Mathematical Notation

Math 113 - Introduction to Applied Statistics

Use Word or WordPerfect to recreate the following documents. Each article is worth 10 points and should be emailed to the instructor at james@richland.edu

Type your name at the top of each document. Include the title as part of what you type. The lines around the title aren't that important, but if you will type ----- at the beginning of a line and hit enter, both Word and WordPerfect will draw a line across the page for you.

For expressions or equations, you should use the equation editor in Word or WordPerfect. The instructor used WordPerfect and a 14 pt Times New Roman font with 0.75" margins, so they may not look exactly the same as your document. The equations were created using 14 pt font and appear a little larger than normal text. This is a quirk of WordPerfect, but it allows you to see what was done with the equation editor versus the word processor.

If there is an equation, put both sides of the equation into the same equation editor box instead of creating two objects. Be sure to use the proper symbols, there are some instances where more than one symbol may look the same, but they have different meanings and don't appear the same as what's on the assignment.

There are instructions on how to use the equation editor in a separate document or on the website. Be sure to read through the help it provides. There are some examples at the end that walk students through the more difficult problems. You will want to read the handout on using the equation editor if you have not used this software before.

If you fail to type your name on the document, you will lose 1 point. Don't type the hints or reminders that appear on the pages.

These notations are due before the beginning of class on the day of the exam for that material. For example, notation 1 is due on the day of exam 1. Late work will be accepted but will lose 20% of its value per class period. If I receive your emailed assignment more than one class period before it is due and you don't receive all 10 points, then I will email you back with things to correct so that you can get all the points. Any corrections need to be submitted by the due date and time or the original score will be used.

Remember to put your name at the top of the page.

Notation 1: Descriptive Statistics

Here are the basic graphs that may be appropriate for the type of data you have.

Number of Variables	Categorical Data (Counts)	Numerical Data (Measurements)
One Variable	Frequency Distribution Pie Chart Bar Chart	Histogram Box Plot Dot Plot Stem and Leaf Plot
Two Variables	Joint Frequency Table	Scatter Plot

Some Common Statistical Measures

- The **mean** is found by adding up all the values and dividing by the number of values. The sample mean is $\overline{x} = \frac{\sum x}{n}$.
- The **median** is the middle number after the data has been put into order. Half of the values are less than the median and half the values are more than the median. With an even number of values, the median will be the value that is halfway between the two middle numbers.
- The **variation** is the sum of the squares (abbreviated SS) of the deviations from the mean. You take each value, subtract the mean from it, square that difference, and then add up those squared deviations. The sample variation is $Variation = SS(x) = \sum (x - \overline{x})^2$. Two shortcut formulas for finding the variation are $SS(x) = \sum x^2 - n\overline{x}^2$ (only the $\overline{x}$ is squared) and $SS(x) = \sum x^2 - \left(\sum x\right)^2 / n$.
- The **variance** is the mean of the squared (abbreviation MS) deviations from the mean. You divide the variation by the degrees of freedom (df), which is $df = n - 1$ for now, to get the sample variance $s^2 = Variance = \frac{Variation}{df}$.
- The **standard deviation** is the average deviation from the mean and is the square root of the variance. The sample standard deviation is $s = \sqrt{Variance}$.

Remember to put your name at the top of the page.

If you transform your data by adding, subtracting, multiplying, or dividing by a constant, the summary statistics for the transformed data can be determined from the summary statistics for the original data. For measures of center or position, you apply the same transformation to the summary statistics that you applied to the data. Addition and subtraction do not affect measures of spread but multiplication and division do. The variance is always the square of the standard deviation.

Notation 2: Probability

The probability of an event is its long-run relative probability. The law of large numbers says that the long-run relative frequency of an event will get closer and closer to the true relative frequency as the number of trials increases. However, there is no law of averages that applies to the short-term.

All probabilities must be between 0 and 1 inclusive, $0 \le P(A) \le 1$. If an event is impossible, then the probability of it happening is 0. If an event must happen, then the probability of it happening is 1. The sum of the probabilities of all the different outcomes is one.

If all the outcomes are equally likely, then the probability of something happening is the number of ways it can occur divided by the total number of outcomes.

The complement of an event is everything that is not that event. The probability that something won't happen is one minus the probability that it will happen, $P(A') = 1 - P(A)$. One case where this is often used is when you want to find the probability of "at least one" of something. The complement of "at least one" is "none" and it is easier to subtract the probability of "none" from 1 than it is to add together all the outcomes that are "at least one."

If two events are mutually exclusive (disjoint), then the probability of one or the other occurring is the sum of their probabilities. If they aren't disjoint, then you add together their probabilities and subtract what they have in common.

The probability of two or more independent (not related) events both happening is found by multiplying their probabilities together.

Tree diagrams are useful for finding probabilities of compound events. To find the probability of reaching a point on the tree, you multiply the probabilities along all the

branches getting to that point. At each point in a tree diagram, the sum of the probabilities of all the branches from a single point must be one.

The mean of a discrete probability distribution is also called its expected value and is $\mu = E(x) = \sum xp(x)$. The variance of a probability distribution is $\sigma^2 = \left(\sum x^2 p(x)\right) - \mu^2$.

A binomial experiment is a fixed number of independent trials each having exactly two possible outcomes. The mean of a binomial distribution is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$.

The normal distribution is unimodal, symmetric, and bell-shaped. The empirical rule states that all normal distributions have approximately 68% of the data within one standard deviation of the mean, 95% of the data within two standard deviations of the mean, and 99.7% of the data within three standard deviations of the mean.

A non-standard model can be standardized by subtracting the mean and dividing by the standard deviation.

Notation 3: Inferential Statistics

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of the sample proportions, $\hat{p}$, is modeled by a normal model with the $Mean(\hat{p}) = p$ and the $SD(\hat{p}) = \sqrt{pq/n}$. When we are creating confidence intervals, we usually do not know the population proportion *p*, so we use the standard error of the proportion, $SE(\hat{p}) = \sqrt{\hat{p}\hat{q}/n}$.

When the values are independent and the sample size is large enough, the sampling distribution of the sample means, $\overline{x}$, is modeled by a normal model with the $Mean(\overline{x}) = \mu$ and the $SD(\overline{x}) = \sigma/\sqrt{n}$. When the population standard deviation is not known, which is most of the time, we find the standard error of the mean, $SE(\overline{x}) = s/\sqrt{n}$, and use the Student's t distribution instead of the normal model.

The Student's t distribution is very similar to the standard normal distribution except the standard deviation is greater than one. When looking at a graph, the t curve is flatter in the middle and wider in the tails than the normal curve. There are actually many t distributions, one for each degree of freedom, but as the sample size increases, the t distributions approach the standard normal distribution.

The critical value is based on the model being used and the level of confidence. The center of the confidence interval for the population parameter is the sample statistic (either the sample proportion $\hat{p}$ or the sample mean $\overline{x}$). The distance from the center of the interval to either endpoint is called the margin of error (ME) or maximum error of the estimate. The margin of error is the critical value times the standard error, $ME = CV \times SE$.

All hypothesis testing is done under the assumption that the null hypothesis is true. If the results we get are too unusual to happen by chance alone, then we reject our assumption that the null hypothesis is true.

The null hypothesis, H_0, is a statement of no change from the normal or assumed condition and always contains the equal sign. Our decision is always based on the null hypothesis and is either to reject or retain the null hypothesis. If the claim involves the null hypothesis, then we will use the word "reject" in our conclusion. We will never accept or support the null hypothesis.

The alternative hypothesis, H_1, is a statement of change from the normal or assumed condition and never contains the equal sign. The alternative hypothesis is used to determine whether the test is a left tail, right tail, or two tail test. If the claim is the alternative hypothesis, then we will use the word "support" in our conclusion.

The critical value is a pre-determined value based on the model, not on the sample data. It separates the critical region, where the null hypothesis is rejected, from the non-critical region, where the null hypothesis is retained. The test statistic is a value that is based on the sample data and the decision to reject or retain the null hypothesis is based on which region where the test statistic falls. The test statistic for a normal distribution is $z = \frac{value - mean}{SD}$ and the test statistic for a t distribution is $t = \frac{value - mean}{SE}$.

The p-value is the probability of getting the results we did if the null hypothesis is true. The level of significance, α, is how unusual we require something to be before saying it's too unusual to happen by chance alone. α is also how willing we are to make a type I

error. We will reject the null hypothesis if the p-value is less than the level of significance and retain the null hypothesis if the p-value is greater than the level of significance.

Besides looking at the test statistic to see whether or not it lies in the critical region, you can also look at the confidence interval to see whether or not it contains the hypothesized or claimed value. Since the confidence intervals are the believable values, we'll retain the null hypothesis if the claimed value falls in the confidence interval.

Those last three paragraphs can be summarized with the following decision rules.

- Reject the null hypothesis if the test statistic falls in the critical region.
- Reject the null hypothesis if the p-value is less than the significance level.
- Reject the null hypothesis if the claimed value does not fall in the confidence interval.

When the null hypothesis is rejected, it is because you had enough evidence to conclude that your results are too unusual to happen by chance. If there is not enough evidence, then you go on retaining the null hypothesis.

When working with two independent samples, we need to know how variables behave when we combine them. The mean of a difference is the difference of the means $Mean(x-y)=Mean(x)-Mean(y)$, but the variance of a difference is the sum of the variances $Var(x-y)=Var(x)+Var(y)$.

When working with two independent means, you can use two formulas, one with a pooled variance and one without. Most of the time, it is safer and easier to assume the variances are not equal and not pool them together. Sometimes it makes sense to assume that the variances are equal and we'll make that assumption when we work Analysis of Variance problems.

If you have paired data, then you create a new variable *d* that is the difference between the two observations and then work it out as a test about a single population mean.

Notation 4: Advanced Topics

The χ^2 goodness of fit test checks the hypothesis of whether the claimed proportions are correct. It does this by comparing the observed frequency of categories to their expected frequencies and seeing how close they are. If the observations are close to what is

expected, then the test statistic is small and we retain the null hypothesis. If the observations are not close to what we expected, then we reject the null hypothesis and say the claimed values are not correct. Because we reject the null hypothesis only when the differences are large, the χ^2 goodness of fit test is always a right tail test. The degrees of freedom is one less than the number of categories and the test statistic is

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}.$$

The contingency table can be used to check for independence between two categorical variables. It uses the same formula for the test statistic as the goodness of fit test, but the degrees of freedom is $df = df_{row} \times df_{col}$. The expected frequency for each cell is Row Total × Column Total ÷ Grand Total.

The one-way ANOVA compares three or more means to see if they're equal. The null hypothesis is that the means are equal and the alternative is that at least one of the means is different. The "one-way" part is because the data is categorized in exactly one way, much like a goodness of fit test.

Here is what a typical one-way ANOVA table looks like.

Source	SS	df	MS	F	p
Between (Factor)	1820	4	455	3.50	0.033
Within (Error)	1950	15	130		
Total	3770	19	198.42		

The values in the SS (Sum of Squares) column are variations and the values in the MS (Mean Square) column are sample variances. To find the MS, you divide the SS by the df. The F test statistic is the ratio of two sample variances and is found by dividing the MS for the source by the MS(Error).

The two-way ANOVA actually has three tests rolled into one. The data is categorized two ways instead of one and two one-way ANOVAs are performed, one for each way of classification. The third test is to see if there is any interaction between the two classification systems, which is similar to the test for independence.

The correlation coefficient is a measure of how well the line fits the data. The correlation coefficient, *r*, is always between -1 and +1. Correlations near zero correspond to weak or

no linear correlation. Changing the order of the x and y variables won't change the value of *r*. Changing the scale on either variable won't change the value for *r* because it is based on the standardized score. There are no units on *r*, but it is sensitive to outliers. You cannot correlate categorical variables and do not make the mistake of assuming that because two variables are correlated that one causes the other.

The correlation coefficient *r* can be found using the formula $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$ or the simpler form $r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$.

When looking at a scatter plot, there are three things you should look for. Scatter: how spread out is the data? Form: is the plot linear, quadratic, or some other shape? and Direction: does the pattern have a positive direction or a negative one?

The best fit line will always pass through the centroid of the data, which is the point $(\bar{x},\bar{y})$, and the equation of the line is $\hat{y} = b_0 + b_1 x$ where b_0 and b_1 are the y-intercept and slope of the line. The slope is related to the correlation coefficient using the equation $b_1 = r\frac{s_y}{s_x} = \frac{SS(xy)}{SS(x)}$. If there is no significant correlation, then the best fit line is horizontal and the estimated value is always the mean of the response variable, $\hat{y} = \bar{y}$.

The coefficient of determination, r^2, is the percent of the variation that can be explained by the regression equation. The higher the coefficient of determination, the better the model, but there is no magic number for how large it should be to consider the model good.

For multiple regression, there is one response variable and several predictor variables. One should look at the adjusted-R^2, rather than the R^2, when determining the best model. The adjusted-R^2 takes into account the sample size and the number of independent variables. The R^2 and adjusted-R^2 have similar formulas, with the R^2 using the variations (SS) while the adjusted-R^2 uses the variances (MS).

$$R^2 = \frac{SS(total) - SS(residual)}{SS(total)} \qquad Adj - R^2 = \frac{MS(total) - MS(residual)}{MS(total)}$$