STATISTICS PRIMER

Dorina Kallogjeri, MD, MPH
Washington University in Saint Louis

Disclosure

I have relevant financial relationship(s) with respect to this educational activity with the following organization(s):

Employee
Washington University in St. Louis

Research/Research Grants
National Institutes of Health (NIH)

Why do I need statistics?

Objectives

- Identify different data types (variables)
- Understand importance of data exploration and description
- Understand ways to identify relationships between variables

Statistics and Clinical Research

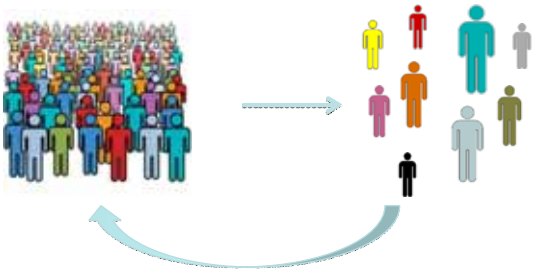
- Research question or Observation
- Theory or Hypothesis to test
- **Collect data to test Hypothesis (study design)**
- **Analyze data**
- **Conclusion**

Objectives

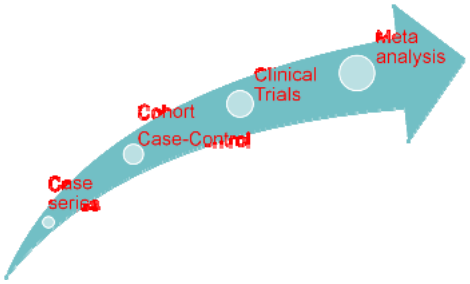
- Identify different data types (variables)
- Understand importance of data exploration and description
- Understand ways to identify relationships between variables

- **Descriptive Statistics**
  Describe and summarize data using numbers or graphs
- **Inferential Statistics**
  Analyze data from a sample representative of a population in order to make inferences about the population

Population versus Sample



Statistical Studies and Their strength



Hypothesis Testing

- Null Hypothesis (nothing is going on)
  - Claims the opposite of what you would like to be true

Example: There is no association between smoking and lung cancer

- Alternative hypothesis
  - Describes the situation when the null hypothesis is false

Example: Smokers have higher risk for lung cancer compared to non-smokers

  - One Sided
  - Two Sided

Norusis 1997, page 207

The Null Hypothesis, H_o

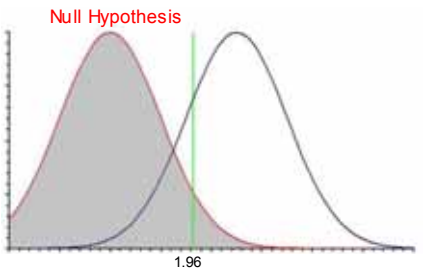
- No real (true) difference between group means or proportions
- Observed differences due to chance
- If null hypothesis is rejected, researcher concludes that the observed differences are statistically significant

The Alpha Level

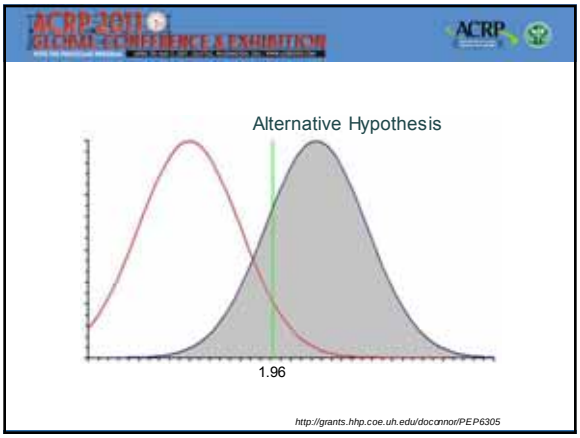
- Alpha is the conditional probability of rejecting a true null hypothesis
- Alpha level (α) - used to determine the value of the test statistic that causes us to reject or retain the null hypothesis
- By custom, the level of alpha is usually set at 0.05
- Type I error

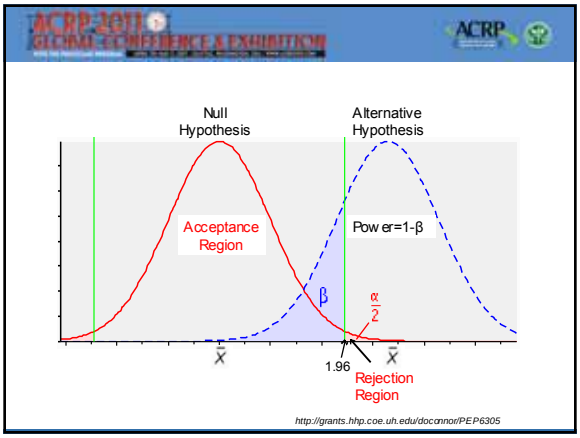
Power

- Power is the ability of a statistical test to detect true differences
- Statistical Power of a test is usually required 80% or more



http://grants.hhp.coe.uh.edu/doconnor/PEP6305





Truth in population and study results

Truth in population

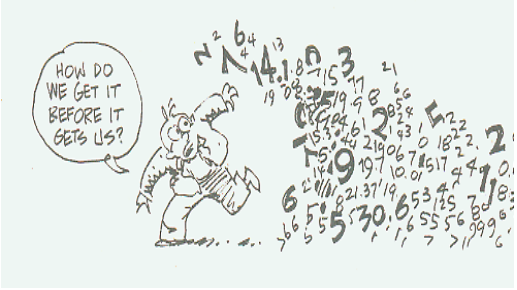
Study Results	Treatment A different from treatment B	No difference between treatments A and B
Treatment A different from treatment B		
No difference between treatments A and B		

a. Power-Correct to reject null hypothesis
b. β-error. Did not reject a false null hypothesis
c. α-error. Did not reject a true null hypothesis
d. Correct. Did not reject a true null hypothesis

Truth in population and study results

Study Results		Truth in population: Treatment A different from treatment B	Truth in population: No difference between treatments A and B
	Treatment A different from treatment B	Power Correct Reject Ho	α Type I error (reject Ho when true)
	No difference between treatments A and B	β Type II error Did not reject Ho when false	Correct Do not reject a true Ho

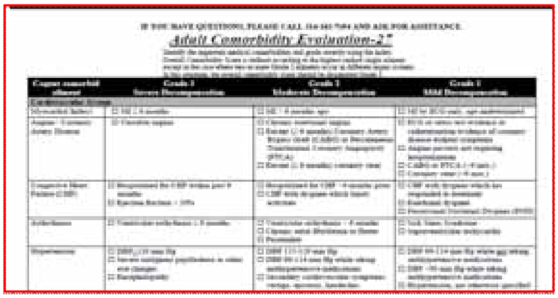
Data Collection



Example Of Data Collection Form

ACRP 2011 GLOBAL CONFERENCE & EXHIBITION

Example Of Data Collection Form



ACRP 2011 GLOBAL CONFERENCE & EXHIBITION

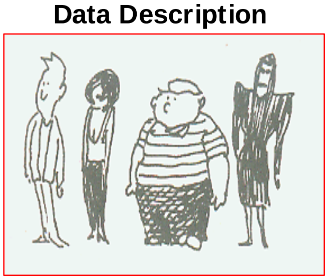
Excel Spreadsheet Example

D49

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	Gender	DOB	Age	Race	ZipCode	InitialDiag	TumorLoc	ClinT	ClinN	ClinM	PathT	PathN
2	1												
3	2												
4	3												
5	4												
6	5												
7	6												
8	7												
9	8												
10	9												
11	10												
12	11												
13	12												
14	13												
15	14												

ACRP 2011 GLOBAL CONFERENCE & EXHIBITION

Data Description



Variables

Characteristics or qualities that can be measured are called variables.

- Categorical
  - Dichotomous
  - Nominal
  - Ordinal
- Continuous
  - Interval
  - Ratio

Examples of variables

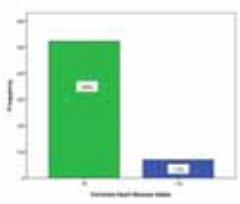
- Gender
- Race
- Having diabetes
- Having cancer
- Pain scale
- Age
- Height
- Weight
- Blood pressure
- Hemoglobin

Description of Categorical Level Variable

- Frequency
- Percent
- Bar graph

Description of Categorical Level Variable

Coronary Heart Disease	N (%)
No	523 (88%)
Yes	69 (12%)
Missing	15



Description of Continuous Level Variable

- Central tendency (center)
  - Mean (Average)
  - Median
  - Mode
- Distribution around center
  - Standard deviation
  - Quartiles
  - Range
- Histogram or Boxplot

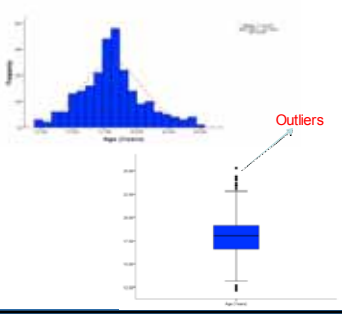
Description of Continuous Level Variable

Statistics

Age (Years)

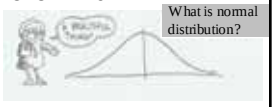
N	Valid	231
	Missing	0
Mean		18.06
Std. Error of Mean		.15
Median		18.01
Mode		12.17[a]
Std. Deviation		2.42
Range		13.12
Minimum		12.17
Maximum		25.29
Percentiles	25	16.56
	50	18.01
	75	19.17

a. Multiple modes exist. The smallest value is shown

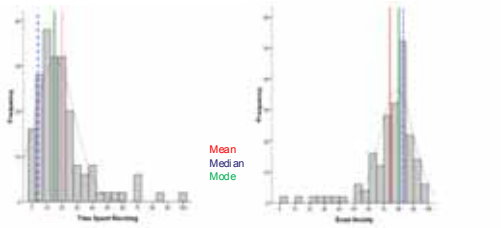


Normal (Gaussian) Distribution

- Variable ranges from – ∞ to +∞
- Two parameters: mean (μ mu); std. dev. (σ sigma)
- Symmetric about its mean, μ (mean μ=mode=median)
- Total area under the curve is = 1.0
- Bell-shaped curve
- Easy to use



Where is the center?



Positively skewed distribution

Negatively skewed distribution

Bivariate Analysis

- Investigates the relationship between 2 variables
  - Does a relationship exists? – statistical tests and p-values
  - How strong is an existing relationship?

p Stands For Probability

- p values - The observed value from a statistical test
  - Probability that differences identified by the test are due to chance alone
- If *p* value is smaller than alpha level, then we reject null hypothesis ($p<0.05$)

Bivariate Analysis - Choosing Correct Test

Dependent variable or outcome

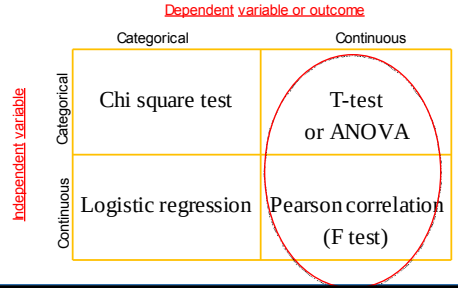
Independent variable	Categorical	Continuous
Categorical		
Continuous		

Bivariate Analysis - Choosing Correct Test

Dependent variable or outcome

Independent variable	Categorical	Continuous
Categorical	Chi square test	T-test or ANOVA
Continuous	Logistic regression	Pearson correlation (F test)

Bivariate Analysis - Choosing Correct Test



Assumptions For Parametric Tests

- Interval or ratio level data
- Normally distributed data
- Variation of data among groups is equal
- Independent and random observations

Note: Sample size is a factor to be considered

Type Of Data And Appropriate Statistical Test to Investigate Presence of Relationship

Chi Square Test

- Relationship between two categorical variables
- Null hypothesis is that population proportions are equal in the 2 or more groups we are studying
- Determine the difference between observed and expected frequencies

Is there a relationship between presence and severity of comorbidities at time of breast cancer diagnosis and 5 year survival rate? (n=1635)

	5-year survival rate
Comorbidity	Alive
None	644 (71%)
Mild	251 (63%)
Moderate	131 (57%)
Severe	26 (29%)

Chi square p-value<0.001, so conclude that there is a relationship

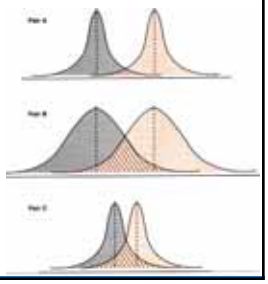
t-test

- Compares the means of the two groups (Relationship between a continuous level variable with a categorical level variable)
  - Independent groups=Independent Samples t-test
  - Matched groups=Paired Samples t-test
- Null hypothesis is that population means are equal
- Formula is different for each test

Example

Which one of the three figures shows means that are significantly different from each other?

1. Pair A
2. Pair B
3. Pair C

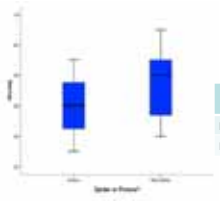


Independent Versus Paired Analysis Examples

Example

Do people with arachnophobia who are exposed to a picture of a spider have the same level of anxiety as people exposed to a real spider?

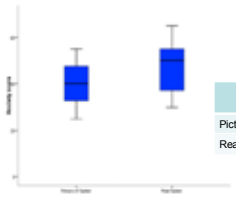
Two different groups:



	Anxiety score Mean	Anxiety score Standard deviation
Picture (n=12)	40	9.3
Real Spider (n=12)	47	11.0

Independent samples t-test p=0.107, conclude that the average anxiety level is the same for the 2 groups

Same group of subjects



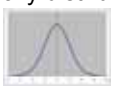
	Anxiety score Mean	Anxiety score Standard deviation
Picture (n=12)	40	9.3
Real Spider (n=12)	47	11.0

Paired samples t-test p=0.031, conclude that the average anxiety level in seeing a real spider is significantly different from anxiety from seeing a picture of it.

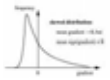
Parametric Versus Non-parametric test

Type Of Data And Appropriate Statistical Test

Are the data normally distributed



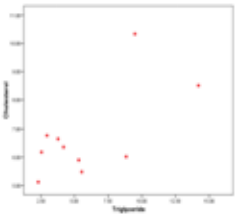
Or not?



Parametric tests more likely to give significant results

Are triglyceride levels linearly correlated with cholesterol levels?

Cholesterol and Triglyceride levels (n=10 patients)



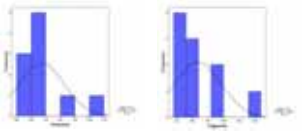
Parametric test

Correlations

		Cholesterol	Triglyceride
Cholesterol	Pearson Correlation	1	.650*
	Sig. (2-tailed)		.042
	N	10	10

Parametric Vs Non-parametric Test

Are Cholesterol and Triglyceride values normally distributed?



Parametric test

Correlations

		Cholesterol	Triglyceride
Cholesterol	Pearson Correlation	1	.650*
	Sig. (2-tailed)		.042
	N	10	10

Nonparametric test

Correlations

		Cholesterol	Triglyceride
Spearman's rho	Cholesterol Correlation Coefficient	1.000	.418
	Sig. (2-tailed)	.	.229
	N	10	10

Precision of Results –Confidence Intervals

- The quantity obtained (i.e. mean difference, proportion difference) is just an estimate of truth.
- 95% CI – 95% of such intervals includes the true value
- CI is wider in a smaller sample pointing to lack of information
- CI conveys more information than *p* values because they indicate a range of values for the true effect that is compatible with the sample observations.
- CI should not contain the point estimate of null hypothesis for the results to be significant

Relationship Between CI and P

A p value of <0.05 will correspond to a 95% confidence interval that excludes the value indicating equality (the null hypothesis value)

Confidence Intervals

Example – Anxiety score looking at a spider or a picture of a spider

Real Spider: Mean anxiety score 47
Picture: Mean anxiety score 40
Difference in mean anxiety scores: 7
95% CI: .77 to 13.23

What is this telling us?

Conclude with 95% confidence, the true difference of mean anxiety scores can be as small as .77 or as large as 13.23

Confidence Intervals

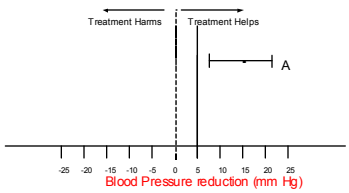
Example -- Cancer Survivorship Study

Two different treatment arms

Difference in mortality: 10%
95% CI: 2% to 18%

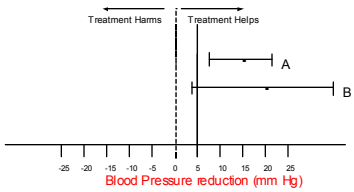
What is this telling us?
Conclude with 95% confidence, the true mortality difference can be as small as 2% or as large as 18%

Statistical and Clinical Significance



Trial A: CI excludes zero so statistical significance and lower bound excludes 5mm Hg reduction in blood pressure so clinically important difference.

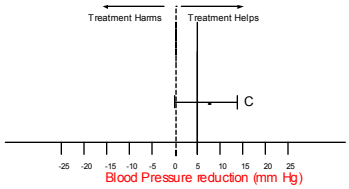
Statistical and Clinical Significance



Trial A: CI excludes zero so statistical significance and lower bound excludes 1% absolute risk reduction so clinically important difference.

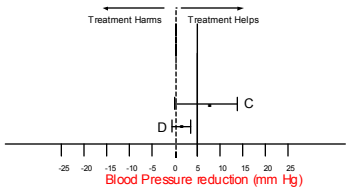
Trial B: CI excludes zero so statistical significance; drug appears to have promise, but lower bound of CI doesn't exclude the possibility of 5mm Hg reduction in blood pressure (clinical importance).

Statistical and Clinical Significance



Trial C: Lower bound of CI includes zero but upper bound suggests the possibility for clinical significance

Statistical and Clinical Significance



Trial C: Lower bound of CI includes zero but upper bound suggests the possibility for clinical significance

Trial D: Lower bound of CI includes zero, because upper bound doesn't cross 5mm Hg reduction in blood pressure there is little chance drug tested in Trial D will be clinically significant.

Variable types

- Independent - characteristics observed (exposure) and used to predict the outcome
- Dependent – The outcome that we want to predict
- Confounder – A variable related to both the exposure and the outcome, but not in the casual pathway

Example of confounding

Relationship of interest: smoking and MI
Gender is a potential confounder because:

- Smoking is more common in men (related to exposure)
- MI is more common in men (related to outcome)
- Gender is not in the causal chain between smoking and MI

* Enc of Epi I:143

Multivariate analysis

- Use a group of variables or characteristics (independent variables) to predict the outcome of interest (dependent)
- Method used depends on type of dependent variable
- Examples

  Logistic regression- Dependent variable is categorical

  Linear regression- Dependent variable is continues

  And many more….

Logistic regression

- Dependent variable is dichotomous
- One or more variables can be used to predict the outcome
- The impact of each variable is presented as Odds Ratios (OR) and their 95% Confidence Intervals

Example -Logistic regression

Question: Is smoking a predictor of coronary heart disease (CHD)? (N

OR=1.9; 95% CI: (1.06 to 3.35)

What does this mean?:

People who smoke are 1.9 times more likely to develop CHD.

We are 95% confident that people who smoke can be from 1.06 to 3.35 times more likely to develop CHD

Example -Logistic regression

Question: Is smoking a predictor of coronary heart disease (CHD) after controlling for age?

OR=2.21; 95% CI: (1.23 to 3.98)

What does this mean?:

We are 95% confident that after controlling for the effect of age people who smoke can be as low as 1.23 times and as high as 3.98 times more likely to develop CHD than people who do not smoke

Linear regression

- Dependent variable is continuous
- One or more variables can be used to predict the outcome
- The impact of each variable is presented as estimates of beta coefficient in a linear equation line (and their 95% Confidence Intervals)

Linear regression example

Is height a predictor of weight?

What else do we need to know?

Linear regression example

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write

H.G. Wells

References

- Larry Gonick, Woollcott Smith: “The Cartoon Guide to Statistics”
- Andy Field: “Discovering Statistics Using SPSS”

Questions