Basic Inference (Ch. 7-10 in textbook)

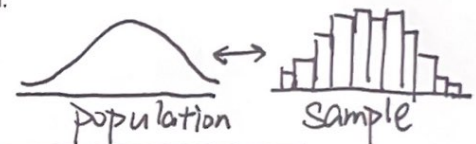
VARIABLES

TYPE OF DATA	SUMMARIZE	Numerical Summaries	Graphs
QUANTITATIVE VARIABLE numbers like height, grades, etc	Mean	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	Histogram Stemplot
	Standard deviation	$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$	Boxplot Dot plot
CATEGORICAL VARIABLE Yes/No type answers	Proportion of successes	$\hat{p} = \frac{\# \text{successes}}{\# \text{Total}}$	Bar chart Pie chart

SAMPLE vs POPULATION → original distribution the sample comes from

We typically have data for a sample or subset of the population. If that sample is random and (SRS) representative, we can use the data to make inferences about the population.

CI   Sig Tests

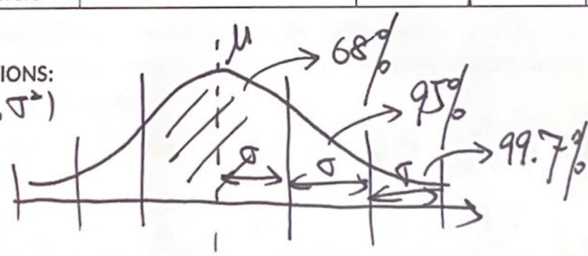


PARAMETERS vs STATISTICS

TYPE OF DATA	SYMBOLS	Population Parameter	Sample Statistic (ESTIMATOR)
QUANTITATIVE VARIABLE numbers like height, grades, etc	Mean	μ	$\bar{X}$
	Standard deviation	σ	S
CATEGORICAL VARIABLE Yes/No type answers	Proportion of successes	p	$\hat{p}$

NORMAL DISTRIBUTIONS:

$N(\mu, \sigma^2)$



NORMAL PROBABILITIES USING Z Table: Women's heights have a Normal Distribution with mean of 65 inches and standard deviation of 3 inches. Find the probability that one woman's height is more than 69 inches.

$$z\text{-score} = \frac{69-65}{3} = 1.33$$

$$P(Z > 1.33) = 1 - 90.82\% = 9.18\%$$

from z-table

SAMPLING DISTRIBUTIONS study the distribution of a sample statistic – that is, how sample statistics vary when the population parameter is known. Knowing this allows us to do Statistical Inference when the parameter is unknown.

CLT

TYPE OF DATA	Sampling Distribution of the STATISTIC	NORMAL Distribution valid IF:	
QUANTITATIVE VARIABLE numbers like height, grades, etc	$\bar{X} \overset{a}{\sim} N(\mu, \frac{\sigma}{\sqrt{n}})$	Original distribution is Normal **OR** $n \geq 30$ (CLT)	eg Roll a die n times
CATEGORICAL VARIABLE Yes/No type answers	$\hat{p} \overset{a}{\sim} N(p, \sqrt{\frac{p(1-p)}{n}})$	$np \geq 15$ **AND ALSO** $n(1-p) \geq 15$	eg Suppose 30% of population would say "yes" to "Do you have at least one cat?"

PROPORTIONS: 70% of students $(p=0.70)$ like white chocolate. We will take a random sample of 100 students and record how many of them like white chocolate. Find the probability that less than 60% of them say yes.

$$\#\text{sample} = 100 \Rightarrow \begin{array}{l} 0.7 \times 100 = 70 \geq 15 \\ 0.3 \times 100 = 30 \geq 15 \end{array}$$

$$\Rightarrow \hat{p} \sim N(0.7, \sqrt{\frac{0.7 \times 0.3}{100}}) = N(0.7, \overset{\text{std}}{0.0458})$$

$$\Rightarrow P(\hat{p} < 0.60) = P(\frac{\hat{p} - 0.7}{0.0458} < \frac{0.6 - 0.7}{0.0458}) = P(Z < -2.18)$$

$$= 0.0146$$

MEANS: Women's heights have a Normal Distribution with mean of 65 inches and standard deviation of 3 inches. Find the probability that the average height of 10 randomly selected women is more than 69 inches.

$$X_i \sim N(65, \overset{\text{std}}{3}) \Rightarrow \bar{X} \sim N(65, \frac{3}{\sqrt{10}}) = N(65, 0.9487)$$

$$\Rightarrow P(\bar{X} > 69) = P(\frac{\bar{X} - 65}{0.9487} > \frac{69 - 65}{0.9487}) = P(Z > 4.22) \approx 0.$$

Draw conclusions about unknown population parameters based on random + representative samples (SRS)

STATISTICAL INFERENCE

In STATISTICAL INFERENCE the parameters are UNKNOWN and we want to estimate them. So we take a random and representative sample from the population of interest and compute the sample statistic (estimator). However, we don't just report that estimator – we know the value of the parameter is close to that, but probably not exactly the same. So Statistical Inference also attaches a measure of reliability to our estimator. **There are two basic types of Statistical Inference:**

CONFIDENCE INTERVALS

A random sample of 500 students at your school reveals that 78% in the sample drink alcohol. What can we say about the population proportion?

- The standard error of the estimator is 0.0185≈ 2%. $\rightarrow \sqrt{\frac{0.78 \times 0.22}{500}}$
- If we want to have a 95% chance of capturing the true population proportion, we need to go about 2 standard error left and right, or about 4%.
- So we can be 95% confident that the true proportion of students who drink alcohol at your school is between **74% and 82%.**

SIGNIFICANCE TESTS

A friend suggests that the average GPA of students at your school is 3.5. You think that is too high, so you collect data for a random sample of 50 students at your school and their average GPA is 3.28 with a standard deviation of 0.826. How much evidence do you have to say your friend was wrong?

- The z-score for this observation is -1.88. $\rightarrow \frac{3.28-3.5}{0.826/\sqrt{50}}$
- The p-value of the test is computed to be 0.03, meaning only 3% of samples would give results as low as yours if the true average GPA was 3.5.
- This is **pretty strong evidence to say the true average is lower than your friend thought.**

In Stats 1 you learned how to make and interpret Confidence Intervals and Significance Tests for:

- Mean of one group
- Proportion of successes for one group
- Comparing Means of two groups
- Comparing Proportion of successes for two groups.

We will focus on MEANS during the review, but interpretations are similar for proportions.

A confidence interval is an interval that gives a reasonable estimate of the unknown parameter

CONFIDENCE INTERVAL for an UNKOWN Population Parameter:

estimator ± margin of error => estimator ± (t or z) * standard error

Margin of error depends on two things:

- **how far we need to go for whatever confidence level we want (typically 95%)**
- **and the standard error of our estimator (as we learned in Sampling Distributions)**

CASE	CI for UNKOWN PARAMETER	ASSUMPTIONS - Need to check FIRST:
ONE MEAN (μ)	$\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}}$	• SRS => Data is random and representative of population of interest • Original distribution is Normal **OR** $n \geq 30$ • **NOTE:** if $n \geq 30$ in practice we can use the Z table. But if $n<30$ we must check there are no outliers and use the t table with $df = n-1$
ONE PROPORTION (p)	$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	• SRS => Data is random and representative of population of interest • We have **at least** 15 successes and 15 failures in data $n\hat{p}$ $n(1-\hat{p})$

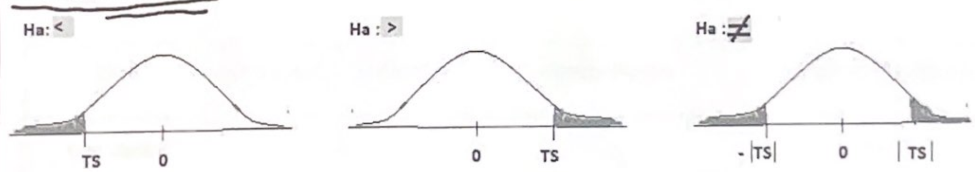
(random is good, but representative is what we really want)

INTERPRETING CONFIDENCE INTERVALS: We are 95% (or whatever%) confident that **the parameter** is between the endpoints of the interval. (Confidence intervals are statements about the POPULATION PARAMETER, not about the sample statistic or about individuals.)

SIGNIFICANCE TESTS for UNKOWN Population Parameters:

ELEMENTS OF A SIGNIFICANCE TEST	SIGNIFICANCE TEST FOR μ
ASSUMPTIONS – need to check first:	• SRS => Data is random and representative of population of interest • Original distribution is Normal **OR** $n \geq 30$ NOTE: if $n \geq 30$ we can use Z But if $n<30$ we must check there are no outliers and use the t table with df = n-1
NULL HYPOTHESIS: what we want to **DISPROVE** H_o : parameter = # **ALTERNATIVE HYPOTHESIS:** what we want to **PROVE** H_a : parameter ≠ / > / < #	$H_o: \mu = \mu_o$ $H_a: \mu \neq \mu_o$ / $>$ / $<$ Need to identify the # we are trying to disprove, and the sign of the Alternative Hypothesis (come from the story)
TEST STATISTIC: z-score (or t-score) summarizes the information from the sample – it measures how far away the estimator is from the value of the parameter specified in the null hypothesis in terms of standard errors	**Test Statistic TS:** $t = \dfrac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$

P-VALUE: "corner" area in the direction of the alternative hypothesis Ha. P-value is the area shaded in green below. It is the probability that the test statistic equals the observed value or a value even more extreme if Ho is true.



CONCLUSIONS: statement based on the p-value in everyday language

- **Small p-values support H_a and lead us to REJECT H_o and determine the results are Statistically Significant.**
- **How small is small?** Compare to **significance level α (alpha level).**
- **Most common α's :** 0.10, 0.05, 0.01 (corresponding to 90%, 95% and 99% confidence in Ha)
- **The smaller the p-value, the more evidence we have to prove Ha.**
- **But if the p-value is bigger than 0.10 we say we FAIL to Reject H_o – never Accept H_o or Reject H_a**

	Not rej H_o	Rej H_o
H_o T	✓	Type I error → indiscreet
H_o F	Type II error ↓ too prudent	✓

p-val < 0.01 — very strong evidence

< 0.05 — pretty strong - -

< 0.10 — some - -

RELATIONSHIP BETWEEN CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS

A confidence interval for mean gives the same interpretation as a two-tailed significance test

- If the confidence interval contains the parameter under H_0, then we should fail to reject H_0.
- If the confidence interval does not contain the parameter under H_0, the we should reject H_0.
- This is especially helpful in comparing TWO GROUPS. If the confidence interval contains 0, then we should conclude that there is **no significant difference** between the two groups. We should fail to reject H_0 that the difference is 0.
- If the confidence interval does not contain 0, then we should conclude that there **is a significant difference** between the two groups. We should reject H_0 and conclude the difference is not 0.

COMPARING TWO GROUPS

Interpreting CI for Comparing Two Groups – Look for ZERO in the interval

- (- , +): If the confidence interval includes zero, then the difference between the two groups could be zero. There is no statistically significant evidence of a significant difference between the (means or proportions) two groups in the population. No Significant Differences
- (+, +): If the confidence interval does not include zero, and the values in the interval are positive (a, b), then the (mean or proportion) for group 1 is between a and b higher than group 2.
- (-, -): If the confidence interval does not include zero, and the values in the interval are negative (-a, -b), then the (mean or proportion) for group 2 is between ~~a~~ b and ~~b~~ a higher than group 1.

Interpreting results of Significance Tests for Two Groups – Look at the p-value

- Typically the Null Hypothesis will say there is **NO DIFFERENCE** in the groups ⇔ difference = 0
- Small p-value → Reject Ho → Significant Differences between the two groups
- Large p-value → Fail to Reject Ho → Not enough evidence to prove a Significant Difference between the two groups

Case	parameter	estimator	standard error	Sampling Distribution
one mean	μ	$\bar{x}$	$s/\sqrt{n}$	t (n-1)
mean of matched pairs difference	μ_d	$\bar{x}_d$	$s_d/\sqrt{n}$	t (n-1)
difference of two independent means	$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	t with df between: smallest of (n_1-1) and (n_2-1) n_1+n_2-2
one proportion	p	$\hat{p}$	Too messy! CI: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ST: $\sqrt{\frac{p_0(1-p_0)}{n}}$	
difference of two independent proportions	p_1-p_2	$\hat{p}_1-\hat{p}_2$	Too messy! CI: $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ ST: $\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	