Biostatistics 101

Kirsten N. Kangelaris DCR

Goals:

1. **Learn the basic considerations in choosing a statistical test**
2. **Identify the appropriate statistical test for your research questions**
3. **Provide a practical guide of how to perform basic statistical tests without a statistician (Chi-square/Fisher's exact, T-test [paired and unpaired], ANOVA, Pearson, descriptive statistics for proportions)**
4. **Know when to get expert help**

Step 1: Identify the types of variables for your predictor and outcome in your research question:

Continuous: Infinitely divisible

1. Normally distributed (parametric): age, laboratory values in a healthy population (WBC count, hemoglobin), height, weight
2. Non-normally distributed (non-parametric): length of stay, troponin (several other inflammatory biomarkers)

Categorical: discrete, indivisible

1. Ordinal: Cancer stage, categories of continuous variables (eg., age categories), Likert scale*, number of hospital beds
2. Nominal: race, floor, disposition of patient on discharge (home, rehab, SNF, hospice, etc.)

Dichotomous: sex, mortality, presence of disease/diagnosis

*Likert scale often treated like a continuous variable for statistical purposes

Step 2: Choose the appropriate test based on your predictor and outcome

		Predictor				
Outcome		**Continuous, normal**	**Continuous, skewed**	**Categorical (ordinal)**	**Categorical (nominal)**	**Dichotomous**
	Continuous, Normal	Pearson correlation	Spearman correlation	Linear regression with test for trend	ANOVA	t-test
	Continuous, skewed	Spearman correlation	Spearman correlation	Spearman correlation *	Kruskal-Wallis	Wilcoxon
	Categorical (ordinal)	Use regression set-up for this combination	Spearman correlation*	Chi-square with test for trend	Chi-square test for trend	Chi-square test for trend
	Categorical (nominal)	ANOVA	Reverse role of predictor and outcome, use Kruskal-Wallis	Chi-square test for trend	Chi-square or Fishers' exact test**	Chi-square or Fishers' exact test**
	Dichotomous	t-test	Reverse role of predictor and outcome, use Wilcoxon	Chi-square test for trend	Chi-square or Fishers' exact test**	Chi-square or Fishers' exact test**

Gray shading—need statistical program to calculate

*If using STATA, better to use non-parametric trend test implemented as *nptrend* command

** Use Fisher's exact test when any of the expected cells counts in contingency table are 5 or less.

Chi square test for trend is *opartchi* command in STATA

Step 3: Bare bone basic tips for data management (there are classes on just this!!)

1. Use a de-identified dataset with anonymous study IDs (separate from key for MRN, names, other patient identifiers, etc.)
2. Store in Excel if not using statistical program, Access database, etc.
3. NO WORDS in datafile!! Label dichotomous variables consistently with 0 for no, 1 for yes describing the variable (eg., male=1/0, where male=1 is for men and male=0 for women)
4. Create a data dictionary to define what the numbers in your datafile mean. This is especially important for categorical variables (eg., race variables 1, 2, 3, 4, 5…easy to forget how you defined them)

Step 4: Use an online calculator or excel "data" and "formula" menus to run simple statistical tests (not those shaded above) or download EPI Info a great data collection and analysis instrument that is free and easy to use!
Helpful online calculators:

1. Best for describing continuous variables (graphical determination of distribution, mean, SD, median, IQR), T-Test (paired and unpaired) and ANOVA: http://www.physics.csbsju.edu/stats/
2. Best for Fisher's exact/Chi-squared tests and for generating 95% Confidence intervals around proportions for descriptive studies: http://graphpad.com/quickcalcs/CatMenu.cfm
3. More on choosing a statistical test: http://www.graphpad.com/www/book/choose.htm
4. Online sample size calculators: http://www.epibiostat.ucsf.edu/dcr/
5. Free, easy to use statistical program from the CDC: http://wwwn.cdc.gov/epiinfo/

Step 5: Biostatistics 102—For multivariable analyses, recommend data analysis program and/or statistician to help.

Select the appropriate regression model for the **outcome type**.

Outcome type	Regression model
Continuous, normal	Linear
Continuous, skewed	Linear, after transformation of outcome or using bootstrap SEs; Gamma, Poisson, or Negative-Binomial generalized linear models.
Ordinal	Proportional odds or continuation ratio model
Nominal categorical	Multinomial logistic
Count (eg. number of readmissions)	Poisson, Zero-inflated Poisson, Negative Binomial, Zero-inflated Negative Binomial
Dichotomous	Logistic
Censored time to event	Cox proportional hazards model

Determine how to include predictors in any of those regression models

Predictor	How to include it as a predictor in regression
Continuous	Check for linearity and influential points. Accommodate non-linearity using log-, polynomial, restricted cubic spline or categorical transformation of predictor. ***Continuous predictors do not need to be Normally distributed.***
Ordinal categorical	Include as multi-category predictor; test for trend in effects across categories
Nominal categorical	Include as multi-category predictor; use test for heterogeneity across categories, possibly followed by tests for pairwise differences between categories
Count	Include as continuous, checking for linearity and influential points. Accommodate non-linearity using methods for other continuous predictors.
Dichotomous	Include as 0-1 indicator variable