SDS358: Applied Regression Analysis

Day 3: Correlation

Dr. Michael J. Mahometa

Agenda for Today:

- Correlation
- Concept
- Assumptions/Properties
- Statistical Testing

Correlation

The Reading Quiz

- Let's see how you guys did…
  - Canvas

Today's Quesiton:

Primary Research Question:
Is there a significant linear relationship between the Tar content and Carbon Monoxide content in typical cigarettes?

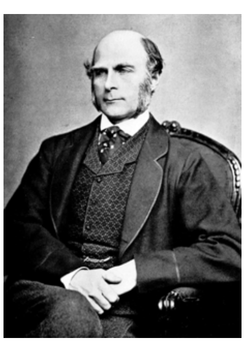
Correlation

A quick refresher

- Variable: any component being measured
- Type: categorical or continuous (quantitative)
- Scale: nominal, ordinal, interval, and ratio
- Univariate: single variable
- Multivariate: containing multiple variables

In the Beginning

- In the beginning there was Britain (and eugenics)
- Sir Francis Galton & Karl Pearson
- And there was data
    - specifically paired data



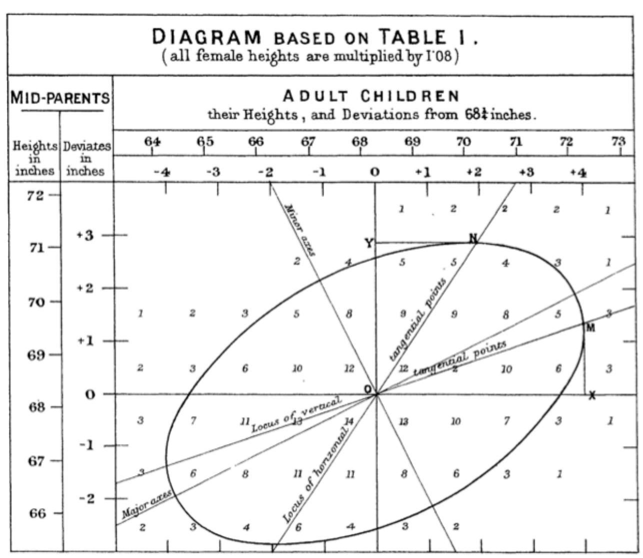

Draw a Picture

(or have the computer do it for you)

- Scatterplots may be the most common and most effective display for data.
- In a scatterplot, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others.
- Scatterplots are the best way to start observing the relationship and the ideal way to picture associations between two quantitative variables.

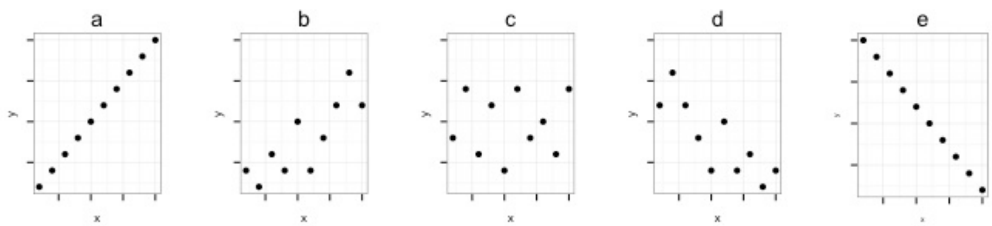
The (First) Scatterplot

- Height of parents (mean values).
- Height of adult children.



The Scatterplot

- When looking at scatterplots, we will look for trend, shape, and strength. (and unusual features)
  - Trend: Gives us direction.
  - Shape: Shows us linearity.
  - Strength: Tells us about the density.



It's All About Relationships

- TWO variables are needed
- It is important to determine which of the two quantitative variables goes on the x-axis and which on the y-axis.
- This determination is made based on the roles played by the variables.
- When the roles are clear, the explanatory or predictor variable goes on the x-axis, and the response variable goes on the y-axis.

Our First (and Last) Relationship

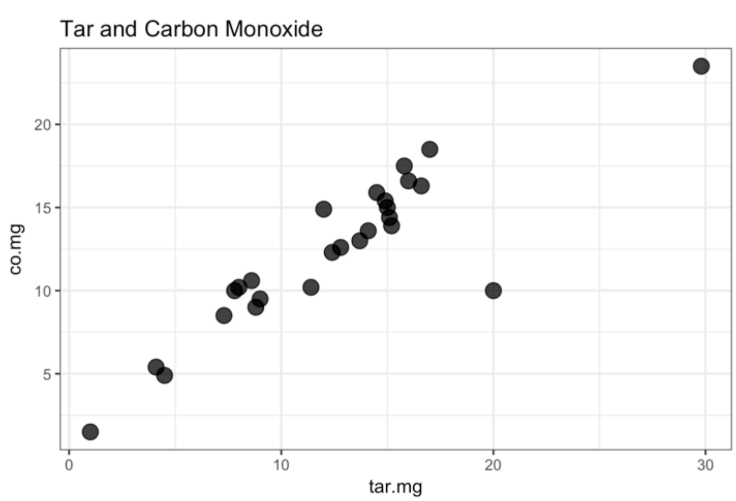
Some Data:

```
cigs <- read_csv("../data/cigarettes.csv")
cigs
```

```
## # A tibble: 26 x 5
##    brand         tar.mg nicotine.mg weight.g co.mg
##    <chr>          <dbl>       <dbl>    <dbl> <dbl>
##  1 Alpine         14.1        0.860    0.985 13.6
##  2 Benson&Hedges  16.0        1.06     1.09  16.6
##  3 BullDurham     29.8        2.03     1.16  23.5
##  4 CamelLights     8.00       0.670    0.928 10.2
##  5 Carlton         4.10       0.400    0.946  5.40
##  6 Chesterfield   15.0        1.04     0.888 15.0
##  7 GoldenLights    8.80       0.760    1.03   9.00
##  8 Kent           12.4        0.950    0.922 12.3
##  9 Kool           16.6        1.12     0.937 16.3
## 10 L&M            14.9        1.02     0.886 15.4
## # ... with 16 more rows
```


Our First (and Last) Relationship

The Scatterplot:



Criterion for "r"

Galton said to Pearson…

- A numeric description for the relationship between two continuous variables.
- Should range from 0 (no relationship) to 1 (perfect relationship).
- Should capture the type of relationship (positive or negative).

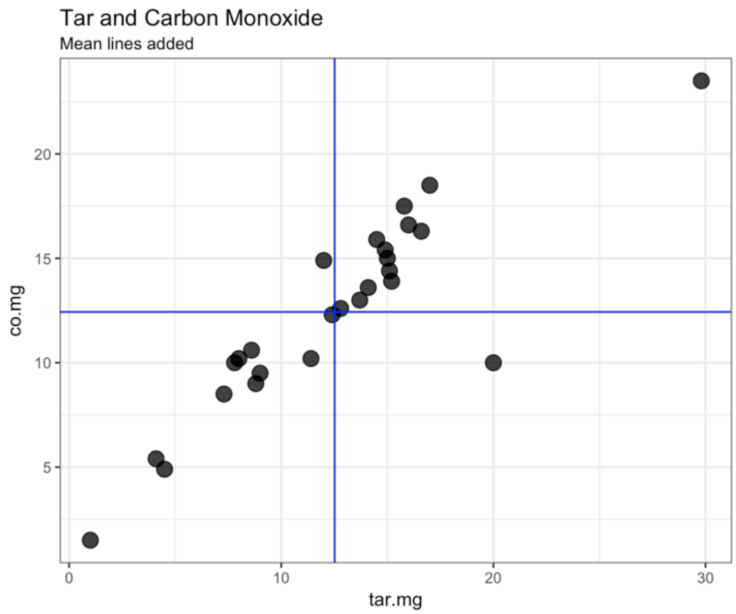
What if…

What if you didn't know the formula for r?

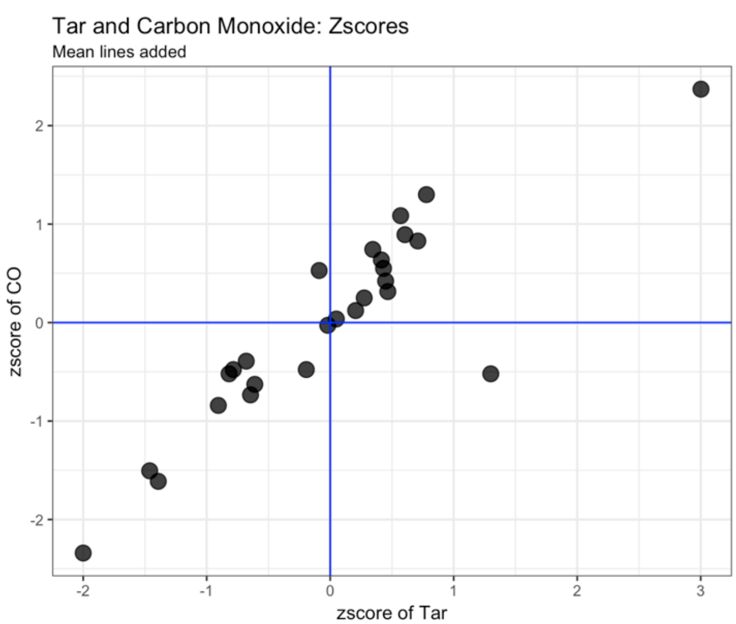
And, you were forced to think about it from scratch (put yourself in Pearson's shoes).

Here's what I think you might find:

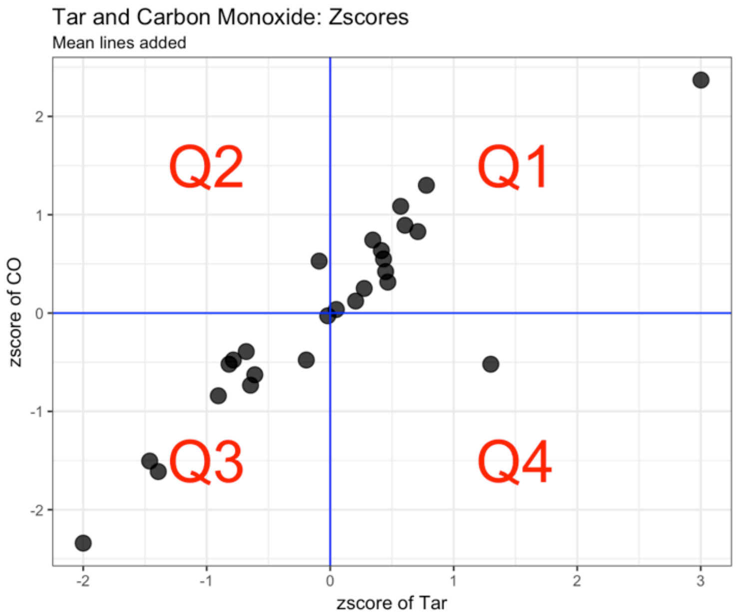
What if…mean lines



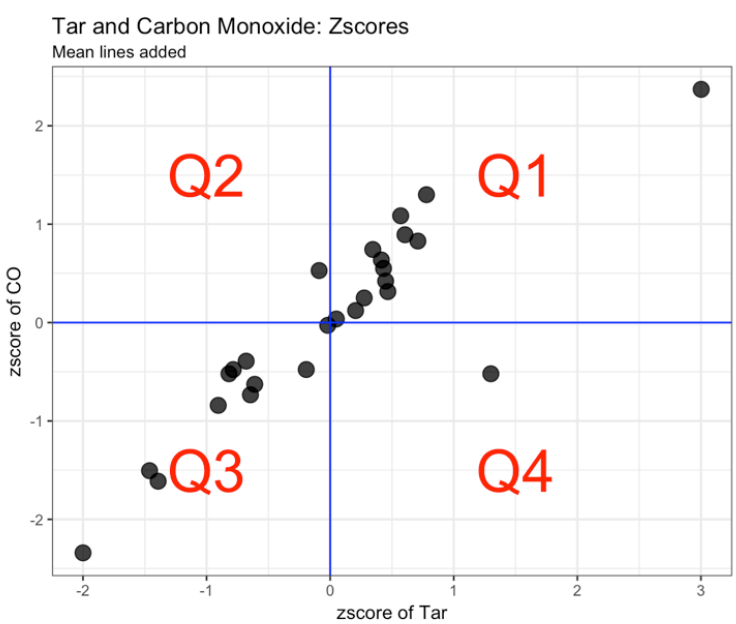
What if…mean lines



What if…mean lines (plus quadrants)



What if…mean lines (plus quadrants)



$$r = \sum \frac{(Z_x Z_y)}{n-1}$$

The Equation We All Love

$$r = \sum \frac{(Z_x Z_y)}{n-1}$$

$$r = \sum \frac{(x-\bar{x})(y-\bar{y})}{(S_x)(S_y)(n-1)}$$

Your first "r" (in R)

```
library(tidyverse)
cor(select(cigs, tar.mg, co.mg))
```

```
##           tar.mg     co.mg
## tar.mg 1.0000000 0.8898509
## co.mg  0.8898509 1.0000000
```

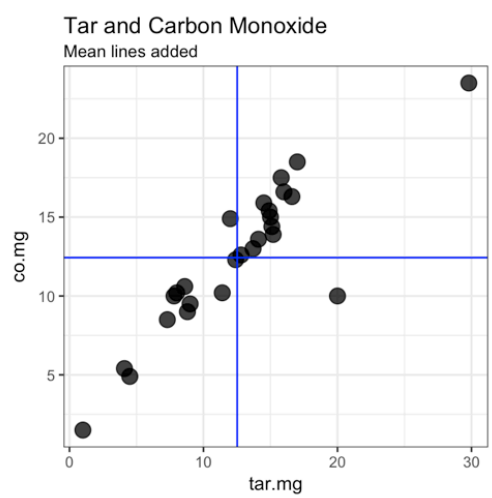
But wait! (Assumptions)

In order for this "idea" of relationship (and the formula) to work:

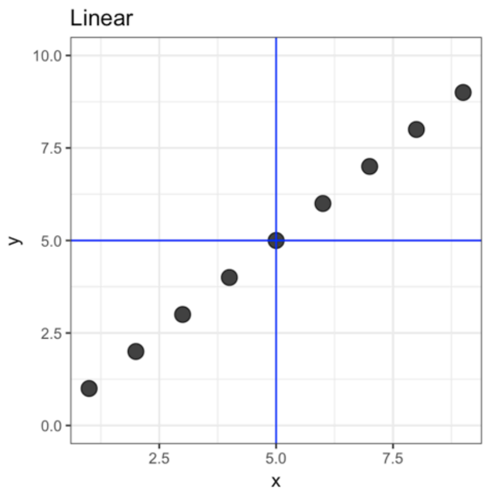
- Observations (rows) should be *independent*
- Variables used in the Pearson r should be *quantitative*
- No *outliers* or *clusters* to the relationship
- There should be a *linear* relationship between the variables

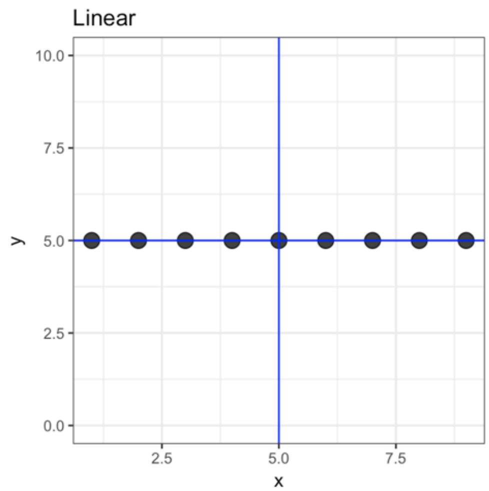
Some Properties of r

- **Linearity:** r measures how close the points in a scatterplot are to a straight line.
- **Scale transformations:** r is unaffected by linear transformations to the data.
- **Outliers hurt**
- **Causation**
- **Interpretation**

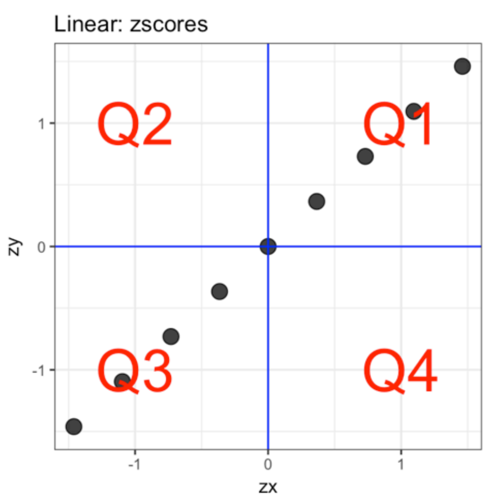


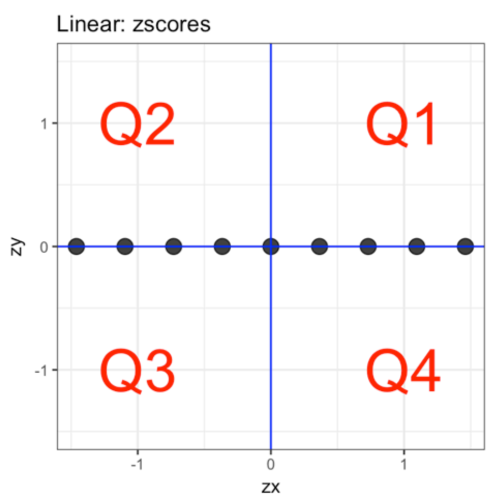
Linearity



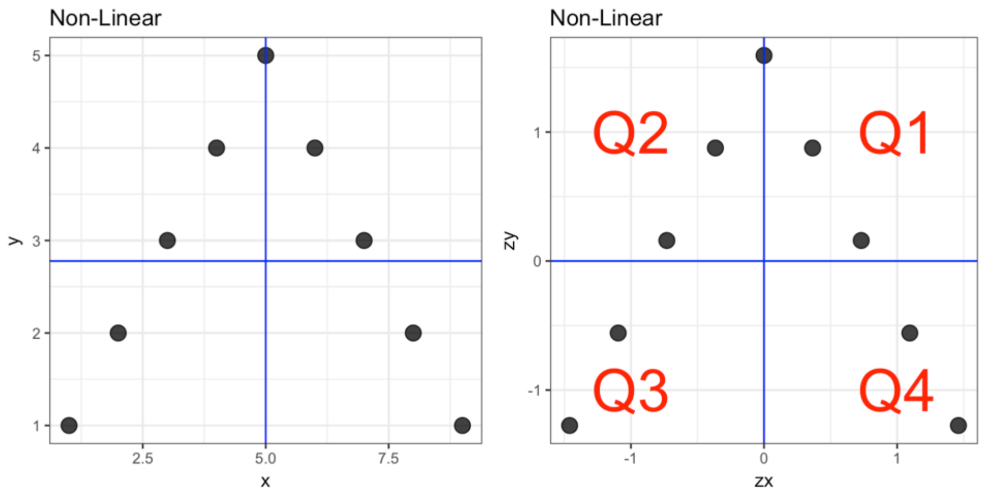


Linearity





Non-Linearity



Scale Transformation

- Linear transformations do not hurt the value of r
- The importance is on the relationship between two variables, not their scale

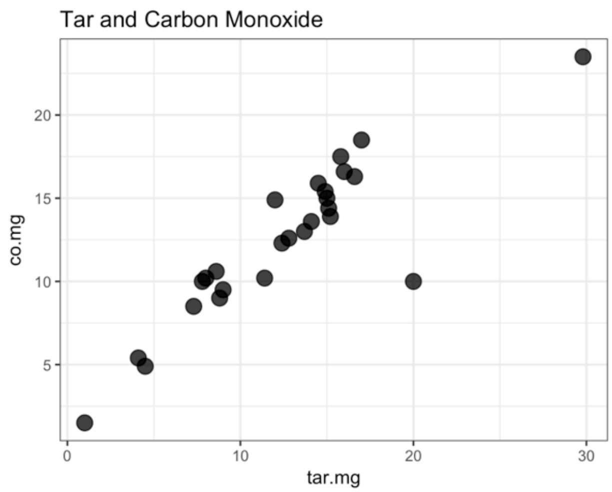
```
cigs <- read_csv("../data/cigarettes.csv")
head(cigs)
```

```
## # A tibble: 6 x 5
##   brand          tar.mg nicotine.mg weight.g co.mg
##   <chr>           <dbl>       <dbl>    <dbl> <dbl>
## 1 Alpine          14.1        0.860    0.985 13.6
## 2 Benson&Hedges   16.0        1.06     1.09  16.6
## 3 BullDurham      29.8        2.03     1.16  23.5
## 4 CamelLights      8.00       0.670    0.928 10.2
## 5 Carlton          4.10       0.400    0.946  5.40
## 6 Chesterfield    15.0        1.04     0.888 15.0
```

Scale Transformation

The Relationship is Key

```
simpleScatter(cigs, tar.mg, co.mg, title="Tar and Carbon Monoxide")
```



Scale Transformation

The Relationship is Key

```
library(tidyverse)
cor(select(cigs, tar.mg, co.mg))
```

```
##           tar.mg     co.mg
## tar.mg 1.0000000 0.8898509
## co.mg  0.8898509 1.0000000
```

Scale Transformation

Add 20…

```
cigs <- cigs %>%
  mutate(tar_20 = tar.mg + 20,
         co_20 = co.mg + 20)
```


Scale Transformation

Add 20…

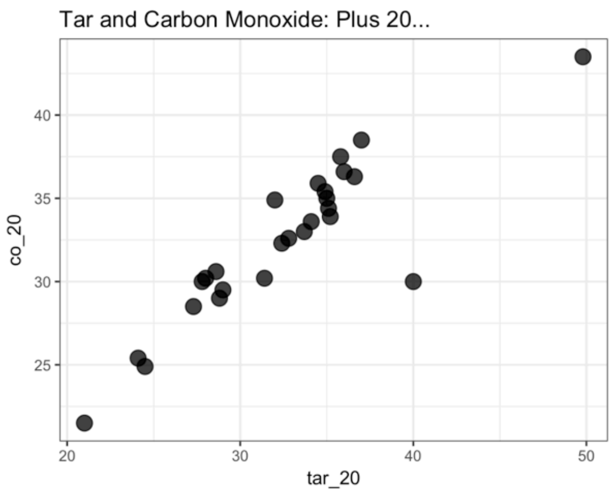
```
select(cigs, 1, 2, 5, 6, 7)
```

```
## # A tibble: 26 x 5
##    brand         tar.mg co.mg tar_20 co_20
##    <chr>          <dbl> <dbl>  <dbl> <dbl>
##  1 Alpine         14.1  13.6    34.1  33.6
##  2 Benson&Hedges  16.0  16.6    36.0  36.6
##  3 BullDurham     29.8  23.5    49.8  43.5
##  4 CamelLights     8.00 10.2    28.0  30.2
##  5 Carlton         4.10  5.40   24.1  25.4
##  6 Chesterfield   15.0  15.0    35.0  35.0
##  7 GoldenLights    8.80  9.00   28.8  29.0
##  8 Kent           12.4  12.3    32.4  32.3
##  9 Kool           16.6  16.3    36.6  36.3
## 10 L&M            14.9  15.4    34.9  35.4
## # ... with 16 more rows
```

Scale Transformation

Add 20…

```
simpleScatter(cigs, tar_20, co_20, title="Tar and Carbon Monoxide: Plus 20...")
```



Scale Transformation

Add 20…

```
cor(select(cigs, tar_20, co_20))
```

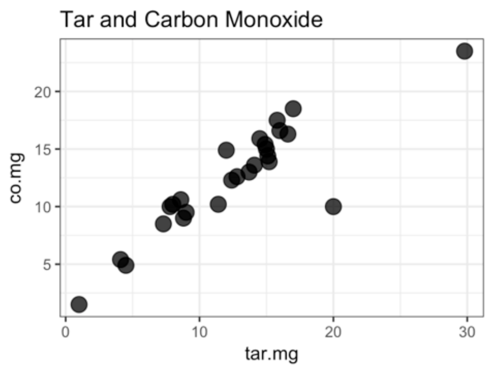
```
##           tar_20     co_20
## tar_20 1.0000000 0.8898509
## co_20  0.8898509 1.0000000
```

Outliers Hurt

```
cor(select(cigs, tar.mg, co.mg))
```

```
##           tar.mg     co.mg
## tar.mg 1.0000000 0.8898509
## co.mg  0.8898509 1.0000000
```

```
simpleScatter(cigs, tar.mg, co.mg, title="Tar and Carbon Monoxide")
```



Outliers Hurt

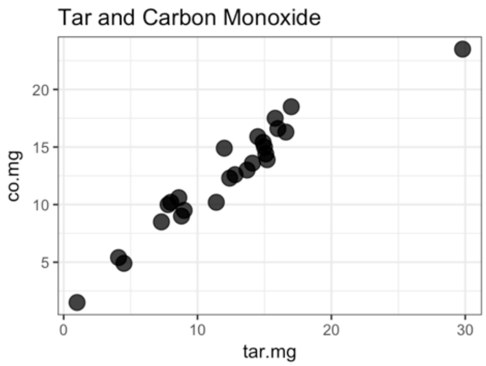
```
cigs <- cigs %>%
  mutate(tar.mg = replace(tar.mg, tar.mg == 20, NA))
```

Outliers Hurt

```
cor(select(cigs, tar.mg, co.mg), use="pairwise.complete.obs")
```

```
##           tar.mg     co.mg
## tar.mg 1.0000000 0.9574853
## co.mg  0.9574853 1.0000000
```

```
simpleScatter(cigs, tar.mg, co.mg, title="Tar and Carbon Monoxide")
```



Visuals are VERY IMPORTANT

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

1. numerical colculations are exact, but graphs are rough;
2. for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
3. performing intricate calculations is virtuous, whereas actually looking at data is cheating.

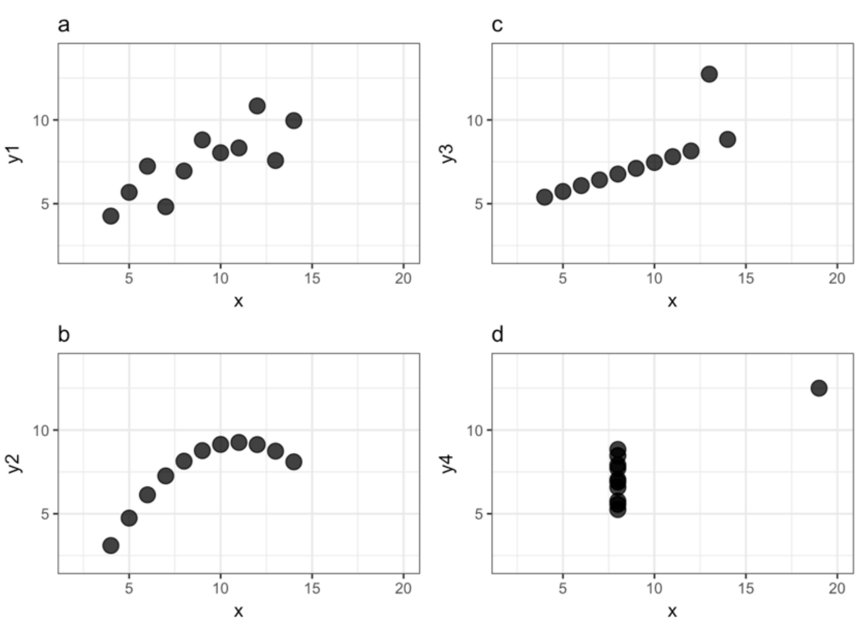
Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21.

Anscombe's Quartet

Anscombe made *four* datasets:

- The number of paired observations in each is 11
- The x-variable has a mean of 9
- The y-variable has a mean of 7.5
- The r value for the relationship between x and y is 0.816

Anscombe's Quartet



Causation

- "Correlation does not imply causation."

Causation

- "Correlation does not imply causation."
- Daily ice cream consumption and daily water consumption for a city.
- Satisfaction and performance.
- APD: more pan-handlers and more traffic accidents.

Interpretation

- When is r big?
  - That depends - on the context.
  - Choice based on the outcome - small r.
  - Reliability - large r.
- How about r2?
  - Proportion of variance accounted for.
  - (Literally) - We'll see more of this with simple regression.

Many r's to Examine

The correlation matrix

```
cor(select(cigs, co.mg, tar.mg, nicotine.mg, weight.g),
    use="pairwise.complete.obs")
```

```
##              co.mg tar.mg nicotine.mg weight.g
## co.mg       1.0000 0.9575      0.9101   0.4647
## tar.mg      0.9575 1.0000      0.9766   0.4908
## nicotine.mg 0.9101 0.9766      1.0000   0.4959
## weight.g    0.4647 0.4908      0.4959   1.0000
```

r and Significance

- History
- Pearson himself had a *Student*
  - William S. Gosset
- Turns out, r can be examined in the face of a t-distribution.

r and Significance

$$t = \frac{\bar{x} - 0}{\sqrt{S^2/n}}$$

$$t = \frac{r - 0}{\sqrt{(1 - r^2)/(n - 2)}}$$

r and Significance

The `corr.test()` function in psych

t-values

```
library(psych)
corr.test(select(cigs, co.mg, tar.mg, nicotine.mg, weight.g))$t
```

```
##               co.mg  tar.mg nicotine.mg weight.g
## co.mg           Inf 15.9176     10.7597   2.5710
## tar.mg      15.9176     Inf     21.7814   2.7013
## nicotine.mg 10.7597 21.7814         Inf   2.7976
## weight.g     2.5710  2.7013      2.7976      Inf
```


r and Significance

The `corr.test()` function in psych

p-values

```
library(psych)
corr.test(select(cigs, co.mg, tar.mg, nicotine.mg, weight.g))$p
```

```
##              co.mg tar.mg nicotine.mg weight.g
## co.mg       0.0000 0.0000        0.00     0.03
## tar.mg      0.0000 0.0000        0.00     0.03
## nicotine.mg 0.0000 0.0000        0.00     0.03
## weight.g    0.0168 0.0127        0.01     0.00
```

Other "Relationship" Measures

- Spearman’s Rho
- Phi Coefficient
- Point-Biserial correlation
- Even chi-square!

Now for the context (summary):

- We use correlation (Pearson Correlation) to:
- Determine the relationship between two *quantitiative* variables.
- Give that relationship definition.
    - Size of r
- However, only *linear* relationships are correctly captured by r
    - That's why we use the scatterplot
- And, we can use the t-distribution to describe "significance" of the relationship.