# Chapter 12
# Correlation


| | | Page |
|---|---|---|
| 1. | Pearson correlation coefficient | 12-2 |
| 2. | Inferential tests on correlation coefficients | 12-9 |
| 3. | Correlational assumptions | 12-13 |
| 4. | Non-parametric measures of correlation | 12-14 |
| 5. | A correlational example | 12-16 |
| 6. | Relationship between correlation and the t-test | 12-27 |
| 7. | Factors that will limit the size of the correlation coefficient | 12-30 |

# Correlation

1. Pearson product moment correlation coefficient

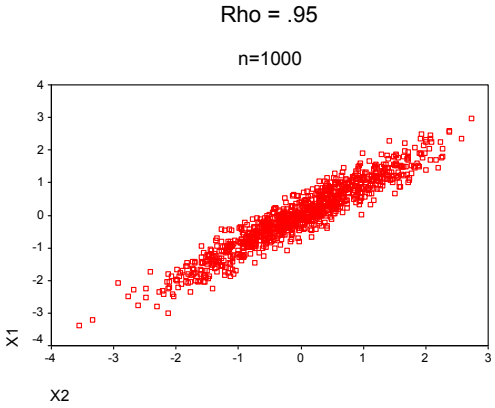
   - One method of describing the relationship between two variables is to indicate the strength of the linear relationship between the two variables.

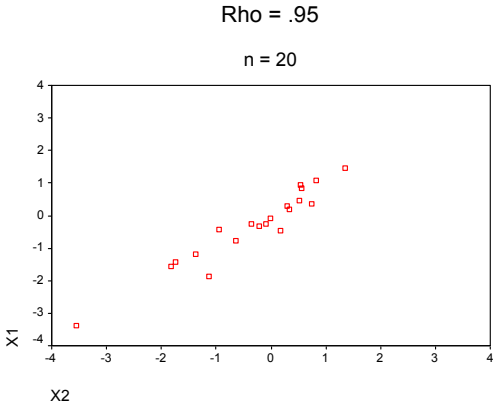
   - The correlation coefficient is a quantitative measure of this linear association between two variables.

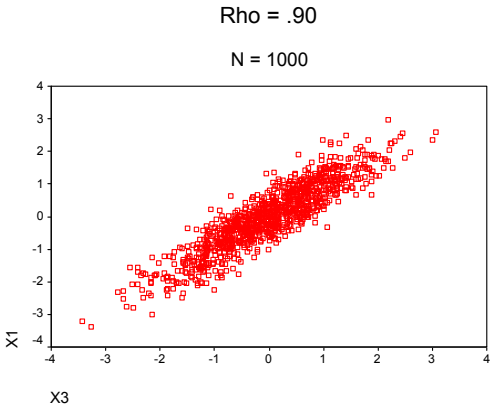
     - Rho, $\rho$, quantifies the linear relationship between two variables in the population.
       - Rho varies between –1 and +1.
       - $\rho = 0$ indicates no linear relationship between two variables.
       - The greater rho deviates from zero, the greater the strength of the linear relationship between the two variables:
         $\rho = +1$ indicates a perfect positive linear relationship between two variables.
         $\rho = -1$ indicates a perfect negative linear relationship between two variables.

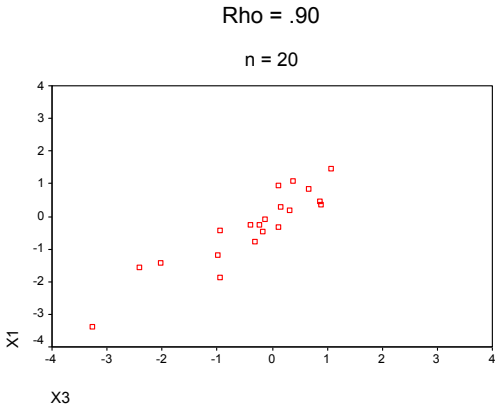
     - The Pearson product moment correlation coefficient, $r$ (1890's), quantifies the linear relationship between two variables in the sample.

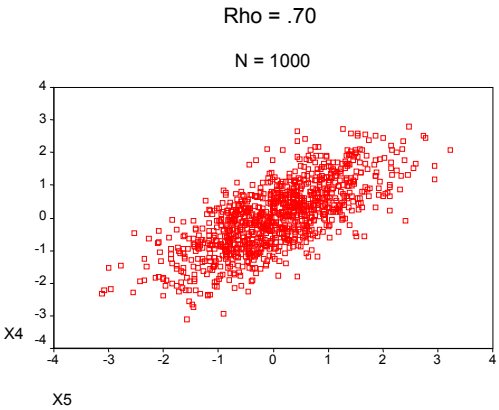
- Examples of linear relationships in large (n = 1000) and small (n = 20) datasets:

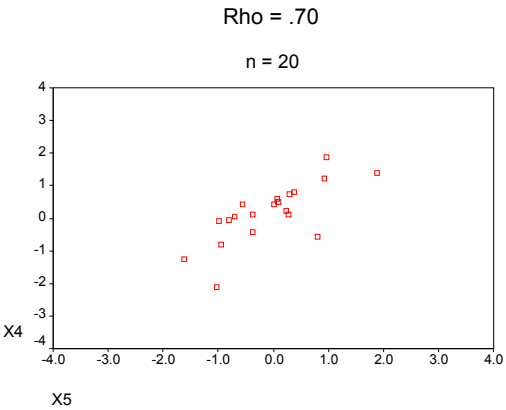


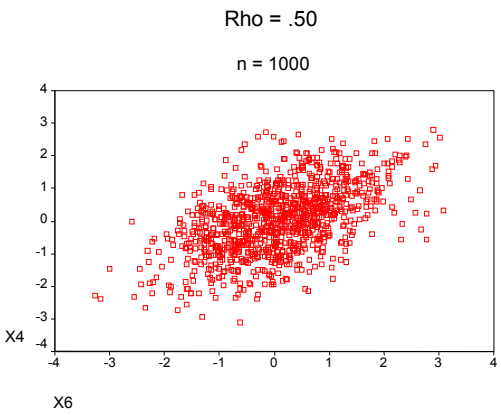



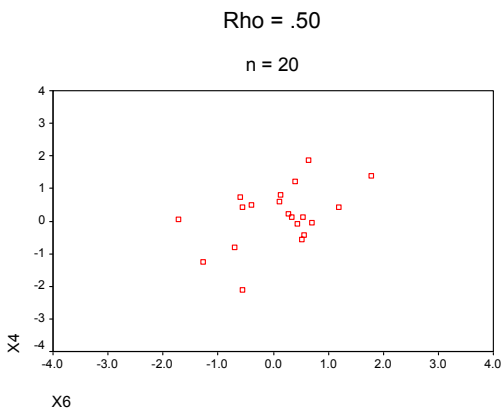



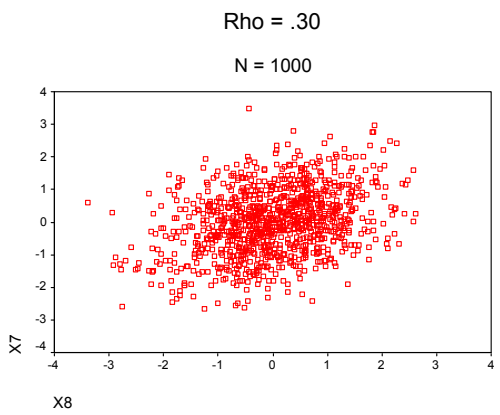



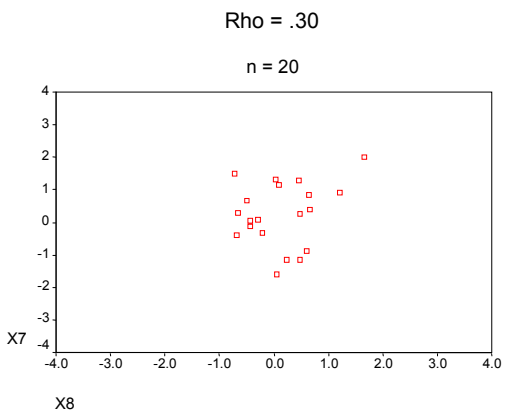



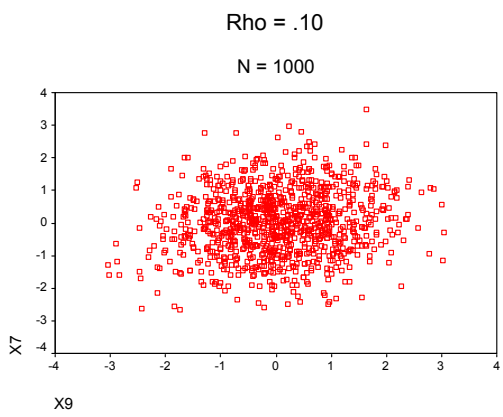



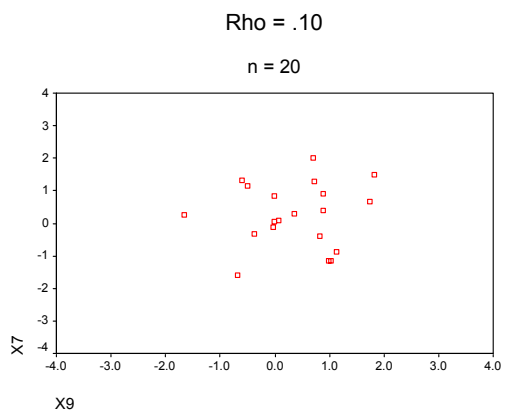


- Understanding the correlation coefficient:

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y))}}$$

- To understand the correlation between two variables, it is useful to consider the correlation as having two parts:
  - Part 1: A measure of the association between two variables
  - Part 2: A standardizing process

- Part 1 of the correlation coefficient: The <u>covariance</u> between two variables is a measure of linear association between two variables.

  - The covariance is similar to the variance, except that the covariance is defined over two variables (X and Y) whereas the variance is defined over one variable (X).

$$Var(X) = \frac{\sum(X_i - \bar{X})^2}{N-1} = \frac{\sum(X_i - \bar{X})(X_i - \bar{X})}{N-1} = \frac{SS_{XX}}{N-1}$$

    When we expand this formula to two variables, we obtain the covariance:

$$Cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N-1} = \frac{SS_{XY}}{N-1}$$

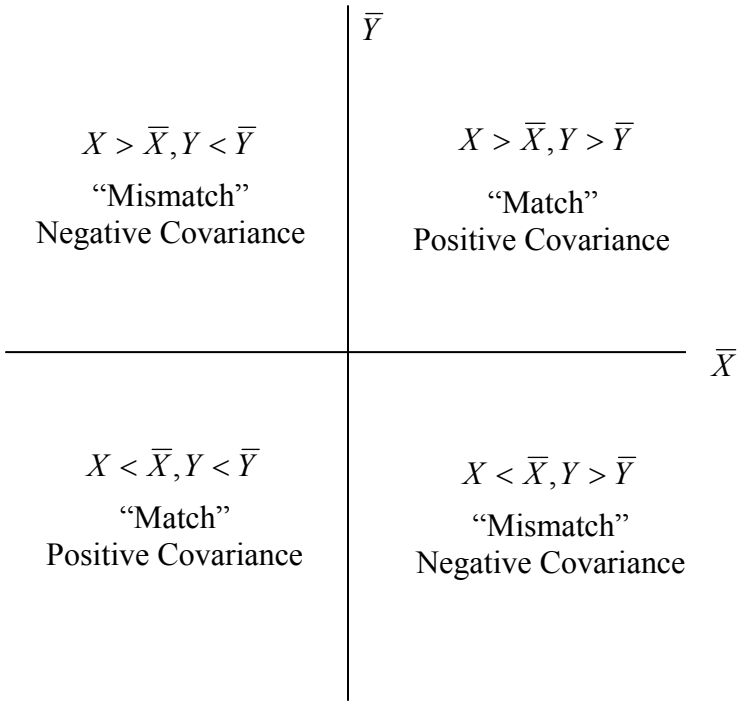
  - The covariance is a measure of the direction and the magnitude of the linear association between $X$ and $Y$.

    - Covariance can be thought of as the sum of <u>matches</u> and <u>mismatches</u> across subjects.

    - A <u>match</u> occurs when both variables are on the same side of the mean.
    - A <u>mismatch</u> occurs when the score on one variable is above the mean and the score on the other variable is below the mean (or vice versa).

  - We can think of the variance as the covariance of a variable with itself:
    - For $Cov_{XX}$, all pairs will be matches.
    - $Cov_{XX}$ (or $Var_X$) then is the extent to which $X$ deviates from its mean.

- For $Cov_{XY}$, some pairs may be matches and some may be mismatches:




- The sign of the covariance tells us the <u>direction</u> of the relationship between $X$ and $Y$.
- The size of the covariance tells us the <u>magnitude</u> of the relationship between $X$ and $Y$.
  - If there is a strong linear relationship, then most pairs will be matches and the covariance will be large.
  - If there is a weak linear relationship, then some mismatches will cancel some of the matches, and the covariance will be small.

- The covariance only describes *linear* relationships between $X$ and $Y$.

- The covariance depends on the scale of the variable, making the magnitude of the relationship difficult to interpret:
  - If responses are on a 1-7 scale, then the covariance will be relatively small.
  - If responses are on a 1-100 scale, then the covariance will be relatively large.

- Part 2 of the correlation coefficient: A standardizing process so that the correlation coefficient will not depend on the scale of a variable.

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y))}}$$

$$r_{XY} = \frac{\frac{SS_{XY}}{N-1}}{\sqrt{\frac{SS_{XX}}{N-1}}\sqrt{\frac{SS_{YY}}{N-1}}} = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}}$$

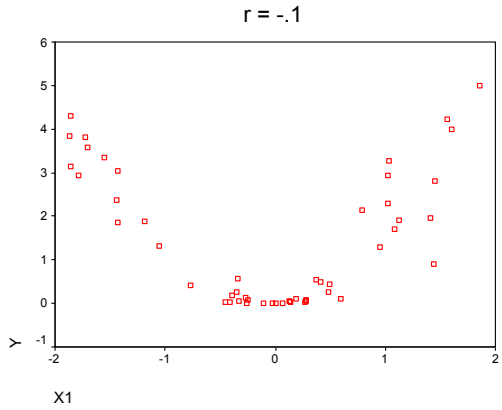
  - By dividing the covariance by the standard deviation of each variable, we standardize the covariance. The correlation ranges between –1 and +1, regardless of scale of the variables.
  - The *N-1* terms in the numerator and denominator cancel. Thus, a correlation can be thought of as a ratio of sums of squares (it is the product of the first moments).

- Properties of the correlation coefficient:
  - If we multiply each variable by a constant, the correlation remains unchanged.
  - If we add a constant to each variable, the correlation remains unchanged.

- Effect size
  - $r_{XY}$ is a measure of effect size. It measures the strength of the linear association between two variables.

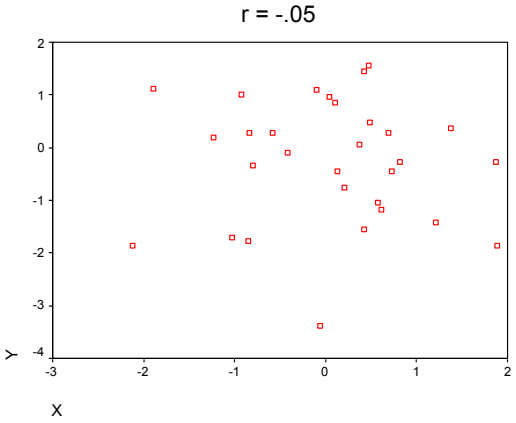
  - $r_{XY}^2$ is known as the coefficient of determination
    - It is a measure of the proportion of the variance in *Y* that is accounted for by the linear relationship between *X* and *Y*.
    - This measure of "variance accounted for" differs from our previous effect size measures because it only measures the variance accounted for by the linear relationship.

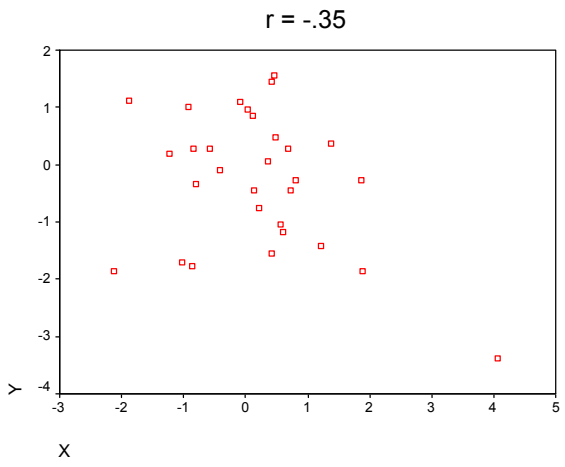
- Cautions regarding correlation coefficients
  - There may be a non-linear relationship between $X$ and $Y$, but $r_{XY}$ will only capture linear relationships. $r_{XY}$ will not be useful in measuring non-linear relationships between $X$ and $Y$.

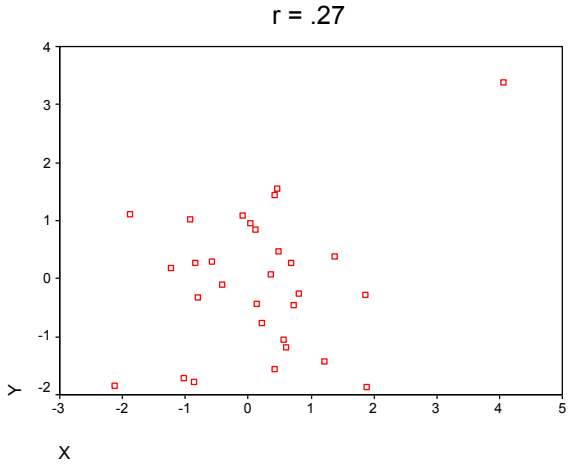



There is no <u>linear relationship</u> between $X$ and $Y$, but it would be misleading to say that $X$ and $Y$ were unrelated.

  - The correlation coefficient is quite sensitive to outliers



2. Inferential tests on correlation coefficients

   - To test if a correlation is different from zero:
     - State null and alternative hypotheses:
       $H_0 : \rho_{XY} = 0$
       $H_1 : \rho_{XY} \neq 0$

     - Construct test statistic:

$$t \sim \frac{estimate}{standard\ error\ of\ the\ estimate}$$

       We can estimate $\rho_{XY}$ with the correlation coefficient, $r_{XY}$

       Without going into all the details, we can compute the standard error of $r_{XY}$ with the following formula:

$$\sqrt{\frac{1-r_{XY}^2}{N-2}}$$

       Putting both parts together, we obtain:

$$t(N-2) \sim \frac{r_{XY}}{\sqrt{\frac{1-r_{XY}^2}{N-2}}}$$

     - Look-up p-value.

     - Alternatively, some people prefer tables of critical values for significant correlations.

       Notice that the t-test for the correlation only depends on the size of the correlation and the sample size. Thus, we can work backwards and determine critical r-values for significance at $\alpha = .05$. However, it is usually preferable to report exact p-value.

- In SPSS

```
CORRELATIONS
 /VARIABLES=x1 x2 x3.
```

  - If you use the pull-down menus, be sure to use "Bivariate correlation" with a two-tailed Pearson correlation coefficient.
    (These are defaults so no extra syntax is required)

  - SPSS outputs a correlation matrix. You only need to examine the top half or the bottom half.

**Correlations**

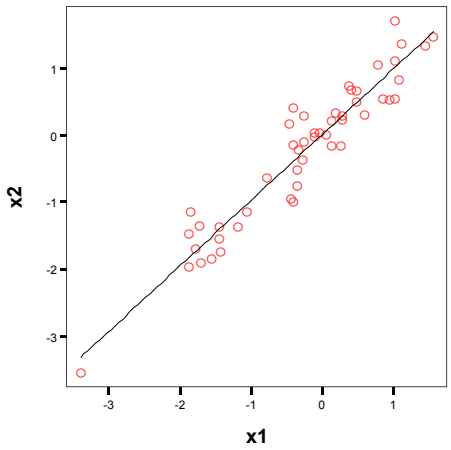
| | | X1 | X2 | X3 |
|---|---|---|---|---|
| X1 | Pearson Correlation | 1 | .954 | .903 |
| | Sig. (2-tailed) | . | .000 | .000 |
| | N | 50 | 50 | 50 |
| X2 | Pearson Correlation | .954 | 1 | .958 |
| | Sig. (2-tailed) | .000 | . | .000 |
| | N | 50 | 50 | 50 |
| X3 | Pearson Correlation | .903 | .958 | 1 |
| | Sig. (2-tailed) | .000 | .000 | . |
| | N | 50 | 50 | 50 |

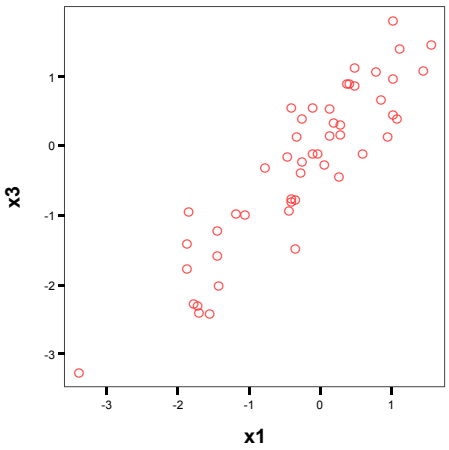
For the correlation between *X1* and *X2*: $r(48) = .95, p < .001$

For the correlation between *X1* and *X3*: $r(48) = .90, p < .001$

For the correlation between *X2* and *X3*: $r(48) = .96, p < .001$

  - Anytime you report a correlation, you should examine the scatterplot between those two variables.
    ⇒ To check for outliers
    ⇒ To make sure that the relationship between the variables is a <u>linear relationship</u>



- These "standard" significance tests only apply for testing for differences from zero.
  - The sampling distribution of the correlation is only symmetric when $\rho = 0$. If $\rho \neq 0$, then the sampling distribution is asymmetric and other methods of inference must be used.

- The Fisher r-to-z transformation for a single correlation:

  - Fisher showed that we can transform a correlation to a z-score:

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

    The sampling distribution of $Z_f$ is asymptotically normal with variance:

$$\frac{1}{N-3}$$

    Where $N$ = the number of participants used to compute the correlation (or the number of pairs of scores)

  - As a consequence, we can compute a test statistic:

$$Z = \frac{Z_f - Z_{null}}{\sqrt{\frac{1}{N-3}}}$$

    Where $Z_{null}$ is the null hypothesis of *r* transformed to the $Z_f$ scale. This test statistic follows a standard normal distribution.

    If $|Z_{obs}| > 1.96$ then $r_{obs}$ differs from $r_{null}$ with a two-tailed $\alpha = .05$
    Or look up exact p-value on z-table

- For example, let's test if $r_{12} = .954$ differs from $r_{null} = .80$, $N = 50$

$$H_0 : \rho_{12} = .80$$
$$H_0 : \rho_{12} \neq .80$$

$$Z_f = \frac{1}{2}\ln\left(\frac{1+.954}{1-.954}\right) = 1.8745 \qquad Z_{null} = \frac{1}{2}\ln\left(\frac{1+.8}{1-.8}\right) = 1.0986$$

$$Z = \frac{1.8745 - 1.0986}{\sqrt{\frac{1}{50-3}}} = \frac{0.7759}{.146} = 5.31, p < .001$$

We conclude that $r_{12} = .954$ differs significantly from $r_{null} = .80$

o Extending the Fisher r-to-z transformation to correlations from two independent samples:

- When correlations come from two independent samples, we may want to know if they differ from each other

$$H_0 : \rho_1 = \rho_2 \qquad H_1 : \rho_1 \neq \rho_2$$
$$H_0 : \rho_1 - \rho_2 = 0 \qquad H_1 : |\rho_1 - \rho_2| > 0$$

Or more generally

$$H_0 : \rho_1 - \rho_2 = k \qquad H_1 : |\rho_1 - \rho_2| > k$$

After transforming both observed correlations to the $Z_f$ scale, we can compute a Z-test statistic:

$$Z = \frac{Z_{f1} - Z_{f2} - k}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Where $n_1$ = the number of participants used to compute the first correlation

$n_2$ = the number of participants used to compute the second correlation

If $|Z_{obs}| > 1.96$ then reject the two-tailed null hypothesis at $\alpha = .05$

- For example, suppose $r_1 = .2$ with $N_1 = 40$ and $r_2 = .8$ with $N_2 = 45$. We would like to test if $r_1$ differs from $r_2$.

  $H_0 : \rho_1 = \rho_2$
  $H_0 : \rho_1 - \rho_2 = 0$

  First transform $r_1 = .2$ and $r_2 = .8$ to the $Z_f$ scale

  $$Z_{f1} = \frac{1}{2}\ln\left(\frac{1+.2}{1-.2}\right) = 0.2027 \qquad Z_{f2} = \frac{1}{2}\ln\left(\frac{1+.8}{1-.8}\right) = 1.0986$$

  Next compute the test statistic, $Z$

  $$Z = \frac{.2027 - 1.0986 - 0}{\sqrt{\frac{1}{40-3} + \frac{1}{45-3}}} = \frac{-0.8959}{.2255} = -3.973$$

  Finally, determine significance

  $|Z_{obs}| = 3.973, p = .000071$
  Reject null hypothesis with 2-tailed test

3. Correlational assumptions

- The assumptions of this of the test for a correlation are that both X and Y are normally distributed (Actually X and Y must jointly follow a bivariate normal distribution).
  - No other assumptions are required, but remember:
    - The correlation coefficient is very sensitive to outliers
    - The correlation coefficient only detects <u>linear relationships</u>
  - These assumptions can be checked visually:
    - Boxplots, histograms & univariate scatterplots of each variable
    - Bivariate scatterplots of the two variables together

- The normality assumption is only required for the significance test of the correlation. It is not necessary if you only want to calculate the correlation coefficient.

- If you have a violation of the normality assumption (or if outliers are present)
  - Use the Spearman rank correlation

4. Non-parametric measures of correlation

- The Spearman rank correlation, $\rho$ (1904), is a correlation performed on the rank of the variables.
  - Rank X from small to large (*rX*).
  - Rank Y from small to large (*rY*).
  - Compute the Pearson correlation coefficient on the rank variables.
    - (Spearman's formula for determining significance is actually a bit different, but for large sample sizes the results are similar.)

```
RANK VARIABLES=x1 x2 x3.
CORRELATIONS
  /VARIABLES=rx1 rx2 rx3.
```

**Correlations**

| | | RANK of X1 | RANK of X2 | RANK of X3 |
|---|---|---|---|---|
| RANK of X1 | Pearson Correlation | 1 | .935 | .885 |
| | Sig. (2-tailed) | . | .000 | .000 |
| | N | 50 | 50 | 50 |
| RANK of X2 | Pearson Correlation | .935 | 1 | .954 |
| | Sig. (2-tailed) | .000 | . | .000 |
| | N | 50 | 50 | 50 |
| RANK of X3 | Pearson Correlation | .885 | .954 | 1 |
| | Sig. (2-tailed) | .000 | .000 | . |
| | N | 50 | 50 | 50 |

For the correlation between *X1* and *X2*: $\rho(48) = .94, p < .001$

For the correlation between *X1* and *X3*: $\rho(48) = .89, p < .001$

For the correlation between *X1* and *X3*: $\rho(48) = .95, p < .001$

- You can also ask for Spearman's *rho* directly

```
NONPAR CORR
  /VARIABLES=x1 x2 x3
  /PRINT=SPEARMAN.
```

**Correlations**

| | | | X1 | X2 | X3 |
|---|---|---|---|---|---|
| Spearman's rho | X1 | Correlation Coefficient | 1.000 | .935 | .885 |
| | | Sig. (2-tailed) | . | .000 | .000 |
| | | N | 50 | 50 | 50 |
| | X2 | Correlation Coefficient | .935 | 1.000 | .954 |
| | | Sig. (2-tailed) | .000 | . | .000 |
| | | N | 50 | 50 | 50 |
| | X3 | Correlation Coefficient | .885 | .954 | 1.000 |
| | | Sig. (2-tailed) | .000 | .000 | . |
| | | N | 50 | 50 | 50 |

- Spearman's *rho* can be used for:
  ⇒ Non-normal data
  ⇒ Data with outliers
  ⇒ Data that arrive in rank format

  However, *rho* still only detects linear relationships.

- If both of *X* and *Y* are dichotomous, then you can conduct a compute phi, $\phi$, as a measure of the strength of association between dichotomous variables (see p. 2-65).

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

  - An example:

| | Candidate Preference | |
|---|---|---|
| | Candidate U | Candidate V |
| Homeowners | 19 | 54 |
| Non-Homeowners | 60 | 52 |

```
CROSSTABS
 /TABLES = prefer by homeown
 /STAT = CHISQ.
```

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 13.704 | 1 | .000 |
| N of Valid Cases | 185 | | |

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{13.704}{185}} = .272$$

5. A correlational example
   - Consider the SENIC (Study on the Efficacy of Nosocomial Infection Control) data set from 1975-76. This data set contains information on 113 hospitals (each "subject" is a hospital) including:

| | |
|---|---|
| length | 'Average length of stay (in days)' |
| infrisk | 'Probability of acquiring an infection in the hospital' |
| age | 'Average age (years)' |
| beds | 'Number of beds in the hospital' |

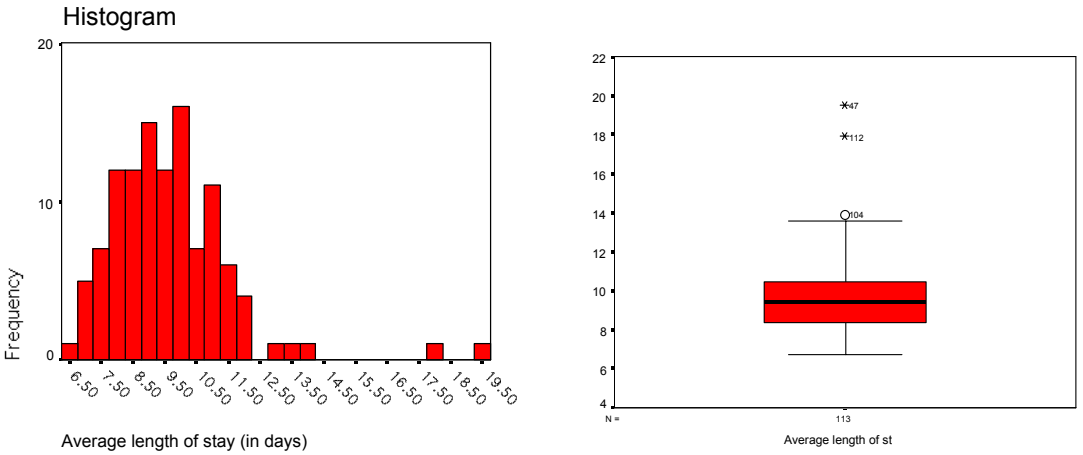
   - We would like to look for relationships between each of the variables and the average length of stay in hospitals.
   - For a correlational analysis to be valid, we need:
     - The relationships between all the variables to be linear.
     - Each variable to be normally (or symmetrically distributed).
     - No outlying observations influencing the analysis.

   - Let's start by checking our key variable, average length of stay, for normality and outliers:

```
EXAMINE VARIABLES=length
  /PLOT BOXPLOT HISTOGRAM NPPLOT.
```

**Descriptives**

| | | Statistic | Std. Error |
|---|---|---|---|
| LENGTH | Mean | 9.6483 | .17981 |
| | 5% Trimmed Mean | 9.4864 | |
| | Median | 9.4200 | |
| | Minimum | 6.70 | |
| | Maximum | 19.56 | |
| | Skewness | 2.069 | .227 |
| | Kurtosis | 8.077 | .451 |



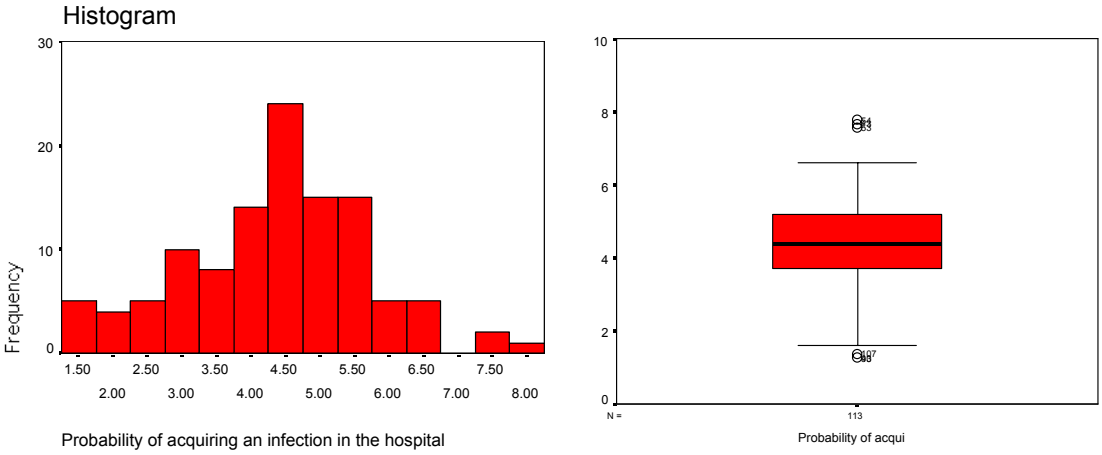

     - It looks like we have a normal distribution with two outliers. We will have to keep track of these outliers and see if they influence our conclusions.

- To examine the relationship between length of hospital visits and the probability of acquiring an infection in the hospital, we need to make sure that this second variable is normally/symmetrically distributed with no outliers.

```
EXAMINE VARIABLES=infrisk
  /PLOT BOXPLOT HISTOGRAM NPPLOT.
```

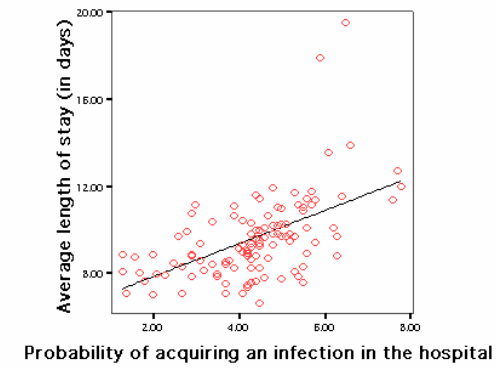
**Descriptives**

| | | Statistic | Std. Error |
|---|---|---|---|
| INFRISK | Mean | 4.3549 | .12614 |
| | 5% Trimmed Mean | 4.3586 | |
| | Median | 4.4000 | |
| | Minimum | 1.30 | |
| | Maximum | 7.80 | |
| | Skewness | -.120 | .227 |
| | Kurtosis | .182 | .451 |




  - Risk of infection looks normal and symmetric.

- The next step is to check for non-linearity in the relationship between length of stay and infection risk




  - The relationship looks linear, but those two outliers are menacing!

- Now that we have done all our back ground work, we can finally examine the correlation between the two variables

```
CORRELATIONS /VARIABLES=length WITH infrisk.
```

**Correlations**

| | | INFRISK |
|---|---|---|
| LENGTH | Pearson Correlation | .533 |
| | Sig. (2-tailed) | .000 |
| | N | 113 |

  - There is a significant positive correlation between infection risk and average length of stay in a hospital, $r(111) = .533, p < .001$.

- But we want to make sure that those outliers are not influencing our conclusions. We have two methods to examine their influence:
  - Conduct a sensitivity analysis.
  - Conduct a rank regression (Spearman's rho).

  - Let's start by conducting a sensitivity analysis:

```
TEMPORARY.
SELECT IF length < 15.
CORRELATIONS /VARIABLES=length WITH infrisk.
```

**Correlations**

| | | INFRISK |
|---|---|---|
| LENGTH | Pearson Correlation | .549 |
| | Sig. (2-tailed) | .000 |
| | N | 111 |

    - Correlation with the outliers: $r(111) = .533, p < .001$
    - Correlation without the outliers: $r(109) = .549, p < .001$

    - In this case the outliers do not influence the magnitude of the correlation or the significance of the correlation. It appears that we can report the correlation on the full data with confidence.

- A second check on the influence of the outliers can be done by computing Spearman's Rho on the data.

  - If Spearman's Rho and the Pearson correlation are similar in magnitude, then we can be more confident the outlier does not influence the analysis.

  - There are two ways to compute Spearman's Rho. First, we can ask for it directly in SPSS.

```
NONPAR CORR /VARIABLES=length WITH infrisk age
 /PRINT=SPEARMAN .
```

**Correlations**

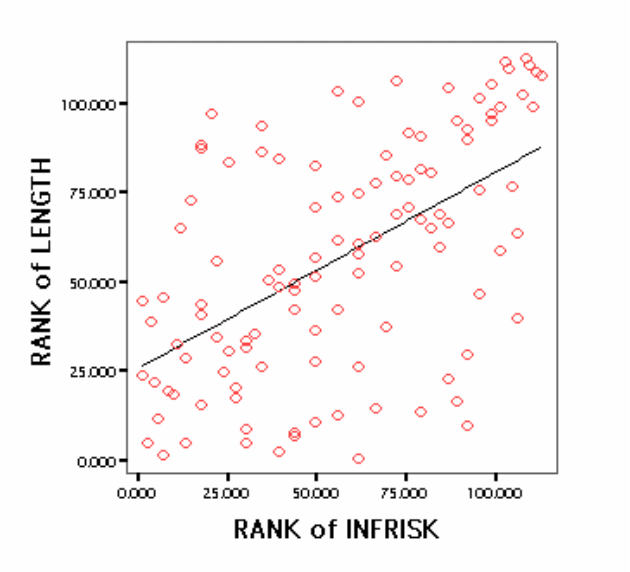
| | | | INFRISK |
|---|---|---|---|
| Spearman's rho | LENGTH | Correlation Coefficient | .549 |
| | | Sig. (2-tailed) | .000 |
| | | N | 113 |

  - Alternatively, we can rank the data manually.

```
RANK VARIABLES= infrisk length.
```

  - The advantage of this method is that we can look at the ranked data

```
GRAPH /SCATTERPLOT(BIVAR)=rinfrisk WITH rlength.
```



- We can also obtain the Rho by performing a Pearson correlation on the ranked data

```
CORRELATIONS /VARIABLES=rlength WITH rinfrisk rage.
```

**Correlations**

| | | RANK of INFRISK |
|---|---|---|
| RANK of LENGTH | Pearson Correlation | .549 |
| | Sig. (2-tailed) | .000 |
| | N | 113 |

- In this case we find nearly identical results for the two methods (Pearson vs. Spearman) of computing the correlation:

$$\rho(111) = .549, p < .001$$
$$r(111) = .533, p < .001$$

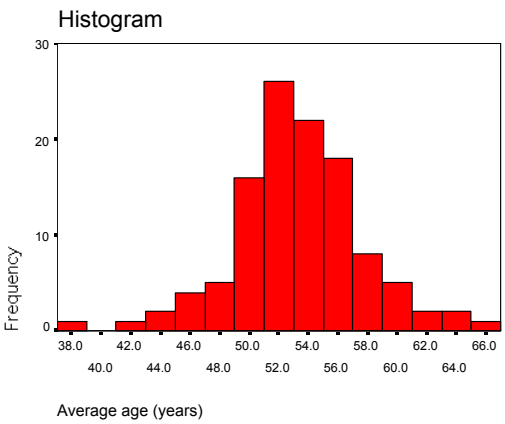
  - This finding gives us additional confidence that the Pearson correlation is an accurate estimate of the linear relationship between the two variables.

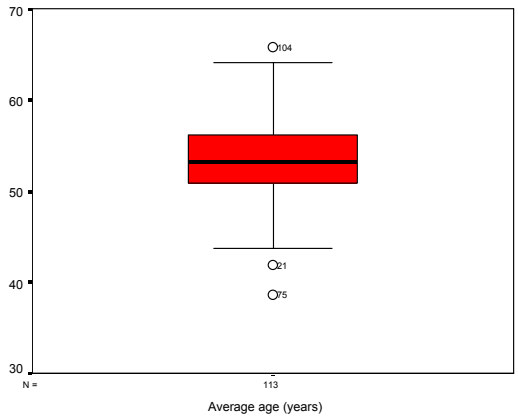
- Now, let's turn to the second variable of interest. To examine the relationship between length of hospital visits and the average age of patients, we need to make sure that are is normally/symmetrically distributed with no outliers and that the relationship between the variables is linear.

```
EXAMINE VARIABLES= age
 /PLOT BOXPLOT HISTOGRAM NPPLOT.
```

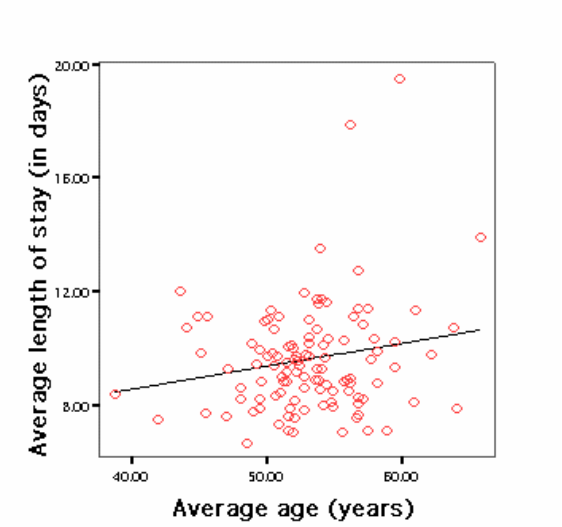
**Descriptives**

| | | Statistic | Std. Error |
|---|---|---|---|
| AGE | Mean | 53.2319 | .41971 |
| | 5% Trimmed Mean | 53.2481 | |
| | Median | 53.2000 | |
| | Minimum | 38.80 | |
| | Maximum | 65.90 | |
| | Skewness | -.104 | .227 |
| | Kurtosis | 1.066 | .451 |




- Normality/symmetry looks OK for the age variable




- This scatterplot does not contain any indications of nonlinearity. However, those two outliers will again require out attention

- Let's look at the correlation between length of stay and average age of patients.

```
CORRELATIONS /VARIABLES=length WITH age.
```

**Correlations**

| | | AGE |
|---|---|---|
| LENGTH | Pearson Correlation | .189 |
| | Sig. (2-tailed) | .045 |
| | N | 113 |

We find a significant linear relationship between average length of stay and average age of patients, $r(111) = .189, p = .045$.

- Now, let's check the influence of the outliers by conducting a sensitivity analysis and by examining Spearman's Rho.

  - A sensitivity analysis:
    TEMPORARY.
    SELECT IF length < 15.
    CORRELATIONS /VARIABLES=length WITH age.

**Correlations**

| | | AGE |
|---|---|---|
| LENGTH | Pearson Correlation | .122 |
| | Sig. (2-tailed) | .201 |
| | N | 111 |

  - This time, we find a very different result (in terms of both magnitude and significance) when we omit the outliers.

$$r(109) = .122, p = .201$$

  - Rank-based correlation:
    NONPAR CORR /VARIABLES=length WITH age
    /PRINT=SPEARMAN .

**Correlations**

| | | | AGE |
|---|---|---|---|
| Spearman's rho | LENGTH | Correlation Coefficient | .113 |
| | | Sig. (2-tailed) | .232 |
| | | N | 113 |

  - Again, we find a very different result (in terms of both magnitude and significance) when we conduct Spearman's Rho

$$\rho(111) = .113, p = .232$$

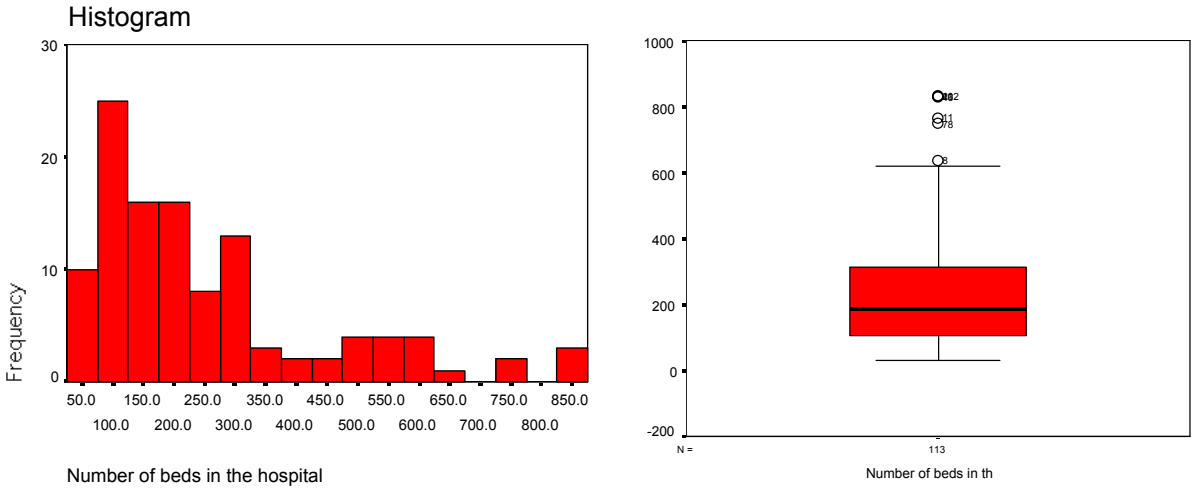
- In this case, the two outliers exert a large influence on the conclusions we draw. We should be very cautious about reporting and interpreting the Pearson correlation on the complete data.

- Finally, to examine the relationship between the third variable, number of hospital beds, and average length of stay in the hospital, we need to examine the assumptions.

```
EXAMINE VARIABLES= beds
 /PLOT BOXPLOT HISTOGRAM NPPLOT.
```

**Descriptives**

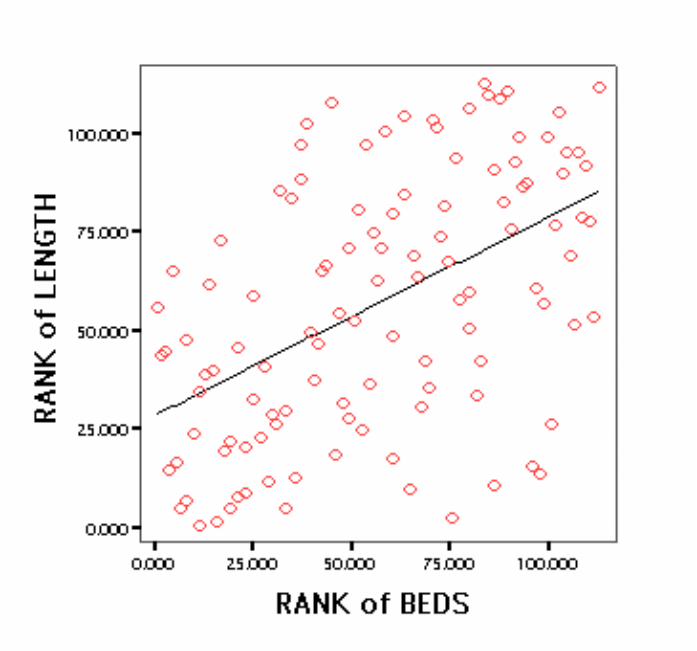
| | | Statistic | Std. Error |
|---|---|---|---|
| BEDS | Mean | 252.1681 | 18.14111 |
| | 5% Trimmed Mean | 233.5772 | |
| | Median | 186.0000 | |
| | Minimum | 29.00 | |
| | Maximum | 835.00 | |
| | Skewness | 1.379 | .227 |
| | Kurtosis | 1.281 | .451 |




  - This looks like a clear case of non-normality. If we run a Pearson correlation on these data, the statistical tests and p-values will be biased.
  - The solution is to compute Spearman's Rho rather than the Pearson correlation.

- But first, we need to make sure that the relationship between rank of average length of stay and rank of number of hospital beds is linear

```
RANK VARIABLES=beds.
GRAPH /SCATTERPLOT(BIVAR)=rbeds WITH rlength.
```




```
CORRELATIONS /VARIABLES=rlength WITH rbeds.
```

**Correlations**

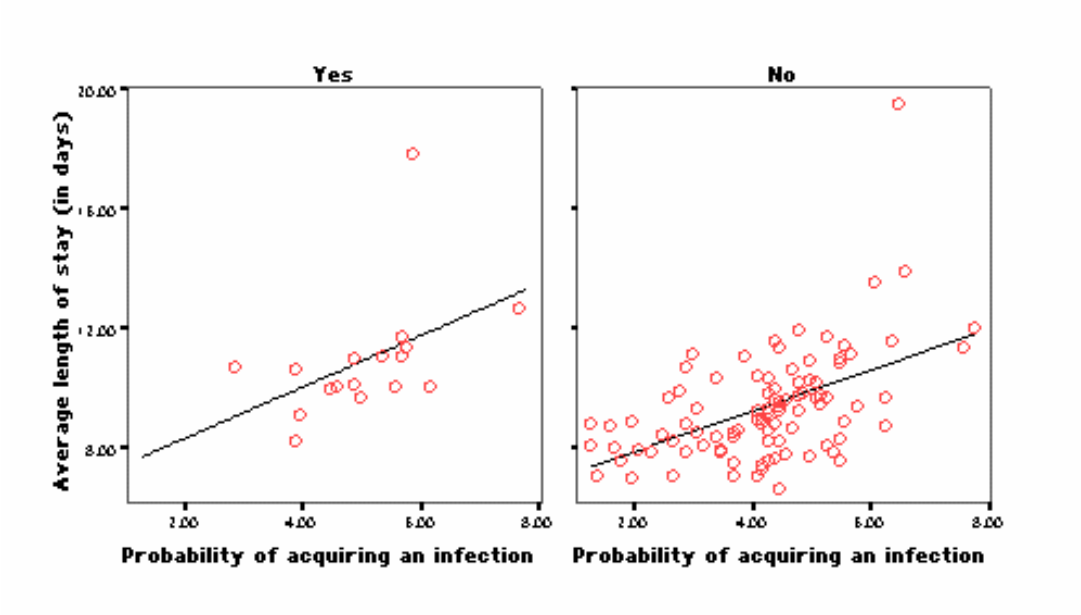
| | | RANK of BEDS |
|---|---|---|
| RANK of LENGTH | Pearson Correlation | .503 |
| | Sig. (2-tailed) | .000 |
| | N | 113 |

$$\rho(111) = .503, p < .001$$

- We find a significant linear relationship between the rank of average length of stay and the rank of number of hospital beds.
- Because of the assumption violation, we should report Spearman's Rho rather than the Pearson correlation.

- One additional variable in the SENIC dataset is affiliation with a medical school or not. Suppose we would like to check if the correlation between average length of hospital stay and risk of hospital infection differs for med school affiliated hospitals compared to non-med school affiliated hospitals

  - First, compute the Pearson correlation separately for each sample.




```
TEMPORARY.
SELECT IF medsch=1.
CORRELATIONS
  /VARIABLES=length WITH infrisk.
```

**Correlations**

| | | INFRISK |
|---|---|---|
| LENGTH | Pearson Correlation | .463 |
| | Sig. (2-tailed) | .061 |
| | N | 17 |

```
TEMPORARY.
SELECT IF medsch=2.
CORRELATIONS
/VARIABLES=length WITH infrisk.
```

**Correlations**

| | | INFRISK |
|---|---|---|
| LENGTH | Pearson Correlation | .510 |
| | Sig. (2-tailed) | .000 |
| | N | 96 |

- There appears to be a similar sized linear relationship in both samples. To statistically test for this difference, we can conduct a Fisher r-to-z transformation and then conduct a z-test.

$$H_0 : \rho_{No} = \rho_{Yes}$$
$$H_1 : \rho_{No} \neq \rho_{Yes}$$

$$Z_{Yes} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+.463}{1-.463}\right) = .501 \qquad Z_{No} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+.510}{1-.510}\right) = .563$$

$$Z = \frac{Z_{No} - Z_{Yes}}{\sqrt{\frac{1}{N_{No}-3} + \frac{1}{N_{Yes}-3}}} = \frac{.563 - .501}{\sqrt{\frac{1}{14} + \frac{1}{93}}} = 0.217, p = .83$$

- We fail to reject the null hypothesis. There is insufficient evidence to conclude the linear relationship between length of stay and risk of infection differs by med school affiliation, $Z = 0.22, p = .83$.

6. Relationship between correlation and the t-test

- Consider the relationship between a dichotomous independent variable and a continuous variable (and for the simplicity, we will assume that we have nice data with all standard parametric assumptions satisfied).

- Previously, we considered the relationship between sleep deprivation and memory (see 2-45).

| Control | | Sleep Deprived | |
|---|---|---|---|
| 55 | 58 | 48 | 55 |
| 43 | 45 | 38 | 40 |
| 51 | 48 | 53 | 49 |
| 62 | 54 | 58 | 50 |
| 35 | 56 | 36 | 58 |
| 48 | 32 | 42 | 25 |

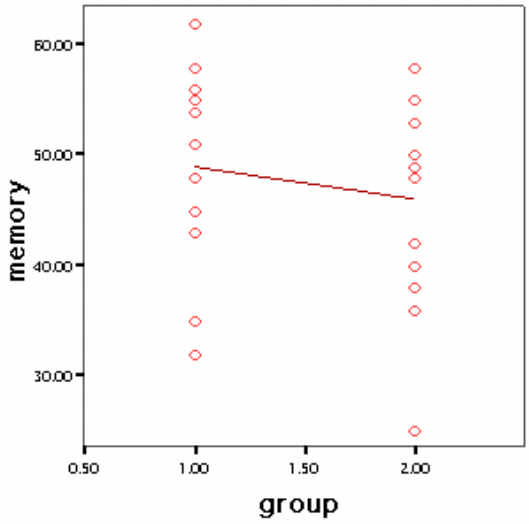
**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| MEMORY | Equal variances assumed | .256 | .618 | .748 | 22 | .462 | 2.9167 | 3.89922 | -5.16982 | 11.00315 |
| | Equal variances not assumed | | | .748 | 21.781 | .462 | 2.9167 | 3.89922 | -5.17453 | 11.00786 |

$$t(22) = 0.75, p = .46$$

  - We concluded that there was no evidence to suggest that recall memory was affected by sleep deprivation.

- What would happen if we tried to analyze these data using a correlation?




  - The t-test examines differences between groups by comparing the average score.

  - A correlation between group (control vs. sleep deprived) and recall would examine the linear relationship between the two variables.
    - If the is no difference between the means of the two groups, then there would be no linear relationship.
    - If there is a difference between the groups, then a linear relationship will be observed, and the greater the difference between the groups, the greater the strength of the linear relationship.
    - Thus, intuitively, it make sense the when the IV is dichotomous and the DV is continuous, then a t-test and a correlation are testing the same hypothesis using different methods.

- When we analyze these data with a correlation, we find the exact same significance value as we observed with the t-test.

```
CORRELATIONS
  /VARIABLES=group WITH memory.
```

**Correlations**

| | | GROUP |
|---|---|---|
| MEMORY | Pearson Correlation | -.157 |
| | Sig. (2-tailed) | .462 |
| | N | 24 |

Correlation: $r(22) = -.16, p = .46$
t-test: $t(22) = 0.75, p = .46$

- This is not a coincidence. When the IV/predictor is dichotomous and the DV/outcome is continuous, a t-test and a correlation will give identical tests of significance.

- Which test should you use?
  - A correlation a measure of the linear relationship between two variables (i.e., as X increases, Y increases).
  - If it makes sense to interpret your effect as a linear relationship, then a correlation is appropriate.
    - In our example, we can say that as sleep deprivation increases, recall does not significantly change. This statement is interpretable, so a correlation is appropriate.

  - If it does not make sense to interpret your effect as a linear relationship, then a correlation is not appropriate.
    - Consider a study of gender differences in recall.
    - In this example, a correlation would mean that as gender increases, the DV increased/decreased/did not change.
    - But it makes no sense to say, "as gender increased. . ."!!! Thus, a t-test would be more appropriate to explore gender differences.

  - In sum, the two approaches to analysis will lead you to identical conclusions. You should choose to present the analysis that is the easiest to interpret.

7. Factors that will limit the size of the correlation coefficient

- The reliability of X and Y

  - What is reliability?
    - The correlation between two equivalent measures of the same variable
    - The extent to which all items on a scale assess the same construct (internal consistency)

$$X = X_t + error$$

$$\text{Reliability} = \frac{Var(\text{True Score})}{Var(\text{Observed Score})} = \frac{Var(X_t)}{Var(X_t + error)}$$

    - Reliability may be interpreted as the proportion of a measure's variance that is left to correlate with other variables (because error is assumed not to correlate with anything).

  - When computing a correlation coefficient, we assume that X and Y are measured without error. If X and Y are measured with error, then it reduces the maximum correlation we can observe between those variables

$$r_{X_t Y_t} = \frac{r_{XY}}{\alpha_X * \alpha_Y}$$

    - Image you find a correlation of $r_{XY} = .44$, but each variable is measured with error: $\alpha_X = .70$ and $\alpha_Y = .80$. What correlation should you have observed between *X* and *Y*?

$$r_{X_t Y_t} = \frac{r_{XY}}{\alpha_X * \alpha_Y} = \frac{.44}{.70 * .80} = .78$$

    - Image you know the true correlation between *X* and *Y* is $\rho_{XY} = .45$, but each variable is measured with error: $\alpha_X = .73$ and $\alpha_Y = .66$. What correlation are you likely to observe between *X* and *Y*?

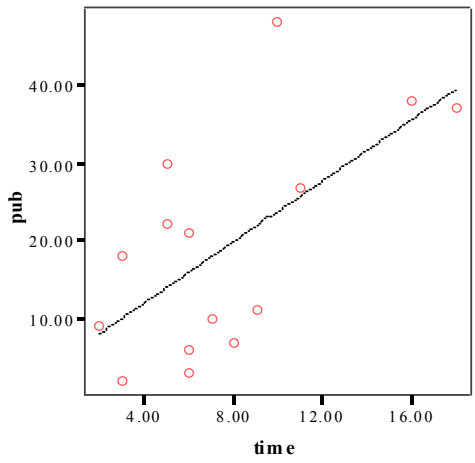
$$r_{X_t Y_t} = \frac{r_{XY}}{\alpha_X * \alpha_Y} \qquad .45 = \frac{r_{XY}}{.73 * .66} \qquad r_{XY} = .22$$

- Structural equation modeling can be used to estimate the error captured in each variable and to estimate the true correlation between two variables in the absence of error
- Unreliability of variables can result in low correlations, but it can not cause correlations to be spuriously high

- Restriction of range
  - When the range of either X or Y is restricted by the sampling procedure, the correlation between X and Y may be underestimated.

  - AN Example: Consider the relationship between time since PhD and number of publications. Data are collected for 15 professors are obtained (range of time since PhD from 3 to 18 years). In an analysis of a subset of these data, only professors with 5 to 11 years since their PhD are considered.

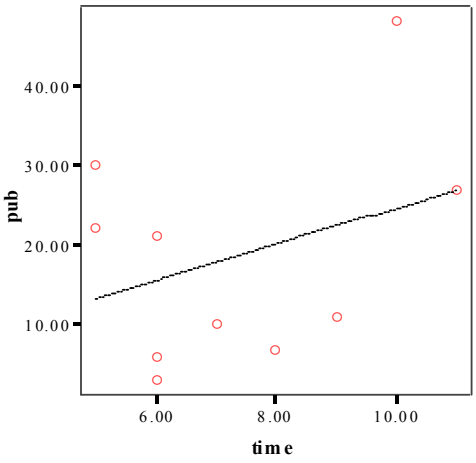
Full Range




**Correlations**

| | | pub |
|---|---|---|
| time | Pearson Correlation | .635 |
| | Sig. (2-tailed) | .011 |
| | N | 15 |

Restricted Range




**Correlations**

| | | pub |
|---|---|---|
| time | Pearson Correlation | .345 |
| | Sig. (2-tailed) | .328 |
| | N | 10 |

  - You need to be very careful in interpreting correlation coefficients with you have a limited range of values.