Correlation

Cohen Chapter 9

EDUC/PSY 6600

"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.

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

Correlation

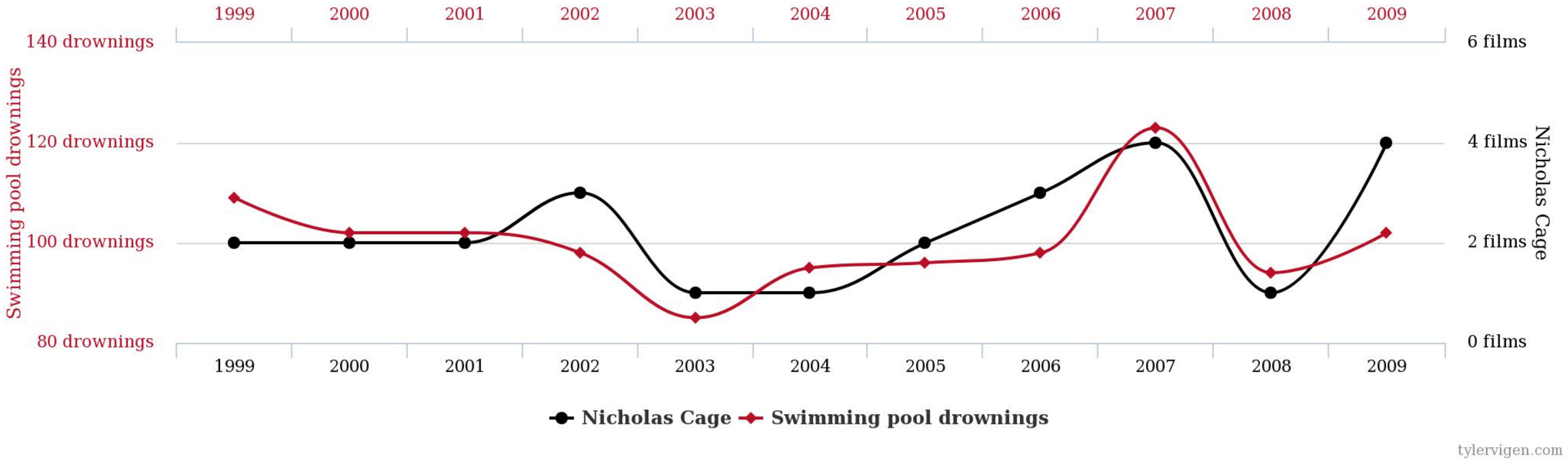
- Interested in **degree** of covariation or co-relation among >1 variables measured on SAME objects/participants
  - Not interested in group differences, per se
- Variable measurements have:
  - Order: Correlation
  - No order: Association or dependence

- 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

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



http://www.tylervigen.com/spurious-correlations

Always **Visualize** Data First

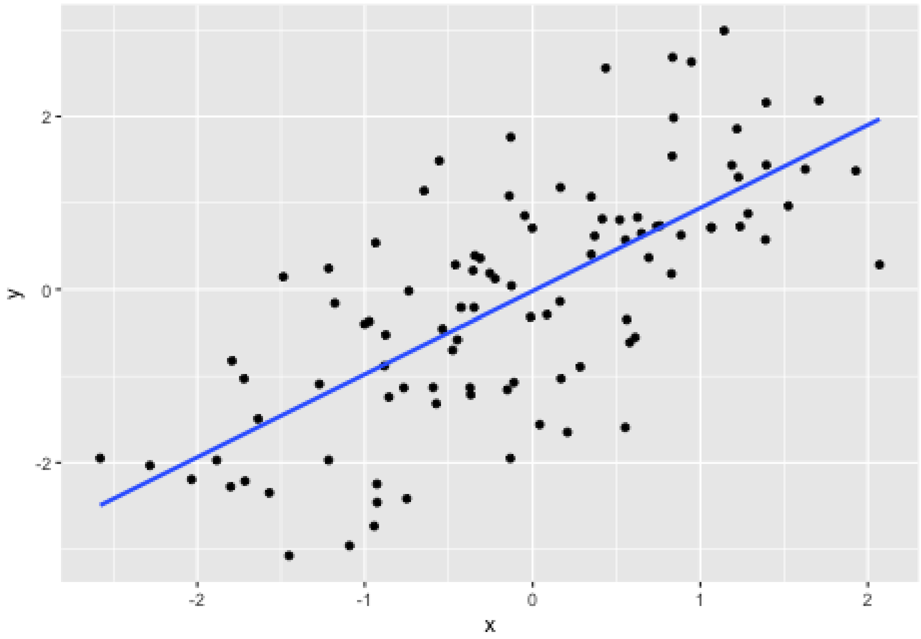
Scatterplots

Aka: scatterdiagrams, scattergrams

Notes:

1. Can stratify scatterplots by subgroups
2. Each subject is represented by 1 dot (x and y coordinate)
3. Fit line can indicate nature and degree of relationship (Regression or prediction lines)

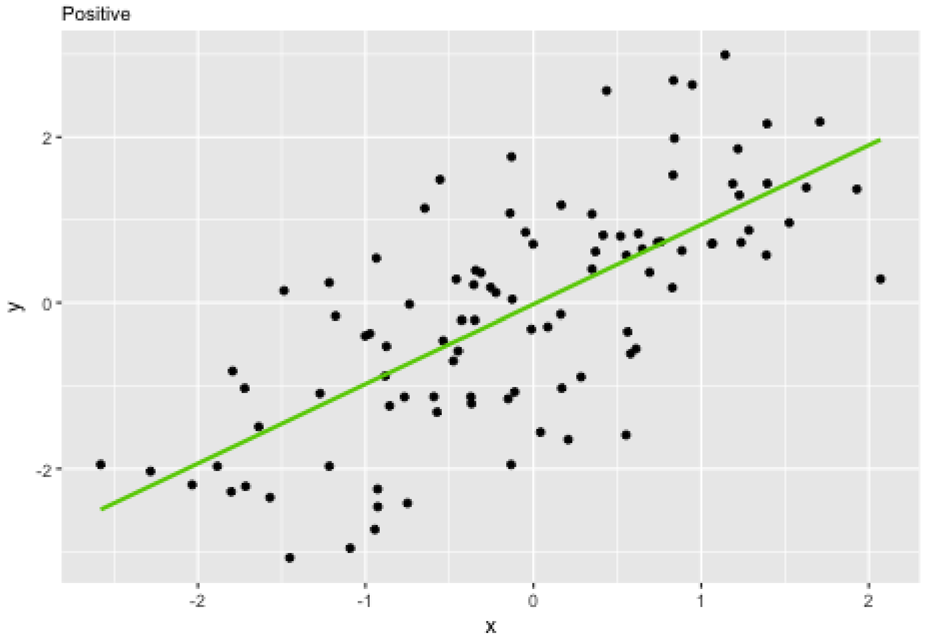
```r
library(tidyverse)
df %>%
  ggplot(aes(x, y)) +
    geom_point() +
    geom_smooth(se = FALSE,
                method = "lm")
```



Correlation: Direction

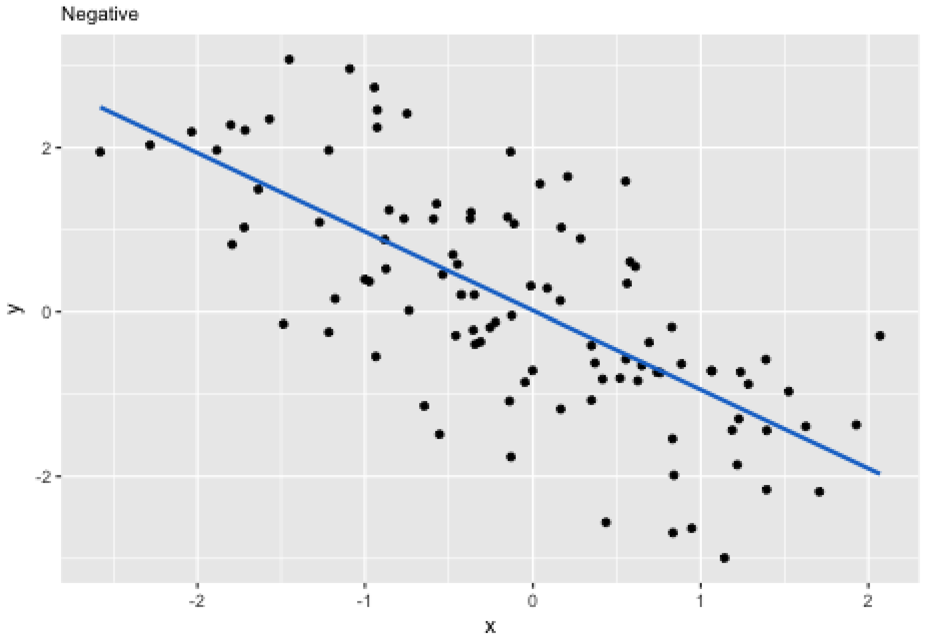
Positive Association

High values of one variable tend to occur with **High values** of the other



Negative Association

High values of one variable tend to occur with **Low values** of the other



Correlation: Strength

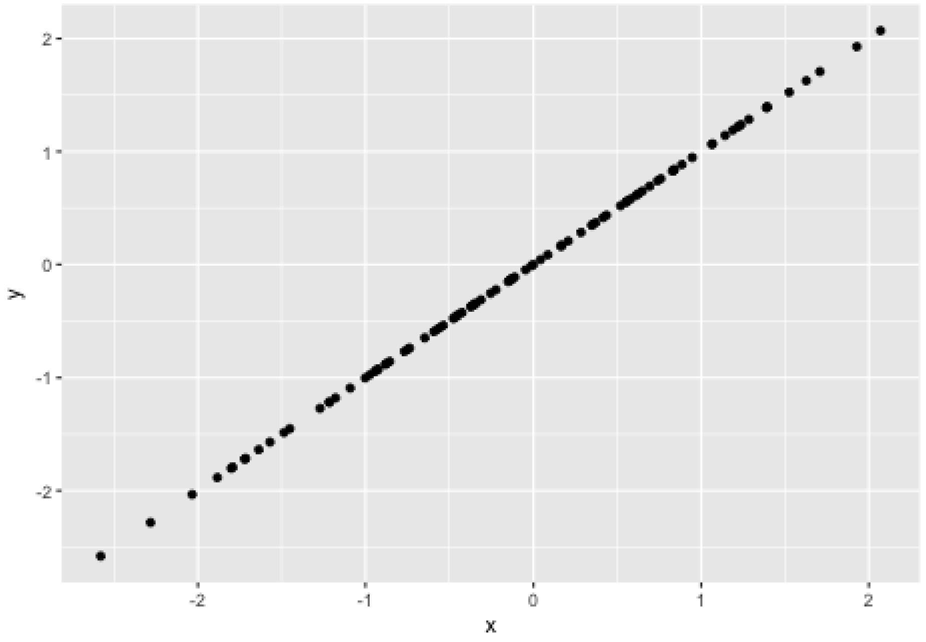
The strength of the relationship between the two variables can be seen by how much variation, or scatter, there is around the main form.

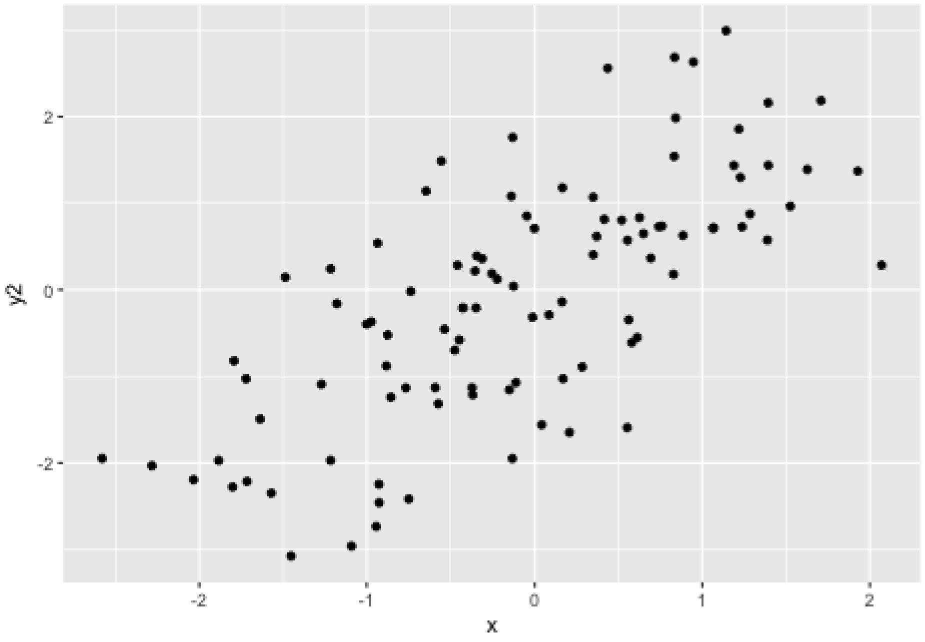
- With a strong relationship, you can get a pretty good estimate of y if you know x.
- With a weak relationship, for any x you might get a wide range of y values.

Correlation: Strength

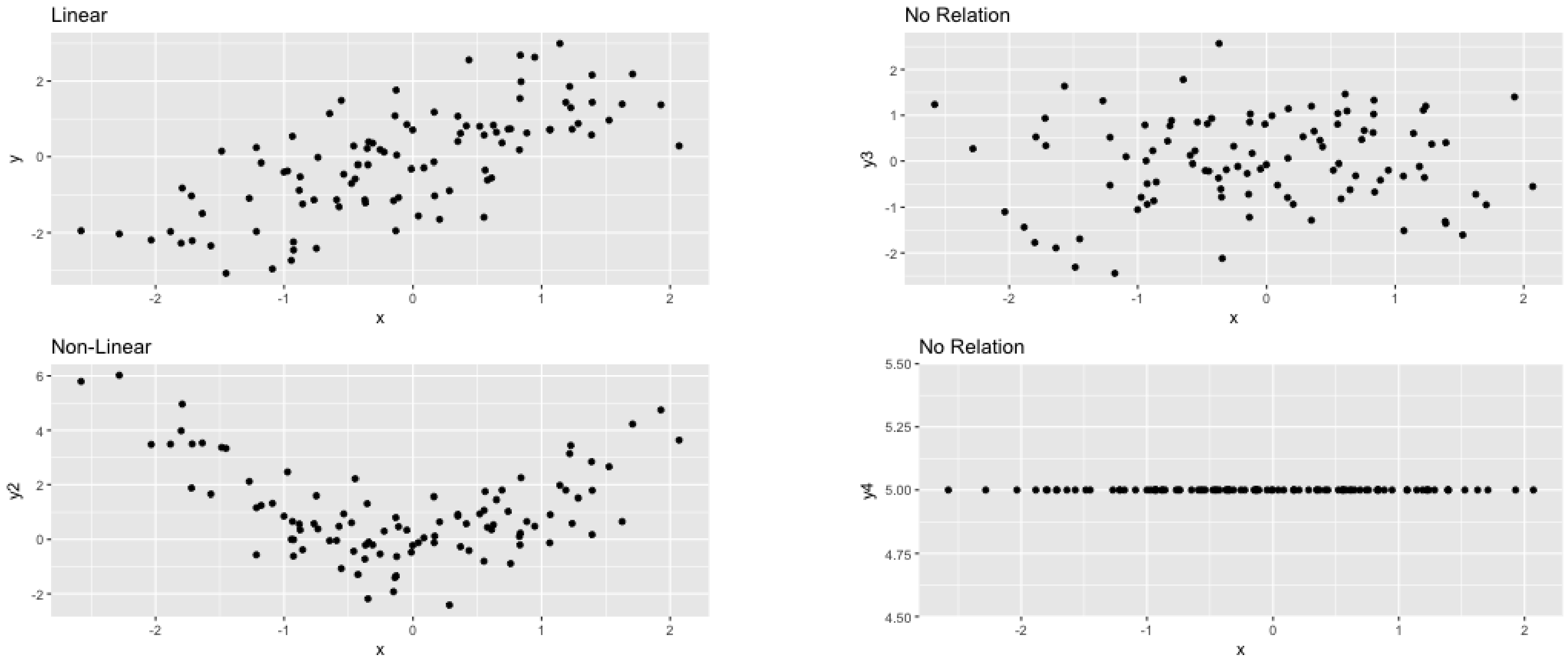
The strength of the relationship between the two variables can be seen by how much variation, or scatter, there is around the main form.

- 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.



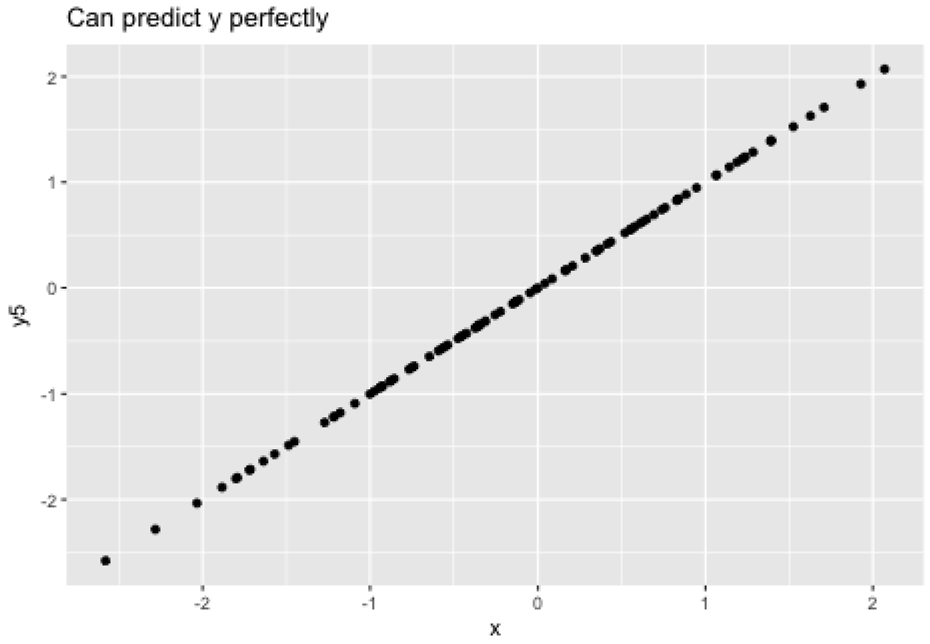


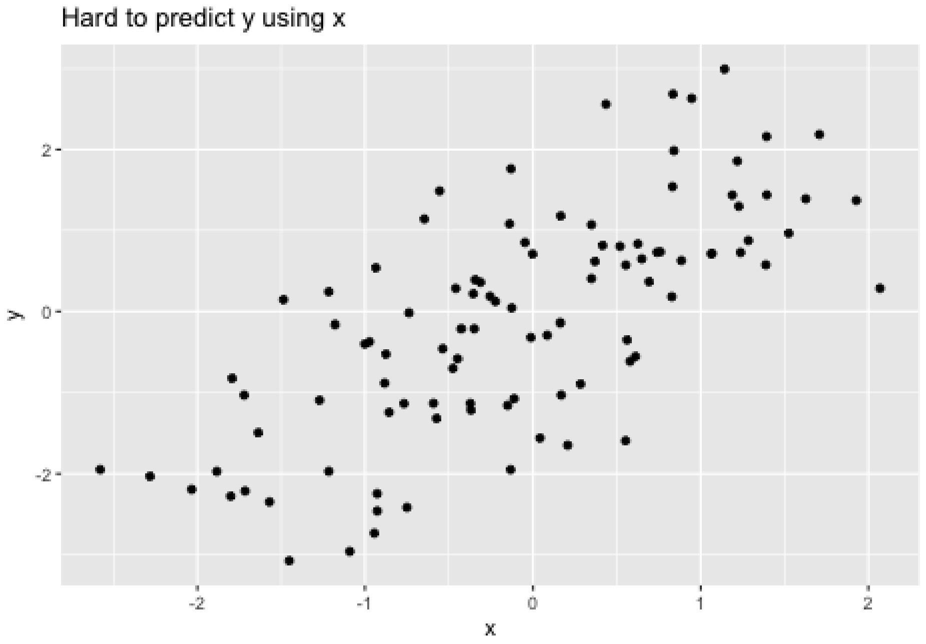
Scatterplot Patterns



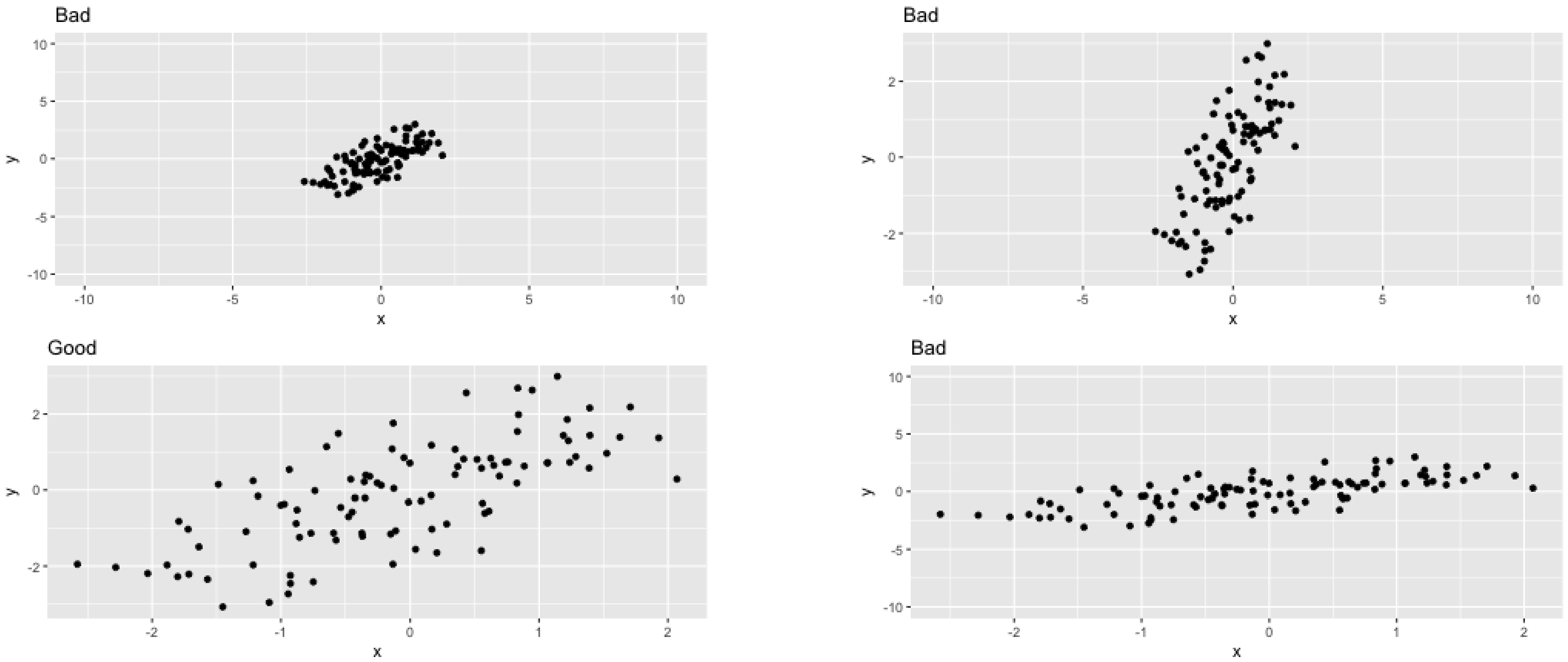
Predictability

The ability to predict `y` based on `x` is another indication of correlation strength:





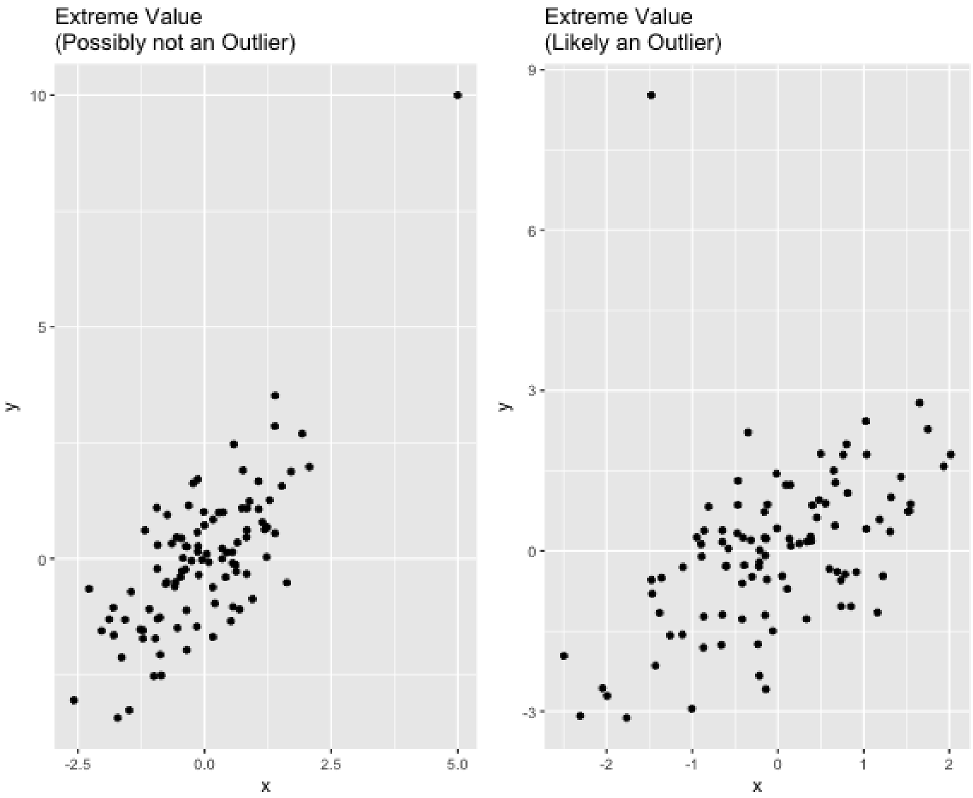
Scatterplot: Scale



Note: all have the same data! Also, `ggplot2`'s defaults are usually pretty good

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 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. Why?

`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!

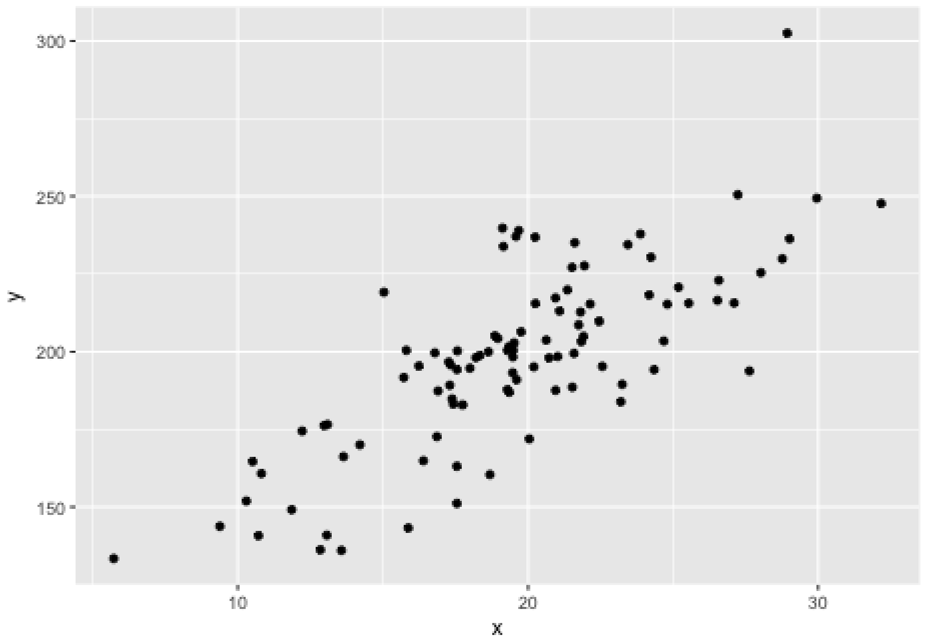
Correlation: Calculating

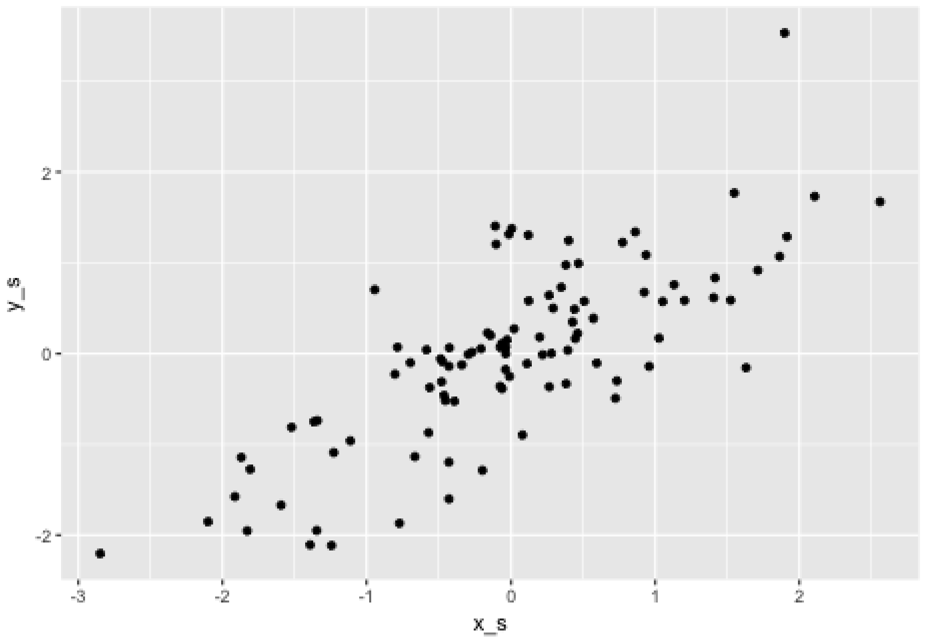
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Anyone want to do this by hand??

Let's use R to do this for us

Correlation: Calculating





Same Plots -- Left is unstandardized, Right is standardized

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].

Correlations in R Code

```
df %>%
  cor.test(~x + y,
           data = .,
           method = "pearson")
```

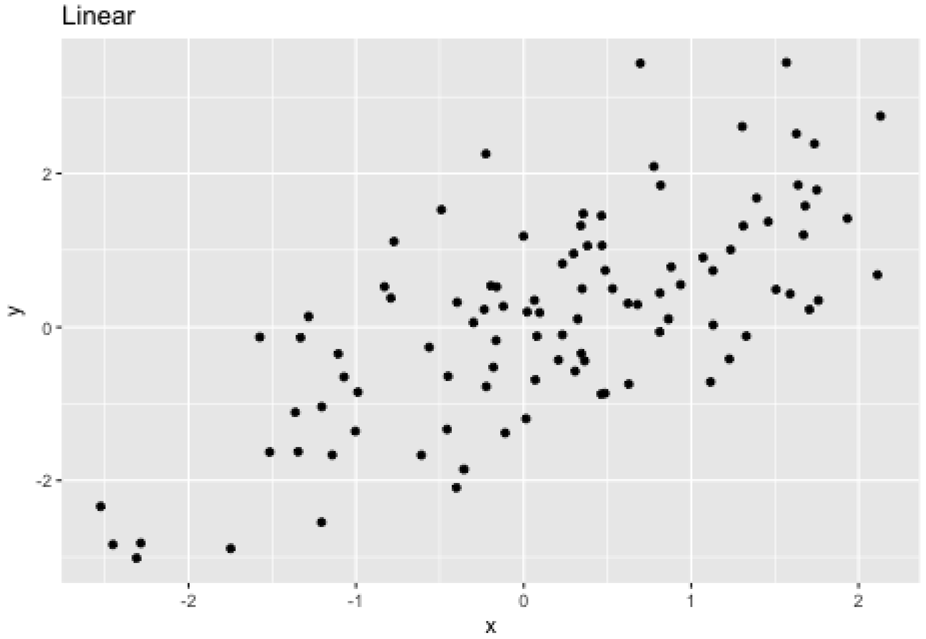
```

	Pearson's product-moment correlation

data:  x and y
t = 0.53442, df = 98, p-value = 0.5943
alternative hypothesis: true correlation is not equal
95 percent confidence interval:
 -0.1440376  0.2477011
sample estimates:
       cor
0.05390564
```

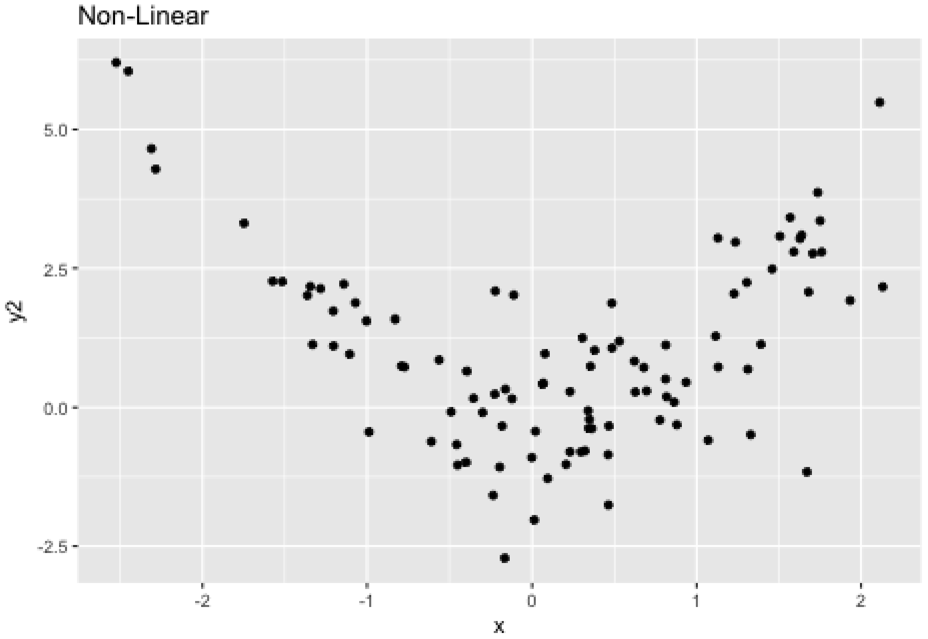
```
df %>%
  furniture::tableC(x, y)
```

```
      [1]           [2]
[1]x  1.00
[2]y  0.054 (0.594) 1.00
```

Relationship Form

Correlations only describe **linear** relationships





Note: You can sometimes *transform* a non-linear association to a linear form, for instance by taking the logarithm.

Let's see it in action

Correlation App

- Influential Points
- Eye-ball the correlation
- Draw the line of the best fit

Why are correlations not resistant to outliers?

When do outliers have more *leverage*?

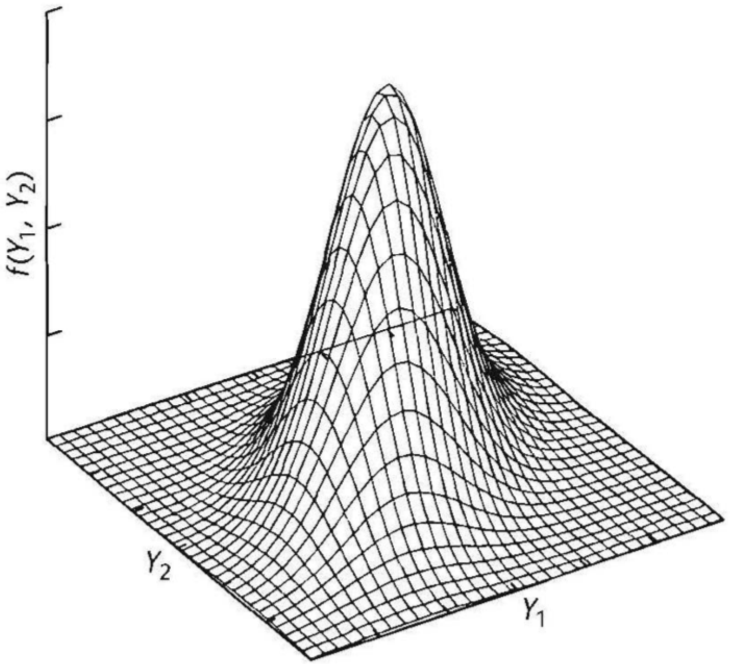
Assumptions

1. Random Sample

2. Relationship is linear (check scatterplot, use transformations)

3. 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`

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

`r` is converted to a t-statistic

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

- Compare to t-distribution with $df = N - 2$
  - Rejection = statistical evidence of relationship
  - Or look up critical values of `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 recal memory

```
# A tibble: 7 x 2
   Mood Recall
  <dbl>  <dbl>
1    45     48
2    34     39
3    41     48
4    25     27
5    38     42
6    20     29
7    45     30
```

```
df %>%
  cor.test(~Mood + Recall,
           data = .)
```

```

    Pearson's product-moment correlation

data:  Mood and Recall
t = 1.8815, df = 5, p-value = 0.1186
alternative hypothesis: true correlation is not equal
95 percent confidence interval:
 -0.2120199  0.9407669
sample estimates:
      cor
0.6438351
```

Power

Want to know N necessary to reject H_0 given an effect ρ (we transform it into a d)

- Determine effect size needed to detect
- Determine delta (δ ; the value from appendix A.4 that would result in given level of power at $\alpha = .05$)
- Solve:

$$\left(\frac{\delta}{d}\right)^2 + 1 = N$$

Example

Based on a pilot study, if we had a pearson correlation of .6, how many observations should I plan to study to ensure I have at least 80% power for an $\alpha = .05$, two-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 and Communcation

Correlation $\neq$ Causation

But, correlation can be 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: $r(df) = -.74$, $p = .006$

APA Style of 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).

Let's Apply This to the Cancer Dataset

Read in the Data

```r
library(tidyverse)     # Loads several very helpful 'tidy' packages
library(haven)         # Read in SPSS datasets
library(furniture)     # for tableC()
```

```r
cancer_raw <- haven::read_spss("cancer.sav")
```

And Clean It

```r
cancer_clean <- cancer_raw %>%
  dplyr::rename_all(tolower) %>%
  dplyr::mutate(id = factor(id)) %>%
  dplyr::mutate(trt = factor(trt,
                             labels = c("Placebo",
                                        "Aloe Juice"))) %>%
  dplyr::mutate(stage = factor(stage))
```

R Code: Basic Correlations

```
cancer_clean %>%
  cor.test(~ totalcin + totalcw2,
           data = .,
           alternative = "two.sided",
           method = "pearson")
```

```
    Pearson's product-moment correlation

data:  totalcin and totalcw2
t = 1.5885, df = 23, p-value = 0.1258
alternative hypothesis: true correlation is not equal
95 percent confidence interval:
 -0.09215959  0.63114058
sample estimates:
     cor
0.314421
```

R Code: Basic Correlations

```
cancer_clean %>%
  cor.test(~ totalcin + totalcw2,
           data = .,
           alternative = "two.sided",
           method = "pearson")
```

```
cancer_clean %>%
  cor.test(~ totalcin + totalcw2,
           data = .,
           alternative = "less",
           method = "pearson")
```

```
cancer_clean %>%
  cor.test(~ totalcin + totalcw2,
           data = .,
           alternative = "greater",
           method = "pearson")
```

```
	Pearson's product-moment correlation

data:  totalcin and totalcw2
t = 1.5885, df = 23, p-value = 0.1258
alternative hypothesis: true correlation is not equal
95 percent confidence interval:
 -0.09215959  0.63114058
sample estimates:
     cor 
0.314421 
```

R Code: Correlation Matrix

```r
cancer_clean %>%
  furniture::tableC(totalcin, totalcw2,
                    totalcw4, totalcw6)
```

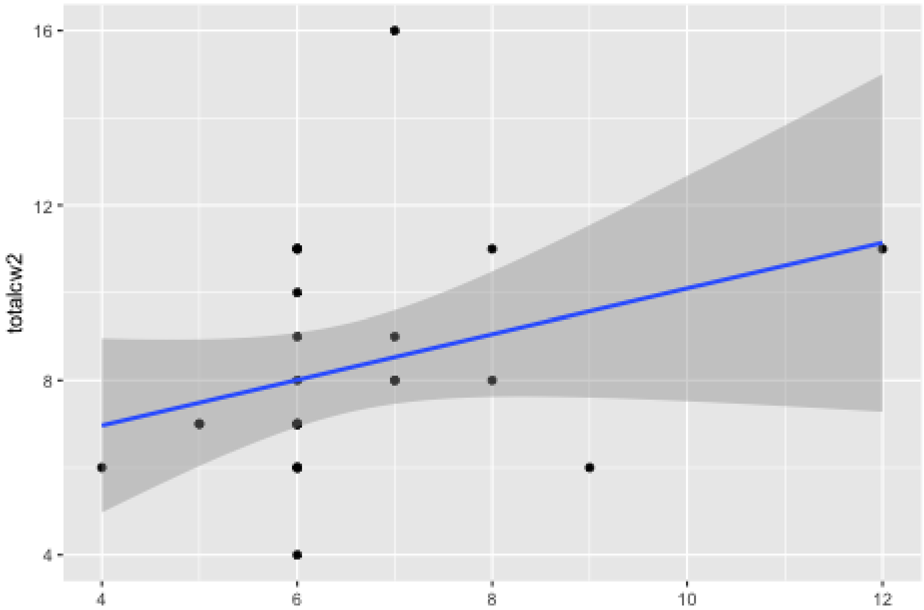
```r
cancer_clean %>%
  furniture::tableC(totalcin, totalcw2,
                    totalcw4, totalcw6,
                    na.rm=TRUE)
```

	[1]	[2]	[3]	[4]
[1]totalcin	1.00			
[2]totalcw2	0.314 (0.126)	1.00		
[3]totalcw4	0.222 (0.287)	0.337 (0.099)	1.00	
[4]totalcw6	NA NA	NA NA	NA NA	1.00

	[1]	[2]	[3]	[4]
[1]totalcin	1.00			
[2]totalcw2	0.282 (0.192)	1.00		
[3]totalcw4	0.206 (0.346)	0.314 (0.145)	1.00	
[4]totalcw6	0.098 (0.657)	0.378 (0.075)	0.763 (<.001)	1.00

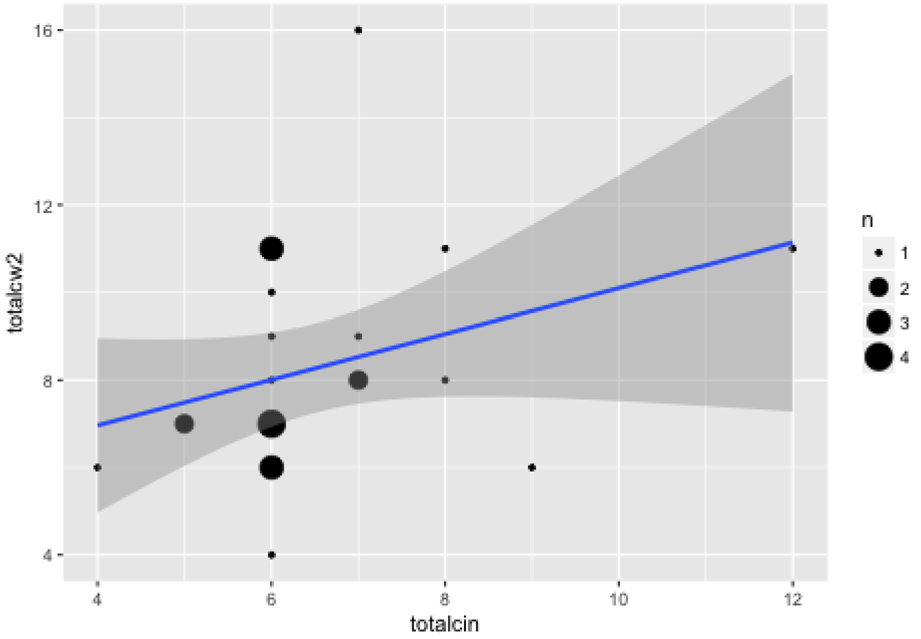
R Code: Scatterplot with Regression Line

```
cancer_clean %>%
  ggplot(aes(totalcin, totalcw2)) +
    geom_point() +
    geom_smooth(method = "lm")
```



R Code: Scatterplot with Count

```
cancer_clean %>%
  ggplot(aes(totalcin, totalcw2)) +
    geom_count() +
    geom_smooth(method = "lm")
```



Questions?

Next Topic

Linear Regression