POLS 203 DS 7

Midterm Recap

Alp Ecevit

hours of studying → score at exam

Independent Variable	Dependent Variable
(Cause)	(Effect)
(Explanatory Variable)	(Response Variable)

- ✓ If a study has many **confounding** variables, that study has low internal validity.
- ✓ **Internal validity:** Is the extent to which you can be confident that a cause-and-effect relationship established in a study cannot be explained by other factors. How good your IV measures your DV.
- ✓ **Confounding:** The situation in which a change in your DV is resulted by some third variable
- ✓ **External validity:** It is about how applicable a research is to the real world. How the results can be applied to other situations than your research.

- ✓ ***Parameter:** A numerical study of the population
- ✓ ***Statistic:** A numerical summary of the sample data

- Measurement scale of variables
  - ✓ **Ordinal scale:** strongly agree, agree… (variables in an ordered fashion)
  - ✓ **Nominal scale:** hair color
  - ✓ **Interval scale:** income = [2500,3000] (quantitative)

- Variable types
  - ✓ **Quantitative variables:** Measures difference in quantity
  - ✓ **Categorical variables:** Measures differences in quality

- Discrete and Continuous Variables
  - ✓ A variable is **discrete** if its possible values form a set of separate numbers like 1,2,3. For example, the money in your bank account
  - ✓ A variable is **continuous** if it can take an infinite continuum of possible real number values. If you try to count these kind of variables, it will take forever. e.g., stars in the universe.

- Methods of sampling
  - ✓ **Systematic random sampling:** n = sample size, N = population. N/n, population divided by sample size. Select a random number from the first x number of subjects and then select every xth subject listed after x. 2,22,42,62…

  - ✓ **Stratified random sampling:** Divides the population into separate groups called strata and select a simple random sample from each stratum. Stratifying according to the political party identification is an example.

  - ✓ **Cluster sampling:** Divide the population into a large number of clusters, such as city blocks or countries (heterogeneousity is an important concept here). Select a simple random sample of the clusters. Use these subjects as the sample. Think of the New York City and its blocks. You divide the city blocks into clusters to sample 2% of the families in the city. You select simple random samples from the clusters and sample every family on each block.

  - ✓ **What's the difference between a stratified sample and a cluster sample?**
  - A stratified sample uses *every* stratum. The strata are usually groups we want to compare. By contrast, a cluster sample uses a *sample* of the clusters, rather than all of them. In cluster sampling, clusters are merely ways of easily identifying groups of subjects. The goal is not to compare the clusters but rather to use them to obtain a sample.

DESCRIPTIVE STATISTICS

- Bell shaped distributions
  - ✓ Mean = 0, sd = 1. Symmetric distributions. Very rare and hard to achieve.

- Mean

Sum of the observations divided by the number of observations. Often called the *average.*

  - ✓ R → mean(), summary(),

- Median
  - ✓ For symmetric distributions such as the bell shaped distribution, mean = median. For *right skewed distribution*, mean > median and for *left skewed distribution,* mean < median.
  - ✓ R → median(), summary()

- Median vs. Mean
  - ✓ The median is usually more appropriate than the mean when the distribution is highly skewed. The mean can be affected greatly by outliers, whereas the median is not.
  - ✓ If symmetric, you can use either because mean and median will be equal.

- Mode
  - ✓ It is the value that occurs most *frequently*.
  - ✓ The mode is appropriate for all kinds of data. Interval, nominal, ordinal.
  - ✓ A frequency distribution is called *bimodal* if two distinct mounds occur in the distribution.
  - ✓ In symmetric distributions, such as the bell shaped distribution, mean = median = mode.

- Standard deviation

The ***standard deviation*** s of n observations is

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size} - 1}}.$$

This is the positive square root of the ***variance*** s^2, which is

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}.$$

Example: Suppose that our sample is A = 1, 2, 2, 4, 1. Calculate standard deviation.

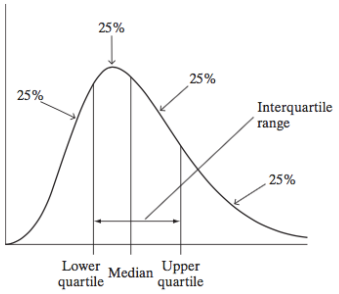
$1 + 2 + 2 + 4 + 1 = 10/5 = 2$

$((1-2)^2 + (2-2)^2 + (2-2)^2 + (4-2)^2 + (1-2)^2) / 5-1 = 1.5$

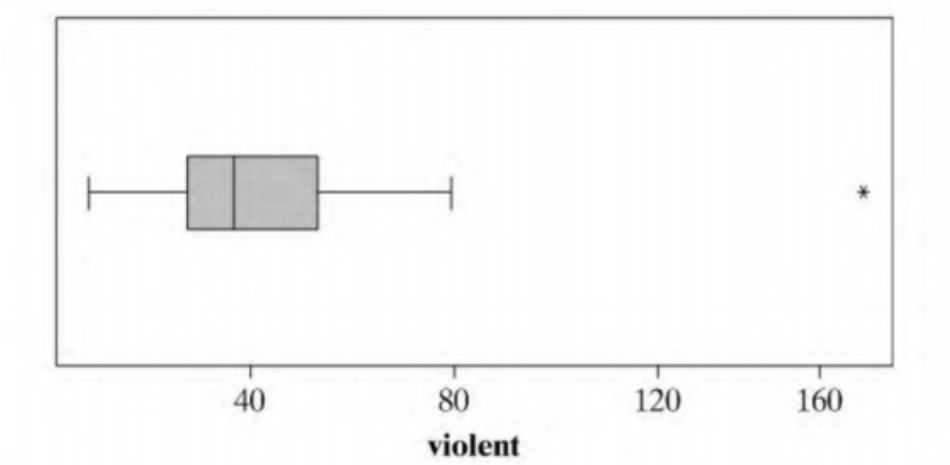
$\sqrt{1.5} = 1.22$

#or simply use sd() in R

- Percentile
  - ✓ The pth percentile is the point such that p% of the observations fall below or at that point and (100-p)% for above it.
  - ✓ Substituting p = 50 gives the 50th percentile. This is the *median*. In proportion terms, a percentile is called a *quantile.* The 50th percentile is the 0.50 quantile.
  - ✓ The 25th percentile is called the *lower quartile.*
  - ✓ The 75th percentile is called the *upper quartile.*
  - ✓ One quarter of the data fall below the lower quartile. One quarter fall above the upper quartile.
  - ✓ The lower quartile is the median for the observations that fall below the median, that is, the bottom half of the data. The upper quartile is the median for the observations that fall above the median, that is, the upper half of the data.
  - ✓ R → summary()



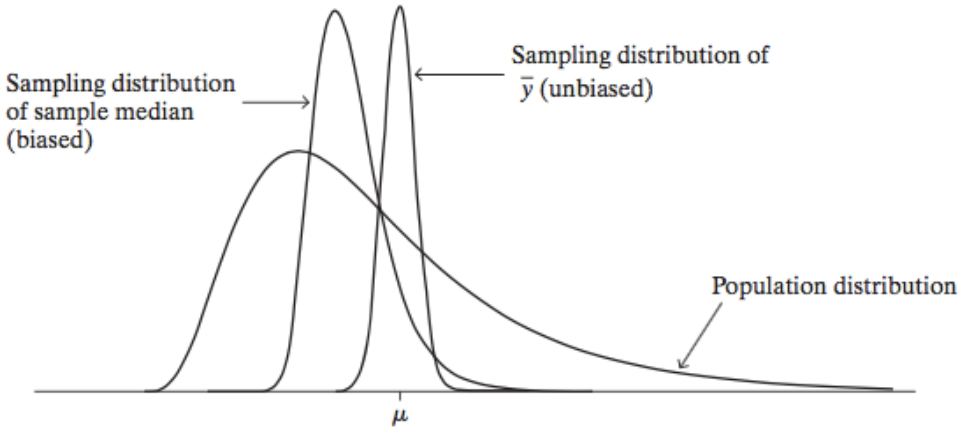
- Box plots



  - ✓ Box plot summarizes center and variability.
  - ✓ Box contains 50% of observations.
  - ✓ Line in the box denotes median.
  - ✓ Beginning and the end of the graph show maximum and minimum values.
  - ✓ Left whisker represents the bottom 25% of the data.
  - ✓ Right whisker represents the upper 75% of the data.
  - ✓ Star on the far right represents an outlier value.
  - ✓ This is a left skewed distribution. Remember mean tends to be pulled to the direction of the longer tail. More observations of lower value and some outliers of larger values.
  - ✓ R → boxplot(), ggplot(……) + geom_boxplot()

ESTIMATION

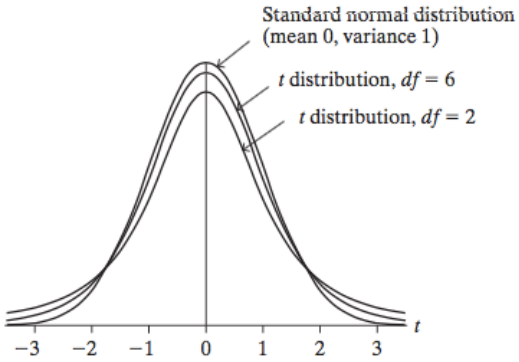
- It should be clear that when you see the word "estimation" you should think about the confidence interval
- A good estimator has **a small standard error** and it is **unbiased.**
- An estimator is unbiased if its sampling distribution is centers around the parameter.
- Maximum likelihood method of estimator (MLE). MLE has 3 properties:
  - ✓ Efficient → small standard error
  - ✓ Little or no bias when the sample size increases
  - ✓ Approximately a normal sampling distribution



- CONFIDENCE INTERVAL (CI)
  - ✓ CI = Point estimate ± margin of error
  - ✓ How many percentage points our estimate differs from the observation.
  - **CI for categorical data**
    - ✓ The proportion of Americans who lack health insurance.
    - ✓ The proportion of Canadians who favor independent status for Quebec.
    - ✓ The proportion of Australian young adults who have taken a "gap year," that is, a break of a year between high school and college or between college and regular employment. $\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi(1-\pi)}{n}}$.
    - ✓ Here, π denotes the population proportion. Its estimator is the sample proportion. $\hat{\pi}$ is the sample proportion. The formula above gives you the **standard error** of your sample; however, the probability is 0.05 that $\hat{\pi}$ is such that $\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$ does *not* contain π .
    - ✓ As your sample size increases, the standard error gets smaller. The sample comes closer to the population.
    - ✓ Formula of CI for categorical data is:
      - ➢ $\hat{\pi} \pm$ se*1.96 (95% of the normal distribution)
      - ➢ The interval $\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$ is an interval estimate for π with confidence level 0.95. It is called a ***95% confidence interval***. In practice, the value of the standard error $\sigma_{\hat{\pi}} = \pi\ (1 - \pi\ )/n$ for this formula is unknown, because it depends on the unknown parameter π. So, we estimate this standard error by,

$$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

- **CI for a mean**
  - ✓ Point estimate ± margin of error
  - ✓ Standard error = s/√n
  - ✓ Here, s is our sample standard error and n is our sample size
- The larger sample size, the smaller margin of error and the confidence interval will be narrower.
- **t-distribution**
  - ✓ t-distributions are a family of distributions that look almost identical to the normal distribution but they are more spread out and have thicker tails.
  - ✓ They are bell-shaped, symmetric.
  - ✓ The standard deviation is larger than 1. The precise value of the sd depends on degrees of freedom (df = n-1).
  - ✓ The t-distribution becomes closer to the z-distribution as the sample increases.
  - ✓ Even if the population is not normal, confidence intervals based on the *t* distribution still work quite well, especially when *n* exceeds about 15. We can say that t-distribution is **robust** to the violations of its normality assumption.



Parameters	Point Estimate	Estimated standard error	Confidence Interval	Sample Size
Mean (μ)	$\bar{y}$	$se = \frac{sd}{\sqrt{n}}$	$\bar{y}$ ± t*(se) t = 1.96 (95%)	$n = \sigma^2 * (\frac{z}{M})^2$ σ is population sd and M is margin of error
Proportion ($\hat{\pi}$)	$\hat{\pi}$	$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	$\hat{\pi}$ ± z*(se) z = 1.96 (95%)	$\pi * (1-\pi) * (\frac{z}{M})^2$

SIGNIFICANCE TESTS

1. Assumptions → random sampling
   → standard distribution of a population parameter
2. Hypothesis → 2 types of hypothesis:
   - Null hypothesis (H_0) → we suggest that parameter takes a **particular value**
   - Alternative hypothesis ($H_{a)}$ → the parameter falls in some alternative **range of values**
3. Test statistic → summarize how far some estimate falls from the parameter value suggested by H_0 .
4. p-value → the probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by H_0.

- If p-value is small (in the .05 significance level) we **reject** H_0 (If p-value if equal to the significance level, we reject H_0).
- If p-value is large (in the .05 significance level) we **fail to reject** H_0.

- Significance test of a mean
  - ✓ Assumption → Hypothesis → Test statistic → p-value → Inference
  - ✓ t-score = $(\bar{y} - \mu_0)/se$
  - ✓ se = $sd/\sqrt{n}$
  - ✓ After finding t-score we find the corresponding p-value. If p-value is smaller than the significance level, we reject H_0; if p-value is larger than the significance level, we fail to reject H_0.

- Significance test for proportion
  - ✓ Assumption → Hypothesis → Test statistic → p-value → Inference

$$z = \frac{\hat{\pi} - \pi_0}{se_0}, \quad \text{where} \quad se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}.$$

  - ✓ $\hat{\pi}$ is the sample proportion, π_0 is the population proportion, se_0 is the standard error of our estimate.
  - ✓ se_0 is used to indicate that it is the standard error under the presumption that H_0 is true.
  - ✓ When H_0 is true, $\pi = \pi_0$, se_0: $se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$.

- Type 1 and Type 2 error
  - ✓ T1 (false positive) → H_0 is true but we reject it
  - ✓ T2 (false negative) → H_0 is false but we fail to reject it
  - ✓ If you try to decrease T1 error, your T2 error might increase. As T1 goes down, T2 goes up.

- The Binomial Distribution
  - ✓ Flipping a coin. For each flip, we observe whether the outcome is head (category 1) or tail (category 2). The probabilities of the outcomes are the same for each flip (0.50 for each if the coin is balanced). The outcome of a particular flip does not depend on the outcome of other flips.
  - ✓ for $n = 5$ coin flips, x = number of heads could equal 0, 1, 2, 3, 4, or 5.

$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

  - ✓ π denotes the population proportion. Choosing any male from a sample; $\pi = .5$

Parameter	**Hypothesis**	**Test Statistic**	**Standard Error**	**p-value**
Mean	H_o H_a	$t = \frac{\bar{y} - \mu_0}{\sqrt{n}}$	$se = \frac{sd}{\sqrt{n}}$	If >.05 fail to reject H_0, If < .05 reject H_0
Proportion	H_o H_a	$z = \frac{\hat{\pi} - \pi_0}{se_0}$	$se_0 = \sqrt{\frac{\pi_0 * (1 - \pi_0)}{n}}$	If >.05 fail to reject H_0, If < .05 reject H_0