# Math 338 Final Exam Study Guide

Disclaimer: This exam is not intended to be a *comprehensive* guide to everything I could possibly ask about on the final exam. 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 1-3: Introduction to Probability

- Define a probability model (sample space and probability of each outcome in the sample space)
- Use axioms of probability, complement rule, and/or general addition rule to calculate a probability
- Classify two events as independent, disjoint, both or neither
- Write the probability mass function (pmf) of a discrete random variable
- Use the pmf to compute the expected value, variance, and standard deviation of a discrete random variable
- Use transformation rules to compute the expected value, variance, and standard deviation of a linear combination of discrete random variables ($aX + bY$)

### 1.2 Lectures 4-6: Sampling Distributions and Data Collection

- Identify the case/unit/subject about which we are recording data
- Identify whether a collection of cases is a population or a sample
- Identify whether a summary value is a parameter or a statistic
- 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
- Explain the difference between bias and variability of a sampling distribution
- Check the four conditions (BINS) that must be satisfied for data to be collected in the binomial setting
- Given parameters $n$ and $p$, compute the expected value and variance for a binomial random variable X
- Given parameters $n$ and $p$, compute the expected value and variance for a sample proportion $\hat{p}$
- Given two variables, identify which variable is most likely the explanatory variable and which is the response variable
- Identify whether a study is an observational study or an experiment
- Identify whether it would be both possible and ethical to perform an experiment to answer a research question
- Identify the levels of a factor and the treatments in an experiment
- Classify two explanatory variables as interacting variables, confounding variables, both or neither
- Given an experiment, identify whether the placebo effect would occur in the treatment group(s) only or in both the treatment and control groups.
- Apply the principles of control, randomization, and replication/repetition to identify potential flaws in an experimental design
- Classify an experimental design as completely randomized design, blocked design, or matched pairs design
- Classify a study as single-blind, double-blind, or not blind
- Critically evaluate real-world reasons why a study might incorrectly "prove" an effect (or fail to show an effect that does exist)

### 1.3 Lecture 7: Two-Way Tables and Conditional Probabilities

- Given diagnostic testing results, identify the number/proportion of true positives, true negatives, false positives, and false negatives in the sample.
- Use conditional proportions and/or probabilities to estimate the sensitivity, specificity, positive predictive value, and negative predictive value of a test.

- Use conditional probability to determine whether two events are independent
- Compute the conditional probability of one event given that a different event is known to have happened (by any means necessary; the simpler the better)
- Given a complicated conditional probability situation, set up the problem using a two-way table, tree diagram, and/or Bayes's Rule, and solve for a conditional probability

## 1.4 Lectures 8-9: Neyman-Pearson Hypothesis Testing

- Write the null hypothesis $H_0$ and the alternative hypothesis $H_1$ in the Neyman-Pearson framework
- Given a testing situation, identify what would be a Type I Error vs. Type II Error
- Given a set of conditional probabilities, identify $\alpha$, $\beta$, and power of the test
- Given $\alpha$ and $\beta$ values, identify whether the power of the test is sufficiently high to detect $H_1$ when it is true
- 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)

## 1.5 Lecture 10: Null Hypothesis Significance Testing

- Write the null hypothesis $H_0$ and the alternative hypothesis $H_a$ in the Null Hypothesis Significance Testing (NHST) framework
- Explain in context the idea of a p-value
- 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)

## 1.6 Lectures 11-12: Fisher's Significance Testing

- Write the (null) hypothesis for a goodness-of-fit test - specifically, I'm looking for the proportion of each category in your model of the population
- Write the (null) hypothesis for a test of independence - specifically, I'm looking for a statement that two categorical variables are not related (remember, you can write $H_0$ for a test of homogeneity exactly like a test of independence by making one variable the population)

- Compute the degrees of freedom parameter for a $\chi^2$ distribution, for both goodness-of-fit test and test of independence
- Decide whether the data represent a meaningful difference from the model or the model is a reasonable representation of reality, 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)
- Evaluate whether the data collection assumptions of the model are reasonable (specifically, this means to critically think about how/whether your sample would differ from other samples due to anything *other* than random chance)

### 1.7 Lectures 13-14: Numerical Variables and Continuous Random Variables

- Sketch the pdf for a uniform random variable and use it to find probabilities
- Use the 68-95-99.7 rule of thumb to estimate probabilities involving normal random variables
- Convert values to z-scores and explain why a z-score is used to compare values from different distributions
- Identify statistics/parameters as measures of center (average) or variability (spread, variation)
- Identify a density curve as skewed left/skewed right/symmetric and unimodal/multimodal
- Use the $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ convention to identify outliers
- Compute the new mean and new variance of a numerical variable after linear transformation
- Compute the new mean and new variance of a linear combination of two numerical variables

### 1.8 Lecture 15: Sampling Distribution of the Sample Mean

- Identify the difference between rounding error, measurement error, and sampling error
- 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
- Explain the difference between bias and variability of a sampling distribution
- Use the Central Limit Theorem to approximate the sampling distribution of a sample mean
- Make an educated guess about whether the Central Limit Theorem approximation is "good enough" given a sample size and the distribution of the sample

## 1.9 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)

## 1.10 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

### 1.11 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 $\alpha$
- 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)

### 1.12 Lecture 23: Scatterplots, Correlation, and Linear Regression

- Given a real-life situation, use "common sense" to identify the explanatory and response variables
- Given a scatterplot, identify the direction, form, and strength of the association, and identify possible outliers (in x, in y, or that don't fit the trend)
- Explain the concept of correlation and what the sign and magnitude of $r$ mean
- Write the equation of the least-squares regression line
- Interpret (give the meaning in context) the slope and intercept of the least-squares regression line and determine whether the statistic ($b_1$ or $b_0$) has a meaningful value
- Predict the value of y (compute $\hat{y}$) given a value of x
- Compute the residual corresponding to a particular (x,y) point and interpret the value
- Explain why extrapolation (computing $\hat{y}$ at x-values outside the original range of x-values used to fit the least-squares regression line) can be silly/dangerous
- Identify whether an outlier is an influential point

### 1.13 Lectures 24-26: Inference for Linear Regression and Multiple Linear Regression

- Write the population model for linear regression ($y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$) and explain what all the parameters and variables represent
- Identify and explain the four main assumptions of the population model for linear regression
- Use scatterplots, residual plots, and normal quantile (q-q) plots to determine whether assumptions of the population model are clearly violated
- Write the (null) hypothesis for an ANOVA F test for linear regression
- Given a linear regression situation, 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)
- Identify the "null model" and explain what it means for a linear model to be "not significantly better" than the nulll model

- Explain what the (Multiple) $r^2$ value represents and compute its value from an ANOVA table
- Write the null and alternative hypothesis for a t-test for slope (in both the simple and multiple linear regression frameworks)
- Identify the appropriate test statistic for a t-test for slope and the distribution it comes from (including degrees of freedom)
- Interpret a confidence interval for the slope of a regression line
- Interpret a confidence interval for the mean of y at a particular x-value
- Interpret a prediction interval for the value of y at a particular x-value
- Given an x-value, explain whether the confidence interval for $\mu_{y|x}$ or the prediction interval for $y$ will be wider and why
- Explain why we cannot compare two multiple regression models by simply comparing their $R^2$ values
- Explain the reasoning behind the technique of backward selection and identify the "least significant predictor" from a Coefficient Estimates table
- Interpret (give the meaning in context) the slopes and intercept of the regression equation in multiple linear regression (this is slightly different from simple linear regression!)
- Predict the value of y (compute $\hat{y}$) given values of *all* explanatory variables and compute the corresponding residual

# 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

The hardest part of every lab exam is *figuring out what the question is asking you to do.* Look in the example problems and lab assignments for tell-tale signs that a question will involve power analysis or a specific type of hypothesis test. Often, deciding the hypothesis test to use 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? (Hint: on Midterm 1, this answer is always "I'm not recording any")
4. How many samples do I have, and are all the cases in my sample(s) independent?

## 2.2 Lab 4

- Download a dataset from Titanium and import it into software
- Create a bar graph to summarize one or two categorical variables

## 2.3 Lab 5

- Compute the probability of getting exactly $X$ successes in the binomial setting
- Compute the probability of getting an interval of successes (e.g., more than 18, less than 6, at least 20, at most 45) in the binomial setting
- Compute the probability of getting exactly $\hat{p}$ proportion of successes in the binomial setting
- Compute the probability of getting an interval for $\hat{p}$ values in the binomial setting

## 2.4 Labs 8-9

- Compute the critical region for a hypothesis test in the Neyman-Pearson framework
- Compute the power and $\beta$ for a hypothesis test in the Neyman-Pearson framework

### 2.5 Labs 10-12

- Perform a binomial hypothesis test in the Neyman-Pearson framework and make an appropriate conclusion
- Perform a binomial hypothesis test in the NHST framework and make an appropriate conclusion
- Perform a goodness-of-fit test (either using a $\chi^2$ distribution or simulation as appropriate) and make an appropriate conclusion
- Perform a test of independence (either using a $\chi^2$ distribution or simulation as appropriate) and make an appropriate conclusion

### 2.6 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)

### 2.7 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)

### 2.8 Labs 18-20

- Perform a one-sample t hypothesis test in the Neyman-Pearson framework and make an appropriate conclusion
- Compute the power and $\beta$ 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

## 2.9 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

## 2.10 Labs 23-26

- Create a scatterplot to graphically display the relationship between two numerical variables (and, potentially, a categorical variable)
- Add the least-squares regression line (possibly for each group) to a scatterplot

- Using the *Linear Regression* module in Rguroo or the *lm* command in R, obtain the coefficient estimates table and use it to write the least-squares regression equation (simple or multiple linear regression)

- Produce a residual plot and a normal quantile (q-q) plot corresponding to a least-squares regression equation, and determine whether any assumptions of the population model are clearly violated

- Identify the values of the following statistics from the software output: coefficient of determination ($r^2$), observed value of the t-statistic for slope, degrees of freedom corresponding to that t-statistic, p-value for the t-test for slope, observed value of the F-statistic for ANOVA, degrees of freedom corresponding to that F-statistic

- Use the output to perform an ANOVA test for the overall model and make an appropriate conclusion

- Use the output to perform a t-test for an individual slope in the model and make an appropriate conclusion

- Construct (and interpret) a 95% confidence interval for the slope of the regression line

- Construct (and interpret) a 95% confidence interval for the mean of the response variable, given a value of the explanatory variable

- Construct (and interpret) a 95% prediction interval for the actual value of the response variable, given a value of the explanatory variable

- Create a scatterplot matrix to graphically display pairwise relationships between multiple numerical variables