NYU Department of Politics - Quant I
Fall 2013 - Prof. Patrick Egan

7 Lecture 7

7.1 Sampling, sample statistics and sampling distributions

- Let's think about our journey through the world of inferential statistics so far. Our goal the entire time has been to come up with the best possible way to make inferences about populations from samples. To do this, we have:
  - recast social phenomena as *experiments* that yield observations for analysis.
    - we called the outcomes of these experiments *events*
    - and defined the *sample space* of an experiment as the set of all the events that are possible
    - and considered carefully how we assign probabilities to events of interest.
  - we then defined a *random variable* as a function mapping a sample space to the real numbers
    - and then discussed the ways we describe the probability distribution of a random variable...
    - and the probability distribution of a *function* of random variables.
- In this context, observed social phenomena (election results, outbreaks of war, passage of legislation) can all be considered realizations of random variables.
  - Let's be a little more clear about this. The phenomena that social scientists study are random events.
  - This may sound odd: in common usage we say something is "random" when it cannot be anticipated.
  - Social scientists use this term differently. When we say an event is random, we mean that it is probabilistic rather than deterministic.

  - For a random event, we attempt to specify the causal processes that alter the chance that it occurs. But we cannot specified the causal process that guarantee the event will occur. If we could, we would be studying a deterministic event.
    * An election is a good example of a random event. The best we can do is develop a theory that specifies factors important to determining election winners and then hypothesize that these factors change the odds of a particular result.
  - Thus, saying that teh variables in our theories are random variables amounts to saying that we expect that the values of these variables that we observe are draws from from an associated probability distribution.

- Finally, we arrived at a very powerful result. If the probability distribution is an accurate representation of the population frequency distribution, then the *expected value* of a random variable is the population mean, μ, where we define expected value as

$$\begin{aligned} E(Y) &\equiv \sum_y yp(y) \text{ in the discrete case and} \\ E(Y) &\equiv \int_y yf(y)\,dy \text{ in the continuous case.} \end{aligned}$$

- Now it's time to put all this theory to work helping us undertake the fundamental challenge we face in statistics: making inferences from samples to populations.

- The relevance of what we've learned previously to this task is that estimates are almost always functions of the n random observations that appear in a sample. Therefore, they are:
  - outcomes of experiments that are themselves random variables with their own probability distributions.

- If we can be very specific about the process giving rise to the sample, we can develop an **estimator** to make inferences from the sample to the population.

- An **estimator** is a *rule*, often expressed as a formula, that tells us how to calculate an estimate from a sample.

- As an example, consider the following as an estimator for the population mean, μ :
  - Draw a random sample of n observations, $y_1, y_2, ...y_n$, from the population and employ the observed sample mean

$$\overline{Y} \equiv \frac{y_1 + y_2 + ... + y_n}{n} = \frac{1}{n}\sum_i y_i$$

- As an estimator for μ, Y-bar seems pretty intuitive. But let's be precise about just how good this estimate is for μ.
  - To do this, we first think of each of the n draws that gave rise to the sample as a realization of a random variable.
  - An example: we want to know the mean income of the American population, μ.
  - To do so, I draw a sample of the American population and ask each person in the sample one question: what is your annual income?
  - Denote the response given by the first person in my sample, draw number 1, as y_1.
  - Now here's the key concept: y_1 is one of literally millions of responses I could have observed in my first survey participant. Therefore it is a realization of the random variable Y_1.
  - Now do this again: pick participant number 2 and ask him or her the same question. My observation y_2 is again one of millions of responses I could have observed for respondent number 2. Therefore it is a realization of the random variable Y_2.
  - So if we think about it this way, a sample of n observations is the realization of how many random variables? n: from $Y_1, Y_2, ...Y_n$.
  - OK, now let's think again about Y-bar.

$$\overline{Y} = \frac{1}{n}\sum_i Y_i,$$

  - In this context, our sample of n observations is just one possible realization of Y-bar.
  - That is, Y-bar is a *function* of random variables, and therefore is *itself* a random variable.

- Note use of capital letters here, we are talking about theoretical, not observed, quantitites. In keeping with our convention, little $\bar{y}$ is a realization of the random variable $\overline{Y}$.
- Because it is a random variable, $\overline{Y}$ has a probability distribution. To specify it, we will stipulate some simple but powerful assumptions that hold whenever we have a **random sample**. (Recall how we defined random sample: each of the $\binom{N}{n}$ different possible samples has an equal probability of being drawn.)

- This generates the canonical case, in which we have a function of the random variables $Y_1, Y_2, ...Y_n$ observed in a random sample from a population of interest.
  - When we have a random sample from a large enough population, we can safely assume that the RVs $Y_1, Y_2, ...Y_n$ are independent and identically distributed ("i.i.d.").
    - To recall: $Y_1, Y_2, ...Y_n$ are said to be **independent** iff

$$F(y_1, y_2, ...y_n) = F_1(y_1)F_2(y_2) \cdot ... \cdot F_n(y_n) \text{ for every } n\text{-tuple } (y_1, y_2, ...y_n), \text{ and}$$

  - Now we introduce another concept: $Y_1, Y_2, ...Y_n$ are said to be **identically distributed** iff

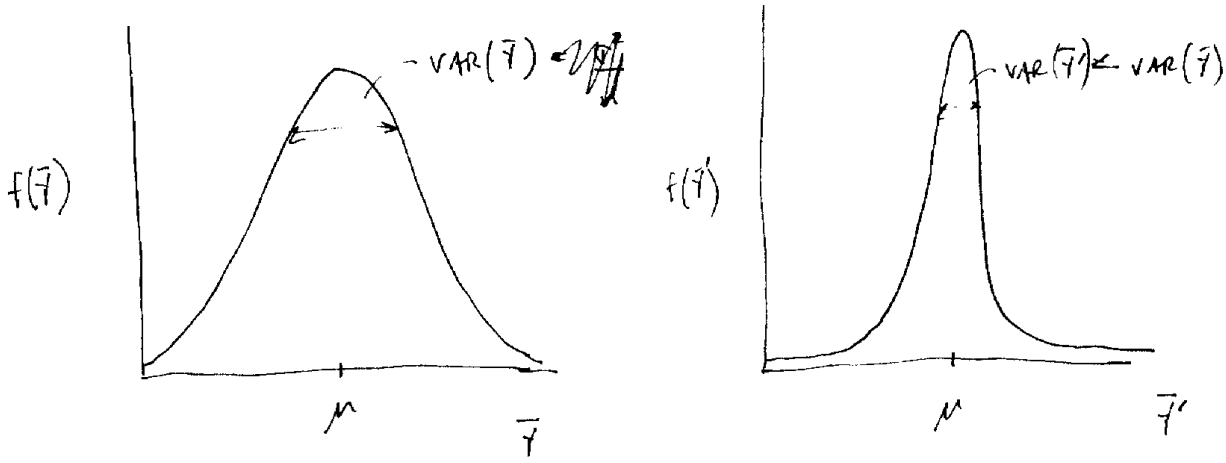
$$F_1(y_1) = F_2(y_2) = ... = F_n(y_n) = F(y) \text{ for } y_1, y_2, ...y_n.$$

  - In this context, $\overline{Y} = \frac{1}{n}\sum_i Y_i$ is a **sample statistic**, which we define as a function of the observable random variables in a sample and known constants.

- We know (from the handout last time) that $E(\overline{Y}) = \mu$, [note that this is due to the identicality assumption] and so we can be assured that, on average, $\overline{Y}$ should equal μ. So now let's ask, how good of an estimate is it?
  - A straightforward measure of "goodness" would be how far off we can expect any realization of $\overline{Y}$ to be from the population mean, μ.
  - And since the mean of $\overline{Y}$ is μ, this quantity is just the standard deviation of $\overline{Y}$,or $\sigma_{\overline{Y}}$. But THIS of course is just $\sqrt{VAR(\overline{Y})}$, which we showed last time to be $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$. [recall that we need BOTH identicality and independence to achieve this result.]

of y-

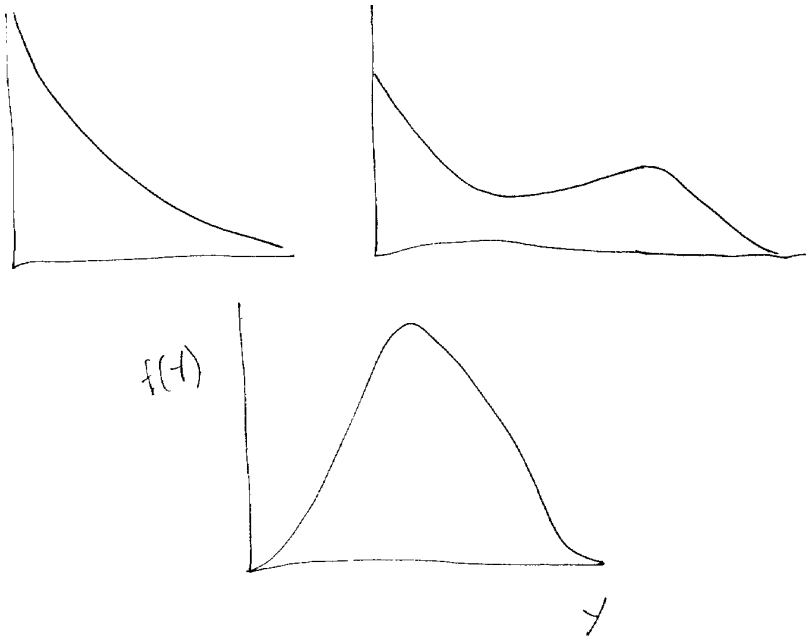
bar



1.*pdf*

- Well, because sample statistics are themselves random variables, they have probability distributions (recall: table, graph, formula). We have a special term for the probability distributions of sample statistics: **sampling distributions**. The sampling distribution of a sample statistic is a theoretical model for the possible values of the statistic we would expect to observe through repeated sampling.
- The expected value and the variance of a sample statistic are important properties of the statistic's sampling distribution. For example, here are graphs depicting two sampling distributions of Y-bar: Y-bar and Y-bar-prime:

- Y-bar and Y-bar-prime have the same expected value: μ. But clearly Y-bar prime, which has a smaller variance, is a better estimate of μ than Y-bar. It's closer, on average, to μ than Y-bar.
- So:
  - we've got $\overline{Y}$, an estimate for μ that, on average, equals μ.
  - we now also know how good an estimate this is: on average, it is $\frac{\sigma}{\sqrt{n}}$ units away from μ.



- But now an additional question arises: what does the distribution of the random variable Y look like? Often, we'll be drawing samples from populations about whose frequency distribution we have no idea. E.g., I want to estimate μ, the mean number of potatoes eaten by the average American per month. Does the distribution of Y look like this? Or this? Or this?

- If possible, we'd really like to avoid making assumptions about the shape of the population distribution of Y in order to describe the distribution of Y-bar.

- And, it turns out, we can...

7.2 The Central Limit Theorem

- ...because if my sample size is large enough, I don't need any assumptions about the distribution of Y to describe the sampling distribution of Y-bar. For it turns out that if $Y_1 ... Y_n$ are i.i.d, then:

  - Y-bar has a sampling distribution that is approximately Normal
  - as the sample size becomes large.
  - This is the **central limit theorem**.

- Before digging into the math (and we won't dig that far), let's have a look at a few distributions that convey the intuition (go over handout: "Probability Distributions of Y and Simulated Sampling Distributions of $\overline{Y}$")

- The CLT is more formally stated as follows:

- Let $Y_1, Y_2, ...Y_n$ be i.i.d. random variables with $E(Y_i) = \mu$ and $VAR(Y_i) = \sigma^2$. Define

$$U_n \equiv \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}.$$

  That is, we standardize Y-bar by (1) subtracting its hypothesized mean (μ) and then (2) dividing this difference by Y-bar's standard deviation $(\frac{\sigma}{\sqrt{n}})$. Then the CDF of U_n converges in probability to (where "converges in probability to" means "as n becomes large it is distributed as") the standard normal CDF. That is,

$$\lim_{n\to\infty} F_{U_n}(u) = \lim_{n\to\infty} P(U_n \leq u) = P(Z \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \text{ for all } u.$$

- Take a moment to appreciate the power of the CLT. It yields the sampling distribution of Y-bar *without requiring any assumptions about the probability distribution of* Y. We will not cover the proof of the CLT in this class, as it requires knowledge of moment generating functions, which itself requires knowledge of Taylor series expansions. Those equipped with these tools who want to satisfy themselves with a proof may see section 7.4 of the text.

7.3 Estimation

- With the sampling distributions yielded by the CLT in hand, we can develop **estimators** of population parameters. An estimator is a rule–often expressed as a formula–that tells us how to calculate an estimate of a population parameter.

  - Two kinds of estimates that we'll focus on here include

    * **point estimates**–in which a single value, or point, is given as the estimate of the parameter

    - ∗ **interval estimates**-in which two values are used to construct an interval that we believe contains/traps/encloses the parameter of interest.

- So the sample mean, $\overline{Y} = \frac{1}{n}\sum_i Y_i$, is one possible point estimator of the population mean μ.
- But there are lots of others. For example, consider the estimator $\overline{Y_B} = \frac{1}{n}\sum_i (Y_i + 1)$.
- Intuitively, we know this is a worse estimator than Y-bar. But can we put some meat on this intuition? We do this by specifying two desirable criteria for evaluating a potential estimator:
    - 1. unbiasedness
    - 2. (relatively) small variance, which is also known as **efficiency** or **precision**.
- First, some terminology. We typically write the population parameter for which we seek a point estimate as θ, and a proposed estimator for this parameter as $\widehat{\theta}$.
- Now, here is why we've spent a fair amount of time learning about the math of expectations: we can use these tools to show whether an estimator is unbiased and to determine how precise it is.
- In fact, we define an estimator as unbiased if–in expectation–it is equal to the parameter it claims to estimate. That is, an estimator is unbiased if the expected value of its distribution is the parameter. Formally,
    - $\widehat{\theta}$ is an **unbiased estimator** for θ if $E(\widehat{\theta}) = \theta$.
    - If $E(\widehat{\theta}) \neq \theta$, then we say $\widehat{\theta}$ is **biased**.
    - The **bias** of a point estimator is given by $B(\widehat{\theta}) = E(\widehat{\theta}) - \theta$.
- E.g. So one way to say that the estimator $\overline{Y_B}$ isn't a good estimator is to show that it is a

biased estimator for μ. We do that by showing that $E(\overline{Y_B}) \neq \mu$, and thus that $B(\overline{Y_B}) \neq 0$.

$$\begin{aligned} E(\overline{Y_B}) &= E\left[\frac{1}{n}\sum_i (Y_i + 1)\right] \\ &= \frac{1}{n}\left\{\left[\sum_i E(Y_i)\right] + nE(1)\right\} \\ &= \frac{1}{n}\{[n\mu] + n\} \\ &= \mu + 1 \neq \mu, \text{ and } B(\overline{Y_B}) = 1. \end{aligned}$$

- So much for the first criterion. Our second criterion is that we'd like our estimator to be as close as possible to θ in repeated sampling. In mathematical terms, we of course want the variance of the sampling distribution of our estimator to be as small as possible. Recalling the definition of the variance of a random variable, we write the variance of an estimator $\widehat{\theta}$ as $VAR(\widehat{\theta}) = E\left\{[\widehat{\theta} - E(\widehat{\theta})]^2\right\}$. In an ideal world, we wish for this to be as small as possible.

- Sometimes we face a tradeoff between reducing an estimator's bias and reducing its variance. One way to evaluate the tradeoff is to minimize an estimator's **mean square error (MSE)**, which is defined as the expected value of the square of the distance between the estimator and the parameter

$$MSE(\widehat{\theta}) = E\left[\left(\widehat{\theta} - \theta\right)^2\right].$$

- It can be shown that the MSE is equal to the sum of an estimator's variance plus the square of its bias:

$$MSE(\widehat{\theta}) = VAR(\widehat{\theta}) + \left[B(\widehat{\theta})\right]^2.$$

  - Ask: Why does this measure include the square of its bias?

7.4 An example

- For example, let us say that we wish estimate the parameter $\mu_1 - \mu_2$, the difference in means of two different populations drawn from independent samples. Is the intuitive estimator,

$\overline{Y_1} - \overline{Y_2}$, unbiased? Let's see:

$$\begin{aligned} E(\overline{Y_1} - \overline{Y_2}) &= E(\overline{Y_1}) - E(\overline{Y_2}) \\ &= \mu_1 - \mu_2. \end{aligned}$$

- Yes, $\overline{Y_1} - \overline{Y_2}$ is an unbiased estimator for $\mu_1 - \mu_2$. Now what is its variance? Well,

$$\begin{aligned} VAR(\overline{Y_1} - \overline{Y_2}) &= VAR(\overline{Y_1}) + VAR(\overline{Y_2}) + 2COV(\overline{Y_1}, \overline{Y_2}) \\ &= VAR(\overline{Y_1}) + VAR(\overline{Y_2}) \ [\overline{Y_1}, \overline{Y_2} \text{ independent}] \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \ [\text{variance of Y-bar}] \end{aligned}$$

  where σ_1^2 and σ_2^2 are the variances of populations 1 and 2 respectively.

- Note that we can talk about the standard error of an estimator $\widehat{\theta}$, which we write $\sigma_{\widehat{\theta}}$. It is just the square root of the variance of the estimator. The standard error of the estimator $\overline{Y_1} - \overline{Y_2}$ is thus $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- Another variant of the CLT (which does not require identicality, but instead requires some other conditions) tells us that the sampling distribution of a sum of independent random variables (like $\overline{Y_1}$ and $\overline{Y_2}$) approximates the Normal as n becomes large. And so

$$\overline{Y_1} - \overline{Y_2} \tilde{} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \text{ as } n_1, n_2 \longrightarrow \infty.$$

8 Lecture 8

8.1 The Intuitive Estimator Is Not Always the Unbiased Estimator!

- You are probably not surprised to learn that $\overline{Y}$ is an unbiased estimator for μ while $\overline{Y}_B$ is a biased estimator. After all, $\overline{Y}$ is the mean of the sample and so intuitively it seems like it should be unbiased estimator for the population mean. The same for $\overline{Y_1} - \overline{Y_2}$ as an estimator of $\mu_1 - \mu_2$.

- HOWEVER, it is not always the case that the intuitive estimator is the unbiased estimator.

We'll illustrate this in a way that explains the distinction statisticians draw between sample variance and population variance.

- It would seem natural to estimate the variance of a population, σ^2, with the sample variance $S^2 = \frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n}$. But we can show that this is in fact a *biased* estimator for σ^2:

$$\begin{aligned}
E\left(S^2\right) &= E\left(\frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n}\right) \\
&= E\left[\frac{\sum_i (Y_i)^2}{n} - \left(\overline{Y}\right)^2\right] \text{ [remember this from PS 1?]} \\
&= \frac{1}{n} E\left[\sum_i (Y_i)^2 - n\left(\overline{Y}\right)^2\right] \text{ [multiplying terms by } \frac{n}{n}\text{]} \\
&= \frac{1}{n}\left\{\left(\sum_i E\left[(Y_i)^2\right]\right) - nE\left[\left(\overline{Y}\right)^2\right]\right\} \text{ [distributing expectations]} \\
&= \frac{1}{n}\left\{\left(\sum_i VAR(Y) + [E(Y)]^2\right) - n\left(VAR(\overline{Y}) + \left[E\left(\overline{Y}\right)\right]^2\right\}\right.
\end{aligned}$$

[using the identity $VAR(Y) = E(Y^2) - [E(Y)]^2$, and identicality of Y_i]

$$\begin{aligned}
&= \frac{1}{n}\left\{\left(\sum_i \sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\} \text{ [substituting identities]} \\
&= \frac{1}{n}\left\{n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\} \\
&= \frac{1}{n}\left\{n\sigma^2 - \sigma^2\right\} = \frac{n-1}{n}\sigma^2 \neq \sigma^2.
\end{aligned}$$

- Now consider an alternate estimator, $S_U^2 = \frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n-1}$. Well, $E\left(S_U^2\right) = \frac{1}{n-1} E\left[\sum_i \left(Y_i - \overline{Y}\right)^2\right] = \frac{n-1}{n-1}\sigma^2 = \sigma^2$ [will be on homework] and hence is unbiased.

- When describing the variance of a sample, we will write S^2 and use the formula we've been using so far this semester. But when describing the unbiased estimator of a population variance, we will write S_U^2 and use this new formula. Note the difference between this practice and your text.

- But we are probably making a mountain out of a molehill. Because what is $B(S^2)$? What is

$\lim_{n\to\infty} B(S^2)$? What is the implication of this?

$$\begin{aligned} B(S^2) &= E\left(S^2\right) - \sigma^2 \\ &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= \left(\frac{n-1}{n} - 1\right)\sigma^2 \\ &= -\sigma^2/n. \\ \text{So} \quad &: \quad \lim_{n\to\infty} B(S^2) = \lim_{n\to\infty} -\sigma^2/n = 0. \end{aligned}$$

- Make sure to talk notation: we'll use the notation S_U^2 to specify the unbiased estimator for the population variance, where $S_U^2 \equiv \frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n-1}$.

8.2 Confidence Intervals

- Last time we talked about point estimators and two desirable properties: unbiasedness and relatively small variance. Today we'll discuss another kind of estimator: interval estimators.
- An **interval estimator** is:
  - a rule
  - specifying how we use the sample to calculate two numbers
  - that form the endpoints of an interval
  - containing/trapping/enclosing a parameter of interest, θ.
- Intervals have two desirable properties. We want them to:
  - Contain the parameter of interest, θ
  - Be relatively narrow (sound familiar?)
- As with point estimators, the length and location of the interval are random quantitites, so our goal is to find an interval estimator that generates narrow intervals with a high probability of trapping θ.
- [Distribute handout about here.]

- Interval estimators are commonly called **confidence intervals (CIs)**.
  - CIs are constructed of two quantities called the **upper** and **lower confidence limits** (or **upper** and **lower bounds).**
  - The probability that a random CI will enclose θ is called the **confidence coefficient:** it is:
    - the fraction of the time,
    - in repeated sampling,
    - that the CI will contain θ.
  - We thus like confidence coefficient associated with our CI to be high.
  - The confidence coefficient is written $(1-\alpha)$. If $\widehat{\theta}_L$ and $\widehat{\theta}_H$ are the random lower and upper confidence limits, then

$$P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U) = (1-\alpha).$$

- We typically call a CI with confidence coefficient $(1-\alpha)$ a "$[100 \cdot (1-\alpha)]-$percent confidence interval." Sometimes we also say a "CI with alpha=[α]."
- Here we discuss how to construct a confidence interval for a sample statistic $\widehat{\theta}$ that is Normally distributed with mean μ and standard error $\sigma_{\widehat{\theta}}$. (What would be an example of such a statistic? The CLT tells us that one example would be Y-bar constructed from a large sample.)
- Let's standardize the statistic as follows:

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}},$$

  Now it is distributed approxmately standard Normal, as we have substracted the estimator's hypothesized mean and divided by its standard deviation.

- To construct a confidence interval for $\widehat{\theta}$, pick two values $-z_{\alpha/2}, z_{\alpha/2}$ such that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = \int_{-z_{\alpha/2}}^{z_{\alpha/2}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 1 - \alpha.$$

- Substituting for Z, we have

$$\begin{aligned} P(-z_{\alpha/2} &\leq \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \leq z_{\alpha/2}) = 1 - \alpha \\ P(-z_{\alpha/2}\sigma_{\widehat{\theta}} &\leq \widehat{\theta} - \theta \leq z_{\alpha/2}\sigma_{\widehat{\theta}}) = 1 - \alpha \\ P(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}} &\leq \theta \leq \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}) = 1 - \alpha. \end{aligned}$$

- And therefore $\widehat{\theta}_L = \widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}}$, and $\widehat{\theta}_H = \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}$.

- And how to find $-z_{\alpha/2}$ and $z_{\alpha/2}$? Well, (Draw Normal curve on board.) Help class figure out that $z_{\alpha/2}$ is the value satisfying $P(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$.

- Interlude for a bit of notation: recall that we write the standard Normal CDF (the probability that the standardized Normal RV Z will be less than or equal to z) as $\Phi(z)$. Looking at the Normal curve, it is evident that $\Phi(-z) = 1 - \Phi(z)$. Furthermore, it is often helpful to talk about the *inverse* of the Normal CDF–that is, a function whose argument is a probability and which returns a value of Z. We write this $\Phi^{-1}(p)$,where p is the probability and $P(Z \leq \Phi^{-1}(p)) = p$. Similarly, $P(Z \geq -\Phi^{-1}(p)) = p$.

- So in this case, if we're trying to find the $z_{\alpha/2}$ that satisfies $P(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$, we can also write $z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$ and $-z_{\alpha/2} = \Phi^{-1}\left(\frac{\alpha}{2}\right)$.

- In Stata, we type

  .**display invnormal(*p*)** to find $\Phi^{-1}(p)$.

- So when I type **display invnormal(.025)**, I get **-1.959964.**

  - This is the z-score associated with a $1 - 2(.025) = .95$ CI. And so if $\alpha = .05$, then $z_{\alpha/2} = 1.96$.
  - and when I type **display invnormal(.05)**, I get **-1.6448536.**

– and when I type **display invnormal(.005), I get -2.5758293.**

- So a 95% confidence interval for an estimator $\widehat{\theta}$ whose sampling distribution is Normally distributed is constructed as

$$[\widehat{\theta} - (1.96)\,\sigma_{\widehat{\theta}}, \widehat{\theta} + (1.96)\,\sigma_{\widehat{\theta}},$$

Now, what is $\sigma_{\widehat{\theta}}$? Why, we've already learned that: it's $\sigma_{\widehat{\theta}} = \sqrt{VAR(\widehat{\theta})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$.

- What about a 99% CI? A 90% CI?

- So let's do an example. [Do Example 8.7 on page 412, but do not specify the variance, leave it as unknown.]

8.3 What is σ^2?

- As you'll recall, the CLT tells us that if $Y_1, Y_2, ...Y_n$ be i.i.d. random variables with $E(Y_i) = \mu$ and $VAR(Y_i) = \sigma^2$,then

$$U_n \equiv \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

converges in probability to the standard normal CDF.

- Note that we've been dancing around a little bit of a problem. We've been using the CLT to construct interval estiamtors for μ, but they remain unquantified because they're in terms of σ^2, the population standard variance.

- This is a problem because it is very unusual to know the value of σ^2. Lots of homework problems and exercises will supply you the value of σ^2, but in practice we almost never have any reason to know what it actually is.

- So we need to estimate it with $S_U^2 \equiv \frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n-1}$, our unbiased estimator for σ^2.

- At first seems intuitive. We have an unbiased estimate of σ^2 and we should be comfortable substituting S_U^2 for σ^2. It turns out that to justify this move, we need a bit more theory.

8.4 Interlude: the property of consistency

- In order to justify the estimation, we need to learn a new property of estimators: **consistency**. An estimator $\widehat{\theta}_n$ constructed from a sample of size n (subscripted to indicate just that) is a *consistent estimator* for θ if for any positive number ε,

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \varepsilon) = 0.$$

- That is, as the sample size used to construct a consistent estimator becomes large, the chance that the error of the estimator is non-zero converges to zero. (Let's say this again...) We sometimes equivalently say that "$\widehat{\theta}_n$ converges in probability to θ," or $\widehat{\theta}_n \xrightarrow{p} \theta$

- Are unbiased estimators consistent estimators? Often, but not necessarily. Intuitively, this is because you could have an unbiased estimator that–even as n gets large–bounces around θ in a random, unbiased fashion but never centers on θ. Mathematically, this can be expressed as the helpful result (we'll omit proof here, it's on p. 450 of your text) that an unbiased estimator $\widehat{\theta}_n$ for θ is a consistent estimator for θ if its variance converges in probability to zero, that is:

$$E(\widehat{\theta}_n) = \theta \text{ and } \lim_{n\to\infty} VAR(\widehat{\theta}_n)=0 \Rightarrow \lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \varepsilon) = 0, \textit{i.e.}$$

$\widehat{\theta}$ *unbiased* for θ and $\lim_{n\to\infty} VAR(\widehat{\theta}_n) = 0 \Rightarrow \widehat{\theta}$ *consistent* for θ.

- This result, for example, means that $\overline{Y}$ is not only an unbiased estimator for μ, but because $VAR(\overline{Y}) = \frac{\sigma^2}{n}$ and thus $\lim_{n\to\infty} VAR(\overline{Y}) = 0$, $\overline{Y}$ is a consistent estimator for μ. The fact that $\overline{Y} \xrightarrow{p} \mu$ is sometimes referred to as the *law of large numbers*. It is the formal way of saying the intuitive idea that the average of many independent measures from a population should be quite close to the true population mean with high probability.

- Note that an estimator $\widehat{\theta}$ can be *consistent* for θ but not *unbiased* for θ.

  - E.g., S^2 as an estimator for σ^2.

- An estimator can be unbiased for θ but not consistent for θ.if its variance does not become

monotonically smaller as n goes to infinity.

- Some helpful results are that if we have two estimators $\widehat{\theta}$ and $\widehat{\theta}'$ (I'm now going to drop the subscripts n to keep notation cleaner) such that $\widehat{\theta} \xrightarrow{p} \theta$ and $\widehat{\theta}' \xrightarrow{p} \theta'$, then [put these on left-hand board for reference later]:

$$\begin{aligned} \widehat{\theta} + \widehat{\theta}' &\xrightarrow{p} \theta + \theta' \\ \widehat{\theta} \times \widehat{\theta}' &\xrightarrow{p} \theta \times \theta' \\ \frac{\widehat{\theta}}{\widehat{\theta}'} &\xrightarrow{p} \frac{\theta}{\theta'}, (\theta' \neq 0). \end{aligned}$$

  Furthermore if $g(\cdot)$ continuous at θ, then $g(\widehat{\theta}) \xrightarrow{p} g(\theta)$.

8.5 Back to σ^2

- Recall that the reason we discuss consistency was to help us justify estimating the population variance with S_U^2.

- Note that if we were simply estimating σ^2 in a vacuum, we'd be perfectly comfortable using S_U^2. After all $E\left(S_U^2\right) = \sigma^2$.

- But our task is a little more complicated here. We typically wish to substitute S_U^2 for σ^2 in the ratio $U_n \equiv \frac{\overline{Y}-\mu}{\sqrt{\sigma^2/n}}$ (which the CLT tells us is distributed standard Normal as n becomes large), and be assured that the resulting ratio, $\frac{\overline{Y}-\mu}{\sqrt{S_U^2/n}}$ (or, as we usually write it, $\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}$) is itself distributed standard Normal as n becomes large.

- That is, we wish to show that $F\left(\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}\right) \xrightarrow{p} \Phi$. (Put box around this to keep eyes on the prize during rest of tedious derivation.)

- Intuitively, it seems like it should. But this is definitely a more complicated question, because where in the original ratio σ^2 was a parameter, S_U^2 is a random variable.

- To begin, let's first show that S_U^2 is not only unbiased for σ^2; it's also consistent for σ^2.

- Rewrite

$$\begin{aligned} S_U^2 &\equiv \frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n-1} \\ &= \frac{1}{n-1}\left[\sum_i Y_i^2 + \sum_i \overline{Y}^2 - \sum_i 2Y_i\overline{Y}\right] \\ &= \frac{1}{n-1}\left[\left(\sum_i Y_i^2\right) + n\overline{Y}^2 - 2n\left(\overline{Y}\overline{Y}\right)\right] \\ &= \frac{1}{n-1}\left[\left(\sum_i Y_i^2\right) - n\overline{Y}^2\right] \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_i Y_i^2 - \overline{Y}^2\right) \end{aligned}$$

- Now let's consider

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}\sum_i Y_i^2 - \overline{Y}^2 &= \lim_{n\to\infty} \frac{1}{n}\sum_i Y_i^2 - \lim_{n\to\infty} \frac{1}{n}\sum_i \overline{Y}^2 \\ &= \mu_{Y^2} - \left(\mu_Y\right)^2, \end{aligned}$$

  where the first result is due to the law of large numbers (as $n \to \infty$, the sample mean converges to the population mean–in this case, the population mean of Y^2, which I write μ_{Y^2}). The second result is due to $g(\widehat{\theta}) \xrightarrow{p} g(\theta)$, since the function $g(x) = x^2$ is continuous and since $\overline{Y} \xrightarrow{p} \mu$, $g(\overline{Y}) \xrightarrow{p} g(\mu)$, or $\lim_{n\to\infty} \frac{1}{n}\sum_i \overline{Y}^2 = \left(\mu_Y\right)^2$.

- Note that $\mu_{Y^2} - \left(\mu_Y\right) = E(Y^2) - \mu^2$, and now we have something that should look familiar once we recall the variance decomposition formula

$$\sigma^2 = E(Y^2) - \mu^2.$$

  Thus

$$\lim_{n\to\infty} \frac{1}{n}\sum_i Y_i^2 - \overline{Y}^2 = \mu_{Y^2} - \left(\mu_Y\right)^2 = \sigma^2.$$

- Now the multiplicand $\frac{n}{n-1}$. But what is $\lim_{n\to\infty} \frac{n}{n-1}$? It is a sequence of numbers converging to

1. Thus

$$\lim_{n\to\infty} \frac{n}{n-1}\left(\frac{1}{n}\sum_i Y_i^2 - \overline{Y}^2\right) = 1\sigma^2 = \sigma^2, \text{and}$$
$$S_U^2 \xrightarrow{p} \sigma^2.$$

- And so S_U^2 is a consistent estimator for σ^2.

9 Lecture 9

9.1 Where we are

Just to quickly review:

- We are keenly interested in identifying a good estimator for the population mean, μ, from a random sample of data from that population.
- $\overline{Y} \equiv \frac{1}{n}\sum_i Y_i$, the sample mean, is an obvious choice for such an estimator.
- We proceed by modeling the sampling process yielding n observations as a series of random variables $Y_1, Y_2, ...Y_n$. They are independent, and they are identically distributed: that is, they all have the some CDF F, the same mean μ and the same variance σ^2. With this in hand, we:
  - established that $\overline{Y}$ is an unbiased estimator of μ, i.e. that $E\left(\overline{Y}\right) = \mu$.
  - we showed that its variance is $VAR\left(\overline{Y}\right) = \sigma_{\overline{Y}}^2 = \frac{\sigma^2}{n}$, and thus its standard deviation $\sqrt{VAR\left(\overline{Y}\right)} = \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$.
- That's good. Now we want to know how close, on average, the estimator Y-bar is to μ.
  - Well, the central limit theroem tells us that the sampling distribution of Y-bar is distributed Normal as n becomes large. We typically find it more useful to write this in terms of the *standardized* version of Y-bar, that is

$$U_n \equiv Z \equiv \frac{\overline{Y} - \mu}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}},$$

– where the CLT tells us that this converges in probability to the *standard* Normal:

$$F\left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right) \xrightarrow{p} \Phi.$$

- This allows us to begin to quantify how close Y-bar is, on average, to μ. Since Y-bar is distributed Normal, when n is large it is generated through a process that yields intervals trapping μ in repeated sampling $1-\alpha$ percent of the time, where α and $z_{\alpha/2}$ satisfy

$$P(\overline{Y} - z_{\alpha/2}\sigma_{\overline{Y}} \leq \mu \leq \overline{Y} + z_{\alpha/2}\sigma_{\overline{Y}}) = 1-\alpha.$$

- For any α we pick, we can find the appropriate $z_{\alpha/2}$ with statistical software or tables; it is the value at with the CDF of the standard Normal is evaluated that yields $\alpha/2$.

- And because $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$, we know that

$$P(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

- But to quantify the distribution of Y-bar, we need one more thing. We need to contend with σ, the standard deviation of Y. To deal with this, we:

  – Identified an estimator $S_U^2 \equiv \frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n-1}$, and showed that it is unbiased for σ^2, the population variance.

  – We also showed that this estimator is *consistent* for σ^2; i.e. that $S_U^2 \xrightarrow{p} \sigma^2$.

  – We want to get to the point where we can justify substituting S_U for σ and saying

$$F\left(\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}\right) \xrightarrow{p} \Phi.$$

  – This is exactly what we are about to do.

9.2 Slutzky's Theorem

- To justify our substitution of S_U for σ, we'll need one more tool: *Slutzky's Theorem* (love that name). [Put this on separate board.] This theorem tells us that:

- if the distribution of some function is such that $F(U_n) \xrightarrow{p} \Phi$ and
- if the distribution of some other function W_n is such that $F(W_n) \xrightarrow{p} 1$, then
- $F\left(\frac{U_n}{W_n}\right) \xrightarrow{p} \Phi$.
- In words, Slutzky's theorem tells us that the ratio of a function that converges to the Standard Normal over a function that converges to 1 itself converges to the Standard Normal.

9.3 Putting it all together

- OK, now we're ready to prove the powerful result we've been seeking:

$$F\left(\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}\right) \xrightarrow{p} \Phi.$$

Proof:

- Begin by re-writing $F\left(\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}\right) = F\left(\frac{\frac{\overline{Y}-\mu}{\sqrt{n}}\cdot\frac{1}{\sigma}}{S_U\cdot\frac{1}{\sigma}}\right) = F\left(\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\frac{S_U}{\sigma}}\right)$. If we can show this final expression converges to the standard Normal, then we know that $\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}$ does, too.
- Note that $\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\frac{S_U}{\sigma}}$ is a ratio of a function that converges to the Standard Normal over the function $\frac{S_U}{\sigma}$:
  - The CLT tells us that
$$F\left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right) \xrightarrow{p} \Phi.$$
  - So if we can show that $\frac{S_U}{\sigma}$ converges to 1, then Slutzky's Theorem implies that
$$F\left(\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\frac{S_U}{\sigma}}\right) \xrightarrow{p} \Phi.$$
- To do this,
- recall that we've shown $S_U^2 \xrightarrow{p} \sigma^2$ [consistency of S_U^2 .]
- Now note that $\frac{S_U}{\sigma} = +\sqrt{\frac{S_U^2}{\sigma^2}}$. Because the function $g(x) = +\sqrt{\frac{x}{c}}$ is continuous if both x, c positive, then we can invoke the rule that if $\widehat{\theta} \xrightarrow{p} \theta$ and $g(\cdot)$ continuous at θ, then

$g(\widehat{\theta}) \xrightarrow{p} g(\theta)$.

- Here $\frac{S_U^2}{\sigma^2} \xrightarrow{p} \frac{\sigma^2}{\sigma^2} = 1$, and $\sqrt{}$is clearly continuous at 1, so $\frac{S_U}{\sigma} = +\sqrt{\frac{S_U^2}{\sigma^2}} \xrightarrow{p} \sqrt{\frac{\sigma^2}{\sigma^2}} = 1$.
- Now we invoke Slutzky's Theorem to show that the distribution of this ratio, and therefore the distribution of $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$, converges in probability to the standard Normal.

- Whew, that was a lot of work! What does it buy us? It tells us that when n is large, $\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}$ is distributed approximately standard Normal, whatever the distribution of the underlying population.

- Therefore it follows that

$$\begin{aligned} P\left[-z_{\alpha/2} \leq \frac{\overline{Y}-\mu}{S_U/\sqrt{n}} \leq z_{\alpha/2}\right] &\approx 1-\alpha \text{ and so} \\ P\left[\overline{Y} - z_{\alpha/2}\left(\frac{S_U}{\sqrt{n}}\right) \leq \mu \leq \overline{Y} + z_{\alpha/2}\left(\frac{S_U}{\sqrt{n}}\right)\right] &\approx 1-\alpha. \end{aligned}$$

- Thus $\overline{Y} \pm z_{\alpha/2}\left(\frac{S_U}{\sqrt{n}}\right)$ forms a valid **large-sample CI** for μ. And this is the challenge we originally faced. We can now substitute $\frac{S_U}{\sqrt{n}}$ for $\sigma_{\widehat{\theta}}$.

9.4 Examples of Large-Sample CIs

- Let's revisit the notion of a large-sample CI with an example.

- The American Community Study (ACS) is a program of the Census Bureau that estimates quantities of interest in the population using a large-sample survey.

- For example, the mean household income of New York State was estimated to be \$76,247 using a sample of about 350,000 households. The unbiased estimate of the population standard deviation is $S_U = 61,427$. What is the 90% CI associated with this estimate?

  - Recall that we write the $100\,(1-\alpha)$ percent CI for the population mean, μ as

$$\overline{Y} \pm z_{\alpha/2}\,(\sigma_{\overline{Y}})\,, \text{where } z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ and } \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}.$$

  - Let's first find $z_{\alpha/2}$.

- What is alpha here? It's 1 minus the confidence coefficient (in this case, .90), or .10 .
- So what is $z_{.10/2} = z_{.05}$? It's $z_{.05} = -\Phi^{-1}(.05)$. Calculate this by typing **di invnormal(.05)** in Stata, obtaining -1.64. So $z_{.05} = 1.64$.
- We're almost there. Our 90% CI can be written

$$\overline{Y} \pm z_{\alpha/2}(\sigma_{\overline{Y}}) = \$76,247 \pm (1.64)\ \sigma_{\overline{Y}}.$$

- Recall that we've shown we can substitute

$$\begin{aligned} S_U &= \sqrt{\frac{\sum (y_i - \overline{y})^2}{n-1}} \text{ for the population standard deviation,} \\ &\text{and thus can rewrite our CI as} \\ \overline{Y} \pm z_{\alpha/2}(\sigma_{\overline{Y}}) &= \$76,247 \pm (1.64)\ \left(\frac{S_U}{\sqrt{n}}\right) \\ &= \$76,247 \pm (1.64)\ \left(\frac{61,427}{\sqrt{350,000}}\right) \\ &= \$76,247 \pm (1.64)\ (103.83) \\ &= \$76,247 \pm 170.28, \text{ or } [\$76,077,\ \$76,417]. \end{aligned}$$

9.5 Another example of a large-sample CI: proportions

- CNN poll, Oct 16-18, 2009 with sample of 1,038 American adults.
- Finding: 64 percent say they have a "favorable" opinion of Michelle Obama; 36% do not.
- Let's construct a 95% large-sample CI around this estimate.
- Before proceeding, let's think:
  - In the previous example, we wrote our CI for the population mean, μ, as

$$\widehat{\mu}_{LB}, \widehat{\mu}_{UB} = \overline{Y} \pm z_{\alpha/2}(\sigma_{\overline{Y}}), \text{where } z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ and } \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}.$$

  - But recall that the CLT tells us we can also write this more generically for *any* estimator

that is a linear combination of random variables that are i.i.d. as

$$\widehat{\theta}_{LB}, \widehat{\theta}_{UB} = \widehat{\theta} \pm z_{\alpha/2} \left(\sigma_{\widehat{\theta}}\right).$$

- And in this example, our parameter of interest is p: the proportion of Americans viewing Michelle Obama favorably. Our estimator is $\widehat{p} = \frac{Y}{n}$, where $Y = 0$ if Obama is viewed unfavorably and $Y = 1$ if she is viewed favorably. We've shown previously that $\widehat{p}$ is unbiased for p. So let's write $\widehat{p} = .64$.
- Now rewrite our CI of interest as

$$\widehat{p}_{LB}, \widehat{p}_{UB} = \widehat{p} \pm z_{\alpha/2} \left(\sigma_{\widehat{p}}\right)$$

- Now think:
  - We have $\widehat{p}$.
  - We'll find $z_{\alpha/2}$ the usual way. (It's equal to - **invnormal**$(.025)$ = 1.96.)
  - What about $\sigma_{\widehat{p}}$?
- A few lectures ago we showed that

$$VAR\left(\widehat{p}\right) = VAR\left(\frac{Y}{n}\right) = \frac{1}{n^2} VAR\left(Y\right) = \frac{np\left(1-p\right)}{n^2} = \frac{p\left(1-p\right)}{n}.$$

- And so

$$\sigma_{\widehat{p}} = \sqrt{\frac{p\left(1-p\right)}{n}}.$$

- We can substitute $\widehat{p}$, our estimate of p, in the formula for $\sigma_{\widehat{p}}$,and so a large-sample CI for a population proportion p can be written

$$\widehat{p}_{LB}, \widehat{p}_{UB} = \widehat{p} \pm z_{\alpha/2} \left(\sqrt{\frac{\widehat{p}\left(1-\widehat{p}\right)}{n}}\right).$$

- To return to our example, we can write the 95% CI about our estimate of the proportion of

the population having a favorable opinion of Michelle Obama as

$$\begin{aligned} & .64 \pm 1.96 \left(\sqrt{\frac{.64\,(1-.64)}{1038}} \right) \\ = \quad & .64 \pm .029 \end{aligned}$$

- This corresponds to the poll's published "Margin of Error" of "plus or minus 3 percentage points." When you see this reported with any poll, it is shorthand for saying how big the 95% CI is around the polling result.

9.6 A large-sample CI for the difference between two proportions

- The same logic underlies the construction of a large-sample confidence interval for the difference between two proportions. Consider this example:
  - * In a Zogby Poll conducted with 1,203 likely voters nationwide between Oct 24-26, 2008, Barack Obama led John McCain, 52.5 percent to 47.5 percent, among those expressing a preference.
  - * This is a tracking poll. In the previous three-day window of the poll (Oct 21-23), Obama led McCain 55.6 to 44.4 percent (N=1,203).
  - * According to the poll, Obama's lead shrunk by about six points in three days. How confident are we that this change is not due to sampling error?
  - * Set it up:
  - * The parameter we seek is now $p_1 - p_2$, where $p_1 =$ Obama's true support in the first poll (Oct 21-23) and $p_2 =$ Obama's true support in the second poll.
  - * The polls may be considered two binomial experiments in which Y_1 is the number of "successes" (here, the # favoring Obama) in the first poll, (no ideological agenda) and Y_2 is the number of of such "successes" in the second poll.
  - * An intuitive estimator for this quantity would be $\widehat{p}_1 - \widehat{p}_2$, where the p-hats are the proportions of respondents favoring Obama in the two polls. Is it an unbiased

estimator for $p_1 - p_2$?

$$\begin{aligned} E(\widehat{p}_1 - \widehat{p}_2) &= E(\widehat{p}_1) - E(\widehat{p}_2) \\ &= E\left(\frac{Y_1}{n_1}\right) - E\left(\frac{Y_2}{n_2}\right) \text{ [}\widehat{p}_1 \text{ and } \widehat{p}_2 \text{ are functions of the RVs } Y_1, Y_2\text{]} \\ &\quad \frac{1}{n_1}E\left(Y_1\right) - \frac{1}{n_2}E\left(Y_2\right) \\ &= \frac{1}{n_1}n_1p_1 - \frac{1}{n_2}n_2p_2 \text{ [}E\left(Y\right) = np \text{ if } Y \text{ is distributed binomial]} \\ &= p_1 - p_2. \end{aligned}$$

* Our next step is to say how precise $\widehat{p}_1 - \widehat{p}_2$ tends to be as an estimator of $p_1 - p_2$.
* We do this by figuring out what the estimator's standard error is. It's

$$\begin{aligned} \sqrt{VAR(\widehat{p}_1 - \widehat{p}_2)} &= \sqrt{VAR(\widehat{p}_1) + VAR(\widehat{p}_2)} \text{ [assume samples drawn independently]} \\ &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \end{aligned}$$

* We make the substitution

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

* Plugging in, we have

$$\begin{aligned} &(55.6 - 52.5) \pm z_{\alpha/2}\sqrt{\frac{(55.6)\,(100-55.6)}{1,203} + \frac{(52.5)(100-52.5)}{1,203}} \\ &3.1 \pm z_{\alpha/2}(2.031). \end{aligned}$$

* Do you recall how we find $z_{\alpha/2}$? We type **display invnormal($\frac{\alpha}{2}$)**, substituting our chosen α**.** You'll remember that $z_{\alpha/2}$ associated with an $\alpha = .05$ is $z_{.025} = -1.96$. So our 95% CI is:

$$3.1 \pm 1.96(2.031) = 3.1 \pm 3.98, \text{ or } [-.9, 7.1].$$

* We are 95% confident that the true change between the two polls was between -.9

and 7.1 percentage points.

- Note that this CI includes zero. So another interpretation of this CI is that we are **not** 95% confident that there was zero change between the two polls. And this, of course, is what we really wanted to know: was there truly any movement between Oct 21-23 and Oct 24-26?
- Now, does the 90% confidence interval about our point estimate include zero?
  - Let's see: our alpha is .10.
  - typing **display invnormal(**.05**)** gives us -1.64. So our 90% CI is:

$$3.1 \pm 1.64(2.031) = 3.1 \pm 3.33, \text{ or } [-.23,\ 6.43].$$

- Still no cigar. At what level of confidence would we be satisfied that there was movement between the two surveys?
- Think: we wish to find some α^* such that the lower bound of the $100 * (1 - \alpha)$ CI is greater than zero. That is, find some α^* meeting this criterion:

$$\alpha^* : 3.1 - z_{\alpha^*/2}(2.031) > 0.$$

- To do this, manipulate the expression

$$\begin{aligned} -z_{\alpha^*/2}(2.031) &> -3.1 \\ z_{\alpha^*/2} &< \frac{3.1}{2.031} \\ z_{\alpha^*/2} &< 1.5263 \end{aligned}$$

- So for any alpha such that $z_{\alpha/2} < 1.5263$, we will be $100 * (1 - \alpha)$ percent confident that the true change was greater than zero. How do we find this α? Well, if

$$\begin{aligned} z_{\frac{\alpha}{2}} &= -\Phi^{-1}\left(\frac{\alpha}{2}\right), \text{ then} \\ \Phi\left(-z_{\frac{\alpha}{2}}\right) &= \frac{\alpha}{2}, \text{and} \\ \alpha &= 2\Phi\left(-z_{\frac{\alpha}{2}}\right). \end{aligned}$$

- So in this particular case, $\alpha = 2\Phi(-1.5263)$.
  - To find this alpha, we now type **di normal(-1.5263)** in Stata, which is the CDF of the standard Normal evaluated at its argument. This returns **.063**.
  - Thus $\alpha/2 = .063$ and alpha is thus .126.
  - And thus if we are working with confidence intervals of $100 * (1 - .126) = 87.4\%$ or smaller, we will conclude that there was true movement between the two polls.
- Keep this in mind: it will connect to other concepts we'll be covering today and next lecture.

9.7 Hypothesis Testing

- This way of framing the question motivates a process known as **hypothesis testing**. A hypothesis test consists of four elements:
  1. A **null hypothesis about a parameter**, which we write as H_0.
     - This is typically either what the "conventional wisdom" says is the the value of the parameter–or that the parameter is equal to zero.
  2. An **alternative hypothesis about the parameter**, H_A.
     - This is typically that the parameter is equal to *something different* than the null hypothesis. It may be more specific: that the parameter is either greater than or less than the null hypothesis.
  3. A **test statistic derived from an estimator of the parameter.**
  4. A **rejection region**.
     - The RR specifies the range of values of the test statistic for which the null H_0 is to be *rejected* in favor of the alternative H_A.
- Choosing the rejection region:
  - RR's are associated with two kinds of error:
    - Type I error (a.k.a. a "false positive") is made if H_0 is rejected when H_O is actually true.

· $Pr(\text{Type I error}) = \alpha$. (Yes, the very same α we've been working with.)

* Type II error (a.k.a. a "false negative") is made if H_0 is accepted when H_A is actually true.

· $Pr(\text{Type II error}) = \beta$.

– α and β are two very practical ways to measure the goodness of a statistical test. We call α the test's **level of significance**. We call the quantity $1 - \beta$ the test's **statistical power**. In the best of all worlds, we want a test's level of significance to be low and its power to be high. In reality, we always face a tradeoff between these two goals.

– To illustrate this tradeoff, consider the data from which we constructed the earlier CI about Obama and McCain. Let's re-pose this question in terms of a hypothesis test, where

$$\begin{aligned} H_0 &: p_1 - p_2 = 0 \\ H_A &: p_1 - p_2 > 0 \end{aligned}$$

– Here our *test statistic* is the difference between our two sample proportions, $\widehat{p}_1 - \widehat{p}_2$. And our rejection region includes the values of the statistic for which we reject the null for our chosen α.

* Here, the rejection region are those values of $\widehat{p}_1 - \widehat{p}_2$ for which the constructed CI does not include zero. This would lead us to say (with $100 * (1 - \alpha)\%$ confidence) that the change between the two polls was greater than zero.

* What is this region? Let's look at our CI again:

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

* Had $(\widehat{p}_1 - \widehat{p}_2)$ been big enough that $(\widehat{p}_1 - \widehat{p}_2) - z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} > 0$, then

our CI would not have incorporated zero. That is, if

$$\begin{aligned}(\widehat{p}_1 - \widehat{p}_2) &> z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \\ \frac{(\widehat{p}_1 - \widehat{p}_2)}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}} &> z_{\alpha/2},\end{aligned}$$

we should reject the null and accept H_A.

* So, a few questions:
* what sample sizes would we have needed for our difference in sample proportions $(\widehat{p}_1 - \widehat{p}_2)$ to have been found statistically different from zero with 95% confidence? (Assume n_1=$n_2 = n$.)

$$\begin{aligned}\frac{3.1}{\sqrt{\frac{(55.6)(100-55.6)}{n} + \frac{(52.5)(100-52.5)}{n}}} &> 1.96 \\ \frac{3.1}{\sqrt{\frac{4962.4}{n}}} &> 1.96 \\ \frac{3.1}{1.96} &> \sqrt{\frac{4962.4}{n}} \\ \left(\frac{3.1}{1.96}\right)^2 &> \frac{4962.4}{n} \\ n &> 1,983.7\end{aligned}$$

* We would have needed two samples of at least 1,984 in size.

9.8 Hypothesis tests vis-a-vis confidence intervals

- Just to be clear about what we're doing here:
  - We have an estimate (typically $\overline{Y}$ for the parameter μ, but let's be agnostic about this and call it the estimator $\widehat{\theta}$ for the parameter θ).
  - We'd like to say something about the confidence we have in our estimate.
  - One way to do this is to construct a confidence interval around the estimate with confidence coefficient $1 - \alpha$. This is the probability that the process used to contruct the CI will trap the parameter in repeated samples.

- Recall that the way we did this was:
  - we know that CLT tells us that the standardized version of *any* estimator $\widehat{\theta}$ that is a linear combination of i.i.d.random variables is distributed standard Normal in large samples: $\frac{\widehat{\theta}-E\left(\widehat{\theta}\right)}{\sigma_{\widehat{\theta}}} \tilde{} N(0,1)$.
    - we have observed $\widehat{\theta}$
    - if $\widehat{\theta}$ unbiased than by definition $E\left(\widehat{\theta}\right) = \theta$.
    - and we've got various ways to get $\sigma_{\widehat{\theta}}$ from our data.
  - so rewrite as $\frac{\widehat{\theta}-\theta}{\sigma_{\widehat{\theta}}} \tilde{} N(0,1)$.
  - Now we can back out a CI for θ, the thing we care about, since:

$$\begin{aligned}
\Pr\left(a \leq \frac{\widehat{\theta}-\theta}{\sigma_{\widehat{\theta}}} \leq b\right) &= \Pr\left(a\sigma_{\widehat{\theta}} \leq \widehat{\theta}-\theta \leq b\sigma_{\widehat{\theta}}\right) \\
&= \Pr\left(a\sigma_{\widehat{\theta}} - \widehat{\theta} \leq -\theta \leq b\sigma_{\widehat{\theta}} - \widehat{\theta}\right) \\
&= \Pr\left(\widehat{\theta} - a\sigma_{\widehat{\theta}} \geq \theta \geq \widehat{\theta} - b\sigma_{\widehat{\theta}}\right) \\
&= \Pr\left(\widehat{\theta} - b\sigma_{\widehat{\theta}} \leq \theta \leq \widehat{\theta} - a\sigma_{\widehat{\theta}}\right)
\end{aligned}$$

  - Then we take advantage of the fact that $a = -b$ in our particular case and use this to choose the $z_{\frac{\alpha}{2}}$ that goes with our confidence coefficient $(1-a)$, and construct the CI as

$$\begin{aligned}
\Pr\left(\widehat{\theta} - z_{\frac{\alpha}{2}}\sigma_{\widehat{\theta}} \leq \theta \leq \widehat{\theta} + z_{\frac{\alpha}{2}}\sigma_{\widehat{\theta}}\right) &= \Phi\left(-z_{\frac{\alpha}{2}}\right) - \Phi\left(z_{\frac{\alpha}{2}}\right), \text{ or} \\
\Pr\left(\widehat{\theta}_{LB} \leq \theta \leq \widehat{\theta}_{UB}\right) &= 1-\alpha.
\end{aligned}$$

- But another way to do this is to pose the following question:
  - Having obtained a point estimate $\widehat{\theta}$, how sure am I now that θ is not equal to some value I care about, θ_0?
  - This is the question we ask when we conduct *hypothesis tests*.
  - Often this value is $\theta_0 = 0$, but in practice it can be any value.
  - The link between the two methods is this: we reject the "null hypothesis," that $\theta = \theta_0$ in

favor of the "alternative hypothesis" $\theta \neq \theta_0$ when the CI we construct does not include θ_0.

- Use diagrams on board.
- In this context, the process by which we decide whether to accept or reject the null is known as a *two-tailed test*. You'll see why in a minute.

9.9 A One-Tailed Hypothesis Test

- Now let's turn it around a bit.
- Consider the case where θ equals some hypothesized value of θ, θ_0.
- If this were true, than our unbiased estimator of θ, $\widehat{\theta}$ would be distributed Normal around θ_0 with standard deviation $\sigma_{\widehat{\theta}}$. (Draw "One-Tailed Test" on diagram on board.)
- Pick any value k. What is $\Pr(\widehat{\theta} > k|\theta = \theta_0)$?
- Well, we know from before that that $\frac{\widehat{\theta}-\theta}{\sigma_{\widehat{\theta}}} \tilde{} N(0,1)$ and so $\Pr\left(\frac{\widehat{\theta}-\theta}{\sigma_{\widehat{\theta}}} \geq k\right) = 1 - \Phi(k)$.
- And so in the case where $\theta = \theta_0$, $\Pr\left(\frac{\widehat{\theta}-\theta_0}{\sigma_{\widehat{\theta}}} \geq k|\theta = \theta_0\right) = 1 - \Phi(k)$.
- That is, given that $\theta = \theta_0$, our estimator will be greater than k $[1 - \Phi(k)] \times 100)\%$ of the time due to chance variation, while it will be less than k $\Phi(k) \times 100\%$ of the time.
- We like to talk about these things in terms of α [add shaded region to diagram].
- And so if we wanted to specify a significance level α, then we already have an expression for k : It's $\theta_0 + z_\alpha \sigma_{\widehat{\theta}}$.
- So if we are considering a null $H_0 : \theta = \theta_0$ and an alternative hypothesis $H_A : \theta_0 > \theta$,then we reject the null with $(1 - \alpha) \times 100\%$ confidence if we see $\widehat{\theta} > \theta_0 + z_\alpha \sigma_{\widehat{\theta}}$.
- Again, the logic here is that it's quite unlikely we would have observed a $\widehat{\theta}$ so high if θ were truly equal to θ_0. In fact, it would only happen $\alpha \times 100$ percent of the time.
- On the other hand if we see $\widehat{\theta} < \theta_0 + z_\alpha \sigma_{\widehat{\theta}}$, we accept the null and rule out the alternative that $\theta_0 > \theta$ with $(1 - \alpha) \times 100\%$ confidence.

- So to recap. In assessing whether it is possible that if $\theta = \theta_0$, we would see H_A sheerly by chance more than (1-α) percent of the time.
  - If so, we reject H_A, because we can't rule out that our result was due to chance variation.
  - If not, we accept H_A, because we are 100*(1-α)% sure that our result was not due to chance variation.
- Now draw ONE-TAILED and TWO-TAILED HYPOTHESIS TESTS diagrams on board and discuss
- So here's the thing. Let's say you have a test statistic (some realization of $\widehat{\theta}$) whose value is greater than θ_0. You make a *ex post* ("based on actual results") hypothesis that $H_A : p_1 - p_2 \geq 0$ and conduct a one-tailed hypothesis test. This hypothesis is not based on theory. Are you cooking the books?
  - Yes. Knowing that $\widehat{\theta} > \theta_0$, to reject H_0 with a one-tailed test, you need
$$\widehat{\theta} > z_{\alpha}.$$
  - But to reject H_0 with a two-tailed test, you need
$$\widehat{\theta} > z_{\frac{\alpha}{2}}.$$
  - Typically political scientists are skeptical of one-tailed tests because they can look awfully post hoc. Most of the hypothesis tests you'll see in journals are two-tailed.

10 Lecture 11

10.1 Calculating the Power of a Hypothesis Test

- It is relatively easy to calculate the power of the the kinds of hypothesis tests we have been discussing.

- You'll recall that a test's statistical power is $1-\beta$. It is the probability that the test will falsely reject a positive result, or the probability of the commission of a Type II error.
- Go over handout called "Type I, Type II error."
- So how do we calculate β?

$$\begin{aligned}\beta &= \Pr\left(\text{Reject } H_0 | H_A \text{ true}\right) \\ &= \Pr\left(\widehat{\theta} < \theta_0 + z_\alpha \sigma_{\widehat{\theta}} | \theta = \theta_A\right)\end{aligned}$$

- Well we know that $\widehat{\theta}$ is distributed Normal with mean θ_A and standard deviation $\sigma_{\widehat{\theta}}$. So we therefore know the probability with which it will take on any value. First standardize $\widehat{\theta}$:

$$\begin{aligned}&= \Pr\left(\frac{\widehat{\theta}-\theta_A}{\sigma_{\widehat{\theta}}} < \frac{\theta_0 + z_\alpha \sigma_{\widehat{\theta}} - \theta_A}{\sigma_{\widehat{\theta}}} | \theta = \theta_A\right) \\ &= \Phi\left(\frac{\theta_0 + z_\alpha \sigma_{\widehat{\theta}} - \theta_A}{\sigma_{\widehat{\theta}}}\right). \\ &= \Phi\left(\frac{\theta_0 - \theta_A}{\sigma_{\widehat{\theta}}} + z_\alpha\right)\end{aligned}$$

- So a test's power is therefore

$$1 - \Phi\left(\frac{\theta_0 - \theta_A}{\sigma_{\widehat{\theta}}} + z_\alpha\right).$$

- Note that we have all the quantities necessary to calculate β and therefore power:
  - We've specified a θ_0 and θ_A.
  - We've also specified a α and therefore a z_α.
  - And we use our usual methods to obtain $\sigma_{\widehat{\theta}} = \frac{\sigma}{\sqrt{n}}$.
    - (So we might actually write:

$$Power = 1 - \Phi\left(\frac{\theta_0 - \theta_A}{\frac{\sigma}{\sqrt{n}}} + z_\alpha\right).$$

- Now for something interesting. What is the sign of $\frac{\partial \text{Power}}{\partial \alpha}$? Of $\frac{\partial \text{Power}}{\partial \sigma}$? Of $\frac{\partial \text{Power}}{\partial n}$? Of $\frac{\partial \text{Power}}{\partial(|\theta_0 - \theta_A|)}$?

- NOTE TO SELF: Note that $\theta_0 - \theta_A < 0$!
- Well, since Φ is monotonically increasing in its argument, we know that:
  - $\frac{\partial z_\alpha}{\partial \alpha} < 0$, and so $\frac{\partial \Phi\left(\frac{\theta_0-\theta_A}{\frac{\sigma}{\sqrt{n}}}+z_\alpha\right)}{\partial \alpha} < 0$, and so $\frac{\partial \text{Power}}{\partial \alpha} > 0$. As we increase α (that is, decrease our confidence coefficient) we increase power.
  - $\frac{\partial \frac{\theta_0-\theta_A}{\frac{\sigma}{\sqrt{n}}}}{\partial \sigma} = \frac{\partial \frac{(\theta_0-\theta_A)\sqrt{n}}{\sigma}}{\partial \sigma} > 0$ (since $\theta_0 - \theta_A < 0$), and so $\frac{\partial \Phi\left(\frac{\theta_0-\theta_A}{\frac{\sigma}{\sqrt{n}}}+z_\alpha\right)}{\partial \sigma} > 0$, and so $\frac{\partial \text{Power}}{\partial \sigma} < 0$. As there is more variance in the population, our estimator becomes less precise, and the power of our statistical tests goes down.
  - Conversely, $\frac{\partial \frac{\theta_0-\theta_A}{\frac{\sigma}{\sqrt{n}}}}{\partial n} < 0$, and so $\frac{\partial \text{Power}}{\partial n} > 0$.
  - And finally $\frac{\partial \text{Power}}{\partial(|\theta_0-\theta_A|)} = \frac{\partial \text{Power}}{\partial(\theta_A-\theta_0)}$(since $\theta_0 - \theta_A < 0$). Since $\frac{\partial \Phi\left(\frac{\theta_0-\theta_A}{\frac{\sigma}{\sqrt{n}}}+z_\alpha\right)}{\partial(\theta_A-\theta_0)} < 0$, $\frac{\partial \text{Power}}{\partial(|\theta_0-\theta_A|)} > 0$. The farther way you specify an alternative hypothesis away from the null, the less likely you are to falsely reject the null.
- Do example 10.8 in book (p. 508).

11 Lecture 12

11.1 Another way to report results of a statistical test: *p*-values

- You'll recall that α, the probability of a Type I ("false positive") error associated with a statistical test, is often called the test's **significance level.**
- In the hypothesis testing regime we've discussed so far, we:
  - pick a significance level
  - determine the critical value(s) of the test statistic associated with the significance level
  - and then determine whether to accept or reject the null hypothesis by comparing our test statistic with the critical value.

- This is all very black-or-white: the result is either significant or it's not. But if report only whether we reject or accept the null, we're actually failure to report a fair amount of information.
- Another way to report results of a statistical test that provides the reader with this additional information is to report what's called its p-value.
  - For any test statistic, the *p-value*, or **attained significance level**, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.
- The smaller a p-value is, the more compelling is the evidence that the null hypothesis should be rejected. That is because the null should be rejected for any value of α *down to and including* the p-value.
- Reporting the p-value permits the reader to make her own choice about whether the observed data justify a rejection of the null.
  - Do example 10.7 (p. 501) and then example 10.11 (p. 515)
- This should be obvious, but:

$$\begin{aligned} \text{Reject } \mathrm{H}_0 &\iff p \leq \alpha \\ \text{Accept } \mathrm{H}_0 &\iff \alpha \leq p \end{aligned}$$

11.2 A final topic: small-sample significance tests

- All of our confidence interval building and hypothesis testing has assumed that we have a sample size large enough to be reasonably sure that the CLT applies and our test statistic's sampling distribution approximates the Normal.
- But what happens when our samples are pretty small? Essentially, we need to make some adjustments to the sampling distribution to account for this.

- For the first time in this class, we'll make the simplifying assumption that we are drawing samples from a Normally distributed population (we'll return in a few moments to consider what happens when this assumption is violated). So we assume:
  - $Y_1, Y_2, ...Y_n$ represent a random sample drawn from a Normal population, with $\overline{Y}$ and S_U^2 as the sample mean and our (unbiased) estimator of the population variance, respectively.
  - Goal is to construct a CI for μ (or, equivalently, to conduct hypothesis tests) when $VAR(Y_i) = \sigma^2$ is unknown and the sample size is small.
  - To do this, we need to be able to say something about the sampling distribution of $\overline{Y}$. Again, because we don't have enough n, we can't appeal to the CLT and conclude that it is distributed Normal.
  - What to do instead?
- Well, start with the theorem (proof omitted, see Ch. 6 if you care) that a linear combination of independent, Normally distributed RVs is itself Normally distributed.
- By linear combination, we mean a random variable composed of the sum of the products of a total of J RVs and scalars:
$$\sum_{i=1}^{J} a_i Y_i$$
- $\overline{Y}$ is, of course, one such linear combination, where the a_i 's are each just equal to $\frac{1}{n}$.
- Thus the sampling distribution of $\overline{Y}$ is Normal, as as before we know that $E\left(\overline{Y}\right) = \mu_{\overline{Y}} = \mu$ and $VAR\left(\overline{Y}\right) = \frac{\sigma^2}{n}$.
- Now let's standardize each of the Y_i's, with $Z_i = \frac{Y_i - \mu}{\sigma}$. Now consider the sum of their squares:
$$\sum_{i=1}^{n} Z_i^2 = \sum_i \left(\frac{Y_i - \mu}{\sigma}\right)^2$$
- This sum of squares takes on what's called a **chi-squared $\left(\chi^2\right)$ distribution with n degrees of freedom.**

- More generally, the sum of the squares of any n i.i.d. standard Normal random variables is distributed chi-squared with n degrees of freedom.

  - Just so you know, the chi-squared distribution–like any probablity distribution–has a density function. It happens to look like this:

$$f(x) = \frac{-2\left(\frac{\nu}{2}\right)}{\Gamma\left[\frac{\nu}{2}\right]} x^{\left(\frac{\nu-2}{2}\right)} e^{-\frac{x}{2}},$$

    where $\Gamma[\alpha]$ is the gamma function and $\Gamma[\alpha] = \int_0^\infty e^{-u} u^{\alpha-1} du$, and here ν ("nu") is the number of degrees of freedom.

  - This is very complicated. To make more simple, here is what you need to know about the chi-square (draw pdf of chi-square on board):

  - * as df increase, chi-square approaches the Normal distribution.
    * The expected value of a chi-square RV is its number of degrees of freedom: $E(X) = E\left(\sum_i Z_i^2\right) = \nu$.
    * For $\nu > 2$, the density function peaks at $\nu - 2$.

11.3 Degrees of freedom

- – By the way, "degrees of freedom" is a characteristic of any statistic that signifies the number of independent pieces of information on which the statistic is based. In general, the degrees of freedom associated with an estimate is equal to the number of pieces of data you have (generally, n) minus the number of parameters needed to generate the estimate.

  - Another way to think about it is that a statistic's degrees of freedom is also equal to the number of values in the final calculation of a statistic that are "free to vary."

  - For example, if I calculate the mean of three observations, Y_1 and Y_2,and Y_3 as follows:

$$\overline{Y} = \frac{Y_1 + Y_2 + Y_3}{n}$$

    this is a statistic with $n = 3$ degrees of freedom. I used three pieces of information to

calculate the statistic.

- Compare this to the canonical example is that an estimate of variance using n observations requires that we first estimate the mean and then calculate

$$\begin{aligned} S_U^2 &= \frac{\sum_i \left(\overline{Y} - Y_i\right)^2}{n-1}. \text{ but this is just:} \\ &= \frac{\left(\frac{Y_2+Y_3}{n}\right)^2 + \left(\overline{Y} - Y_2\right)^2 + \left(\overline{Y} - Y_3\right)^2}{n-1}. \end{aligned}$$

- Thus my calculation of S_U^2 really only uses two pieces of information, because once I've used the three observations to calculate Y-bar, the equation for S_U^2 is fully identified with any two of the observations. So the statistic S_U^2 has d.f. of $n-1$ associated with this estimate.

- Now to return:

- Recall that we are considering the case where $Y_1, Y_2, ...Y_n$ represent a random sample drawn from a Normal population, with $\overline{Y}$ and S_U^2 as the sample mean and our (unbiased) estimator of the population variance, σ^2, respectively. It turns out that the ratio

$$\frac{(n-1)\,S_U^2}{\sigma^2} = \frac{1}{\sigma^2}\sum\left(Y_i - \overline{Y}\right)^2$$

also has a χ^2 distribution with $n-1$ degrees of freedom (proofs omitted).

- Now consider a situation where we divide a standard Normal RV, Z, by the square root of the ratio of a chi-squared RV (call this W) divided by its degrees of freedom, ν: $\frac{Z}{\sqrt{W/\nu}}$. This is itself a random variable, and it turns out that when Z and W are independent, the quantity

$$T = \frac{Z}{\sqrt{W/\nu}}$$

has what's called a t distribution with ν degrees of freedom. Why do we care about the t

distribution? Well, doing a little math gets us:

$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{\frac{\left(\overline{Y}-\mu\right)}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S_U^2}{\sigma^2}/\left(n-1\right)}} = \sqrt{n}\left(\frac{\overline{Y}-\mu}{S_U}\right) = \frac{\overline{Y}-\mu}{\frac{S_U}{\sqrt{n}}}.$$

- But this of course is Y-bar minus its hypothesized mean divided by its estimated standard error. We now can say how the combination of three numbers we know–n, Y-bar, and our unbiased estimate of the standard deviation, S_U–are distributed around the hypothesized mean. $\frac{\overline{Y}-\mu}{\frac{S_U}{\sqrt{n}}}$ has a "t distribution" with $(n-1)$ d.f.
- That is, we can now fully describe the sampling distribution of Y-bar when we have a small sample of Normally distributed random variable.
- This is the key that allows us to now conduct hypothesis tests with small samples.
  - Again, just so you see it, the t-distribution has a density function that looks like this:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}\left(\nu+1\right)\right]\left(\sigma^2\nu\pi\right)^{-\frac{1}{2}}}{\Gamma\left[\frac{\nu}{2}\right]}\left(1+\frac{\left(x-\mu\right)^2}{\nu\sigma^2}\right)^{-\frac{1}{2}(\nu+1)},$$

    where μ and σ^2 are the population mean and variance of the RV Y.
- As ν approaches infinity, the t distribution approaches the standard Normal distribution. Draw picture on p. 360.
- These findings allow us to:
  - construct $100(1-\alpha)\%$ CI's around estimates of μ drawn from small samples;
  - perform hypothesis tests with these estimates; and
  - to do the same with estimates of $\mu_1 - \mu_2$.

- For example, for a CI around $\overline{Y}$, an estimate of μ, we proceed as follows:

$$\begin{aligned} P\left(-t_{\frac{\alpha}{2},\nu} \leq T \leq t_{\frac{\alpha}{2},\nu}\right) &= 1-\alpha, \text{ or} \\ P\left(-t_{\frac{\alpha}{2},\nu} \leq \frac{\overline{Y}-\mu}{\frac{S_U}{\sqrt{n}}} \leq t_{\frac{\alpha}{2},\nu}\right) &= 1-\alpha \\ P\left(-t_{\frac{\alpha}{2},\nu}\frac{S_U}{\sqrt{n}} - \overline{Y} \leq -\mu \leq t_{\frac{\alpha}{2},\nu}\frac{S_U}{\sqrt{n}} - \overline{Y}\right) &= 1-\alpha \\ P\left(t_{\frac{\alpha}{2},\nu}\frac{S_U}{\sqrt{n}} + \overline{Y} \geq \mu \geq -t_{\frac{\alpha}{2},\nu}\frac{S_U}{\sqrt{n}} + \overline{Y}\right) &= 1-\alpha \end{aligned}$$

- Thus the endpoints of the $100(1-\alpha)\%$ CI are: $\overline{Y} \pm t_{\frac{\alpha}{2},\nu}\frac{S_U}{\sqrt{n}}$. Note how the logic is the same as that which we use to construct the CI used for a large-sample, which is $\overline{Y} \pm z_{\frac{\alpha}{2}}\frac{S_U}{\sqrt{n}}$.

- How about hypothesis tests ? Well, I won't bore you with the details - everything's very similar to the large-sample test:

$$\begin{aligned} H_0 &: \mu = \mu_0, \\ H_A &: \mu > \mu_0, \mu < \mu_0 \text{ (one tailed)} \\ \mu &\neq \mu_0 \text{ (two-tailed)} \\ \text{Test statistic is:} & \quad T = \frac{\overline{Y}-\mu}{\frac{S_U}{\sqrt{n}}} \\ \text{Reject } H_0 \text{ if } t &> t_{\alpha,\nu} \text{ or } t > -t_{\alpha,\nu} \text{ (one-tailed)} \\ |t| &> t_{\frac{\alpha}{2},\nu} \text{ (two-tailed)} \end{aligned}$$

- And tests for $\mu_1 - \mu_2$?

- Very similar. We assume that our two samples are independent (as we do for large-sample tests). But we typically make an important additional assumption: that the variances of our two populations are the same. That is, $\sigma_1^2 = \sigma_2^2$. Two reasons:

  - a matter of convenience: with small samples it is difficult to get good estimates of small population variances (remember how we need lots of n for consistency property of S_U^2 to kick in)

- it's reasonable, since we're already assuming that both populations are Normal; we might as well assume that they have the same variance.

- You'll recall that the test statistic in the large-sample case was

$$Z = \frac{\overline{y}_1 - \overline{y}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

- But if we assume $\sigma^2 = \sigma_1^2 = \sigma_2^2$, then we can construct a consistent estimate σ^2 by taking a weighted average of the two sample variances. This weighted average is called "s-squared pooled" and calculated as

$$s_p^2 = \frac{(n_1 - 1)\, S_{U1}^2 + (n_2 - 1)\, S_{U2}^2}{n_1 + n_2 - 2}.$$

Thus our test statistic is calculated:

$$\begin{aligned} H_0 \quad &: \quad \mu_1 - \mu_2 = 0, \\ H_A \quad &: \quad \mu_1 - \mu_2 > 0,\ \mu_1 - \mu_2 < 0 \ \text{(one tailed)} \\ \mu_1 - \mu_2 \quad &\neq \quad 0 \ \text{(two-tailed)} \end{aligned}$$

$$\text{test statistic is: } T = \frac{\overline{y}_1 - \overline{y}_2 - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

$$\begin{aligned} \text{Reject } H_0 \text{ if } t \quad &> \quad t_{\alpha,\nu} \text{ or } t > -t_{\alpha,\nu} \ \text{(one-tailed)} \\ |t| \quad &> \quad t_{\frac{\alpha}{2},\nu} \ \text{(two-tailed),} \\ &\qquad \text{where we have } n_1 + n_2 - 2 \ \text{d.f.} \end{aligned}$$

- What if the underlying population isn't Normal?

- Statisticians have resorted to empirical studies, where they sample from populations of (known) nonnormal distributions.

  - Moderate departures from normality have little effect on the probability distribution of

the test statistic.

  - BUT we are in somewhat treacherous waters here.

- Two more things:

  - because the t is indistinguishable from the Normal at high d.f., a t-test is indisinguishable from a z-test at most levels of n with which we are used to working. This has led to the ubiquitousness of calling hypothesis tests about μ and $\mu_1 - \mu_2$ "t-tests," even though in most cases they are indistinguishable from z-tests.
  - We will revisit t-tests in the context of multivariate regression.

11.4 Summing up inference with one variable

We've spent the past few weeks thinking carefully about the inferences regarding one variable that we can make from a sample to a population. It's helpful to recap the steps we've taken to do this. Let's recap in terms of the goal of making inferences about a population mean μ from a sample statistic Y-bar.

- Assume (big assumption # 1) that we have a random sample, which yields independent, identically distributed observations.

  - *Identicality* assures us that our sample mean, Y-bar, is an unbiased estimator of μ.

- Now we want to know how precise our estimate is. We phrase this question as: how far off, on average, is our estimate typically going to be from the true mean?

- To do this, we need to know the (1) distribution of our estimator and (2) the parameters of the distribution of Y-bar.

- (1) As N becomes large, the Central Limit Theorem tells us that Y-bar is distributed Normal.

- (2) The Normal has two parameters: its mean μ and variance σ^2.

  - Because Y-bar is an unbiased estimator of μ, the mean of Y-bar is μ.

- If we make the assumption of *independence*, Y-bar's variance is $\frac{\sigma^2}{n}$ and its standard deviation is $\frac{\sigma}{\sqrt{n}}$. We have a consistent, unbiased estimator of σ, which is S_U. A combination of theories allows us to substitute S_U in our estimate of the estimator's standard deviation, that is, $\frac{S_U}{\sqrt{n}}$ for $\frac{\sigma}{\sqrt{n}}$. The resulting quantity $\frac{\overline{Y}-\mu}{\frac{S_U}{\sqrt{n}}}$ converges in probability to the Standard Normal.

- What if we have a small sample? We then make big assumption #2: that Y is distributed Normal. This yields a Y-bar that is distributed T–a distribution that approaches the Normal as N becomes large.

- We now know the pdf and cdf of Y-bar. It is Normal if N is large. And it is T if Y is Normal.

- We can now answer several related questions.

  - *How confident am I about my estimate of Y-bar?* To do this, I identify the values of Y located on either side of Y-bar by an equal distance that between them incorporate (confidence)% of the probability density. I am then "(confidence)% confident" that μ falls in the interval I've created.

  - *How confident that μ is not equal to some hypothesized value μ_0?* To answer this question with "(confidence)% confidence" , I see if μ_0 is found within the confidence interval associated with that level of confidence. If it is, then I am not "(confidence)% confident" that I can reject the null that $\mu = \mu_0$. If it's not, I reject the null at that level of confidence.

  - *At what level of confidence can I be sure that μ is not equal to some hypothesized value μ_0?* To answer this question, I find the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected. To do this, I find the largest confidence interval around Y-bar that does not contain μ_0. The proportion of the density under the curve contained in this interval is the p-value associated with the hypothesis test.