Math 338 Midterm 1 Study Guide

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 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)

  A probability model allows for us to map each event occurrence to a probability value. This is considered a hash table. The sample space is the area of interest and there is an associated probability of each event taking place in said sample space.

- Use axioms of probability, complement rule, and/or general addition rule to calculate a probability

  Axioms are statements so obvious we don't need to prove. The compliment rule denotes the probability of event $P(A^c)$ occurring when knowing the probability of P(A). It's formula is denoted as $P(A^c) = 1 - P(A)$. Event probabilities **must** add up to 1.

- Classify two events as independent, disjoint, both or neither

  Independent events are when two events do not influence the outcome of one another. Disjoint events are the opposite.

- Write the probability mass function (pmf) of a discrete random variable

  The formula is $p(x) = P(X = x)$. This translates to "The probability of x is equal to an event denoted as X"

- Use the pmf to compute the expected value, variance, and standard deviation of a discrete random variable

  All of these can be calculated using the cheat sheet attached below.

- Use transformation rules to compute the expected value, variance, and standard deviation of a linear combination of discrete random variables (aX + bY)

  All of these can be calculated using the cheat sheet attached below.

1.2 Lectures 4-6: Sampling Distributions and Data Collection

- Identify the case/unit/subject about which we are recording data

  A case/unit/subject is an individual or entity of interest

- Identify whether a collection of cases is a population or a sample

  A collection of cases represents a population when it includes all ($\forall$) individuals of interest. A sample is a subset ($\subset$) of individuals of a given **population**.

- Identify whether a summary value is a parameter or a statistic

  A summary value is a scalar that represents the larger population.

- Identify whether a distribution is the distribution of a variable or the sampling distribution of a statistic

  A distribution of a variable is the relative number of times each possible outcome will occur in a given number of trials. A sampling distribution is the range of different outcomes that could possibly occur for a statistic (quantifying value of a population $\uparrow$).

- Identify whether a statistic is a biased or unbiased estimator of a parameter

  When the estimator is biased, it will deviate from the true value of the parameter. In turn, when the bias is low or non-existent, the true value of the parameter is known.

- Explain the difference between bias and variability of a sampling distribution

  Bias will influence the outcome of a parameter whereas the variability is how far the range of values diverge from the mean.

- Check the four conditions (BINS) that must be satisfied for data to be collected in the binomial setting

  Binary outcome, Independent, Number of trials, and Success is likely across **all** events.

- Given parameters n and p, compute the expected value and variance for a binomial random variable X

  The formula for expected value is going to be nP. Variance would be found applying the following: $nP(1-P)$.

- Given parameters n and p, compute the expected value and variance for a sample proportion $\hat{p}$

  The expected value is going to be nP and the variance of $\hat{p}$ would be $\frac{P(1-P)}{n}$.

- Given two variables, identify which variable is most likely the explanatory variable and which is the response variable

  The explanatory variable is an independent variable that by nature is not certain to be independent. A response variable is a dependent variable.

- Identify whether a study is an observational study or an experiment

  An observational study is when things are seen and data is recorded (hands-**off**) whereas an experiment is when variables are changed and those changes are recorded (hands-**on**)

- Identify whether it would be both possible and ethical to perform an experiment to answer a research question

  If it's something you and the people are around you are okay with doing. Testing if plants will die because of a lack of water is okay but doing it to dogs is **not**

- Identify the levels of a factor and the treatments in an experiment

  Levels of a factor are the number of variations of the factor that were used in the experiment. Treatment is the number of combinations of factor levels that can occur.

- Classify two explanatory variables as interacting variables, confounding variables, both or neither

  Interacting variables will directly influence other variables in the experiment. A confounding variable however, influences both the independent and dependent variable during the course of the experiment.

- Given an experiment, identify whether the placebo effect would occur in the treatment group(s) only or in both the treatment and control groups.

  Placebo effect is when the "fake" treatment can have no know benefits and but it does. This is due to the patient's *belief* in the treatment.

- Apply the principles of control, randomization, and replication/repetition to identify potential flaws in an experimental design

  Control allows for there to be consistency with the experiment, allowing it to be further tested. Replication is when similar data is acquired through the act of repeating the experiment by the same or different people. Randomization allows us to mitigate lurking variables other bias from permeating into the experiment.

- Classify an experimental design as completely randomized design, blocked design, or matched pairs design

  Completely random means there are no predefined requirements to create sample spaces from a given sample. Block design is when a sample is broken into two or more groups based on a characteristic and then given random assignment to a treatment. Matched pairs design: special case of block design with blocks size of two (only looking at one treatment with two levels)

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 a two way table.

- Use conditional proportions and/or probabilities to estimate the sensitivity, specificity, positive predictive value, and negative predictive value of a test.

  Sensitivity is on top of the tree and specificity is on the bottom

- Use conditional probability to determine whether two events are independent

  Conditional probability is when an event occurs because of another event. Independent events are not influenced by such behavior.

- 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)

  The formula for conditional probability is $P(B|A) = \frac{P(A \cap B)}{P(A)}$. This is read as "the probability of B given A"

- 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

  USE A TREE DIAGRAM. Just multiply down each branch.

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

  Null Hypothesis is the original hypothesis and our alternative hypothesis is what we fall back onto when the null hypothesis is not true.

- Given a testing situation, identify what would be a Type I Error vs. Type II Error

  A Type I Error is when we reject the null hypothesis when in fact the null hypothesis is true. A Type II Error is when the alternative hypothesis is false but it is actually true.

- Given a set of conditional probabilities, identify α, β, and power of the test

- Given α and β values, identify whether the power of the test is sufficiently high to detect H_1 when it is true

  If power is above 80

- 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 χ^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)

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 β 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 χ^2 distribution or simulation as appropriate) and make an appropriate conclusion

- Perform a test of independence (either using a χ^2 distribution or simulation as appropriate) and make an appropriate conclusion