1 Basic Statistics

Statistics are used everywhere.

Weather forecasts estimate the probability that it will rain tomorrow based on a variety of atmospheric measurements. In statistics, especially in regression analysis, we call them explanatory variables, while the weather forecast itself is called a dependent variable. The same goes to everwhere we use statstics. Our email clients estimate the probability that incoming email is spam using features found in the email message. By querying a relatively small group of people, pollsters can gauge the pulse of a large population on a variety of issues, including who will win an election. Although the details of the polling are and can be questioned in a variety of aspects, what is surprising is the election outcome is not far from the pollsters' expectation.

On top of this, the past decade or so has seen an explosion in the amount of data we collect across many fields. For example,

- The Large Hadron Collider, the world's largest particle accelerator, produces 15 petabytes of data about particle collisions every year: that's 10^{15} bytes, or a million gigabytes.
- Biologists are generating 15 petabytes of data a year in genomic information.
- The internet is generating 1826 petabytes of data every day. The NSA's analysts claim to look at 0.00004% of that traffic, which comes out to about 25 petabytes per year!

And those are just a few examples! Statistics plays a key role in summarizing and distilling data (large or small) so that we can make sense of it. Obviously, there are certain set of data that many experts regard useless, but as with the growth of aggregate amount of data, data scientists find ways to apply statistical models into unknown territory of data research.

While statistics is an essential tool for justifying a variety of results in research projects, many researchers lack a clear grasp of statistics, mis-using its tools and producing all sorts of bad science! The goal of the course is to help you avoid falling into that trap: I'll arm you with the proper tools to produce sound statistical, known as data scientific, analyses.

In particular, I'll do this by presenting important statistical tools and techniques while emphasizing their underlying principles and assumptions.

Let me start with a motivating example of how powerful statistics can be when they're used properly, and then dive into definitions of basic statistical concepts, exploratory analysis methods, and an overview of some commonly used probability distributions.

EXAMPLE: UNCOVERING DATA FAKERS

In 2008, a polling company called Research 2000 was hired by Daily Kos to gather approval data on top politicians (shown below). Do you see anything odd?

	Favorable		Unfavorable		Undecided	
Topic	Men	Women	Men	Women	Men	Women
Obama	43	59	54	34	3	7
Pelosi	22	52	66	38	12	10
Reid	28	36	60	54	12	10
McConnell	31	17	50	70	19	13
Boehner	26	16	51	67	33	17
Cong.(D)	28	44	64	54	8	2
Cong.(R)	31	13	58	74	11	13
Party(D)	31	45	64	46	5	9
Party(R)	38	20	57	71	5	9

Several amateur statisticians noticed that within each question, the percentages from the men almost always had the same parity (odd-/even-ness) as the percentages from the women. In other words, all numbers are odd, if men's approval rating is odd, and if one number is even for one politician, all other

numers are even. If they truly had been sampling people randomly, this should have only happened about half the time. This table only shows a small part of the data, but it happened in 776 out of the 778 pairs they collected. The probability of this happening by chance is less than 10^{-228}!

What can be expected, at the back of the scene, is that the pollster probably had hard time collecting answers from targeted people. They not only rounded numbers, but it is most likely they manipulated the outcome, to look "clean". The table may give us a tendency, if collected for weeks, but the number itself is no longer trustworthy.

As with the above example, almost all summary data that are available in real world are "touched". Some of them are just too simple to be visible as discussed in the example. What is required to unwrap the decoration is not a simple coding library that claims machine learning can do everything automatically. You need a scientific tool to do the job robustly. The very tool for that has been called statistics, and only recently re-named as data science.

One final comment on the set size of data is that it is a commonly accepted term that the sheer size is an important, and mostly a critical factor to determine whether the data is "BigData" or not. Although the formal definition of "BigData" will come in later chapters, it is important to emphasize that the summary statistics in the above example does not qualify to be a big data, even if it is polled by the entire population of the U.S. The key reason for disqualification is the absence of multiple patterns in the poll. The data set represents a single pattern of the U.S. citizens' political division.

1.1 Introduction

We start with some informal definitions:

- **Probability** is used when we have some model or representation of the world and want to answer questions like "what kind of data will this truth produce?"
- **Statistics** is what we use when we have data and want to discover the "truth" or model underlying the data. In fact, some of what we call statistics today used to be called "inverse probability".

I'll focus on situations where we observe some set of particular outcomes, and want to figure out "why did we get these points?" It could be because of some underlying model or truth in the world (in this case, we're usually interested in understanding that model), or because of how we collected the data (this is called bias, and we try to avoid it as much as possible).
There are two schools of statistical thought:

- Loosely speaking, the **Frequentist** viewpoint holds that the parameters of probabilistic models are fixed, but we just don't know them. These notes focuses on classical frequentist statistics.
- The **Bayesian** viewpoint holds that model parameters are not only unknown, but also random. In this case, we'll encode our prior belief about them using a probability distribution.

To give an example of the difference, assume you have tossed a coin for 100 times to see if it is a fair coin. If the head counts 48 and tail 52, you put 0.48 - 0.5 divided by standard error. Then, if you feel the amount of trial is insufficient, you toss the coin twenty times more. Now with 120 coin toss, the frequentists give the same weight to 120 trials, but the Bayesian statisticians may give different weights to the most recent 20 trials. For a coin toss, it does not seem worth differentiating both set of events, but if it is a traffic time from home to work, as time goes by, there could be other factors affecting the total travel time. In other words, Bayesian can be applied to cases where background states are changing overtime.
Data comes in many types. Here are some of the most common:

- Categorical: discrete, not ordered (e.g., 'red', 'blue', etc.). Binary questions such as polls also fall into this category.
- Ordinal: discrete, ordered (e.g., survey responses like 'agree', 'neutral', 'disagree')
- Continuous: real values (e.g., 'time taken').

- Discrete: numeric data that can only take on discrete values can either be modeled as ordinal (e.g., for integers), or sometimes treated as continuous for ease of modeling.

A **random variable** is a quantity (usually related to our data) that takes on random values. For a discrete random variable, **probability distribution** p describes how likely each of those random values are, so $p(a)$ refers to the probability of observing value a. The **empirical distribution** of some data (sometimes informally referred to as just the distribution of the data) is the relative frequency of each value in some observed dataset. We'll usually use the notation $x_1, x_2, ..., x_n$ to refer to data points that we observe. We'll usually assume our sampled data points are *independent* and *identically distributed*, or i.i.d., meaning that they're independent and all have the same probability distribution.

If you can't grasp what it means by i.i.d., rethink a situation where i.i.d. is broken. If sampled data are dependent to each other, can you trust the mean, variance, thus t-statistics? If it is not identically distributed, would the variance be the same across data? These types of extensions will be discussed in later lecture, but statistics, as with any science, start from the most simple and basic setting. Because extensions are complicated, and often requires in-depth understanding of the simple logic first.
The **expectation** of a random variable is the average value it takes on:

$$\mathbb{E}(x) = \sum_{all\ a} p(a) \cdot a$$

We'll often use the notation μ_x to represent the expectation of random variable x, which is also known as "Mean" or "Average", but the more precise expression for this case is "Equal-weighted Mean". Later in the course, you will see unequal-weighted statistics.
Expectation is linear: for any random variables x, y and constants c, d,

$$\mathbb{E}(cx + dy) = c\mathbb{E}(x) + d\mathbb{E}(y).$$

This is a useful property, and it's true even when x and y aren't independent!

The **variance** of a random variable is a measure of how spread out it is:

$$var(x) = \sum_{all\ a} p(a) \cdot (a - \mathbb{E}(x))^2$$

Note that $p(a)$ is multiplied to the square term, which represent any potential unequal weighting. As mentioned for mean, it does not have to be equal-weighted.

For any constant c, $var(cx) = c^2 var(x)$. If random variables x and y are independent, then $var(x + y) = var(x) + var(y)$; if they are not independent then this is not necessarily true!

The standard deviation is the square root of the variance. We'll often use the notation σ_x to represent the standard deviation of random variable x.

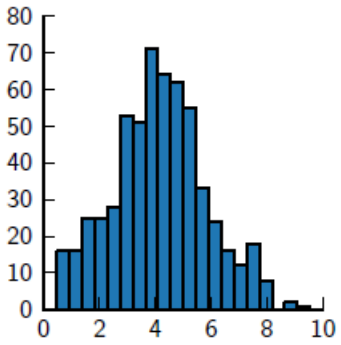
Another common terminology is to describe the mean and variance as moments. Moments are, in simple terms, a value-weighted averaging, which is basically expectation operation. Given square operation, or in mathematical terms, the second order operation, the variance is categorized as a second-moment operator. Similarly, mean is a first-moment operator.

1.2 Exploratory Analysis

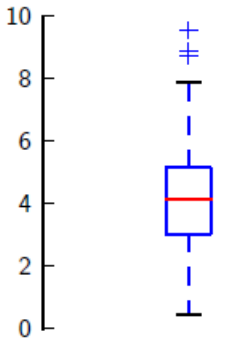
This section lists some of the different approaches we'll use for exploring data. This list is not exhaustive but covers many important ideas that will help us find the most common patterns in data.

Some common ways of plotting and visualizing data are shown in Figure 1.1. Each of these has its own strengths and weaknesses, and can reveal different patterns or hidden properties of the data.

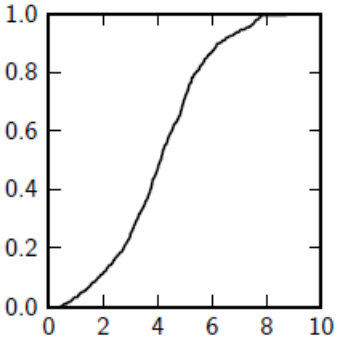
Much of the analysis we'll look at in this class makes assumptions about the data. It's important to check for complex effects; analyzing data with these issues often requires more sophisticated models. For example,



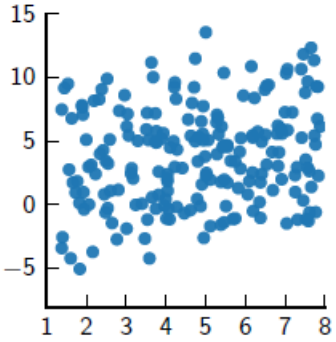
(a) Histogram: this shows the distribution of values a variable takes in a particular set of data. It's particularly useful for seeing the shape of the data distribution in some detail.



(b) Boxplot: this shows the range of values a variable can take. It's useful for seeing where most of the data fall, and to catch outliers. The line in the center is the median, the edges of the box are the 25th and 75th percentiles, and the lone points by themselves are outliers.

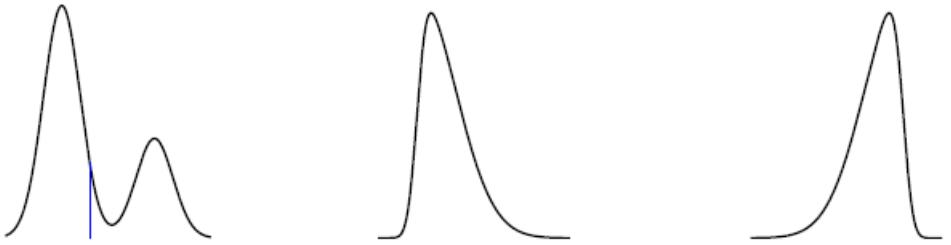


(c) Cumulative Distribution Function (CDF): this shows how much of the data is less than a certain amount. It's useful for comparing the data distribution to some reference distribution.



(d) Scatterplot: this shows the relationship between two variables. It's useful when trying to find out what kind of relationship variables have.

Figure 1: Figure 1.1: Different ways of plotting data



(a) A distribution with two modes. The mean is shown at the blue line.

(b) A right-skewed distribution (positive skew); the tail of the distribution extends to the right.

(c) A left-skewed distribution (negative skew); the tail of the distribution extends to the left.

Figure 2: Figure 1.2: Different irregularities that can come up in data

- Are the data multi-modal? In Figure 1.2a, the mean is a bad representation of the data, since there are two peaks, or in statistical term, modes, of the distribution.
- Are the data skewed? Figures 1.2b and 1.2c show the different kinds of skew: a distribution skewed to the right has a longer tail extending to the right, while a left-skewed distribution has a longer tail extending

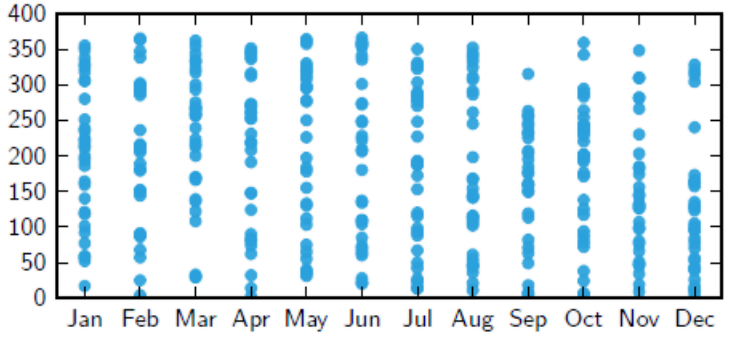
to the left.

Before we start applying any kind of analysis (which will make certain assumptions about the data), it's important to visualize and check that those properties are satisfied.

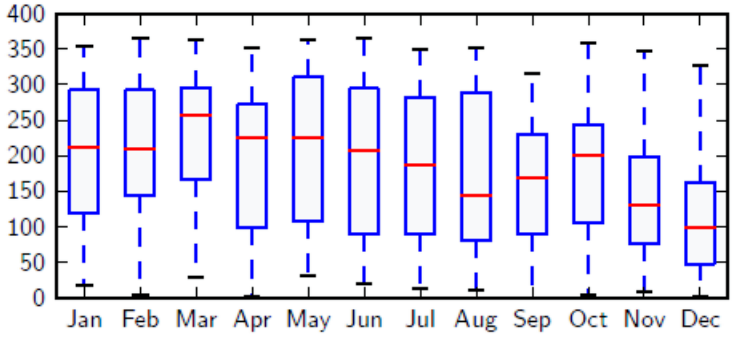
EXAMPLE: VISUALIZING BIAS IN THE VIETNAM DRAFT LOTTERY, 1970

In 1970, the US military used a lottery to decide which young men would be drafted into its war with Vietnam. The numbers 1 through 366 (representing days of the year) were placed in a jar and drawn one by one.

The number 258 (representing September 14) was drawn first, so men born on that day would be drafted first. The lottery progressed similarly until all the numbers were drawn, thereby determining the draft order. The following scatter plot shows draft order (lower numbers indicate earlier drafts) plotted against birth month. Can you spot a pattern?



There seem to be a lot fewer high numbers (later drafts) in the later months and a lot fewer low numbers (earlier drafts) in the earlier months. The following boxplot shows the same data:



The boxplot shows varying weights on different days in a month, and it is definitely the case that a major chunk of data points are distributed in the middle.
In fact, the lottery organizers hadn't sufficiently shuffled the numbers before the drawing, and so the unlucky people born near the end of the year were more likely to be drafted sooner.

1.3 Problem setup

Suppose we've collected a few randomly sampled points of data from some population. If the data collection is done properly, the sampled points should be a good representation of the population, but they won't be perfect. From this random data, we want to estimate properties of the population.

We'll formalize this goal by assuming that there's some "true" distribution that our data points are drawn from, and that this distribution has some particular mean μ and variance σ^2. We'll also assume that our data points

are i.i.d. according to this distribution.

For the rest of the class, we'll usually consider the following data setup:

- We've randomly collected a few samples $x_1, ..., x_n$ from some population. We want to find some interesting properties of the population (we'll start with just the mean, but we'll explore other properties as well).
- In order to do this, we'll assume that all data points in the whole population are randomly drawn from a distribution with mean μ and standard deviation σ (both of which are usually unknown to us: the goal of collecting the sample is often to find them). We'll also assume that our data points are independent.

1.4 Quantitative measures and summary statistics

Here are some useful ways of numerically summarizing sample data:

- **Sample Mean**: $\bar{x} = \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Sample Variance**: $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Median**: the middle value when the data are ordered, so that 50% of the data are above and 50% are below.
- **Percentiles**: an extension of median to values other than 50%.
- **Inter-Quartile range (IQR)**: the difference between the 75th and 25th percentile
- **Mode**: the most frequently occuring value
- **Range**: The minimum and maximum values

Notice that most of these fall into one of two categories: they capture either the center of the distribution (e.g., mean, median, mode), or its spread (e.g., variance, IQR, range). These two categories are often called measures of central tendency and measures of dispersion, respectively.

How accurate are these quantitative measures? Suppose we try using the sample mean $\widehat{\mu}$ as an estimate for μ. $\widehat{\mu}$ is probably not going to be exactly the same as μ, because the data points are random. So, even though μ is fixed, $\widehat{\mu}$ is a random variable (because it depends on the random data). On average, what do we expect the random variable $\widehat{\mu}_x$ to be? We can formalize this question by asking "What's the expectation of $\widehat{\mu}_x$, or $\mathbb{E}(\widehat{\mu}_x)$?"

$$\begin{aligned}
\mathbb{E}(\widehat{\mu}_x) &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} x_i\right] && \text{(definition of } \widehat{\mu}\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] && \text{(linearity of expectation)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mu = \mu
\end{aligned}$$

This result makes sense: while $\widehat{\mu}_x$ might sometimes be higher or lower than than the true mean, on average, the bias (i.e., the expected difference between these two) will be 0.

Deriving the formula for the sample variance $\widehat{\sigma}^2$ requires a similar (but slightly more complicated) process; we obtain $\mathbb{E}[\widehat{\sigma}^2] = \sigma^2$. Notice that we divide by $n-1$ in the denominator and not n. Intuitively, we have to do this because $\bar{x}$, which is not the true mean μ but is instead an estimate of the true mean, is "closer" to each of the observed values of x's compared to the true mean μ. Put another way, the distance between each observed value of x and $\bar{x}$ tends to be smaller than the distance between each observed value of x and μ. In the case of expectation, some such errors were positive and others were negative, so they cancelled out on average. But, since we're squaring the distances, our values, $(x_i - \bar{x})^2$, will be systematically lower than the true ones, $(x_i - \mu)^2$. So, if we divide by n instead of $n-1$, we'll end up underestimating our uncertainty. For a more rigorous derivation, see the supplementary materials at the course website.

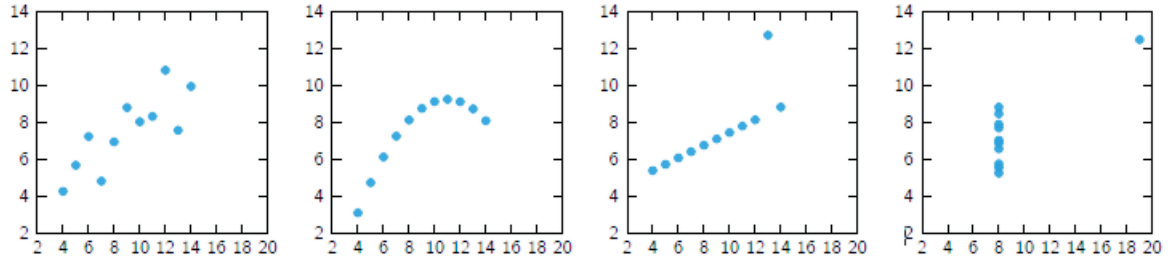
It's often tempting to compute quantitative measures like mean and variance and move on to analyzing them, but these summary statistics have important limitations, as the next example illustrates.

EXAMPLE: ANSCOMBE'S QUARTET

Suppose I have 4 datasets of (x,y) pairs, and they all have the following properties:

- For random variable x, the estimate $\bar{x}$ for the mean and the estimate $\widehat{\sigma}_x^2$ for the variance are 9 and 11 respectively
- For random variable y, the estimate $\bar{y}$ for the mean and the estimate $\widehat{\sigma}_y^2$ for the variance are 7.50 and 4.12 respectively
- The correlation between x and y is 0.816. We'll explain precisely what this means in a couple lectures, but roughly speaking, it's a measure of how well y and x predict each other.

At this point, it would be easy to declare the datasets all the same, or at least very similar, and call it a day. But, if we make scatterplots for each of these datasets, we find that they're actually very different:



These datasets were constructed by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing and visualizing data. It's easy to get lost in crunching numbers and running tests, but the right visualization can often reveal hidden patterns in a simple way. We'll see these again in a few lectures when we discuss regression.

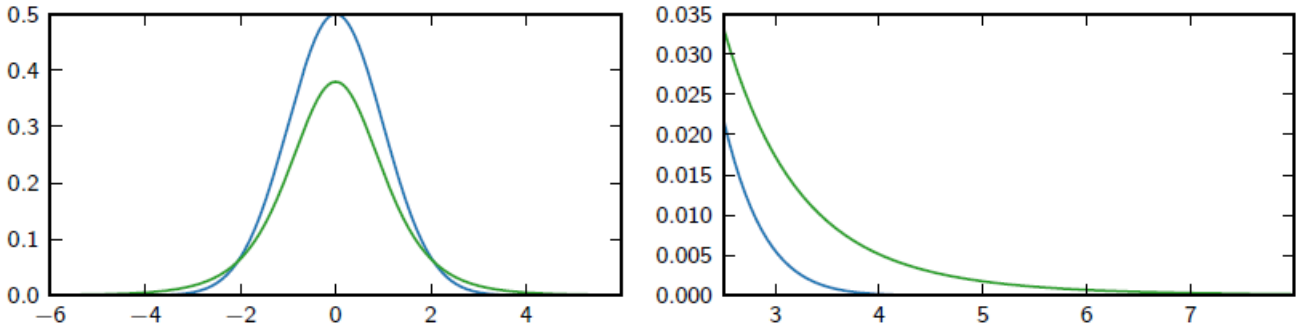


Figure 3: Figure 1.3: The standard normal distribution (blue) and Student t distribution with 5 degrees of freedom (green). The second plot zooms in on the x-axis from 2.5 to 8. Notice that the t distribution has *heavier tails*: that is, the probability of obtaining a value far away from the mean is higher for the t than for the normal.

1.5 Important Distributions

Here are some important probability distributions we'll use to model data. As we use these distributions to model data, we'll want to understand their properties and be able to compute probabilities based on them.

1. **Gaussian/Normal:** This is the common "bell curve" that you've probably heard about and seen before. We'll use it often for continuous data. We say $x \sim N(\mu, \sigma^2)$ to mean that x is drawn from a Gaussian (or Normal) distribution with mean μ and variance σ^2 (or equivalently standard deviation σ). We'll often use the standard normal distribution, or $N(0, 1)$ (i.e., mean 0 and variance 1).

   Here are some useful facts about Gaussian random variables:

   - If $x \sim N(\mu, \sigma^2)$, and we define $y = (x - \mu)/\sigma$, then $y \sim N(0, 1)$. y is a standardized version of x.
   - They're very unlikely to be too far from their mean. The probability of getting a value within 1 standard deviation of the mean is about 68%. For 2 standard deviations, it's about 95%, and for 3 standard deviations it's about 99%. This is sometimes called the "68-95-99 rule".
   - Computing probabilities with Gaussian random variables only requires knowing the mean and variance. So, if a we're using a Gaussian approximation for some distribution (and we know the approximation works reasonably well), we only have to compute the mean and variance of the distribution that we're approximating.

   Figure 1.3 illustrates the Gaussian distribution along with the Student t distribution (described below).

2. **Bernoulli:** A Bernoulli random variable can be thought of as the outcome of flipping a biased coin, where the probability of heads is p. To be more precise, a Bernoulli random variable takes on value 1 with probability p and value 0 with probability $1 - p$. Its expectation is p, and its variance is $p(1 - p)$.

   Bernoulli variables are typically used to model binary random variables.

3. **Binomial:** A binomial random variable can be thought of of as the number of heads in n independent biased coinflips, where each coinflip has probability p of coming up heads. It comes up often when we aggregate answers to yes/no questions. Suppose we have Bernoulli-distributed random variables $x_1, ...x_n$, where each one has probability p of being 1 and probability $1 - p$ of being 0. Then $b = \sum_{i=1}^{n} x_i$ is a binomial random variable. We'll use the notation $b \sim B(n, p)$ as shorthand for this.

   Since the expectation of each flip is p, the expected value of b is np:

   $$\mathbb{E}(b) = \sum_{i=1}^{n} \mathbb{E}(x_i) = \sum_{i=1}^{n} p = np$$

   Since the variance of each flip is $p(1 - p)$ and they're all independent, the variance of b is $np(1 - p)$:

   $$Var(b) = \sum_{i=1}^{n} var(x_i) = \sum_{i=1}^{n} p(1 - p) = np(1 - p)$$

4. **Chi-Squared** (χ^2): We'll sometimes see random variables that arise from summing squared quantities, such as variances or errors. This is one motivation for defining the chi-squared random variable as a sum of several standard normal random variables.
   To be a bit more formal, suppose we have $x_1, ..., x_r$ that are i.i.d., and $x_i \sim N(0, 1)$. If we define $y = \sum_{i=1}^{r} x_i^2$, then y is a chi-squared random variable with r degrees of freedom: $y \sim \chi^2(r)$.

   Note that not every sum of squared quantities is chi-square!

5. **Student t distribution:** When we want to draw conclusions about a Gaussian variable for which the standard deviation is unknown, we'll use a student t distribution (we'll see why in more detail in the next chapter).

Formally, suppose $z \sim N(0, 1)$ and $u \sim \chi^2(r)$. The quantity $t = \frac{z}{\sqrt{u/r}}$ is distributed according to the Student's t-distribution. Figure 1.3 illustrates the t distribution along with the normal distribution.

EXAMPLE: WARNING OF THE DAY: ECOLOGICAL FALLACY

It's important to be careful about aggregate data and individual data. In particular, aggregate data can't always be used to draw conclusions about individual data!

For example, in 1950, a statistician named William S. Robinson looked at each of the 48 states and for each one computed the literacy rate and the fraction of immigrants. These two numbers were positively correlated: the more immigrants a state had, the more literate that state was. Here's a graph of his data:

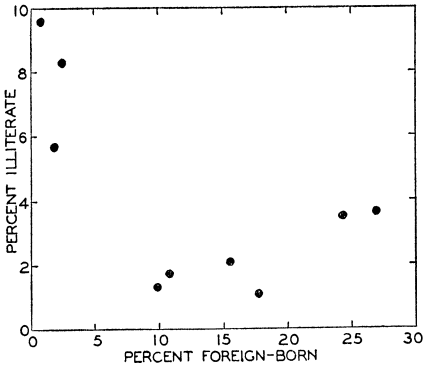


FIG. 3

You might immediately conclude from this that immigrants in 1950 were more literate than non-immigrants, but in fact, the opposite was true! When he went through and looked at individuals, immigrants were on average less literate:

	Foreign Born	Native Born	Total
Illiterate	1304	2614	3918
Literate	11913	81441	93354
Total	13217	84055	97272

The reason he'd made the first finding about the states was that immigrants were more likely to settle in states that already had high literacy rates. So even though they were on average less literate, they ended up in places that had higher literacy rates.

In the 2004 U.S. presidential election, George W. Bush won the 15 poorest states, and John Kerry won 9 of the 11 richest states. But, 64% of poor voters (voters with an annual income below $15,000) voted for Kerry, while 62% of rich voters (with an annual income over $200; 000) supported Bush. This happened because income affected voting preference much more in poor states than in rich states. So, when Kerry won rich states, the rich voters in those states were the few rich voters who leaned Democratic. On the other hand, in the poorer states where Bush won, the rich voters leaned heavily Republican and therefore gave him the boost in those states.

In other words, a variable has varying level of effects to the outcome, depending on the surrounding conditions. This tells us that statistical analysis that is based on a single variable explanation, for example, wage difference in gender, is exposed to fallacy due to unobserved or unidentified factors.
Here's a more concrete simple example: suppose we have datasets $x = \{1, 1, 1, 1\}$ and $y = \{2, 2, 2, -100\}$. $\bar{x} = 1$ and $\bar{y} = -23.5$, so in aggregate, $\bar{x} > \bar{y}$. But the x values are usually smaller than the y values when examined individually. More thorough discussion of omitted variable bias will be presented in later chapters.

2 Confidence intervals and hypothesis tests

This chapter focuses on how to draw conclusions about populations from sample data. We'll start by looking at binary data (e.g., polling), and learn how to estimate the true ratio of 1s and 0s with confidence intervals, and then test whether that ratio is significantly different from some baseline value using hypothesis testing. Then,

we'll extend what we've learned to continuous measurements.

2.1 Binomial data

Suppose we're conducting a yes/no survey (thus boolean type) of a few randomly sampled people, and we want to use the results of our survey to determine the answers for the overall population.

2.1.1 The estimator

The obvious first choice is just the fraction of people who said yes. Formally, suppose we have n samples $x_1, ..., x_n$ that can each be 0 or 1, and the probability that each x_i is 1 is p (in frequentist style, we'll assume p is fixed but unknown: this is what we're interested in finding). We'll assume our samples are indendepent and identically distributed (i.i.d.), meaning that each one has no dependence on any of the others, and they all have the same probability p of being 1. Then our estimate for p, which we'll call $\widehat{p}$, or "p-hat" would be

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Notice that $\widehat{p}$ is a random quantity, since it depends on the random quantities x_i. In statistical lingo, $\widehat{p}$ is known as an estimator for p. Also notice that except for the factor of $\frac{1}{n}$ in front, $\widehat{p}$ is almost a binomial random variable (that is, $(n\widehat{p}) \sim B(n, p)$). We can compute its expectation and variance using the properties we reviewed:

$$\mathbb{E}(\widehat{p}) = \frac{1}{n}np = p, \tag{1}$$

$$var(\widehat{p}) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}. \tag{2}$$

Since the expectation of $\widehat{p}$ is equal to the true value of what $\widehat{p}$ is trying to estimate (namely p), we say that $\widehat{p}$ is an unbiased estimator for p. Reassuringly, we can see that another good property of $\widehat{p}$ is that its variance decreases as the number of samples increases.

2.1.2 Central Limit Theorem

The Central Limit Theorem, one of the most fundamental results in probability theory, roughly tells us that if we collect mean of a bunch of independent random variables that all have the same distribution, the result will be approximately Gaussian.

A lot of people, even experienced statisticians, are confused that it is a sum of means from each random sample from the same population that converges to Gaussian, but no more than that. It is NOT a Gaussian distribution if you just add up all independent random variables, even if all are from the same distribution. Since it is mean value only, whichever the starting distribution you have, it does not matter, and the sum of each sample mean goes to approximately Gaussian.

We can apply this to our case of a binomial random variable, which is really just the sum of a bunch of independent Bernoulli random variables. As a rough rule of thumb, if p is close to 0.5, the binomial distribution will look almost Gaussian with $n = 10$. If p is closer to 0.1 or 0.9 we'll need a value closer to $n = 50$, and if p is much closer to 1 or 0 than that, a Gaussian approximation might not work very well until we have much more data. And it is an extreme probability, we often assian a Poisson distribution to that. See below derivation.

$$\begin{aligned} P[X = i] &= \frac{n!}{(n-i)!i!}p^i(1-p)^{n-i} \\ &= \frac{n!}{(n-i)!i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\cdot(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} \end{aligned}$$

Note that below conditions are satisfied if $n \to \infty$,

$$\left(1 - \frac{\lambda}{n}\right)^n \approxeq e^{-\lambda}$$
$$\frac{n(n-1) \cdot (n-i+1)}{n^i} \approxeq 1$$
$$\left(1 - \frac{\lambda}{n}\right)^i \approxeq 1$$

Then, combining all three terms, we have a Poisson probability density function.

$$P\left[X = i\right] \approxeq e^{-\lambda} \frac{\lambda^i}{i!}$$

This is useful for a number of reasons. One is that Gaussian variables are completely specified by their mean and variance: that is, if we know those two things, we can figure out everything else about the distribution (probabilities, etc.). So, if we know a particular random variable is Gaussian (or approximately Gaussian), all we have to do is compute its mean and variance to know everything about it.
And for Poisson distribution, since mean and variance are the same, only one variable is needed to gauge of the form of the distribution.

2.1.3 Sampling Distributions

Going back to binomial variables, let's think about the distribution of $\widehat{p}$ (remember that this is a random quantity since it depends on our observations, which are random). Figure 2.1a shows the sampling distribution of $\widehat{p}$ for a case where we flip a coin that we hypothesize is fair (i.e. the true value p is 0.5). There are typically two ways we use such sampling distributions: to obtain confidence intervals and to perform significance tests.

2.1.4 Confidence intervals

Suppose we observe a value $\widehat{p}$ from our data, and want to express how certain we are that $\widehat{p}$ is close to the true parameter p. We can think about how often the random quantity $\widehat{p}$ will end up within some distance of the fixed but unknown p. In particular, we can ask for an interval around $\widehat{p}$ for any sample so that in 95% of samples, the true mean p will lie inside this interval. Such an interval is called a con
dence interval. Notice that we chose the number 95% arbitrarily: while this is a commonly used value, the methods we'll discuss can be used for any confidence level.

We've established that the random quantity $\widehat{p}$ is approximately Gaussian with mean p and variance $p(1-p)/n$. We also know from last time that the probability of a Gaussian random variable being within about 2 standard deviations of its mean is about 95%. This means that there's a 95% chance of $\widehat{p}$ being less than $2\sqrt{p(1-p)/n}$ away from p. So, we'll define the interval

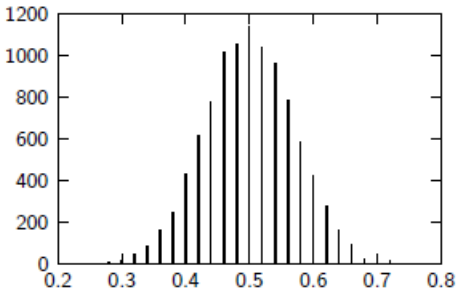
$$\widehat{p} \pm \underbrace{2}_{coeff.} \cdot \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{std.dev.} \tag{3}$$

With probability 95%, we'll get a $\widehat{p}$ that gives us an interval containing p.

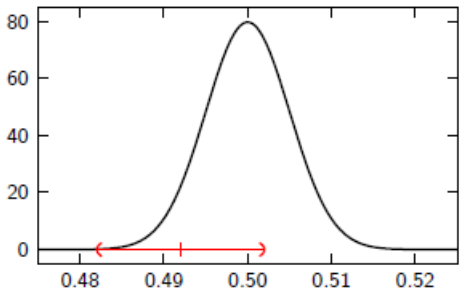
What if we wanted a 99% confidence interval? Since $\widehat{p}$ is approximately Gaussian, its probability of being within 3 standard deviations from its mean is about 99%. So, the 99% confidence interval for this problem would be

$$\widehat{p} \pm \underbrace{3}_{coeff.} \cdot \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{std.dev.} \tag{4}$$

We can define similar confidence intervals, where the standard deviation remains the same, but the coefficient depends on the desired confidence. While our variables being Gaussian makes this relationship easy for 95% and 99%, in general we'll have to look up or have our software compute these coefficients.



(a) The sampling distribution of the estimator $\hat{p}$: i.e. the distribution of values for $\hat{p}$ given a fixed true value $p = 0.5$.



(b) The 95% confidence interval for a particular observed $\hat{p}$ of 0.49 (with a true value of $p = 0.5$). Note that in this case, the interval contains the true value p. Whenever we draw a set of samples, there's a 95% chance that the interval that we get is good enough to contain the true value p.

Figure 4: Figure 2.1

But, there's a problem with these formulas: they requires us to know p in order to compute confidence intervals! Since we don't actually know p (if we did, we wouldn't need a confidence interval), we'll approximate it with $\widehat{p}$, so that (3) becomes

$$\widehat{p} \pm 2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{5}$$

This approximation is reasonable if $\widehat{p}$ is close to p, which we expect to normally be the case. If the approximation is not as good, there are several more robust (but more complex) ways to compute the confidence interval.

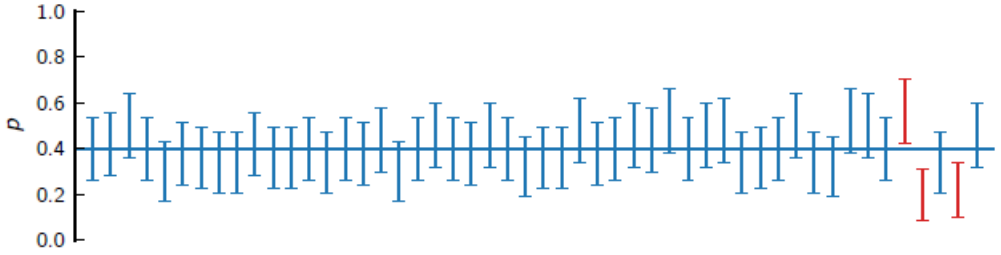


Figure 2.2: Multiple 95% confidence intervals computed from different sets of data, each with the same true parameter $p = 0.4$ (shown by the horizontal line). Each confidence interval represents what we might have gotten if we had collected new data and then computed a confidence interval from that new data. Across different datasets, about 95% of them contain the true interval. But, once we have a confidence interval, we can't draw any conclusions about where in the interval the true value is.

Interpretation

It's important not to misinterpret what a confidence interval is! This interval tells us nothing about the distribution of the true parameter p. In fact, p is a fixed (i.e., deterministic) unknown number! Imagine that we sampled n values for x_i and computed $\widehat{p}$ along with a 95% confidence interval. Now imagine that we repeated this whole process a huge number of times (including sampling new values for x_i). Then about 5% of the con dence intervals constructed won't actually contain the true p. Furthermore, if p is in a confidence interval, we don't know where exactly within the interval p is.

Furthermore, adding an extra 4% to get from a 95% confidence interval to a 99% confidence interval doesn't mean that there's a 4% chance that it's in the extra little area that you added! The next example illustrates this.

In summary, a 95% confidence interval gives us a region where, had we redone the survey from scratch, then 95% of the time, the true value p will be contained in the interval. This is illustrated in Figure 2.2.

2.1.5 Hypothesis testing

Suppose we have a hypothesized or baseline value p and obtain from our data a value $\widehat{p}$ that's smaller than p. If we're interested in reasoning about whether $\widehat{p}$ is "significantly" smaller than p, one way to quantify this would be to assume the true value were p and then compute the probability of getting a value smaller than or as small as the one we observed (we can do the same thing for the case where $\widehat{p}$ is larger). If this probability is "very low", we might think the hypothesized value p is incorrect. This is the hypothesis testing framework.

We begin with a null hypothesis, which we call H_0 (in this example, this is the hypothesis that the true proportion is in fact p) and an alternative hypothesis, which we call H_1 or H_a (in this example, the hypothesis that the true mean is significantly smaller than p).

Usually (but not always), the null hypothesis corresponds to a baseline or boring finding, and the alternative hypothesis corresponds to some interesting finding. Once we have the two hypotheses, we'll use the data to test which hypothesis we should believe.

"Significance" is usually defined in terms of a probability threshold α, such that we deem a particular result significant if the probability of obtaining that result under the null distribution is less than α. A common value for α is 0.05, corresponding to a 1/20 chance of error. Once we obtain a particular value and evaluate its probability under the null hypothesis, this probability is known as a p-value.

This framework is typically used when we want to disprove the null hypothesis and show the value we obtained is significantly different from the null value. In the case of polling, this may correspond to showing that a candidate has significantly more than 50% support.

In the case of a drug trial, it may correspond to showing that the recovery rate for patients given a particular drug is significantly more than some baseline rate.

Here are some definitions:

- In a one-tailed hypothesis test, we choose one direction for our alternative hypothesis: we either hypothesize that the test statistic is "significantly big", or that the test statistic is "significantly small".
- In a two-tailed hypothesis test, our alternative hypothesis encompasses both directions: we hypothesize that the test statistic is simply different from the predicted value.
- A false positive or Type I error happens when the null hypothesis is true, but we reject it. Note that the probability of a Type I error is α.
- A false negative or Type II error happens when the null hypothesis is false, but we fail to reject it
- The statistical power of a test is the probability of rejecting the null hypothesis when it's false (or equivalently, 1-(probability of type II error).

  Power is usually computed based on a particular assumed value for the quantity being tested: "if the value is actually ( ), then the power of this test is ( )." It also depends on the threshold determined by α.

  It's often useful when deciding how many samples to acquire in an experiment, as we'll see later.

Example

The concepts above are illustrated in Figure 2.3. Here, the null hypothesis H_0 is that $p = p_0$, and the alternative hypothesis H_a is that $p > p_0$: this is a one-sided test. In particular, we'll use the value p_a as the alternative value so that we can compute power. The null distribution is shown on the left, and an alternative distribution is shown on the right. The $\alpha = 0.05$ threshold for the alternative hypothesis is shown as p^*.

- When the null hypothesis is true, $\widehat{p}$ is generated from the null (left) distribution, and we make the correct decision if $\widehat{p} < p^*$ and make a Type I error (false positive) otherwise.

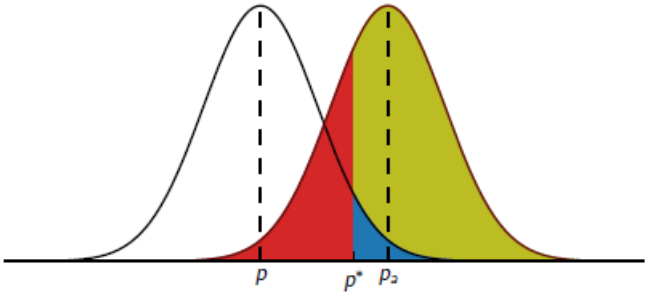


Figure 2.3: An illustration of statistical power in a one-sided hypothesis test on variable p.

- When the alternative hypothesis is true, and if the true proportion p is actually p_a, $\widehat{p}$ is generated from the right distribution, and we make the correct decision when $\widehat{p} > p^*$ and make a Type II error (false negative) otherwise.

The power is the probability of making the correct decision when the alternative hypothesis is true. The probability of a Type I error (false positive) is shown in blue, the probability of a Type II error (false negative) is shown in red, and the power is shown in yellow and blue combined (it's the area under the right curve minus the red part).

Notice that a threshold usually balances between Type I and Type II errors: if we always reject the null hypothesis, then the probability of a Type I error is 1, and the probability of a Type II error is 0, and vice versa if we always fail to reject the null hypothesis.

EXAMPLE: DRUG THERAPHY RESULTS: A WARNING ABOUT DATA COLLECTION

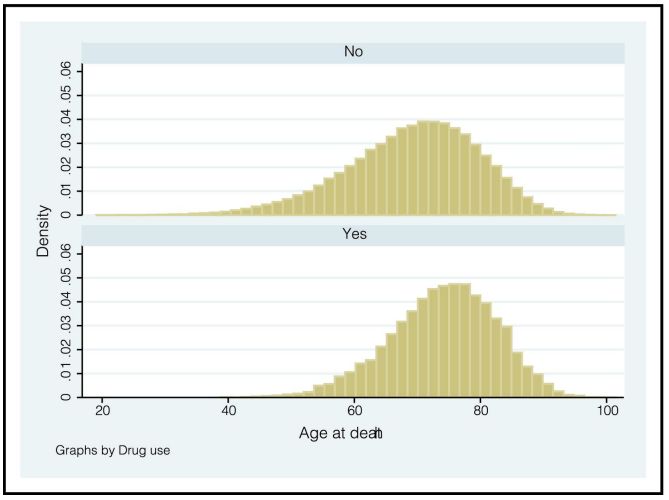


Figure 2.4: Results of a simulated drug trial measuring the effects of statin drugs on lifespan. The top gure shows the lifespan of subjects who did not receive treatment, and the bottom figure shows the lifespan of subjects who did receive it.

Figure 2.4 shows results from a simulated drug triala. At first glance, it seems clear that people who received the drug (bottom) tended to have a higher lifespan than people who didn't (top), but it's important to look at hidden confounds! In this simulation, the drug actually had no effect, but the disease occurred more often in older people: these older people had a higher average lifespan simply because they had to live longer to get the drug.

Any statistical test we perform will say that the second distribution has a higher mean than the first one, but this is not because of the treatment, but instead because of how we sampled the data!

2.2 Continuous random variables

So far we've only talked about binomial random variables, but what about continuous random variables? Let's focus on estimating the mean of a random variable given observations of it. As you can probably guess, our estimator will be $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

We'll start with the case where we know the true population standard deviation; call it σ. This is somewhat unrealistic, but it'll help us set up the more general case.

2.2.1 When σ is known

Consider random i.i.d. Gaussian samples $x_1, ..., x_n$, all with mean μ and variance σ^2. We'll compute the sample mean $\widehat{\mu}$, and use it to draw conclusion about the true mean μ.

Just like $\widehat{p}$, $\widehat{\mu}$ is a random quantity. Its expectation, which we computed in Chapter 1, is μ. Its variance is

$$\begin{aligned} Var(\widehat{\mu}) &= var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} var(x_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}. \end{aligned} \tag{6}$$

This quantity (or to be exact, the square root of this quantity) is known as the standard error of the mean. In general, the standard deviation of the sampling distribution of the a particular statistic is called the standard error of that statistic.

Since $\widehat{\mu}$ is the sum of many independent random variables, it's approximately Gaussian. If we subtract its mean μ and divide by its standard deviation $\sigma/\sqrt{n}$ (both of which are deterministic), we'll get a standard normal random variable. This will be our test statistic:

$$z = \frac{\widehat{\mu} - \mu}{\sigma/\sqrt{n}} \tag{7}$$

Hypothesis testing

In the case of hypothesis testing, we know μ (it's the mean of the null distribution), and we can compute the probability of getting z or something more extreme. Your software of choice will typically do this by using the fact that z has a standard normal distribution and report the probability to you. This is known as a z-test.

Confidence intervals

What about a confidence interval? Since z is a standard normal random variable, it has probability 0.95 of being within 2 standard deviations of its mean. We can compute the confidence interval by manipulating a bit of algebra:

$$\begin{aligned} P(-2 \le z \le 2) &\approx 0.95 \\ P(-2 \le \frac{\widehat{\mu} - \mu}{\sigma/\sqrt{n}} \le 2) &\approx 0.95 \\ P(-2\frac{\sigma}{\sqrt{n}} \le \widehat{\mu} - \mu \le 2\frac{\sigma}{\sqrt{n}}) &\approx 0.95 \\ P(\widehat{\mu} - \underbrace{2}_{coeff.}\underbrace{\frac{\sigma}{\sqrt{n}}}_{std.dev} \le \mu \le \widehat{\mu} + \underbrace{2}_{coeff.}\underbrace{\frac{\sigma}{\sqrt{n}}}_{std.dev}) &\approx 0.95 \end{aligned}$$

This says that the probability that μ is within the interval $\widehat{\mu} \pm 2\frac{\sigma}{\sqrt{n}}$ is 0.95. But remember: the only thing that's random in this story is $\widehat{\mu}$! So when we use the word "probability" here, it's referring only to the randomness in $\widehat{\mu}$. Don't forget that μ isn't random!

Also, remember that we chose the confidence level 0.95 (and therefore the threshold 2) somewhat arbitrarily, and we could just as easily compute a 99% confidence interval (which would correspond to a threshold of about 3) or an interval for any other level of confidence: we could compute the threshold by using the standard normal distribution.

Finally, note that for a two-tailed hypothesis test, the threshold at which we declare significance for some particular α is the same as the width of a confidence interval with confidence level $1-\alpha$. Can you show why this is true?

Statistical power

If we get to choose the number of observations n, how do we pick it to ensure a certain level of statistical power in a hypothesis test? Suppose we choose α and a corresponding threshold x^* How can we choose n, the number of samples, to achieve a desired statistical power? Since the width of the sampling distribution is controlled by n, by choosing n large enough, we can achieve enough power for particular values of the alternative mean.

The following example illustrates the effect that sample size has on significance thresholds.

EXAMPLE: FERTILITY CLINICS

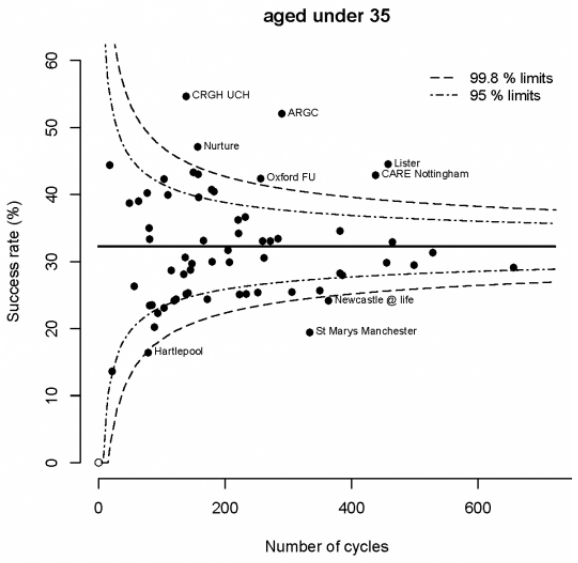


Figure 2.5: A funnel plot showing conception statistics from fertility clinics in the UK. The x-axis indicates the sample size; in this case that's the number of conception attempts (cycles). The y-axis indicates the quantity of interest; in this case that's the success rate for conceiving. The funnels (dashed lines) indicate thresholds for being significantly different from the null value of 32% (the national average).

Figure 2.5 is an example of a funnel plot. We see that with a small number of samples, it's difficult to judge any of the clinics as significantly different from the baseline value, since exceptionally high/low values could just be due to chance. However, as the number of cycles increases, the probability of consistently obtaining large values by chance decreases, and we can declare clinics like Lister and CARE Nottingham significantly better than average: while other clinics have similar success rates over fewer cycles, these two have a high success rate over many cycles. So, we can be more certain that the higher success rates are not just due to chance and are in fact meaningful.

2.2.2 When σ is unknown

In general, we won't know the true population standard deviation beforehand. We'll solve this problem by using the sample standard deviation. This means using $\widehat{\sigma}^2/n$ instead of σ^2/n for $var(\widehat{\mu})$. Throughout these notes, we'll refer to this quantity as the standard error of the mean (as opposed to the version given in Equation (6)).

But once we replace the fixed σ with the random $\widehat{\sigma}$ (which we'll also write as s), our test statistic (Equation (7)) becomes

$$t = \frac{\widehat{\mu} - \mu}{\widehat{\sigma}/\sqrt{n}} \tag{8}$$

Since the numerator and denominator are both random, this is no longer Gaussian. The denominator is roughly χ^2-distributed quantity, and the overall statistic is t-distributed. In this case, our t distribution has $n-1$ degrees of freedom.

Confidence intervals and hypothesis tests proceed just as in the known-σ case with only two changes: using $\widehat{\sigma}$ instead of σ and using a t distribution with $n-1$ degrees of freedom instead of a Gaussian distribution. The confidence interval requires only $\widehat{\mu}$ and the standard error s, while the hypothesis test also requires a hypothesis, in the form of a value for μ.

For example, a 95% confidence interval might look like

$$\widehat{\mu} \pm t^* \frac{\widehat{\sigma}}{\sqrt{n}} \tag{9}$$

To determine the coefficient t^*, we need to know the value where a t distribution has 95% of its probability. This depends on the degrees of freedom (the only parameter of the t distribution) and can easily be looked up in a table or computed from any software package. For example, if $n = 10$, then the t distribution has $n - 1 = 9$ degrees of freedom, and $k = 2.26$. Notice that this produces a wider interval than the corresponding Gaussian-based confidence interval from before. If we don't know the standard deviation and we estimate it, we're then less certain about our estimate $\widehat{\mu}$.

To derive the t-test, we assumed that our data points were normally distributed. But, the t-test is fairly robust to violations of this assumption.

2.3 Two-sample tests (or A/B Tests)

So far, we've looked at the case of having one sample and determining whether it's significantly greater than some hypothesized amount. But what about the case where we're interested in the dfference between two samples? We're usually interested in testing whether the difference is significantly different from zero. There are a few different ways of dealing with this, depending on the underlying data.

- In the case of matched pairs, we have a "before" value and an "after" value for each data point (for example, the scores of students before and after a class). Matching the pairs helps control the variance due to other factors, so we can simply look at the differences for each data point, $x_i^{post} - x_i^{pre}$ and perform a one-sample test against a null mean of 0.
- In the case of two samples with pooled variance, the means of the two samples might be different (this is usually the hypothesis we test), but the variances of each sample are assumed to be the same. This assumption allows us to combine, or pool, all the data points when estimating the sample variance. So, when computing the standard error, we'll use this formula:

  $$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

  Our test statstics is then

  $$t = \frac{\widehat{\mu}^{(1)} - \widehat{\mu}^{(2)}}{s_p\sqrt{(1/n_1) + (1/n_2)}}$$

  This test still provides reasonably good power, since we're using all the data to estimate s_p. In this setting, where the two groups have the same variance, we say the data are homoskedastic.

- In the general case of two samples with separate (not pooled) variance, the variances must be estimated separately. The result isn't quite a t distribution, and this variant is often known as Welch's t-test. It's important to keep in mind that this test will have lower statistical power since we are using less data to estimate each quantity. But, unless you have solid evidence that the variances are in fact equal, it's best to be conservative and stick with this test.
  In this setting, where the two groups have different variances, we say the data are heteroskedastic.

2.4 Some important warnings for hypothesis testing

- **Correcting for multiple comparisons (very important)**: suppose you conduct 20 tests at a significance level of 0.05. Then on average, just by chance, even if the null hypothesis is wrong, one of the tests will show a significant difference. There are a few standard ways of addressing this issue:
  - Bonferroni correction: If we're doing m tests, use a significance value of α/m instead of α. Note that this is very conservative, and will dramatically reduce the number of acceptances.
  - False discovery rate (Benjamini-Hochberg): this technique guarantees α overall error by using the very small signi

    cances to allow slightly larger ones through as well.
- **Rejecting the null hypothesis**: You can never be completely sure that the null hypothesis is false from using a hypothesis test! Any statement stronger than "the data do not support the null hypothesis" should be made with extreme caution.
- **Practical vs statistical significance**: with large enough n, any minutely small difference can be made statistically significant. The first example below demonstrates this point. Sometimes small differences like this matter (e.g., in close elections), but many times they don't.
- **Independent and identically distributed**: Many of our derivations and methods depend on samples being independent and identically distributed. There are ways of changing the methods to account for dependent samples, but it's important to be aware of the assumptions you need to use a particular method or test.

EXAMPLE: PRACTICAL VS STATISTICAL SIGNIFICANCE

Suppose we are testing the fairness of a coin. Our null hypothesis might be $p = 0.5$. We collect 1000000 data points and observe a sample proportion $\widehat{p} = 0.501$ and run a significance test. The large number of samples would lead to a p-value of 0.03. At a 5% significance level, we would declare this significant. But, for practical purposes, even if the true mean were in fact 0.501, the coin is almost as good as fair. In this case, the strong statistical significance we obtained does not correspond to a "practically" significant difference. Figure 2.6 illustrates the null sampling distribution and the sampling distribution assuming a proportion of $p = 0.501$.

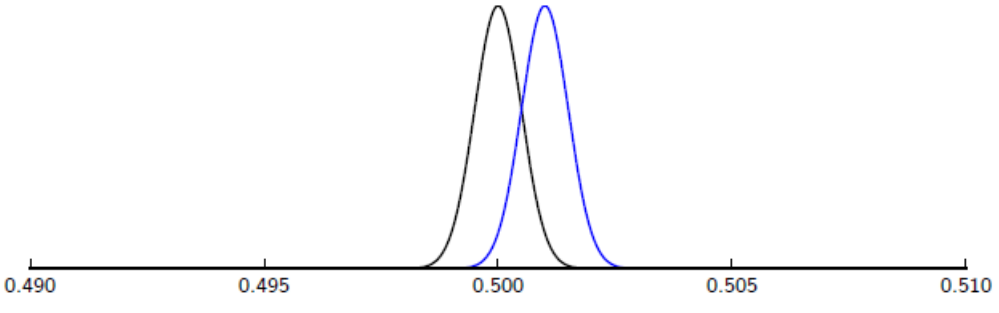


Figure 2.6: Sampling distributions for $p = 0.5$ (black) and $p = 0.501$ (blue) for $n = 1000000$. Note the scale of the x-axis: the large number of samples dramatically reduces the variance of each distribution.

EXAMPLE: PITFALL OF THE DAY: INTERPRETATION FALLACIES AND SALLY CLARK

In the late 1990s, Sally Clark was convicted of murder after both her sons died suddenly within a few weeks of birth. The prosecutors made two main claims:

- The probability of two children independently dying suddenly from natural causes like Sudden Infant Death Syndrome (SIDS) is 1 in 73 million. Such an event would occur by chance only once every 100 years, which was evidence that the death was not natural.
- If the death was not due to two independent cases of SIDS (as asserted above), the only other possibility was that they were murdered.

The assumption of independence in the first item was later shown to be incorrect: the two children were not only genetically similar but also were raised in similar environments, causing dependence between the two events. This wrongful assumption of independence is a common error in statistical analysis. The probability then goes up dramatically.

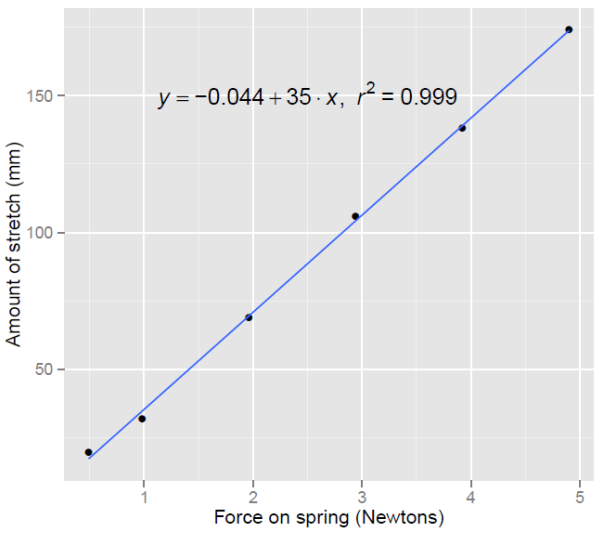
Also, showing the unlikeliness of two chance deaths does not imply any particular alternative! Even if it were true, it doesn't make sense to consider the "1 in 73 million claim" by itself: it has to be compared to the probability of two murders (which was later estimated to be even lower). This second error is known as the $prosecutor's\ fallacy$. In fact, tests later showed bacterial infection in one of the children!

3 Linear regression

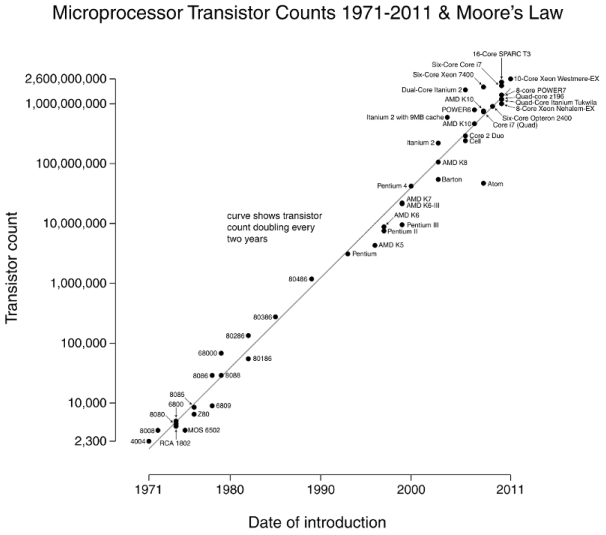
Once we've acquired data with multiple variables, one very important question is how the variables are related. For example, we could ask for the relationship between people's weights and heights, or study time and test scores, or two animal populations. Regression is a set of techniques for estimating relationships, and we'll focus on them for the next two chapters.

In this chapter, we'll focus on fnding one of the simplest type of relationship: linear. This process is unsurprisingly called linear regression, and it has many applications. For exam- ple, we can relate the force for stretching a spring and the distance that the spring stretches (Hooke's law, shown in Figure 3.1a), or explain how many transistors the semiconductor industry can pack into a circuit over time (Moore's law, shown in Figure 3.1b).

Despite its simplicity, linear regression is an incredibly powerful tool for analyzing data. While we'll focus on the basics in this chapter, the next chapter will show how just a few small tweaks and extensions can enable more complex analyses.



(a) In classical mechanics, one could empirically verify Hooke's law by dangling a mass with a spring and seeing how much the spring is stretched.



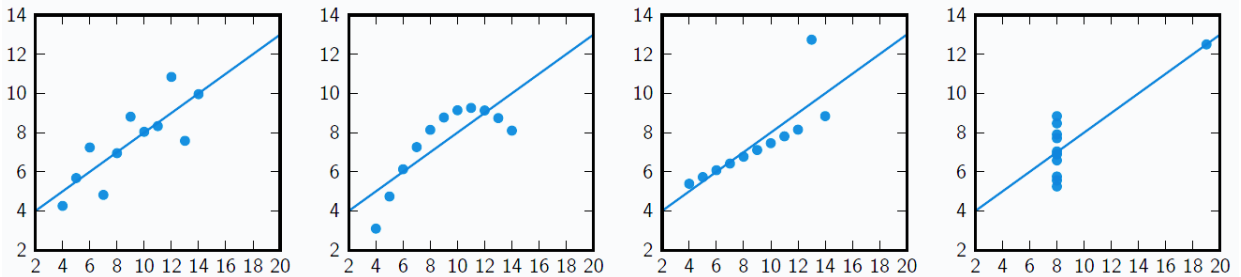
(b) In the semiconductor industry, Moore's law is an observation that the number of transistors on an integrated circuit doubles roughly every two years.

Figure 3.1: Examples of where a line fit explains physical phenomena and engineering feats.

But just because fitting a line is easy doesn't mean that it always makes sense. Let's take another look at Anscombe's quartet to underscore this point.

Example 1. Anscombe's Quartet Revisited

Recall Anscombe's Quartet: 4 datasets with very similar statistical properties under a simple quantitative analysis, but that look very different. Here they are again, but this time with linear regression lines fitted to each one:



For all 4 of them, the slope of the regression line is 0.500 (to three decimal places) and the intercept is 3.00 (to two decimal places). This just goes to show: visualizing data can often reveal patterns that are hidden by pure numeric analysis!

We begin with simple linear regression in which there are only two variables of interest (e.g., weight and height, or force used and distance stretched). After developing intuition for this setting, we'll then turn our attention to multiple linear regression, where there are more variables.

Disclaimer: While some of the equations in this chapter might be a little intimidating, it's important to keep in mind that as a user of statistics, the most important thing is to understand their uses and limitations. Toward this end, make sure not to get bogged down in the details of the equations, but instead focus on understanding how they fit in to the big picture.

3.1 Simple linear regression

We're going to fit a line $y = \beta_0 + \beta_1 x_1$ to our data. Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable.

Before we talk about how to do the fit, let's take a closer look at the important quantities from the fit:

- β_1 is the slope of the line: this is one of the most important quantities in any linear regression analysis. A value very close to 0 indicates little to no relationship; large positive or negative values indicate large positive or negative relationships, respectively. For our Hooke's law example earlier, the slope is the spring constant.
- β_0 is the intercept of the line.

In order to actually fit a line, we'll start with a way to quantify how good a line is. We'll then use this to fit the "best" line we can.

One way to quantify a line's "goodness" is to propose a probabilistic model that generates data from lines. Then the "best" line is the one for which data generated from the line is "most likely". This is a commonly used technique in statistics: proposing a probabilistic model and using the probability of data to evaluate how good a particular model is. Let's make this more concrete.

A probabilistic model for linearly related data

We observe paired data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where we assume that as a function of x_i, each y_i is generated by using some true underlying line $y = \beta_0 + \beta_1 x_1$ that we evaluate at x_i, and then adding some Gaussian noise. Formally,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{3.1}$$

Here, the noise ϵ_i represents the fact that our data won't fit the model perfectly. We'll model ϵ_i as being Gaussian: $\epsilon \sim \mathcal{N}(0, \sigma^2)$ Note that the intercept β_0, the slope β_1, and the noise variance σ^2 are all treated as fixed (i.e., deterministic) but unknown quantities.

Solving for the fit: least-squares regression

Assuming that this is actually how the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we observe are generated, then it turns out that we can find the line for which the probability of the data is highest by solving the following

optimization problem:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{3.2}$$

where $\min_{\beta_0,\beta_1}$ means "minimize over β_0 and β_1". This is known as the least-squares linear regression problem. Given a set of points, the solution is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \tag{3.3}$$

$$= r\frac{s_y}{s_x}, \tag{3.4}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{3.5}$$

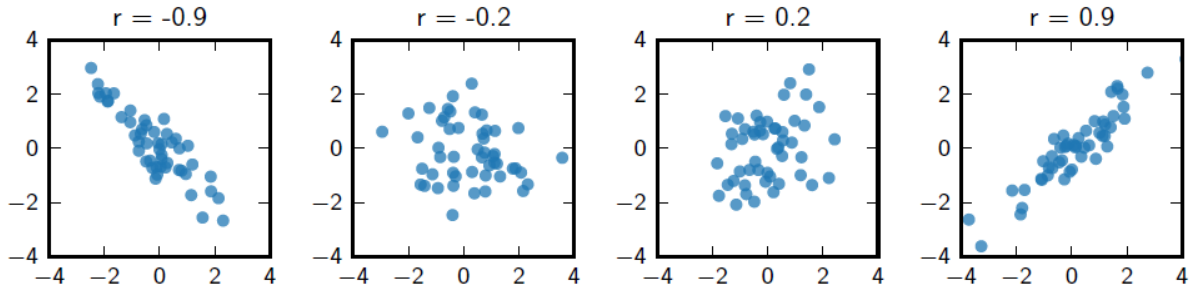


Figure 3.2: An illustration of correlation strength. Each plot shows data with a particular correlation coefficient r. Values farther than 0 (outside) indicate a stronger relationship than values closer to 0 (inside). Negative values (left) indicate an inverse relationship, while positive values (right) indicate a direct relationship.

where $\bar{x}$, $\bar{y}$, s_x and s_y are the sample means and standard deviations for x values and y values, respectively, and r is the correlation coefficient, defined as

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \tag{3.6}$$

By examining the second equation for the estimated slope $\hat{\beta}_1$, we see that since sample standard deviations s_x and s_y are positive quantities, the correlation coefficient r, which is always between -1 and 1, measures how much x is related to y and whether the trend is positive or negative. Figure 3.2 illustrates different correlation strengths.

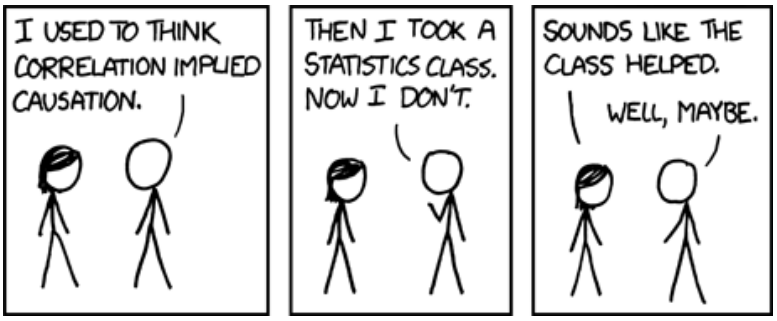
The square of the correlation coefficient r_2 will always be positive and is called the coefficient of determination. As we'll see later, this also is equal to the proportion of the total variability that's explained by a linear model.

As an extremely crucial remark, correlation does not imply causation! This is also known as "Spurious regression", which indicates the regression does not provide any causal interpretation.

An untold number of machine learning models that are in fact an extension of regression have been coined Artificial Intelligence arguing that the "AI" model has a capacity to find a new variable that traditional statistics have failed for centuries. The real truth, often times, is that the brute force machine learning model or blinded regression is nothing more than a spurious regression where variables do NOT have any causal meaning.

We devote the entire next page to this point, which is one of the most common sources of error in interpreting statistics.

Example 2. Correlation and Causation



Just because there's a strong correlation between two variables, there isn't necessarily a causal relationship between them. For example, drowning deaths and ice-cream sales are strongly correlated, but that's because both are affected by the season (summer vs. winter). In general, there are several possible cases, as illustrated below:

$x \longrightarrow y$
$x \longleftarrow y$

(a) **Causal link:** Even if there is a causal link between x and y, correlation alone cannot tell us whether y causes x or x causes y.

z
$x \quad y$

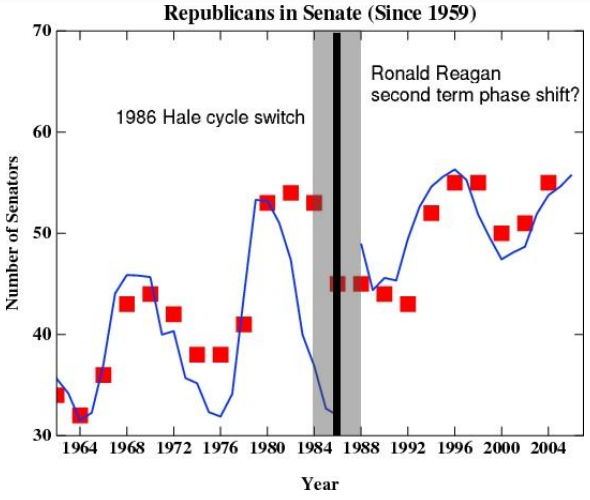
(b) **Hidden Cause:** A hidden variable z causes both x and y, creating the correlation.

z
$x \rightarrow y$

(c) **Confounding Factor:** A hidden variable z and x both affect y, so the results also depend on the value of z.

$x \quad y$

(d) **Coincidence:** The correlation just happened by chance (e.g. the strong correlation between sun cycles and number of Republicans in Congress, as shown below).



(e) The number of Republican senators in congress (red) and the sunspot num- ber (blue, before 1986)/inverted sunspot number (blue, after 1986).

3.2 Tests and Intervals

Recall that in order to do hypothesis tests and compute confidence intervals, we need to know our test statistic, its standard error, and its distribution. We'll look at the standard errors for the most important quantities and their interpretation. Any statistical analysis software can compute these quantities automatically, so we'll focus on interpreting and understanding what comes out.

Warning: All the statistical tests here crucially depend on the assumption that the observed data actually comes from the probabilistic model defined in Equation (3.1)!

Slope

For the slope β_1, our test statistic is

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} \tag{3.7}$$

which has a Student's t distribution with $n-2$ degrees of freedom. The standard error of the slope s_{β_1} is

$$s_{\beta_1} = \frac{\hat{\sigma}}{\sqrt{\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{how close together x values are}}}} \tag{3.8}$$

and the mean squared error $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{\overbrace{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}^{\text{how large the errors are}}}{n-2} \tag{3.9}$$

These terms make intuitive sense: if the x-values are all really close together, it's harder to fit a line. This will also make our standard error s_{β_1} larger, so we'll be less confident about our slope. The standard error also gets larger as the errors grow, as we should expect it to: larger errors should indicate a worse fit.

Intercept

$$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{s_{\beta_0}} \tag{3.10}$$

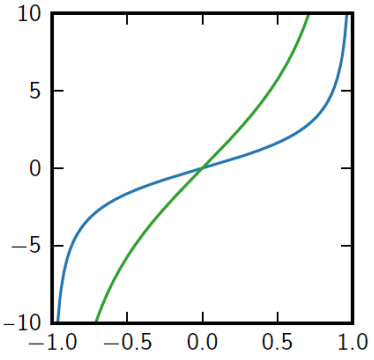


Figure 3.4: The test statistic for the correlation coefficient r for $n = 10$ (blue) and $n = 100$ (green).

which is also t-distributed with $n-2$ degrees of freedom. The standard error is

$$s_{\beta_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}, \tag{3.11}$$

and $\hat{\sigma}$ is given by Equation (3.9).

Correlation

For the correlation coefficient r, our test statistic is the standardized correlation

$$t_r = \hat{\sigma}\sqrt{\frac{n-2}{1-r^2}}, \tag{3.12}$$

which is t-distributed with $n-2$ degrees of freedom. Figure 3.4 plots t_r against r.

Prediction

Let's look at the prediction at a particular value x^*, which we'll call $\hat{y}(x^*)$. In particular:

$$\hat{y}(x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^*,$$

We can do this even if x^* wasn't in our original dataset.

Let's introduce some notation that will help us distinguish between predicting the line versus predicting a particular point generated from the model. From the probabilistic model given by Equation (3.1), we can similarly write how y is generated for the new point x^*:

$$y(x^*) = \underbrace{\beta_0 + \beta_1 x^*}_{\text{defined as } \mu(x^*)} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.13}$$

Then it turns out that the standard error $s_{\hat{\mu}}$ for estimating $\mu(x^*)$ (i.e., the mean of the line at point x^*) using $\hat{y}(x^*)$ is:

$$s_{\hat{\mu}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})}{\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{distance from comfortable prediction region}}}}$$

This makes sense because if we're trying to predict for a point that's far from the mean, then we should be less sure, and our prediction should have more variance. To compute the standard error for estimating a particular point $\hat{y}(x^*)$ and not just its mean $\mu(x^*)$, we'd also need to factor in the extra noise term ϵ in Equation (3.13):

$$s_{\hat{\mu}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \underbrace{+1}_{\text{added}}}$$

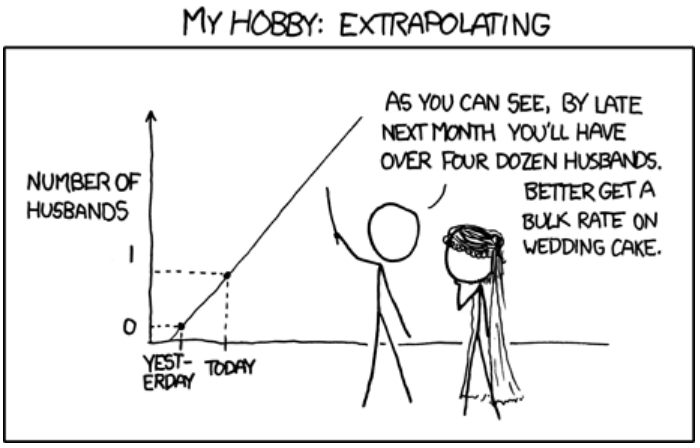
While both of these quantities have the same value when computed from the data, when analyzing them, we have to remember that they're different random variables: $\hat{y}$ has more variation because of the extra ϵ.

3.3 Interpolation vs. extrapolation

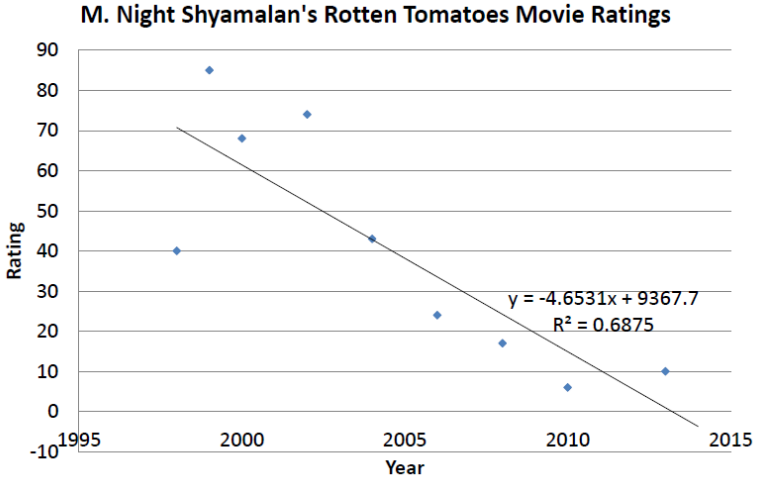
As a reminder, everything here crucially depends on the probabilistic model given by Equation (3.1) being true. In practice, when we do prediction for some value of x we haven't seen before, we need to be very careful. Predicting y for a value of x that is within the interval of points that we saw in the original data (the data that we fit our model with) is called interpolation. Predicting y for a value of x that's outside the range of values we actually saw for x in the original data is called extrapolation.

For real datasets, even if a linear fit seems appropriate, we need to be extremely careful about extrapolation, which can often lead to false predictions!

Example 3. The perils of extrapolation



By fitting a line to the Rotten Tomatoes ratings for movies that M. Night Shyamalan directed over time, one may erroneously be led to believe that in 2014 and onward, Shyamalan's movies will have negative ratings, which isn't even possible!



3.4 Multiple Linear Regression

Now, let's talk about the case when instead of just a single scalar value x, we have a vector $(x_1, \cdots, x_p)$ for every data point i. So, we have n data points (just like before), each with p different predictor variables or features. We'll then try to predict y for each data point as a linear function of the different x variables:

$$y = \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p. \tag{3.14}$$

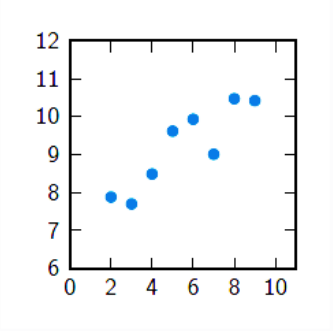
Even though it's still linear, this representation is very versatile; here are just a few of the things we can represent with it:

- Multiple dependent variables: for example, suppose we're trying to predict medical outcome as a function of several variables such as age, genetic susceptibility, and clinical diagnosis. Then we might say that for each patient, x_1 = age; x_2 = genetics, x_3 = diagnosis, and y = outcome.

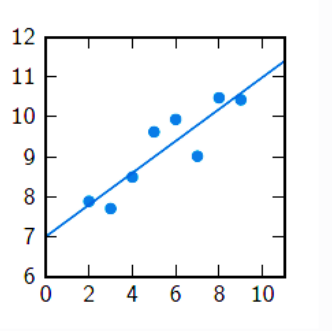
- Non-linearities: Suppose we want to predict a quadratic function $y = ax^2 + bx + c$: then for each data point we might say $x_1 = 1$, $x_2 = x$, and $x_3 = x_2$. This can easily be extended to any nonlinear function we want.

One may ask: why not just use multiple linear regression and fit an extremely high-degree polynomial to our data? While the model then would be much richer, one runs the risk of overfitting, where the model is so rich that it ends up fitting to the noise! We illustrate this with an example; it's also illustrated by a song.

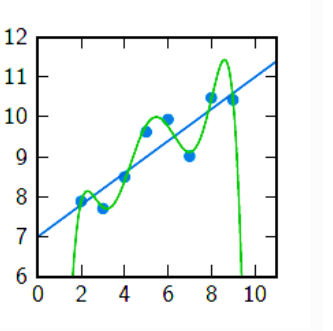
Example 4. Overfitting
Using too many features or too complex of a model can often lead to overfitting. Suppose we want to fit a model to the points in Figure 3.3(a). If we fit a linear model, it might look like Figure 3.3(b). But, the fit isn't perfect. What if we use our newly acquired multiple regression powers to fit a 6th order polynomial to these points? The result is shown in Figure 3.3(c). While our errors are definitely smaller than they were with the linear model, the new model is far too complex, and will likely go wrong for values too far outside the range.



(a) A set of points with a simple linear relationship.



(b) The same set of points with a linear fit (blue).



(c) The same points with a 6th-order polynomial fit (green). As before, the linear fit is shown in blue.

We'll talk a little more about this in later chapters.

We'll represent our input data in matrix form as X, an $x \times p$ matrix where each row corresponds to a data point and each column corresponds to a feature. Since each output y_i is just a single number, we'll represent the collection as an n-element column vector y. Then our linear model can be expressed as

$$y = X\beta + \epsilon. \tag{3.15}$$

where β is a p-element vector of coefficients, and ϵ is an n-element matrix where each element, like ϵ_i earlier, is normal with mean 0 and variance σ^2. Notice that in this version, we haven't explicitly written out a constant term like β_0 from before. We'll often add a column of 1s to the matrix X to accomplish this (try multiplying things out and making sure you understand why this solves the problem). The software you use might do this automatically, so it's something worth checking in the documentation.

This leads to the following optimization problem:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2, \tag{3.16}$$

where $\min_\beta$. just means "find values of β that minimize the following", and X_i refers to row i of the matrix X.

We can use some basic linear algebra to solve this problem and find the optimal estimates:

$$\hat{\beta} = (X^T X)^{-1} X^T y, \tag{3.17}$$

which most computer programs will do for you. Once we have this, what conclusions can we make with the help of statistics? We can obtain confidence intervals and/or hypothesis tests for each coefficient, which most statistical software will do for you. The test statistics are very similar to their counterparts for simple linear regression.

It's important not to blindly test whether all the coefficients are greater than zero: since this involves doing multiple comparisons, we'd need to correct appropriately using Bonferroni correction or FDR correction as

described in the last chapter. But before even doing that, it's often smarter to measure whether the model even explains a significant amount of the variability in the data: if it doesn't, then it isn't even worth testing any of the coefficients individually.

Typically, we'll use an analysis of variance (ANOVA) test to measure this. If the ANOVA test determines that the model explains a significant portion of the variability in the data, then we can consider testing each of the hypotheses and correcting for multiple comparisons.

We can also ask about which features have the most effect: if a feature's coefficient is 0 or close to 0, then that feature has little to no impact on the final result. We need to avoid the effect of scale: for example, if one feature is measured in feet and another in inches, even if they're the same, the coefficient for the feet feature will be twelve times larger. In order to avoid this problem, we'll usually look at the standardized coefficients $\frac{\hat{\beta}_k}{s_{\hat{\beta}_k}}$

3.5 Model Evaluation

How can we measure the performance of our model? Suppose for a moment that every point y_i was very close to the mean $\bar{y}$: this would mean that each y_i wouldn't depend on x_i, and that there wasn't much random error in the value either. Since we expect that this shouldn't be the case, we can try to understand how much the prediction from x_i and random error

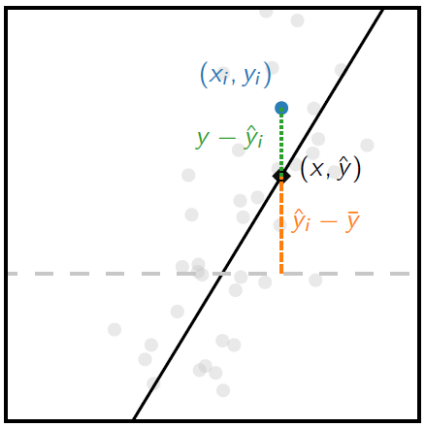


Figure 3.5: An illustration of the components contributing to the difference between the average y-value $\bar{y}$ and a particular point (x_i, y_i) (blue). Some of the difference,$\hat{y}_i - \bar{y}$, can be explained by the model (orange), and the remainder, $y_i - \hat{y}_i$is known as the residual (green).

contribute to y_i. In particular, let's look at how far y_i is from the mean $\bar{y}$. We'll write this difference as:

$$y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{difference explained by model}} + \underbrace{(y_i - \hat{y}_i)}_{\text{difference not explained by model}} \tag{3.18}$$

In particular, the residual is defined to be $y_i - \hat{y}_i$: the distance from the original data point to the predicted value on the line. You can think of it as the error left over after the model has done its work. This difference is shown graphically in Figure 3.5. Note that the residual $y_i - \hat{y}$ isn't quite the same as the noise ϵ! We'll talk a little more about analyzing residuals (and why this distinction matters) in the next chapter.

If our model is doing a good job, then it should explain most of the difference from $\bar{y}$, and the first term should be bigger than the second term. If the second term is much bigger, then the model is probably not as useful. If we square the quantity on the left, work through some algebra, and use some facts about linear regression, we'll find that

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{SS_{total}} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{SS_{explained}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{SS_{residualerror}} \tag{3.19}$$

where "SS" stands for "sum of squares". These terms are often abbreviated as TSS, ESS, and RSS respectively. If we divide through by TSS, we obtain

$$1 = \underbrace{\frac{ESS}{TSS}}_{r^2} + \underbrace{\frac{RSS}{TSS}}_{1-r^2}$$

where we note that r^2 is precisely the coefficient of determination mentioned earlier. Here, we see why r^2 can be interpreted as the fraction of variability in the data that is explained by the model.

One way we might evaluate a model's performance is to compare the ratio ESS/RSS. We'll do this with a slight tweak: we'll instead consider the mean values, ESS$/(p-1)$ and RSS$/(n-p)$, where the denominators correspond to the degrees of freedom. These new variables have χ^2 distributions, and their ratio

$$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} \tag{3.20}$$

has what's known as an F distribution with parameters $p-1$ and $n-p$. The widely used ANOVA test for categorical data, which we'll see in Chapter 6, is based on this F statistic: it's a way of measuring how much of the variability in the data is from the model and how much is from random error, and comparing the two.