"Statistics is not a discipline like physics, chemistry, or biology where we study a subject to solve problems in the same subject.

We study statistics with the main aim of solving problems in other disciplines."

– C.R. Rao, Ph.D.

COHEN CHAP 9. LINEAR CORRELATION

For EDUC/PSY 6600

MOTIVATING EXAMPLE

Dr. Mortimer is interested in knowing whether people who have a ***positive view of themselves*** *in one aspect of their lives also tend to have a* ***positive view of themselves in other*** *aspects of their lives.*

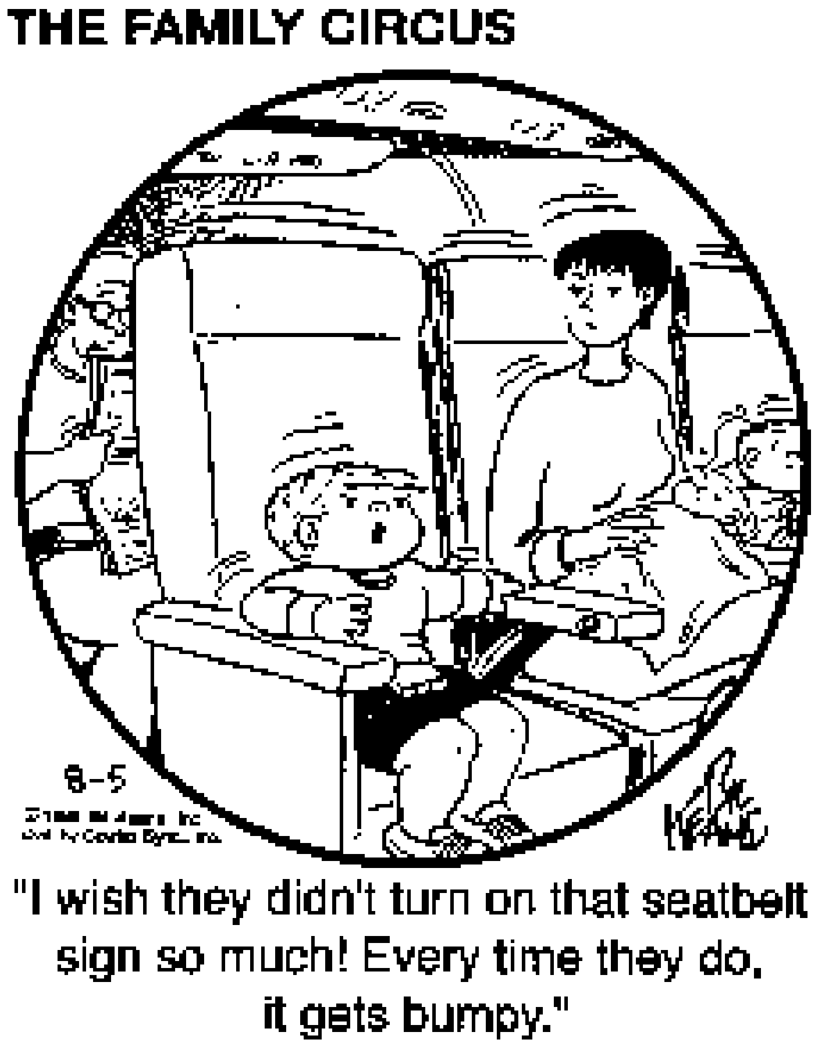
He has 80 men complete a self-concept inventory that contains 5 scales. Four scales involve questions about how competent respondents feel in the areas of intimate relationships, relationships with friends, common sense reasoning and everyday knowledge, and academic reasoning and scholarly knowledge.

The 5th scale includes items about how competent a person feels in general.

10 correlations are computed between all possible pairs of variables.

CORRELATION

- Interested in degree of covariation or co-relation among >1 variables measured on SAME objects/participants
  - Not interested in group differences, *per se*
- Variable measurements have
  - Order: Correlation
  - No order: Association or dependence
- Level of measurement for each variable determines type of correlation coefficient
- Data can be in raw or standardized format
  - Correlation coefficient is scale-invariant
- Statistical significance of correlation?
  - H_0: population correlation coefficient = 0



SCATTERPLOTS

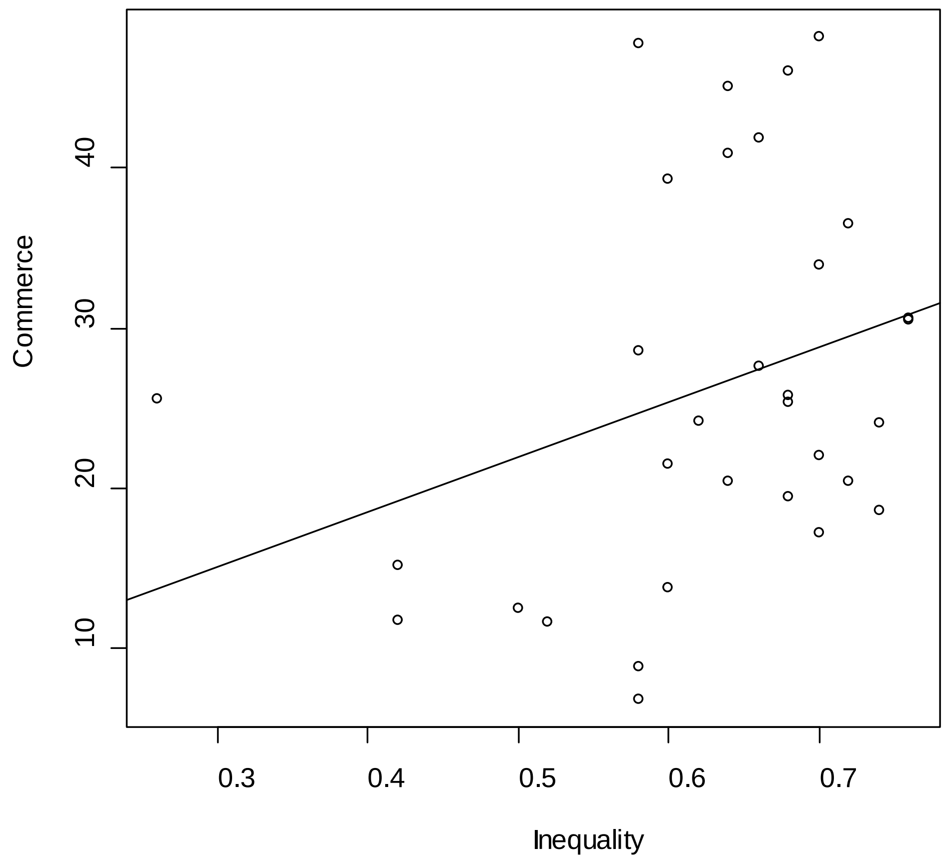
ALWAYS VISUALIZE DATA 1st

- Scatterplots, scatterdiagrams, or scattergrams
- Can stratify scatterplots by subgroups
- Each subject is represented by 1 dot
  - (x and y coordinate)

Fit line can indicate nature

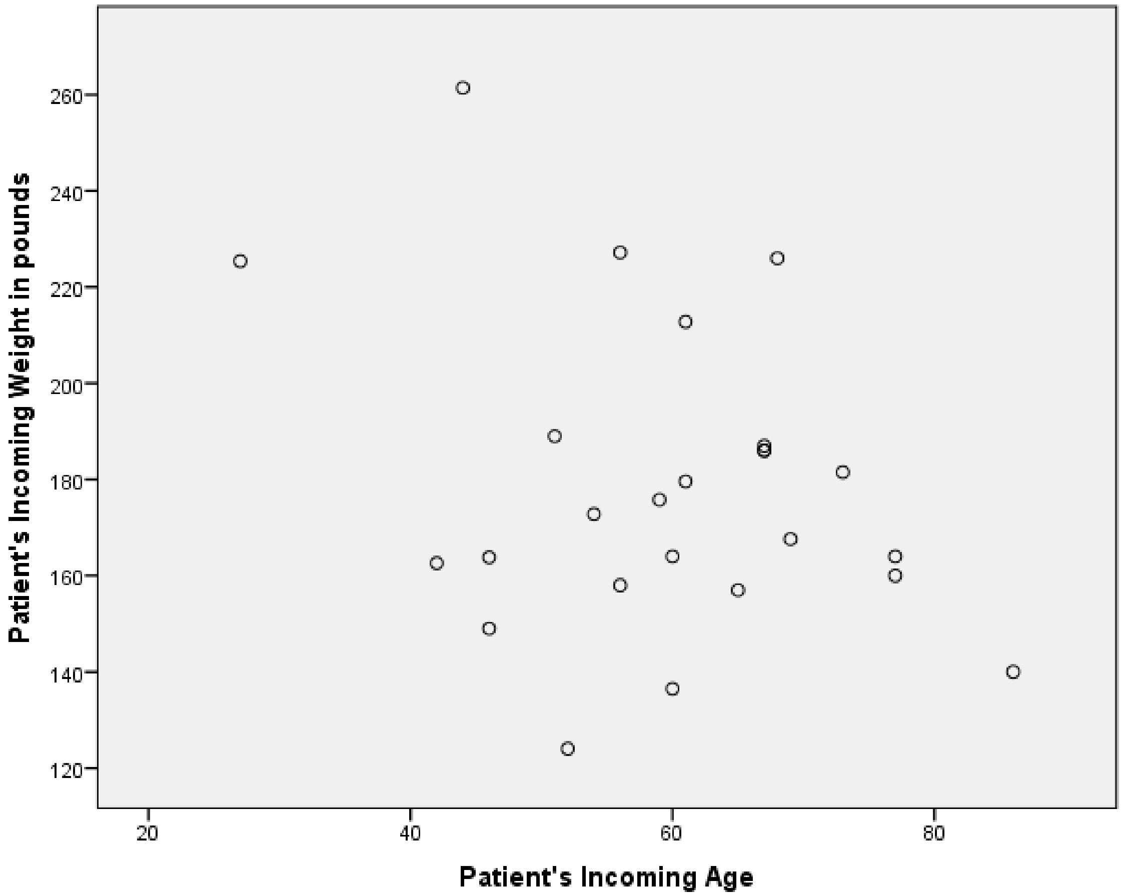
and degree of relationship

- Regression or prediction lines



SPSS: BASIC SCATTERPLOT

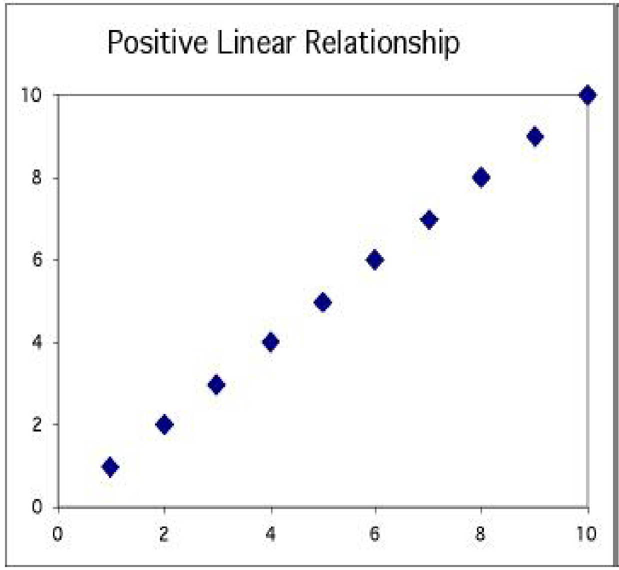
```
GRAPH /SCATTERPLOT(BIVAR)= AGE WITH WEIGHIN.
```

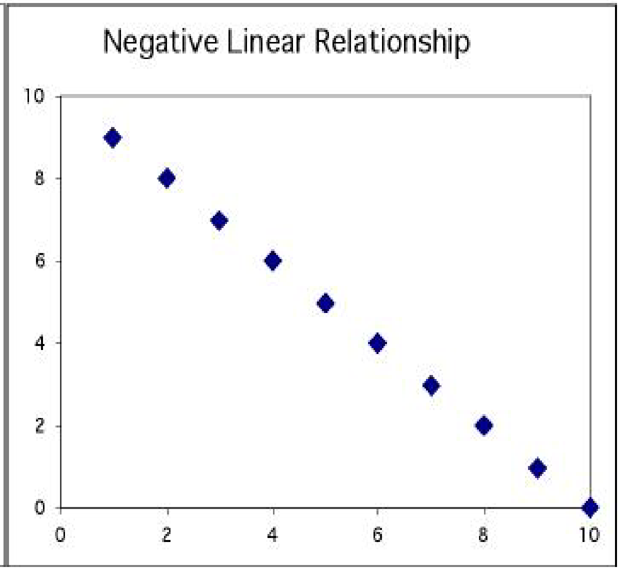


CORRELATION: DIRECTION

Positive association:
High values of one variable
tend to
occur together with
High values of the other variable

Negative association:
High values of one variable
tend to
occur together with
Low values of the other variable



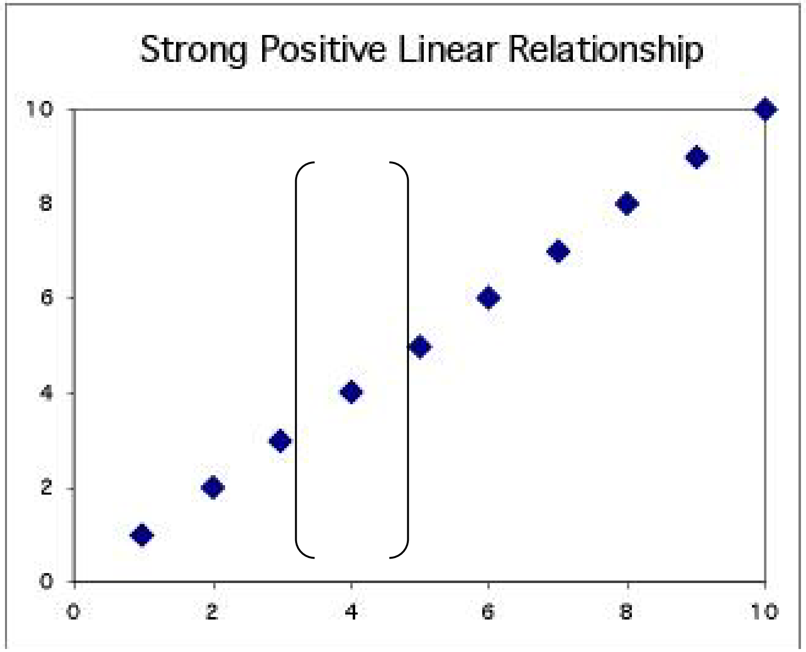


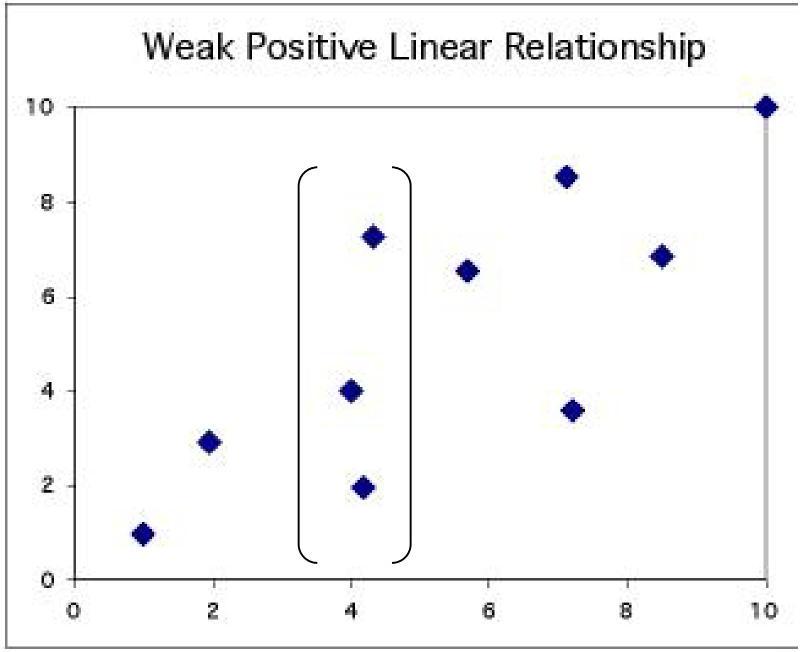
CORRELATION: STRENGTH

The **strength** of the relationship between the two variables can be seen by how much variation, or **scatter,** there is around the main form.

With a strong relationship, you can get a pretty good estimate of *y* if you know *x*.

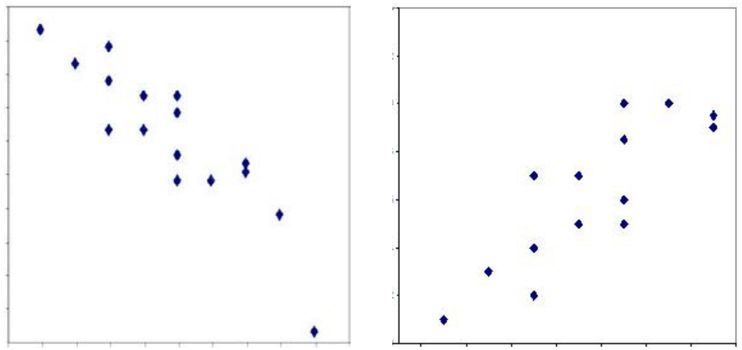
With a weak relationship, for any *x* you might get a wide range of *y* values.



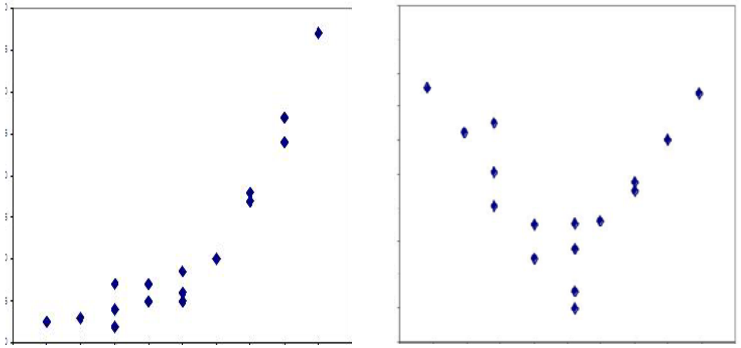


SCATTERPLOT PATTERNS

Linear



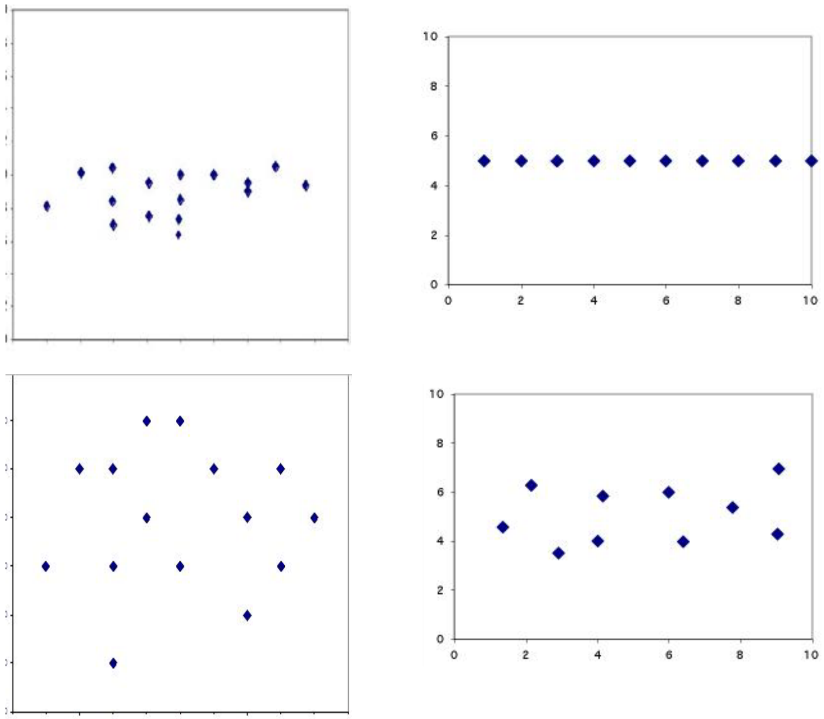
Nonlinear



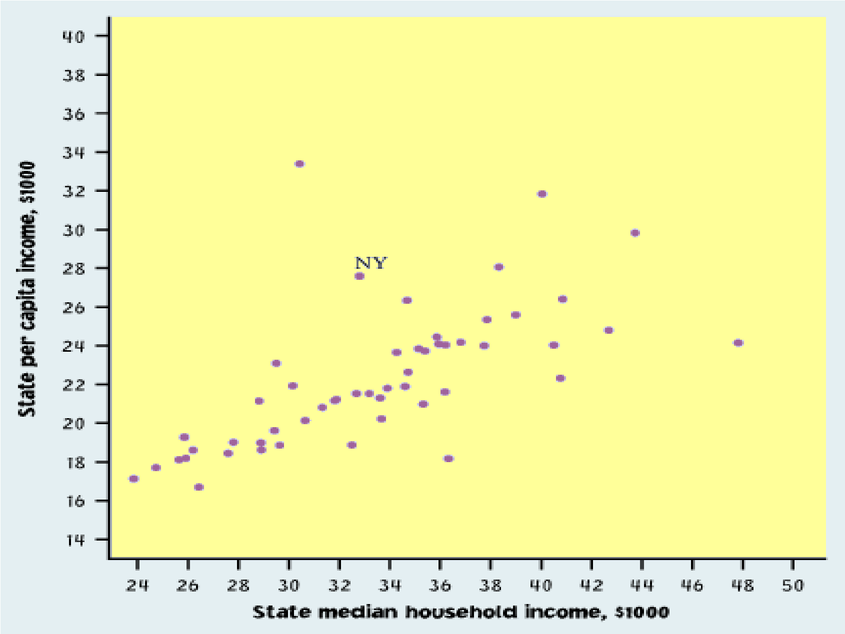
No relationship:

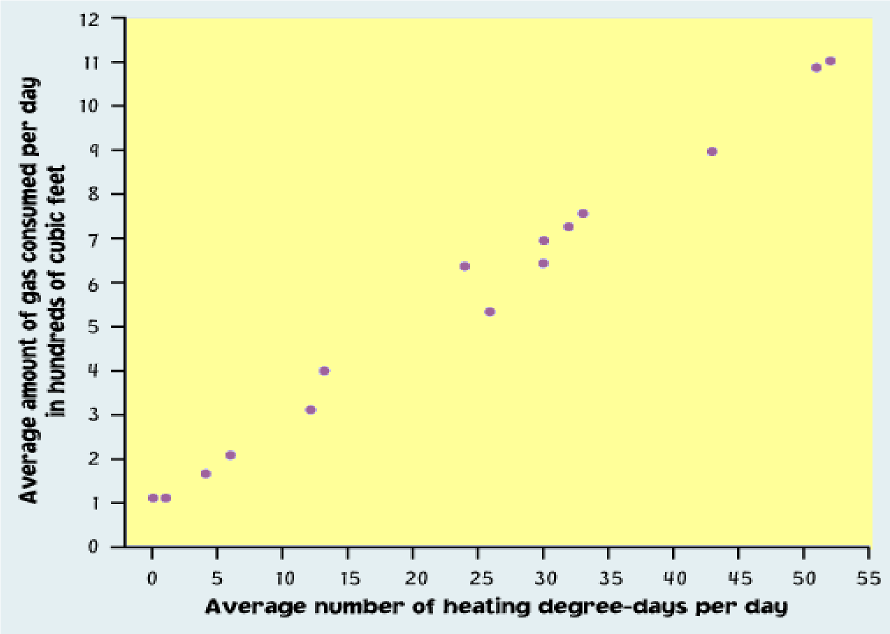
X and *Y* vary independently.

Knowing *X* tells you nothing about *Y*.



CORRELATION: EXAMPLES





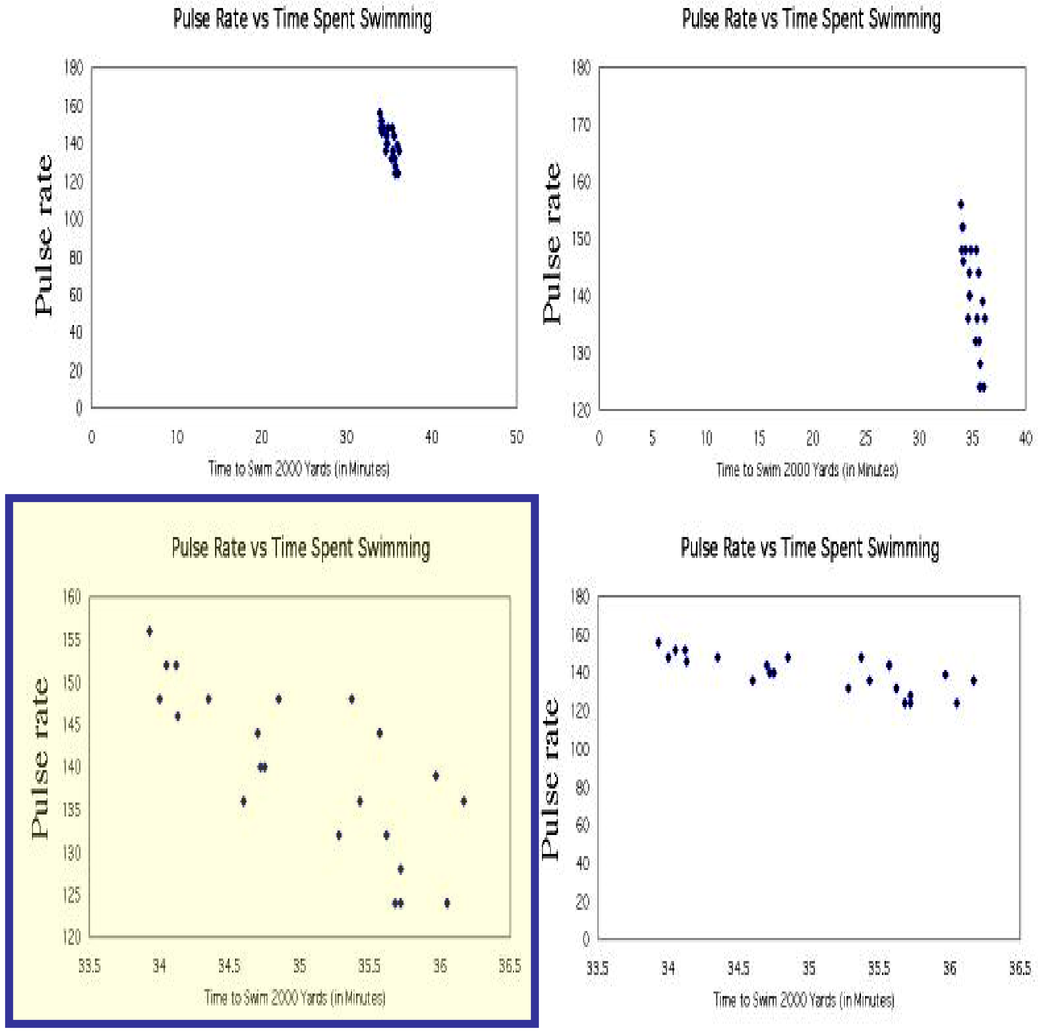
This is a **weak** relationship.

For a particular state median household income, you **can't predict** the state per capita income very well.

This is a **very strong** relationship.

The daily amount of gas consumed can be predicted **quite accurately** for a given temperature value.

SCATTERPLOT: SCALE



Same data in all four plots

Using an inappropriate scale for a scatterplot can give an incorrect impression.

Both variables should be given a similar amount of space:

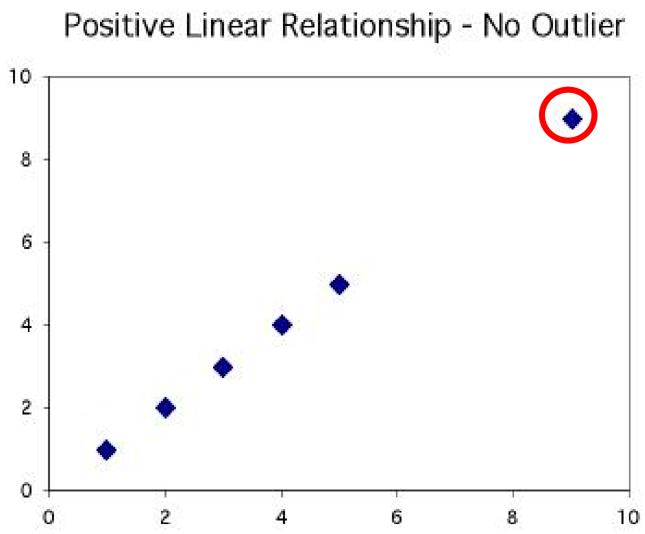
- Plot roughly square
- Points should occupy all the plot space (no blank space)

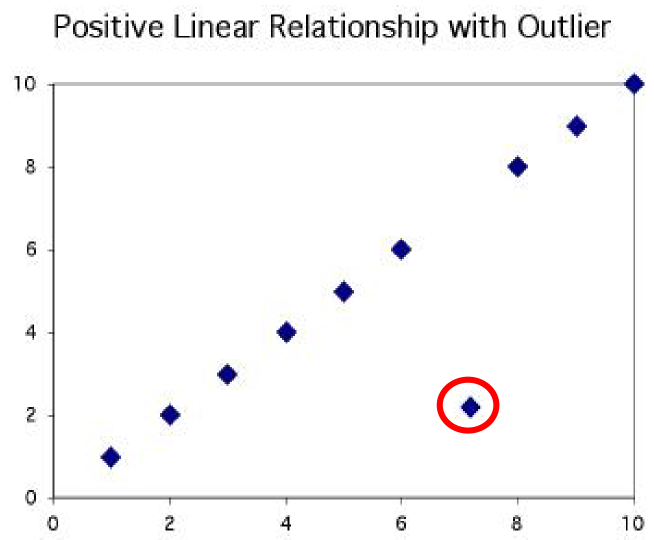
OUTLIERS

An **outlier** is a data value that has a very low probability of occurrence (i.e., it is unusual or unexpected).

In a scatterplot, BIVARIATE outliers are points that fall outside of the overall pattern of the relationship.

Not all extreme values are outliers.



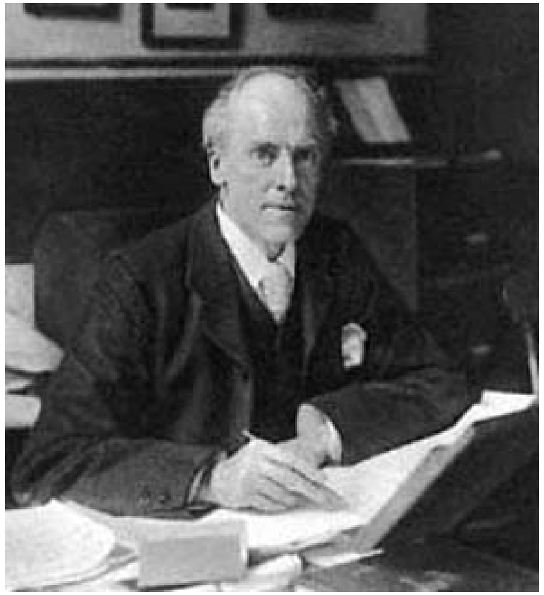


PEARSON "PRODUCT MOMENT" CORRELATION COEFFICIENT (R)

Used as a measure of

- Magnitude (strength) and direction of relationship between two continuous variables
  - Degree to which coordinates cluster around STRAIGHT regression line

- Test-retest, alternative forms, and split half reliability
- Building block for many other statistical methods



Population: ρ (rho)

Sample: r

PEARSON "PRODUCT MOMENT" CORRELATION COEFFICIENT (R)

The correlation coefficient is a measure of the direction and strength of a **linear** relationship.

It is calculated using the **mean** and the **standard deviation** of both the x and y variables.

Correlation can only be used to describe **quantitative** variables.

Categorical variables don't have means and standard deviations.

r* does not distinguish between *x* and *y

***r* has no units of measurement**

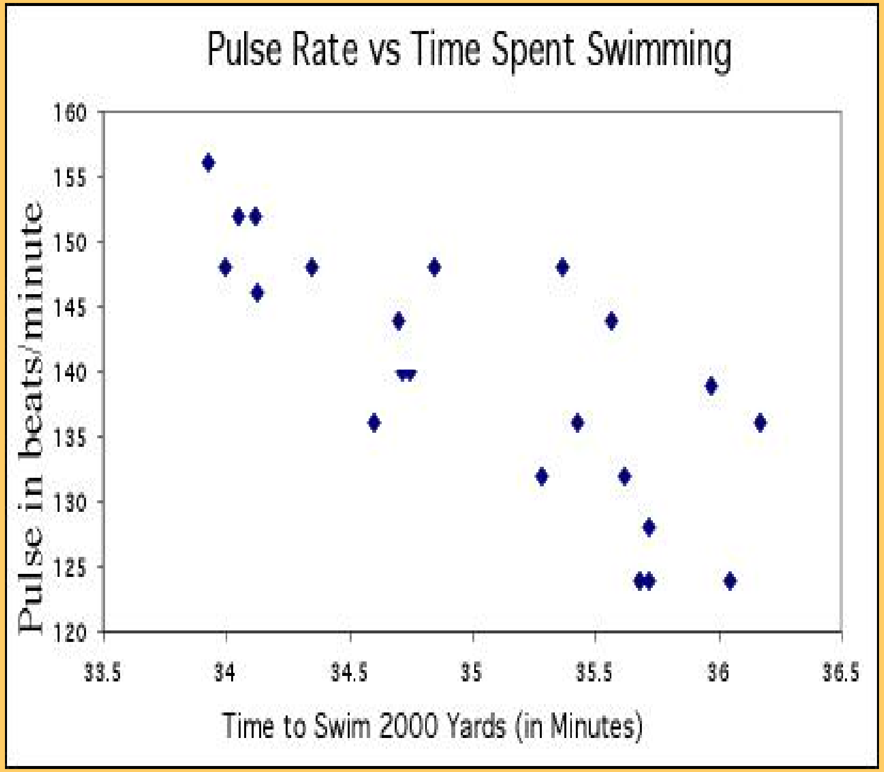
***r* ranges from -1 to +1**

Influential points…can change 'r' a great deal!

CORRELATION: CALCULATING

Time to swim: x-bar = 35, s_x = 0.7

Pulse rate: x-bar = 140 s_y = 9.5



How to find "r"?

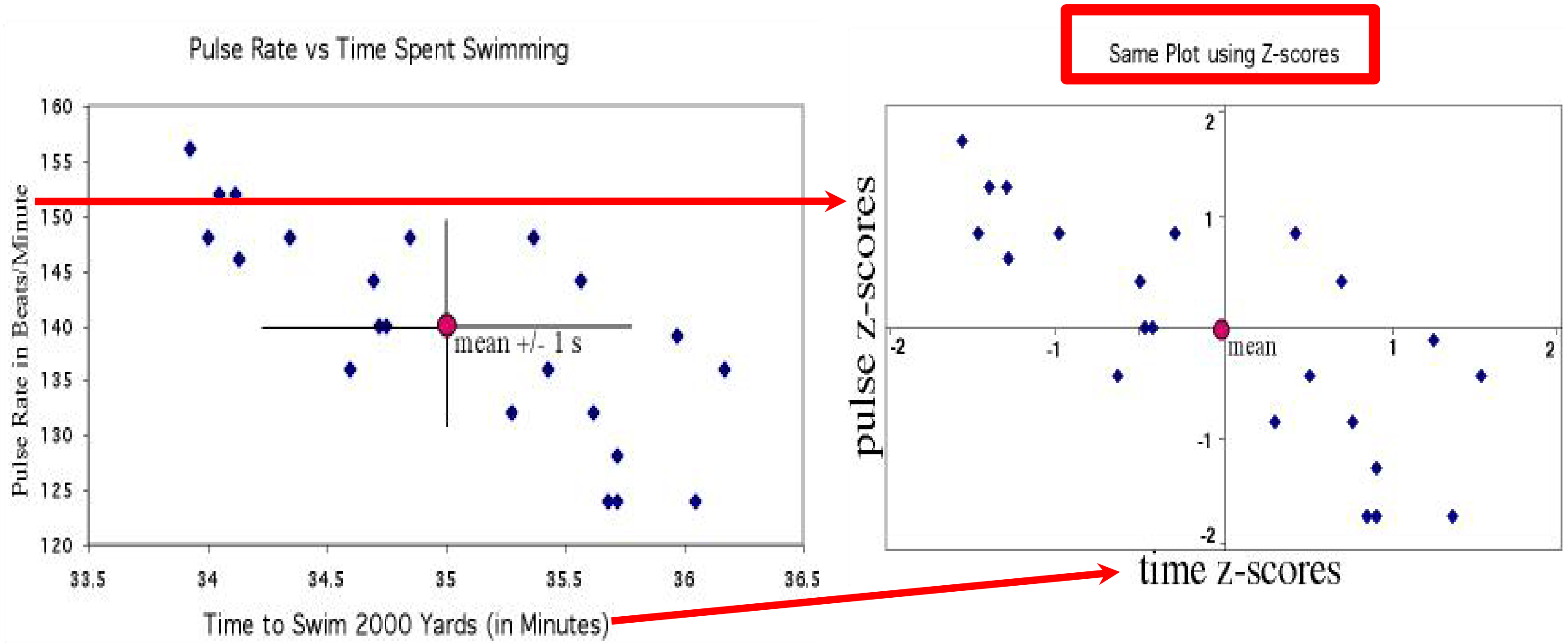
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

You DON'T want to do this by hand!

Make sure you learn how to use your calculator or software.

CORRELATION: CALCULATING

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$



Standardization:

Allows us to compare correlations between data sets where variables are measured in different units or when variables are different.

For instance, we might want to compare the correlation between [swim time and pulse], with the correlation between [swim time and breathing rate].

SPSS: CORRELATION - BASIC

```
* two variables only.

CORRELATIONS AGE WEIGHIN.
```

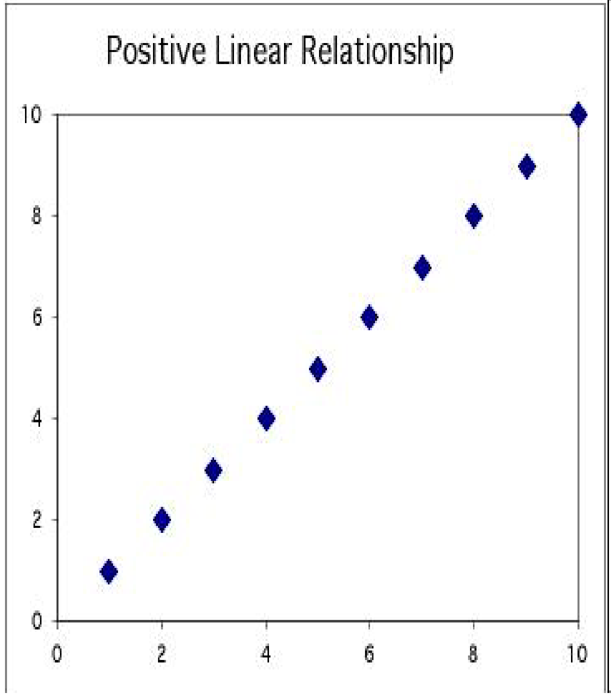
Correlations

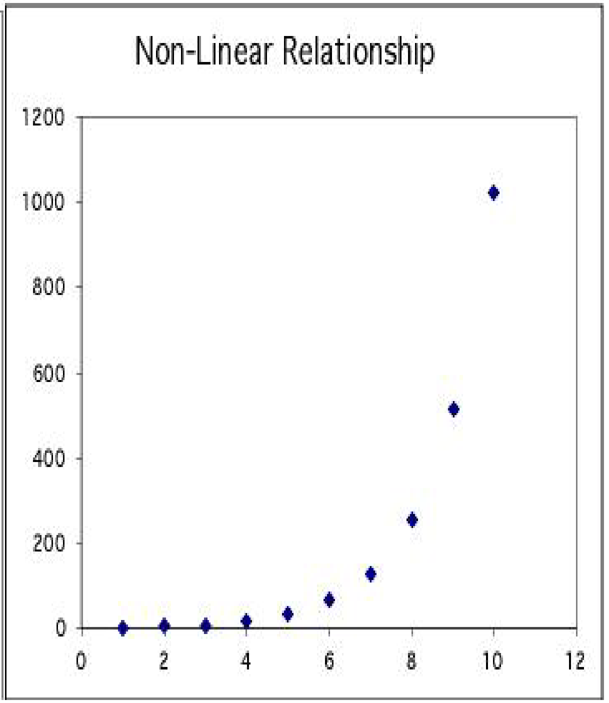
		AGE Patient's Incoming Age	WEIGHIN Patient's Incoming Weight in pounds
AGE Patient's Incoming Age	Pearson Correlation	1	-.288
	Sig. (2-tailed)		.163
	N	25	25
WEIGHIN Patient's Incoming Weight in pounds	Pearson Correlation	-.288	1
	Sig. (2-tailed)	.163	
	N	25	25

CORRELATION: RELATIONSHIP FORM

Correlation only describes linear relationships

No matter how strong the association, *r* does not describe curved relationships.



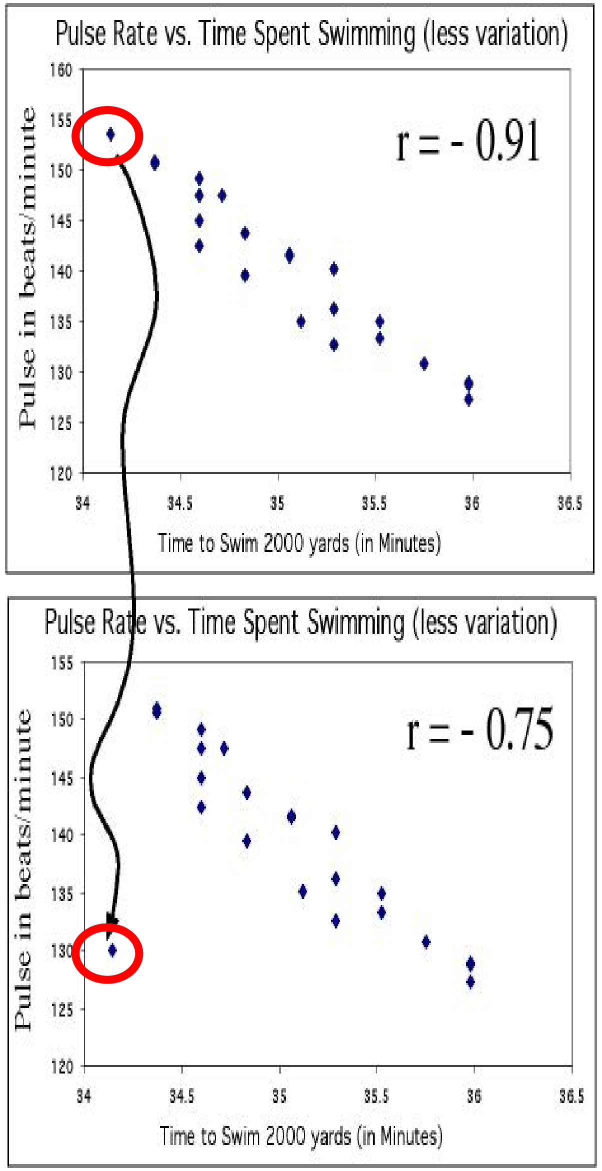


Note: You can sometimes transform a non-linear association to a linear form, for instance by taking the logarithm. You can then calculate a correlation using the transformed data.

CORRELATION: INFLUENTIAL POINTS

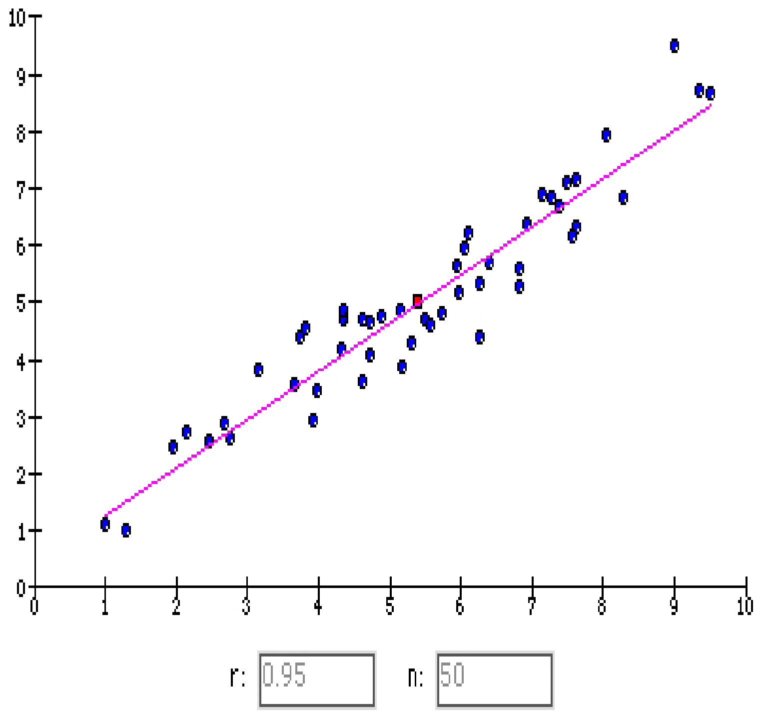
Correlations are calculated using **means** and **standard deviations**, and thus are **NOT** resistant to outliers.

Just moving one point away from the general trend here decreases the correlation from -0.91 to -0.75

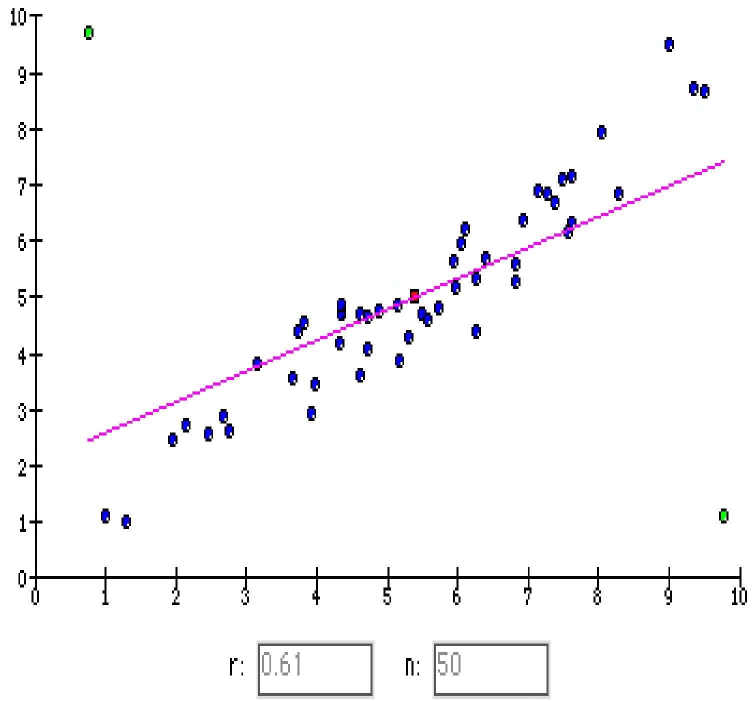


CORRELATION: INFLUENTIAL POINTS

http://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html



Adding two outliers decreases *r* from 0.95 to 0.61.

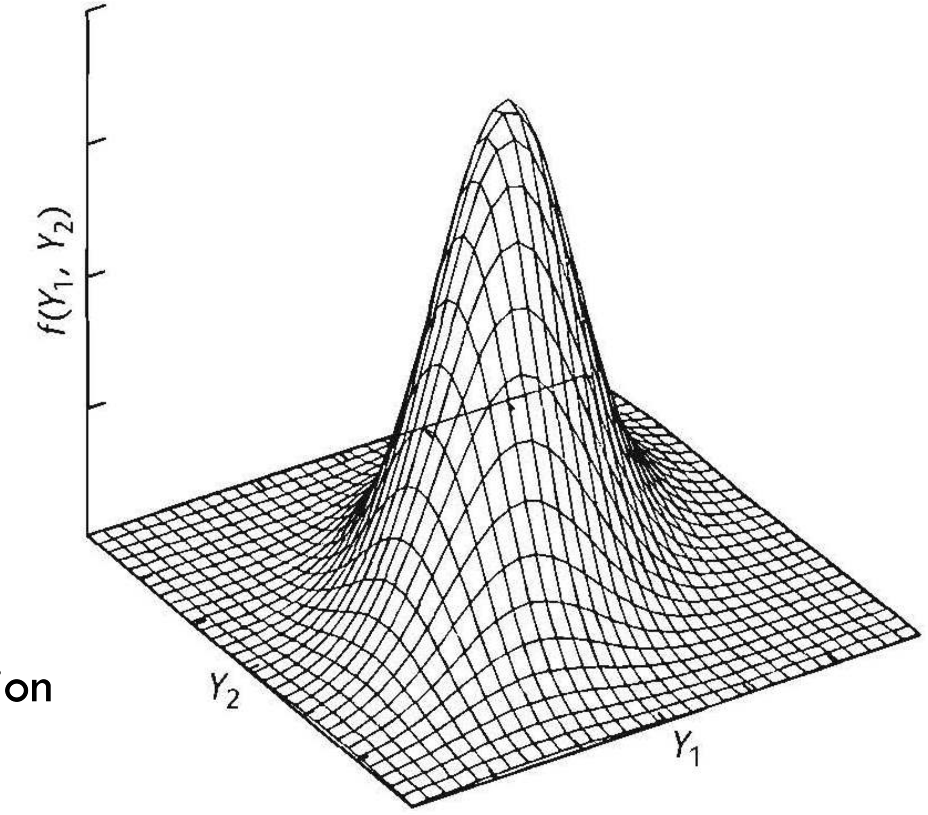


ASSUMPTIONS

- Random sample

- Relationship between variables is linear
  - Check scatterplot
  - Transform data or use alternative methods

- Bivariate normal distribution
  - Each variable should be normally distributed in population
  - Joint distribution should be bivariate normal
  - Curvilinear relationships = violation
  - Less important as *N* increases



SAMPLING DISTRIBUTION OF "RHO"

Normal distribution about 0

Becomes non-normal as ρ gets larger and deviates from H_0 value of 0 in the population

- Negatively skewed with large, positive null hypothesized ρ
- Positively skewed with large, negative null hypothesized ρ

Leads to

- Inaccurate p-values
- No longer testing H_0 that $\rho = 0$

Fisher's solution: transform sample r coefficients to yield normal sampling distribution, regardless of ρ

- We will let the computer worry about the details…

HYPOTHESIS TESTING FOR 1-SAMPLE "R"

H_0: $\rho = 0$

H_1: $\rho \neq 0$ (2-tailed)

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

r converted into t-statistic

- No r transformation as ρ is at center (0)

Compare to t-distribution with $df = N - 2$

- Rejection: statistical evidence for co-relationship
- Or, see table of critical values for r

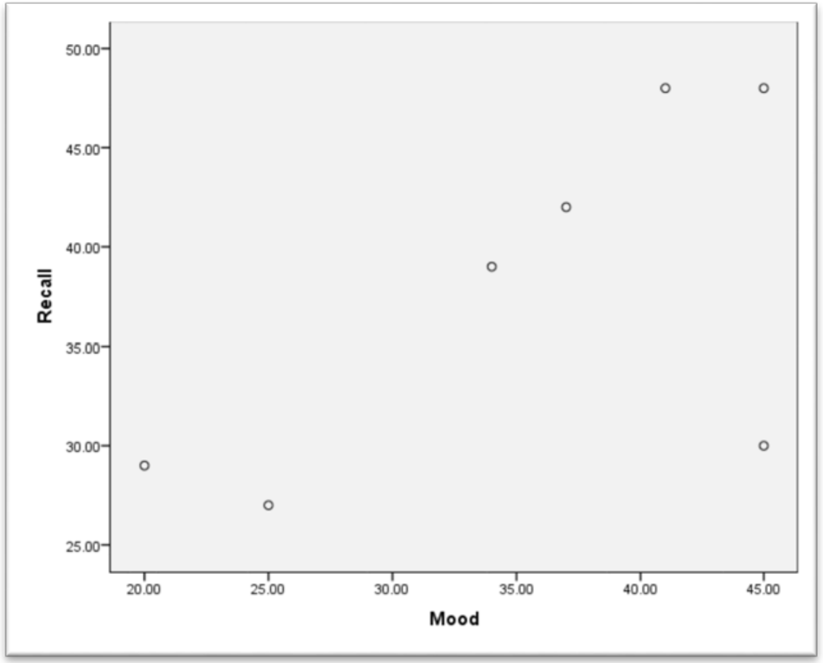
	Levels of Significance for a One-Tailed Test			
	.05	.025	.01	.005
	Levels of Significance for a Two-Tailed Test			
df	**.10**	**.05**	**.02**	**.01**
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.755	.833	.875
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.498	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.533	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.623
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.529	.575
18	.379	.444	.516	.562
19	.369	.433	.503	.549

EXAMPLE

Researcher wishes to correlate scores from 2 tests: current mood state and verbal recall memory

Compute r, test for significance (H_0: $\rho = 0$), construct 95% CI

Pearson Correlation	.638
Sig. (2-tailed)	.123
N	7



Mood	Recall
45	48
34	39
41	48
25	27
38	42
20	29
45	30

POWER

N necessary to reject H_0 given an effect ρ

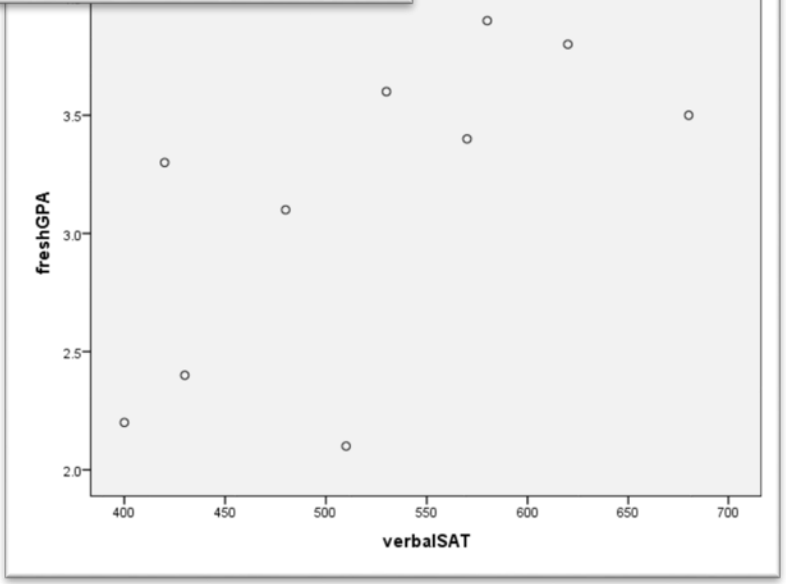
- Determine d (value of ρ [or r] to detect as significant)
- Determine delta (δ) (value from **appendix A.4** that would result in given level of power at $\alpha = .05$)
- Solve: $$\left(\frac{\delta}{\mathrm{d}}\right)^2 + 1 = N$$

Example: Chap9A #8

A college admissions officer is tracking the relationship between students' verbal SAT scores and their first-year grade point average (GPA).

Verbal SAT	GPA
510	2.1
620	3.8
400	2.2
480	3.1
580	3.9
430	2.4
530	3.6
680	3.5
420	3.3
570	3.4

Pearson Correlation	.661
Sig. (2-tailed)	.038



Based on this pilot data, how many students should I plan to study to ensure I have at least 95% power for an alpha = .01, one-tailed test?

FACTORS AFFECTING VALIDITY OF R

Range restriction (variance of X and/or Y)

- r can be inflated or deflated
  - May be related to small N

Outliers

- r can be heavily influenced

Use of heterogeneous subsamples

- Combining data from heterogeneous groups can inflate correlation coefficient or yield spurious results by stretching out data

INTERPRETATION & COMMUNICATION

Correlation ≠ Causation

Can infer strength and direction; not form or prediction from r

- Can say that prediction will be better with large r, but cannot predict actual values

Statistical significance

- p-value heavily influenced by N
- Need to interpret size of r-statistic, more than p-value

APA format

- $r(18) = -.74, p < .01$

APA STYLE REPORTING

Correlations: Correlations provide a measure of statistical relationship between two variables. Note that correlations can be tested for statistical significance (and that this information should be summarized if it is available and of interest).

For the nine students, the scores on the first quiz (M = 7.00, SD = 1.23) and the first exam (M = 80.89, SD = 6.90) were strongly and significantly correlated, $r(8) = .70$, $p = .038$.

"A Pearson product-moment correlation coefficient was computed to assess the relationship between the amount of water that one consumed and rating of skin elasticity. There was a positive correlation between the two variables, r(5) = 0.985, p = 0.002. A scatterplot summarizes the results (Figure 1) Overall, there was a strong, positive correlation between water consumption and skin elasticity. Increases in water consumption were correlated with increases in rating of skin elasticity."

Table 3. Correlation coefficients values (Spearman's rho) between demographic variables, psychopathology, and neuroimaging parameters of the whole sample.

	Age	Age of onset	Duration	Positive symptoms	Negative symptoms	Desorganization symptoms	PFAI	VBR
Age								
Age of onset	0.82**							
Duration	0.24	–0.26						
Positive symptoms	0.85*	0.72	–0.01					
Negative symptoms	–0.53	–0.32	–0.07	–0.70				
Desorganization symptoms	–0.69	–0.63	0.21	–0.79*	0.84*			
PFAI	0.31	0.35	–0.07	0.46	–0.14	–0.34		
VBR	0.07	0.07	–0.13	0.005	0.50	0.10	0.26	

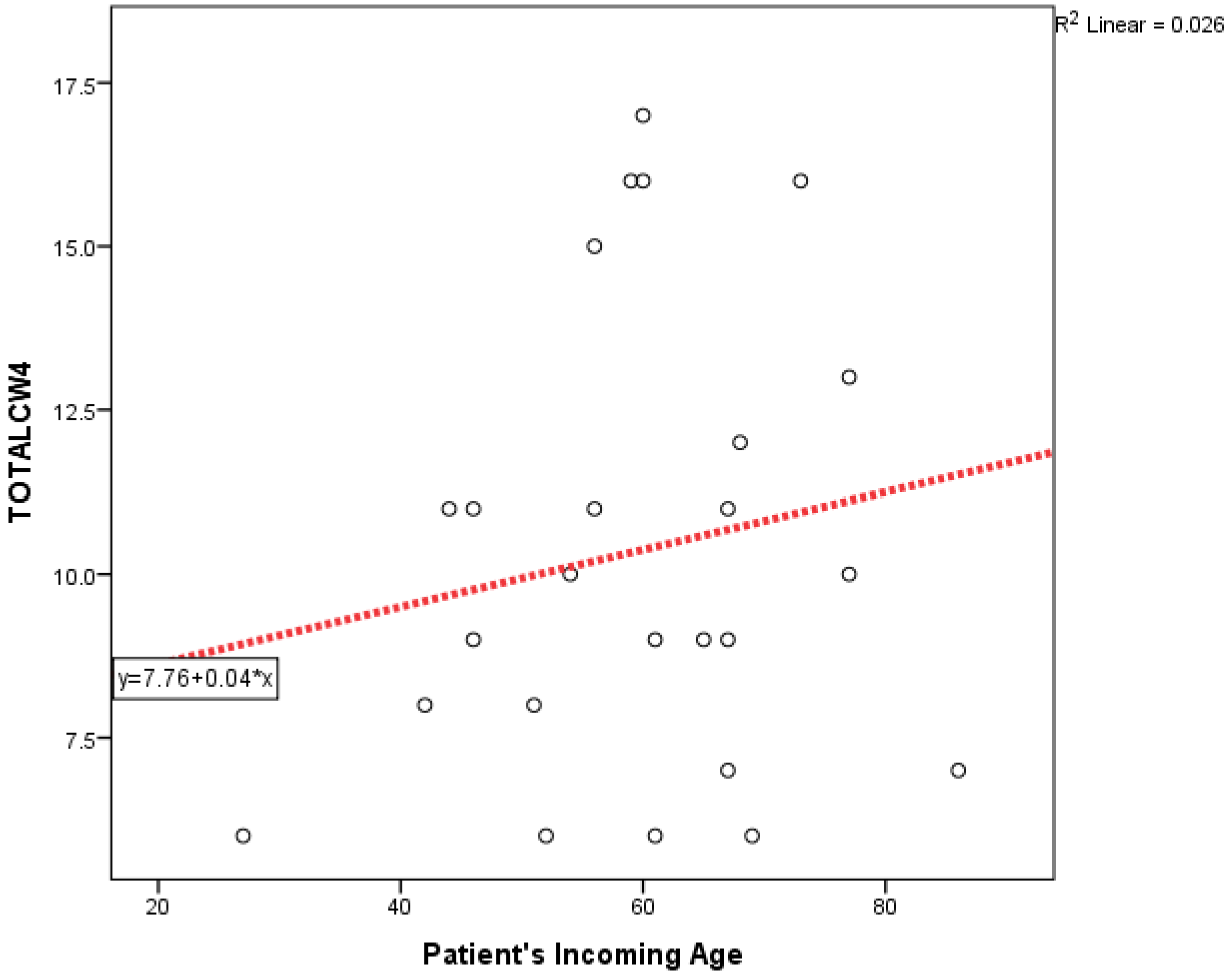
VBR, ventricle to brain ratio; PFAI, pre-frontal sulcal prominence index.
Correlation coefficients that reached significance are displayed in bold. *The level of significance (p<0.01) was obtained after Bonferroni adjustment (0.05/64=0.0008).

SPSS: SCATTERPLOT – ADD REGRESSION LINE

```
GRAPH /SCATTERPLOT(BIVAR)= AGE WITH TOTALCW4.
```

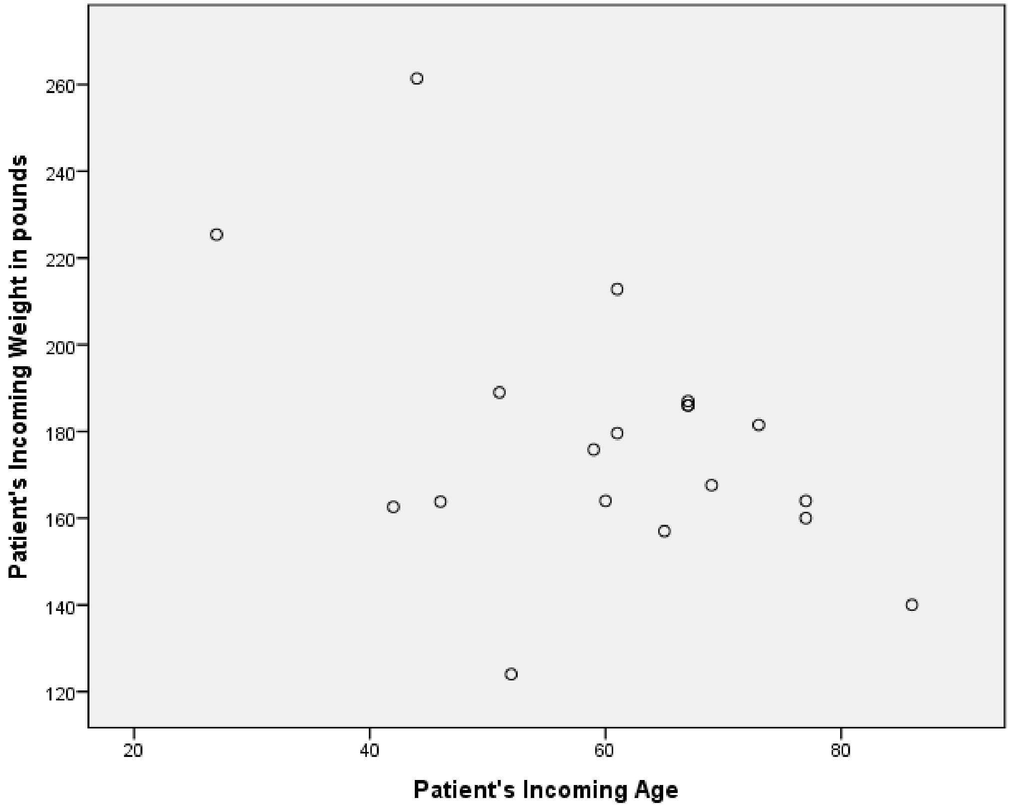
To add a regression line:

1. Double-click on the graph in the output window
2. At the top of the newly opened "chart editor" window with the plot, click the word "Elements" and select the phrase "Fit Line at Total"
3. The default is a "LINEAR" fit method, so you can click the "Close" button and be done, or you can play with the other options in the various tabs.
4. When you done making changes, click the Red "x" or "dot" in the upper right corner of the "Chart Editor" to paste your edited plot back in the Output Window.



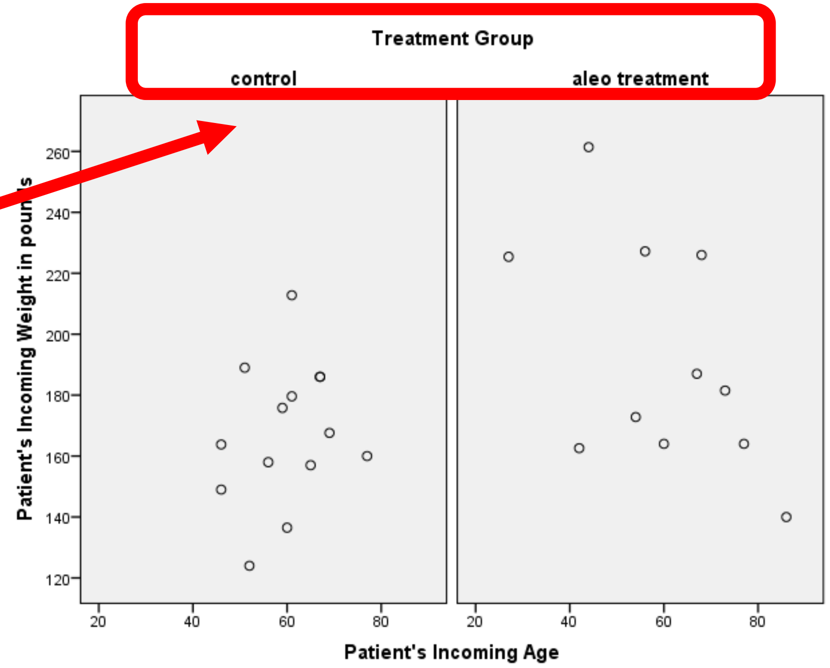
SPSS: SCATTERPLOTS – SELECT CASES

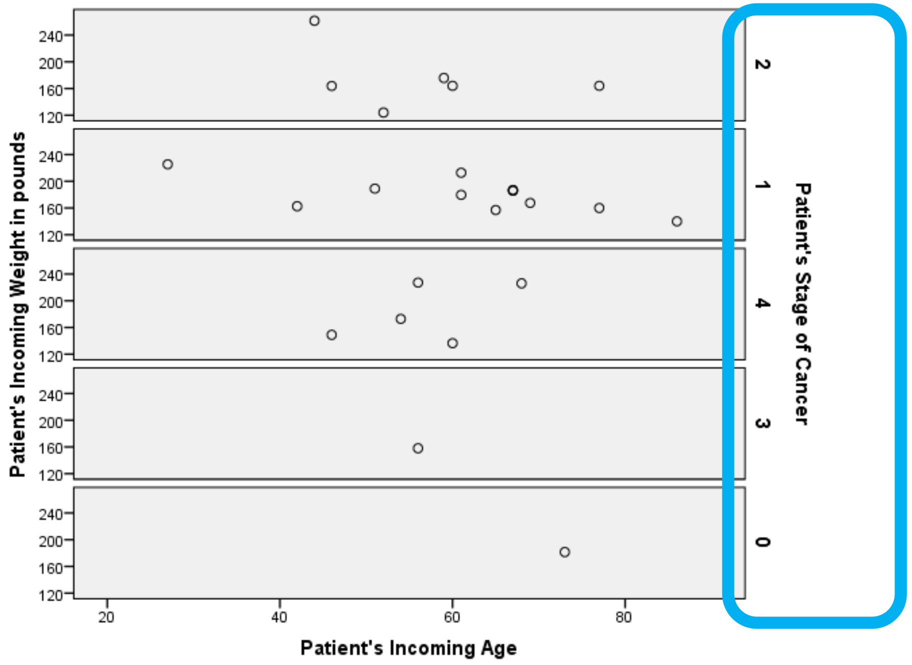
```
* use "SELECT IF" to restrict to only cases in stage 0, 1 or 2.

TEMPORARY.
SELECT IF STAGE <= 2.
GRAPH /SCATTERPLOT(BIVAR)= AGE WITH WEIGHIN.
```



SPSS: SCATTERPLOTS – PANELS

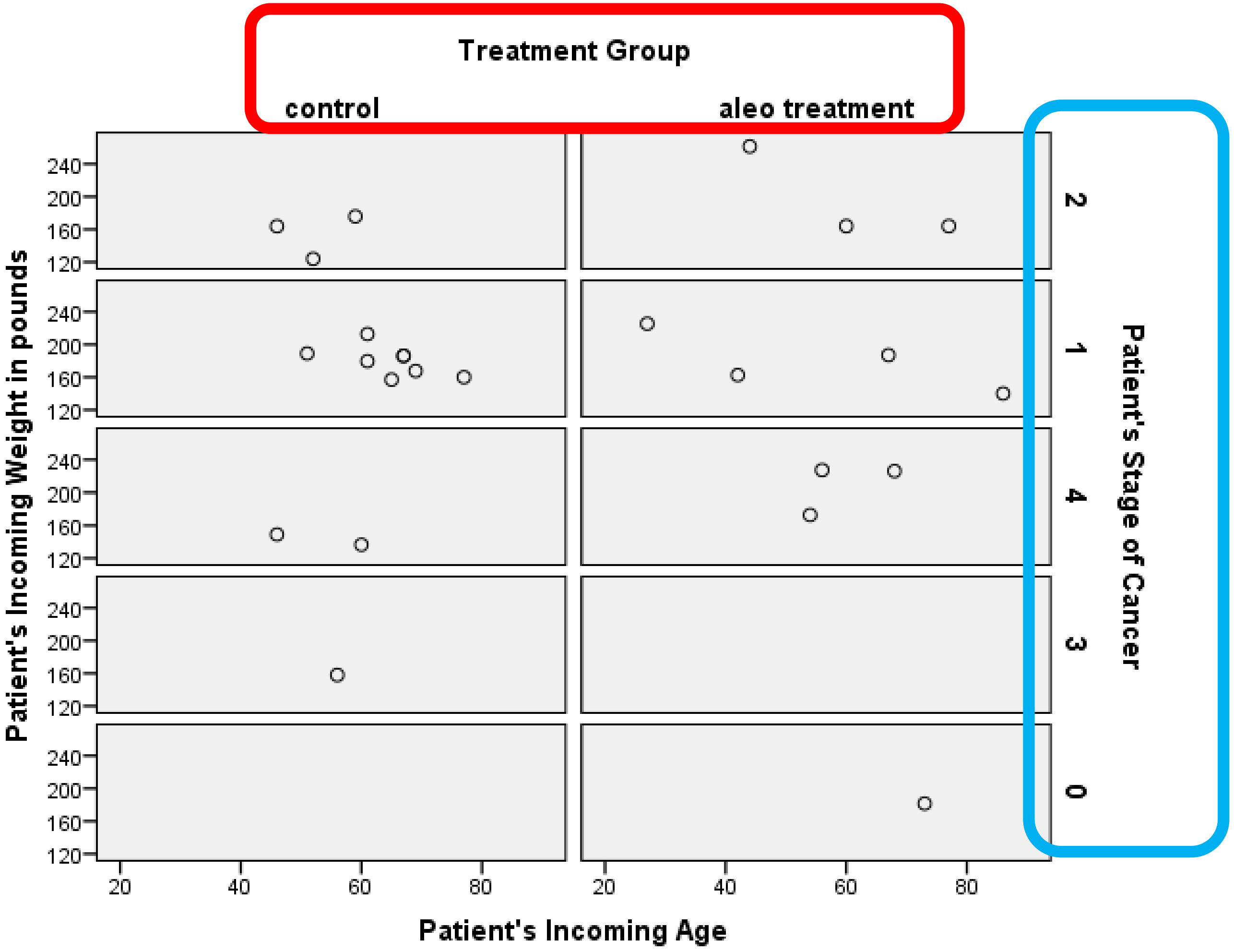
```
* seperate treatment/control into panels (columns).

GRAPH
  /SCATTERPLOT(BIVAR)=AGE WITH WEIGHIN
  /PANEL COLVAR=TRT.


* seperate stage into panels (rows).

GRAPH
  /SCATTERPLOT(BIVAR)=AGE WITH WEIGHIN
  /PANEL ROWVAR=STAGE.
```





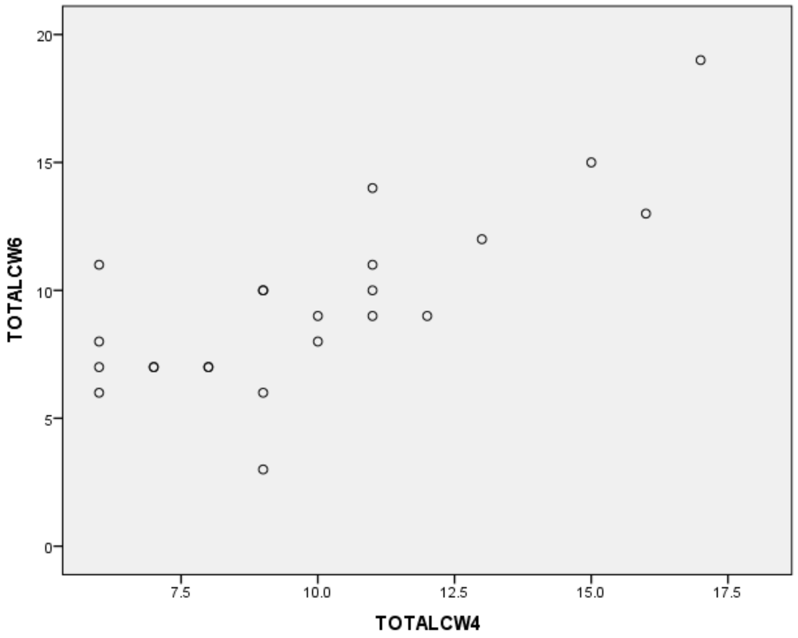
SPSS: SCATTERPLOTS – PANELS

```
* seperate treatment/control into panels (columns)
  AND stage into panels (rows).

GRAPH
  /SCATTERPLOT(BIVAR)=AGE WITH WEIGHIN
  /PANEL COLVAR=TRT ROWVAR=STAGE
```



SPSS: CORRELATION – MATRIX (SYMMETRICAL)

```
* more than two variables = "correlation matrix".

CORRELATIONS TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6.
```



Correlations

		TOTALCIN	TOTALCW2	TOTALCW4	TOTALCW6
TOTALCIN	Pearson Correlation	1	.314	.222	.098
	Sig. (2-tailed)		.126	.287	.657
	N	25	25	25	23
TOTALCW2	Pearson Correlation	.314	1	.337	.378
	Sig. (2-tailed)	.126		.099	.075
	N	25	25	25	23
TOTALCW4	Pearson Correlation	.222	.337	1	.763
	Sig. (2-tailed)	.287	.099		.000
	N	25	25	25	23
TOTALCW6	Pearson Correlation	.098	.378	.763	1
	Sig. (2-tailed)	.657	.075	.000	
	N	23	23	23	23

SPSS: CORRELATION – "WITH" OPTION

```
* Use the "with" option to create a smaller matrix.

CORRELATIONS
  /VARIABLES = TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6 WITH AGE WEIGHIN.
```

Correlations

		AGE Patient's Incoming Age	WEIGHIN Patient's Incoming Weight in pounds
TOTALCIN	Pearson Correlation	.256	.170
	Sig. (2-tailed)	.217	.418
	N	25	25
TOTALCW2	Pearson Correlation	-.106	.274
	Sig. (2-tailed)	.615	.185
	N	25	25
TOTALCW4	Pearson Correlation	.162	-.095
	Sig. (2-tailed)	.438	.651
	N	25	25
TOTALCW6	Pearson Correlation	.030	-.078
	Sig. (2-tailed)	.891	.725
	N	23	23

SPSS: CORRELATION – W/ "SELECT IF"

```
* use "SELECT IF" to restrict to only cases in stage 3 & 4.

TEMPORARY.
SELECT IF STAGE = 3 or STAGE = 4.
CORRELATIONS TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6.

* use "SELECT IF" to restrict to only cases in stage 0, 1 or 2.

TEMPORARY.
SELECT IF STAGE <= 2.
CORRELATIONS AGE WEIGHIN.
```

SPSS: CORRELATION – "SPLIT FILE BY"

```
* use "SPLIT FILE" to calculate on subgroups.

SORT CASES by TRT.
TEMPORARY.
SPLIT FILE by TRT.
CORRELATIONS AGE WEIGHIN.
SPLIT FILE off.
SORT CASES by ID.
```

Correlations

TRT Treatment Group			AGE Patient's Incoming Age	WEIGHIN Patient's Incoming Weight in pounds
0 control	AGE Patient's Incoming Age	Pearson Correlation	1	.235
		Sig. (2-tailed)		.418
		N	14	14
	WEIGHIN Patient's Incoming Weight in pounds	Pearson Correlation	.235	1
		Sig. (2-tailed)	.418	
		N	14	14
1 aleo treatment	AGE Patient's Incoming Age	Pearson Correlation	1	-.534
		Sig. (2-tailed)		.091
		N	11	11
	WEIGHIN Patient's Incoming Weight in pounds	Pearson Correlation	-.534	1
		Sig. (2-tailed)	.091	
		N	11	11