

# **Week 4:**

# **Parameter estimation**

phase 2

# Key Ideas

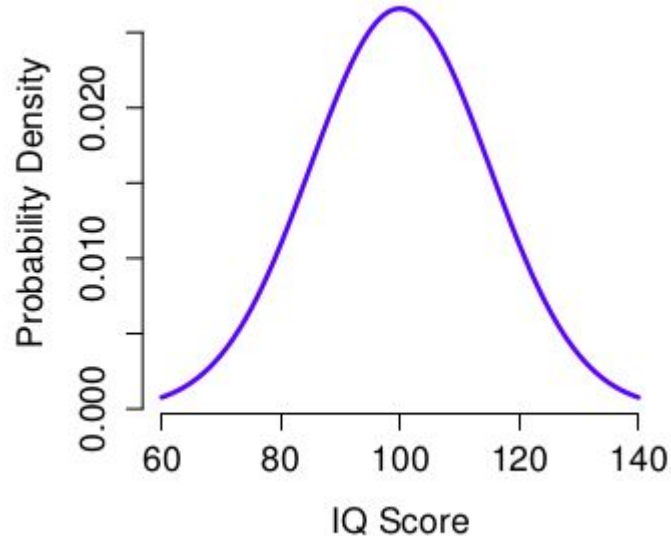
- We generally don't want to make claims about samples, but rather, do estimations about the **population**.
- We use **randomization** to ask what inferences our sample tells about the population
- We are always talking about **degrees of evidence**. Our estimations will never have total certainty.

# Samples vs Populations

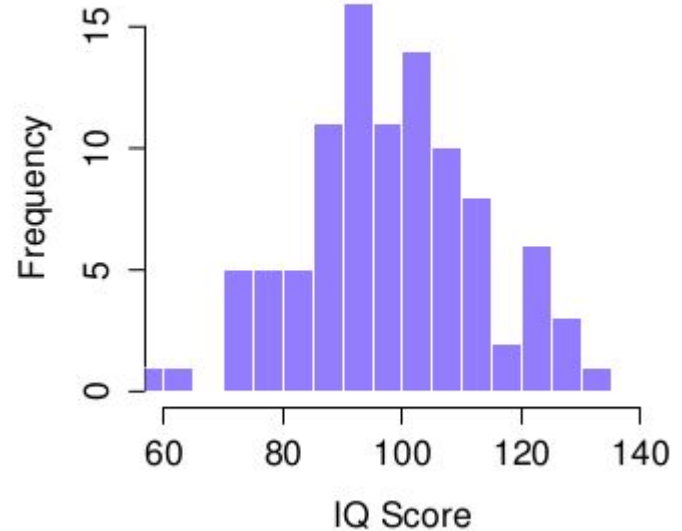
- A **population** is the entire group that you want to draw conclusions about.
  - The population depends on the study.
- 
- A **sample** is a part of the population that we actually examine (i.e., our data) to gather information.
  - The size of the sample is always less than the total size of the population.

# Samples vs Populations: Example

IQ population distribution:  
 $\mu=100$ ,  
 $\sigma = 15$

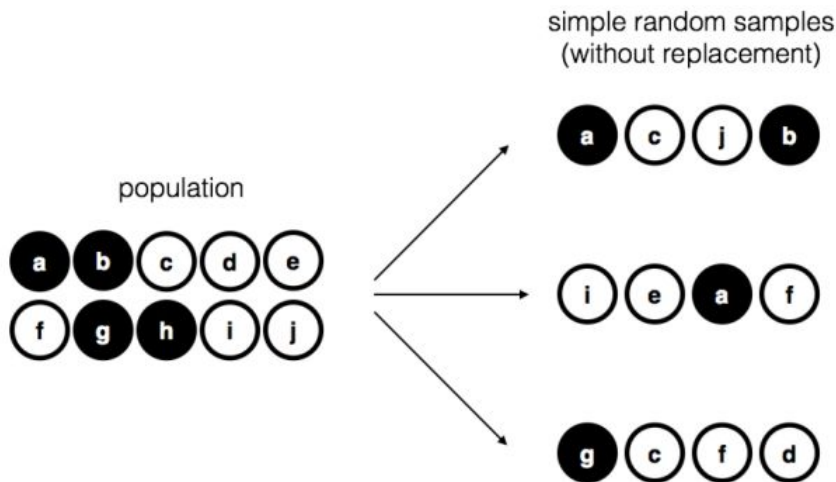


IQ sample distribution:  
 $\mu=98.5$ ,  
 $\sigma = 17$



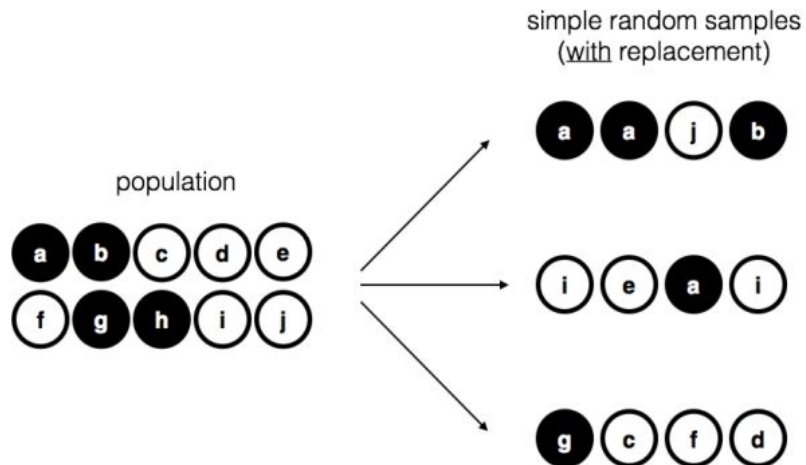
# Sampling

- The way in which we take samples from the population is called **sampling**.
- The simplest way of doing this is by taking a simple random sample.



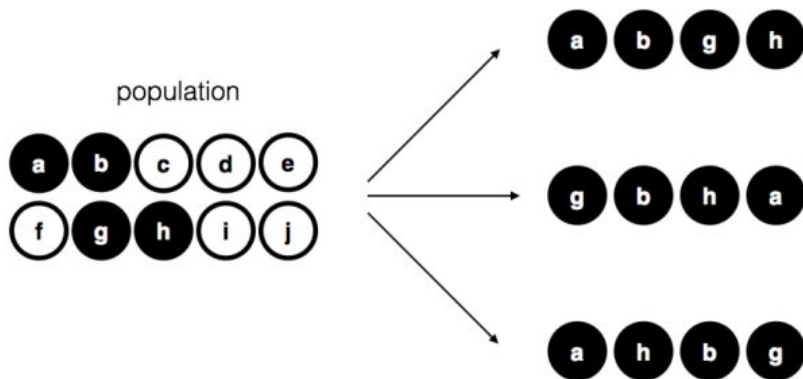
# Sampling

- The way in which we take samples from the population is called **sampling**.
- The simplest way of doing this is by taking a simple random sample.



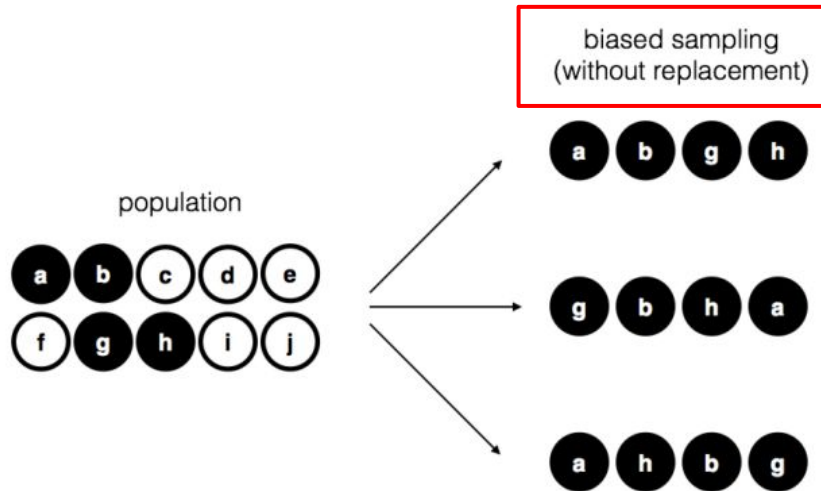
# Why random sampling?

Question: What is wrong with this kind of sampling?



# Why random sampling?

Question: What is wrong with this kind of sampling?





# Why random sampling?

- Random sampling circumvents this bias, since it gives all observations the same probability to be chosen.

# Why random sampling?

- Random sampling circumvents this bias, since it gives all observations the same probability to be chosen.
- Anyway, there are times in which incorporating prior knowledge of the population to the sampling procedure can be beneficial (e.g. stratified sampling)

# The law of large numbers

As the sample size increases, the sample mean tends to the population mean.

Why? We'll see this in the tutorial for this lesson!

# The law of large numbers

As the sample size increases, the sample mean tends to the population mean.

## Takeway:

Sample sizes in experiments are important!



# Sampling distributions

- A **sampling distribution** of **any** statistic (e.g. the mean, median, etc) shows how it would vary in identical repeated data collections.
- It answers the question: “What would happen if we did this experiment or sampling many times?”
- The **sampling distribution of the sample mean** is very useful because it can tell us the probability of getting any specific mean from a random sample.

# Sampling distributions

[https://onlinestatbook.com/stat\\_sim/sampling\\_dist/](https://onlinestatbook.com/stat_sim/sampling_dist/)

# The Central Limit Theorem

1. The mean of the sampling distribution of the mean ( $\mu_{<x>}$ ) is equal to the mean of the population ( $\mu$ )

$$\mu = \mu_{<x>}$$

2. The standard deviation of the sampling mean  $\sigma_{<x>}$  (also called the standard error) gets smaller as the sample size  $N$  increases

$$\text{SEM} \equiv \sigma_{<x>} = \sigma/\sqrt{N}$$

3. The shape of the sampling distribution of the mean becomes gaussian as the sample size increases, **no matter the population distribution.**  
(wait, really?? Yes → tutorial and assignments!)

# The Central Limit Theorem: Why is important?

- Most of the measured quantities in real life involve averages (e.g. IQ).



# The Central Limit Theorem: Why is important?

- Most of the measured quantities in real life involve averages (e.g. IQ).
- Doing statistical inference using gaussian distributions is lot easier!

# The Central Limit Theorem: Why is important?

- Most of the measured quantities in real life involve averages (e.g. IQ).
- Doing statistical inference using gaussian distributions is lot easier!
- We want to have **large experiments**, as they are more reliable than small ones (Related: they tend to be more powerful; wait for next week's lecture).

# Estimating population parameters

- Why do we sample at the end of the day? → To estimate about the population!

# Estimating population parameters

- Why do we sample at the end of the day? → To estimate about the population!
- **The estimate of the population mean is just the sample mean.** It's the best guess that we can make!

$$\hat{\mu} = \langle X \rangle = \frac{1}{N} \sum_i^N X_i$$

# Estimating population parameters

- Why do we sample at the end of the day? → To estimate about the population!

- **The estimate of the population mean is just the sample mean.** It's the best guess that we can make!

$$\hat{\mu} = \langle X \rangle = \frac{1}{N} \sum_i^N X_i$$

- The standard deviation estimation is (almost) similarly computed from the sample standard deviation.

$$\hat{\sigma} = \sqrt{\frac{\sum_i^N (X_i - \langle X \rangle)^2}{N - 1}}$$

# Estimating population parameters

- Why do we sample at the end of the day? → To estimate about the population!

- **The estimate of the population mean is just the sample mean.** It's the best guess that we can make!

$$\hat{\mu} = \langle X \rangle = \frac{1}{N} \sum_i^N X_i$$

- The standard deviation estimation is (almost) similarly computed from the sample standard deviation.

$$\hat{\sigma} = \sqrt{\frac{\sum_i^N (X_i - \langle X \rangle)^2}{N - 1}}$$

Question: Why N-1?  
(Tutorial!)

# Estimating with confidence

- Every time we sample from our population a different answer is obtained, i.e. estimates are never perfectly accurate.
- **Confidence intervals** quantifies the amount of uncertainty attached to (any) estimates.
- They are computed as a  $100 \cdot (1 - \alpha)\%$ , such that if we replicate the experiment many times and compute a  $100 \cdot (1 - \alpha)\%$  confidence interval for each replication, the  $100 \cdot (1 - \alpha)\%$  of those intervals would contain the true estimate.

## Example: Confidence intervals (CI) for the mean

- If data are **normally** distributed and the population variance  $\sigma^2$  is **known**:

$$< X > \pm z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

- If data are **normally** distributed and the population variance  $\sigma^2$  is **unknown**:

$$< X > \pm t_{N-1, \alpha/2} \frac{\hat{\sigma}}{\sqrt{N}}$$

- If data are **not normally** distributed:

$$< X > \pm t_{N-1, \alpha/2} \frac{\hat{\sigma}}{\sqrt{N}}$$



## Example: Confidence intervals (CI) for the mean

- If data are **normally** distributed and the population variance  $\sigma^2$  is **known**:

$$\langle X \rangle \pm z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

In R:  
*qnorm* ( $\alpha/2$ ,  $\mu=0$ ,  $sd=1$ )

- If data are **normally** distributed and the population variance  $\sigma^2$  is **unknown**:

$$\langle X \rangle \pm t_{N-1, \alpha/2} \frac{\hat{\sigma}}{\sqrt{N}}$$

**QUANTILES!**  
(see previous week's slides)

- If data are **not normally** distributed:

$$\langle X \rangle \pm t_{N-1, \alpha/2} \frac{\hat{\sigma}}{\sqrt{N}}$$

In R:  
*qt*( $\alpha/2$ ,  $df=N-1$ )

# Example: Confidence intervals (CI) for the mean

- If data are **normally** distributed and the population variance  $\sigma^2$  is **known**:

$$\langle X \rangle \pm z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

- If data are **normally** distributed and the population variance  $\sigma^2$  is **unknown**:

$$\langle X \rangle \pm t_{N-1, \alpha/2} \frac{\hat{\sigma}}{\sqrt{N}}$$

- If data are **not normally** distributed:

$$\langle X \rangle \pm t_{N-1, \alpha/2} \frac{\hat{\sigma}}{\sqrt{N}}$$

Degrees of freedom in the t-distribution. ( $N$  is the sample size)

# Example: 95% CI for the sample mean

Assuming normality and known population variance  $\sigma$ :

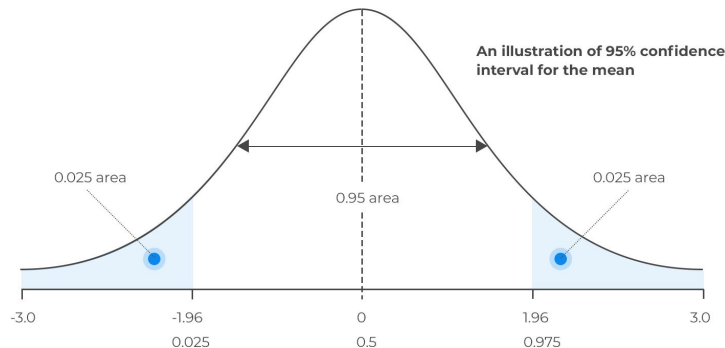
$$95\% \equiv 100 \times (1 - 0.05)\%$$

$$\rightarrow \alpha = 0.05$$

$$\rightarrow z_{0.05/2} \approx 1.96$$



95% Interval



$$\langle X \rangle - (1.96 \times SEM) \leq \mu \leq \langle X \rangle + (1.96 \times SEM)$$

# Recap

- We use **samples** to infer about the **population**.
- **Large** sample sizes are **important**: they provide more **precise** estimations, and concerning the mean, they allow us to work with **gaussian** distributions.
- We are always talking about **degrees of evidence**. Our estimations will usually be expressed within **confidence intervals**.