Inference Notes

Shiro Kuriwaki

Harvard University
kuriwaki@g.harvard.edu

Last updated November 2, 2018

> "We are so lucky that we live in a world where the Central Limit Theorem is true"
>
> My first stats professor

(These notes are designed to accompany a social science statistics class. They emphasize important ideas and tries to connect them with verbal explanation and worked examples. However, they are not meant to be comprehensive, and they may contain my own errors, which I will fix as I find them. I rely on multiple sources for explanation and examples[1]. Thanks to the students of API-201Z (2017) for their feedback and Matt Blackwell for inspiration.)

WHERE ARE WE? WHERE ARE WE GOING?

We have now covered the world of probability: How we quantify the likelihood of complex events by extrapolating from what we know (or assume) about how simple events work. For example, you now can compute the probability that of the event that out of 10,000 fair coin flips 3,472 of them would be heads, and you now can compute the probability a Normally distributed random variable of any shape would generate outcomes in any range (like between -2 and 1). But the social phenomena we analyze are not coin flips, and the assumption that a r.v. comes from a certain distribution might turn out to be wrong. Inference is all about trying to make good guesses about the not-fully-observable world. Relying on a few proven theorems (e.g., the Central Limit Theorem), we can make pretty good guesses.

[1] DeGroot and Schervish (2012), Blitzstein and Morris (unpublished manuscript), Imai (2017), Diez, Barr, and Cetinkaya-Rundel (2015), Moore, McCabe, and Craig (2002)

CONTENTS

CHECK YOUR UNDERSTANDING

- What is an estimator? What is an estimate?
- What is a sampling distribution?
- What is a standard error?
- How do we know that an expectation of the sample mean estimator equals its true mean?
- What is the variance of a sample mean estimator? How do we know that?
- What is the intuition for the Law of Large Numbers?
- What, according to the Central Limit Theorem, converges to a Normal distribution?
- What is the rejection region?
- What is the interpretation of a confidence interval?

ESTIMATORS AND ESTIMATES

In a word where we only have data and no knowledge about the probability distribution of events, we now must construct estimators. Estimators (1) take data and (2) apply one set of manipulations to that data.

Definition 1 (Estimator)**.** An estimator is a function of observed data. The observed data are realization of random variables, so the estimator as function is itself a random variable. ■

The most important (and probably most intuitive) estimator is the sample mean estimator. Given a pile of data, the sample mean estimator adds them all up and divides by the number of observations, taking the mean. Another important estimator that we will use in the section is the sample variance estimator. This takes as its components the difference between each observation and the sample mean squared, adds those squared differences up, and divides by the number of observations minus one.

Definition 2 (Sample mean and sample variance)**.** Given a sequence of random variables (which in our case will almost always be data observations) $X_1, X_2, ... X_n$, the sample mean estimator is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

and the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

And here $\sqrt{s^2} = s$ is the sample standard deviation ■

Perhaps the only non-intuitive part here is why, in sample variance, we divide a sum of n squared difference by $n-1$ and not n. The short and non-intuitive answer is that $E(s^2)$ is exactly σ^2 for i.i.d. random variables, but $E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)$ is different from σ^2 and thus it is "biased"[2]. A perhaps more intuitive answer is that because we have used $\bar{X}_n$ as an estimate of $E(X)$ in our formula, we use up one degree of freedom in our data. The effective number of observations contained in our sum of squared differences is $n-1$, because if one knows $X_1, X_2, ...X_{n-1}$ and $\bar{X}_n$, then one automatically knows X_n.

THE LAW OF LARGE NUMBERS

As we feed more and more data to our estimator, the estimator's estimates eventually converge on the correct answer.

Theorem 1 (Law of Large Numbers (LLN))**.** *Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) draws from a distribution with expected value μ and variance σ^2. Let $\bar{X}_n$ be the sample mean estimator, $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$. Then*

[2] Derivation here: http://dawenl.github.io/files/mle_biased.pdf

$$\bar{X}_n \to \mu$$

where $\to$ is shorthand for convergence: as n (the number of random variables that comprise the sample mean) increases to infinity, the left-hand side becomes the right-hand side. We read the statement as $\bar{X}_n$ converges to μ. ■

The key contribution of the LLN is that it tells the analyst more concretely about the behavior of estimator as we collect more data.
In contrast, we actually *didn't* need the LLN to tell us that the expected value of $\bar{X}_n$ is μ. We only had to use the definition of expectation to figure that out

$$\begin{aligned}
E(\bar{X}_n) &= E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\
&= \frac{1}{n} E(X_1 + X_2 + \ldots + X_n) \\
&= \frac{1}{n}\left(E(X_1) + E(X_2) + \ldots + E(X_n)\right) \quad \because \text{expectation is linear} \\
&= \frac{1}{n}(\mu + \mu + \ldots + \mu) \quad \because \text{each random variable has an identical distribution} \\
&= \frac{n\mu}{n} = \mu
\end{aligned}$$

We never used LLN to show the above. But the LLN tells us that as we increase n to a very large number, then we have a guarantee that the value of our estimator will reach $E(\bar{X}_n)$.
It may be easier to see this with a simple example: Five coin flips. You flip five coins, each with a 0.5 probability of flipping a Heads. Then, we take the *total* number of Heads (out of 5) that occurred. That means this experiment can be expressed as a Binomial random variable:

$$X \sim \text{Binomial}(\underbrace{5}_{\text{trials}}, \underbrace{0.5}_{\text{prob.}})$$

This exercise is a familiar one, from using Binomials. The new twist is that we do this experiment several times, with the intuition that more experiments are better than one. Call this number of experiments n. Then, denote the random variable at each of those experiments $i = 1, 2, ...n$ as X_i.
Suppose we wanted an estimator that accurately predicted the expected value of heads in a given trial. We know that the answer is $5 \times 0.5 = 2.5$ analytically, but for pedagogical purposes let's numerically simulate the situation the LLN would ask for. As we increase n, the number of the 5-coin-flip experiments, how does the value of our estimator change?
Here is one sequence of $X_1, X_2, X_3, ...X_n$, for $n = 100$. That is, there are 500 coin flips, 5 at a time.

```
Xs <- rbinom(n = 100, size = 5, prob = 0.5)
```

the sample mean is

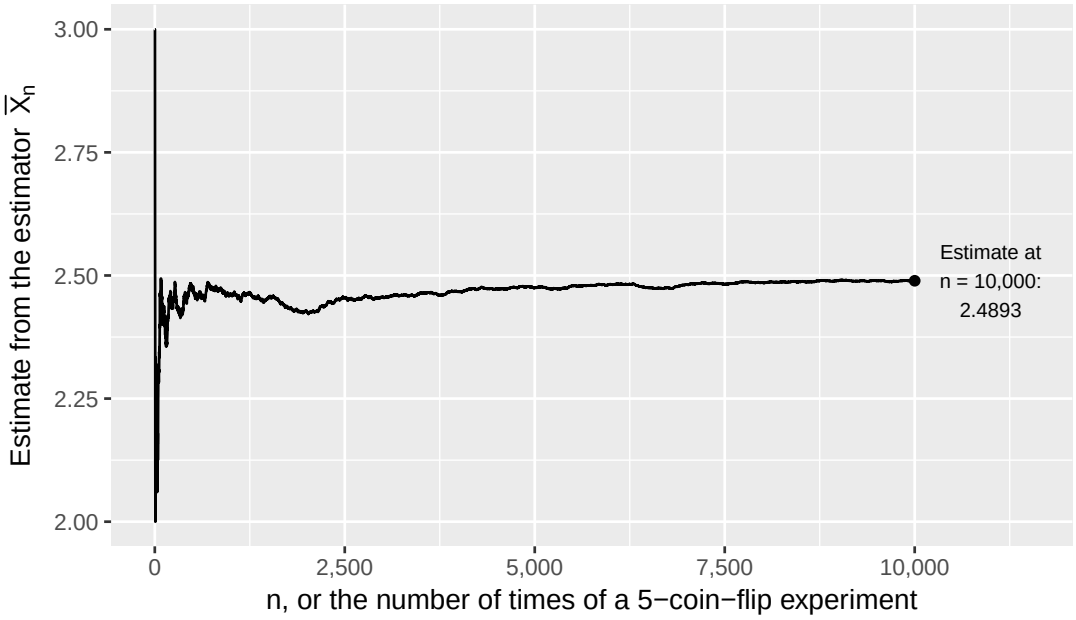


Figure 1: Law of Large Numbers at Work: Tracking the sample mean for the number of heads in 5 coin flips

```
mean(Xs)
```

```
## [1] 2.72
```

Suppose we did this for $n = 1, 2, ...10,000$, and we plotted the sample mean each time. We get Figure 1.

In this graph, see how our estimates is very noisy with small n but then stabilizes as n grows larger. Interestingly, even with 10,000 experiments our sample estimation does not get to 2.5000 exactly. But, LLN tells us that as n goes to infinity, it will.

This statement might seem trivial: We know that the expected value of a sample mean is the population mean without LLN; why do we need another law to tell us how we reach that point? Actually, the key benefit of the LLN comes when we *don't know the value of* $E(X)$. That situation is basically in any real life situation outside a computer.

In the example of the coin flip, the experiment was sufficiently sterile that we could say for certainty that $X \sim \text{Binomial}(5, 0.5)$. But for social phenomena, rarely do we know the distribution of our random variable. In this case, the LLN is telling us that whatever the sample mean approaches with large n *is* the expected value.

Here's another simulation example that highlights why the LLN is helpful. Suppose we know that X is distributed in the following complicated distribution with multiple nested mechanisms:

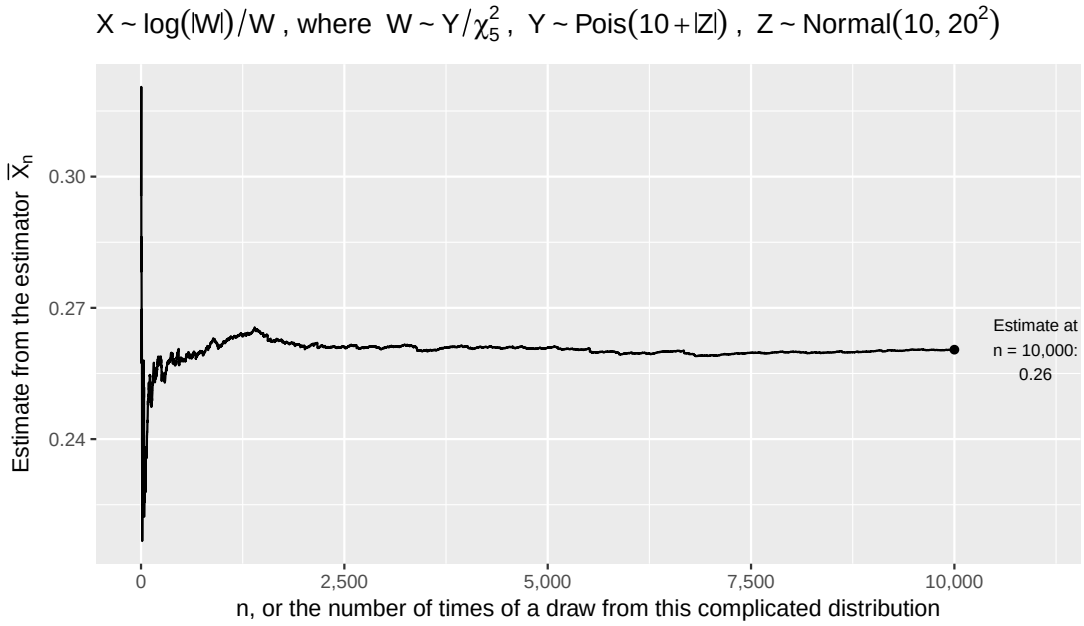


Figure 2: For a distribution whose expectation is unknown or hard to compute, the Law Large Numbers tells us that whatever the sample mean converges to is it.

$$
\begin{aligned}
Z &\sim \text{Normal}(\mu = 10, \sigma^2 = 400) \\
Y &\sim \text{Poisson}(\lambda = 10 + |Z|) \\
W &\sim \frac{Y}{\chi^2_5} \\
X &\sim \frac{\log(|W|)}{W}
\end{aligned}
$$

What is $E(X)$? Although it may be possible to apply the definitions of expectation and work through a lot of calculus, this is a lot of work. The Law of Large Numbers is useful here. It says that what ever the sample mean is after many draws from X, that is the expectation. This simulation is an example of where simulating many draws is free but the properties of the distribution is complicated. You can imagine that in the real world, collecting observations is cheap but the underlying distribution is unknown. Figure 2 shows that the sample mean converges to around 0.26 with a lot of observations.
The result of LLN justifies the use of a random number generator like the one we used here. This is called a Monte Carlo method.

THE CENTRAL LIMIT THEOREM

The LLN tells us that the estimate of the sample mean estimator will eventually reach the population mean. But it does not tell us how fast it will get there, or how noisy a given estimate for $n = 100$ is. Remember, we never get an infinite n so in practice analysts need to defend how much n gives a good enough estimate.
These shortcomings point to a need for an estimate of variance. That is, we know that

$$E(\bar{X}_n) = \mu$$

But what is

$$\text{Var}(\bar{X}_n)$$

?
Moreover, the variance is only one measure of uncertainty. We don't need the Central Limit Theorem to tell us the variance of the sample mean of Normals, but really what we'd *like* to know is the entire distribution (such as PDF or CDF) of the r.v.. The distribution tells us essentially everything about the r.v., including its expectation and variance.
Remember that expectation and variances are functions of random variables, and $\bar{X}_n$ is a random variable. Thus, $\text{Var}(\bar{X}_n)$ should give us a constant number. If $\bar{X}_n$ is a random variable, then it must have a distribution. We call this commonly used distribution as a sampling distribution (the distribution of the sampling estimator)

Definition 3 (Sampling Distribution)**.** The sampling distribution is the distribution of a estimator (that takes a sample.) ■

The Central Limit Theorem is a remarkable result that tells us that the sampling distribution (which is different depending on n), will approximate a Normal distribution as n increases.

Theorem 2 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) draws from a distribution with mean μ and variance σ^2. Let $\bar{X}_n$ be the sample mean, $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$. Then*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \textit{Normal}(0, 1)$$

where the arrow $\xrightarrow{d}$ means that the distribution of the random variable on the left-hand side approaches the distribution on the right-hand side as n increases to infinity.[3] ■

The Central Limit Theorem is beautiful because it applies to all kinds of random variables (distributions), and gives a single answer about its sampling distribution. However, in practice we do not have an infinite number of samples n, thus we may never observe this clean Normal distribution. Yet, even an approximation is useful:

---

[3] Even though the variance of $\bar{X}_n$ approaches 0 as $n \to \infty$, we need not worry because it does not move to zero "fast enough": the standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$, so the values of $\bar{X}_n$ shrink at the rate $\sqrt{n}$, smaller than the rate n.

Theorem 3 (Approximation via Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) draws from a distribution with mean μ and variance σ^2. Then the distribution of $\bar{X}_n$ is approximately*

$$\textit{Normal}\left(\underbrace{\mu}_{\textit{mean}}, \underbrace{\frac{\sigma^2}{n}}_{\textit{variance}}\right)$$

■

Just because any sequence of r.v.'s sums converge to the same thing (a Normal), that doesn't mean that the distribution of those component r.v.'s are irrelevant. Highly skewed distributions and distributions that are noisy to begin with converge slower, so n needs to be very large in order for the Normal approximation to be a good one. On the other end, if you start with a set of Normal distributions, then no convergence is necessary: We know the exact distribution of the sample mean:

$$\bar{X}_n \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{for } X_1, X_2, \ldots X_n \text{ i.i.d. Normal}(\mu, \sigma^2)$$

Let's illustrate the Central Limit Theorem with a Monte Carlo example[4]. Imagine four random variables: Binomial$(10, p = 0.9)$, Poisson$(\lambda = 2)$, Unif$(a = -1, b = 1)$, Beta$(\alpha = 0.8, \alpha = 0.8)$. The particulars of these named distributions are not important — the point is that they all look different from each other.

Suppose we observe a string of these random variables, and for a given sample size we take the sample mean. The question is, what is the distribution of the sample mean for a given sample size n?

It is possible to write out the distribution by math, but plotting a histogram is more intuitive. If we generate a lot (say, 5,000) of sample means for a given sample size n, we could take a histogram of that which approximates the distribution. Let's also consider a couple of n's: $n = 1$, $n = 5$, $n = 30$, and $n = 100$. The result is in Figure 3 and serves as our picture fort the Central Limit Theorem.

Example 1 (Measurements)**.** We want to know the prevalence of a virus in a river. The local government is interested in knowing whether there is over 15 parts per milliliter of the virus — a sign of a dangerous contamination. We cannot measure the entire river, so a virologist decides to take several samples of measurements from the river and make an inference from that sample. Let the r.v. X be the prevalence of the virus (in parts/mm).

(a) Suppose that we know the mean of X is 13 and the variance of X is 16. Now suppose the virologist takes one sample $n = 1$, call it X_1. What can we say about the distribution of this sample?

[4] Inspired from Blitzstein and Huang., p.437

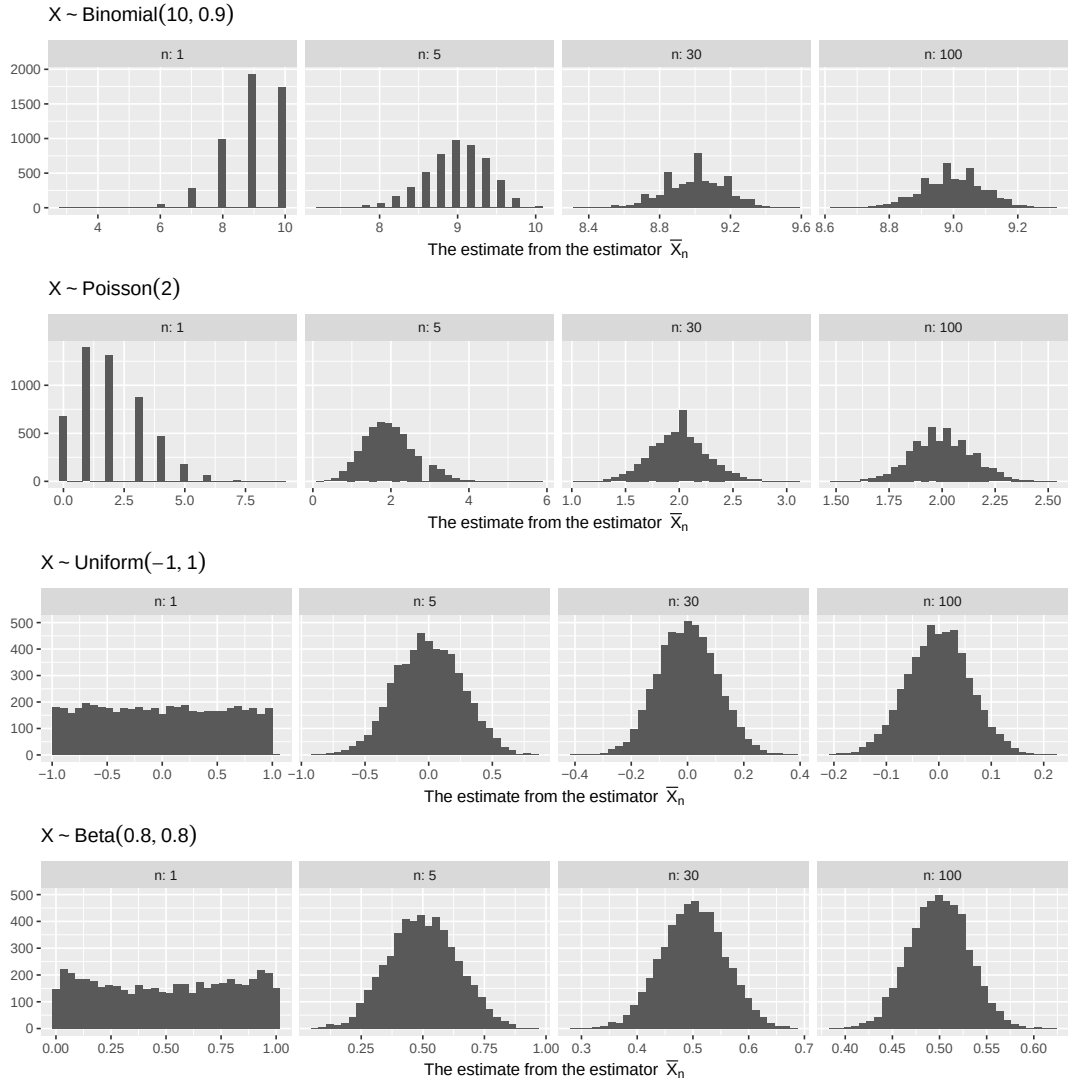


Figure 3: With Large enough n, $\bar{X}_n$ becomes approximately Normal, regardless of the distribution of each X

Answer Not much. The expected value of the one sample is 13, but the sample size may be too small use the Central Limit Theorem.

(b) Suppose that we know that X is distributed Normal, in addition to the facts in part (b). Now what do we know about X_1?
Answer Now we know pretty much everything about X_1. $X_1 \sim \text{Normal}(\mu = 13, \sigma^2 = 16)$.

(c) Given the assumptions in parts (a) and (b), what is the probability that the virologist's first measurement will be more than 15, the danger threshold?
Answer With the distribution in hand, answering $P(X_1 > 15)$ is simply using the Z-score method.

$$\begin{aligned} P(X_1 > 15) &= P\left(\frac{X_1 - 13}{4} > \frac{15 - 13}{4}\right) \\ &= P\left(Z > \frac{1}{2}\right) \\ &= 0.308 \end{aligned}$$

(d) Still with the assumptions in parts (a) and (b), what is the probability that the sample mean of 40 observations will be more than 15, the danger threshold?
Answer Using the CLT might be defensible here given the fairly large sample size. Actually, in this particular case we don't need to rely on the CLT because the sum of Normals is also Normal. In this case we have an "exact" result, where the sample mean $\bar{X}_{40}$ follows

$$\bar{X}_{40} \sim \text{Normal}\left(\mu = 13, \sigma^2 = \frac{16}{40}\right)$$

without any convergence. With this distribution in hand, we now can compute

$$\begin{aligned} P(\bar{X}_{40} > 15) &= P\left(\frac{\bar{X}_{40} - 13}{4/\sqrt{40}} > \frac{15 - 13}{4/\sqrt{40}}\right) \\ &= P(Z > -3.162) \\ &= 0.0007827011 \end{aligned}$$

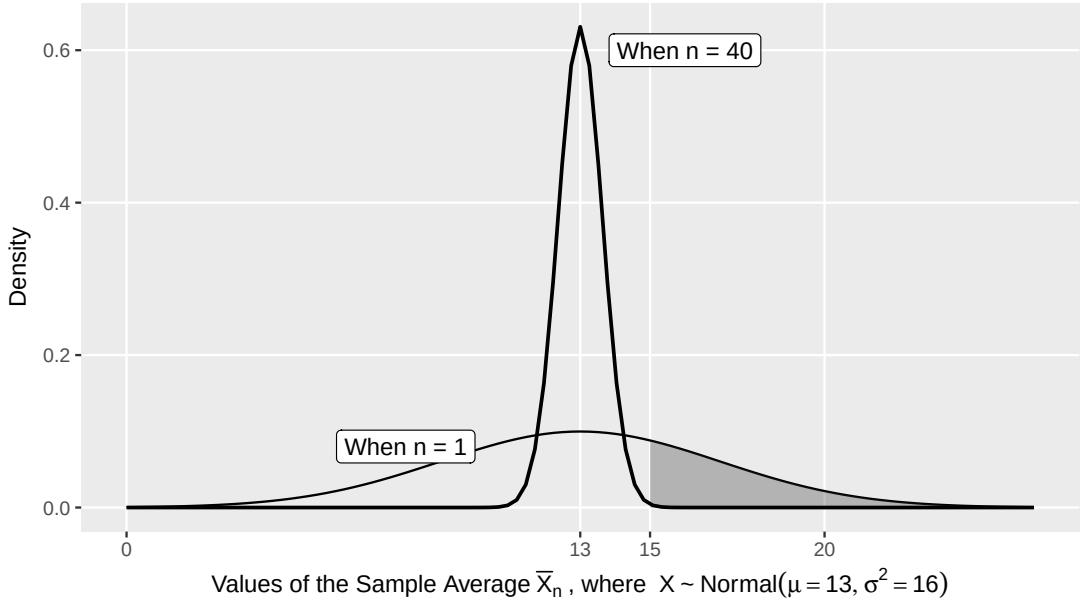
(e) Why is the quantity in part (d) much smaller than the quantity in part (c), even though we were interested in the probability that a measurement is over the same threshold and the virologist was sampling from the same population?
Answer Because we had more sample in part (d), and the true mean happened to be different from our threshold of interest. In Figure 4 we can draw the distribution of X_1 and $\bar{X}_{40}$. Both distributions are Normal, and both distributions have the same mean. Because the sampling distribution with the larger n had a smaller variance, there is much less area under the curve larger than 15 in that distribution. In other words, in distributions with smaller variance, values that are away from the mean become increasingly less likely to occur (almost by definition of variance).

(f) How would the virologist go about making inferences if she did not know the true distribution of X?

Answer Here finally the CLT comes to the rescue. For independent and most reasonable distributions, the sum (or average) of measurements ($\bar{X}_n$) will become a Normal distribution, regardless of the underlying distribution of X. To make inferences about particular events, we just need to know two more things: the mean and variance of the $\bar{X}_n$. We can make a guess about this by using the large sample we collected to estimate the mean (with the sample mean) and the variance (with the sample variance).

■

Figure 4: The distribution of the sample mean shrinks when n (the number of elements that comprise the mean) is large, which then reduces the probability of seeing $\bar{X}_n > 15$.



A function of a random variable is also a random variable, and we can use the rules of expectation and variance to make inferences on such a transformed r.v. Here is one example where we care about the sum rather than the mean of random variables:

Example 2 (Total Wait Time)**.** A bank teller serves customers standing in the queue one by one. Suppose that the service time X_i for customer i is independent and has mean 2 minutes and variance 1 minute. Let Y be the total time serving 50 customers. What is the probability that Y is between 90 minutes and 110 minutes?

Answer

We know that $\bar{X}_n \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. But, now we're not interested in $\bar{X}_n$, we're interested in

$$Y = X_1 + X_2 + ... + X_n$$

$$Z = \frac{\frac{X_1+X_2+...X_n}{n} - \mu}{\sigma/\sqrt{n}}$$

Once we create Y in this equation, the rest is a "Z-score" problem. To do this, multiply both top and bottom by n

$$Z = \frac{(X_1 + X_2 + ... + X_n) - n\mu}{\sigma\sqrt{n}}$$

This implicitly tells us that

$$E(Y) = n\mu, \;\; SD(Y) = \sigma\sqrt{n}$$

Although we could have gotten this by applying the rules of expectation and variance:

$$\begin{aligned} E(Y) &= E(X_1 + X_2 + ...X_n) \\ &= E(X_1) + E(X_2) + ...E(X_n) \\ &= nE(X_1) \\ \text{Var}(Y) &= \text{Var}(X_1 + X_2 + ...X_n) \\ &= n\text{Var}(X_1) \end{aligned}$$

In this problem, $n = 50$ and we want to know $P(90 < Y < 110)$. Now that we know the distribution of Y, we can back out:

$$\begin{aligned} P(90 < Y < 110) &= P\left(\frac{90 - n\mu}{\sigma\sqrt{n}} < Z < \frac{110 - n\mu}{\sigma\sqrt{n}}\right) \\ &= P\left(\frac{90 - 50(2)}{\sqrt{50}} < Z < \frac{110 - 50(2)}{\sqrt{50}}\right) \\ &= P(-\sqrt{2} < Z < \sqrt{2}) \\ &= P(Z < \sqrt{2}) - P(Z < -\sqrt{2}) = 0.8427 \end{aligned}$$

■

THE T-DISTRIBUTION

As an analyst, increasing n is often simply not possible. In this situation, invoking the CLT or its approximation is not quite tenable, because our approximations will be quite bad and thus it is a harder to defend the use of sample means and sample variances. The t distribution is a new distribution that summarizes the distribution of *sample mean of Normals* for any sample size, *and* without having to know the true variance σ^2.

Definition 4 (t-distribution)**.** A random variable, call it T, has the t distribution with parameter ν if it is distributed as

$$T \sim \frac{Z}{\sqrt{\chi^2_\nu/\nu}}$$

where the parameter ν is an integer and called the degrees of freedom, and χ^2_ν is a Chi-squared distribution with parameter ν.[5] ■

Why bother with this distribution here? It turns out that the sample mean estimator $\bar{X}_n$ of any size n follows a t distribution, under some conditions:
Suppose $X_1, X_2, X_3, ...X_n$ are Normal random variables (independent), each with mean μ and variance σ^2. We already know that the sample mean $\bar{X}_n$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$). Then, the standardize sample mean using the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, the statistic

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}}$$

follows a t-distribution with the degree of freedom (df) parameter $n-1$.

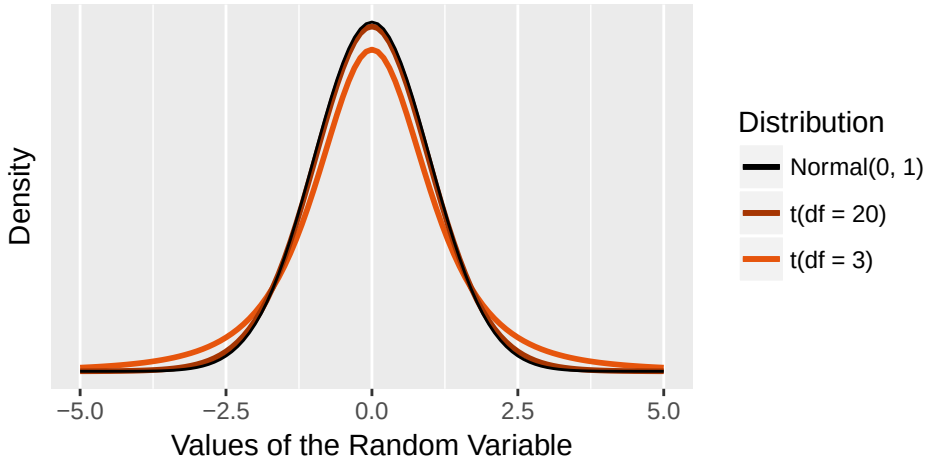
$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}} \sim t_{df=n-1}$$

The t distribution is quite similar to a Normal distribution, especially as n gets large. This makes sense, because the CLT tells us that as n gets large the sample mean of *any* random variables will approach a Normal distribution, and the t distribution is concerned about sample means. For example, Figure 5 is the distribution of t_5, t_{20}, and $Z \sim \text{Normal}(0,1)$. The larger the degrees of freedom of the t distribution, it approaches a normal.
How does the t differ from a Normal? The assumptions we don't need to make. Notice that in the definition above we actually did not make guesses about the σ^2, we just used something similar to the σ^2 (s^2) and used that directly in the formula. Notice also that we do not have the "$\xrightarrow{d}$" symbol and instead used a "$\sim$", which means that we know the distribution of the left-hand side exactly.
Why is it such a big deal that we have only one parameter instead of one? Isn't the sample variance a good enough approximation? Conceptually, leaving open an unknown σ^2 leaves the danger for a vicious cycle of guessing: With only a guess for the variance, our guess of the Normal's mean is effectively reliant on another guess, and that second guess (of variance) is reliant on another guess, and so on[6].
Of course, there is no free lunch. The cost we had to pay to say something is t is that we needed to say that the underlying X r.v. was Normally distributed. The CLT, in contrast, could deal

[5] We have not covered the χ^2 distribution yet, but roughly it is the sum of squared Normal distributions.

[6] This blog post by statistician Xiao-Li Meng has a nice motivation of for the problem. http://bulletin.imstat.org/2013/07/the-xl-files-from-t-to-t/: "Perhaps a reasonable analogy is to consider that in order to know which rank list (for example, who is the most opinionated statistician) to trust, we need to know which ranker is the most trustworthy. This would then require a rank list of the rankers. But then we need to know how trustworthy is this ranker of the rankers, leading to a Catch 22 situation (at least, in theory)."

Figure 5: A t distribution approximates a Normal distribution as its parameter (degrees of freedom) increases



with any random variable. In the defense of the t, though, if we conceive of our individual observations as sums or averages of a smaller phenomena, then it is reasonable to assume each is Normal (by CLT). For example, standardized SAT scores tend to be distributed Normal, so when calculating the distribution of the sample average of SAT scores among $n = 10$ students, it is reasonable to use the t distribution.

PRINCIPLES OF INFERENCE AND HYPOTHESES

How do we infer something about the world (the data generating process) if we never observe it? The LLN and CLT are the links between observed data and the underlying data generation process that allows us to do this.

Generically, CLT and the t distribution tells us pretty much everything we know how a sample mean $\bar{X}_n$ behaves. By this we mean we can quantify the probability of observing any range of values of the r.v. *if* we know that it is a Normal or t (with distribution). But remember that in both z-scores for the CLT and t distributions, we still had a parameter μ that was unknown. That is, in the sample mean of Normal random variables,

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

In practice, we know $\bar{X}_n$ (the mean of our data), we know s (the sample standard deviation), and we know n (the number of observations). But, we don't know μ. With an unknown μ, we still know the distribution of the left-hand side.

With only one unknown, we will no finally get to the task of inference: What is μ? Although we will never know for sure, what we know about probability and distributions will give us some probabilistic answer.

In probability, we spent many exercises asking questions of the form, "If X was distributed in a certain way, what is the probability that we observe a certain range of values of X (for example, $P(-1 < X < 1)$)?" If we assume a certain μ in the above z-score, we can give a good probabilistic answer to these questions. For example, we can make statements like: If we assume $\mu = 0$, then the probability that $\frac{\bar{X}_n - \mu}{s/\sqrt{n}} = \frac{\bar{X}_n}{s/\sqrt{n}}$ is more than 10 is 0.01 (numbers are just examples). This is equivalent to computing the following the probability that we observe the value of $\bar{X}_n$ that we observed if we assume $\mu = 0$. If that probability is low, it suggests the evidence for that assumption is weak.
Hypotheses are simply names for these assumptions.

Definition 5 (Hypothesis)**.** A hypothesis (in statistics) is a statement or assertion about a distributional property or a parameter. Depending on our research question, we conceive of a null hypothesis that we often want to reject and call it H_0. We test the null against the alternative, H_a. ■

Seeing if we can reject the null hypothesis, we typically take our data and compute the probability that we would see certain ranges of estimates assuming the null hypothesis is true. The range we often care about is "more extreme". If our null hypothesis is that the average number of days for pregnancy is 297 days, the finding that the average is actually 310 is unfavorably for the null hypothesis, as is finding that the average is actually 305. If the probability that we would observe the estimate we observed or more extreme than our estimate is, under the null hypothesis, very low, then that is sufficient grounds to reject the null hypothesis. This is exactly the motivation of the p-value, as we will see below.

ERROR RATES

We want to reject the null hypothesis when the null hypothesis is unlikely, but how do we quantify what is unlikely enough? We have measures like the following:

Definition 6 (Type I error, Type II error, and Power)**.** Type I error is the event one rejects the null hypothesis given that the null hypothesis is true (and thus should have not been rejected). Type II error is the event one does not reject the null hypothesis given that the null hypothesis was false (and thus should have been rejected).
The Type I error rate (often referred to as α) and Type II error rate (sometimes referred to as β) is the conditional probability of making Type I errors and Type II errors, respectively.
Finally, the **Power** of a test is the probability of rejecting the null hypothesis given that the null is in fact false (and thus should have been rejected). Notice

$$\text{Power } = 1 - P(\text{Type II Error})$$

because of complements:

$$\underbrace{P(\text{Reject } H_0 \mid H_a)}_{\text{Power}} = 1 - \underbrace{P(\{\text{Reject } H_0\}^c \mid H_a)}_{\text{Type II Error Rate}}$$

■

As stated these terms are only definitions of things that plausibly occur. But we will see that we can use these definitions as a kind of standard quality measure for any single test. Because the Type I and Type II error rates are probabilities, the common use is to *first set some error rate* and then back out the test criteria in terms of the data associated with that rate. Because our probability statements about our data may not be exactly right, these error rates are "nominal" – kind of like a sticker price for a test. When a user picks a test, she is implicitly agreeing to a ex ante standard of decision-making that always makes some mistakes.

Clearly we always want a lower error rate. Why not then only use tests with a Type I and Type II error rate of 0 or at least very close to 0? Considering the definitions of error immediately points out that this is not possible because there is a trade-off between Type I and Type II Error. A test that *never* rejects the null hypothesis has a Type I Error rate of 0 percent (for a truly null hypothesis, it never makes the error to reject). But it has a Type II Error rate of 100 percent — very bad — because for any time a null hypothesis is false, your test never gets to rejecting the null. The conditional probabilities are such that one error rate is not one minus the other (i.e., $\alpha + \beta \neq 1$), but usually there is a trade-off. The analyst can choose to control his test at any level of Type I Error rate or Type II error rate he wants, but cannot control both arbitrarily.

POWER ANALYSIS

Power analysis is another way to show this trade-off, but in terms that are usually more tangible and over which the researcher has some control. Statistical power is intuitively how likely it is your *detects* a true effect. This is perhaps more natural for practical use because many research questions are driven by the motivation (if implicitly) to show that some effect exists, rather than make a statement about the world without making errors.

Power analysis is process of backing out either the *sample size* or *magnitude of an effect* that is required to ensure your test nominally (i.e. "if all my assumptions are correct") has high power (e.g. 0.80). Large samples are more expensive than small samples and large-magnitude effects are harder to come by than small-magnitude effects. If your only goal is to detect an effect if it exists (after all, it would be disappointing if there was an effect in the population but the evidence wasn't strong enough to claim that it does) and not waste money, then you want to do is to data collection design that has just enough sample size to detect a hypothesized effect.

Practically, power calculations are a bit harder to compute analytically (tools[7] exist to do the computation under the hood for you). Another caveat is that you need to provide a bit more than your intended power level. You also need to provide the data-generating model (with unknown parameters of interest).

Example 3 (Power to detect lead levels[8])**.** A lab is developing a test that can measure the amount of lead, in parts per billion (ppb), in a sample of water. The test is calibrated to have $\alpha = 0.01$, that is it has a Type I Error rate of at most 0.01. Assume that repeated measurements follow a Normal distribution with unknown mean μ and known variance $\sigma^2 = (0.25)^2$. We want to run a hypothesis test of whether or not the mean μ is 6 (ppb). Thus,

---

[7] For example https://egap.shinyapps.io/Power_Calculator/

[8] Example 6.30 in Moore, McCabe, and Craig

$$H_0 : \mu = 6$$

Suppose we observe three values ($n = 3$). If our alternative hypothesis was that $\mu = 6.5$, what is the Power of this test? How does Power change as the value of the alternative hypothesis moves farther away from 6?

Answer

The Power is at the unit of a test, so we first need to identify our test. We are told that the test should have $\alpha = 0.01$; that is it should reject the null hypothesis whenever the null hypothesis is true at most 1 percent of the time. Under the null hypothesis, $\mu = 6$ so

$$\frac{\bar{X} - 6}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

We would like to reject extreme values of this distribution because those indicate a $\bar{X}$ further away from 6. What is the range of values such that z the $P(Z < -z) + P(Z > z) = 0.01$? This is 2.57. Thus, the test we were looking for is one that says reject H_0 whenever

$$\frac{\bar{X} - 6}{\sigma/\sqrt{n}} < -2.57., \text{ or } \frac{\bar{X} - 6}{\sigma/\sqrt{n}} > 2.57$$

the only random variable in this expression is $\bar{X}$. Although we know that one draw of $n = 3$ gave us $\bar{X} = 6.70$, we find a general formula and simplify in terms of $\bar{X}$, giving us:

$$\text{reject when } \bar{X} < 5.63, \text{ or } \bar{X} > 6.37$$

Only now do we consider the alternative hypothesis. Suppose that the alternative hypothesis is true, so $\mu = 6.5$ and data is generated as $X \sim \text{Normal}(6.5, (0.25)^2)$. Then, what is the probability that $\bar{X} < 5.64$, or $\bar{X} > 6.37$? This probability is by definition your power, because you've assumed the alternative hypothesis and the event that you reject a null is, in this test, equivalent to this inequality.

$$\begin{aligned} P(\bar{X} < 5.64) &= P\left(\frac{\bar{X} - 6.5}{0.25/\sqrt{3}} < \frac{5.64 - 6.5}{0.25/\sqrt{3}}\right) \\ &= P(Z < -5.96) \\ &\approx 0 \end{aligned}$$

$$\begin{aligned} P(\bar{X} > 6.37) &= P\left(\frac{\bar{X} - 6.5}{0.25/\sqrt{3}} > \frac{6.37 - 6.5}{0.25/\sqrt{3}}\right) \\ &= P(Z > -1.52) \\ &\approx 0.816 \end{aligned}$$

So the power is the sum of those two, around 0.816.

This example can be seen with the following densities in Figure 6.
The power is the total area of the shaded area. The area under the solid line but *not* shaded is the region of no rejection, which under the alternative will be a Type II Error. The key things to remember from this procedure is that the threshold is determined by the Type I Error rate. This need not be the case, but it is customary to first have a test that controls the Type I error at a certain rate, and then considers power at different alternative values and sample sizes.
Notice that our alternative hypothesis of $\mu = 6.5$ was a single point, which made it feasible to compute the Power by hand. Typically, a power *function* is one that computes the power for any range of alternative values. Everything else constant, the further the alternative value (the hypothesized true effect) is from the null, the higher the power. And also everything else constant, the higher the sample size n, the higher the power (Figure 6). The intuition is that the unobserved truth has more signal that is easier to detect with a conservative decision rule. ■

p-values

We want a cutoff of "too extreme" that has a low enough Type I error rates but high enough power. In the above example, we want a cutoff c such that, for example,

$$\alpha = P(\underbrace{|\bar{X}_n| > c}_{\text{criteria for rejection}} \mid H_0) = \underbrace{0.05}_{\text{a low probability}}$$

As we make our tests more and more stringent by increasing our threshold, our Type I error rate decreases (because we simply make it harder to reject anything). What is the smallest level of α a rejection rule could push itself to? This is the p-value:

Definition 7 (p-value)**.** A p-value of a test is formally the smallest Type I error rate α we could achieve by rejecting the null hypothesis with our estimate. In other words, an analyst who rejects a null hypothesis if and only if the p-value is at most some level α_0 is by definition capping his Type I error at less than α_0. Equivalently the p value is also the probability of seeing a estimate as or more extreme than the observed data if the null hypothesis were true (this is sometimes given as the primary definition). ■

The p-value is one type of result from a test: If someone reports a p-value of 0.001, the reader knows that the test with a Type I error rate larger than 0.001 (more tolerant of Type I error) would still have rejected the null. Thus the lower the p-value, the more certain you are of rejecting the null.
One thing that the p-value is *not* some probabilistic statement on the hypothesis itself. To make probabilistic statements, we need a distribution, and to get a distribution, we need to assume a certain hypothesis. This is equivalent to *conditioning* on a hypothesis (which is a claim about a parameter) being true. The probability of seeing the data given the hypothesis is not the same thing as, and is almost always different from, the probability of the hypothesis being true given the data[9].

[9] Bayesian analysis gets some traction on this by Bayes rule and assuming a prior.

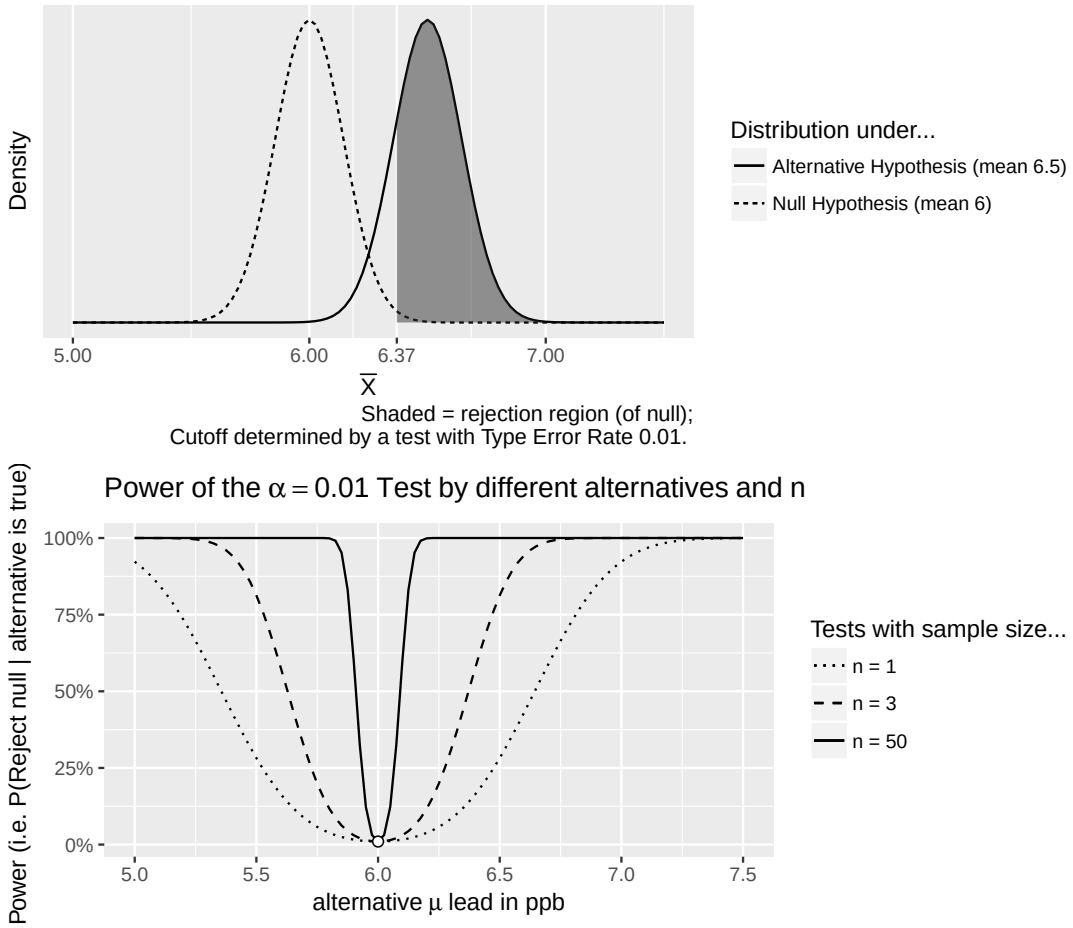


Figure 6: Visual versions of power analysis in Example 3. The top plot shows the power of a $\alpha = 0.01$ test when the alternative hypothesis is $\mu = 6.5$, in the shaded area. The bottom plot shows the power of the same $\alpha = 0.01$ test but with varying alternative hypothesis values (on the x-axis) and sample sizes (in lines), on the y-axis.

Common Error 1 (Interpretation of a p-value)**.** The p-value is not $P(H_0 \text{ is true} \mid \text{observation})$. It is rather the opposite $P(\text{observation} \mid H_0 \text{ is true})$. ■

See the Prosecutor's fallacy example in the probability notes for how mixing up these two conditional probabilities drastically misleads.

HYPOTHESIS TEST WITH MEANS

We are now in a place to look at an example of how might use a test and quantify the strength of our hypotheses. When we want to make inference on the mean parameter of a data generating process, we tend to use the t distribution because we know that the sample mean of random

variables standardized by its true mean μ and data follows a t distribution. The line of thinking sticks to the principles of inference

ONE-SAMPLE INFERENCE WITH MEANS

In some cases, we have one sample of data and we want to compare it to a pre-specified benchmark.

Example 4 (Healthcare load)**.** The number of in-patient days a nursing home accrues (in hundreds) is normally distributed[10]. Suppose that the number 200×100 in-patient days is a relevant benchmark, for example the government will subsidize all nursing homes if the in-patient load is larger than 20,000 in-patient days. We do not know the mean (μ) nor the variance (σ^2) of this normal random variable. However, we sample 18 nursing homes and find the following observations:

$$128, 281, 291, 238, 155, 148, 154, 232, 316, 96, 146, 151, 100, 213, 208, 157, 48, 214$$

Sample statistics here are

$$\bar{X}_{18} = 182, \;\; s^2 \equiv \frac{1}{18-1}\sum_{i=1}^{18}(X_i - \bar{X}_{18})^2 = (72.1)^2$$

(a) We would like to test the hypothesis that

$$H_0 : \mu = 200$$
$$H_a : \mu \neq 200$$

and maintain a Type I error rate of at most 0.05. What would be our test?

Answer

We first find the distribution of $\bar{X}_n$. We know that each X is distributed normal, so according to the t distribution definition we know that the standardized version of the sample mean is a test statistic that has the following distribution:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{18}} \sim t_{17}$$

Now, *if* we assume the null hypothesis H_0, what is the probability that we would have observed our estimate $\bar{X}_{18} = 182.17$ or an estimate as extreme as it? Under the null hypothesis, we have the distribution t_{17}. To find the cutoff, we find a threshold at which "more extreme values" would constitute at most 0.05 of the distribution. We find this by a table or

[10] comes from DeGroot and Schervish

```
qt(p = 0.05/2, df = 17, lower.tail = TRUE)
```

```
## [1] -2.109816
```

Thus, a test that rejects if $T < -2.11$ or $T > 2.11$ would have a Type I error rate or less (0.05).

(b) What is the p-value of this test?

Answer

Under the null, our data gives us

$$T = \frac{182 - 200}{72.1/\sqrt{18}} = -1.06$$

What Type I error rate would allow us to reject the null?

```
pt(q = -1.06, df = 17, lower.tail = TRUE)
```

```
## [1] 0.1519867
```

For this two-sided test, if we were comfortable with a Type I error rate of say 0.31, we could reject the null. But if the largest Type I error rate we could take was 0.29, rejecting at $T = -1.06$ would exceed our tolerance. The p-value is

```
pt(q = -1.06, df = 17, lower.tail = TRUE)*2
```

```
## [1] 0.3039734
```

(c) Suppose our hypothesis test was actually

$$H_0 : \mu \geq 200$$
$$H_a : \mu < 200$$

With the same observed values, what is the p-value of the test now?

Answer

All our test-statistic stays the same, but now values that are larger than 1.06 consistent with the null and is no longer considered "extreme". Thus we only reject when we get values of T smaller than -1.06. The probability that we do this under the null is

```
pt(q = -1.06, df = 17, lower.tail = TRUE)
```

```
## [1] 0.1519867
```

and thus our p-value is 0.1519867
Conveniently, built in programs do much of the calculation for us.
For a two-sided test with this example:

```
dat <- c(128, 281, 291, 238, 155, 148, 154, 232, 316,
         96, 146, 151, 100, 213, 208, 157, 48, 214)
t.test(dat, mu = 200, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  dat
## t = -1.0586, df = 17, p-value = 0.3046
## alternative hypothesis: true mean is not equal to 200
## 95 percent confidence interval:
##  146.1242 217.8758
## sample estimates:
## mean of x
##       182
```

If your alternative is a "less than 200" test,

```
t.test(dat, mu = 200, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  dat
## t = -1.0586, df = 17, p-value = 0.1523
## alternative hypothesis: true mean is less than 200
## 95 percent confidence interval:
##      -Inf 211.5807
## sample estimates:
## mean of x
##       182
```

shows the result of a one-sided t-test. ■

TWO-SAMPLE INFERENCES WITH MEANS

In other cases, we don't really have a pre-specified benchmark but we have data from two groups and we want to compare these means. The principle of testing whether two means are different is the same as previous examples. The only complication is that instead of using the standardized sample mean ($\bar{X}$) as the basis of our test statistic, we need to consider the *difference in means*.

Let's refer to our two groups as 1 and 2. Inconveniently, we will now have to switch around notation to define $\bar{X}_1$ as the sample mean in group 1. And μ_1 is the mean of the r.v. X in group 1, n_1 is the number of observations in group 1, etc.. Instead of looking at $\bar{X}_1$ we now compare the two sample means

$$\bar{X}_1 - \bar{X}_2$$

and if this is large enough, we reject the null hypothesis that the two groups come from a distribution with the same mean. That is we set up

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

Like the one sample test, we need to standardize to get a distribution. With quite a bit of math, we find that

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

approximates a t distribution with a degree of freedom that is approximately $\min\{n_1 - 1, n_2 - 1\}$[11].
Practically, this approximation is not ideal but good enough. Statistical programs employ a complex (and a bit better) approximation of the degree of freedom[12]

Example 5 (Age gap in Trump-Clinton support)**.** A NYT Upshot-Siena poll surveyed a sample of voters in October of 2016. This poll included 338 white women (registered voters) in the battleground state of Pennsylvania. $n_1 = 146$ of them supported Donald Trump and $n_2 = 157$ supported Hilary Clinton. We would like to know if the population age between the two groups differed. Given the data, conduct a test in difference in means of age.

Answer

Setup the variables. Let μ_1, μ_2 be the population age for Trump white women PA voters ("voters", hereafter) and Clinton voters, respectively. Let $\bar{X}_1, \bar{X}_2$ be the sample mean of age in these two groups. Now, read in the data

---

[11] If we assume the population variance in the two are equal, we can estimate that variance by the pooled sample variance statistic $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ and get an exact degree of freedom, $n_1 + n_2 - 2$. The good part about this is that we get an exact degree of freedom, the bad part is that we assume equal variance

[12] In case you are interested, the Welch modification:

$$\nu' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

```
pa <- read_csv("data/upshot_PA-white-women.csv")
```

To use a t-test, we need to assume that the underlying age of each population is a Normal distribution. We can't tell whether this is the case for sure, but we can look at the observed sample in Figure 7.

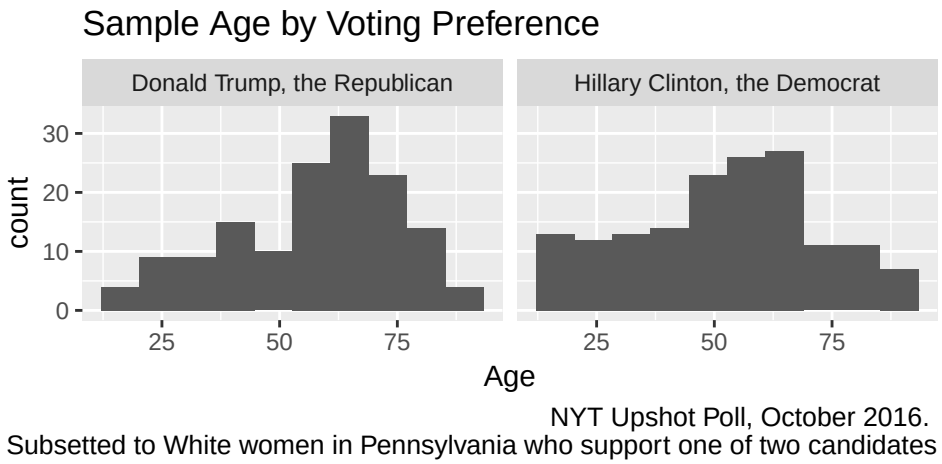


Figure 7: For a sample mean to be distributed t, components should be Normal

Given that we are comfortable making the normality assumption, we can conduct a two sample t test by using the values of age in the two groups:

```
ages_trump <- pa$file_age[pa$vt_pres_2 == "Donald Trump, the Republican"]
ages_clinton <- pa$file_age[pa$vt_pres_2 == "Hillary Clinton, the Democrat"]
```

These have the following sample statistics

```
##                n sample mean sample sd standard error
## age_trump    146    57.48630  17.71235       1.465885
## age_clinton 157    52.10191  19.33152       1.542823
```

We enter these into your statistical program's t-test function, with the null hypothesis value of the difference (0).

```
t.test(ages_trump, ages_clinton, mu = 0, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  ages_trump and ages_clinton
```

```
## t = 2.5301, df = 300.94, p-value = 0.01191
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.196405 9.572376
## sample estimates:
## mean of x mean of y
##  57.48630  52.10191
```

We need to rely on the statistical program because of the complicated degree of freedom issue, but if we had to do this analytically, we would have computed the test statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx \frac{57.5 - 52.1 - 0}{\sqrt{(1.47)^2 + (1.54)^2}} \approx 2.53$$

And to get a p-value for this two-sided test, with the (conservative) degree of freedom approximation:

```
2*pt(q = 2.53, df = min(146 - 1, 157 -1), lower.tail = FALSE)
```

```
## [1] 0.01247442
```

which is close enough to the p-value of the program's t-test. ■

HYPOTHESIS TEST WITH PROPORTIONS

When we talk about proportions instead of means, we introduce formulas that look new. But in fact, the underlying logic is all the same, and it would be counterproductive to treat them as separate types of hypotheses. I try to summarize the underlying source of the differences and their implications in this section.

THE FUNDAMENTAL LINK BETWEEN PROPORTION AND BERNOULLIS

Instead of making inference on the mean of random variable, we want to make inferences on the proportion — the fraction of the size of one set over the size of its enclosing set. A proportion thus always ranges from 0 to 1, like probability. Conceptually, a probability can be said to be a type of proportion plus the idea of experiments and sample space (see probability notes). But in many ways the two are the same.

Proportions are easier to deal with if we take them as the mean of a Bernoulli random variable, where a Bernoulli r.v.'s success is defined as the event of interest and 0 as its complement. This is always a valid thing to say – if an experiment has only two outcomes 0 and 1, then its underlying distribution is Bernoulli. The parameter of a Bernoulli, referred to as p, is the underlying population proportion.

Some ground facts from probability to remember. If $X \sim \text{Bernoulli}(\pi)$, then

$$E(X) = \pi$$
$$\text{Var}(X) = \pi(1-\pi)$$

we can derive this by noting that the PMF of a Bernoulli is $P(X = x) = \pi^x(1-\pi)^{1-x}$ where x is either 0 or 1. Then, use the weighted average definition of Expectation and Variances.
Similarly, if we have a *sequence* of Bernoulli random variables $X_1, \ldots X_n$, its sample mean $\bar{\pi}_n$ has *a* distribution with the following mean and variance:

$$\begin{aligned}
E(\bar{X}_n) &= \frac{1}{n}E(X_1 + X_2 + ...X_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\
&= \pi \\
\text{Var}(\bar{X}_n) &= \text{Var}\left(\frac{1}{n}(X_1 + X_2 + ...X_n)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) \\
&= \frac{\pi(1-\pi)}{n}
\end{aligned}$$

Let's take a moment to interpret the first statement, $E(\bar{X}_n)$. This says that the expected value of the sample mean of Bernoullis (i.e.., a series of 1s and 0s) is in fact the population proportion. Indeed, the Law of Large Numbers tells us that the sample proportion (things like 0.121, 0.213, etc..), will become closer and closer to the true mean as we increase the number of trials.
In some ways, this is the same old expectation-variance calculation. In other ways, the neat thing here is that the variance of a Bernoulli ($\pi(1-\pi)$) is a simple transformation of the mean of a Bernoulli (π). That means knowing (π) gets us both the mean and variance at the same time. In contrast, for a Normal distribution, knowing the mean (μ) essentially gave us no traction on knowing the variance (σ^2).
With these facts in hand, let's add on another layer of helpful theory – the Central Limit Theorem (CLT). The CLT tells us that the sample mean of *any* random variable approaches a normal distribution. Using the same notation where $X \sim$ Bernoulli,

$$\bar{X}_n \xrightarrow{d} \text{Normal}(E(\bar{X}_n),\ \text{Var}(\bar{X}_n))$$

To make it explicit that we are dealing with proportions, or that our underlying X r.v.s are Bernoulli, let's refer to $\bar{X}_n$ as $\widehat{\pi}$. The hat denotes that this is an estimator and a statistic computable from our data. The π denotes that we are trying to estimate the p parameter in a Bernoulli.

Definition 8 (sample proportion)**.** The sample proportion, denoted $\widehat{\pi}$ (or $\widehat{p}$), is the sample proportion of interest in your data. It is equivalent to the estimate of a sample mean from an independent and identically distributed sequence of Bernoulli random variables. As an estimator,

this simply takes the average

$$\widehat{\pi} = \frac{1}{n}(X_1 + X_2 + ...X_n)$$

where X is a Bernoulli r.v.. ■

Example 6 (Sampling distribution of proportions)**.** Suppose there is an equal number of men and women in a population of interest. You construct an estimator that randomly selects 100 people from the population and computes the proportion of women. Call this estimator $\widehat{\pi}$.
(a) Let X be the random variable which is 1 if any one person selected is a woman, 0 otherwise. What is the distribution of X?
Answer
Because this is a 0/1 variable where the population proportion of 0.5, $X \sim$ Bernoulli(0.5).
(b) What is the distribution of $\widehat{\pi}$, with $n = 100$?
Answer
If the sample size of n is deemed sufficient, we can approximate the distribution of the sample mean as a Normal r.v. by CLT. The mean is the population proportion 0.5 and the variance is a function of the mean, i.e. $\frac{0.5*(1-0.5)}{100} = 0.0025$. Thus,

$$\widehat{\pi} \sim \text{Normal}(0.5, 0.0025)$$

■

TEST STATISTICS WHEN x IS BERNOULLI

Summarizing our findings from the previous section, we can say that

$$\frac{\widehat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \xrightarrow{d} \text{Normal}(0, 1)$$

Our approximation with finite n will be, then, that for large enough n the sample mean estimator $\widehat{\pi}$ is approximately

$$\text{Normal}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Again, this is similar to the standard CLT approximation we discussed earlier. In fact, π is basically the same thing as $E(X) = \mu$. The main difference is in the variance – whereas we took σ^2 and divided by n before, now we have $\pi(1-\pi)$. This makes our lives easier because there is one fewer parameter to guess around about. We just make inferences about π and then the variance follows without any assumption. Did we need to make any new assumptions to reach to this simplification? Not really, the only change is that we said we were interested in estimating a proportion rather than a generic mean. A proportion is a special type of mean where the component outcomes are limited to 1 or 0 quantities. It then follows (without additional assumptions) that the underlying outcome is a Bernoulli r.v., and the variance estimates follow. To reformulate this in a way that is useful for hypothesis test, We now have a test statistic,

$$Z = \frac{\widehat{\pi} - \pi}{\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}}$$

which (a) follows a standard Normal distribution, and (b) allows analysts to give an estimate if we assume the parameter π (for example, under the null hypothesis). The rest is the same as a Z-test, as in the following example.

Example 7 (Effects of Medicaid on Diagnosis)**.** Researchers from the Oregon Healthcare Experiment[13] surveyed $20,745$ wait-list participants in Portland, OR, for a lottery to gain access to Medicaid. $n = 12,229$ respondents answered the survey (assume a random sample), and there among them $n_C = 6,387$ lost the lottery and $n_T = 5,842$ won the lottery. The survey asked respondents a series of health questions; treatment for depression being one of them. Suppose that the sample had the following distribution among respondents[14]:

Table: Results from Oregon Healthcare Experiment

	Lost Medicaid Lottery	Won Medicaid Lottery
Obtained treatment for depression	307	334
Did not obtain treatment for depression	6080	5508

(number of observations in cells)

Conduct a two-sided hypothesis test of proportions on whether the treatment rate is different between treatment and control.

Answer

Let π_T and π_C be the proportion of those who obtained depression treatment in the treatment and control groups, respectively. The hypotheses are then

$$H_0 : \pi_T - \pi_C = 0$$
$$H_a : \pi_T - \pi_C \neq 0$$

The sample proportions are

$$\widehat{\pi_T} = 334/(334 + 5508) = 0.0572$$
$$\widehat{\pi_C} = 307/(307 + 6080) = 0.0481$$

The estimate of our difference in sample means is

$$\widehat{\pi_T} - \widehat{\pi_C} = 0.0091$$

[13] Baicker, Katherine et al. 2013. "The Oregon Experiment — Effects of Medicaid on Clinical Outcomes." *New England Journal of Medicine* 368(18): 1713–22. http://www.nejm.org/doi/10.1056/NEJMsa1212321.

[14] These numbers are backed out of the analysis in the paper and may contain rounding error.

or 0.9 percentage points.
The variance of the difference in means is

$$\text{Var}(\widehat{\pi_T} - \widehat{\pi_C}) = \frac{\text{Var}(\widehat{\pi_T})}{n_T} + \frac{\text{Var}(\widehat{\pi_C})}{n_C}$$

and we estimate the variance of our sample proportion by using the sample proportion estimate,

$$\text{Approximate } \frac{\text{Var}(\widehat{\pi_T})}{n_T} + \frac{\text{Var}(\widehat{\pi_C})}{n_C} \text{ as } \frac{0.0572 * (1 - 0.0572)}{5842} + \frac{0.0481 * (1 - 0.0481)}{6387} = 0.0000164$$

Thus the Standard error of the difference in means is approximated by

$$SE(\widehat{\pi_T} - \widehat{\pi_C}) = \sqrt{0.0000164} = 0.00405$$

Therefore,

$$Z = \frac{0.0091 - (\pi_T - \pi_C)}{0.00405} \sim \text{Normal}(0, 1)$$

Under the null hypothesis, $\pi_T - \pi_C = 0$ so the Z-score becomes 0.0091/0.00405 = 2.2469.
The probability that a value as extreme as this occurs under the null (i.e. under the null distribution) is

```
2*pnorm(2.2469, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 0.02464642
```

which is the p-value for this Hypothesis test. ■

HYPOTHESIS TESTS WITH PAIRED DATA

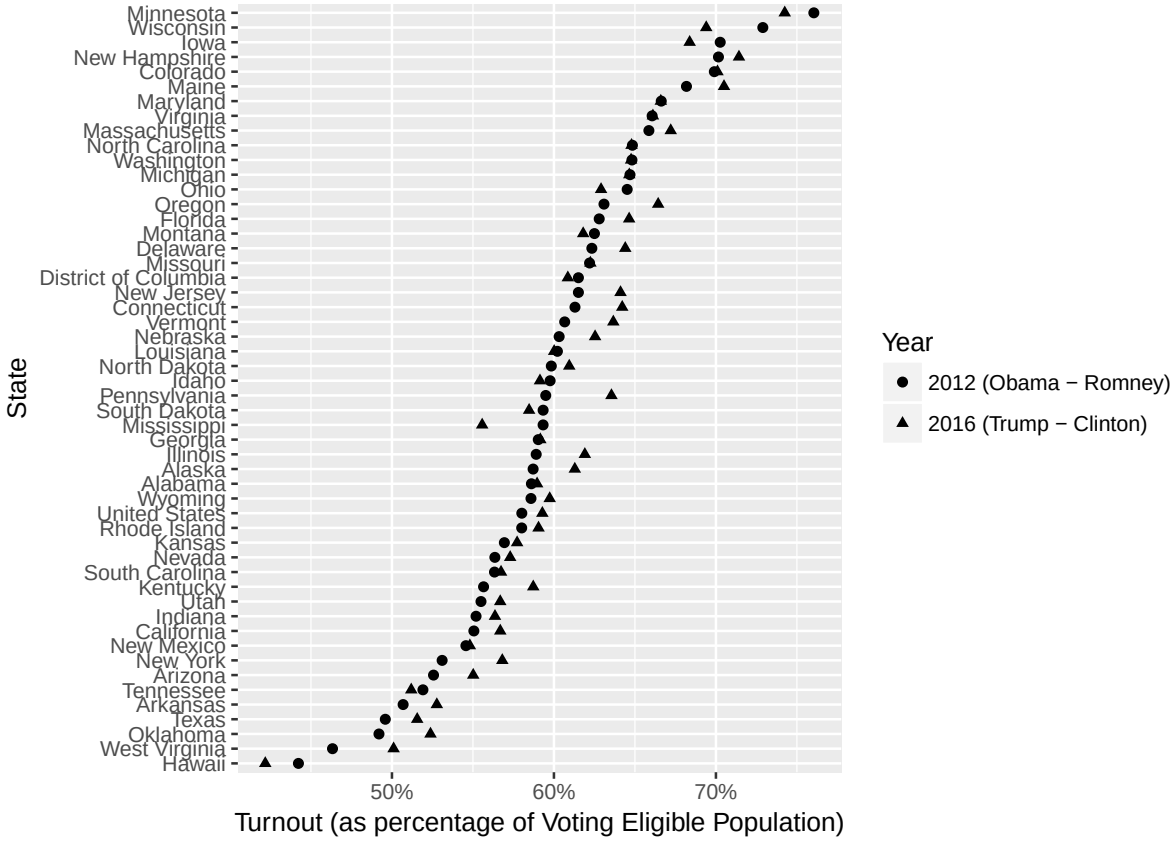
As the reasoning behind tests of proportions hopefully showed, all hypothesis tests have the same logic. When the structure of the data changed (e.g. means of continuous r.v.'s became proportions of binary r.v.'s), we would want to change our test statistic accordingly. This may lead to making more or fewer assumptions, or ending up with a slightly distribution. All these adjustments are important because we want to end up with the *correct distribution*, which the allows us to specify a rejection region that will keep a fixed Type I error rate (typically $\alpha = 0.05$). The value of a test is its α; the lower, the better. But the advertised value is only "nominal" – a test (e.g., reject if the sample mean is more than 10) can be advertised as having a Type I error rate of 0.01, but we need some sound statistical theory to tells us that said test *actually* has a Type I error rate of 0.01. The nominal and actual error rates will diverge if the conditions for the CLT or t distribution are not met.
A good example of this is when our data structure is paired. For CLT to hold, one of the conditions was that the observations are independent[15]. However, if two data come from the same

[15] Variants of the CLT will actually still hold under dependent random variables, as long as the dependency is not too large and $n \to \infty$. Nevertheless, the speed at which the sampling distribution converges to a Normal gets worth with dependent data

person, simply at different time periods, then the r.v. for each of those two data points are highly *dependent*. Ignoring the dependence and Plowing through the hypothesis test will still "work" in terms of the math, but will give you an actual Type I error is completely off from the nominal Type I error.

The fix, in this paired example case, is to difference out individual characteristics and testing the distribution of the within-pair differences. This allows us to use a normal "one-sample" t-test like as before. Because we condense two observations per pair to only one, we halve our sample size. This is generally not good for power (lower sample size $\rightsquigarrow$ lower power), but the reduction in variance (and thus the narrowing of the rejection reason) compensates.

Example 8 (Turnout, 2012 vs. 2016)**.** Did state-level turnout change between 2012 and 2016? Here are the turnout levels for all 50 states and DC:



Although this is a two-sample test, constructing the two-sample test statistic

$$t_{df=(102-1)} = \frac{\bar{X}_{2016} - \bar{X}_{2012} - (\mu_{2016} - \mu_{2012})}{\sqrt{\frac{s^2_{2016}}{n_{2016}} + \frac{s^2_{2012}}{n_{2012}}}}$$

and then trying to get the rejection region for $\alpha = 0.05$ will be invalid, at least with this limited sample size, because the observations are not independent (Minnesota's 2012 turnout is dependent with Minnesota's 2016 turnout). Invalid doesn't mean that your formula will break – it

will still give you a number — but it would no longer have a t distribution with the degrees of freedom proposed in previous sections.
Instead, if we consider the difference in turnout from 2016 to 2012 for a given state $X_d = X_{2016} - X_{2012}$, then the assumption of independence across states is a bit more plausible. Now our test is effectively one-sample,

$$t_{df=(51-1)} = \frac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}}$$

and the rejection region for this test is around $\mid t \mid > 2$.

The test-statistic is

```
diff <- t2016$trn - t2012$trn
mean(diff) / (sd(diff) / sqrt(length(diff)))
```

```
## [1] 3.541021
```

which easily passes the rejection threshold.
The t-tests using paired data show different test statistics and different degrees of freedom.

```
t.test(t2016$trn, t2012$trn, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  t2016$trn and t2012$trn
## t = 3.541, df = 51, p-value = 0.0008615
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.003827568 0.013849713
## sample estimates:
## mean of the differences
##             0.008838641
```

```
t.test(t2016$trn, t2012$trn, paired = FALSE) # Incorrect specification
```

```
##
##  Welch Two Sample t-test
##
## data:  t2016$trn and t2012$trn
## t = 0.72005, df = 101.58, p-value = 0.4731
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.01550990  0.03318718
## sample estimates:
## mean of x mean of y
## 0.6077123 0.5988736
```

■

CONFIDENCE INTERVALS

The confidence interval is another form of inference, as an alternative to the rejection of null hypotheses and p-values. Confidence intervals give a range of values that are likely to include the true value of the parameter. It turns out that these two methods – hypothesis testing and confidence intervals – are closely linked mathematically. Under some mild conditions, one could take a hypothesis test and "invert" it to get a confidence interval.

DERIVATION

The mechanics of a confidence interval directly reveals itself if we start from the central limit theorem.
Suppose a sample mean of interest, $\bar{X}_n$, has mean μ (unknown), and standard error $\sqrt{\frac{\sigma^2}{n}}$, where σ^2 is the variance of the component r.v. and n is the sample size. Then s^2 is our estimator for σ^2. Then, we know that $\bar{X}$ is approximately Normally distributed with mean μ and variance $\frac{s^2}{n}$. Another way to write that is,

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}} \sim \text{Normal}(0, 1)$$

Once we know that the left-hand side is a Normal distribution with known mean and variance, we can make any probabilistic statement about the probability that the left-hand side falls in a certain range. The easiest quantity to remember is the $\alpha = 0.05 \rightsquigarrow z = \pm 1.96$ rule, which is to say

$$P\left(-1.96 < \frac{\bar{X}_n - \mu}{s/\sqrt{n}} \leq 1.96\right) = 0.95$$

We get a confidence interval if we re-arrange the terms above, without making additional assumptions, as follows

$$
\begin{aligned}
0.95 &= P\left(-1.96 < \frac{\bar{X}_n - \mu}{s/\sqrt{n}} < 1.96\right) \\
&= P\left(-1.96\left(\frac{s}{\sqrt{n}}\right) < \bar{X}_n - \mu < 1.96\left(\frac{s}{\sqrt{n}}\right)\right) \quad \because \text{multiply both sides by } s/\sqrt{n} \\
&= P\left(-1.96\left(\frac{s}{\sqrt{n}}\right) < \mu - \bar{X}_n < 1.96\left(\frac{s}{\sqrt{n}}\right)\right) \quad \because \text{multiply both sides by -1} \\
&= P\left(\bar{X}_n - 1.96\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{X}_n + 1.96\left(\frac{s}{\sqrt{n}}\right)\right)
\end{aligned}
$$

We then define two estimators: the Lower bound (LB) and the Upper bound (UB). Together, these two numbers[16] form a confidence interval.

$$
P\left(\underbrace{\bar{X}_n - 1.96\left(\frac{s}{\sqrt{n}}\right)}_{\text{Lower Bound}} < \mu < \underbrace{\bar{X}_n + 1.96\left(\frac{s}{\sqrt{n}}\right)}_{\text{Upper Bound}}\right) = 0.95
$$

Definition 9 (Confidence Interval)**.** A confidence interval for a given parameter, such as μ, is an interval given by two statistics, LB, and UB

$$[LB, UB]$$

we compute these as

$$
\begin{aligned}
LB &= \bar{X}_n - z_{\alpha/2} \times \widehat{SE}(\bar{X}_n) \\
UB &= \bar{X}_n + z_{\alpha/2} \times \widehat{SE}(\bar{X}_n)
\end{aligned}
$$

where $z_{\alpha/2}$ is called the critical value: The quantile of the standard Normal distribution such that the $P(Z > z_{\alpha/2}) = 1 - \alpha/2$. (under the null, achieves a level-α test from a two-sided test), and SE is the standard error estimator.
In particular, a confidence interval using the value α as in the formula above is called a $100(1-\alpha)$ percent confidence interval. ■

The same definition holds if we want to use the t distribution as our approximation. Simply replace $z_{\alpha/2}$ with $t_{\alpha/2}$ with degrees of freedom $n-1$.
It is no coincidence that we used the symbol α, which is the Type I Error rate in hypothesis testing. Again, confidence intervals are a distinct exercise from hypothesis testing, but we could invert one into the other. Under mild conditions, the $100(1-\alpha)$ percent CI captures the values of the sample mean that would lead to a failed rejection of the null for a α level test.

[16] Although we speak of an "interval", we are not technically estimating a range of values. Rather, we are estimating two quantities and then letting those two numbers be the bounds of our interval.

Definition 10 (Coverage)**.** The Coverage of a confidence interval is the probability it contains the true mean, thus

$$P\left(\bar{X}_n - 1.96\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{X}_n + 1.96\left(\frac{s}{\sqrt{n}}\right)\right)$$

Nominally, this probability is the α we used to get the critical value (in this case, $\alpha = 0.05$ got us probability of 0.95). But because the CLT is an approximation and other aspects of the model may not be true, the actual coverage might differ from nominal coverage. ■

INTEPRETATION

Confidence intervals are easy to misinterpret. The key to avoiding misinterpretation is to remember the definitions about estimators. Remember that estimator is a function of the data, and we conceptualize "data" as an outcome of a random variable. Thus the estimator itself is a r.v. In the CI equation, both $\bar{X}_n$ and s are estimators (random variables). On the other hand μ is not a random variable – it is an unknown parameter. $z_{\alpha/2}$ is both fixed and known.
A common error is to fall into the temptation to interpret the exact opposite – μ as a r.v. and $\bar{X}$ as a fixed quantity.

Common Error 2 (CIs are not a probability statement about μ)**.** Probabilities are most often statements about random variables (otherwise, they are statements about events). μ is not a random variable, so it is incorrect to say that the probability that μ is between an *estimate* of $\bar{X}_n \pm z_{\alpha/2} \times \widehat{SE}(\bar{X}_n)$ is $(1-\alpha)$. Instead, the correct statement is that the probability that the estimates of LB and UB (not μ) capture μ (that is, μ lies in the interval $[LB, UB]$), is $1-\alpha$.
That is, once we draw a sample and give an estimate to the estimator $\bar{X}_n \pm z_{\alpha/2} \times \widehat{SE}(\bar{X}_n)$, then the probability statement loses its probabilistic meaning. ■

Common Error 3 (Confidence is over repeated samples)**.** The "confidence" of a CI is a statement about repeated estimations of the intervals, not about any particular interval. If a 95 percent CI is valid, then 95 percent of its CI estimates from repeated samples will contain the true proportion μ. However, we have no way of knowing *which* 95 percent contain the true proportion, and moreover we don't know *where* in those 95 percent the true proportion lies. ■

TESTING COVERAGE

It is easy to see this in this pedagogical example, where we actually know the population parameter but we pretend as if we didn't.

Example 9 (CIs from Census)**.** Suppose our population is the 10 percent U.S. Census (30,871,077 people). We want to estimate the average age of this population. Call this μ. To do so, we randomly sample $n = 5$ people from this population and ask their age. Refer to the ages of these sample observations as X_1, X_2, X_3, X_4, X_5
(a) Construct a 95 percent confidence interval estimator for μ based on this sample. Answer

$$\left[\bar{X} - 1.96\frac{s}{\sqrt{5}}, \bar{X} + 1.96\frac{s}{\sqrt{5}}\right]$$

where $\bar{X} = \frac{1}{5}(X_1 + X_2 + X_3 + X_4 + X_5)$, and $s^2 = \frac{1}{5-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
(b) Suppose our realization of the 5 observations are

$$X_1 = 6, \;\; X_2 = 68, \;\; X_3 = 22, \;\; X_4 = 40, \;\; X_5 = 30$$

what is our *estimate* of the 95 percent confidence interval?
Answer
We plug in the data to the estimator in part (a) and get

$$[12.95, \;\; 53.45]$$

(c) What is the probability that the true mean μ lies in the interval you computed?
Answer
We have no way of knowing this; the confidence interval is the probability that the LB and UB estimator capture the true mean, not the probability that the true mean lies in a certain range.
(d) If we repeated this procedure (sample $n = 5$ and compute the CI) 100 times, how many of the 100 CIs should include the true mean?
Answer
About 95 of them, by definition of confidence intervals (see simulation below).

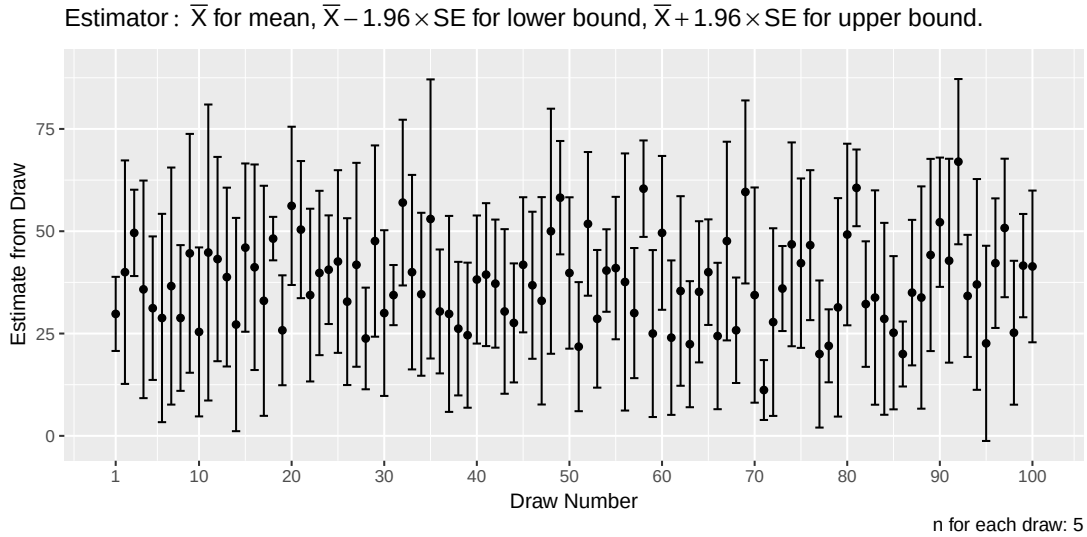
■

We can see test whether our last answer is indeed true by bringing out the actual population data, and simulating many draws of $n = 5$. We can easily do this because we have the actual population data at hand, but it's worth emphasizing why this is merely a pedagogical example (albeit a useful one). In reality, we only have a sample, not the "population". If we had the data on the population (e.g. a census), we don't need to bother sampling from it.Moreover, conceptually we only get one and only one sample. Yes, we can sample from the public many times, but for the purpose of building a confidence interval, we would rather treat all those samples as one sample.
Anyhow, in this example IPUMS has a 10 percent extract of the entire 2010 U.S. Census.

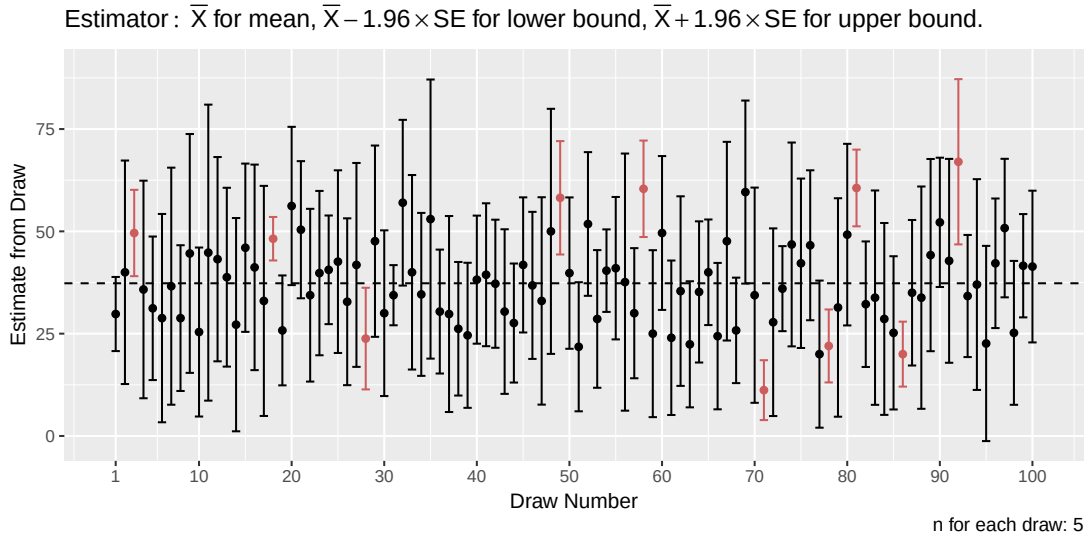
```
cen <- read_dta("usa_00010_fmt_sex-age-race.dta")
```

One sample gives us one confidence interval. But we can sample again, say, 100 times. Each sample gets us a new set of 5 people, with different ages. Take a look at the CI estimator again to verify that the CI estimates will change as well. In this example, we get the following 100 confidence intervals[17]:

[17] Code: https://gist.github.com/kuriwaki/35e9b6caf58e60c80e396735ac561d99



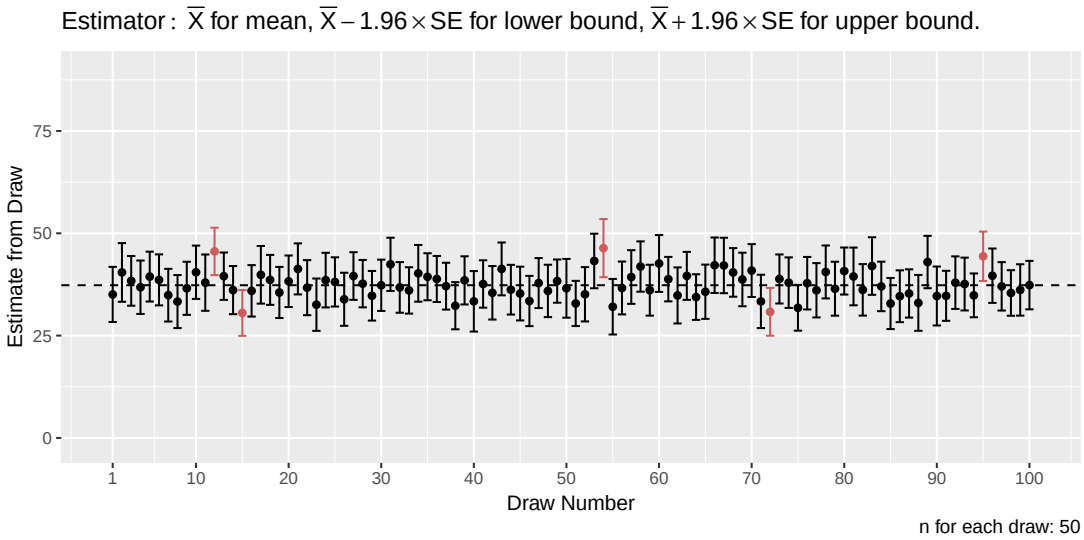
If our confidence intervals are constructed correctly, then roughly 95 out of 100 of them should include the population mean. Is this really the case? In this pedagogical example, we have the entire population so we can compute the population mean (which turns out to be 37.3). We mark this value with a line and color the confidence intervals that do not include 37.3 in them.



In this figure, only 90 out of the 100 confidence intervals cross the true mean. This is a bit lower than 95 – why is this? Although the method to compute a 95 percent confidence interval was not wrong, perhaps the assumptions underlying how we built a confidence interval were not warranted. In particular, perhaps the Central Limit Theorem could not have been applied with such small sample ($n = 5$. What's relevant is the size of the sample, not the size of the population, which is over 30 million here).

Indeed, if we increase our sample to $n = 50$ and repeat the entire procedure again, we get 95 out of 100 confidence intervals covering the population mean. Notice also that, as we increased the n,

1. The width of the confidence intervals shrink considerably – this is because the standard error of a sample estimator is decreasing as n gets large
2. The sample means became closer to the population mean – this can be explained by LLN and also the fact that the variance of the (unbiased) sample mean is decreasing in n.



Again, this is pedagogical example to verify if a confidence interval estimator does what it is supposed to do. In practice,

- We don't know the true population mean (i.e. the 37.3 in this Census example), and thus
- We don't know which of the confidence intervals contain the population mean (i.e. the red intervals in this example), and moreover
- We only get effectively one draw from the sample, not 100, and because of this,
- We cannot be completely certain whether our nominal coverage is true.

ANOVA

We introduce a different test called analysis of variance (ANOVA), to run a global, or omnibus test of difference in means. Yet another test! While ANOVA introduces a different test statistic and different null distribution to evaluate for good reason, at this point it can be helpful to emphasize the shared logic across tests.

MOTIVATION

We can think of ANOVA as a more general form of the two-sample difference in means test. Instead of testing only two means,

$$H_0 : \mu_1 = \mu_2$$

What is a test that can handle multiple groups

$$H_0 : \mu_1 = \mu_2 .... = \mu_k$$

given this problem, let's recall the general procedure we have used in the past tests.

1. First, we come up with a summary of data. This is called the test statistic, and it is a function only of the data.
2. We use our knowledge about CLT and other distributions to identify the *distribution* of the test statistic.
3. We further specify the distribution and its parameters by taking the claims of the null hypothesis as given.
4. With the distribution in hand, we compute the probability that our estimate (or an estimate more extreme) of the test statistic would have occurred under the null hypothesis (the p-value).

The test statistic in the two sample case was essentially a difference in means,

$$\bar{X}_1 - \bar{X}_2$$

divided by functions of variance and sample size in order to match it up with the CLT. We then found that this normalized random variable was approximately distributed Normal, or a t distribution.
We cannot use the same type of sample mean difference when there are 3 or more groups. The first approach might be to take all the pairwise differences and add them together. This does not give us an estimator with the property we want, however. We want differences in sample means (which can each be positive, negative, or zero) to *add up*, rather than *cancel out*. Thus, a natural way to aggregate differences will be to square them and sum the squares:

$$\text{Squared Sum (of Differences) Between Groups} = \sum_{i=1}^{k} n_i(\bar{X}_i - \overline{X})^2$$

where the index $i = 1, 2, ...k$ counts the different groups in the data, $\bar{X}_i$ shows the sample mean in group $\bar{X}_i$, and $\overline{X}$ is the mean across all observations regardless of group. Because considering all $\binom{k}{2}$ pairs is redundant, we take each of the k sample means and compare them against the global mean. We multiply by n_i to maintain the information that this deviation measure results from $\sum_{i=1}^{k} n_i$ observations.
The intuition of the Squared Sum between groups is that the more a particular group (as measured by its average) differs from the entire sample, the more it should count against the null hypothesis.

Of course, whether or not this difference (squared then summed) is big or small is relative measure. The baseline we compare to is the overall noisiness of the sample means themselves. Within each group i, suppose there are \$j = \$, $1, 2, ...n_j$ observations. Then the variability within group i is another sum of squares:

$$\text{Squared Sum (of Differences) within a Group} = \sum_{j=1}^{n_j}(X_{ij} - \bar{X}_i)^2$$

And we can later add k of these together.
Finally, the total variance of the sample is each observation compared against the grand mean:

$$\text{Total sum of Squares} : \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2$$

This is the total amount of variation around the mean in the sample. Now, it turns out that the math works out nicely to show that the Total Sum of Squares can be *decomposed* exactly into the between-group sum of squares and the within-group sum of squares:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2}_{\text{Total Sum of Squares}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{X}_i - \overline{X})^2}_{\text{Between Group Sum of Squares}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_j}(X_{ij} - \bar{X}_i)^2}_{\text{Sum of Within Group Sum of Squares}}$$

This variance decomposition applies to all sets of data and will be useful to interpret regressions, when the group means are replaced with fitted values of regression coefficients. For now, the fact that we can decompose the total variance into between and within group variations leads to the idea that, with any set of data, *the ratio* of the between group variances (normalized by the number of groups) and the sum of the within group variances (normalized by the number of observations) shows the relative weight each of the two decomposed variances hold. And, further, it is worth noting that the more this ratio increases (i.e. the numerator is larger), the variation between groups outweighs the average variation we would expect from within group. At this point we have been interchanging "squared sums" with "variances". This holds because roughly variance is the squared sum divided by the number of observations. Moreover, this quantity, the sum of squares within group i,

$$\sum_{j=1}^{n_j}(X_{ij} - \bar{X}_i)^2$$

becomes the sample standard deviation estimator S^2 when divided by $n_j - 1$.
Thus, when we normalize by the degree of freedoms associated with our s^2 estimates, the ratio becomes, exactly stated,

$$\text{F-statistic} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \overline{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_j}(X_{ij} - \bar{X}_i)^2/(\sum_{i=1}^{k} n_i - k)}$$

and now this is a ratio of two types of variances. This is the reason why the procedure will be called "Analysis of Variances" even though it is motivated by analyzing difference in means.

Mathematically, the way we keep score of the size of difference in means is very similar to how we compute variances.
Where has this led us? All these sums turn out to be useful to us, because, once set up in this way, probability theory tells us that this statistic follows a F distribution in large samples. The F-distribution is a ratio of two normalized chi-squared distributions. A chi-squared distribution is a sum of squared standard Normals, and has parallels with the squaring we repeatedly performed to make sure variances add up.

Definition 11 (F-distribution and Chi-squared distributions)**.** A F distribution is a continuous distribution with two degrees of freedom. Its PDF is too complicated to write here, but it is the ratio of two Chi-squared statistics each divided by its degree of freedom. Because a chi-squared is always Positive, a F-distribution also has positive outcomes.

$$F_{m,n} \sim \frac{\chi^2_m/m}{\chi^2_n/n}$$

where χ^2_n is a Chi-squared distribution with degree of freedom. *Its* distribution can be in turn represented as a sum of squared standard Normals,

$$\chi^2_n \sim Z_1^2 + Z_2^2 + ... + Z_n^2$$

where $Z_1, Z_2, ...Z_n$ is a sequence of standard Normals (Normal$(0, 1)$). ■

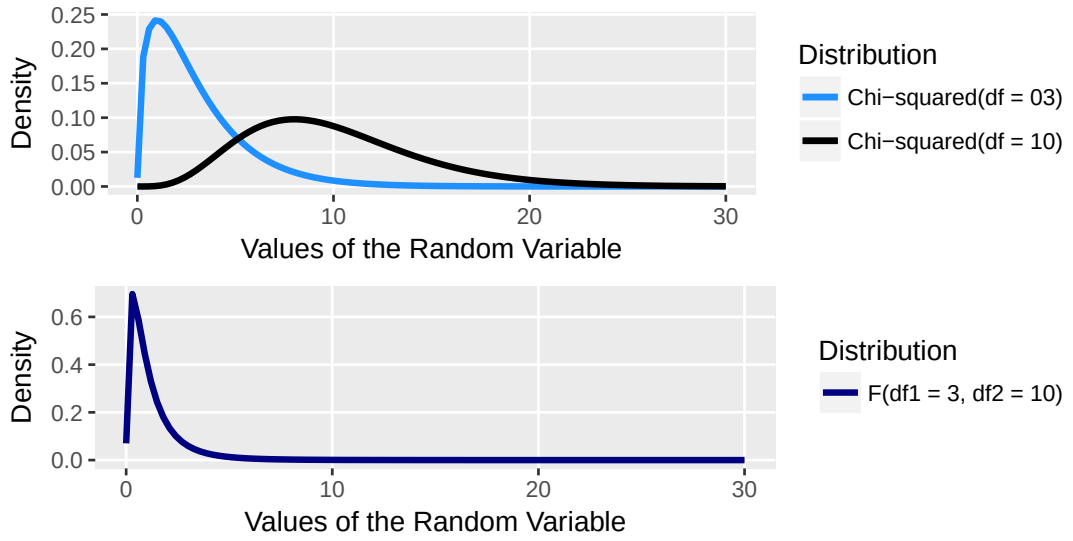
Figure 8 some examples of the shape of the distribution with certain degree of freedom values. Like the normal, larger values become increasingly unlikely. The key difference between the Normal distribution is that both the Chi-squared and F-distributions are always positive, and are (thus) not symmetric.
The key takeaway here is that at its core, both F-distributions and chi-squared distributions are composed of Normal distributions. And why do we like Normal distributions? Because from the Central Limit Theorem, we know that sums and means of all kinds of random variables will increasingly become a Normal distribution. Again, the general strategy of inference is to use our data (whose underlying distribution is **unknown**) to come up with a statistic whose distribution **is known**. Then, we can back out (or infer) the properties of the underlying unknown distribution.
Interestingly, we could see how this F-statistic is a more general version of the t statistic of difference in means. In a simple case of pooled variance and equal sample size,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

Now, anticipating that we need to keep track of multiple differences instead of just one difference, square this value and we get

$$t^2 = \frac{\frac{n}{2}(\bar{X}_1 - \bar{X}_2)^2}{s_p^2}$$

Figure 8: Examples of F and χ^2 distributions. The particular examples also illustrate how $F_{n,m} = \chi^2_n/\chi^2_m$.



Which has the same interpretation as the F-statistic: between-groups in numerator, within-groups in denominator.

To summarize, when we care about comparing arbitrarily many differences, then it turns out that the F-statistic defined above is appropriate for inference because it both (a) behaves the way we want it to (i.e. gets higher values when groups are different), and (b) has a known (approximate) distribution regardless of the distribution of X.

SIMULATION EXAMPLE ON WHY VARIANCE MATTERS

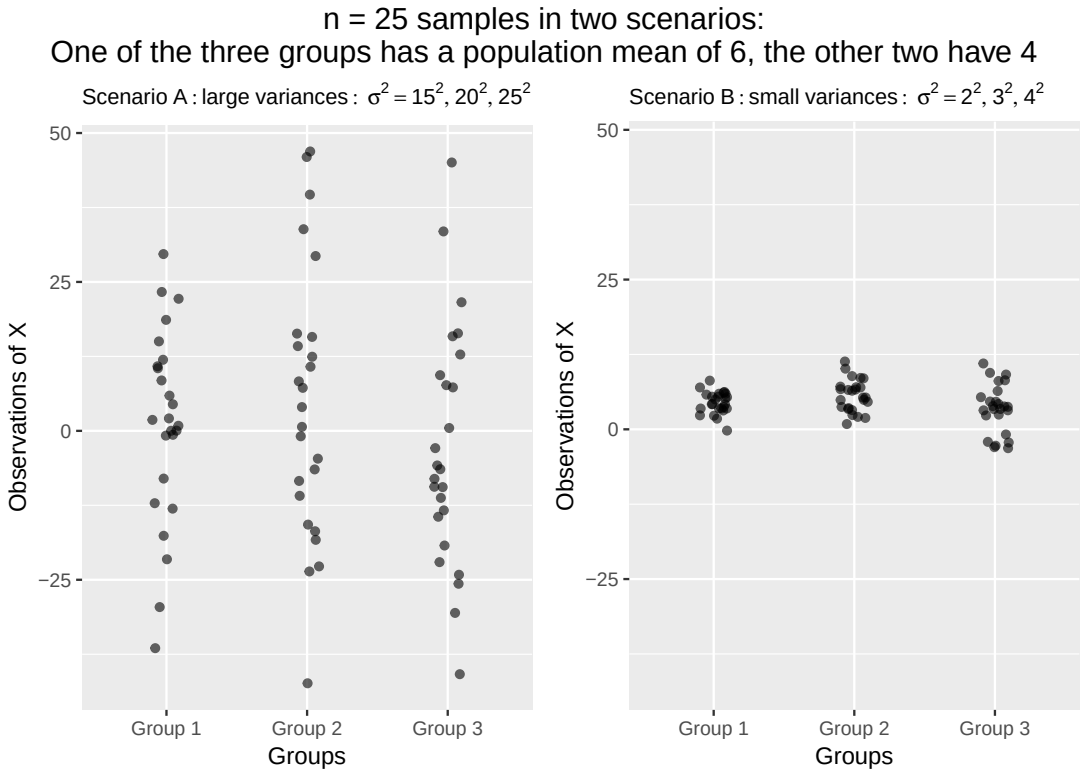
We saw in the previous section how variance becomes a key part of a global difference in means test. An example from simulated data shows another intuition for why variance plays a key role in inference, and shows ANOVA in action.

Consider an example for a continuous outcome variable X. There are three groups, each of $n = 25$. In two of the three groups, X is a Normal distribution with mean 4, whereas the X in another group comes from a Normal with mean 6. Can you tell from the sample of data that the means were not the same? Figure 9 shows two scenarios.

In the first (left-hand) scenario, the underlying variances of the Normal distributions are large. Even though the mean is actually not all equivalent, it is hard to tell that from the particular sample we have partly because the data is so scattered. The lesson here is that whether or not a difference of 2 units (mean 6 - mean 4) is meaningfully large to be detectable depends in part on the variability of the data.

In contrast, the second (right-hand) scenario keeps the means of the three groups the same, but shrinks the variance and puts the sample on the same scale. Now, can you tell whether the underlying means are different? Most people can now detect some difference, and perhaps can

Figure 9: "Large" difference in means depends on variance. Simulation example



tell which of the three groups have a different mean[18].
This visually apparent difference shows up when we do an actual ANOVA analysis.

```
lm1 <- lm(X ~ group, df_long1)
anova(lm1)
```

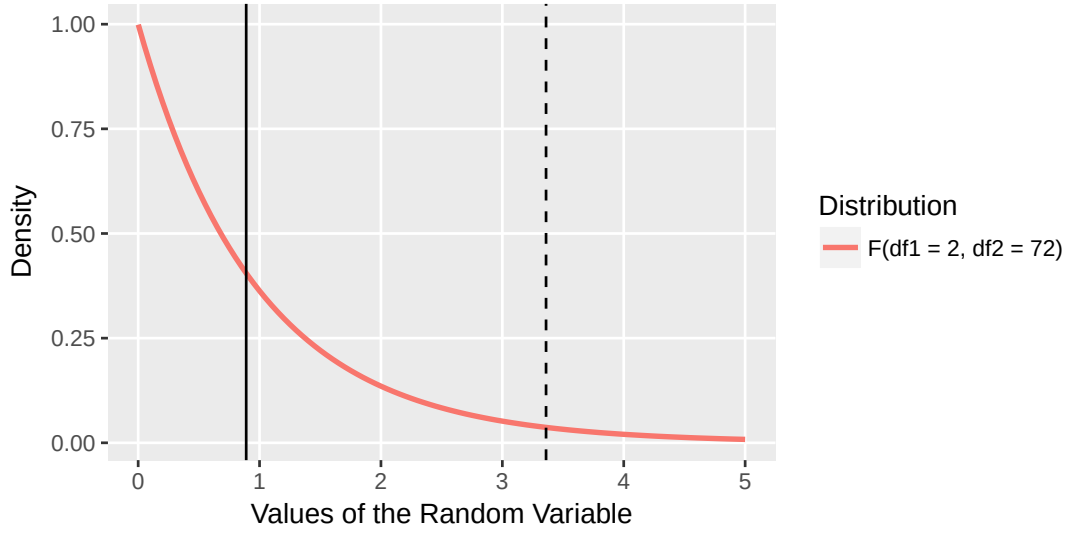
```
## Analysis of Variance Table
##
## Response: X
##           Df  Sum Sq Mean Sq F value Pr(>F)
## group      2   708.1  354.04  0.8899 0.4152
## Residuals 72 28643.6  397.83
```

```
lm2 <- lm(X ~ group, df_long2)
anova(lm2)
```

```
## Analysis of Variance Table
```

[18] Group 2 has $\mu = 6$

Figure 10: p-values from a ANOVA is the area under the curve to the right (more extreme) of the respective F-statistic




```
##
## Response: X
##           Df Sum Sq Mean Sq F value  Pr(>F)
## group      2  60.00 30.0001  3.3601 0.04026 *
## Residuals 72 642.85  8.9284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that in both ANOVA outputs, the degrees of freedom is the same because we have the same amount of observations ($n_1 = n_2 = n_3 = 25$) and groups ($k = 3$). Yet both types of sums of squares are much smaller in the second case, where the underlying variance is much smaller. Accordingly, the mean sum of squares is smaller in the second case as well. Moreover, the F-statistic, which is simply the between group sums of square divided by the within group sums of squares, becomes higher in the second case, is higher.
We can see in Figure 10 that the p-values – the results of our hypothesis test – corresponds to the area under the curve above the F-statistic. That is, under the null hypothesis our F-statistic would be distributed $F_{2,72}$, and the probability that a value of F that is more extreme than 3.360069 is the p-value (and the area under the curve to the right of 3.360069).

COMPUTING ANOVA FROM SUMMARY STATISTICS

We can work out exactly how the program computes each of the squared sums with the following example[19].

[19] Moore, McCabe, Craig, Example 12.3

Example 10 (Likability Ratings)**.** A study[20] showed a total of 134 students fake Facebook profiles and asked the likability rating (on a score from 1 to 7) of the person in the profile. The study randomly assigned students to five groups, varying only the number of Facebook friends that profile had. The results were below. What would an ANOVA tell you about the null hypothesis of equal means?

i:	1 102 friends	2 302 friends	3 502 friends	4 702 friends	5 902 friends
$\bar{X}_i$	3.82	4.88	4.56	4.41	3.99
s_i	1.00	0.85	1.07	1.43	1.02
n_i	24	33	26	30	21

Answer

Let's continue to use the notation X_{ij} to refer to the jth observation in the ith group (out of k total groups, in this case $k = 5$). Also let n with no subscript be the total number of observations ($n = 134$).

The global mean $\overline{X}$ would be the sum of all values of X_{ij} divided by the total number of observations n. Noticing that in general, the mean multiplied by the number of observations gives you the raw sum of observations,

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_j} X_{ij} \\
&= \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_i \\
&= \frac{1}{134}(3.82 * 24 + 4.88 * 33 + ...3.99 * 21) \\
&= 4.38
\end{aligned}$$

The between sum of squares is basically adding up differences in the group mean and the global mean,

$$\begin{aligned}
SS_B &= \sum_{i=1}^{k} n_i(\bar{X}_i - \overline{X})^2 \\
&= 24 * (3.82 - 4.38)^2 + 33 * (4.88 - 4.38)^2 + ...21 * (3.99 - 4.38)^2 \\
&= 19.84
\end{aligned}$$

[20] Tong, S. T. et. al. (2008), Too Much of a Good Thing? The Relationship Between Number of Friends and Interpersonal Impressions on Facebook. *Journal of Computer-Mediated Communication*, 13: 531–549

Backing out the within sum of squares is a bit trickier, but using the fact that

$$s_i = \sqrt{\frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_j - \bar{X}_j)^2}$$

we can do

$$\begin{aligned} SS_W &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_j - \bar{X}_i)^2 \\ &= \sum_{i=1}^{k} s_i^2(n_i - 1) \\ &= 1.00^2(24-1) + 0.85^2(33-1) + ...1.02^2(21-1) \\ &= 154.85 \end{aligned}$$

The rest is fairly straightforward. The degree of freedom for SS_W is $134 - 5 = 129$, and the degree of freedom for SS_B is $5 - 1 = 4$. Thus the mean sum of squares are

$$\begin{aligned} MS_B &= 19.84/4 = 4.96 \\ MS_W &= 154.85/129 = 1.20 \end{aligned}$$

And thus the F-stat is

$$F = \frac{4.96}{1.20} = 4.13$$

Under the null distribution, this statistic should be distributed $F_{4,129}$. The probability that we get a value of 4.13 or more is

```
pf(q = 4.13, df1 = 4, df2 = 129, lower.tail = FALSE)
```

```
## [1] 0.00350563
```

which corresponds to the p-value.

■

CHI-SQUARE TESTS

Finally, Chi-squared tests are tests designed to detect differences in discrete distributions, typically count data. Typically, we would be interested in seeing whether the observed breakdown of our sample into certain categories is as expected from a given baseline.
A discrete set of categorizations gives us data that is in counts. To test if the differences in observed counts is different from some expected value (if the null were true), we again come

up with a test statistic that square the differences between expected counts (so that they don't cancel out).
In order to account for the different size of groups, we divide by the expected amount of counts. This is a good statistic not only because it highlights differences in proportions but also because we know its distribution: A chi-squared distribution.

Theorem 4 (Standardized sum of counts converge to Chi-squared distribution)**.** *Let the null hypothesis of a population proportion be $\pi_1, \pi_2, ..\pi_k$. Then, for a total sample of size n, and with observed counts $n_1^{obs}, n_2^{obs}, ...n_k^{obs}$,*

$$\frac{(\textit{Observed Counts of } i - \textit{Exepcted Counts of } i)^2}{\textit{Exepcted Counts of } i} = \frac{(n_i^{obs} - n\pi_i)^2}{n\pi_i} \xrightarrow{d} Z^2$$

where Z is a standard Normal. It follows that the sum of these is a Chi-squared, because a Chi-square is defined as a sum of squared standard Normals.

$$\sum_{i=1}^{k} \frac{(\textit{Observed Counts of } i - \textit{Exepcted Counts of } i)^2}{\textit{Exepcted Counts of } i} = \sum_{i=1}^{k} \frac{(n_i^{obs} - n\pi_i)^2}{n\pi_i} \xrightarrow{d} \chi^2_{k-1}$$

■

Proving that this test statistic indeed converges to a Chi-squared[21] is actually quite involved and beyond the scope of this text. The main complication is because for fixed n, whether or not an observation falls into group i is *dependent* with falling into another group. Thus we cannot use "nice" i.i.d. distributions, but have to work with a covariance matrix. Despite the complicated proof, the intuition is still that of "standardizing" the data by differencing and dividing by the values that one would expect. Standardizing leads to a Normal distribution with mean 0 and variance 1.

Common Error 4 (Normalized *counts* converge to Chi-squared, not proportions)**.** Although

$$\frac{(\text{Observed Counts of } i - \text{Exepcted Counts of } i)^2}{\text{Exepcted Counts of } i} = \frac{(n_i^{\text{obs}} - n\pi_i)^2}{n\pi_i} \xrightarrow{d} Z^2$$

it is *not* true that

$$\frac{(\text{Observed Proportion of } i - \text{Exepcted Proportion of } i)^2}{\text{Exepcted Proportion of } i} = \frac{(\pi_i^{\text{obs}} - \pi_i)^2}{\pi_i}$$

converges to a Chi-squared test. The quantity of proportions is something quite different. The CLT is very general, but unfortunately this form is not exactly the CLT form as in Theorem 2. Using this incorrect test statistic will give you "a" number for the test-statistic, so it is tempting to get your p-value that way. But it will be incorrect to line that number up against a χ^2 distribution table and conclude that the p-value is $P(\chi^2_{df} > \text{test-stat})$ ■

[21] See for example http://sites.stat.psu.edu/~drh20/asymp/lectures/p175to184.pdf

CHI-SQUARED TESTS FOR GOODNESS OF FIT (ONE-WAY)

The Chi-squared test is also referred to as a goodness of fit test, because one could set the baseline as any kind of model. The observed, discrete breakdown of sample data is then compared against the predicted breakdown to quantify the goodness of fit between the test and data.

Example 11 (Composition of juries)**.** Consider the racial composition of a random sample of 275 jurors. We would like to know if this composition (first row) is different enough from the population racial composition (second row) to be not due to chance.

Race:	White	Black	Hispanic	Other	Total
Observed	205	26	25	19	275
Population Proportion	0.72	0.08	0.12	0.08	1.00
Expected if H_0	198	22	33	22	275

Answer

The third row of expected observations is generated simply by multiplying the sample size by the population proportion (e.g. $205 \times 0.72 = 198$).
Then, for each category we make a test-statistic, we do so by standardize the proportions for each group i:

$$Z_i = \frac{(\text{Observed Count} - \text{Expected Count under Null})^2}{\text{Expected Count under Null}}$$

Remember we use this formula because CLT tells us that this statistic, suggestively labelled "Z", will be distributed as a standardized Normal random variable.
And we need to just have one number to summarize the test-statistic, so we add the four Z_i's together:

$$\begin{aligned} Z &= \frac{(205-198)^2}{198} + \frac{(26-22)^2}{22} + \frac{(25-33)^2}{33} + \frac{(19-22)^2}{22} \\ &= 0.247 + 0.727 + 1.939 + 0.409 \\ &= 3.323 \end{aligned}$$

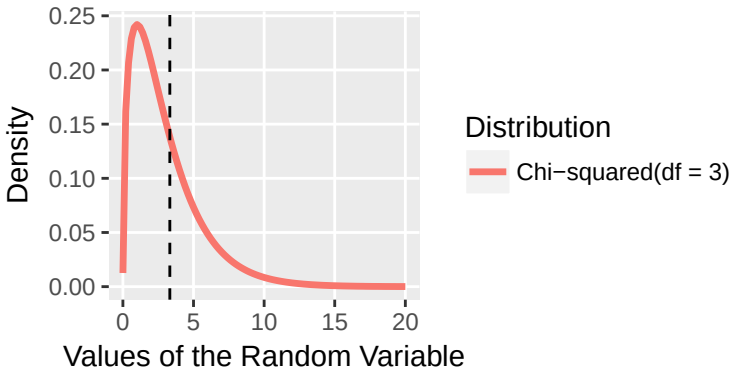
And, we know that this has a distribution of χ^2_{4-1}.
The p-value is the area to the right of 3.23 in a χ^2_3, pictured below.
And the p-value is directly

```
pchisq(q = 3.323, df = 3, lower.tail = FALSE)
```

```
## [1] 0.3444543
```

Thus, we would fail to reject the null hypothesis of representatives at conventional α levels of 0.05 or 0.10.

■



CHI-SQUARED TESTS FOR INDEPENDENCE (TWO-WAY)

In practice Chi-squared tests come up most frequently as a test to infer, from observed data, whether or not two variables are independent from each other. The assumption of two variables being independent enables a host of causal inferences, so being able to infer it from observed data is critical. In the earlier probability section, we learned that if two events A And B were independent,

$$P(A \cap B) = P(A)P(B)$$

Using random variables, if r.v.'s X and Y are independent, that means their realizations (events) satisfy

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

In a Chi-squared test of independence, we will set the relationship expressed in equation **??** as the null hypothesis. Observing evidence that contradicts this claim (as expressed by a small p-value) will be a sign that the two variables are dependent with each other.
Often times our data are counts, as we see in tables and cross-tabs. As long as we have a pair of observed counts of an event and the expected count if the same event under some hypothesis, then Chi-squared tests can be deployed pretty flexibly. In a test of Chi-squared tests of independence, there is nothing inherent about Chi-squared statistics that says "independence", but because have clear expectations about the frequency of discrete events *if* their r.v.'s are independent, we can use a Chi-squared test to quantify how likely the data would have been observed under the null hypothesis.
The actual steps for a Chi-squared test is best seen through an actual example.

Example 12 (Different Perceptions)**.** A survey asked 189 respondents their income and whether or not their financial status is worse, same, or better compared to the past two years. The Table below is a cross-tab of how many individuals ended up in each cell. From these observed counts, would you infer that one's income class and their perception of their change in finance status is independent?

	Personal financial status			
Income range	Worse	Same	Better	Total
Under $20,000	20	15	12	47
$20,000 - $35,000	24	27	32	83
Over $35,000	14	22	23	59
Total	58	64	67	189

Answer

Each of the nine cells in the cross-tab is the count of a joint occurrence of two events. The row sums and column sums on the side is the count of the an occurrence of a marginal (in the sense of marginal probabilities, not in the sense of unimportant) event. Remembering our implication of independence, if our two discrete variables – income and perception – were independent, the nine relationships below should hold:

$$P(\text{Income} = \text{Under \$20,000}, \text{Status} = \text{Worse}) = P(\text{Income} = \text{Under \$20,000})P(\text{Status} = \text{Worse})$$
$$P(\text{Income} = \text{\$20,000 - \$35,000}, \text{Status} = \text{Worse}) = P(\text{Income} = \text{\$20,000 - \$35,000})P(\text{Status} = \text{Worse})$$
$$\dots$$
$$P(\text{Income} = \text{Over \$35,000}, \text{Status} = \text{Better}) = P(\text{Income} = \text{Over \$35,000})P(\text{Status} = \text{Better})$$

For inference we need to make a comparison between a expected value and the observed value. So in this case, what next? In a Chi-square test, we *take the marginal counts as given* and see if the observed distribution of joint counts are close enough to the counts implied by independence. That is, let's set aside the actual data in the 9 cells and examine the marginal distribution of the two variables. Converted in terms of proportions, we get:

(Observed Proportion)				
	Personal financial status			
Income range	Worse	Same	Better	Total
Under $20,000				0.24
$20,000 - $35,000				0.43
Over $35,000				0.32
Total	0.30	0.34	0.36	

Now it's time to explicitly state our null hypothesis:

$$H_0 : \text{Income and Status are independent}$$

Now we populate the empty cells with counts under H_0: that is, holding the marginal proportions fixed and assuming that the two events are independent. Remember independence implies that joint probabilities are equal to the product of the marginal probabilities, so

(Expected Counts)				
		Personal financial status		
Income	Worse	Same	Better	Total
Under 20k	$0.24 * 0.30 * 185 = 13.6$	$0.24 * 0.34 * 185 = 15.3$	$0.24 * 0.36 * 185 = 16$	0.24
20 - 35k	$0.43 * 0.30 * 185 = 23.9$	$0.43 * 0.34 * 185 = 27.0$	$0.43 * 0.36 * 185 = 28.7$	0.43
35k +	$0.32 * 0.30 * 185 = 17.8$	$0.32 * 0.34 * 185 = 20.1$	$0.32 * 0.36 * 185 = 21.3$	0.32
Total	0.30	0.34	0.36	

Now that we have nine observed values and their nine expected counterparts, we can summarize their deviation in a way so that we could make use of the Chi-square distribution theorem. The computation is tedious, but we sum up to get one number.

$$\text{Test stat} = \underbrace{\frac{(13.6-20)^2}{13.6}}_{3.00} + \underbrace{\frac{(15.3-15)^2}{15.3}}_{0.006} + ... = 7.75$$

Is 7.75 big enough to reject the null hypothesis? It depends on the parameters of the distribution we will compare this against. For a two-way table, the appropriate degrees of freedom for the Chi-squared distribution under the null happens to be the product of the number of rows – minus one – and the number of columns – also minus one.
In this example there are three categories of the row variable and three categories of the column variable, so under null our Chi-squared distribution should be distributed: $\chi^2_{(3-1)*(3-1)} = \chi^2_4$
So our p-value is

```
pchisq(q = 7.75, df = 4, lower.tail = FALSE)
```

```
## [1] 0.1011774
```

There is roughly a 10 percent probability that we will observe deviations from the implications of independence that are as extreme as the ones we observe. In simpler words, the likelihood that the two variables are in-truth independent seems unlikely.

■