Gov 2000: 5. Estimation and Statistical Inference

Matthew Blackwell

Fall 2016

1. Point Estimation
2. Properties of Estimators
3. Interval Estimation
4. Where Do Estimators Come From?*
5. Wrap up

Housekeeping

- This Thursday, 10/6: HW 3 due, HW 4 goes out.
- Next Thursday, 10/13: HW 4 due, HW 5 goes out.
- Thursday, 10/20: HW 5 due, Midterm available.
- Midterm:
    - Check-out exam: you have 8 hours to complete it once you check it out.
    - Answers must be typeset, as usual.
    - You should have more than enough time.
    - We'll post practice midterms in advance.
- Evaluations: we'll be fielding an anonymous survey about the course this week.

Where are we? Where are we going?

- Last few weeks: probability, learning how to think about r.v.s
- Now: how to estimate features of underlying distributions with real data.
- Build on last week: if the sample mean will be "close" to μ, can use it as a best guess for μ?

1/ Point Estimation

Motivating example

- Gerber, Green, and Larimer (APSR, 2008)

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______

Motivating Example

```
load("../data/gerber_green_larimer.RData")
## turn turnout variable into a numeric
social$voted <- 1 * (social$voted == "Yes")
neigh.mean <- mean(social$voted[social$treatment ==
    "Neighbors"])
neigh.mean
```

```
## [1] 0.378
```

```
contr.mean <- mean(social$voted[social$treatment ==
    "Civic Duty"])
contr.mean
```

```
## [1] 0.315
```

```
neigh.mean - contr.mean
```

```
## [1] 0.0634
```

- Is this a “real”? Is it big?

Why study estimators?

- Goal 1: Inference
  - What is our best guess about some quantity of interest?
  - What are a set of plausible values of the quantity of interest?
- Goal 2: Compare estimators
  - In an experiment, use simple difference in sample means $(\overline{Y} - \overline{X})$?
  - Or the post-stratification estimator, where we estimate the estimate the difference among two subsets of the data (male and female, for instance) and then take the weighted average of the two ($\overline{Z}$ is the share of women):

$$(\overline{Y}_f - \overline{X}_f)\overline{Z} + (\overline{Y}_m - \overline{X}_m)(1 - \overline{Z})$$

  - Which (if either) is better? How would we know?

Samples from the population

- Our focus: $Y_1, \ldots, Y_n$ are i.i.d. draws from $f(y)$
  - e.g.: $Y_i = 1$ if citizen i votes, $Y_i = 0$ otherwise.
  - i.i.d. can be justified through random sampling from a population.
  - $f(y)$ is often called the population distribution
- Statistical inference or learning is using data to infer $f(y)$.

Point estimation

- Point estimation: providing a single "best guess" as to the value of some fixed, unknown quantity of interest, θ.
  - θ is a feature of the population distribution, $f(y)$
  - Also called: estimands, parameters.
- Examples of quantities of interest:
  - $\mu = \mathbb{E}[Y_i]$: the mean (turnout rate in the population).
  - $\sigma^2 = \mathbb{V}[Y_i]$: the variance.
  - $\mu_y - \mu_x = \mathbb{E}[Y] - \mathbb{E}[X]$: the difference in mean turnout between two groups.
  - $r(x) = \mathbb{E}[Y|X = x]$: the conditional expectation function (regression).
- These are the things we want to learn about.

Estimators

Estimator

An estimator, $\hat{\theta}_n$ of some parameter θ, is a function of the sample: $\hat{\theta}_n = h(Y_1, \ldots, Y_n)$.

- $\hat{\theta}_n$ is a r.v. because it is a function of r.v.s.
  - $\rightsquigarrow$ $\hat{\theta}_n$ has a distribution.
  - $\{\hat{\theta}_1, \hat{\theta}_2, \ldots\}$ is a sequence of r.v.s, so we can think about convergence in probability/distribution.
- An estimate is one particular realization of the estimator/r.v.

Examples of Estimators

- For the population expectation, μ, we have many different possible estimators:
  - $\hat{\theta}_n = \overline{Y}_n$ the sample mean
  - $\hat{\theta}_n = Y_1$ just use the first observation
  - $\hat{\theta}_n = \max(Y_1, \ldots, Y_n)$
  - $\hat{\theta}_n = 3$ always guess 3

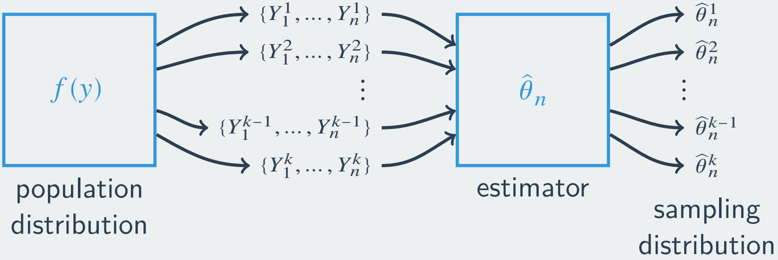
Understanding check

- **Question** Why is the following statement wrong: “My estimate was the sample mean and my estimator was 0.38”?

The three distributions

- Population Distribution: the data-generating process
  - Bernoulli in the case of the social pressure/voter turnout example)
- Empirical distribution: $Y_1, \dots, Y_n$
  - series of 1s and 0s in the sample
- Sampling distribution: distribution of the estimator over repeated samples from the population distribution
  - the 0.38 sample mean in the "Neighbors" group is one draw from this distribution

Sampling distribution, in pictures



Sampling distribution

```r
## now we take the mean of one sample, which is one
## draw from the **sampling distribution**
my.samp <- rbinom(n = 10, size = 1, prob = 0.4)
mean(my.samp)
```

```
## [1] 0.2
```

```r
## let's take another draw from the population dist
my.samp.2 <- rbinom(n = 10, size = 1, prob = 0.4)
```

```r
## Let's feed this sample to the sample mean
## estimator to get another estimate, which is
## another draw from the sampling distribution
mean(my.samp.2)
```

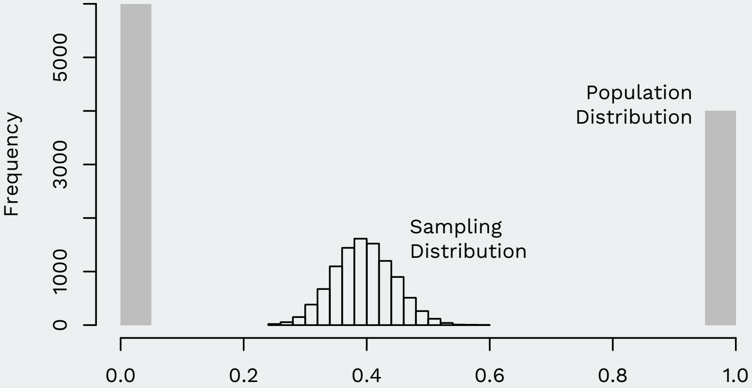
```
## [1] 0.4
```

Sampling distribution by simulation

- Let's generate 10,000 draws from the sampling distribution of the sample mean here when $n = 100$.

```r
nsims <- 10000
mean.holder <- rep(NA, times = nsims)
for (i in 1:nsims) {
    my.samp <- rbinom(n = 100, size = 1, prob = 0.4)
    mean.holder[i] <- mean(my.samp)  ## sample mean
    first.holder[i] <- my.samp[1]  ## first obs
}
```

Sampling distribution versus population distribution



Question The sampling distribution refers to the distribution of θ, true or false.

2/ Properties of Estimators

Properties of estimators

- We only get one draw from the sampling distribution, $\hat{\theta}_n$.
- Want to use estimators whose distribution is “close” to the true value.
- There are two ways we evaluate estimators:
    - Finite sample: the properties of its sampling distribution for a fixed sample size n.
    - Large sample: the properties of the sampling distribution as we let $n \to \infty$.

Running example

- Two independent random samples (treatment/control):
  - $Y_1, \ldots, Y_{n_y}$ are i.i.d. with mean μ_y and variance σ_y^2
  - $X_1, \ldots, X_{n_x}$ are i.i.d. with mean μ_x and variance σ_x^2
  - Overall sample size $n = n_y + n_x$
- Parameter is the population difference in means, which is the treatment effect of the social pressure mailer: $\mu_y - \mu_x$
- Estimator is the difference in sample means:

$$\widehat{D}_n = \overline{Y}_{n_y} - \overline{X}_{n_x}$$

Finite-sample properties

Let $\hat{\theta}_n$ be a estimator of θ. Then we have the following definitions:

- bias$[\hat{\theta}_n] = \mathbb{E}[\hat{\theta}_n] - \theta$
  - $\hat{\theta}_n$ is unbiased if bias$[\hat{\theta}_n] = 0$
  - Last week: $\overline{X}_n$ is unbiased for μ since $\mathbb{E}[\overline{X}_n] = \mu$
- Sampling variance is $\mathbb{V}[\hat{\theta}_n]$.
  - Example: $\mathbb{V}[\overline{X}_n] = \sigma^2/n$
- Standard error is se$[\hat{\theta}_n] = \sqrt{\mathbb{V}[\hat{\theta}_n]}$
  - Example: se$[\overline{X}_n] = \sigma/\sqrt{n}$

Diff-in-means finite sample properites

- Unbiasedness from unbiasedness of sample means:

$$\mathbb{E}[\overline{Y}_{n_y} - \overline{X}_{n_x}] = \mathbb{E}[\overline{Y}_{n_y}] - \mathbb{E}[\overline{X}_{n_x}] = \mu_y - \mu_x$$

- Sampling variance, by independent samples:

$$\mathbb{V}[\overline{Y}_{n_y} - \overline{X}_{n_x}] = \mathbb{V}[\overline{Y}_{n_y}] + \mathbb{V}[\overline{X}_{n_x}] = \frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}$$

- Standard error:

$$\text{se}[\widehat{D}_n] = \sqrt{\frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}}$$

Mean squared error

- Mean squared error or MSE is

$$\text{MSE} = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$$

- The MSE assesses the quality of an estimator.
  - How big are (squared) deviations from the true parameter?
  - Ideally, this would be as low as possible!
- Useful decomposition result:

$$\text{MSE} = \text{bias}[\hat{\theta}_n]^2 + \mathbb{V}[\hat{\theta}_n]$$

- $\rightsquigarrow$ for unbiased estimators, MSE is the sampling variance.
- Might accept some bias for large reductions in variance for lower overall MSE.

Consistency

- An estimator is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$.
  - Distribution of $\hat{\theta}_n$ collapses on θ as $n \to \infty$.
  - WLLN: $\overline{X}_n$ is consistent for μ.
  - Inconsistent estimator are bad bad bad: more data gives worse answers!
- Theorem: If $\text{bias}[\hat{\theta}_n] \to 0$ and $\text{se}[\hat{\theta}_n] \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent.
- Example: Difference-in-means.
  - $\widehat{D}_n$ is unbiased with $\mathbb{V}[\widehat{D}_n] = \frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}$
  - $\rightsquigarrow \widehat{D}_n$ consistent since $\mathbb{V}[\widehat{D}_n] \to 0$
- NB: Unbiasedness does not imply consistency, nor vice versa.

Unbiased versus consistent

- Unbiased, not consistent: "first observation" estimator, $\hat{\theta}_n^f = Y_1$.
    - Unbiased because $\mathbb{E}[\hat{\theta}_n^f] = \mathbb{E}[Y_1] = \mu_y$
    - Not consistent: $\hat{\theta}_n^f$ is constant in n so its distribution never collapses.
    - Said differently: the variance of $\hat{\theta}_n^f$ never shrinks.
- Consistent, but biased: sample mean with n replaced by $n-1$:

$$\frac{n}{n-1}\overline{Y}_n = \frac{1}{n-1}\sum_{i=1}^{n} Y_i$$

- Bias: $\mathbb{E}[\frac{n}{n-1}\overline{Y}_n] - \mu_y = \frac{1}{n-1}\mu_y$
- Consistent because bias and se $\to 0$ as $n \to \infty$.

Asymptotic normality

- An estimator is asymptotically normal if

$$\frac{\hat{\theta}_n - \theta}{\text{se}[\hat{\theta}_n]} \xrightarrow{d} N(0, 1)$$

  - Allows us to approximate the probability of $\hat{\theta}_n$ being far away from θ in large samples.
- Many, many, many estimators will be asymptotically normal by some version of the Central Limit Theorem.
  - CLT: $\overline{X}_n$ is asymptotically normal
- By an extension of the CLT for independent samples:

$$\frac{\widehat{D}_n - (\mu_y - \mu_x)}{\sqrt{\sigma_y^2/n_y + \sigma_x^2/n_x}} \xrightarrow{d} N(0, 1)$$

Help, I don't know the SE

- But we don't know $\text{se}[\hat{\theta}_n]$?!
- $\rightsquigarrow$ plug in a consistent estimator $\widehat{\text{se}}[\hat{\theta}_n]$!
- If $\hat{\theta}_n$ is asymptotically normal and $\widehat{\text{se}}[\hat{\theta}_n] \xrightarrow{p} \text{se}[\hat{\theta}_n]$, then:

$$\frac{\hat{\theta}_n - \theta}{\widehat{\text{se}}[\hat{\theta}_n]} \xrightarrow{d} N(0, 1)$$

- Using the true vs. estimated standard error doesn't matter in large samples.

Estimating the Sampling Variance/Standard Error

- Diff-in-means variance: $\mathbb{V}[\widehat{D}_n] = \frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}$
  - Need to estimate these dang unknown population variances, σ_y^2 and σ_x^2.
- Use the sample variances: $S_y^2 = \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (Y_i - \overline{Y}_{n_y})^2$
  - Consistent for population variance: $S_y^2 \xrightarrow{p} \sigma_y^2$
- Estimated diff-in-means variance is consistent:

$$\widehat{\mathbb{V}}[\widehat{D}_n] = \frac{S_y^2}{n_y} + \frac{S_x^2}{n_y} \xrightarrow{p} \frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x} = \mathbb{V}[\widehat{D}_n]$$

Putting it all together

- If $\widehat{\mathbb{V}}[\widehat{D}_n] \xrightarrow{p} \mathbb{V}[\widehat{D}_n]$ then $\widehat{\text{se}}[\widehat{D}_n] = \sqrt{\widehat{\mathbb{V}}[\widehat{D}_n]} \xrightarrow{p} \text{se}[\widehat{D}]$
  - Challenge question: prove this.
- Since we know $\widehat{D}_n$ is asymptotically normal and $\widehat{\text{se}}[\widehat{D}_n]$ is consistent, then we know that:

$$\frac{\widehat{D}_n - (\mu_y - \mu_x)}{\sqrt{S_y^2/n_y + S_x^2/n_x}} \xrightarrow{d} N(0,1)$$

- Now we can make approximate probability statements about how far $\widehat{D}_n$ will be from the truth!

3/ Interval Estimation

Interval estimation - what and why?

- $\overline{Y}_n - \overline{X}_n$ is our best guess about $\mu_y - \mu_x$
- But $\mathbb{P}(\overline{Y}_n - \overline{X}_n = \mu_y - \mu_x) = 0$!
- Alternative: produce a range of values that will contain the truth with some fixed probability
- An interval estimate of the population difference in means, $\mu_y - \mu_x$, consists of two bounds within which we expect $\mu_y - \mu_x$ to reside:
$$a \leq \mu_y - \mu_x \leq b$$
- How can we possibly figure out such an interval? We'll rely on the distributional properties of estimators. Ideas extend to all estimators, including regression.

What is a confidence interval?

Confidence interval

A $100(1-\alpha)\%$ confidence interval for a population parameter θ is an interval $C_n = (a, b)$, where $a = a(Y_1, \dots, Y_n)$ and $b = b(Y_1, \dots, Y_n)$ are functions of the data such that

$$\mathbb{P}(a \leq \theta \leq b) \geq 1 - \alpha.$$

- The random interval (a, b) will bound θ $100(1-\alpha)\%$ of the time.
    - An estimator just like $\overline{X}_n$ but with two values.
- $1 - \alpha$ is the coverage of the confidence interval.
- Extremely useful way to represent our uncertainty about our estimate.

Deriving a probabilistic bound

- Let $\widehat{\text{se}} = \sqrt{S_y^2/n_y + S_x^2/n_x}$, so that:

$$\frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{\text{se}}} \xrightarrow{d} N(0, 1)$$

- Because of the CLT, we can use this to derive a confidence interval such that: $(\mu_y - \mu_x)$:

$$\mathbb{P}\left(a \le (\mu_y - \mu_x) \le b\right) = 0.95$$

- We want to find a value so that in 95% of random samples, it will between these two bounds.
- Use the following fact. For large n:

$$\mathbb{P}\left(-1.96 \le \frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{\text{se}}} \le 1.96\right) \approx 0.95$$

Deriving the interval

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

$$
\begin{aligned}
0.95 &\approx \mathbb{P}\left(-1.96 \le \frac{\widehat{D}_n - (\mu_y - \mu_x)}{\widehat{\mathsf{se}}} \le 1.96 \right) \\
&= \mathbb{P}\left(-1.96 \times \widehat{\mathsf{se}} \le \widehat{D}_n - (\mu_y - \mu_x) \le 1.96 \times \widehat{\mathsf{se}} \right) \\
&= \mathbb{P}\left(-\widehat{D}_n - 1.96 \times \widehat{\mathsf{se}} \le -(\mu_y - \mu_x) \le -\widehat{D}_n + 1.96 \times \widehat{\mathsf{se}} \right) \\
&= \mathbb{P}\left(\widehat{D}_n - 1.96 \times \widehat{\mathsf{se}} \le (\mu_y - \mu_x) \le \widehat{D}_n + 1.96 \times \widehat{\mathsf{se}} \right)
\end{aligned}
$$

- Lower bound: $\widehat{D}_n - 1.96 \times \widehat{\mathsf{se}}$
- Upper bound: $\widehat{D}_n + 1.96 \times \widehat{\mathsf{se}}$
  - Usually written as $\widehat{D}_n \pm 1.96 \times \widehat{\mathsf{se}}$
- Bounds are random! Not $(\mu_y - \mu_x)$!

CI for social pressure effect

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	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

```r
neigh_var <- var(social$voted[social$treatment == "Neighbors"])
neigh_n <- 38201
civic_var <- var(social$voted[social$treatment == "Civic Duty"])
civic_n <- 38218

se_diff <- sqrt(neigh_var/neigh_n + civic_var/civic_n)

## lower bound
(0.378 - 0.315) - 1.96 * se_diff
```

```
## [1] 0.0563
```

```r
## upper bound
(0.378 - 0.315) + 1.96 * se_diff
```

```
## [1] 0.0697
```

Interpreting the confidence interval

- Caution! An often recited, but incorrect interpretation of a confidence interval is the following:
  - "I calculated a 95% confidence interval of [0.05,0.13], which means that there is a 95% chance that the true difference in means in is that interval."
  - This is WRONG.
- The true value of the population difference in means, $\mu_y - \mu_x$, 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: $\widehat{D}_n \pm 1.96 \times \widehat{\text{se}}[\widehat{D}_n]$. 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

- Draw samples of size 500 (pretty big) from $N(1, 10)$
- Calculate confidence intervals for the sample mean:

$$\overline{Y}_n \pm 1.96 \times \widehat{\text{se}}[\overline{Y}_n] \rightsquigarrow \overline{Y}_n \pm 1.96 \times S_n / \sqrt{n}$$

```
set.seed(2143)
sims <- 10000
cover <- rep(0, 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
    }
}
mean(cover)
```

```
## [1] 0.95
```

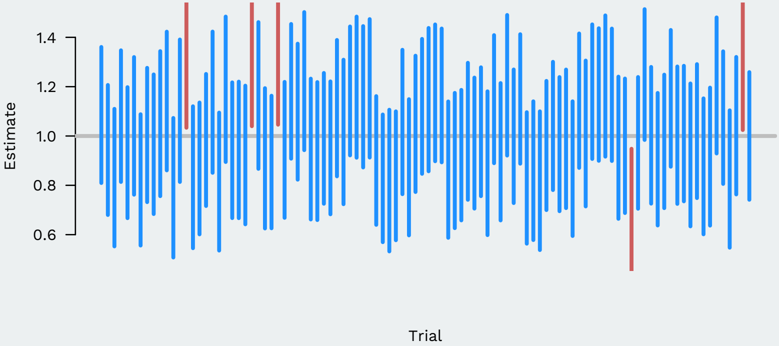
Plotting the CIs

Plotting the CIs

Plotting the CIs

Plotting the CIs

Plotting the CIs



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

More general confidence intervals

- Let $\widehat{\theta}_n$ be an asymptotically normal estimator for θ.
  - Any aysmp. normal estimator! $\overline{X}_n$, $\widehat{D}_n$, or whatever!
- A general formula for a $100(1-\alpha)\%$ confidence interval is:

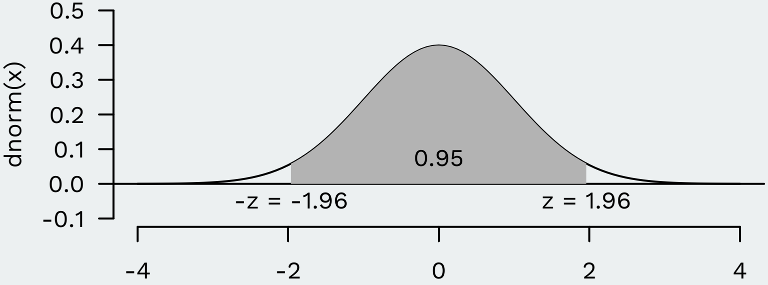
$$\widehat{\theta}_n \pm z_{\alpha/2} \times \widehat{\mathsf{se}}[\widehat{\theta}_n]$$

- $z_{\alpha/2}$ comes from a similar derivation as earlier:

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}[\widehat{\theta}_n]} \leq z_{\alpha/2}\right) = (1-\alpha)$$

- Remember! Asymptotics are approximations!

Finding the z values



- How do we figure out what $z_{\alpha/2}$ will be? Need to find the values such that for $Z \sim N(0, 1)$:

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

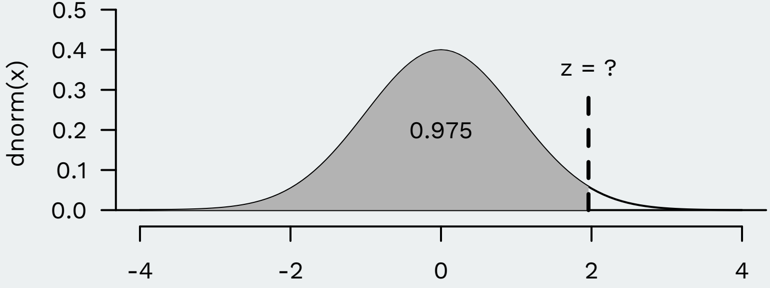
- Intuitively, we want the z values that puts $\alpha/2$ in each of the tails.
- For example, with $\alpha = 0.05$ for a 95% confidence interval, we want the z values that put 0.025 (2.5%) in each of the tails.

Putting it in the tails

- How to get the z values? Put α probability in the tails:

$$\begin{aligned}
\mathbb{P}(\{Z < -z_{\alpha/2}\} \cup \{Z > z_{\alpha/2}\}) &= \alpha & \\
\mathbb{P}(Z < -z_{\alpha/2}) + \mathbb{P}(Z > z_{\alpha/2}) &= \alpha & \text{(additivity)} \\
2 \times \mathbb{P}(Z > z_{\alpha/2}) &= \alpha & \text{(symmetry)} \\
\mathbb{P}(Z < z_{\alpha/2}) &= 1 - \alpha/2 &
\end{aligned}$$

- Find the z-value that puts probability $1 - \alpha/2$ below it:



Calculating z-values in R

- Inverse of the CDF (quantile) of the standard Normal evaluated at $1 - \alpha/2$!
- Procedure for a 90% confidence interval:
    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 $z_{\alpha/2}$:

```r
qnorm(0.95)
```

```
## [1] 1.64
```

- 90% CI: $\hat{\theta}_n \pm 1.64 \times \widehat{\text{se}}[\hat{\theta}_n]$

Question

- **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?
- **Answer** Wider!
- Decreases $\alpha \rightsquigarrow$ increases $1 - \alpha/2 \rightsquigarrow$ increases $z_{\alpha/2}$

4/ Where Do Estimators Come From?*

Statistical models

- A statistical model, $\mathbb{F}$, is a set of distributions we will consider that could have possibly generated the data.
- A parametric model is a set that can be parameterized by a finite number of parameters.
  - Bernoulli distribution:

$$\mathbb{F} = \left\{f(y; p) = y^p (1-y)^{1-p} : 0 \leq p \leq 1\right\}$$

  - Normal distribution:

$$\mathbb{F} = \left\{f(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} : \mu \in \mathbb{R}, \sigma^2 > 0\right\}$$

- Pros: easy to work with and explicit answers often exist
  - Basis of maximum likelihood, Bayesian inference, etc.
- Cons: inferences are model dependent
  - $\rightsquigarrow$ if our choice of model is wrong, our inferences might be wrong

Nonparametric models

- A nonparametric model is a set that cannot be parameterized by a finite set of parameters.
    - All distributions with finite mean:

$$\mathbb{F} = \{f(y) : \mathbb{E}[Y] < \infty\}$$

    - All distributions with finite mean and variance:

$$\mathbb{F} = \{f(y) : \mathbb{E}[Y] < \infty, \mathbb{V}[Y] < \infty\}$$

- Pros: no modeling assumptions beyond what we need.
- Cons: can be difficult to work with and difficult to interpret.

Where do estimators come from?

- Parametric models: maximum likelihood, Bayesian estimation, method of moments.
    - Derive estimators from the assumed p.m.f./p.d.f. $f(y)$.
    - Gov 2001 and beyond.
- Nonparametric models: plug-in estimation/analogy principle.
    - Quantities of interest are usually made up of expectations: $\mathbb{E}[g(Y)]$ for some function $g()$
    - Analogy principle: replace any population expectations, $\mathbb{E}[g(Y)]$ with sample means, $\frac{1}{n}\sum_{i=1}^{n} g(Y_i)$

Plug-in estimators, examples

- Expectation:

$$\mu = \mathbb{E}[Y_i] \rightsquigarrow \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Variance:

$$\sigma^2 = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] \rightsquigarrow \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Covariance:

$$\mathrm{Cov}[X_i, Y_i] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_i - \mathbb{E}[Y_i])] \rightsquigarrow \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

5/ Wrap up

Wrap up

- Generalized discussion of sample means to any estimator of any parameter.
- Unbiasedness, consistency, confidence intervals, etc will be with you for almost any statistical procedure moving forward.
- These properties give us an expectation about how far away our estimates will be from the truth.
- Next time: Testing hypotheses about parameters