PSY 503: Foundations of Statistical Methods in Psychological Science

Estimation and Confidence intervals

Suyog Chandramouli

311 PSH (Princeton University)

3rd November, 2025

Recap: Why Do We Need Inferential Statistics?

- **The Population Problem**
  - Usually impossible or impractical to measure entire population *(Population parameters: μ, σ, ρ)*
  - Must rely on samples to draw conclusions about the population *(Sample statistics: $\bar{x}$, s, r)*
    - However,
      - Different samples from the same population give different results
      - We need to account for this uncertainty
- **Example: Political Polling**

Recap: Two Main Tasks in Inference

- **Hypothesis Testing**
  - Is there an effect?
  - Are groups different?
  - Did intervention work?

- **Estimation**
  - How large is the effect?
  - What's the population parameter?
  - How precise is our estimate?

Dealing with sampling error

- **Hypothesis Testing** [binary decisions]
  - Is there an effect?
  - Are groups different?
  - Did intervention work?

- **Estimation**
  - How large is the effect?
  - What's the population parameter?
  - How precise is our estimate?

Dealing with sampling error

- **Hypothesis Testing** [binary decisions] → Set Type 1 Error (α)
  - Find p-values

  Accept/Reject the Null
  - Is there an effect?
  - Are groups different?
  - Did intervention work?

- **Estimation**
  - How large is the effect?
  - What's the population parameter?
  - How precise is our estimate?

Dealing with sampling error

- **Hypothesis Testing** [binary decisions] → Set Type 1 Error (α)
  - Find p-values

  Accept/Reject the Null
  - Is there an effect?
  - Are groups different?
  - Did intervention work?

- **Estimation** [quantifies precision: range of plausible values]
  - How large is the effect?
  - What's the population parameter?
  - How precise is our estimate?

What's the Difference?

- **Hypothesis Testing asks:**
  - Is the effect different from zero?
- **Estimation asks:**
  - What is the effect?
- **Example:** Does sleep improve memory?
  - HT: "Yes, sleep helps" ($p < .05$)
  - Estimation: "Sleep improves recall by 12-18 points (95% CI)”
- Which tells you more?
  Is there additional information in NHST?

Why Test at All?

- **Historical reasons.**
  - Fisher, Neyman-Pearson era
  - Binary decisions were needed
- **Practical reasons.**
  - Quick screening
  - Journals want p-values
  - Some fields require it
- But —
  - More statisticians now say: Report CIs first.

Information in CI report

- APA format says:
  - M = 15, 95% CI [10, 20]

- You have
  - point estimate
  - uncertainty around point estimate
  - implicit hypothesis test decision (if you need it)

What’s CI and how to find it?

What’s CI and how to find it?

- Let’s back up a little

CI: Core problem

- We have: One sample mean ($\bar{X}$)
- We want: The population mean (μ)

- The question: How confident can we about our estimate?

Frequentist thinking

- **The Frequentist Idea:**
  - What if we repeated this experiment 1000 times?
  - Each repetition → different sample mean
  - All those means → sampling distribution

- **Why this matters?**
  - Shows how much our estimate could vary
  - Let's us quantify uncertainty
  - → This is how we get confidence intervals

Different questions → different sampling distributions

- **NHST asks:**
  - Is μ = 0?
  - Build distribution at 0
  - Check if data fits
- Yes/no decision

- NHST: Test specific hypothesis

- **CI asks:**
  - What μ values work with our data?
  - Start with your **sample mean**
  - Choose confidence level (e.g. 95%)
    - SE= standard deviation / sqrt(N)
  - Find critical values corresponding to confidence levels
    - $CI=\bar{x}\pm(\text{critical value})\times SE$

- CI: Find compatible parameters

Intepreting CI

- We're *not* saying there's a 95% chance the true mean is in the CI
  - the mean is fixed; the interval varies
- Then what are we saying?

Intepreting CI

- We're *not* saying there's a 95% chance the true mean is in the CI
  - the mean is fixed; the interval varies
- Rather, *if we repeated this process infinitely many times, 95% of those ranges would capture the true mean*.

Robust misinterpretation of confidence intervals

Rink Hoekstra [1], Richard D Morey, Jeffrey N Rouder, Eric-Jan Wagenmakers

Affiliations + expand

PMID: 24420726 DOI: 10.3758/s13423-013-0572-3

Abstract

Null hypothesis significance testing (NHST) is undoubtedly the most common inferential technique used to justify claims in the social sciences. However, even staunch defenders of NHST agree that its outcomes are often misinterpreted. Confidence intervals (CIs) have frequently been proposed as a more useful alternative to NHST, and their use is strongly encouraged in the APA Manual. Nevertheless, little is known about how researchers interpret CIs. In this study, 120 researchers and 442 students-all in the field of psychology-were asked to assess the truth value of six particular statements involving different interpretations of a CI. Although all six statements were false, both researchers and students endorsed, on average, more than three statements, indicating a gross misunderstanding of CIs. Self-declared experience with statistics was not related to researchers' performance, and, even more surprisingly, researchers hardly outperformed the students, even though the students had not received any education on statistical inference whatsoever. Our findings suggest that many researchers do not know the correct interpretation of a CI. The misunderstandings surrounding p-values and CIs are particularly unfortunate because they constitute the main tools by which psychologists draw conclusions from data.

Why CIs Work

- Sampling
- Sampling Distributions
- Law or Large Numbers
- Central Limit theorem

Why CIs Work

- **CIs assume your sample represents the population. How do we ensure that?**

Why CIs Work

- **CIs assume your sample represents the population. How do we ensure that?**

Random sampling

Discuss

- **Context:** Imagine you're trying to find the average height of all the students in a large school.
  - You could take a sample of, say, 30 students and calculate their average height.
- **Question:**
  - What happens under these different sampling scenarios:
    - we just decide to sample 30 students who visit the basketball court, as it is conveniently located?
    - we email a survey link, that students can voluntarily fill?
    - randomly measured a bunch of people?
    - try to ensure representativeness by sampling 3 students from each age group?
    - try to ensure representativeness by sampling students who have different first letters of their last names?

Random sampling

- A process for generating a sample (taking things from a population)
- Random samples ensure that each value in a sample is drawn **independently** from other values
- all values in the population have a chance of being in the sample

The population problem and the sampling solution

- Unknown: The population
- Solution: Take a sample of the population
  - Samples will tend to look the population they came from, especially when sample-size (N) is large.
  - We can use the sample to **estimate** the population.

Random Sampling

- **The problem:** Can't measure entire population
- **The solution:** Take a random sample (experimentation)

- **Why it works:** Random samples tend to look like the population they came from (especially with large N)

- **The key:** Every member must have a chance of being selected

Law of large numbers

- As sample size increases, statistics of the sample tend to approach population parameters
  - i.e. the properties of the sample become more like properties of the population.
  - $\text{As } n \to \infty, \bar{X}_n \to \mu$
- Larger samples = better estimates
- This is why we prefer larger sample sizes / more data in research
  - Remember:
    - Power – with larger n, variance of test-statistic reduces, and power increases
    - Effect size – larger n lets you be sensitive to smaller effects
    - .. And so on

    →*Provides the foundation for statistical inference*

Sampling Distribution

Discuss

- **Context:** Imagine you're trying to find the average height of all the students in a large school
  - Instead, you could take a ***random sample*** of, say, 30 students and calculate their average height.

- **Question:** What does the sampling distribution of student height means constitute?

Sampling distribution

- Distribution of a statistic across many repeated samples
- Shows variation in sample statistics
- Key to understanding estimation uncertainty

- Generation Process
  - Take sample of size N from population
  - Calculate statistic (e.g., mean)
  - Repeat many times (say J times)
  - Plot distribution of statistics

  →This is the ***"empirical sampling distribution"***

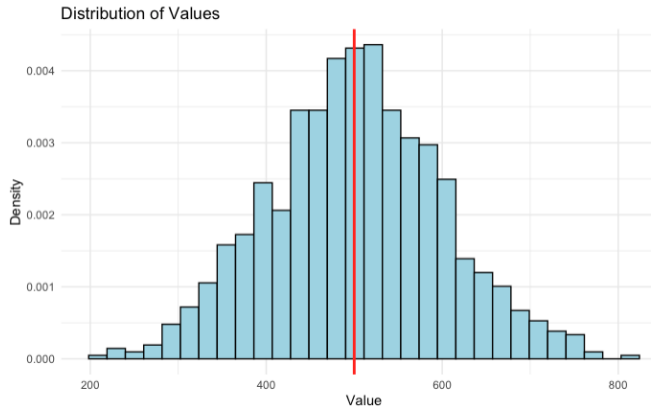
  - When J is infinity, we have the ***"theoretical sampling distribution"***

Example – sampling distribution of mean

- Imagine that instead of a single experiment, we are repeating the same experiment several times

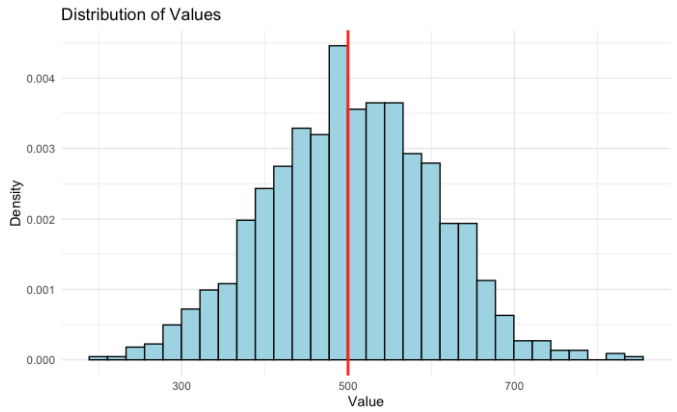
```r
library(tidyverse)

data_df <- tibble(
  y = rnorm(n = 1000, mean = 500, sd = 100)
  #A single large sample of data from a specified population.
)

print (mean(data_df$y))
```

```
[1] 497.9887
```




```r
library(tidyverse)

data_df <- tibble(
  y = rnorm(n = 1000, mean = 500, sd = 100)
  #A single large sample of data from a specified population.
)

print (mean(data_df$y))
```

```
[1] 499.084
```



- 1000 experiments
  - We are simulating sampling 1000 times

```r
mu <- 500
sigma <- 100
n <- 1000  # sample size
k <- 100  # number of experiments

# Create data
experiment_data <- expand_grid(
  experiment = 1:k,
  observation = 1:n
) %>%
  group_by(experiment) %>%
  mutate(
    value = rnorm(n, mean = mu, sd = sigma)
  )

print (experiment_data)
```

A tibble: 100,000 × 3 | Groups: experiment [100]

experiment <int>	observation <int>	value <dbl>
1	1	448.8396
1	2	523.6938
1	3	445.8411
1	4	621.9228
1	5	517.4136
1	6	438.4732
1	7	319.3107
1	8	435.6319
1	9	704.6019
1	10	443.9238

```r
exp_means <- experiment_data %>%
  #get means for each experiment
  group_by(experiment) %>%
  summarise(experiment_mean = mean(value))
```

- 1000 experiments
  - We are simulating sampling 1000 times

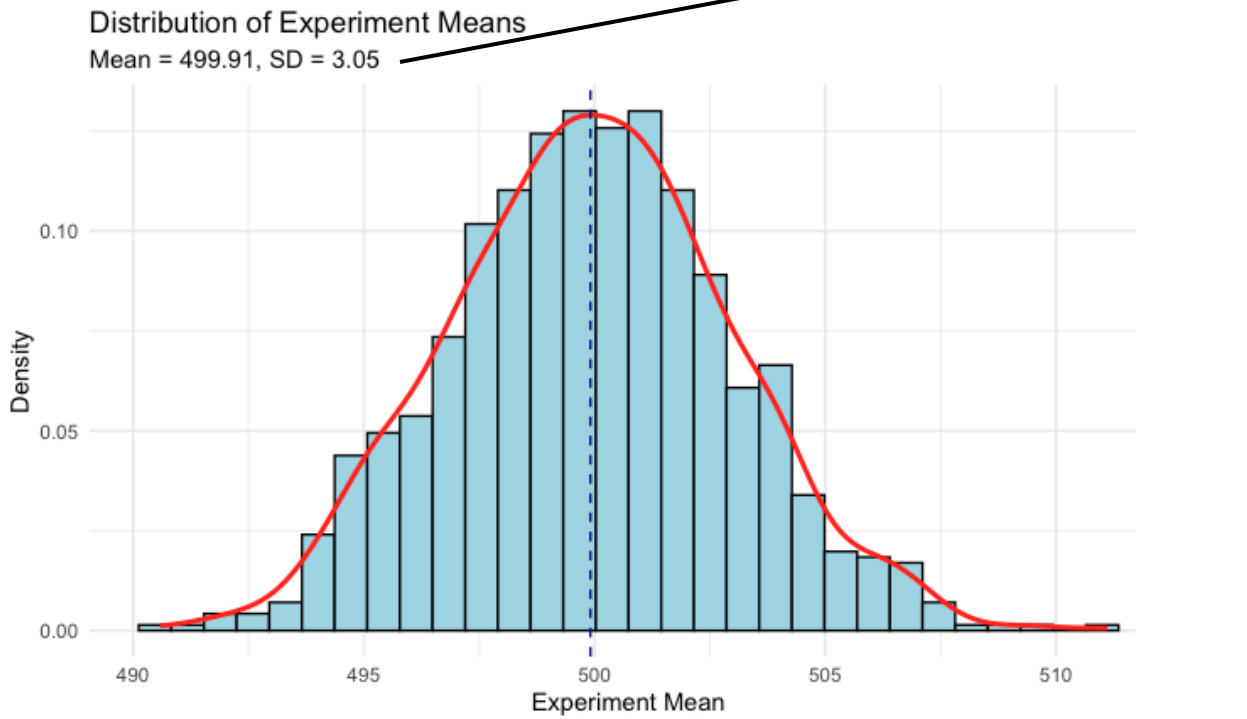
```{r}
mu <- 500
sigma <- 100
n <- 1000  # sample size
k <- 100  # number of experiments

# Create data
experiment_data <- expand_grid(
  experiment = 1:k,
  observation = 1:n
) %>%
  group_by(experiment) %>%
  mutate(
    value = rnorm(n, mean = mu, sd = sigma)
  )

print (experiment_data)
```

A tibble: 100,000 × 3 | Groups: experiment [100]

experiment <int>	observation <int>	value <dbl>
1	1	448.8396
1	2	523.6938
1	3	445.8411
1	4	621.9228
1	5	517.4136
1	6	438.4732
1	7	319.3107
1	8	435.6319
1	9	704.6019
1	10	443.9238

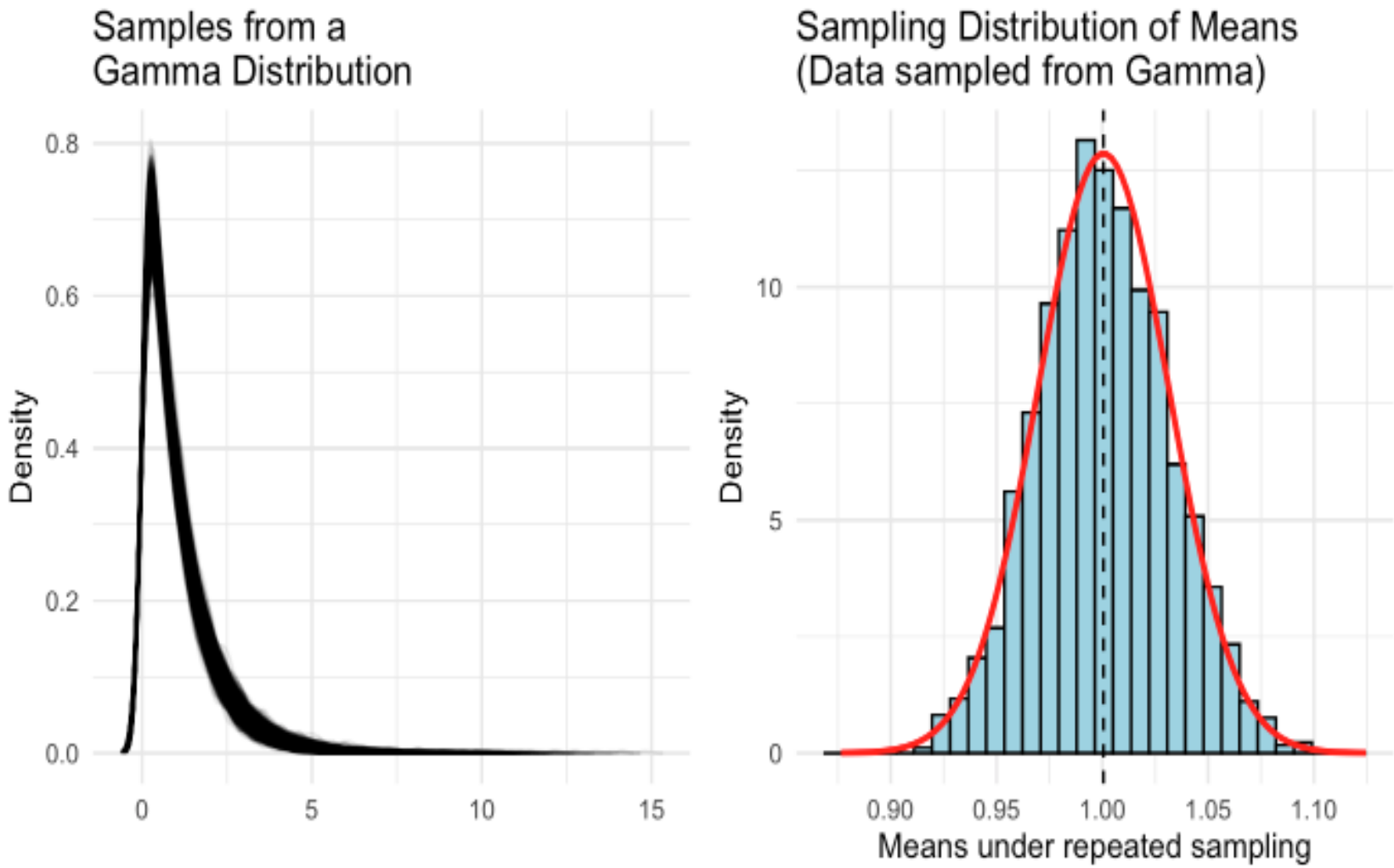
```{r}
exp_means <- experiment_data %>%
  #get means for each experiment
  group_by(experiment) %>%
  summarise(experiment_mean = mean(value))
```

Standard Error

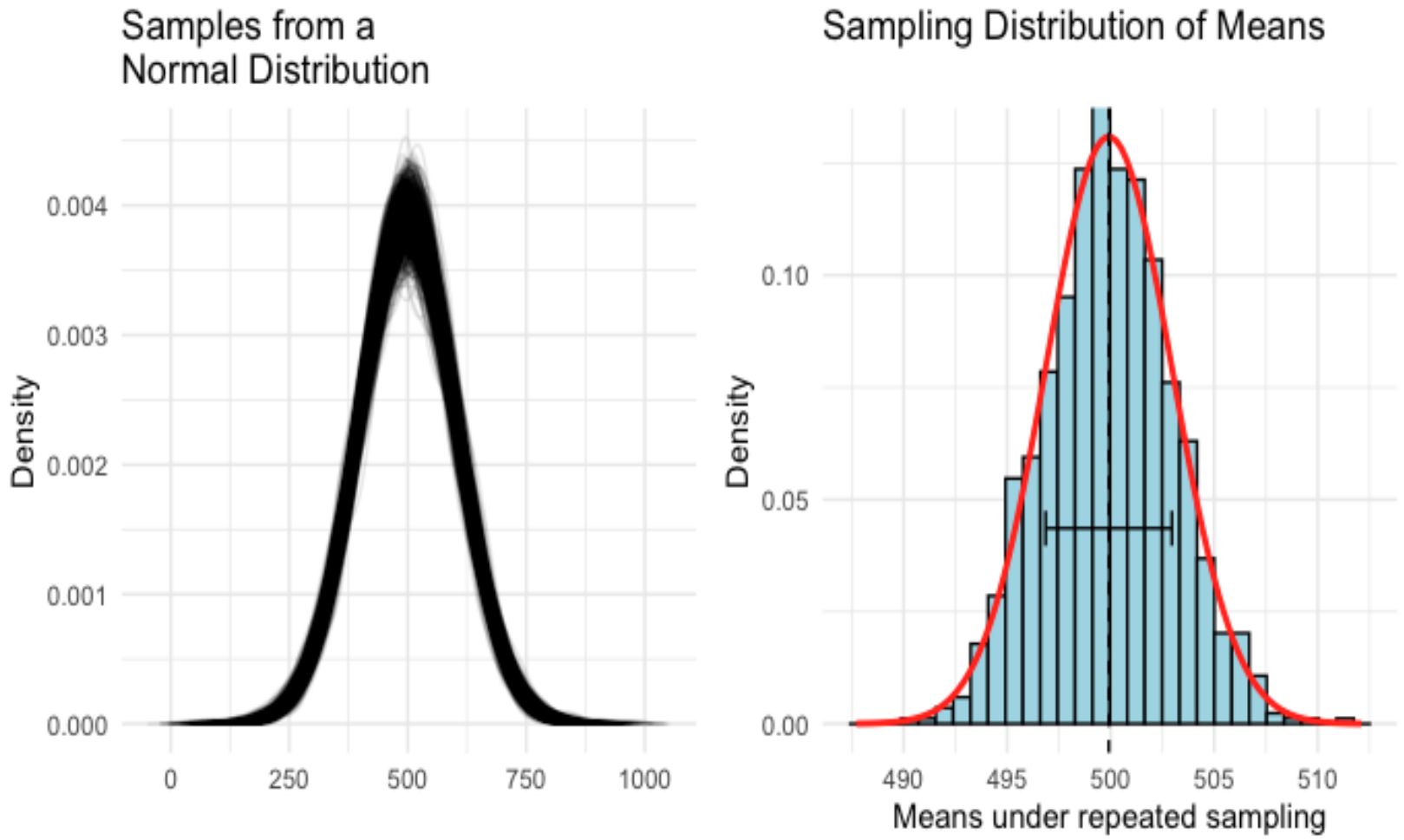


Distribution of means under repeated sampling approaches, is roughly normal

Distribution of means under repeated sampling approaches, is roughly normal

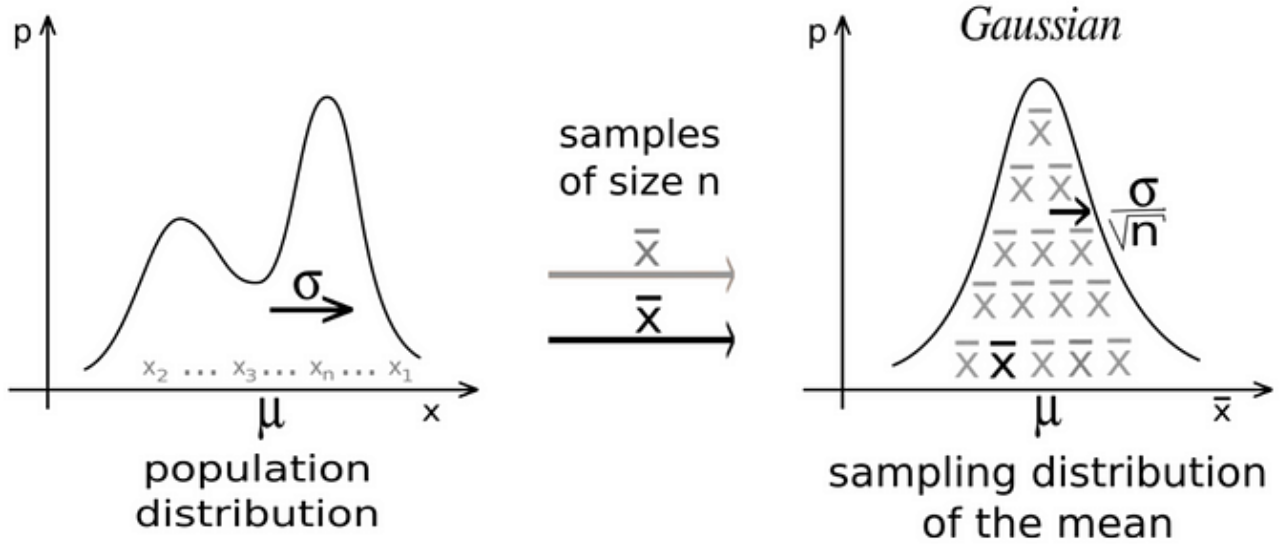


Distribution of means under repeated sampling approaches, is roughly normal



Central Limit Theorem (CLT)

- as you take more and more samples from the population (and the sample size is sufficiently large), the sampling distribution of the mean will approach a normal distribution



Requirements for CLT

- Population distribution has
  - Finite/defined mean
  - Finite/defined variance
  - Samples are independently and identically distributed (random sampling)
  - Samples should be large enough (n>=30)

- Some Exceptions
  - When population is a
    - Cauchy distribution
    - Other heavy tailed distributions w/o defined moments..

    , etc.

CLT is a cornerstone of frequentist statistics

- Makes statistical inference possible regardless of underlying population distribution
- Justifies the use of normal distribution in hypothesis testing
- Allows us to quantify uncertainty in sample estimates

- 1000 experiments === empirical sampling distribution
- Infinite experiments === theoretical sampling distribution

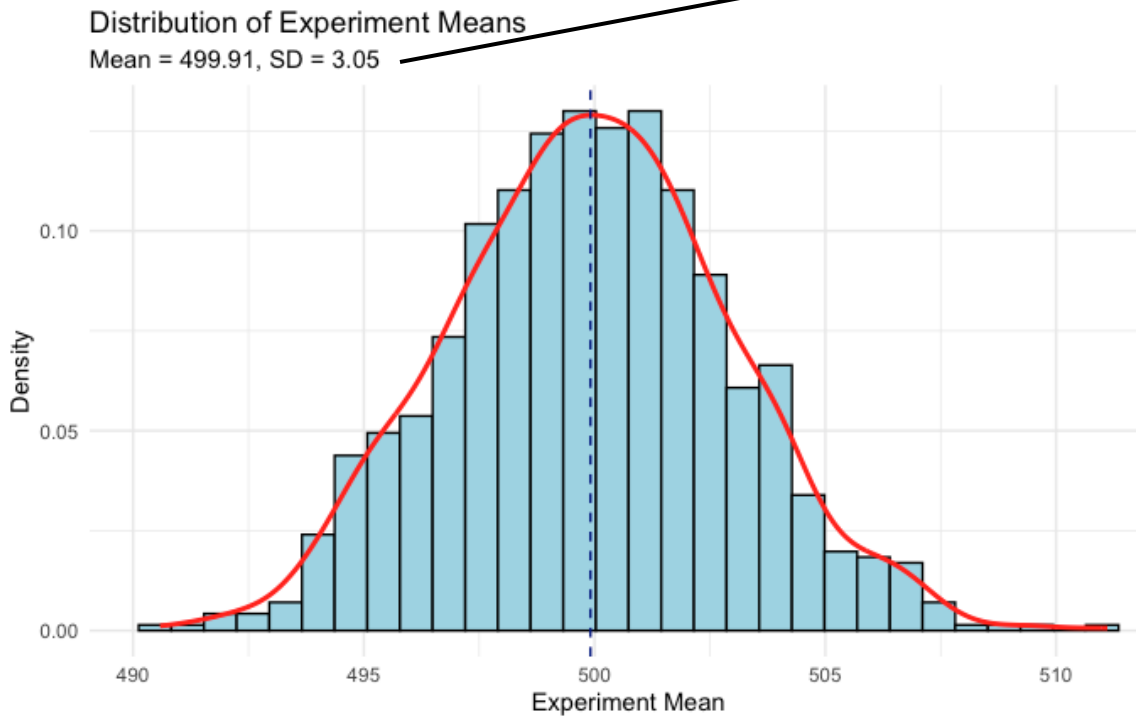
```{r}
mu <- 500
sigma <- 100
n <- 1000  # sample size
k <- 100  # number of experiments

# Create data
experiment_data <- expand_grid(
  experiment = 1:k,
  observation = 1:n
) %>%
  group_by(experiment) %>%
  mutate(
    value = rnorm(n, mean = mu, sd = sigma)
  )

print (experiment_data)
```

A tibble: 100,000 × 3 | Groups: experiment [100]

experiment <int>	observation <int>	value <dbl>
1	1	448.8396
1	2	523.6938
1	3	445.8411
1	4	621.9228
1	5	517.4136
1	6	438.4732
1	7	319.3107
1	8	435.6319
1	9	704.6019
1	10	443.9238

```{r}
exp_means <- experiment_data %>%
  #get means for each experiment
  group_by(experiment) %>%
  summarise(experiment_mean = mean(value))
```

Standard Error



This next graph is worth understanding

Population

μ = mean

σ = standard deviation

σ^2 = variance

Sample

$\overline{X}$ = mean

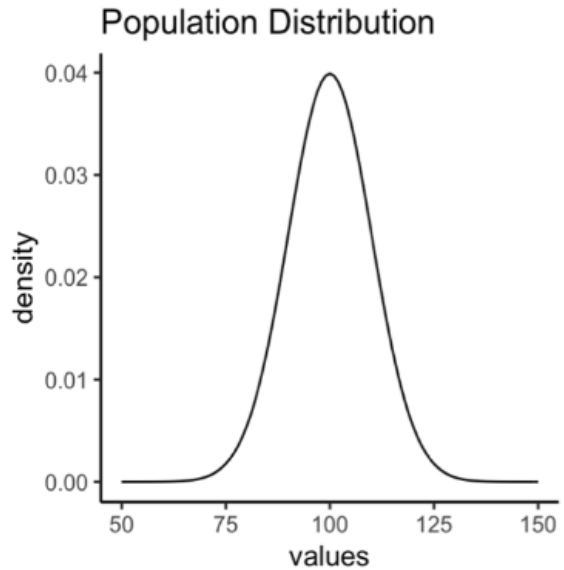
s = standard deviation

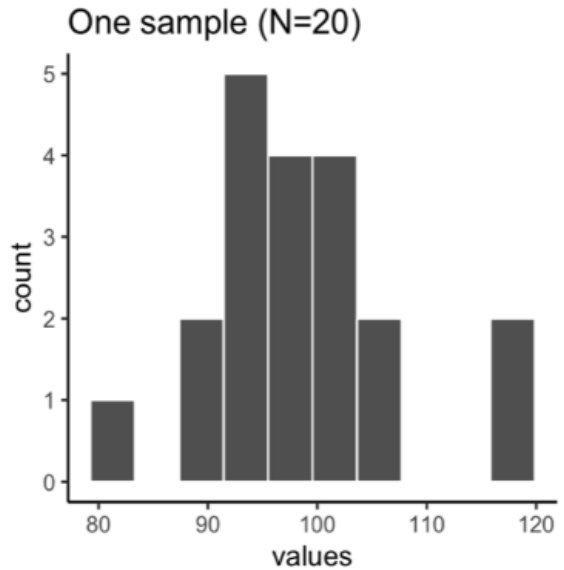
s^2 = variance

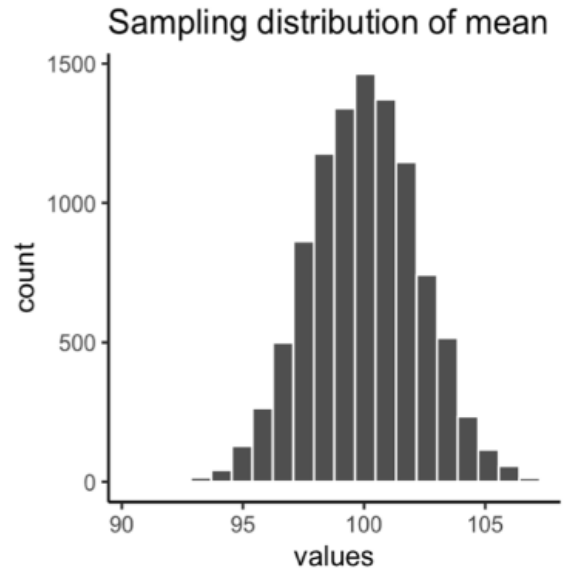
Sampling Distribution

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

SEM= s of sampling distribution







Population Estimate

$\hat{\mu}$ Estimate of population mean =

$\hat{\sigma}$ Estimate of population standard deviation =

Sample Statistic

$\overline{X}$ = sample mean

$$\sqrt{\frac{\sum (x_i - \bar{X})^2}{N-1}}$$

SEM (estimated)

$$\frac{\hat{\sigma}}{\sqrt{N}}$$

Properties of Sampling distributions

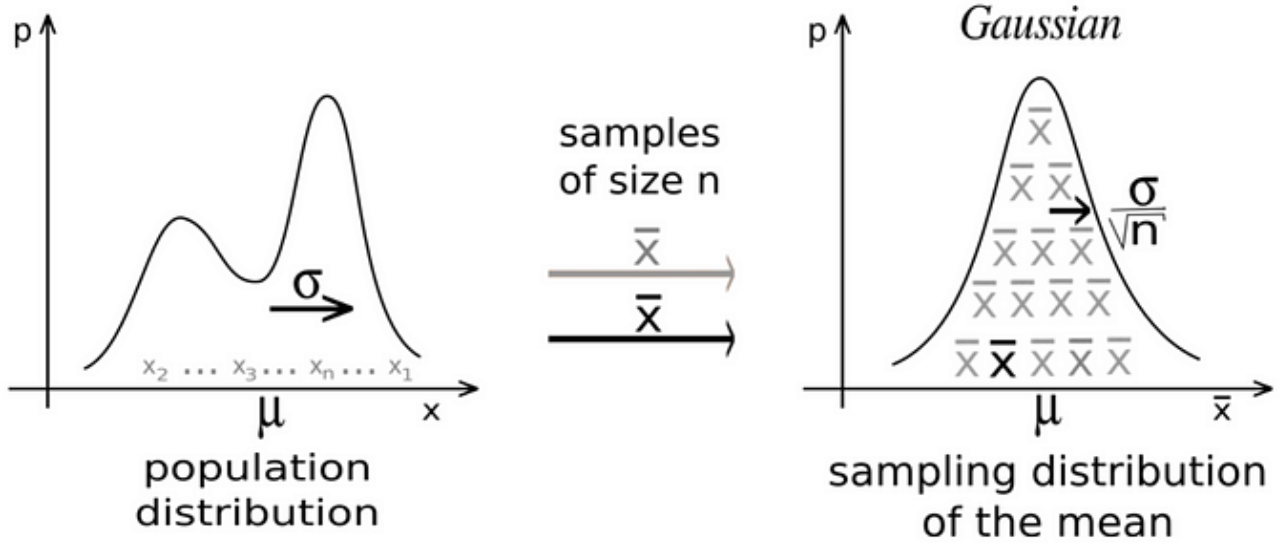
- Center
  - Mean of empirical distribution = *Estimate* of μ , the population parameter
    - Varies with each set of samples
  - Mean of theoretical distribution = μ , the population parameter
    - i.e. As number of samples → ∞, Empirical center → True population mean
    - recognize **LLN**

Properties of Sampling distributions

- Spread
  - Variance of empirical distribution = s^2/n
  - Standard deviaiton of empirical distribution = $\sqrt{(s^2/n)}$ = standard error
  - To be more accurate
    - n-1 in the denominators to reduce underestimatino
  - Variance of theoretical sampling distribution = σ^2/n

Properties of Sampling distributions

- Shape
  - Central Limit theorem
    - as you take more and more samples from the population (and the sample size is sufficiently large), the sampling distribution of the mean will approach a normal distribution



Population

μ = mean

σ = standard deviation

σ^2 = variance

Sample

$\overline{X}$ = mean

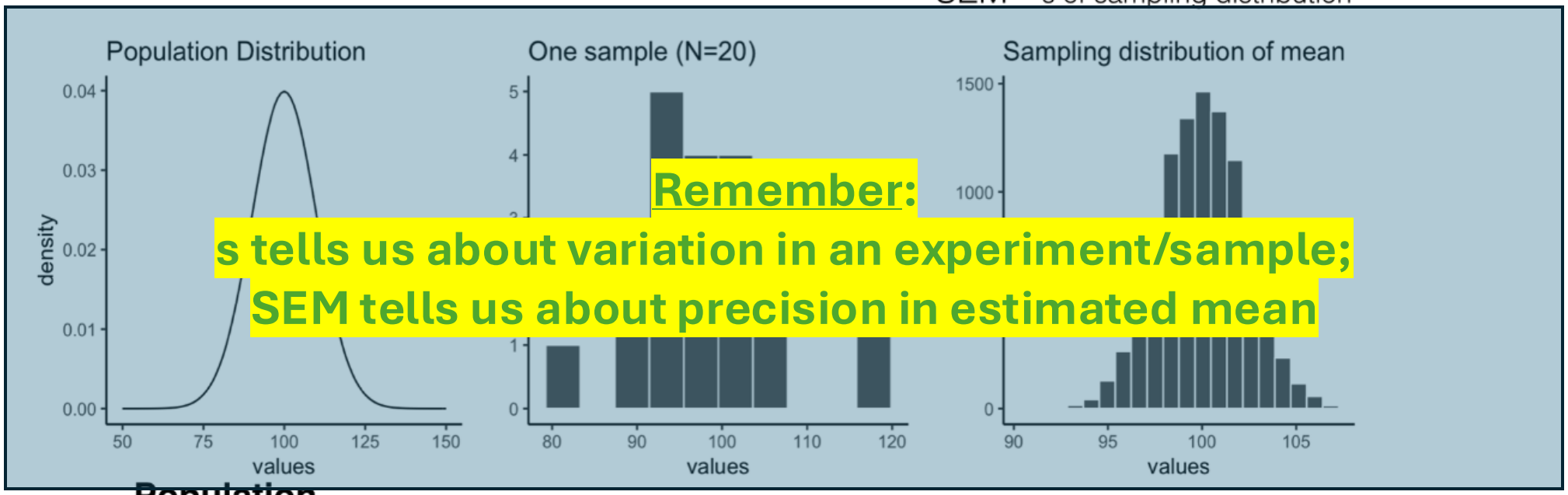
s = standard deviation

s^2 = variance

Sampling Distribution

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

SEM= s of sampling distribution



Remember:

s tells us about variation in an experiment/sample;

SEM tells us about precision in estimated mean

Population Estimate

$\hat{\mu}$ Estimate of population mean = $\overline{X}$ = sample mean

$\hat{\sigma}$ Estimate of population standard deviation = $\sqrt{\frac{\sum (x_i - \bar{X})^2}{N-1}}$

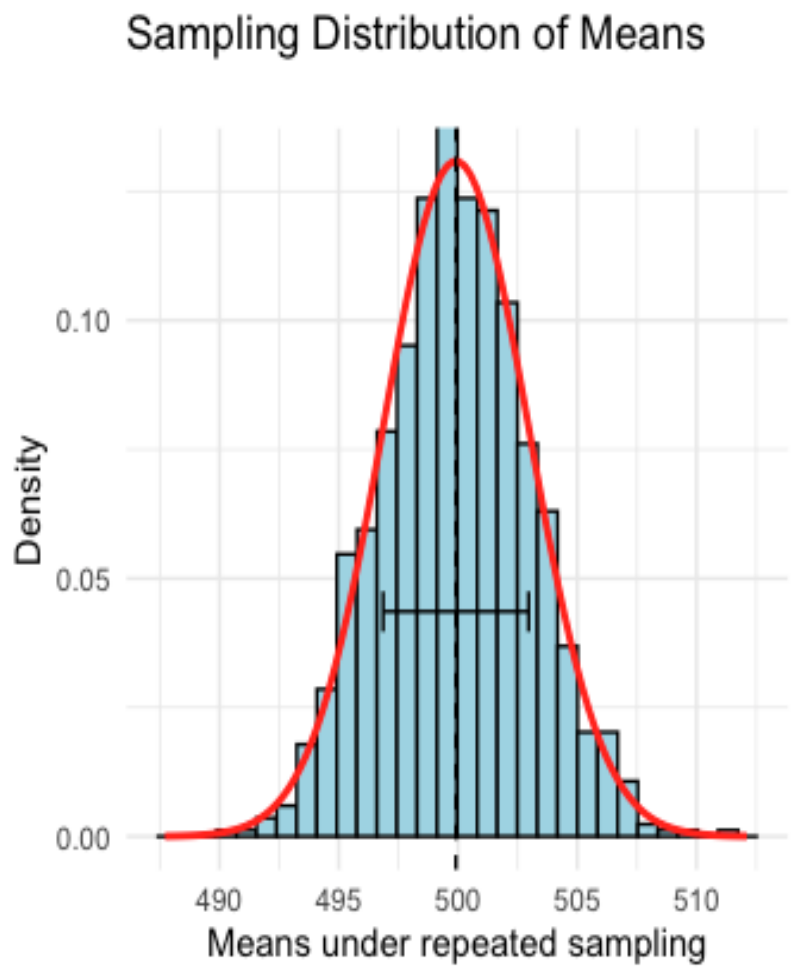
Sample Statistic

SEM (estimated)

$$\frac{\hat{\sigma}}{\sqrt{N}}$$

Confidence Intervals

- CI = Sample Mean ± (critical value × SEM)
- Intuition
  - Gives us a range of values within which we believe that the true population parameter lies.
  - Critical value is z-score related to a given significance level
    - Sample Mean ± (1.96 × SEM), for a 95% confidence
      - About 95% of our intervals contains the true mean



Interpreting Confidence Intervals

- Do not say "95% probability parameter is in interval"
- Correct: "95% of similarly constructed intervals would contain true parameter"

- Width affected by:
  - Sample size (N)
  - Population variability (σ)
  - Confidence level (e.g., 95% vs 99%)

Confidence intervals for decision-making

- Based on the confidence interval we can rule out 0 (and therefore the simple model) as a DGP for which our data is likely.
  - Confidence intervals and model comparison approaches have direct correspondence between them.
  - If 0 if contained within the confidence interval, we will retain the simple model (fail to reject simple model)
  - If 0 is not contained within the confidence interval, we will reject the simple model (adopt the complex model)
- A confidence interval provides a range of hypotheses (in this case about β1) for which we would reject the null hypothesis.

Details

Linking Timelines in Brain and Behavior: Modeling as an Iterative Bridge Between Data and Theory

How does the brain represent time in ways that support learning and adaptive behavior? In this talk, I present a case for iterative modeling that bridges neural population dynamics and cognitive-level theory. Analyses of neural recordings reveal that populations encode both elapsed time since past events and remaining time until future actions across a continuum of timescales, compressed in a way consistent with the Weber–Fechner law of time perception. These observations motivate a biologically inspired computational framework in which scale-invariant neural timelines are embedded within simple cognitive decision rules. The model accounts for canonical timing behaviors across tasks with continuously updated past and future timelines. Because future prediction must itself unfold over time, the model learns to track evolving reward probabilities within a trial, supporting flexible, adaptive decisions in dynamic environments. This work illustrates how combining cognitive theory with population-level neural data can uncover shared computational principles across levels of analysis.

Wednesday

- Bootstrapping
- Papaja