Computational Public Health Statistics Formulas (Part 1)

Version: March 2007

"Old Fashioned" Descriptive Statistics – You should come into the course already knowing these

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.	• Accompany descriptive statistics w/ EDA when possible. • Mean and standard deviation are best when distribution is bell-shaped or at least symmetrical. • Assuming normal distribution, 68% of data points lie within ± 1 standard deviation from the mean, 95% within ± 2. For other data use 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 squared	
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 - You should come into the course already knowing these

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 spread (look at IQR's), relative to each other. Also discuss 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. Fences are never drawn. • Remember to plot dots for outside values. • Good method to compare groups, esp. when distribution is asymmetrical.
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 -- always interpret in context of descriptives and EDA; remember fallacies of sig. testing; remember validity assumptions trump distributional conditions

Test	***Hypothesis***	***Formulas***	***Notes / Discussion***
Equal variance *t* test "Student's *t*" "Pooled t" (not recommended in practice)	$H_o: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$s_p^2 = \frac{(df_1)(s_1^2) + (df_2)(s_2^2)}{df}$ $se_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ $t_{stat} = \frac{\bar{x}_1 - \bar{x}_2}{se_{\bar{x}_1 - \bar{x}_2}}$, $df = df_1 + df_2$	• A common error is forgetting to square the standard deviations before pooling. 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. • Conditions: independent samples, Normality, equal variance
Unequal variance *t* test ("Behrens-Fisher problem;" Welch's solution)	$H_o: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	Standard error of the mean difference $se_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ <u>t</u> statistic $t_{stat} = \frac{\bar{x}_1 - \bar{x}_2}{se_{\bar{x}_1 - \bar{x}_2}}$ $df' = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$	• This *t* test can 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. • Conditions: independent samples, Normality
ANOVA	$H_o: \mu_1 = \mu_2 = \mu_3 .....$ $H_1: H_o$ is false (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 → $F_{stat} = \frac{s_B^2}{s_W^2}$	• ANOVA doesn't tell you which means differs, so you might need to do post-hoc comparisons. • ANOVA has all the limitations of significance testing. • Conditions: independent samples, Normality, equal variance

Computational Public Health Statistics Formulas (Part 1)

Version: March 2007

Post-hoc Comparisons: These tests are performed following ANOVA

Test	***K groups / Assumption***	***Hypothesis***	***Formulas***	***Notes / Discussion***
Least Significance Difference (LSD)	2 groups for each comparison Number of comparisons: $m={}_kC_2$	Example: for $k = 3$ $H_o : \mu_1 = \mu_2$ $H_o : \mu_1 = \mu_3$ $H_o : \mu_2 = \mu_3$	$se_{\bar{x}_i - \bar{x}_j} = \sqrt{s_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$, where s_w^2 is obtained from ANOVA $t_{stat} = \frac{\bar{x}_i - \bar{x}_j}{se_{\bar{x}_i - \bar{x}_j}}$, $df = N - k$	• Multiple *t* tests to infer which mean differences are significant. • Bonf. adjusts for the Problem of Multiple Comparisons • Look at summary statistics and EDA to put comparisons in context
Bonferroni's Method	Same as LSD Method	Same as LSD Method	For each *P* value obtained using LSD, $p_{Bonf} = p \times m$	

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) $H_o : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$	$F_{stat} = \frac{s_1^2}{s_2^2}$ 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	$H_o : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 ...$ $H_1 : H_o$ is false	Compute with SPSS, no hand calculations ☺	

Computational Public Health Statistics Formulas (Part 1)

Version: March 2007

Scatterplots

<table>
<tr><th>Action</th><th>Look for</th><th>Additional notes</th></tr>
<tr><td>Determine X and Y</td><td>X (explanatory; independent)
Y (response; dependent)</td><td rowspan="4">• Label axes with variable names and units
• Watch for relations that are curved, U shaped, or otherwise non-linear.
• It is difficult to assess strength of association visually (too dependent on the scale of the plot)
• Outliers may be important or influential</td></tr>
<tr><td>Judge Linearity</td><td>Look at scatter plot. Can trend be described with a straight line?</td></tr>
<tr><td>Assess Direction</td><td>Look at the direction of the slope (positive, negative, flat)</td></tr>
<tr><td>Look for Outliers</td><td>Points outside scatter cloud (i.e., with large residuals)</td></tr>
</table>

Correlation

<table>
<tr><th>Statistic</th><th>Parameter</th><th>Point Estimate</th><th>Formula</th><th>Notes / Discussion</th></tr>
<tr><td>Sum of squares</td><td></td><td>SS</td><td>$SS_{xx} = \sum_{i=1}^{n}(x-\bar{x})^2$
$SS_{yy} = \sum_{i=1}^{n}(y-\bar{y})^2$
$SS_{xy} = \sum_{i=1}^{n}\left[(x-\bar{x})(y-\bar{y})\right]$</td><td rowspan="3">• 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</td></tr>
<tr><td>Correlation Coefficient</td><td>ρ</td><td>r</td><td>$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$</td></tr>
<tr><td>Testing</td><td colspan="2">$H_0 : \rho = 0$ vs.
$H_1 : \rho \neq 0$</td><td>$se_r = \sqrt{\frac{1-r^2}{n-2}}$, $t_{stat} = r / se_r$
$df = n-2$</td></tr>
</table>

Computational Public Health Statistics Formulas (Part 1)

Version: March 2007

Linear Regression

Statistic	***Parameter***	***Point Estimate***	***Formula***	***Notes***
Slope Coefficient:	β	b	$b = \frac{SS_{XY}}{SS_{XX}} = r\frac{s_Y}{s_X}$	• The slope indicates predicted change in Y per unit change in X (key statistic) • Y-intercept is the value on the Y-axis when X is equal to 0 (needed to anchor the line; not often interpreted. • Distributional conditions: **L**inearity between X and Y **I**ndependence of each bivariate observation **N**ormality of the residuals **E**qual variance of the residual ▪ Validity conditions (good info, good sample, no confounding) trump distributional conditions.
Intercept Coefficient	α	a	$a = \bar{y} - b\bar{x}$	
Regression Model	$\hat{y} = a + bx$ where $\hat{y}$ is the predicted Y at level x, a is the intercept estimate and b is the slope estimate			
Confidence Interval for β	$b \pm (t_{n-2, 1-\alpha/2})(se_b)$ where $se_b = \frac{se_{Y\mid x}}{\sqrt{SS_{xx}}}$ is the standard error of the slope $se_{Y\mid x}$ = standard error of the regression = Residual Mean Square Compute se_b and/or $se_{Y\mid x}$ with SPSS (too tedious to do by hand)			
Significance Test	H_0: β = 0 vs H_0: β ≠ 0 Null hypothesis claims population slope = 0 (no association)		Use t test or ANOVA, as described in Lab Workbook	

A few additional general comments:

- Interpretation starts with understanding what you hope to accomplish. Make your statistics valid and meaningful!
- Confidence intervals are usually *more uses* than significance test because of their ability to estimate effect size.
- Understand each step, not just the conclusion. Step are: 1. Hypotheses statements, 2.Test statistics, 3. P value 4.Conclusion. The P value is a measure of evidence against H_0 (typical thresholds: 0.10, 0.05, and 0.01) but provides no information about the strength, direction, or importance of the relationship.
- Use computational tools when available. Don't feel compelled to calculate everything by hand.