Session 5: Associations

Li (Sherlly) Xie

http://www.nemoursresearch.org/open/StatClass/February2013/

Session 5 Flow

1. Bivariate data visualization
   - Cross-Tab
   - Stacked bar plots
   - Box plot
   - Scatterplot
2. Correlation
   - Correlation coefficient (r)
   - Coefficient of determination (R-squared)
   - Simple, Multiple, Adjusted R-squared

Usual Bivariate Graphic Displays

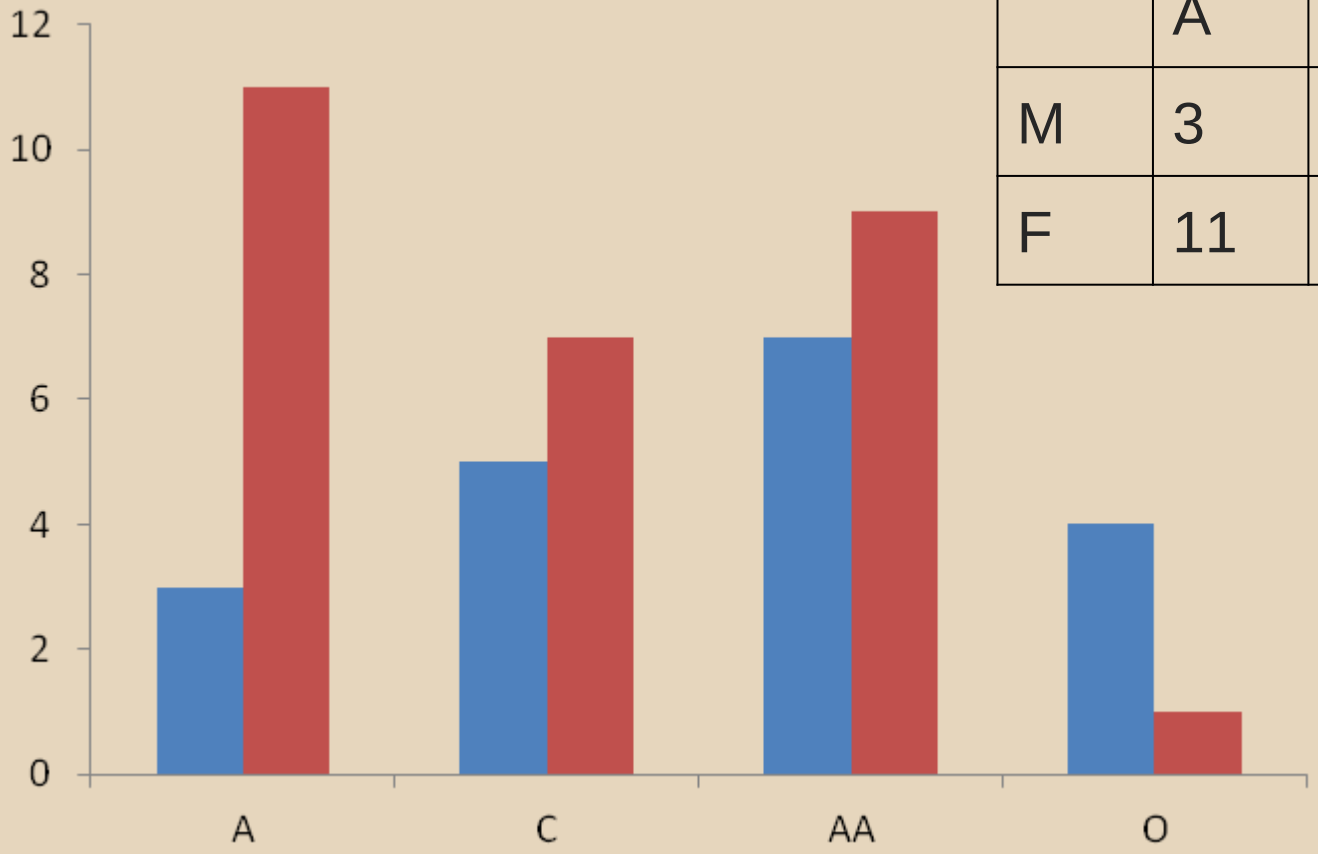
Variable Pairs	**Display**
Categorical, Categorical	Crosstabs; stacked/clustered/stacked and clustered bar graph
Categorical, Quantitative	Box plot
Quantitative, Quantitative	Scatterplot

Stacked Clustered Bar Graph

Showing the relationship between 2 categorical variables (e.g. gender vs race, FHDM vs gender, FHDM vs race, etc)

Display the SAME information that could be displayed using a cross-tab

Race vs Gender, Clustered (by race) Bar Plot



	A	C	AA	O
M	3	5	7	4
F	11	7	9	1

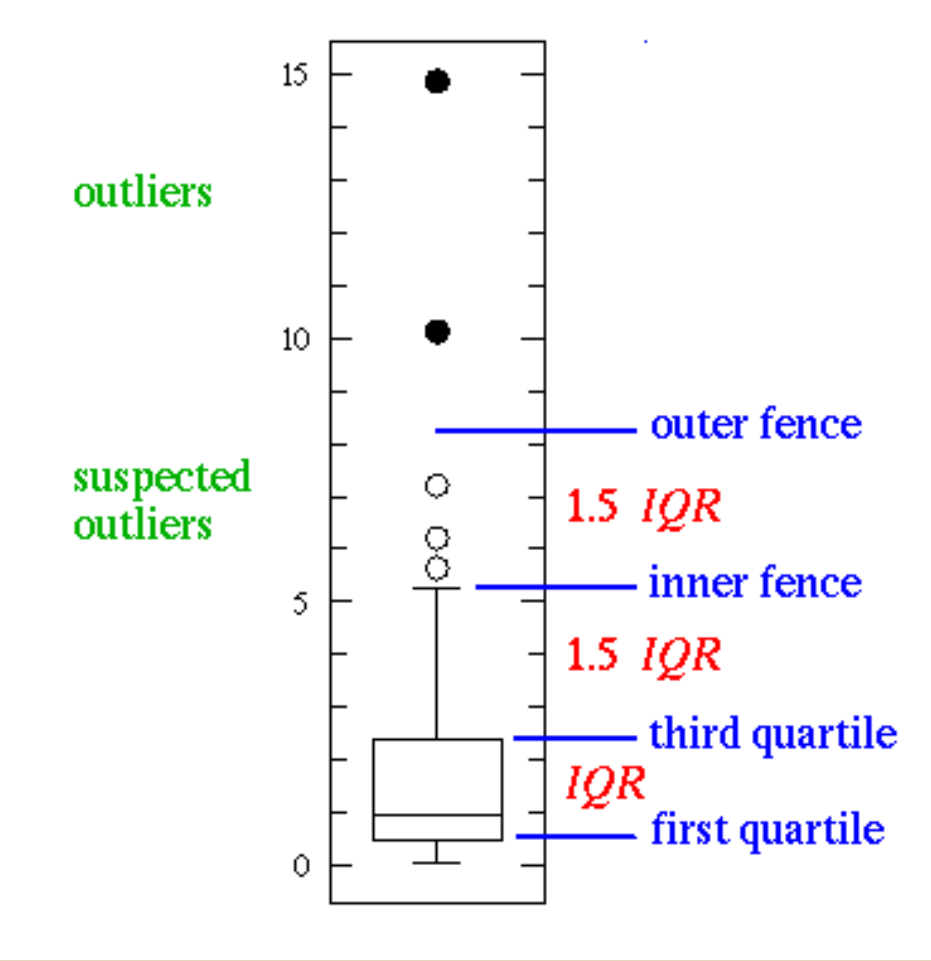
Box Plots

Excel macro download link:
http://www.vertex42.com/Files/download/excel.php?file=box-plot.xls

Shows distribution of continuous variable by categories—

Years of friendship by personality type;

times being your job reference by whether you remember his/her phone



http://www.physics.csbsju.edu/stats/box2.htm

Median, IQR, Fences

Example: 21 data points for the age variable:

1,1,1,2,3,4,8,5,3,2,8,7,5,8,3,7,5,4,7,11,18

Find median: sort from smallest to largest:

1,1,1,2,2,3,3,3,4,4,5,5,5,7,7,7,8,8,8,11,18

median is the number in middle: 5

find first quartile (25th percentile): 3

find third quartile (75th percentile): 7

IQR=7-3=4

Inner fence: 1.5 times IQR from first or third quartile

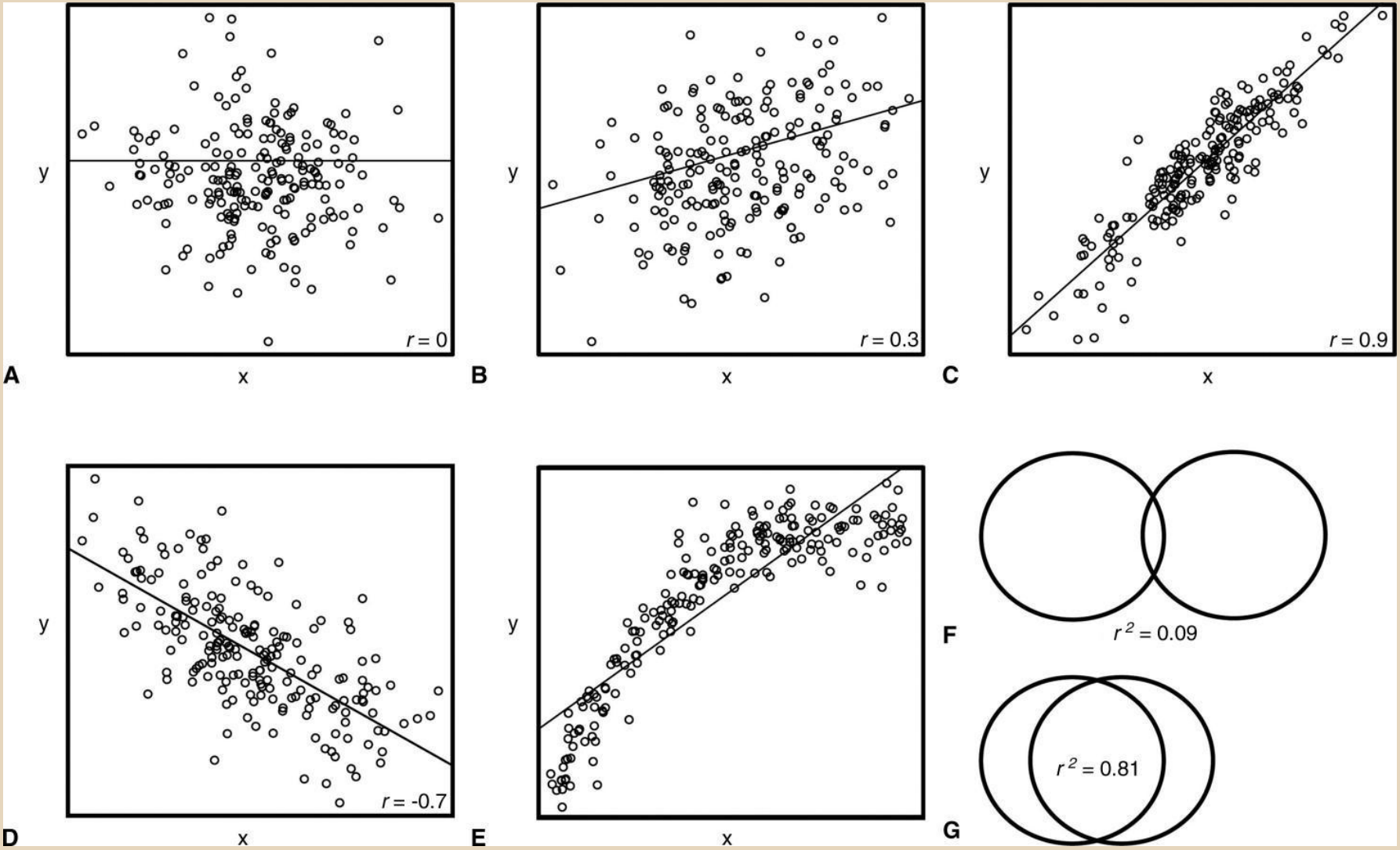
Outer fence: 1.5 times IQR from inner fences (or, equivalently, 3 times IQR from first or third quartile)

Scatterplot

A scatterplot reveals relationships or association between two variables.

ASSUMPTION OF SCATTERPLOT: all data points weigh the same

Type & Strength of Associations; r is NOT r-squared



r, R, r-squared, R-squared, adjusted R-squared

r (Correlation Coefficient, Pearson's r): Measures the strength of **LINEAR association** between two CONTINUOUS variables.

R (Coefficient of determination): A measure of the proportion of variance in one variable accounted for by the variance(s) in one (or more) explanatory variable(s)

r-squared, R-squared, adjusted R-squared

r-squared (Coefficient of Simple Determination): The percent of the variance in the dependent variable that can be explained by one independent variable.

R-squared (Coefficient of Multiple Determination): The percent of the variance in the dependent variable that is explained by all of the independent variables taken together.

R-Squared Adjusted (Adjusted R-Squared): A version of R-Squared that has been adjusted for the number of predictors in the model. R-Squared tends to overestimate the strength of the association especially if the model has more than one independent variable. *Adj R-squared always less than or equal to BUT NEVER EXCEEDS R-squared.*

Rules of Thumb

1. r-squared, R-squared and adjusted R-squared) range from 0 to 1 (percentage of variation in one variable explained by another variable or other variables
2. Pearson's r ranges from -1 to 1 (magnitude and direction of correlation)

   0 to < 0.3: weak correlation

   0.3 to <0.5: moderate correlation

   0.5 to 1: strong correlation
3. Interpretation of r-squared's and Pearson's r is DUAL:

   (Simple/unadjusted/adjusted) amount of variation in Y explained by THE GROUP of factors ABC IS THE SAME as (simple/unadjusted/adjusted) amount of TOTAL variation in ABC explained by Y.

   X is as strongly negatively correlated with Y as Y is with X.

** In some studies (e.g.: field, observational), a "practically good/strong" r may be as low as 0.3*

R-squared Applied

If r-squared between # years of friendship and # times the person being your job reference is 0.5, it means

1. 50% of the VARIATION in # years of friendship IS explained by # times the person being your job reference;
2. 50% of the VARIATION in # years of friendship IS NOT explained by # times the person being your job reference;
3. 50% of the VARIATION in # times the person being your job reference IS explained by # years of friendship;
4. 50% of the VARIATION in # times the person being your job reference IS NOT explained by # years of friendship

r Applied and Conventional Cutoffs

If r = 0.1 between length of friendship and # times the person being your job reference, this POSITIVE and WEAK LINEAR correlation coefficient means as length of friendship increases # times the person being your job reference also increases, but the LINEAR trend is not that strong.

If r = 0.5, it means they are MODERATELY POSITIVELY linearly correlated.

If r = -0.9, it means they are HIGHLY NEGATIVELY LINEARLY CORRELATED--as length of friendship increases # times the person being your job contact decreases.