

Lecture Notes part B

PSY Vo500, Statistical Methods in Psychology

Kevin R Foster, the Colin Powell School at the City College of New York, CUNY

Spring 2024

Table of Contents

PSY Vo500, Statistical Methods in Psychology1

Probability.....3

 Think Like a Statistician3

 Randomness in Games.....3

 Some math.....3

 Independent Events.....5

 Terms and Definitions.....11

 Counting Rules.....13

Discrete and Continuous Random Variables13

 Common Distributions:14

 Uniform14

 Bernoulli15

 Binomial15

 Poisson16

 Continuous Random Variables17

 The PDF and CDF.....17

 Normal Distribution17

 Motivation: Sample Averages are Normally Distributed20

 Hints on using Excel or R to calculate the Standard Normal cdf.....26

 Excel26

 Google26

 R26

Is That Big?.....28

 Get a central parameter29

 Variation around central mean32

 How can we try to guard against seeing relationships where, in fact, none actually exist?32

 Law of Large Numbers.....32

 Standard Error of Average.....33

 ⚡ A bit of Math:.....34

Hypothesis Testing.....35

Confidence Intervals	37
Find p-values.....	38
Type I and Type II Errors.....	38
Examples	39
P-values	42
Confidence Intervals for Polls.....	42
Complications from a Series of Hypothesis Tests.....	43
Issues with Canned Tests.....	44
Bayesian Stats.....	44
Details of Distributions T-distributions, chi-squared, etc.....	45
T-tests	45
T-tests with two samples	47
Other Distributions	47

Probability

Beyond presenting some basic measures such as averages and standard deviations, we want to try to understand how much these measures can tell us about the larger world. How likely is it, that we're being fooled, into thinking that there's a relationship when actually none exists? To think through these questions we must consider the logical implications of randomness and often use some basic statistical distributions (discrete or continuous).

Think Like a Statistician

The basic question that a Statistician must ask is "How likely is it, that I'm being fooled?" Once we accept that the world is random (rather than a manifestation of some god's will), we must decide how to make our decisions, knowing that we cannot guarantee that we will always be right. There is some risk that the world will seem to be one way, when actually it is not. The stars are strewn randomly across the sky but some bright ones seem to line up into patterns. So too any data might sometimes line up into patterns.

Statisticians tend to stand on their heads and ask, suppose there were actually no relationship? (Sometimes they ask, "suppose the conventional wisdom were true?") This statement, about "no relationship," is called the **Null Hypothesis**, sometimes abbreviated as H_0 . The Null Hypothesis is tested against an **Alternative Hypothesis**, H_A .

Before we even begin looking at the data we can set down some rules for this test. We know that there is some probability that nature will fool me, that it will seem as though there is a relationship when actually there is none. The statistical test will create a model of a world where there is actually no relationship and then ask how likely it is that we could see what we actually see, "How likely is it, that I'm being fooled?" What if there were actually no relationship, is there some chance that I could see what I actually see?

Randomness in Games

As an example, consider games or sports events. As any sports fan knows, a team or individual can get lucky or unlucky. The baseball World Series, for example, has seven games. It is designed to ensure that, by the end, one team or the other wins. But will the better team always win?

First make a note about subjectivity: if I am a fan of the team that won, then I will be convinced that the better team won; if I'm a fan of the losing team then I'll be certain that the better team got unlucky. But fans of each team might agree, if they discussed the question before the Series were played, that luck has a role.

Will the better team win? Clearly a seven-game Series means that one team or the other will win, even if they are exactly matched (if each had precisely a 50% chance of winning). If two representatives tossed a coin in the air seven times, then one or the other would win at least four tosses – maybe even more. We can use a computer to simulate seven coin-tosses by having it pick a random number between zero and one and defining a "win" as when the random number is greater than 0.5.

Or instead of having a computer do it, we could use a bit of statistical theory.

Some math

Suppose we start with just one coin-toss or game (baseball and basketball use 7 games to decide a champion; global football and American football use just one). Choose to focus on one team so that we can talk about "win" and "loss". If this team has a probability of winning that is equal to p , then it has a probability

of losing equal to $(1-p)$. So even if p , the probability of winning, is equal to 0.6, there is still a 40% chance that it could lose a single game. In fact unless the probability of winning is 100%, there is some chance, however remote, that the lesser team will win.

What about if they played two games? What are the outcomes? The probability of a team winning both games is $p \cdot p = p^2$. If the probability were 0.5 then the probability of winning twice in a row would be 0.25.

A table can show this:

	Win Game 1 { p }	Lose Game 1 { $1-p$ }
Win Game 2 { p }	outcome: W,W	L,W
Lose Game 2 { $1-p$ }	W,L	L,L

This is a fundamental fact about how probabilities are represented mathematically: if the probabilities are not related (i.e. if the tossed coin has no memory) then the probability of both events happening is found by multiplying the probabilities of each individual outcome. (What if they're not unrelated, you may ask? What if the first team that wins gets a psychological boost in the next so they're more likely to win the second game? Then the math gets more complicated – we'll come back to that question!)

The math notation for two events, call them A and B, both happening is:

$$\Pr\{A \text{ and } B\} = \Pr\{A \cap B\}$$

The fundamental fact of independence is then represented as:

$$\Pr\{A \cap B\} = \Pr\{A\} \Pr\{B\} \quad \text{if } A \text{ and } B \text{ are independent}$$

where we use the term "independent" for when there is no relationship between them.

The probability that a team could lose both games is $(1-p) \cdot (1-p) = (1-p)^2$. The probability that the teams could split the series (each wins just one) is $p \cdot (1-p) + (1-p) \cdot p = 2p(1-p)$. There are two ways that each team could win just one game: either the series splits (Win, Loss) or (Loss, Win).

For three games the outcomes become more complicated: now there are 8 combinations of win and loss:

(W,W,W)	(W,W,L)	(W,L,W)	(L,W,W)	(W,L,L)	(L,W,L)	(L,L,W)	(L,L,L)
$p \cdot p \cdot p$	$p \cdot p \cdot (1-p)$	$p \cdot (1-p) \cdot p$	$(1-p) \cdot p \cdot p$	$p \cdot (1-p) \cdot (1-p)$	$(1-p) \cdot p \cdot (1-p)$	$(1-p) \cdot (1-p) \cdot p$	$(1-p) \cdot (1-p) \cdot (1-p)$

and the probabilities are in the row below.

The team will win the series in any of the left-most 4 outcomes so its overall probability of winning the series is

$$p^3 + 3p^2(1-p)$$

while its probability of losing the series is

$$3p(1-p)^2 + (1-p)^3.$$

Clearly if p is 0.5 so that $p=(1-p)$ then the chances of either team winning the three-game series are equal. If the probabilities are not equal then the chances are different, but as long as there is a probability not equal to one or zero (i.e. no certainty) then there is a chance that the worse team could win.

If you keep on working out the probabilities for longer and longer series you might notice that the coefficients and functional forms are right out of Pascal's Triangle. This is your first notice of just how "normal" the Normal Distribution is, in the sense that it jumps into all sorts of places where you might not expect it. The terms of Pascal's Triangle begin (as N becomes large) to form a normal distribution! We'll come back to this again...

Independent Events

A is independent of B if and only if $P\{A|B\} = P\{A\}$

If we have multiple random variables then we can consider their **Joint Distribution**: the probability associated with each outcome in both sample spaces. So a coin flip has a simple discrete distribution: a 50% chance of heads and a 50% chance of tails. Flipping 2 coins gives a joint distribution: a 25% chance of both coming up heads, a 25% chance of both coming up tails, and a 50% chance of getting one head and one tail.

The probability of multiple independent events is found by multiplying the probabilities of each event together. So the chance of rolling two 6 on two dice is $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. The probability of getting to the computer lab on the 6th floor of NAC from the first floor, without having to walk up a broken escalator, can be found this way too. Suppose the probability of an escalator not working is p ; then the probability of it working is $(1-p)$ and the probability of five escalators each working is $(1-p)^5$. So even if the probability of a breakdown is small (5%), still the probability of having every escalator work is just $(1-5\%)^5 = (95\%)^5 = (0.95)^5 = \left(\frac{95}{100}\right)^5 = 0.7738 = 77.38\%$ so this implies that you'd expect to walk more than once a week.

A simple representation of the joint distribution of two coin flips is a table:

	coin 1 Heads	coin 1 Tails
coin 2 Heads	H,H at 25%	H,T at 25%
coin 2 Tails	T,H at 25%	T,T at 25%

Where, since the outcomes are independent, we can just multiply the probabilities.

The Joint Distribution tells the probabilities of all of the different outcomes. A **Marginal Distribution** answers a slightly different question: given some value of one of the variables, what are the probabilities of the other variables?

When the variables are independent then the marginal distribution does not change from the joint distribution. Consider a simple example of X and Y discrete variables. X takes on values of 1 or 2 with

probabilities of 0.6 and 0.4 respectively. Y takes on values of 1, 2, or 3 with probabilities of 0.5, 0.3, and 0.2 respectively. So we can give a table like this:

	X=1 (60%)	X=2 (40%)
Y=1 (50%)	(1,1) at probability 0.3	(2,1) at probability 0.2
Y=2 (30%)	(1,2) at probability 0.18	(2,2) at probability 0.12
Y=3 (20%)	(1,3) at probability 0.12	(2,3) at probability 0.08

On the assumption that X and Y are independent. The probabilities in each box are found by multiplying the probability of each independent event.

If instead we had the two variables, A and B, not being independent then we might have a table more like this:

	A=1	A=2
B=1	(1,1) at probability 0.25	(2,1) at probability 0.13
B=2	(1,2) at probability 0.23	(2,2) at probability 0.12
B=3	(1,3) at probability 0.17	(2,3) at probability 0.1

We will examine the differences.

If we add up the probabilities along either rows or columns then we get the **marginal probabilities** (which we write in the *margins*, appropriately enough). Then we'd get:

	X=1 (60%)	X=2 (40%)	
Y=1 (50%)	(1,1) at probability 0.3	(2,1) at probability 0.2	0.5
Y=2 (30%)	(1,2) at probability 0.18	(2,2) at probability 0.12	0.3
Y=3 (20%)	(1,3) at probability 0.12	(2,3) at probability 0.08	0.2
	0.6	0.4	

Which just re-states our assumption that the variables are independent – and shows that, where there is independence, the probability of either variable alone does not depend on the value that the other variable takes on. In other words, knowing X does not give me any information about the value that Y will take on, and vice versa.

If instead we do this for the A,B case we get:

	A=1	A=2	
B=1	(1,1) at probability 0.25	(2,1) at probability 0.13	0.38
B=2	(1,2) at probability 0.23	(2,2) at probability 0.12	0.35
B=3	(1,3) at probability 0.17	(2,3) at probability 0.1	0.27
	0.65	0.35	

Where we double check that we've done it right by seeing that the sum of either of the marginals is equal to one ($65\% + 35\% = 100\%$ and $38\% + 35\% + 27\% = 100\%$).

So the marginal distributions sum the various ways that an outcome can happen. For example, we can get A=1 in any of 3 ways: either (1,1), (1,2) or (1,3). So we add the probabilities of each of these outcomes to find the total chance of getting A=1.

But if we want to understand how A and B are related, it might be more useful to consider this as a prediction problem: would knowing the value that A takes on help me guess the value of B? Would knowing the value that B takes on help me guess the value of A?

These are abstract questions but they have vitally important real-life analogs. In airport security, is the probability that someone is a terrorist independent of whether they are Muslim? Is the probability that someone is pulled out of line for a thorough search independent of whether they are Muslim? (*The TSA might have different beliefs than you or me!*) In medicine, is the probability that someone gets cancer independent of whether they eat lots of vegetables? In poker, if my opponent just raised the bid, what is the probability that her cards are better than mine?

For these questions we want to find the conditional distribution: what is the probability of some outcome, given a particular value for some other random variable?

Just from the phrasing of the question, you should be able to see that if the two variables are independent then the conditional distribution should not change from the marginal distribution – as is the case of X and Y. Flipping a coin does not help me guess the outcome of a roll of the dice. (Cheering in front of a sports game on TV does not affect the outcome, for another example – although plenty of people act as though they don't believe that!)

How do we find the conditional distribution? Take the value of the joint distribution and divide it by the marginal distribution of the relevant variable.

For example, suppose we want to find the probability of B outcomes, conditional on A=1. Since we know that A=1, there is no longer a 65% probability of A – assume that it happened. So we divide each joint probability by 0.65 so that the sum will be equal to 1. So the probabilities are now:

	A=1	A=2	
B=1	(1,1) at probability 0.25/.65	(2,1) at probability 0.13	0.38
B=2	(1,2) at probability 0.23/.65	(2,2) at probability 0.12	0.35
B=3	(1,3) at probability 0.17/.65	(2,3) at probability 0.1	0.27
	0.65/.65	0.35	

so now we get the conditional distribution:

	A=1	A=2	
B=1	(1,1) @ 0.3846	(2,1) at probability 0.13	0.38
B=2	(1,2) @ 0.3538	(2,2) at probability 0.12	0.35
B=3	(1,3) @ 0.2615	(2,3) at probability 0.1	0.27
		0.35	

We could do the same to find the conditional distribution of B, given that A=2:

	A=1	A=2	
B=1	(1,1) at probability 0.25	(2,1) @ 0.13/.35 =.3714	0.38
B=2	(1,2) at probability 0.23	(2,2) @ 0.12/.35 =.3429	0.35
B=3	(1,3) at probability 0.17	(2,3) @ 0.1/.35 = .2857	0.27
	0.65		

These conditional probabilities are denoted as $\Pr\{B|A=2\}$ for example. We could find the expected value of B given that A equals 2, $E[B|A=2]$, just by multiplying the value of B by its probability of occurrence, so $E[B|A=2] = (1 \cdot .3714) + (2 \cdot .3429) + (3 \cdot .2857)$.

We could find the conditional probabilities of A given B=1 or given B=2 or given B=3. In those cases we would sum across the rows rather than down the columns.

More pertinently, we can get crosstabs on two variables the BRFSS data. We've got information on each person's health insurance, which are the columns, and then whether they've ever been diagnosed with depressive disorder as rows. We can ask, "are these statistically independent?" Which is equivalent to asking, are these uncorrelated? Now you're probably answering something like, "Duh!" but this is an example of how we "think like a statistician."

The table is:

	Ever had depressive disorder	
	yes	No
Private insurance	36912	160504
Medicare	25555	109589
Medicaid	10626	18198
Other govt ins	10861	29798
No insurance	4432	18374

But these are raw numbers of people not fractions – so divide by the total number of observations (easy in Excel or can be done in R, depending on your preference); I also show the marginals:

	yes	No	<i>marginals</i>
Private insurance	0.087	0.378	0.465
Medicare	0.060	0.258	0.318
Medicaid	0.025	0.043	0.068
Other govt ins	0.026	0.070	0.096
No insurance	0.010	0.043	0.054
<i>marginals</i>	0.208	0.792	

Some R code to do those tables:

```
table(brfss22$ADDEPEV3,brfss22$PRIMINSR)
brfss22$recode_insurance <- recode_factor(brfss22$PRIMINSR,
  "health ins thr employer or union" = "private",
  "private plan" = "private",
  "Medicare" = "Medicare",
  "Medigap" = "other govt",
  "Medicaid" = "Medicaid",
  "CHIP" = "other govt",
  "CHAMPUS" = "other govt",
  "Indian Health Svc" = "other govt",
  "State sponsored plan" = "other govt",
  "other govt prog" = "other govt",
  "dont know not sure" = "NA",
  "no coverage of any type" = "no ins",
  "refused" = "NA")

table(brfss22$ADDEPEV3,brfss22$recode_insurance) # vs
table(brfss22$recode_insurance,brfss22$ADDEPEV3)
```

These numbers are rough to interpret; the conditionals might be easier. So can ask, what is the likelihood of different insurance types, conditional on mental health diagnosis?

	yes	No
Private insurance	0.418	0.477
Medicare	0.289	0.326

Medicaid	0.120	0.054
Other govt ins	0.123	0.089
No insurance	0.050	0.055

Note each column sums to 1.

Conditional on a person having had a depressive diagnosis, 41.8% had private insurance, 28.9% had Medicare, 12% had Medicaid, 12.3% had other government insurance and 5% were uninsured. Compare with those who did not have that diagnosis, where 47.7% had private insurance, 32.6% had Medicare, lower levels of other government programs and 5.5% uninsured.

The other conditional is asking, conditional on a certain insurance status, what are the proportion who did or did not have that diagnosis? That table is found by dividing each row in the original table by its marginal:

	yes	No
Private insurance	0.187	0.813
Medicare	0.189	0.811
Medicaid	0.369	0.631
Other govt ins	0.267	0.733
No insurance	0.194	0.806

Now each row sums to one.

So this shows that, among those with private insurance, 81.3% had never been diagnosed while 18.7 had. That's not far different from those without insurance. But quite different from those with Medicaid, where almost 37% had. Although it's often tempting, let me caution against stating that any of these show causation – that's not true here and not in general. It is likely the case here that the causality goes in both directions. Someone with a history of mental health issues might be more likely to be on Medicaid, but someone without insurance might be less likely to be diagnosed.

Both of these conditioning sets help understand how the variables are interrelated – there is not necessarily one better than the other.

Conditional probabilities can also be calculated with what is called **Bayes' Theorem**:

$$P\{B|A\} = \frac{P\{A|B\} \cdot P\{B\}}{P\{A\}}.$$

This can be understood by recalling the definition of conditional probability, $P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}$, so

$P\{B|A\} = \frac{P\{A \cap B\}}{P\{A\}}$, that the conditional probability equals the joint probability divided by the marginal probability.

The power of Bayes' Theorem can be understood by thinking about medical testing. Suppose a genetic test screens for some disease with 99% accuracy. Your test comes back positive – how worried should you be? The surprising answer is not 99% worried; in fact often you might be more than likely to be healthy! Suppose that the disease is rare so only 1 person in 1000 has it (so 0.1%). So out of 1000 people, one person has the disease and the test is 99% likely to identify that person. Out of the remaining 999 people, 1% will be

misidentified as having the disease, so this is 9.99 – call it 10 people. So eleven people will test positive but only one will actually have the disease so the probability of having the disease given that the test comes up positive, $P\{sick|test+\}$, is $\frac{P\{test+|sick\}P\{sick\}}{P\{test+\}} = \frac{0.99 \cdot 0.001}{0.01} = .099$.

The test is not at all useless – it has brought down an individual's likelihood of being sick by orders of magnitude, from one-tenth of one percent to ten percent. But it's still not nearly as accurate as the "99%" label might imply.

Many healthcare providers don't quite get this and explain it merely as "don't be too worried until we do further tests." But this is one reason why broad-based tests can be very expensive and not very helpful. These tests are much more useful if we first narrow down the population of people who might have the disease. For example home pregnancy tests might be 99% accurate but if you randomly selected 1000 people to take the test, you'd find many false positives. Some of those might be guys (!) or women who, for a variety of reasons, are not likely to be pregnant. The test is only useful as one element of a screen that gets progressively finer and finer. (Occasionally politicians think it might be a good idea to have, for example, every welfare recipient tested for drugs, without discussion of how many false positives and false negatives would be produced.)

Terms and Definitions

Some basics: a sample space is the entire list of possible outcomes (can be whole long list or even mathematical sets such as real numbers); events are subsets of the sample space. Simple event is a single outcome (one dice comes up 6); a compound event is several outcomes (both dice come up 6). Notate an event as A. The complement of the event is the set of all events that are not in A; this is A'.

The events must be **mutually exclusive and exhaustive**, so a good deal of the hard work in probability is just figuring out how to list all of the events.

Mutually exclusive means that the events must be clearly defined so that the data observed can be classified into just one event. Exhaustive means that every possible data observed must fit into some event. The "mutually exclusive" part means that probabilities can be added up, so that if the probability of rolling a "1" on a dice is 1/6 and the probability of rolling a 6 is 1/6, then the probability of rolling either a 1 or 6 is 2/6 = 1/3. The "exhaustive" part of defining the events means that the sum of all the events must equal one.

For example, suppose we roll two dice. We might want to think of "die #1 comes up as 6" as one event [in English, the singular of "dice" is "die" – how morbid gambling can be!]. But the other die can have 6 different values without changing the value of the first die. So a better list of events would be the integers from 2 to 12, the sum of the dice values – with the note that there are many ways of achieving some of the events (a 7 is a 6 & 1 or a 5 & 2, or 4 & 3, or 3 & 4, or 2 & 5, or 1 & 6) while other events have only one path (each die comes up 6 to make 12).

A **sample space** is the set of all possible events. The sum of the probability of all of the events in the sample space is equal to one. There is a 100% chance that something happens (provided we've defined the sample space correctly). So if a lottery brags that there is a 2% chance that "you might be a winner!" this is equivalent to stating that there is a 98% chance that you'll lose.

Events have **probability**; this must lie between zero and one (inclusive); so $0 \leq P \leq 1$. The probability of all of the events in the sample space must sum to one. This means that the probability of an event and its complement must sum to one: $P\{A\} + P\{A'\} = 1$.

Probabilities come from empirical results (relative frequency approach) or the classical (a priori or postulated) assignment or from subjective beliefs that people have.

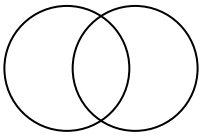
In empirical approach, the **Law of Large Numbers** is important: as the number of identical trials increases, the estimated frequency approaches its theoretical value. You can try flipping coins and seeing how many come up heads (*flip a bunch at a time to speed up the process*); it should be 50%.

We are often interested in finding the probability of two events both happening; this is the "**intersection**" of two events; the logical "and" relationship; two things both occurring. In the PUMS data we might want to find how many females have a college degree; in poker we might care about the chance of an opponent having an ace as one of her hole cards and the dealer turning up a king. We notate the intersection of A and B as $A \cap B$ and want to find $P\{A \cap B\}$. In SPSS this is notated with "&".

The "**union**" of two events is the logical "or" so it is either of two events occurring; this is $A \cup B$ so we might consider $P\{A \cup B\}$ or, in SPSS, "|". In the PUMS data we might want to combine people who report themselves as having race "black" with those who report "black – white". In cards, it is the probability that any of 3 opponents has a better hand.

Married people can buy life insurance policies that pay out either when the first person dies or after both die – logical *and* vs *or*.

Venn Diagrams (Ballantine)



General Law of Addition

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$$

$$\text{and so } P\{A \cap B\} = P\{A\} + P\{B\} - P\{A \cup B\}$$

Mutually Exclusive (Special Law of Addition),

$$\text{If } A \cap B = \emptyset \text{ then } P\{A \cap B\} = 0 \text{ and } P\{A \cup B\} = P\{A\} + P\{B\}$$

Conditional Probability

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}} \text{ if } P\{B\} \neq 0. \text{ See Venn Diagram.}$$

Counting Rules

If A can occur as N_1 events and B can be N_2 events then the sample space is $N_1 \cdot N_2$ (visualize a contingency table with N_1 rows and N_2 columns).

Factorials: If there are N items then they can be arranged in $N! = (n)(n-1)(n-2)\dots(1) = \prod_{i=0}^{N-1} (N-i)$ ways.

Permutations: n events that can occur in r items (where order is important) have a total of $nPr = \frac{n!}{(n-r)!}$ possible outcomes.

Combinations: n events that can occur in r items (where order is not important) have $nCr = \frac{n!}{r!(n-r)!}$ possible outcomes – just the permutation divided by $r!$ to take care of the multiple ways of ordering.

So to apply these, consider computer passwords (see NYTimes article below).

The article reports:

Mr. Herley, working with Dinei Florêncio, also at Microsoft Research, looked at the password policies of 75 Web sites. ... They reported that the sites that allowed relatively weak passwords were busy commercial destinations, including PayPal, Amazon.com and Fidelity Investments. The sites that insisted on very complex passwords were mostly government and university sites. What accounts for the difference? They suggest that "when the voices that advocate for usability are absent or weak, security measures become needlessly restrictive."

Consider the simple mathematics of why a government or university might want complex passwords. How many permutations are possible if passwords are 6 numerical digits? How many if passwords are 6 alphabetic or numeric characters? If the characters are alphabetic, numeric, and fifteen punctuation characters (, . _ - ? ! @ # \$ % ^ & * ' ")? What if passwords are 8 characters? If each login attempt takes 1/100 of a second, how many seconds of "brute-force attack" does it take to access the account on average? If there is a penalty of 10 minutes after 3 unsuccessful login attempts, how long would it take to break in? (Of course, the article notes, if password requirements are so arcane that employees put their passwords on a Post-It attached to the monitor, then the calculations above are irrelevant.)

Discrete and Continuous Random Variables

For any discrete random variable, the mean or expected value is:

$$E(X) = \mu = \sum_{i=1}^N x_i P(x_i)$$

and the variance is

$$\sigma^2 = \sum_{i=1}^N (x_i - \mu)^2 P(x_i) \text{ so the standard deviation is the square root.}$$

These can be described by PDF or CDF – probability density function or cumulative distribution function. The PDF shows the probability of events; the CDF shows the cumulative probability of an event that is smaller than or equal to that event. The PDF is the derivative of the CDF.

Linear Transformations:

- If $Y = aX + b$ then Y will have mean $\mu_Y = a\mu_X + b$ and standard deviation $\sigma_Y = a\sigma_X$.
- If $Z = X + Y$ then $\mu_Z = \mu_X + \mu_Y$; $\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}}$ (and if X and Y are independent then the covariance term drops out)

WARNING: These statements DO NOT work for non-linear calculations! The propositions above do NOT tell about when X and Y are multiplied and divided: the distributions of $X \cdot Y$ or X/Y are not easily found. Nor is $\ln X$, nor e^X . We might wish for a magic wand to make these work out simply but they **don't** in general.

Common Distributions:

Uniform

- depend on only upper and lower bound, so all events are in $[a, b]$
- mean is $\frac{a+b}{2}$; standard deviation is $\sqrt{\frac{[b-a+1]^2 - 1}{12}}$
- Many null hypotheses are naturally formulated as stating that some distribution is uniform: e.g. stock picks, names and grades, birth month and sports success, etc.

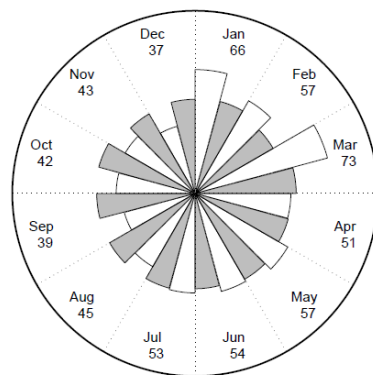


Figure 1: Circular plot of the observed and expected number of AFL players' births. The observed values are shown in white segments and the expected value in grey. The numbers around the outside of the plot are the observed number of births in each month. The expected number of births are based on national data.

from: Barnett, Adrian G. (2010) The relative age effect in Australian Football League players. Working Paper.

Although note that distribution of births is not quite uniform; certainly among animal species humans are unusual in that births are not overwhelmingly seasonal.

Benford's Law: not really a law but an empirical result about measurements, that looking at the first digit, the value 1 is much more common than 9 – the first digit is not uniformly distributed. Originally stated for tables of logarithms. Second digit is closer to uniform; third digit closer still, etc. See online R program. This is a warning that sometimes our intuition about how we might think numbers are distributed is actually wrong.

Bernoulli

- depend only on p , the probability of the event occurring

$$E(X) = \mu = \sum_{i=1}^N x_i P(x_i) \quad \text{and}$$

$$\sigma^2 = \sum_{i=1}^N (x_i - \mu)^2 P(x_i)$$

. Where there are only 2 values, 0 and 1, this is easy to calculate. $E(X)$ here is $1 \cdot P(x=1) + 0 \cdot P(x=0) = 1 \cdot p + 0 \cdot (1-p) = p$. Variance is $(1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = \{\text{some algebra to write out}\} = p - p^2$.

- mean is p ; standard deviation is $\sqrt{p(1-p)}$
 - *Where is the maximum standard deviation? Intuition: what probability will give the most variation in yes/no answers? Or use calculus; note that has same maximum as $p(1-p)$ so take derivative of that, set to zero. Then hit your forehead with the palm of your hand, realizing that calculus gave you the same answer as simple intuition.*
- Used for coin flips, dice rolls, events with "yes/no" answers: Was person re-employed after layoff? Did patient improve after taking the drug? Did company pay out to investors from IPO?

Binomial

- have n Bernoulli trials, each independent; record how many were 1 not zero
- $\mu = np$; $\sigma = \sqrt{np(1-p)}$
 - These formulas are easy to derive from rules of linear combinations. If B_i are independent random variables with Bernoulli distributions, then what is the mean of $B_1 + B_2$? What is its std dev?
 - What if this is expressed as a fraction of trials? Derive.
- what fraction of coin flips came up heads? What fraction of people were re-employed after layoff? What fraction of patients improved? What fraction of companies offered IPOs?
- questions about opinion polls – the famous "plus or minus 2 percentage points" – get margin of error depending on sample size (n)

Some students are a bit puzzled by two different sets of formulas for the binomial distribution – the standard deviation is listed as either $\sqrt{np(1-p)}$ or $\sqrt{\frac{p(1-p)}{n}}$. Which is it?!

It depends on the units. If we measure the **number** of successes in n trials, then we multiply by n . If we measure the **fraction** of successes in n trials, then we don't multiply but divide.

Consider a simple example: the probability of a hit is 50% so $\sqrt{p(1-p)} = \sqrt{\frac{1}{2} \cdot \frac{1}{2}} = \sqrt{\frac{1}{4}} = \frac{1}{2}$. If we have 10 trials and ask, how many are likely to hit, then this should be a different number than if we had 500 trials. The standard error of the raw number of how many, of 10, hits we would expect to see, is $\sqrt{10} \cdot \frac{1}{2}$ which is 1.58, so with a 95% probability we would expect to see 5 hits, plus or minus $1.96 \cdot 1.58 = 3.1$ so a range between 2 and 8. If we had 500 trials then the raw number we'd expect to see is 250 with a standard error or $\sqrt{500} \cdot \frac{1}{2} = 11.18$ so the 95% confidence interval is 250 plus or minus

22 so the range between 228 and 272. This is a bigger range (in absolute value) but a smaller part of the fraction of hits.

With 10 draws, we just figured out that the range of hits is (in fractions) from 0.2 to 0.8. With 500 draws, the range is from 0.456 to 0.544 – much narrower. We can get these latter answers if we take the earlier result of standard deviations and divide by n . The difference in the formula is just this result, since $\frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}$. You could think of this as being analogous to the other "standard error of the average" formulas we learned, where you take the standard deviation of the original sample and divide by the square root of n .

Alternately instead of memorizing formulas for different distributions, you can derive this one easily from our rules of linear combinations. (Try it!)

Poisson

- model arrivals per time, assuming independent
- depends only on λ which is also mean
- PDF is $\frac{\lambda^x e^{-\lambda}}{x!}$
- model how long each line at grocery store is, how cars enter traffic, how many insurance claims

Continuous Random Variables

The PDF and CDF

Where discrete random variables would sum up probabilities for the individual outcomes, continuous random variables necessitate some more complicated math. When X is a continuous random variable, the probability of it being equal to any particular value is zero. If X is continuous, there is a zero chance that it will be, say, 5 – it could be 4.99998 or 5.000001 and so on. But we can still take the area under the PDF by taking the limit of the sum, as the horizontal increments get smaller and smaller – the Riemann method, that you remember from Calculus. So to find the probability of X being equal to a set of values we integrate the PDF between those values, so

$$P\{a \leq X \leq b\} = \int_a^b p(x) dx.$$

The CDF, the probability of observing a value less than some parameter, is therefore the integral with $-\infty$ as the lower limit of integration, so $P\{X \leq b\} = \int_{-\infty}^b p(x) dx.$

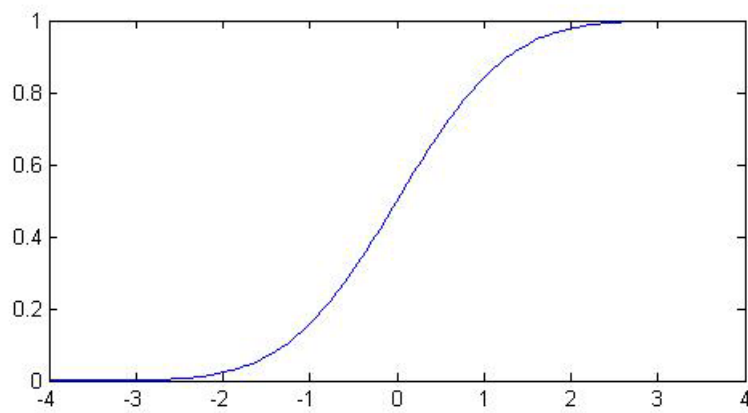
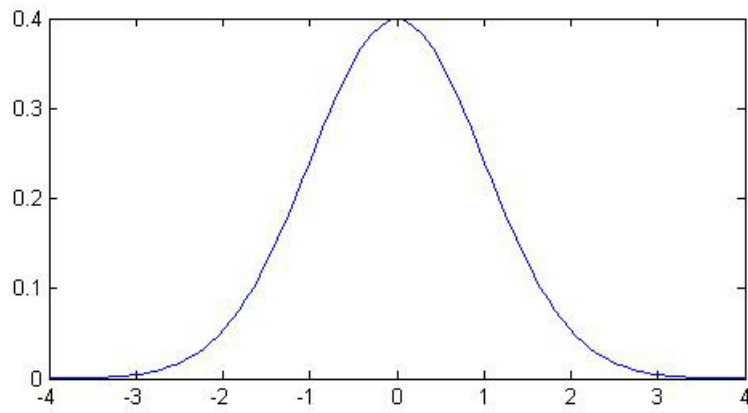
For this class you aren't required to use calculus but it's helpful to see why somebody might want to use it. (Note that many of the statistical distributions we'll talk about come up in solving partial differential equations such as are commonly used in finance – so if you're thinking of a career in that direction, you'll want even more math!)

Normal Distribution

We will most often use the Normal Distribution – but usually the first question from students is "Why is that crazy thing normal?!!!" You're not the only one to ask. Be patient, you'll see why; for now just remember e^{-x^2} .

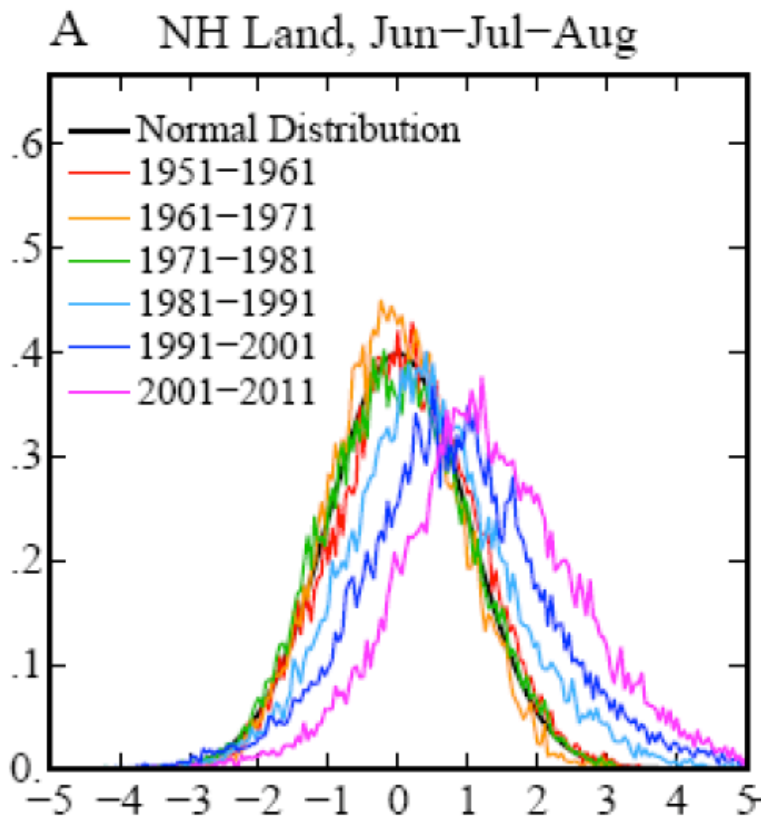
In statistics it is often convenient to use a normal distribution, the bell-shaped distribution that arises in many circumstances. It is useful because the (properly scaled) mean of independent random draws of many other statistical distributions will tend toward a normal distribution – this is the Central Limit Theorem.

Some basic facts and notation: a normal distribution with mean μ and standard deviation σ is denoted $N(\mu, \sigma)$. (The variance is the square of the standard deviation, σ^2 .) The Standard Normal distribution is when $\mu=0$ and $\sigma=1$; its probability density function (pdf) is denoted $\text{pdf}_N(x)$; the cumulative density function (CDF) is $\text{cdf}_N(x)$ or sometimes $\text{Nor}(x)$. This is a graph of the PDF (the height at any point) and CDF of the normal:



Example of using normal distributions:

A paper by Hansen, Sato, & Ruedy (2012) showed these decadal distributions of temperature anomalies:



This shows the rightward spread of temperature deviations. The x-axis is in standard deviations, which makes the various geographies easily comparable (a hot day in Alaska is different from a hot day in Oklahoma). The authors define extreme heat as more than 3 standard deviations above the mean and note that the probability of extreme heat days has risen from less than 1% to above 10%.

One of the basic properties of the normal distribution is that, if X is distributed normally with mean μ and standard deviation σ , then $Y = A + bX$ is also distributed normally, with mean $(A + b\mu)$ and standard deviation $b\sigma$. We will use this particularly when we "standardize" a sample: by subtracting its mean and dividing by its standard deviation, the result should be distributed with mean zero and standard deviation 1.

In some machine learning situations, data might be standardized, i.e. subtract the mean and divide by standard deviation, so $Z = \frac{X - \bar{X}}{s_X}$; or scaled to unit interval, so $W = \frac{X - X_{min}}{(X_{max} - X_{min})}$. Since these are linear transformations, we understand how these affect the distributions.

Oppositely, if we are creating random variables with a normal distribution, we can take random numbers with a $N(0,1)$ distribution, multiply by the desired standard deviation, and add the desired mean, to get normal random numbers with any mean or standard deviation. In Excel, you can create normally distributed random numbers by using the `RAND()` function to generate uniform random numbers on $[0,1]$, then `NORMSINV(RAND())` will produce standard-normal-distributed random draws.

In R, just use `rnorm()` to get random numbers from a normal distribution; you can multiply and add to get other mean/stdev or you can use the canned procedure, `rnorm(n, mean = 0, sd = 1)`.

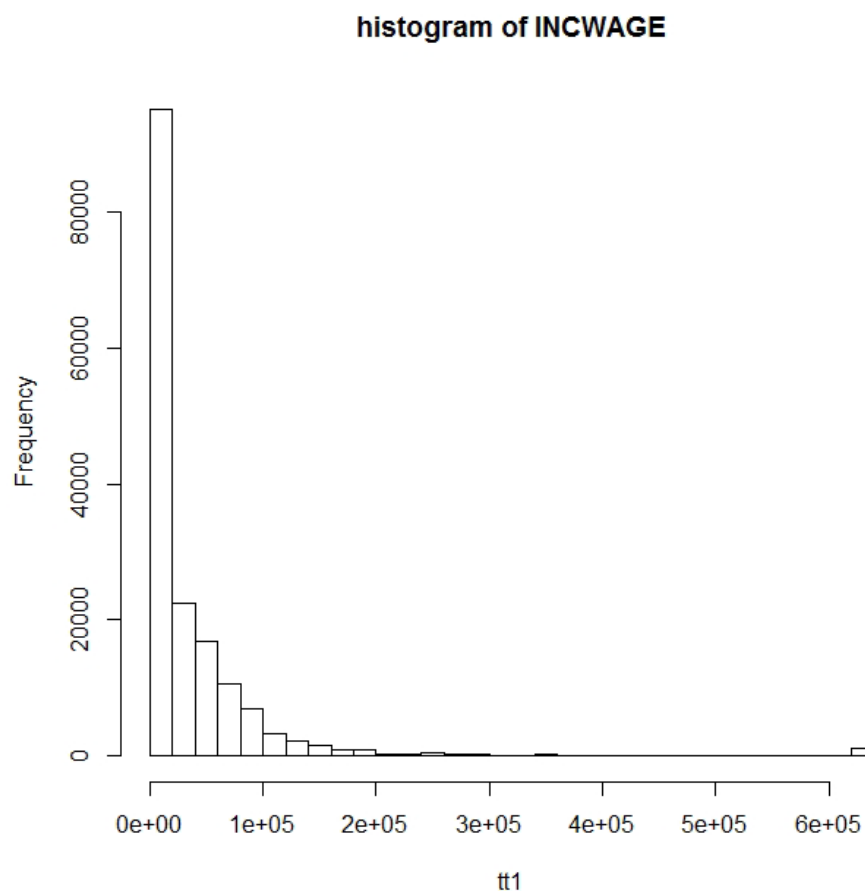
Motivation: Sample Averages are Normally Distributed

Before we do a long section on how to find areas under the normal distribution, I want to address the big question: Why the heck would anybody ever want to know those?!?!

Consider a case where we have a population of people and we sample just a few to calculate an average. Before elections we hear about these types of procedures all of the time: a poll that samples just 1000 people is used to give information about how a population of millions of people will vote. These polls are usually given with a margin of error ("54% of people liked Candidate A over B, with a margin of error of plus or minus 2 percentage points"). If you don't know statistics then polls probably seem like magic. If you do know statistics then polls are based on a few simple formulas.

For class we're using the PUMS NY data with 196,585 observations and for now concentrate on the income from wages (INCWAGE) data. The true average of all of those people (omitting the na values) is \$33,795.55. (Not quite; the top income value is cut at \$638,000 – people who made more are still just coded with that amount. But don't worry about that for now.) The standard deviation of the full data is 66,170.

A histogram of the data shows that most people report zero (zero is the median value), which is reasonable since many of them are children or retired people. However some report incomes up to \$638,000!



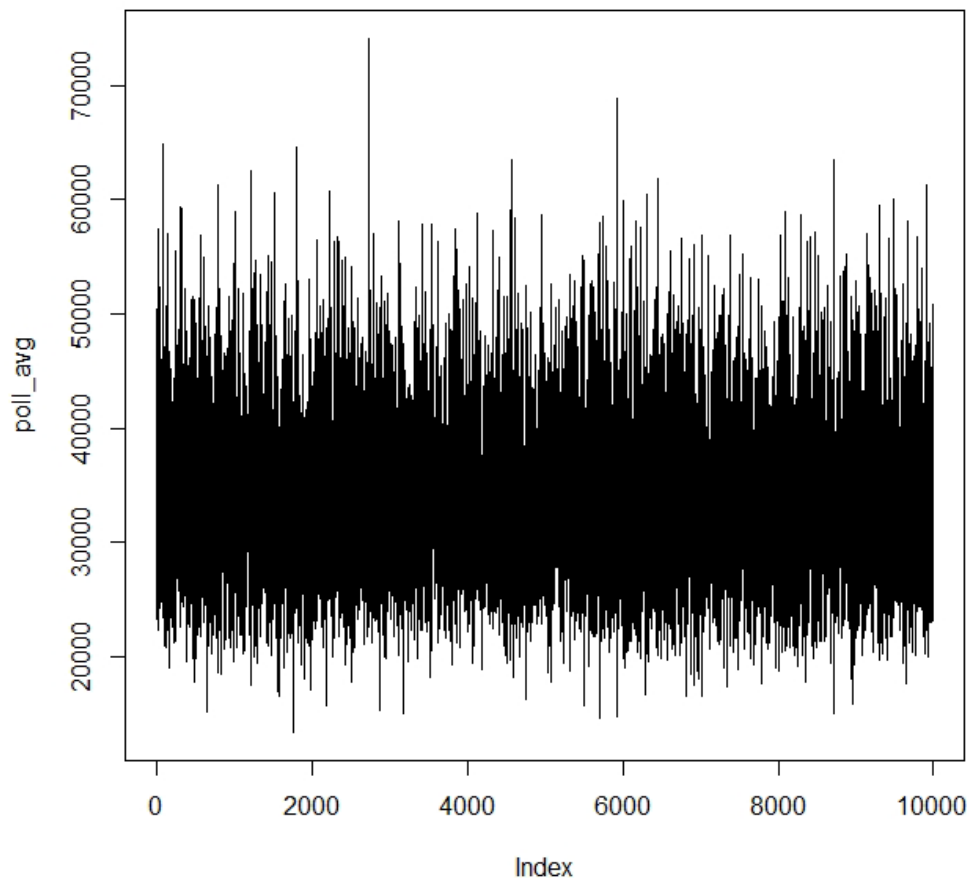
Taking an average of a population with such extreme values would seem to be difficult.

Suppose that I didn't want to calculate an average for all 196,585 people – I'm lazy or I've got a real old and slow computer or whatever. I want to randomly choose just 100 people and calculate the sample average. Would that be "good enough"?

Of course the first question is "good enough for what?" – what are we planning to do with the information?

But we can still ask whether the answer will be very close to the true value. In this case we know the true value; in most cases we won't. But this allows us to take a look at how the sampling works.

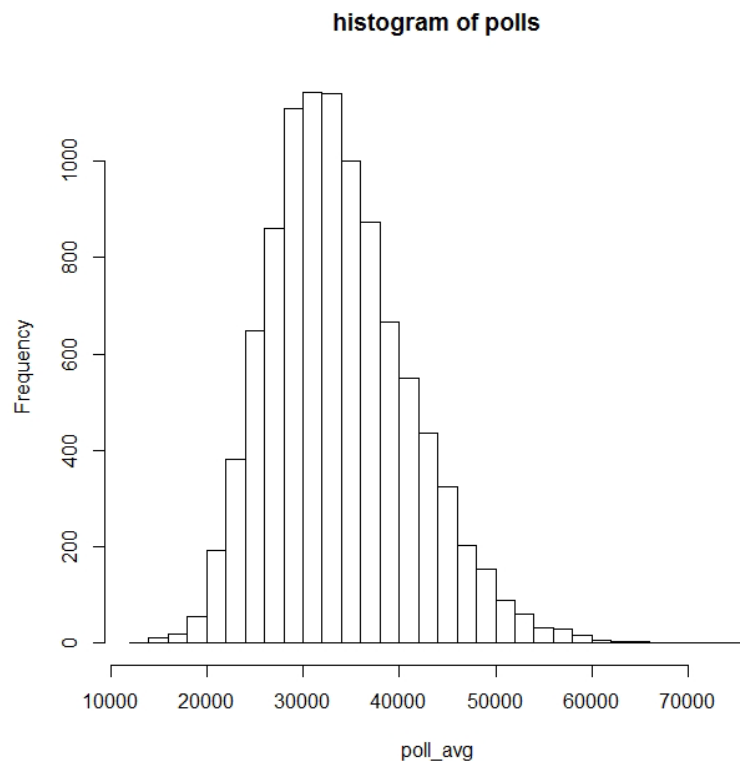
Here is a plot of values for 10000 different polls (each poll with just 100 people).



We can see that, although there are a few polls with averages as low almost 10,000 and a few with averages as high as 60,000, most of the polls are close to the true mean of \$33,796.

In general the average of even a small sample is a good estimate of the true average value of the population. While a sample might pick up some extreme values from one side, it is also likely to pick extreme values from the other side, which will tend to balance out.

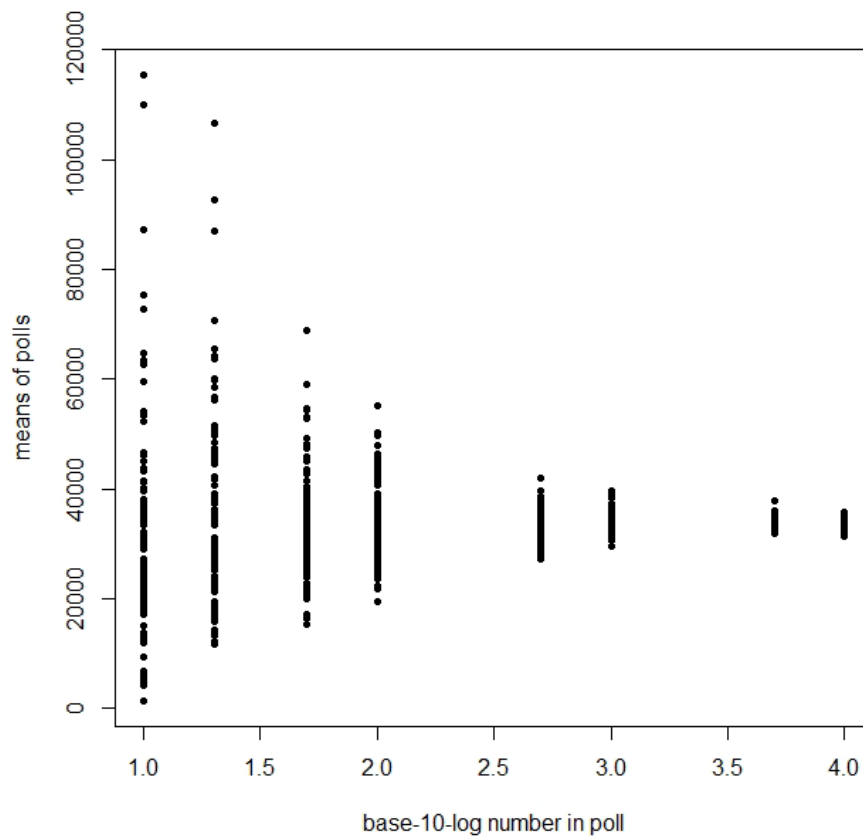
A histogram of the 10000 poll means is here:



This shows that the distribution of the sample means looks like a Normal distribution – another case of how "normal" and ordinary the Normal distribution is.

Of course the size of each sample, the number of people in each poll, is also important. Sampling more people gets us better estimates of the true mean.

This graph shows the results from 100 polls, each with different sample sizes.

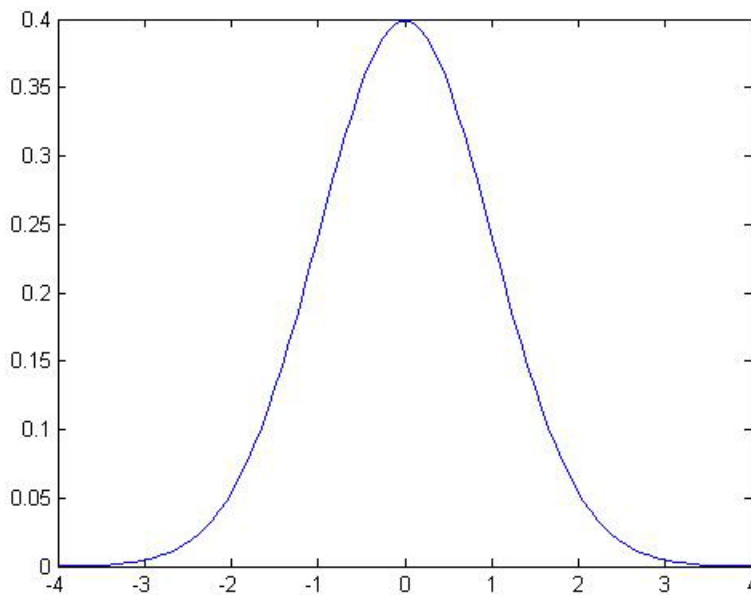


In the first set of 100 polls, on the left, each poll has just 10 people in it, so the results are quite varied. The next set has 20 people in each poll, so the results are closer to the true mean. By the time we get to 100 people in each poll (10^2 on the log-10-scale x-axis), the variation in the polls is much smaller. (Note that if you used the formulas from above instead of this Monte Carlo procedure, you would miss the asymmetry for the small polls.) As economists we would immediately see that there are diminishing marginal returns to sample size (and much of the business of polling derives from that).

Each distribution has a bell shape, but we have to figure out if there is a single invariant distribution or only a family of related bell-shaped curves.

If we subtract the mean, then we can center the distribution around zero, with positive and negative values indicating distance from the center. But that still leaves us with different scalings: as the graph above shows, the typical distance from the center gets smaller. So we divide by its standard deviation and we get a "Standard Normal" distribution.

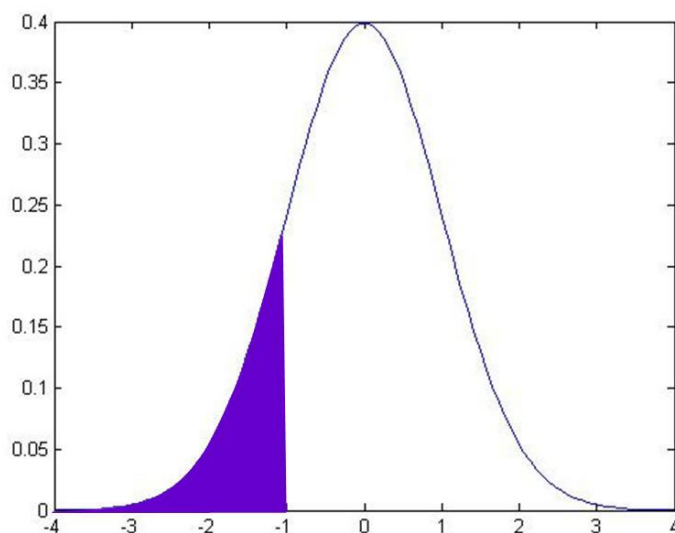
The Standard Normal graph is:



Note that it is symmetric around zero. Like any histogram, the area beneath the curve is a measure of the probability. The total area under the curve is exactly 1 (probabilities must add up to 100%). We can use the known function to calculate that the area under the curve, from -1 to 1, is 68.2689%. This means that just over 68% of the time, I will draw a value from within 1 standard deviation of the center. The area of the curve from -2 to 2 is 95.44997%, so we'll be within 2 standard deviations over 95.45% of the time.

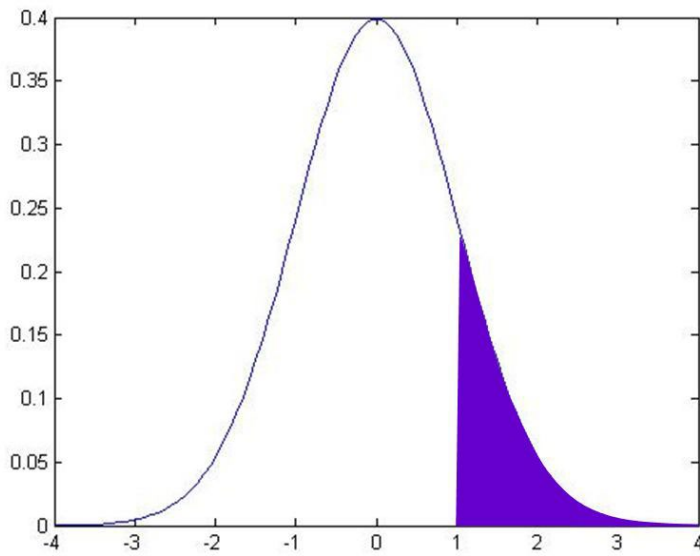
It is important to be able to calculate areas under the Standard Normal. For this reason people used to use big tables (statistics textbooks still have them); now we use computers. But even the computers don't always quite give us the answer that we want, we have to be a bit savvy.

So the normal CDF of, say, -1, is the area under the pdf of the points to the left of -1:

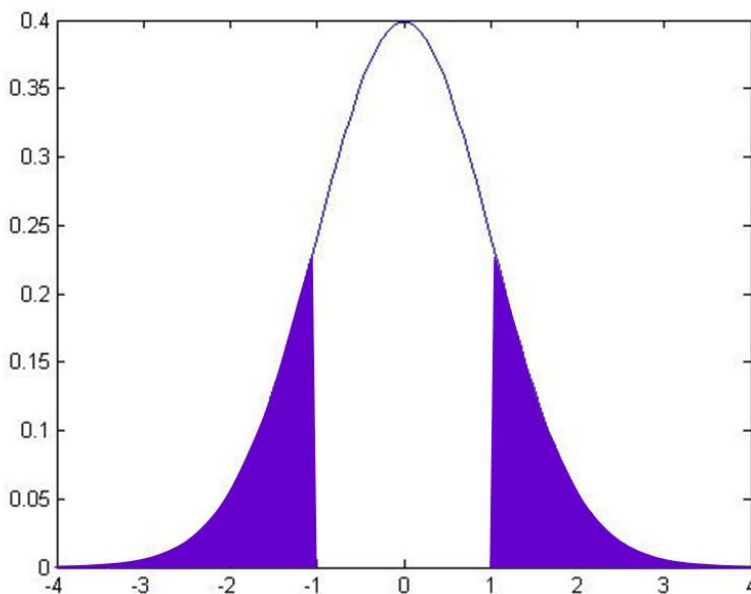


This area is 15.87%. How can I use this information to get the value that I earlier told you, that the area in between -1 and 1 is 68.2689%? Well, we know two other things (more precisely, I know them and I wrote them just 3 paragraphs up, so you ought to know them). We know that the total area under the pdf is 100%.

And we know that the pdf is symmetric around zero. This symmetry means that the area under the other tail, the area from +1 all the way to the right, is also 15.87%.



So to find the area in between -1 and +1, I take 100% and subtract off the two tail areas:



And this middle area is $100 - 15.87 - 15.87 = 68.26$.

Sidebar: you can think of all of this as "adding up" without calculus. On the other hand, calculus makes this procedure much easier and we can precisely define the cdf as the integral, from negative infinity to some

point Z , under the pdf:

$$cdf(Z) = \int_{-\infty}^Z pdf(x) dx$$

So with just this simple knowledge, you can calculate all sorts of areas using just the information in the CDF.

Hints on using Excel or R to calculate the Standard Normal cdf

Excel

Excel has `norm.s.dist` that assumes the mean is zero and standard deviation is one so you just use `norm.s.dist(X, TRUE)`. Read the help files to learn more. The final argument of the `normdist` function, "Cumulative" is a true/false: if true then it calculates the cdf (area to the left of X); if false it calculates the pdf. *[Personally, that's an ugly and non-intuitive bit of coding, but then again, Microsoft has no sense of beauty.]*

To figure out the other way – what X value gives me some particular probability, we use `norm.s.inv`.

All of these commands are under "Insert" then "Function" then, under "Select a Category" choose "Statistical".

Google

Mistress Google knows all. When I google "Normal cdf calculator" I get a link to http://www.uvm.edu/~dhowell/StatPages/More_Stuff/normalcdf.html. This is a simple and easy interface: put in the z-value to get the probability area or the inverse. Even ask Siri!

R

R has functions `pnorm()` and `qnorm()`. If you have a Z value and want to find the area under the curve to the left of that value, use `pnorm(X)`. If you don't tell it otherwise, it assumes mean is zero and standard deviation is one. If you want other mean/stdev combinations, add those – so leaving them out is same as `pnorm(X, mean = 0, sd = 1)` or change 0 and 1 as you wish. If you have a probability and want to go backwards to find X, then use `qnorm(p)`.

Side Note: *The basic property, that the distribution is normal whatever the time interval, is what makes the normal distribution {and related functions, called Lévy distributions} special. Most distributions would not have this property so daily changes could have different distributions than weekly, monthly, quarterly, yearly, or whatever!*

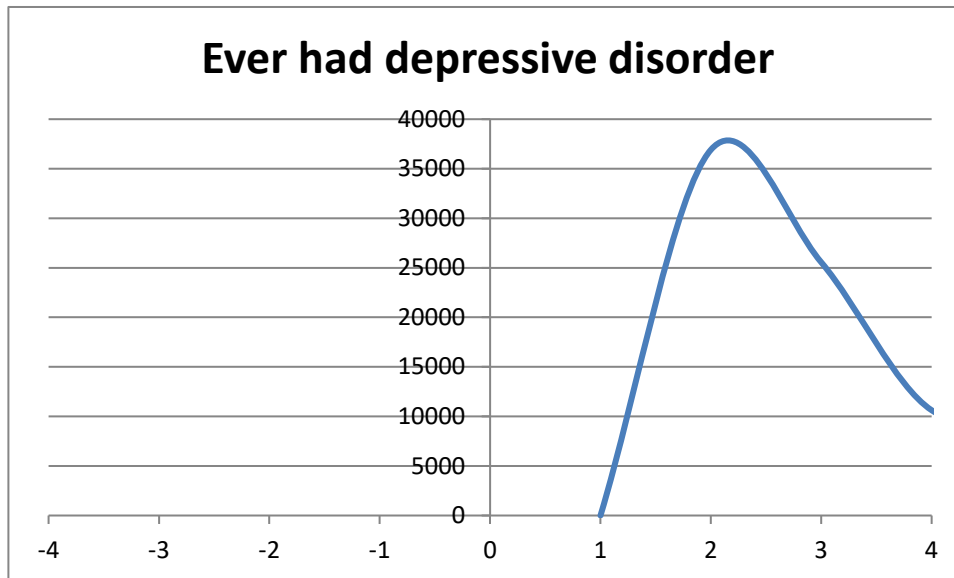
Recall from calculus the idea that some functions are not differentiable in places – they take a turn that is so sharp that, if we were to approximate the slope of the function coming at it from right or left, we would get very different answers. The function, $y = |x|$, is an example: at zero the left-hand derivative is -1; the right-hand derivative is 1. It is not differentiable at zero – it turns so sharply that it cannot be well approximated by local values. But it is continuous – it can be continuous even if it is not differentiable.

Now suppose I had a function that was everywhere continuous but nowhere differentiable – at every point it turns so sharply as to be unpredictable given past values. Various such functions have been derived by mathematicians, who call it a Wiener process; it generates Brownian motion. (When Einstein visited CCNY in 1905 he discussed his paper using Brownian motion to explain the movements of tiny particles in water, that are randomly bumped around by water molecules.) This function has many interesting properties – including an important link with the Normal distribution. The Normal distribution gives just the right degree of variation to allow continuity – other distributions would not be continuous or would have infinite variance.

Note also that a Wiener process has geometric form that is independent of scale or orientation – a Wiener process showing each day in the year cannot be distinguished from a Wiener process showing each

minute in another time frame. As we noted above, price changes for any time interval are normal, whether the interval is minutely, daily, yearly, or whatever. These are fractals, curious beasts described by mathematicians such as Mandelbrot, because normal variables added together are still normal. (You can read Mandelbrot's 1963 paper in the *Journal of Business*, which you can download from JStor – he argues that Wiener processes are unrealistic for modeling financial returns and proposes further generalizations.)

The Normal distribution has a pdf which has a formula that looks ugly but isn't so bad once you break it down. It is proportional to e^{-x^2} . This is what gives it a bell shape:



To make this a real probability we need to have all of its area sum up to one, so the probability density function (PDF) for a standard normal (with zero mean and standard deviation of one) is

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

To allow a mean, μ , different from zero and a standard deviation, σ , different from one, we modify the formula to this:

$$pdf_N = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

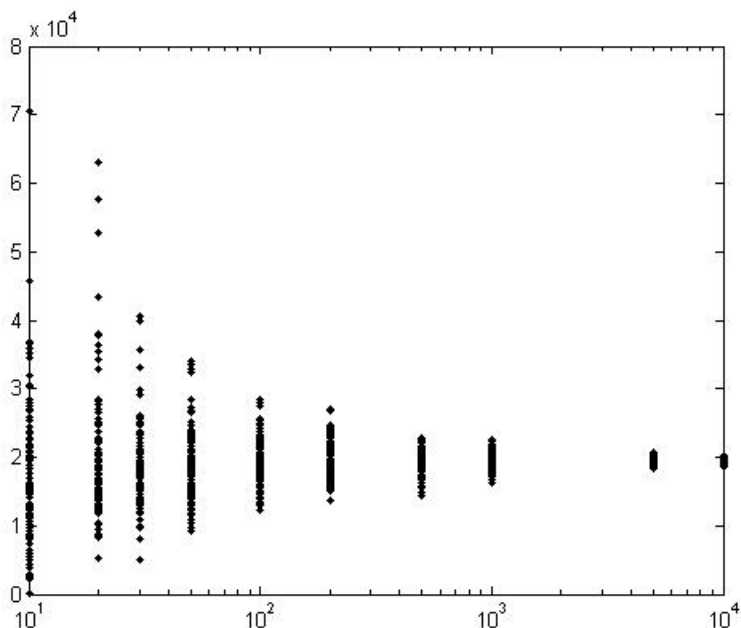
The connection with e is useful if it reminds you of when you learned about "natural logarithms" and probably thought "what the heck is 'natural' about that ugly thing?!" But you learn that it comes up everywhere (think it's bad now? wait for differential equations!) and eventually make your peace with it. So too the 'normal' distribution.

If you think that the PDF is ugly then don't feel bad – its discoverer didn't like it either. Stigler's History of Statistics relates that Laplace first derived the function as the limit of a binomial distribution as $n \rightarrow \infty$ but couldn't believe that anything so ugly could be true. So he put it away into a drawer until later when Gauss derived the same formula (from a different exercise) – which is why the Normal distribution is often referred to as "Gaussian". The Normal distribution arises in all sorts of other cases: solutions to partial differential equations; in physics Maxwell used it to describe the diffusion of gases or heat (again Brownian motion; video here <http://fuckyeahfluidynamics.tumblr.com/post/56785675510/have-you-ever-noticed-how-motes-of-dust-seem-to>); in information theory

where it is connected to standard measures of entropy (Kullback Liebler); even in the distribution of prime factors in number theory, the Erdős–Kac Theorem.

I'll note the statistical quincunx, which is a great word since it sounds naughty but is actually geeky (google it or I'll try to get an online version to play in class).

Final note on stratified sampling: Look again at this picture,



You can see, from the perspective of an economist, that the "production function" of accuracy as a function of the number of observations has diminishing returns – doubling the number of observations has a progressively smaller impact on accuracy. This is why many government data sets have weights for over-sampling of smaller populations. Suppose there are two groups of people; one makes up 90% of the population. Then if we randomly sample from the population, a sample of 1000 people would be expected to have 900 from one group (getting quite small standard errors) while just 100 from the other group (larger standard errors). The marginal increase in accuracy, from increasing the sample size, is very far from equal in the two groups. So many datasets oversample smaller populations – the equivalent of sampling 800 from the big group and 200 from the small group, then using the weights to fix the fact that the smaller group is oversampled. The exact procedures of weighting vary with the dataset. For this class, we will ignore the problem and not worry about the weights, but if you go on to do more stats, you can figure it out.

Is That Big?

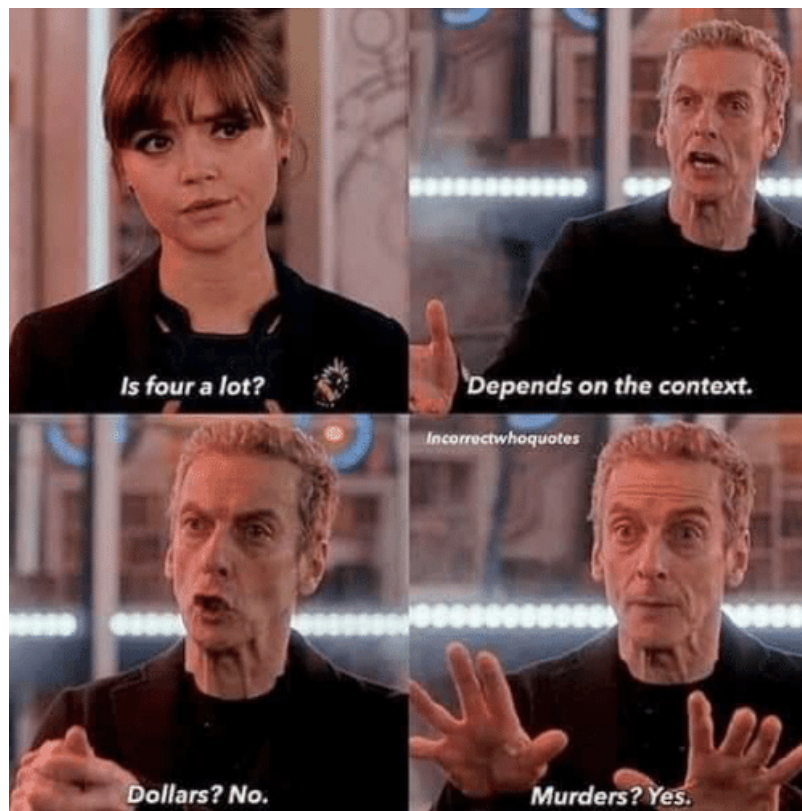
Learning Outcomes (from CFA exam Study Session 3, Quantitative Methods)

Students will be able to:

- define the standard normal distribution, explain how to standardize a random variable, and calculate and interpret probabilities using the standard normal distribution;

The sample average has a normal distribution. This is hugely important for two reasons: one, it allows us to estimate a parameter, and two, because it allows us to start to get a handle on the world and how we might be fooled.

You calculate some statistic, maybe it's a difference between means of two groups. But you immediately have to answer: is that big? Is it a big difference?







Well, it's about the standard error... We have to understand the issues of sampling.

Get a central parameter

The basic idea is that if we take the average of some sample of data, this average should be a good estimate of the true mean. For many beginning students this idea is so basic and obvious that you never think about when it is a reasonable assumption and when it might not be. For example, one of the causes of the Financial Crisis was that many of the 'quants' (the quantitative modelers) used overly-optimistic models that didn't seriously take account of the fact that financial prices can change dramatically. Most financial returns are not normally distributed! But we'll get more into that later; for now just remember this assumption. Later we'll talk about things like bias and consistency.

Return to the example of loading the dice, that we tried in the first homework assignment. Suppose we rolled 2 dice, and want to distinguish if either one is loaded. Call them "A" and "B". These are the results:

	A	B
Number of times roll 1 	4	2
... 2 	2	2
... 3 	5	4
... 4 	1	2

... 5 	4	4
... 6 	4	6

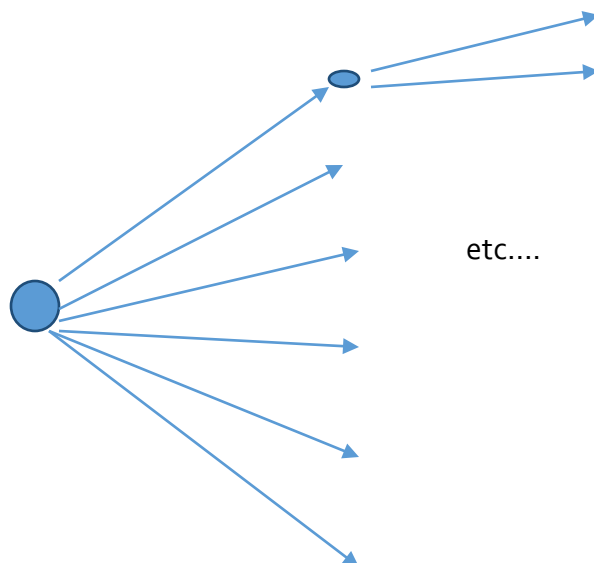
You might guess that B is loaded. But how likely is it? Could they both be fair?

A comes up with a 6 on $4/20 = 0.2$; B comes up as 6 on $6/20 = 0.3$. Both are higher than the expected value of 0.167. They are different but is that a big difference? (How big is 'big'?)

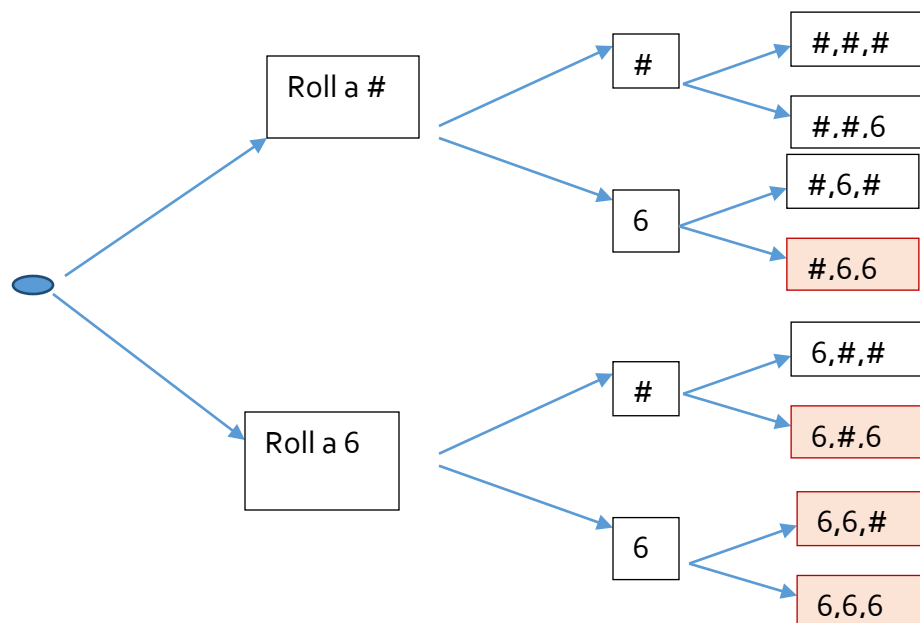
Thinking about Sampling Distributions

We tried to load dice, to get them to come up 6 more often. Suppose we want to test a dice to see if it actually comes up 6 more often, we could roll it once. If it comes up 6 then does that prove it's loaded? Well we know that a 6 comes up $1/6$ of the time even with a fair dice, so that's not too improbable. What about if the first 2 rolls come up 6 – how likely is that, if the dice were fair? Well the likelihood of getting 2 rolls of 6 is $(1/6) * (1/6) = 1/36$, so that gets less likely, under 3%. The likelihood of getting a 6 three times in a row is even less, $1/6^3 = 1/216 = .0046$. So if we keep rolling and keep getting a 6 each and every time, the likelihood of the dice being fair just keeps falling and falling. At some point we would decide that the likelihood of the dice being fair is just too low, and end the experiment.

But what if the dice came up 6 twice out of the first 3 rolls – would that be the same level of evidence? Again we might want to figure out how likely it would be, for a fair dice to come up with $2/3$ of the rolls as a 6. This is a bit more of a complicated permutation since either the first, second, or third roll could be the non-6 roll. Recall that we can represent it (as if in extended form of game) as:



But quickly I get lazy and don't want to draw 6 choices, each with 6 choices, each with 6 choices, but instead represent the choice of rolling either a 6 or not-a-6, so



Then figure the probabilities of each outcome, where probability of rolling 6 is $1/6$ and probability of rolling another number is $5/6$.

Now I don't know about you, but I don't have the patience to do that for too many more rounds. If I roll the dice 10 times and want to see how likely it is, that at least 3 of the 10 rolls will come up 6 ... that's just too much!

Fortunately we have a tool that is optimized for repeatedly doing very simple math problems, the computer. So fire up R!

```

# do one set of 10 rolls:
set.seed(12345)
x <- sample(6,10, replace = TRUE)
sum(x == 6)

# -----
NN = 100000
num_in_sampl <- rep(0,NN)
set.seed(12345)
for(indx in 1:NN) {
  x <- sample(6,10, replace = TRUE)
  num_in_sampl[indx] <- sum(x == 6)
}

h_s <- hist(num_in_sampl, breaks = c(-1,0,1,2,3,4,5,6,7,8,9,10))
prop.table(h_s$counts)

```

The next step is to ask, "do I have to do thousands of simulations every time?" Answer: "No, that's the power of stats!" Rather than doing a lot of simulations you can just find a formula. Sure the formula is a bit ugly but you've seen the program, it's not so easy either. (As you get more sophisticated you will find that there are tradeoffs to each method.)

Variation around central mean

Knowing that the sample average has a normal distribution also helps us specify the variation involved in the estimation. We often want to look at the difference between two sample averages, since this allows us to tell if there is a useful categorization to be made: are there really two separate groups? Or do they just happen to look different?

How can we try to guard against seeing relationships where, in fact, none actually exist?

To answer this question we must think like statisticians. To "think like a statistician" is to do mental handstands; it often seems like looking at the world upside-down. But as you get used to it, you'll discover how valuable it is. (There is another related question: "What if there really is a relationship but we don't find evidence in the sample?" We'll get to that.)

The first step in "thinking like a statistician" is to ask, What if there were actually no relationship; zero difference? What would we see? A big difference would be evidence in favor of different means; a small difference would be evidence against. But, in the phrase of Dierdre McCloskey, "How big is big?"

Law of Large Numbers

Probability and Statistics have many complications with twists and turns, but it all comes down to just a couple of simple ideas. These simple ideas are not necessarily intuitive – they're not the sort of things that might, at first, seem obvious. But as you get used to them, they'll become your friend.

One basic idea of statistics is the "Law of Large Numbers" (LLN). The LLN tells us that certain statistics (like the average) will very quickly get very close to the true value, as the size of the random sample increases. This means that if I want to know, say, the fraction of people who are right-handed or left-handed, or the fraction of people who will vote for Politician X versus Y, I don't need to talk with every person in the population.

This is strenuously counter-intuitive. You often hear people complain, "How can the pollsters claim to know so much about voting? They never talked to me!" But they don't have to talk to everyone; they don't even have to talk with very many people. The average of a random sample will "converge" to the true value in the population, as long as a few simple assumptions are satisfied.

With computers we can take much of the complicated formulas and derivations and just do simple experiments. Of course an experiment cannot replace a formal proof, but for the purposes of this course you don't need to worry about a formal proof.

R makes this easy. Run this little program, kind of like the dice example but for polling now:

```
# create the population of people
set.seed(1)
prob_of_yes <- 0.45
population_values <- runif(1000)
pop_yes <- (population_values < prob_of_yes)

# check that value should be near 0.45 although not exactly
mean(pop_yes)

# now do this the long way, for a sample of size 30 from the population
sampl_size <- 30
s1 <- sample(pop_yes, sampl_size)
mean(s1)
```



```
# you could go through and create s2, s3, etc or get lazy and do this...

# number of times to do this
NN <- 100

samples_from_pop <- matrix(data = NA, nrow = 1, ncol = NN)
for (i in 1:NN){
  samples_from_pop[i] <- mean(sample(pop_yes, 30))
}
hist(samples_from_pop)

# you can go through and play with sample size, population size, and how many different
samples to take (NN)
```

You could do this with a spreadsheet, lots of formulas like "`=if(RAND()<0.45,1,0)`" but that's ugly! And it doesn't make it easy to replicate, but with "`set.seed`" you should be able to replicate the same results each time on R. (Read R's help on random numbers if you want to learn about pseudo-random number generation.)

In the problem set, you will be asked to do some similar calculations.

So we can formulate many different sorts of questions once we have this figured out.

First the question of polls: if we poll 500 people to figure out if they approve or disapprove of the President, what will be the standard error?

Standard Error of Average

With some math (⚡) we can figure out a formula for the standard error of the sample average. It is just the standard deviation of the sample divided by the square root of the sample size. So the sample average is distributed normally with mean of μ and standard error of $se = \frac{s}{\sqrt{N}}$. This is sometimes written compactly as $\bar{X} \sim N\left(\mu, \frac{s}{\sqrt{N}}\right)$.

Sometimes this causes confusion because in calculating the standard error, s , we divided by the square root of $(N-1)$, since $s = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1}}$, so it seems you're dividing twice. But this is correct: the first division gets us an estimate of the sample's standard deviation; the second division by the square root of N gets us the estimate of the sample average's standard error.

The standardized test statistic (sometimes called Z-score since Z will have a standard normal distribution) is the mean divided by its standard error, $\frac{\bar{X}}{se} = \frac{\bar{X}}{\frac{s}{\sqrt{N}}} = \sqrt{N} \frac{\bar{X}}{s}$. This shows clearly that a larger sample size (bigger N) amplifies differences of \bar{X} from zero (the usual null hypothesis). A small difference, with only a few observations, could be just chance; a small difference, sustained over many observations, is less likely to be just chance.

One of the first things to note about this formula is that, as N rises (as the sample gets larger) the standard error gets smaller – the estimator gets more precise. So if N could rise towards infinity then the

sample average would converge to the true mean; we write this as $\bar{X} \xrightarrow[p]{\text{p}} \mu$ where the $\xrightarrow[p]{\text{p}}$ means "converges in probability as N goes toward infinity".

So the sample average is **unbiased**. This simply means that it gets closer and closer to the true value as we get more observations. Generally "unbiased" is a good thing, although later we'll discuss tradeoffs between bias and variance.

Return to the binomial distribution, and its normal approximation. We know that std error has its maximum when $p = 1/2$, so if we put in $p = 0.5$ then the standard error of a poll is, at worst, $\frac{1}{2\sqrt{n}}$, so more observations give a better approximation. See Excel sheet *poll_examples*. We'll return to this once we learn a bit more about the standard error of means.

A bit of Math:

We want to use our basic knowledge of linear combinations of normally-distributed variables to show that, if a random variable, X , comes from a normal distribution then its average will have a normal distribution with the same mean and the standard deviation of the sample divided by the square root of the sample size,

$$\bar{X} \sim N\left(\mu, \frac{s}{\sqrt{N}}\right).$$

The formula for the average is $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Consider first a case where there are just 2 observations.

This case looks very similar to our rule about, if $W = CX + DY$, then

$W \sim N\left(C\mu_X + D\mu_Y, \sqrt{C^2\sigma_X^2 + D^2\sigma_Y^2 + 2CD\sigma_{XY}}\right)$. With $N=2$, this is $\bar{X} = \frac{1}{2}X_1 + \frac{1}{2}X_2$, which has mean

$\frac{1}{2}\mu_{X1} + \frac{1}{2}\mu_{X2}$, and since each X observation comes from the same distribution then $\mu_{X1} = \mu_{X2}$ so the mean is μ_X (it's unbiased). You can work it out when there are n observations.

Now the standard error of the mean is

$\sqrt{\left(\frac{1}{2}\right)^2\sigma_{X1}^2 + \left(\frac{1}{2}\right)^2\sigma_{X2}^2 + 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\sigma_{XY}} = \sqrt{\frac{1}{4}\sigma_{X1}^2 + \frac{1}{4}\sigma_{X2}^2} = \frac{1}{2}\sqrt{\sigma_{X1}^2 + \sigma_{X2}^2}$. The covariance can be set to zero because we assume that we're making an independent random sample. Again since they come from the same distribution, $\sigma_{X1} = \sigma_{X2}$, the standard error is $\frac{1}{2}\sqrt{\sigma_X^2 + \sigma_X^2} = \frac{1}{2}\sqrt{2\sigma_X^2} = \frac{\sqrt{2}}{2}\sqrt{\sigma_X^2} = \frac{\sqrt{2}}{2}\sigma_X = \frac{1}{\sqrt{2}}\sigma_X$.

With n observations, the mean works out the same and the standard error of the average is

$$\sqrt{\left(\frac{1}{n}\right)^2 \sum_{i=1}^n \sigma_x^2} = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sigma_x^2} = \sqrt{\frac{n}{n^2} \sigma_x^2} = \frac{\sigma_x}{\sqrt{n}}.$$

Hypothesis Testing

One of the principal tasks facing the statistician is to perform hypothesis tests. These are a formalization of the most basic questions that people ask and analyze every day – just contorted into odd shapes. But as long as you remember the basic common sense underneath them, you can look up the precise details of the formalization that lays on top.

The basic question is "How likely is it, that I'm being fooled?" Once we accept that the world is random (rather than a manifestation of some god's will), we must decide how to make our decisions, knowing that we cannot guarantee that we will always be right. There is some risk that the world will seem to be one way, when actually it is not. The stars are strewn randomly across the sky but some bright ones seem to line up into patterns. So too any data might sometimes line up into patterns.

A formal hypothesis sets a mathematical condition that I want to test. Often this condition takes the form of some parameter being zero for no relationship or no difference.

Statisticians tend to stand on their heads and ask: What if there were actually **no** relationship? (Usually they ask questions of the form, "suppose the conventional wisdom were true?") This statement, about "no relationship," is called the **Null Hypothesis**, sometimes abbreviated as H_0 . The Null Hypothesis is tested against an **Alternative Hypothesis**, H_A .

Before we even begin looking at the data we can set down some rules for this test. We know that there is some probability that nature will fool me, that it will seem as though there is a relationship when actually there is none. The statistical test will create a model of a world where there is actually no relationship and then ask how likely it is that we could see what we actually see, "How likely is it, that I'm being fooled?"

The "likelihood that I'm being fooled" is the p-value.

For a scientific experiment we typically first choose the level of certainty that we desire. This is called the significance level. This answers, "How low does the p-value have to be, for me to accept the formal hypothesis?" To be fair, it is important that we set this value first because otherwise we might be biased in favor of an outcome that we want to see. By convention, economists typically use 10%, 5%, and 1%; 5% is the most common.

A five percent level of a test is conservative, it means that we want to see so much evidence that there is only a 5% chance that we could be fooled into thinking that there's something there, when nothing is actually there. Five percent is not perfect, though – it still means that of every 20 tests where I decide that there is a relationship there, it is likely that I'm being fooled in one of those – I'm seeing a relationship where there's nothing there.

To help ourselves to remember that we can never be truly certain of our judgment of a test, we have a peculiar language that we use for hypothesis testing. If the "likelihood that I'm being fooled" is less than 5% then we say that the data allow us to *reject* the null hypothesis. If the "likelihood that I'm being fooled" is more than 5% then the data *do not reject* the null hypothesis.

Note the formalism: we never "accept" the null hypothesis. Why not? Suppose I were doing something like measuring a piece of machinery, which is supposed to be a centimeter long. The null hypothesis is that it is not defective and so is one centimeter in length. If I measure with a ruler I might not find any difference to the eye. So I cannot reject the hypothesis that it is one centimeter. But if I looked with a microscope I might find that it is not quite one centimeter! The fact that, with my eye, I don't see any

difference, does not imply that a better measurement could not find any difference. So I cannot say that it is truly exactly one centimeter; only that I can't tell that it isn't.

Or again with the example of dice – the 6 might come up slightly more than $1/6$ of the time, maybe if I rolled a million times I might finally distinguish a difference. But our hypothesis testing is much more limited, all we can say is that given the available tests we can't find a difference.

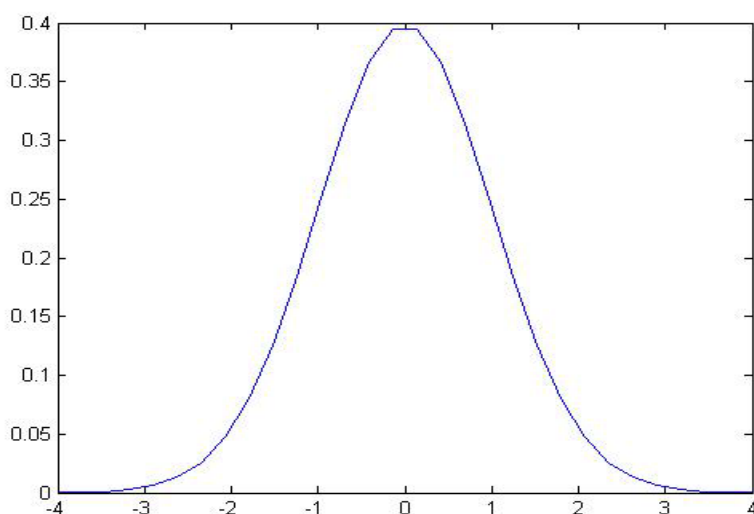
So too with statistics. If I'm looking to see if some portfolio strategy produces higher returns, then with one month of data I might not see any difference. So I would not reject the null hypothesis (that the new strategy is no improvement). But it is possible that the new strategy, if carried out for 100 months or 1000 months or more might show some tiny difference.

Not rejecting the null is saying that I'm not sure that I'm not being fooled. (Read that sentence again; it's not immediately clear but it's trying to make a subtle and important point.)

To summarize, Hypothesis Testing asks, "What is the chance that I would see the value that I've actually got, if there truly were no relationship?" If this p-value is lower than 5% then I reject the null hypothesis of "no relationship." If the p-value is greater than 5% then I do not reject the null hypothesis of "no relationship."

The rest is mechanics.

The null hypothesis would tell that a parameter has some particular value, say zero: $H_0 : \mu = 0$; the alternative hypothesis is $H_A : \mu \neq 0$. Under the null hypothesis the parameter has some distribution (often normal), so $H_0 : \mu \sim N(0, \sigma_{std\ err})$. Generally we have an estimate for $\sigma_{std\ err}$, which is se (for small samples this inserts additional uncertainty). So I know that, under the null hypothesis, $\frac{\mu}{se}$ has a standard normal distribution (mean of zero and standard deviation of one). I know exactly what this distribution looks like, it's the usual bell-shaped curve:



So from this I can calculate, "What is the chance that I would see the value that I've actually got, if there truly were no relationship?," by asking what is the area under the curve that is farther away from zero

than the value that the data give. (I still don't know what value the data will give! I can do all of this calculation beforehand.)

A particular estimate of μ is generally going to be \bar{X} . So the test statistic is formed with $\frac{\bar{X}}{se}$.

Looking at the standard normal pdf, a value of the test statistic of 1.5 would not meet the 5% criterion (go back and calculate areas under the curve). A value of 2 would meet the 5% criterion, allowing us to reject the null hypothesis. For a 5% significance level, the standard normal **critical value** is 1.96: if the test statistic is larger than 1.96 (in absolute value) then its p-value is less than 5%, and vice versa. (You can find critical values by looking them up in a table or using the computer.)

Sidebar: Sometimes you see people do a one-sided test, which is within the letter of the law but not necessarily the spirit of the law (particularly in regression formats). It allows for less restrictive testing, as long as we believe that we know that there is only one possible direction of deviation (so, for example, if the sample could be larger than zero but never smaller). But in this case maybe the normal distribution is inapplicable. Personally whenever I read a paper where the authors do a one-sided test, I immediately become suspicious.

The test statistic can be transformed into measurements of μ or into a confidence interval.

If I know that I will reject the null hypothesis of $\mu = 0$ at a 5% level if the test statistic, $\frac{\bar{X}}{se}$, is greater than 1.96 (in absolute value), then I can change around this statement to be about \bar{X} . This says that if the estimated value of \bar{X} is less than 1.96 standard errors from zero, we cannot reject the null hypothesis. So cannot reject if:

$$\frac{|\bar{X}|}{se} < 1.96$$

$$|\bar{X}| < 1.96se$$

$$-1.96se < \bar{X} < 1.96se.$$

This range, $(-1.96se, 1.96se)$, is directly comparable to \bar{X} . If I divide \bar{X} by its standard error then this ratio has a normal distribution with mean zero and standard deviation of one. If I don't divide then \bar{X} has a normal distribution with mean zero and standard deviation, se .

If the null hypothesis is not zero but some other number, μ_{null} , then under the null hypothesis the estimator would have a normal distribution with mean of μ_{null} and standard error, se . To transform this to a standard normal would mean subtracting the mean and dividing by se , so cannot reject if $\frac{|\bar{X} - \mu_{null}|}{se} < 1.96$, i.e. cannot reject if \bar{X} is within the range, $(\mu_{null} - 1.96se, \mu_{null} + 1.96se)$.

Confidence Intervals

We can use the same critical values to construct a confidence interval for the estimator, usually expressed in the form $\bar{X} \pm 1.96se$. This shows that, for a given sample size (therefore se , which depends on

the sample size) that there is a 95% likelihood that the interval formed around a given estimator contains the true value.

This relates to hypothesis testing because if the confidence interval includes the null hypothesis then we cannot reject the null; if the null hypothesis value is outside of the confidence interval then we can reject the null.





Find p-values

We can also find p-values associated with a particular null hypothesis by turning around the process outlined above. If the null hypothesis is zero, then with a 5% significance level we reject the null if $\frac{\bar{X}}{se}$ is greater than 1.96 in absolute value. What if the ratio $\frac{\bar{X}}{se}$ were 2 – what is the smallest significance level that would still reject? (Check your understanding: is it more or less than 5%?)

We can compute the ratio $\frac{\bar{X}}{se}$ and then convert this number to a p-value, which is the smallest significance level that would still reject the null hypothesis (and if the null is rejected at a low level then it would automatically be rejected at any higher levels).

Type I and Type II Errors

Whenever we use statistics we must accept that there is a likelihood of errors. In fact we distinguish between two types of errors, called (unimaginatively) Type I and Type II. These errors arise because a null hypothesis could be either true or false and a particular value of a statistic could lead me to reject or not reject the null hypothesis, H_0 . A table of the four outcomes is:

	H_0 is true	H_0 is false
Do not reject H_0	 good!	oops – Type II 
Reject H_0	oops – Type I 	 good!

Our chosen significance level (usually 5%) gives the probability of making an error of Type I. We cannot control the level of Type II error because we do not know just how far away H_0 is from being true. If our null hypothesis is that there is a zero relationship between two variables, when actually there is a tiny, weak relationship of 0.0001%, then we could be very likely to make a Type II error. If there is a huge, strong relationship then we'd be much less likely to make a Type II error.

There is a tradeoff (as with so much else). If I push down the likelihood of making a Type I error (using 1% significance not 5%) then I must be increasing the likelihood of making a Type II error.

Edward Gibbon notes that the emperor Valens would "satisfy his anxious suspicions by the promiscuous execution of the innocent and the guilty" (chapter 26). This rejects the null hypothesis of "innocence"; so a great deal of Type I error was acceptable to avoid Type II error.

Every email system fights spam with some sort of test: what is the likelihood that a given message is spam? If it's spam, put it in the "Junk" folder; else put it in the inbox. A Type I error represents putting good mail into the "Junk" folder; Type II puts junk into your inbox.

People play with setting the null hypothesis:

- There is an advertisement for gas, "no other brand has been proven to be better";
- Rand Paul offered a law that would allow a drug maker to publish any claim about drug efficacy that has not been proven false – does this mean that the claims will be true?;
- Regulators of chemicals face this problem: policy of prohibit use of chemicals proved to be unsafe vs. policy of only allow chemicals proved to be safe.

Examples

Assume that the calculated average is 3, the sample standard deviation is 15, and there are 100 observations. The null hypothesis is that the average is zero. The standard error of the average is

$se = \frac{15}{\sqrt{100}} = 1.5$. We can immediately see that the sample average is more than two standard errors away from zero so we can reject at a 95% confidence level.

Doing this step-by-step, the average over its standard error is $\frac{\bar{X}}{se} = \frac{3}{1.5} = 2$. Compare this to 1.96 and see that $2 > 1.96$ so we can reject. Alternately we could calculate the interval, $(-1.96s, 1.96s)$, which is $((-1.96 \cdot 1.5), (1.96 \cdot 1.5)) = (-2.94, 2.94)$, outside of which we reject the null. And 3 is outside that interval. Or calculate a 95% confidence interval of $3 \pm 2.94 = (0.06, 5.94)$, which does not contain zero so we can reject the null. The critical value for the estimate of 3 is 4.55% (found from Excel either $2 \cdot (1 - \text{NORMSDIST}(2))$ if using the standard normal distribution or $2 \cdot (1 - \text{NORMDIST}(3, 0, 1.5, \text{TRUE}))$ if using the general normal distribution with a mean of zero and standard error of 1.5).

If the sample average were -3, with the same sample standard deviation and same 100 observations, then the conclusions would be exactly the same.

Or suppose you find that the average difference between two samples, X and Y, (i.e.

$\bar{X} - \bar{Y} = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i)$) is -0.0378. The sample standard deviation is 0.357. The number of observations is 652.

These three pieces of information are enough to find confidence intervals, do t-tests, and find p-values.

How?

First find the standard error of the average difference. This standard error is 0.357 divided by the square root of the number of observations, so $\frac{.357}{\sqrt{652}} = 0.01398$.

So we know (from the Central Limit Theorem) that the average has a normal distribution. Our best estimate of its true mean is the sample average, -0.0378. Our best estimate of its true standard error is the sample standard error, 0.01398. So we have a normal distribution with mean -0.0378 and standard error 0.01398.

We can make this into a standard normal distribution by adding 0.0378 and dividing by the sample standard error, so now the mean is zero and the standard error is one.

We want to see how likely it would be, if the true mean were actually zero, that we would see a value as extreme as -0.0378. (Remember: we're thinking like statisticians!)

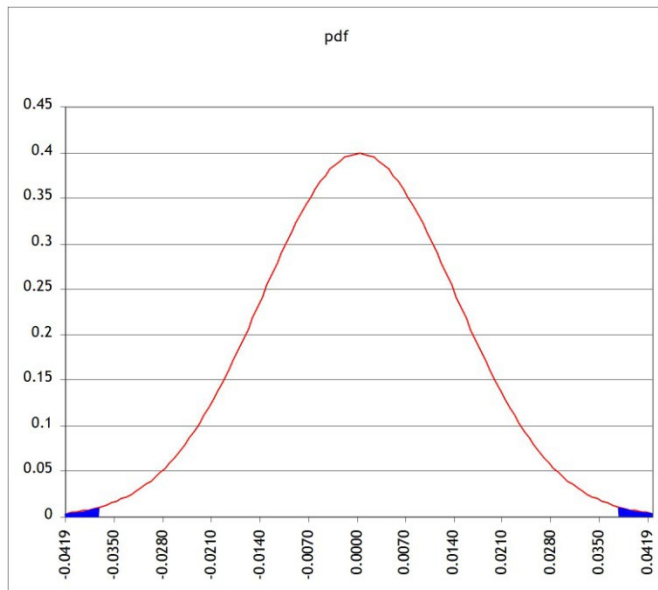
The value of -0.0378 is $\frac{-0.0378}{0.01398} = -2.70$ standard deviations from zero.

From this we can either compare this against critical t-values or use it to get a p-value.

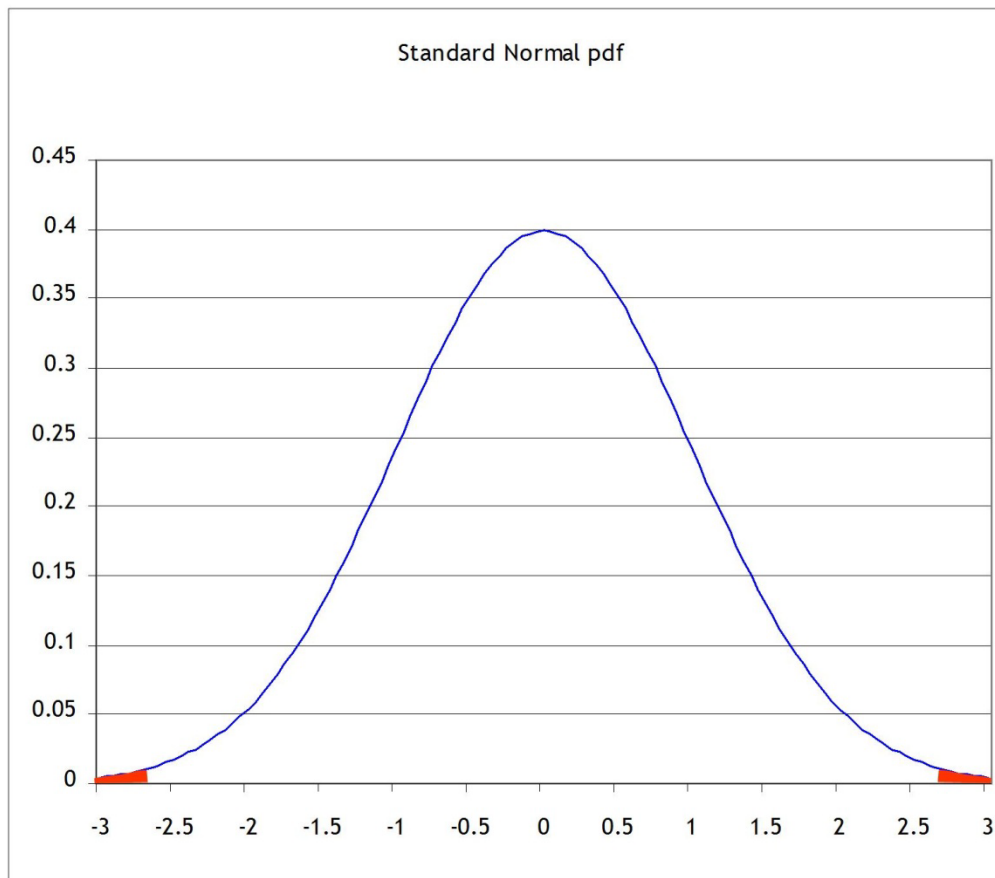
To find the p-value, we can use Excel. If we have a standard normal distribution, what is the probability of finding a value as far from zero as -2.27, if the true mean were zero? This is $2*(1-NORMSDIST(-2.27)) = 0.6\%$. The p-value is 0.006 or 0.6%. If we are using a 5% level of significance then since 0.6% is less than 5%, we reject the null hypothesis of a zero mean. If we are using a 1% level of significance then we can reject the null hypothesis of a zero mean since 0.6% is less than 1%.

Or instead of standardizing we could have used Excel's other function to find the probability in the left tail, the area less than -0.0378, for a distribution with mean zero and standard error 0.01398, so $2*NORMDIST(-0.0378,0,0.01398,TRUE) = 0.6\%$.

Standardizing means (in this case) zooming in, moving from finding the area in the tail of a very small pdf, like this:



to moving to a standard normal, like this:



But since we're only changing the units on the x-axis, the two areas of probability are the same.

We could also work backwards. We know that if we find a standardized value greater (in absolute value) than 1.96, we would reject the null hypothesis of zero at the 5% level. (You can go back to your notes and/or HW1 to remind yourself of why 1.96 is so special.)

We found that for this particular case, each standard deviation is of size $\frac{.357}{\sqrt{652}} = 0.01398$. So we can multiply 1.96 times this value to see that if we get a value for the mean, which is farther from zero than $0.01398 \times 1.96 = 0.0274$, then we would reject the null. Sure enough, our value of -0.0378 is farther from zero, so we reject.

Alternately, use this 1.96 times the standard error to find a confidence interval. Choose 1.96 for a 95% confidence interval, so the confidence interval around -0.0378 is plus or minus 0.0274, -0.0378 ± 0.0274 , which is the interval (-0.0652, -0.0104). Since this interval does not include zero we can be 95% confident that we can reject the null hypothesis of zero.

Sometimes we want to compare groups and ask, are they statistically significantly different from each other? Our formula that we learned previously has only one n – what do we do if we have two samples?

We want to figure out how to use the two separate standard errors to estimate the joint standard error; otherwise we'll use the same basic strategy to get our estimate, subtract off the null hypothesis (usually zero), and divide by its standard error. We just need to know, what is that new standard error?

To do this we use the sum of the two sample variances: if we are testing group 1 vs group 2, then a test of just group 1 would estimate its variance as $\frac{s_1^2}{n_1}$, a test of group 2 would use $\frac{s_2^2}{n_2}$, and a test of the group would estimate the standard error as $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

We can make either of two assumptions: either the standard deviations are the same (even though we don't know them) or they're different (even though we don't know how different). It is more conservative to assume that they're different (i.e. don't assume that they're the same) – this makes the test less likely to reject the null.

Assuming that the standard errors are different, we compare this test statistic against a t-distribution with degrees of freedom of the minimum of either $(n_1 - 1)$ or $(n_2 - 1)$.

P-values

If confidence intervals weren't confusing enough, we can also construct equivalent hypothesis tests with p-values. Where the Z-statistic tells how many standard errors away from zero is the observed difference (leaving it for us to know that more than 1.96 implies less than 5%), the p-value calculates this directly. So a p-value for the difference above, between time spent by those with a college degree and those with an advanced degree, is found from $-4.7919/1.6403 = -2.92$. So the area in the tail to the left of -2.92 is $\text{NORMSDIST}(-2.92) = .0017$; the area in both tails symmetrically is .0034. The p-value for this difference is 0.34%; there is only a 0.34% chance that, if the true difference were zero, we could observe a number as big as -4.7919 in a sample of this size.

Confidence Intervals for Polls

I promised that I would explain to you how pollsters figure out the " ± 2 percentage points" margin of error for a political poll. Now that you know about Confidence Intervals you should be able to figure these out. Remember (or go back and look up) that for a binomial distribution the standard error is $\sqrt{\frac{p(1-p)}{N}}$, where p is the proportion of "one" values and N is the number of respondents to the poll. We can use our estimate of the proportion for p or, to be more conservative, use the maximum value of $p(1-p)$ where $p = \frac{1}{2}$. A bit of quick math shows that with $p = \frac{1}{2}$, $\sqrt{p(1-p)} = \sqrt{\frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2} = 0.5$. So a poll of 100 people has a maximum standard error of $\frac{.5}{\sqrt{100}} = \frac{.5}{10} = .05$; a poll of 400 people has maximum standard error half that size, of .025; 900 people give a maximum standard error 0.0167, etc.

A 95% Confidence Interval is 1.96 (from the Standard Normal distribution) times the standard error; how many people do we need to poll, to get a margin of error of ± 2 percentage points? We want

$$1.96 \sqrt{\frac{p(1-p)}{N}} < .02 \text{ so this is, at maximum where } p = \frac{1}{2}, 2401.$$

A polling organization therefore prices its polls depending on the client's desired accuracy: to get ± 2 percentage points requires between 2000 and 2500 respondents; if the client is satisfied with just ± 5 percentage points then the poll is cheaper. (You can, and for practice should, calculate how many

respondents are needed in order to get a margin of error of 2, 3, 4, and 5 percentage points. For extra, figure that a pollster needs to only get the margin to ± 2.49 percentage points in order to round to ± 2 , so they can get away with slightly fewer.)

Here's a devious problem:

1. You are in charge of polling for a political campaign. You have commissioned a poll of 300 likely voters. Since voters are divided into three distinct geographical groups (A, B and C), the poll is subdivided into three groups with 100 people each. The poll results are as follows:

	total	A	B	C
number in favor of candidate	170	58	57	55
number total	300	100	100	100

Note that the standard deviation of the sample (not the standard error of the average) is given.

- a. Calculate a t-statistic, p-value, and a confidence interval for the main poll (with all of the people) and for each of the sub-groups.
- b. In simple language (less than 150 words), explain what the poll means and how much confidence the campaign can put in the numbers.
- c. Again in simple language (less than 150 words), answer the opposing candidate's complaint, "The biased media confidently says that I'll lose even though they admit that they can't be sure about any of the subgroups! That's neither fair nor accurate!"

Complications from a Series of Hypothesis Tests

Often a modeler will make a series of hypothesis tests to attempt to understand the inter-relations of a dataset. However while this is often done, it is not usually done correctly. Recall from our discussion of Type I and Type II errors that we are always at risk of making incorrect inferences about the world based on our limited data. If a test has an significance level of 5% then we will not reject a null hypothesis until there is just a 5% probability that we could be fooled into seeing a relationship where there is none. This is low but still is a 1-in-20 chance. If I do 20 hypothesis tests to find 20 variables that significantly impact some variable of interest, then it is likely that one of those variables is fooling me (I don't know which one, though). It is also likely that my high standard of proof meant that there are other variables which are more important but which didn't seem it.

Sometimes you see very stupid people who collect a large number of possible explanatory variables, run hundreds of regressions, and find the ones that give the "best-looking" test statistics – the ones that look good but are actually entirely fictitious. Many statistical programs have procedures that will help do this; help the user be as stupid as he wants.

Why is this stupid? It completely destroys the logical basis for the hypothesis tests and makes it impossible to determine whether or not the data are fooling me. In many cases this actually guarantees that, given a sufficiently rich collection of possible explanatory variables, I can run a regression and show that some variables have "good" test statistics – even though they are completely unconnected. Basically this is the infamous situation where a million monkeys randomly typing would eventually write Shakespeare's plays. A million earnest analysts, running random regressions, will eventually find a regression that looks great – where all of the proposed explanatory variables have test statistics that look great. But that's just due to persistence; it doesn't reflect anything about the larger world.

Consider the logical chain of making a number of hypothesis tests in order to find one supposedly-best model. When I make the first test, I have 5% chance of making a Type I error. Given the results of this test, I make the second test, again with a 5% chance of making a Type I error. The probability of not making an error on either test is $(.95)(.95) = .9025$ so the significance level of the overall test procedure is not 5% but $1 - .9025 = 9.75\%$. If I make three successive hypothesis tests, the probability of not making an error is .8574 so the significance level is 14.26%. If I make 10 successive tests then the significance level is over 40%! This means that there is a 40% chance that the tester is being fooled, that there is not actually the relationship there that is hypothesized – and worse, the stupid tester believes that the significance level is just 5%.

Issues with Canned Tests

Students often use a pre-written statistical test to declare that some difference is or is not statistically significant. This is a great efficiency! But it shouldn't come at the expense of understanding. What is being measured? To state that something is statistically significant is to state that it is "big" – so you'd better make sure that you know what in fact is big!

Let me give an example from an old exam. Take a moment to do this problem. In a medical study (reference below), people were randomly assigned to use either antibacterial products or regular soap. In total 592 people used antibacterial soap; 586 used regular soap. It was found that 33.1% of people using antibacterial products got a cold; 32.3% of people using regular soap got colds.

- a. Test the null hypothesis that there is no difference in the rates of sickness for people using regular or antibacterial soap. (What is the p-value?)

Standard deviation : $\sqrt{p(1-p)}$: $\sqrt{.331(1-.331)}$

Standard error: $\sqrt{p(1-p)/n}$: $\sqrt{.331(1-.331)/592}$

Difference .331 - .323

Standard error of difference: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- b. Create a 95% confidence interval for the difference in sickness rates. What is the 90% confidence interval? The 99% interval?

E.L.Larson, S.X. Lin, C. Gomez-Pichardo, P. Della-Latta, (2004). "Effect of Antibacterial Home Cleaning and Handwashing Products on Infectious Disease Symptoms: A Randomized Double-Blind Trial," Ann Intern Med, 140(5), 321-329.

Many students obliged by forming a statistical test to show whether there was a significant difference, but without ever noticing the counter-intuitive direction! In this case a test of statistical significance is useless and irrelevant – certainly you don't need to do any calculations to assert that this study shows no beneficial effect of antibacterial soap!

On many homework assignments, I've observed similar answers. Students rush into the mechanics of the test without any assessment. A statistical test is an important component of an argument but it is not the alpha and omega. Much more of the time and mental effort needs to go into thinking about the other factors – why might you observe these values? Have you got the right measure in the first place? Have you got a reasonable sample? What are some of the possible hypotheses that explain the difference? Is there a way to eliminate some of these hypotheses or to reduce the variation?

Once you've done the hard thinking and got an interesting measure, you can ask whether it is statistically significant. And this class will help you be more adroit with those tests.

Bayesian Stats

A reminder about basic stats – and illustration of the power of Bayesian statistics.

We did this example before: a 99% accurate test reveals that a person tests positive for a disease. How likely does the patient actually have the disease?

It depends.

If population overall has prevalence of 0.1%, then testing 1000 people will find the one person with disease plus 10 who don't have it (1% error of 99% test; 1% of 1000 = 10) – so a positive test for the disease means a 1/11 chance of actually having it.

On the other hand, if a subgroup of the population has a higher prevalence (say 1%) then putting together this prior information with the fact of a positive test implies that a positive test means about a 50% chance that the patient actually has the disease (10 people who have it plus 10 false positives).

So in the first case, the expected value of whether the person has the disease is 0.09 (=1/11); in the second case the expected value is 0.5. So the expected value depends on the empirical information (positive test result) but also the prior expectation (what is your guess of prevalence in subgroup).

In much of stats you can see this tradeoff between data and prior. In this case, with one data point, the prior is very important. With more data the importance of the prior recedes, but there are many important cases where people's priors remain a key determinant.

Details of Distributions T-distributions, chi-squared, etc.

Take the basic methodology of Hypothesis Testing and figure out how to deal with a few complications.

T-tests

The first complication is if we have a small sample and we're estimating the standard deviation. In previous examples, we used a large sample. For a small sample, the estimation of the standard error introduces some additional noise – we're forming a hypothesis test based on an estimation of the mean, using an estimation of the standard error.

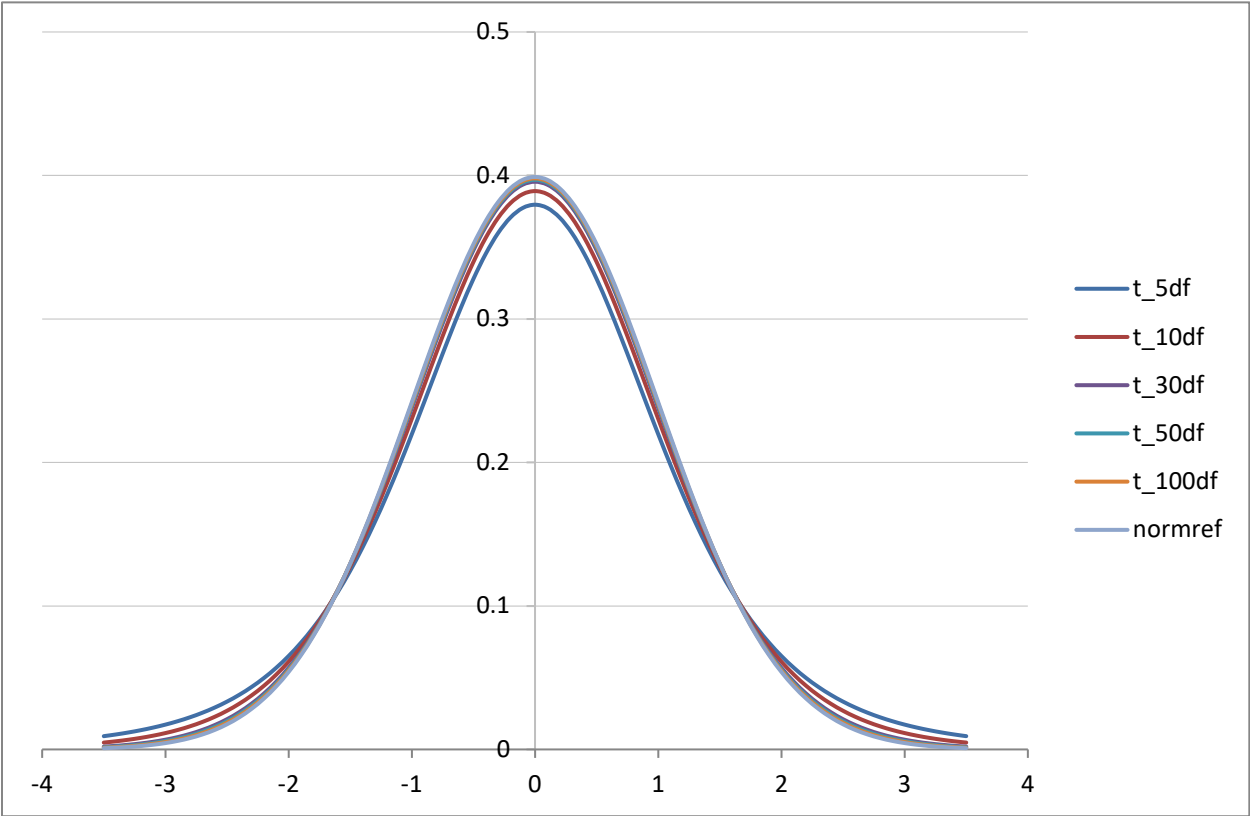
How "big" should a "big" sample be? Evidently if we can easily get more data then we should use it, but there are many cases where we need to make a decision based on limited information – there just might not be that many observations. Generally after about 30 observations is enough to justify the normal distribution. With fewer observations we use a t-distribution.

To work with t-distributions we need the concept of "Degrees of Freedom" (df). This just takes account of the fact that, to estimate the sample standard deviation, we need to first estimate the sample average, since the standard deviation uses $\sum_{i=1}^N (X_i - \bar{X})^2$. So we don't have as many "free" observations. You might remember from algebra that to solve for 2 variables you need at least two equations, three equations for three variables, etc. If we have 5 observations then we can only estimate at most five unknown variables such as the mean and standard deviation. And "degrees of freedom" counts these down.

If we have thousands of observations then we don't really need to worry. But when we have small samples and we're estimating a relatively large number of parameters, we count degrees of freedom.

The family of t-distributions with mean of zero looks basically like a Standard Normal distribution with a familiar bell shape, but with slightly fatter tails. There is a family of t-distributions with exact shape depending on the degrees of freedom; lower degrees of freedom correspond with fatter tails (more variation; more probability of seeing larger differences from zero).

This chart compares the Standard Normal PDF with the t-distributions with different degrees of freedom.



This table shows the different critical values to use in place of our good old friend 1.96:

Critical Values for t vs N

df	95%	90%	99%
5	2.57	2.02	4.03
10	2.23	1.81	3.17
20	2.09	1.72	2.85
30	2.04	1.70	2.75
50	2.01	1.68	2.68
100	1.98	1.66	2.63
Normal	1.96	1.64	2.58

The higher numbers for lower degrees of freedom mean that the confidence interval must be wider – which should make intuitive sense. With just 5 or 10 observations a 95% confidence interval should be wider than with 1000 or 10,000 observations (even beyond the familiar \sqrt{N} term in the standard error of the average).

T-tests with two samples

When we're comparing two sample averages we can make either of two assumptions: either the standard deviations are the same (even though we don't know them) or they could be different. It is more conservative to assume that they're different (i.e. don't assume that they're the same) – this makes the test less likely to reject the null.

Assuming that the standard errors are different, we compare this test statistic against a t-distribution with degrees of freedom of the minimum of either $(n_1 - 1)$ or $(n_2 - 1)$.

Sometimes we have paired data, which can give us more powerful tests.

We can test if the variances are in fact equal, but a series of hypothesis tests can give us questionable results.

Note on the t-distribution:

Talk about the t distribution always makes me thirsty. Why? It was originally called "Student's t distribution" because the author wanted to remain anonymous and referred to himself as just a student of statistics. William Gosset worked at Guinness Brewing, which had a policy against its employees publishing material based on their work – they didn't want their brewing secrets revealed! It's amusing to think that Gosset, who graduated top of his class from the one of the world's top universities in 1899, went to work at Guinness – although at the time that was a leading industrial company doing cutting-edge research. A half-century later, the brightest students from top universities would go to GM; after a century the preferred destinations would be Google or Goldman Sachs. The only thing those companies have in common is the initial G.

Other Distributions

There are other sampling distributions than the Normal Distribution and T-Distribution. There are χ^2 (Chi-Squared) Distributions (also characterized by the number of degrees of freedom); there are F-Distributions with two different degrees of freedom. For now we won't worry about these but just note that the basic procedure is the same: calculate a test statistic and compare it to a known distribution to figure out how likely it was, to see the actual value.

(On Car Talk they joked, "I once had to learn the entire Greek alphabet for a college class. I was taking a course in ... Statistics!")