Math 338 Midterm 2 Study Guide

Disclaimer: This exam is not intended to be a guide to everything I could possibly ask about on the midterm. However, if you understand the computational procedures and terms below, concepts related to those terms/procedures, and how to interpret your results, you are probably in good shape for the exam.

1 Lecture Portion

1.1 Lectures 13-14: Numerical Variables and Continuous Random Variables

- Sketch the pdf for a uniform random variable and use it to find probabilities

  This is a rectangular curve with area of one. $\sqcap$ represents the area under the curve. Formula for said area is $\frac{1}{b-a} \times (b-a)$.

- Use the 68-95-99.7 rule of thumb to estimate probabilities involving normal random variables

  Under a normal distribution, the number of standard deviations (σ) away from the mean (μ) will determine the area (amount of probability) it will be.

- Convert values to z-scores and explain why a z-score is used to compare values from different distributions

  The formula for converting z-scores is $z = \frac{\bar{x}-\mu}{\sigma}$. A value's relationship to the mean (μ) of a group of values, measured in terms of standard deviations (σ) from the mean. Z-Score allows us to have a universal standard for density curves with different scales.

- Identify statistics/parameters as measures of center (average) or variability (spread, variation)

  Parameters as a measure of $\mu \rightarrow$: The average age in this class is 21 years old.

  Parameters as a measure of $\sigma \rightarrow$: I scored one σ from the average, which means I did better than average (μ)

- Identify a density curve as skewed left/skewed right/symmetric and unimodal/multimodal

  Skewed Left: long left tail. Sloping $\rightarrow$

  Skewed Right: long right tail: Sloping $\leftarrow$

  Symmetric: perfectly distributed curve. Think bell shaped curve.

  Unimodal: one hump

  Bimodal: two humps

- Use the $Q_1 \; - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ convention to identify outliers

  Five number summary:

  - Min = 39
  - Q_1 = 55
  - Median = 63
  - Q_3 = 69
  - Max = 85
  - $IQR = 69 - 55.5 = 13.5$
  - Lower fence: $55.5 - (1.5)(13.5) = 35.25$
  - Upper fence: $69 + (1.5)(13.5) = 89.25$

  In this data set we have no outliers because our data falls between the fences.

- Compute the new mean and new variance of a numerical variable after linear transformation

  Linear Transformation Article

  A linear transformation is a change to a variable characterized by one or more of the following operations: adding a constant to the variable, subtracting a constant from the variable, multiplying the variable by a constant, and/or dividing the variable by a constant

  $\bar{y} = m\bar{x} \pm b$ is an example of calculating the new μ

  $\sigma(x) = \sqrt{\sigma^2(x)} \times b$ is an example of calculating the new σ

  Where b is the constant being appended to the base equation

- Compute the new mean and new variance of a linear combination of two numerical variables

  Linear Combination Article

  Mean differences formula: $\mu_{\mathrm{x}} + \mu_{\mathrm{y}} = \mu_{\mathrm{x}} + \mu_{\mathrm{y}}$

  ↑ The above formula also applies to expected values for random variables: $E(X+Y) = E(X) + E(Y)$

  Two separate quantities can be treated as one unified entity

  ↑ ↑ Please get some clarification about this

1.2 Lecture 15: Sampling Distribution of the Sample Mean

- Identify the difference between rounding error, measurement error, and sampling error
  - Rounding error: When a number collected in an experiment is rounded to n too many places, therefore losing precision.
  - Measurement error: When data is collected by a faulty piece of equipment or personnel.
  - Sampling error: When the sample being selected has an underlying problem (not random, misrepresentative of the population, etc.)
- Identify the shape and mean of a distribution used to model rounding error, measurement error, and sampling error
- Identify whether a distribution is the distribution of a variable or the sampling distribution of a statistic
- Identify whether a statistic is a biased or unbiased estimator of a parameter

  Biased: When there is a difference between the expected value and the actual recorded value during an experiment.

  Unbiased: When there is no difference between the expectation and the observed value.
- Explain the difference between bias and variability of a sampling distribution

  Bias is something that the experimenter can control (to a degree) and variability is something that comes directly from the data. Variability can be the **result** of bias but bias is not variability. More generally, variability is the moment in which data points diverge from one another.
- Use the Central Limit Theorem to approximate the sampling distribution of a sample mean

  $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

  - $\bar{X}$: is the value it converges to
  - μ: is the population mean
  - σ: is the population standard deviation
  - n: is the sample size
- Make an educated guess about whether the Central Limit Theorem approximation is "good enough" given a sample size and the distribution of the sample

  The Central Limit Theorem states that the sampling distribution will eventually converge to a single value if the sample size is "big enough". This is determined by how good the approximation needs to be and the shape of the distribution. This theorem plays with the notion of Calculus limit, where in as the function approaches ∞, the function will converge to a single value. This is the same here, where n (sample size) approaches an infinite amount.

1.3 Lectures 17-19: t-Statistics and t-Tests

- Given summary statistics for a sample, compute the standard error of the sample mean
- Identify the appropriate degrees of freedom in the t-distribution the t-statistic comes from (one-sample and matched pairs only)
- Write the null hypothesis H_0 and the alternative hypothesis H_1 for a t-test in the Neyman-Pearson framework (one-sample, matched pairs, and two-sample)
- Compute the t statistic under the null hypothesis H_0 (one-sample and matched pairs only)
- Decide whether to accept H_1 or to accept H_0, and explain in real-world context what your decision means (you will be given sufficient information to do this; I won't ask you to compute a critical region by hand)
- Given a testing situation, explain what would be a Type I Error vs. Type II Error and explain what the power of the test represents
- Write the null hypothesis H_0 and the alternative hypothesis H_a in the Null Hypothesis Significance Testing (NHST) framework (one-sample, matched pairs, and two-sample)
- Explain in context the idea of a p-value (one-sample, matched pairs, and two-sample)
- Decide whether to reject H_0 (and accept H_a) or to fail to reject H_0, and explain in real-world context what your decision means (you will be given sufficient information to do this; I won't ask you to compute a p-value by hand)

1.4 Lecture 20: One-Way ANOVA

- Given the description of an experiment, write the (null) hypothesis for a one-way ANOVA F test
- Given the description of an experiment, identify the correct DF values (all of them) for the ANOVA table
- Given sufficient information to complete the Sum of Squares column, complete the ANOVA table (except for the p-value)
- Check the assumptions of ANOVA (normal distribution in each group, equal population sd in each group) using our rules of thumb
- Identify the appropriate degrees of freedom parameters in the F-distribution the F-statistic comes from
- Decide whether to reject the hypothesis and explain in real-world context what your decision means (you will be given sufficient information to do this; I won't ask you to compute a p-value by hand)
- Explain when/why you do *post hoc* procedures

1.5 Lectures 21-22: Confidence Intervals

- Explain what a confidence interval is and what it means to be "95% confident"
- Explain the relationship between the confidence level and α
- Given a confidence interval situation, define the parameter to be estimated

- Given a t^{**} critical value, compute a confidence interval for the parameter (one-sample and matched pairs only)
- Given an arbitrary confidence interval, write a sentence interpreting it
- Given an arbitrary confidence interval, identify the values of the point estimate and margin of error
- Explain how the center and/or width of the confidence interval would change as the following change: sample mean, sample standard deviation, sample size, confidence level (one-sample and matched pairs only)
- Given a confidence interval for a population mean of paired differences or difference of population means, decide which population is larger on average
- Given an arbitrary confidence interval, decide whether to accept H_0 or H_1 (N-P), or decide whether to reject H_0 or fail to reject H_0 (NHST)

2 Lab Portion

Disclaimer: This exam is not intended to be a *comprehensive* guide to everything I could possibly ask about on the midterm. However, if you understand how to perform and interpret results of each procedure below, you are probably in good shape for the exam.

2.1 General Lab Hints

Save all of your scripts/dialogs and give them informative names! This means you will just have to open up the right script/dialog and follow its example. Look in the example problems and Sapling/lab assignments for tell-tale signs that a question will involve power analysis or a specific type of hypothesis test/confidence interval. Often, deciding the type of hypothesis test/confidence interval can be solved by answering four simple questions:

1. What is a case/unit/subject in this study?
2. What categorical variable(s) am I recording for each case, and how many possible values does each variable have?
3. What numerical variables am I recording for each case?
4. How many samples do I have, and are all the cases in my sample(s) independent?

2.2 Lab 14

- Create a histogram to graphically display a numerical variable
- Create a boxplot to graphically display a numerical variable
- Linearly transform a numerical variable (using *Transform* function in Rguroo or *mutate* command in R)

2.3 Labs 13, 15, and 17

- For a normal random variable/normal population distribution, find the probability of obtaining an *individual value* below a given value/above a given value/between two given values
- For a sampling distribution of sample mean, find the probability of obtaining a *sample mean value* below a given value/above a given value/between two given values
- For a t-distributed random variable, find the probability of obtaining a *t-statistic* below a given value/above a given value/between two given values
- Perform those procedures "in reverse" to find cumulative proportions/upper tail probabilities (i.e., using qnorm/qt or Probability $\rightarrow$ Values)

2.4 Labs 18-20

- Perform a one-sample t hypothesis test in the Neyman-Pearson framework and make an appropriate conclusion
- Compute the power and β for a one-sample t hypothesis test in the Neyman-Pearson framework (using Rguroo's Mean Inference $\rightarrow$ Details $\rightarrow$ Power Analysis or R's power.t.test function)
- Perform a one-sample t hypothesis test in the NHST framework and make an appropriate conclusion

- Add a variable to the dataset containing paired differences (using *Transform* function in Rguroo or *mutate* command in R)
- Perform a matched pairs t hypothesis test in the NHST framework and make an appropriate conclusion
- Create a set of histograms showing the distribution of a numerical variable in two or more groups
- Perform a two-sample t hypothesis test in the NHST framework and make an appopriate conclusion
- Create a set of boxplots showing the distribution of a numerical variable in two or more groups
- Perform a One-Way ANOVA hypothesis test (Fisher framework) and make an appropriate conclusion
- If the null hypothesis for a One-Way ANOVA hypothesis test is rejected, perform *post hoc* procedures and make an appropriate conclusion

2.5 Labs 21-22

- Construct a t confidence interval for population mean and interpret it
- Construct a t confidence interval for population mean of paired differences and interpret it
- Construct a t confidence interval for difference of population means and interpret it (in particular, which population mean is bigger and by how much)
- Determine whether a specific null hypothesis can be accepted (N-P framework) or rejected (NHST framework) based on the confidence interval