Statistical Foundations: Estimating Population Values

19 February 2020

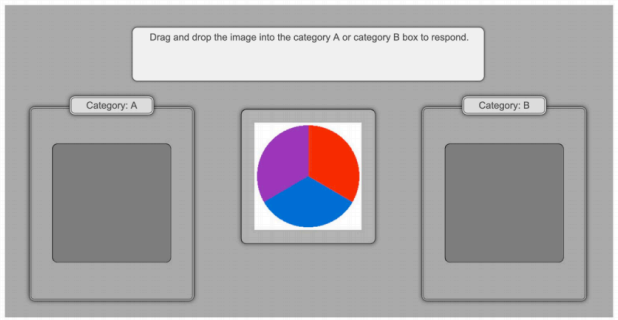
Modern Research Methods

Early course feedback focus groups

- Thanks!
- A couple of your suggestions:
  - More active stuff in lab
  - More R resources: At beginning (helpful for next time!) and more resources for more advanced R

Replicating Zettersten and Lupyan (2020)

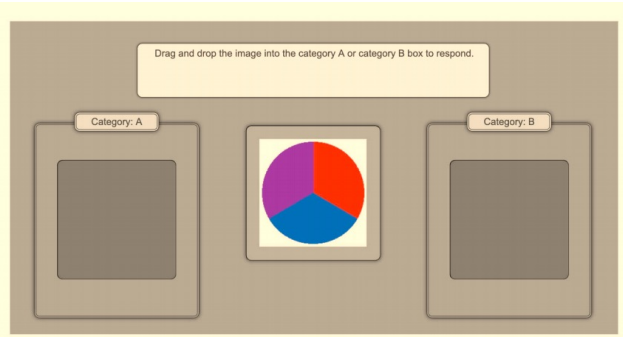
Original



predicting participants' trial-by-trial accuracy on training trials from condition, including a by-subject random intercept.[3] We used the lme4 package version 1.1-21 in R (version 3.6.1) to fit all models (D. Bates & Maechler, 2009; R Development Core Team, 2019). Participants in the High Nameability condition (M = 84.0%, 95% CI = [78.6%, 89.4%]) were more accurate than participants in the Low Nameability Condition (M = 67.7%, 95% CI = [59.9%, 75.4%]), b = 1.02, 95% Wald

Replication

[You]



High Nameability Condition = 75%
Low Nameability Condition = 69%

Should you expect to replicate the original finding? Did you replicate it? What would convince you?

Scientific goal as making inferences about the population based on the sample

Population

N = a lot

μ σ

(mean) (standard deviation)

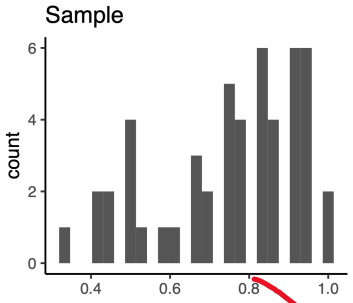
We want to estimate (best guess) the mean of the distribution in this box. $\hat{\mu}$ $\hat{\sigma}$

$\bar{X}$ = .84

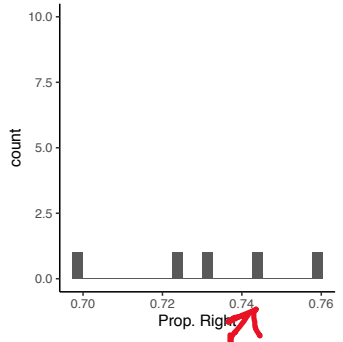
Sample

N = 50

$\bar{X}$ s

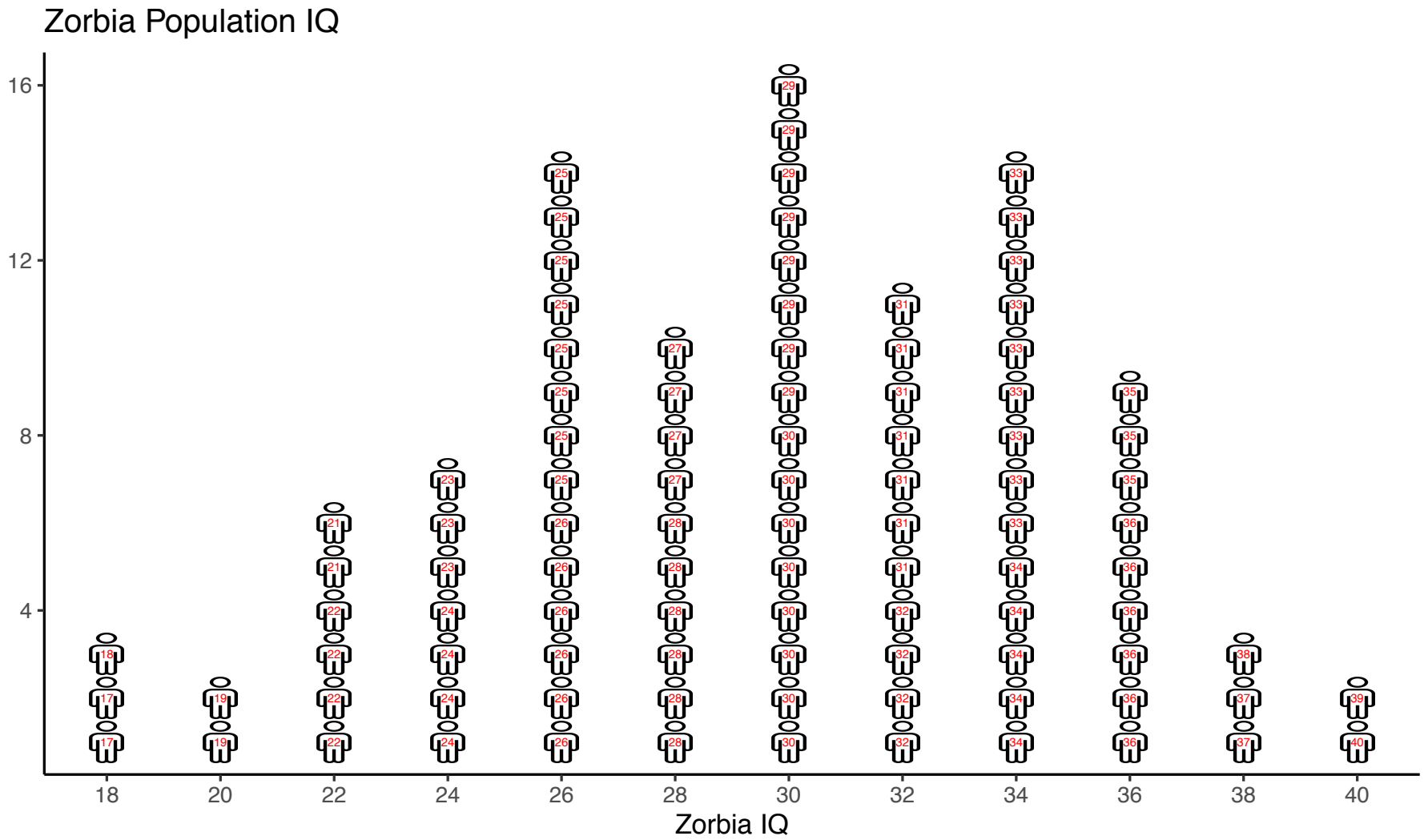


Sampling Distribution



Can describe central tendency and dispersion of distribution with **mean** and **variance/standard deviation**.

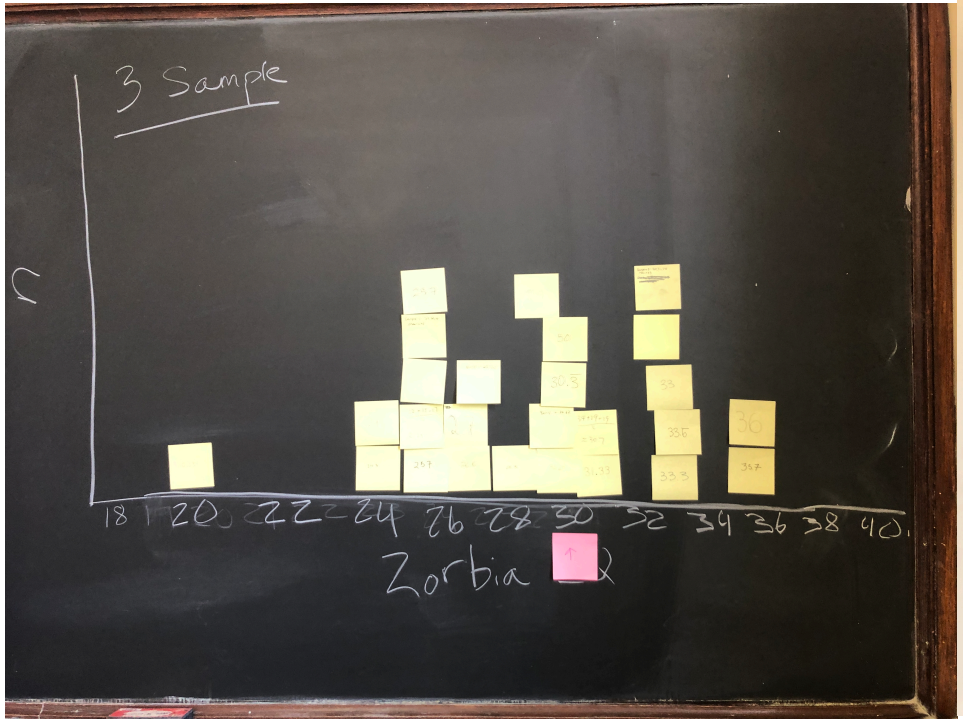
What's the mean IQ of Zorbia?

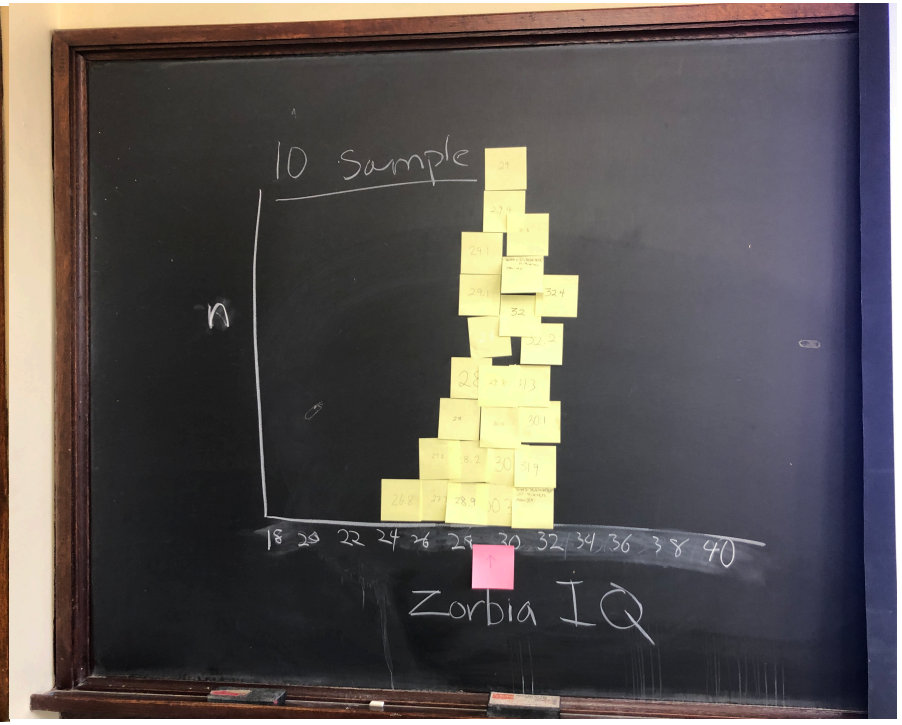


N = 97
Mean = 29

In class simulation results

Sampling Distributions:



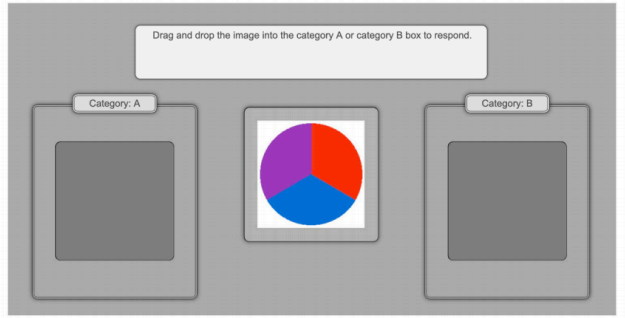


If you understand the idea behind this simulation, you understand the core of Null Hypothesis Testing!

Two samples from the same population will tend to have somewhat different means.

The bigger the sample size the narrower the sampling distribution gets

Estimating population mean



Population

N = a lot

μ σ

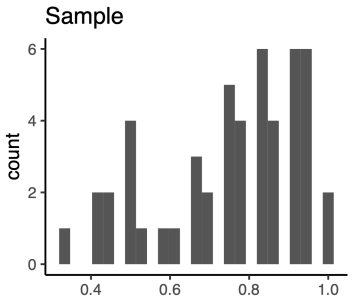
We want to estimate (best guess) the mean of the distribution in this box.

$\hat{\mu}$ $\hat{\sigma}$

Sample

N = 50

$\bar{X}$ s



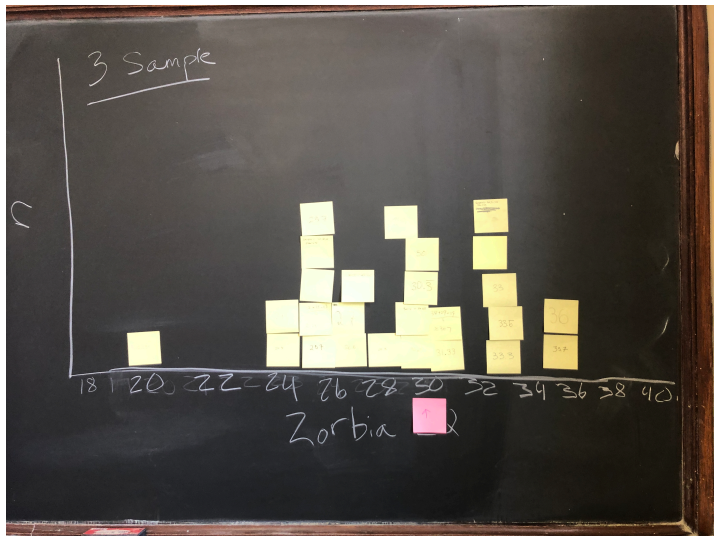
Estimate population mean is the sample mean $\bar{X}$

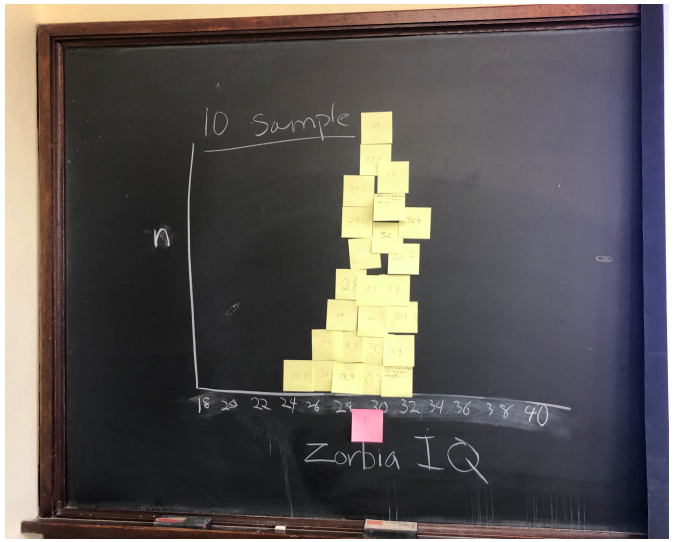
$\hat{\mu} = \bar{X} = .84$

Point estimates vs. uncertainty

- .84 is a **point estimate**
- But, we know that some estimates of the mean are going to be more accurate that others.
- It might be nice to know how much **uncertainty** there is by giving a range for our value of .84

Sampling Distributions





1. Have a special property: They are <u>normally</u> distributed.

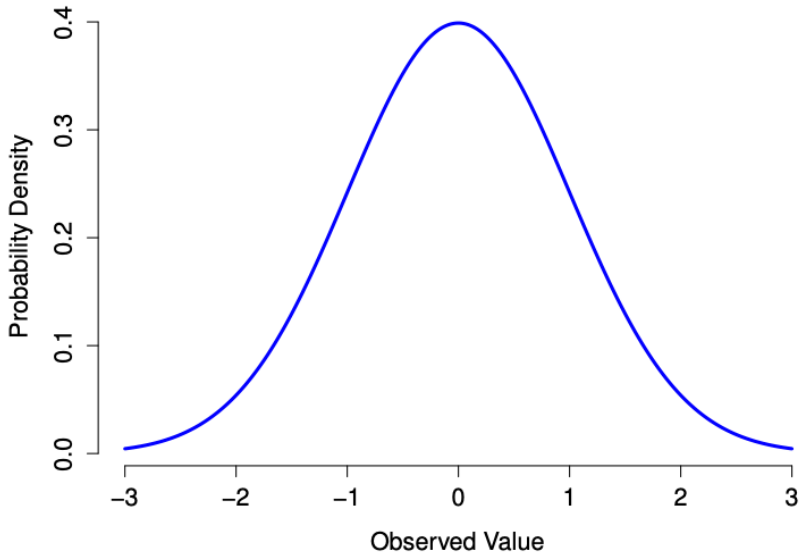
2. We can measure the dispersion with the <u>standard error of the mean</u> (SEM).

The Normal Distribution

- Bell curve/Gaussian distribution
- Two parameters: Central tendency and standard deviation

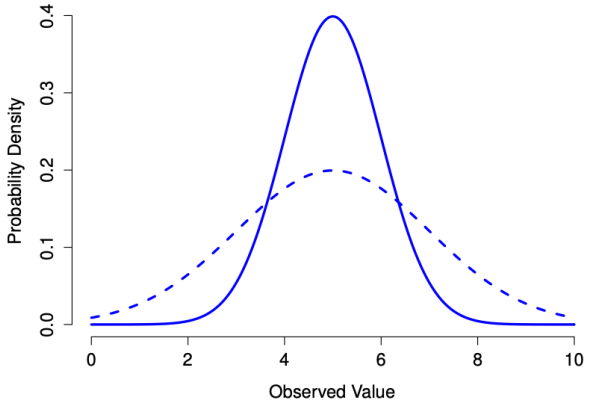
$$X \sim \text{Normal}(\mu, \sigma)$$

- x-axis = value of some variable
- y-axis = how likely are we to observe a value?

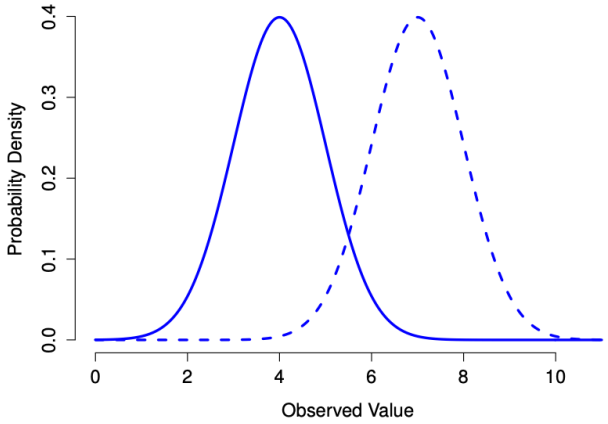


(Thanks to Danielle Navarro, LSR https://learningstatisticswithr.com/)

Probability Density and the Normal Distribution

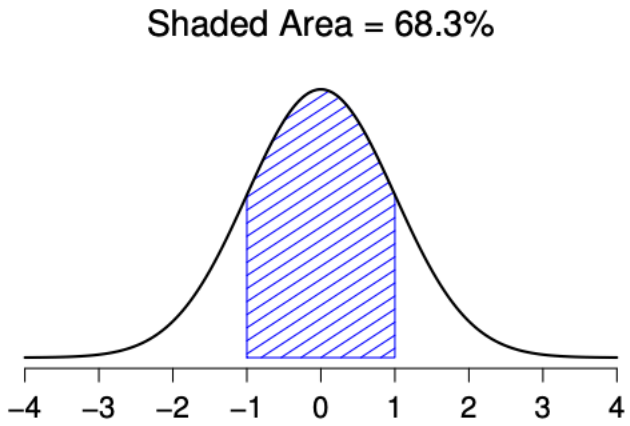


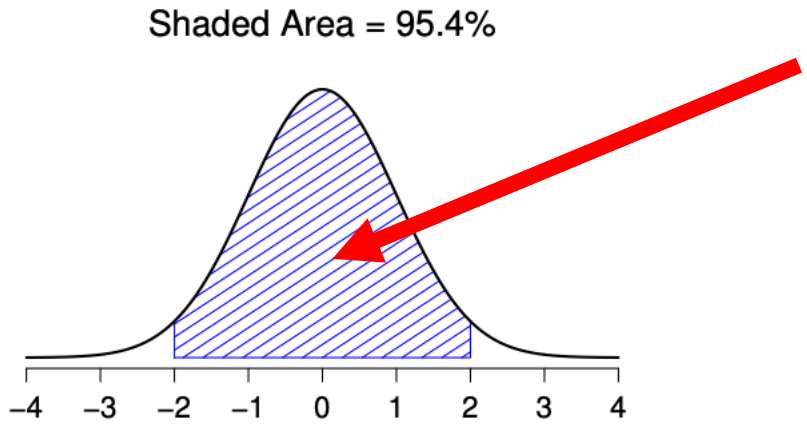
Same mean,
different standard deviations



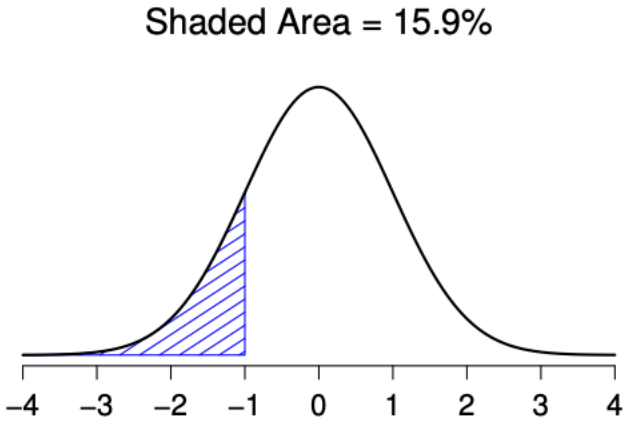
Same standard deviation,
different means

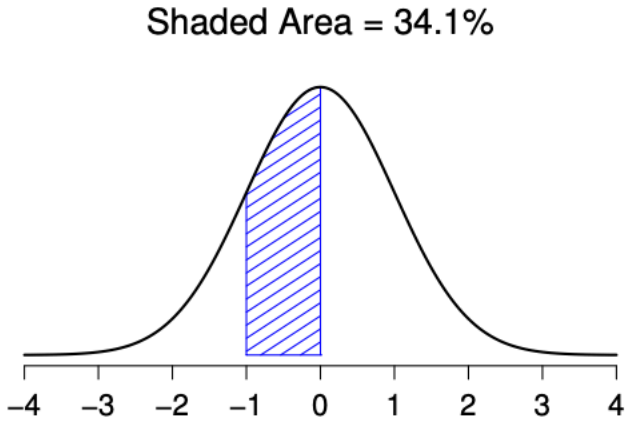
Probability and the Normal Distribution



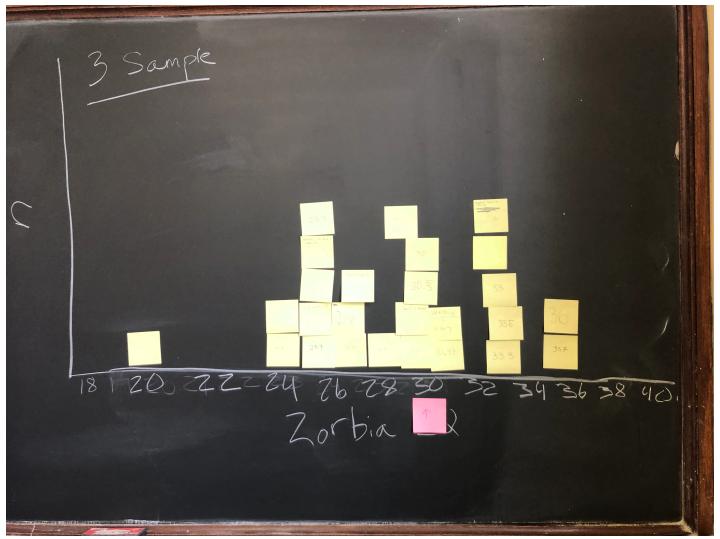


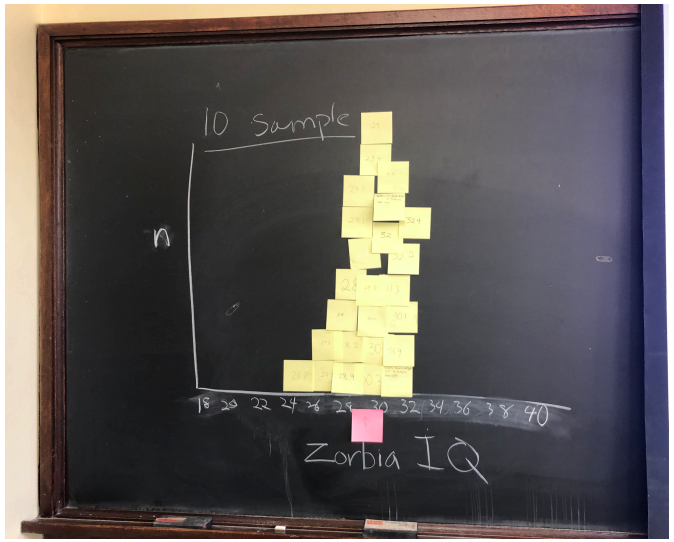
Blue shading = probability that value falls in between a range.





Sampling Distributions





1. Have a special property: They are normally distributed.

2. We can measure the dispersion with the standard error of the mean (SEM).

Calculating variance/standard deviation

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Variance is the average squared deviation from the mean of a dataset.

Standard deviation is the square root of variance.

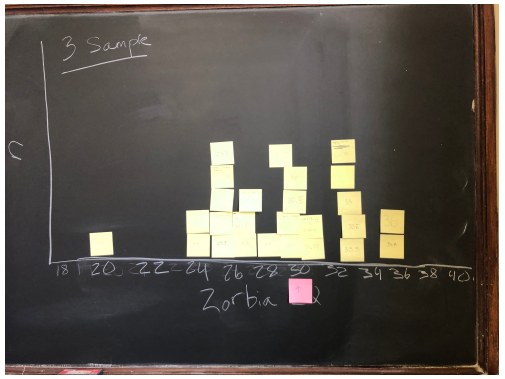
i [which game]	X_i [value]	$X_i - \bar{X}$ [deviation from mean]	$(X_i - \bar{X})^2$ [absolute deviation]
1	56	19.4	376.36
2	31	-5.6	31.36
3	56	19.4	376.36
4	8	-28.6	817.96
5	32	-4.6	21.16

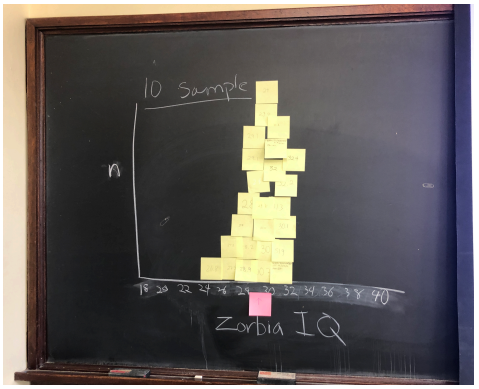
```
( 376.36 + 31.36 + 376.36 + 817.96 + 21.16 ) / 5
```

```
## [1] 324.64
```

```
var( afl.margins )
```

Standard Error of the Mean



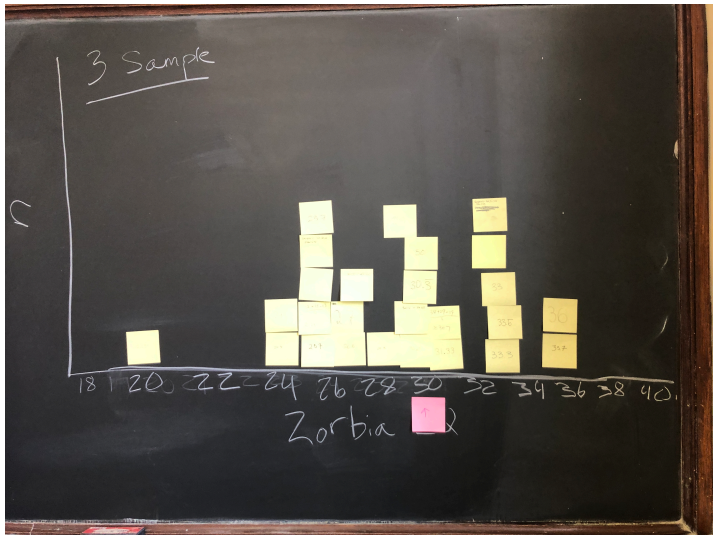


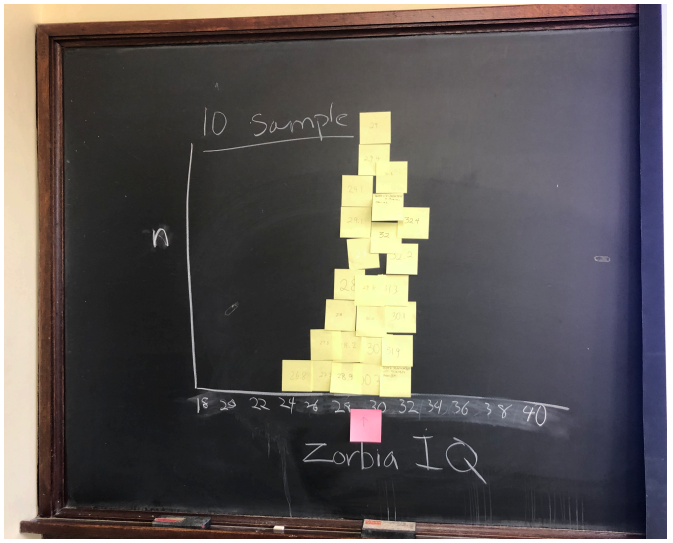
Standard deviation has a special name for the sampling distribution "**Standard Error**" (SE)

"**Standard Error of the Mean**" (SEM)

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{N}}$$

Sampling Distributions





1. Have a special property: They are normally distributed.

2. We can measure the dispersion with the standard error of the mean (SEM).

Great – we have all of this machinery…How do we use it to estimate uncertainty around the mean?

Population

N = a lot

μ σ

(mean) (standard deviation)

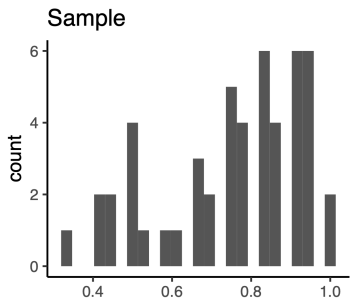
We want to estimate (best guess) the mean of the distribution in this box.

$\hat{\mu}$ $\hat{\sigma}$

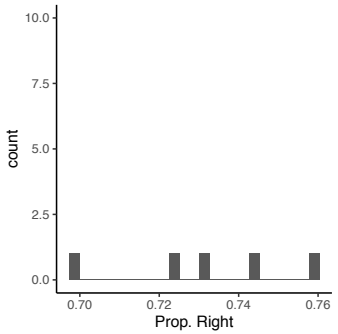
Sample

N = 50

$\bar{X}$ s

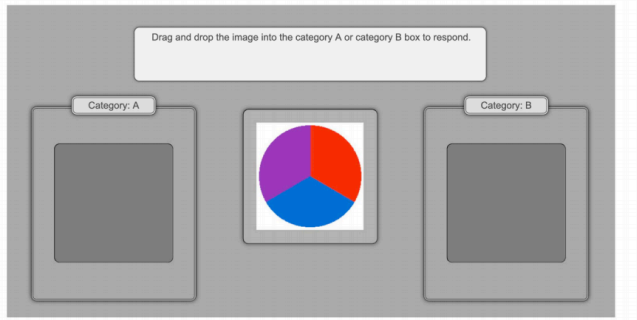


Sampling Distribution



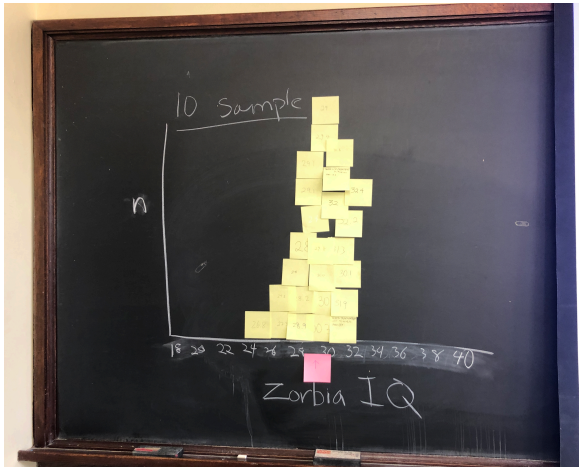
1. Has a special property: Is normally distributed.
2. We can measure the dispersion with the standard error of the mean (SEM).

What does information does the uncertainty around the mean depend on?



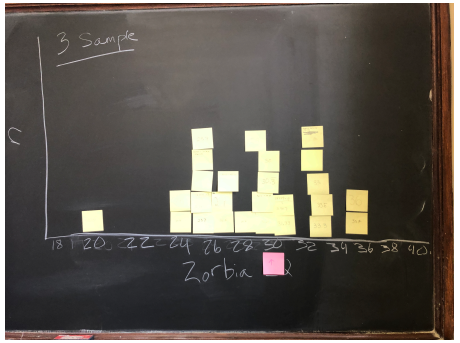
$$\hat{\mu} = \bar{X} = .84$$

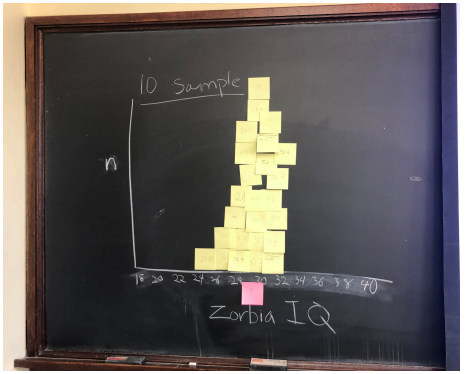
Talk to the person next to you to try to answer this question.



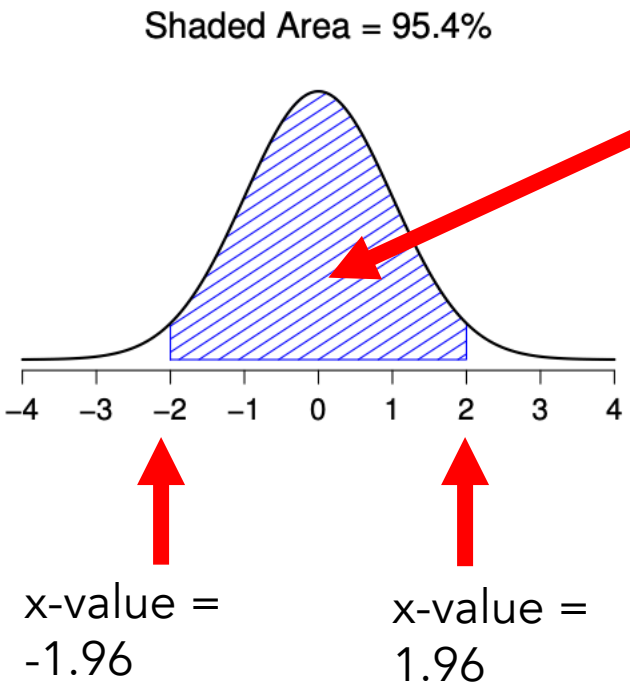
Answer: Take into account sample size.

Bigger sample (*N*) -> more certainty





We know the sampling distribution is normally distributed.



Blue shading = probability that value falls in between a range.

x-value = -1.96

x-value = 1.96

$$\bar{X} - (1.96 \times \text{SEM}) \leqslant \mu \leqslant \bar{X} + (1.96 \times \text{SEM})$$

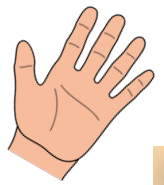
$$\text{CI}_{95} = \bar{X} \pm \left(1.96 \times \frac{\sigma}{\sqrt{N}}\right)$$

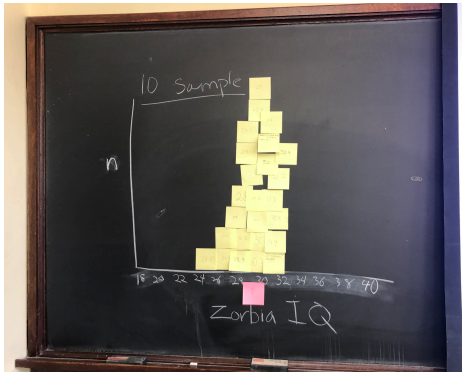
Interpreting confidence intervals

- Interpretation little tricky
- "if we replicated the experiment over and over again and computed a 95% CI for each replication, then 95% of those *intervals* would contain the true mean"
- "Our CI is a range of plausible values for. Values outside the CI are relatively implausible. " (Cumming & Finch, 2005)
- Not about your beliefs about the population

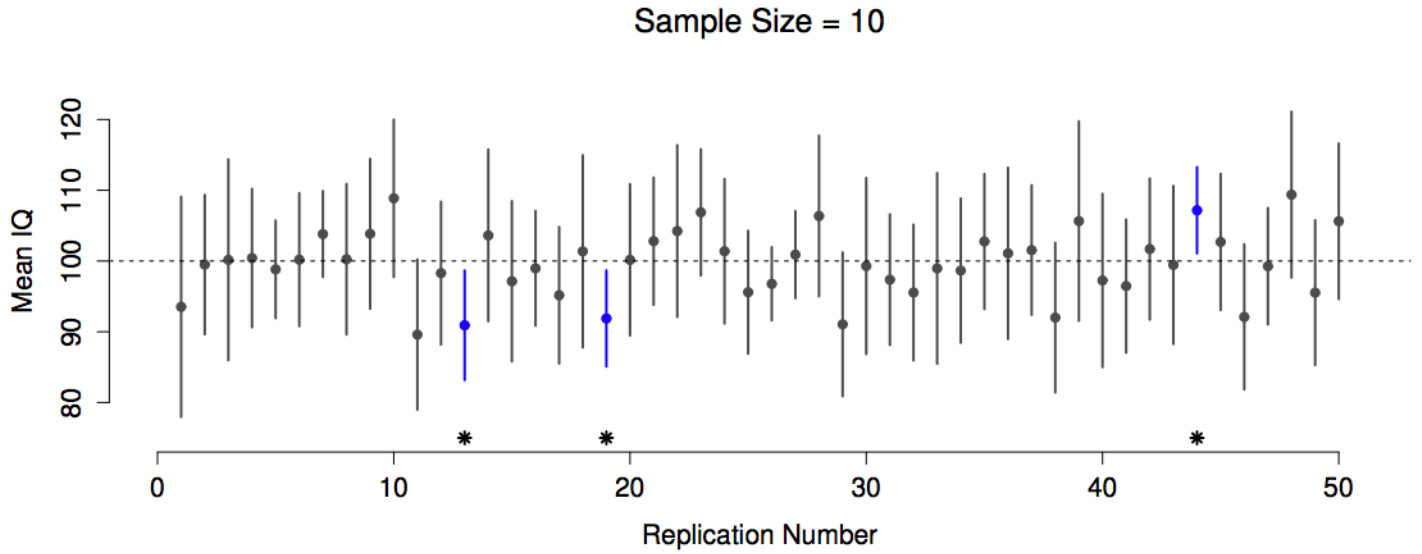
- In an alternative framework – Bayesian Statistics – there's a related idea called "credible intervals". Credible intervals concern beliefs.

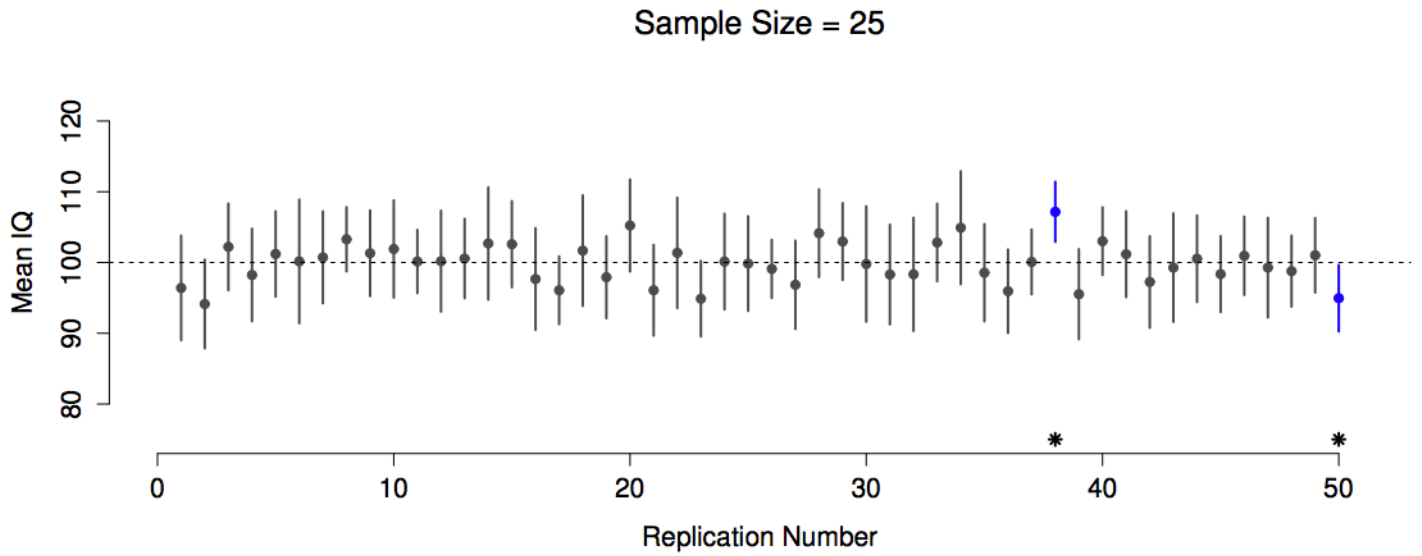
Every time you do an experiment…





.84





SEM as another way to represent uncertainty

$$\bar{X} - (1.96 \times \mathrm{SEM}) \leqslant \mu \leqslant \bar{X} + (1.96 \times \mathrm{SEM})$$

- But, much smaller.
- Not as good for inference because not related to probability
- Make sure you know which one you're looking at in a plot
- Rule of thumb: For N at least 10, SE bars can be doubled in length to get, approximately, the 95% CI *(Cumming & Finch, 2005).*

How to plot confidence intervals in R

`geom_pointrange()`

[see CI plotting slide slides]

Explore this app: https://bit.ly/325h4tW

Chapter 4: Frequentist Inference

Point Estimation

One of the main goals of statistics is to estimate unknown parameters. To approximate these parameters, we choose an estimator, which is simply any function of randomly sampled observations.

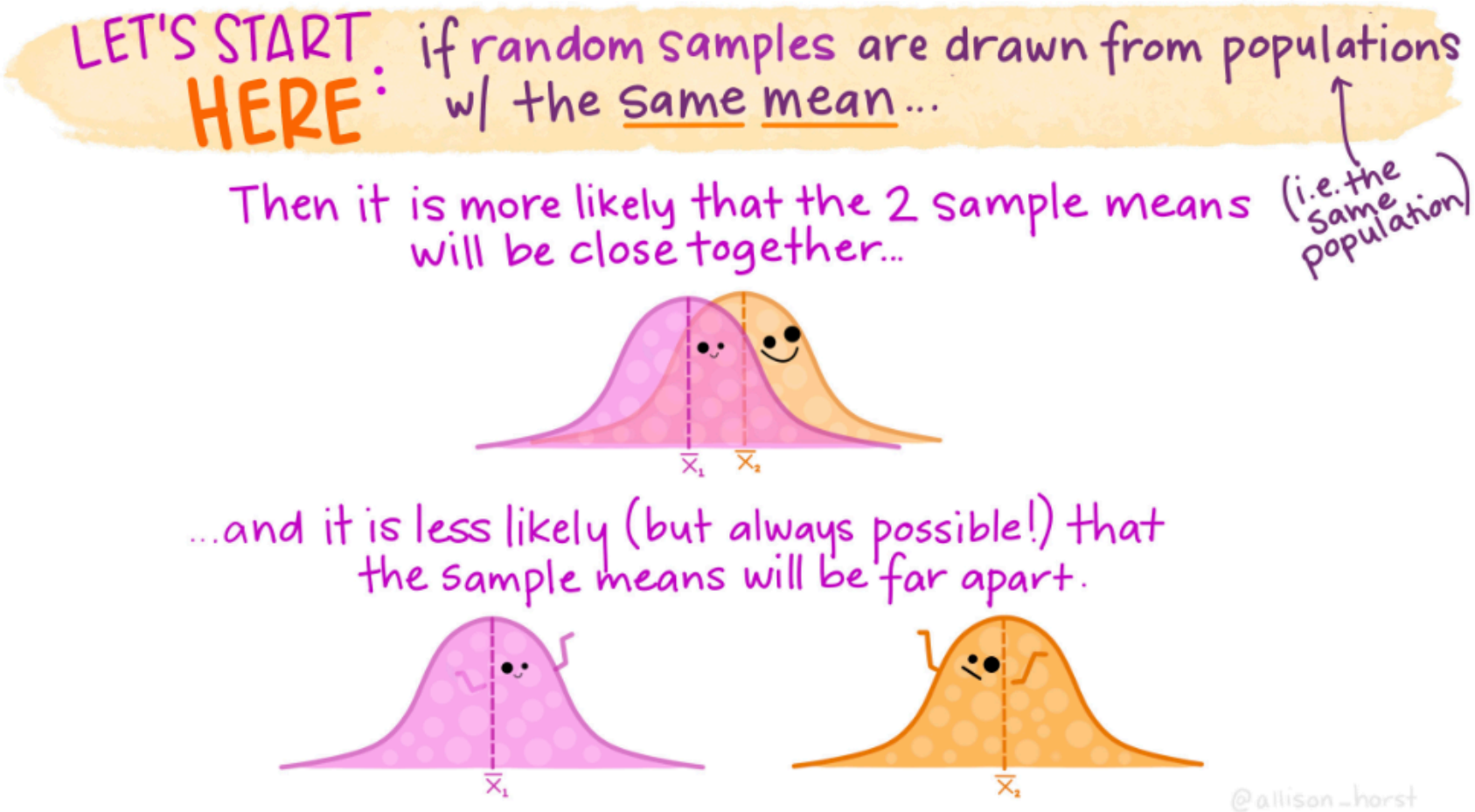
To illustrate this idea, we will estimate the value of π by uniformly dropping samples on a square containing an inscribed circle. Notice that the value of π can be expressed as a ratio of areas.

$$\begin{aligned} S_{circle} &= \pi r^2 \\ S_{square} &= 4r^2 \end{aligned} \implies \pi = 4\frac{S_{circle}}{S_{square}}$$

We can estimate this ratio with our samples. Let m be the number of samples within our circle and n the total number of samples dropped. We define our estimator $\hat{\pi}$ as:

$$\hat{\pi} = 4\frac{m}{n}$$

Next Time: Two sample *t*-test and effect sizes



Artwork by @allison_horst

Acknowledgements

- Slides 10-22 have content adapted from Danielle Navarro, Learning Statistics with R (https://learningstatisticswithr.com/)