Gov 50: 23. Inference with Mathematical Models

Matthew Blackwell

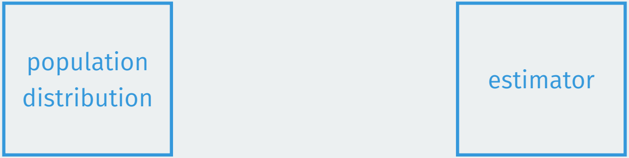
Harvard University

Roadmap

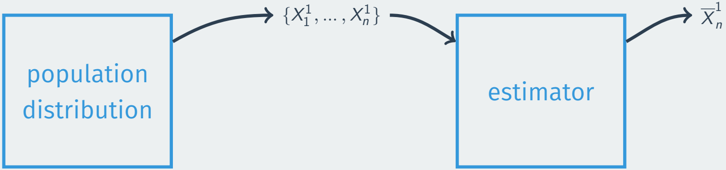
1. Central limit theorem
2. Normal distribution
3. Using the Normal for inference

1/ Central limit theorem

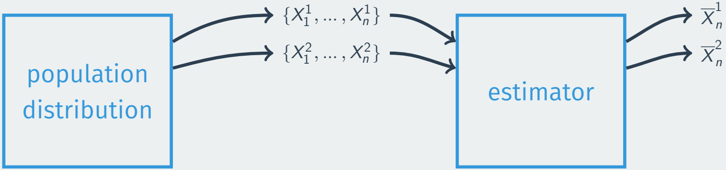
Sampling distribution, in pictures



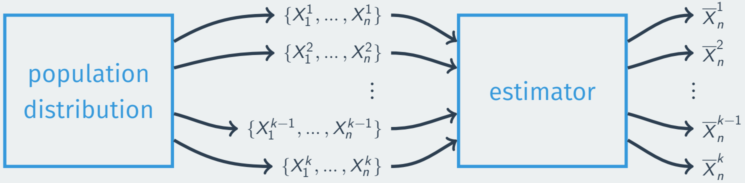
Sampling distribution, in pictures



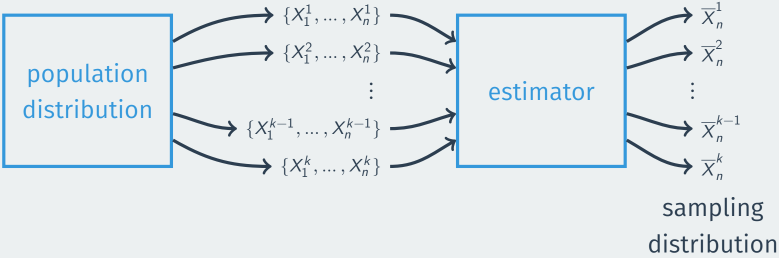
Sampling distribution, in pictures



Sampling distribution, in pictures



Sampling distribution, in pictures



Sampling distribution of the sample proportion

$$\text{sample mean} = \text{population mean} + \text{chance error}$$
$$\overline{X} = \mu + \text{chance error}$$

Sampling distribution of the sample proportion

$$\text{sample mean} = \text{population mean} + \text{chance error}$$
$$\overline{X} = \mu + \text{chance error}$$

Then $\overline{X}$ centered at μ.

Sampling distribution of the sample proportion

$$\text{sample mean} = \text{population mean} + \text{chance error}$$
$$\overline{X} = \mu + \text{chance error}$$

Then $\overline{X}$ centered at μ.

Spread: standard deviation of the sampling distribution is the **standard error**

Spread of the sample mean

- **Standard error**: how big is the chance error on average?

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
    - This is the standard deviation of the estimator **across repeated samples**.

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
    - This is the standard deviation of the estimator **across repeated samples**.
    - With random samples, we can get a formula for the SE for many estimators.

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
  - This is the standard deviation of the estimator **across repeated samples**.
  - With random samples, we can get a formula for the SE for many estimators.
- Standard error for the sample mean:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
    - This is the standard deviation of the estimator **across repeated samples**.
    - With random samples, we can get a formula for the SE for many estimators.
- Standard error for the sample mean:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$

- Two components:

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
    - This is the standard deviation of the estimator **across repeated samples**.
    - With random samples, we can get a formula for the SE for many estimators.
- Standard error for the sample mean:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$

- Two components:
    - Population SD: more spread of the variable in the population → more spread of sample means

Spread of the sample mean

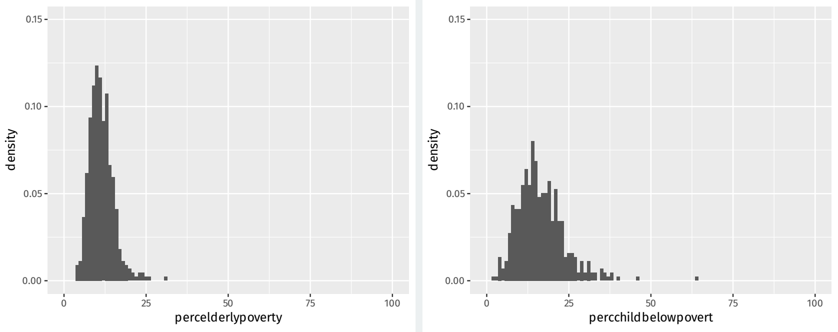
- **Standard error**: how big is the chance error on average?
  - This is the standard deviation of the estimator **across repeated samples**.
  - With random samples, we can get a formula for the SE for many estimators.
- Standard error for the sample mean:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$

- Two components:
  - Population SD: more spread of the variable in the population → more spread of sample means
  - Size of the sample: larger sample → smaller spread of the sample means

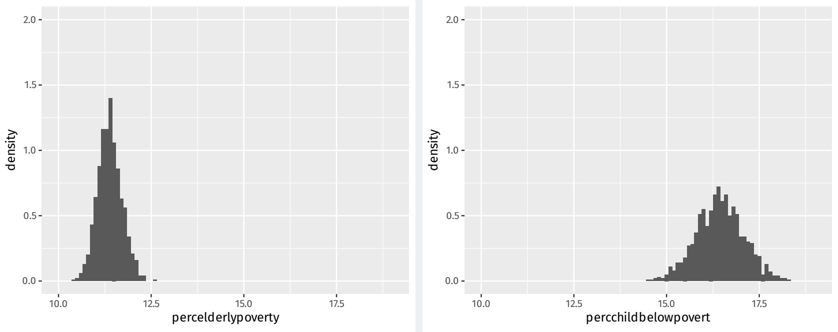
Midwest counties

Population distributions:



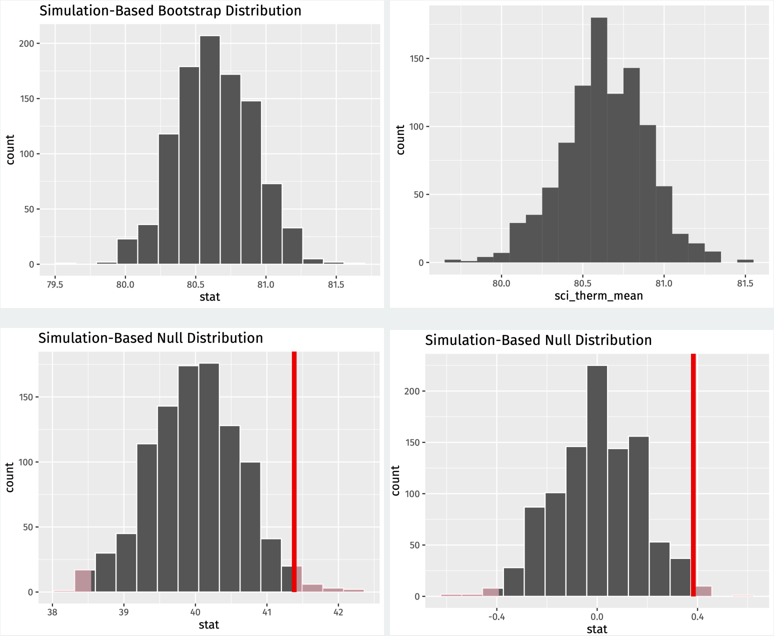
Midwest counties

Sampling distributions with $n = 100$

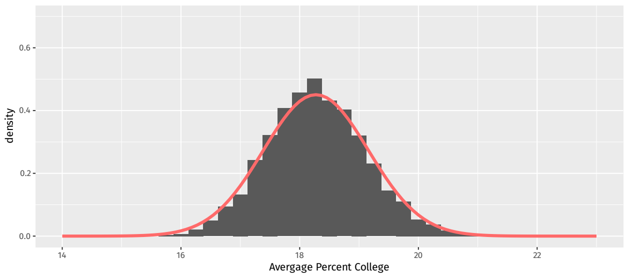


More population spread → higher SE

Similarity in the bootstrap/null distributions

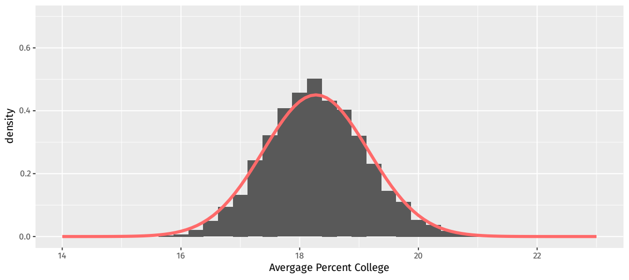


Conditions for the CLT



Central limit theorem: sums and means of **random samples** tend to be normally distributed as the **sample size grows**.

Conditions for the CLT

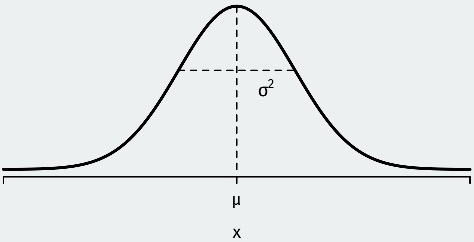


Central limit theorem: sums and means of **random samples** tend to be normally distributed as the **sample size grows**.

Many, many estimators will follow the CLT and have a normal distribution and will be easier to use this to do inference rather than doing increasingly complicated simulations.

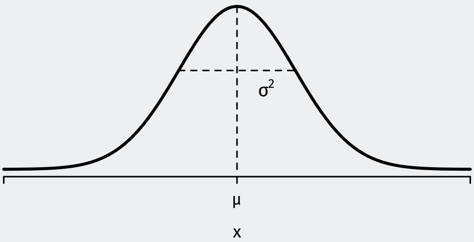
2/ Normal distribution

Normal distribution



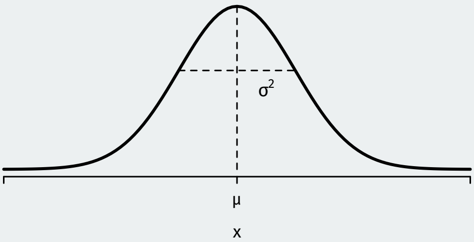
- A normal distribution can be affect by two values:

Normal distribution



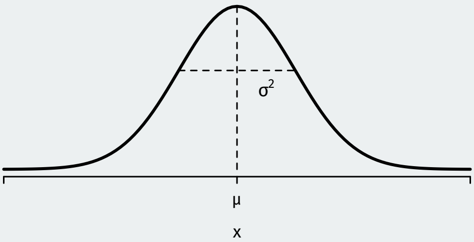
- A normal distribution can be affect by two values:
    - **mean/expected value** usually written as μ

Normal distribution



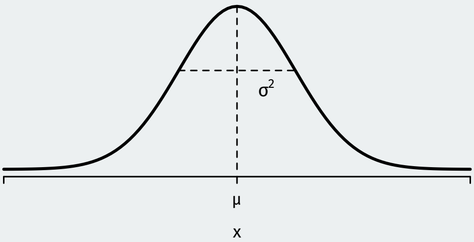
- A normal distribution can be affect by two values:
  - **mean/expected value** usually written as μ
  - **variance** written as σ^2 (standard deviation is σ)

Normal distribution



- A normal distribution can be affect by two values:
    - **mean/expected value** usually written as μ
    - **variance** written as σ^2 (standard deviation is σ)
    - Written $X \sim N(\mu, \sigma^2)$.

Normal distribution



- A normal distribution can be affect by two values:
    - **mean/expected value** usually written as μ
    - **variance** written as σ^2 (standard deviation is σ)
    - Written $X \sim N(\mu, \sigma^2)$.
- **Standard normal distribution**: mean 0 and standard deviation 1.

Reentering and scaling the normal

- How do transformations of a normal work?

Reentering and scaling the normal

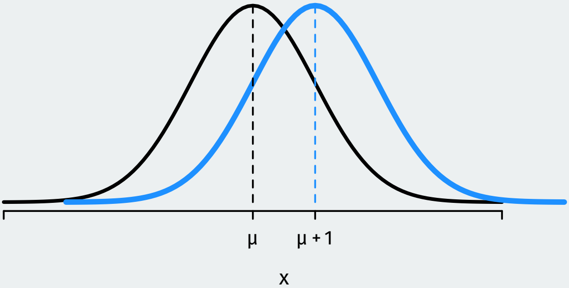
- How do transformations of a normal work?
- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.

Reentering and scaling the normal

- How do transformations of a normal work?
- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.
- If $Z = X + c$, then $Z \sim N(\mu + c, \sigma^2)$.

Reentering and scaling the normal

- How do transformations of a normal work?
- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.
- If $Z = X + c$, then $Z \sim N(\mu + c, \sigma^2)$.
- Intuition: adding a constant to a normal shifts the distribution by that constant.



Recentering and scaling the normal

- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.

Recentering and scaling the normal

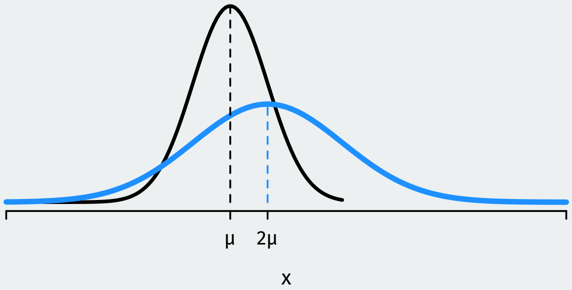
- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.
- If $Z = cX$, then $Z \sim N(c\mu, (c\sigma)^2)$.

Recentering and scaling the normal

- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.
- If $Z = cX$, then $Z \sim N(c\mu, (c\sigma)^2)$.
- Intuition: multiplying a normal by a constant scales the mean and the variance.

Recentering and scaling the normal

- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.
- If $Z = cX$, then $Z \sim N(c\mu, (c\sigma)^2)$.
- Intuition: multiplying a normal by a constant scales the mean and the variance.



Z-scores of normals

- These facts imply the **z-score** of a normal variable is a standard normal:

$$z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Z-scores of normals

- These facts imply the **z-score** of a normal variable is a standard normal:

$$z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- Subtract the mean and divide by the SD $\rightsquigarrow$ standard normal.

Z-scores of normals

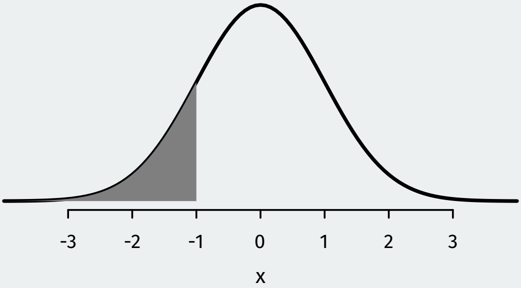
- These facts imply the **z-score** of a normal variable is a standard normal:

$$z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- Subtract the mean and divide by the SD $\rightsquigarrow$ standard normal.
- z-score measures how many SDs away from the mean a value of X is.

Normal probability calculations

What’s the probability of being below -1 for a standard normal?



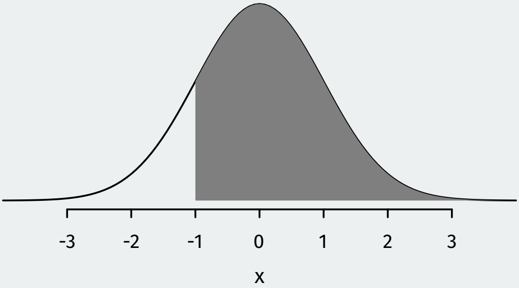
This is the area under the normal curve, which `pnorm( )` function gives us this:

```
pnorm(-1, mean = 0, sd = 1)
```

```
## [1] 0.159
```

Normal probability calculations

What's the probability of being **above** -1 for a standard normal?



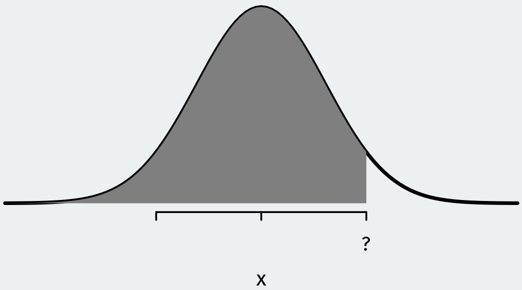
Total area under the curve (1) minus the area below -1:

```
1 - pnorm(-1, mean = 0, sd = 1)
```

```
## [1] 0.841
```

Normal quantiles

What if we want to know the opposite? What value of the normal distribution puts 95% of the distribution below it?



This is a **quantile** and we can get it using `qnorm( )`:

```
qnorm(0.95, mean = 0, sd = 1)
```

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

3/ Using the Normal for inference

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?

How popular is Joe Biden?

- What proportion of the public approves of Biden’s job as president?
- Latest Gallup poll:

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
    - Oct 2nd-23rd

How popular is Joe Biden?

- What proportion of the public approves of Biden’s job as president?
- Latest Gallup poll:
    - Oct 2nd-23rd
    - 1,009 adult Americans

How popular is Joe Biden?

- What proportion of the public approves of Biden’s job as president?
- Latest Gallup poll:
    - Oct 2nd-23rd
    - 1,009 adult Americans
    - Telephone interviews

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
  - Oct 2nd-23rd
  - 1,009 adult Americans
  - Telephone interviews
  - Approve (37%), Disapprove (59%)

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
  - Oct 2nd-23rd
  - 1,009 adult Americans
  - Telephone interviews
  - Approve (37%), Disapprove (59%)
- Define r.v. Y_i for Biden approval:

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
  - Oct 2nd-23rd
  - 1,009 adult Americans
  - Telephone interviews
  - Approve (37%), Disapprove (59%)
- Define r.v. Y_i for Biden approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Biden, 0 otherwise.

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
    - Oct 2nd-23rd
    - 1,009 adult Americans
    - Telephone interviews
    - Approve (37%), Disapprove (59%)
- Define r.v. Y_i for Biden approval:
    - $Y_i = 1 \rightsquigarrow$ respondent i approves of Biden, 0 otherwise.
    - $p = \mathbb{P}(Y_i = 1)$ the population proportion of Biden approvers.

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
    - Oct 2nd-23rd
    - 1,009 adult Americans
    - Telephone interviews
    - Approve (37%), Disapprove (59%)
- Define r.v. Y_i for Biden approval:
    - $Y_i = 1 \rightsquigarrow$ respondent i approves of Biden, 0 otherwise.
    - $p = \mathbb{P}(Y_i = 1)$ the population proportion of Biden approvers.
    - $\overline{Y} = 0.37$ is the sample proportion.

Standard errors for sample proportions

How variable will our sample proportion be? Depends on the **standard error**.

Standard errors for sample proportions

How variable will our sample proportion be? Depends on the **standard error**.

Special rule for SEs of sample proportion $\overline{Y}$:

$$SE \text{ for } \overline{Y} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(\text{pop. proportion}) \times (1 - \text{pop. proportion})}{\text{sample size}}}$$

Standard errors for sample proportions

How variable will our sample proportion be? Depends on the **standard error**.

Special rule for SEs of sample proportion $\overline{Y}$:

$$SE \text{ for } \overline{Y} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(\text{pop. proportion}) \times (1 - \text{pop. proportion})}{\text{sample size}}}$$

Because we don't know p, we replace it with our best guess, $\overline{Y}$:

$$\widehat{SE} = \sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}$$

CLT for confidence intervals

$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?

CLT for confidence intervals

$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?
    - Find a range of plausible chance errors and add them to $\overline{Y}$

CLT for confidence intervals

$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?
    - Find a range of plausible chance errors and add them to $\overline{Y}$
    - With **bootstrap**, we used resampling to simulate chance error.

CLT for confidence intervals

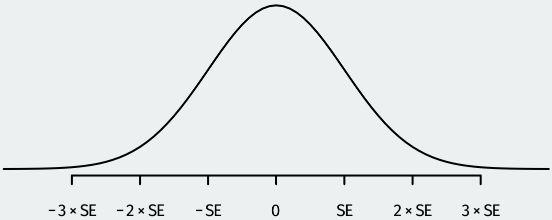
$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?
  - Find a range of plausible chance errors and add them to $\overline{Y}$
  - With **bootstrap**, we used resampling to simulate chance error.
- Central limit theorem implies

$$\overline{Y} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

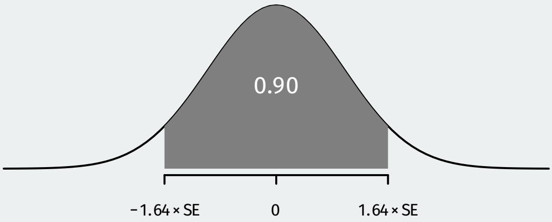
Chance error: $\overline{Y} - p$ is approximately normal with mean 0 and SE equal to $\sqrt{p(1-p)/n}$

Chance errors



If $\overline{Y} \sim N(p, SE^2)$, then chance errors are $\overline{Y} - p \sim N(0, SE^2)$ so:

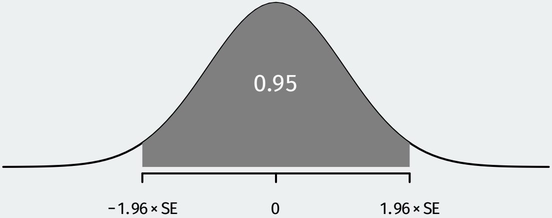
Chance errors



If $\overline{Y} \sim N(p, SE^2)$, then chance errors are $\overline{Y} - p \sim N(0, SE^2)$ so:

- ≈ 90% of chance errors $\overline{Y} - p$ are within 1.64 SEs of the mean.

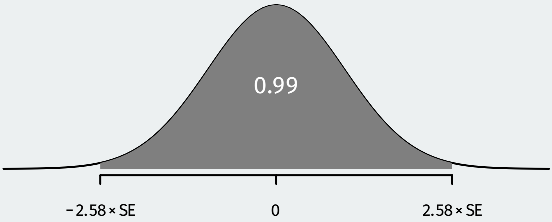
Chance errors



If $\overline{Y} \sim N(p, SE^2)$, then chance errors are $\overline{Y} - p \sim N(0, SE^2)$ so:

- ≈ 90% of chance errors $\overline{Y} - p$ are within 1.64 SEs of the mean.
- ≈ 95% of chance errors $\overline{Y} - p$ are within 1.96 SEs of the mean.

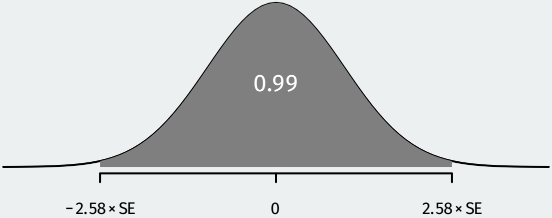
Chance errors



If $\overline{Y} \sim N(p, SE^2)$, then chance errors are $\overline{Y} - p \sim N(0, SE^2)$ so:

- ≈ 90% of chance errors $\overline{Y} - p$ are within 1.64 SEs of the mean.
- ≈ 95% of chance errors $\overline{Y} - p$ are within 1.96 SEs of the mean.
- ≈ 99% of chance errors $\overline{Y} - p$ are within 2.58 SEs of the mean.

Chance errors



If $\overline{Y} \sim N(p, SE^2)$, then chance errors are $\overline{Y} - p \sim N(0, SE^2)$ so:

- ≈ 90% of chance errors $\overline{Y} - p$ are within 1.64 SEs of the mean.
- ≈ 95% of chance errors $\overline{Y} - p$ are within 1.96 SEs of the mean.
- ≈ 99% of chance errors $\overline{Y} - p$ are within 2.58 SEs of the mean.

This implies we can build a 95% confidence interval with $\overline{Y} \pm 1.96 \times SE$

How did we get those values?

- First, choose a **confidence level**.

How did we get those values?

- First, choose a **confidence level**.
    - What percent of chance errors do you want to count as "plausible"?

How did we get those values?

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as “plausible”?
  - Convention is 95%.

How did we get those values?

- First, choose a **confidence level**.
    - What percent of chance errors do you want to count as "plausible"?
    - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

How did we get those values?

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**

How did we get those values?

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0, 1)$ z-score that would put $\alpha/2$ in the upper tail:

How did we get those values?

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0, 1)$ z-score that would put $\alpha/2$ in the upper tail:
  - $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = \alpha$

How did we get those values?

- First, choose a **confidence level**.
    - What percent of chance errors do you want to count as "plausible"?
    - Convention is 95%.
- $100 \times (1-\alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

    - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0,1)$ z-score that would put $\alpha/2$ in the upper tail:
    - $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = \alpha$
    - 90% CI $\rightsquigarrow \alpha = 0.1 \rightsquigarrow z_{\alpha/2} = 1.64$

How did we get those values?

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0,1)$ z-score that would put $\alpha/2$ in the upper tail:
  - $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = \alpha$
  - 90% CI $\rightsquigarrow \alpha = 0.1 \rightsquigarrow z_{\alpha/2} = 1.64$
  - 95% CI $\rightsquigarrow \alpha = 0.05 \rightsquigarrow z_{\alpha/2} = 1.96$

How did we get those values?

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0, 1)$ z-score that would put $\alpha/2$ in the upper tail:
  - $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = \alpha$
  - 90% CI $\rightsquigarrow \alpha = 0.1 \rightsquigarrow z_{\alpha/2} = 1.64$
  - 95% CI $\rightsquigarrow \alpha = 0.05 \rightsquigarrow z_{\alpha/2} = 1.96$
  - 99% CI $\rightsquigarrow \alpha = 0.01 \rightsquigarrow z_{\alpha/2} = 2.58$

Standard normal z-scores in R

`qnorm(x, lower.tail = FALSE)` will find the quantile of $N(0,1)$ that puts x in the upper tail:

```
qnorm(0.05, lower.tail = FALSE)
```

Standard normal z-scores in R

`qnorm(x, lower.tail = FALSE)` will find the quantile of $N(0, 1)$ that puts x in the upper tail:

```r
qnorm(0.05, lower.tail = FALSE)
```

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

Standard normal z-scores in R

`qnorm(x, lower.tail = FALSE)` will find the quantile of $N(0, 1)$ that puts x in the upper tail:

```r
qnorm(0.05, lower.tail = FALSE)
```

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

```r
qnorm(0.025, lower.tail = FALSE)
```

Standard normal z-scores in R

`qnorm(x, lower.tail = FALSE)` will find the quantile of $N(0, 1)$ that puts x in the upper tail:

```r
qnorm(0.05, lower.tail = FALSE)
```

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

```r
qnorm(0.025, lower.tail = FALSE)
```

```
## [1] 1.96
```

Standard normal z-scores in R

`qnorm(x, lower.tail = FALSE)` will find the quantile of $N(0, 1)$ that puts x in the upper tail:

```
qnorm(0.05, lower.tail = FALSE)
```

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

```
qnorm(0.025, lower.tail = FALSE)
```

```
## [1] 1.96
```

```
qnorm(0.005, lower.tail = FALSE)
```

Standard normal z-scores in R

`qnorm(x, lower.tail = FALSE)` will find the quantile of $N(0, 1)$ that puts x in the upper tail:

```
qnorm(0.05, lower.tail = FALSE)
```

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

```
qnorm(0.025, lower.tail = FALSE)
```

```
## [1] 1.96
```

```
qnorm(0.005, lower.tail = FALSE)
```

```
## [1] 2.58
```