Foundations of Data Analytics

October 30, 2018

1. Math Background

In Data Science (DS), data is typically presented in the form of *records*.

■ **Example 1.1** Consider the following records of some visits in a supermarket. ■

Gender	Age	Height
F	10	130
M	10	135
M	15	140
F	20	160

Table 1.1: Supermarket visits

Exercise 1.1

Q: Why would it be technically incorrect to think of data as a *set of records?*

A: Because duplicates are allowed and often they should not be removed. Consider a data set capturing temperature measurements at some location; one shouldn't remove duplicates! Nevertheless, with abuse of language, we will occasionally refer to data as 'data set' (with the understanding that record duplicates are allowed). ■

A common and convenient way to *model* data is through *random variables*. While the attribute *variable* should be familiar (data values are in general '*variable*' (can be different)), the attribute *random* captures the unknown nature of data. Note however that in Example 1.1, data is quite known; the 'random' attribute in this case would be appropriate for some *yet* to be determined record.

Exercise 1.2

Q: What are the random variables in Example 1.1?

A: Quite a lot! The most natural are the *univariate (single)* random variables G:=Gender, A:=Age, and H:=Height. For instance, the values of A are $\{10,10,15,20\}$. Oops! But this is a set, should I remove the duplicates? (answer later ...)

There are also the so-called *multivariate* random variables or *random vectors*. Quite a few of those: $(G,A),(G,H),(A,H),(G,A,H)$. Note that the last models the whole data.

R Order doesn't matter! Although (G,A) and (A,G) are technically '*different*', it is just a matter of choice to pick some convenient order.

Besides, we could define as many other random variables as imagination allows (e.g., the product of age and height).

■

But what are random variables anyway? Essentially, a random variable X is a *function*

$$X:\Omega\to\mathbb{R}$$

One can think of Ω as the '*(sample) space of what could happen*'. Note however that when data is already there (as in Example 1.1), then Ω is really the '*(sample) space of what has already happened*'. This remark is important as in the former case we deal with *theoretical* random variables, whereas in the latter case we deal with *empirical* random variables.

R As random variables take *real* values, the 'random variable' G mentioned earlier is not really a random variable (r.v.) as it takes categorical values. Yet, for convenience, we shall think of it as a r.v. as well.

Exercise 1.3

Q: What is Ω in Example 1.1?

A: A natural choice would be $\Omega=\{John,Mary,Ben,Susie\}$, hypothetically corresponding to the four customers. ■

One can also consider $\Omega=$ 'People' but note that there is no record corresponding to some hypothetical customer 'Rob'. To better understand such issues we have to introduce some necessary background.

1.1 Probability Measure

Random variables are intimately related to the concept of *probability* which is also a real function, but defined on the *power set* of Ω

$$\mathbb{P}:\mathscr{P}(\Omega)\to[0,1]$$

This function is used to *measure* (quantify) the likelihood (chance) of *events*, which are simply sets of *elementary events* (i.e., the very elements of Ω). More formally

Definition 1.1.1 — Probability.

The function $\mathbb{P}:\mathscr{P}(\Omega)\to[0,1]$ is called a probability measure if it satisfies

1. $$\mathbb{P}(\Omega)=1$$
   .
2. If $A_1,A_2,A_3,\ldots$ is a sequence of pairwise mutually exclusive events ($\forall i,j:A_i\cap A_j=\emptyset$),

then

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \cdots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \cdots$$

Note the tacit assumption that $\mathbb{P}$ must take values in $[0,1]$ (a mere yet very convenient convention). The two properties can be thought of as *axioms of probability*. The latter is a bit annoying (as it involves a countable (possibly infinite) sequence of events) but is there for good reasons.

Exercise 1.4

Starting from the axioms of probability prove the following

1. $$\mathbb{P}(\emptyset) = 0$$
2. Denoting A^c as the complementary event of A then
$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$
.
3. For two events A and B such that $A \subseteq B$ then
$$\mathbb{P}(A) \leq \mathbb{P}(B)$$
4. For two events A and B
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$
Note that if $A \cap B = \emptyset$ (the events are mutually exclusive) then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
5. Generalize the previous result to n events $A_1, A_2, \ldots, A_n$.

■

Let us consider few examples of *random experiments* involving the space Ω and measure $\mathbb{P}$.

■ **Example 1.2 — Tossing one coin.** The space is

$$\Omega = \{H, T\}$$

A *choice* for defining $\mathbb{P}$ would be

$$\mathbb{P}(H) = 1/2$$

It is actually sufficient to only specify $\mathbb{P}(H)$, as $\mathbb{P}(T)$ can be inferred from the axioms. There are also (uncountable) many choices of defining $\mathbb{P}(H)$, depending on the amount of information known about the experiment. As a side remark, the correct way of defining $\mathbb{P}$ is by writing $\mathbb{P}(\{H\})$ (see the definition of $\mathbb{P}$), but it is OK to abuse notation and drop the extra brackets when there are no ambiguities. ■

■ **Example 1.3 — Tossing two coins.** The space is

$$\Omega = \{HH, HT, TH, TT\}$$

Do keep in mind that each point in Ω is called an elementary event (e.g., $\{HT\}$) whereas the associated events are subsets of Ω (e.g., $\{HT, TT\}$).

A *choice* for defining $\mathbb{P}$ would be

$$\mathbb{P}(HH) = 1/4,\ \mathbb{P}(HT) = 1/4,\ \mathbb{P}(TH) = 1/4,\ \mathbb{P}(TT) = 1/4$$

Again, defining only three values would suffice; also, there are (uncountable) many alternative ways of defining $\mathbb{P}$.

What happens if only $\mathbb{P}(HH) = 1/4$ and $\mathbb{P}(HT) = 1/4$ were assigned? Then what is called as the *probability space* (space + assigned probabilities to the events) would be incompletely defined and it would not have any or much use. ■

R The last example hints at why $\mathbb{P}$ is defined on the powerset of Ω; for instance, one may be interested in the *probability* of the event $\{HT, TH, TT\}$ that at least one tail shows, which can be computed using the axioms. The power set captures all the possible (appropriate) questions we could ask about the experiment!

Exercise 1.5

Q: For the same experiment, one may also be interested in the probability that '*nothing shows*', which is immediately provided by the axioms, i.e., $\mathbb{P}(\emptyset) = 0$. While that makes sense, one may argue that if the random experiment was performed on the Moon then $\mathbb{P}(\emptyset) > 0$. How do we resolve this apparent contradiction of the axioms themselves?

A: Simply, we would be dealing with a different random experiment, with a different space

$$\Omega = \{\text{still-levitating}, HH, HT, TH, TT\}$$

One can of course include the elementary events $\{HL\}$ (first coin lands as H whereas the second is still levitating), etc. For this space, one should reassign the probabilities to the elementary events. ■

■ **Example 1.4 — A student answering a question.** The space is

$$\Omega = \{\text{student}_1,\ \text{student}_2,\ \text{etc.}\}$$

whereas the probabilities should be properly assigned.

Q: Is there another space? Yes, many, e.g.,

$$\Omega = \{(\text{student}_1, \text{good-answer}), (\text{student}_1, \text{bad-answer}),\ \text{etc.}\}$$

It depends on the sort of events one needs to study (i.e., measure their chance of occurrence). ■

1.2 Probability and Random Variables

We can now more formally put together the two concepts. As we have already mentioned, random variables are defined on a sample space Ω, defined itself for some random experiment. Moreover, random variables model numerical data (keep in mind that it is often more convenient to work with numbers than with records, as we could take full advantage of the power of mathematics in terms of what we could do with the (data) records). Here are some basic examples:

■ **Example 1.5** In the 'Tossing two coins' experiment, one may interested to model the earnings, assuming that each heads yields 10 (£) whereas each tails yields nothing

$$X := \text{'earnings', i.e., } X(HH) = 20,\ X(HT) = 10,\ X(TH) = 10,\ X(TT) = 0$$

One can also consider whatever other function on the elementary events (e.g., 'the square-root of the number of heads plus 10').

The intimate connection between probability and random variables becomes obvious when we ask, for instance, to quantify

'What is the probability of earning at least 7?′

This can be nicely expressed as

$$\mathbb{P}(X \geq 7) = \mathbb{P}(\{HH, HT, TH\}) = 3/4$$

A perhaps more interesting information that we may interested in is

$$\text{'What is the everage earnings?'}$$

One has to be very careful with the precise meaning of the attribute *average*. In this context it means that the same experiment is repeated over and over again. If one is interested in the *theoretical average* (denoted by $E[X]$), then that corresponds to the average earnings had the experiment been repeated infinitely over, i.e.,

$$E[X] = 20\,\mathbb{P}(HH) + 10\,\mathbb{P}(HT) + 10\,\mathbb{P}(TH) + 0\,\mathbb{P}(TT) = 10$$

If the experiment has already taken place, say for n times, then we would be dealing with an *empirical average*

$$E[X] = \frac{e_1 + e_2 + \cdots + e_n}{n}$$

where e_i is the earning in the i^{th} experiment (i.e., the entry in the data set). ■

■ **Example 1.6** In the 'A student answering a question' experiment, one may define

$X :=$ 'right or wrong', i.e.,, $X((\text{student}_1, \text{good-answer})) = 1$, $X((\text{student}_1, \text{bad-answer})) = 0$, etc.

Asking for the 'probability of receiving a correct answer', that is

$$\mathbb{P}(X = 1) = \mathbb{P}\left(\{(\text{student}_1, \text{good-answer}), (\text{student}_2, \text{good-answer}),\ \text{etc.}\}\right)$$

Certainly, this random variable may seem a bit superfluous (we could answer questions by circumventing it!).

Q: What is the (theoretical) average of X? That is simply $\mathbb{P}(X = 1)$.

We could also define more involved random variables, e.g., $X :=$ 'reward', e.g.,

$$X((\text{student}_i, \text{good-answer})) = 10 \text{ if student}_i \text{ is in an MEng degree}$$
$$X((\text{student}_i, \text{good-answer})) = 15 \text{ if student}_i \text{ is in an MSc degree}$$
$$X((\text{student}_i, \text{bad-answer})) = -5\ .$$

This random variable would certainly simplify calculations involving events related to earnings.

Q: Give an example in which $E[X] = -5$! Take $\Omega = \{(\text{J}, \text{bad-answer})\}$. Give another example in which $E[X] = -10$! Impossible!

■

1.3 Independence

This is arguably the concept from probability theory being by far responsible for most errors. Ironically, this fame is directly linked to the belief that the concept is seemingly very intuitive.

Exercise 1.6

Q1: In the 'Tossing two coins example', are the events HH and TT independent?

A1: While intuitively they appear to be, perhaps because they are mutually exclusive (?), they are not! In fact, *any* mutually exclusive events are not independent because the mere occurrence of one would tell us for sure that the other can't happen. It is important to keep in mind that we are dealing with a one-shot experiment, i.e., the two events HH and TT are relative to the same experiment, and not two successive ones!).

Q2: What about the events Ω and HH?

A2: Following the hinted logic from the previous question, does Ω tell something about the occurrence of HH (for the same one-shot experiment!). It doesn't, simply because Ω is 'anything can happen'; in other words, if one is told that 'anything can happen' in the experiment, the chance of occurrence for HH remains unchanged.

Q3: What about the events $\emptyset$ and HH?

A3: Following the same logic, knowing that 'nothing can happen' (i.e., $\emptyset$) appears to give us an indication that HH can't simply happen. The two events are however still independent (to be shown shortly) and the previous logic appears to be broken (we shall see why!).

Q4: Give an example of independent events neither involving Ω nor $\emptyset$ which are independent!

A4: $\{HH,HT\}$ and $\{HH,TH\}$ (we are now forced to use the brackets). One can certainly try to explain the independence through the above logic (knowing the either HH or HT happens does not give any indication whether either HH or TH happens), but it is arguably hard. ■

As the above logic is seemingly broken, we need to properly define the concept of independence

Definition 1.3.1 — Independence of Events. Two events A and B defined on some sample space Ω are (statistically) independent if

$$\mathbb{P}(A,B) = \mathbb{P}(A)\mathbb{P}(B)$$

R We prefer to write 'A,B' as a shorthand for $A \cap B$; in other words, $\mathbb{P}(A,B) = \mathbb{P}(A \cap B)$.

One can now immediately see that both Ω and $\emptyset$ are independent of any possible event.

R Keep in mind that the space Ω characterizes a one-shot experiment! But what if I tossed a coin twice, in succession? Then you'd be dealing with a new space $\{HH,HT,TH,TT\}$ and *not* two separate spaces $\{H,T\}$. For any random experiment, there is always a single space! If you ran multiple random experiments, each with its own space, then the whole random experiment would have a single space (roughly the cartesian product of the individual ones).

Exercise 1.7 **Q**: Say you toss a coin twice. The space for each sub-experiment is $\Omega = \{H,T\}$. Give a simple explanation avoiding the math why $\{HH,HT\}$ and $\{HH,TH\}$ are independent relative to the whole experiment, whose sample space is $\Omega \times \Omega = \{HH,HT,TH,TT\}$.

Q: Because $\{HH,HT\}$ corresponds to getting *heads* in the first run, whereas $\{HH,TH\}$ corresponds to getting *heads* in the second run; these two (successive) events are arguably independent! ■

The concept of events' independence is used to define the concept of independence for random variables

Definition 1.3.2 — Independence of Random Variables. Two random variables X and Y defined on the same sample space are (statistically) independent if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

for all possible values x and y.

In other words, the independence of r.v.'s corresponds to the independence of the events $X = x$ and $Y = y$ for all the values x and y. Recall that the event $\{X = x, Y = y\}$ is a shorthand for $X = x \cap Y = y$; do keep in mind that $X = x$ is actually a set, i.e., the event $\{\omega \in \Omega \mid X(\omega) = x\}$.

Exercise 1.8 Consider the simplified data from Example 1.1, i.e.,

$$(A,H) \in \{(10,130),(10,135),(15,140),(20,160)\}$$

Q1: Are A and H independent?

A1: No! do the math for $A = 10$ and $H = 130$.

Q2: What is an intuitive indication that A and H are not independent?

A2: The older the taller. In other words, the data gives an indication that older customers are more likely to be taller.

Q3: Give a non-empty data set for which A and H are independent!

A3: The simplest example is

$$(A,H) \in \{(10,130)\}$$

Another would be combining all pairs of ages and heights, e.g.,

$$(A,H) \in \{(10,130),(10,170),(20,130),(20,170)\}$$

■

R In all the previous examples the data set can be considered as the sample space Ω because the entries are different. Moreover, the assigned probabilities are assumed to be uniform, e.g., $\mathbb{P}((20,170)) = 1/4$ in the last example. If the data set had duplicates, i.e.,

$$(A,H) \in \{(10,130),(10,130),(10,170),(20,130),(20,170)\}$$

then the sample space would get rid of duplicates, whereas the probabilities would be skewed, e.g., $\mathbb{P}((10,130)) = 2/5$ and $\mathbb{P}((20,170)) = 1/5$.

Exercise 1.9 You roll two dice whose outcomes are X and Y. The sample space is $\Omega = \{11,12,\ldots,66\}$ and hence $X(\omega)$ is defined as 'the first component of ω' (e.g., $X(54) = 5$).

Q1: Is $X+Y = 4$ independent of $X+Y = 10$?

A1: Note that we are dealing with the independence of the corresponding events, i.e., $\{13,22,31\}$ and $\{46,55,64\}$; they are not independent simply because they are mutually exclusive.

Q2: What about $X+Y = 7$ and $X+Y = 9$?

A2: One needs to do the math ...

Q3: What about X and $X+Y$?

A3: (independence of r.v.). Intuitively, they are not independent because a big value for X would mean a big (enough) value for $X+Y$; such an argument needs to be backed up though by proper computations.

Q4: Give an example of two independent r.v.'s.

A4: The simplest is X and Y. Another would be X^2 and Y^{10}, and more generally $f(X)$ and $g(Y)$ for some functions $f(\cdot)$ and $g(\cdot)$. What about $X+Y$ and $X-Y$? ■

1.4 Conditional Probability

This is a crucial concept in data science.

Let us consider the toy data set

$$(A,H) \in \{(10,130),(10,135),(15,140),(20,160)\} \ .$$

Asking for the probability of getting $H = 130$ should be quite clear, i.e., $\mathbb{P}(H = 130) = 1/4$ as exactly one of the four records satisfies the $H = 130$ property (note the tacit assumption that all the records are 'equally likely').

What about asking for the probability that a 10-year old has $H = 130$? The crucial difference from the previous probability is the *additional constraint/condition*, i.e., $A = 10$. In other words, we are asking for $\mathbb{P}(H = 130)$ constrained, or conditioned, on $A = 10$. This is really the conditional probability, denoted in this case by $\mathbb{P}(H = 130 \mid A = 10)$.

What would then be a 'sensible' value for $\mathbb{P}(H = 130 \mid A = 10)$? Do keep in mind that a conditional probability is in the end a probability which must be defined on a space and must satisfy all the conditions (axioms) from Definition 1.1.1! The key feature of a conditional probability is that it is defined on another space, i.e., the conditional space (subject of course to the conditioning on the given constraint/condition). In our case, given the condition $A = 10$, the conditional space is simply the subset of records for which $A = 10$, i.e.,

$$\Omega_c := \{(10,130),(10,135)\} \ .$$

On this new space, one can simply compute $\mathbb{P}_c(H = 130) = 1/2$ (again, we are relying on the tacit assumption that all records are equally likely); we remark that we wrote $\mathbb{P}_c$ on purpose to denote a probability measure which applies to the conditional space (do recall again that this probability function must satisfy the 'axioms', in particular $\mathbb{P}_c(\Omega_c) = 1$!). Therefore $\mathbb{P}(H = 130 \mid A = 10)$ from the original space satisfies

$$\mathbb{P}(H = 130 \mid A = 10) = P_c(H = 130) = \frac{1}{2}$$

Let us now rewrite that as

$$\mathbb{P}(H = 130 \mid A = 10) = P_c(H = 130) = \frac{1}{2} = \frac{\frac{1}{4}}{\frac{2}{4}} = \frac{\mathbb{P}(H = 130, A = 10)}{\mathbb{P}(A = 10)} \ .$$

Note that in the 3rd equally we divided both ratio's terms by the number of samples 4. It turns out that this formula (first and last term) is exactly the general formula for the conditional probability.

Definition 1.4.1 — Conditional Probability. For two events A and B such that $\mathbb{P}(B) > 0$, the probability of A conditioned on B is defined as

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)} \ .$$

It should be quite clear why the condition $\mathbb{P}(B) > 0$ is needed; also, we can now better see why the logic from Exercise 1.6 (A3) does not generally work. Do note the implicit assumption on some underlying space Ω; what is important is that both A and B are defined on the *very same* space. Also, keep in mind that B is just an ordinary event which can be as complicated as possible (e.g., involving unions and intersections of other events).

A very important and useful result is the following

Proposition 1.4.1 — Bayes' Formula. For two events A and B it holds

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{P(B)} .$$

Note the implicit assumption that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Let us next give some examples on computing conditional probabilities.

■ **Example 1.7** Throw two fair dice and let the random variables X and Y represent the outcomes. What is the probability that the sum is greater than 6 given that the first die displays 3, i.e, $\mathbb{P}(X+Y > 6 \mid X = 3)$? The corresponding events are

$$A := \{X+Y = 6\} = \{(x,y) : x+y > 6\},\ B := \{X = 3\} = \{(3,y) : 1 \leq y \leq 6\} .$$

Observing that $A \cap B = \{(3,4),(3,5),(3,6)\}$ we immediately get

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)} = \frac{\frac{3}{36}}{\frac{6}{36}} = \frac{1}{2}$$

Obviously, we assumed that the dice are fair. Do note that if required to also compute $\mathbb{P}(B|A)$, then one can follow the same steps, or, simply, apply Bayes' formula. ■

Exercise 1.10 **Q**: Why is Bayes' formula needed anyway? (at least, in the previous example, it seems that we don't really need it.)

A: It depends on the complexity of the conditioning event B. Consider the data set

$$(A,H) \in \{(10,130),(10,135),(15,140),(20,160),\dots\}$$

If the event A is 'simple' (e.g., $H = 130$) whereas B is too complicated (e.g., some complicated function of A and H), then it is conceivably easier to think of $\mathbb{P}(B \mid A)$ than of $\mathbb{P}(A \mid B)$ (in the former it is much easier (for humans!) to visualize the conditional space, i.e., only the rows for which $H = 130$). ■

■ **Example 1.8** Toss two fair coins. Denote the corresponding space $\Omega = \{HH,HT,TH,TT\}$ and let the event $B = \{HH,HT,TH\}$, i.e., 'at least one *heads*'. Compute $\mathbb{P}(HH \mid B)$.

A quite popular and careless answer is $\mathbb{P}(HH \mid B) = \frac{1}{2}$! Let us *carefully* do the computations:

$$\mathbb{P}(HH \mid B) = \frac{\mathbb{P}(\{HH\} \cap \{HH,HT,TH\})}{\mathbb{P}(B)} = \frac{\mathbb{P}(HH)}{\mathbb{P}(B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3} .$$

An alternative way is to explicitly consider the conditional space

$$\Omega_c = \{HH,HT,TH\} .$$

On this space, $\mathbb{P}_c(HH)$ (i.e., the conditional probability from the original space $\mathbb{P}(HH \mid B)$) is just $\mathbb{P}_c(HH) = \frac{1}{3}$. This alternative way of thinking should illustrate the benefits of reasoning with conditional spaces; by 'benefits' I mean getting probability computations right without a pen! ■

■ **Example 1.9** Toss a golden and a silver coin. The space is exactly the same as in the previous example, i.e.,

$$\Omega = \{HH,HT,TH,TT\}$$

but with the important distinction that 'order now matters', e.g., the event $\{TH\}$ represents the event 'the golden coin shows *tails* whereas the silver coin shows *heads*'. Denote the event $B = \{HH, HT\}$, i.e., the 'golden coin shows *heads*'. Compute $\mathbb{P}(HH \mid B)$:

$$\mathbb{P}(HH \mid B) = \frac{\mathbb{P}(\{HH\} \cap \{HH, HT\})}{\mathbb{P}(B)} = \frac{\mathbb{P}(HH)}{\mathbb{P}(B)} = \frac{1}{2} \ .$$

Recall that it is acceptable to omit brackets when clear from the context, in order to simplify notation, i.e., it is advisable to avoid writing $\mathbb{P}(\{HH\} \mid B)$. ■

We wrap up the chapter with another important (very useful!) property:

Proposition 1.4.2 — Law of Total Probability. For some event A and a partition $B_1, B_2, \ldots, B_n$ of the underlying sample space Ω it holds

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i) \ .$$

Recall the two essential properties of a partition:

$$\begin{cases} \cup_{i=1}^{n} B_i = \Omega \\ B_i \cap B_j = \emptyset \text{ for all } i \neq j \end{cases}$$

.

We next illustrate the usefulness of this property through some examples:

■ **Example 1.10** ■

2. Heavy-Tailed Distributions

We devote a whole chapter to heavy-tailed distributions (laws), due to their prevalence in real-world data. We shall focus in particular on some concrete examples, the underlying causes driving the “heavy-tailed" behavior, and some insights into how to properly fit heavy-tailed distributions.

It is instructive to first reflect on what does the word association “*heavy-tailed*" actually mean. The word “*tail*" is typically associated to a distribution function, and more concretely to values far away from the mean. More concretely, the tail of a distribution of some r.v. X would be the probability value (aka *tail probability*)

$$\mathbb{P}(X > x)$$

for some large value x; as we'll be mostly dealing with positive r.v.'s, we are really considering values *far to the right* (i.e., both positive and large, relative to the mean). In principle one can also consider values *far to the left* (i.e., both negative and small, relative to the mean), but we are mostly consider positive r.v.'s.

Take the familiar exponential distribution with rate λ and CCDF

$$\mathbb{P}(X > x) = e^{\lambda x} .$$

Pick $\lambda = 1$. A value *far to the right*, relative to $E[X] = \frac{1}{\lambda} = 1$, would be 100, or 1000, or any arbitrarily larger. One can immediately see that the tail at those points is simply too small to be even considered. In other words, the tail is “*light*", i.e., not “*heavy*". So *heavy-tail* means roughly non-negligible probabilities at large values *far to the right*.

Consider next an empirical toy data: Denote by X_n as the number of tweeter followers for student n in a class of 100 students and assume that the (observed) data satisfies

$$X_1 + X_2 + \cdots + X_{100} = 1,000,000 .$$

(As a side remark, you should think of $X_1, X_2, \ldots, X_{100}$ as a sequence of *empirical* random variables, following the same distribution as some generic r.v. X; as an analogy, recall the “throwing a die multiple times" random experiment, in which case we had the random variables $X_1, X_2, \ldots,$ where

X_i denoted the i'th die occurrence.) The likely scenario is that one student has ridiculously many followers whereas all the others have just very few (say for the sake of the argument less than 10). The empirical mean will roughly be $1,000,000/100 \approx 10K$. If we inspect the tail, i.e., values *far to the right* relative to the mean $10K$, we would have the tail probability

$$P(X > x) = 0.01$$

for values of x very close and smaller than $1M$, which is an indication that the *tail* is quite *heavy* (because 0.01, as a probability, is not a negligible number but on the contrary).

Can we now state that the distribution of X is *heavy-tailed*, based on the empirical data? While with the above knowledge the distribution appears to be heavy-tailed, it is actually not; we shall see why. In our toy data, the $1M$ value ought to be interpreted as an *outlier* (i.e., in sharp contrast with pretty much *all the others*).

2.1 Real Histograms

Let us us now give some examples of real-data histograms; for each we shall discuss the 'apparent' type of distribution and its most relevant aspects. By the way, a *histogram* is simply the plot of the density (DF) or (probability) mass function (PMF); in other words, we would plot values $P(X = x)$ against the value of x, for all the values of the r.v. X. As we'll be dealing with *real* data, the underlying r.v.'s are by default discrete. You should also keep in mind that a histogram is sometimes called a *frequency plot*, obviously because the word *frequency* closely relates to the probability values $P(X = x)$!

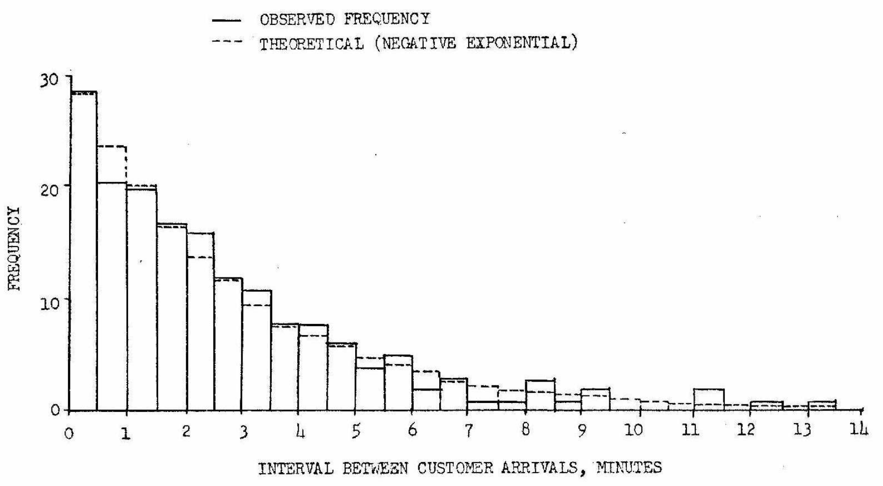


Figure 2.1: Frequency histogram for the time intervals between customer arrivals, on a Tuesday day, at some retail store, and more specifically at the "Ladies' Belt Counter"; plot extracted from a 1958 PhD thesis by William Archie Gresham Jr., GA Tech

We start with an example (see Fig. 2.1 + description) which seemingly indicates an underlying exponential distribution (because the plot looks very much like an exponential!); the author properly proved that the exponential is indeed an "excellent fit" using statistical tests. What is relevant to mention about the exponential distribution is that the mass function is decreasing (i.e., smaller interarrivals are more likely than larger interarrivals), and that the frequency of larger interarrivals decays very fast (i.e., it is very unlikely to have a large interarrival, e.g., 15 minutes, especially at the Ladies' Belt Counter!).

What is somewhat unfortunate about the exponential distribution is that supporting real-world examples are quite rare (I at the very least had to dig up the literature all the way to the 1950s and read about how frequently belts were being bought on Tuesdays at some retail store in Southern US!). That means that the exponential distribution, as convenient as it may be from an analytical point of view, it is not that common. However, is is typically argued to model several phenomena, e.g., the time to failure (e.g., for some device/appliance to fail) or battery lifetime, but empirical evidence (i.e., data supporting the claim) is quite modest to my knowledge.

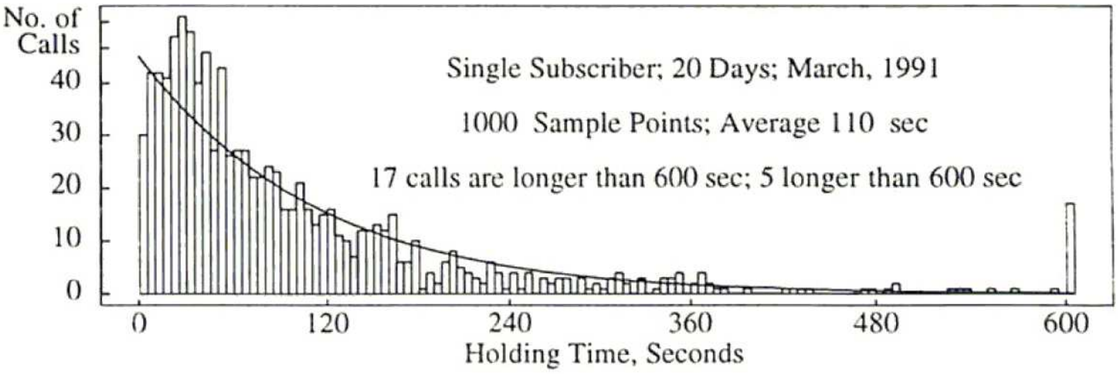


Figure 2.2: Frequency histogram of holding times at some telephone subscriber; plot from an ITC 1994 paper by Bolotin

A typically cited phenomena characterized by an exponential distribution is the holding times of phone calls. Consider the frequency plot from Fig. 2.2, which indicates an exponential distribution (it looks like!) for the holding time. It turns out however that the superposed exponential fit (i.e., the exponential function which best approximates the empirical data) performs very poorly from a statistical point of view (bad fit!); at the very least, note that the exponential misses the initial rising part in the histogram. Reasonable fits could be achieved with a lognormal distribution (look it up!) or a mixture of distributions. As it is outside the scope of the chapter, we won't give details about how good/bad of a fit an exponential distribution is; this topic, for general distributions, is referred to as 'goodness of fit' and examples of related statistical tests include Kolmogorov-Smirnov test (K-S test) or Pearson's chi-squared test (the latter used in the belts example!).

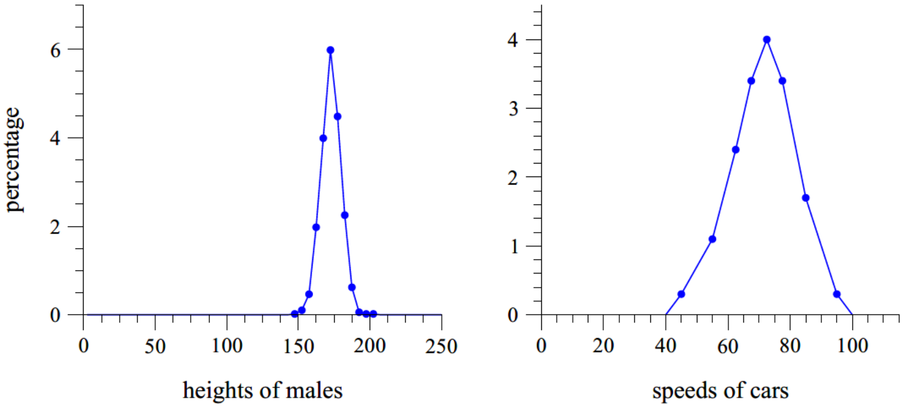


Figure 2.3: Left: histogram of heights in centimetres of American males. Data from the National Health Examination Survey, 1959. 1962 (US Department of Health and Human Services). Right: histogram of speeds in miles per hour of cars on UK motorways. Data from Transport Statistics 2003 (UK Department for Transport); plot+description from [New05]

Let us now consider the histograms from Fig. 2.3 which quite clearly indicate some underlying normal variables (yet one has to properly prove that using statistical tests). What is important to emphasize about the normal is that pretty much ‘*all the action*’ happens around the mean; recall that a normal r.v. $X \sim \mathcal{N}(\mu, \sigma)$ satisfies $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.97$! Moreover, values beyond $\mu \pm 3\sigma$, let alone $\mu \pm 4\sigma$, are very unlikely.

Exercise 2.1

Q: Take a standard normal $X \sim \mathcal{N}(0,1)$. What is roughly $\mathbb{P}(X > 10)$?

Q1: Ridiculously small: 10^{-24}! That is much less than the value of a (similar) exponential, because the density $f(x)$ of the (standard) normal is negatively exponential in x^2 and not in x, i.e., $f(x) \approx e^{-\frac{x^2}{2}}$; the comparison is not really fair, as the exponential has a positive mean, whereas $X \sim \mathcal{N}(0,1)$ has zero mean, but the point should be quite clear. Keep in mind that as the normal decays much faster than the exponential, it is sometimes referred to as a *thin-tailed* distribution; recall that the purpose of the chapter is the opposite, i.e., *heavy-tailed* distributions. ■

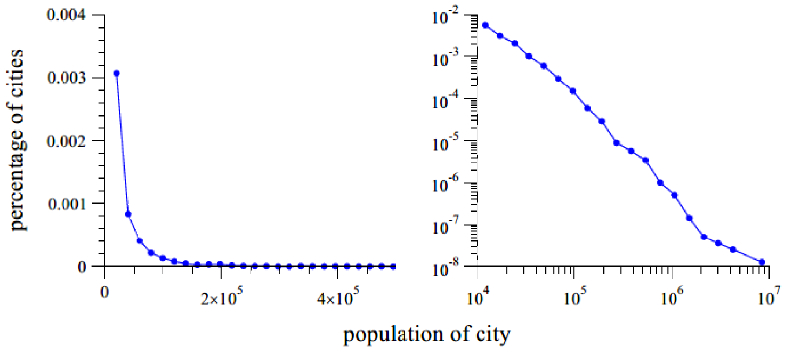
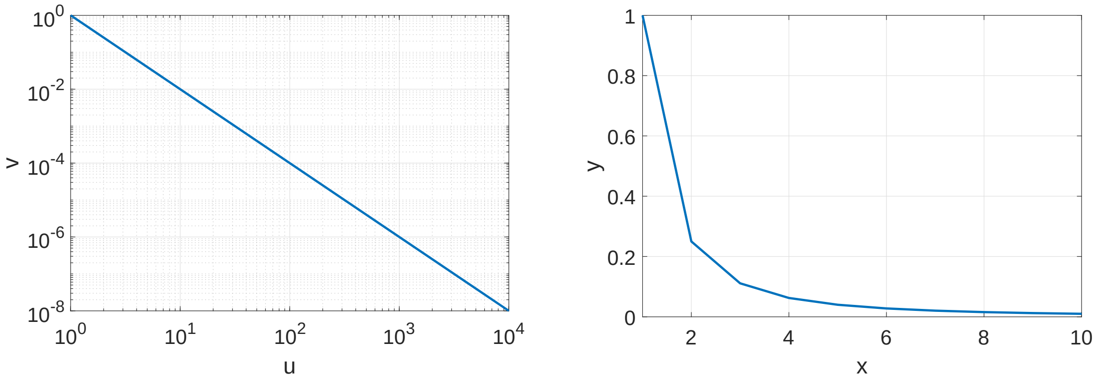


Figure 2.4: Left: histogram of the populations of all US cities with population of 10 000 or more. Right: another histogram of the same data, but plotted on logarithmic scales. The approximate straight-line form of the histogram in the right panel implies that the distribution follows a power law. Data from the 2000 US Census; plot+description from [New05]

Consider next the frequency-plot from Fig. 2.4 (ignore the right figure for now); note that the *frequencies* are expressed as probabilities, i.e., percentage of cities with some number (from the x-axis) of inhabitants. The first observation we could possible make is obvious: the plot is exponential! But that is unfortunately not the case; in fact, the true underlying distribution is fundamentally different than the exponential (which it is really unfortunate for mathematicians, as the exponential is truly easy to mathematically handle in all sorts of problems).

If the data in the plot is not exponential then what could it be? Note that just by starring at a plot we cannot say much more than the obvious! A quite remarkable trick to get out of the apparent impasse is to re-plot the data on a *log-log* scale (see the Right figure). By log-log it is simply meant that we would deal with the same values on both axes, but the (visual) distances between say two points u and v would change from $u - v$ to $\log u - \log v$; in other words, every point u relocates to position $\log u$! For instance, on the x-axis, the distance between 10^5 and 10^4 is the same as the distance between 10^6 and 10^5. As a side remark, log-log means that on *both* axes the points are relocated according to the new distance constraint; in turn, a log-linear plot means that only the points on the x-axis are relocated (btw, for some authors, the log-linear plot corresponds to relocating the points on the y-axis; to avoid the ambiguity, it is a good practice to be specific what log-linear or linear-log refers to).

To better see what we could get out of a log-log plot consider the toy data plotted in Fig 2.5.(a).



(a) log-log (b) linear-linear

Figure 2.5: Some toy data (the function $y = x^{-2}$) represented in two ways

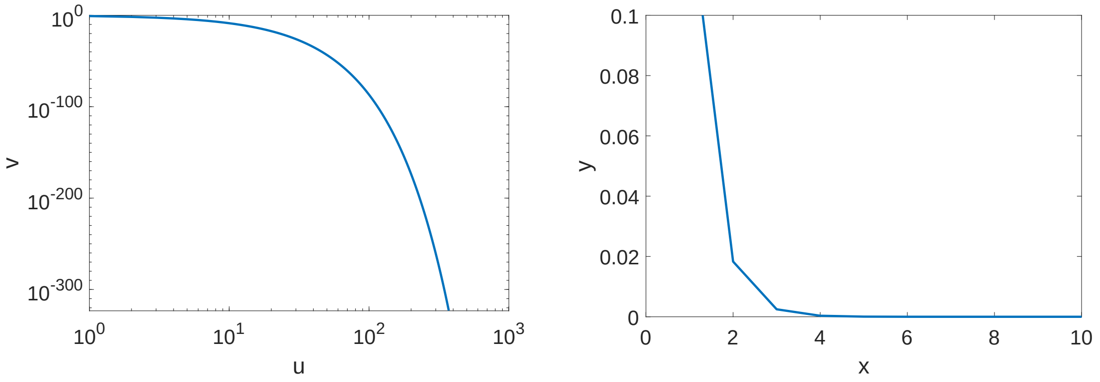
In this log-log (visual) representation it is quite clear that the data satisfies

$$v = -2u\ .$$

We do know however that every point u corresponds to some *true* point $\log x$, as *that* x was relocated to $\log x$; similarly, every v corresponds to some $\log y$. Rewriting the previous equation yields

$$\log y = -2\log x \Leftrightarrow y = x^{-2}\ ,$$

which represents the actual relation between the visually unaltered data points; see the (b) plot. We mention that we purposely restricted the x-axis in (b) to values less than 10 as otherwise it would simply be impossible to visualize anything meaningful (the blue line would appear as completely overlapping the x-axis); do note that we don't have this visualization annoyance of the linear-linear plot in (a)! In other words, the log-log plot is also a great trick to visualize data spanning a wide range of data.



(a) log-log (b) linear-linear

Figure 2.6: Some toy data (the function $y = e^{-2x}$) represented in two ways

To further stress these points let us also plot an exponential function ($y = e^{-2x}$) in both ways; see Fig. 2.6. The plot on the right should be quite obvious: a boring exponential with nothing much useful to see (note that we again had to restrict the range of x to $1\ldots10$). Let us next see what is the plot on the left, i.e., the relationship between u (corresponding to $\log x$) and v (corresponding to $\log y$). Taking the 'log' in $y = e^{-2x}$ we get

$$\log y = -2x = -2e^{\log x}$$

so the sought relation is

$$v = -2e^u \ ,$$

which is what's plotted on the left; in other words, if you plotted the function $y = -2e^y$ on the normal (linear) axes you'd get the behavior from the log-log plot. As a side remark, while the log-log plot were shown with a base-10 logarithm, the derivations are in base e as it's easier. Also note that the exponential does have a ridiculously light *tail*, i.e., the probabilities far-to-the right are just negligible; for instance, when $x = 150$, the tail (probability) is 10^{-300}.

Exercise 2.2 **Q**: How would you get a (visually looking) linear plot from the function $y = e^{-2x}$?

A: See the first line in the previous derivation, i.e., taking the 'log' which yielded

$$\log y = -2x \ .$$

By rewriting $\log y = v$ note that

$$v = -2x$$

which is a linear relation! To visualize that, you'd need to keep the x-axis unaltered and relocate all the points from the y-axis y to their 'log' value $v = \log y$; that's the linear-log plot! ■

We wrap up this section with several log-log real-data plots (see Fig. 2.12 at the end of the chapter) which very clearly depict the quite unexpected linear relation between the $\log x$ and $\log y$ values, or, equivalently, a relation of the form

$$y = x^{-\alpha}$$

for some value α; such a function is again fundamentally different than the exponential because the tails are *heavy*, i.e., they cannot be just ignored as for the exponential but on the contrary. As a side remark, do note that the plots are not frequency-plot but "*rank/frequency plots*"! The difference is that while the former plots the DF the latter plots the CCDF (for a very good reason we shall see later). The extra word "*rank*" has to do with the fact that the CCDF $\mathbb{P}(X > x)$ of some r.v. X is a non-decreasing function; in other words, the tails are ordered/ranked in a non-increasing order, i.e., the largest first, the second largest second, and so on.

2.2 Pareto Distribution

Having enough convincing evidence that plenty of real-data is heavy-tailed, we now formally define one of the most common heavy-tailed distributions, i.e., the Pareto one:

Definition 2.2.1 — The Pareto Distribution. A random variable X has a Pareto(α, x_{m}) distribution if its CDF is

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 1 - \left(\frac{x_{\mathrm{m}}}{x}\right)^{\alpha}, & x \geq x_{\mathrm{m}} \\ 0, & \text{otherwise} \end{cases}$$

or, alternatively, its DF is

$$f(x) = \begin{cases} \frac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}}, & x \geq x_{\mathrm{m}} \\ 0, & \text{otherwise} \end{cases} .$$

The Pareto distribution is a continuous one. A discrete version can be similarly defined; do keep in mind however that while real-data is discrete it is often more convenient to work with continuous approximations; the conceptual reason is that it is often easier to integrate than to take (discrete) sums!

The important parameter of the Pareto distribution is its *shape* $\alpha > 0$, which essentially determines the slope of the log-log plots (the slope is actually $-\alpha$). The other *scale* parameter $x_\mathrm{m} > 0$ essentially determines where the distribution starts.

Exercise 2.3 As a necessary exercise to get a good feeling about the Pareto distribution, do check the

$$\int_{x_\mathrm{m}}^{\infty} f(x) = 1 \; ,$$

i.e., the sanity-check condition that $f(x)$ is indeed DF. Do also further refresh your integration skills by fully deriving

$$\mathbb{E}[X] = \begin{cases} \infty, & \alpha \leq 1 \\ \frac{\alpha x_\mathrm{m}}{\alpha - 1}, & \alpha > 1 \end{cases} \qquad \mathrm{Var}[X] = \begin{cases} \infty, & \alpha \leq 2 \\ \frac{\alpha x_\mathrm{m}^2}{(\alpha-1)^2(\alpha-2)}, & \alpha > 1 \end{cases} .$$

■

Exercise 2.4 Let us now briefly revisit the Tweeter example from the start: most of the students have just few followers whereas one has $1M$.

Q: Is the distribution of X, from which the 100 samples were drawn, heavy-tailed (e.g., Pareto)?

A: It is not! To see why, let us start from the very far-right, i.e., the value $1M$. We roughly know that

$$\mathbb{P}(X \geq 1M) = \frac{1}{100} \; ,$$

from where we could try to fit the α (i.e., the truly important shape parameter). We could then try to also fit the scale parameter by accounting for all the other data values, and constrained of course by the integrability condition $\int_{x_\mathrm{m}}^{\infty} f(x) = 1$! The true problem is that, according to the data, the CCDF also satisfies

$$\mathbb{P}(X \geq x) = \frac{1}{100}$$

for a lot of values smaller than $1M$ (because there is only one guy with at least say 100 followers). Therefore, any possible fit of α and x_m would simply yield a bogus distribution to reasonably capture that whole tail from 100 to $1M$; a convincing reason is that while the tail of the real-data is constant, the Pareto tail does decrease (not exponentially fast but fast enough, more exactly polynomially!). The way one should actually look at the Tweeter data is to categorize the *popular* guy as an *outlier* (i.e., roughly to ignore...). ■

Let us next attempt to get more insights into the Pareto distribution. Let us frame the discussion around *residual waiting times*, i.e., the time one has to wait for something to happen *given* that one has already waited for some *fixed* time. Take three real-life scenarios:

1. Waiting at some check-in counter.
2. Waiting for the bus.
3. Waiting for a response from someone.

Which scenario would be prone to some heavy-tailed behavior (of the whole waiting time)? Ignoring catastrophic situations, the first two scenarios are quite obvious: no heavy-tails as those events are likely to happen in some reasonable amount of time. In the last scenario, however, one can argue that if a response hasn't come back in a *reasonable* amount of time then very likely something unfortunate must have happened (e.g., that someone has forgotten to reply!).

Exercise 2.5 **Q**: Does a one data point (someone has forgotten to reply to an e-mail) indicate a Pareto law?

A: No! Same reason as in the Tweeter example. That point would simply be an outlier. To do get a heavy-tail law, one needs many more data points (which would arguably be the case in some hypothetical real-data, as many do forget to reply to e-mails for a while); that *many* data points are necessary to get a statistically reasonably fit. Do keep in mind that a theoretical CCDF $\mathbb{P}(X > x)$ has values for all x; the empirical data would accordingly need sufficiently many values spread around (i.e., forgetting/late times of 24hr, 25hr, 30hr, 50hr, etc.). ■

It is instructive at this point to take a closer look at the residual waiting time which is analytically defined as

$$R_x(y) := \mathbb{P}(X \geq x+y \,|\, X \geq x) = \frac{1-F(x+y)}{1-F(x)} \,,$$

where $F(x)$ is the CDF of the r.v. X (representing/modelling the *whole* waiting time). In other words, we are interested in the following: given that we have already waited for x (time units), what is the probability of waiting for at least y (time units) more? Let us give two examples:

- Exponential Distribution ($X \sim \text{Exp}(\lambda)$)

$$R_x(y) = \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} = e^{-\lambda y}$$

- Pareto ($X \sim \text{Pareto}(\alpha, x_{\text{m}})$)

$$R_x = \frac{\left(\frac{x_{\text{m}}}{x+y}\right)^{\alpha}}{\left(\frac{x_{\text{m}}}{x}\right)^{\alpha}} = \left(1 + \frac{x_{\text{m}}}{x}\right)^{-\alpha}$$

Exercise 2.6 **Q1**: What is truly interesting about the above examples?

A1a: In the exponential case, the residual waiting time does not depend on x! In other words, no matter how much we have already waited for something to happen, the probability for happening from that point onwards is the same as at the very beginning! (for sure we don't want that when waiting for buses!) This rather counterintuitive behavior is generally referred to as the *memoryless property* of the exponential distribution. Do try to argue whether waiting for your phone to fail obeys such a property.

A1b: In the Pareto case, the residual waiting time is increasing in x! In other words, the more we have waited for something to happen, the more unlikely it will happen. Again, we have a rather counterintuitive behavior; arguably, it sort of makes sense in the waiting for an e-mail response example.

Q2: In the Pareto case we have just argued "The more we have waited for something to happen, the more unlikely it will happen.". But will *it* eventually happen?

A2: Yes, because the tail of X (the whole waiting time), i.e., $\mathbb{P}(X > x)$ converges to zero as $x \to \infty$ (i.e., the probability for not-happening at all is zero). ■

2.3 What Drives Heavy-Tailed Laws?

Here we analytically explain the occurrence of heavy tails from a theoretical point of view. Concretely, we shall give examples of random experiments whose outcomes lend themselves to heavy-tailed laws/distributions.

■ **Example 2.1 — Multiplicative Processes.** Consider two (independent) exponential r.v.'s X_1 and X_2 with parameter λ. Their product X_1X_2 satisfies in terms of CCDF

$$\mathbb{P}\left(X_1X_2 > x\right) \geq \mathbb{P}\left(X_1 > \sqrt{x}, X_2 > \sqrt{x}\right) = \mathbb{P}\left(X_1 > \sqrt{x}\right)\mathbb{P}\left(X_2 > \sqrt{x}\right) = e^{-2\lambda\sqrt{x}} \, .$$

As a side remark, we used the inequality because $X_1 > \sqrt{x} \cap X_2 > \sqrt{x} \subseteq X_1X_2 > x$; recall that two events A and B satisfy $P(A) \leq P(B)$ if $A \subseteq B$. We could have also *exactly* computed the CCDF of X_1X_2, instead of a lower bound, but it would have looked a bit more complicated (try it out using the Law of Total Probability for continuous r.v.!).

What is important though is the structure of $e^{-2\lambda\sqrt{x}}$ which is *heavier* than the exponential (does not decrease that fast, in this case because of the square root in the exponential)[1]. Importantly, do reflect upon the fact that note that what has really led to the heavy-tail was the underlying *multiplicative structure* of the process.

Exercise 2.7 Q1: Give some real-life examples, subject to a *multiplicative* structure, and consequently lending themselves to power laws.

A2: The Tweeter example, and more exactly the number of followers as a r.v. X: popular guys have a lot of direct followers, of whom some are also popular, and hence generating a lot of additional followers, of whom again some are popular, and so on. The r.v. X has clearly a multiplicative structure. This phenomena is usually better expressed through the aphorism "Rich get richers".

Q2: Which of the data from Fig. 2.12 is subject to an underlying multiplicative process? ■

■

■ **Example 2.2 — Time to Ruin.** Consider going to the casino with 10£ and always betting 1£ on the *heads* or *tails* (fair) game, as long as your funds allow of course. Denote by T the *time to ruin*, i.e., the number of games you can play before you end up with 0£; note that T is a r.v. which can take *any* possible value between 10 (you can certainly play for at least 10 times) and ∞. Without going into details (quite nice actually) one can show that

$$\mathbb{P}(T > n) \approx \frac{\sqrt{\frac{2}{\pi}}}{n^{\frac{1}{2}}} \, .$$

Do note that we are now dealing with a sort of discrete Pareto distribution (the symbol n was used on purpose, as it typically denotes discrete values). Do also note that the corresponding shape is $\alpha = \frac{1}{2}$. By making the analogy with the Pareto distribution (having infinite mean when $\alpha \leq 1$), we obtain that

$$E[T] = \infty \, ,$$

i.e., on average, you may indefinitely get stuck in the casino (assuming that you can only leave once ruined). ■

2.4 Identifying Heavy-Tails Laws

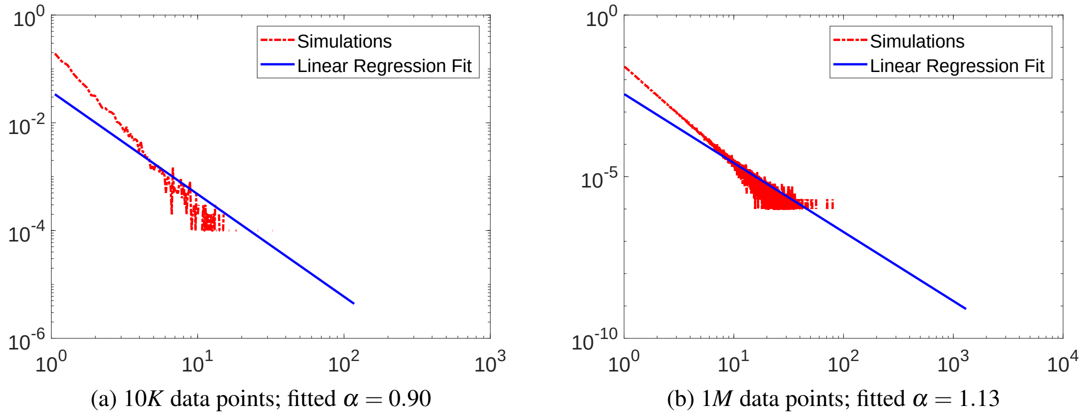
Having established that heavy-tails are ubiquitous in real-world data and that they can be theoretically explained, let us now address the following issue:

Having some real data, what is the actual distribution?

[1] While we have only formally introduced the Pareto law, as an example of a heavy-tailed distribution, the wider class of such distributions is huge.

We shall present two *informal* methods (one bad, one good) for fitting theoretical distributions to real-data. We shall also give some insights into another informal method to check whether some theoretical distribution is a good fit for some data; as we have already mentioned, we omit formal methods investigating statistical tests for 'goodness of fit'.

2.4.1 Bad Method: Frequency Plot



(a) $10K$ data points; fitted $\alpha = 0.90$

(b) $1M$ data points; fitted $\alpha = 1.13$

Figure 2.7: Frequency Plot: PDF fitting using linear regression; data (numbers) simulated according to $\text{Pareto}(2,1)$

In Fig. 2.7 we illustrate the frequency log-log plot (the red line) of a data set constructed by generating either $10K$ or $1M$ $\text{Pareto}(2,1)$ numbers. The blue line is obtained by invoking linear regression, i.e., a model trying to predict the cdf value subject to minimizing the sum of squared errors (see the next Chapter!). Unfortunately, this method performs extremely poorly as seen visually at the very least. We note in particular that while the slope of the CDF is 3, the fitted slope is as low as 1.9 in the case of $10K$ points; recall that the slope for the PDF in the case of a Pareto distribution is $\alpha + 1$. The other interesting observation is that generating only $10K$ numbers is not sufficient to build a longer tail by simulations only (there are very few numbers greater than 100). In turn, generating $1M$ numbers yields a significantly longer tail. For the Pareto distribution, in particular, large numbers do show but by invoking a very large number of draws.

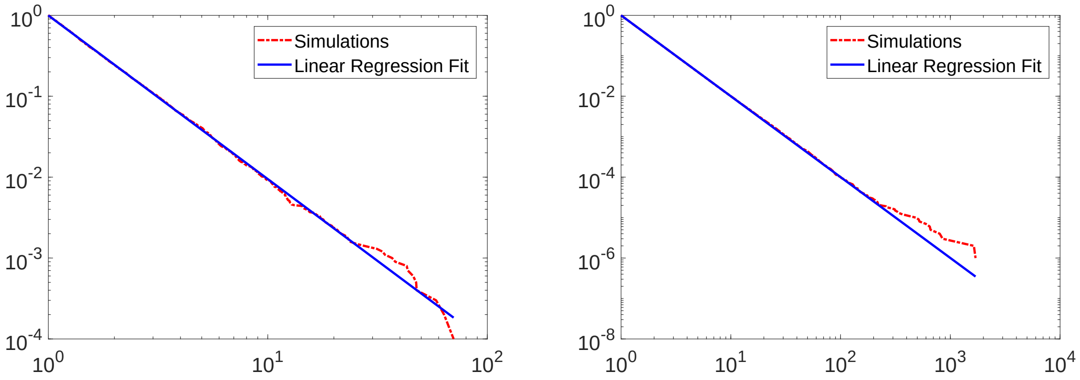
A common observation in both (a) and (b) is the extreme variation in the tail (i.e., the cdf values alternate frequently, creating a visual brushing effect). It is this variance which lends itself to the poor fit by linear regression. We should point that, despite its poor performance, the underlying method of running linear regression on a frequency-plot has been very common in the literature!

2.4.2 Good (Enough) Method: Frequency/Rank Plot

The key weakness of the frequency plot is usually addressed by running the same procedure but on the rank-frequency plot; see Fig. 2.8. One reason for the sharp drop in variance in the tail is the fact that the CCDF is a monotonous function. Although the fitted values of α are almost the true one (i.e., $\alpha = 2$), we point out that this method is still an informal one; recall the previous discussion about proper (statistical) fitting methods.

2.4.3 Q-Q Plot

Another informal and yet very powerful method to check whether some data follows some specific distribution is the so-called Q-Q plot. Let us first introduce the following concept



(a) $10K$ data points; fitted $\alpha = 2.03$

(b) $1M$ data points; fitted $\alpha = 1.99$

Figure 2.8: Rank-Frequency Plot: CCDF fitting using linear regression; data (numbers) simulated according to Pareto$(2,1)$

Definition 2.4.1 — Quantile. For some r.v. X and $0 \leq \varepsilon \leq 1$ the corresponding quantile is a value q_ε such that

$$\mathbb{P}(X \leq q_\varepsilon) = \varepsilon \ .$$

Exercise 2.8 **Q1**: What is q_0?

A1: That would be any value (strictly) less than minimum of the random variable X. Similarly, q_1 would be any value larger (or equal) than the maximum of X.

Q2: Consider the r.v.

$$X : \begin{pmatrix} 3 & 5 & 6 \\ \frac{1}{5} & \frac{3}{5} & \frac{1}{5} \end{pmatrix}$$

i.e., $\mathbb{P}(X=3) = \frac{1}{5}$ and so on. Note that $q_{\frac{4}{5}} = 5$, i.e., $\mathbb{P}(X \leq 5) = \frac{4}{5}$. What about $q_{\frac{2}{5}}$?

A2: According to the definition such a value doesn't exist. One would need a more robust definition of a quantile, e.g.,

$$q_\varepsilon := \sup\{x : \mathbb{P}(X \leq x) \leq \varepsilon\} \, ,$$

in which case $q_{\frac{2}{5}} = 3$. ■

Let us know give the main definition of this section

Definition 2.4.2 — Q-Q plot. Given two random variables X and Y, either theoretical or empirical, the Q-Q plot consists of the points $(q_\varepsilon^X, q_\varepsilon^Y)$ for a range of values ε; q_ε^X and q_ε^Y are the quantiles corresponding to X and Y, respectively.

The first important aspect of a Q-Q plot is that it typically looks like the one from Fig. 2.9, i.e., the points are between the pairs of the minimum and maximum values of X and Y.

The highlight of the Q-Q plot is that as long as the plot itself is *closed* to the $y = x$ line, then X and Y have *roughly* the same distribution. Fig. 2.10.(a,b) illustrate this point: in (a), the two distributions are *roughly* the same. In turn, in (b), note that the values of Y span from roughly 2 to 5; this range alone is in sharp contrast to the range of X, which is an immediate indication that the distributions of X and Y are vastly different. As an example, X can represent some empirical data, whereas Y can represent a subset of the same and *sorted* empirical data.

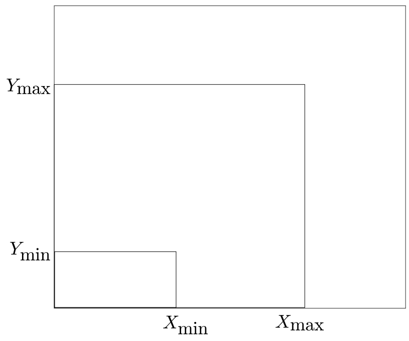
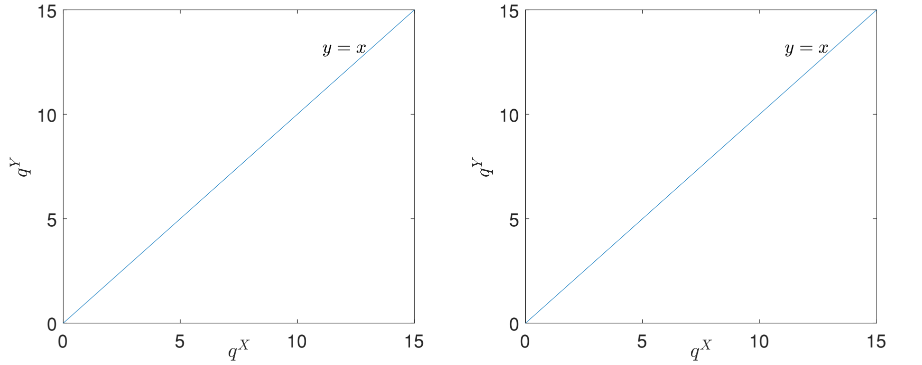


Figure 2.9: A Q-Q plot for two r.v. X and Y; the plot's points are between the corresponding minimum and maximum pairs.



(a) X and Y have *roughly* the same distributions (b) X and Y have largely different distributions

Figure 2.10: Examples of Q-Q plots for two r.v. X and Y; q^X and q^Y correspond to the quantiles of each

Exercise 2.9 **Q**: Give an example of X and Y (either theoretical or empirical) such that the corresponding Q-Q plot would look like the one from Fig. 2.11.

A: It is not possible, because the Q-Q curve is non-decreasing. Assume by contradiction that such a plot would be possible. According to the CDF's monotonicity:

$$x_1 \leq x_2 \Rightarrow \mathbb{P}(X \leq x_1) \leq \mathbb{P}(X \leq x_2)$$

which implies from the definition of the Q-Q plot that

$$\mathbb{P}(Y \leq y_1) \leq \mathbb{P}(Y \leq y_2) \Rightarrow y_1 \leq y_2$$

which is a *visual* contradiction. ■

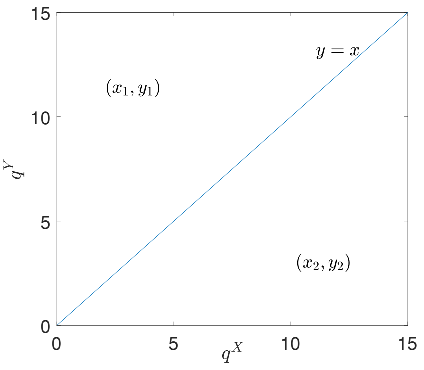


Figure 2.11: A hypothetical Q-Q plot

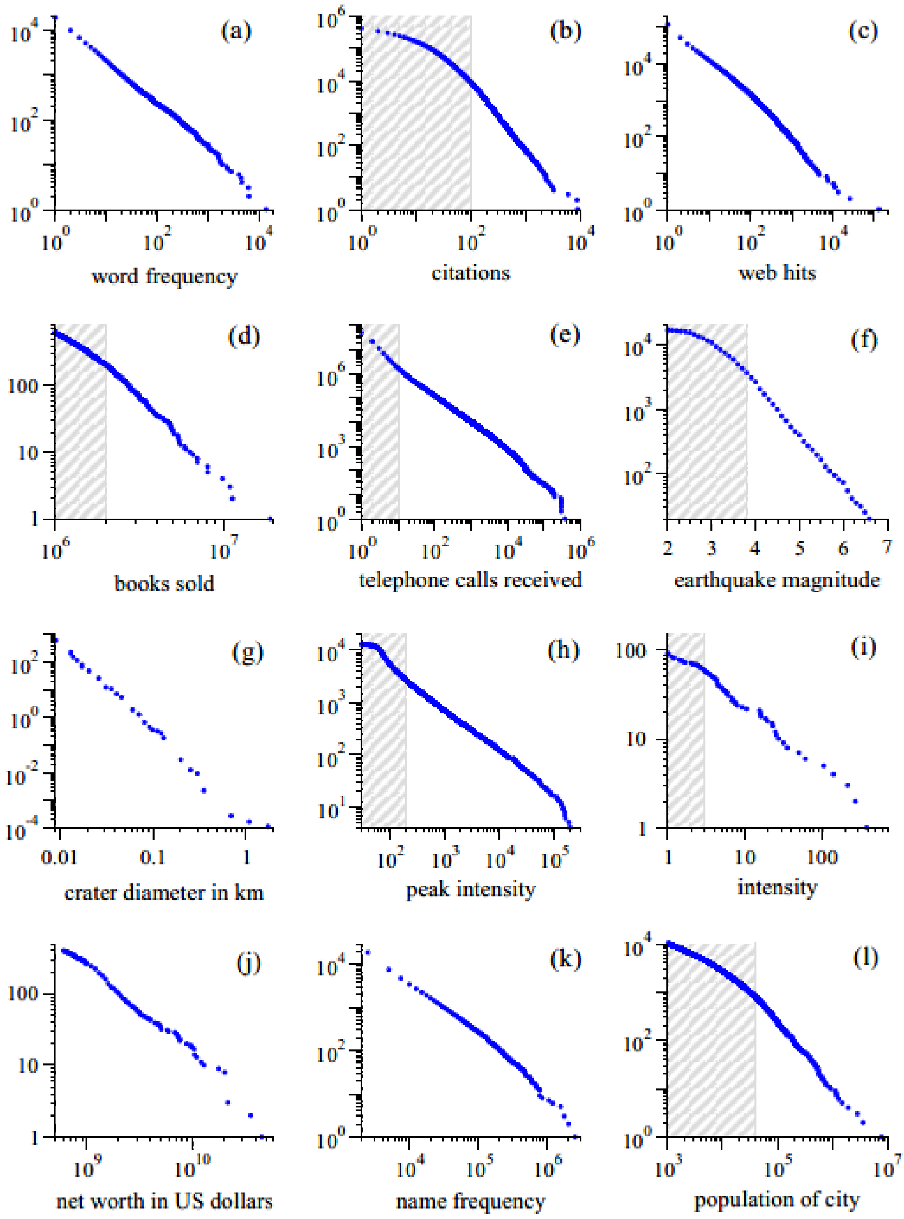


Figure 2.12: Cumulative distributions or "rank/frequency plots" of twelve quantities reputed to follow power laws. The distributions were computed as described in Appendix A. Data in the shaded regions were excluded from the calculations of the exponents in Table I. Source references for the data are given in the text. (a) Numbers of occurrences of words in the novel Moby Dick by Hermann Melville. (b) Numbers of citations to scientific papers published in 1981, from time of publication until June 1997. (c) Numbers of hits on web sites by 60 000 users of the America Online Internet service for the day of 1 December 1997. (d) Numbers of copies of bestselling books sold in the US between 1895 and 1965. (e) Number of calls received by AT&T telephone customers in the US for a single day. (f) Magnitude of earthquakes in California between January 1910 and May 1992. Magnitude is proportional to the logarithm of the maximum amplitude of the earthquake, and hence the distribution obeys a power law even though the horizontal axis is linear. (g) Diameter of craters on the moon. Vertical axis is measured per square kilometer. (h) Peak gamma-ray intensity of solar flares in counts per second, measured from Earth orbit between February 1980 and November 1989. (i) Intensity of wars from 1816 to 1980, measured as battle deaths per 10 000 of the population of the participating countries. (j) Aggregate net worth in dollars of the richest individuals in the US in October 2003. (k) Frequency of occurrence of family names in the US in the year 1990. (l) Populations of US cities in the year 2000. plot+description from [New05]

3. Regression

Consider a r.v. Y and a random sample $Y_1, Y_2, \ldots, Y_n$; the Y_i's are r.v's themselves and have the same distribution as Y. One can think of them as multiple observation points of the same experiment (e.g., the value of a die thrown multiple times, or the temperatures recorded at some location and fixed time over multiple days).

Definition 3.0.1 — Statistics. A statistics is simply a function

$$f(Y_1, Y_2, \ldots, Y_n) \ .$$

For instance, $Z = Y_1 + Y_2 + \cdots + Y_n$ is a *statistics*.

Assume now that Y depends on some parameter θ and let us pose the following problem:

How could one estimate θ from $Y_1, Y_2, \ldots Y_n$?

The *estimation* will take the form of a *statistics*, which we would denote it by Z_θ.

Definition 3.0.2 — Unbiased Estimator. An estimator Z_θ for a parameter θ is called *unbiased* if

$$E\left[Z_\theta\right] = \theta \; .$$

Note that the estimator Z_θ is itself a r.v. (as a function of the Y_i's, and hence taking the expectation seems reasonable).

Let us give some specific examples:

■ **Example 3.1** Assume that $\theta = \mu := E[Y]$, i.e., we would like to estimate the mean of Y in terms of the sample Y_i's. Take

$$Z_\theta := \overline{Y} := \frac{Y_1 + \cdots + Y_n}{n} \; .$$

As a side remark r.v's written with a bar on top typically denote averages (in this case a *sample* average).

Applying the properties of the expectation we get

$$E\left[Z_\theta\right] = E\left[\frac{Y_1 + \cdots + Y_n}{n}\right] = \frac{nE\left[Y_1\right]}{n} = E[Y] \; ,$$

i.e., $\overline{Y}$ is an unbiased estimator for μ. ■

■ **Example 3.2** Assume that $\theta = \sigma^2 := \mathrm{Var}[Y]$, i.e., we would like to estimate the variance of Y in terms of the sample Y_i's. Take

$$Z_\theta := \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \; .$$

and let us observe that

$$\begin{aligned}
E\left[Z_\theta\right] &= \frac{1}{n-1} E\left[\sum_i Y_i^2 - 2\sum_i \overline{Y} Y_i + n\overline{Y}^2\right] = \frac{1}{n-1} E\left[\sum_i Y_i^2 - n\overline{Y}^2\right] \\
&= \frac{1}{n-1}\left(nE\left[Y^2\right] - nE\left[\overline{Y}^2\right]\right) = \frac{1}{n-1}\left(nE\left[Y^2\right] - n\left(\mathrm{Var}\left[\overline{Y}\right] + E\left[\overline{Y}\right]^2\right)\right) \\
&= \frac{1}{n-1}\left(n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\
&= \sigma^2 \; ,
\end{aligned}$$

i.e., Z_θ is an unbiased estimator for σ^2. We note that we applied twice the property that $\mathrm{Var}[Z] = E\left[Z^2\right] - E[Z]^2$ for some r.v. Z; we also used $\mathrm{Var}[aZ] = a^2\mathrm{Var}[Z]$ for some constant a.

Let us also point out that the reason for having $\frac{1}{n-1}$ as a factor in Z_θ, and not $\frac{1}{n}$, is not only a *technical one*, i.e., the estimator is unbiased (by taking $\frac{1}{n}$ then the estimator would *underestimate* σ^2). If one lists the elements in the sum of Z_θ, i.e.,

$$Y_1 - \overline{Y}, Y_2 - \overline{Y}, \ldots, Y_n - \overline{Y}$$

then one may notice that any $n-1$ are random, whereas the other one can be computed from the rest (i.e., it is no longer random). For instance

$$Y_n - \overline{Y} = \sum_{i=1}^{n-1} \left(\overline{Y} - Y_i\right) \; .$$

■

In other words, the n elements are subject to $n-1$ *degrees of freedom* which is what conceptually introduces the bias in the estimator; dividing by $n-1$ instead of n cancels out this bias.

Bibliography

[New05] MEJ Newman. “Power laws, Pareto distributions and Zipf’s law”. In: *Contemporary Physics* 46.5 (2005), pages 323–351 (cited on pages 14, 15, 25).