

Gov 2000 - 5. Univariate Inference (II)

Matthew Blackwell

Harvard University

mblackwell@gov.harvard.edu

Where are we? Where are we going?

- Last few weeks = how to produce a single, best estimate of some population parameter, drawing on our knowledge of probability.
- Now: what if we want to express uncertainty about estimates? What if we want to give a range of values instead of a single value? (Interval estimation)
- Also, what if we want to assess the plausibility of a particular hypothesis about our data (Hypothesis testing).
- This last goal is basically what we did in the Lady Tasting Tea example.

INTERVAL ESTIMATION

Interval estimation - what and why?

- Last week we talked about producing a single estimate of some population parameter. For example, we estimated the population difference in means using the sample difference in means. But that gave us one number. That might be our “best guess” about the true value, but what is the probability that our estimate will be equal to the true value? Zero!
- A different (complementary) approach to estimation is to produce a range of values that will contain the truth with some fixed probability.

- An **interval estimate** of the population mean, μ , consists of two bounds within which we expect μ to reside:

$$LB \leq \mu \leq UB$$

- A confidence interval is a kind of interval estimator that guarantees that the true value of the parameters will be in interval in some fixed proportion of random samples.
- How can we possibly figure out such an interval? We'll rely on the distributional properties of estimators that we learned last week. Today we'll focus on developing these types of interval estimators for the sample mean, but the same ideas apply to all estimators. Later in the course, we'll see how they apply to regression coefficients.

Estimating the standard error of the sample mean

- We're going to make heavy use of the sampling distribution of the sample mean to derive interval estimates, called confidence intervals, for the sample mean.
- By the CLT, we know that in large samples, we know that the sample mean will be approximately Normally distributed:

$$\bar{Y} \sim N(\mu, SE[\bar{Y}]^2)$$

- By the properties of probability limits from last week, if we have a consistent estimator, $\widehat{SE}[\bar{Y}] \xrightarrow{p} SE[\bar{Y}]$, then when n is large, we know can replace the true SE with the estimate in the approximating distribution:

$$\bar{Y} \sim N(\mu, \widehat{SE}[\bar{Y}]^2)$$

- For the sample mean, remember that the sampling variance is $\frac{\sigma^2}{n}$. Remember that we know n and we can estimate σ^2 with the sample variance:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- To get an estimate of the population standard deviation, we just use the sample standard deviation: $S_n = \sqrt{S_n^2}$
- We can plug this into our formula for the standard error the sample mean:

$$\widehat{SE}[\bar{Y}] = \frac{S}{\sqrt{n}}$$

Derivation of the Confidence Interval for the Population Mean

- **Definition** A $100(1 - \alpha)\%$ **confidence interval** with confidence level $1 - \alpha$ is an interval estimator for a population parameter θ that will contain/cover the true value, θ , $100(1 - \alpha)\%$ of the time.
- We going to use the above distributional results to come up with a rule for constructing a confidence interval. This rule will be an estimator just like the sample mean or the sample variance, but it will produce two values instead of one: the upper and lower values of the intervals.
- If Y_1, \dots, Y_n are iid and n is large, then we have

$$\begin{aligned}\bar{Y} &\sim N(\mu, \widehat{SE}[\bar{Y}]^2) \\ \bar{Y} - \mu &\sim N(0, \widehat{SE}[\bar{Y}]^2) \\ \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} &\sim N(0, 1)\end{aligned}$$

- One fact that you should memorize is that that 95% of the probability mass of a standard Normal falls between -1.96 and 1.96, so we know that:

$$\mathbb{P}(-1.96 \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq 1.96) = 0.95$$

- Let's work backwards to derive the confidence interval:

$$\begin{aligned}\mathbb{P}\left(-1.96 \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq 1.96\right) &= 0.95 \\ \mathbb{P}\left(-1.96 \times \widehat{SE}[\bar{Y}] \leq \bar{Y} - \mu \leq 1.96 \times \widehat{SE}[\bar{Y}]\right) &= 0.95 \\ \mathbb{P}\left(-1.96 \times \widehat{SE}[\bar{Y}] - \bar{Y} \leq -\mu \leq 1.96 \times \widehat{SE}[\bar{Y}] - \bar{Y}\right) &= 0.95 \\ \mathbb{P}\left(\bar{Y} - 1.96 \times \widehat{SE}[\bar{Y}] \leq \mu \leq \bar{Y} + 1.96 \times \widehat{SE}[\bar{Y}]\right) &= 0.95\end{aligned}$$

Different confidence intervals

- Confidence intervals take the following form: $\bar{Y} \pm 1.96 \times \widehat{SE}[\bar{Y}]$, which is shorthand for $[LB, UB]$ where these are:

$$\begin{aligned}LB &= \bar{Y} - 1.96 \times \widehat{SE}[\bar{Y}] \\ UB &= \bar{Y} + 1.96 \times \widehat{SE}[\bar{Y}]\end{aligned}$$

- Where did the 1.96 come from?

$$\mathbb{P}\left(-1.96 \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq 1.96\right) = 0.95$$

- What about confidence intervals other than 95%? Say we want an $(1 - \alpha)\%$ confidence interval:

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq z_{\alpha/2}\right) = (1 - \alpha)$$

$$\mathbb{P}\left(\bar{Y} - z_{\alpha/2} \times \widehat{SE}[\bar{Y}] \leq \mu \leq \bar{Y} + z_{\alpha/2} \times \widehat{SE}[\bar{Y}]\right) = (1 - \alpha)$$

- General formula for a $(1 - \alpha)\%$ confidence interval:

$$\bar{Y} \pm z_{\alpha/2} \times \widehat{SE}[\bar{Y}]$$

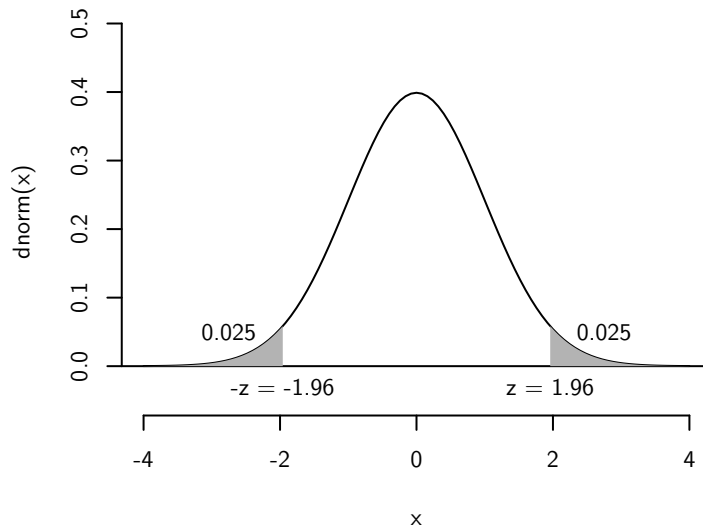
- Here we call the $z_{\alpha/2}$ values the z-values for the particular confidence intervals.

Finding the z values

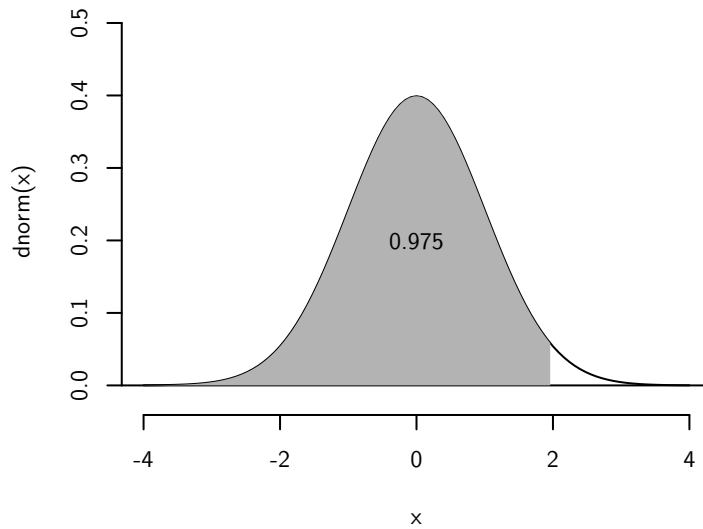
- How do we figure out what $z_{\alpha/2}$ will be? We will need to find the values such that the following holds. Given a standard Normal variable, $Z \sim N(0, 1)$:

$$\mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

- Thus, we want the values that puts $\alpha/2$ in each of the tails.
- For example, with $\alpha = 0.05$ for a 95% confidence interval, we want the z values that put 0.025 (2.5%) in each of the tails:



- But how did we find the 1.96 in this case? For these types of two-sided confidence intervals, we will simply find the z -value that puts $\alpha/2$ of the probability mass above it or, equivalently, $1 - \alpha/2$ below it. Thus, we want to know the value z , such that $\mathbb{P}(Z \leq z) = 1 - \alpha/2$:



- This is just the CDF of the standard Normal evaluated at $1 - \alpha/2$! We can find this using the `qnorm()` function in R.
- Procedure:

1. Choose a value α (0.1 for example) for a $100(1 - \alpha)\%$ confidence interval (90% in this case)
2. Convert this to $1 - \alpha/2$ (0.95 in this case)
3. Plug this value into `qnorm()` to find $z_{\alpha/2}$:

```
qnorm(0.95)
```

```
## [1] 1.644854
```

- **Question** What happens to the size of the confidence interval when we increase our confidence, from say 95% to 99%? Do confidence intervals get wider or shorter?

Confidence intervals you should probably memorize

- For a 90% CI, we have $\alpha = 0.1$ and $z_{0.05} = 1.64$
- For a 95% CI, we have $\alpha = 0.05$ and $z_{0.025} = 1.96$
- For a 99% CI, we have $\alpha = 0.01$ and $z_{0.005} = 2.58$

Interpreting the confidence interval

- **Caution** An often recited, but **incorrect** interpretation of a confidence interval is the following: “I calculated a 95% confidence interval for the sample mean of [0.05,0.13], which means that there is a 95% chance that the true population mean in is that interval.” This is WRONG.
- The true value of the population mean, μ , is **fixed**. It is either in the interval or it isn't—there's no room for probability at all. The randomness is in the interval: $\bar{Y} \pm 1.96 \times \hat{SE}[\bar{Y}]$. This is what varies from sample to sample.
- Correct interpretation: across 95% of random samples, the constructed confidence interval will contain the true value.

Confidence interval simulation

- Let's do a simulation to show how this works. We're going to draw samples of size 500 from a $N(1, 10)$ and then calculate the confidence interval for population mean and then calculate the proportion of the intervals that contain the true value of the population mean.

```

set.seed(02143)
sims<- 10000
cover <- rep(NA, times = sims)
low.bound <- up.bound <- rep(NA, times = sims)
for(i in 1:sims){
  draws <- rnorm(500, mean=1, sd=sqrt(10))
  low.bound[i] <- mean(draws) - sd(draws)/sqrt(500) * 1.96
  up.bound[i] <- mean(draws) + sd(draws)/sqrt(500) * 1.96
  if (low.bound[i] < 1 & up.bound[i] > 1) {
    cover[i] <- 1
  } else {
    cover[i] <- 0
  }
}
mean(cover)

```

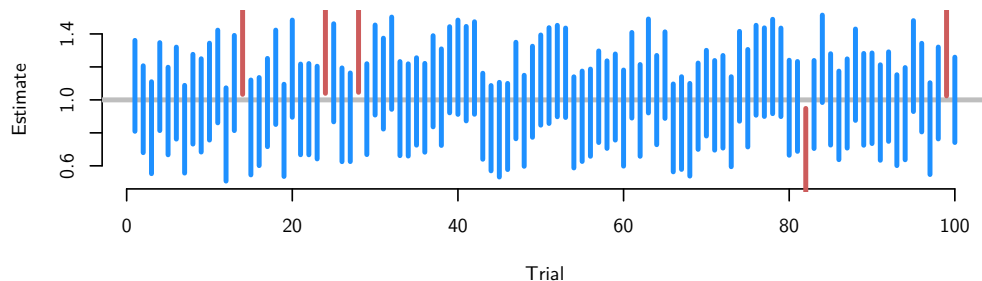
```
## [1] 0.9498
```

- We can see that roughly 95% of the confidence intervals (across the samples) contain the true value. Let's plot the first 100 of these interval estimates from the first 100 samples and then color code the intervals by whether they cover the true value:

```

cols <- ifelse(cover == 1, "dodgerblue", "indianred")
plot(x = NULL, y = NULL, xlim = c(1,100), ylim = c(0.5, 1.5), bty = "n",
     xlab = "Trial", ylab = "Estimate")
abline(h = 1, col = "grey", lwd = 3)
segments(x0 = 1:100, y0 = low.bound[1:100], y1 = up.bound[1:100],
        col = cols[1:100], lwd = 3)

```



- You can see that in these 100 samples, exactly 95 of the calculated confidence intervals contains the true value.

- **Question** What happens to the intervals when we increase n , the sample size? Do the intervals get shorter or longer? Do more or fewer of them contain the true value?

What makes a good confidence interval?

Small-sample confidence intervals

- What if n is not large? Then we can't rely on the large-sample properties of the sample mean like consistency and asymptotic Normality. We can only rely on finite sample properties unbiasedness and knowing that the standard error is σ/\sqrt{n} .
- To take a somewhat canonical example, let's say that we are trying to estimate the percent alcohol by volume of Guinness beer, which let's say that we know is $N(4.2, 0.09)$. Imagine that we could only take samples of six packs, so we can only use $n = 6$. What happens when we apply the usual, Normal-based CI formula to that data?

```
set.seed(02143)
sims<- 10000
cover <- rep(NA, times = sims)
low.bound <- up.bound <- rep(NA, times = sims)
for(i in 1:sims){
  draws <- rnorm(6, mean=4.2, sd=sqrt(0.09))
  low.bound[i] <- mean(draws) - sd(draws)/sqrt(6) * 1.96
  up.bound[i] <- mean(draws) + sd(draws)/sqrt(6) * 1.96
  if (low.bound[i] < 4.2 & up.bound[i] > 4.2) {
    cover[i] <- 1
  } else {
    cover[i] <- 0
  }
}
mean(cover)
```

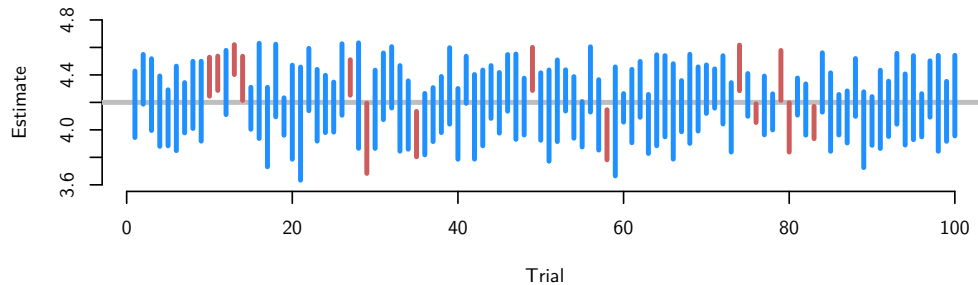
```
## [1] 0.8903
```

- Let's plot the first 100 confidence intervals again:


```

cols <- ifelse(cover == 1, "dodgerblue", "indianred")
plot(x = NULL, y = NULL, xlim = c(1,100), ylim = c(3.6, 4.8), bty = "n",
     xlab = "Trial", ylab = "Estimate")
abline(h = 4.2, col = "grey", lwd = 3)
segments(x0 = 1:100, y0 = low.bound[1:100], y1 = up.bound[1:100],
        col = cols[1:100], lwd = 3)

```



- Obviously, we can see here that far fewer than 95 of the 100 confidence intervals contains the true value. We call this **undercoverage**.
- But what is the distribution of \bar{Y} in small samples: $\bar{Y} \sim ?(\mu, \sigma^2/n)$
- If we're given no other information we actually don't know. But if know that the Y_1, \dots, Y_n are iid samples from $N(\mu, \sigma^2)$. With samples from the Normal distribution, then we know what the sampling distribution of the mean will be:

$$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

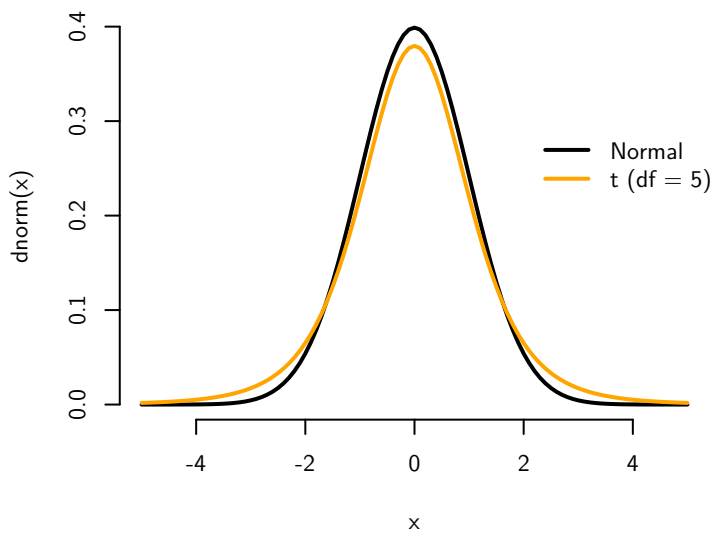
- Usually we have to estimate σ because we don't know it. Remember that in large samples this didn't affect the distribution of the estimator, but here it will:

$$\frac{\bar{Y} - \mu}{\frac{s_n}{\sqrt{n}}} \sim t_{n-1}$$

Student's t distribution

- Here, t_{n-1} is the **Student's t -distribution** (usually just called the t distribution) with $n - 1$ degrees of freedom (df).
- Named after **William Sealy Gossett** who published under the pen name, Student, while he was an employee at Guinness Brewery in Ireland. He developed the distribution while working on quality control problems in the brewery, where sample sizes could be quite small.

- The t distribution is completely summarized by its degrees of freedom, which here is dictated by the sample size. As the degrees of freedom increases, the t looks more and more like the standard Normal distribution.
- Similar shape to the Normal, but with fatter tails—it has slightly higher variance than the Normal. You can think of this extra variance as coming from the extra variance of estimating the denominator of $\frac{\bar{Y}-\mu}{\frac{s_n}{\sqrt{n}}}$. Sometimes the “extra” sampling error will push our estimates out toward the extremes.



- Let $t_{n-1, \alpha/2}$ be the value that put $\alpha/2$ of the mass of the t distribution above that value: $\mathbb{P}(T \geq t_{n-1, \alpha/2}) = \alpha/2$, where $T \sim t_{n-1}$. We can get this from the `qt()` function in R.
- For example, if we have a sample size of 6 and we wanted $\alpha = 0.05$. We know that we want to find the value such that 0.025 of the probability mass is above that value. Now, we know that this is the points that puts 0.975 below that value (`qt()` is the CDF, so it gives us the value such that a certain mass is below that value):

```
qt(0.975, df = 6 - 1)
```

```
## [1] 2.570582
```

- Note that this is different the value we would get with a Normal:

```
qnorm(0.975)
```

```
## [1] 1.959964
```

- Then,

$$\mathbb{P}\left(-t_{n-1,\alpha/2} \leq \frac{\bar{Y} - \mu}{\widehat{SE}[\bar{Y}]} \leq t_{n-1,\alpha/2}\right) = (1 - \alpha)$$

$$\mathbb{P}\left(\bar{Y} - t_{n-1,\alpha/2} \times \widehat{SE}[\bar{Y}] \leq \mu \leq \bar{Y} + t_{n-1,\alpha/2} \times \widehat{SE}[\bar{Y}]\right) = (1 - \alpha)$$

- Thus, when have small samples, but we know that the sample is from the Normal distribution, we can formulate confidence intervals like so: $\bar{Y} \pm t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}}$.
- Revisiting the simulation, let's change how we calculate the confidence interval, using the `qt()` function to get the critical value of the t distribution.

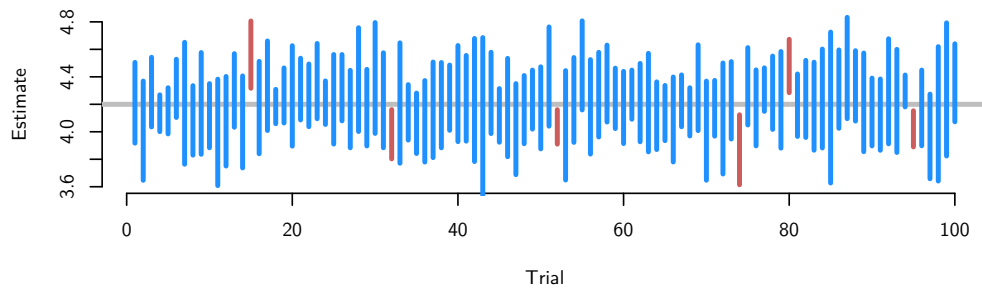
```
sims<- 10000
cover <- rep(NA, times = sims)
low.bound <- up.bound <- rep(NA, times = sims)
for(i in 1:sims){
  draws <- rnorm(6, mean=4.2, sd=sqrt(0.09))
  tval <- qt(0.975, df = 6 - 1)
  low.bound[i] <- mean(draws) - sd(draws)/sqrt(6) * tval
  up.bound[i] <- mean(draws) + sd(draws)/sqrt(6) * tval
  if (low.bound[i] < 4.2 & up.bound[i] > 4.2) {
    cover[i] <- 1
  } else {
    cover[i] <- 0
  }
}
mean(cover)
```

```
## [1] 0.9508
```

```

cols <- ifelse(cover == 1, "dodgerblue", "indianred")
plot(x = NULL, y = NULL, xlim = c(1,100), ylim = c(3.6, 4.8), bty = "n",
     xlab = "Trial", ylab = "Estimate")
abline(h = 4.2, col = "grey", lwd = 3)
segments(x0 = 1:100, y0 = low.bound[1:100], y1 = up.bound[1:100],
        col = cols[1:100], lwd = 3)

```



- Here we can see that our coverage is back on track even though we have really small samples!

Interval estimation of the difference in means

- Suppose we have two iid samples (also independent from each other) from possibly different distributions:
 - $Y_{11}, Y_{12}, \dots, Y_{1n_1}$ with population mean μ_1 and population variance σ_1^2
 - $Y_{21}, Y_{22}, \dots, Y_{2n_2}$ with population mean μ_2 and population variance σ_2^2
- Last week we saw that the sample difference in means is unbiased so we know that $\mathbb{E}[\bar{Y}_1 - \bar{Y}_2] = \mu_1 - \mu_2$.
- Let's calculate the sampling variance of the difference in means:

$$\begin{aligned}
 \text{Var}[\bar{Y}_1 - \bar{Y}_2] &= \text{Var}[\bar{Y}_1] + \text{Var}[\bar{Y}_2] && \text{(independence + variance properties)} \\
 &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \text{Var}[Y_{1i}] + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \text{Var}[Y_{2i}] \\
 &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sigma_1^2 + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sigma_2^2 \\
 &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
 \end{aligned}$$

- Putting all of this together, we know that in large samples, we have the following:

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

- Using the logic from the last section, we can replace the true population variances with sample estimates:

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})$$

- Combining this with the arguments above, we can construct the following confidence interval for the difference in means in large samples:

$$(\bar{Y}_1 - \bar{Y}_2) \pm z_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Interval estimation of the population proportion

- Let's say that we have a sample of iid Bernoulli random variables, Y_1, \dots, Y_n , where each takes $Y_i = 1$ with probability π . Note that this is also the **population proportion** of ones. We have shown in previous weeks that the expectation of one of these variable is just the probability of seeing a 1: $\mathbb{E}[Y_i] = \pi$.
- The **variance of a Bernoulli random variable** is a simple function of its mean: $\text{Var}(Y_i) = \pi(1 - \pi)$.
- **Problem** Show that the sample proportion, $\hat{\pi} = \frac{1}{n} \sum_{i=1}^n Y_i$, of the above iid Bernoulli sample, is unbiased for the true population proportion, π , and that the sampling variance is equal to $\frac{\pi(1-\pi)}{n}$.
- Note that if we have an estimate of the population proportion, $\hat{\pi}$, then we also have an estimate of the sampling variance: $\frac{\hat{\pi}(1-\hat{\pi})}{n}$.
- Given the facts from the previous problem, we just apply the same logic from the population mean to show that the following confidence interval:

$$P\left(\hat{\pi} - z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = (1 - \alpha)$$

Gerber, Green, and Larimer experiment

- Let's go back to the Gerber, Green, and Larimer experiment from last week. Here are the results of their experiment:

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Let's use what we have learned up until now and the information in the table to calculate a 95% confidence interval for the difference in proportions voting between the Neighbors group and the Civic Duty group.
- You may assume that the samples within each group are iid and the two samples are independent.
- First, we know that the sample proportion for the Civic Duty $\hat{\pi}_C$ group will be approximately Normal around the population mean π_C with sampling variance $(\pi_C(1 - \pi_C))/n$. Same goes for the sample proportion in the Neighbors group, $\hat{\pi}_N$.
- Given that these are just sample means (of the underlying Bernoulli voter turnout variables), we can apply the difference in means results from above to show that the difference in proportions has the following distribution:

$$\hat{\pi}_N - \hat{\pi}_C \sim N \left(\pi_N - \pi_C, \frac{\pi_N(1 - \pi_N)}{n_N} + \frac{\pi_C(1 - \pi_C)}{n_C} \right)$$

- We can replace the variances with our estimates from the sample proportions:

$$\hat{\pi}_N - \hat{\pi}_C \sim N \left(\pi_N - \pi_C, \frac{\hat{\pi}_N(1 - \hat{\pi}_N)}{n_N} + \frac{\hat{\pi}_C(1 - \hat{\pi}_C)}{n_C} \right)$$

- Now, we can use this in the same way as every other example to calculate the 95% confidence interval:

$$(\hat{\pi}_N - \hat{\pi}_C) \pm 1.96 \times \sqrt{\frac{\hat{\pi}_N(1 - \hat{\pi}_N)}{n_N} + \frac{\hat{\pi}_C(1 - \hat{\pi}_C)}{n_C}}$$

- Let's plug in the values into R and go!

```
n.n <- 38201
samp.var.n <- (0.378 * (1 - 0.378))/n.n
n.c <- 38218
samp.var.c <- (0.315 * (1 - 0.315))/n.c
se.diff <- sqrt(samp.var.n + samp.var.c)
se.diff
```

```
## [1] 0.003435201
```

```
## lower bound
lb <- (0.378 - 0.315) - 1.96 * se.diff
lb
```

```
## [1] 0.05626701
```

```
## upper bound
ub <- (0.378 - 0.315) + 1.96 * se.diff
ub
```

```
## [1] 0.06973299
```

- Thus, the confidence interval for the effect is $[0.056267, 0.069733]$.

Interval estimation of the population variance

- **Challenge Problem** Suppose that you have an iid sample, Y_1, \dots, Y_n , from a population with variance, σ^2 . Derive a $(1 - \alpha)\%$ confidence interval for the sample variance, S_n^2 . (Hint: you may have to look in Wooldridge and learn about a new distribution.)

HYPOTHESIS TESTS

What is a hypothesis test?

- A hypothesis test is just an evaluation of a particular hypothesis about the population distribution.
- These are just **statistical thought experiments** where we assume that we know how the data generating process works (we know the population distribution) and we compare our observed data to the data we would expect in that thought experiment.
- Would our data be common or rare in the statistical thought experiment? The hypothesis test will formulate exactly how unusual our observed data would have to be in the thought experiment in order to “reject” the thought experiment as plausible.
- This is exactly like the Lady Tasting Tea.

What is a hypothesis?

- **Definition** A **hypothesis** is just a statement about population parameters.
- We might have hypotheses about causal inferences
 - Does social pressure induce higher voter turnout? (mean turnout higher in social pressure group compared to Civic Duty group?)
 - Do daughters cause politicians to be more liberal on women’s issues? (voting behavior different among members of Congress with daughters?)
 - Do treaties constrain countries? (behavior different among treaty signers?)
- We might also have hypotheses about other parameters:
 - Is the share of Hillary Clinton supporters more than 50%?
 - Are traits of treatment and control groups different?
 - Is there evidence of electoral manipulation (Iran HW question)?

Null and alternative hypotheses

- **Defintion** The **null hypothesis** is a proposed, conservative value for a population parameter. This is usually “no effect/difference/relationship.” We denote this hypothesis as $H_0 : \mu = a$.

- **Definition** The **alternative hypothesis** for a given null hypothesis is the research claim we are interested in supporting. Usually, “there is a relationship/difference/effect.” We denote this as $H_a : \mu \neq a$.
- Always mutually exclusive
- Causal inference: Does social pressure (Neighbors) mailer affect turnout?
 - H_0 : Social pressure doesn’t affect turnout (No difference in mean turnout between Neighbors group and Civic Duty group)
 - H_a : Social pressure affects turnout (Difference in mean turnout between Neighbors group and Civic Duty group)
- Parameters: presence of electoral manipulation?
 - H_0 : no electoral manipulation (7s in the last digit occur with probability 1/10)
 - H_a : electoral manipulation (proportion of 7s higher than 1/10)

General framework

- We need a set of rules to adjudicate between the null and alternative hypotheses.
- **Definition** We base these rules on a **test statistic**, $T(Y) = T(Y_1, \dots, Y_n)$, which is just some function of the data (a *statistic*) calculated to perform a hypothesis *test*.
- **Definition** The **rejection region** to the values of $T(Y)$ for which we would reject the null hypothesis.
- If $T(Y)$ is outside the rejection region, we retain (or fail to reject) the null hypothesis.
- The rejection region represents values of the test statistic that cast doubt on the validity of the null hypothesis

The Lady Tasting Tea redux

- Let’s go all the way back to the Lady Tasting Tea. Remember that the goal there was to determine if the Lady was able to determine if the tea was poured first or the milk was poured first.
- Simplification of the problem: the Lady tastes n cups of tea and either guesses the i th cup correctly $Y_i = 1$ or incorrectly, $Y_i = 0$.
- Parameter of interest: $\pi =$ probability of a correct guess.
 - $\pi = 0.5$ guessing at random
 - $\pi > 0.5$ able to discern tea-first vs milk-first

- $\pi < 0.5$ worse than random
- Null and alternative hypotheses:
 - $H_0 : \pi \leq 0.5$
 - $H_a : \pi > 0.5$
- Assume she tastes LOTS of cups, so n is large, so that we have $\bar{Y} \sim N(\pi, \sigma^2/n)$.
- If we have an estimate of the variance, like the sample variance S_n^2 , then we know (by the usual arguments in the confidence intervals section), that:

$$\frac{\bar{Y} - \pi}{S_n/\sqrt{n}} \sim N(0, 1)$$

- But now our statistic depends on the true parameter, π , so what do we do?
- Statistical thought experiment time! Calculate the test statistic *as if the null hypothesis were true*.
- **Definition** The distribution of the test statistic under the null is call the **null distribution**.
- We can use this distribution to tell how unusual/weird our sample was compared to the null distribution.

Type I and Type II errors

- **Definition** A **Type I** error is when we reject the null hypothesis when it is in fact true.
- Guess the Lady is discerning when she is just guessing.
- **Definition** A **Type II** error is when we fail to reject the null hypothesis when it is false.
- Guess that the Lady is just guessing when she is truly discerning.

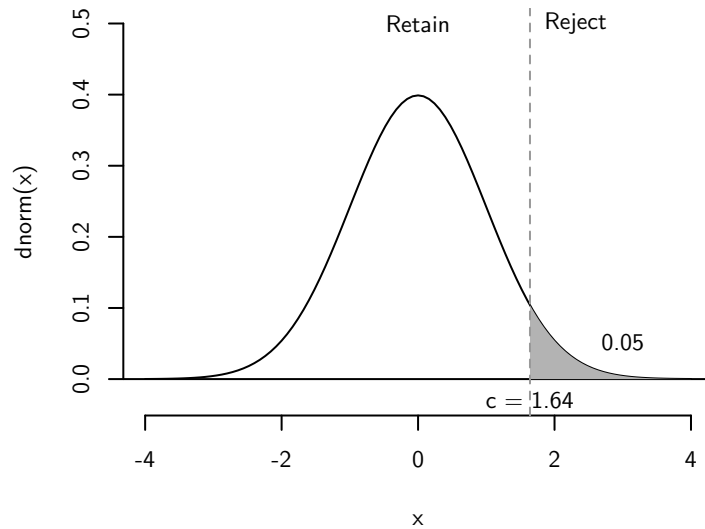
Size, power, and errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Definition** The **level of the test**, or α , is the probability of a Type I error.
- **Definition** The **rejection region** are the values of the test statistic for which we would reject the null hypothesis.
- Choose a level α based on aversion to false discovery:
 - Convention in social sciences is $\alpha = 0.05$, but nothing magical there
 - Particle physicists at CERN use $\alpha \approx \frac{1}{1,750,000}$
 - Higher values guard against “flukes” but increase barriers to discovery

Deriving the rejection region

- Let's go to our statistical thought experiment where we are assuming that the null is true, so that $\mu = 0.5$.
- Then with the Lady Tasting Tea: $T(Y) = \frac{\bar{Y}-0.5}{S_n/\sqrt{n}} \sim N(0, 1)$.
- When should we reject our null ($H_0 : \pi \leq 0.5$)? When the test statistic is consistent with the alternative ($H_a : \pi > 0.5$)! Thus, when $T(Y)$ big.
- We want to choose a rejection region that rejects the null in α of the cases when the null is true. So we want to find the rejection region that will contain α of the random draws of the test statistic.
- Are the guesses of the Lady sufficient good to make us reject the null?
- **Definition** The **critical value**, c is the value that marks the start of the rejection region.
- Let's say that we were happy with $\alpha = 0.05$ in this case. We would need to find the value above which there would only be 0.05 chance of seeing the test statistic: $P(T(Y) > c) = 0.05$. Since we know this is normal, we can plot it:



- To determine the critical value here, use the fact that $\mathbb{P}(T(Y) \leq c) = 1 - \alpha$, so that we can use the CDF functions in R to calculate the critical values:

```
qnorm(0.95)
```

```
## [1] 1.644854
```

- Note that we are assuming large samples, but if we have small samples, then we have to rely on the t distribution in a similar way to with confidence intervals. There, we just use the t instead of the Normal to derive critical values.

What happens under the null (stays under the null?)

```
samp.teststats <- rnorm(100, mean = 0, sd = 1)
c.val <- qnorm(0.95)
mean(samp.teststats > c.val)
```

```
## [1] 0.07
```

p-values

- Definition** The **p-value** for an observed test statistic is the probability of seeing a test statistic at least as extreme as observed statistic, *given that the null hypothesis is correct*.

- “Extreme” is determined by the alternative hypothesis.
- Alternative: the smallest value of α that we could still reject the null.
- What p-values are **not**:
 - An indication of a large substantive effect
 - The probability that the null hypothesis is false
 - The probability that the alternative hypothesis is true
- Just the $\mathbb{P}(T(Y) > c|H_0)$.

One-sided tests

- In the Lady Tasting Tea (redux) example, we are interested in a one-sided test.
- **Definition A one-sided test** is a test of an alternative hypothesis that only goes in one direction.
- What does this mean? It means that only deviations from the null hypothesis in one direction cast doubt on the null hypothesis. In the LTT(r) example, the rejection region only ended up being large values of $T(Y)$.
- But if the null is that she’s randomly guessing, why not allow for her being discerning in the opposite direction as evidence against the null? That is, why don’t we have the following alternative: $H_a : \pi \neq 0.5$.
- This **two-sided alternative** is almost always the right thing to do in the social sciences. One-sided tests are usually used to cheat—easier to reject without changing α .

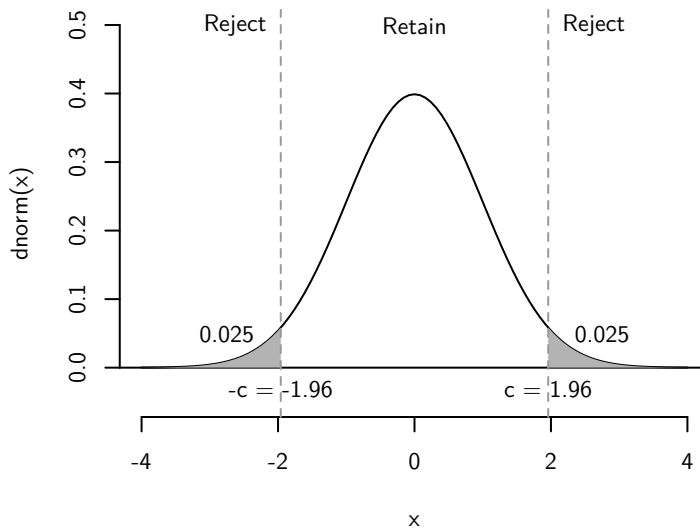
Example of a two-sided test: Social pressure experiment

- Let’s go back to the social pressure mobilization experiment. Again, remember that we are trying to see if there was an effect of the Neighbors treatment relative to the Civic Duty treatment on turnout.
- π_N is the proportion of the Neighbors group that turned out and π_C is the proportion of the Civic Duty group that turned out.
- $H_0 : \pi_N - \pi_C = 0$ (no effect/difference in turnout)
- $H_a : \pi_N - \pi_C \neq 0$ (some effect/difference)
- We’ll set $\alpha = 0.05$ to be standard.

- Let's go to the statistical thought experiment—under the null, we have $\pi_N = \pi_C$ so that

$$\frac{(\hat{\pi}_N - \hat{\pi}_C) - 0}{SE[\hat{\pi}_N - \hat{\pi}_C]} \sim N(0, 1)$$

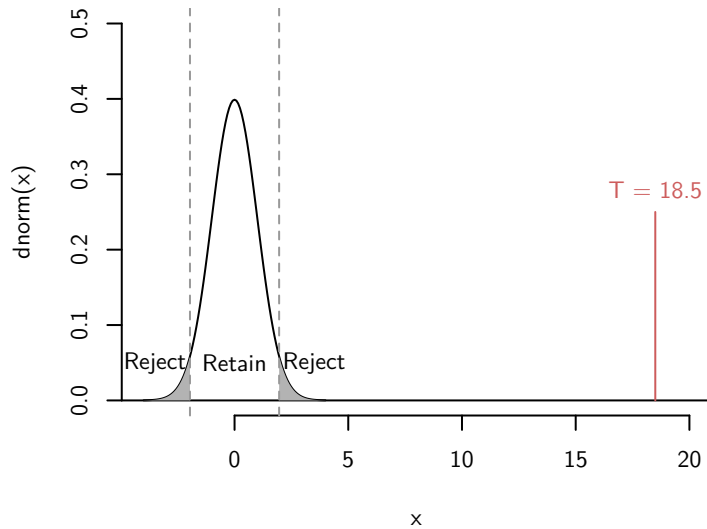
- What's the rejection region? Because the alternative is two-sided, we have to allow for very positive values of T and very negative values of T to be evidence against the null. Thus, we'll derive two equally sized rejection regions on either side:



- Under the null, we have a 0.05 probability of being more than 1.96 standard errors away from 0. Thus, our critical value becomes 1.96
- Earlier we got an estimate of the standard error, $\widehat{SE} = 0.0034$ and we have $\hat{\pi}_N = 0.378$ and $\hat{\pi}_C = 0.315$. Then our observed test statistic is:

$$\frac{(0.378 - 0.315)}{0.0034} = 18.5294118$$

- Is 18.5 large relative to the null distribution (which is a standard Normal)?



- Quite obviously, we would reject the null in this case. How unlikely would it be to get a difference this extreme or more extreme (either above or below 0)? That is, what is the p-value?
- Use the `pnorm()` function and the fact that $\mathbb{P}(T(Y) > 18.5) = \mathbb{P}(T(Y) < -18.5)$:

```
2 * pnorm(-18.5)
```

```
## [1] 2.06474e-76
```

Connection between confidence intervals and hypothesis tests

- Reject if:

$$\frac{\bar{Y} - \mu_0}{\frac{S}{\sqrt{n}}} < -z_{\alpha/2} \quad \text{or} \quad \frac{\bar{Y} - \mu_0}{\frac{S}{\sqrt{n}}} > z_{\alpha/2}$$

$$\bar{Y} - \mu_0 < -z_{\alpha/2} \times \frac{S}{\sqrt{n}} \quad \text{or} \quad \bar{Y} - \mu_0 > z_{\alpha/2} \times \frac{S}{\sqrt{n}}$$

$$\bar{Y} + z_{\alpha/2} \times \frac{S}{\sqrt{n}} < \mu_0 \quad \text{or} \quad \bar{Y} - z_{\alpha/2} \times \frac{S}{\sqrt{n}} > \mu_0$$

- What does this say? Basically we reject μ_0 if is outside the $(1 - \alpha)\%$ confidence interval: $\bar{Y} \pm z_{\alpha/2} \times \frac{S}{\sqrt{n}}$.

- This means that all rejected null hypotheses at the α level fall outside the $(1 - \alpha)\%$ confidence interval. And all values inside the $(1 - \alpha)\%$ confidence interval must be those that are not rejected at the α level.

Pitfalls of hypothesis testing

1. Size of the effect often more important than whether it exists or not
 2. Only relevant when there is a specific hypothesis to test
- **Definition** We say that an effect/difference is **statistically significant** if we reject the null hypothesis of no effect/difference.
 - Just because an effect is statistically significant doesn't make it substantively significant! How big is the effect relative to other effects or relative to the cost of intervening.
 - Often, a point estimate and a confidence interval will be better tools than a hypothesis test.