Module 7: Estimating unknown quantities from a sample

California State University - Northridge

Ovande Furtado, Jr., Ph.D.

March 5, 2022

CSUN | CALIFORNIA STATE UNIVERSITY NORTHRIDGE

```
xaringanExtra::use_webcam()
```

Samples, populations and sampling

- Sampling theory plays role in specifying the assumptions upon which your statistical inferences rely
- We’re drawing inferences from (the sample) and what it is that we’re drawing inferences about (the population).
- We can’t possibly get every person in the world to do our experiment

Defining a population

It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about and is generally much bigger than the sample.[1]

Population is more abstract than the sample

- Not always clear to the researcher what the population is
- To circumvent this issue, researchers often define a target population and a study population
  - target pop. --> see definition above
  - study pop. --> a subset of the target population from which the sample is actually selected[2]

Example

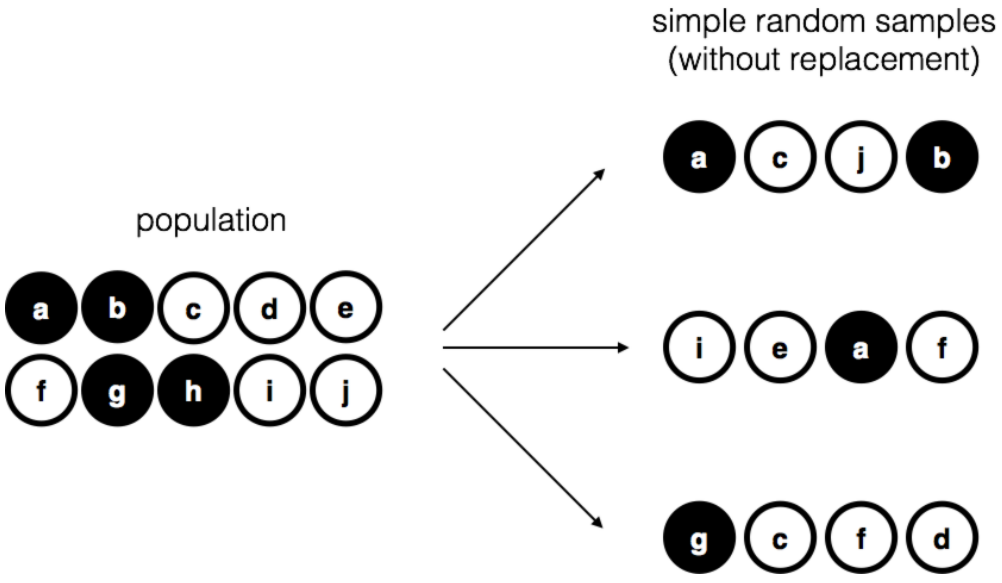
- Survey undergraduate students on their perception regarding physical activity
- target pop. --> undergraduate students all over the world
- study pop. --> undergraduate students at CSUN

[1] Navarro and Foxcroft (2019)

[2] Hu, S. (2014). Study Population. Encyclopedia of Quality of Life and Well-Being Research. Springer. doi: 10.1007/978-94-007-0753-5

Simple random samples - with no replacement

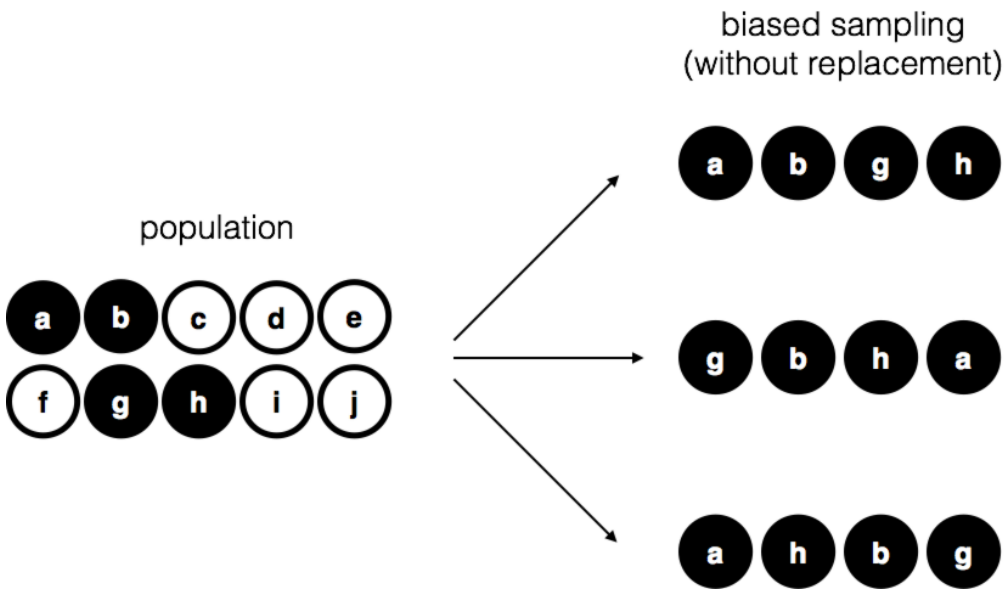
A procedure in which every member of the population has the same chance of being selected is called a simple random sample.



The fact that we did not put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled without replacement.

Simple random samples - biased with no replacement

A procedure in which every member of the population has the same chance of being selected is called a simple random sample.



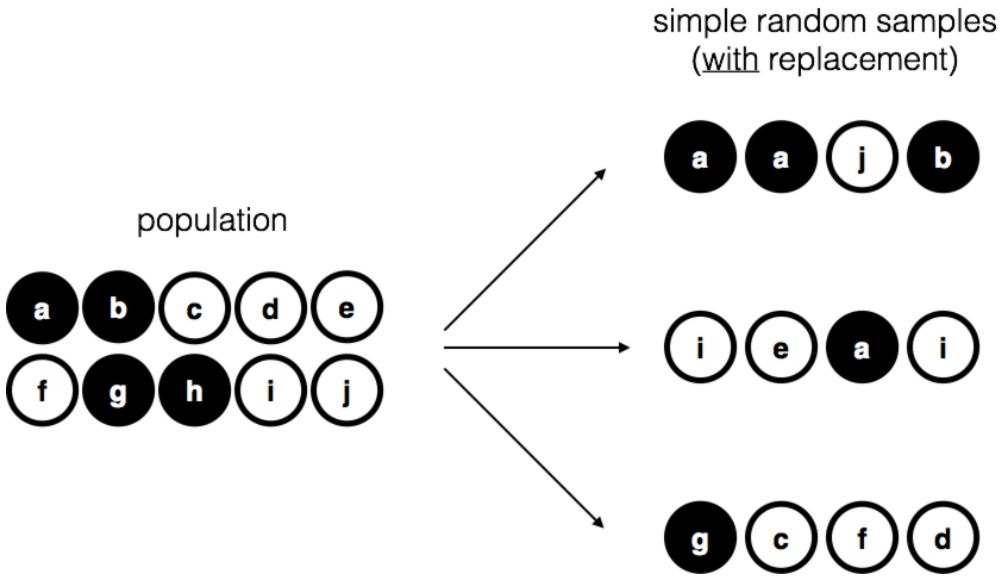
Only black chips were picked

If you know that the sampling scheme is biased to select only black chips then a sample that consists of only black chips doesn't tell you very much about the population!

Example: The study population is **undergraduate students** at CSUN, but you decided to collect data (interview students) at the Rec Center only.

Simple random samples - with replacement

A procedure in which every member of the population has the same chance of being selected is called a simple random sample.



Data sets generated in this way are still simple random samples, but because we put the chips back in the bag immediately after drawing them it is referred to as a sample with replacement.[1]

[1] Experiments in kinesiology tend to be sampling without replacement, because the same person is not allowed to participate in the experiment twice.

Sampling types

A full discussion on the types samples is beyond the scope of this presentation, but you should now that:

"The generalizability of clinical research findings is based on multiple factors related to the internal and external validity of the research methods. The main methodological issue that influences the generalizability of clinical research findings is the sampling method."[1]

- Probability sampling
  - simple random, stratified random, systematic, clustered random
- Non-probability sampling
  - convenience, judgmental, snow-ball

Refer to Elfil and Negida (2017)[1] for a full review.

[1] Elfil, M., & Negida, A. (2017). Sampling methods in Clinical Research; an Educational Review. Emergency, 5(1). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325924

What if I don't have a random sample?

- it can matter if your data are not a simple random sample
- there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this book!)
- Convenience sample - biased
    - A bias in your sampling method is only a problem if it causes you to draw the wrong conclusions.

The Central Limit Theorem (CLT)

In probability theory, the central limit theorem (CLT) establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed. [1]

The CLT states that as N becomes large, the distribution of the **sample means** becomes closer and closer to a normal distribution with mean μ and diminishing standard deviation.

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \longrightarrow Z \sim N(0,1) \tag{1}$$

[...] the result above is independent of the underlying distribution of the random variable X^2.

[1] Contributors to Wikimedia projects. (2022, February 23). Central limit theorem - Wikipedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Central_limit_theorem&oldid=1073557211

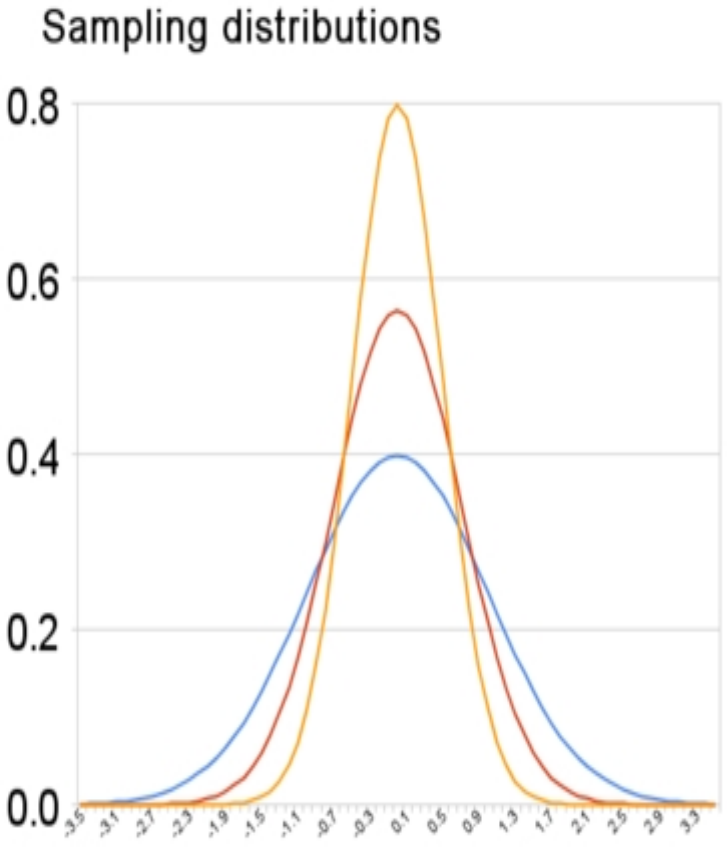
[2] Chapter 8 Sampling Distributions - Statistics. (2022, March 07). Retrieved from http://statistics.wikidot.com/ch8

The Central Limit Theorem (CLT)

The mean remains unchanged and as n increases, the standard deviation of of the sample means $\overline{X}_n$ diminishes as 1/ $\sqrt{n}$

- this means that the weight gets concentrated on the center around the mean

if the blue line represents the case of the original sample size n, the red is then 2n and the yellow is when the sample is 4 times its original size.



The Central Limit Theorem (CLT)

We can see the CLT in action by playing with an online Java applet

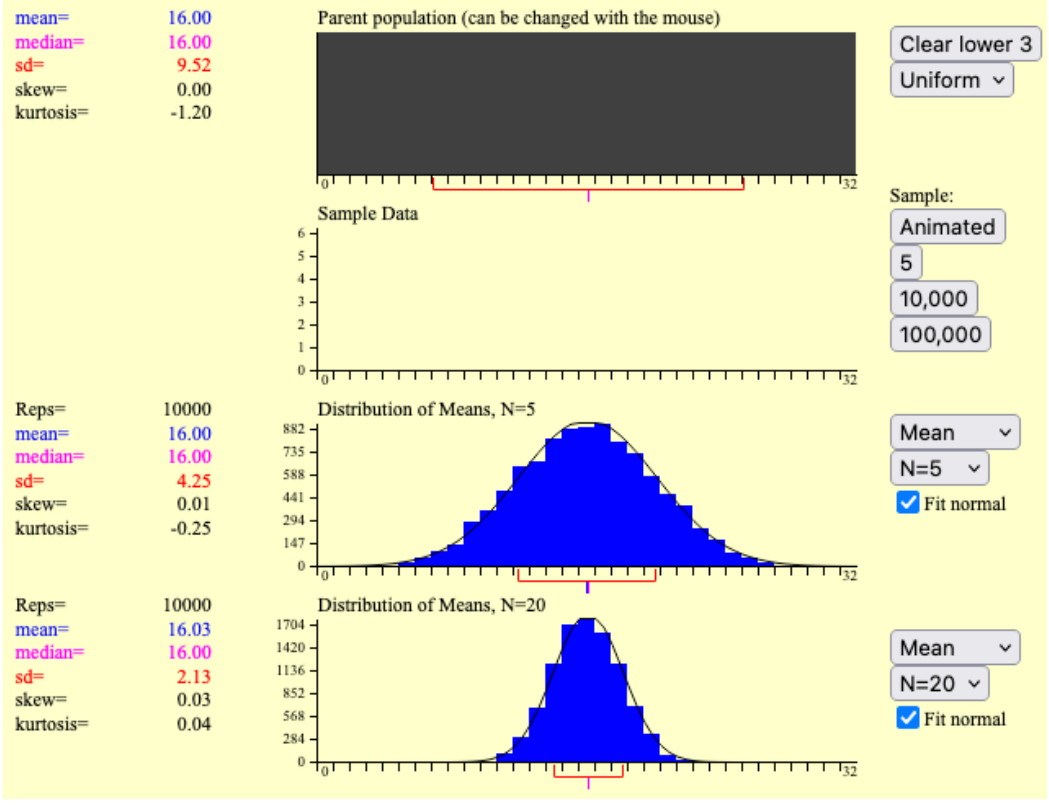
- Visit the this link online
- Click on *Begin* - top left of the site to get the Java applet running
- Set the distribution on the top to be uniform
- in the 3rd and 4th panel the mean and N=5 and N=20 respectively
- Check **Fit normal** on both panels
- Click on animate in the second panel once or twice, then on the number of repetitions 5, 10000, and 100000

The Central Limit Theorem (CLT)

Observe how the mean becomes closer and closer to a normal distribution.

Also notice that for n = 5, the standard deviation is about twice as large than for n = 20, as the number of repetitions increases.[1]

No matter which distribution[2] we take, if we draw n random (large enough) samples, the mean of the samples will be normally distributed.



[1] Chapter 8 Sampling Distributions - Statistics. (2022, March 07). Retrieved from http://statistics.wikidot.com/ch8

[2] Even not normal (i.e., badly skewed) - give it a try with the java applet introduced in the previous slide.

The Central Limit Theorem (CLT)

On the basis of the sample means distributions (previous slide), it seems like we have evidence for all of the following claims about the sampling distribution of the mean.

- The mean of the sampling distribution is the same as the mean of the population
- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases
- The shape of the sampling distribution becomes normal as the sample size increases

The statements above are true and the CLT proves it.

The Central Limit Theorem (CLT)

As per the CLT,

if the population distribution has mean μ and standard deviation σ, then the **sampling distribution** of the mean also has mean μ and the **standard error of the mean** is

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

Because we divide the population standard deviation (σ) by the square root of the sample size (n), the SE_M gets smaller as the sample size increases.

It also tells us that the shape of the sampling distribution becomes normal.

Predicting populaiton parameters

- Definitions
  - Value of Statistic: characteristics related to a sample
  - Value of Parameter: characteristics related to a population
- Measuring an entire population is rarely possible.
- Researchers take a sample from the population and assume that it is representative (must be random).
- Inferential statistics = the process of estimating population parameters based on sample statistics.

Sampling error

- Regardless of how the sample was collected (target or study population), sampling error will occur
- Sampling error = amount of error in the estimate of a population parameter that is based on a sample statistic.
- Even if a sample is randomly drawn, the population mean would still unknown
- Never know the true population mean since not all members are measured
    - Need a way to determine how accurate the sample mean is and what are the odds that it is deviant from the population mean by a given amount

Estimating Sampling Error

- Standard Error: index of how variable the sample statistic is when multiple samples of the same size are drawn from the same population.
- Use ONE sample to estimate the STANDARD ERROR
- If I know the estimated standard error, I can estimate how much difference there is between my sample statistic and the population parameter (i.e., Mean, SD, Corr. Coefficient)

Estimating Sampling Error

The standard error is a general term, SE_M is **specific** and applies to the sample mean (x)

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The SE_M is:

- a numeric value that indicates the amount of error that may occur when a random sample mean is used as a predictor of a population mean

The **population mean** (μ) is assumed to exist between some set limits and the chance of this assumption being correct is stated as odds such as:

90 to10 (p = .10); 95 to 5 (p = .05); 99 to 1 (p = .01)

Estimating Sampling Error ($\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$)

Example: Sit-and-reach for HS girls

Estimate where the mean of the population lies (within certain limits)

Desc. Stats: n = 50; $\bar{x}$ = 35 cm; s = 10 cm

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where:

$\sigma_{\bar{x}}$ refers standard deviation of the sample means[1] indicated by $\bar{x}$.

n refers to the sample size.

s refers to the sample standard deviation.

Working this out we have:

$\sigma_{\bar{x}}$ = 1.4 cm^2

[1] Recall that it is impossible to collect data on 10000 people. So we try to estimate the average of the sample means.

[2] Can be interpreted as any other standard deviation on a normal curve

Estimating Sampling Error ($\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$)

Making inference using confidence intervals (CI)]

$$\mu = \bar{x} \pm 1 \frac{s}{\sqrt{n}}$$

where:

`(\bar{x})` refers for the sample mean.

`(s)` refers to the sample standard deviation.

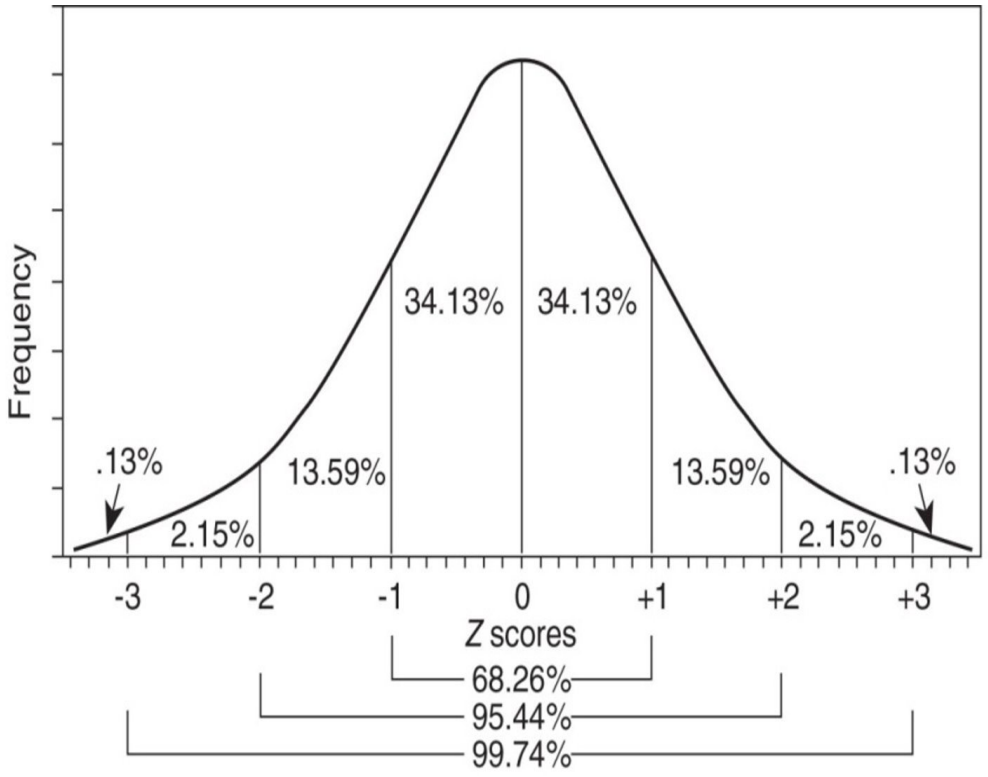
`(n)` refers to the sample size.

Solution: 35 ± (1 x 1.4)

33.6 and **36.4** or **33.6** ≤ μ ≤ **36.4** (limits)

- Accurate at the 68% level of confidence (LOC)
  - Because we used 1Z (SEM)
- 68% chance being correct
- 32% change being incorrect (probability of error - $p < .32$)

Estimating Sampling Error ($\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$)



Note that trying to be accurate at the 68% level of confidence (LOC) is too risky.

The 32% of being incorrect is not acceptable when making inferences about population parameters; i.e., μ, σ, etc.

Estimating Sampling Error ($\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$)

Making inference using confidence intervals (CI)

$$\mu = \bar{x} \pm 1 \frac{s}{\sqrt{n}}$$

Now, substitute **1** by **1.96**. By doing so, you accurately represent the 95% LOC or *p* = 0.05

Solution: 35 ± (**1.96** x 1.4)

33.2 and **37.7** or **32.2** ≤ μ ≤ **37.7** (limits)

- Range of values: Confidence Interval (CI)
  - The 68% CI is about 33.6 to 36.4
  - The 95% CI is about 32.3 to 37.7 cm
- If 99% LOC (p= .01)
  - use Z ± 2.58 (CI = 31.4 to 38.6 cm)

Estimating Sampling Error ($\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$)

The more precise (narrow) the estimate, the lower the odds of being correct

As the estimate become more general (broad), the odds of being correct improve

Solution: 35 ± (**1.96** x 1.4)

33.2 and **37.7** or **32.2** ≤ μ ≤ **37.7** (limits)

- Range of values: Confidence Interval (CI)
  - The 68% CI is about 33.6 to 36.4
  - The 95% CI is about 32.3 to 37.7 cm
- If 99% LOC (p= .01)
  - use Z ± 2.58 (CI = 31.4 to 38.6 cm)

Estimating Sampling Error ($\sigma_{\overline{x}} = \frac{s}{\sqrt{n}}$)

Your Turn!!!

You goal is to find the average height of 2nd graders in a school district (μ). Random selection was used and data were collected for a sample comprised of 60 participants:

$\bar{x}$= 100 cm; s= 10 cm; n= 60

Using the information above, estimate the population mean at the 95% LOC, *p* = 0.05

LOC, Probability Error, and Confidence Intervals

Summary

- Level of Confidence (LOC): percentage figure that establishes the probability that a statement is correct
  - Derived from s (e.g. +/-1 s = 68%)
- Probability of Error = percentage figure that establishes the probability that a statement is incorrect
  - If 68% correct then 32% incorrect ($p < .32$)
  - The area under the normal curve representing error is called alpha (α)
  - 95% LOC or $p < .05$ is more common in Kinesiology research
- Confidence Interval (CI): the range of values associated with LOC

Thanks!

CSUN | CALIFORNIA STATE UNIVERSITY NORTHRIDGE

ovandef@csun.edu

@ofurtado

@drfurtado