MATH 147B Exam 1 + Post-Exam 3 Study Guide

David Collins and Thu Quan

December 6, 2022

Exam 1 Material

1 Chapter 1

- Data are numbers that correspond to some person or subject.
- Categorical/Nominal/Qualitative variables hold information about categories of a certain thing (age *groups*, race, gender, country of origin, etc.).
- Quantitative variables are numbers that represent the number of measured things (height, weight, age, etc.)
- A subject/participant is a human experimental unit.

2 Chapter 2

- Visual Display of Data
- We want to represent data visually to help understand them.
- The Area Principle says that each data value should be represented by the same amount of area in a graph or model.
- A frequency table is used for categorical variables: it gives the count/frequency of the observations in each box of the table. We can use this to find a relative frequency table.
- Bar charts, histograms, pie charts, dot plots, etc. are all models used to visually represent data. Histograms and bar charts will be used the most frequently in this class.
- The center of a distribution is represented by its mean or median, which we will define soon.
- The mode is a LOCAL high point in the distribution of a variable. Unimodal means having one mode, bimodal means having two modes, etc.
- A distribution is symmetric if the left side of the center looks roughly the same as the right side of the center.

- Distributions can have tails, where they trail off to a side. A left skew means that the distribution has a tail on the left side, and similarly for a right skew.
- An outlier is an extreme value that doesn't seem to belong with the rest of the data.
- Measures of Center and Spread
- The median is the middle value of the data, with half the data above and below it. *ALWAYS make sure to arrange the data in numerical order!!!* For an odd number of data points, it is exactly the middle number. For an even number of data points, add up the two middle values and divide by 2. For example, for the list 1,2,3,4, the median is (2+3)/2=2.5
- The mean (or average) is obtained by adding all the data values and dividing by the number of data values.

$$\overline{x} = \frac{\text{total}}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Spread is a numerical summary of how tightly the data values are clustered around the center. Common measures of spread are the IQR and standard deviation.
- The range is the maximum data value - minimum data value.
- A quartile represents the percentage of data BELOW it. For example, Q1 is the 25% quartile/percentile, where 25% of the data is below it.
- The IQR (interquartile range) is Q3-Q1, a.k.a. the 75% percentile - 25% percentile, so we get the middle 50% of the data.
- We define the standard deviation as

$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}} \tag{1}$$

- The variance is the square of the standard deviation, aka $Var(x) = s(x)^2$.

3 Chapter 3

- A contigency table shows the counts of two variables. There are also row and column totals, and an overall total in the corner.
- We can make the marginal distribution by focusing only on the row or column totals of a contigency table.
- A table percent is a given entry divided by the table total.
- Conditional distributions show the distribution of one variable given that they satisfy a condition on another variable.
- Two variables are independent if the distribution of one variable is the same for all categories of another variable (will come up again in Chapter 22!).

- Simpson's Paradox is when the averages taken across groups seem to contradict the overall averages.

- A lurking variable is one that affects the relationships between variables being studied but is not explicitly there. For example, COVID is a lurking variable in basically every social science experiment for the past 2 years.

4 Chapter 4

- We can make a boxplot from a five number summary of the data. The minimum is the lowest value of the dataset, and the maximum is the highest value. The median is the midpoint of the data. Similarly, 25% of the data is below Q1 and 75% of the data is below Q3.
- The IQR is Q3-Q1. We make the fences by taking Q1-1.5IQR for lower fence and Q3+1.5IQR for the upper fence.
- The fences are the boundary points. The whiskers are drawn at the lowest data point WITHIN the fences.
- Outliers are values that are below the lower fence or above the outer fence.

5 Chapter 5

- We let s be the standard deviation and $\overline{x}$ be the sample mean. To standardize values, we use this formula:
$$z = \frac{x - \overline{x}}{s} \tag{2}$$
- If we add a constant value to all the data values, we do not change the spread of the data, but it does affect the mean. For example, if the mean is 3 and we add 5 to every value, the mean is now 8. However, if we multiply all the data values, it multiples the mean and SD by that value.
- The 68-95-99.7 Rule refers to a Normal model. 68% of the data is contained within ± 1 standard deviation, 95% are within ± 2 SD, and 99.7% of the data are within 3 SD.
- We want Normal probability plots to look like a straight line. It's okay if the ends of the plots trail off a bit.

6 Chapter 6

- A scatterplot is a plot that shows the relationship between two quantitative variables.
- We can use scatterplots to find the association between variables.
- The direction of the association is either positive or negative. Form refers to whether the data looks like a linear, a curve or a bunch of points. Strength refers to how tightly clustered the data are.
- The variable of interest (the y variable) is called the response variable. The x variable (or variables) are called predictors or explanatory variables.
- The correlation coefficient is a measure of the strength of the relationship between the predictors and the response. This value is always between -1 and 1, and the formula is
$$r = \frac{\sum z_x z_y}{n-1} = \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2 (y-\overline{y})^2}} = \frac{\sum (x-\overline{x})(y-\overline{y})}{(n-1)s_x s_y} \tag{3}$$

- Before we can interpret correlation, we need to check 3 conditions: Quantitative Variables Condition (are the variables quantitative or categorical?), Straight Enough Condition (does the scatterplot look roughly linear?), and the No Outliers Condition (are there outliers or not?).
- Outliers can greatly affect correlation!
- IMPORTANT: Correlation is NOT causation. The only way to determine causation is by running an experiment!

7 Chapter 7

- A residual is a measure of error. The formula is

$$e = y - \hat{y} \tag{4}$$

where y is the observed value (the data) and $\overline{y}$ is the predicted value from the regression line.

- The regression equation is

$$\hat{y} = b_0 + b_1 x \tag{5}$$

where b_0 is the y-intercept and b_1 is the slope.

- The equation for the slope is

$$b_1 = r\frac{s_y}{s_x} \tag{6}$$

and the equation for the y-intercept is

$$b_0 = \overline{y} - b_1\overline{x} \tag{7}$$

since the point $(\overline{x}, \overline{y})$ is always a point on the regression line.

- As a result, we sometimes call it regression to the mean because of this. We also call it least squares regression, as we're trying to minimize the sum of the squares (minimize the distance between ALL data points and the regression line).
- We measure the standard deviation of the residuals by

$$s_e = \sqrt{\frac{\sum e^2}{n-2}} \tag{8}$$

- R^2 is the square of the correlation, r. It is the fraction of variability accounted for by the model.
- Make sure to turn R^2 into a decimal before square-rooting it. $\sqrt{R^2} = r$ and $r^2 = R^2$. Like correlation, R^2 is between 0% and 100% (we usually write R^2 as a percentage rather than a decimal). The higher the R^2 the better!

8 Chapter 8

- Interpolation is when we try to predict data within the existing range of data. This is usually very accurate and easy to do. Extrapolation is when we try to go further outside of the data range and predict new data. The further we go away from the existing range of data, the less accurate the extrapolation is.
- We predict the response by plugging into a given regression line.
- An outlier is a data point that is very far away from the rest of the data (in the x and/or y directions).
- A leverage point is a data point that is far away from the rest of the data in the x direction.

- An influential point changes the slope of the regression line when it is removed.
- Leverage and influential points can be outliers!
- A lurking variable changes the association between the predictor and response but is not present in the scenario. A low R^2 variable could be an indicator of the presence of a lurking variable.

9 Chapter 10

A *simple random sample* (SRS) is a sample in which each combination of elements in the population has an equal chance of being selected. The *sampling frame* is a list of individuals from which the sample is drawn. There are several ways to draw an SRS from a sampling frame. Two simple ideas are:

- Assign a random number with several digits (say, from 0 to 10,000) to each individual. Then sort the random numbers into numerical order, keeping each name with its number. The first n names are then a random sample of that size.
- Assign each individual a number (say, from 1 to 100), then choosing randomly 20 specific numbers in that 100 numbers and survey people who have that chosen numbers.

Other sampling designs are

- Stratified sampling: Sometimes the population is divided into several subpopulations, or strata first, then random samples are drawn from each stratum.
- Cluster Sampling: If there are smaller groups which are representative of the whole population, then each group is called a cluster. Selecting one or a few clusters at random, then drawing a random sample within each cluster.
- Multistage sampling: Sampling schemes that combine several methods.
- Systematic sample: A sample drawn by selecting individuals systematically from a sampling frame. For example, you might survey every 5th person *on an alphabetical list* of students.

Common sampling mistakes, or How to sample badly

- Mistake 1: Sample volunteers

  Voluntary response samples are often biased toward those with strong opinions or those who are strongly motivated. People with very negative opinions tend to respond more often than those with equally strong positive opinions.

- Mistake 2: Sample conveniently

  In convenience sampling, we simply include the individuals who are convinient for us to sample, which may not be representative of the population.

- Mistake 3: Use a bad sampling frame

  An SRS from an incomplete sampling frame introduces bias because the individuals included may differ from the ones not in the frame.

- Mistake 4: Undercoverage

  Some portion of the population is not sampled at all or has a smaller representation in the sample than it has in the population.

10 Chapter 11

11.1. Observational study

Observational studies includes: Retrospective study and Prospective study.

- Restropective study: An observational study in which subjects are selected and then their previous conditions or behaviours are determined.
- Prospective study: An observational study in which subjects are followed to observe future outcomes.

11.2. Randomized, comparative experiments

Experiments study the relationship between two or more variables. Many properties of an experiment

- Individuals take part in the experiment are called subjects/ participants/ experimental unit.
- An experiment must have at least one explanatory variable, called *factor*, to manipulate and at least one response variable to measure.
- A specific value is assigned for each factor is called *level*; the combination of specific levels from all the factors that an experimental unit receives is called *treatment*.

11.3. The four principles of experimental design

- Control: We control the factors in an experiment in two ways:
  Factors under study: decide on their levels and how they allocated to subjects.
  Factors affect outcome: control their levels so they don't vary.
- Randomize: assign subjects to treatment at random.
- Replicate: Two kinds of replication:
  apply each treatment to a number of subjects, or the entire experiment is repeated on a different population.
- Block: grouping similar individuals into homogenenous groups or blocks.

11.4. Control Groups

- Blinding: Who can affect the outcome of the experiment
  - those who could influence the results (subjects, treatment administrators, or technicians)
  - those who evaluate the results (judges, treating physicians, etc.)

  When all the individuals in either one of these classes are blinded, an experiment is said to be *single-blind*. When everyone in both classes is blinded, we call the experiment double-blinded.
- Placebo: A treatment known to have no effect, administratered so that all groups experience the same consitions.

Post Exam 3 Material

11 Chapter 9

9.1. What is Multiple Regression?

- A regression model with two or more predictor variables is called **multiple** regression. A regression on a single predictor is called a **simple** regression.
- In the multiple regression model, $R^2 (= r^2, r$ is the correlation) is still the fraction of the variability of y-variable accounted for by the multiple regression model.
- s_e (or s) is the standard deviation of residuals, which still shows how much the data points vary about the linear regression line. It also indicates how much the variation of the residuals in the scatterplot of residuals.

9.2. Interpreting Multiple Regression Coefficients

- The multiple regression model is

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k,$$

where $\hat{y}$ = dependent variable = response variable = variable of interest,

$x_1, \ldots, x_k$ = independent variables = predictors = explanatory variable.

($b_0, b_1, \ldots, b_k$ are taken from the coefficients column).

9.3. The Multiple Regression Model - Assumptions and Conditions

These assumptions and conditions are about the same as which we for a simple regression model, we will check these conditions for each of the predictors.

- Linearity Assumption: We'll check the Straight Enough Condition for each of the predictors.

  Straight Enough Condition: Scatterplots of y against each of the predictors are reasonably straight, here we'll just check that there isn't a bend or other nonlinearity.

  It is also a good idea to check the residuals for linearity, that the scatterplot of residuals against x-variable should not show any particular patterns, trends or clumping, especially bends or other nonlinearities.
- Equal Variance Assumption, or **Does the Plot Thicken? Condition**: The scatter of the residuals should be about the same everywhere. The scatterplot of residuals should should look patternless. Check in particular for any bend (which would also suggest that the data weren't straight after all) and for any thinkening, be alert for any tendency for the variability to grow or shrink in one part of the scatterplot.
- Normal Population Assumption (or Nearly Normal Condition and the Outlier Condition): The histogram of residuals should look unimodal, symmetric, and without outliers. Or the ***Normal probability plot*** (the qq plot) ***looks straight***.

12 Chapter 23.5,23.8

23.5. Multiple Regression Inference

- In a multiple regression, R^2 is still the fraction of the variability of y accounted for by the regression model. s is still the standard deviation of residuals.

- ***Degrees of freedom*** $= n$***(the sample size)*** $-$ ***#coefficients.***

  (The number of degrees of freedom in this model is equal the sample size n minus the number coefficients we have in the table.)

- ***Standard error*** is used for estimating the standard deviation of the sampling distribution for each coefficient; ***t-ratios*** are the same as t-statistic; ***P-value*** is still the value which we based our conclusion for the null hypothesis of the hypothesis test for each coefficient. Because we don't know the underlying population standard deviations, we use the standard errors which estimates the standard deviations.

Dependent variable is: %Body Fat
R squared = 71.3%
s = 4.460 with 250 − 3 = 247 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−3.10088	7.686	−0.403	0.6870
Waist	1.77309	0.0716	24.8	<0.0001
Height	−0.601539	0.1099	−5.47	<0.0001

Figure 1: The multiple regression table

Collinearity

- When two or more predictors are linearly related (or highly correlated), they are said to be **colinear**. The general problem of the predictors with linear relationships is called the problem of **collinearity** (check out tables on page 721 in the textbook for an example).

- So what we hope to see is the predictors are unrelated to each other, each provides new information to help account for more of the variation in y. A more stable model can be built when predictors have low correlation, the points are spread out, and the predictors vary in different ways so that the multiple regression has a stable base.

23.8. More about regression

Adjusted R^2

- R^2 tells us how much of the variance of y is accounted for by the regression model. In other words, R^2 is an overall measure of how successfully the model linearly relates y to x, a model with a higher R^2 is more desirable.

- Adjusted R^2: attempts to adjust R^2 by adding a *penalty* for each predictor in the model so that adding more predictors and *adjusting* R^2 *doesn't make the value bigger.*

The ANOVA Table

- Many computer regression tables include an additional table that looks like this:

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3.83413	1	3.83413	490
Residual	0.438420	56	0.007829	

Figure 2: The ANOVA Table for multiple regression

Each "Mean Square" is a Sum of Squares divided by its associated df. And the "F-ratio" is the ratio of the Mean Square for regression divided by the Mean Square residual. Specifically, the F-ratio is a statistic used to test the null hypothesis that all the coefficients (except the intercept) are equal1ly zero.

- In a simple regression model, we have just one predictor, so we can use the t-test for one slope coefficient, and get P-value for the test. Meanwhile in a multiple regression, the F-statistic tests the null hypothesis that *all* the coefficients are zero.

13 Chapter 25

25.1. Testing whether the Means of Several Groups are equal

- F-test: tests the null hypothesis that all the group means are equal. So the null hypothesis here is that *all the group means are equal*. The alternative hypothesis (always one-sided) is that *the group means are not equal* or *at least two group means in all group means are different.*

- We have two different estimates of variance:
  - The Error Mean Square (MS_E): is the estimate of variance using pooling, based only on the variation within groups around each of their own means, and *does not* depend at all on the null hypothesis being true.
  - The Treatment Mean Square (MS_T): is the estimate of variance under the assumption that the treatment means (or group means) are all equal.

- F-statistic: is the ratio $\text{MS}_\text{T}/\text{MS}_\text{E}$. When the null hypothesis is true, or the means between groups are equal, this ratio should be closed to 1. But, when the null hypothesis is false, the MS_T will be *larger*, or the numerator in the ratio $\text{MS}_\text{T}/\text{MS}_\text{E}$ tend to be larger than the denominator, and the ratio tend to be bigger than 1.

- F-distribution: is the sampling distribution for the ratio $\text{MS}_\text{T}/\text{MS}_\text{E}$ when the null hypothesis that the treatment means are equal.

  It has ***two degree of freedoms*** parameters:
  - Numerator df (comes from MS_T) $= k - 1$;
  - Denominator df (comes from MS_E) $= k(n-1)$,

  where n = sample size of each group (or observations in each group), k = the number of groups.

25.2. The ANOVA Table

Analysis of Variance Table

Source	DF	Sum of Squares	Mean Square	F-Ratio	P-Value
Method	3	29882	9960.64	7.0636	0.0011
Error	28	39484	1410.14		
Total	31	69366			

Figure 3: The ANOVA Table

The most important quantities are the F-statistic (F-ratio) and its associated P-value. When the F-statistic is large, the Treatment (*here is Method*) Mean Square (MS_T) is larger than the Error Mean Square (MS_E). The null hypothesis is false, and provides evidence that the means of the groups are not all equal.

25.3. Assumptions and Conditions

- Independence Assumptions: The groups must be independent of each other, the data within each treatment group must be independent as well. So we check the Randomization Condition: Were the data collected with suitable randomization?
- Equal Variance Assumption (or Similar Spread Condition): The ANOVA requires that the variances of the treatment groups be equal. There are some ways to check this condition.
  - Look at side-by-side boxplots of groups to see whether they have roughly the same spread. Look at the original boxplots of the response values, do the spreads seem to change *systematically* with the centers?
  - Look at the residuals plotted against the predicted value. This is about the same as Does the Plot Thicken? Condition.
- Normal Population Assumption (or Nearly Normal Condition): Check Normality with a histogram or a Normal Probability plot of all residuals together. The histogram should look unimodal, symmetric, without outliers. And the Normal Probability Plot should look roughly straight.

25.4. Comparing Means

We know how to compare two means with a t-test. But now we want to do several t-tests, comparing several– or even all–pairs of group means. One method for this multiple comparisons is the Bonferroni method. In this method, instead of using significance level α, it uses $\frac{\alpha}{J}$. That makes the confidence intervals wider because we find the interval at the confidence level $1-\frac{\alpha}{J}$ instead of the original $1-\alpha$, where J is the number of pairs we want to compare their means.

14 Conclusion

Whether you're reading this before exam 1 or when prepping for finals, just know that we're all here for you and we know that you'll all succeed on these exams!