Chapter 1

Basic statistics

Statistics are used everywhere.

Weather forecasts estimate the probability that it will rain tomorrow based on a variety of atmospheric measurements. 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. In fact, during the 2012 US presidential election, Nate Silver successfully aggregated such polling data to correctly predict the election outcome of all 50 states![1]

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[2]: that's 10^{15} bytes, or a million gigabytes.
- Biologists are generating 15 petabytes of data a year in genomic information[3].
- 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.

While statistics is an essential tool for justifying a variety of results in research projects, many researchers lack a clear grasp of statistics, misusing its tools and producing all sorts of bad science![4] The goal of these notes is to help you avoid falling into that trap: we'll arm you with the proper tools to produce sound statistical analyses.

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

[1] See Daniel Terdiman, "Obama's win a big vindication for Nate Silver, king of the quants," CNET, November 6, 2012

[2] See CERN's Computing site

[3] See Emily Singer, "Biology's Big Problem: Theres Too Much Data to Handle," October 11, 2013

[4] See The Economist, "Unreliable research: trouble at the lab", October 19, 2013.

We'll 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[a]). 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. 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}!

Another anomaly they found: in normal polling data, there are many weeks In Research 2000's data, this almost never happened: they were probably afraid to make up the same number two weeks in a row since that might not "look random". These problems (and others) were caught thanks to statistical analysis!

[a] Data and a full description at Daily Kos: Research 2000: Problems in plain sight, June 29, 2010..

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".

We'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 (see this relevant xkcd[5]):

- Loosely speaking, the **frequentist** viewpoint holds that the parameters of probabilistic models are fixed, but we just don't know them. These notes will focus 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.

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[6]. 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[7]. 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 $x_1, x_2, \ldots, x_n$ 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.

The **expectation** of a random variable is the average value it takes on:

$$\mathbb{E}[x] = \sum_{\text{poss. values } a} p(a) \cdot a$$

We'll often use the notation μ_x to represent the expectation of random variable x.

Expectation is *linear*: for any random variables x, y and constants c, d,

$$\mathbb{E}[cx + dy] = c\mathbb{E}[x] + d\mathbb{E}[y].$$

[5]Of course, this comic oversimplifies things: here's (Bayesian) statistician Andrew Gelman's response.

[6]Formally, a random variable is a function that maps random outcomes to numbers, but this loose definition will suit our purposes and carries the intuition you'll need.

[7]If the random variable is continuous instead of discrete, $p(a)$ instead represents a *probability density function*, but we'll gloss over the distinction in these notes. For more details, see an introductory probability textbook, such as *Introduction to Probability* by Bertsekas and Tsitsiklis.

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

Intuition for linearity of expectation

Suppose that we collect 5 data points of the form (x, y): $(1,3), (2,4), (5,3), (4,3), (3,4)$. Let's write each of these pairs along with their sum in a table:

x	y	$x+y$
1	3	4
2	4	6
5	3	8
4	3	7
3	4	7

To estimate the mean of variable x, we could just average the values in the first column above (i.e., the observed values for x): $(1+2+5+4+3)/5 = 3$. Similarly, to estimate the mean of variable y, we average the values in the second column above: $(3+4+3+3+4)/5 = 3.4$. Finally, to estimate the mean of variable $x+y$, we could just average the values in the third column: $(4+6+8+7+7)/5 = 6.4$, which turns out to be the same as the sum of the averages of the first two columns.

Notice that to arrive at the average of the values in the third column, we could've reordered values within column 1 and column 2! For example, we scramble column 1 and, separately, column 2, and then we recompute column 3:

x	y	$x+y$
1	3	4
2	3	5
3	3	6
4	4	8
5	4	9

The average of the third column is $(4+5+6+8+9)/5 = 6.4$, which is the same as what we had before! This is true even though x and y are clearly not independent. Notice that we've reordered columns 1 and 2 to make them both increasing in value, effectively making them more correlated (and therefore less independent). But, thanks to linearity of expectation, the average of the sum is still the same as before.

In summary, linearity of expectation says that the ordering of the values within column 1, and separately within column 2 don't actually matter in computing the average of the sum of two variables, which need not be independent.

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

$$\text{var}[x] = \sum_{\text{poss. values } a} p(a) \cdot (a - E[x])^2$$

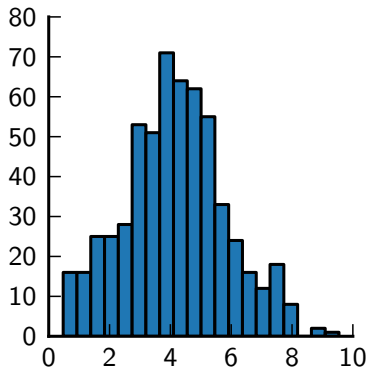
For any constant c, $\text{var}[cx] = c^2 \,\text{var}[x]$. If random variables x and y are independent, then $\text{var}[x+y] = \text{var}[x] + \text{var}[y]$; if they are not independent then this is not necessarily true!

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.

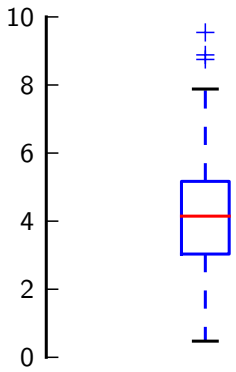
1.2 Exploratory Analysis

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.

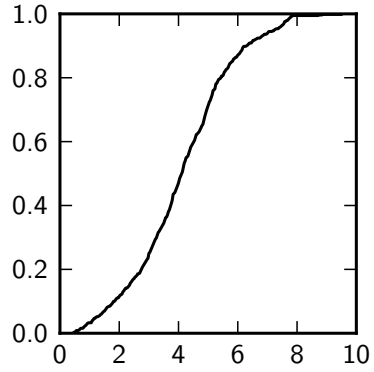
Some common ways of plotting and visualizing data are shown in Figure 1.1. Each of these has its own strengths and weaknesses, and can reveal different patterns or hidden properties of the data.



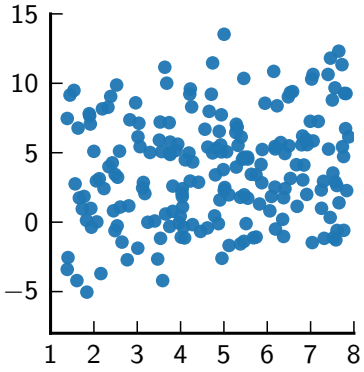
(a) Histogram: this shows the distribution of values a variable takes in a particular set of data. It's particularly useful for seeing the shape of the data distribution in some detail.



(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.

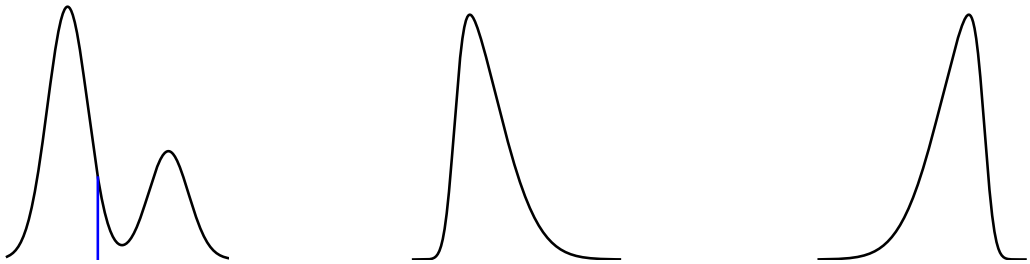


(c) Cumulative Distribution Function (CDF): this shows how much of the data is less than a certain amount. It's useful for comparing the data distribution to some reference distribution.



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

Figure 1.1: 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 1.2: Different irregularities that can come up in data

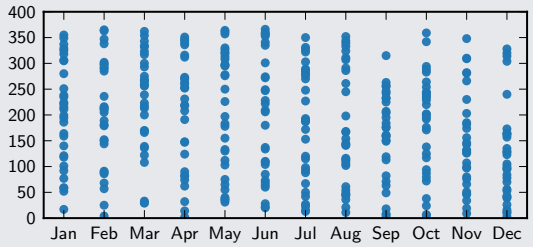
Much of the analysis we'll look at in this class makes assumptions about the data. It's important to check for complex effects; analyzing data with these issues often requires more sophisticated models. For example,

- Are the data *multimodal*? In Figure 1.2a, the mean is a bad representation of the data, since there are two peaks, or 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.

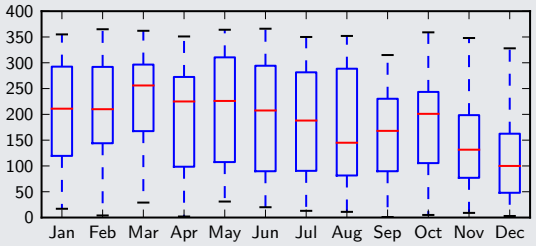
Before we start applying any kind of analysis (which will make certain assumptions about the data), it's important to visualize and check that those properties are satisfied. This is worth repeating: *it's always a good idea to visualize before testing!*

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[a]. Do you see 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:



It's now clearer that our hunch was correct: 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.

[a] Data from the Selective Service: http://www.sss.gov/LOTTER8.HTM

■ 1.2.1 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 $x_1, \ldots, x_n$ 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.

■ 1.2.2 Quantitative measures and summary statistics

Here are some useful ways of numerically summarizing sample data:

- **Sample Mean**: $\bar{x} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- **Sample Variance**: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **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%.
- **Interquartile 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 $\hat{\mu}$ as an estimate for μ. $\hat{\mu}$ is probably not going to be exactly the same as μ, because the data points are random. So, even though μ is fixed, $\hat{\mu}$ is a random variable (because it depends on the random data). On average, what do we expect the random variable $\hat{\mu}_x$ to be? We can formalize this question by asking "What's the expectation of $\hat{\mu}_x$, or $\mathbb{E}[\hat{\mu}_x]$?"

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

This result makes sense: while $\hat{\mu}_x$ 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 $\hat{\sigma}^2$ requires a similar (but slightly more complicated) process; we obtain $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$. 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.

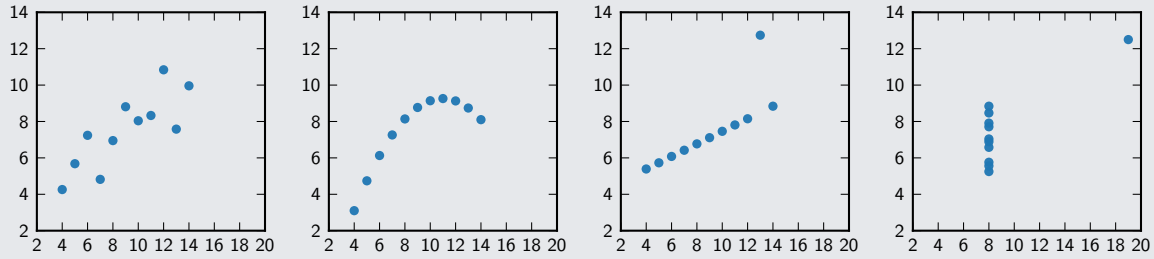
It's often tempting to compute quantitative measures like mean and variance and move on to analyzing them, but these summary statistics have important limitations, as the next example illustrates.

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 $\hat{\sigma}_x^2$ for the variance are 9 and 11 respectively
- For random variable y, the estimate $\bar{y}$ for the mean and the estimate $\hat{\sigma}_y^2$ 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.

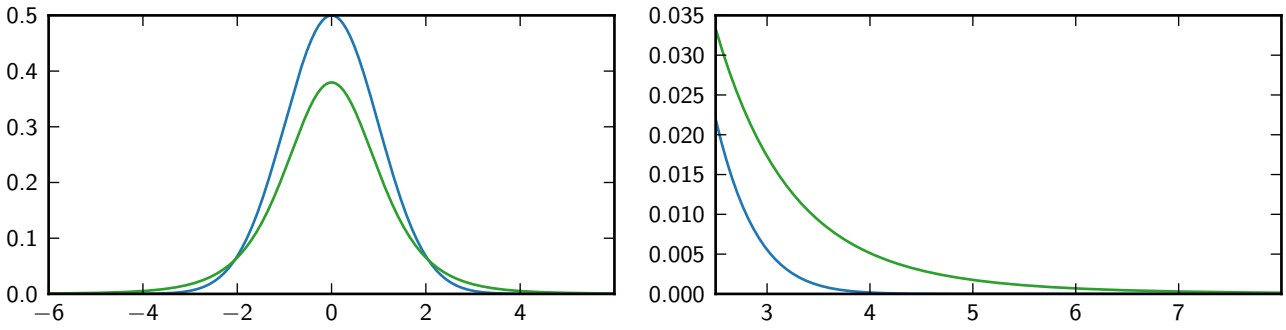


Figure 1.3: The standard normal distribution (blue) and Student t distribution with 5 degrees of freedom (green). The second plot zooms in on the x-axis from 2.5 to 8. Notice that the t distribution has *heavier tails*: that is, the probability of obtaining a value far away from the mean is higher for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

■ 1.3 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 $x \sim \mathcal{N}(\mu, \sigma^2)$ 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 $\mathcal{N}(0, 1)$ (i.e., mean 0 and variance 1).

   Here are some useful facts about Gaussian random variables:

   - If $x \sim \mathcal{N}(\mu, \sigma^2)$, and we define $y = (x - \mu)/\sigma$, then $y \sim \mathcal{N}(0, 1)$. y is usually referred to as a *standardized* version of x. We'll take advantage of this fact in the next lecture.
   - 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 $x_1, \ldots, x_n$, 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 $np(1-p)$:

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

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 $x_1, \ldots, x_r$ that are i.i.d., and $x_i \sim \mathcal{N}(0, 1)$. 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 $z \sim \mathcal{N}(0, 1)$ 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:

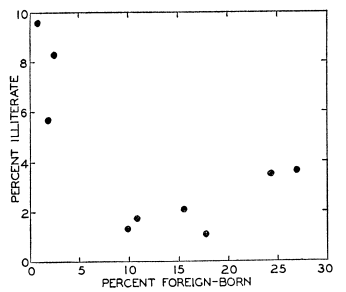


Fig. 3

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[a].

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[b]. 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.

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.

We'll see this issue come up again when we discuss Simpson's Paradox.

[a] see Robinson. Ecological Correlations and Behavior of Individuals. American Sociological Review, 1950.

[b] see statistician Andrew Gelman's book, **Red State, Blue State, Rich State, Poor State**.