Unit 6: CLT based inference

1. Central Limit Theorem

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

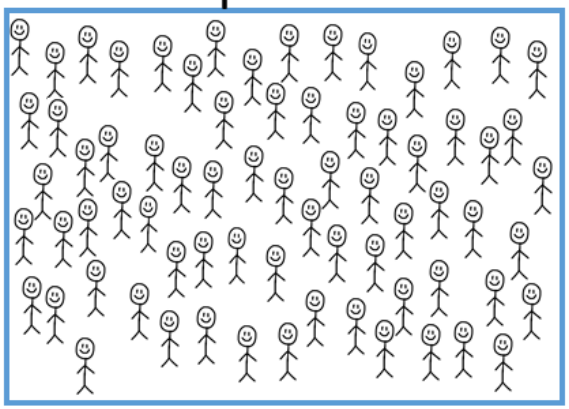
Outline

- EA 04
  - Remember our goal: tell your story with data
  - Describe what you saw in the figures and summary table, and explain how it helped illustrate the main point in your story
  - Re-submission (up to 50% lost points); submit with EA05
- EA 05
  - Due Friday (Oct 2) 6pm
- EA 06/07 (Oct 9)
  - EA 06: final paper (grade based on group work)
  - EA 07: presentation (grade based on individual performance)
- Optional final oral exam

Outline

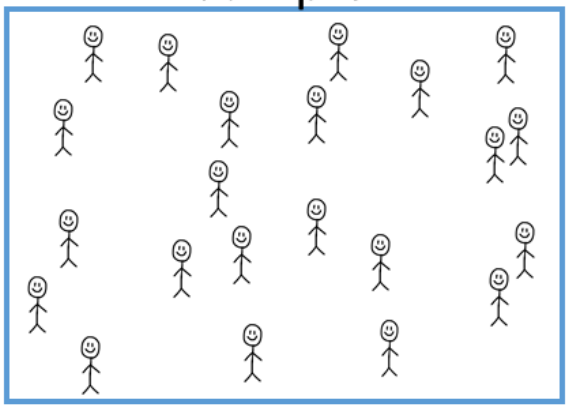
So far: estimate population parameter from sample statistics

Population



Take Sample

Sample



Calculate Sample Statistic

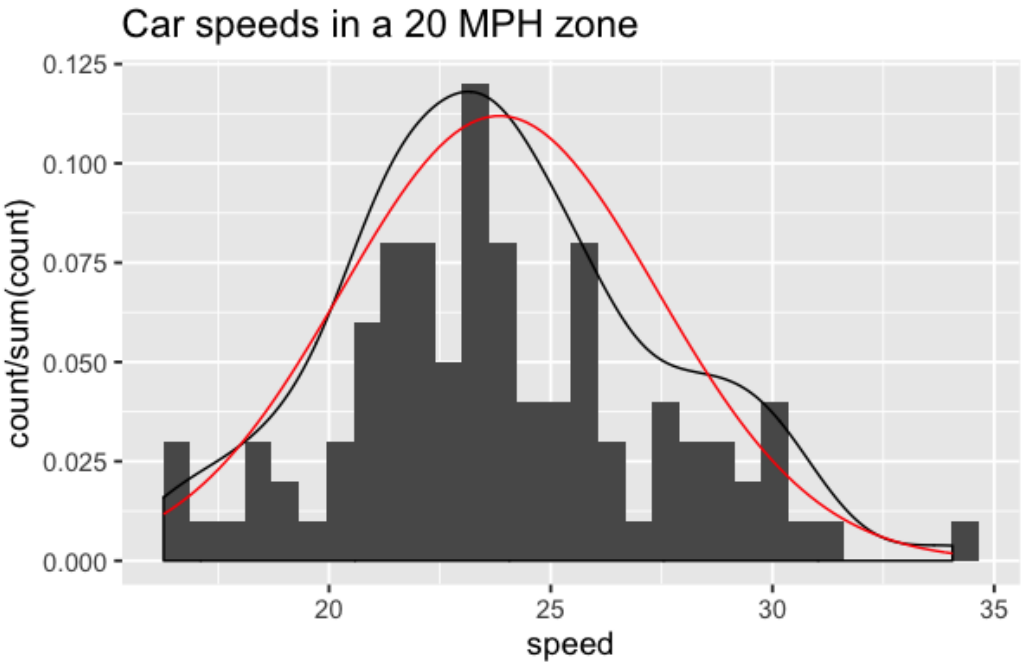
$\hat{p}$: **sample proportion** – what proportion of people in the **sample** support the president?

Inference

Our Goal: Estimate the Population Parameter

p: **population proportion** – what proportion of people in the **population** support the president?

- The symmetric, bell-shaped curve arises everywhere
- A normal distribution $\mathcal{N}(\mu, \sigma)$ is determined by two things
  - The **mean** μ of a normal distribution shows where it is centered.
  - The **standard deviation** σ of a normal distribution shows how spread out the normal is.



In March 2011, a random sample of 1000 US adults were asked

"Do you think exercise is important"

753 adults responded they think exercise is important.

Question

Give a 95% Confidence Interval (CI) for the proportion of all US adults that think exercise is important.

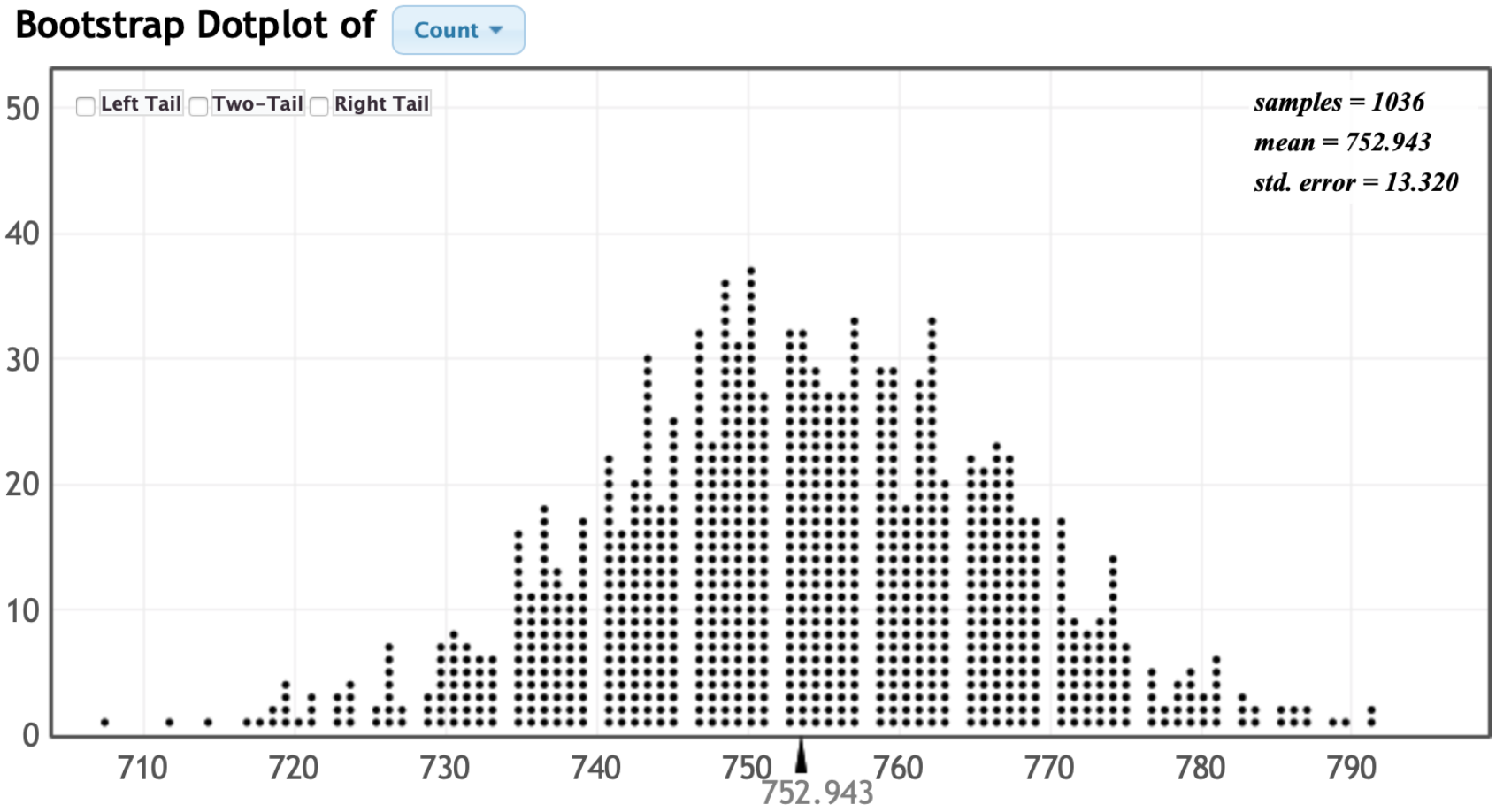
We first identify ...

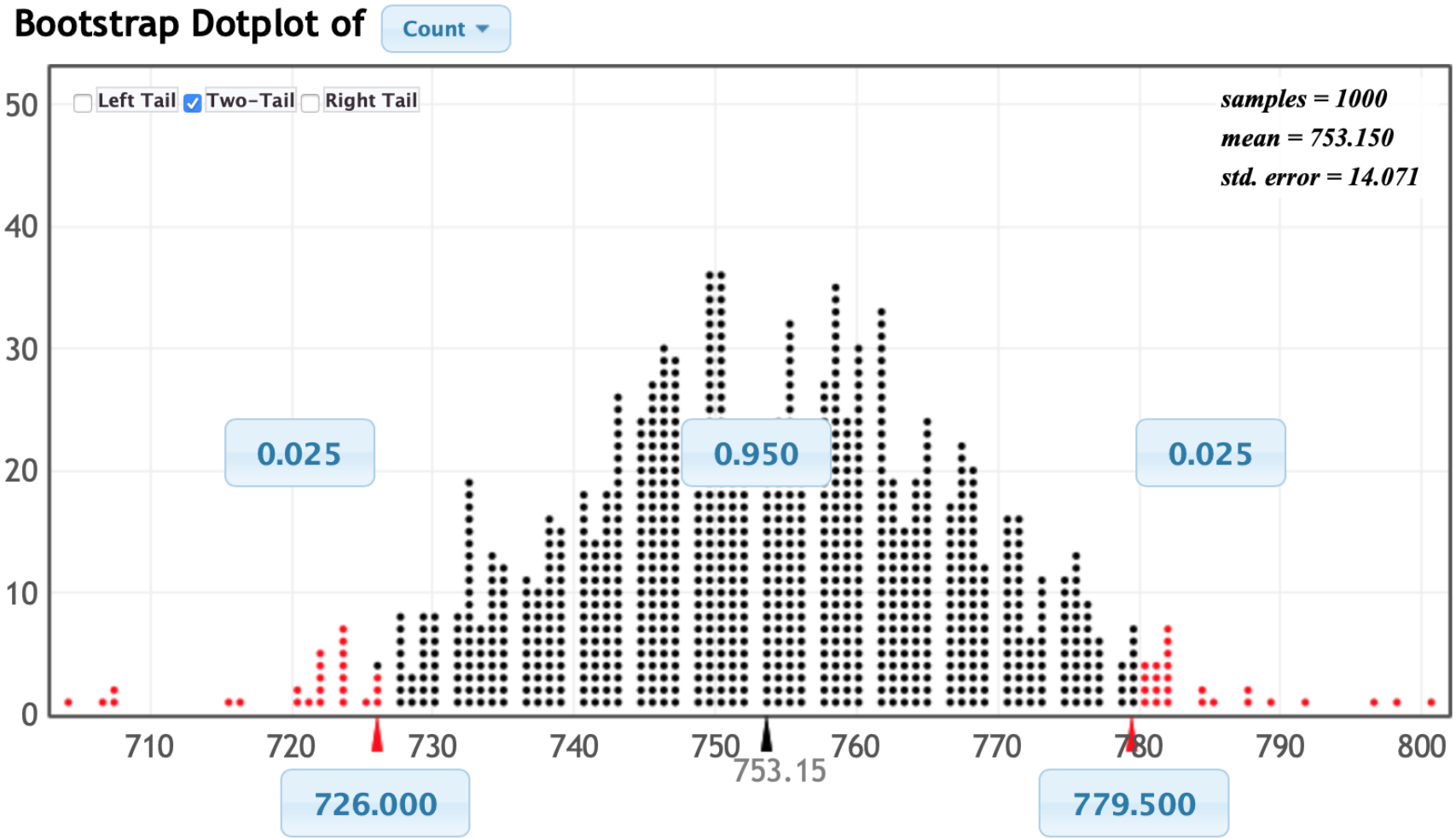
- Population: all adults in US
- Sample: 1000 adults in the survey
- Population parameter: p which denotes the proportion of all adults that think exercise is important
- Sample statistic: $\hat{p}$ which denotes the proportion of the adults that think exercise is important in a sample of sample size 1000

There is only one sample.
Let's bootstrap

Let's play together...

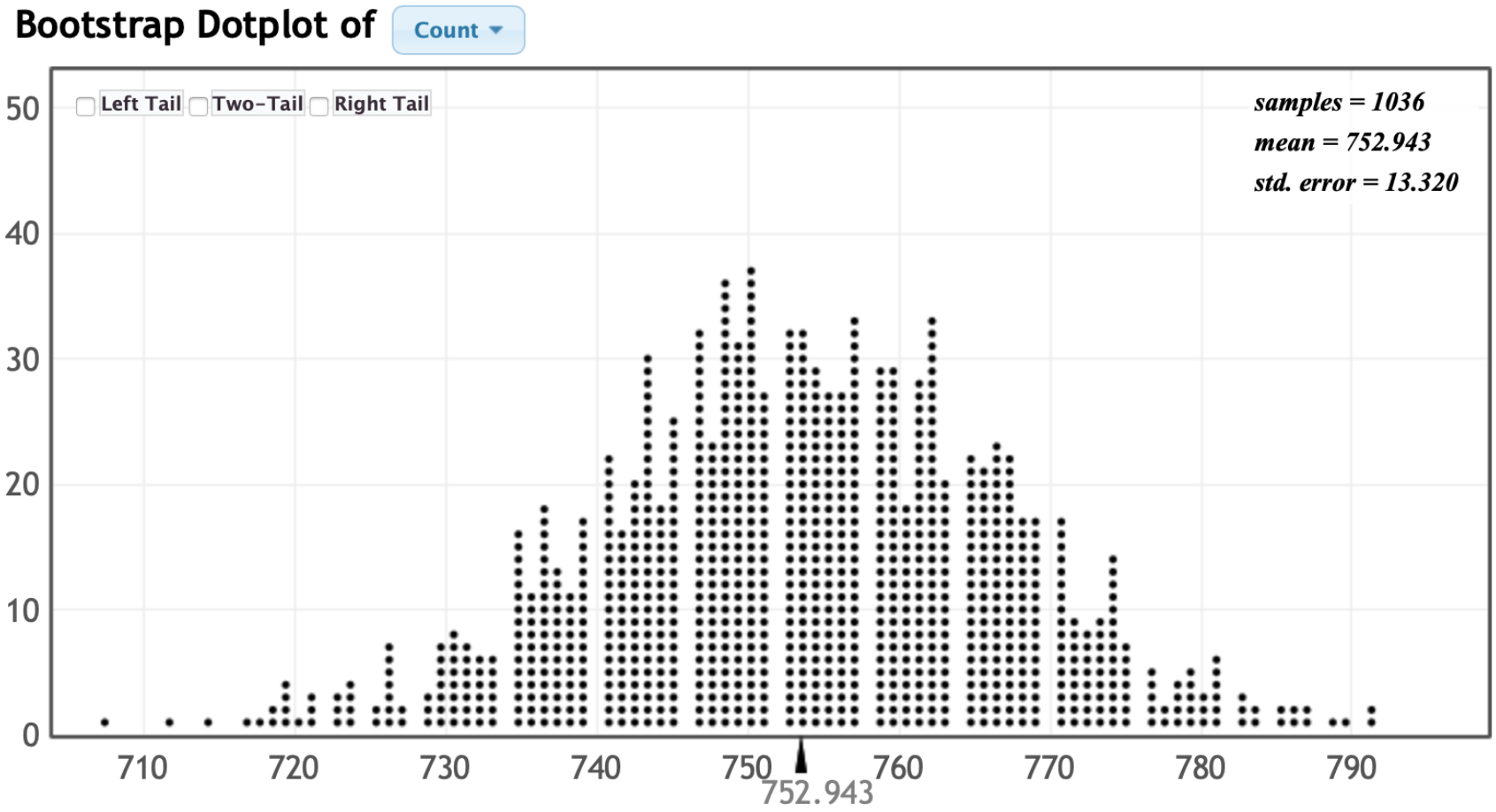
`http://www.lock5stat.com/StatKey/bootstrap_1_cat/bootstrap_1_cat.html`


Bootstrap Dotplot of
Count
Left Tail
Two-Tail
Right Tail
samples = 1036
mean = 752.943
std. error = 13.320
50
40
30
20
10
0
710
720
730
740
750
760
770
780
790
752.943



We compute the proportion of the distribution falling within the middle 95% of the bootstrap distribution.

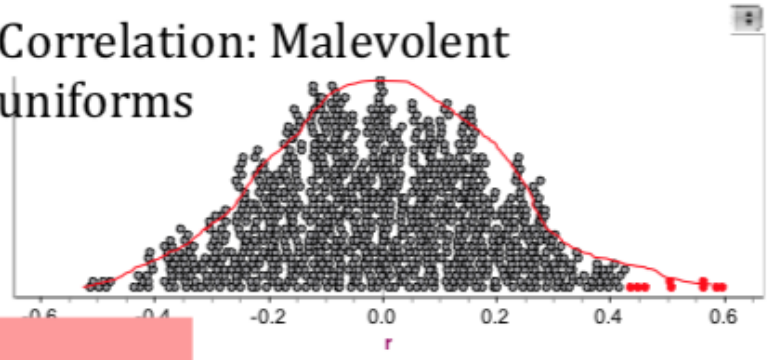
Put a normal curve on the bootstrap distribution



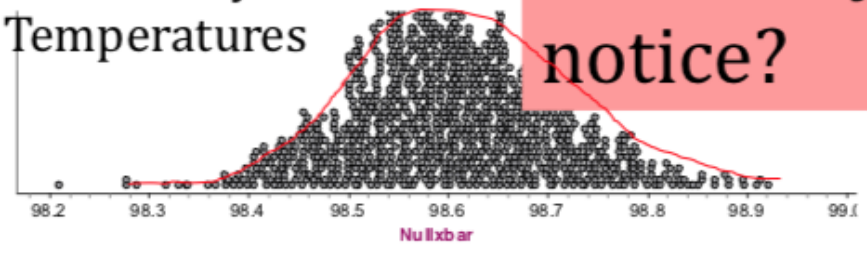
Is the sample proportion normally distributed?

Slope :Restaurant tips

Correlation: Malevolent uniforms

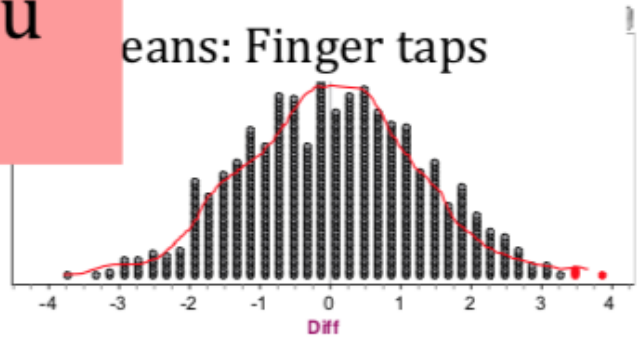


Mean :Body Temperatures

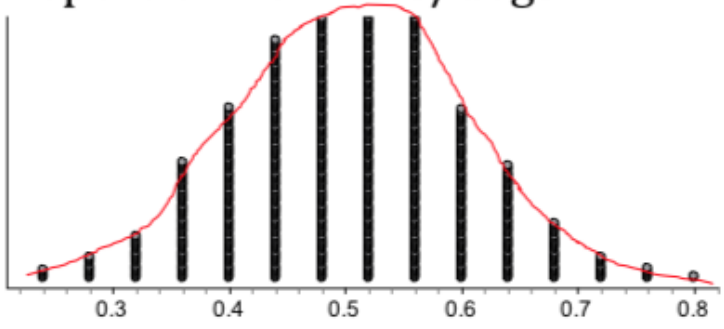


What do you notice?

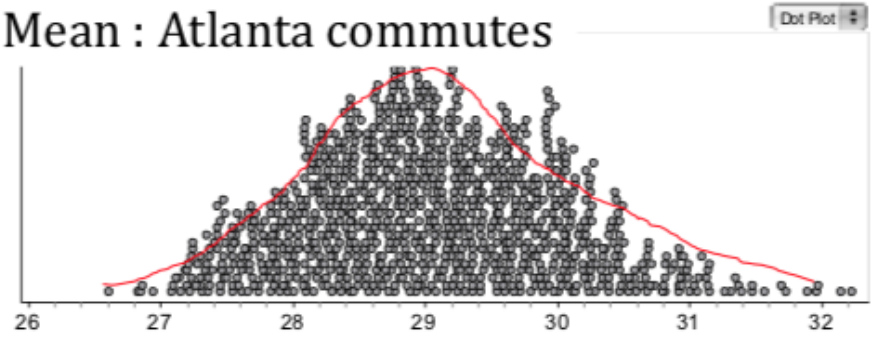
eans: Finger taps



Proportion : Owners/dogs



Mean : Atlanta commutes



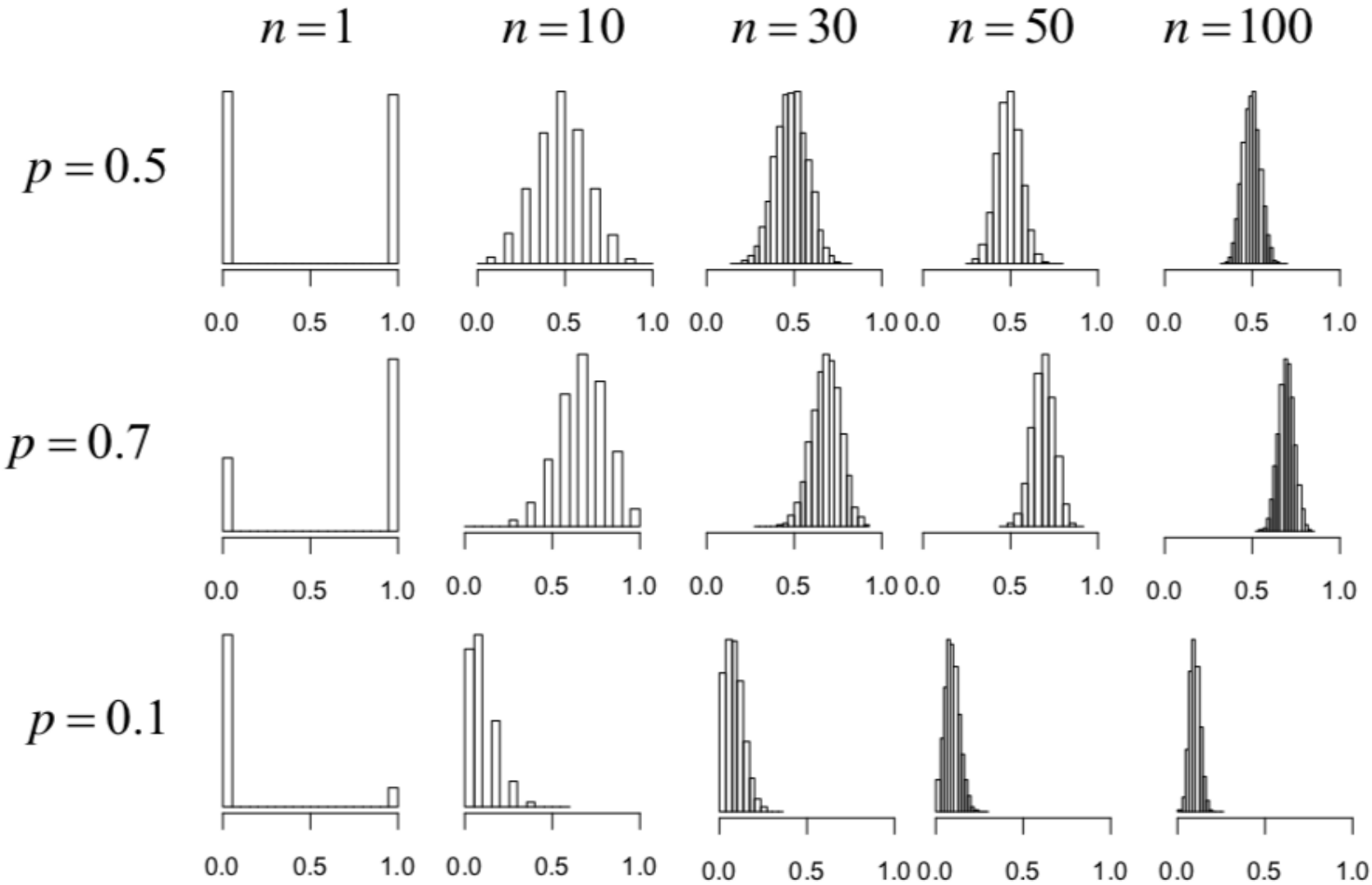
Outline

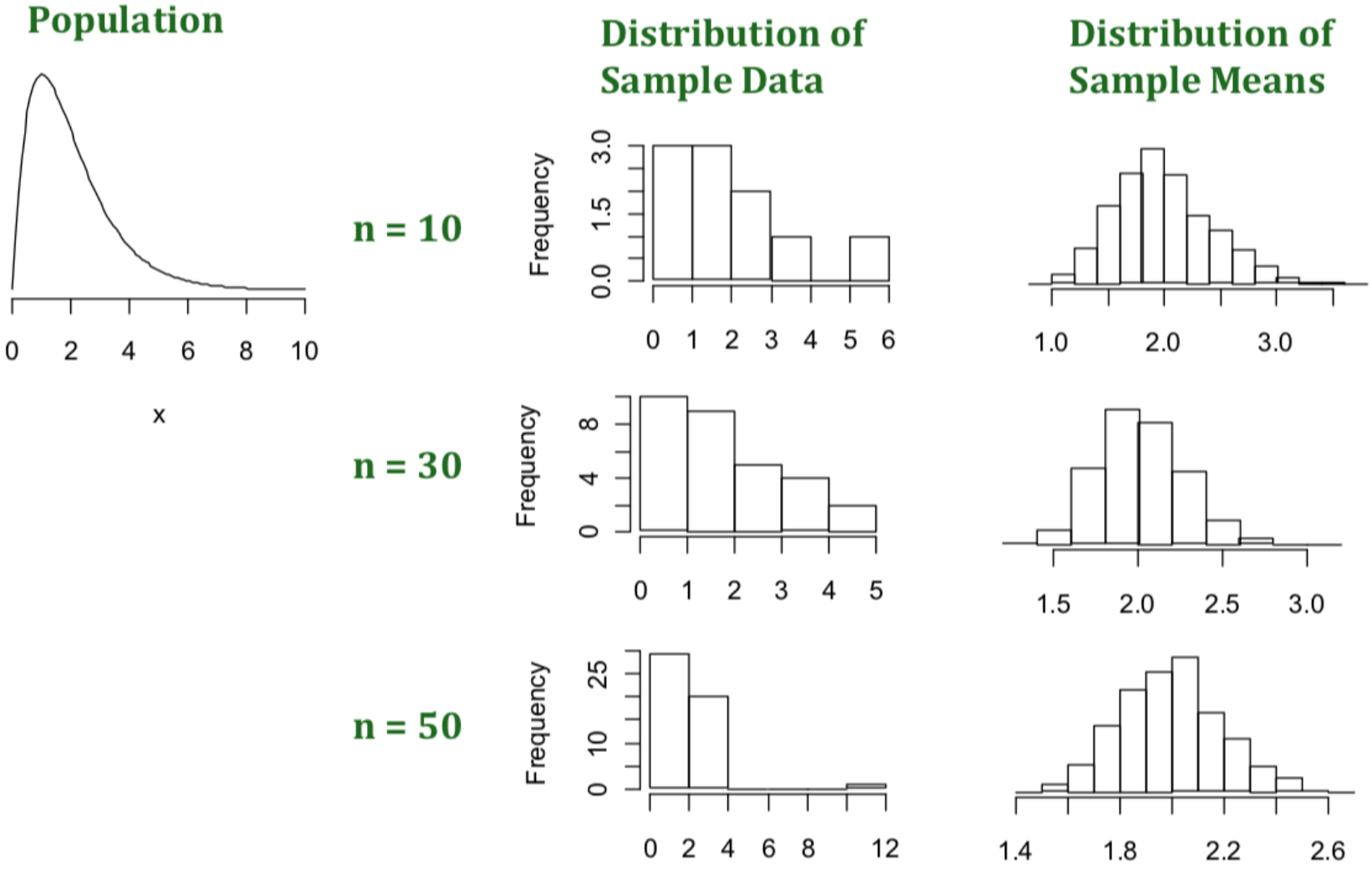
For a sufficiently large sample size, the distribution of sample proportion is normal.

For a sufficiently large sample size, the distribution of sample proportion is normal.

Also true for

- sample mean
- difference in sample mean
- difference in sample proportion
- ...


$n = 1$
$n = 10$
$n = 30$
$n = 50$
$n = 100$
$p = 0.5$
$p = 0.7$
$p = 0.1$
0.0
0.5
1.0


Population
Distribution of Sample Data
Distribution of Sample Means
n = 10
n = 30
n = 50
Frequency
x
0 2 4 6 8 10
0 1 2 3 4 5 6
1.0 2.0 3.0
0 1 2 3 4 5
1.5 2.0 2.5 3.0
0 2 4 6 8 12
1.4 1.8 2.2 2.6

- The central limit theorem holds for ANY original distribution, although "sufficiently large sample size" varies
- The more skewed the original distribution is (the farther from normal), the larger the sample size has to be for the CLT to work
- For small samples, it is more important that the data itself is approximately normal

"The theory of probabilities is at bottom nothing but common sense reduced to calculus."
– Laplace, in *Théorie analytique des probabilités*, 1812

We need to check the following two conditions

- Independence: Sampled observations must be independent. This is difficult to verify, but is more likely if
  - random sampling/assignment is used, and,
  - if sampling without replacement, $n < 10\%$ of the population.
- Sample size/skew: Either
  - the population distribution is normal or
  - $n > 30$ and the population dist. is not extremely skewed, or
  - n is much larger than 30 (approx. gets better as n increases).

- For distributions of a quantitative variable that are not very skewed and without large outliers, $n \geq 30$ is usually sufficient to use the CLT
- For distributions of a categorical variable, counts of at least 10 within each category is usually sufficient to use the CLT

The central limit theorem says ...

- For a sample proportion

$$\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

where p is the population proportion, and n is the sample size

- For a sample mean

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

where μ is the population mean, σ is the population standard deviation and n is the sample size

The central limit theorem says ...

- **For a sample proportion**

$$\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

where p is the population proportion, and n is the sample size

- For a sample mean

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

where μ is the population mean, σ is the population standard deviation and n is the sample size

In March 2011, a random sample of 1000 US adults were asked

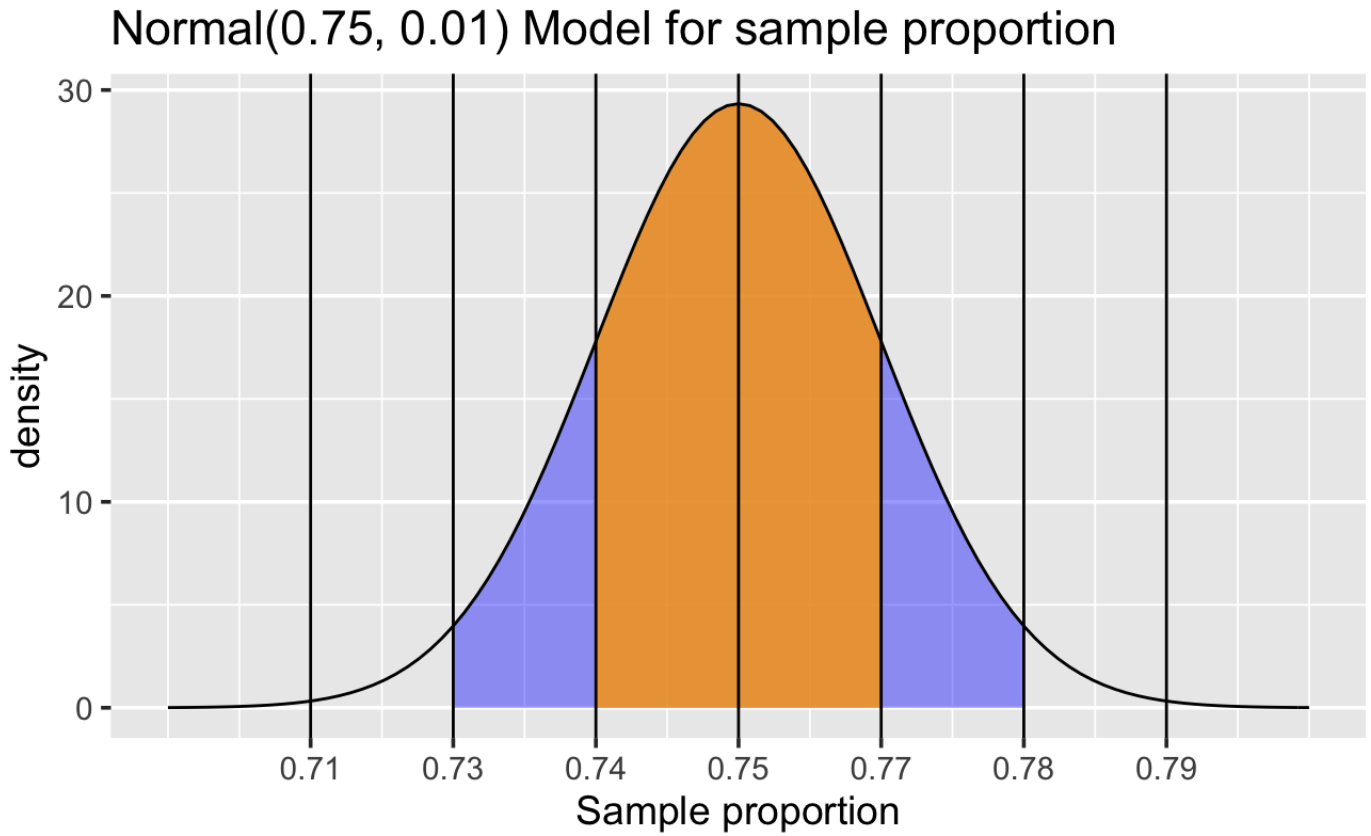
"Do you think exercise is important"

753 adults responded they think exercise is important

- Use $753/1000 = 0.753$ to approximate p
- Sample size $n = 1000$

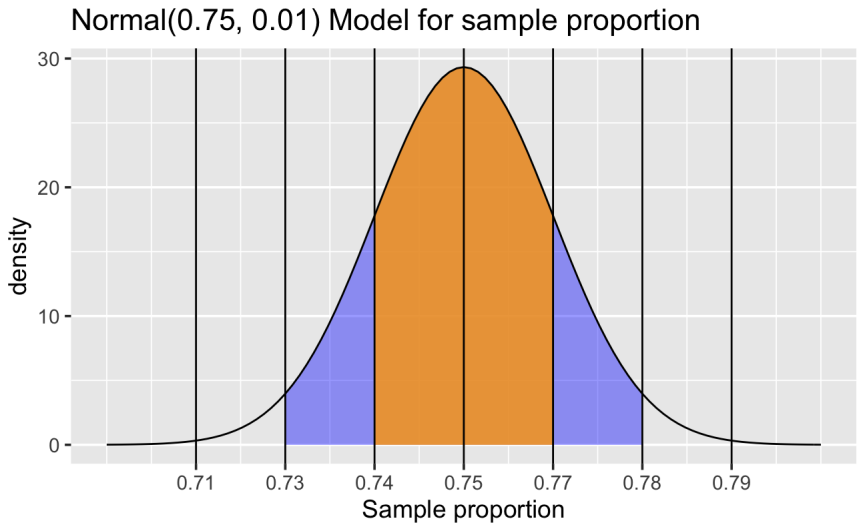
Plugging in $\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, we get

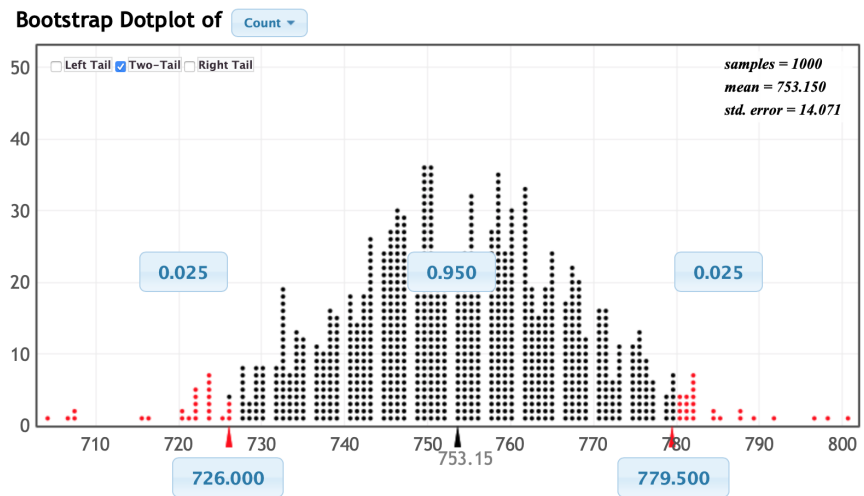
$$\hat{p} \sim \mathcal{N}\left(0.75, 0.01\right)$$



We compute the area under the curve within two standard deviation away from the mean.

The area under the curve of a normal distribution is equal to the proportion of the distribution falling within that range





CI : point estimate $\pm$ margin of error

If the parameter of interest is the population proportion, and the point estimate is the sample proportion,

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

z^* is called the critical value, which comes from a standard normal model $\mathcal{N}(0, 1)$.

Common confidence levels:

- 95%: $z^* = 1.96$ (but 2 is close enough)
- 90%: $z^* = 1.645$
- 99%: $z^* = 2.576$

Outline

The slasher horror film has been deplored based on claims that it depicts eroticized violence against predominately female characters as punishment for sexual activities.

Is **survival** for female characters in slasher films associated with **sexual activity**?

Sexual activity	Outcome of physical aggression		n
	Survival	Death	
Present	13.3% (n=11)	86.7% (n=72)	83
Absent	28.1% (n=39)	71.9% (n=100)	139

Let

- p_{present} denote the survival rate when there is sexual activity present in the movie
- p_{absent} denote the survival rate when there is no sexual activity present in the movie

Is $p_{\text{present}} = p_{\text{absent}}$?
Compute a 95% confidence interval for $p_{\text{present}} - p_{\text{absent}}$.

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

where p_1, p_2 denote the two population proportion and n_1, n_2 denote the two sample size

We use the observed sample statistic to approximate p_1, p_2.

To compute the confidence interval

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

We found,

- $\hat{p_1} = 0.133$
- $\hat{p_2} = 0.281$
- $n_1 = 83$
- $n_2 = 139$
- $z^* = 2$

The confidence interval is therefore

$$0.148 \pm 0.1 = (0.048, 0, 248)$$