CHRISTOPHER PRENER, PH.D.
FALL, 2017

WEEK 06
LECTURE 08

QUANTITATIVE ANALYSIS

FOUNDATIONS FOR INFERENCE

AGENDA

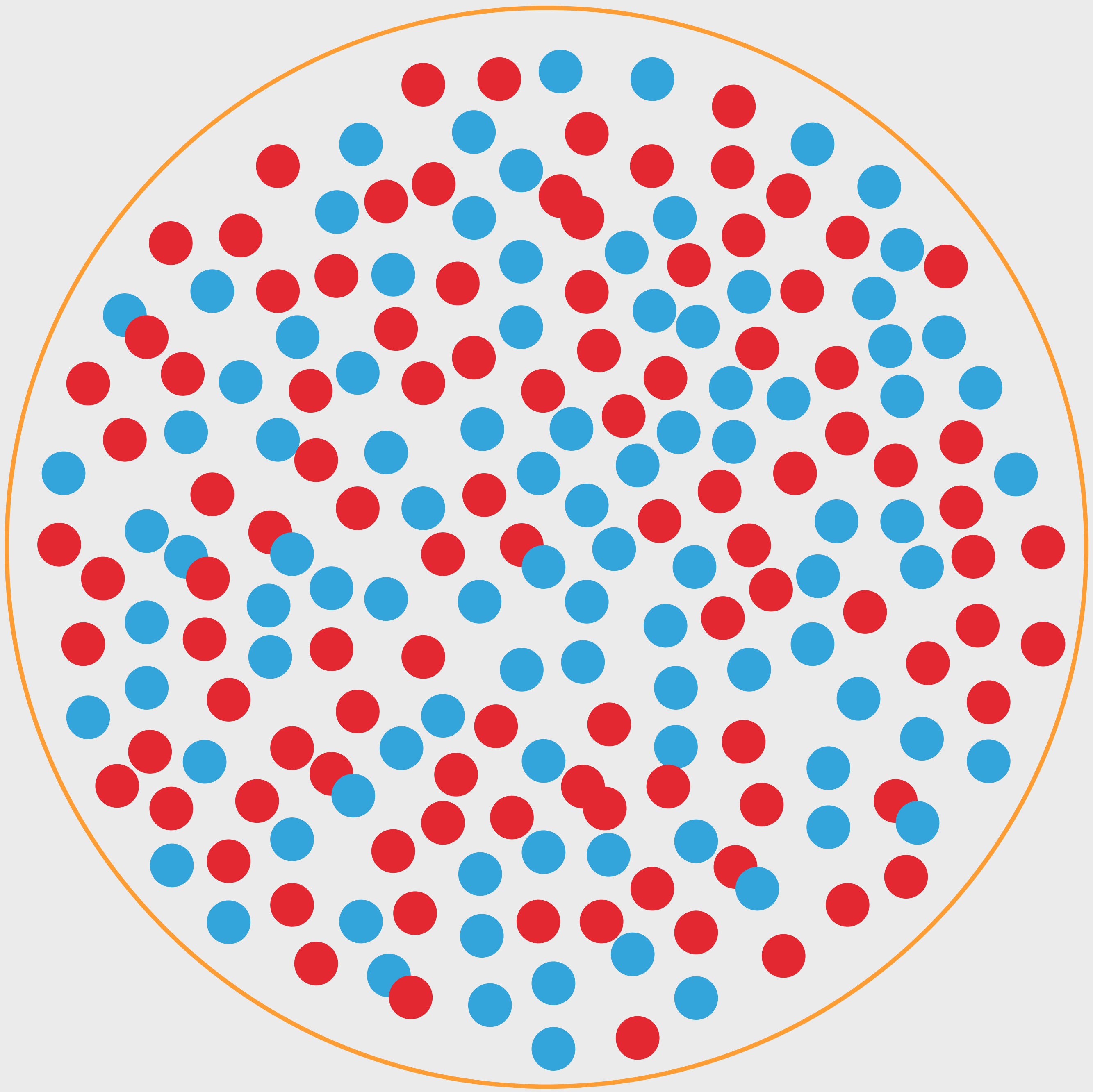
1. Follow-up
2. Inferential Goals
3. Central Limit Theorem
4. Confidence Intervals
5. Hypothesis Testing

1 FOLLOW-UP

2 INFERENTIAL GOALS

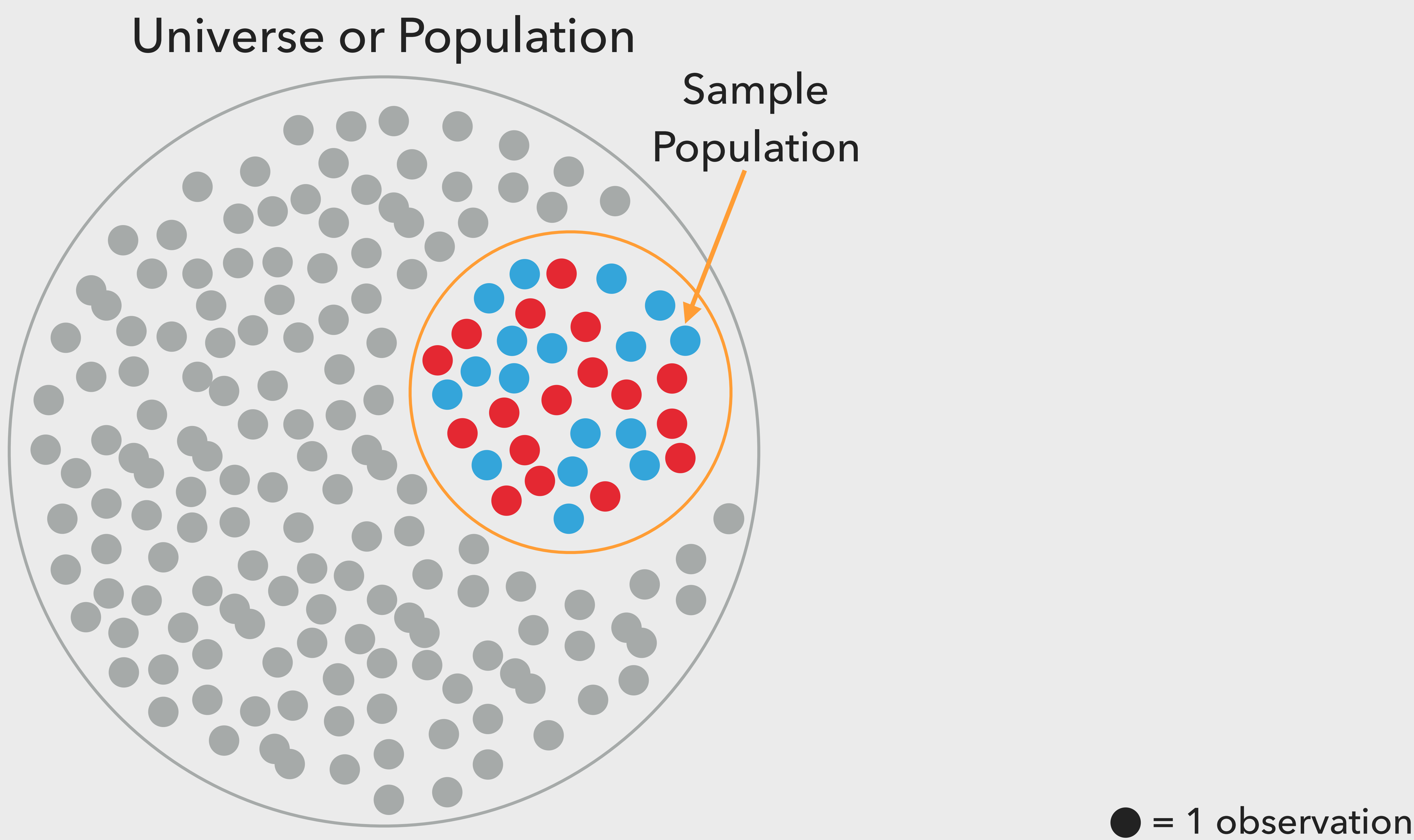
DRAWING INFERENCE

Universe or Population

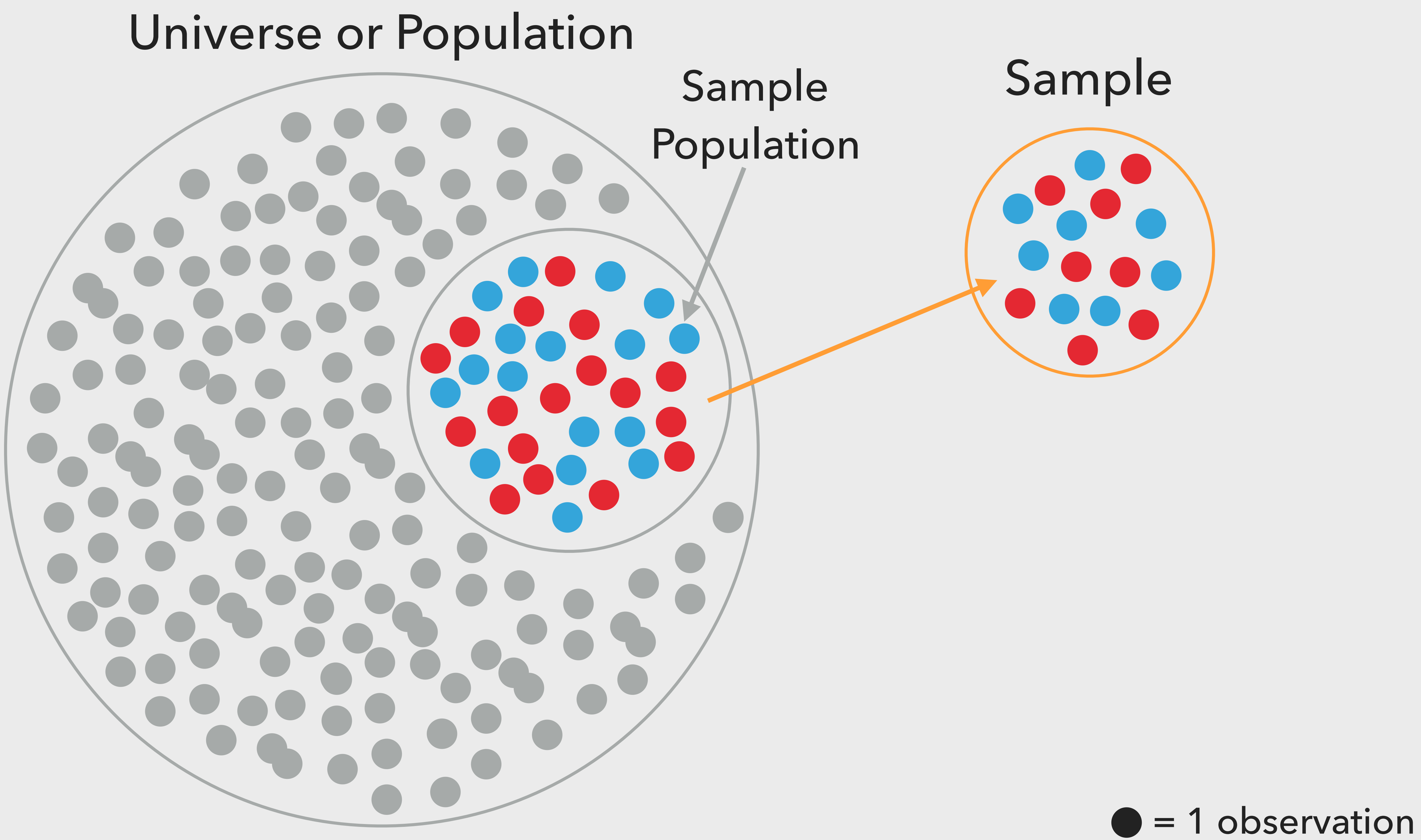


● = 1 observation

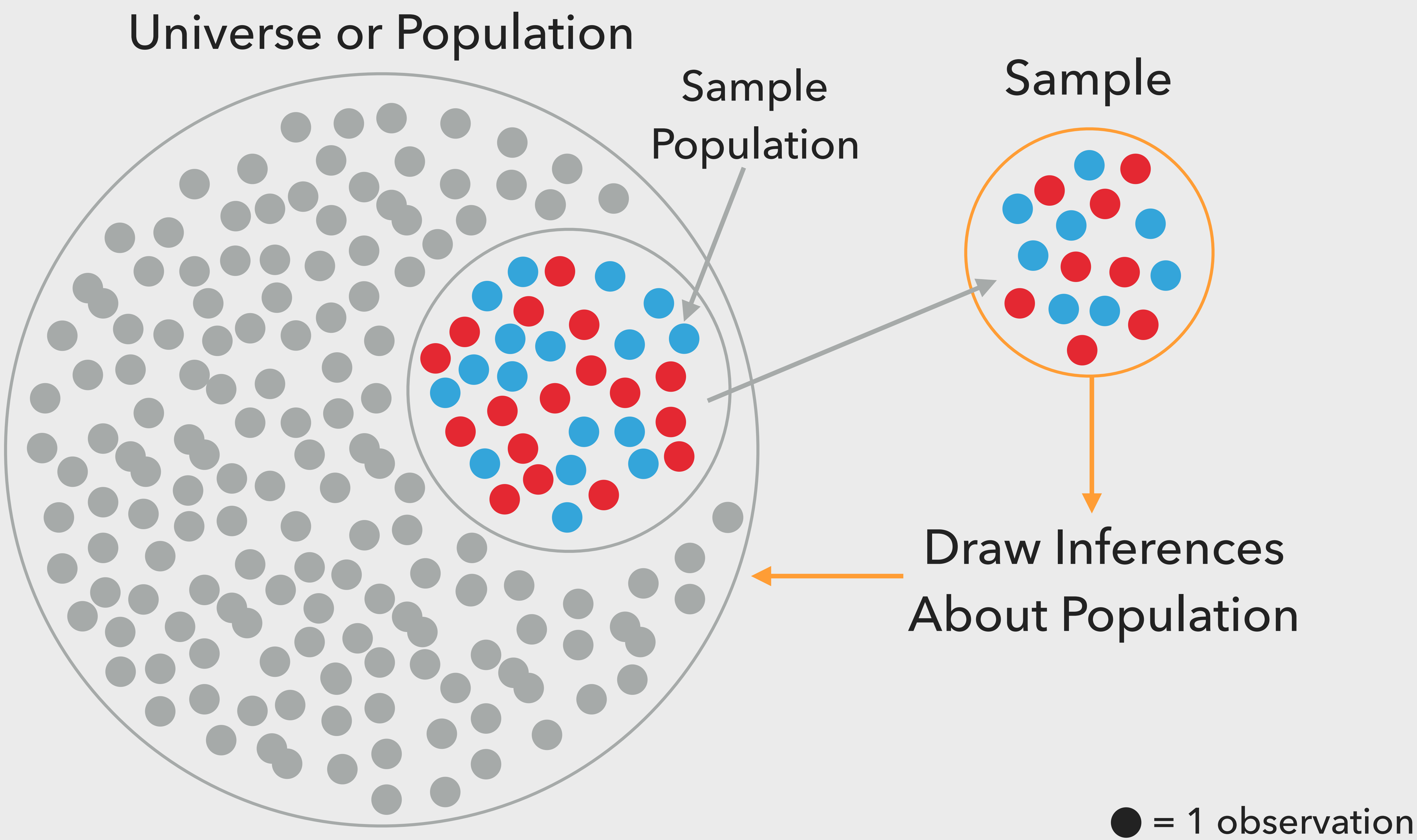
DRAWING INFERENCE



DRAWING INFERENCE



DRAWING INFERENCE



SAMPLE SIZE

Polling Data								
Poll	**Date**	**Sample**	**MoE**	**Clinton (D)**	**Trump (R)**	**Johnson (L)**	**Stein (G)**	**Spread**
RCP Average	**9/21 - 9/29**	**--**	**--**	**43.4**	**41.1**	**7.0**	**2.4**	**Clinton +2.3**
FOX News	9/27 - 9/29	911 LV	3.0	43	40	8	4	Clinton +3
Rasmussen Reports	9/27 - 9/29	1500 LV	2.5	43	42	6	2	Clinton +1
PPP (D)	9/27 - 9/28	933 LV	3.2	44	40	6	1	Clinton +4
Rasmussen Reports	9/26 - 9/28	1500 LV	2.5	42	41	7	2	Clinton +1
Reuters/Ipsos	9/22 - 9/26	1041 LV	3.5	42	38	7	2	Clinton +4
Quinnipiac	9/22 - 9/25	1115 LV	2.9	44	43	8	2	Clinton +1
Bloomberg	9/21 - 9/24	1002 LV	3.1	41	43	8	4	Trump +2
Monmouth	9/22 - 9/25	729 LV	3.6	46	42	8	2	Clinton +4
Economist/YouGov	9/22 - 9/24	948 RV	3.8	44	41	5	2	Clinton +3
NBC News/SM	9/19 - 9/25	13598 LV	1.1	45	40	10	3	Clinton +5
ABC News/Wash Post	9/19 - 9/22	651 LV	4.5	46	44	5	1	Clinton +2
Rasmussen Reports	9/20 - 9/21	1000 LV	3.0	39	44	8	2	Trump +5
Gravis	9/20 - 9/20	1560 LV	2.5	44	40	5	2	Clinton +4
Economist/YouGov	9/18 - 9/19	936 RV	4.0	40	38	7	2	Clinton +2
Reuters/Ipsos	9/15 - 9/19	1111 LV	3.4	37	39	7	2	Trump +2
McClatchy/Marist	9/15 - 9/20	758 LV	3.6	45	39	10	4	Clinton +6
NBC News/Wall St. Jrnl	9/16 - 9/19	922 LV	3.2	43	37	9	2	Clinton +6
Associated Press-GfK	9/15 - 9/16	1251 LV	--	45	39	9	2	Clinton +6
NBC News/SM	9/12 - 9/18	13320 LV	1.2	45	40	10	4	Clinton +5
FOX News	9/11 - 9/14	867 LV	3.0	41	40	8	3	Clinton +1

3 CENTRAL LIMIT THEOREM

3. CENTRAL LIMIT THEOREM


ABRHAM DE MOIVRE


PIERRE-SIMON LAPLACE

A POPULATION

```
> library("testDriveR")

> autoData <- auto17

> nrow(autoData)

[1] 1216

> summary(autoData$combFE)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  11.00   19.00   23.00   23.27   26.00   58.00

> sd(autoData$combFE)

[1] 5.83503
```

A RANDOM SAMPLE

```
> library(“dplyr”)

> sample1 <- dplyr::sample_n(autoData, 500)

> summary(sample1$combFE)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  12.00   19.00   23.00   23.38   26.25   56.00

> sd(sample1$combFE)

[1] 5.814742
```

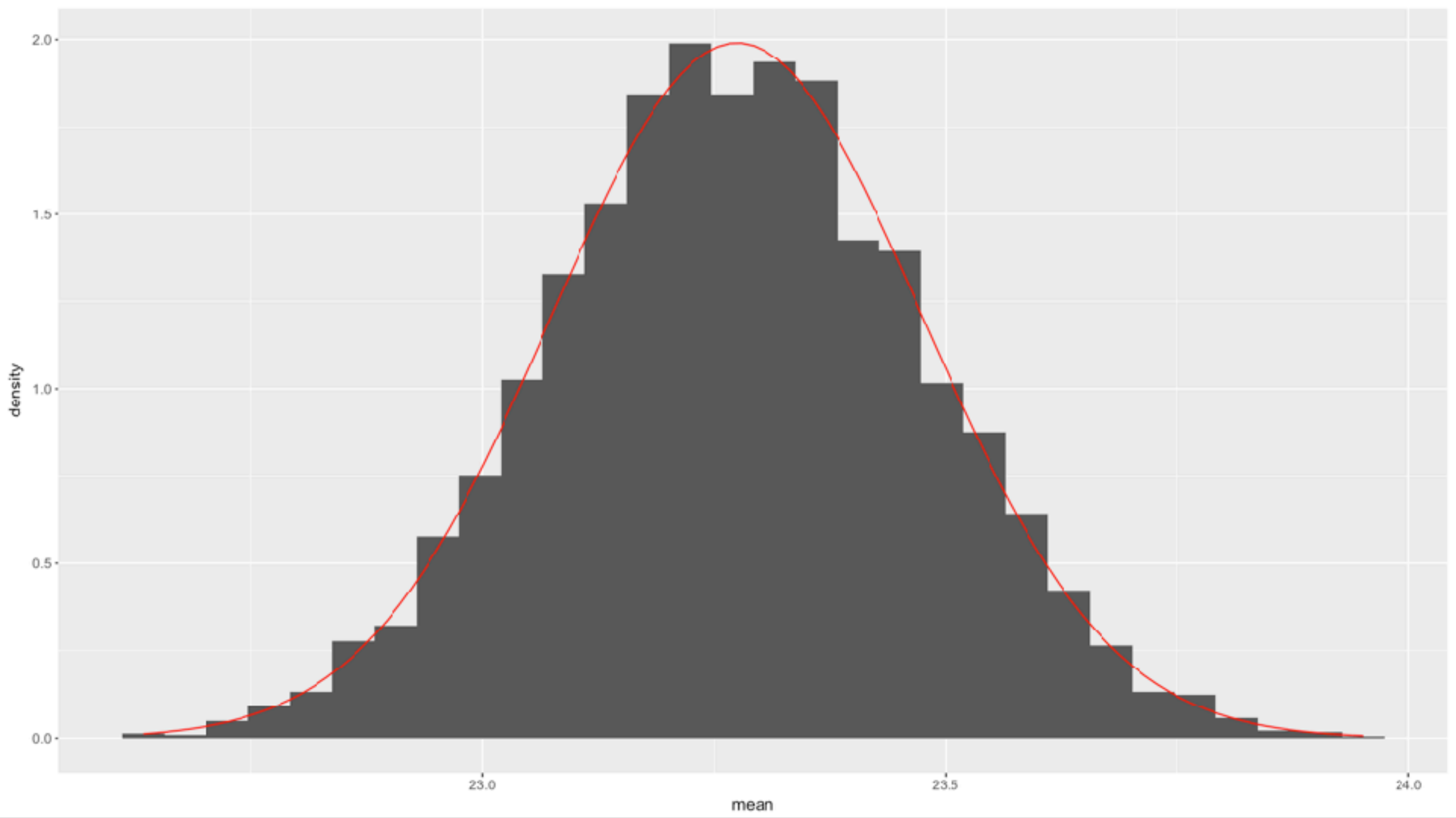
A SECOND RANDOM SAMPLE

```
> sample2 <- dplyr::sample_n(autoData, 500)
> summary(sample2$combFE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   12.0    19.0    23.0    23.3    26.0    58.0
> sd(sample2$combFE)
[1] 6.263133
```

ANOTHER 4,998 RANDOM SAMPLES

```
> summary(mpgSample$mean)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.63   23.14   23.27   23.27   23.41   23.95 
> sd(mpgSample$mean)
[1] 0.2003956
```

DISTRIBUTION OF K=5000 MEANS



THE “MAGIC” OF THE CLT

- This holds up for any population regardless of its underlying distribution.

https://goo.gl/qYaZlx

DEFINITION

- Population:
  - Parameters of μ , **σ**
  - Sample size of *n*
  - Sample means of $\overline{x}_1, \overline{x}_2, \overline{x}_3, \ldots \overline{x}_k$
- Distribution of $\overline{X}$:
  - Has mean of μ
  - Has a standard deviation of $\dfrac{\sigma}{\sqrt{n}}$
  - Normal as *n* → ∞

COMPARING A POPULATION AND RELATED SAMPLES

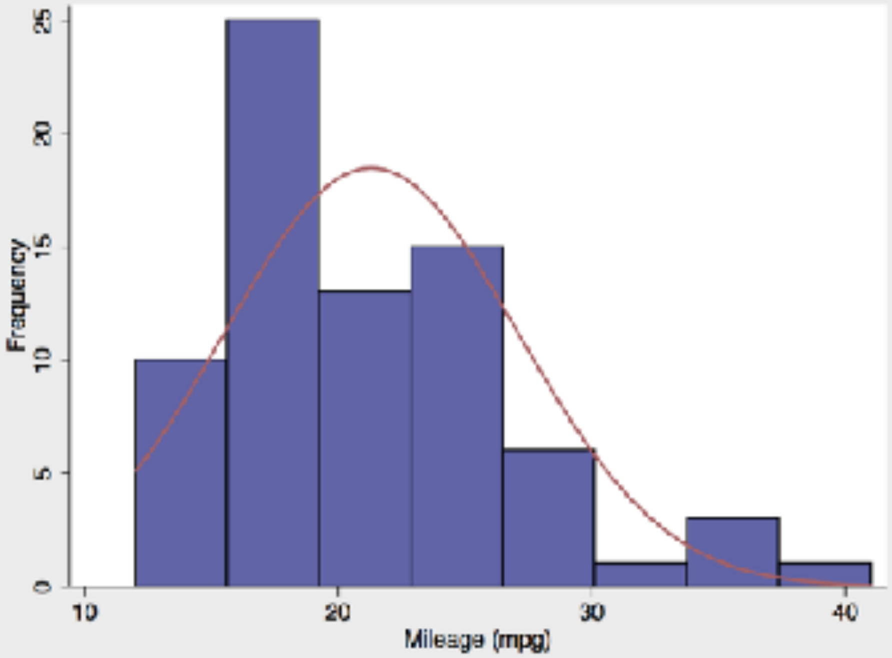
```
> mean(autoData$combFE)
[1] 23.27385
> sd(autoData$combFE)
[1] 5.83503


> mean(mpgSample$mean)
[1] 23.27471
> sd(mpgSample$mean)
[1] 0.2003956
```

STANDARD ERROR

- The standard deviation of the distribution of sample means ($\overline{X}$) is known as the *standard error*.
- A means for assessing the reliability of a particular statistic by estimating the difference between the sample statistic and the population statistic.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

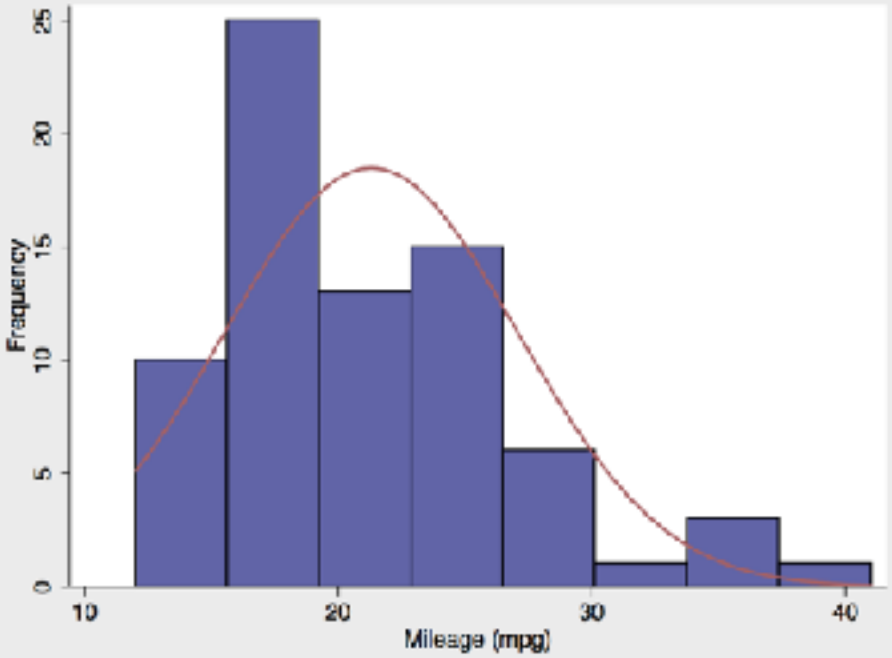
Population



$$\sigma_x = \sqrt{\frac{\sum (x-\mu)^2}{n}}$$

$$\mu = \frac{\sum x}{n}$$

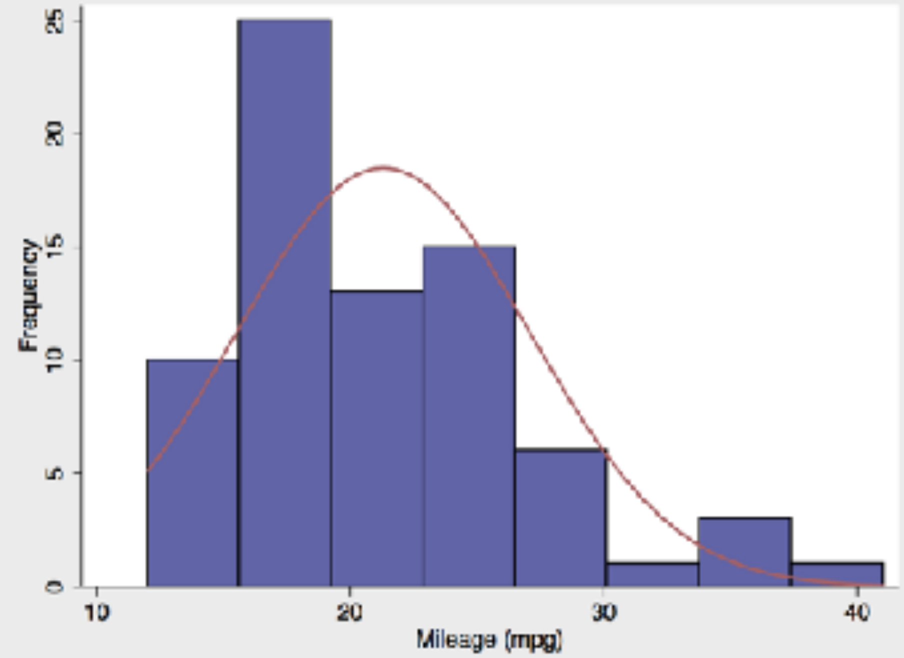
Population



$$\sigma_x = 5.83263$$

$$\mu = 23.27385$$

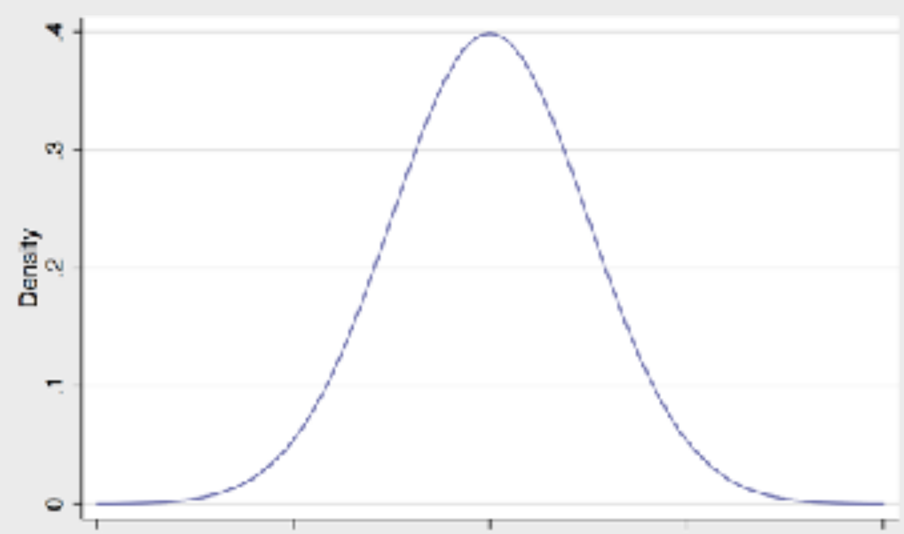
Population



$$\sigma_x = 5.83263$$

$$\mu = 23.27385$$

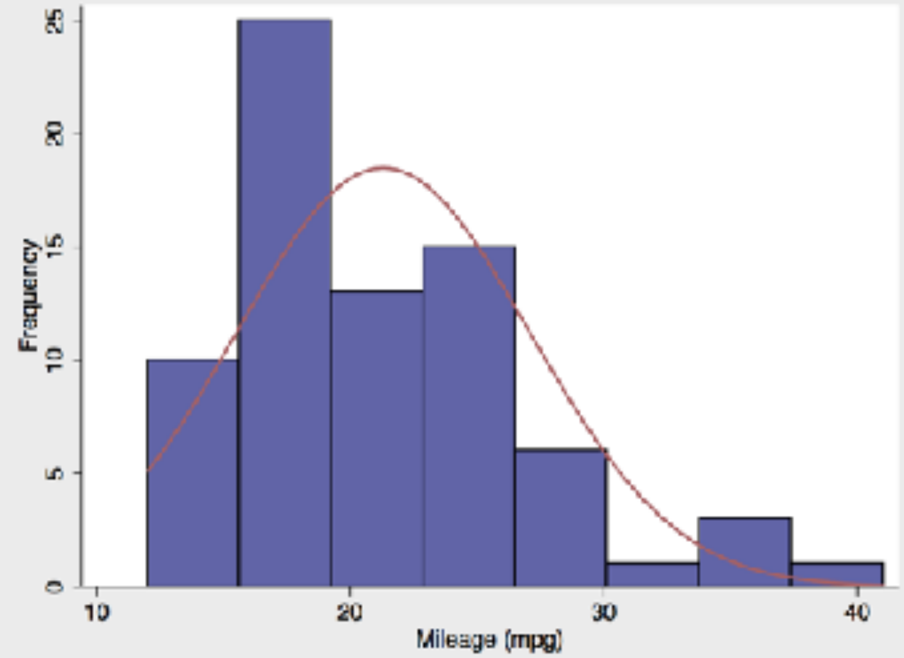
Samples of $\bar{X}$



$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

$$\mu = 23.27385$$

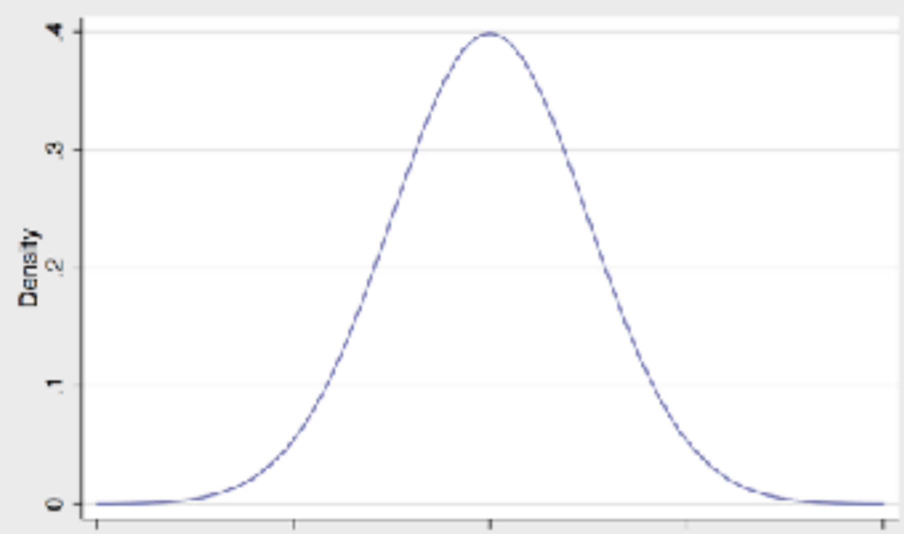
Population



$$\sigma_x = 5.83263$$

$$\mu = 23.27385$$

Samples of $\bar{X}$



$$\sigma_{\bar{x}} = \frac{5.83263}{\sqrt{500}} = 0.260$$

$$\mu = 23.27385$$

Z-SCORES

- The value of an observation expressed in standard deviations.

$$z = \frac{x - \mu}{\sigma}$$

Z-SCORES FOR SAMPLING DISTRIBUTIONS

- The value of an observation expressed in standard deviations.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Z-SCORES FOR SAMPLING DISTRIBUTIONS

- Taking repeated samples of *n*=500 from this population, what proportion of these samples will have means ≥ 25?

$$z = \frac{25 - 23.27385}{\frac{5.83263}{\sqrt{500}}} = \frac{1.72615}{0.260} = 6.639$$

Z-SCORES FOR SAMPLING DISTRIBUTIONS

- Taking repeated samples of *n*=500 from this population, what proportion of these samples will have means ≥ 25?

$$z = \frac{25 - 23.27385}{\frac{5.83263}{\sqrt{500}}} = \frac{1.72615}{0.260} = 6.639$$

```
> pnorm(6.6389, mean = 0, sd = 1, lower.tail = FALSE)
[1] 1.580164e-11
```

- The likelihood of obtaining a sample mean that is ≥ 25 from that population is very, very small.

ESTIMATING SAMPLE SIZES

- The CLT can be used to estimate sample sizes based on how close we want our sample to be to the population. This is one version of what we call *power analyses*.

$$\left(\frac{1.96\sigma}{\Delta}\right)^2$$

- The Greek uppercase letter Δ ("Delta") is used to represent the amount of error we are willing to tolerate.
- We want our sample to be within ± Δ of the population mean.

ESTIMATING SAMPLE SIZES

- Given the population parameters we have been using in this case for miles per gallon, what sample size would we need to have sample mean that is within 3 miles per gallon of the population's?

$$\left(\frac{1.96\sigma}{\Delta}\right)^2$$

ESTIMATING SAMPLE SIZES

- Given the population parameters we have been using in this case for miles per gallon, what sample size would we need to have sample mean that is within 3 miles per gallon of the population's?

$$\left(\frac{(1.96)\,(5.83263)}{3}\right)^2 = \left(\frac{11.4319548}{3}\right)^2 = (3.8106516)^2 = 14.521$$

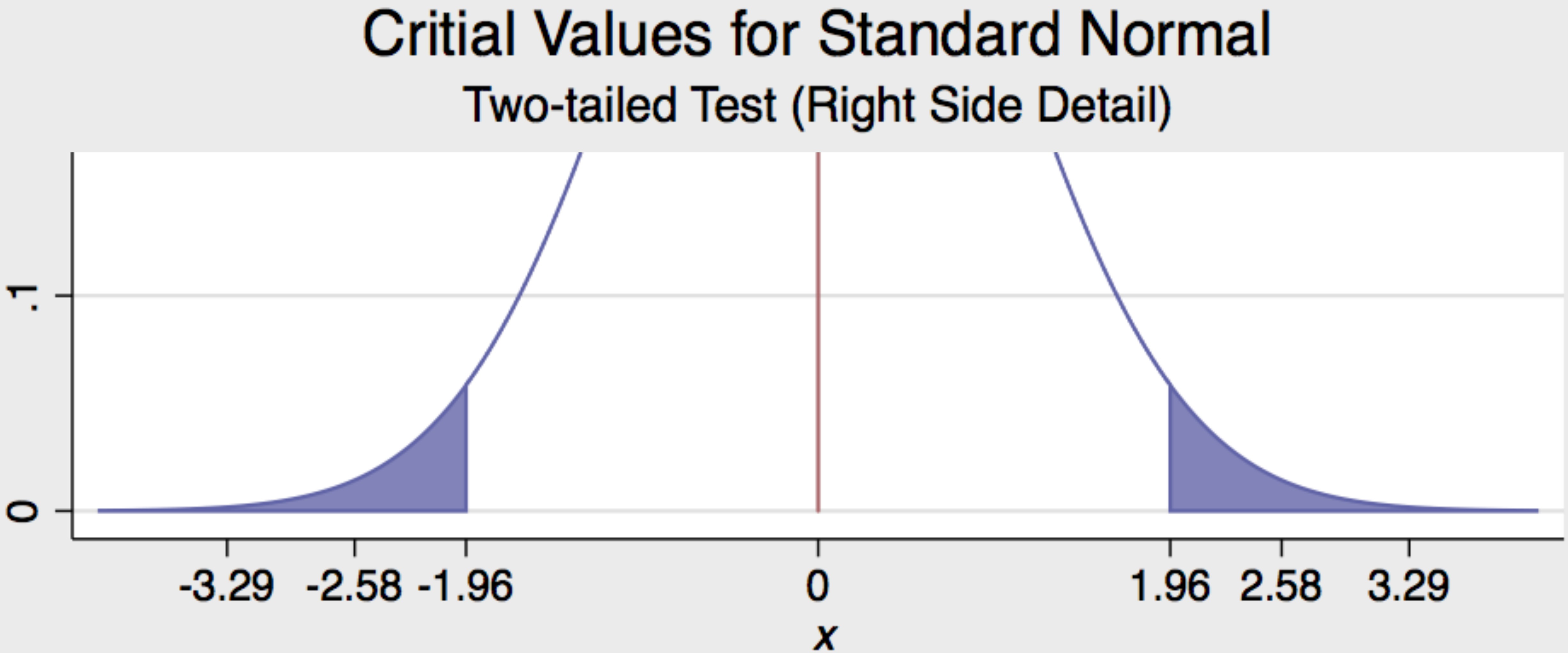
- We need to have a sample size of at least 15 vehicles to have a sample mean within 3 miles per gallon of the population's.
- To be within 2 miles per gallon, we need *n*=32.
- To be within 1 miles per gallon, we need *n*=127.

4 CONFIDENCE INTERVALS

THE PREDICTIVE INTERVAL

- Related to the confidence interval.
- Can be used prior to sampling to estimate a value for both *x* and $\bar{x}$.
- Use z-scores from two-sided critical values.

$$(\mu - 1.96\sigma, \mu + 1.96\sigma)$$



z	1.96	2.58	3.29
p <	0.05	0.01	0.001
% of scores inside	95%	99%	99.9%
% of scores outside	5%	1%	0.1%

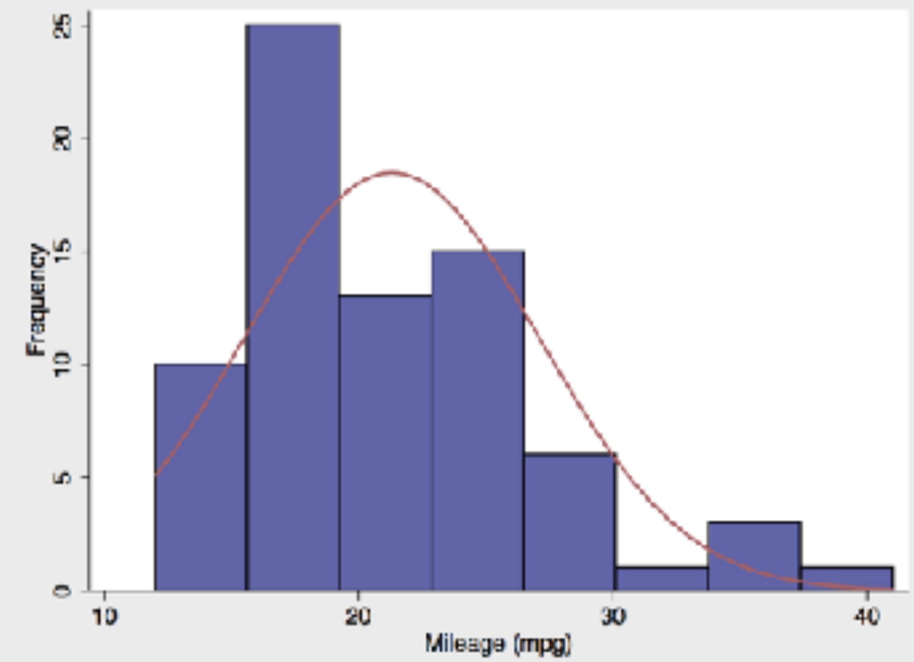
THE PREDICTIVE INTERVAL

$$(\mu - 1.96\sigma, \mu + 1.96\sigma)$$

$$(21.297 - (1.96)(5.746), 21.297 + (1.96)(5.746))$$

$$(21.297 - 11.26216, 21.297 + 11.26216)$$

$$(10.03484, 32.55916)$$



$$\mu = 21.297$$
$$\sigma_x = 5.746$$

- Based on the predictive interval, a given value of *x* selected at random will fall between 10.035 and 32.559 95% percent of the time.

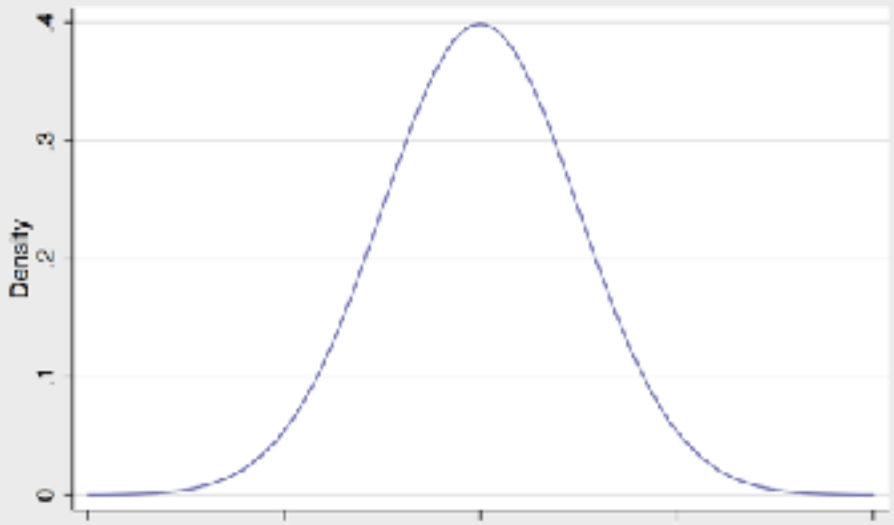
THE PREDICTIVE INTERVAL

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$(21.297 - (1.96)(0.909), 21.297 + (1.96)(0.909))$$

$$(21.297 - 1.78164, 21.297 + 1.78164)$$

$$(19.51536, 23.07864)$$



$$\mu = 21.297$$

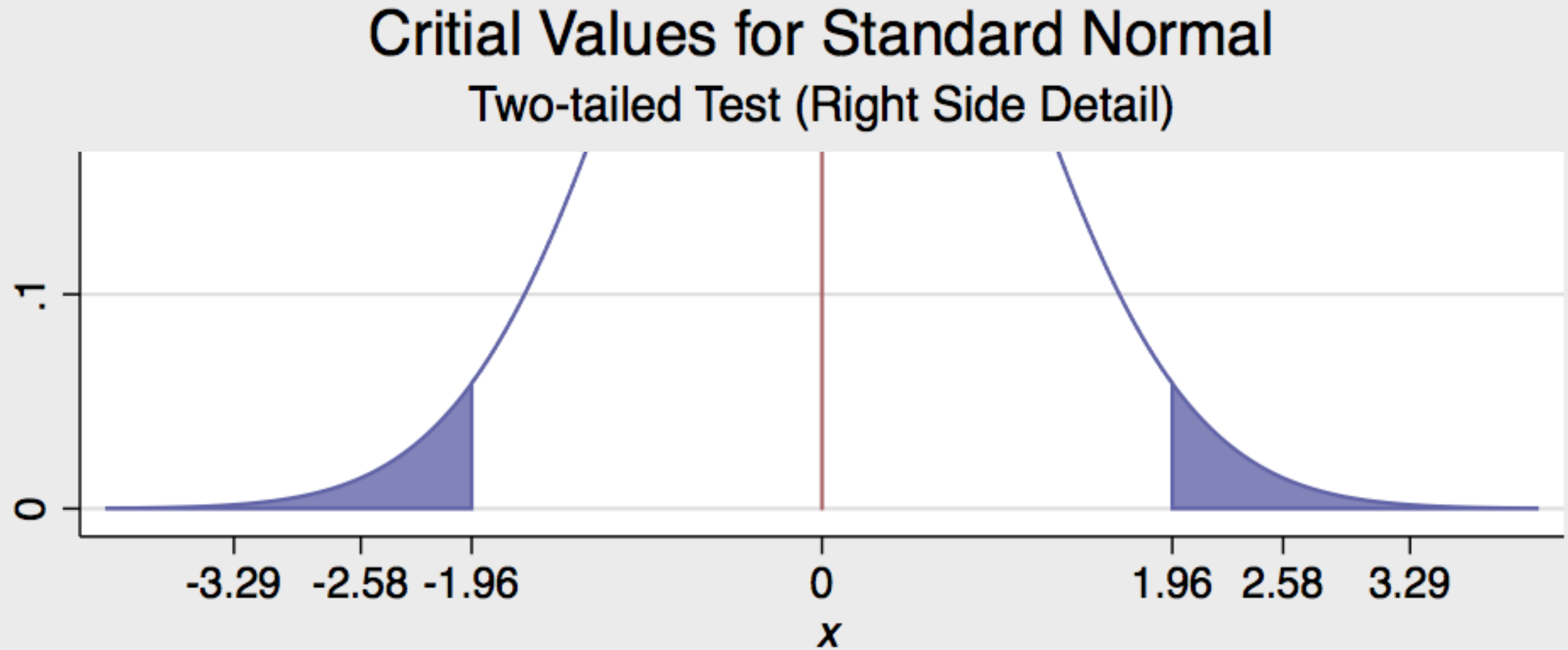
$$\sigma_{\bar{X}} = \frac{5.746}{\sqrt{40}} = 0.909$$

- Based on the predictive interval, a sample mean will fall between 19.515 and 23.079 95% percent of the time.

THE CONFIDENCE INTERVAL

- Used after sampling to the amount of possible error between the given sample mean (for example) and the population sample mean.
- Like predictive intervals, use z-scores from two-sided critical values.

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

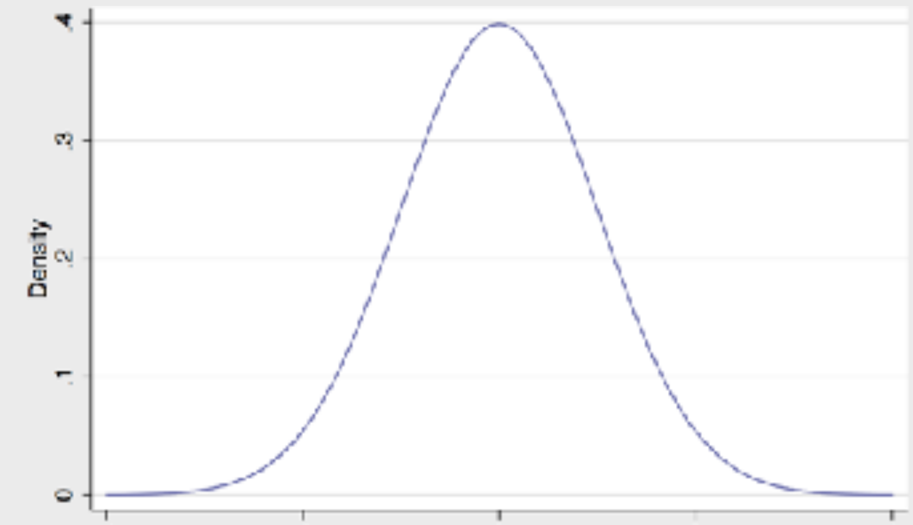


z	1.96	2.58	3.29
p <	0.05	0.01	0.001
% of scores inside	95%	99%	99.9%
% of scores outside	5%	1%	0.1%

THE CONFIDENCE INTERVAL

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$(\bar{x} - 1.78164, \bar{x} + 1.78164)$$



$$\mu = 21.297$$

$$\sigma_{\bar{X}} = \frac{5.746}{\sqrt{40}} = 0.909$$

- If we take a sample of size n=40 from our population, the the interval of the sample mean ± 1.782 has a 95% chance of covering μ.

WIDTH OF CONFIDENCE INTERVALS

Confidence Interval	Formula	Width
95%	$\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$	$3.92 \frac{\sigma}{\sqrt{n}}$
99%	$\overline{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$	$5.16 \frac{\sigma}{\sqrt{n}}$

CONFIDENCE INTERVALS & N

n	95% CI for μ	Width
10	$\overline{X} \pm 0.620\sigma$	1.240σ
100	$\overline{X} \pm 0.196\sigma$	0.392σ
1000	$\overline{X} \pm 0.062\sigma$	0.124σ

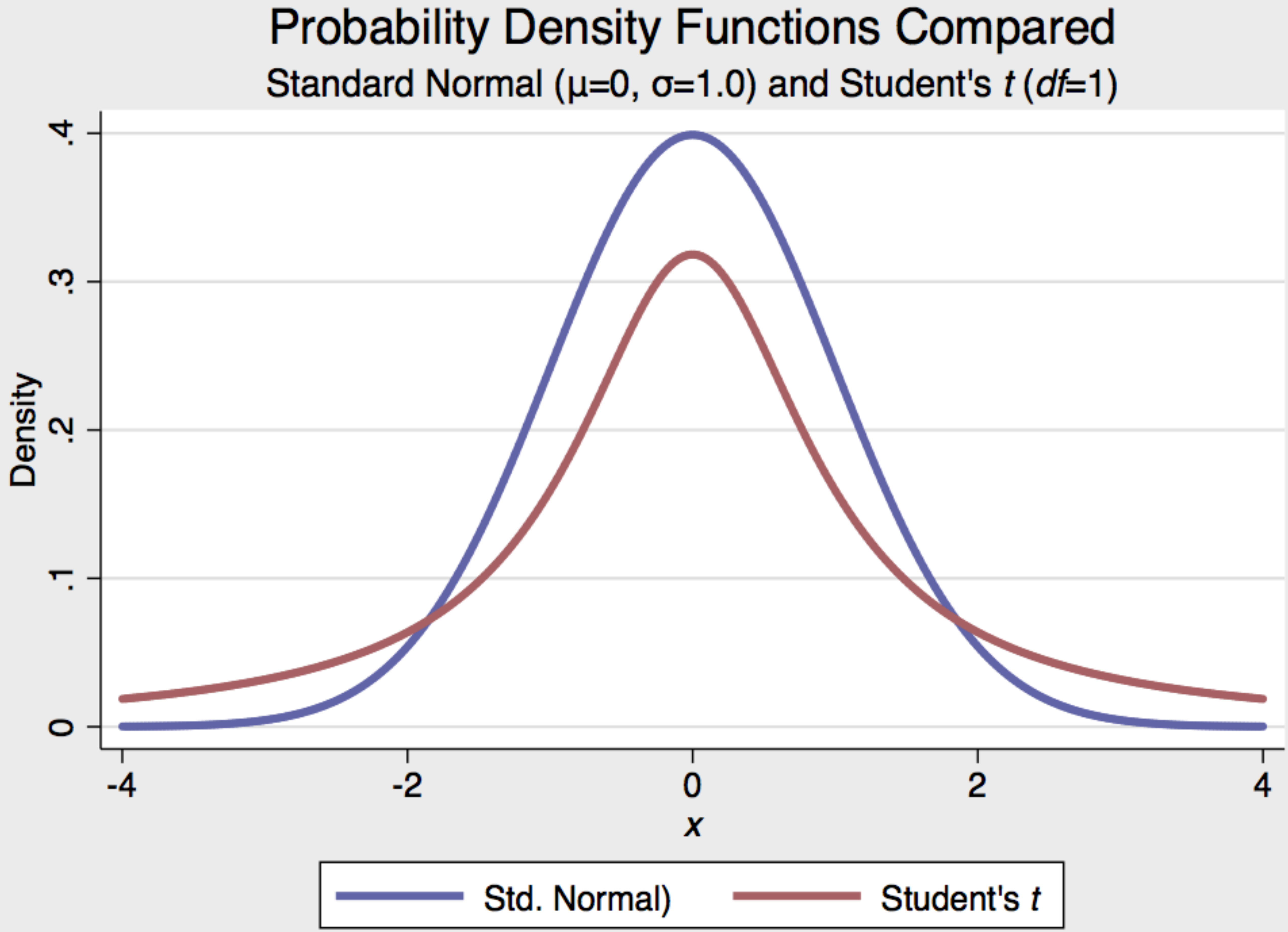
5 HYPOTHESIS TESTING

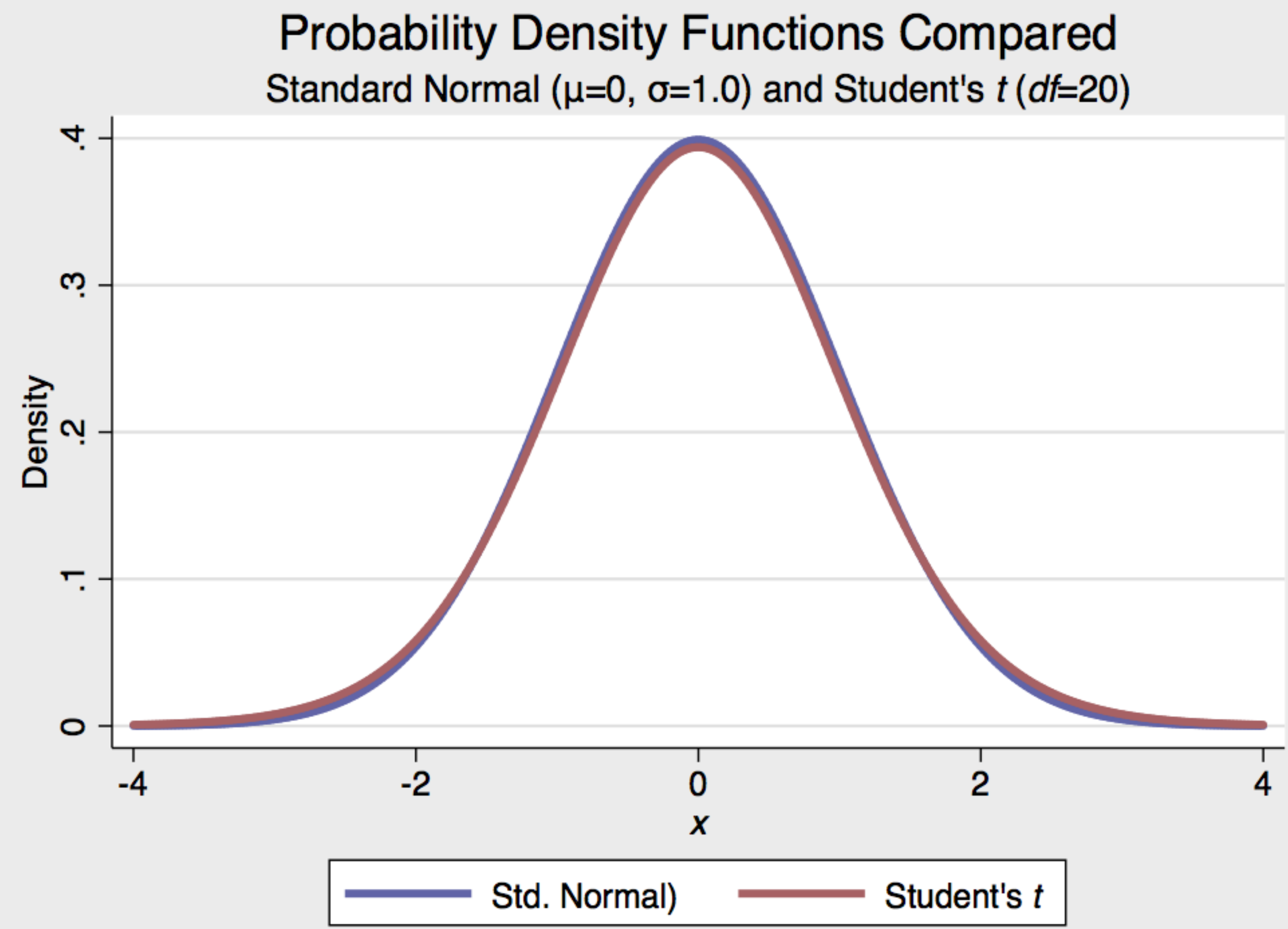
THE PROBLEM: EVERYTHING WE HAVE DONE SO FAR ASSUMES WE KNOW THE POPULATION PARAMETERS

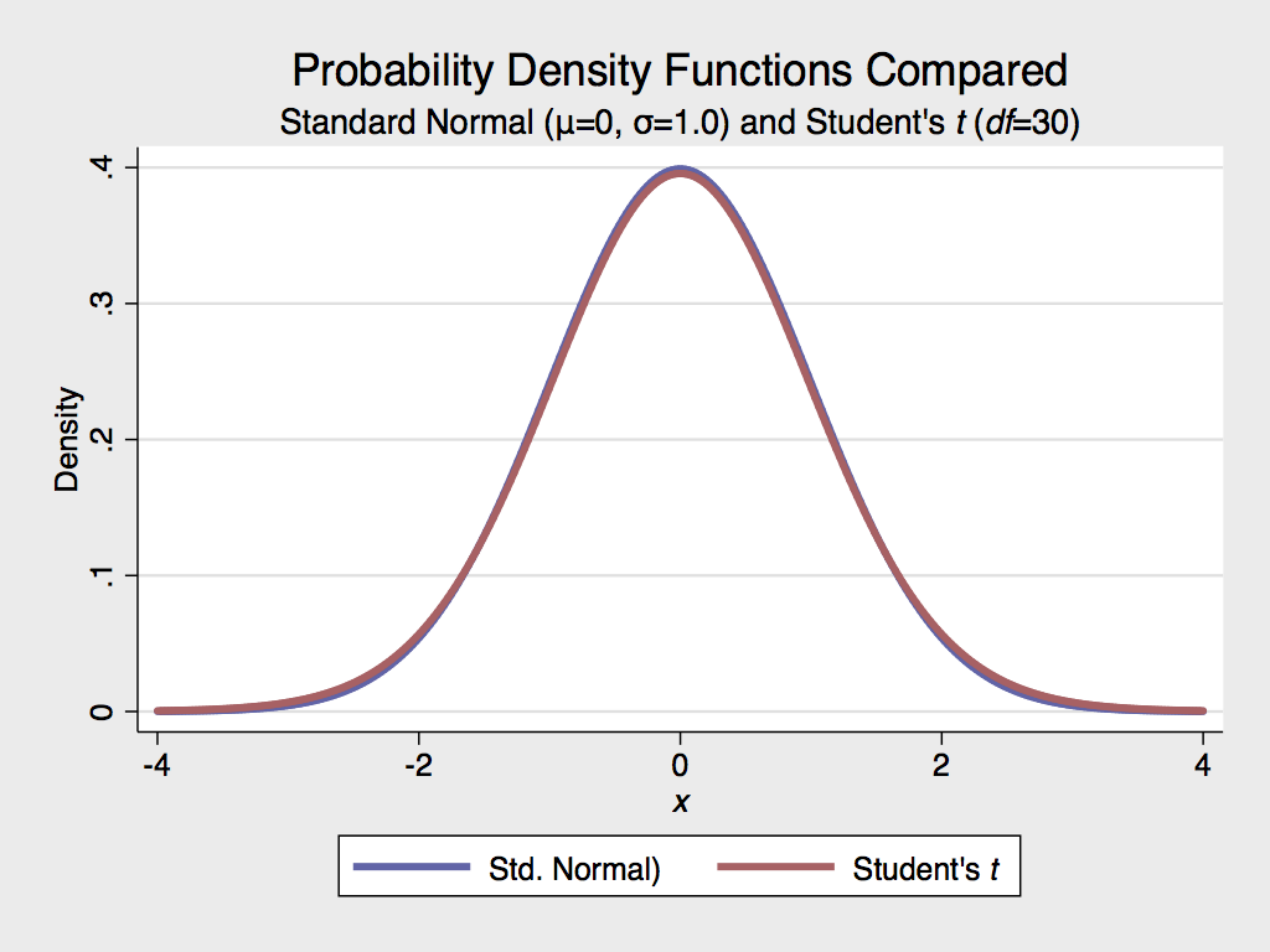
AN ALTERNATIVE

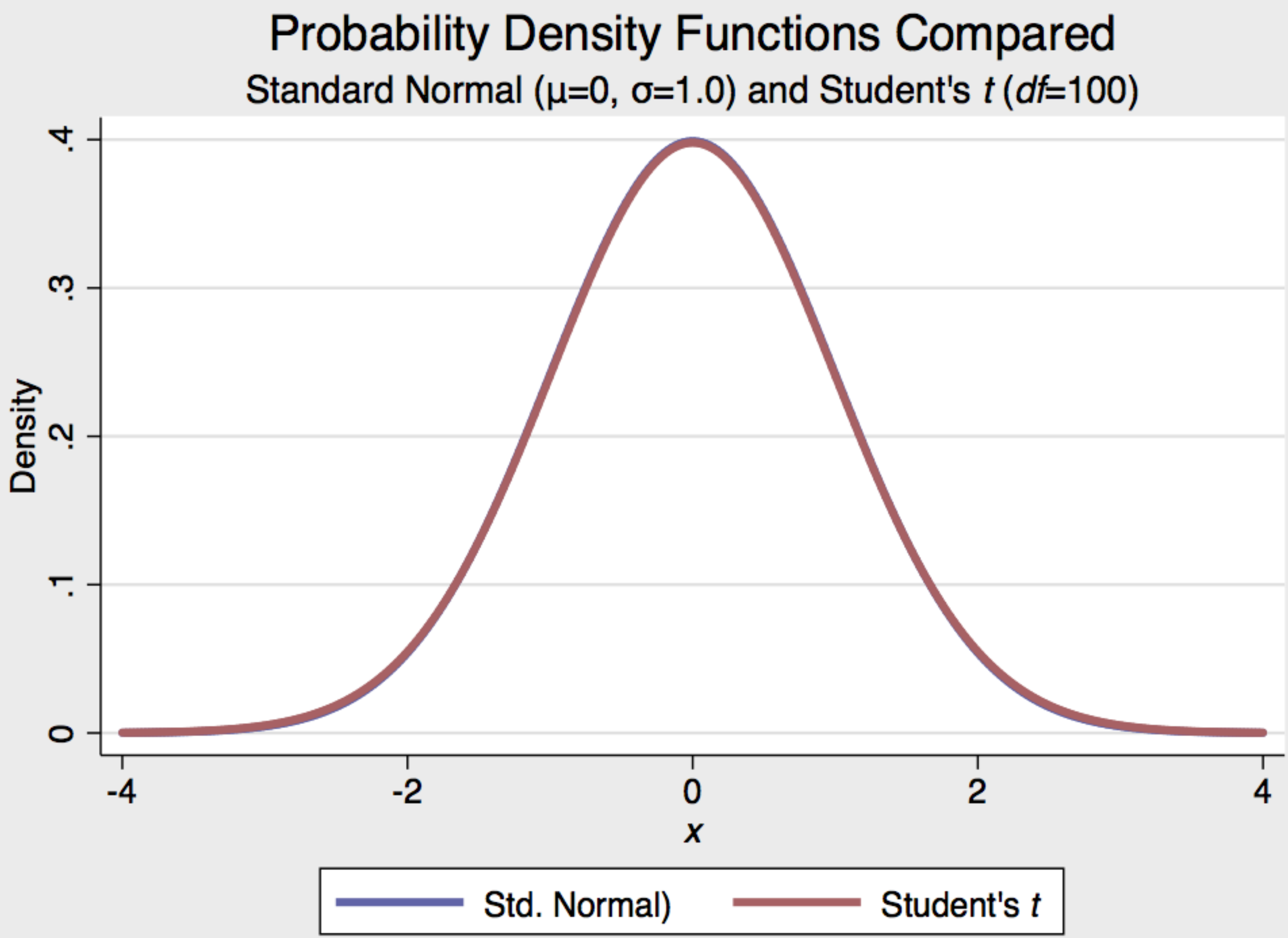
- As part of his work with Guisness, Gosset identified a solution to the problem of not knowing the population parameters.
- The Student’s *t* distribution approximates normal once the degrees of freedom (*n*-1) is ≥ 30.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

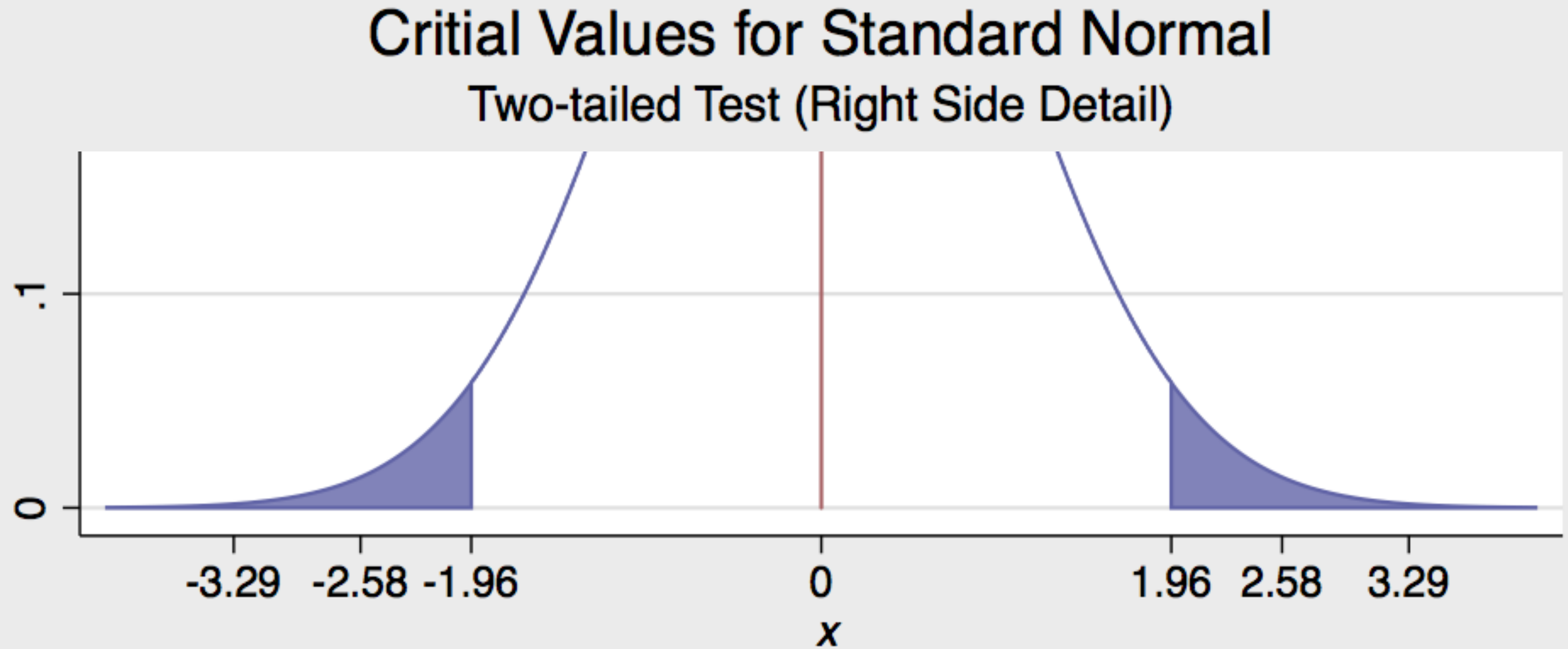

Probability Density Functions Compared
Standard Normal (μ=0, σ=1.0) and Student's t (df=1)
Density
.4
.3
.2
.1
0
-4
-2
0
2
4
x
Std. Normal)
Student's t


Probability Density Functions Compared
Standard Normal (μ=0, σ=1.0) and Student's t (df=20)
Density
0
.1
.2
.3
.4
-4
-2
0
2
4
x
Std. Normal)
Student's t


Probability Density Functions Compared
Standard Normal (μ=0, σ=1.0) and Student's t (df=30)
Density
.4
.3
.2
.1
0
-4
-2
0
2
4
x
Std. Normal)
Student's t



Probability Density Functions Compared
Standard Normal (μ=0, σ=1.0) and Student's t (df=100)
Density
.4
.3
.2
.1
0
-4
-2
0
2
4
x
Std. Normal)
Student's t



z	1.96	2.58	3.29
p <	0.05	0.01	0.001
% of scores inside	95%	99%	99.9%
% of scores outside	5%	1%	0.1%

ERROR

Sample	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not Reject	yes	Type II
Reject	Type I	yes

*The null hypothesis is that $\mu=\mu_0$

ERROR

Pr(Type II)=β

1-β=power

Sample	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not Reject	yes	Type II
Reject	Type I	yes

*The null hypothesis is that $\mu=\mu_0$

Pr(Type I)=α

THE PROBABILITY OF GETTING RESULTS AT LEAST AS EXTREME AS THE ONES YOU OBSERVED, GIVEN THAT THE NULL HYPOTHESIS IS CORRECT

Christie Aschwanden

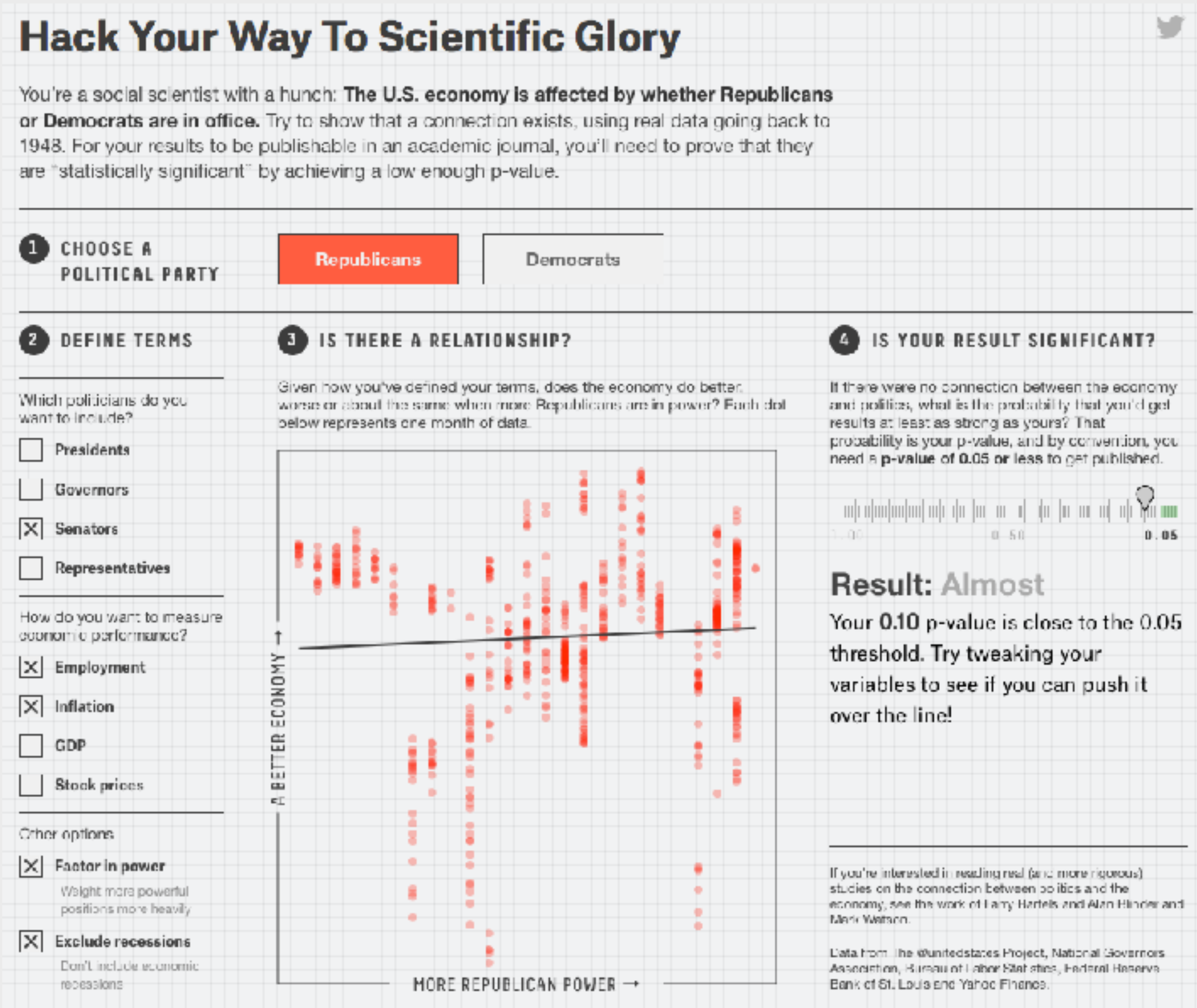
FiveThirtyEight's [p-value story]

AES STATEMENT ON P-VALUES

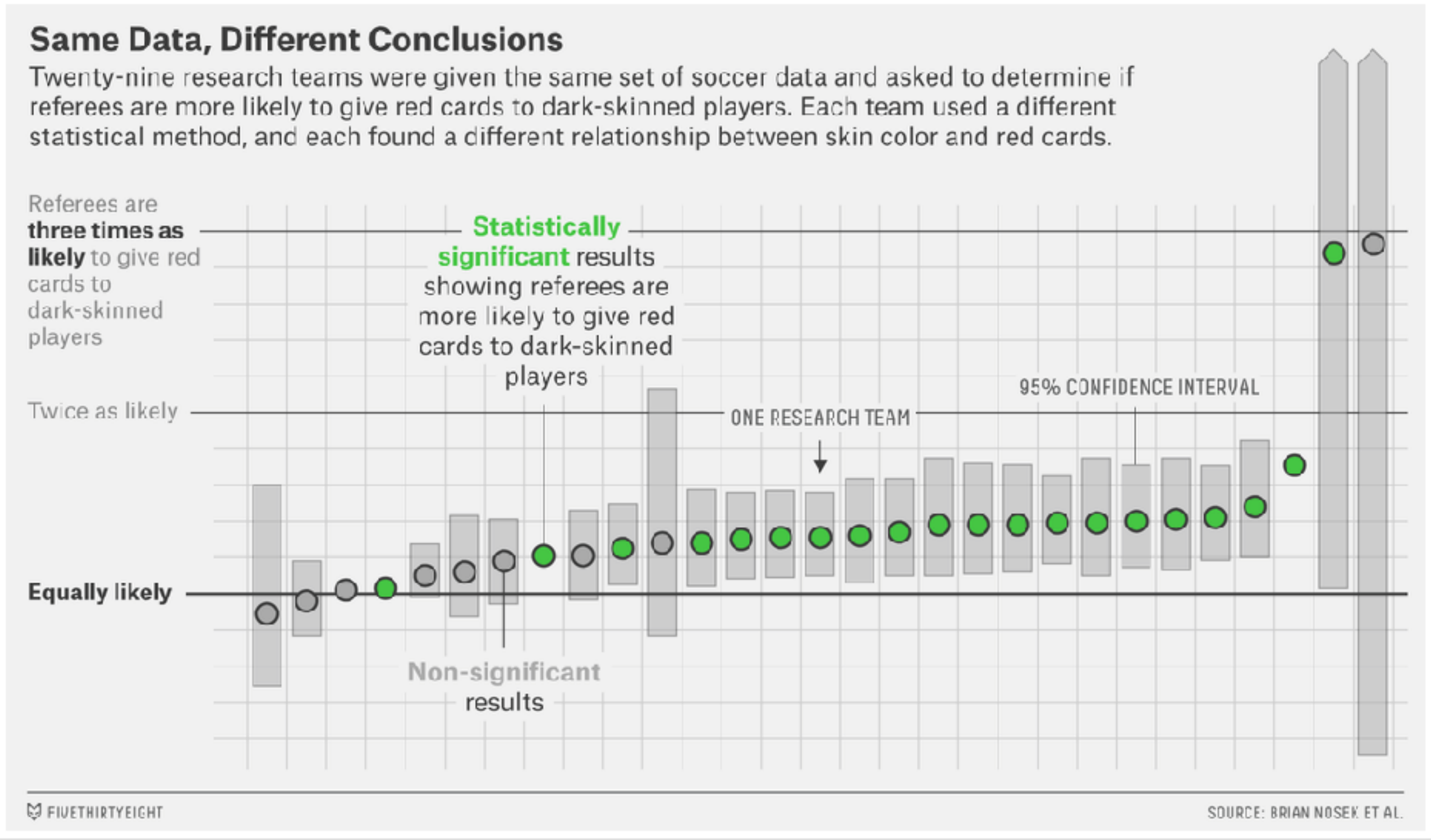
1. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
2. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
3. Proper inference requires full reporting and transparency.
4. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
5. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

P-HACKING

https://goo.gl/3oVKaP



P-HACKING


Same Data, Different Conclusions
Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.
Referees are three times as likely to give red cards to dark-skinned players
Statistically significant results showing referees are more likely to give red cards to dark-skinned players
Twice as likely
ONE RESEARCH TEAM
95% CONFIDENCE INTERVAL
Equally likely
Non-significant results
FIVETHIRTYEIGHT
SOURCE: BRIAN NOSEK ET AL.