Handout 1: Data Analysis 2

September 24, 2017

Table of Contents

Basic Statistical Concepts

This chapter introduces the basic statistical concepts that will be necessary for the rest of the textbook. This is a very dense chapter intended for refreshing material that has already been studied instead of learning all things anew. We introduce the concept of the distribution and its visualization; basic statistics of central value, spread and skewness; some theoretical distributions and their properties; the concept of probability, independence, conditional and joint probabilities; inverse conditional probabilities and Bayes' rule; joint and conditional distributions, covariance and correlation. Throughout the chapter we provide definitions, some of the most important formulae as well as short explanations to highlight the intuitive content of the statistics.

Introduction

How do prices of the same product vary across stores, how do they vary within stores across days or weeks? Are there times when prices are exceptionally high or exceptionally low? Are prices more often in the low range with few highs or, conversely, are they more often in the high range with a few lows? Is the price of a product more or less likely to be high, or is it expected to be the same, when the price of another product is high? Answering some of these questions require visualizing prices, while answering others need some quantitative measures. This chapter starts with the tools, both visual and quantitative, to describe the distribution of variables to answer such questions.

1 Basic summary statistics

A statistic is something meaningful that one can compute from the data at hand. Basic summary statistics are the most widely used statistics to describe certain aspects of distributions of single variables. The most used statistic is the mean, or average:

$$\overline{x} = \frac{\sum x_i}{n} \tag{1}$$

where x_i is the variable under consideration. In this case, we are interested in the hotel prices.

Sometimes the name average and the corresponding $\overline{x}$ notation are reserved for the mean value of a variable computed from actual data, and the name mean, as well as the name expected value and the notation $E[x]$ is used for something more abstract. We do not make that distinction here and use the mean and average as interchangeable names. When we want to make the distinction we shall give more context instead of using different names for the mean.

An important feature of the mean is that it changes the exact same way as the variable if we transform the variable in a linear fashion. If we add a number to the variable the mean of this new variable is the old mean plus the number we added to it:

$$\overline{x+a} = \overline{x} + a \tag{2}$$

We rarely do this in practice though. In contrast, we often multiply variables with numbers, such as when we express money values in 000s or in constant prices or in some other currency. If we multiply a variable with a number, say b, its mean value gets multiplied by the same number b:

$$\overline{x \cdot b} = \overline{x} \cdot b \tag{3}$$

These properties are intuitive; they are also straightforward to derive from the formula for the mean.

The median is the middle value of the distribution in the sense that exactly half of the observations have lower value and the other half have higher value. The median is one of the many quantiles: a quantile is the value that divides the observations in the dataset to two parts in specific proportions. The first quartile has one quarter of the observations below and three quarters above; the second quartile has two quarters of the observations below and two quarters above (this is also the median); the third quartile has three quarters of the observations below and one quarter above. Tercile (thirds) and quintiles (fifths) are defined similarly. So are percentiles: the first percentile is the value below which one percent of the observations are and 99 percent above, and so on.

The mode is the value with the highest frequency in your data. Some data have one mode and are called unimodal, while others may have more. A data with two modes, sometimes called a bi-modal, has two values that are apart from each other with the same frequency, or approximately the same frequency. Distributions with two high-frequency values right next to each other are considered unimodal because they are easily put into the same bin. The mean, median and mode are different statistics for the central value of the distribution. The mode is the most frequent value; the median is the middle value; the mean is the value that one can expect for a randomly chosen observation. Statistics that measure the spread of distributions are the range, inter-quantile ranges, the variance and the standard deviation.

The range is the difference between the highest value (the maximum) and the lowest value (the minimum) of a variable. Inter-quantile ranges are related measures: they tell the difference between two quantiles. The inter-quartile range gives the difference between the third quartile (the 25^{th} percentile) and the first quartile (the 25^{th} percentile). The 90- 10 percentile range gives the difference between the 90^{th} percentile and the 10^{th} percentile. And so on.

The most widely used measure of spread is the standard deviation. Its square is the variance, which is the average squared difference of the observations from the mean. The standard deviation captures the typical difference between a randomly chosen observation and the mean. The variance is a less intuitive measure. At the same time, the variance is easier to work with, because it is a mean value itself. The formulae are respectively:

$$\text{Var}(x) = \frac{\sum(x_i - \overline{x})^2}{n} \tag{4}$$

$$\text{Std}(x) = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}} \tag{5}$$

The standard deviation is often used to re-calculate differences between values in the distribution in order to express such differences in terms of typical distance. Formulaically, this amounts to divide the difference with the standard deviation. Such measures are called standardized differences. A standardized difference expresses the difference in units of standard deviation: a measure one means a difference of one standard deviation, and so on.

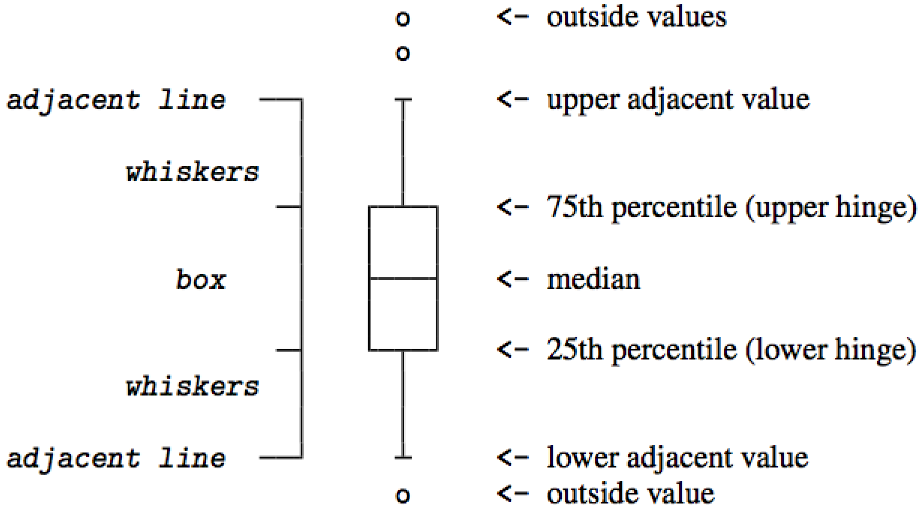
What happens to the standard deviation and the variance when we add a number to the variable or multiply it with a number is more complicated than for the mean. When we add a number to a variable its variance and standard deviation remain the same. This is intuitive: the spread of the variable does not change if we add a number to it. It can be seen in the formula, too, by noting that when we add a number to variable x its mean, $\overline{x}$, increases with the same number so the difference of the two remains unchanged. When we multiply a variable with a number the variance is multiplied by the square of the number, and the standard deviation is multiplied by the absolute value of that number (the square root of its square). Intuitively, multiplying a variable changes its scale so the spread should change the same way as now we measure it using a different scale. One can also see how things work out this way in the formulae.

We consider one more summary statistic, the mean –median measure of skewness. A distribution is skewed if it isn't symmetric. It may be skewed in two ways, having a long left tail or having a long right tail. When the distribution is symmetric its mean and median are the same. When it is skewed with a long right tail the mean is larger than the median: the few very large values in the right tail tilt the mean further to the right. Conversely, when a distribution is skewed with a left right tail the mean is smaller than the median: the few very small values in the left tail tilt the mean further to the left. The mean –median measure of skewness captures this intuition, and it standardizes the mean –median difference by dividing it with the standard deviation.

$$\text{Skewness} = \frac{(\overline{x} - \text{med}(x))}{\text{Std}(x)} \tag{6}$$

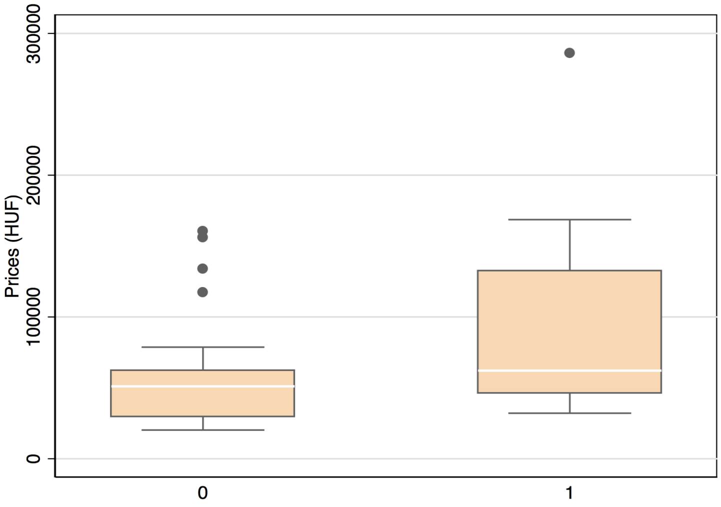
Finally, we introduce the box plot, a visual representation of some of the summary statistics. The Box Plot is really a one-dimensional vertical graph, only it is shown with some width so it looks better. The center of a box plot is a horizontal line at the median value of the variable, placed within a box. The upper side of the box is the third quartile (the 75^{th} percentile) and the lower side is the first quartile (the 25^{th} percentile). Vertical line segments on both the upper and lower side of the box capture most of the rest of the distribution. The ends of these line segments are usually drawn at 1.5 times the inter-quartile range added to the third quartile and subtracted from the first quartile. Observations with values not contained within those values are usually added to the box plot as dots with their respective values. The box plot conveys some important features of distributions such as their skewness and show some of the quantiles in an explicit way. Figure 1 shows the example that Stata manual gives for the box plot.

Figure 1: Box Plot



Notes: Illustration of a box plot using Stata Help command.

Figure 2: Box Plot of Hotel Prices in Vienna



Notes: Box plot of Vienna hotel prices over a dummy variable taking value one if the hotel has 4 stars or more and zero otherwise.

This box plot is telling us that the hotels with more than five prices have higher price variability. From the graph we can clearly see that hotels with more stars have a higher median price. From the box plot we can also see potential skewness of the dataset. Indeed, it seems that hotels with higher prices are skewed with a right long tail (This is because the portion of the box to the right of the median is larger than the portion of the box to the left of the median.)

Summary statistics to know:

- Measures of central value: Mean (average), median, other quantiles (percentiles), mode.
- Measures of spread: Range, inter-quantiles ranges, variance, standard deviation.
- Measures of skewness: the mean – median statistic.
- The Box Plot is a visual representation of many quantiles and may include extreme values.

Case study.

Assume that you are arranging a trip to Vienna for New Years, and you are looking at hotel prices on a major major hotel website. Suppose also that your only constraint: You are interested to stay strictly less than two metres from the city center. You give this parameters to your search and the website sorts you the hotels by the closeness the city centre and gives you the following hotels as in Table 2 below:

Table 1: Hotel Prices less than 2 km from the city centre

Hotel Name	Rating	Stars	Price (huf)	Dist center km
Hotel Lamee	4.5	4	148127	.2
Hotel Topazz	4.3	4	148127	.2
CH- Wellness Apartments	3.7	3	78702	.6
Hilton Vienna am Stadtpark	4.3	4	155528	.8
Derag Livinghotel An der Oper	4.6	4	131825	.8
Le Meridien Wien	4.4	5	285718	.8
Palais Hansen Kempinski Vienna	4.8	5	168646	.9
Hilton Vienna Plaza	4.6	4	155528	.9
Das Capri - Ihr Wiener Hotel	4.5	3	133526	1.2
Citadella Residence Appartments Vienna	5	4	90883	1.2
Royal Resort Apartments Urania	4.3	3.5	117116	1.3
Ruby Sofie Hotel Vienna	4.3	3.5	54342	1.4
Royal Resort Apartments Blattgasse	3.4	3.5	59339	1.4
NH Wien City	4	4	80888	1.7
Magdas Hotel	4	2	35697	1.8

Compute the mean price of a room in the city center?
Now compute median of these data. What are they? Compute the range and the standard deviation. Which of these measures of dispersion is better in your opinion?

Answer: To compute the mean of the hotel prices, then we apply the formula in Equation 1 and we have that the mean price is equal to 122932.8.

In order to get the median, mode, range and quartiles we sort the data from the lowest value to the highest. The median is 131825 which is slightly higher than the mean. Note that the mean and the median do not necessarily coincide. They only coincide in symmetric distributions.

While the mean is not necessarily a value from the original list the median, in cases where we haves odd number of observations is exactly the middle value, as in this case the 8-th element in the list which parts the data in half. In this case there is "middle" number, because we are dealing with odd observations then the median will be mean (that is, the usual average) of the middle two values within the list of prices.

In this particular example the mean and the mode are close in values as the mean is also similar to the median value in the list of prices. Usually, as a statistics the median is not as strongly influenced by the skewed values as the mean. The mean is very useful centrality statistics in "well behaved" data. When the data become skewed, toward the right or the left, the mean loses its ability to provide the best central location for the data because the extreme values drag it away from the typical value.

Applying the formula in Equation 2 we get that the standard deviation is: 60195.89. In order to find the range and, the quartiles and the range of these prices we sort them from the lowest to the highest value that these prices take. The range is the difference between the maximum and the minimum value in the data in this case is 250021.
Again when comparing the range with the standard deviation, a point worth stressing is that the range as a statistics is extremely sensitive to outliers. Usually, as a measure of dispersion in statistics we use the standard deviation.

2 Probability

A probability is a measure of the likelihood of an event. The concept of probability is a generalization of proportions in datasets. The "event" is that we look at an observation and see its value. The probability of a value of a variable is the proportion of such events among all instances of looking at the variable, which is the number of observations in the dataset.

Probabilities are always between zero and one. Sometimes they are expressed in percentage terms so they are between 0% and 100%. We denote the probability of an event as $p(event)$, so that $0 \leq p(event) \leq 1$. The probability of joint occurrence of events is called their joint probability and is denoted as $p(event1\&event2)$. The probability of the joint occurrence of two exclusive events is zero: $p(event1\&event2) = 0$ (exclusive events never happen at the same time).

An event either happens or it does not; these two “events” (the event happening and it not happening) are exclusive. We denote an event not happening as $\sim event$ so $p(event\ \&\ \sim event) = 0$. The probability of one of two exclusive events happening, or, in other words, the probability that one event or another, exclusive event happens, is the sum of the two probabilities: $p(event1|event2) = p(event1) + p(event2)$ if $p(event1\&event2) = 0$. When two events are not mutually exclusive, they may occur contemporaneously: $p(event1\&event2) > 0$. Then $p(event1|event2) = p(event1)+p(event2)-p(event1\&event2)$.

Conditional probability is the probability of an event if another event happens. That another event is sometimes called the conditioning event. Conditional probabilities are denoted as $p(event1|event2)$. The conditional probability can be expressed as the corresponding joint probability divided by the probability of the conditioning event:

$$p(event1|event2) = \frac{p(event1\&event2)}{p(event2)} \tag{7}$$

Two event are said independent if their joint probability equals the product of their individual probabilities: $p(event1\&event2) = p(event1)p(event2)$. This also means that the conditional probabilities are the same as the individual (unconditional) probabilities: $p(event1|event2) = p(event1)$ and $p(event2|event1) = p(event2)$. In words: two events are independent if the probability of one of the events is the same regardless of whether or not the other event occurs.

- Probability generalizes proportions of values in data. It is a number representing the likelihood of an event happening.
$0 \leq p(event) \leq 1$

- Joint probability of two events is the probability that both happen at the same time.

- Exclusive events never happen at the same time; their joint probability is zero:
p(event1 & event2) = 0.

- The probability or one or another event happening is:
$p(event1|event2) = p(event1) + p(event2) - p(event1\&event2)$.

- Conditional probability of an event, conditional on another event is the probability of an event (the conditional event) happening when the other event (the conditioning event) happens. $p(event1|event2) = p(event1\&event2)/p(event2)$.

- Two events are independent if the probability of one of the events is the same regardless of whether the other event occurs or not:
$p(event1|event2) = p(event1)$ and $p(event2|event1) = p(event2)$.

Case Study

Suppose now that you allow the website to give you a recommendation about a hotel in Vienna. In this case, the website will use all the hotels in the database.

Table 2: Hotel Prices less than 2 km from the city centre

	≥2 km	< 2 km	Total
< 4 stars	24	6	30
≥ 4 stars	23	9	32
Total	47	15	62

What is the unconditional probability that the hotel recommended from the website is less than 2 km from the city centre? What is the conditional probability that the hotel recommended is of at least four stars or more if it is less than 2 km away from the city center? What is the joint probability of the website recommends a hotel less than two km from the city center and a hotel with at least 4 starts?

3 Inverse Conditional Probabilities, Bayes Rule

Inverse conditional probabilities are two conditional probabilities, in which the role of the conditioning event and the conditional event are switched: $p(event1|event2)$ and $p(event2|event1)$.

Oftentimes in real life and in data analysis we face inverse conditional probabilities. Imagine that we want to know if an athlete used illegal substance (doping). For this case we collect lab tests. Does the positive result of the test indicates that there is illegal substance in the body of the athlete? We we are interested in is whether the athlete has doped given the positive test result, $p(doped|positive)$, or its complement, that the athlete did not dope even though the test shows positive results: $p(\sim doped|positive)$. Tests are imperfect in real life, and they sometimes give positive results even if athletes don't dope: $p(positive| \sim doped) > 0$. The relation of inverse conditional probabilities tells us how the imperfect nature of a doping test determines how confident we can be concluding that an athlete doped if the result of the test is positive.

The two inverse conditional probabilities are related although their relation might seem complicated. We can derive one from the other using the formula that links conditional probabilities and joint probabilities as both are related to the same joint probability, the probability of both $event1$ and $evet2$ occurring. The relation is called Bayes' rule after the reverend Bayes which formulated this formula first in the 17^{th} century.

$$p(event2|event1) = \frac{p(event1|event2)p(event2)}{p(event1)} \tag{8}$$

Which, in turn, can be rewritten as:

$$p(event2|event1) = \frac{p(event1|event2)p(event2)}{p(event1|event2)p(event2) + p(event1| \sim event2) * p(\sim event2)} \tag{9}$$

The most important message of these formulae is that inverse conditional probabilities are not the same in general. Instead of memorizing these formulae we suggest using a different approach that uses frequencies and proportions in place of abstract probabilities. Consider our doping example: whats the likelihood that an athlete is a doper (or a non-doper) if they receive a positive test result? Start with assuming that a fifth of the athletes dope. Out of, say, 1000 athletes that means 200 doping and 800 not doping. Consider a test that is imperfect but not bad: it always shows a positive result when the athlete dopes, but it also shows positive results 10 percent of the times if an athlete does not dope. The former means that the test will be positive for all 200 dopers.

The latter means that the test will also be positive for 10% of the non-dopers, which would be 80 out of the 800. In total we have 280 positive tests out of 1000. Of these 280 positives 200 are dopers and 80 non-dopers. We don't know which 200 is a doper and which 80 is a non-doper, but we can use these figures to calculate probabilities. The probability that an athlete is a doper if their test is positive is 200/280 = 71% approximately. The probability that an athlete is not a doper if their test is positive is 40/240 = 29% approximately. This may look surprising: a relatively small imperfection (10% of positive results for non-dopers) results in a much larger drop in our confidence: the chance that a positive tester did not dope in fact is 29%. As we shall see this is because we started with the assumption that only 20% of athletes dope.

Working through the formulae gives the same results. $event1$ is a positive result for the test, and $event2$ is the event that the athlete dopes. $p(event2) = 0.2$ (20 % of the athletes are dopers). P(event1—event2)=1: the test gives positive result for dopers with 100% likelihood. $p(event1| \sim event2) = 0.1$: the test gives positive results for non-dopers with 10% likelihood. Therefore we have that $p(event1) = [p(event1|event2)p(event2) + p(event1| \sim event2)p(\sim event2)] = [1{*}0.2{+}0.1{*}0.8] = 0.28$. So that $p(event2|event1) = 1{*}(0.2/0.28) = 0.71$, and, thus, $p(\sim event2|event1) = 1 - p(event2|event1) = 0.29$.

This example highlights several important things. First, working through the frequencies is not super-easy, but it is doable. With some practice it can become a relatively straightforward exercise. It is not easier to work with the formulae, but many of us find that route a lot less intuitive and thus less easy to remember. Second, however we carry out the calculation we need the probability that the test comes out positive for each of the groups, dopers and non-dopers (these are $p(event1|event2)$ and $p(event1| \sim event)$). Third, we need the overall fraction of athletes that dope. That is $P(event2)$ in the formulae above. This proportion is sometimes called the base rate. Without the base rate we can't compute the inverse probability.

- Bayes' rule shows the relation of inverse conditional probabilities: $p(event1|event2)$and$p(event2|event1)$

- $p(event2|event1) = \frac{p(event1|event2)p(event2)}{p(event1|event2)p(event2)+p(event1|\sim event2)*p(\sim event2)}$

- It is more intuitive to work this out with the help of frequencies instead of abstract probabilities.

4 Distributions

All variables have a distribution. The distribution of a variable tells the number of times each possible value of the variable occurs in the data. Besides the frequency of each value, the distribution may be expressed in terms of percentage of each value among all observations. The distribution of a variable completely describes the variable as it occurs in the data. It does so in isolation from other variables: the distribution of a single variable tells the frequency of each possible value but it does not tell whether certain values are more frequent when some other variable, or variables, have certain values. That information is contained in the joint distribution of variables, a concept we shall cover later in this chapter.

4.1 Theoretical Distributions

Theoretical distributions are distributions of variables with idealized properties. Some theoretical distributions occur in real data frequently, while others occur rarely and are better seen as approximations of actual distributions. It is important to learn some theoretical distributions and their properties because it helps understand features of real data. Comparing the distribution of actual variables to these theoretical distributions can help learn more about them. Some theoretical distributions are useful also in understanding what happens when we manipulate data or want to generalize from the data we have. Of the many theoretical distributions known in statistics we focus on a few important ones here that we shall use later in the textbook.

Theoretical distributions are fully captured by a few parameters: these are statistics that determine the distributions. The most important feature of theoretical distribution is their parameters. Among the other important features we shall look at the range of values they may take, the shape of the histograms they display, and their mean and variance, and how these are related to the basic parameter, or parameters, of the distribution.

4.1.1 Bernoulli Distribution

The distribution of a binary variable coded zero-one is called Bernoulli. The name comes from Jacob Bernoulli, a mathematician from the 1600's who first examined it. The Bernoulli distribution is a theoretical distribution that we observe over an over: all zero-one variables are distributed Bernoulli. (Note the use of words: if the distribution of a variable is Bernoulli, we say that variable is distributed Bernoulli and so will be the use of words for other theoretical distributions.) Examples include whether a customer makes a purchase (1 if yes 0 if no), whether the CEO of a firm is young, or whether a portfolios produces a large negative loss. The Bernoulli distribution has one parameter: p, the probability of observing value one (instead of value zero).

With only two possible values zero and one the range of the Bernoulli distribution is zero to one, and its histogram consists of two bars: the frequency of observations with value zero, and the frequency of observations with value one. If, instead of frequency, the histogram shows the proportion of each values, the height of the bar at value one is equal to p, and the height of the bar at zero equals $1p$. The mean of a Bernoulli variable is simply p, the proportion of ones. (To verify this try $p = 0, p = 1, p = 0.5$.) Its variance is $p(1 - p)$ so its standard deviation is $\sqrt{p(1 - p)}$.

4.1.2 Binomial Distribution

The Binomial distribution is based on the Bernoulli distribution. A variable is distributed Binomial if it can be viewed as the sum of many independent Bernoulli variables with the same p parameter (what independence precisely means will be discussed later in this chapter). Some actual variables that may be distributed Binomial include the number the number of car accidents, or the number of times number number one shows up in the winning lottery ticket. Binomial variables have two parameters, p, the probability of one for each Bernoulli variable and n, the number of Bernoulli variables that are added up.

The possible values of a Binomial variable are zero, one, and all other integer numbers up to n. Its range is therefore zero through n. The histogram of a Binomial variable has $n + 1$ bars (zero, one, through n). The Binomial distribution has one mode in the middle, and it is symmetric so its median, mean and mode are the same. With large n the histogram of a Binomial variable is bell shaped. The mean of a Binomial variable is np, and its variance is $np(1 - p)$, so its standard deviation is $\sqrt{np(1 - p)}$.

The other three distributions we cover in this section are for continuous variables: variables that can take on many values that may include fractions as well as irrational numbers such as π or the square root of two. In real data few variables can take on such values. Even variables that may be in principle continuous such as distance or time are almost always recorded with countable values such as integers or fractions rounded to a few decimal points. Continuous variables and their theoretical distributions are best seen potential approximations of variables that can take on many values even if those values do not include all fractions and irrational numbers.

4.1.3 Uniform Distribution

The uniform distribution characterizes continuous variables with values that are equally likely to occur within a minimum value and a maximum value. Examples of real life variables that may be approximately uniformly distributed are rare; the uniform distribution is more often used as a benchmark to which other distributions may be compared. The uniform distribution has two parameters, the minimum value a and the maximum value b. The histogram of the uniform distribution is completely flat between a and b with zero frequency below a and above b. It is therefore symmetric. Somewhat strangely, it has no mode: any value is just as frequent as any other value. The mean of a uniformly distributed variable is $\frac{a+b}{2}$, the variance is $\frac{(b-a)^2}{12}$, the standard deviation is $\sqrt{\frac{(b-a)^2}{12}}$.

4.1.4 The Normal Distribution

The Normal distribution the best known and most widely used theoretical distribution of continuous variables. It is a pure theoretical construct in the sense that it was derived mathematically from another distribution, the binomial. It can be thought of as a generalization of the binomial with infinitely many Bernoulli variables added up. For each of these Bernoulli variables the value one has to have tiny probability p. Variables with a normal distribution can take on any value in principle from negative infinity to positive infinity (negative values may come in as further generalizations to the infinite-binomial construct). The histogram of the normal distribution is bell shaped. For that reason the normal distribution is sometimes called the bell curve. The normal distribution has two parameters, usually denoted as μ and σ. They also refer to the mean and the standard deviation, respectively. The variance is the square of the standard deviation, σ.

Quite a few variables in real data are close to be normally distributed. The height of people in a population is usually approximately normal, and so is their IQ, a measure of intelligence (although that is in part because the tests behind the IQ measure are constructed that way). Percentage returns on a broad portfolio of liquid assets such as the S&P can be also approximately normally distributed if the returns are measured in a long enough horizon. The fact that many variables in real life are well approximated by the normal distribution has to do to the fact that the normal is a generalization of the binomial, which is derived as the sum of Bernoulli variables. It a variable is the result of several inputs, it may be distributed binomial. If it is the result of many, many inputs, each of which has a tiny contribution in itself, it may be distributed normal.

A variation on the normal is the standard normal distribution. It is a normal distribution with parameters $\mu = 0$ and $\sigma = 1$: its mean is zero and its standard deviation is one (and thus its variance is also one). If a variable, x, is normally distributed with mean μ and standard deviation σ, $x \sim \mathcal{N}(\mu, \sigma^2)$ its transformed version is distributed standard normal if we take out μ and divide this difference by σ: $\frac{(x-\mu)}{\sigma}$.

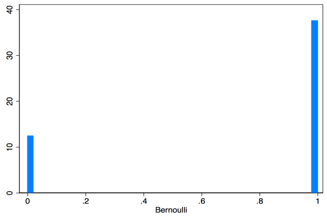
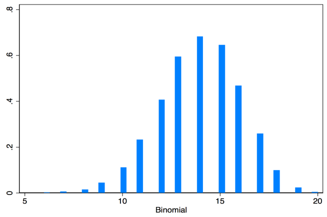
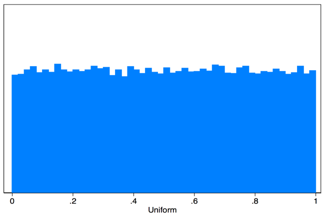
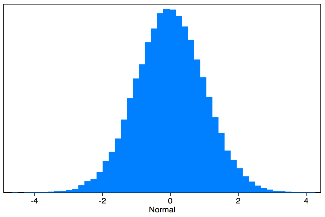
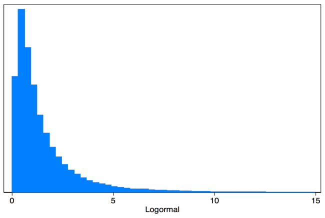
That the mean of this variable is zero and the standard deviation is one should not be surprising; recall what happens to the mean and the variance of variables when they are transformed by addition and multiplication. The fact that a variable transformed like this is normally distributed is more unique a phenomenon, guaranteed by the properties of the normal distribution. Note that other, nonlinear transformations don't yield a normally distributed variable. For example, the square of a normally distributed variable is not normally distributed.

4.1.5 The Lognormal Distribution

Finally, we consider one more variation on the normal: the lognormal distribution. If we take a variable that is distributed normal (x) and take e to its power (exp^x) the resulting variable is distributed lognormal. The old variable is the natural logarithm of the new variable, therefore the name of the new distribution (the log of which is normal). Because we raised e to the power of the original variable the resulting lognormal variable is always positive. It ranges between zero and positive infinity (never reaching any of them). By convention the parameters of the lognormal are the mean μ and standard deviation σ of the original variable, which is the logarithm of the new variable. Thus the mean and the standard deviation of the lognormal are complicated functions of these parameters, they are: $exp(\mu + \frac{\sigma^2}{2})$ and $\sqrt{exp(\mu + \frac{\sigma^2}{2}) \cdot (exp(\sigma^2) - 1)}$

There are also real life variables that are approximately lognormally distributed. These include distributions of prices, incomes, firm size. The reason is that differences in the natural log of a variable approximate relative differences: for example, a difference of 0.1 of log variables is approximately 10 % of the values of the original variable. Variables whose percentage differences are distributed normally are lognormally distributed.

Important theoretical distributions:

Distribution	histogram	parameters	range	mean	variance
Bernoulli		p	$[0, 1]$	p	$p(1-p)$
Binomial		p, n	$[0, n]$	np	$np(1-p)$
Uniform		a, b	$[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal		μ, σ	$(-\infty, +\infty)$	μ	σ
Lognormal		μ, σ	$[0, +\infty)$	μ	σ

4.2 Distributions in Practice

Many features of distributions are well captured visually. The simplest and most popular way to visualize a distribution is the histogram. The histogram is a bar chart that shows the frequency of each value. With binary variables the distribution is the frequency of the two possible values and thus the histogram consists of two bars. With variables that take on a few values the histogram shows as many bars as the number of possible values. With variables with many potential values such visualizations are usually uninformative. They may show some tall bars next to holes and short bars with little pattern. For variables with many values we need to group the values in bins. The histogram with binned variables shows bars with the number of observations within each bin. It is good practice to create bins of equal width so each bin covers the same range of values. As the case study demonstrates the size of bins can have important consequences for how the histogram looks like.

Visual inspection of a histogram reveals many interesting properties of a distribution. It can inform us about the number and location of modes: these are the peaks in the distribution that stand out from their immediate neighborhood but may have different height. Most distributions with many values have a center and tails, and the histogram tells us the approximate regions for the center and the tails. Some have extreme values that are values very different from the rest with low frequency. Some distributions are more symmetric than others.

Some of these properties may look somewhat different for different choices of bin size of the histogram (or different parameters for the kernel density). Very wide bins may lump together multiple modes. Statistical software usually compute a recommended bin size and shows the corresponding histogram by default for variables that have many values in the data. It is good practice to start with these default histograms but then experiment with a few alternative bin sizes to make sure that important features of the distributions don't remain hidden.

Kernel densities are an alternative to histograms for variables with many potential values. Instead of bars kernel densities show continuous curves. An intuitive way to think about kernel densities as curves that wrap around the corresponding histograms. Similarly to histograms, there are details to set for kernel densities that may make them look different. In fact there are more things that need to be set for a kernel density. The most important parameter to set is the bandwidth, which is the closer thing to bin size for a histogram. Kernel densities also require setting the type of the kernel. The details of drawing appropriate kernel densities are beyond the scope of this textbook. We advise not to draw kernel densities alone, only perhaps as complementing histograms, unless one knows much about kernel densities.

- A distribution of a variable contains the frequency of each potential value of the variable in the data.

- The histogram is a bar graph showing the frequency, or percentage, of each value of a variable if the variable has few potential values.

- Kernel densities are continuous lines that can be viewed as wrapped around corresponding histograms. Drawing kernel densities is complicated.

4.3 Joint and Conditional Distributions, Covariance, Correlation

The distribution of a variable is nothing else than a list of the probabilities of its potential values. The concepts of joint and conditional probabilities and independence have their counterparts in distributions. The joint distribution of two variables shows the probabilities of each value combination of the two variables. Joint distributions are sometimes visualized as three-dimensional histograms or three-dimensional kernel densities. Such graphs are not always easy to comprehend. A less ambitious but often more informative visualization of joint distributions is the scatterplot. A scatterplot is a two-dimensional graph with the values of each of the two variables measured on its two axes, and dots entered for the value-combinations that occur in the dataset.

Conditional distributions are distributions of one variable for each and every separate value of another variable. If two variables are independent the conditional distribution of one is the same for each and every value of the other one, and it is thus the same as the simple (unconditional) distribution. Two variables are positively dependent if larger, more positive values of one variable are more likely if the other variable has larger, more positive values than if that other variable has smaller, less positive values. Two variables are positively dependent if smaller, less positive values of one variable are more likely if the other variable has larger, more positive values than if that other variable has smaller, less positive values.

The covariance, and its relative, the correlation coefficient, are measures of dependence. The correlation coefficient, sometimes simply called the correlation is a standardized version of the covariance. While the covariance may be any positive or negative number the correlation is bound to be between negative one and positive one. When two variables are independent the covariance and the correlation are both zero. When two variables are positively dependent the covariance and the correlation are positive. When two variables are negatively dependent the covariance and the correlation are negative.

Two variables are perfectly linearly dependent when one is a linear function of the other.

In that case the correlation coefficient is positive one or negative one. The formula for the covariance between to variables x and y both observed in a dataset with n observations is:

$$\text{Cov}(x, y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n} \tag{10}$$

Spending a little time staring at this formula can help understand why the covariance measures dependence the way it does. x and y are independent if observations for which y is larger than its $\overline{y}$ average x is as likely to be larger than its average $\overline{x}$ as smaller than it. Those equally likely events average to zero so the covariance is zero. If y and x are positively dependent we have that x_i tends to deviate from its mean $\overline{x}$ in the same direction as y_i deviates from its mean $\overline{y}$. Therefore when we multiply these two deviations we tend to get positive numbers, so the average of them is positive. If y and x are negatively dependent we have that x_i tends to deviate from its mean $\overline{x}$ in the opposite direction as y_i deviates from its mean $\overline{y}$. Therefore when we multiply these two deviations we tend to get negative numbers, so the average of them is negative. When y and x are linear functions of each other, such as $y_i = a + b * x_i$. In this case, the covariance is equal to b times the variance of x, which can be more formally written as:

$$\text{Cov}(x, y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y}) \cdot b}{n} \tag{11}$$

The correlation coefficient is the standardized version of the covariance, dividing it by the product of the standard deviations of the two variables featured in it:

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{[\text{Std}(x) \cdot \text{Std}(y)]} \tag{12}$$

This puts the maximum value at one and the minimum at negative one, both occurring with perfect linear dependence.

When $y_i = a + b \cdot x_i$ then the standard deviation of y can be written as: $\text{Std}(y) = b \cdot \text{Std}(x)$. Thus, the denominator $[\text{Std}(x) \cdot Std(y)]$ becomes the variance of x multiplied by the absolute value of b. If b is positive, the resulting correlation coefficient is a positive one; if b is negative the resulting correlation coefficient is negative one.

There is nothing in these formulae that tells what kinds of variables x and y should be. One can insert into them binary variables, variables that can take few values, and variables that can take many values. The covariance and the correlation coefficient will always show whether the two variables are independent, positively dependent or negatively dependent. Some argue that the way the correlation coefficient standardizes the covariance is less intuitive for binary variables and variables with few values. For such variables variations on the correlation are also used, but those are beyond the scope of this textbook.

• The joint distribution of two variables shows the probabilities of each value combination of the two variables.

• The scatterplot is a good way to visualize joint distributions.

• The covariance measures the dependence of two variables. It is zero of the variables are independent, positive if the variables are positively dependent and negative if they are negatively dependent.
$\text{Cov}(x, y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n}$

•The correlation coefficient is a standardized version of the covariance and it ranges between -1 and 1.
$\text{Corr}(x, y) = \frac{\text{Cov}(x,y)}{[\text{Std}(x) \cdot \text{Std}(y)]}$

Corr(x, y)=1 indicates a perfect positive linear relationship between x and y. Points lie on an increasing straight line. Corr(x, y)=0 indicates no relationship between the variables. Corr(x, y)=-1 indicates a perfect negative linear relationship between x and y. Points lie on an decreasing straight line.

Practice questions.

1. Why don't we take as a measure of dispersion just the sum of the deviations from the mean? Is this a good measure of variability of the data? What is this value? How does this value change from one dataset to the other?

2. What can you say about the prices of hotels with less than four stars plotted in Figure 2? Is the distribution skewed? To what side is the distribution skewed? Are there extreme values to deal with?

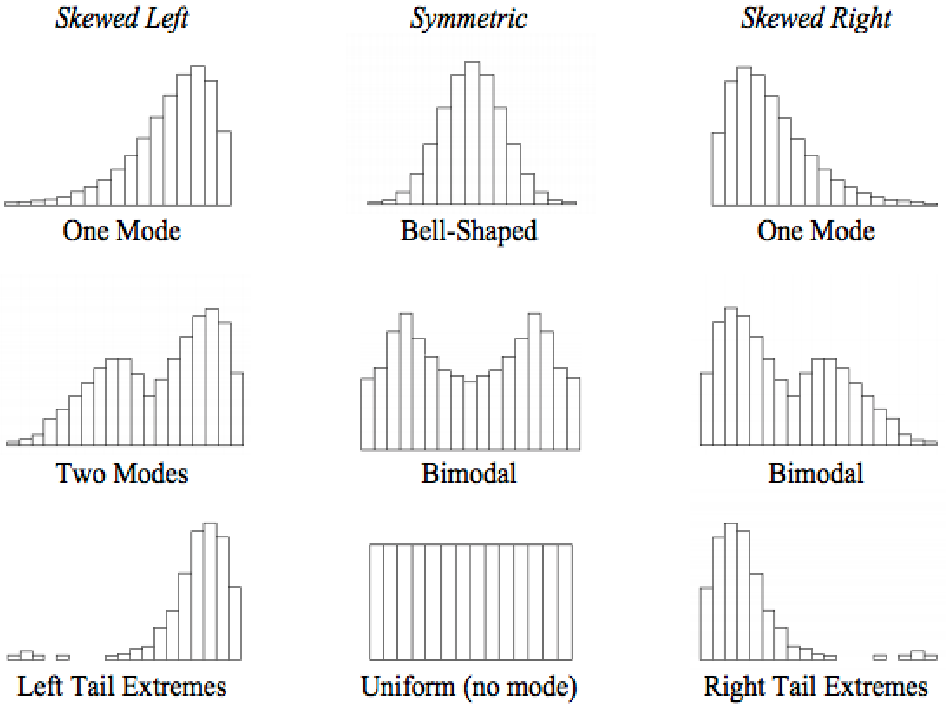
3. In Case Study 1, the hotel price data has has different values occurring twice, namely hotels Hotel Lamee and Topazz with price value 148127 and Hilton Vienna am Stadtpark and Hilton Vienna Plaza with price values 155528. Is this variable bimodal? Yes? No? Why? When is a variable bi-modal?

4. State two independent events. Can independent events happen at the same time?

5. State two mutually exclusive events. Can mutually exclusive events happen at the same time?

6. What is the difference between mutually exclusive events and independent events?
7. State the inverse conditional probability of: $p(\text{go to the gym}|\text{rains})$
8. When does $p(A|B)$=$p(B|A)$.
9. Is it correct that in case *mean*=*median* the distribution is symmetric? Give a simple counter-example.
10. Is the uniform distribution a symmetric one? Why?

Figure 3: Skewness and Distributions



Notes: Examples taken from ”Measuring Skewness: A Forgotten Statistic?”.