

# Introduction to Biostatistics And Epidemiology

David Dayya, D.O., M.P.H.

Saint Barnabas Hospital

Department of Family Practice

# Statistics

- The analysis an interpretation of data (s. datum) with a view toward objective evaluation of the reliability of the conclusions based on the data.

# Data Types

- Nonparametric Data
- Nominal or Categorical Data – Count data, devoid of rank/order or magnitude i.e. sex, ethnicity, survival.
- Ordinal Data – has rank/order and magnitude but does not maintain equidistant intervals between successive values i.e. small, medium, large; grade 1-5, stage 1-5.
- Parametric Data
- Interval Data – number-line data has rank/order, magnitude, and maintains equidistant intervals between successive values i.e. blood pressure, age, cholesterol, blood sugar, weight, temperature, time.
- Ratio scale – Interval data with a meaningful 0-value i.e. blood pressure, age, cholesterol, blood sugar, weight.

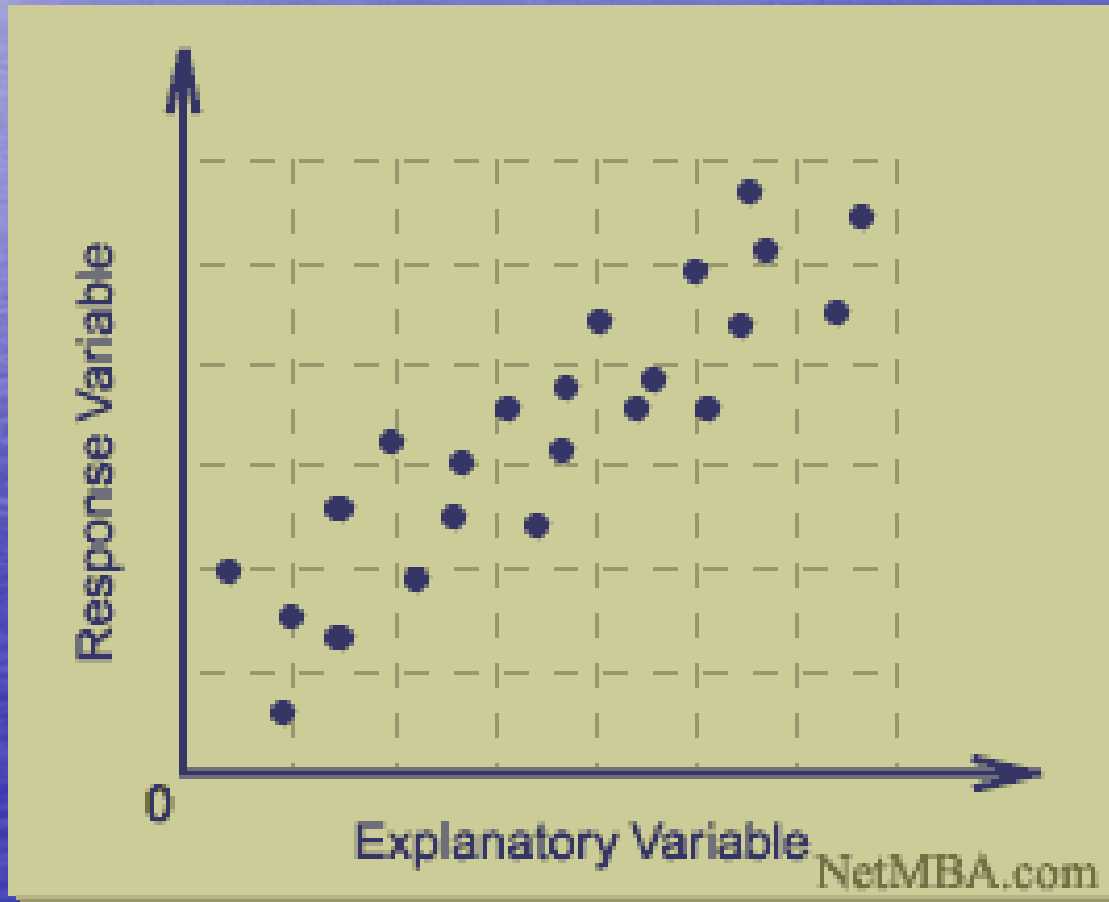
# Data Types

- Continuous
- A possible value exists between any other values.
- An infinite range of values exists between any other range of values.
- 50 kg, 50.1 kg or 50.1111 kg
- Ratio, Interval or Ordinal
- Discrete (discontinuous)
- A variable that can only take on specific values.
- Usually integers
- Ratio, Interval, Ordinal, or Nominal

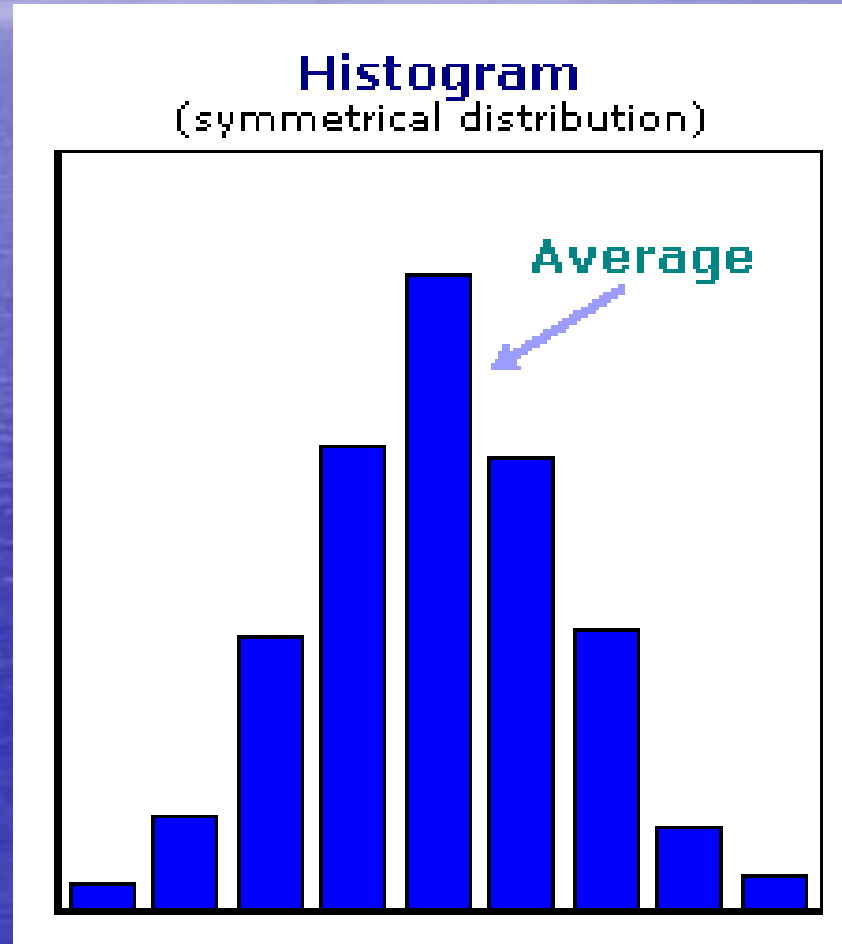
# Graphical Representation

- Scatter plot
- Box Plot
- Histogram
- Survival Curve

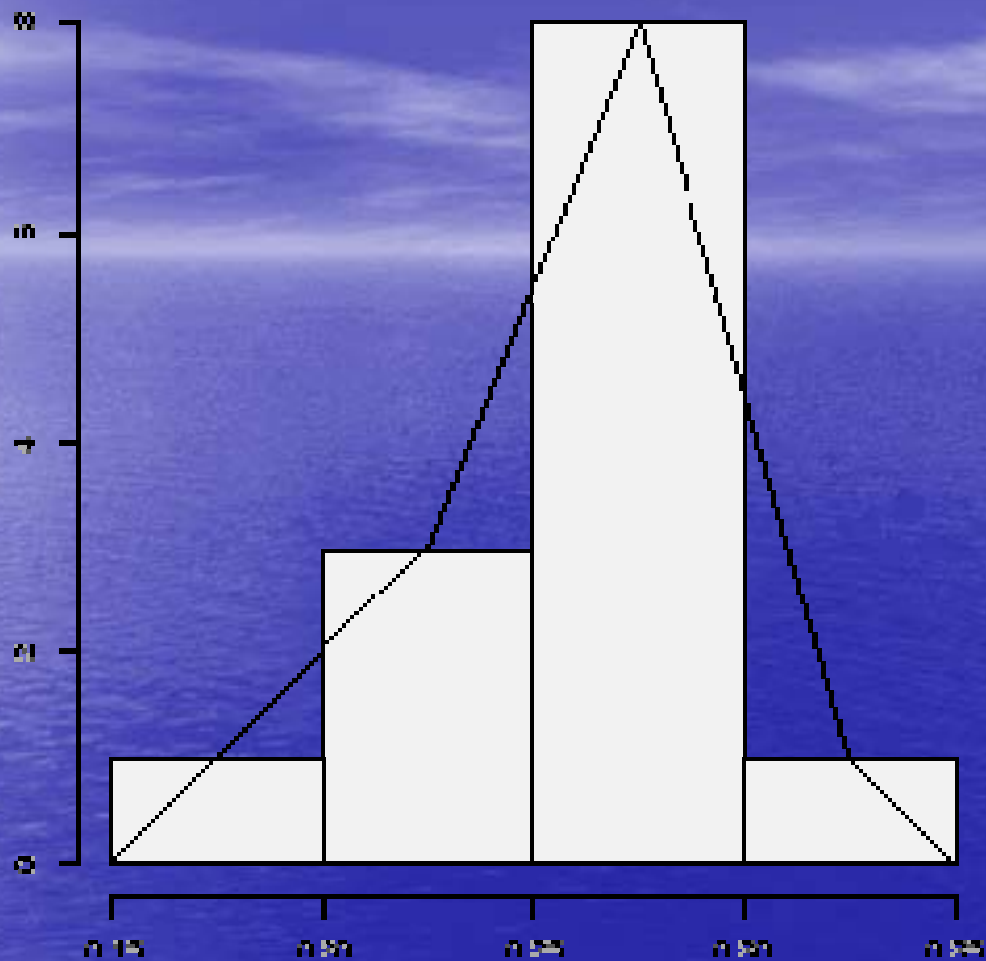
# Scatter Plot



# Histogram

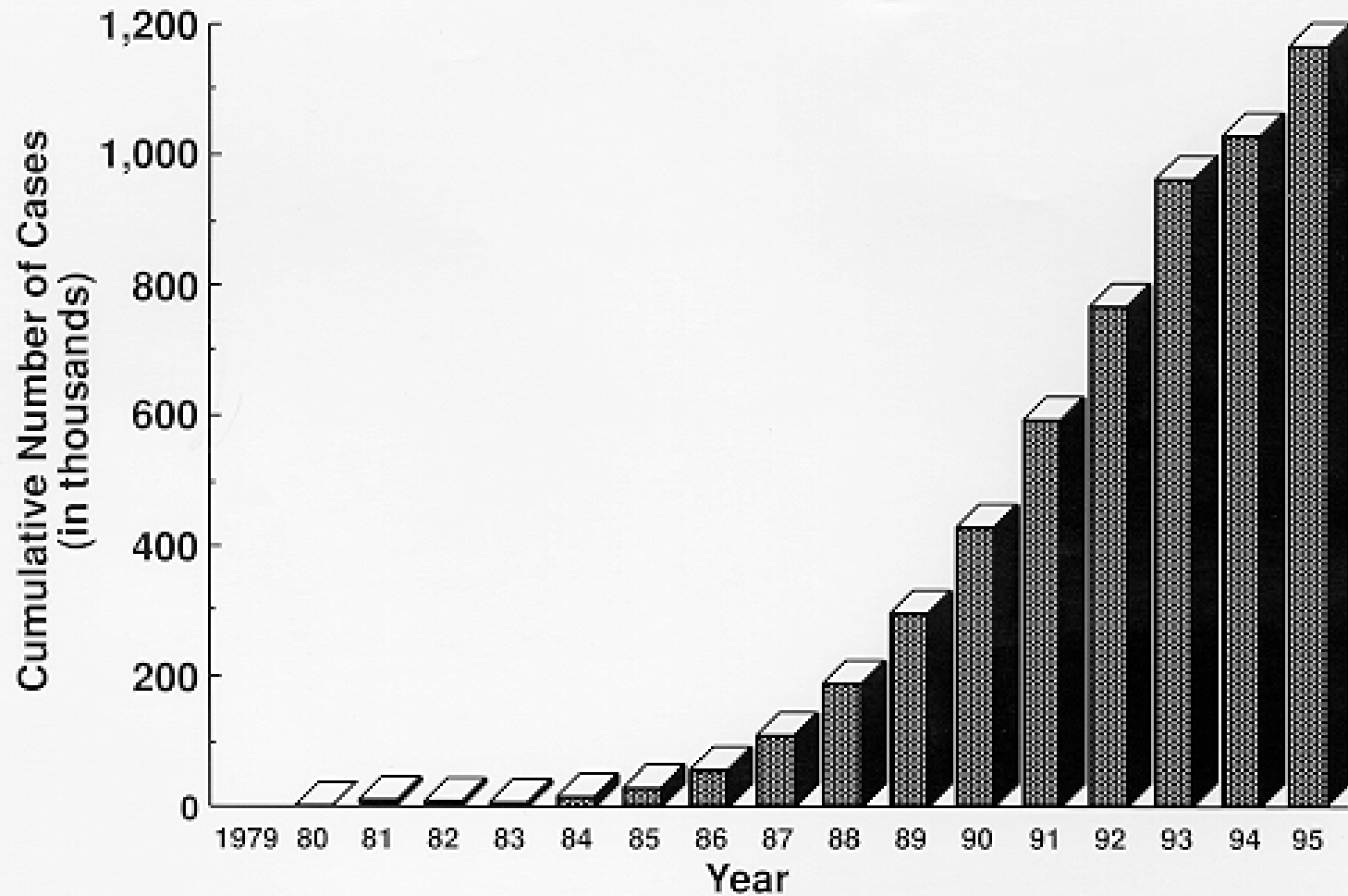


histogram with frequency polygon



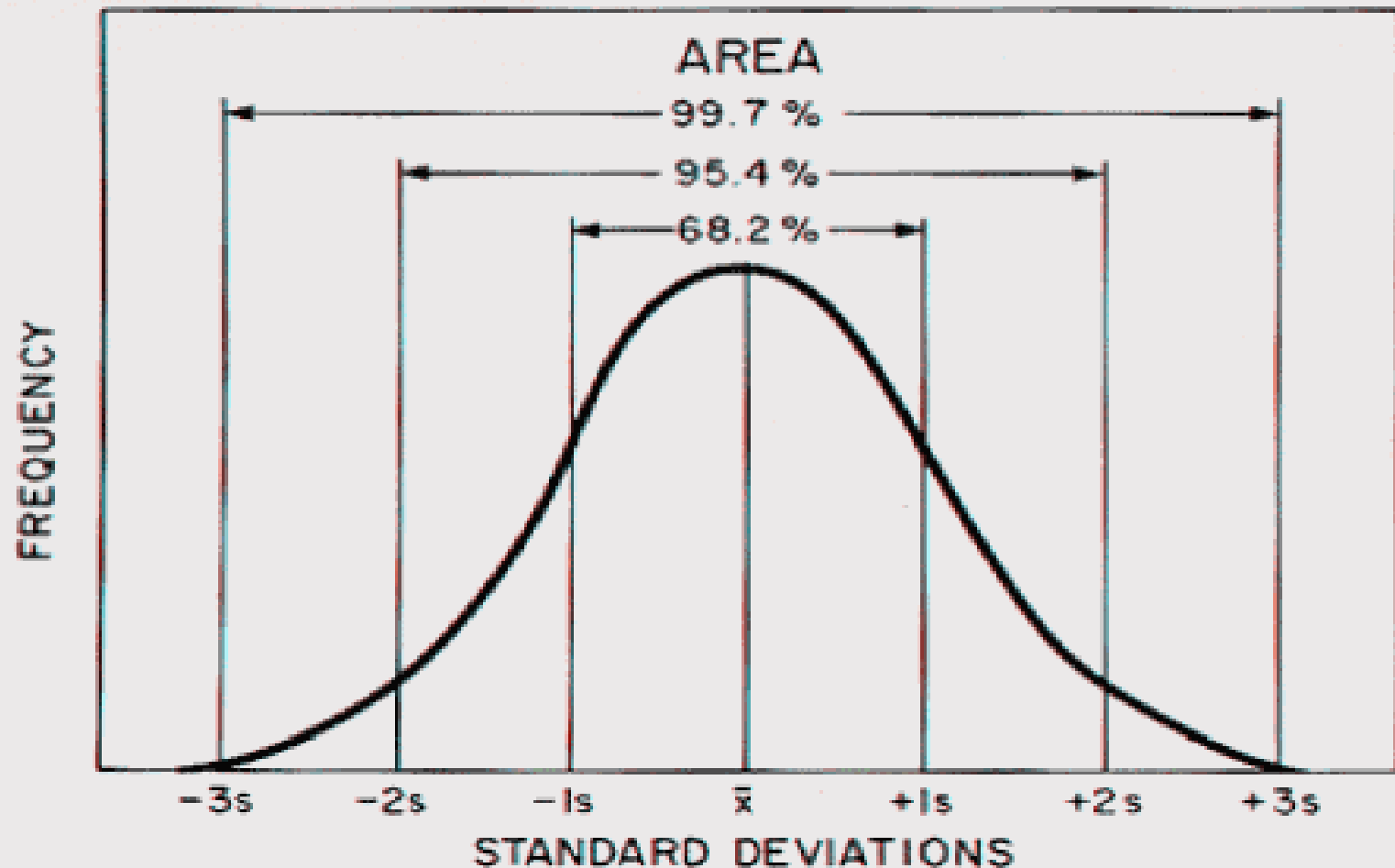
Frequency Polygon – CUNY-CSI

## Cumulative Frequency distribution

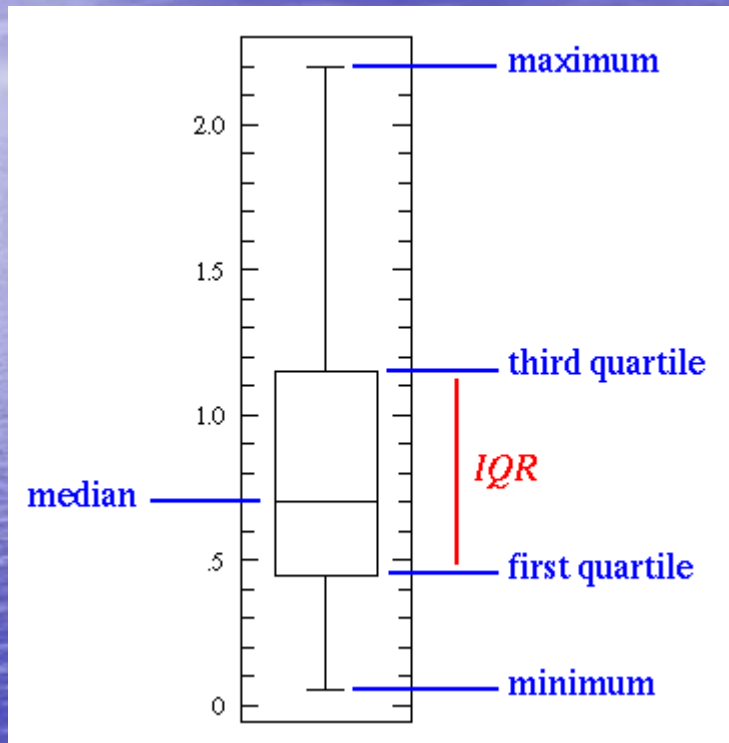


# Standard Normal Distribution

## Gaussian Curve

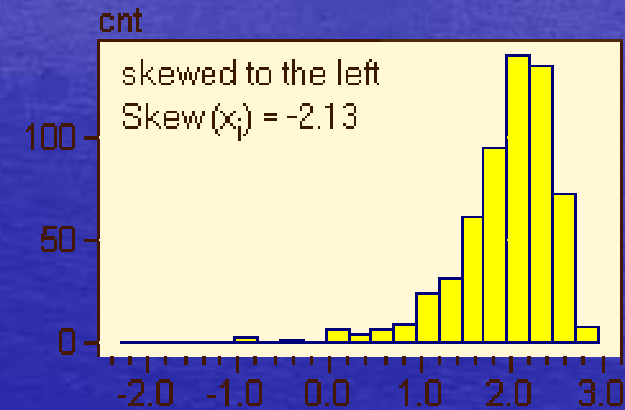
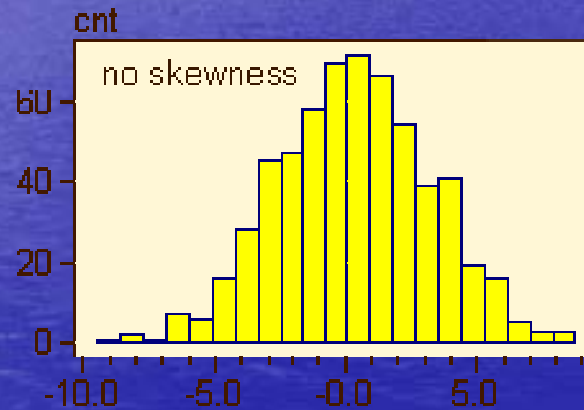
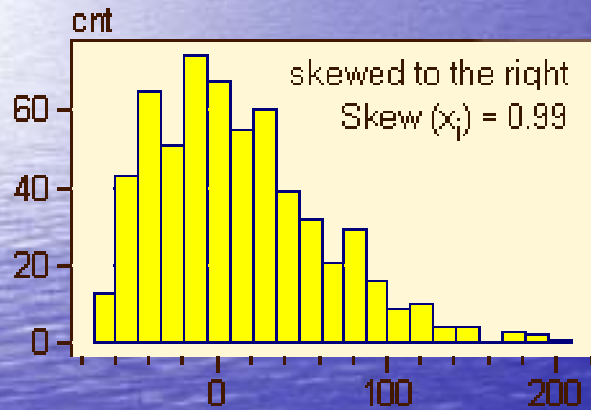


# 5-number summary Box Plot



maximum  
Q3  
Q2 or Median  
Q1  
minimum

# Skewed Distributions



# Population and Sample symbols

- Population Parameter
- TRUTH
- Greek Symbols
- $\mu$  mean
- $\sigma$  standard deviation
- $\sigma^2$  population variance
- $\sigma_{\bar{x}}$  population standard error of the mean
- N population size
- P population proportion
- B population parameter
- Sample Statistics
- ESTIMATE OF THE TRUTH
- Roman Symbols
- $\bar{x}$  mean
- s standard deviation
- $s^2$  sample variance
- $s_{\bar{x}}$  sample standard error of the mean
- n sample size
- p sample proportion
- b sample coefficient

# Assumptions

- Random Sampling – each member of the population has an equal and independent chance of being selected. The selection of one in no way influences the other.
- Unbiased estimate - long run average of statistic will equal population parameter.
- Precision, efficiency, or reliability - prefer the statistic from any single sample be close to the parameters actual value.
- Consistent Statistic – is one that as we take larger samples it will better estimate of the parameter.

# Descriptive

- Statistics that describe data in terms of their central tendencies and their dispersions.

# Inferential

- Statistics that enable the researcher to draw conclusions about data and the relationships between variables.

# Descriptive Statistics

# Measures of Central Tendency or the Middle of the Data

- Mean
- Median
- Mode

# Mean (Average)

$$\text{Mean} = \frac{\sum_{i=1}^n x_i}{n}$$

A measure of central tendency affected by extreme values in the data, also known as outliers.

# Median Observation or Datum

$$\textit{Median} = \frac{(n + 1)}{2}$$

ranked data point

“Middle Observation” or 50<sup>th</sup> percentile.  
No more than half the observations are above  
and half are below the observation.

# Mode

The most frequently occurring observation(s).

# Measures of Central Tendency

- Mean – Fulcrum
- Median – middle measurement that divides the data into two equal parts.
- Mode – unimodal, bimodal

# Normal Distribution

Mean = Median = Mode

# Measures of Dispersion or Spread in Data

- Range
- Interquartile Range
- Variance
- Standard Deviation

# Range

A difference between the highest and the lowest values. Not practical as it is influenced by a single extreme value. The greater the number of observations the larger it gets.  
Given by (Highest-Lowest)

# Intertquartile Range

Sometimes an interquartile range  
(25-75 percentiles)  
is used to limit the influence of extreme  
values.

Rank the orders from lowest to highest and use  $(n+1)/4$  ranked value as the  
Q1,  $(n+1)/2$  ranked value as the Q2,  $3(n+1)/4$  ranked value as the Q3.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

# Sum Of Squares

$$SS = \left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

# Mean Absolute Deviation

$$MAD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

# Variance

$$\text{Variance} = s^2 = \frac{\left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]}{n-1}$$

A measurement of dispersion about the mean. An average of squared deviations from the mean.

# Standard Deviation

$$\text{Standard Deviation} = s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

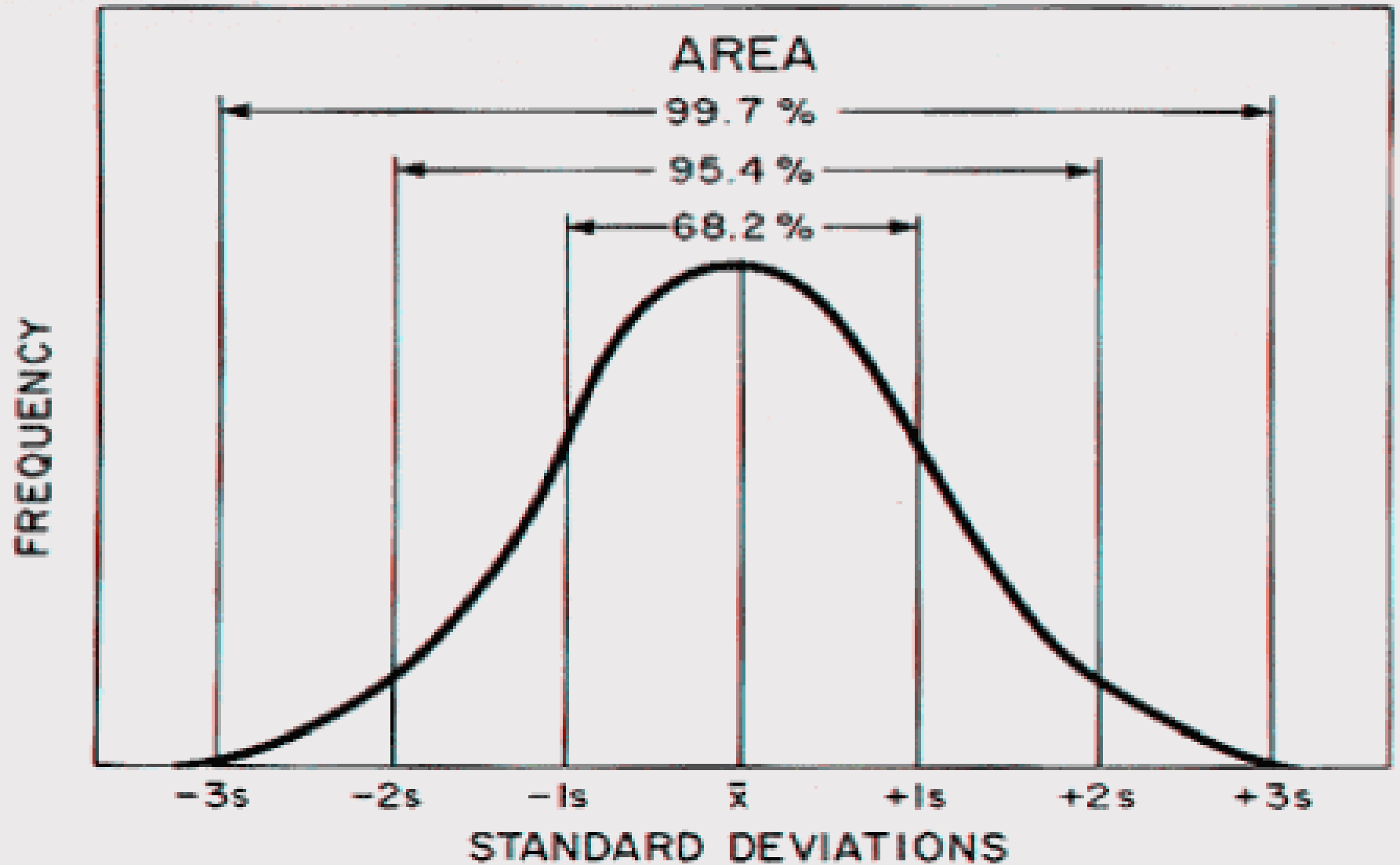
The average of the square root of the squared deviations from the mean.

1 standard deviation on either side of the mean will include about 68% of data values.

2 standard deviations, 95%.

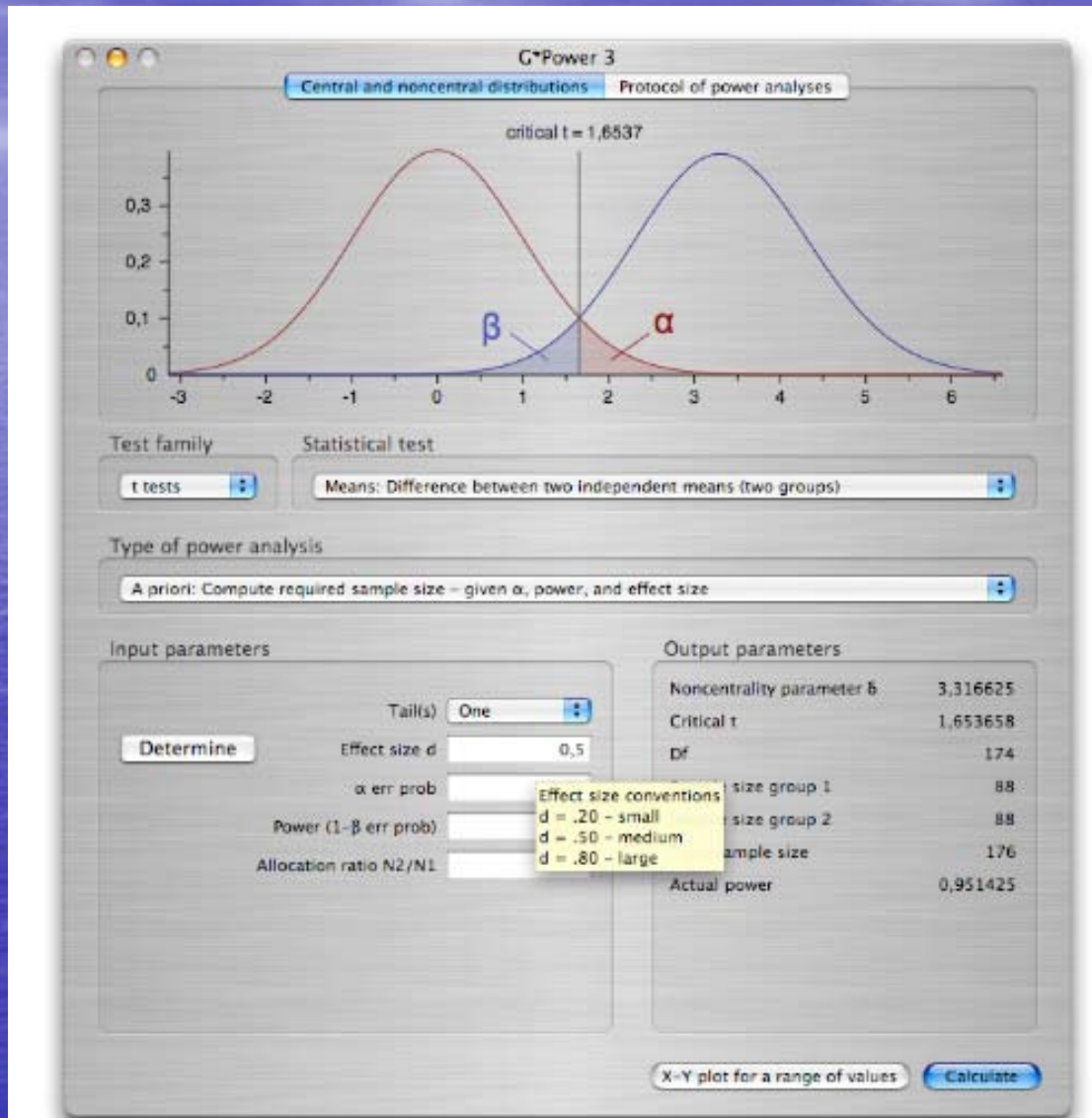
3 standard deviations, 99%.

# Gaussian Curve



$$Y_i = 1/\sigma^2\pi [e^{-(X_i - \mu)^2 / 2\sigma^2}]$$

# Effect Size



# Hypothesis Testing

- A mathematical formula that generates a test statistic from the data which is then compared to a table or manipulated by computer software to generate a p-value and/or a confidence interval.

# Toss a quarter

- $\frac{1}{2} = .5$
- $\frac{1}{2} * \frac{1}{2} = \frac{1}{4} = .25$
- $\frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{8} = .125$
- $\frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{16} = .06125$
- $\frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{32} = 0.030625$

# Alpha and False Positive Rate

The threshold below which you can reject random error or chance as the explanation for the observed group difference (effect size) or outcome.

Two possibilities  $P \geq \alpha$  or  $p < \alpha$

False positive rate =  $\alpha$

Also approximately  $n * \alpha$  where  $n$  in this case is the number of hypothesis tests performed.  $\{FPR = 1 - (1 - \alpha)^n\}$

# The P Value and alpha

- Alpha = The probability cut off below which we can reject the null hypothesis and conclude that our findings are statistically significant. ( $p < \alpha$ ).
- The probability that the observed group difference was due to random error or chance. ( $p \geq \alpha$ ).
- $p$  = The probability you would obtain an outcome (or group difference) as large as the one observed or more extreme than the one observed, if you repeated your experiment multiple times.
- A key objective behind all hypothesis testing is to generate a  $p$  value and/or confidence interval.
- See table of hypothesis tests and choosing the right test.

# The Confidence Interval

- If you repeated your experiment numerous times with random sampling and generated a confidence intervals for each experiment than the true population difference will be contained within 95 % of those confidence intervals generated. The true population difference will not be contained within 5% of the confidence intervals.
- The advantage of the confidence interval is it not only allows us the opportunity to accept or reject the null hypothesis and conclude our findings are statistically significant but also permits us to see how large the observed group difference might be. Whereas this is not available with the p value alone.

# Confidence Interval

Test statistic = (Effect Size)/(standard deviation)

Effect Size +/- (test statistic) \* (standard deviation)

# Type 1 and Type 2 Errors

- Type 1 Error
- Concluding that the observed group difference was a true effect when in fact it was due to chance or systematic error.
- Convicting the innocent man.
- False Positive rate
- Type 2 Error
- Concluding that the observed group difference was due to chance or random error, when in fact it was a true effect.
- Absolving the guilty man.
- False Negative rate

# POWER

- Probability of detecting an effect =  $\beta$
- Probability of type 2 error =  $1 - \beta$  = False Negative Rate

# Scale Used for the Dependent Variable

	Non-parametric	Non-parametric	Parametric
	Nominal Data	Ordinal Data	Interval Data
One Sample Test	Binomial (binomial equation or Z approximation)  Chi-Square Goodness of Fit Test	Kolmogorov-Smirnov One-Sample  Runs	Z test
Two Sample Test			
Related Samples	McNemar	Sign	Paired t
Unrelated Samples	Contingency Chi Square  Fishers Exact	Wilcoxon Two Sample Rank	Unpaired t
K Sample Test			
Related Samples	Cochran Q	Friedman Chi-square test	Randomized Block Analysis Of Variance
Unrelated Samples	Contingency Chi Square	Kruskal-Wallace	Analysis Of Variance (ANOVA) (followed by Tukey or SNK))

# Examples of Test Statistics

- Nonparametric Data
- U value – Wilcoxon Rank Sum
- Chi square – Chi Square test
- Parametric Data
- T value – Paired and Unpaired t-tests
- Z value – Z test

# Advanced Statistical Hypothesis Tests

- One way ANOVA
- Two way ANOVA
- Post Hoc Analysis
- ANCOVA
- MANOVA
- Simple and Multiple Linear Regression
- Logistic Regression
- Survival Analysis Cox Hazard Proportions Analysis
- Nonlinear Regression

# Statistical Software

- SPSS
- STATA
- SAS
- Systat
- GIS

# Anatomy of a Research Article

- Abstract
  - A brief summary of the Study and its findings.
- Background/Introduction
  - Historic overview information related to the research question and its relevance.
- Methods
  - A detailed discussion of the research design including limitations.
- Results
  - Descriptive statistics, graphical representation of data, and results of hypothesis testing and findings.
- Discussion/Conclusion
  - An interpretation of the findings.

# Interpreting the Medical Literature

- Methodical Approach
- Study Design i.e. bias, confounding, internal/external validity, and other forms of systematic error.
- Hypothesis testing and appropriate test selection.
- P value, critical value, alpha value, minimum effect size, confidence interval, OR, RR, power calculation.
- Type 1 or Type 2 error probability.
- Methods most important not limited to the abstract and conclusion.
- Problems with Summary information
- Problems with Consensus panels

# References

- **Epidemiology/Research Methods**

- Gehlbach SH. Interpreting the Medical Literature. Practical Epidemiology For Clinicians. 5th Ed. 2006.
- Gordis L. Epidemiology. 3rd Ed. 2004.
- Hulley SB, Designing Clinical Research. An Epidemiologic Approach. 3rd Ed. 2006.

- **Dataset/Database Management**

- Maran R. Microsoft Office 2000 Simplified. 1999.

- Maran R. Maran Illustrated Office 2003. 1st Ed. 2005.
- Maran R. Maran Illustrated Access 2003. 1st Ed. 2005.
- Maran R. Maran illustrated Excel 2003. 1st Ed. 2005.
- George D, Mallory P. SPSS For Windows: Step-by-Step. 7th Ed. 2006
- Hinton PR, Brownlow C, McMurray I. et. al. SPSS Explained. 1st Ed. 2004.
- Delwiche LD, Slaughter SJ. The Little SAS Book: A Primer. 3rd Ed. 2003
- Acock AC. A Gentle Introduction to Stata. 1st. Ed. 2005.

- **Scholarly Research Paper Publication/Bibliography Software**

- Agrawal A. EndNote 1-2-3 Easy! Reference Management For the Professional. 1st ed. 2005.
- Maran R. Microsoft Office 2000 Simplified. 1999.
- Maran R. Maran Illustrated Office 2003. 1st Ed. 2005.

# References

- **Supplementary and Advanced Level References**

- **Statistics/Biostatistics**

- Glantz SA, Slinker BK. Primer of Applied Regression and Analysis of Variance. 2nd Ed. 2000.
- Kleinbaum DG, Kupper LL, Nizam A, Muller KE. Applied Regression Analysis and Multivariable Methods. 4th Ed.. 2007.
- Winer BJ, Brown DR, Michels KM. Statistical Principles in Experimental Design, 3rd Ed.
- Snedecor GW, Cochran WG. Statistical Methods. 8th Ed. 1989.
- Maxwell SE, Delaney HD. Designing Experiments and Analyzing Data: A Model Comparison Approach. 2nd Ed.
- Keppel G, Wickens TD. Design And Analysis. A Researchers Handbook. 4th Ed. 2004.
- McMahan D. Linear Algebra Demystified. 1st Ed. 2005.
- Lay DC. Linear Algebra and its Applications. 3rd Ed. 2005.
- Clark-Carter D. Quantitative Psychological Research: A Students Handbook. 2004.
- Russo R. Statistics for the Behavioral Sciences. 2003.

- **Epidemiology/Research Methods**

- (1) Evans JS, Evans BT. How To Do Research. 2005.
- (2) Boynton PM. The Research Companion. A Practical Guide for the Social and Health Sciences. 2005.

- **Medical Health Informatics**

- Englehardt SP. Health Care Informatics: An Interdisciplinary Approach
- Medical Informatics: Knowledge Management and Data Mining in Biomedicine

# References

## – Useful WWW Online Resources

- Rice University Virtual Lab in Statistics (online multimedia tutorial and textbook) [www.onlinestatbook.com/rvls/](http://www.onlinestatbook.com/rvls/)
- UCLA Statistical Computing Online Tutorial on SAS, STATA, and SPSS. [www.ats.ucla.edu/stat/overview.htm](http://www.ats.ucla.edu/stat/overview.htm)
- Practice Datasets [www.vetmed.wsu.edu/appliedregression/](http://www.vetmed.wsu.edu/appliedregression/)

## – Video Instruction Resources

- Against All Odds: Inside Statistics [www.learner.org/resources/series65.html](http://www.learner.org/resources/series65.html)
- Statistics [www.videoaidedinstruction.com/](http://www.videoaidedinstruction.com/)