QS101: Introduction to Quantitative Methods in Social Science

Week 9: Sampling Distributions

Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression

November 27, 2014

Probability Distributions

Probability Distributions for Discrete Variables

- Probability Distribution assigns a probability to each possible value of the variable
- Each probability is between 0 and 1
- Sum of all probabilities is equal to 1

$$0 \leq P(y) \leq 1 \text{ and } \sum_{all\ y} P(y) = 1$$

Example

y	$P(y)$
0	0.01
1	0.03
2	0.60
3	0.23
4	0.12
5	0.01
Total	1.0

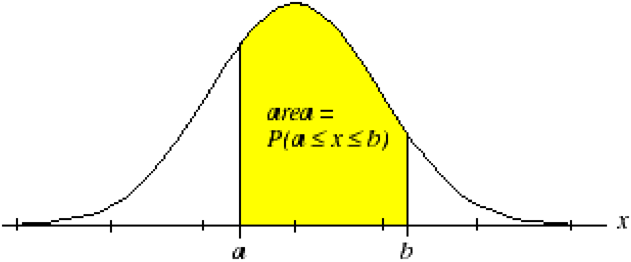
Table: Probability Distribution of y = Ideal Number of Children for a Family (Agresti and Finally, 2014, p. 76)

Probability Distributions for Continuous Variables

- Probabilities are assigned to *intervals* of numbers
- Probability for any interval is between 0 and 1
- Probability of the interval containing all possible numbers equals 1

Example

- Probability equals a particular area under the probability distribution



Parameters to Describe Probability Distributions

- Parameter values are the values measures would assume, in the long run, if a randomised experiment or random sample repeatedly took observations on the variable y
- **Mean**: Sum of possible outcomes times their probabilities

$$\mu = \Sigma y P(y) = E(y)$$

- This is also called the *expected value*
- **Standard Deviation**: is denoted by the Greek letter sigma (σ)

The Normal Probability Distribution

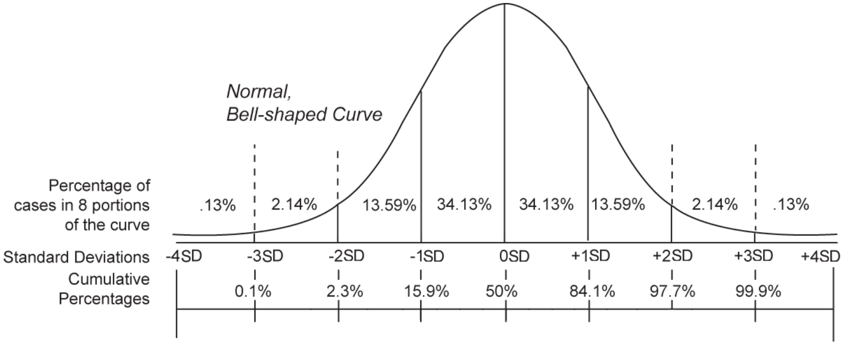
The Normal Distribution

- Probably the most useful and most frequently used distribution
- Has a familiar bell shape
- It is even useful when the sample data are *not* bell shaped

The Normal Distribution – Definition

"The **normal distribution** is symmetric, bell shaped, and characterised by its mean μ and standard deviation σ. The probability within any particular standard deviations of μ is the same for all normal distributions. This probability equals 0.68 within 1 standard deviation, 0.95 within 2 standard deviations, and 0.997 within 3 standard deviations." (Agresti and Finlay, 2014, p. 79)

The Normal Distribution



z-values

- For the normal distribution, for each fixed number of z, the probability of falling within z standard deviations of the mean depends only on the value of z.
- This is the area under the curve between $\mu - z\sigma$ and $\mu + z\sigma$
- For example: the probability is 0.68 for $z = 1$
- z does not need to be a whole number
- Many inferential methods use z-values, so we will encounter this again

z-scores

- The **z-score** for a value y of a variable is the number of standard deviations that y falls from μ. It equals

$$z = \frac{\text{Observation}-\text{Mean}}{\text{Standard Deviation}} = \frac{y-\mu}{\sigma}$$

- This is what you will usually find in normal tables.

Normal Right Tail Probabilities

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856

An Example

- Marks on QS101

An Example

- Marks on QS101
- Assume $\mu = 60$ and $\sigma = 10$

An Example

- Marks on QS101
- Assume $\mu = 60$ and $\sigma = 10$
- A mark of 45 has a z-score of $z = \frac{y-\mu}{\sigma} = -1.5$

An Example

- Marks on QS101
- Assume $\mu = 60$ and $\sigma = 10$
- A mark of 45 has a z-score of $z = \frac{y-\mu}{\sigma} = -1.5$
- Look up $z = 1.5$ in the normal table

An Example

- Marks on QS101
- Assume $\mu = 60$ and $\sigma = 10$
- A mark of 45 has a z-score of $z = \frac{y-\mu}{\sigma} = -1.5$
- Look up $z = 1.5$ in the normal table
- Value is 0.93319

An Example

- Marks on QS101
- Assume $\mu = 60$ and $\sigma = 10$
- A mark of 45 has a z-score of $z = \frac{y-\mu}{\sigma} = -1.5$
- Look up $z = 1.5$ in the normal table
- Value is 0.93319
- This means that fewer than 7% $(1 - 0.93319)$ of the marks are below 45

Extra Special: The Standard Normal Distribution

- The standard normal distribution is the normal distribution with a mean $\mu = 0$ and standard deviation $\sigma = 1$
- Then $\mu + z\sigma = 0 + z(1) = z$
- Therefore, the number falling z standard deviations above the mean is simply the z-score

Sampling Distributions

Why Sampling Distributions?

- We have now learned about probability distributions
- We have also assumed that we know the distribution in question
- This is rarely the case in practice
- Therefore, in practice we make inferences about the parameters of these distributions
- Probability distributions with fixed parameter values are useful for many of these inferential methods

Definition

- A **sampling distribution** of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take (Agresti and Finlay, 2014, p. 87)

Examples of Statistcs

- Sample mean
- Sample proportion
- Sample median
- ...

Examples of Statistcs

- Sample **mean**
- Sample proportion
- Sample median
- ...

The Sample Mean

- The sample mean is usually denoted $\bar{y}$
- In practice, we do not know how close it falls to the population mean μ, because we don't know μ
- We can predict, however, how close it will fall

The Sample Mean (contd.)

- The sample mean $\bar{y}$ is a variable, because its value varies from sample to sample we draw
- It fluctuates around the true mean of the population μ
- The mean of the sampling distribution $\bar{y}$ equals μ

The Standard Error

- The standard deviation of the of $\bar{y}$ of the sampling distribution is called the **standard error**
- It is denoted as $\sigma_{\bar{y}}$
- We could take samples repeatedly to find $\sigma_{\bar{y}}$ out, or we can use a simple formula:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

where n is the sample size.

Example

- Assume we want to know about the average age at Warwick
- The population distribution has $\mu = 36$ and $\sigma = 10$
- We take a sample of $n = 100$
- Therefore, $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$

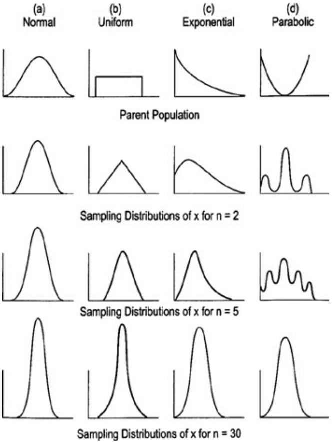
Sampling Error

- This process naturally involves an error, because we only sample part of the population
- This error is called the **sampling error**
- Due to the formula $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ it decreases with increasing sample size n.

The Central Limit Theorem

- Whatever the shape of the population distribution, the sampling distributions of $\bar{y}$ become increasingly bell shaped with increasing n

The Central Limit Theorem (contd.)



The Central Limit Theorem (contd.)

- For random sampling with a large sample size n (usually $n = 30$ is sufficient), the sampling distribution of the sample mean $\bar{y}$ is approximately a normal distribution. (Agresti and Finlay, 2014, p. 93)

Recap on Terminology

Population Distribution This is the distribution from which we select the sample. It is usually unknown. We can make inferences about its characteristics, such as the parameters μ and σ that describe its centre and spuread. The population size is usually denoted as N.

Sample Data Distribution This is the distribution of data that we actually observe; that is the sample observations $y_1, y_2, \ldots y_n$. We can describe it by statistics such as the sample mean $\bar{y}$ and sample standard deviation s. The larger the sample size n, the closer the sample data distribution resembles the population distribution, and the close the sample statistics such as $\bar{y}$ fall to the population parameters such as μ

Recap on Terminology

Sampling Distribution of a statistic: This is the probability distribution for the possible values of a sample statistic, such as $\bar{y}$. A sampling distribution describes the variability that occurs in the statistic's value among samples of a certain size.