Mathematical Notation

Math 113 - Introduction to Applied Statistics

Name : _____

Use Word or WordPerfect to recreate the following documents. Each article is worth 10 points and can be printed and given to the instructor or emailed to the instructor at james@richland.edu. If you use Microsoft Works to create the documents, then you must print it out and give it to the instructor as he can't open those files.

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 documents were created using a 14 pt Times New Roman font with standard 1" margins.

For individual symbols (μ, σ, etc), you can insert symbols. In Word, use "Insert / Symbol" and choose the Symbol font. For WordPerfect, use Ctrl-W and choose the Greek set. For more complex expressions you should use the equation editor. If there is an equation, put both sides of the equation into the same equation editor box.

There are instructions on how to use the equation editor in a separate document. 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 page, you will lose 1 point.

These notations are due at the beginning of class on the day of the exam for that unit. That is, the unit 1 notation is due on the day of the unit 1 test. Late work will be accepted but will lose 10% per class period.

Chapters 1-9

Always ask who, what, where, when, why, and how to determine the context for the data values.

The three rules of data analysis are 1) make a picture, 2) make a picture, and 3) make a picture.

Finding the standard deviation by hand is a four step process.

1. The sample mean, $\overline{y}$, is found by adding up all the values and dividing by the number of values.
2. The variation of a sample is also known as the sum of the squares (SS) because you find how far each value deviates from the mean, square that deviation, and then add them up.
3. The sample variance, s^2, is also known as a mean square (MS) and is found by dividing the variation by the degrees of freedom (which is *n*-1 in this case).
4. The standard deviation, s, is the found by taking the square root of the variance.

Consider the five numbers 3, 8, 2, 4, and 7. The sum of the original values is 24, so divide that by 5 to get the mean $\overline{y} = 4.8$.

y	3	8	2	4	7	24
$y - \overline{y}$	-1.8	3.2	-2.8	-0.8	2.2	0
$(y - \overline{y})^2$	3.24	10.24	7.84	0.64	4.84	26.8

In the table above, the 26.8 is the variation. Divide that by 5-1=4 to get the variance of $s^2 = 6.7$. The standard deviation is $s = \sqrt{6.7} \approx 2.59$.

As you can see, that's pretty involved, so most of the time, we'll just let technology find the standard deviation for us.

To standardize a variable, we take the deviation from the mean and divide it by the standard deviation. This is called a z-score and is found using the formula $z = \frac{y - \overline{y}}{s}$.

The z-score is a ruler for how many standard deviations a value is from the mean. Standardizing doesn't change the shape of a distribution, but it makes the mean 0 and the standard deviation 1.

Scatter plots are the best way to start observing the relationship between two quantitative variables. When we make a scatter plot, we want to look for direction, form, and scatter.

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*. Correlation is sensitive to outliers.

Do not correlate categorical variables, check the scatter plot to make sure the association is linear, and beware of outliers. Correlation does not imply causation.

The best fit line will always pass through the centroid of the data. For standardized scores, the centroid will always be $(0,0)$ and the equation of the line is $\hat{z}_y = rz_x$. If the scores aren't standardized, then centroid is the point $(\overline{x}, \overline{y})$, and the equation of the line is $\hat{y} = b_0 + b_1 x$ where b_0 and b_1 are the constant and slope of the line and are best found by the computer.

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.

Chapters 11-17

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 have to 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 disjoint events must be one.

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, $P(A') = 1 - P(A)$.

If two events are mutually exclusive (disjoint), then the probability of one or the other occurring is the sum of their probabilities. If two events are independent (not related), then the probability of both occurring is the product of their probabilities.

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

Tree diagrams are useful for finding probabilities of compound events. To find the probability of reaching the end of a branch, you multiply the probabilities on 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 probability distribution is its expected value $\mu = E(x) = \sum xp(x)$. 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.

Chapters 18-25

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 $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$. If we don't know the population proportion *p*, then we'll use the standard error of the proportion, $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$.

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of the sample means, $\overline{y}$, is modeled by a normal model with the $Mean(\overline{y}) = \mu$ and the $SD(\overline{y}) = \frac{\sigma}{\sqrt{n}}$. If we don't know the population standard deviation σ, then we'll use the standard error of the mean, $SE(\overline{y}) = \frac{s}{\sqrt{n}}$.

The center of the confidence interval for the population parameter is the sample statistic. The distance from the center of the interval to either endpoint is called the margin of error or maximum error of the estimate. The margin of error is the critical value times the standard error.

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 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 and never contains the equal sign. If the claim is the alternative hypothesis, then we will use the word "support" in our conclusion.

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. We will reject the null hypothesis is 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 value. Since the confidence intervals are the believable values, if the claimed mean falls in the confidence interval, we'll retain it.

The Student's t distribution is very similar to the standard normal distribution except the standard deviation is greater than one. Graphically, this means that the 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 student's t distributions approach the standard normal distribution.

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 not assume equal variances or pool the variances. Sometimes it makes sense to assume that the variances are equal and we'll make that assumption when we work Analysis of Variance problems in Chapter 28.

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

Chapters 26-29

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. A small difference would mean that we retain the null hypothesis while a large difference would mean that the claim is wrong and we would reject the null hypothesis. The goodness of fit test is therefore always a right tail test. The degrees of freedom is one less than the number of categories and the test statistic is

$$\chi^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}.$$

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.

Looking at the standardized residuals can often help us figure out which values are different (if there are any). The standardized residual is $c = \frac{\text{obs} - \text{exp}}{\sqrt{\text{exp}}}$.

The one-way ANOVA compares more than two 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 row by the MS(Error).

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

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).

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

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