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 license about the population

- We are always talking about **degrees of evidence.** Our estimations will never have total certainty.

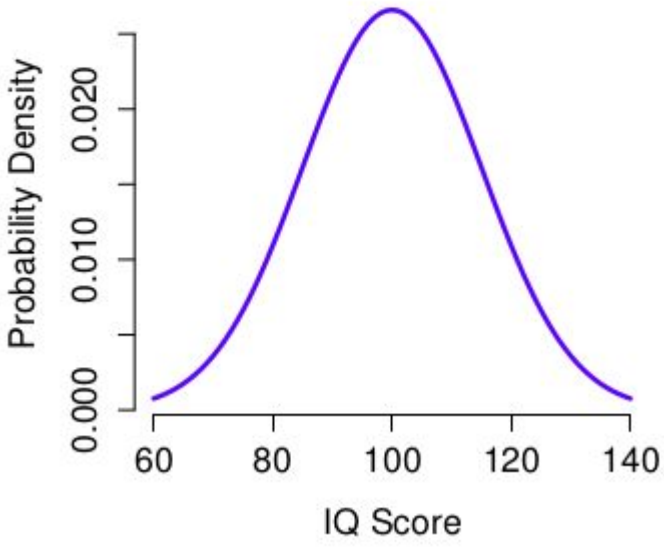
Samples vs Populations

- A **population** is the entire group that you want to draw conclusions about.
- The population depends on the study.

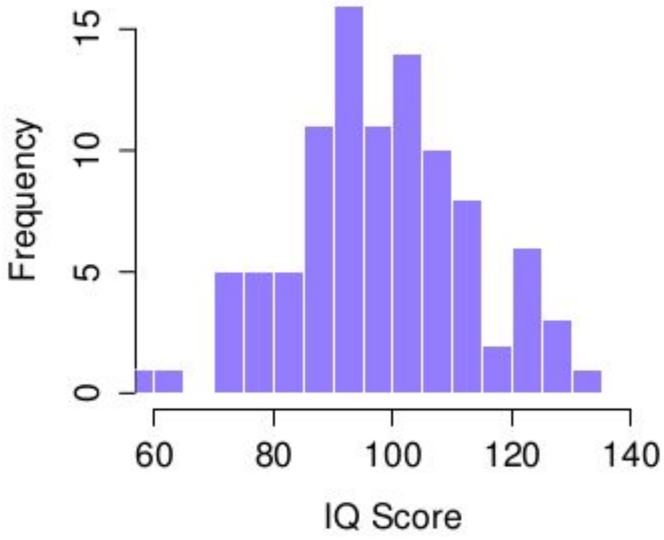
- A **sample** is a part of the population that we actually examine (i.e., our data) to gather information.
- The size of the sample is always less than the total size of the population.

Samples vs Populations: Example

IQ population distribution:
mu=100,
sigma = 15



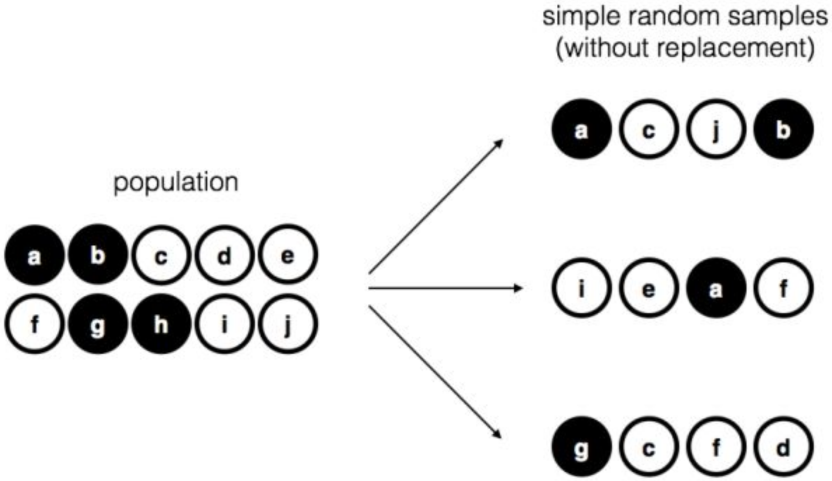
IQ sample distribution:
mu=98.5,
sigma = 17



Sampling

- The way in which we takes samples from the population is called **sampling**.

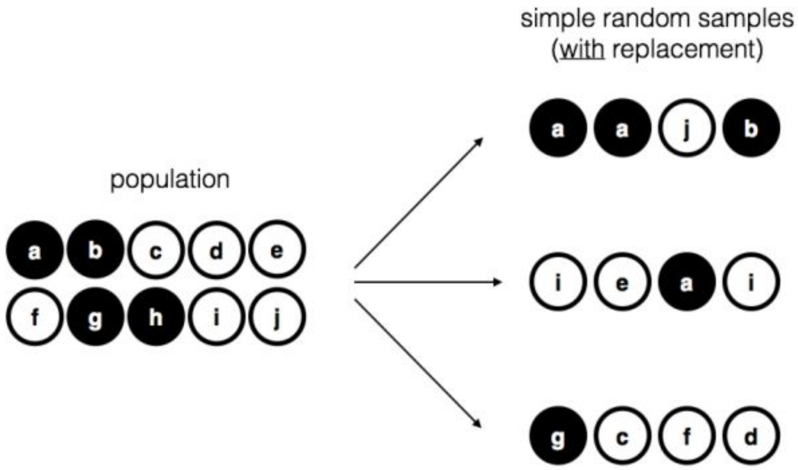
- The simplest way of doing this is by taking a simple random sample.



Sampling

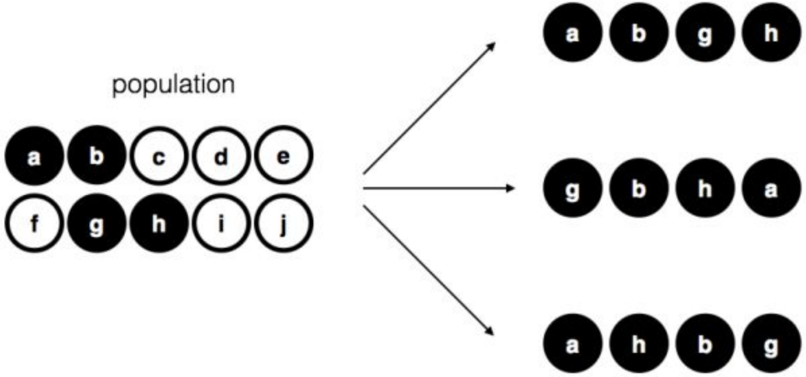
- The way in which we takes samples from the population is called **sampling**.

- The simplest way of doing this is by taking a simple random sample.



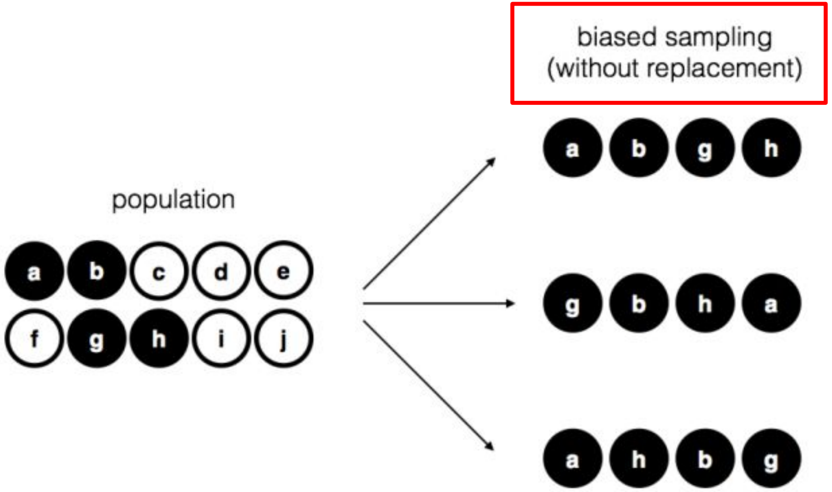
Why random sampling?

<u>Question</u>: What is wrong with this kind of sampling?



Why random sampling?

<u>Question</u>: 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.

<u>Takeway:</u>

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/

The Central Limit Theorem

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

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

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

$$SEM \equiv \sigma_{<X>} = \sigma/N$$

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!

$$\hat{\mu} = < X > = \frac{1}{N}\sum_{i}^{N} X_i$$

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!

$$\hat{\mu} = < X > = \frac{1}{N} \sum_{i}^{N} X_i$$

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

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

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!

$$\hat{\mu} = < X > = \frac{1}{N} \sum_{i}^{N} X_i$$

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

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

<u>Question</u>: Why N-1? (Assignments!)

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 $100*(1-\alpha)\%$, such that if we replicate the experiment many times and compute a $100*(1-\alpha)\%$ confidence interval for each replication, the $100*(1-\alpha)\%$ of those intervals would contain the true estimate.

Example: Confidence intervals (CI) for the mean

- If data are **normally** distributed and the population variance **σ** is **known**:

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

- If data are **normally** distributed and the population variance **σ** 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 **σ** is **known**:

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

In R: ***qnorm* *(mu=0, sd=1)***

- If data are **normally** distributed and the population variance **σ** is **unknown**:

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

QUANTILES!
(see previous week's slides)

In R: ***qt***

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

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

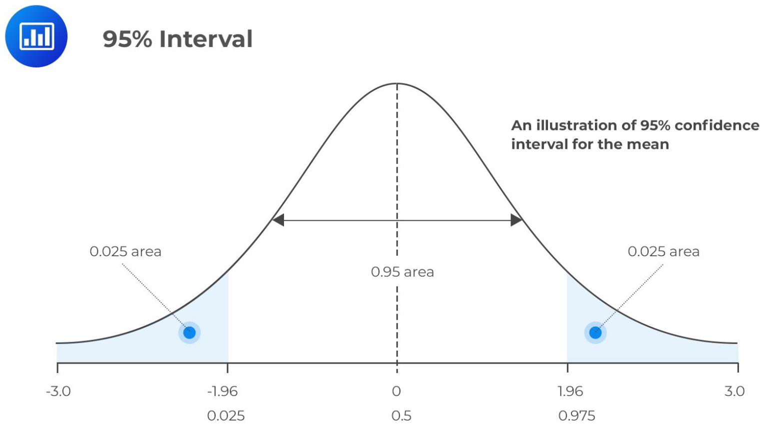
Example: 95% CI for the sample <u>mean</u>

Assuming normality and known population variance **σ:**

95% ≡ 100*(1-0.05)%

→ alpha = 0.05

→ $z_{0.05/2} \cong 1.96$



$$< X > -(1.96 \times SEM) \leq \mu \leq < X > +(1.96 \times SEM)$$

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**.