PSY 216 Exam 2 Review

Correlation is a descriptive statistic that tells you if two (or more) variables are related to each other

- Perfect correlations (magnitude of 1) allow you to predict perfectly
- Less than perfect correlations will not have perfect predictions in general
- Source of variance – a variable that can explain some of the variation in another variable. Some possible sources of variance of GPA include IQ, class load, work load, number of times per week you go out, etc.
- Pearson's Product Moment Correlation Coefficient (r)
    - must be in the range of –1 to +1
    - sign states the direction of the relation
        - positive sign = direct relation = as one variable increases, the other variable tends to increase as well
        - negative sign = indirect relation = as one variable increases, the other variable tends to decrease
    - magnitude – the absolute value of r – shows the strength of the relation
        - closer the magnitude is to 1, the more strongly (perfectly) the variables are related and the more perfectly you will be able to predict one given the other
            - magnitude = 1, then all points on scatterplot fall on straight line
        - close to 0 either implies no relation between the variables or the assumptions of r have been violated
            - magnitude = 0, (assuming no relation), scatterplot is circular
    - assumptions:
        - linear relation
            - the equation of the regression line is of the form Y' = slope * X + intercept
            - violations of this assumption can decrease the magnitude of r
            - can sometimes be corrected by taking appropriate mathematical transformation
        - non truncated range
            - the variability of each variable is reasonably close to its maximum possible value
            - violations of this assumption can decrease the magnitude of r
        - sample size
            - small samples can lead to spurious relations (relations that do not really exist in the population)
            - large samples can make very small rs statistically significant

z-scores transform the scale of a distribution so that the mean = 0 and the standard deviation = 1. This allows you to more easily compare values from two different distributions

- $z = \frac{X - \overline{X}}{s}$ where X = score being converted, $\overline{X}$ = mean, and s = standard deviation
- Three important properties of a set of z-scores
    - mean of a set of z scores = 0 = μ_z
    - standard deviation of a set of z-scores = 1 = s_z
    - sum of the squared z-scores = N = Σz^2 where N = number of scores in the distribution

Pearson's r is defined as $r = \frac{\sum z_X \cdot z_Y}{N}$ which is the mean of the product of the z scores

- The computational formula is algebraically equivalent, but is less subject to round off errors and can be faster to calculate than the above conceptual / theoretical formula

The coefficient of determination

- is the proportion of variability in one variable that is explainable by variation in the other variable
- tells us how well we can predict – the larger the value, the better the prediction will be in general
- equals r^2

The coefficient of non-determination

- is the proportion of variability in one variable that is not explainable by variation in the other variable
- plus the coefficient of determination must equal 1
- equals $1 - r^2$

Correlation cannot show causation because

- correlation cannot establish the direction of the causal link (does X cause Y or does Y cause X)

- a third variable could cause both X and Y to covary

There are numerous special correlation coefficients; which to use is primarily determined by the level of measure of the variables being correlated

Regression tells us the slope and intercept of the line that predicts the value of one variable given the value of the other

- Y' = slope * X + intercept
- Slope = how steep the line is – change in Y / change in X: slope for predicting Y given X = $r\, s_y / s_x$
- Intercept = where the line strikes the Y axis when X = 0: Intercept for predicting Y given X = $\overline{Y} - \text{slope} \cdot \overline{X}$
- The regression line (line of best fit, least squares line) minimizes the squared errors in prediction $\Sigma(Y - Y')^2$

Types of variation in regression

- Total variation – mean of the squared deviate scores $\frac{\sum(Y - \overline{Y})^2}{N}$
- Explained variation – mean of the squared predicted deviate scores $\frac{\sum(Y' - \overline{Y})^2}{N}$
- Unexplained variation – mean of the squared error deviate scores $\frac{\sum(Y - Y')^2}{N}$
- Total variation = explained variation + unexplained variation
- r^2 = explained variation / total variation

Probability

- ratio of the occurrences of a given event and the occurrence of all events
- mutually exclusive events – the occurrence of one event precludes the occurrence of the other event
- addition rule for mutually exclusive events: p(A or B) = p(A) + p(B)
- addition rule for non-mutually exclusive events: p(A or B) = p(A) + p(B) – p(A and B)
- independent events – knowing that one event has occurred tells you nothing about whether the other had occurred
- multiplication rule for independent events: p(A and B) = p(A) X p(B)
- joint probability – probability that two (or more) events happen together
- marginal probability – probability that just one (of several) events happens
- conditional probability – probability that one event occurs given that another event has already occurred
    - p(B | A) = p(A and B) / p(A) = probability of event B given that event A has occurred = joint probability of events A and B occurring divided by the marginal probability of event A occurring
- multiplication rule for non-independent events = p(A and B) = p(A) X p(B | A)

Continuous Probability

- Cannot use counting rules; must use area under the probability distribution (area under the curve between two points / total area under the curve)
- Unit normal curve is a normal distribution with a standard deviation of 1 and an area under it of 1
- If variable is normally distributed, convert the scores to z-scores, consult a table of areas under the unit normal, and apply typical probability rules