# Statistics for CSAI II

## Sampling Theory

Dr. Travis J. Wiltshire |

TILBURG UNIVERSITY

# Outline

1. Populations vs samples
2. Sampling Methods
3. Sample statistics vs population parameters
4. Estimating population parameters from sample statistics
5. Confidence intervals

# Modules

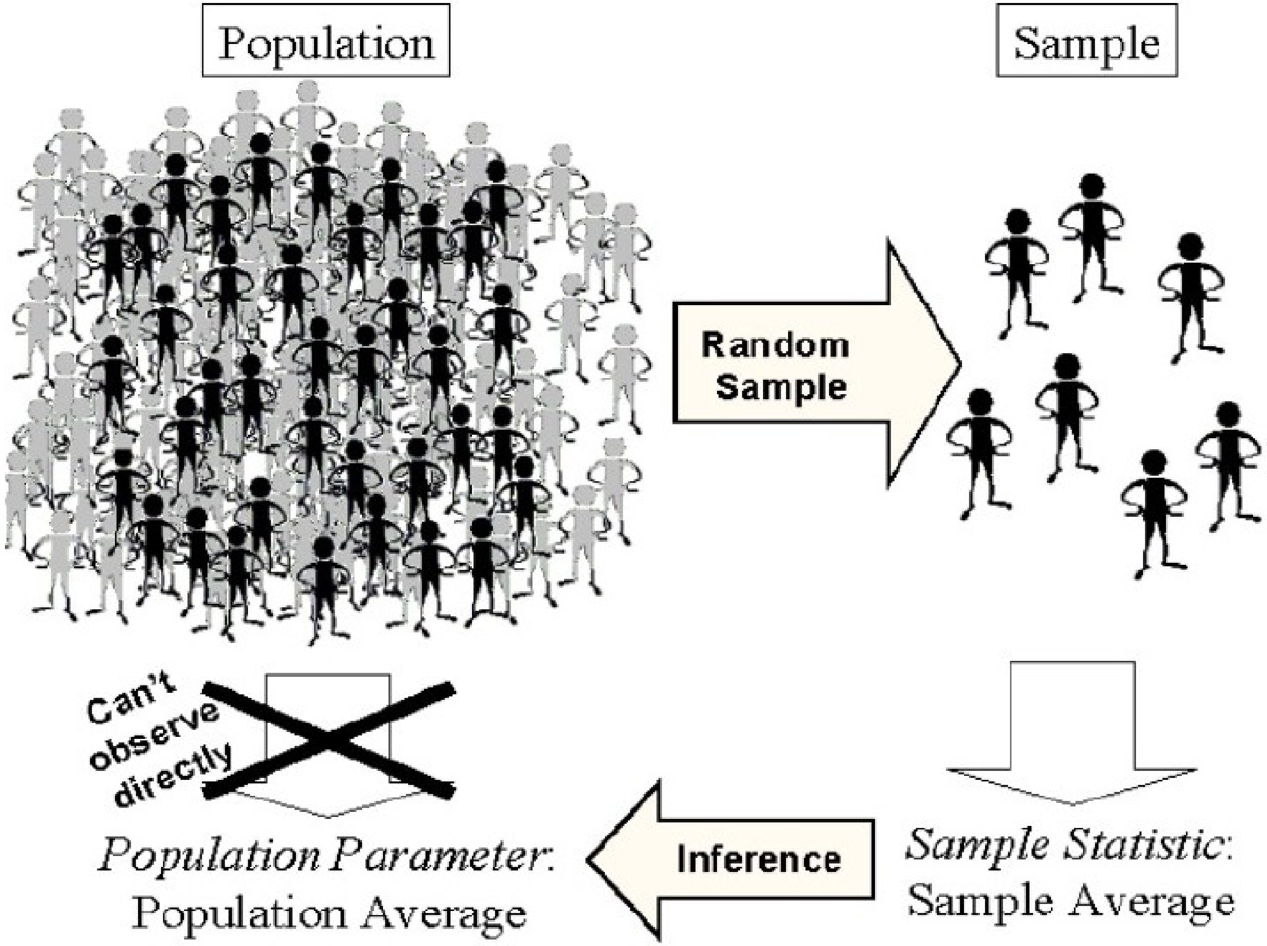
1. Introduction and Probability
2. *Sampling Theory*
3. Revisiting Hypothesis Testing & Intro to Correlation
4. Correlation
5. Intro to Regression
6. More Regression Centering and Checking Assumptions
7. Multiple Regression and Assumptions
8. Interactions
9. Multiple Regression with Categories
10. Multiple Regression with Polynomials
11. Mixed Models
12. Growth Curve Analysis

# Defining a population

- What is a population we might be interested in?
  - All of the undergraduate CS&AI students at Tilburg University?
  - Undergraduate CS&AI students in general, anywhere in the world?
  - All Dutch people currently living?
  - People who play an instrument?
  - Anyone currently alive?
  - Youths that play video games in VR?
  - Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?

# Populations and Samples

- **Population**
  - The collection of units (be they people, plankton, plants, cities, etc.) to which we want to generalize a set of findings or a statistical model
- **Sample**
  - A smaller (but hopefully representative) collection of units from a population used to determine truths about that population
  - Finite



# Empirical work in CSAI

**Goal:** To use our knowledge of the sample to draw **inferences** about the target population

- Descriptive statistics questions: What was the average survey response to question 7?
- Inferential statistics: What can we say about the average response to question 7 for the population

# Sampling Methods

- **Simple random sampling:** idea for statistical inference

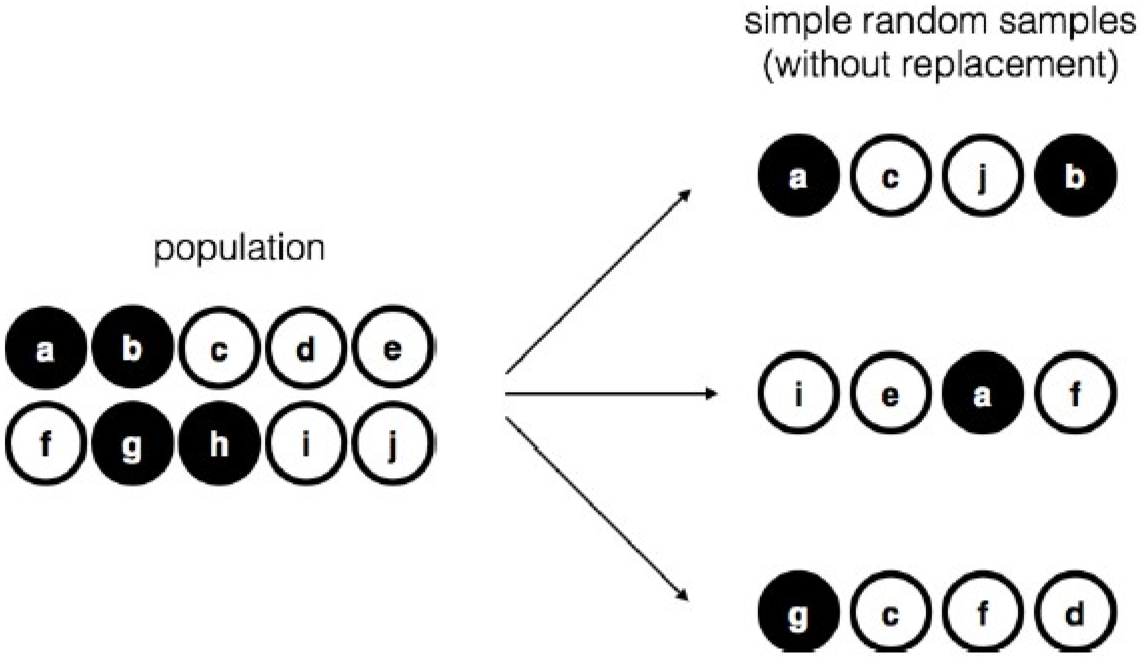


Figure 10.1: Simple random sampling without replacement from a finite population

# Side note on randomness

- Random vs pseudo random
  - **<u>Random sampling or a random process</u>** using the same procedure that can lead to different results each time with each person having equal chance of being selected
  - But, R relies on a deterministic process to generate ‘random’ numbers

DILBERT By Scott Adams
TOUR OF ACCOUNTING
OVER HERE WE HAVE OUR RANDOM NUMBER GENERATOR.
NINE NINE NINE NINE NINE NINE
ARE YOU SURE THAT'S RANDOM?
THAT'S THE PROBLEM WITH RANDOMNESS: YOU CAN NEVER BE SURE.
www.dilbert.com
scottadams@aol.com
© 2001 United Feature Syndicate, Inc.

For more, check out: https://www.random.org/randomness/

# Sampling Methods

- **Biased sampling:** sample does not reflect population parameters

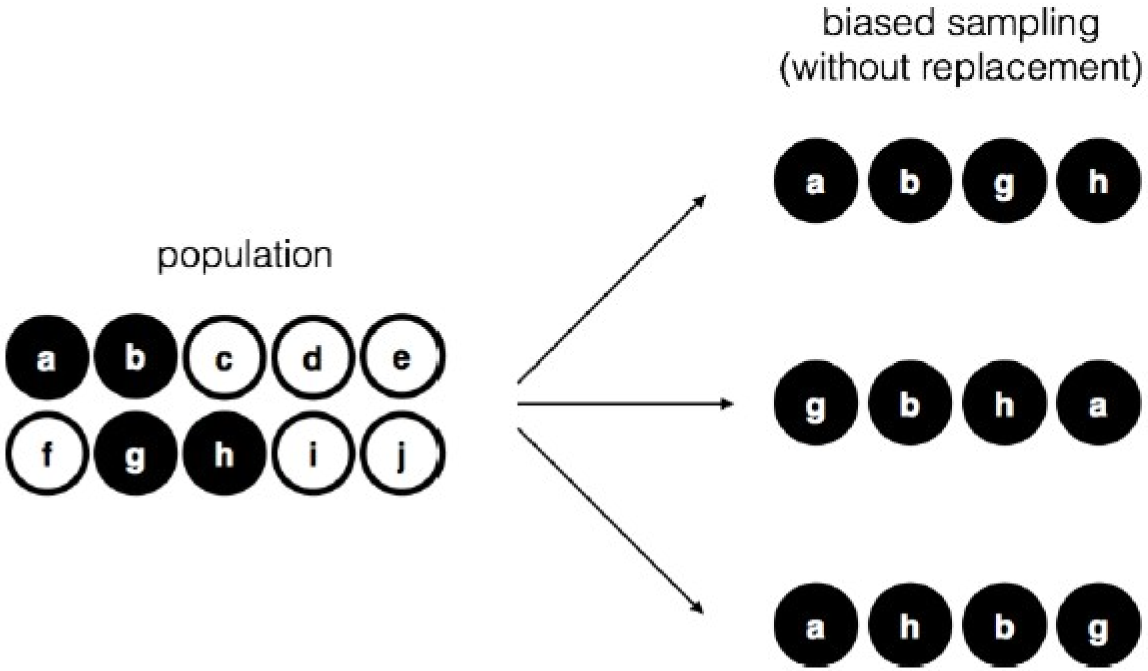


Figure 10.2: Biased sampling without replacement from a finite population

# Sampling Methods

- **Simple random sampling with replacement:** possible to observe the sample population member multiple times

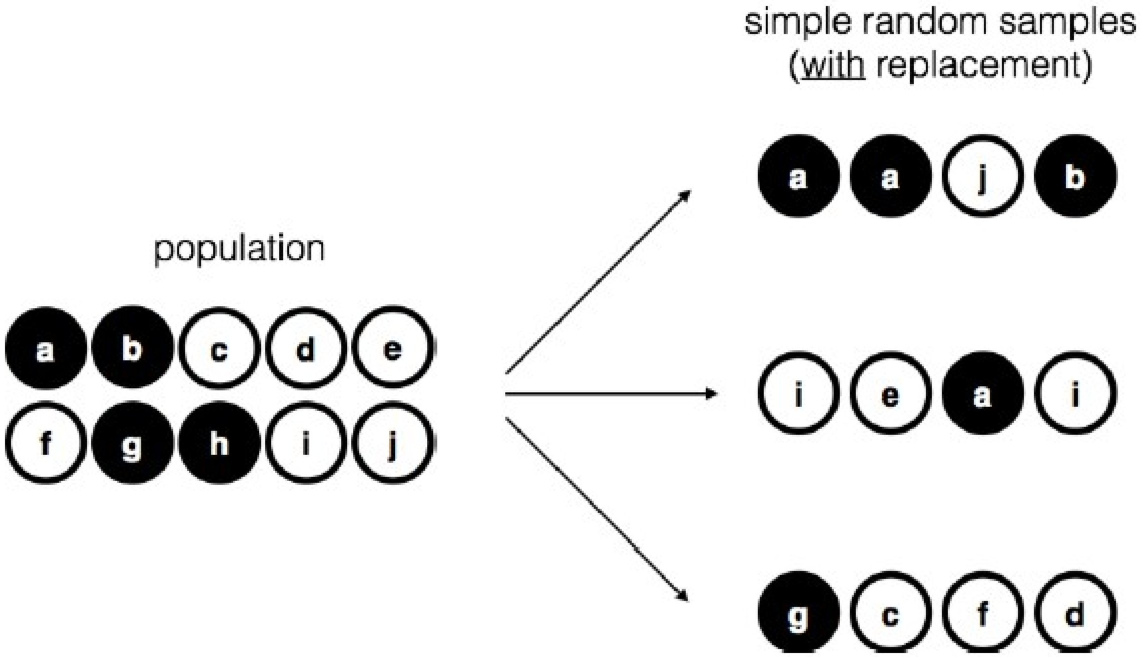


Figure 10.3: Simple random sampling *with* replacement from a finite population

# Sample in R (Exercise)

- Create a variable called `pop` that includes the letters a-j
- Create a variable called `samp1` by sampling the population using the sample() function
- Create a variable called `samp2` by sampling the population using the sample() function *with replacement*
- Try to figure out how to add a bias to your sample that only draws the 'black' letters  a, b, g, and h and then make a `samp3`  variable with the biased sample.

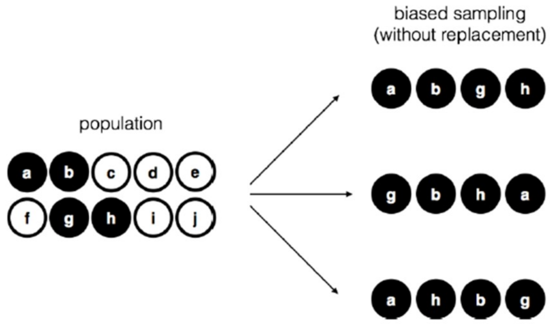
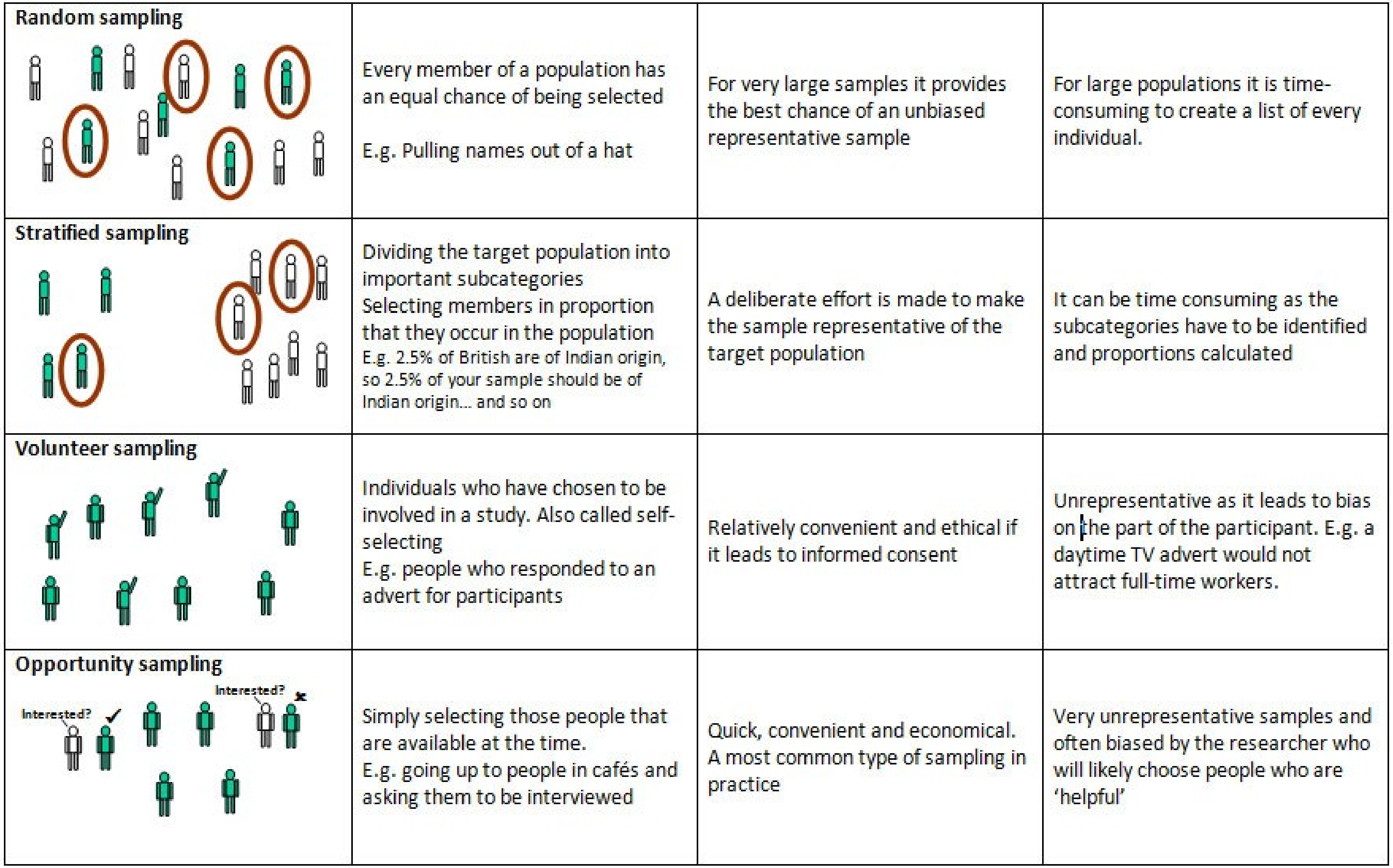


Figure 10.2: Biased sampling without replacement from a finite population

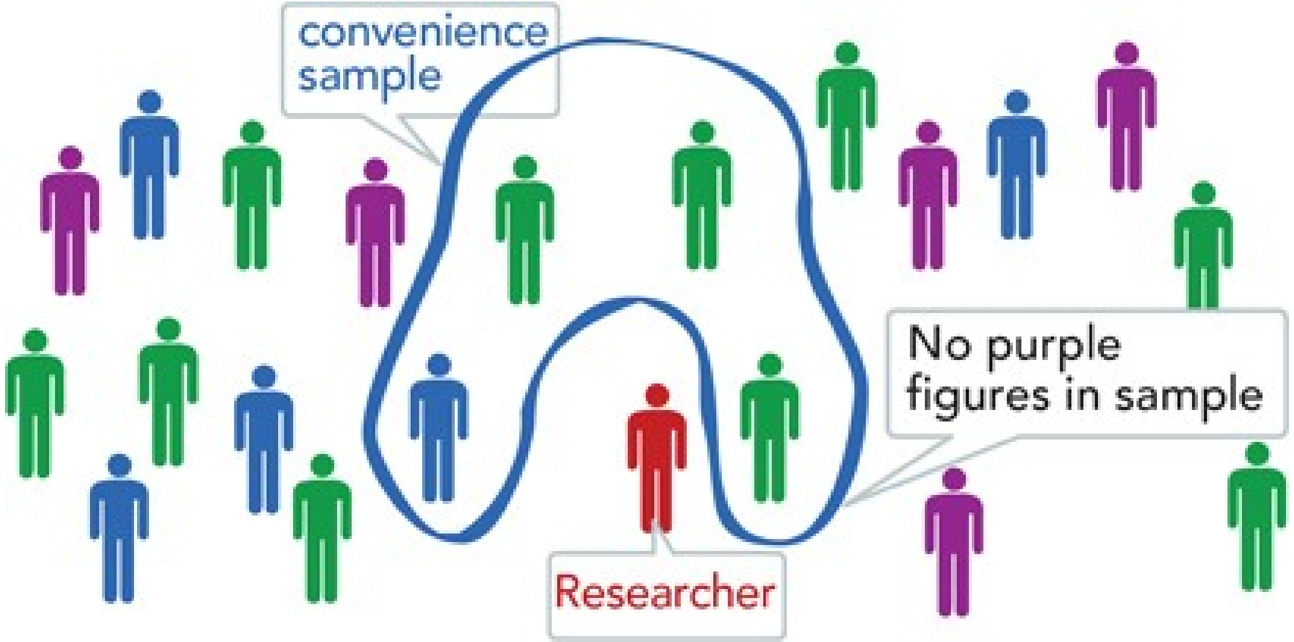
# Types of sampling in practice

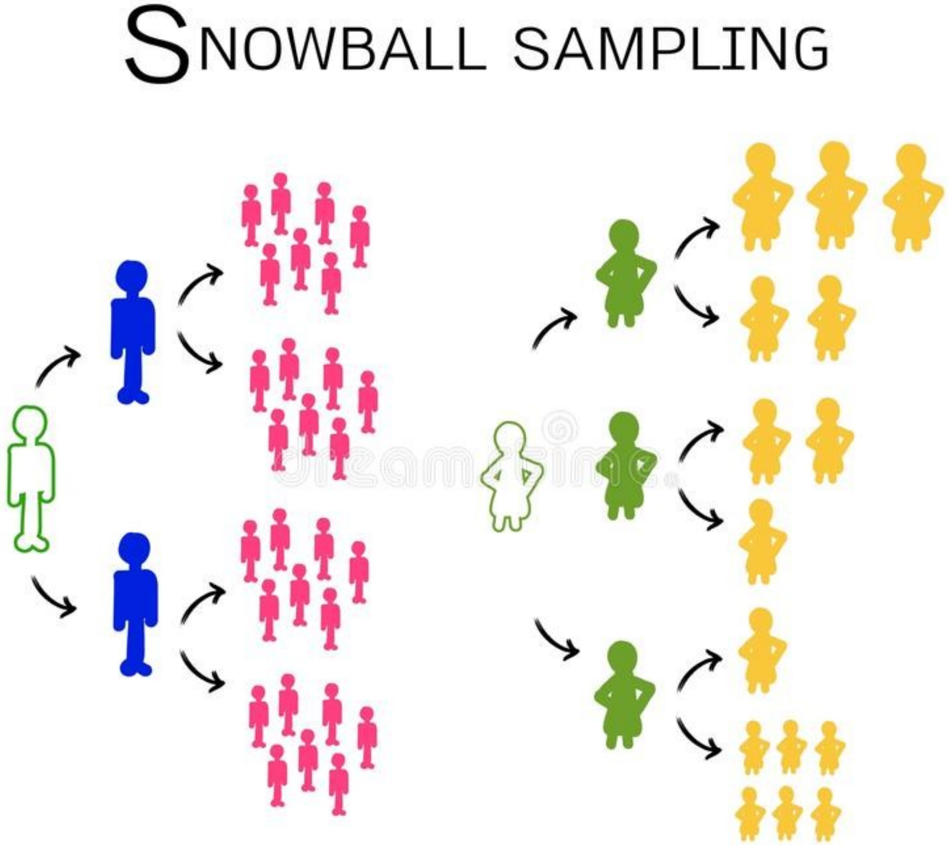


Interested?
✓
Interested?
✗

| | | | |
|---|---|---|---|
| Random sampling | Every member of a population has an equal chance of being selected<br>E.g. Pulling names out of a hat | For very large samples it provides the best chance of an unbiased representative sample | For large populations it is time-consuming to create a list of every individual. |
| Stratified sampling | Dividing the target population into important subcategories<br>Selecting members in proportion that they occur in the population<br>E.g. 2.5% of British are of Indian origin, so 2.5% of your sample should be of Indian origin… and so on | A deliberate effort is made to make the sample representative of the target population | It can be time consuming as the subcategories have to be identified and proportions calculated |
| Volunteer sampling | Individuals who have chosen to be involved in a study. Also called self-selecting<br>E.g. people who responded to an advert for participants | Relatively convenient and ethical if it leads to informed consent | Unrepresentative as it leads to bias on the part of the participant. E.g. a daytime TV advert would not attract full-time workers. |
| Opportunity sampling | Simply selecting those people that are available at the time.<br>E.g. going up to people in cafés and asking them to be interviewed | Quick, convenient and economical.<br>A most common type of sampling in practice | Very unrepresentative samples and often biased by the researcher who will likely choose people who are 'helpful' |

# Convenience sampling and snowball sampling





# Does your sampling method matter?

- It depends!
  - What is the target population?
  - Could your sample be biased? In what way?
  - Even if there is a bias, is the phenomena of interest likely to vary randomly in your sample?
- Often essential to:
  - Measure and report **diversity** in your sample
  - Be **transparent** about your sample and how they were recruited

# Sampling methods activity

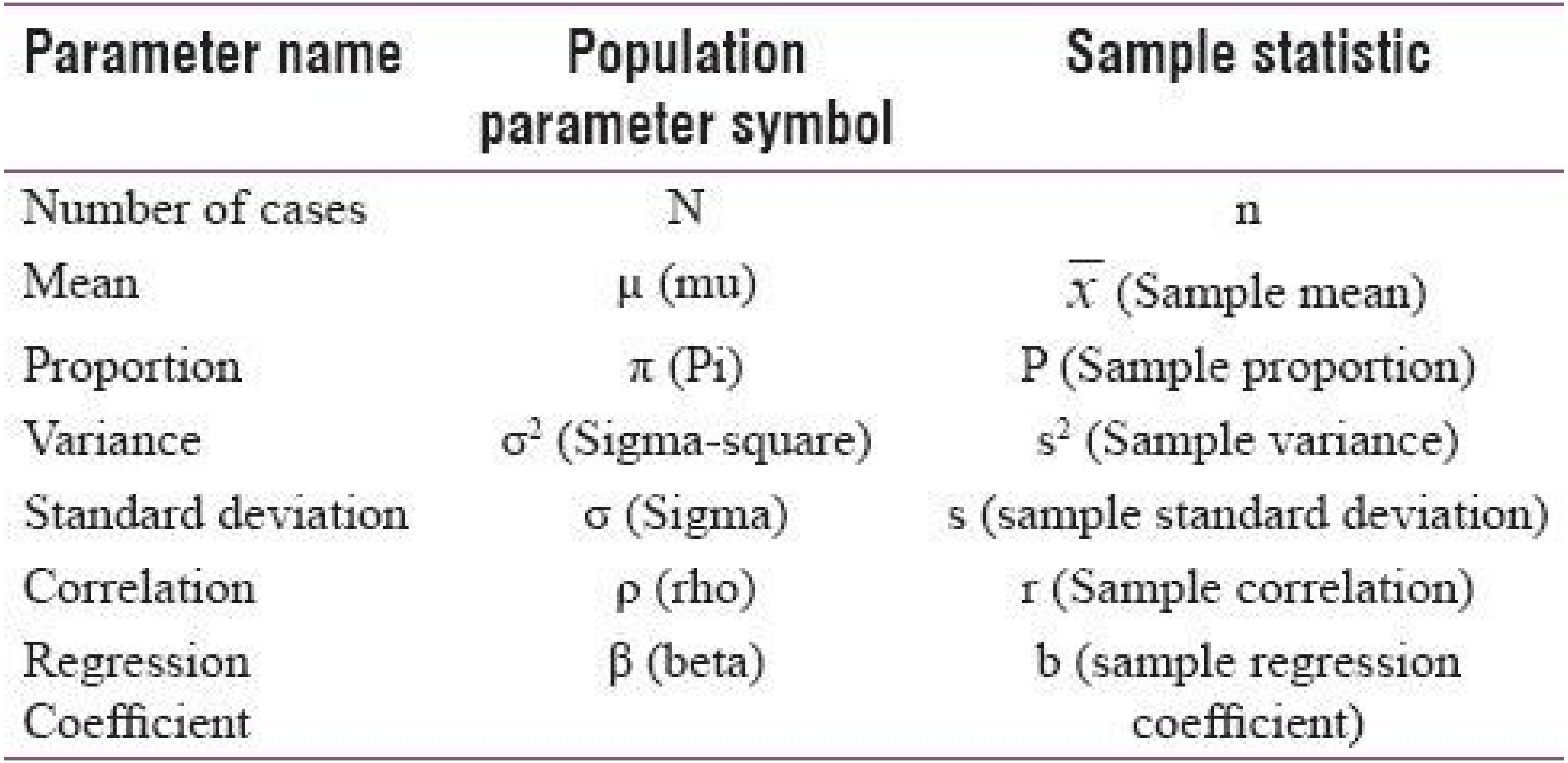
**Discuss with a partner which sampling method might be most realistic in the cases below, why, what details you might need to report, and how the sample could be biased.**

- The perceptions of different aged populations of a virtual agent helping with health issues
- How children with autism spectrum disorder learn social skills from a robot
- What cognitive capacities are associated with better problem solving performance
- People’s levels of stress after arriving to work after taking public transit

# Sample statistics vs. Population parameters

- **Sample**
  - Mean and SD describe only the sample from which they were calculated.
- **Population**
  - Mean and SD are intended to describe the entire population (very rare in cog sci).
- **Sample to Population**
  - Mean and SD are obtained from a sample, but are used to **estimate** the mean and SD of the population (very common in cog sci).

# Notation



| Parameter name | Population parameter symbol | Sample statistic |
|---|---|---|
| Number of cases | N | n |
| Mean | μ (mu) | $\overline{x}$ (Sample mean) |
| Proportion | π (Pi) | P (Sample proportion) |
| Variance | $\sigma^2$ (Sigma-square) | $s^2$ (Sample variance) |
| Standard deviation | σ (Sigma) | s (sample standard deviation) |
| Correlation | ρ (rho) | r (Sample correlation) |
| Regression Coefficient | β (beta) | b (sample regression coefficient) |

# Sample statistics vs. Population parameters

- Lets generate some data

```
>IQ.1 <- round(rnorm(n=100, mean=100, sd=15))
>IQ.2 <- round(rnorm(n=1000,mean=100,sd=15))
>IQ.3 <- round(rnorm(n=10000,mean=100,sd=15))
```

Now compare the mean and standard deviation of both of these.

What do you observe? And why?

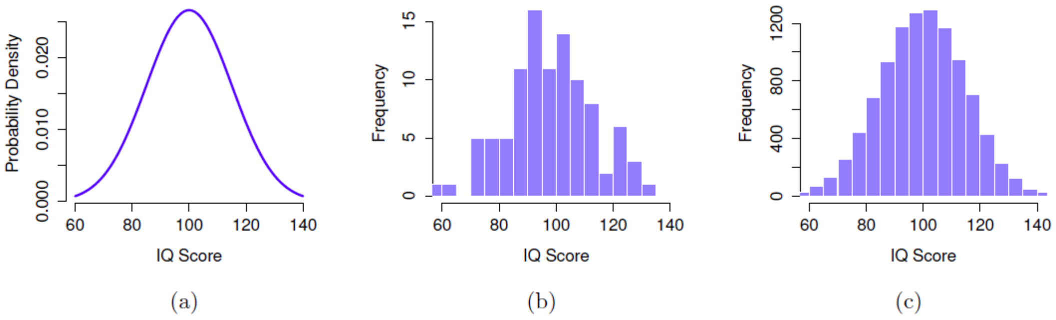
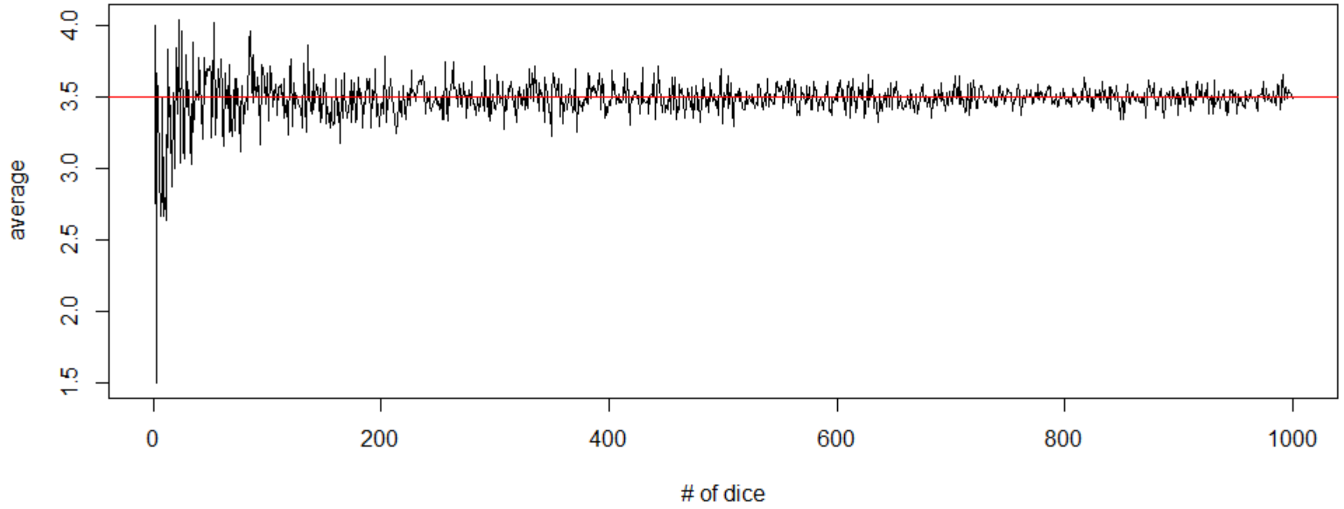


Figure 10.4: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

# Law of large numbers

- We get more ‘accurate’ numbers from larger samples.
- With the roll of a six sided die we see: ---------------------------→
- As our sample size increases (N -> Inf), the sample mean approaches the population mean.
- Our sample statistics often have a degree of inaccuracy, but this tells us with more data, we can move closer and closer to the true population parameters



# Sampling distribution of the mean

- Replicating an experiment over and over again results in a **distribution of sample means**
- Gives information about the behavior of the samples

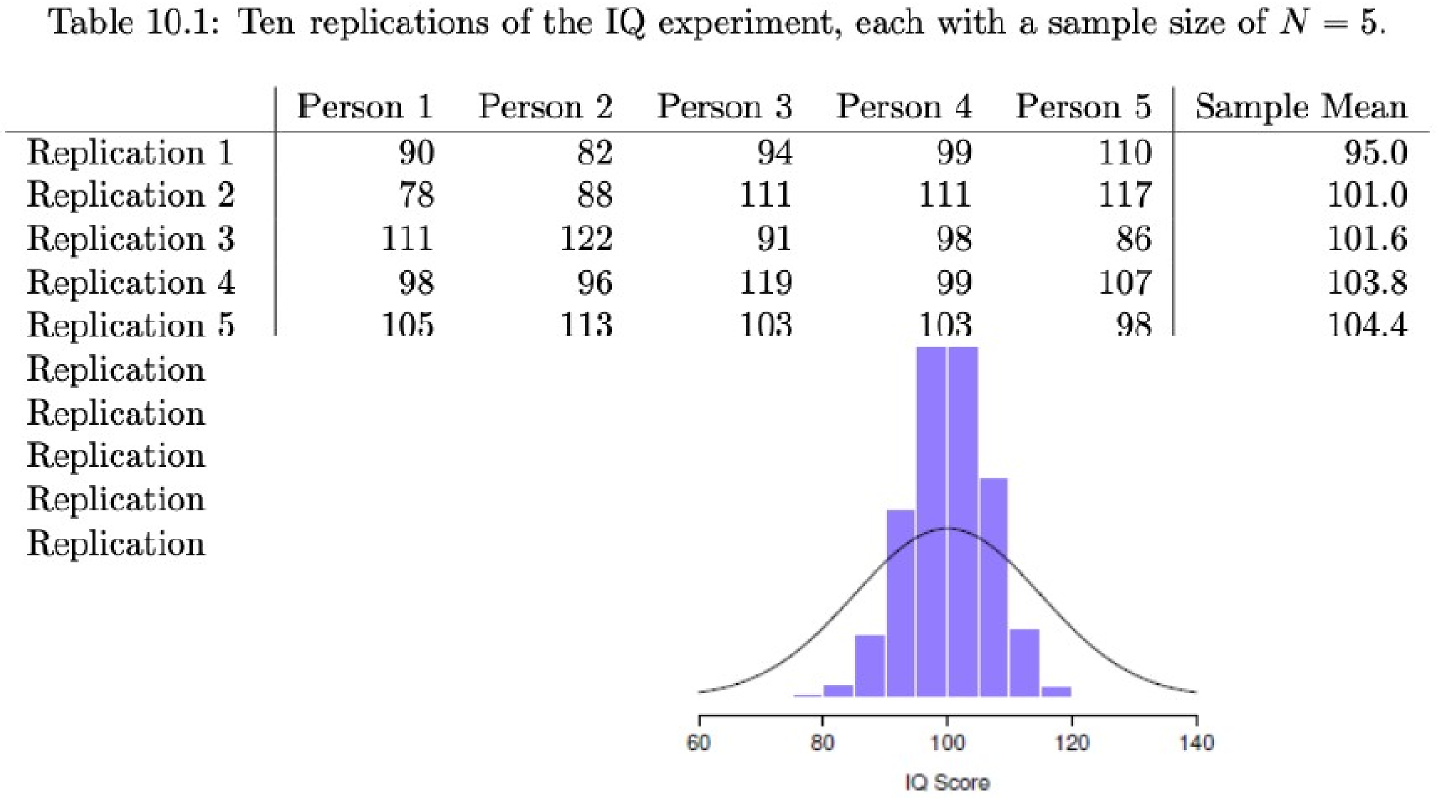


Table 10.1: Ten replications of the IQ experiment, each with a sample size of $N = 5$.

| | Person 1 | Person 2 | Person 3 | Person 4 | Person 5 | Sample Mean |
|---|---|---|---|---|---|---|
| Replication 1 | 90 | 82 | 94 | 99 | 110 | 95.0 |
| Replication 2 | 78 | 88 | 111 | 111 | 117 | 101.0 |
| Replication 3 | 111 | 122 | 91 | 98 | 86 | 101.6 |
| Replication 4 | 98 | 96 | 119 | 99 | 107 | 103.8 |
| Replication 5 | 105 | 113 | 103 | 103 | 98 | 104.4 |
| Replication | | | | | | |
| Replication | | | | | | |
| Replication | | | | | | |
| Replication | | | | | | |
| Replication | | | | | | |

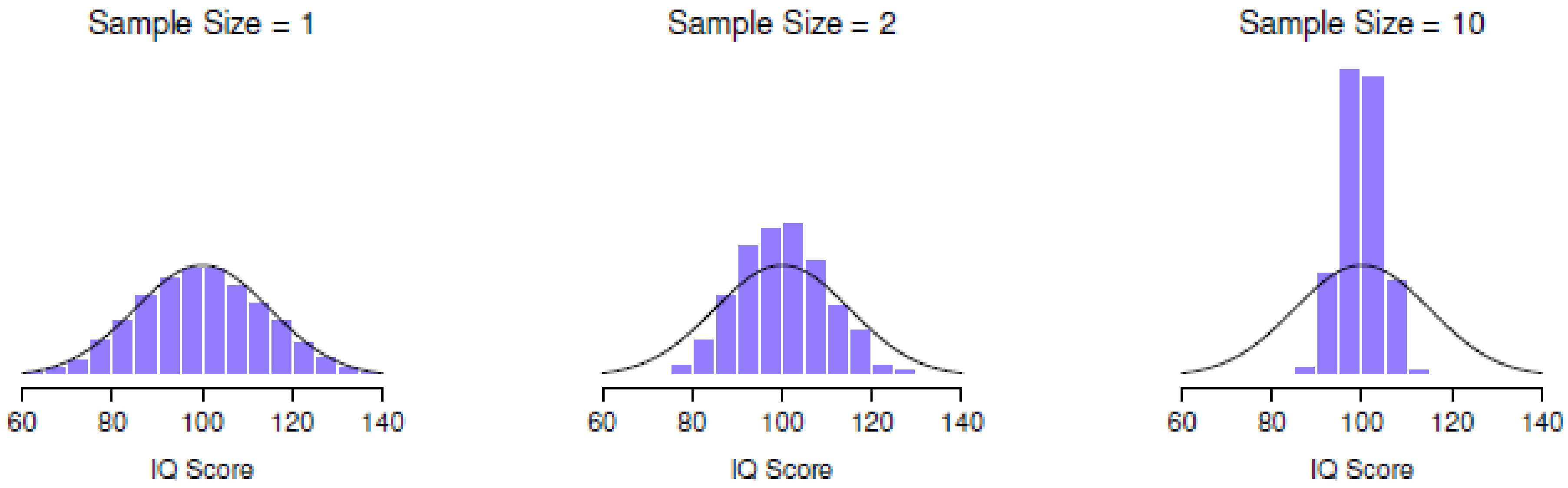
Figure 10.5: The sampling distribution of the mean for the "five IQ scores experiment". If you sample 5 people at random and calculate their *average* IQ, you'll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population distribution of IQ scores.

# Sampling distribution of the mean

- Using the same sample information from before (M = 100, SD = 15)
  - Calculate the sampling distribution of the mean for 100, 1000, and 10000 samples with size 5
  - Plot a histogram for each of these
  - Calculate the sampling distribution of the mean for 100, 1000, and 10000 samples with size 10000
  - Plot a histogram for each of these
  - HINT > You may want to use a for loop

# Sampling distribution of the mean

- Higher sample size results in sample means closer to population means
- The standard deviation of the sampling distribution is referred to as the standard error (SE)
  - Typically interested in the standard error of the mean (SEM)
  - As sample size N increases, the SEM decreases



# Central limit theorem.

- Given a **sufficiently sized sample**, the following claims are typically true:
  - The mean of the sampling distribution is the same as the mean of the population
  - The standard deviation of the sampling distribution (the standard error) gets smaller as the sample size increases
  - The shape of the sampling distribution becomes normal as the sample size increases.
- What can this tell us?
  - Why larger experiments are more reliable than smaller ones
  - And how much more (in terms of standard error)
  - Why the **normal** distribution is normal

$$SE = \frac{\sigma \longleftarrow \text{Standard deviation}}{\sqrt{n} \longleftarrow \text{Number of samples}}$$

# Standard Error

- Using the same sample information from before (M = 100, SD = 15)
  - Calculate the standard error of the mean for 10000 samples with size 5
  - Plot a histogram for each of these

- Using the same sample information from before (M = 100, SD = 15)
  - Calculate the standard error of the mean for 10000 samples with size 10000
  - Plot a histogram for each of these

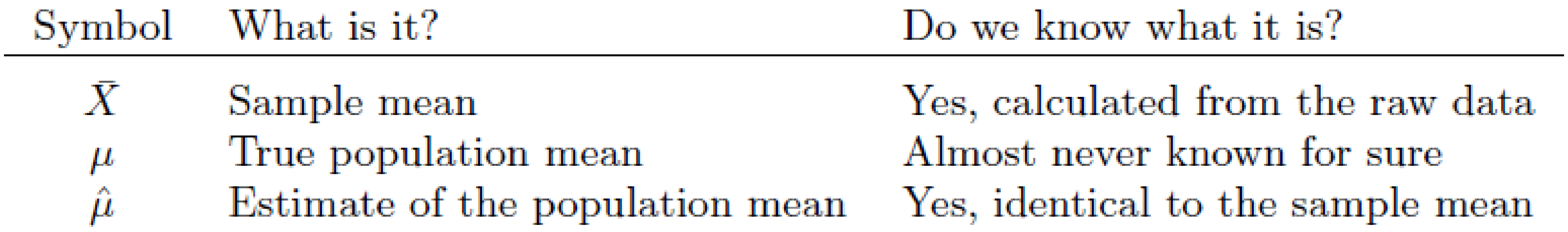
$$SE = \frac{\sigma}{\sqrt{n}}$$

← Standard deviation

← Number of samples

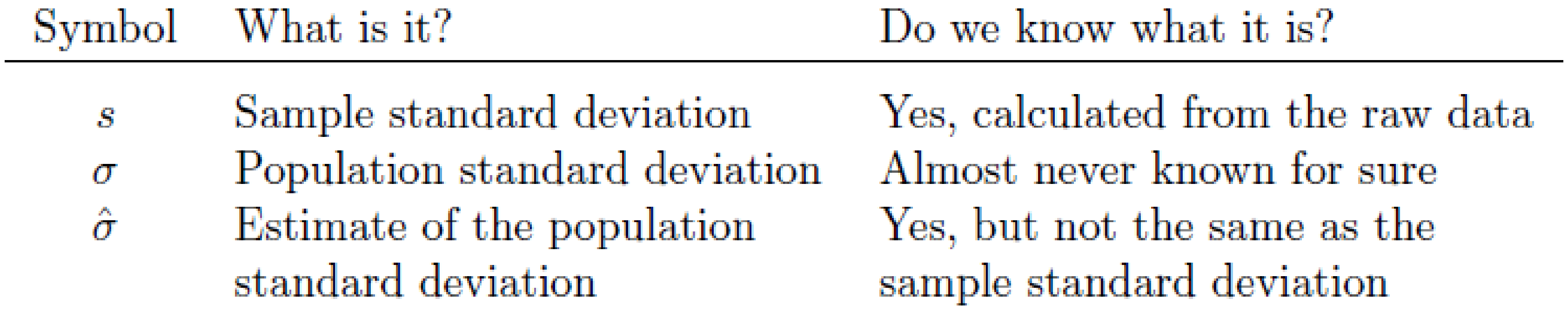
# Estimating population parameters

- Often, we just give our best guess

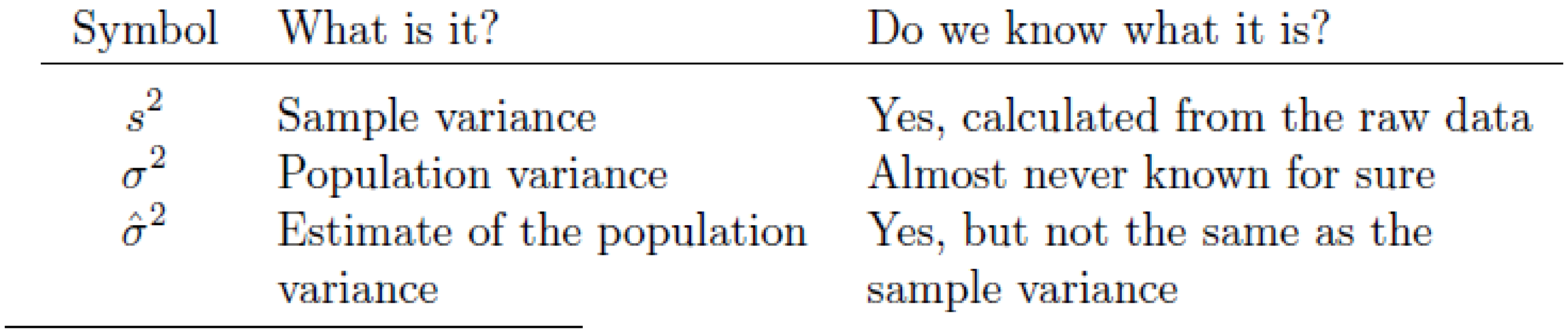


| Symbol | What is it? | Do we know what it is? |
|---|---|---|
| $\bar{X}$ | Sample mean | Yes, calculated from the raw data |
| $\mu$ | True population mean | Almost never known for sure |
| $\hat{\mu}$ | Estimate of the population mean | Yes, identical to the sample mean |

# Estimating population parameters



| Symbol | What is it? | Do we know what it is? |
|---|---|---|
| $s$ | Sample standard deviation | Yes, calculated from the raw data |
| $\sigma$ | Population standard deviation | Almost never known for sure |
| $\hat{\sigma}$ | Estimate of the population standard deviation | Yes, but not the same as the sample standard deviation |



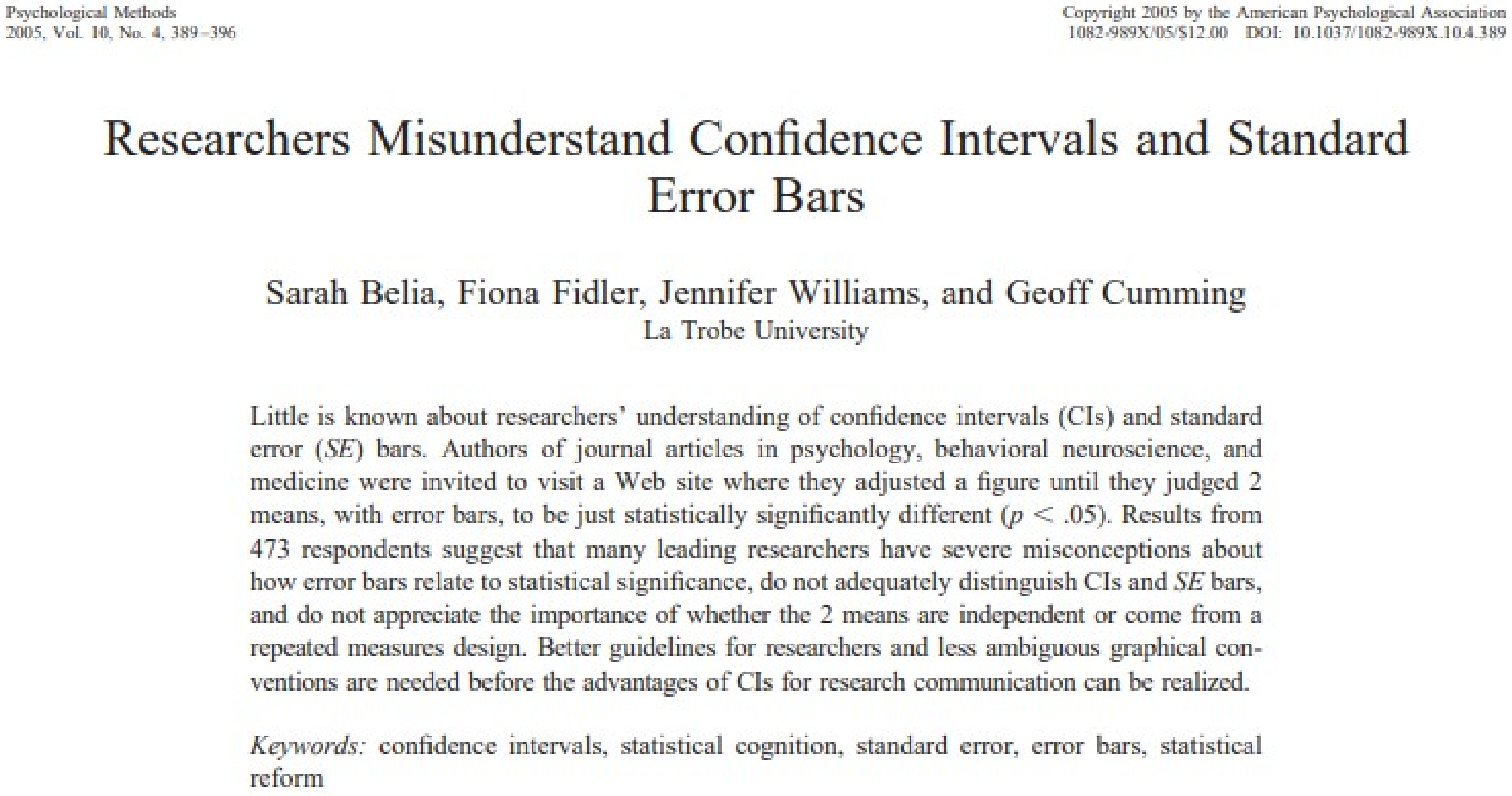
| Symbol | What is it? | Do we know what it is? |
|---|---|---|
| $s^2$ | Sample variance | Yes, calculated from the raw data |
| $\sigma^2$ | Population variance | Almost never known for sure |
| $\hat{\sigma}^2$ | Estimate of the population variance | Yes, but not the same as the sample variance |

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

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

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

- Be careful in R as the standard functions give us the estimates of the population and not the sample statistics



Psychological Methods
2005, Vol. 10, No. 4, 389–396

1082-989X/05/$12.00 DOI: 10.1037/1082-989X.10.4.389

# Researchers Misunderstand Confidence Intervals and Standard Error Bars

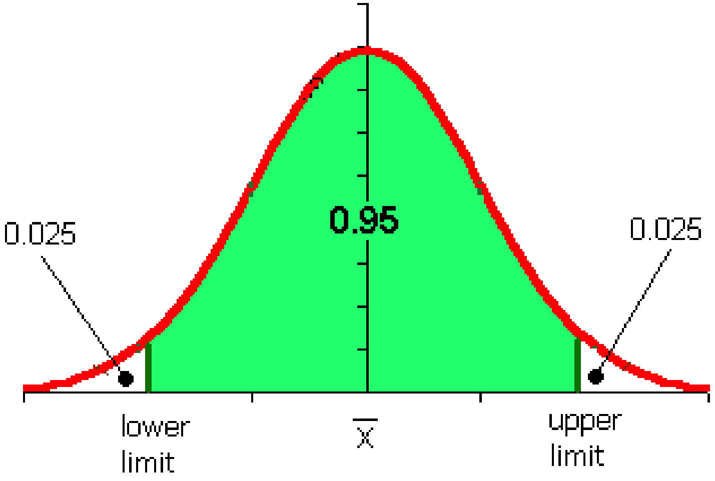
Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming
La Trobe University

Little is known about researchers' understanding of confidence intervals (CIs) and standard error (*SE*) bars. Authors of journal articles in psychology, behavioral neuroscience, and medicine were invited to visit a Web site where they adjusted a figure until they judged 2 means, with error bars, to be just statistically significantly different ($p < .05$). Results from 473 respondents suggest that many leading researchers have severe misconceptions about how error bars relate to statistical significance, do not adequately distinguish CIs and *SE* bars, and do not appreciate the importance of whether the 2 means are independent or come from a repeated measures design. Better guidelines for researchers and less ambiguous graphical conventions are needed before the advantages of CIs for research communication can be realized.

*Keywords:* confidence intervals, statistical cognition, standard error, error bars, statistical reform

# Confidence intervals

- Need to quantify the amount of **uncertainty** that is associated with our estimates of population parameters.
- Typically we want to say there is a 95% chance the true population mean lies within a certain window of values (e.g., between 105 and 114).
- There is a 95% chance that a normally-distributed quantity lies within ~2 (1.96) standard deviations of the true mean.



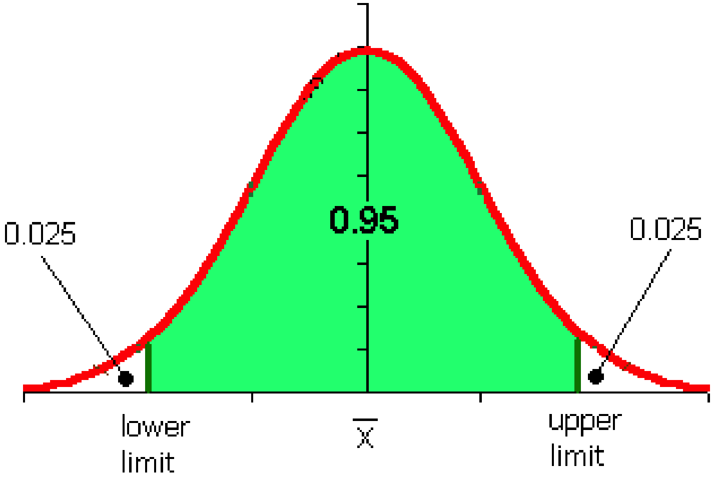
# Confidence intervals

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

- Range of values that has a 95% probability of containing the population mean is defined as :

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

- Upper and lower limit from normal distribution:
  - qnorm( p = c(.025, .975) )
  - [1] -1.959964 1.959964



- Actually need to rely on a t-distribution (because we don't know the true population SD
  - N <- 10000 # suppose our sample size is 10,000
  - > qt( p = .975, df = N-1)
  - [1] 1.960201
- Check this video out here

# How to interpret a confidence interval

- If we replicated the experiment over and over again and computed a 95% confidence interval for each replication, then 95% of those intervals would contain the true population mean.

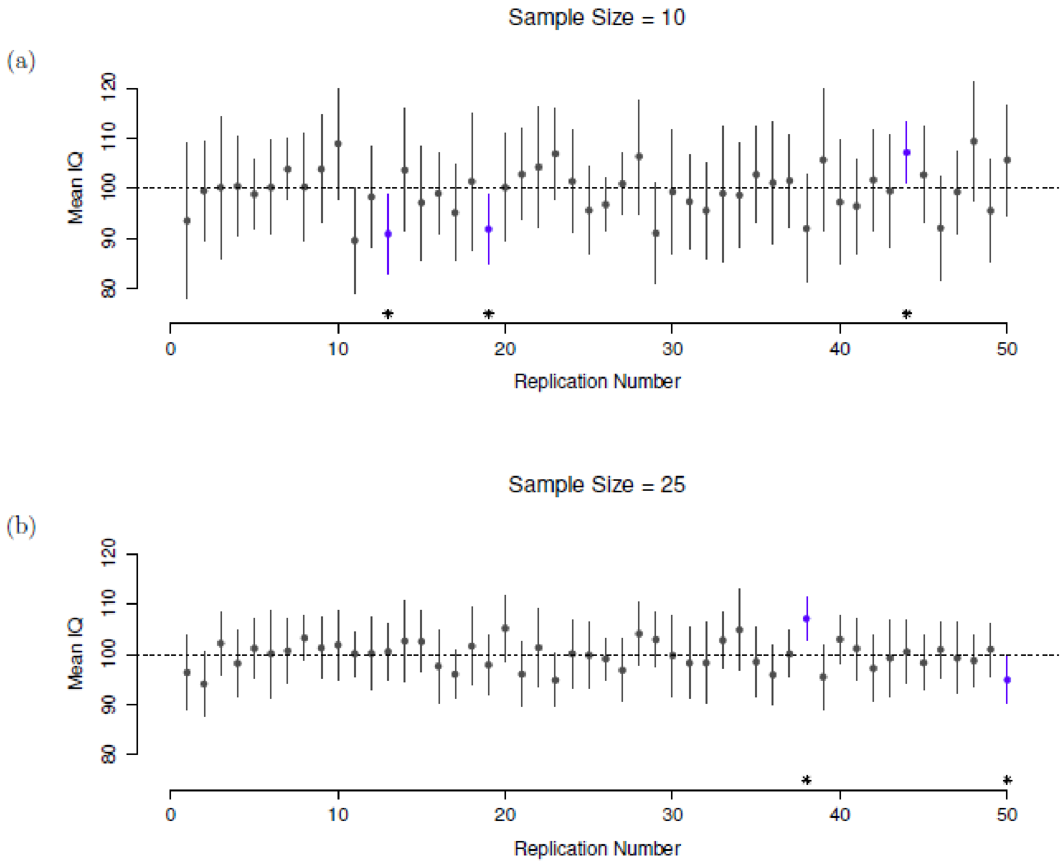


Figure 10.11: 95% confidence intervals. The top (panel a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean, and the line shows the 95% confidence interval. In total 47 of the 50 confidence intervals do contain the true mean (i.e., 100), but the three intervals marked with asterisks do not. The lower graph (panel b) shows a similar simulation, but this time we simulate replications of an experiment that measures the IQs of 25 people.

# Calculating the CI in R

- Install and load the lsr package
- Install and load the R datasets package and load the ‘trees’ and ‘discoveries’ data sets
- Use Help in R to find out what these datasets contain
- For each variable in the **trees** data set (Girth, Height, Volume)
  - Estimate the population mean and sd
  - Use the ciMean() function to find out 95% CI of the population mean
  - Use the ciMean() function to find out 99% CI of the population mean
  - How do these CIs compare?
- For the **discoveries** data set:
  - Do you think this is  sample or a population? Why?
  - Assume it is a sample and  estimate the population mean and sd
  - Use the ciMean() function to find out 95% CI of the population mean
  - Assume it is your population and draw a sample of various sizes (e.g., 5, 10, 30, 60), without replacement and see how the sample mean and SD compare to the original estimates you got.

# Summing Up

- Sampling Methods
- Sample statistics vs population parameters
- Estimating population parameters from sample statistics
- Confidence intervals

# Preparing for Module 3: Hypothesis Testing/Intro to Correlation

- Cumming & Calin-Jageman - CH 6
- *Additional reading (optional):*
- Navarro – CH 11

- No practical session this week!

# Thanks! See you next week! Questions?

TILBURG UNIVERSITY