Hartford Hospital Research Program Research Methods Lecture Series Part III :

Choosing the Appropriate Statistic (Part I)

Dec 7, 2009

Yes! We're going to talk about research methods!

OVERVIEW:

- October: Basic concepts of research design
- November: Concepts of inferential statistics
- December: Choosing the right statistic Part I
- January: Choosing the right statistic Part II
- February: Meta analysis and clinical trials
- March: Grant-writing

Presenters:

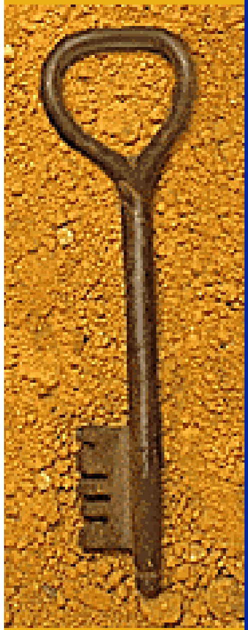
- Tara McLaughlin, PhD, MPE

  Senior Scientist/ Grantwriter (tmclaug@harthosp.org)
- Jyoti Chhabra, PhD

  Senior Scientist (jchhabr@harthosp.org)
- Ilene Staff, PhD

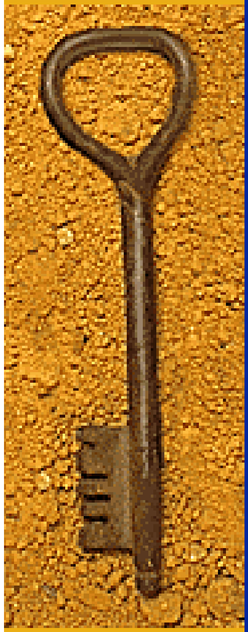
  Senior Scientist (istaff@harthosp.org)
- David O'Sullivan, PhD

  Senior Scientist (dosulli@harthosp.org)



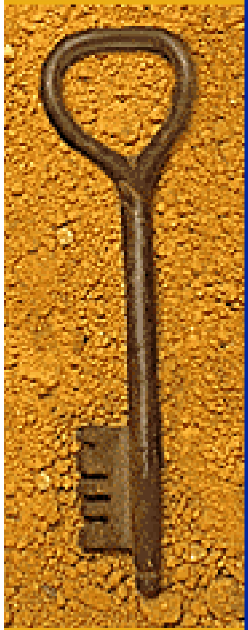
Contents of the Presentation

- Data Distribution Assumptions, Parametric or Nonparametric
- Descriptive stats:
  - Measure of central tendency and variance
  - Measures of frequency and proportion
- Univariate and Multivariate analysis
- Measures of association or Correlation
- Linear and multiple regression



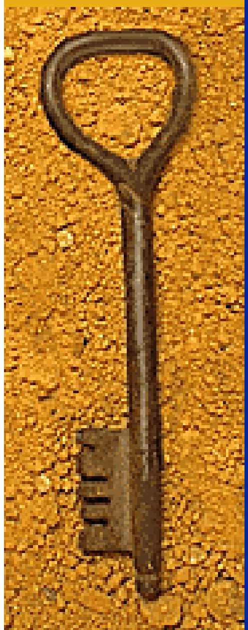
Data Distribution Assumptions

- Many statistical tests and procedures are based on specific data distribution assumptions.
- The assumption of normality is common in classical statistical tests.
- A normal data distribution pattern occurs in many natural phenomena, e.g. ht, wt.

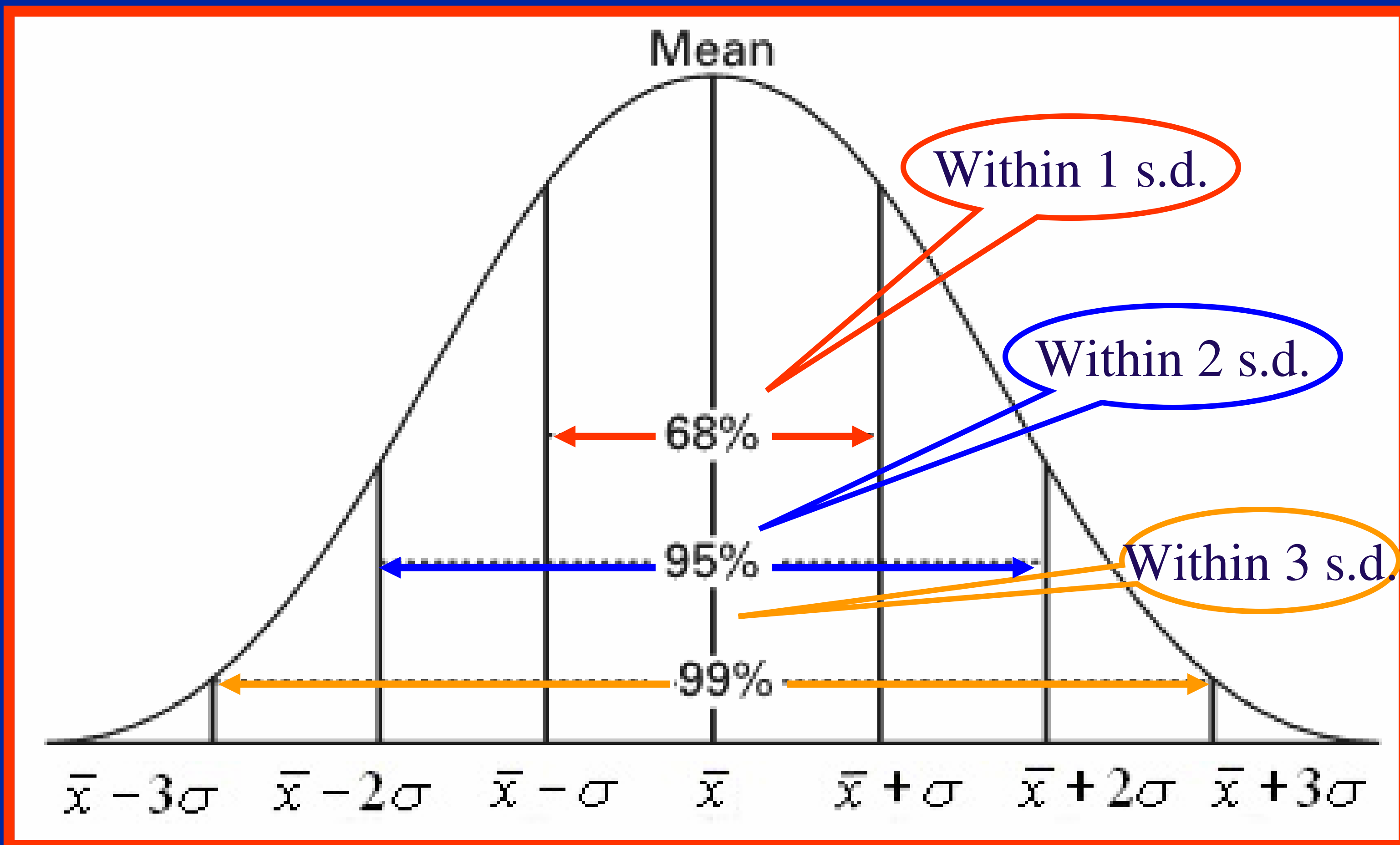


Normal Distribution Assumption

- Normal distributions form symmetrical bell shaped curve with a single central peak at the mean of the data.
- Spread of a normal distribution is dependent on standard deviation. The smaller the standard deviation the more concentrated the data.
- The mean & median are the same in a normal distribution.



Example: Normal Distribution



Much reliability modeling is based on the assumption that the distribution of the data is normal.

Parametric and Non-Parametric Techniques

- PARAMETRIC: Statistical techniques based on the assumption that the data is normally distributed are called parametric techniques.

- NON-PARAMETRIC: Statistical techniques that do not assume that the data are normally distributed are called non-parametric techniques.

Non Parametric vs. Parametric Techniques

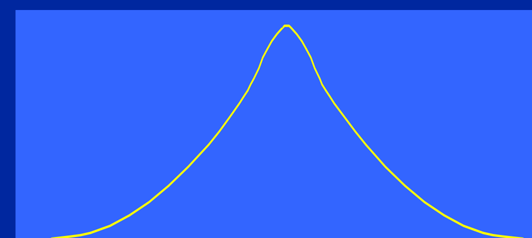
- Non parametric techniques are robust. They perform well under a wide range of distributional assumptions.
- However, techniques based on specific distributional assumptions are more powerful than non-parametric & robust techniques.
- Powerful - the ability to detect a difference when that difference actually exists.
- Therefore, if the distributional assumption can be confirmed, the parametric techniques are generally preferred.

Need to Examine the Distribution to Help Decide What Technique to Use

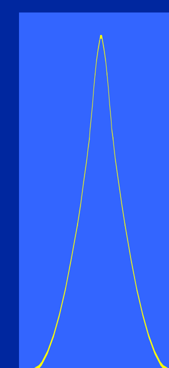
- If you are using a technique that makes a normality assumption, it is important to confirm that this assumption is in fact justified.
- If it is, the more powerful parametric techniques can be used.
- If the distributional assumption is not justified, a non-parametric technique may be required.

Parametric Analysis

- Requires continuous variables



- Assumes that variables are normally distributed (i.e., have a bell-shaped curve)



- Assumes equality of variance between groups (e.g. when using T-test or ANOVA)
- When appropriate, offers greater statistical power

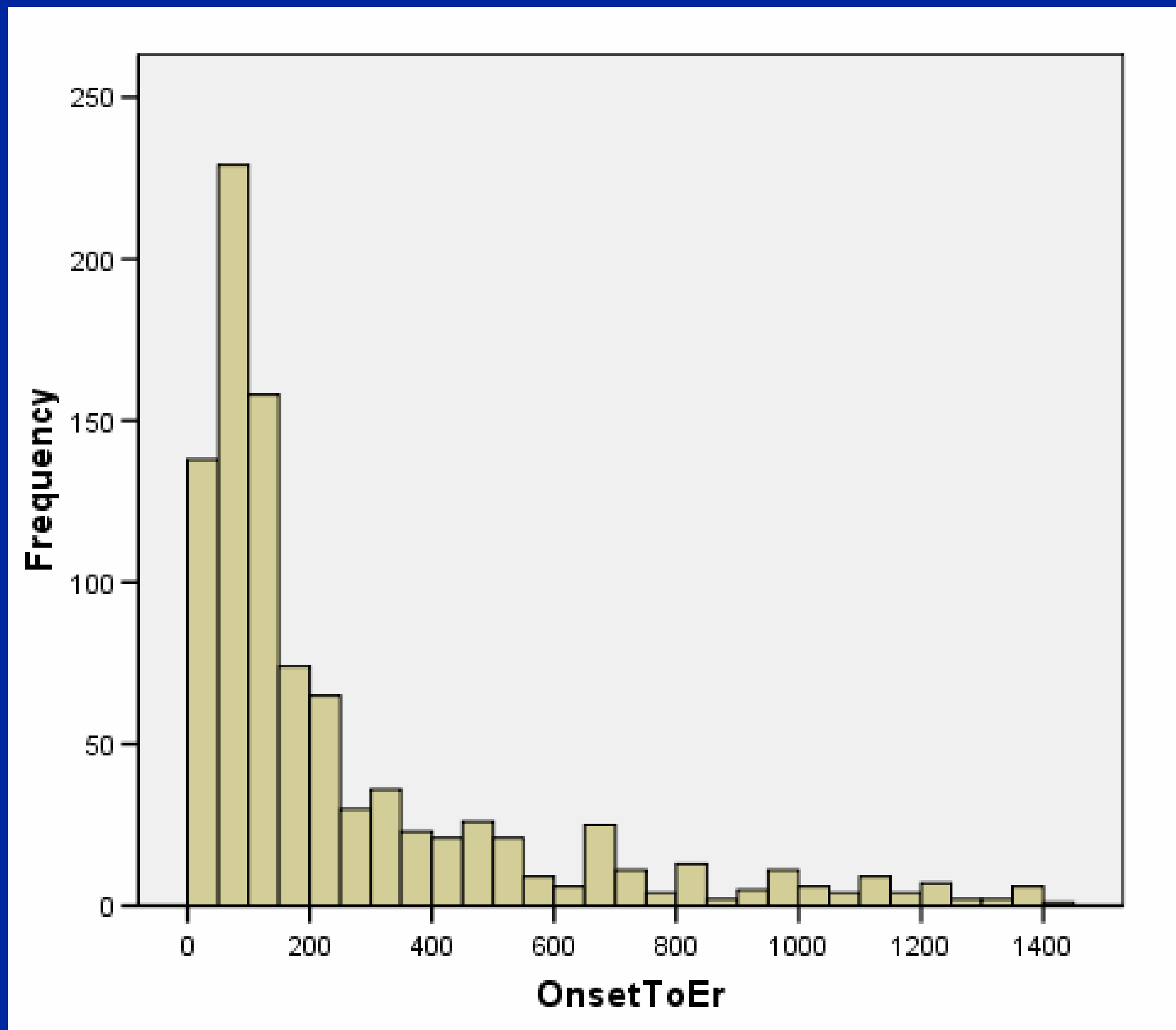
Non - Parametric Analysis

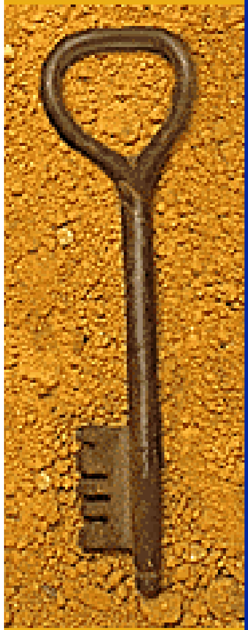
- Used for
  - nominal (categorical-#s assigned for convenience) e.g. race
  - ordinal (sequence along a continuum) level of measurement e.g. satisfaction score (1-5)

- Also used if continuous or ratio level measures cannot meet assumptions for parametric distribution

- Non-parametric tests are “distribution free”; i.e., there are no assumptions about data distribution
  - skewed
  - bi-modal

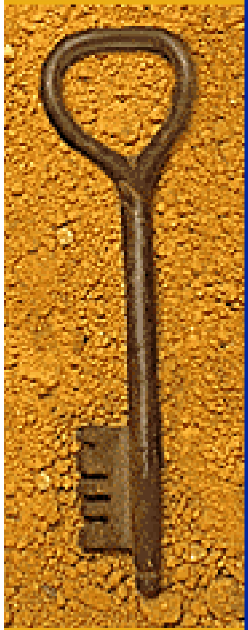
Examples of Non-Normal Data Distribution



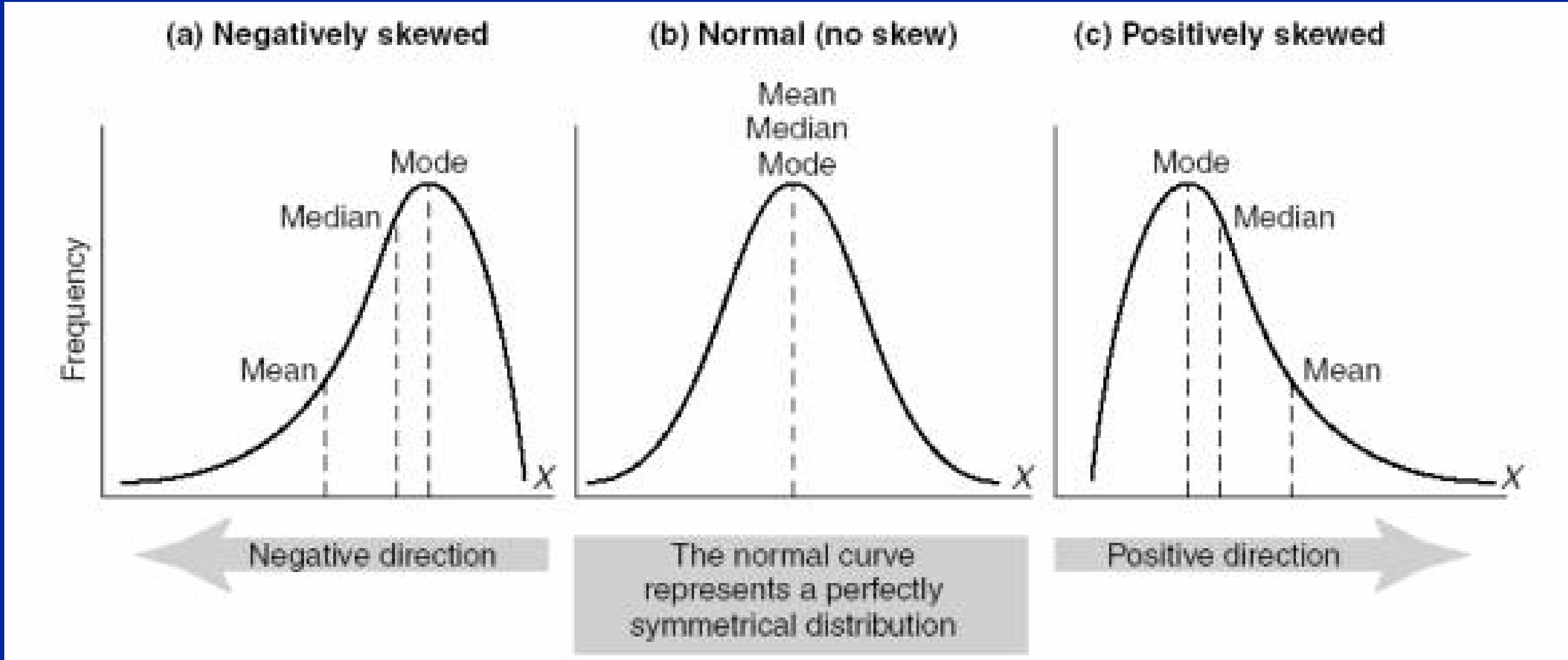


Understanding Violations of Normality

- Normality has two dimensions
  - Skewness
  - Kurtosis



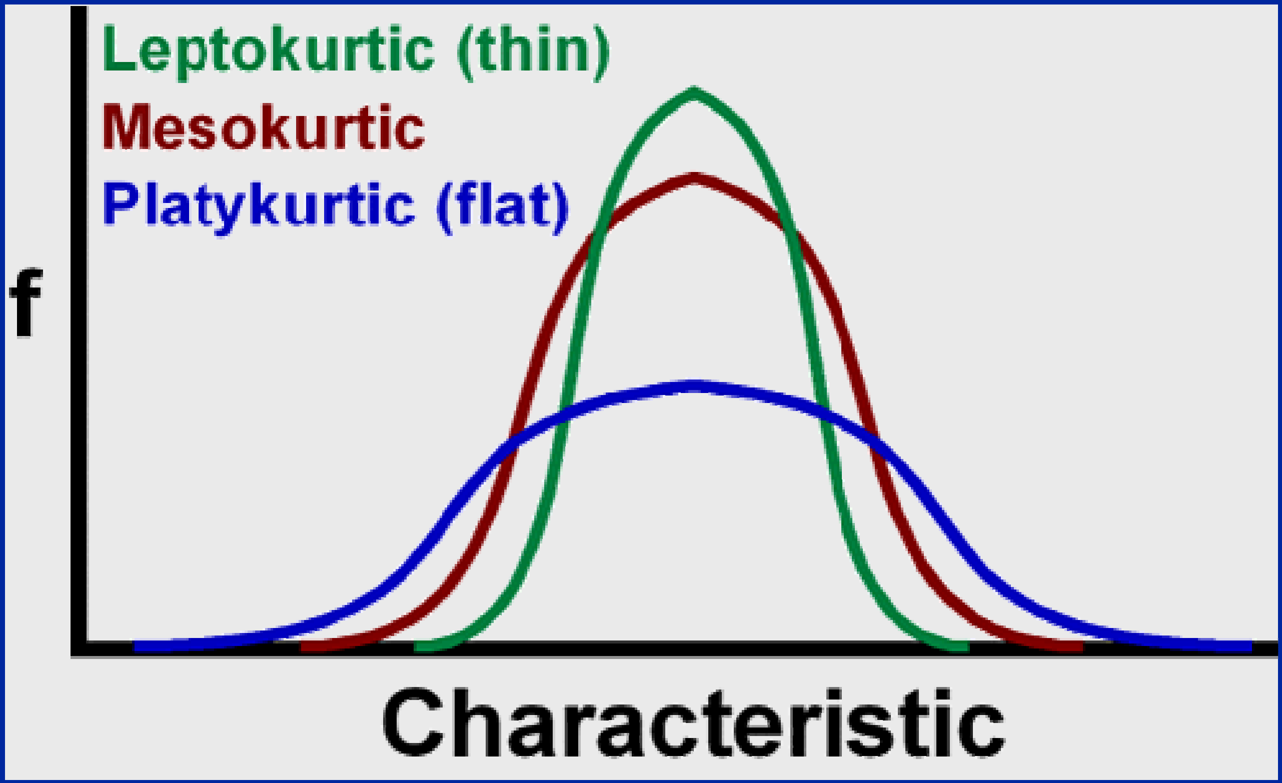
Skewness



Skewness of 1 or -1 is a sizable departure from normality
Skewness of normal distribution =0

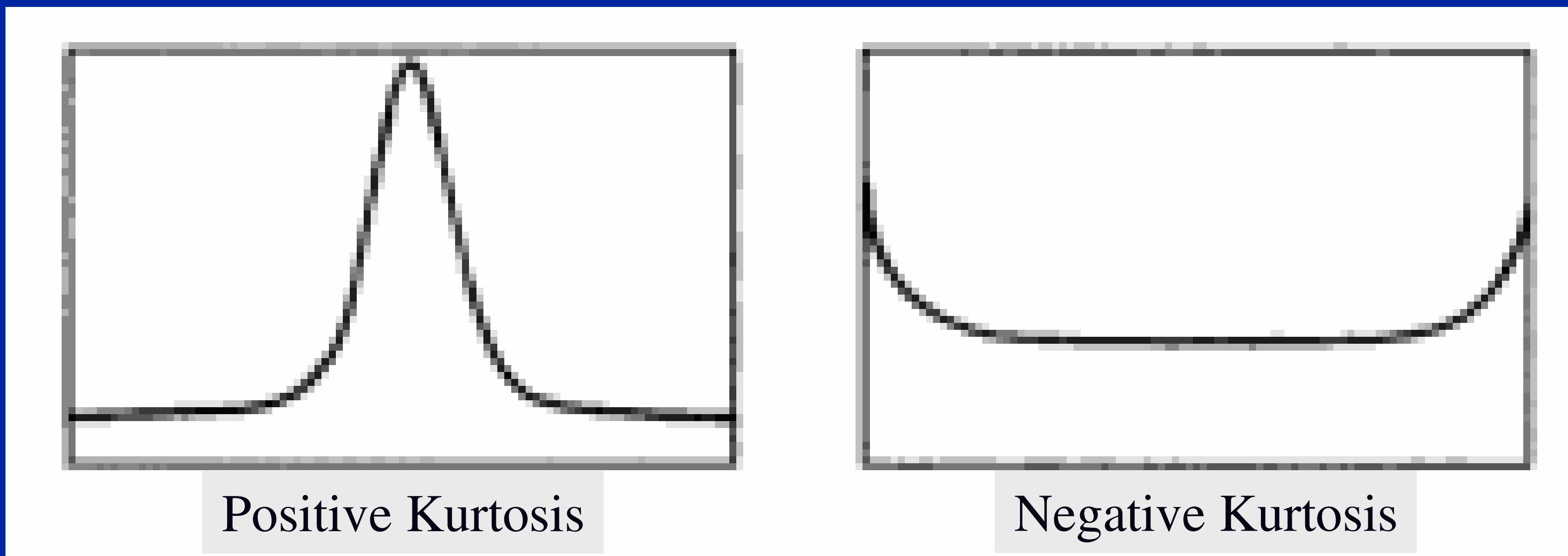
- Mean – average value
- Median – value of the point which has 1/2 the values on the lower side & 1/2 on the higher side.
- Mode – value with greatest frequency in distribution

Kurtosis



- Kurtosis of a normal distribution is 3.
- For interpretation ease, most statistical packages subtract 3 from the kurtosis, thus making the kurtosis for a normal distribution =0.

Examples where kurtosis is not=0

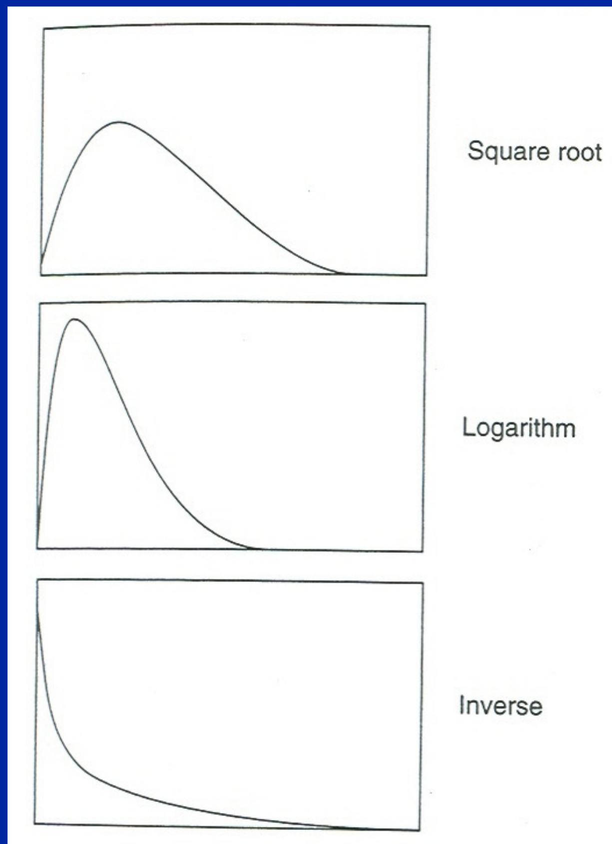


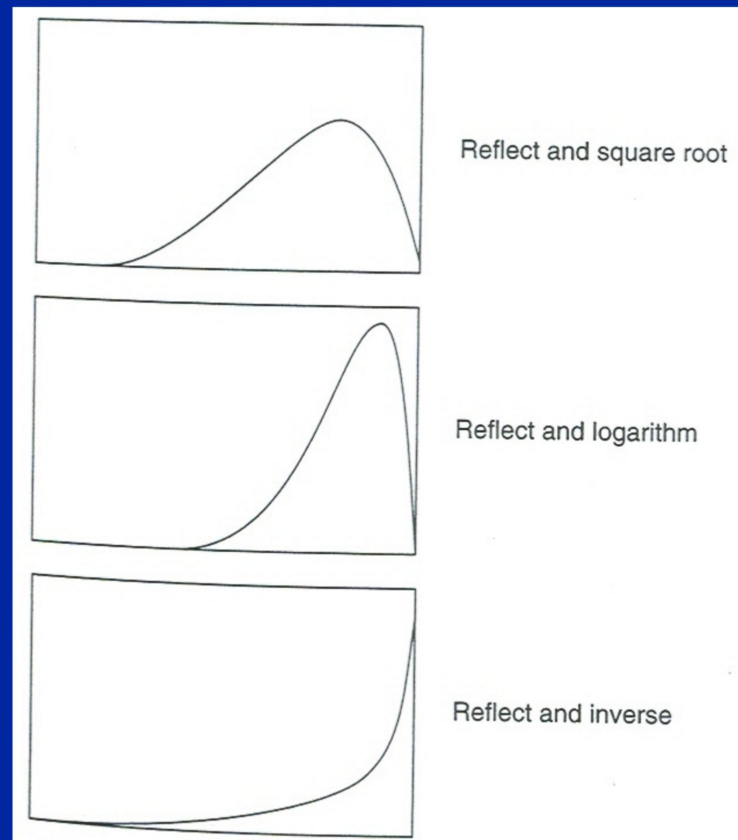
From Tabachnick & LS Fidell (Eds.), (2001)
Using multivariate statistics (4th ed.)

Limits to Violations of Assumptions for Parametric Analysis

- Parametric tests are said to be quite tolerant in terms of withstanding some violations of assumptions
- How tolerant are they? What is acceptable for using parametric?
- For normality: need symmetry and similar shape
- For variance: largest variance no more than 4x smallest

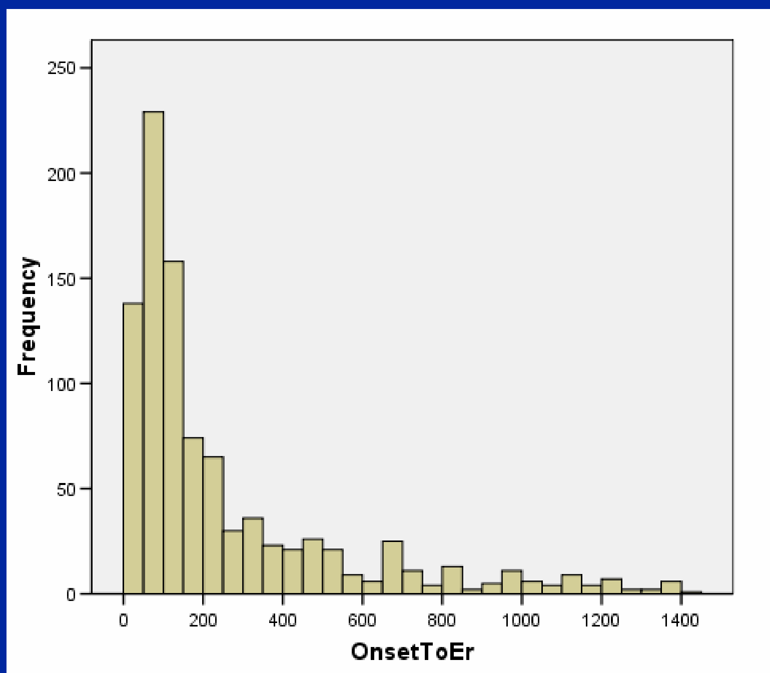
When assumptions are violated . . . "Getting Back to Normal"



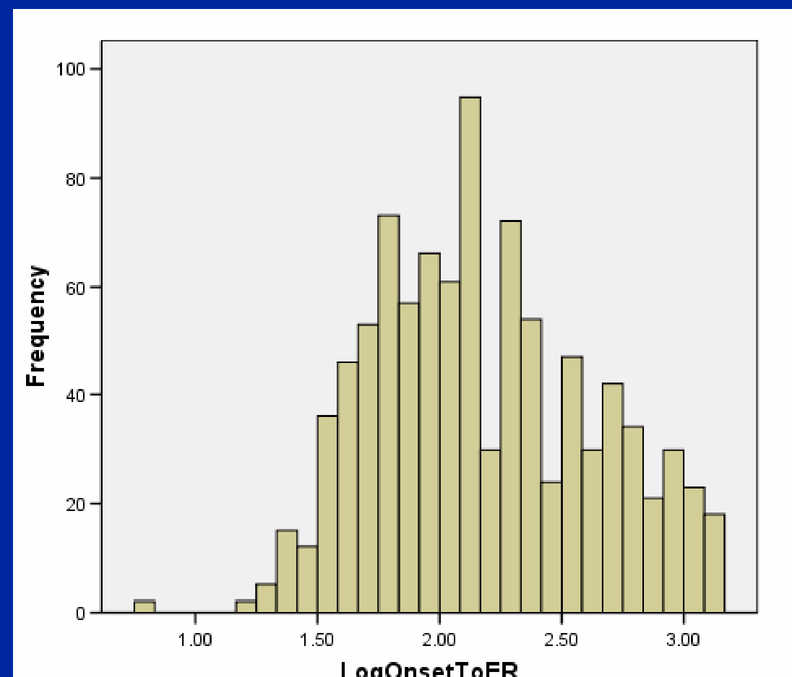


Tabachnick BG, Fidell, LS: Using Multivariate Statistics, 3rd Ed., HarperCollins, 1996.

Example of before and after Log Transformation



BEFORE
Skewness 1.5, Kurtosis 3.2



AFTER
Skewness 0.3, Kurtosis -0.6

Descriptive Statistics

- Frequency distributions
- Are continuous variables normally distributed?
- Measures of central tendency: mean, median, mode
- Measures of variability: variance, standard deviation, range
- Measures of association: crosstabulation, correlation

Univariate Statistics

- One dependent measure – differences among groups, comparison to population estimates

- Example:
  - Is BMI significantly different between patients with DM vs those without a diagnosis of DM?
  - Is BP significantly different between patients with BMI $\geq$ 30 vs < 30?

Bivariate Statistics

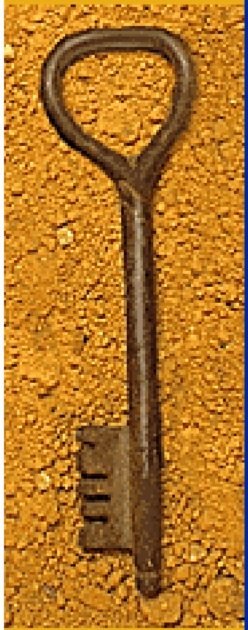
- Association between two variables, e.g., correlation
- Examples:
  - Is BMI related to FPG?
  - Is FPG related to HbA1c?
  - Is BMI related to BP?
  - Is BMI related to TC/TG?

Multivariate Statistics

- Techniques used when there are multiple independent variables and/or multiple dependent variables all correlated with each other to varying degrees
- Better for complex interrelationships
- Example:
  - Relationship between BP, BMI, TG, TC, HDL-c, LDL-c, and FPG, HbA1c.

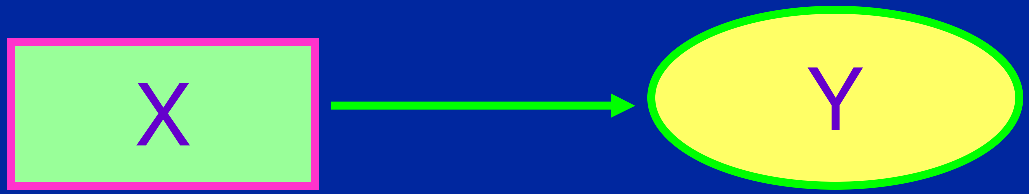
Association

- Correlation indicates relative strength and direction of relationship between two variables.
- Correlation coefficient (-1 to 1)
  - positive number indicates *direct* relationship (e.g., as one variable ↑, the other variable ↑)
  - negative number indicates *inverse* relationship (e.g., as one variable ↑, the other variable ↓ [or vice versa])
  - size of number (value 0 to 1) indicates strength



Prediction

- Linear Regression used to establish this relationship as a basis for prediction
  - x = independent (predictor) variable
  - y = dependent (criterion) variable

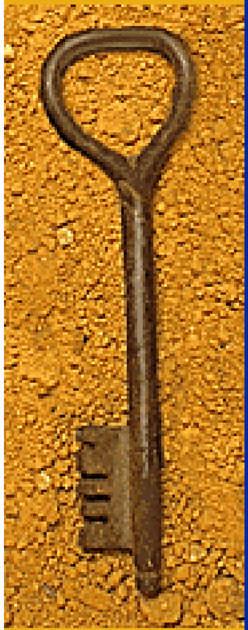


Analysis of Proportions

- Nominal level of measurement (categorical variables)
- Two or more groups, or two or more repetitions
- Example – Is the proportion of patients who received a certain treatment different by treatment site?

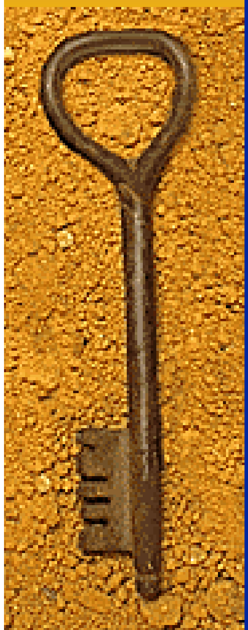
To Examine the difference between 2 Proportions

- Chi square (χ^2) test
  - Used in analysis of difference between two proportions
  - Chi-square test is used to calculate the probability of the difference between data observed and data expected. If there’s no difference; express as simple %
  - Chi square test gives an estimate of true chi-square and is not reliable when cell size is < 5
  - Fisher's Exact test (used when n<5) calculates an exact probability value for the relationship between two dichotomous variables



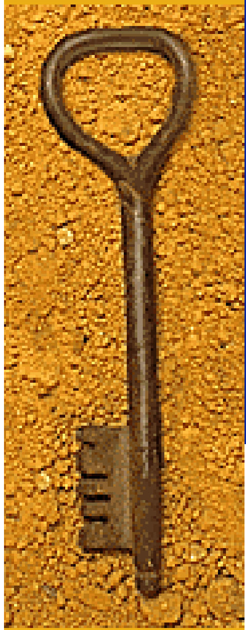
To analyze difference between two proportions- Chi Square Example

- Question: Was Thrombolytic administration (Yes/No) significantly different by gender (M/F)?



Thrombolytic Administration by Gender

			Gen		Total
			Male	Female	
Was Thrombolyti Administere	No	Count	354	377	731
		% within Gen	77.0%	80.6%	78.8%
		% of Total	38.1%	40.6%	78.8%
	Yes	Count	106	91	197
		% within Gen	23.0%	19.4%	21.2%
		% of Total	11.4%	9.8%	21.2%
Total		Count	460	468	928
		% within Gen	100.0%	100.0%	100.0%
		% of Total	49.6%	50.4%	100.0%



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Squ	1.797	1	.180		
Fisher's Exact Te				.199	.104
N of Valid Cases	928				

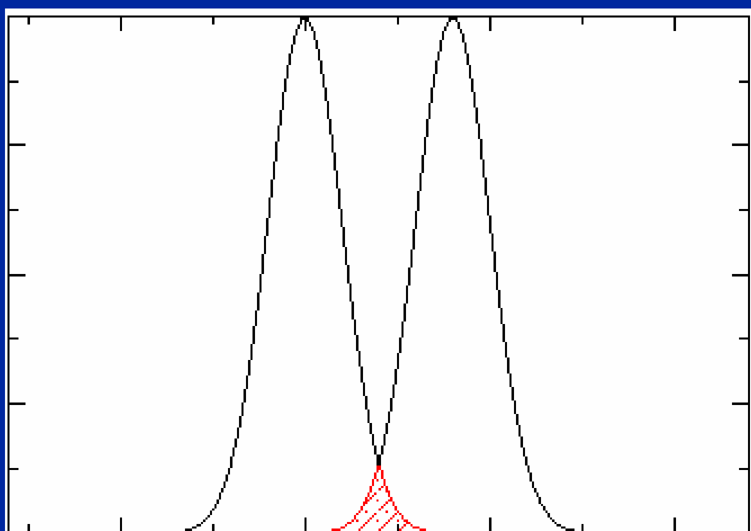
Analyzing Differences in Means Which Test(s) to Use

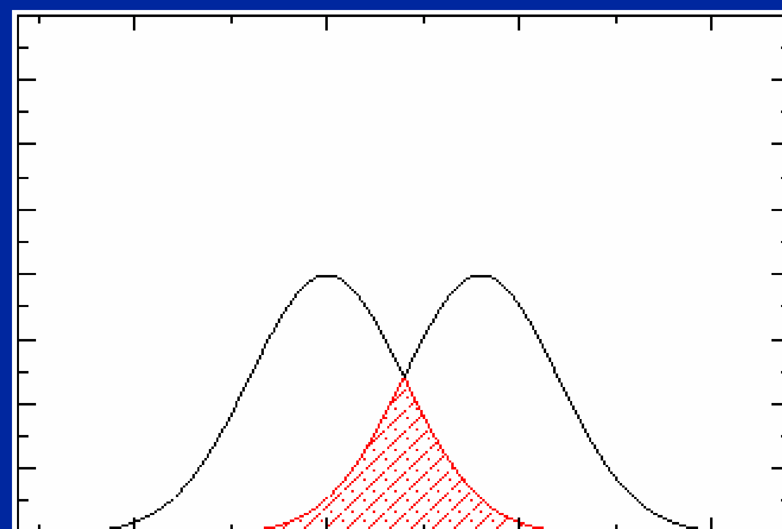
	Parametric Measures	Non-parametric Measures
Independent Groups	T-test (2 groups) One way ANOVA (>2 groups) Factorial ANOVA (2 or more independent variables)	Mann-Whitney or Wilcoxon Ranked Sum (2 groups) Kruskal-Wallis (>2 groups)
Repeated Measures	Paired t-test (2 groups) Repeated Measures ANOVA or GLM	Wilcoxon Signed Rank (2 groups) Friedman (>2 groups)

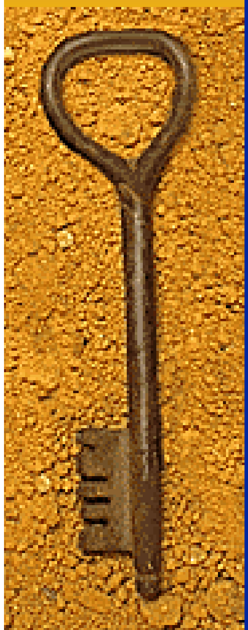
Parametric ANOVA or GLM can be expanded to include covariates (ANCOVA), multiple dependent vars (MANOVA)

E.g. Testing difference between means in 2 independent grps – One way ANOVA

- Does long term glucose control (HbA1c) vary by gender?
- M=9.1, F=9.6

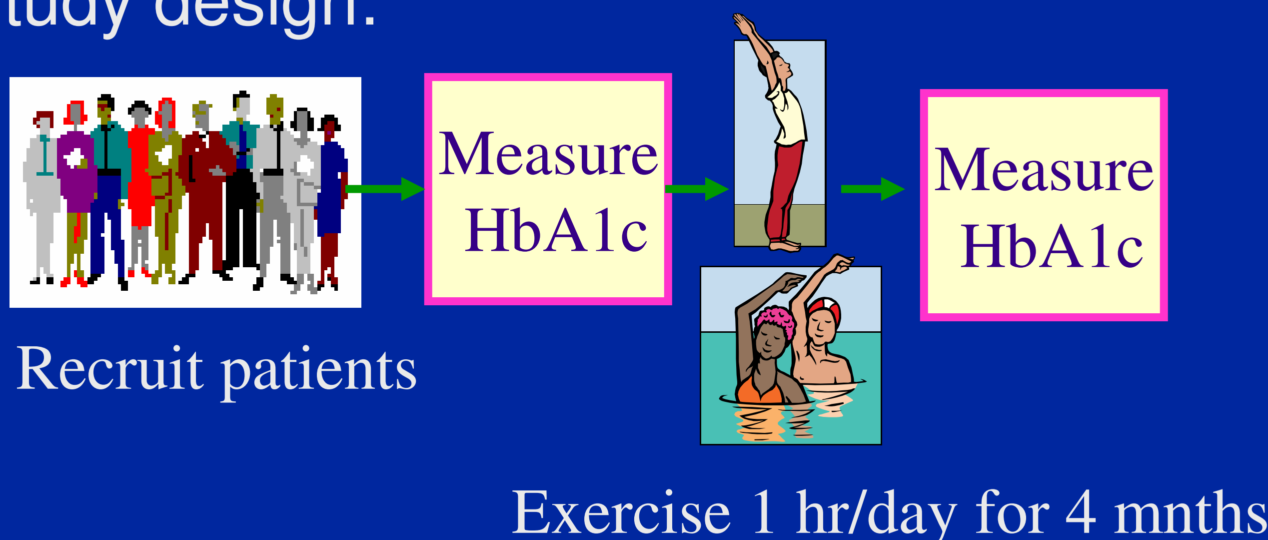


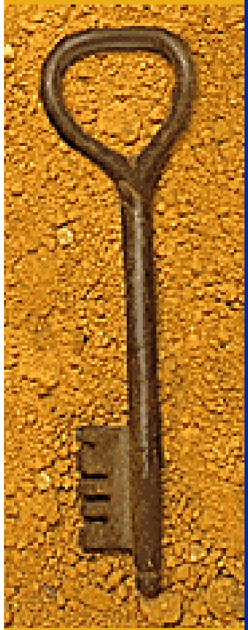




For testing whether means are different at two time points within same individuals - Repeated Measures ANOVA

- Dataset where all participants of a random sample are measured under a number of different conditions.
- Example: Does exercise influence HbA1c levels?
- Study design:





Corresponding Non-parametric Test for each Parametric Test

Parametric	Nonparametric
2 independent sample t-test	Mann-Whitney U Test
2 dependent sample t-test	Wilcoxon Rank Sum Test
One Way Anova	Kruskal-Wallis H Test
Pearson correlation	Spearman Rank Correlation

Involve Matched Pairs & Repeated Measures

Multiple Regression

- An extension of simple linear regression analysis
- Analyzes the strength of the relationship between one dependent variable and a set of predictor variables
- Can provide information concerning the relative strength among the predictors - but be careful
- Does not demonstrate causality

Multiple Regression: Standard, Hierarchical, and Stepwise

- In a (standard) multiple regression analysis, the researcher decides how many predictors to enter and all the predictors enter the regression model simultaneously.

- In a hierarchical multiple regression, the researcher decides not only how many predictors to enter but also the order in which they enter. Usually, the order of entry is based on logical or theoretical considerations.

- In a stepwise multiple regression analysis, the number of predictors to be selected and the order of entry are both decided by **statistical criteria** (e.g., entry or removal criterion).

Coming Attractions – next month. . .

Choosing the appropriate statistics (Part II)

Questions??