Standard errors and confidence intervals

Modern Research Methods
Guest Lecture: Roderick Seow

6th October 2021

Population

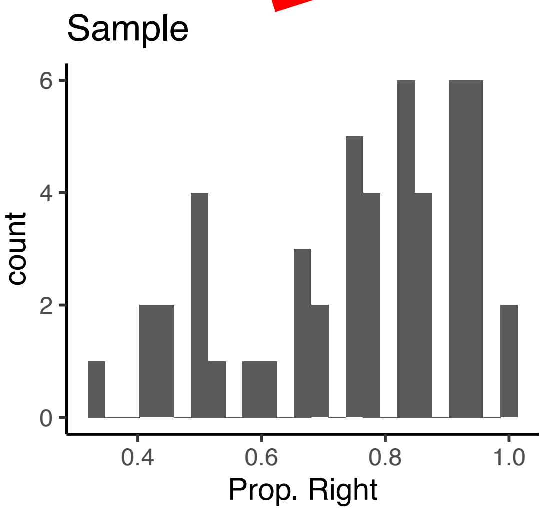
N = a lot

1. Collect a sample of size n from the population

2. What is the mean of the population?

Sample

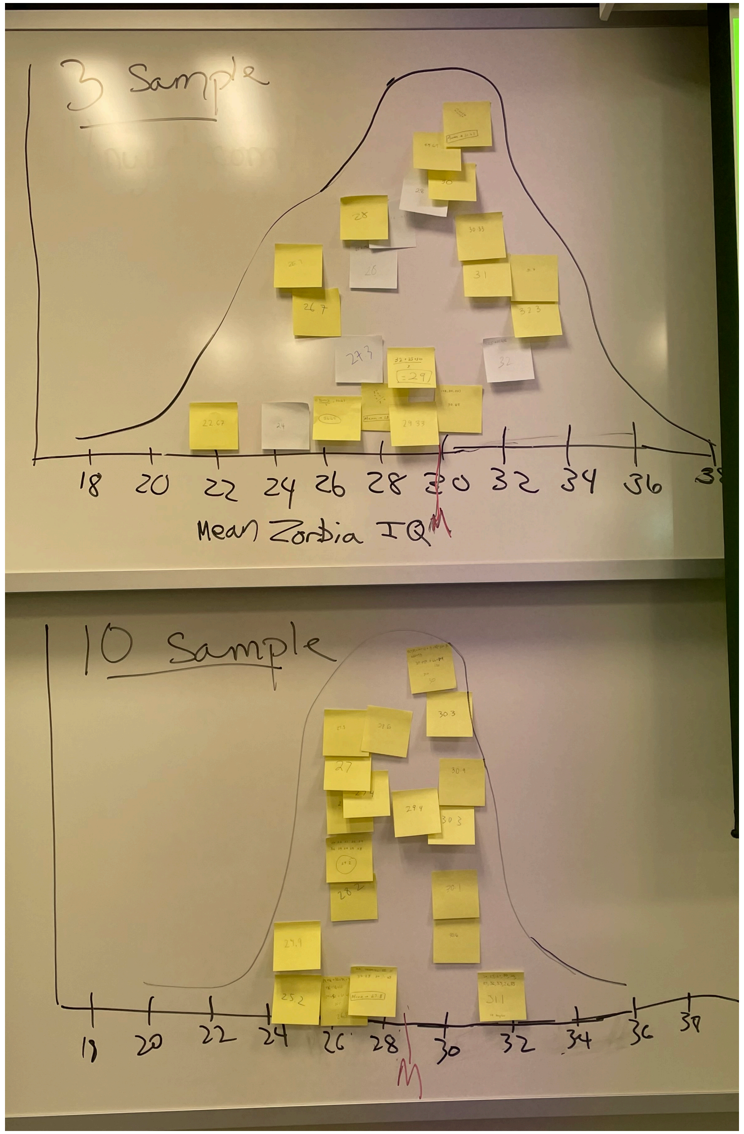
N = 50

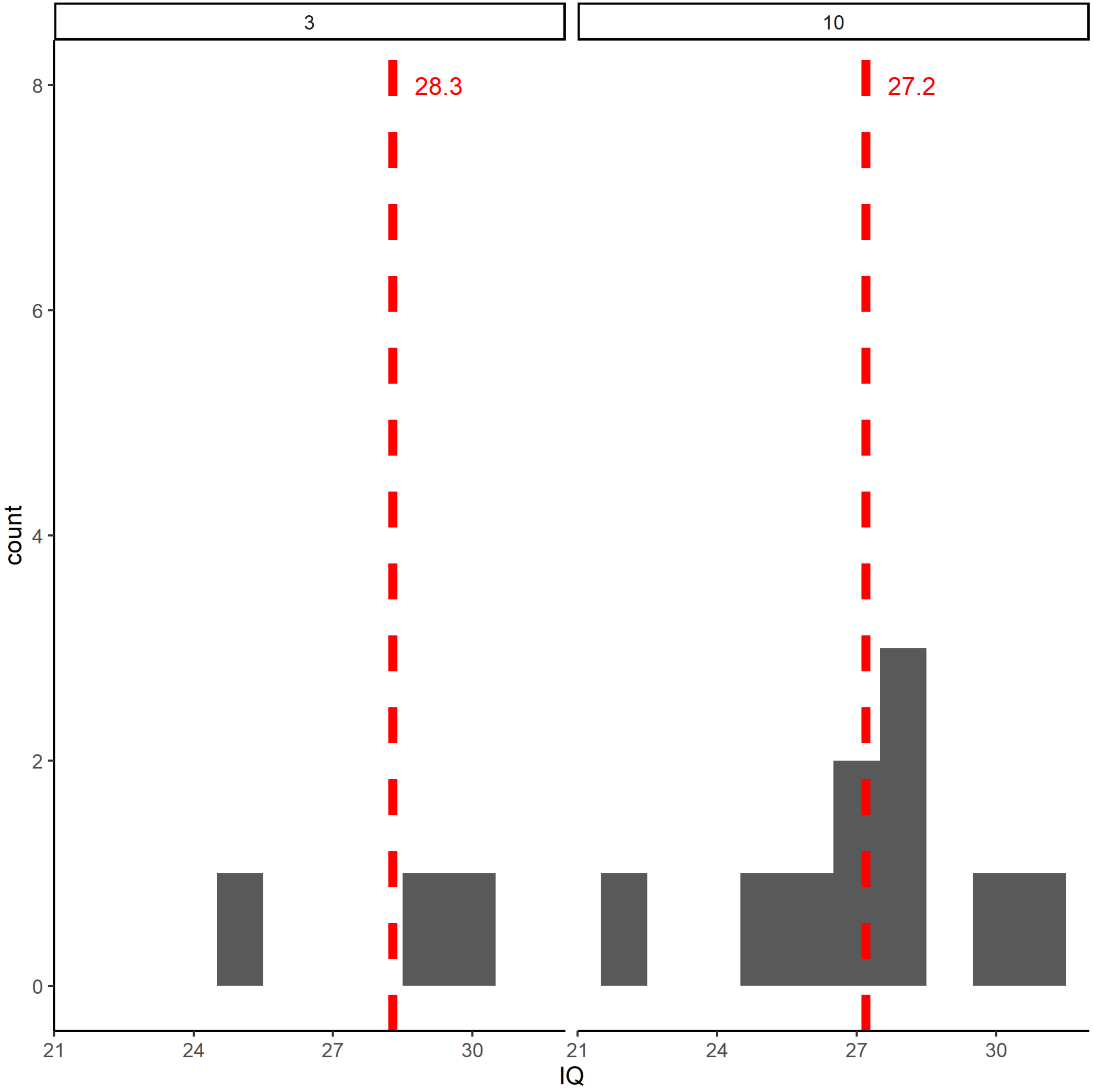


Estimating the mean Zobria IQ

- Repeatedly sampled from the same population
- Means of samples make up the sampling distribution

- But usually, we only take one sample
  - For a single sample, best point estimate of population mean is the sample mean




3
10
28.3
27.2
count
0
2
4
6
8
21
24
27
30
21
24
27
30
IQ

Why care about the certainty of estimates?

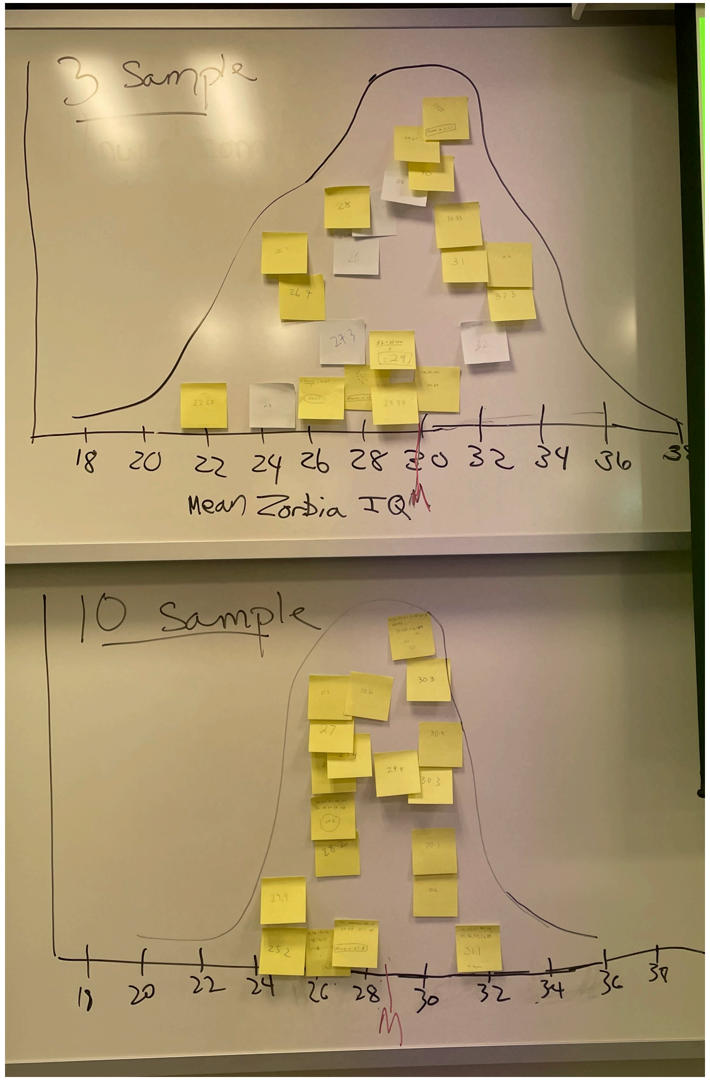
- Deciding between two flights:
  - Flight A: Departs at 8pm, punctual
  - Flight B: Departs at 7pm, has been known to be delayed for up to 3 hours
- Decision making relies on both the value and certainty of the estimate
- Additional description of data

Learning Objectives

- How to quantify certainty/uncertainty about estimate? (Confidence intervals!)
- What do confidence intervals depend on?
- How to interpret confidence intervals?
- How to plot confidence intervals in R using ggplot?

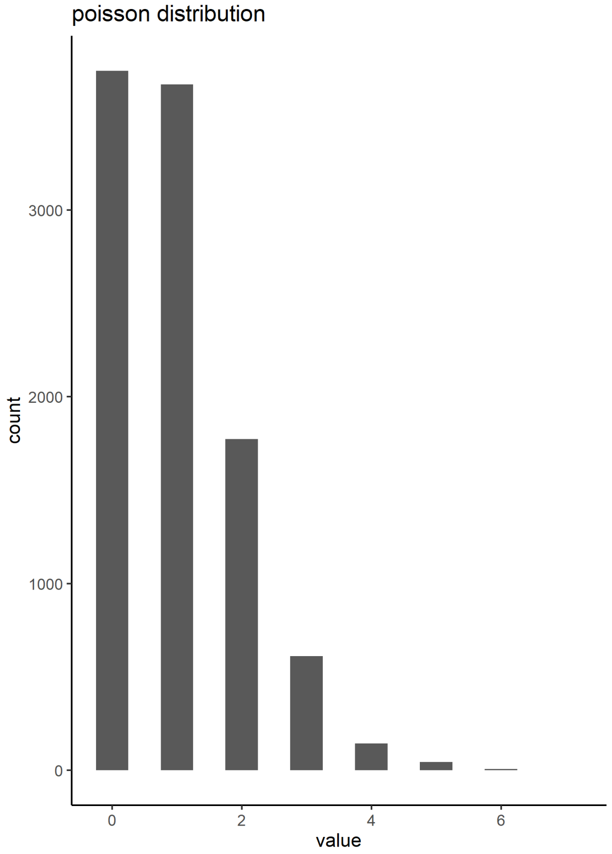
Sampling Distribution

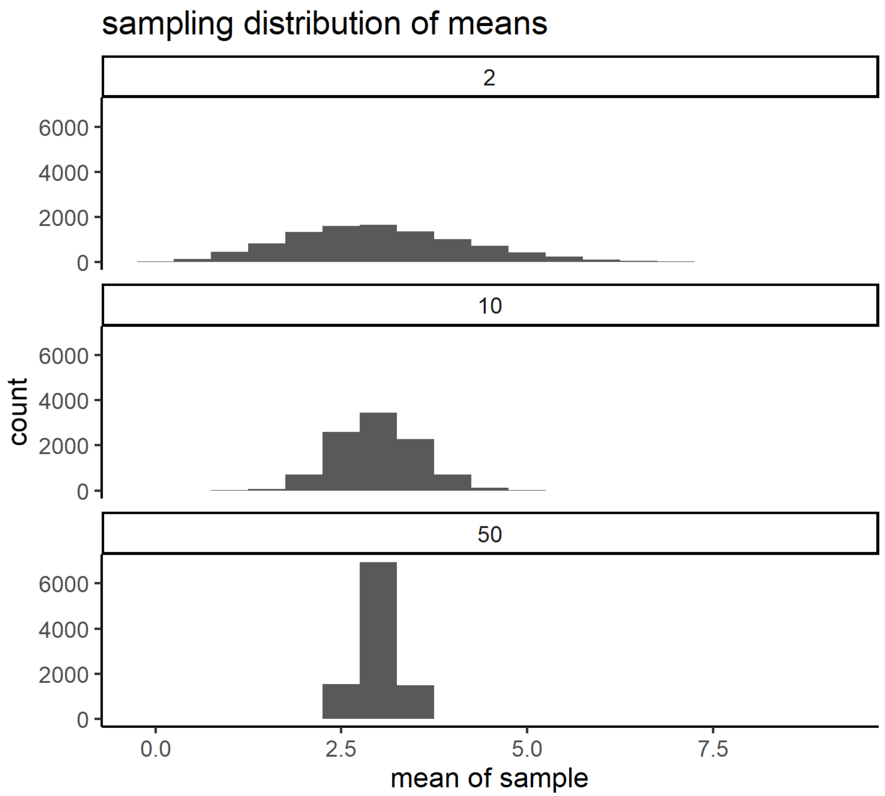
- Theoretical distribution of sample means
- Central Limit Theorem
  - Approaches normal distribution with increasing sample sizes



Sampling Distribution: Variance

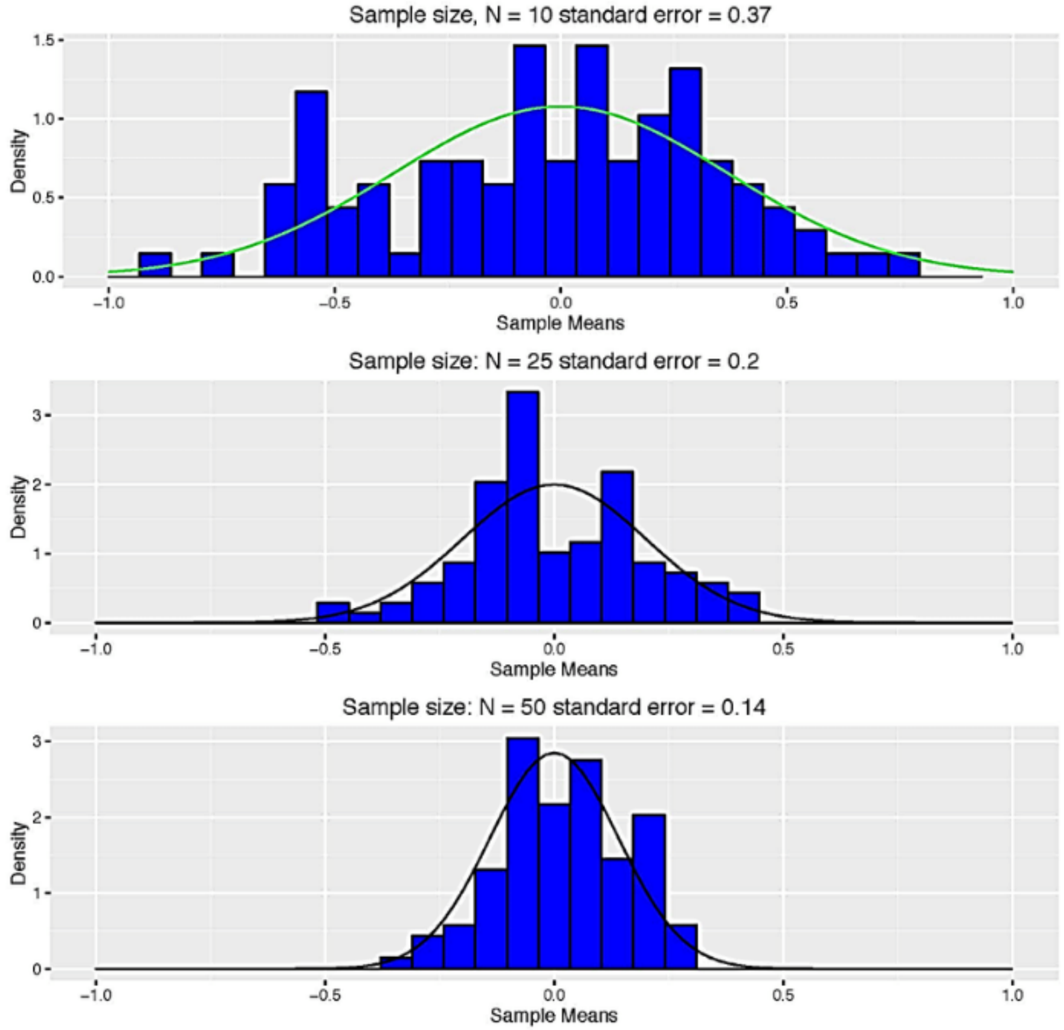
- Depends on the population distribution
  - Highly skewed population distributions lead to skewed sampling distributions
- Depends on sample size





Standard error of the mean (SEM)

- How certain are we that our estimate represents the mean of the population?
- SEM = standard deviation of the sampling distribution



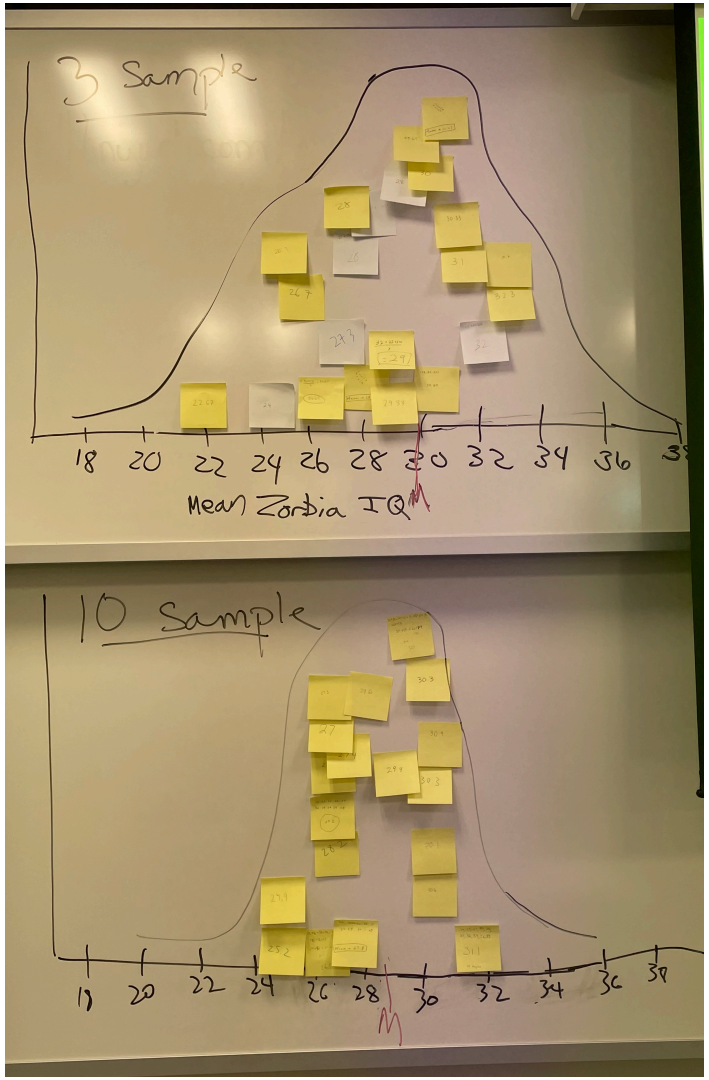
Standard error of the mean

- Estimate of population's standard deviation (σ) divided by square root of sample size (n)

$$SEM = \frac{\sigma}{\sqrt{n}}$$

- What does a smaller SEM tell us about our estimate?

- Smaller SEM = estimate is likely to be closer to population mean



From a point estimate to an interval

- Mean and SEM as point estimates
- What if we could create an interval that we are “reasonably confident” contains the true population mean?

- Average score from sample: 7/10
- What is a range of scores that definitely includes the population average?

Confidence Intervals

- A range of values that captures the population mean with x% confidence, typically set at 95%

- Imagine if we could take multiple samples from the population
- For each sample, we can construct a 95% confidence interval
- Then, 95% of the constructed intervals will include the true population mean

Confidence Intervals - Activity

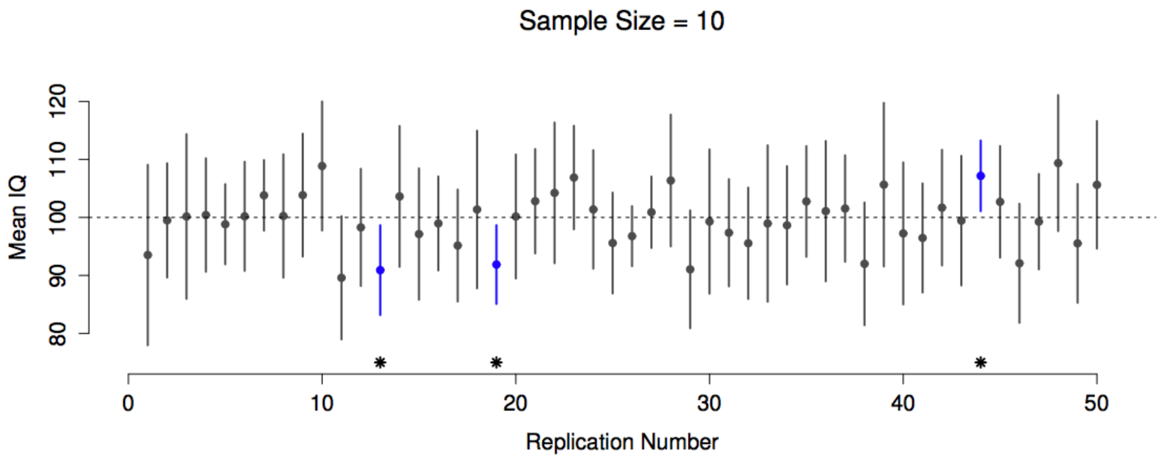
- https://bit.ly/3o9qHoF

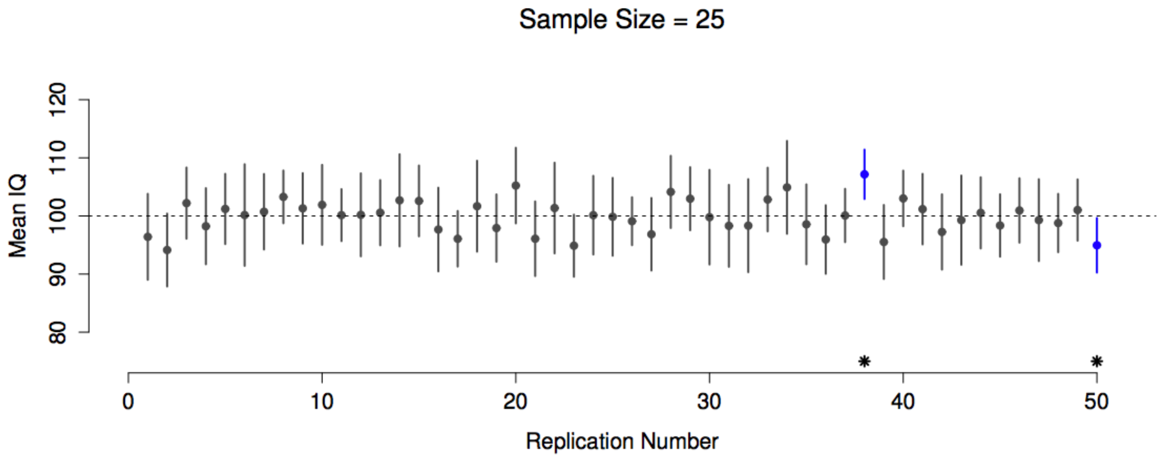
- Start with a normal distribution to sample from

1. How do the lengths of the confidence intervals change with sample size?
2. With confidence level?
3. Try again with the exponential distribution
   - Do your observations hold?

Confidence Intervals - Activity

- Length of confidence interval decreases with increasing sample size
  - Sample means are closer to population mean
- Length of CI increases with increasing confidence level
  - Larger intervals capture more possible parameter values
- Principles apply to all types of population distributions (thanks to CLT)





Source: Danielle Navarro; Estimating a Confidence Interval. (2020, August 11).; Retrieved October 7, 2021, from https://stats.libretexts.org/@go/page/4004

Computing the CI

- Let's assume that the sampling distribution is normal
  - Is this always a valid assumption? When is this assumption inappropriate?

$$CI_x = \bar{X} \pm \left(\boxed{z_x} * \boxed{\frac{\sigma}{\sqrt{n}}}\right)$$

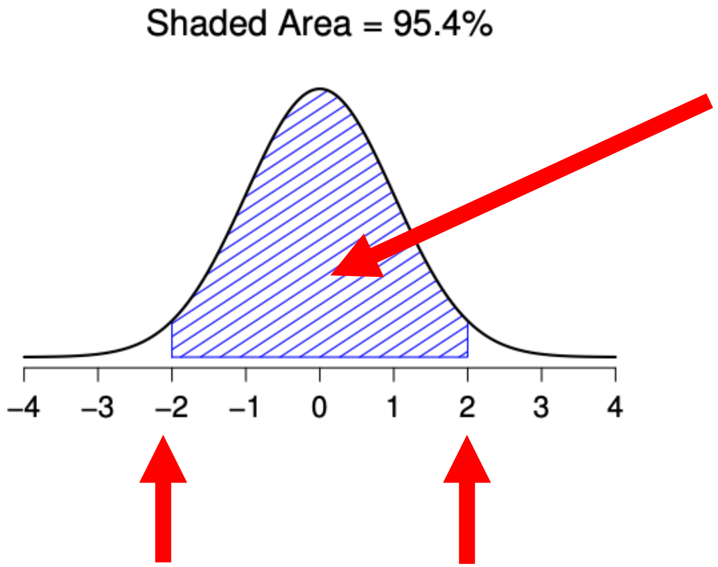
SEM

Two-sided z-critical value

Computing the CI

- Let's assume that the sampling distribution is normal
  - Is this always a valid assumption? When is this assumption inappropriate?

$$CI_{95} = \bar{X} \pm (1.96 * \frac{\sigma}{\sqrt{n}})$$



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

x-value = -1.96

x-value = 1.96

Computing the CI

- What if sampling distribution cannot be assumed to be normal?
  - Small sample size and unknown population variance

Computing the CI

- What if sampling distribution cannot be assumed to be normal?
  - Small sample size and unknown population variance

- Use the student's t-distribution instead!

$$CI_x = \bar{X} \pm (\boxed{t_x} * \frac{\sigma}{\sqrt{n}})$$

Two-sided t-critical value (with n-1 degrees of freedom)

Computing the CI

- What if we want a CI for the difference of means?
  - Same procedure! But need to compute variance of the difference sampling distribution

2 independent samples

$$CI_x = (\bar{X}_1 - \bar{X}_2) \pm (\boxed{t_x} * \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

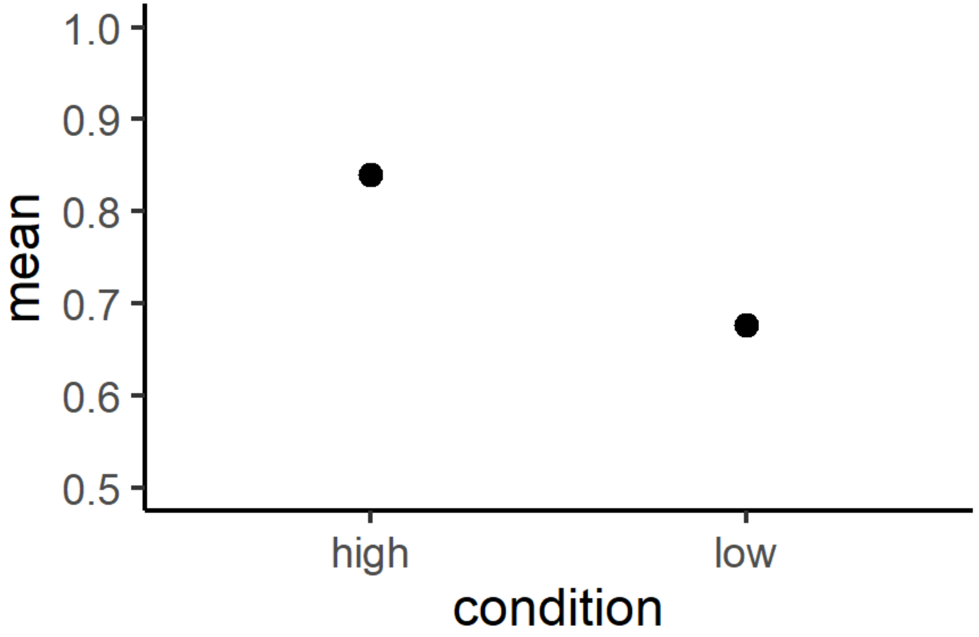
Two-sided t-critical value
(with the smaller of n_1-1 and n_2-1
as the degrees of freedom)

What exactly does the CI mean?

- “Our CI is a range of plausible values for the population mean. Values outside the CI are relatively implausible.” (Cumming & Finch, 2005)

- Is about how much precision our sampling process affords us
- Not about our beliefs about the population
  - Check out credible intervals in Bayesian statistics

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 CI=[0.47, 1.56], z=3.65, p<.001 (see Fig. 4A).



- Less overlap = Smaller p-value
- Presents a more graded picture than p<0.05 or p>0.05
  - Not just whether means are statistically different
  - "Consider interpretations of lower and upper limits and compare these with interpretations of the mean" (Cumming & Finch, 2005)

Plotting confidence intervals with ggplot

We're going to calculate a confidence intervals for the means on accuracy reported in Zettersten and Lupyan (2020), Experiment 1A

Let's start by loading the data.

```
DATA_PATH <- "https://osf.io/a4dzb/download"

zl_data <- read_csv(DATA_PATH)

zl_clean <- zl_data %>%
     clean_names() %>%
     select(experiment, subject, age, condition, block_num, is_right)

zl_exp1a <- zl_clean %>%
     filter(experiment == "1A")
```

experiment	subject	age	condition	block_num	is_right
1A	p150212	29	low	1	1
1A	p150212	29	low	1	1
1A	p150212	29	low	1	1
1A	p150212	29	low	1	1
1A	p150212	29	low	1	0

We start by getting by-subject by-condition means

```
ms_by_overall<- zl_exp1a %>%
  group_by(subject, condition) %>%
  summarize(prop_right = sum(is_right)/n())
```

```
## `summarise()` has grouped output by 'subject'. You can override using the `.groups` arg
```

subject	condition	prop_right
p150212	low	0.8750000
p157080	low	0.7083333
p191463	low	0.9583333
p20905	high	0.9583333
p213384	high	1.0000000
p25634	low	0.6666667
p269913	low	0.4583333
p270949	low	0.9166667
p299672	high	0.8333333

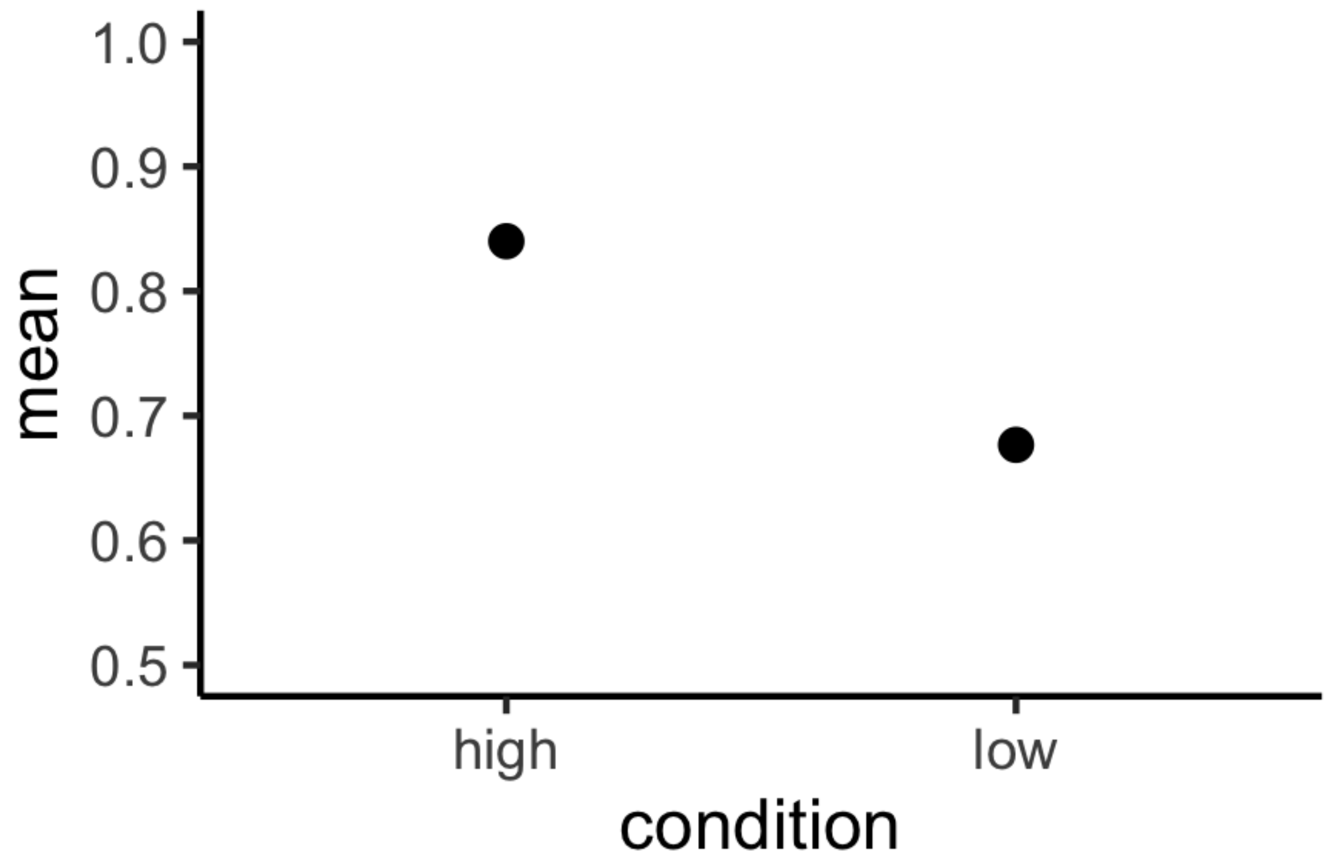
Next, let's calculate a **point estimate** for the mean in each condition.

```
means_by_condition  <- ms_by_overall %>%
  group_by(condition) %>%
  summarize(mean = mean(prop_right))
```

condition	mean
high	0.8400000
low	0.6766667

Plot the point estimates with geom_point.

```
ggplot(means_by_condition, aes(x = condition, y = mean)) +
  geom_point(size = 2) +
  ylim(.5, 1) +
  theme_classic()
```



Next let's calculate a confidence interval around our estimate.

To start we need the sample size in each condition.

```
sample_size <- ms_by_overall %>%
  group_by(condition) %>%
  summarize(n = n())
```

condition	n
high	25
low	25

Now, let's calculate the the CI

```
means_by_condition_with_ci <- ms_by_overall %>%
  group_by(condition) %>%
  summarize(mean = mean(prop_right),
            sd = sd(prop_right),
            n = n()) %>%
  mutate(ci_range_95 =  1.96 * (sd/sqrt(n)),
         ci_lower = mean - ci_range_95,
         ci_upper = mean + ci_range_95)
```

condition	mean	sd	n	ci_range_95	ci_lower	ci_upper
high	0.8400000	0.1304817	25	0.0511488	0.7888512	0.8911488
low	0.6766667	0.1876080	25	0.0735423	0.6031243	0.7502090

Plotting the confidence intervals

```
ggplot(means_by_condition_with_ci, aes(x = condition, y = mean)) +
  geom_point(size = 2) +
  geom_linerange(aes(ymin = ci_lower, ymax = ci_upper)) +
  ylim(.5, 1) +
  theme_classic()
```

There's actually a single geom that plots both points and ranges: geom_pointrange.

```
ggplot(means_by_condition_with_ci, aes(x = condition, y = mean)) +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +
  ylim(.5, 1) +
  theme_classic()
```

There's one small complexity that I've glossed over.

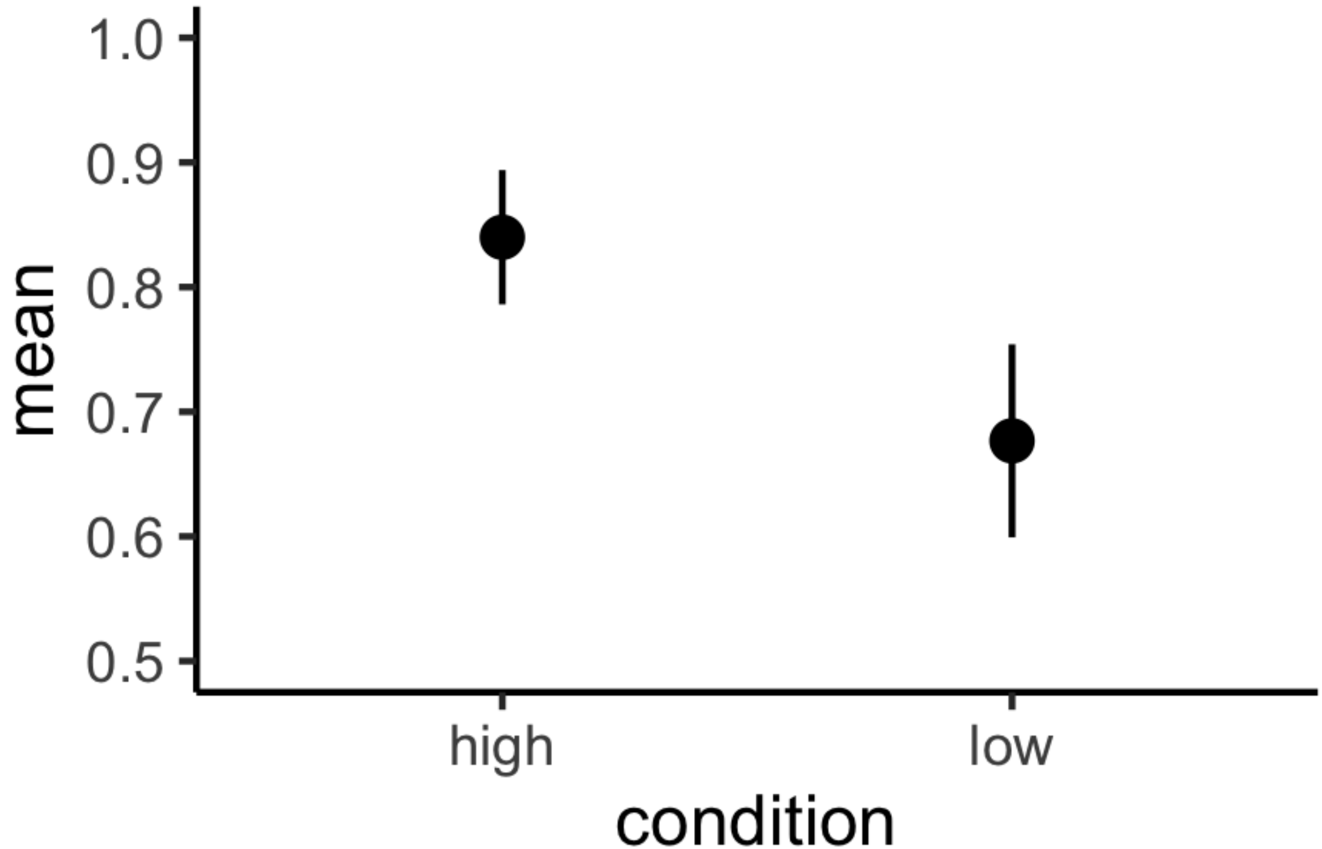
Because we don't actually know the SD for the population distribution we have to estimate from a distribution called the t-distribution.

```
means_by_condition_with_ci_t <- ms_by_overall %>%
  group_by(condition) %>%
  summarize(mean = mean(prop_right),
            sd = sd(prop_right),
            n = n()) %>%
  mutate(ci_range_95 =  qt(1 - (0.05 / 2), n - 1) * (sd/sqrt(n)),
         ci_lower = mean - ci_range_95,
         ci_upper = mean + ci_range_95)
```

condition	mean	sd	n	ci_range_95	ci_lower	ci_upper
high	0.8400000	0.1304817	25	0.0538602	0.7861398	0.8938602
low	0.6766667	0.1876080	25	0.0774408	0.5992259	0.7541074

Point estimates with ranges calculated from the t-distribution.

```
ggplot(means_by_condition_with_ci_t, aes(x = condition, y = mean)) +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +
  ylim(.5, 1) +
  theme_classic()
```



Summary

- Confidence intervals quantify uncertainty about our estimates of the population mean based on a sample
  - Captures precision of the sampling process, not about our beliefs about the value of the true population parameter
  - Encourages thinking about plausible range of values instead of a point estimate

- Larger samples, populations with smaller variances, and lower confidence levels lead to smaller intervals