Computational Public Health Statistics (HS 267) : Formulas (Part 1)

(Initially compiled by Jane Pham; modified by B. Gerstman; Version of March 6, 2005)

“Old Fashioned” Descriptive Statistics

Statistic	***Parameter***	***Point Estimate***	***Formula***	***Interpretation***	***Notes / Discussion***
Sum of squares	$\sigma^2 \times df$	SS	$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2$	No easy interpretation.	• When comparing two or more groups, address central locations, variability and sample size(s). • Mean and standard deviation are best when distribution is symmetrical. • Assuming normal distribution, 68% of data points lie within ±1 standard deviation from the mean, 95% within ±2, and nearly all data within ±3 standard deviations from the mean. • For all data, Chebychev’s rule, at least 75% of data lie within ±2 standard deviations from the mean.
Mean	μ	$\bar{x}$	$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	A measure of central location; “expectation”	
Variance	σ^2	s^2	$s^2 = \frac{SS}{n-1}$	A measure of “spread”; expressed in units <u>squared</u>	
Standard Deviation	σ	s	$s = \sqrt{s^2}$ or $\sqrt{\frac{SS}{n-1}}$	A measure of variability, expressed in data units. More appropriate for descriptive purposes.	

5-point summaries and boxplots

Statistic	***Formula***	***Interpretation***	***5-point Summary***	***Notes / Discussion***
Median	Median has depth of $\frac{n+1}{2}$	A measure of central location	Q0 – Minimum Q1 – First Quartile Q2 – Median Q3 – Third quartile Q4 – Maximum	• When comparing side-by-side boxplots, discuss their locations (look at box, whiskers, and medians), discuss their spread (look at IQR’s), relative to each other. Also discuss their ranges (whisker-spread), overlaps, and outside values. • The box contains 50% of data, drawn from Q1 to Q3, and the height of this box is the IQR (hinge-spread). The line inside the box is Q2 (median). • The lower whisker is drawn from Q1 to the lower inside value. The upper whisker is drawn from Q3 to the upper inside value. NOTE: fences are <u>never</u> formally drawn. • Remember to plot dots for outside values. • Good method to compare group locations and spreads, esp. when distribution is assymetrical.
Interquartile Range (IQR)	$IQR = Q3 - Q1$	A measure of spread, aka “hinge-spread”		
Lower Fence (F_l)	$F_l = Q1 - 1.5(IQR)$	Use to determine: Lower inside value Lower outside value(s)		
Upper Fence (F_u)	$F_u = Q3 + 1.5(IQR)$	Used to determine: Upper inside value Upper outside value(s)		

Testing MEANS

Test	***K groups / Assumption***	***Hypothesis***	***Formulas***	***Notes / Discussion***
Independent *t* –test (Student)	*k* = 2 groups Equal variances assumed	H_o assumes there's no difference in means $H_o : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	Pooled estimate of variance $s_p^2 = \frac{(df_1)(s_1^2)+(df_2)(s_2^2)}{df}$ Standard error of the mean difference $se_{\bar{x}_1-\bar{x}_2} = \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ t statistic , degrees of freedom $t_{stat} = \frac{\bar{x}_1-\bar{x}_2}{se_{\bar{x}_1-\bar{x}_2}}$, $df = df_1 + df_2$	• A common error is forgetting to square the standard deviations in the equation for pooled estimate of variance. Check to see whether variances or standard deviations are given. • When we pool the variances, we suppress non-uniformity of s_1 and s_2 .
Independent t –test (Behrens-Fisher)	K = 2 groups Equal variances NOT assumed	H_o assumes there's no difference in means $H_o : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	Standard error of the mean difference $se_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ t statistic $t_{stat} = \frac{\bar{x}_1-\bar{x}_2}{se_{\bar{x}_1-\bar{x}_2}}$ degrees of freedom, the integer below df' $df' = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$	• This t-test can also be used when equal variances are assumed or not assumed. • Consider using this test if you performed a significance test for variances and concluded heteroscedasticity.
ANOVA	*k* = 2 or more groups Equal variances assumed	H_o assumes there's no difference in means $H_o : \mu_1 = \mu_2 = \mu_3$..... $H_1 : H_o$ is false or (at least one pop. mean differs)	Variance Between groups $SS_B = \sum_{i=1}^{k} n_i\left(\bar{x}_i - \bar{x}\right)^2$, $s_B^2 = \frac{SS_B}{df_B}$, $df_B = k-1$ Variance Within groups $SS_W = \sum_{i=1}^{k}(n_1-1)s_i^2$, $s_W^2 = \frac{SS_W}{df_W}$, $df_W = N-k$ F statistic → $F_{stat} = \frac{s_B^2}{s_W^2}$	• Keep in mind that ANOVA can only reveal whether at least one mean differs or not. It doesn't tell you which means differ, that's why we might need to do post-hoc comparisons.

Post-hoc Comparisons : These tests are performed following ANOVA

Test	***K groups / Assumption***	***Hypothesis***	***Formulas***	***Notes / Discussion***
Least Significance Difference (LSD)	K = 2 groups for each comparison Number of comparisons: $m={}_kC_2$	Ex: 3 groups, 3 comparisons Test 1: $H_o : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ Test 2: $H_o : \mu_1 = \mu_3$ vs. $H_1 : \mu_1 \neq \mu_3$ Test 3: $H_o : \mu_2 = \mu_3$ vs. $H_1 : \mu_2 \neq \mu_3$	Standard error of the mean difference $se_{\bar{x}_i - \bar{x}_j} = \sqrt{s_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$, where s_w^2 is obtained from ANOVA $t_{stat} = \frac{\bar{x}_i - \bar{x}_j}{se_{\bar{x}_i - \bar{x}_j}}$, $df = N - k$	• LSD performs multiple t-tests between two groups each time to infer where the mean difference is significant. • Another alternative (to the Bonf. Method) to adjust for the problem of multiplicity is to apply a more stringent error or α level. • Consider looking at descriptive/summary statistics or using EDA to explain where the difference exists, p values are not always necessary.
Bonferroni’s Method	Same as LSD Method	Same as LSD Method	For each p-value obtained using LSD or independent t-test, that p-value is adjusted by using the following formula: $p_{Bonf} = p \times m$	

Testing VARIANCES

Test	***K groups / Assumption***	***Hypothesis***	***Formulas***	***Notes / Discussion***
F-ratio test	K = 2 groups	H_o assumes there’s no difference in variances (homocedasticity) $H_o : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$	F statistic $F_{stat} = \frac{s_1^2}{s_2^2}$ or $\frac{s_2^2}{s_1^2}$, whichever is larger Degrees of freedom $df_1 = n_1 - 1$, $df_2 = n_2 - 1$	• When equal variances rejected, consider the following methods to compare means: EDA, descriptive/summary statistics, or unequal variance t-test. • Keep the numerator and the denominator separate. The larger variance goes in the numerator and uses the df for this particular group.
Levene’s test	K = 2 or more groups	$H_1 : H_o$ is false	Compute with SPSS, no hand calculations ☺	

Scatterplots

Action	***Examples***		***Notes / Discussion***
	X (independent)	Y (dependent)	
Determine X and Y	Sodium Intake Cigarette Consumption Incoming ACT Scores Age % Births Attended by Pro Fluoride Concentration % Reduced Fee Lunches Weight	Systolic Blood Pressure Lung Cancer Mortality % Graduating Height Maternal Mortality Cavities % Bicycle Helmet Use Eggs Produced	• X (independent) is often called the <u>predictor</u> variable. We hope to use X to predict changes in Y (regression). • Y (dependent) is often called the <u>response</u> variable, because we are asking the question: will Y's value respond to changes in X? • Label X and Y with variable names and units • Watch out for data that appears to be curved, U shaped, or otherwise non-linear. • Should be able to tell the <u>direction</u> of correlation Positive correlation: high X corresponds to high Y Negative correlation: high X corresponds to low Y No correlation: changes in X don't affect Y (horizontal line or vertical cluster) • <u>AVOID</u> trying to assess <u>strength</u> of correlation (weak/strong) as it is too dependent on the scale of the plot. • Outliers may be important, or they may just show an error in data
Judge Linearity	See if the scattered points can be described by a straight line		
Assess Correlation (Direction)	Look at the direction of the slope		
Look for Outliers	Points that are outside scatter cloud		

Correlation

Statistic	***Parameter***	***Point Estimate***	***Formula***	***Notes / Discussion***
Sum of squares		SS	$SS_{xx} = \sum_{i=1}^{n}(x-\bar{x})^2$ $SS_{yy} = \sum_{i=1}^{n}(y-\bar{y})^2$ $SS_{xy} = \sum_{i=1}^{n}\left[(x-\bar{x})(y-\bar{y})\right]$	• SS_{xx} quantifies the spread of variable X SS_{yy} quantifies the spread of variable Y SS_{xy} quantifies the extent in which two variables "go together" • The strength of the correlation is in the absolute value of *r*, and <u>depends on the application</u>. General guidelines: weak if |*r*|<0.3, moderate if 0.3<|*r*|<0.7, strong if |*r*|>0.7 • The coefficient of determination (r^2) is the proportion of Y's variability that can be "explained" by changes in X. • Null hypothesis assumes no correlation. If we reject, then we are saying that *r* is significant, and X and Y are correlated. • Interpretation works best if there is a linear relationship, and inference/testing on ρ assumes that X and Y are bivariate normal
Correlation Coefficient	ρ	r	$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$	
Testing	$H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$		$se_r = \sqrt{\frac{1-r^2}{n-2}}$, $t_{stat} = r / se_r$ $df = n-2$	

Linear Regression

Statistic	***Parameter***	***Point Estimate***	***Formula***	***Notes / Discussion***
Slope Coefficient:	β	*b*	$b = \frac{SS_{xy}}{SS_{xx}}$	• The slope indicates a change in Y per unit change in X. • Y-intercept is the value on the Y-axis when X is equal to 0. • Null hypothesis assumes a slope of 0 (no correlation). If we reject, then we are saying that X and Y are linearly related, with a change in Y per unit change in X. • Linear regression assumes… **L**inearity between X and the expected value of Y **I**ndependence of each observation (each X,Y pair is a single observation) **N**ormality of the residuals **E**qual variance of the residual normal distributions (centered at a point on the line for each value of X)
Y-Intercept Coefficient		*a*	$a = \bar{y} - b\bar{x}$	
Linear Regression Model for Expected Y	$\hat{y} = a + bx$ where $\hat{y}$ is the predicted Y at level of *x*, *a* is the intercept and *b* is the slope.			
Confidence Interval for **β**	$b \pm (t_{n-2\,,1-\alpha/2})(se_b)$			
Hypothesis Testing	$H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$		$se_b = \frac{se_Y}{\sqrt{SS_{xx}}}$, $t_{stat} = (b-0)/se_b$ $df = n-2$	

Few last comments about hypothesis testing:

- Listing of steps is recommended for beginners:
  1.State hypotheses $(H_o$ and $H_1)$
  2.Error threshold (α level) is used *only* for fixed-level testing, but is discouraged at this level and *not* used in practice.
  3.Calculate test statistic and draw the curve to help determine the *p* value
  4.Conclusion: report precise *p* value if available; if not available, report closest approximation (e.g., $p < .001$ or what-ever)
  How do we talk about the results? At the *simplest* level, $p \leq .05$ are considered significant (by convention) in practice; in practice the smaller the *p* value, the better the evidence against H_o and the more confidence we can put in rejecting the null.
- After performing the test, take a step back and view *p* in relation to summary statistics and EDA. Make your conclusions meaningful!