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.

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.

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.

The due date for each of these documents is the day of the exam for that chapter. Late work will be accepted but will lose 10% per class period.

Part 1: Exploring and Understanding Data

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.

The mean of a sample is found by adding up all the values and dividing by the number of values. Mathematically, this can be written as $\bar{y} = \frac{\sum y}{n}$.

The variation of a sample is also known as the sum of the squares because you find how far each value deviates from the mean, square that deviation, and then add them up. Mathematically, this can be written as $\text{Variation} = \sum (y - \bar{y})^2$

The variance of a sample is found by dividing the variation by the degrees of freedom. The variance can be written as $s^2 = \frac{\sum (y - \bar{y})^2}{n-1}$

The standard deviation is the square root of the variance, so it can be found by the formula $s = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}$. Please note that the formula in your book on page 65 is missing the square root. We almost always find the standard deviation using technology.

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 $z = \frac{y - \bar{y}}{s}$. 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.

Part 2: Exploring Relationships Between Variables

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

The correlation coefficient is a measure of how well the line fits the data. It can be found by several formulas, one of involves the standardized scores and is

$$r = \frac{\sum z_x z_y}{n-1}$$

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 scatterplot 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.

Part 3: Gathering Data

When performing simulations, follow these seven steps.

1. Identify the component to be repeated
2. Explain how you will model the outcome
3. Explain how you will simulate the trial
4. State clearly what the response variable is
5. Run several trials
6. Analyze the response variable
7. State your conclusion in the context of the problem

The population is the entire group of objects and is often impossible or impractical to examine. Instead, we'll look at a smaller group of individuals called a sample.

We usually use Greek letters to represent population parameters and Latin letters to represent sample statistics. The one main difference is for proportions, where using the Greek letter π would be too confusing for people.

Name	Statistic	Parameter
Mean	$\overline{y}$	μ
Standard Deviation	s	σ
Correlation	r	ρ
Regression coefficient	b	β
Proportion	$\hat{p}$	p

Statistically, the best type of sampling to use is the simple random sample (SRS) where each individual has an equal chance of being selected. However, stratified sampling, cluster sampling, and systematic sampling are also used. Unfortunately, convenience sampling is used too often because it's the worst type of sampling.

The four principles of experimental design are control, randomize, replicate, and block.

Part 4: From Randomness to 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 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 outcomes each having exactly two possible outcomes.

Part 5: From the Data at Hand to the World at Large

Provided that the sampled values are independent and teh sample size is large enough, the sampling distribution of $\hat{p}$ is modeled by a normal model with a mean of $\mu(\hat{p}) = p$ and a standard deviation of $\sigma(\hat{p}) = SD(\hat{p}) = \sqrt{\frac{pq}{n}}$.

As the sample size, *n*, increases, the mean of *n* independent values has a sampling distribution that tends toward a normal model with a mean, $\mu(\overline{y})$, equal to the population mean, μ, and standard deviation $\sigma(\overline{y}) = SD(\overline{y}) = \frac{\sigma}{\sqrt{n}}$.

The center of the confidence interval for the population proportion is the sample proportion. The distance from the center of the interval to either endpoint is called the margin of error or maximum error of the estimate.

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 fail to reject the null hypothesis if the p-value is greater than the level of significance.

Part 6: Learning About the World

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.

If you know the population standard deviation, σ, then you can use the normal distribution and your test statistic will be $z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$. If you have to estimate teh population standard deviation with the sample standard deviation, then you should use the student's t distribution and you get a test statistic of $t = \frac{\bar{y} - \mu}{s / \sqrt{n}}$ with *n*-1 degrees of freedom.

The confidence interval for the population mean is centered about the sample mean. The margin of error or maximum error of the estimate is the distance from the center to either one of the endpoints.

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. The decisions are different, if the test statistic is in the critical region, the confidence interval will not contain the claimed value.

When working with two independent means, you can use two formulas, one with a pooled variance and one without. Most of the time, its safer and easier to not assume equal variances and 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 in Chapter 28.

If you have paired data, then you can just find the difference between the two observations and then work it out as a test about a single population mean.

Part 7: Inferences When Variables Are Related

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 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 are the degrees of freedom for the row times the degrees of freedom for the column. The expected frequency for each cell is the product of the row and column totals divided by the grand total.

Looking at the residuals can often help us figure out which values are different (if there are any).

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.

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 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 are several predictor variables and only one response variable. 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.