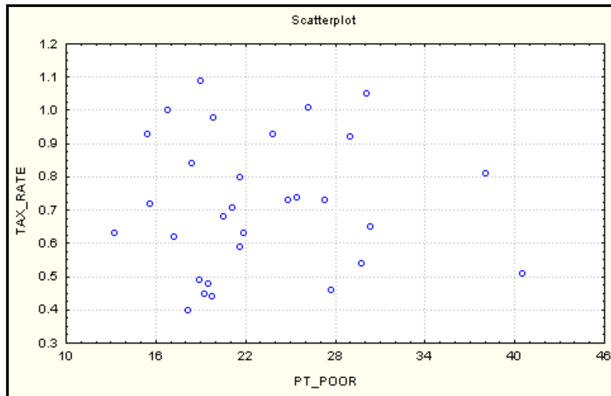
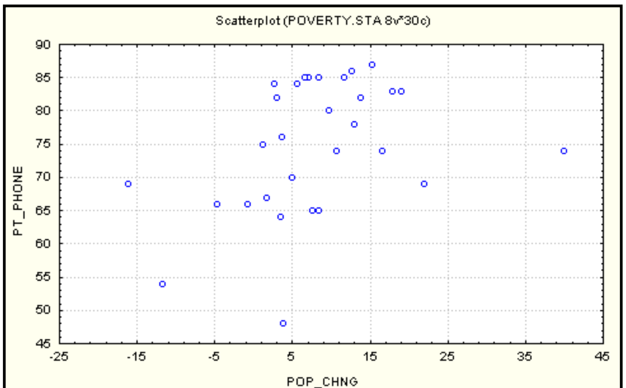
Regression and Correlation

Finding the line that fits the data

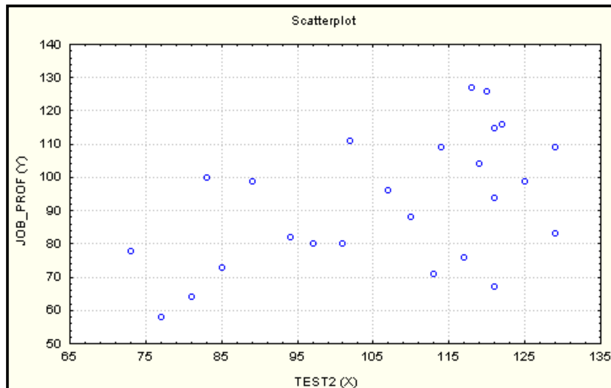


Working with paired data

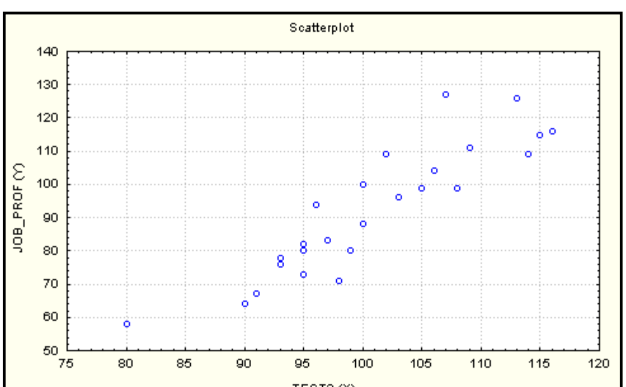
1



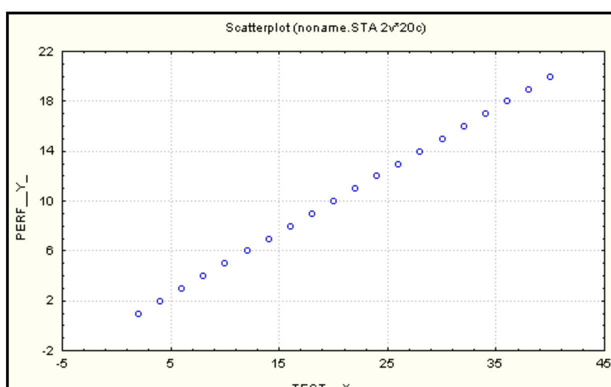
2



3



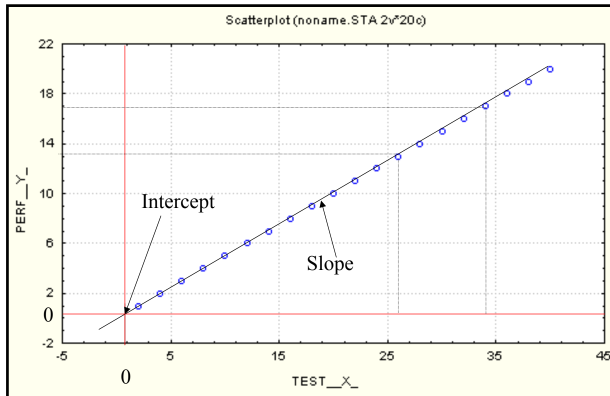
4



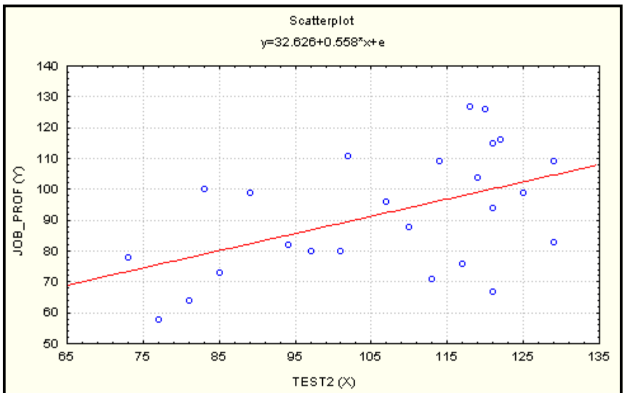
5

Correlations

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



Y= 0 + .5 X



Regression

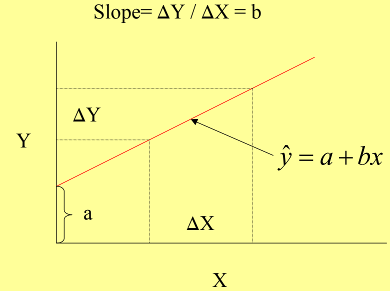
- The regression line is the line that best fits the data. The idea is to capture the relationship between the X and Y variables.
- The line is identified by its intercept and slope.
- The intercept is called “a”
- The slope is called “b”
- So, the line is: line= a + b X

Correlation

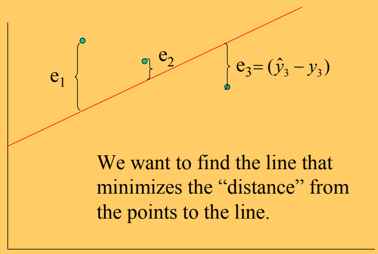
- Once we have identified the best line, then we need to assess how well the line fits the data.
- The correlation tells us “how well the line fits the data.”
- A correlation of one is a perfect fit; whereas, a correlation of zero is the worse fit.

The Line

Slope= $\Delta Y / \Delta X = b$



Finding the Regression Line



We want to find the line that minimizes the "distance" from the points to the line.

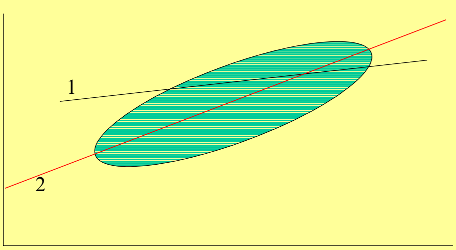
Least Squares Criterion

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

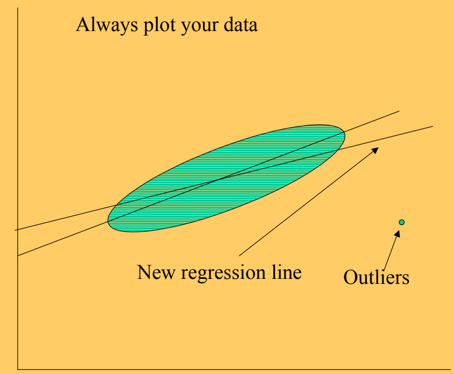
$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{9(1606)-(108)(123)}{9(1522)-(108)^2} = .575$$

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

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

Using the Regression Line

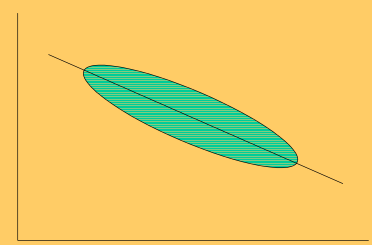
- When using a regression equation for prediction stay within the range of the available data.
- Don't make predictions about a population that is different from the population from which the sample were drawn.
- A regression equation based on old data may be no longer valid.



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

$\Sigma x=108$	$\Sigma y=123$	$\Sigma xy=1606$	$\Sigma x^2=1522$	$\Sigma y^2=1807$

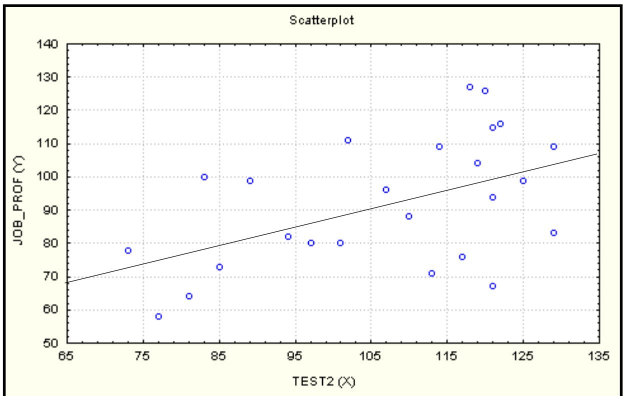
$$r=\frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

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

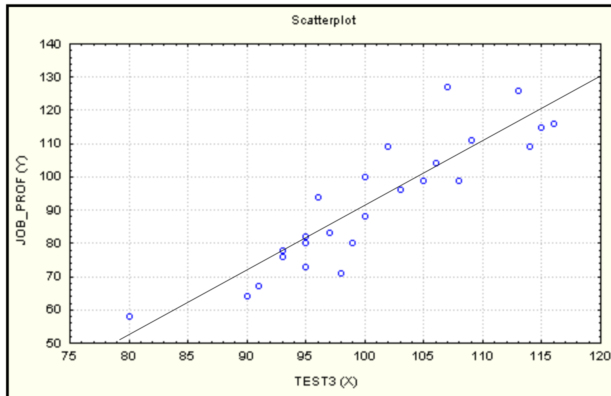
Correlation and Regression

- The regression line is the line that best fits the data: $\hat{y}=a+bx$
- 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

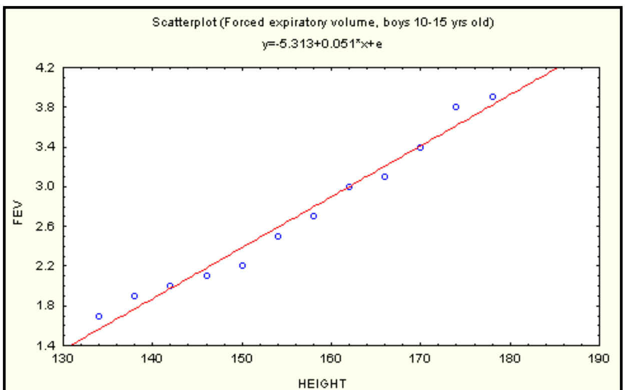
$$r=b\frac{S_x}{S_y}$$



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



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

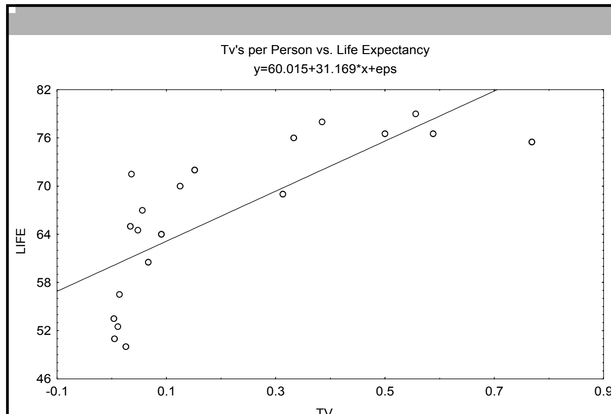


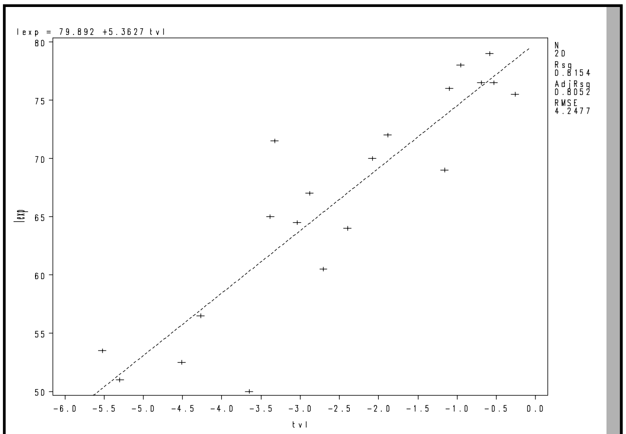
American Review of Respiratory Disease,
r=.98, n=12.

Another Example

- Looking at the relationship between number of tv sets per person and life expectancy over a number of countries.

Australia	0.500	76.5
Canada	0.588	76.5
China	0.125	70
Egypt	0.067	60.5
France	0.385	78
Haiti	0.004	53.5
Iraq	0.056	67
Japan	0.556	79
Madaga	0.011	52.5
Mexico	0.152	72
Morocco	0.048	64.5
Pakistan	0.014	56.5
Russia	0.313	69
South Afr	0.091	64
Sri Lanka	0.036	71.5
Uganda	0.005	51
United K	0.333	76
United S	0.769	75.5
Vietnam	0.034	65
Yemen	0.026	50





Taking the Log(tv)

```
data life; input country $ tv lexp; tvl=log(tv);cards;
Austra       0.500 76.5
Canada       0.588 76.5
;
proc reg data=life;
model lexp=tv;
plot lexp*tv; run;
proc reg data=life; model lexp=tvl;
plot lexp*tvl; run;
```

Correlation and Regression with SAS

Analysis of Variance (tv vs. life expectancy)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1037.46783	1037.46783	25.86	<.0001
Error	18	722.16967	40.12054		
Corrected Total	19	1759.63750			

Root MSE	6.33408	R-Square	0.5896 (r=.77)
Dependent Mean	66.42500	Adj R-Sq	0.5668
Coeff Var	9.53568		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	60.01514	1.89602	31.65	<.0001
tv	1	31.16879	6.12937	5.09	<.0001

$\sqrt{v(b_o)}$

$\sqrt{v(b_1)}$

SAS Output

Analysis of Variance (tv vs log of life expectancy)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1434.87186	1434.87186	79.53	<.0001
Error	18	324.76564	18.04254		
Corrected Total	19	1759.63750			

Root MSE	4.24765	R-Square	0.8154 (r=.90)
Dependent Mean	66.42500	Adj R-Sq	0.8052
Coeff Var	6.39466		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	79.89222	1.78401	44.78	<.0001
tvl	1	5.36272	0.60135	8.92	<.0001

SAS Output

A bit of Sampling Theory

$$MSE = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

$$\hat{v}(b_o) = MSE\left(\frac{\sum x^2 / n}{\sum (x-\bar{x})^2}\right)$$

$$\hat{v}(b_1) = \left(\frac{MSE}{\sum (x-\bar{x})^2}\right)$$

Test of Hypotheses and Confidence Intervals

$$\frac{b_i}{\sqrt{v(b_i)}} \sim t_{n-2}$$

Confidence Interval

$$b_i \pm t_{\alpha/2, n-2} * \sqrt{v(b_i)}$$

Variance Partitioning

- Just as in ANOVA, we can partition variability as follows:

SST=SSR+SSE

$$SST = \sum (y_i - \bar{y})^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

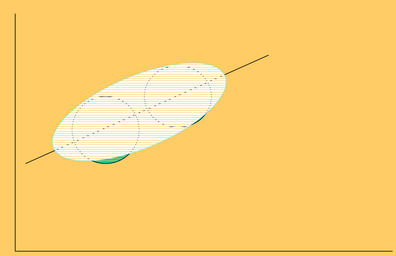
Coefficient of Determination

- R^2=1-(SSE/SST)=SSR/SST
- The proportional reduction in error from using the linear prediction equation (the variable) instead of the mean (of y) is called the coefficient of determination.

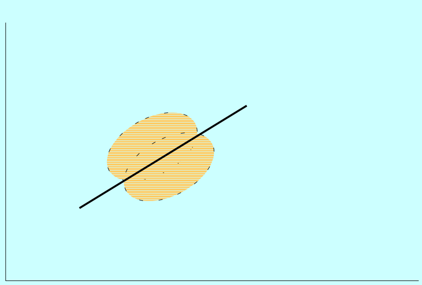
Factors Affecting the Correlation

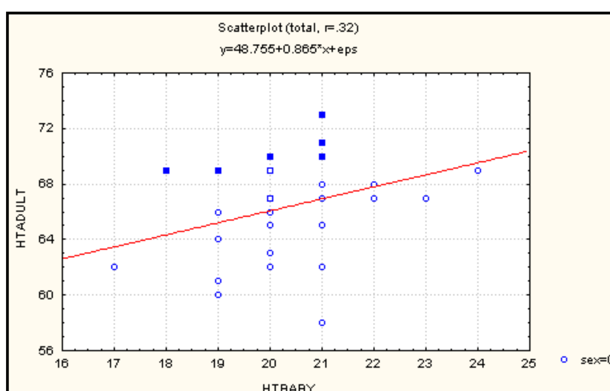
- Correlation is not causation
- Combined Groups
- Outliers
- Restriction in range
- By the way the correlation is invariant under linear transformation

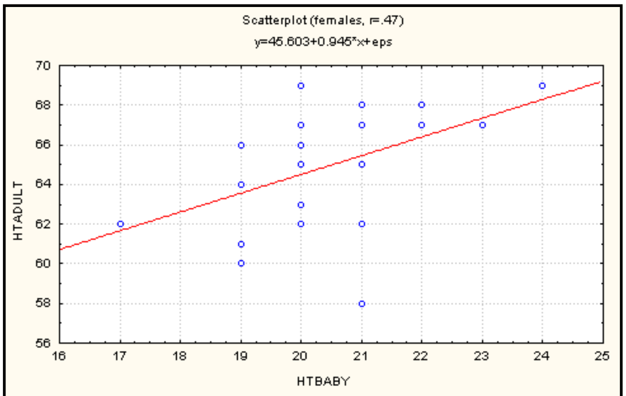
Combined Groups



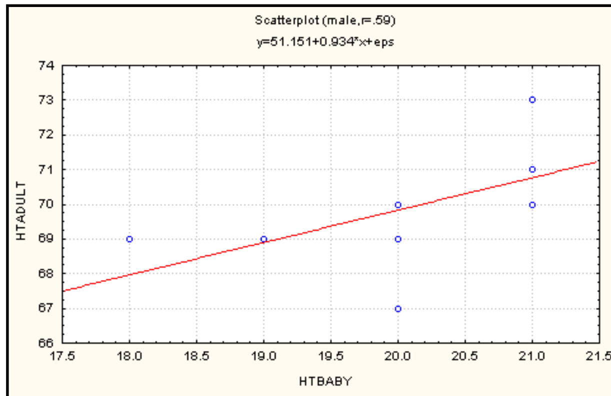
Combined Groups





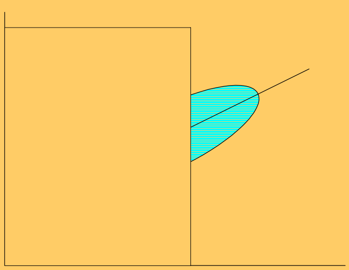


r=.472, n=27.



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

Assumptions for Test of Hypotheses

- X is normally distributed (or fixed)
- The conditional distribution of y given x is normal. (x and y follow a bivariate normal distribution)

Testing a hypothesis about the population correlation

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

The test is based on the t-test.

By the way, note that if* r*=0, then* b*=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 a t-table, computer output, or a table of correlations. To use this table, 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 $\rho \le .3$.
- 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

$$Ho: \rho \le .3$$

$$Ha: \rho > .3$$

$$z_r = \frac{1}{2}\ln\left(\frac{1+.472}{1-.472}\right) = .5126$$

$$z_\rho = \frac{1}{2}\ln\left(\frac{1+.3}{1-.3}\right) = .3095$$

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.

Spearman’s Rank Order

Correlation

Rank Correlation

Definition

- **uses the ranks of sample data and it is more forgiving than Pearson’s r.**
- **used to test for an association between two variables**
- **H_o: $\rho_s = 0$ (There is no correlation between the two variables.)**
- **H_1: $\rho_s \neq 0$ (There is a correlation between the two variables.)**

Assumptions

1. **The sample data have been randomly selected.**
2. **Unlike the parametric methods of, there is no requirement that the data follows a bivariate normal distribution. There is no requirement of a normality at all.**

Advantages

1. **The nonparametric method of rank correlation can be used in a wider variety of circumstances than the parametric method of linear correlation. With rank correlation, we can analyze paired data that are ranks (ordinal) or can be converted to ranks.**
2. **Rank correlation can be used to detect some (not all) relationships that are not linear.**

Notation

r_s = rank correlation coefficient for sample paired data (r_s is a sample statistic)

ρ_s = rank correlation coefficient for all the population data (ρ_s is a population parameter)

n = number of pairs of data

d = difference between ranks for the two values within a pair

r_s is often called Spearman’s rank correlation coefficient

Test Statistic for the Rank Correlation Coefficient

Test Statistic for the Rank Correlation Coefficient

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where each value of d is a difference between the ranks for a pair of sample data

Critical values:

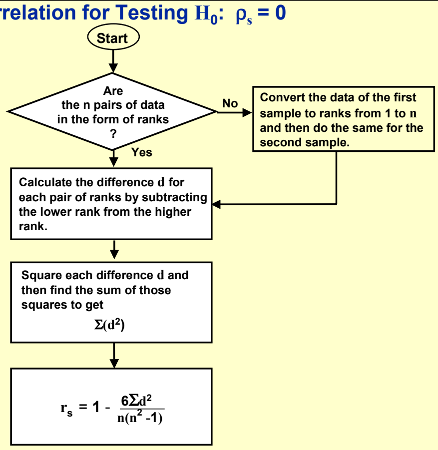
- If $n \le 30$, refer to a Table for Rank Order correlation
- If $n > 30$, use the Z-approximation

Critical Value for Spearman's Rank correlation when N>30

$$r_s = \frac{\pm z}{\sqrt{n - 1}}$$

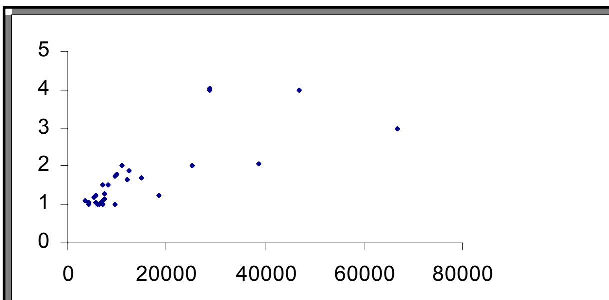
where the value of z corresponds to the significant level from the normal table

Rank Correlation for Testing H_0: $\rho_s = 0$



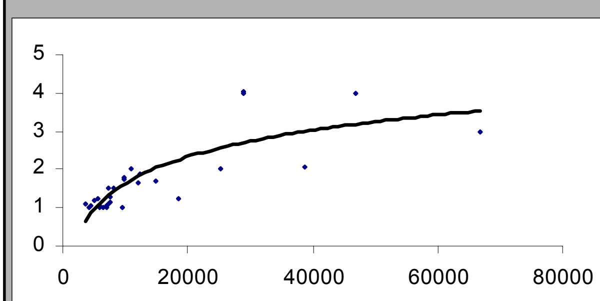
The Price of a Diamond

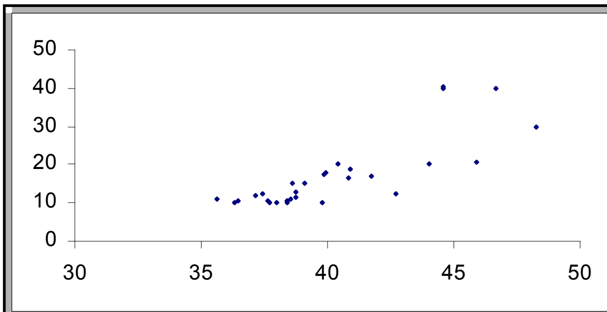
PRICE	CARAT	COLOR
6958	1	3
5885	1	5
6333	1.01	4
4299	1.01	5
9589	1.02	2
6921	1.04	4
4426	1.04	5
6885	1.07	4
5826	1.07	5
3670	1.11	9
7176	1.12	2
7497	1.16	5
5170	1.2	6
5547	1.23	7
18596	1.25	1
7521	1.29	6
7260	1.5	6
8139	1.51	6
12196	1.67	3
14998	1.72	4
9736	1.76	8
9859	1.8	5
12398	1.88	6
25322	2.03	2
11008	2.03	8
38794	2.06	2
66780	3	1
46769	4.01	3
28800	4.01	6
28868	4.05	7



Carat vs. Price

Adding a Logarithm Trend





Carat*10 vs. Log(price)

Rank Correlation

- Spearman's rank correlation between carat (weight) and price:
  r=.83, cv= $\pm$.364
- Sperman's rank correlation between price and color:
  r=.-.33, cv= $\pm$ .364

Applying the Rank Correlation to the tv example.

```
Spearman Correlation=0.8830, ASE=0.0385
Recall that the correlation without the
  transformation was r=.77.
And with the transformation it was r=.90
  transformation.
SAS Setup for Spearman's r:
proc freq data=life;
tables tv*lexp / chisq measures;run;
```

Other Relations that can be examined with the correlation

- With the correlation you can examine:
  1. Y=a + b x
  2. Y=a e^{bx} , transform y by taking the Log(y)
     Log(y)= Log(a) + b x
  3. Y=a + b Log(x)
  4. Y= a x^b, take the Log of both y and x
  5. Y= a + b x^2, put x^2 in the dataset