Review of Statistics

And
Experimental Design

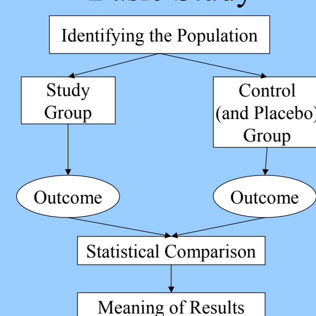
Basics of Experimental Design

- Scientists study relation between variables
- In the context of experiments these variables are called independent and dependent variables
- The purpose of an experiment is to establish a cause and effect relationship between the independent and dependent variables.
- In the process we have to be concerned with counfounding (intervening) variables.

Experimental Design

- Internal validity
  - Isolation of cause and effect
  - Randomization
  - Control (and placebo) group
- External validity
  - Ability to generalize results
  - Random sample
  - Theoretical perspective

Basic Study



When an Experiment is Not Possible

- We study relations among variables
- However, when studying relations among variables we have to be careful of making causal inferences.
- In the social sciences, we often use regression and correlation to explore relations between variables.

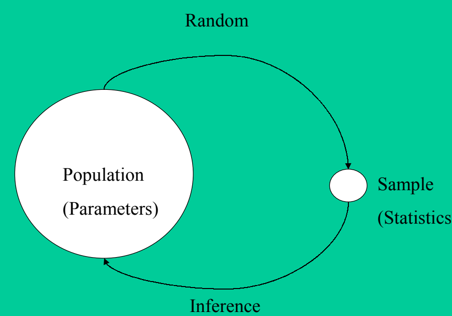
Studying Relations

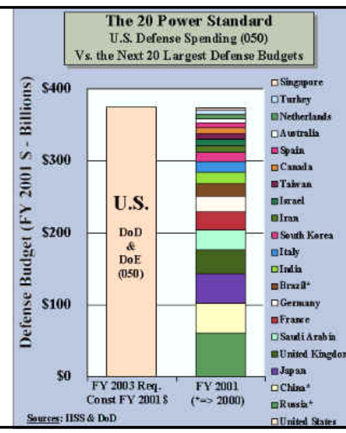
- Theoretical framework
- Simon’s self- containment
- Structural Equation Modeling
- Shoes and reading ability

Issues to Consider in Designing a Study

- The target population
  - Inclusion criteria
  - Exclusion criteria
- Sample size– is there an adequate number of individuals to allow a reasonable chance of demonstrating statistically a difference between the treatment and control groups.
- Study hypotheses– you must have specific questions in mind.

Inferential Statistics





Purposes of Statistics

- Summarizing and describing data
  - Frequency distributions
  - Central tendency and variability
  - Graphs
- Inferences
  - Estimation
  - Hypothesis testing

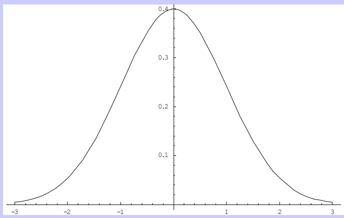
Comparing Distributions

- Shape
  - Skewness (mean-median)/ Std. Dev
  - Kurtosis (peakedness)
    - Leptokurtic (too peaked to be normal)- positive value
    - Platykurtic (too flat to be normal)- negative value

Comparing Distributions

- Central Tendency
  - Mean (arithmetic average)
  - Median (middle score)
  - Mode (most frequent score)
- Variability
  - Variance (Standard deviation)
  - Range
  - Interquartile Range

Normal Distribution



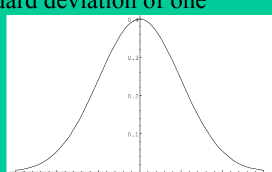
Normal Distribution

- Unimodal
- Symmetric
- From negative infinity to positive infinity
- Density function:

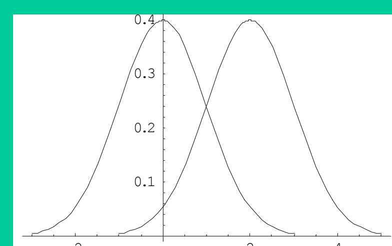
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

Standard Normal

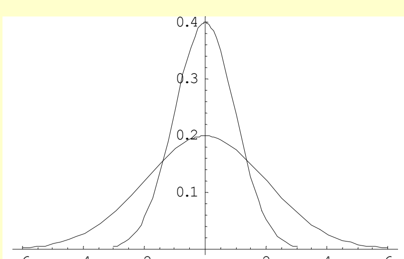
- Mean of zero
- Standard deviation of one



Normal Distributions with Different Means

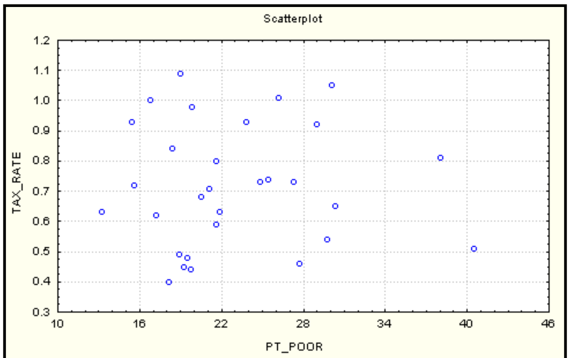


Normal Distributions with Different Variance



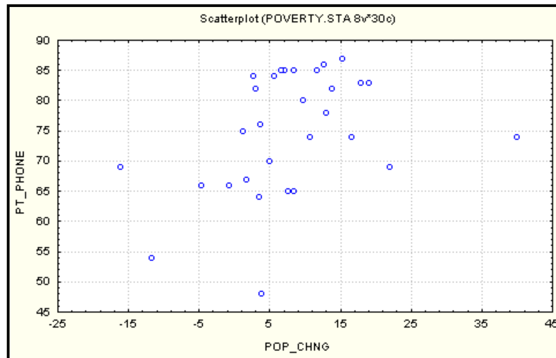
Regression and Correlation

Finding the line that fits the data

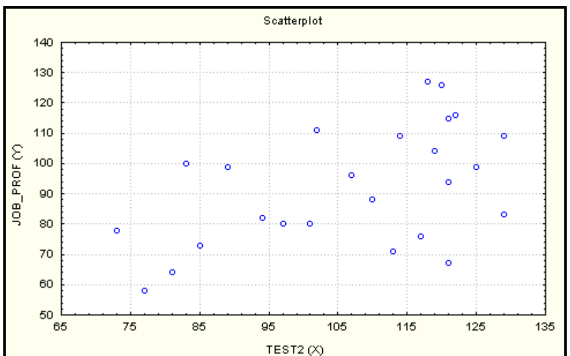


Working with paired data

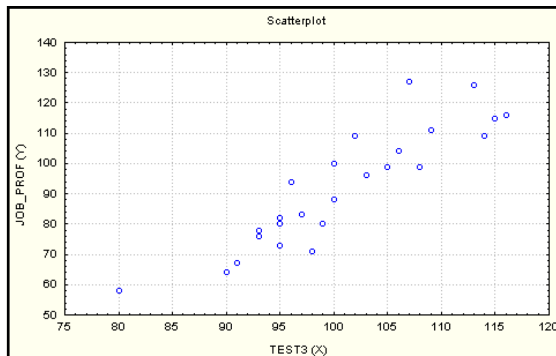
1



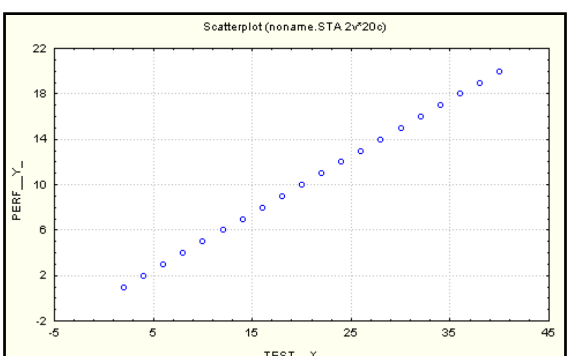
2



3



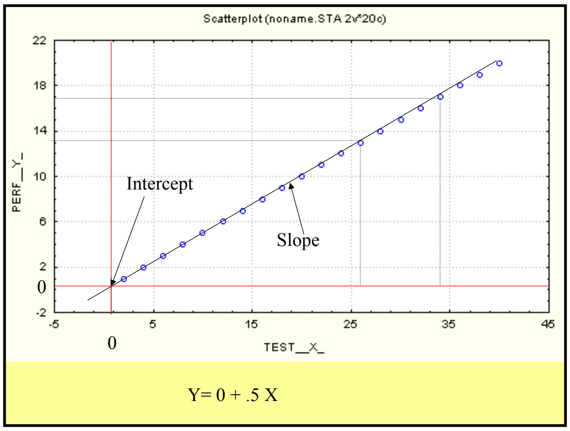
4

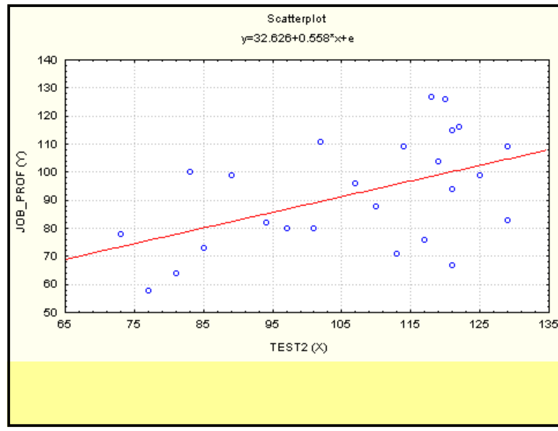


5

Correlations

- Slide 1, r=.01
- Slide 2, r=.38
- Slide 3, r=.50
- Slide 4, r=.90
- Slide 5, r=1.0





Regression

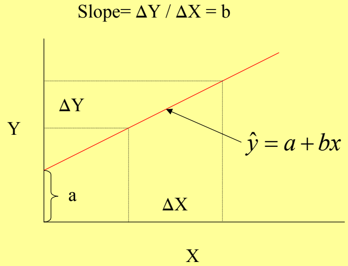
- The regression line is the line that best fits the data. The idea is to capture the relationship between the X and Y variables.
- The line is identified by its intercept and slope.
- The intercept is called "a"
- The slope is called "b"
- So, the line is: line= a + b X

Correlation

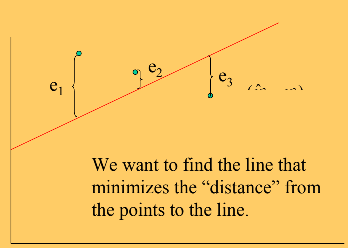
- Once we have identified the best line, then we need to assess how well the line fits the data.
- The correlation tells us "how well the line fits the data."
- A correlation of one is a perfect fit; whereas, a correlation of zero is the worse fit.

The Line

Slope= ΔY / ΔX = b



Finding the Regression Line



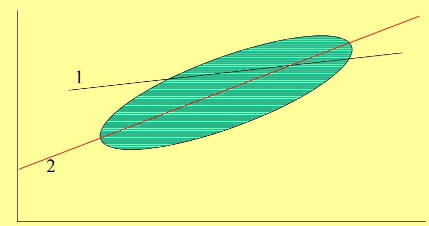
Least Squares Criterion

Find "a" and "b" such that the sum of squares error is the smallest it can be.

$$\min = \sum_{i=1}^{n} e_i^2$$

The line that minimizes the sum of squares error is the best line.

Line Fit



The Best Line

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \bar{y} - b\bar{x}$$

X	Y	XY	X^2	Y^2
4	6	24	16	36
6	12	72	36	144
8	14	112	64	196
11	10	110	121	100
12	17	204	144	289
14	16	224	196	256
16	13	208	256	169
17	16	272	289	256
20	19	380	400	361
Σx=108	Σy=123	Σxy=1606	Σx^2=1522	Σy^2=1807

Finding the Regression Line

Σx=108	Σy=123	Σxy=1606	Σx^2=1522	Σy^2=1807

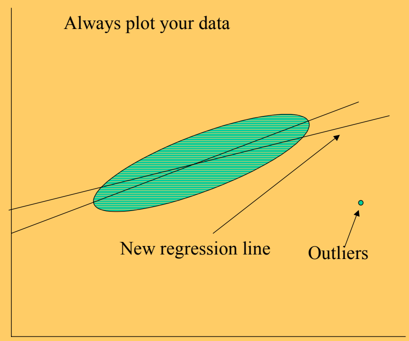
$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{9(1606) - (108)(123)}{9(1522) - (108)^2} = .575$$

$$a = \bar{y} - b\bar{x} = \frac{123}{9} - .575\frac{108}{9} = 6.767$$

$$\hat{y} = 6.767 + .575x$$

Using the Regression Line

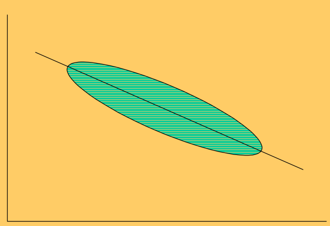
- When using a regression equation for prediction stay within the range of the available data.
- Don't make predictions about a population that is different from the population from which the sample were drawn.
- A regression equation based on old data may be no longer valid.



Correlation

- Tells you how well the line fits the data.
- The correlation ranges from –1 to 1.
- A negative correlation has a negative regression line (slope).
- A correlation of 1 (or –1) indicates a perfect fit between the line and the data.
- A correlation of zero indicates a very poor fit.

A Negative Correlation



Computing the Correlation

Σx=108	Σy=123	Σxy=1606	Σx^2=1522	Σy^2=1807

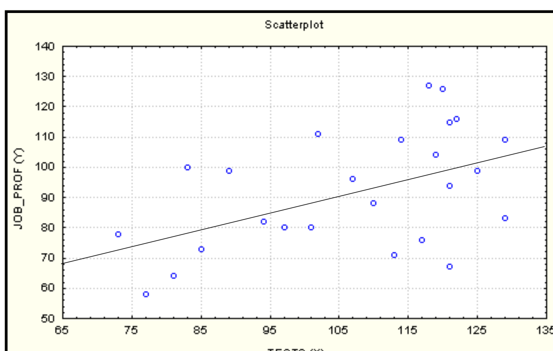
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{9(1606) - (108)(123)}{\sqrt{[9(1522) - (108)^2][9(1807) - (123)^2]}} = .77$$

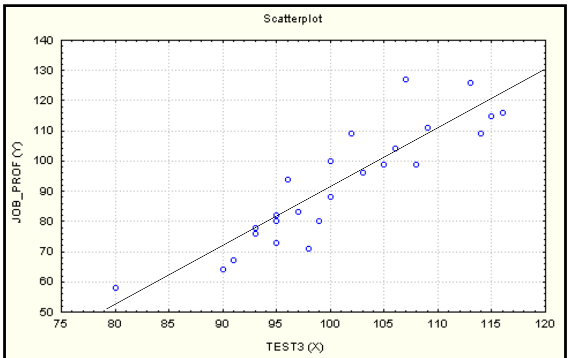
Correlation and Regression

- The regression line is the line that best fits the data:
- The correlation tells us how well the regression line fits the data, r.
- The relationship between the correlation and the slope of the regression line is given by

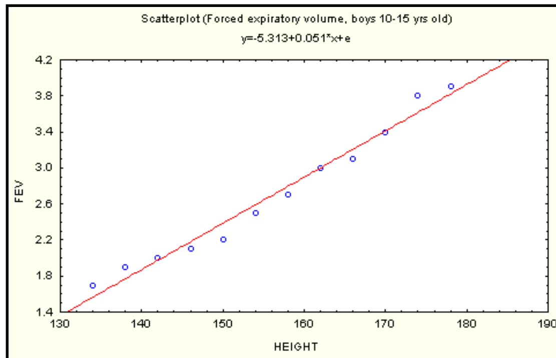
$$r = b\frac{S_x}{S_y}$$



r=.50 $\hat{y} = 32.6 + .56x$



r=.90 $\hat{y} = 57.55 + 1.58x$

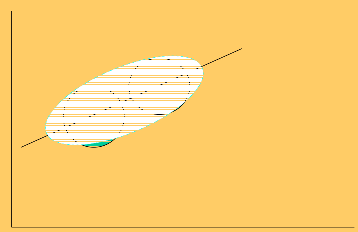


American Review of Respiratory Disease,
r=.98, n=12.

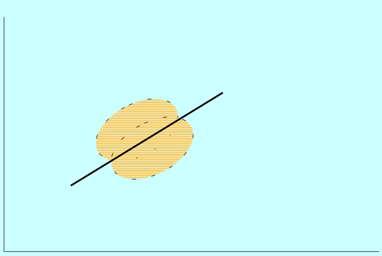
Factors Affecting the Correlation

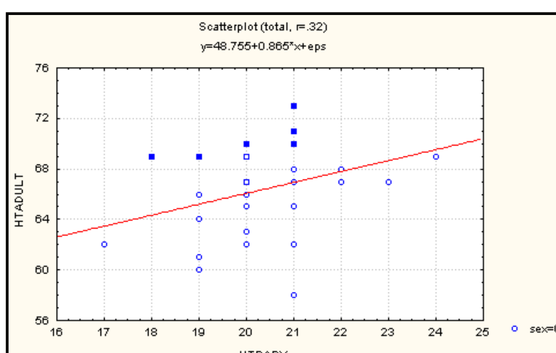
- Correlation is not causation
- Combined Groups
- Outliers
- Restriction in range
- By the way the correlation is invariant under linear transformation

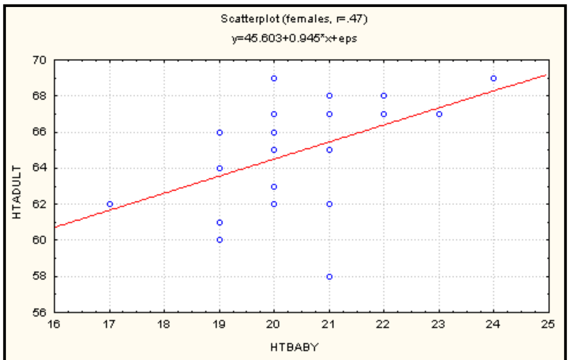
Combined Groups



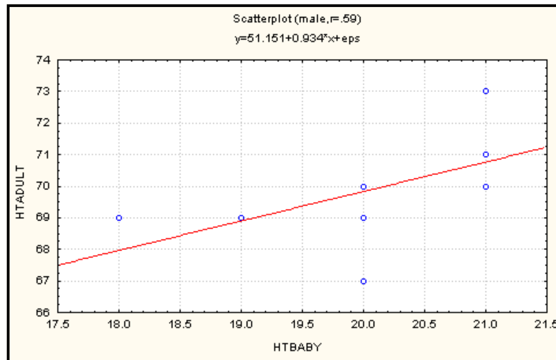
Combined Groups





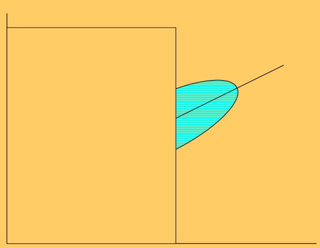


r=.472, n=27.



n=11

Range Restriction Generally Reduces the Correlation



Testing Hypothesis about the Population Correlation

- Two procedures
  - When the Null involves zero
    - Based on the t test
  - When the Null involves a value other than zero
    - A z test on the transformed correlation

Testing a hypothesis about the population correlation

$$Ho: \rho = 0$$

$$Ha: \rho \neq 0$$

The test is based on the t-test. However, if we use Table A-6 (p. 774) the test is very easy to carry out.

By the way, note that if ρ=0, then β=0.

An Example

- Suppose that we are interested in testing the claim that there is a linear relationship (correlation) between height at birth and adult height for females. If we can consider our previous sample to be a random sample from the population of American women, we can conduct the test using the data. Recall that r=.472, and n=27. Set alpha at .05

Solution

$$Ho: \rho = 0$$
$$Ha: \rho \neq 0$$

To test the claim we look at Table A-6. We need to know the sample size and to find the critical value. Here n=27. For a two-tail test (with n=25) the critical value is ±.396. Because r(=.472) is larger than .396, we reject the Null. The data support the claim that there is a relationship between height at birth and adult height.

Testing the Hypothesis that ρ is other than zero.

- If we want to test the hypothesis that the population correlation is other than zero, we must use Fisher's r to z transformation,

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

We can obtain the z transformation using a calculator or a table.

Example

- Suppose that we are interested in testing the Null hypothesis that $\rho \leq .3$.
- Against the alternative that $\rho > .3$
- Let's consider our class data again: r=.472, n=27. Again, set alpha at the .05 level.
- Note that this is a one-tail test.

Solution

$$Ho: \rho \leq .3$$
$$Ha: \rho > .3$$
$$z_r = \frac{1}{2}\ln\left(\frac{1+.472}{1-.472}\right) = .5126$$
$$z_\rho = \frac{1}{2}\ln\left(\frac{1+.3}{1-.3}\right) = .3095$$

Next, we use these z scores to construct a z-test.

The Z test

$$z = \frac{z_r - z_\rho}{\frac{1}{\sqrt{n-3}}} = \frac{.5126 - .3095}{\frac{1}{\sqrt{27-3}}} = \frac{.2031}{.2041} = .99$$

The critical z value for a one-tail test at the .05 level is 1.645. So, we can't reject the Null.