Biostatistics: Types of Data Analysis

Theresa A Scott, MS

Vanderbilt University
Department of Biostatistics
theresa.scott@vanderbilt.edu
http://biostat.mc.vanderbilt.edu/TheresaScott

Goals of data analysis

▷ (1) *To describe* (summarize) the population of interest by describing what was observed in the (study) sample.

- Employs *descriptive statistics*, which involves
  - Summarizing continuous variables using the mean, standard deviation, range, and percentiles (including the median).
  - Summarizing categorical variables using raw and relative frequencies.

▷ (2) To use patterns in the (study) sample data *to draw inferences* about the population represented.

- Employs *inferential statistics*, which involves
  - Confidence intervals.
  - Hypothesis tests & p-values.
  - Correlation and determining associations.
  - Determing relationships, estimating effects, and making predictions using regression analysis.

Recall some key terms

▷ *Population* of interest and (study) *sample*.
- Key to being able to generalize your findings to the population – how representative your study sample is of the population.

▷ *Roles* of collected variables:
- Outcome
- Predictor
- Confounder
- Additional descriptor

▷ *Types* of collected variables:
- Continuous, which includes discrete numeric.
- Categorical, which includes binary and ordinal.

▷ Definitions given in the 'Biostatistics and Research' lecture.

Revisiting specific aim(s)/objective(s)

▷ Nice if the wording of the specific aim(s)/objective(s) conveys the statistical analysis that is/will be used.

▷ Some examples:
- *To describe* the distributions of risk factors among a cohort of women with breast cancer.
- *To compare* the presentation, evaluation, diagnosis, treatment, and follow-up of...
- *To estimate* the incidence of skin cancer among elderly smokers and non-smokers.
- *To determine* whether a significant association exists between cigarette smoking and pancreatic cancer.
- *To determine* the effect of X on Y once adjusted for Z.
- *To predict* the probability of surviving one-year post-surgery ...

Descriptive Statistics

Summarizing individual **continuous variables**

▷ Mean (average) ± standard deviation (SD).

- SD = measure of variability (dispersion) around the mean.
- *Empirical rule*: If the distribution of a variable approximates a bell-shaped curve (ie, is normally distributed), approximately 95% of the variable's values lie within 2 SDs of the mean.
- Both influenced by *outliers* – bad descriptors if not bell-shaped.

▷ Range (minimum and maximum values).

- Also influenced by outliers.

▷ *Percentiles* – values that divide an ordered continuous variable into 100 groups with at most 1% of the values in each group.

- The p-th percentile is the value that p% of the data are less than or equal to (ie, p% of the data lie below it).
  - Follows that (100-p)% of the data lie above it.

Continuous variables, *cont'd*

▷ Percentiles, *cont'd*:

- *Example*: if the 85% percentile of household income is $60,000, then 85% of households have incomes of $\leq$$60,000 and the top 15% of households have incomes of $\geq$$60,000.
- *Not* influenced by outliers – great descriptors no matter shape.
- *Good 3-number summary*: lower quartile (25th percentile), median (50th percentile), and the upper quartile (75th percentile), which describe
  - Central tendency = median (ie, the value in the middle when the data is arranged in order).
  - Spread = difference between the upper and lower quartiles (ie, the 'inter-quartile range', IQR).
  - Symmetry (ie, *skewness*) – compare the difference between the upper quartile and the median with the difference between the median and the lower quartile.

Summarizing individual **categorical variables**

▷ *Raw and relative frequencies*

- Raw: counts; number of times a particular value is obtained in the sample.
- Relative: proportions or percentages; frequency of a particular value divided by the total number of observations.

▷ Often presented in a *frequency table*.

▷ *Example*: Distribution of blood types in a sample of 25 people.

Blood Type	**% (N)**
A	20% (5/25)
B	32% (8/25)
O	32% (8/25)
AB	16% (4/25)

Summarizing **combinations of variables**

▷ Calculate the median, quartiles, mean, SD, etc of a continuous variable among the subsets with each value of a categorical variable.

▷ *Cross-tabulate* two (or more) categorical variables using *contingency tables* (allow you to report *marginal frequencies*).

▷ Example: Height and race of a sample of N = 2735 subjects receiving either a new drug or placebo.[1]

	N	**Drug** $N = 2165$	**Placebo** $N = 570$
Weight (lbs)	2661	191±50 (148 196 233)	188±48 (149 194 229)
Race	2696		
Afr American		41% (868/2134)	38% (215/562)
Caucasian		47% (996/2134)	50% (283/562)
Other		13% (270/2134)	11% (64/562)

[1] The number of *missing* values is inferred via the 'N' column and the denominator frequencies.

Inferential Statistics

Confidence intervals

▷ Wish to calculate/determine a characteristic of or a fact about a population – a population *parameter*.

▷ Because impossible to collect data from the entire population, must use the data collected in the sample to *estimate* the population parameter – a *point estimate*.

▷ Unlikely that the value of the point estimate will be equal to the value of the population parameter because have only collected one sample from the population.

▷ Therefore, the value of the point estimate is used to construct an *interval estimate* for the population parameter.

▷ Will be able to state, with some confidence, that the population parameter lies within this interval – thus, a *confidence interval*.

Confidence intervals, *cont'd*

▷ *Definition*: a y% confidence interval (CI) for an unknown population parameter Y is an interval calculated from sample values by a procedure such that if a large number of independent samples is taken, y% of the intervals obtained will contain Y.

▷ Most often report 95% confidence intervals.

▷ *Interpretation via an example*: "We are 95% confident that mean total cholesterol on this new statin will be 10 to 20 mg/dl lower than on the old formulation."

- *CANNOT STATE*: "There's a 95% probability that mean total cholesterol on this new statin will be 10 to 20 mg/dl lower than on the old formulation."

▷ A narrow 95% CI indicates high precision, whereas a wide 95% CI indicates inadequate sample size.

Hypothesis testing

▷ Wish to test a hypothesis about the value of a population parameter – eg, that it equals a specific value.

▷ In order to do so, we sample the population and compare our observations with theory.

- If the observations disagree with the theory, the hypothesis is rejected.
- If not, we conclude either that the theory is true or that the sample did not detect the difference between the real and hypothesized values of the population parameter.

▷ IMPORTANT: Hypothesis testing involves *proof by contradiction*.

- Support for our hypothesis is obtained by showing that the converse is false.

Hypothesis testing, *cont'd*

▷ Elements of a hypothesis test:

1. Null hypothesis (H_0): hypothesis under test; referred to as the 'straw man' – something set up solely to be knocked down.
2. Alternative hypothesis (H_a): usually the hypothesis we seek to support on the basis of the information contained in the sample.
3. Test statistic: a function of the sample measurements upon which the statistical decision will be based.
4. Rejection region: the values of the test statistic for which the null hypothesis is rejected.

▷ Decision:

- If for a particular sample, the computed value of the test statistic falls in the rejection region, we reject the null hypothesis and accept the alternative hypothesis.

Hypothesis testing, *cont'd*

▷ Question: How do we choose the values that mark the boundaries of the rejection region?

▷ Answer: Determined by the choice of α – the probability of rejecting the null hypothesis when the null hypothesis is true (ie, the probability of a *Type I error*).

▷ The value of α is also called the *significance level* of the test.

▷ Although small values of α are recommended, the actual size of α to use in an analysis is chosen somewhat arbitrarily.

- Two commonly used values are $\alpha = 0.05$ and $\alpha = 0.01$.

▷ NOTE: *Type II error* can also be made – failing to reject the null hypothesis when it is false.

- β (*probability* of a Type II error) and *power* (1-β; probability of correctly rejecting the null hypothesis) are discussed in the 'Sample Size' lecture.

Hypothesis testing, *cont'd*

▷ Still have a problem:

- One person may choose to implement a hypothesis test with $\alpha = 0.05$, whereas another person might prefer $\alpha = 0.01$.
- In turn, it is possible for these 2 people to analyze the same data and reach opposite conclusions – one concluding that the null hypothesis should be rejected at $\alpha = 0.05$; the other deciding the null hypothesis should *not* be rejected with $\alpha = 0.01$.

▷ Solution:

- Calculate the *p-value* – the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected (ie, the *attained significance level*).
- The smaller a *p*-value becomes, the more compelling the *evidence* that the null hypothesis should be rejected.
  - NOTE: a smaller *p*-value does not indicate greater *significance*.

Hypothesis testing, *cont'd*

▷ Assuming a specific value of α, the *p*-value can be used to implement an α-level hypothesis test:

- If the *p*-value $\leq \alpha$, then you reject the null hypothesis.
- If the *p*-value $>$ alpha, then you *fail to reject* the null hypothesis – *null hypothesis is never accepted*.

▷ REMEMBER: P-values provide evidence against a hypothesis, never evidence in favor of it.

▷ Quickly, back to hypotheses:

- The null hypothesis is usually stated as the absence of a difference or an effect.
- The alternative hypothesis can be *one-* or *two-sided*.
  - Two-sided – states only that a difference/effect exists (ie, the effect $\neq 0$).
  - One-sided – specifies the *direction* of the difference/effect (ie, the difference between group A and group is > 0.)

Hypothesis testing, *cont'd*

▷ Example:

- H_0: There is *no* difference between the true mean reaction times (to a stimulus) for men and women.
  - $\bar{x}_{men} = \bar{x}_{women} \rightarrow \bar{x}_{men} - \bar{x}_{women} = 0$.
- H_a: There is a difference (ie, $\bar{x}_{men} - \bar{x}_{women} \neq 0$).
- Data: Independent random samples of 50 men and 50 women – $\bar{x} \pm SD$ is 3.6 ± 0.18 seconds for the men, while 3.8 ± 0.14 seconds for the women.
- Based on the observed data, *p*-value $= 0.0124$.
- Thus, if $\alpha = 0.05$, we reject the null hypothesis and conclude that there is a *significant* difference between the true mean reaction times for men and women.
- However, if $\alpha = 0.01$, we *fail* to reject the null hypothesis and conclude that we failed to find a significant difference.

Hypothesis testing, *cont'd*

▷ Problems with *p*-values:

- *Statistical* significance does not indicate *clinical significance*.
- A small *p*-value by itself only tells half the story – it gives you no information about magnitude (and depending, direction).
- You can't make any conclusion from a large *p*-value (only perhaps that your sample size was too small).
  - 'Absence of evidence is not the evidence of absence'.

▷ Alternatives:

- Report estimated confidence intervals (CIs) in addition to *p*-values – can glean *clinical* significance.
- Perform hypothesis tests with CIs – look to see whether the CI contains the null value.
  - Example: With a CI for a difference, does it contain 0?
  - Example: With a CI for an odds ratio, does it contain 1?

Correlation

▷ A quantitative measure of association between 2 *continuous* variables (ie, the degree to which they change together).

- *Pearson correlation* describes the direction and relative *strength* of a *linear* relationship.
  - Always between -1 and 1.
  - The closer it is to ± 1, the closer to a *perfect* linear relationship.
  - Can use hypothesis test to determine if significant (ie, $\neq 0$).
- *Spearman's rank correlation* does not assume a linear relationship; only a *monotonic* one – when X increases, Y always increases or stays flat, or Y always decreases or stays flat.
  - Can also use hypothesis test to determine if significant.

▷ IMPORTANT:

- Neither is an estimate of the *slope*.
- Correlation $\not\Rightarrow$ Causation
- Correlation $\not\Rightarrow$ Agreement

Determining if an **association** exists

▷ Use *tests of association* to determine if two variables (one continuous and one categorical or both categorical) are *independent*.

- Two variables are associated if one variable affects the value/distribution of the values of the other.
- Example: Association between race (categorical predictor) and presence of disease (categorical outcome).

▷ Testing for a *difference* in the distribution of a variable *between* groups $\Leftrightarrow$ testing for an association between a group (predictor) variable and an outcome variable.

- Example: Difference in the distribution of cholesterol (continuous outcome) between genders (categorical predictor).
- Also includes *paired* data (eg, difference in the distribution of test scores before and after a didactics class).

Determining if an **association** exists, *cont'd*

▷ Incorporates hypothesis testing:

- *Null* hypothesis – proposes the variables are independent.
  - Example: There is no difference in the frequency of drinking well water between subjects who develop peptic ulcer diseases and those who do not.
- *Alternative* hypothesis – proposes that they are associated.
  - Can be either *one-sided* (eg, drinking well water is *more* common among subjects who develop peptic ulcers) or *two-sided* (eg, the frequency of drinking well water is different in subjects who develop peptic ulcers).

▷ If test's p-value $\leq \alpha$ (eg, 0.05), conclude that the two variables are significantly associated.

▷ p-value $> \alpha$ *does* not mean that there is no association *in the population*; only means you failed to find one *in your study sample*.

Determining if an **association** exists, *cont'd*

▷ When testing for a difference in a continuous outcome, commonly used tests of association (ie, t-test) assume the continuous outcome is normally distributed – *parametric* tests.

- *Non-parametric* tests don't assume normality; test is performed on the raw values converted to *ranks*.
- If normality holds, a non-par test is 95% as efficient as its par equivalent.
- If normality *does not* hold, non-par test can be arbitrarily more efficient and powerful than its par equivalent.
- Result of par test (ie, *p*-value) can be highly influenced by outliers; non-par tests are not (because based on ranks).
- Sometimes see others use tests/graphics to assess normality and run a par or non-par test based on the result.
  - Not a good approach – test of normality may not have adequate power to detect non-normality.

Determining if an **association** exists, *cont'd*

▷ Available tests of association:

Purpose	**Type of outcome**	**Test to use**
Compare paired responses	Continuous	Paired t-test [Wilcoxon signed-rank test]
Compare 2 (independent) groups	Continuous	Student's t-test [Wilcoxon rank-sum/ Mann-Whitney U test]
Compare >2 (independent) groups	Continuous	1-way ANOVA [Kruskal Wallis test]
Compare ≥2 (independent) groups	Categorical (≥2 levels)	Chi-square test (Fisher's exact test when cell counts <5)

- 'Group' defined by a categorical variable (≥2 levels).
- Non-parametric equivalent test given in [ ] brackets.

Determining if an **association** exists, *cont'd*

▷ IMPORTANT: Association $\not\Rightarrow$ Causation.

▷ Limitations to just performing tests of association:

- Blur the distinction between statistical & *clinical* significance.
  - Possible for a difference of little clinical importance to achieve a high degree of statistical significance.
  - Cannot conclude clinical relevance from small p-value.
- Very often, not only would you like to determine if an association or difference exists, but would also like to estimate the *magnitude* and *direction/shape* of the effect.
- Can only look at one pair of variables at a time (a single predictor variable and the primary outcome variable); cannot incorporate *confounders*. (True of correlation too!)
- Cannot incorporate other possible complexities of your data (eg, repeated measurements within subjects and cluster sampling).

Regression analysis

▷ Determining relationships and making predictions – an extension of testing for associations:

- Allows you to estimate the significance, direction/shape, and magnitude of the *effect* of $\geq$1 predictor variables on the outcome variable – ie, determine if the outcome is significantly affected by $\geq$1 of the predictors.
- Allows you to incorporate confounders – by *adjusting* the predictor-outcome association for the predictor-confounder and confounder-outcome relationships.
- Results may be used to *predict* the outcome of subjects that were not sampled but are from the same population.
- Specific regression analysis used based on type of outcome – Multiple Linear (continuous), Logistic (binary), Proportional Odds (ordinal), or Cox Proportional Hazards (time to event).

Regression analysis, *cont'd*

▷ IMPORTANT:

- All regression analyses make *assumptions* (eg, the observations are independent; variance of the error is constant).
  - Must assess whether the assumptions are violated.
- All regression analyses (by default) assume a *linear* relationship between each *continuous* predictor and the outcome.
  - Most can be extended to fit *non-linear* relationships (eg, by using *restricted cubic splines*).
- Interpretation of effect estimates depends on type of regression:
  - Linear example: Holding all other predictors constant, the outcome increases/decreases Y units per 1-unit increase of X.
  - Logistic example: The *odds* of the outcome occurring are Y times higher/lower for group X1 compared to group X2, holding all other predictors constant.
- 95% CIs should be reported for all effect estimates.

Pitfalls to avoid

▷ Pitfalls in regression modeling:

- Casewise deletion of missing data.
- Categorizing continuous variables.
- Not using clinical knowledge to specify the model.
- Inappropriate linearity assumptions.
- Using stepwise variable selection (ie, deciding based on p-values).
- Fitting more complex model than data allows (ie, overfitting).
- Lack of model validation (if appropriate).

▷ Pitfalls in reporting & analysis in general:

- Reporting only favorable results.
- Deleting 'outliers' based on observed response values.
- Non-reproducible analyses/results.

References

▷ "The Little Handbook of Statistical Practices" – Gerard Dallal.

- `http://www.tufts.edu/~gdallal/LHDP.HTM`

▷ *Mathematical Statistics with Applications* (5th edition) by Wackerly, Mendenhall, and Schaeffer.