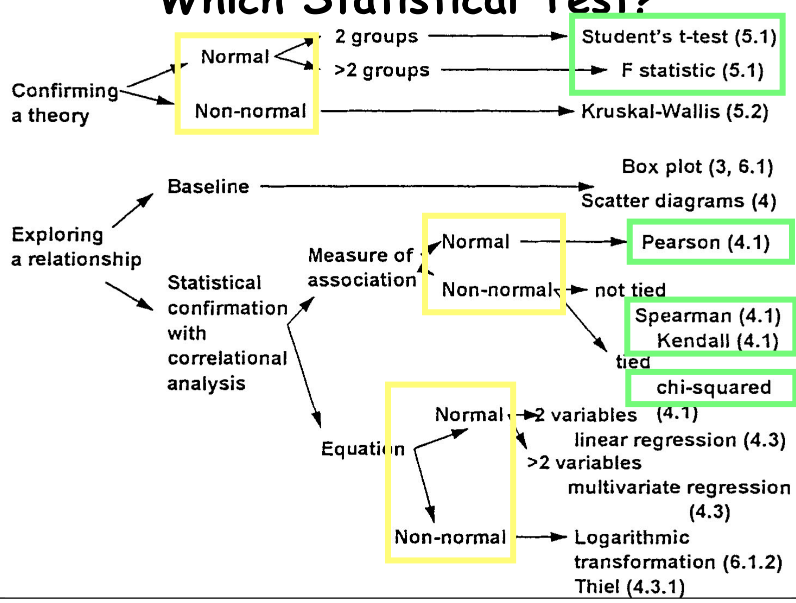
Reminder: Measurement scales

Type	Meaning	Admissible Operations
Nominal Scale	Unordered classification of objects	=
Ordinal Scale	Ranking of objects into ordered categories	=, <, >
Interval Scale	Differences between points on the scale are meaningful	=, <, >, difference, mean
Ratio Scale	Ratios between points on the scale are meaningful	=, <, >, difference, mean, ratio
Absolute Scale	No units necessary - scale cannot be transformed	=, <, >, difference, mean, ratio

Which Statistical Test?



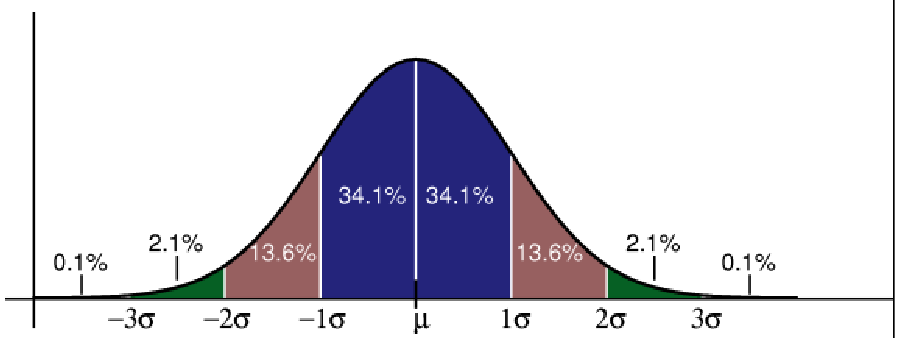
Basics

- Relationships between two variables:
  - Magnitude - how strong is the relationship?
  - Reliability - how well does the relationship in the sample represent the relationship in the population?
- Note:
  - Strong relationships can be detected more reliably
  - Larger sample sizes produce more reliable results

Hypothesis Testing

- Set up some hypotheses
  - Null hypothesis (H_0) - asserts that a relationship does not hold
    - In many cases, this is the same as saying there is no difference in the the means of two different treatment groups
  - Alternative hypotheses (H_1, …) - each asserts a specific relationship
  - Type I error: A false positive (rejecting H_0 when it's true)
  - Type II error: A false negative (accepting H_0 when it's false)
- For the statistical tests
  - P value (we calculate this) - probability that a relationship observed in the sample happened by chance
  - Alpha level (selected a priori) - a threshold for p at which we will accept that a relationship did not happen by chance (typically 0.1 or 0.05)
    - This allows us to fix the probability of a type I error in advance
    - If $p < \alpha$, we say the result was significant

Normal Distribution



Source: wikipedia - http://en.wikipedia.org/wiki/Image:Standard_deviation_diagram.png

Checking your data is normal

- → Draw a Histogram
- → Compute the mean and standard deviation
- → Superimpose the expected normal curve over the histogram

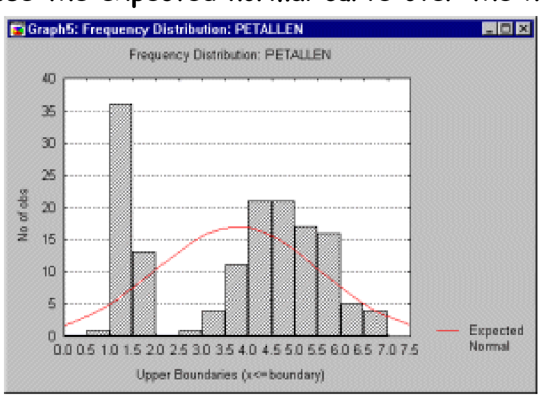
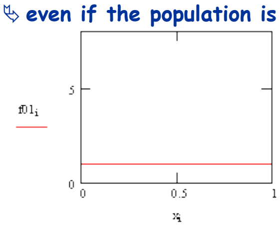


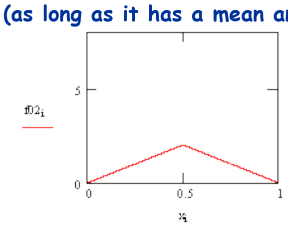
Image source:Statsoft (http://www.statsoft.com/textbook/stathome.html)

Central Limit Theorem

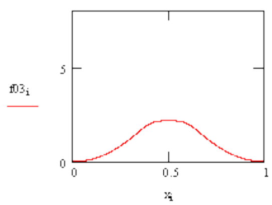
- Average of samples tend to normal distribution
  - …as sample size increases
  - even if the population is not normal (as long as it has a mean and SD)



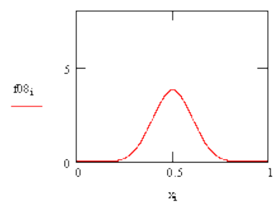
NonNormal Distribution of X



Distribution of Xbar when sample size is 2



Distribution of Xbar when sample size is 3



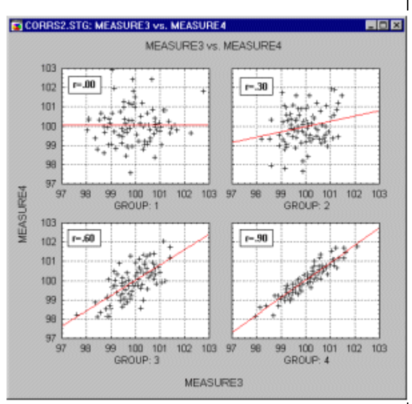
Distribution of Xbar when sample size is 8

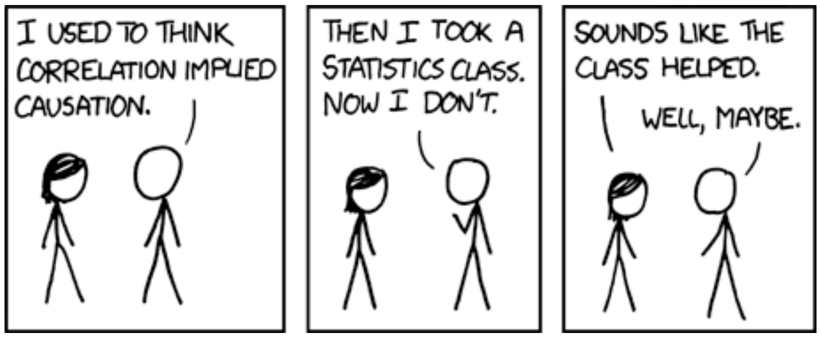
Correlations

- Measure of the relation between two variables:
  - -1 variables perfect inverses (negative correlation)
  - 0 no correlation at all
  - +1 variables are perfectly correlated (they appear on a straight line with positive slope)
- Pearson's r
  - Computed as:

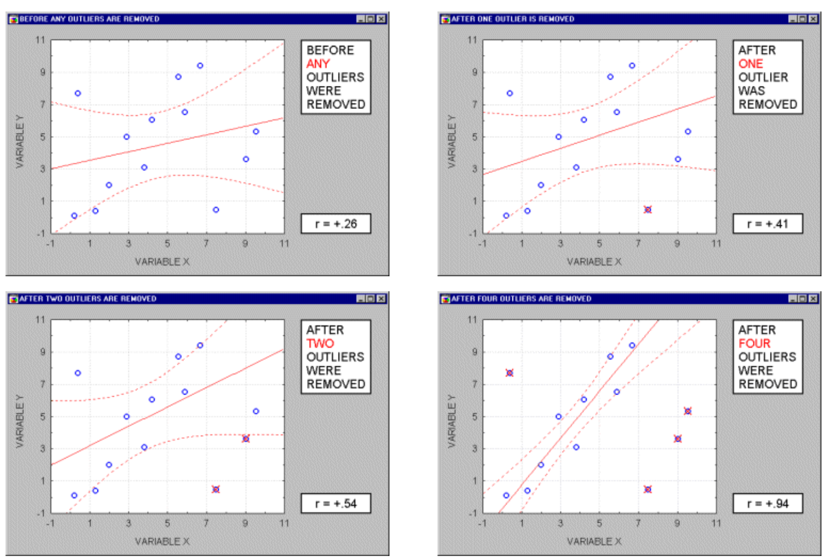
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

    - $\bar{x}$ and $\bar{y}$ are the sample means
    - s_x and s_y are the sample standard deviations
    - n is the sample size
  - Assumes variables are interval or ratio scale
  - Is independent of the measurement unit





Removal of outliers



Correlations for Ordinal Scales

- **Spearman's Rank Coefficient (ρ):**
  - Convert each variable into a ranked list
  - Compute:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

  - D = difference between the ranks for corresponding X and Y values
  - N = Number of pairs of X,Y values
  - Note: assumes no tied ranks
- **Kendall's Robust Rank Correlation (τ)**
  - n - number of items (X,Y pairs)
  - P - sum (over all items) of the items ranked after the given item by both rankings

$$\tau = \frac{2P}{\frac{1}{2}n(n-1)} - 1$$

  - Robust in the face of tied rankings

Student's t test

- **For testing whether two samples really are different**
  - given: two experimental treatments, one dependent variable
  - Assumes:
    - the variables are normally distributed in each treatment
    - the variances for the treatments are similar
    - the sample sizes for the treatments do not differ hugely
  - Basis: difference between the means of samples from two normal distributions is itself normally distributed.
  - The t-test checks whether the treatments are significantly different
- **Procedure:**
  - H_0: "There is no difference in the population means from which the samples are drawn"
  - Choose a significance level (e.g. 0.05)
  - Calculate t as $t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{(SE_A)^2 + (SE_B)^2}}$ where $SE = \frac{SD}{\sqrt{N}}$
  - Look up the value for t, with degrees of freedom $df = (n_A + n_B) - 2$
  - If calculated value of t is greater than the lookup value, reject H_0

Analysis of Variance (ANOVA)

- Generalization of t-test for >2 treatments
  - given: n experimental treatments, one dependent variable
  - Assumes:
    - the variables are normally distributed in each treatment
    - the variances for the treatments are similar
    - the sample sizes for the treatments do not differ hugely
    - (Okay to deviate slightly from these assumptions for larger samples sizes)
  - Works by analyzing how much of the total variance is due to differences within groups, and how much is due to differences across groups.
- Procedure:
  - H_0: "There is no difference in the population means across all treatments"
  - Compute the F-statistic:
    - *F*=(found variation of the group averages)/(expected variation of the group averages)
    - (don't do this by hand!)
  - If H_0 is true, we would expect F=1
  - Note: ANOVA tells you whether there is a significant difference, but does not tell you which treatment(s) are different.

Chi-squared test

- "ANOVA for non-interval data"
  - Given: data in an n x m frequency table (e.g. n treatments, m variables)
  - Assumes:
    - Non-parametric, hence no assumption of normality
    - Reasonable sample size (pref >50, although some say >20)
    - Reasonable numbers in each cell
  - Calculates whether the data fits a given distribution
  - Basis: computes the sum of the Observed-Expected values
- Procedure:
  - Calculate an expected value (mean) for each column
  - Calculate χ^2:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

    - Where O_i is an observed frequency
    - E_i is the expected frequency asserted by the null hypothesis
  - Compare with lookup value for a given significance level and d.f.