Section 6: SAMPLING

Types of studies

Statistic: Mean, standard deviation, proportion, etc., for a sample

Parameter: Mean, standard deviation, proportion, etc., for a population. We use statistics to estimate corresponding parameter values.

Observational studies: In an observational study, we're just looking at the information that's already there, or measuring it in some way, but we're adding nothing to the population that will change it in any way.

Treatment: Something that changes a population

Correlation: Two variables are correlated when they move together predictably. The variables are positively correlated when they increase together or decrease together. Variables are negatively correlated when they increase and decrease in opposite directions: one goes down while the other goes up, or one goes up while the other goes down.

Causation: One variable causes another variable to change. Showing correlation doesn't prove causation.

Confounding variable: A third variable that leads to both of the variables that were correlated

Control group: The group that does nothing, receives nothing, or isn't manipulated

Treatment / experimental group: The group that does something, receives something, or is treated in some way

Explanatory and response variables: In an experiment, we're looking to see whether one or more explanatory variables (the treatment) has an effect on the response variable (whatever is expected to be effected).

Blind experiment: When the participants don't know whether they're in the control group or the treatment group

Double-blind experiment: When neither the participants nor the people administering the experiment know which group anyone is in

Blocking: When researchers separate participants into like groups

Matched-pairs experiment: A more specific kind of blocking where we make sure that the participants in our experimental group and control group are matched based on similar characteristics

Sampling and bias

Representative / unbiased sample: When the sample data "scales up" to the population, or when the sample does a good job representing the population

Bias: When something skews our results and makes them inaccurate

Measurement bias: When there's something wrong with the tool we're using to collect the data, so our method of collecting observations or responses from the sample results in false values

Social desirability bias: When our survey asks something in a way that discourages people from responding truthfully

Leading questions: Questions that are framed in a way that push respondents toward a particular response

Selection bias, undercoverage: When we don't collect data from an entire group of subjects that should have been included in our data

Voluntary response sampling: When people voluntarily respond to or participate in the study

Convenience Sampling: When we choose a sample simply because it's convenient, instead of prioritizing getting a good, random representative sample

Non-response bias: When we get a large number of people who don't respond to our survey

Simple random sample: When we assign subjects to groups in a totally random way

Stratified random sample: When we put some parameter on the sample where we require an even number of subjects from different groups

Clustered random sample: Where we break our population into clusters, and then either

1) take a random sample within each cluster to be our total sample, or
2) randomly pick some clusters an then sample everyone in those clusters

Systematic sampling: When we assign numbers to individuals in a population and choose them at some specified interval

Sampling distribution of the sample mean

Sampling with replacement: When we pick a random sample, and then "put that sample back" into the population before choosing another sample

Sample distribution of the sample mean (SDSM): The probability distribution of all possible sample means for a certain sample size n

The Central Limit Theorem (CLT): Even if a population distribution is non-normal, as long as we use a large enough sample ($n \geq 30$), then we can make inferences using our sample statistics, because of the fact that the SDSM will be a normal distribution.

Mean, variance, and standard deviation of the SDSM:

Mean:

$$\mu_{\bar{x}} = \mu$$

Variance:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \approx \frac{s^2}{n}$$

Standard deviation (standard error):

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

Finite population correction factor (FPC): If we're sampling without replacement or sampling from more than %5 of a finite population, we should multiply standard deviation by $\sqrt{(N-n)/(N-1)}$, and multiply variance by $(N-n)/(N-1)$

Conditions for inference with the SDSM

Conditions for inference: In order to use the sampling distribution of the sample mean, we need to meet the random, normal, and independent conditions. We meet the normal condition if the original population is normal, and/or when $n \geq 30$. We meet the independent condition if we sample with replacement, and/or if we keep our sample size to %10 or less of the total population.

Sampling distribution of the sample proportion

Population proportion: The number of subjects in our population that meet a certain condition

Sample proportion: The proportion of subjects in our sample that meet a certain condition

$$\hat{p} = \frac{x}{n}$$

Sampling distribution of the sample proportion (SDSP): The probability distribution of all possible sample proportions for a certain sample size n

Mean, variance, and standard deviation of the SDSP:

Mean: $\mu_{\hat{p}} = p$

Variance: $\sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$

Standard deviation (standard error): $SE = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

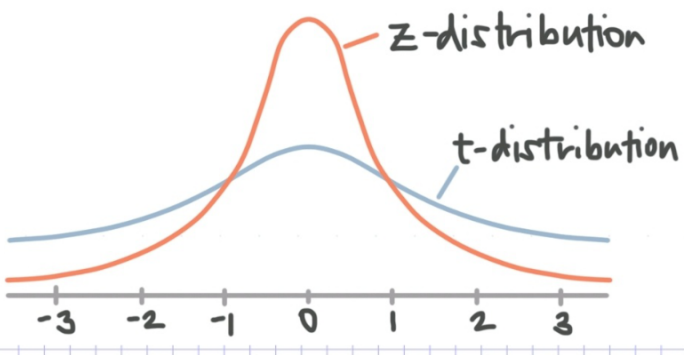
Conditions for inference with the SDSP

Conditions for inference: In order to use the sampling distribution of the sample proportion, we need to meet the random, normal, and independent conditions. We meet the normal condition when $np \geq 5$ and $n(1-p) \geq 5$. We meet the independent condition if we sample with replacement, and/or if we keep our sample size to 10% or less of the total population.

The student's t-distribution

Standard normal distribution, z-distribution: The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$

Student's t-distribution: Similar to the standard normal distribution in the sense that it's symmetrical, bell-shaped, and centered around the mean $\mu = 0$, but flatter and wider. The exact shape depends on the number of degrees of freedom, $df = n - 1$



Confidence interval for the mean

Point estimate: An estimator for a singular value, like the sample mean for the population mean, or the sample standard deviation for the population standard deviation

Interval estimate: A range of values that estimate the interval in with some parameter may lie

Confidence level: The probability that an interval estimate will include the population parameter

Alpha value, α, level of significance, probability of making a Type I error: $\alpha = 1 -$ confidence level

Region of rejection: The region outside the confidence interval, in the tail(s) of the probability distribution. If the z-value we calculate falls in this region, we will reject the null hypothesis.

Confidence interval:

When σ is known: $(a,b) = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$

When the FPC applied: $(a,b) = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

When σ is unknown and/or the sample size is small: $(a,b) = \bar{x} \pm t\frac{s}{\sqrt{n}}$

Required sample size for a fixed margin of error:

$$n = \left(\frac{z \cdot \sigma}{ME}\right)^2$$

Confidence interval for proportion

Confidence interval for proportion:

$$(a,b) = \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$