (1), test the null hypothesis that for the population of analysts, there is no difference in the mean performance of one stock over the other, using the Wilcoxon Signed Rank Test.

Question 4

A consultant id interested in the impact of the introduction of a quality management program on job satisfaction of employees. A random sample of 30 employees was asked to assess level of satisfaction on a scale of 1 (very dissatisfied) to 10 (very satisfied) 3 months before the introduction of the program. These same individuals were then asked to make this assessment again 3 months after the introduction of the program. The 30 differences in the pairs of rating were calculated and the absolute differences ranked. The smaller of the rank sums, which was for those more satisfied before the introduction of the program was 169. What can be concluded from these findings?

Question 5

A random sample of 15 male and an independent random sample of 15 female students were asked to write essays at the conclusion of their writing module. Essays were then ranked from 1 (best) to 30 (worst) by the module leader as:

Males	26	24	15	16	8	29	12	6	18	11	13	19	10	28	7
Females	22	2	17	25	14	21	5	30	3	9	4	1	27	23	20

Question 6

A random sample of 520 customers were asked about the importance of quality of food as a factor in choosing a hospital. Sample members were asked to respond as “not important”, “important”, or “very important”. Respective numbers selecting these answers were: 199, 136 and 167. Test the null hypothesis that a randomly chosen consumer is equally likely to select each of these answers.

Question 7

In a series of surveys, 55 forecasters were asked whether they thought inflation would increase over the next 12 months from its current level. It was also noted whether or not actual inflation increased. The results are reported in the table below:

Outcome	Forecast	
	Increase	No increase
Increase	18	11
No increase	6	20

Test the null hypothesis of no association between forecast and outcome.

Sample Questions (with Answers)

Question 1

A random sample of twelve financial analysts was asked to predict the percentage increases in the prices of two common stocks over the next year. The results obtained are shown in the table below. Use the sign test to test the null hypothesis that for the population of analysts, there is no overall preference for one stock over the other:

Analyst	Stock 1	Stock 2	Analyst	Stock 1	Stock 2
A	6.8	7.1	G	9.3	10.1
B	9.8	12.3	H	1.0	2.7
C	2.1	5.3	I	-0.2	1.3
D	6.2	6.8	J	9.6	9.8
E	7.1	7.2	K	12.0	12.0
F	6.5	6.2	L	6.3	8.9

Answer

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$

n=11 with 1 + value and 10 – values, we want

$$2 \times \Pr(W \leq 1) = 2 \times [\Pr(W = 0) + \Pr(W = 1)] = 2 \times [0.0005 + 0.0054] = 0.0118$$

and so we reject H0 at significance levels in excess of 1.18%.

Question 2

In a random sample of 130 voters, 44 favoured tax increases to raise funding for education, 68 opposed the tax increase, and 18 expressed no opinion. Test against a 2-sided alternative the null hypothesis that voters in the state are evenly divided on the issue of a tax increase.

Answer

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$

n=130-18=112, $W / n = \hat{p} = 44/112 = 0.3929$

$$z = \frac{0.3929 - 0.5}{\sqrt{0.25/112}} = -2.27$$

p-value=2[1-Φ(2.27)]=0.0232 and so we reject H_0 at significance levels in excess of 2.32%

Question 3

Using the data in (1), test the null hypothesis that for the population of analysts, there is no difference in the mean performance of one stock over the other, using the Wilcoxon Signed Rank Test.

Answer

$H_0 : \mu_d = 0$

$H_1 : \mu_d \neq 0$

n=11 with 1 + value and 10 – values, we want:

Analyst	Stock 1	Stock 2	Stk1-Stk2	$\phi_i^+ R_i$	$\phi_i^+ R_i$
A	6.8	7.1	-0.3		3.5
B	9.8	12.3	-2.5		9
C	2.1	5.3	-3.2		11
D	6.2	6.8	-0.6		5
E	7.1	7.2	-0.1		1
F	6.5	6.2	+0.3	3.5	
G	9.3	10.1	-0.8		6
H	1.0	2.7	-1.7		8
I	-0.2	1.3	-1.5		7
J	9.6	9.8	-0.2		2
K	12.0	12.0	0.0		
L	6.3	8.9	-2.6		10
				3.5	62.5

From this we have *T*=3.5, with critical value of 10 (at the 5% significance level for a 2-sided test), we reject H_0.

Question 4

A consultant id interested in the impact of the introduction of a quality management program on job satisfaction of employees. A random sample of 30 employees was asked to assess level of satisfaction on a scale of 1 (very dissatisfied) to 10 (very satisfied) 3 months before the introduction of the program. These same individuals were then asked to make this assessment again 3 months after the introduction of the program. The 30 differences in the pairs of rating were calculated and the absolute differences ranked. The smaller of the rank sums, which was for those more satisfied before the introduction of the program was 169. What can be concluded from these findings?

Answer

$H_0 : \mu_d = 0$

$H_1 : \mu_d < 0$

T=169, $E(T) = \frac{30(31)}{4} = 232.5, V(T) = \frac{30(31)(61)}{24} = 2363.75$

$z = \frac{169 - 232.5}{\sqrt{2363.5}} = -1.31 \Rightarrow$ p-value=1-Φ(1.31)=0.0951 and so we reject H_0 at significance levels in excess of 9.51%.

Question 5

A random sample of 15 male and an independent random sample of 15 female students were asked to write essays at the conclusion of their writing module. Essays were then ranked from 1 (best) to 30 (worst) by the module leader as:

Males	26	24	15	16	8	29	12	6	18	11	13	19	10	28	7
Females	22	2	17	25	14	21	5	30	3	9	4	1	27	23	20

Answer

n_{m}=15, R_{m}=242, n_{f}=15 R_{f}=223.

$H_0 : \mu_d = 0$

$H_1 : \mu_d \neq 0$

$U = 15(15) + \frac{15(16)}{2} - 242 = 103$

$E(U) = \frac{15(15)}{2} = 112.5, V(U) = \frac{15(15)(15+15+1)}{12} = 581.25$

$z = \frac{103 - 112.5}{\sqrt{581.25}} = -0.39 \Rightarrow$ p-value=2[1-Φ(0.39)]=0.6966 and so we reject H_0 at significance levels in excess of 69.66%.

Question 6

A random sample of 520 customers were asked about the importance of quality of food as a factor in choosing a hospital. Sample members were asked to respond as “not important”, “important”, or “very important”. Respective numbers selecting these answers were: 199, 136 and 167. Test the null hypothesis that a randomly chosen consumer is equally likely to select each of these answers.

Answer

H_0 : All outcomes equally likely

H_1 : otherwise

	Not imp	Imp	Very imp	Total
Observed	199	136	167	502
Prob (under *H*0)	0.333	0.333	0.333	1
Expected number	167.33	167.33	167.33	502

$$\sum_{i=1}^{K}\frac{O_i^2}{E_i}-502=\frac{199^2}{167.33}+\frac{136^2}{167.33}+\frac{167^2}{167.33}-502=11.86$$

$\chi^2_{2,0.01}=9.21\Rightarrow$ Reject H_0 at 1% signficance level.

Question 7

In a series of surveys, 55 forecasters were asked whether they thought inflation would increase over the next 12 months from its current level. It was also noted whether or not actual inflation increased. The results are reported in the table below:

Outcome	Forecast	
	Increase	No increase
Increase	18	11
No increase	6	20

Test the null hypothesis of no association between forecast and outcome.

Answer

H_0 : No association between forecast and outcome

H_1 : otherwise

Under H_0 (independence) the probability of being in each category is:

$P(Increase, Increase) = \frac{24}{55} \times \frac{29}{55} = 0.230$, $P(Increase, No\ Increase) = \frac{24}{55} \times \frac{26}{55} = 0.206$,

$P(No\ Increase, Increase) = \frac{31}{55} \times \frac{29}{55} = 0.297$, $P(Increase, No\ Increase) = \frac{31}{55} \times \frac{26}{55} = 0.266$

,

and expected number of observations in each category is:

Outcome	Forecast		
	Increase	No increase	
Increase	0.230×55	0.206×55	29
No increase	0.297×55	0.266×55	26
	24	31	55

Outcome	Forecast		
	Increase	No increase	
Increase	12.65	16.35	29
No increase	11.35	14.65	26
	24	31	55

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{O_{ij}^{\ 2}}{E_{ij}} - 55 = \frac{18^2}{12.65} + \frac{11^2}{16.35} + \frac{6^2}{11.35} + \frac{20^2}{14.65} - 55 = 8.48$$

$\chi^2_{1,0.01} = 6.63 \Rightarrow$ Reject H_0 at 1% signficance level.