Basic Statistcs Formula Sheet

Steven W. Nydick

May 25, 2012

This document is only intended to review basic concepts/formulas from an introduction to statistics course. Only mean-based procedures are reviewed, and emphasis is placed on a simplistic understanding is placed on when to use any method. After reviewing and understanding this document, one should then learn about more complex procedures and methods in statistics. However, keep in mind the assumptions behind certain procedures, and know that statistical procedures are sometimes flexible to data that do not necessarily match the assumptions.

Descriptive Statistics

Elementary Descriptives (Univariate & Bivariate)

Name	Population Symbol	Sample Symbol	Sample Calculation	Main Problems	Alternatives
Mean	μ	$\bar{x}$	$\bar{x} = \frac{\sum x}{N}$	Sensitive to outliers	Median, Mode
Variance	σ_x^2	s_x^2	$s_x^2 = \frac{\sum(x-\bar{x})^2}{N-1}$	Sensitive to outliers	MAD, IQR
Standard Dev	σ_x	s_x	$s_x = \sqrt{s_x^2}$	Biased	MAD
Covariance	σ_{xy}	s_{xy}	$s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{N-1}$	Outliers, uninterpretable units	Correlation
Correlation	ρ_{xy}	r_{xy}	$r_{xy} = \frac{s_{xy}}{s_x s_y}$ $r_{xy} = \frac{\sum(z_x z_y)}{N-1}$	Range restriction, outliers, nonlinearity	
z-score	z_x	z_x	$z_x = \frac{x-\bar{x}}{s_x}$; $\bar{z} = 0$; $s_z^2 = 1$	Doesn't make distribution normal	

Simple Linear Regression (Usually Quantitative IV; Quantitative DV)

Part	Population Symbol	Sample Symbol	Sample Calculation	Meaning
Regular Equation	$y_i = \alpha + \beta x_i + \epsilon_i$	$y_i = a + bx_i + e_i$	$\hat{y}_i = a + bx_i$	Predict y from x
Slope	β	b	$b = \frac{s_{xy}}{s_x^2} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$	Predicted change in y for unit change in x
Intercept	α	a	$a = \bar{y} - b\bar{x}$	Predicted y for $x = 0$
Standardized Equation	$z_{y_i} = \rho_{xy} z_{x_i} + \epsilon_i$	$z_{y_i} = r_{xy} z_{x_i} + e_i$	$\hat{z}_{y_i} = r_{xy} z_{x_i}$	Predict z_y from z_x
Slope	ρ_{xy}	r_{xy}	$r_{xy} = \frac{s_{xy}}{s_x s_y} = b\left(\frac{s_x}{s_y}\right)$	Predicted change in z_y for unit change in z_x
Intercept	None	None	0	Predicted z_y for $z_x = 0$ is 0
Effect Size	P^2	R^2	$r_{\hat{y}y}^2 = r_{xy}^2$	Variance in y accounted for by regression line

Inferential Statistics

t-tests (Categorical IV (1 or 2 Groups); Quantitative DV)

Test	Statistic	Parameter	Standard Deviation	Standard Error	*df*	t-obt
One Sample	$\bar{x}$	μ	$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{N-1}}$	$\frac{s_x}{\sqrt{N}}$	$N-1$	$t_{\text{obt}} = \frac{\bar{x}-\mu_0}{\frac{s_x}{\sqrt{N}}}$
Paired Samples	$\bar{D}$	μ_D	$s_D = \sqrt{\frac{\sum(D-\bar{D})^2}{N_D-1}}$	$\frac{s_D}{\sqrt{N_D}}$	$N_D - 1$	$t_{\text{obt}} = \frac{\bar{D}-\mu_{D0}}{\frac{s_D}{\sqrt{N_D}}}$
Independent Samples	$\bar{x}_1 - \bar{x}_2$	$\mu_1 - \mu_2$	$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$	$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	n_1+n_2-2	$t_{\text{obt}} = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$
Correlation	r	$\rho = 0$	NA	NA	$N-2$	$t_{\text{obt}} = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$
Regression (FYI)	a & b	α & β	$\hat{\sigma}_e = \sqrt{\frac{\sum(y-\hat{y})^2}{N-2}}$	s_a & s_b	$N-2$	$t_{\text{obt}} = \frac{a-\alpha_0}{s_a}$ & $t_{\text{obt}} = \frac{b-\beta_0}{s_b}$

t-tests Hypotheses/Rejection

Question	One Sample	Paired Sample	Independent Sample	When to Reject
Greater Than?	$\text{H}_0 : \mu \leq \#$ $\text{H}_1 : \mu > \#$	$\text{H}_0 : \mu_D \leq \#$ $\text{H}_1 : \mu_D > \#$	$\text{H}_0 : \mu_1 - \mu_2 \leq \#$ $\text{H}_1 : \mu_1 - \mu_2 > \#$	Extreme positive numbers $t_{\text{obt}} > t_{\text{crit}}$ (one-tailed)
Less Than?	$\text{H}_0 : \mu \geq \#$ $\text{H}_1 : \mu < \#$	$\text{H}_0 : \mu_D \geq \#$ $\text{H}_1 : \mu_D < \#$	$\text{H}_0 : \mu_1 - \mu_2 \geq \#$ $\text{H}_1 : \mu_1 - \mu_2 < \#$	Extreme negative numbers $t_{\text{obt}} < -t_{\text{crit}}$ (one-tailed)
Not Equal To?	$\text{H}_0 : \mu = \#$ $\text{H}_1 : \mu \neq \#$	$\text{H}_0 : \mu_D = \#$ $\text{H}_1 : \mu_D \neq \#$	$\text{H}_0 : \mu_1 - \mu_2 = \#$ $\text{H}_1 : \mu_1 - \mu_2 \neq \#$	Extreme numbers (negative and positive) $\lvert t_{\text{obt}}\rvert > \lvert t_{\text{crit}}\rvert$ (two-tailed)

t-tests Miscellaneous

Test	Confidence Interval: $\gamma\% = (1-\alpha)\%$	Unstandardized Effect Size	Standardized Effect Size
One Sample	$\bar{x} \pm t_{N-1;\,\text{crit(2-tailed)}} \times \frac{s_x}{\sqrt{N}}$	$\bar{x} - \mu_0$	$\hat{d} = \frac{\bar{x}-\mu_0}{s_x}$
Paired Samples	$\bar{D} \pm t_{N_D-1;\,\text{crit(2-tailed)}} \times \frac{s_D}{\sqrt{N_D}}$	$\bar{D}$	$\hat{d} = \frac{\bar{D}}{s_D}$
Independent Samples	$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2;\,\text{crit(2-tailed)}} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$\bar{x}_1 - \bar{x}_2$	$\hat{d} = \frac{\bar{x}_1-\bar{x}_2}{s_p}$

One-Way ANOVA (Categorical IV (Usually 3 or More Groups); Quantitative DV)

Source	Sums of Sq.	*df*	Mean Sq.	F-stat	Effect Size
Between	$\sum_{j=1}^{g} n_j(\bar{x}_j - \bar{x}_G)^2$	$g-1$	SSB/dfB	MSB/MSW	$\eta^2 = \frac{SSB}{SST}$
Within	$\sum_{j=1}^{g}(n_j-1)s_j^2$	$N-g$	SSW/dfW		
Total	$\sum_{i,j}(x_{ij} - \bar{x}_G)^2$	$N-1$			

1. We perform ANOVA because of family-wise error -- the probability of rejecting **at least one** true H_0 during multiple tests.
2. G is "grand mean" or "average of all scores ignoring group membership."
3. $\bar{x}_j$ is the mean of group j; n_j is number of people in group j; g is the number of groups; N is the total number of "people".

One-Way ANOVA Hypotheses/Rejection

Question	Hypotheses	When to Reject
Is at least one mean different?	$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ H_1 : At least one μ is different from at least one other μ	Extreme positive numbers $F_{\text{obt}} > F_{\text{crit}}$

- Remember Post-Hoc Tests: LSD, Bonferroni, Tukey (what are the rank orderings of the means?)

Chi Square (χ^2) (Categorical IV; Categorical DV)

Test	Hypotheses	Observed	Expected	*df*	χ^2 Stat	When to Reject
Independence	H_0 : Vars are Independent H_1 : Vars are Dependent	From Table	Np_jp_k	(Cols - 1)(Rows - 1)	$\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(f_{O\,ij}-f_{E\,ij})^2}{f_{E\,ij}}$	Extreme Positive Numbers $\chi^2_{\text{obt}} > \chi^2_{\text{crit}}$
Goodness of Fit	H_0 : Model Fits H_1 : Model Doesn't Fit	From Table	Np_i	Cells - 1	$\sum_{i=1}^{C}\frac{(f_{O\,i}-f_{E\,i})^2}{f_{E\,i}}$	Extreme Positive Numbers $\chi^2_{\text{obt}} > \chi^2_{\text{crit}}$

1. Remember: the sum is over the number of cells/columns/rows (not the number of people)
2. For Test of Independence: p_j and p_k are the marginal proportions of variable j and variable k respectively
3. For Goodness of Fit: p_i is the expected proportion in cell i if the data fit the model
4. N is the total number of people

Assumptions of Statistical Models

Correlation

1. Estimating: Relationship is linear
2. Estimating: No outliers
3. Estimating: No range restriction
4. Testing: Bivariate normality

Regression

1. Relationship is linear
2. Bivariate normality
3. Homoskedasticity (constant error variance)
4. Independence of pairs of observations

One Sample t-test

1. x is normally distributed in the population
2. Independence of observations

Paired Samples t-test

1. Difference scores are normally distributed in the population
2. Independence of pairs of observations

One-Way ANOVA

1. Each group is normally distributed in the population
2. Homogeneity of variance
3. Independence of observations within and between groups

Independent Samples t-test

1. Each group is normally distributed in the population
2. Homogeneity of variance (both groups have the same variance in the population)
3. Independence of observations within and between groups (random sampling & random assignment)

Chi Square (χ^2)

1. No small expected frequencies
   - Total number of observations at least 20
   - Expected number in any cell at least 5
2. Independence of observations
   - Each individual is only in ONE cell of the table

Central Limit Theorem

Given a population distribution with a mean μ and a variance σ^2, the sampling distribution of the mean using sample size N (or, to put it another way, the distribution of **sample means**) will have a mean of $\mu_{\bar{x}} = \mu$ and a variance equal to $\sigma^2_{\bar{x}} = \frac{\sigma^2}{N}$, which implies that $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$. Furthermore, the distribution will approach the normal distribution as N, the sample size, increases.

Possible Decisions/Outcomes

	H_0 True	H_0 False
Rejecting H_0	Type I Error (α)	Correct Decision ($1-\beta$; Power)
Not Rejecting H_0	Correct Decision ($1-\alpha$)	Type II Error (β)

Power **Increases** If: $N\uparrow$, $\alpha\uparrow$, $\sigma^2\downarrow$, Mean Difference $\uparrow$, or One-Tailed Test