

STATISTICAL PROCEDURE TEMPLATE

Basics

Summary of types of research designs and their characteristics

<p>Measurement Model</p> <p>Nominal: Different values indicate a difference in the characteristic being measured.</p> <p>Ordinal: different values indicate a difference in relative amount of the characteristic being measured.</p> <p>Interval: Equal intervals indicate equal differences in amount of the characteristic being measured.</p> <p>Ratio: Ratios of variable values indicate proportional amounts of the characteristic being measured</p> <p>Mathematical Model</p> <p>Continuous: No boundaries between adjoining values. Includes most interval and ratio variables that do not involve counting, and ordinal variables that are not rank orders.</p> <p>Discrete: Clear boundaries between values. Includes nominal variables, counting variables, and rank orders.</p> <p>Measures of Frequency</p> <p>frequency: the frequency of occurrence for each variable value</p> <p>proportion, percent: the relative frequency of occurrence for each variable value</p> <p>Measures of Central Tendency</p> <p>mode: the most commonly occurring value</p> <p>mean: an indicator of central tendency sensitive to the exact position of each score in the distribution</p> <p>median: the middle score in a distribution</p> <p>Measures of Variability</p> <p>range: the difference between the lowest and highest scores</p> <p>variance: the average squared deviation from the mean</p> <p>standard deviation: an indicator of average deviation from mean in the same units of measure as the original scores</p> <p>Measures of Linear Relationship</p> <p>Pearson product-moment correlation coefficient: the strength and direction of the relationship between two variables</p> <p>phi coefficient: a variant of the Pearson statistic for two dichotomous variables</p> <p>point-biserial correlation coefficient; a variant of the Pearson statistic where one variable is dichotomous</p> <p>Spearman rank-order correlation coefficient: a variant of the Pearson statistic for two rank-ordered Variables</p> <p>Simple regression: indicates the best estimate of one variable from the values of another</p>	<p>Parametric Tests of pattern Hypotheses</p> <p>one-group z test: whether a population μ differs from some predefined value where σ is known</p> <p>one-group t test: whether a population μ differs from some predefined value; knowing σ is Unnecessary</p> <p>Nonparametric Tests of pattern Hypotheses</p> <p>Binomial test: whether a population probability for a dichotomous variable differs from some predefined value</p> <p>χ^2 goodness of fit test: whether a set of population probabilities differ from a set of predefined values</p> <p>Parametric Tests of Relationship Hypotheses</p> <p>dependent groups t test: whether a dichotomous variable is related to a quantitative variable in the population when groups defined by the dichotomous variable are dependent</p> <p>independent groups t test: whether a dichotomous variable is related to a quantitative variable in the population when groups defined by the dichotomous variable are independent</p> <p>t test for a correlation coefficient: whether the population ρ_{xy} between two variables differs from 0</p> <p>one-way independent groups ANOVA: whether a discrete variable is related to a quantitative variable in the population when groups defined by the discrete variable are independent</p> <p>one-way dependent groups ANOVA: whether a discrete variable is related to a quantitative variable in the population when groups defined by the discrete variable are dependent</p> <p>two-way independent groups ANOVA: whether the combination of two discrete variables is related to an interval or ratio variable in the population when groups defined by the discrete variables are independent</p> <p>Nonparametric Tests of Relationship Hypotheses</p> <p>χ^2 test of independence: whether two variables demonstrate dependent outcomes in the population</p> <p>Wilcoxon T test: whether a dichotomous variable is related to an ordinal, interval, or ratio variable in the population when groups defined by the dichotomous variable are dependent</p> <p>Mann-Whitney U test: whether a dichotomous variable is related to an ordinal, interval, or ratio variable in the population when groups defined by the dichotomous variable are independent</p> <p>Kruskal-Wallis H test: whether a discrete variable is related to an ordinal, interval, or ratio variable in the population when groups defined by the discrete variable are independent</p>
---	---

Design	Research question	Process	Causal inference
Descriptive	What are the characteristics of their group or this environment?	Each variable analyzed separately	No
Correlation	What is the relation between the two variables _ and _?	At least two measures per subject are needed\	No
Intact group comparison	Do the two groups differ in terms of a specific variable _?	Subjects assigned to groups based on subject variable. Subjects come into the study with that characteristic. They are not assigned that characteristic	No
True experiment	What is the effect of _ (the independent variable) on _ (the dependent variable)?	Subjects are randomly assigned to a treatment group (the independent	Yes. Only in a true experiment is it possible to conclude that one variable has an effect on the dependent measure

		variable) to assess the effect of the variable on a measure of interest (the dependent variable)	
--	--	--	--

Measurement scales and their characteristics		
Scale	Characteristic	Examples
Nominal	Unordered category	Sex of subject: male=1 female=2. DSM-IV diagnosis. Religion of subject. Political party
Ordinal	Order	Rank in class: 1 st , 2 nd , 3 rd . Rank on personality test: high vs. low self-esteem. Respect scale: 10=high respect, 1=low respect
Interval	Order, equal intervals, and arbitrary zero point	Temperature of 98.6F. SAT verbal score of 540. Score on an IQ test. Raw score on a statistics test
Ratio	Order, equal intervals and real zero point	Height, weight, age, running speed. Number of words recalled in a memory experiment. Response latency. Grams of food consumed. Number of problems solved

Brief summary of procedures for calculating univariate, bivariate, and multivariate statistics	
1.	Descriptive statistics make sense out of raw observations. By grouping the data or isolating certain aspects of the data, we can discover the essential amidst the irrelevant. Univariate analysis summarizes and presents single variable data. A frequency distribution can portray one variable in visual form such as the histogram or frequency polygon. Descriptive statistics usually present the central tendency (mode, median, and mean) and dispersion (range, standard deviation).
2.	Other descriptive statistics include skewness and kurtosis. The z score provides a convenient measure of one observation's relative standing. Bivariate analysis describes relationships between two variables. If the measures are nominal or ordinal, we can portray the relationship in a cross-tabulation. When the data are ordered and take on many different values, we commonly use the scattergram to picture the relationship. Coefficients of correlation condense information about association into single numbers. These numbers typically approach 1 (plus or minus) as the relationship becomes stronger and zero as the relationship becomes weaker. Some correlation coefficients derived from the idea of PRE-proportional reduction of error. PRE coefficients tell the percentage of improvement in predicting one variable that comes from knowing the other variable.
3.	Multivariate statistics, which can study two or more predictors in relation to an outcome variable. Regression analysis begins with the simple, two-variable case and followed with multivariate analysis. Time series analysis represents more advanced multivariate procedures that can control for confounding but at the cost of added complexity.

If you want to	Caution
Identify a score associated with a particular percentile. Choose the percentile. Add 1 to the sample size. Determine the location. Look in that location to identify the score	Don't confuse the location with the value of the score. When the location is a fraction, take the average of the two adjacent X values
Determine the percent of cases that fall a or below a given score. Prepare a frequency distribution with a cum f column. Find the cum f associated with the score. Divide that cum f by N and multiply by 100.	To avoid errors, make sure that the top value in cum f equals N. Be sure to multiply by 100. The decimal value you calculate is a proportion, not a percentage. Be careful in your interpretation, Evaluate the percentile in terms of the reference group

Summary of frequently used measures of central tendency		
Measure	When used	Caution
Mode	Easy to obtain measure for nominal data	Not precise, gives very little information, can be misleading
Median	Useful when you want to know the midpoint of a distribution or if distribution is skewed	Not sensitive to extreme scores
Mean	The arithmetic average, easily obtained, frequently used, and widely understood measure of central tendency	Very sensitive to extreme scores, can be misleading if outliers are present or if distribution is skewed
Weighted mean	When overall mean of several groups of different sizes is needed	Very important to use weighting, especially with much variation in size of subgroups

Summary of frequently used measures of dispersion		
Measure	When used	Caution
Range	Easy to obtain	Not precise, gives very little information, can be misleading
Semi-interquartile range	Useful when you want to know roughly the middle 50% of distribution	Used frequently when reporting research
Sum of squared deviations (SS)	The basis of many useful descriptive	By itself not a measure of spread, both spread and

	statistics and statistical tests	sample size affect the size of SS
Variance	Also the basis of some very useful statistical techniques	Examine the distribution; look to see if it is symmetrical
Standard deviation	Frequently used statistic for reporting where approximately two thirds of the distribution lie	Examine the distribution; look to see if it is symmetrical
Several different types of correlation coefficients and numerical scales with which they are used		
Scale	Symbol	Used with
Nominal	r_{phi} (phi coefficient). . . r_b (biserial r). r_t (tetrachoric)	Two dichotomous variables. . . One dichotomous variable with underlying continuity assumed; one variable that can take on more than two values. . . Two dichotomous variables in which underlying continuity can be assumed
Ordinal	r_s (spearman r). . . . T (kendall's tau or rank order correlation)	Ranked data. Both measures must be at least ordinal. They must be expressed as ranks prior to calculating Spearman r.. . Ranked data
Interval or ratio	Pearson r. . . Multiple R	Both scales interval or ratio. . . Three or more interval or ratio scaled variables

STATISTICAL SYMBOL, FORMULA, PROCEDURE			
Symbol	Definition	Symbol	Definition
\neq	not equal to	s^2_{within}	Within-group variance estimate
$>$	greater than	s_{estY}	Standard error of estimate when prediction Y from X
$<$	less than	$SS = \sum (X - \bar{X})^2$	Sum of squares
\leq	less than or equal to	$SIR = \frac{Q_3 - Q_1}{2}$	Semi-interquartile range
\geq	greater than or equal to	SS_{total}	Sum of squares total
\approx	approximately equal to	T	Transformed scores
X^2	X squared	T	Sum of ranks
X^a	X raised to the ath power	T	Sum of observations in a contingency table
\sqrt{X}	Square root of X	t	Students t-ratio
$X!$	Factorial of X; multiply all integers between 1 and X	\hat{t}	Adjusted t-ratio when N1 does not equal N2 and o1 does not equal o2
$\sum X$	Sum of all the numbers in the set of X	U, U'	Statistics calculated for the Mann-Whitney U-test
$\sum X^2$	Sum of squared scores	v	Used to calculate the standard normal deviate
$(\sum X)^2$	Sum of scores, quantity squared	X,Y	Variables
$-X$	Negative value of X	X_i, Y_i	Specific quantities indicated by the subscript, i
$XY, X(Y), X \times Y$	Multiplication	\bar{X}	Arithmetic mean
$X/Y, \frac{X}{Y}$	Division	$(X - \bar{X})$	Deviation score

Term	Formula	Term	Formula
One-sample z test	$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$	Chi Square Test	$\chi^2_{observed} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$ Where O= observed frequency, Where E= expected frequency
One-sample t test	Estimation of standard error $\hat{s}_{\bar{X}} = \frac{\hat{s}}{\sqrt{N}}$ $t = \frac{\bar{X} - \mu}{\hat{s}_{\bar{X}}}$	Cramer's V Lambda	$V = \sqrt{\frac{\chi^2}{(N) \text{Minimum of } (r-1, c-1)}}$ where minimum of (r-1, c-1)= the minimum value of r-1 (number of rows minus 1) or c-1 (number of columns minus 1)
Pearson Product-Moment	$r = \frac{\text{covariance}(X,Y)}{[stdev(X)][stdev(Y)]}$ which $= \frac{S_{YX}}{S_X S_Y}$	Cumulative	$cf = x + N$ where x= data frequency N= all frequencies above x

Correlation Coefficient		frequency	
Percentage	$p = \frac{n}{100}$ where n = data value	Cumulative percentage	$cp = \frac{x}{N}$ where x = individual frequency and N = total number of frequencies below x
Percentile Score	$Q = X_u + \frac{i(cf - cf_{ul})}{f_i}$ where : X_u = value of the lower real limit of the interval containing the cf of interest, i= size of interval, cf= cumulative frequency corresponding to the percentile of interest, cf_{ul} = cumulative frequency at the upper real limit of the interval below the interval containing cf	Deciles	Decile position= $\frac{(a/100)}{(n+1)}$ Where a= Decile number of 10, 40, 50, 60, 70, 80, 90 instead of 25, 50, and 75, n= sample size
Phi	$\Phi = \sqrt{\frac{\chi^2}{N}}$	Eta-squared	$\eta^2 = \frac{SS_{EXPLAINED}}{SS_{TOTAL}}$ where $SS_{EXPLAINED} = SS_{TOTAL} - SS_{ERROR}$ and $SS_{TOTAL} = SS_{EXPLAINED} + SS_{ERROR}$ and $SS_{ERROR} = \sum X_n^2 - \frac{(\sum X_n)^2}{N}$ where $\sum X_n$ = sum of individual scores, N= total number of scores
Point-Biserial Correlation Coefficient	$r_{pbi} = \frac{M_p - M_q}{S_t} \sqrt{pq}$ Where: r_{pbi} = point-biserial correlation coefficient, M_p = whole-test mean for students answering item correctly, M_q = whole-test mean for students answering item incorrectly, S_t = standard deviation for whole test, p = proportion of students answering correctly, q = proportion of students answering incorrectly	F-ratio	$F - \text{distribution} = \frac{V_m}{V_g}$, V_m and V_g can be found with the One way Dependent ANOVA test
Proportions	$p = \frac{f}{N}$ where f = Sum of scores, N= total sample score of a distribution.	F-test: 1way ANOVA	$F_{observed} = \frac{MS_{between}}{MS_{within}}$ where MS=mean square
Quartiles	Quartile Position = $\frac{(a/100)}{(n+1)}$ where a= percentile number of 25, 50, or 75 and n= sample size	F-test: In general	$F_{observed} = \frac{MS_{effect}}{MS_{appropriateErrorTerm}}$ where MS=mean square
Range	$R = X_{highest} - X_{lowest}$	Intraclass Correlation $\hat{\rho}_I$ One way ANOVA	$\hat{\rho}_I = \frac{MS_B - MS_W}{MS_B + (n-1)MS_W}$ where MS_B = mean square between groups, MS_W = mean square within groups, n=number of subjects within a group
Rate	$\frac{n}{N}$ where n= quantity of specific event occurrences, N= totals number of event possibilities	Kruskal-Wallis test	$H = \frac{12}{n(n+1)} \left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \dots + \frac{R_k^2}{n_k} \right) - 3(n+1)$ degree of freedom is k-1 where n=total number of observations in all samples combined, R_1 =sum of ranks for first sample, R_2 =m of ranks r the second sample, R_k = sum of ranks for the kth sample, k= number of samples
Ratio	$\frac{N_1}{N_2}$ where N_1 = number of cases in one category N_2 = number of cases in another category	Lambda	$\lambda = \frac{E_1 - E_2}{E_1}$ where E_1 = (the number of errors made while ignoring X), and E_2 = (the number of errors made when taking the independent variable into account)
Regression Slope	$b_{yx} = \frac{Cov_{XY}}{S_x^2} = r_{xy} \frac{S_y}{S_x}$ where: $Cov_{XY} =$	Linear regression n: 1	$\hat{Y} = \bar{Y} + b_{yx}(X - \bar{X})$ where \hat{Y} = predicted Y score, \bar{Y}

	covariance, $S_X^2 =$ variance of X, $r_{XY} =$ correlation, S = standard deviation	predictor	=mean of a set of scores on variable Y, $b_{YX} =$ regression slope for predicting Y from X, $(\bar{X}) =$ mean of a set of scores on variable X
Relative frequency	$rf = \frac{n}{N}$ where n= number of times a named outcome happened N= number of times an activity was done	Linear regression n: 2 predictors	$\hat{Y} = a + b_1X_1 + b_2X_2$ where $\hat{Y} =$ predicted Y score, $a =$ intercept, $b_1 =$ partial regression coefficient for variable 1, $b_2 =$ partial regression coefficient for variable 2
Scheffe's test 1 way ANOVA	$t_{observed} = \frac{\hat{C}}{\sqrt{MS_w \left(\frac{w_1^2}{n_1} + \frac{w_2^2}{n_2} + \dots + \frac{w_k^2}{n_k} \right)}}$ $t_{observed} = \sqrt{(k-1)F_{critical(\alpha, k-1, df_w)}}$ where: $\hat{C} =$ comparison of interest, $MS_w =$ mean square within group, w = weight, n = number of subjects within a group, k = number of groups, $df_w =$ degree of freedom within groups	Mann-Whitney U test	Formula A $U_1 = n_L n_S + \frac{n_L(n_L+1)}{2} - T_L$ where TL= the larger sum of ranks. L ans S are largest and Smallest respectively Formula B $U_1 = n_L n_S - U_1$ Formula C $Z_U = \frac{U - \left(\frac{n_1 n_2}{2} \right)}{S_U}$
Spearman's Rank Correlation	$r_s = \frac{Cov_{XY}}{S_X S_Y}$ where: X and Y= ranks, $Cov_{XY} =$ covariance, $S_X =$ standard deviation of X, $S_Y =$ standard deviation Y	Multiple correlation coefficient	$R = \frac{C_{ov}(Y, \hat{Y})}{S_Y S_{\hat{Y}}}$ where: $C_{ov}(Y, \hat{Y}) =$ covariance of Y and predicted Y, s = standard deviation
Standard Deviation	σ for standard deviation whole populations, s for standard deviation samples $\sqrt{\frac{(x_1^2 \times X_1) + \dots + (x_n^2 \times X_n)}{N}}$ where x= data value1 - mean, X= data value2, N= total number of data samples	Multiple correlation square	$R^2 = \frac{r_{Y1}^2 + r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - r_{12}^2}$ where: r = correlation, 1,2= predictor
Mode	Mode = n when n is the most frequent occurring number	Omega square ($\hat{\omega}^2$) One way ANOVA	$\hat{\omega}^2 = \frac{SS_B - (k-1)MS_w}{SS_T + MS_w}$ where: $SS_B =$ sum of squares between groups, $MS_w =$ mean square within groups, $SS_T =$ total sum of squares
Two-way Independent ANOVA F Test	$SS_A = \left[\frac{(\sum 1stRow)^2}{nRow} + \dots + \frac{(\sum LastRow)^2}{nRow} \right] - \frac{\sum (score)^2}{N}$ / $SS_B = \left[\frac{(\sum 1stColumn)^2}{nColumn} + \dots + \frac{(\sum LastColumn)^2}{nColumn} \right] - \frac{\sum (score)^2}{N}$ / $SS_{A \times B} = \left[\frac{(\sum 1stgroup)^2}{n} + \dots + \frac{(\sum LastGroup)^2}{n} \right] - \left[\frac{(\sum 1stRow)^2}{nRow} + \dots + \frac{(\sum LastRow)^2}{nRow} \right] - \left[\frac{(\sum 1stColumn)^2}{nColumn} + \dots + \frac{(\sum LastColumn)^2}{nColumn} \right] + \sum (score)^2$ $SS_{WG} = \sum (squarescores) - \left[\frac{(\sum 1stgroup)^2}{n} + \dots + \frac{(\sum Lastgroup)^2}{n} \right]$ $df_A = k - 1 / \quad df_B = q - 1 /$	One-way Dependent ANOVA F Test One-way Independent ANOVA F Test	$Variance1 = Vm = \frac{n_1(x_1 - x)^2 + \dots + n_c(x_c - x)^2}{c - 1}$ $Variance2 = Vg = \frac{(n_1 - 1)s_1^2 + \dots + (n_c - 1)s_c^2}{N - c}$ $SS_{BG} = \left[\frac{(\sum 1stGroup)^2}{n} + \dots + \frac{(\sum LastGroup)^2}{n} \right] - \frac{\sum (score)^2}{N}$ $SS_{WG} = \sum (squarescores) - \left[\frac{(\sum 1stgroup)^2}{n} + \dots + \frac{(\sum Lastgroup)^2}{n} \right]$ $df_{BG} = k - 1 / \quad df_{WG} = N - k /$ $MS_{BG} = \frac{SS_{BG}}{df_{BG}} \quad / \quad MS_{WG} = \frac{SS_{WG}}{df_{WG}} /$ $F = \frac{MS_{BG}}{MS_{WG}}$

	$df_{A \times B} = (k - 1)(q - 1) /$ $df_{WG} = N - k(q) / \quad MS_A = \frac{SS_A}{df_A}$ $MS_B = \frac{SS_B}{df_B} \quad MS_{A \times B} = \frac{SS_{A \times B}}{df_{A \times B}}$ $MS_{WG} = \frac{SS_{WG}}{df_{WG}} \quad F_A = \frac{MS_A}{MS_{WG}}$ $F_B = \frac{MS_B}{MS_{WG}} \quad F_{A \times B} = \frac{MS_{A \times B}}{MS_{WG}}$		
Variance	σ^2 or S^2 the first is for variance whole populations, the second is for variance samples	Mean	$\bar{x} = \frac{f}{N}$ where f= sum of data values, N= total number of data values
Wilcoxon signed-ranks test for two dependent samples	$z = \frac{T - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$	Measures of Dispersion IQV	$\frac{ov}{pv}$ where OV= observed variation and PV= Possible variation, IQV should be between 0 and 1
Z test	$Z - score = \frac{DataPoint - Mean}{StandardDeviation}$	Median	$X_u + \frac{1(\frac{n}{2} - cf_u)}{f_u}$ where X_u =lower real limit of the interval that contains the median, i = interval size, N = number of scores, cf_u = cumulative frequency of the interval below the interval containing the median, f_u = frequency of the interval containing the median.