# Exam 2 Review

## Format of the Exam

Read a dataset and be able to answer the following questions involving this dataset. See the attached documents for specific question examples of this nature. The attached Jupyter notebook is not comprehensive, in that it does not cover ALL things talked about in this document and what might be on the test.

## 1. Descriptive Analytics

### 1.1. Two numerical variables

1.1.1. Visualize a scatterplot between two numerical variables.

1.1.2. Describe the relationship between two numerical variables using a scatter plot

1.1.3. Summary statistics:

1.1.3.1. Calculate and interpret the covariance

1.1.3.2. Calculate and interpret the correlation.

1.1.3.2.1. When are you allowed and not allowed to use the correlation to summarize the association and direction between two numerical variables?

1.1.3.3. Calculate and interpret the R^2, when you have a given linear regression model.

### 1.2. Plotting three or more variables:

1.2.1. Plot two numerical variables in a scatter plot and color code (and/or change the marker type) by some categorical variable.

1.2.2. Plot a series of linear regression best fit lines for each color of point in a scatterplot.

### 1.3. Plotting two or more plots:

1.3.1. Plot multiple scatterplots for all numerical variables in a dataframe all in the same plot.

## 2. Sampling Distributions

### 2.1. What is the mean and standard deviation of each following types of sampling distributions? Under what conditions will they be normal?

2.1.1. Sampling distribution of sample means

2.1.2. Sampling distribution of sample proportions

2.1.3. Sampling distribution of sample mean differences

2.1.4. Sampling distribution of sample proportion differences

## 3. Inference on a Population Parameter

### 3.1. Conduct inference on the following population parameters.

3.1.1. Single population mean.

3.1.2. Difference between two population means

3.1.3. Single population proportion

3.1.4. Difference between two population proportions.

### 3.2. For each type of population parameter (mentioned above in 3.1) know how to do the following inference related things.

3.2.1. Check the necessary conditions needed for conducting inference on that particular population parameter.
3.2.2. Calculate the critical value. (Know whether the critical value comes from a z-distribution of t-distribution for this particular parameter).
3.2.3. Create a XX% confidence interval for this population parameter.
3.2.4. Interpret the confidence interval (ie. put the two numbers into words).
3.2.5. What does "XX%" mean in a confidence interval?
3.2.6. Calculate a test statistic.
- 3.2.6.1. Know what distribution this test statistic comes from for this particular population parameter.
- 3.2.6.2. If degrees of freedom are needed for this particular distribution (know what degrees of freedom you should use).

3.2.7. Calculate a p-value using a test statistic.
3.2.8. Set up null and alternative hypotheses to test a claim.
3.2.9. Form a conclusion about the null and alternative hypotheses using:
- 3.2.9.1. A confidence interval
- 3.2.9.2. A p-value

## 4. Linear Regression

### 4.1. Modeling

4.1.1. Set up a linear regression equation in Python.
4.1.2. Write out the linear regression equation using the Python output, **using the right notation!**
4.1.3. Types of explanatory variables to know how to interpret/setup.
- 4.1.3.1. Numerical explanatory variables
- 4.1.3.2. Categorical explanatory variables.
  - 4.1.3.2.1. If your categorical explanatory variable has *p* levels, how many slopes will represent it in the regression equation?
- 4.1.3.3. Interaction term explanatory variables.

### 4.2. Predictions

4.2.1. Make predictions using this linear regression equation by hand.
4.2.2. Make predictions using this linear regression equation using Python.
4.2.3. Calculate the residual of an observation by hand.

### 4.3. Linear Regression Conditions

4.3.1. How to check the conditions for linear regression modeling and inference.

### 4.4. Linear Transformation

4.4.1. Why might we want to transform one or more of our variables before setting up a linear regression model? How do we do this?

### 4.5. Inference

4.5.1. Individual Non-Zero Population Slope Hypothesis Testing
- 4.5.1.1. Set up the hypotheses.
- 4.5.1.2. Find the test statistic and the p-value in the summary output.
- 4.5.1.3. Calculate the test statistic and the p-value by hand.
- 4.5.1.4. Calculate a confidence interval for this population slope by using the summary output.
- 4.5.1.5. Make a conclusion using the confidence interval and p-value.

4.5.2. Hypothesis Testing for at least one non-zero population slope
- 4.5.2.1. Set up the hypotheses.
- 4.5.2.2. Find the test statistic and the p-value in the summary output.
- 4.5.2.3. Calculate the p-value by hand, using the test statistic.
- 4.5.2.4. Make a conclusion using the p-value.

4.5.3. Hypothesis Testing for at least one non-zero population slope that is in the full model but not the reduced model.
- 4.5.3.1. Set up the hypotheses.
- 4.5.3.2. Find the test statistic and the p-value in the Python.
- 4.5.3.3. Calculate the p-value by hand, using the test statistic.
- 4.5.3.4. Make a conclusion using the p-value.

### 4.6. Summary Statistics
4.6.1. Interpret the R^2.

4.6.2. Calculate the R^2 by hand (for a simple linear regression).

4.6.3. Find the R^2 in the summary output table (for multiple linear regression.

## 5. ANOVA
### 5.1. Inference for the more than 2 population means, testing if at least two are different.
5.1.1. Set up the hypotheses.

5.1.2. Find the test statistic and the p-value in Python.

5.1.3. Calculate the p-value by hand, given the test statistic.

5.1.4. Make a conclusion using the p-value.

5.1.5. **Why would we conduct this hypothesis test?**

## 6. Logistic Regression
### 6.1. Definitions
6.1.1. What are the odds of a success?

6.1.2. What is the relationship between (ie. equation relating) an odds and a probability?

### 6.2. Logistic Regression Modeling
6.2.1. Set up a logistic regression equation in Python.

6.2.2. Write out the logistic regression equation using the Python output.

6.2.3. Types of explanatory variables to know how to interpret/setup.

6.2.3.1. Numerical explanatory variables

6.2.3.2. Categorical explanatory variables.

6.2.3.2.1. If your categorical explanatory variable has *p* levels, how many slopes will represent it in the regression equation?

6.2.3.3. Interaction term explanatory variables.

## 6.3. Logistic Regression Predictions

6.3.1. Make predictions using this logistic regression equation by hand.

6.3.1.1. Predict the log odds of the success level given explanatory variable values.

6.3.1.2. Predict the odds of the success level given explanatory variable values.

6.3.1.3. Predict the probability of the success level given explanatory variable values.

## 6.4. Logistic Regression Inference

6.4.1. Individual Non-Zero Population Slope Hypothesis Testing

6.4.1.1. Set up the hypotheses.

6.4.1.2. Find the test statistic and the p-value in the summary output.

6.4.1.3. Calculate the test statistic and the p-value by hand.

6.4.1.4. Calculate a confidence interval for this population slope by using the summary output.

6.4.1.5. Make a conclusion using the confidence interval and p-value.

## 6.5. Logistic Regression Slope Interpretation

6.5.1.1. What does the slope for a numerical explanatory variable represent in a logistic regression?

6.5.1.2. What does the slope for a categorical explanatory variable (with 2 levels) represent in a logistic regression?

# 7. Probability Theory/Properties

## 7.1. Distributions

### 7.1.1. Binomial Distribution

7.1.1.1. What is the mean and standard deviation of a binomial random variable?

7.1.1.2. How to calculate P(X=#), where X is a binomial random variable.

7.1.1.3. How to recognize when a random variable X is a random variable.

### 7.1.2. Standard Normal Distribution

7.1.2.1. What is the mean and standard deviation of the standard normal distribution?

7.1.2.2. Fill in the blank: The distribution of z-scores of observations from a population will follow the standard normal distribution when the distribution of observations from the population is _____.

### 7.1.3. T-Distribution vs. Z-Distribution (ie. Standard Normal Distribution)

7.1.3.1. What is the difference between the t-distribution and the z-distribution? (In terms of shape).

7.1.3.2. What does the shape of the t-distribution do as we increase the degrees of freedom?

### 7.1.4. F-Distribution

7.1.4.1. What are some properties of the F-distribution?

# 8. Additional Code

## 8.1. groupby function

## 8.2. How to create new columns in a dataframe.

## 8.3. How to rename the names of your levels for a given column in your dataframe.

# Population Parameter Confidence Interval "Cheat Sheet"

$$(point\ estimate) \pm (crit.value)SE$$

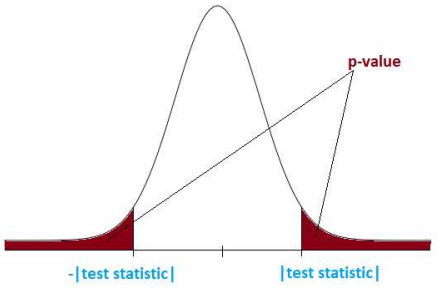
| Types of Variable(s) Involved | Population Parameter | Point Estimate | Standard Error (SE) | The confidence level should be the area between -critical value and critical value under the _____. | Conditions for Inference |
|---|---|---|---|---|---|
| Single **Numerical** Variable | $\mu$ | $\bar{x}$ | $\frac{\sigma}{\sqrt{n}}$ | Know $\sigma$<br>Standard normal distribution<br>Don't Know $\sigma$<br>t-distribution with df=n-1 | 1. Random sample<br>2. n<10% of population size<br>3. Either n>30 or sample distribution is normal. |
| Single **Categorical** Variable (2 levels) | $p$ | $\hat{p}$ | For conf. intervals (no hypothesis test)<br>$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$<br>For Hyp. Test<br>$\sqrt{\frac{p_0(1-p_0)}{n}}$ | Standard normal distribution | 1. Random sample<br>2. n<10% of population size<br>3. $n\hat{p} \geq 10\ and\ n(1-\hat{p}) \geq 10$<br>1. Random sample<br>2. n<10% of population size<br>3. $np_0 \geq 10\ and\ n(1-p_0) \geq 10$ |
| **Numerical** Variable and **Categorical** Variable (with **2 levels**) | $\mu_1 - \mu_2$ | $\bar{x}_1 - \bar{x}_2$ | $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ | Know $\sigma_1$ and $\sigma_2$<br>Standard normal distribution<br>Don't Know $\sigma_1\ and\ \sigma_2$<br>t-distribution with $df = \min\{n_1 - 1, n_2 - 1\}$ | 1. Sample 1 is random<br>2. Sample 2 is random<br>3. $n_1 < 10\%\ of\ population\ 1\ size$<br>4. $n_2 < 10\%\ of\ population 21\ size$<br>5. Either $n_1 > 30$ or sample 1 distribution is normal.<br>6. Either $n_2 > 30$ or sample 2 distribution is normal.<br>7. Observations in sample 1 and sample 2 are independent. |
| **Two Categorical** Variables (each with 2 levels) | $p_1 - p_2$ | $\hat{p}_1 - \hat{p}_2$ | $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ | Standard normal distribution | 1. Sample 1 is random<br>2. Sample 2 is random<br>3. $n_1 < 10\%\ of\ population\ 1\ size$ |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | 4. $n_2 < 10\% \; of \; population 21 \; size$<br>5. $n_1 \hat{p}_1 \geq 10 \; and \; n_1(1-\widehat{p_1}) \geq 10$<br>6. $n_2 \hat{p}_2 \geq 10 \; and \; n_2(1-\widehat{p_2}) \geq 10$<br>7. Observations in sample 1 and sample 2 are independent. |
| **Regression** with **Numerical Response Variable** | $\beta_i$ | $\hat{\beta}_i$ | (See summary output table) | t-distribution with $df = \mathrm{n} - \mathrm{p} - 1$ (where p=# of slopes in the model) | 1. Linearity condition<br>2. Constant variance of residuals condition.<br>3. Normality of residuals (mean of 0) condition.<br>4. Independence of residuals condition.<br>5. No Multicollinearity Condition |
| **Regression** with a **Categorical Response Variable (2 levels)** | $\beta_i$ | $\hat{\beta}_i$ | (See summary output table) | Standard normal distribution | 1. Random sample<br>2. n<10% of population size<br>3. Linearity condition<br>4. No Multi-Collinearity Condition |

# Population Parameter Hypothesis Test (with p-value) "Cheat Sheet"

$Ho: pop.param = null\ value$
$Ha: pop.param \neq null\ value$

$$Test - Stat = \frac{(point\ estimate) - null\ value}{SE}$$




| Types of Variable(s) Involved | Population Parameter | Point Estimate | Standard Error (SE) | The test statistic is an observation from _____. | Conditions for Inference |
|---|---|---|---|---|---|
| Single **Numerical** Variable | $\mu$ | $\bar{x}$ | $\frac{\sigma}{\sqrt{n}}$ | Know $\sigma$<br>Standard normal distribution<br>Don't Know $\sigma$<br>t-distribution with df=n-1 | 1. Random sample<br>2. n<10% of population size<br>3. Either n>30 or sample distribution is normal. |
| Single **Categorical** Variable (2 levels) | $p$ | $\hat{p}$ | For conf. intervals (no hypothesis test)<br>$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$<br>For Hyp. Test<br>$\sqrt{\frac{p_0(1-p_0)}{n}}$ | Standard normal distribution | 1. Random sample<br>2. n<10% of population size<br>3. $n\,\hat{p} \geq 10\ and\ n(1-\hat{p}) \geq 10$<br>1. Random sample<br>2. n<10% of population size<br>3. $np_0 \geq 10\ and\ n(1-p_0) \geq 10$ |
| **Numerical** Variable and **Categorical** Variable (with **2 levels**) | $\mu_1 - \mu_2$ | $\bar{x}_1 - \bar{x}_2$ | $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ | Know $\sigma_1$ and $\sigma_2$<br>Standard normal distribution<br>Don't Know $\sigma_1\ and\ \sigma_2$<br>t-distribution with $df = \min\{n_1 - 1, n_2 - 1\}$ | 1. Sample 1 is random<br>2. Sample 2 is random<br>3. $n_1 < 10\%\ of\ population\ 1\ size$<br>4. $n_2 < 10\%\ of\ population 21\ size$<br>5. Either $n_1 > 30$ or sample 1 distribution is normal.<br>6. Either $n_2 > 30$ or sample 2 distribution is normal.<br>7. Observations in sample 1 and sample 2 are independent. |

| | | | | | |
|---|---|---|---|---|---|
| **Two Categorical** Variables (each with 2 levels) | $p_1 - p_2$ | $\hat{p}_1 - \hat{p}_2$ | $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ | Standard normal distribution | 1. Sample 1 is random<br>2. Sample 2 is random<br>3. $n_1 < 10\% \; of \; population \; 1 \; size$<br>4. $n_2 < 10\% \; of \; population 21 \; size$<br>5. $n_1 \hat{p}_1 \geq 10 \; and \; n_1(1-\widehat{p_1}) \geq 10$<br>6. $n_2 \hat{p}_2 \geq 10 \; and \; n_2(1-\widehat{p_2}) \geq 10$<br>7. Observations in sample 1 and sample 2 are independent. |
| **Regression** with **Numerical Response Variable** | $\beta_i$ | $\hat{\beta}_i$ | (See summary output table) | t-distribution with $df = \mathrm{n} - \mathrm{p} - 1$ (where p=# of slopes in the model) | 1. Linearity condition<br>2. Constant variance of residuals condition.<br>3. Normality of residuals (mean of 0) condition.<br>4. Independence of residuals condition.<br>5. No Multicollinearity Condition |
| **Regression** with a **Categorical Response Variable (2 levels)** | $\beta_i$ | $\hat{\beta}_i$ | (See summary output table) | Standard normal distribution | 1. Random sample<br>2. n<10% of population size<br>3. Linearity condition<br>4. No Multi-Collinearity Condition |