STAT 201 - Final Exam Review Sheet

I. Confidence Intervals $\underline{PE \pm MULT \times SE}$

(Proportions use z^*; means use t^*)

Parameter to Estimate	PE	Multipler	SE
p (one proportion)	$\widehat{p}$	$z^* =$ invnorm()	$\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
$p_1 - p_2$ (two proportions)	$\widehat{p}_1 - \widehat{p}_2$	z^* =invnorm()	$\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$
μ (one mean)	$\bar{x}$	$t^* =$ invT(,df) (or use table)	$s/\sqrt{n}$
$\mu_1 - \mu_2$ (two means)	$\bar{x}_1 - \bar{x}_2$	$t^* =$ invT(,df) (or use table)	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Comments:

- Interpretation: We are xx% confident that the population parameter is between $\underline{PE - MOE}$ and $\underline{PE + MOE}$.
- If H_0 is in the confidence interval, there is little to no evidence for H_A.
- If H_0 is not in the confidence, there is evidence for H_A.

II. Hypothesis Tests

1. State the Hypotheses

Always in terms of parameters; data never goes in hypotheses

Parameter	Null Hypothesis	Alternative Hypothesis
p (one proportion)	$H_0 : p = p_0$	$H_A : p <, >, \neq p_0$
p_1 vs. p_2 (two proportions)	$H_0 : p_1 = p_2$ $H_0 : p_1 - p_2 = 0$	$H_A : p_1 <, >, \neq p_2$ $H_A : p_1 - p_2 <, >, \neq 0$
μ (one mean)	$H_0 : \mu = \mu_0$	$H_A : \mu <, >, \neq \mu_0$
μ_1 vs. p_2 (two means)	$H_0 : \mu_1 = \mu_2$ $H_0 : \mu_1 - \mu_2 = 0$	$H_A : \mu_1 <, >, \neq \mu_2$ $H_A : \mu_1 - \mu_2 <, >, \neq 0$

2. Calculate the Test Statistic

Tells us how far the data is from the initial assumption in H_0. In the form of Test Stat. $= \frac{data-null}{SE}$; proportions z_c; means t_c

Parameter	Test Statistic
p (one proportion)	$z_c = \frac{\widehat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
p_1 vs. p_2 (two proportions)	$z_c = \frac{(\widehat{p}_1-\widehat{p}_2)-0}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
μ (one mean)	$t_c = \frac{\bar{x}-\mu_0}{(s/\sqrt{n})}$
μ_1 vs. p_2 (two means)	$t_c = \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

3. Calculate the P-value

Tells us the probability of seeing our data/test statistic under the null hypothesis; Find area according to H_A; it is a probability so answers should always be between zero and one; means use $tcdf(lwr, upr, df)$, proportions use $normalcdf(lwr, upr)$.

4. Interpret the P-value

"If H_0 is true, we would see data like ours, or more extreme, $p - value \times 100\%$ of the time."

5. State the conclusion.

"There is ____________ evidence for H_A"

- $p > 0.10 \rightarrow$ "Little to no"
- $0.05 < p < 0.10 \rightarrow$ "Some"
- $0.01 < p < 0.05 \rightarrow$ "Strong"
- $p < 0.01 \rightarrow$"Very strong"

Comments:

- If there is little to no evidence for H_A, then H_0 is still in the confidence interval.
- If there is a degree of evidence for H_A, then H_0 is not in the confidence interval.

III. Binomial Distribution

X = number of successes in n trials, $X \sim Binomial(n, p)$

Criteria for a Binomial distribution

- **B**inary outcomes (success/failure)
- **I**ndependent trials (ex. random sample?)
- **N**umber of trials fixed ($n = ?$)
- **S**ame probability of success ($p = ?$)

Mean and Standard Deviation

- $\mu = np$
- $\sigma = \sqrt{np(1-p)}$

Finding probabilities

- $P(X = k)$ =binom**p**df(n, p, k)
- $P(X \leq k)$ =binom**c**df(n, p, k)

IV.Test for Outliers

- Five number summary: min, Q_1, M, Q_3, max
- $IQR = Q_3 - Q1$
- $step = 1.5 \times IQR$
- $LF = Q_1 - step$
- $UF = Q_3 + step$
- Any observations beyond LF, UP are outliers

V. Correlation and Regression

Properties of Correlation

- r has no units and does not imply causation
- measures strength of linear relationships between two quantitative variables
- $r = -1$ implies perfect line with negative slope
- $r = +1$ implies perfect line with positive slope

Regression Calculations

- y = response; the variable we wish to estimate or predict
- x = explanatory variable; the variable that influences the response
- Slope: $b = r(s_y/s_x)$
- Intercept: $a = \bar{y} - b\bar{x}$
- LSR Line: $\hat{y} = a + bx$
- Residual: $e = y - \hat{y}$ (observed - predicted y)
- Coefficient of determination: r^2 = proportion of variation in y explained by the line

Interpretations

- Slope: As x increases by one unit, the predicted y increases by slope units.
- Intercept: When x is zero, the predicted y is intercept units.

VI. Probability

General Formulas

Key Word	Formula
NOT	$P(A^c) = 1 - P(A)$
OR	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
OR	$P(A \cup B) = P(A) + P(B)$ (Special case: Disjoint)
AND	$P(A \cap B) = P(A) \times P(B)$ (Special case: Independent)
AND	$P(A \cap B) = P(A\|B)P(B) = P(A)P(B\|A)$
GIVEN	$P(A\|B) = P(A \cap B)/P(B)$

Tests for Independence

If the two sides are equal, then the events are independent. If unequal, then dependent.
Choose any of the three formulas to test.

- $P(A \cap B) = P(A) \times P(B)$
- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

Test for Disjointedness

If $P(A \cap B) = 0$, the events are disjoint.

VII. Sampling Distributions

Sampling distributions of means

- For $n \geq 30$, the sampling distribution of the sample mean is

$$\bar{x} \sim N\left(\mu, \sigma/\sqrt{n}\right)$$

- To find probabilities related to $\bar{x}$, the corresponding z-score is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

then use `normalcdf()` accordingly to find the probability

Sampling distributions of proportions

- For $np \geq 10$ and $n(1-p) \geq 10$, the sampling distribution of the sample proportion is

$$\widehat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- To find probabilities related to $\widehat{p}$, the corresponding z-score is

$$z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

then use `normalcdf()` accordingly to find the probability