THE UNIVERSITY of EDINBURGH

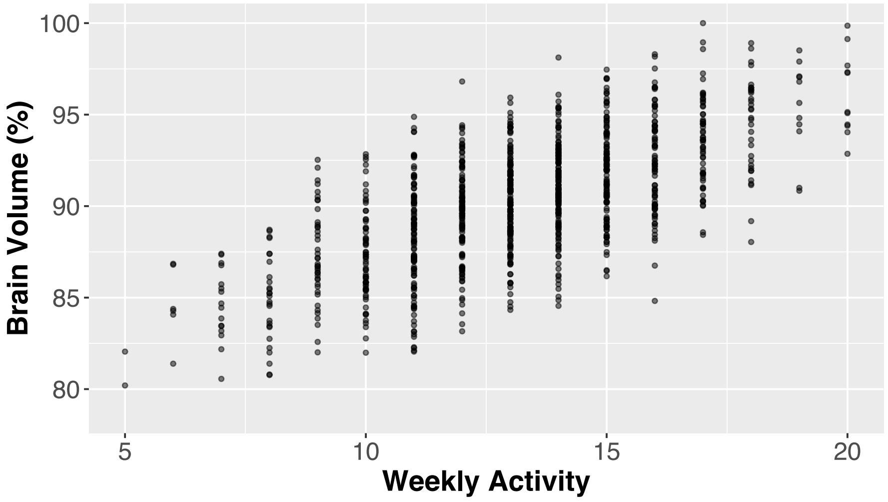
Week 5: Correlations

Univariate Statistics and Methodology using R

Department of Psychology
The University of Edinburgh

Part 1: Correlation

Brain Volume & Activity Level

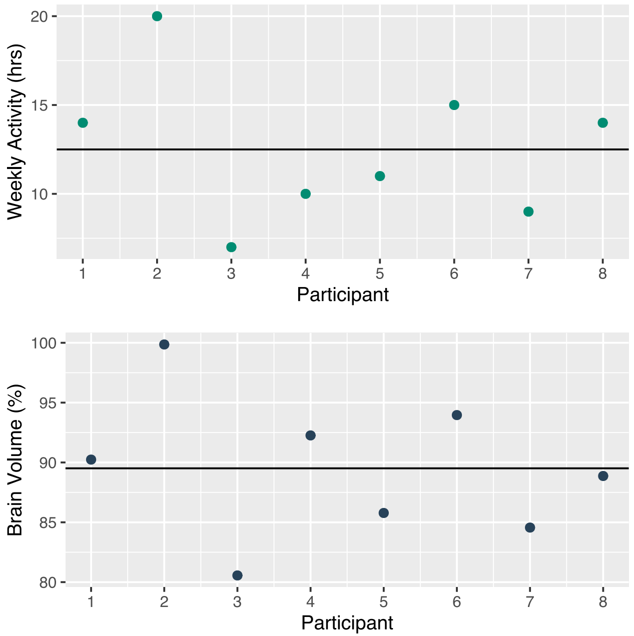


Correlation

- A measure of the relationship between two **continuous** variables
- Does a linear relationship exist between x and y ?
- Specifically, do two variables **covary**?
  - A change in one equates to a change in the other

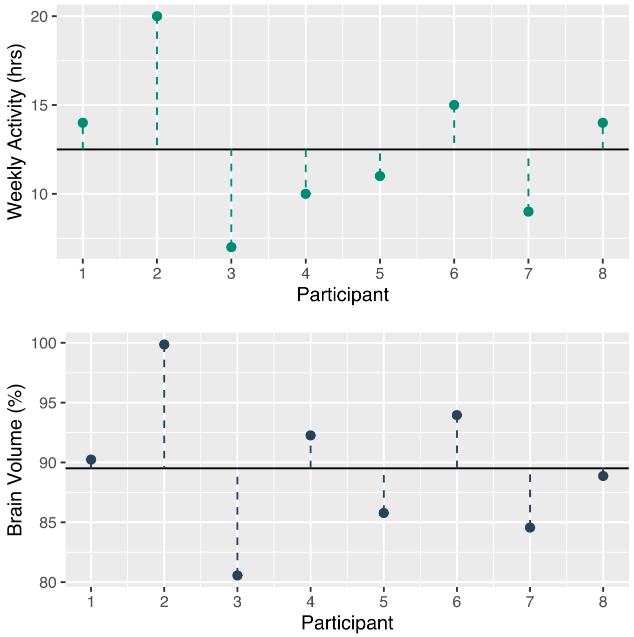
Correlation

- A measure of the relationship between two continuous variables
- Does a linear relationship exist between x and y ?
- Specifically, do two variables covary?
  - A change in one equates to a change in the other
- Does y vary with x ?
- Equivalent to asking "does y differ from its mean in the same way x does?"



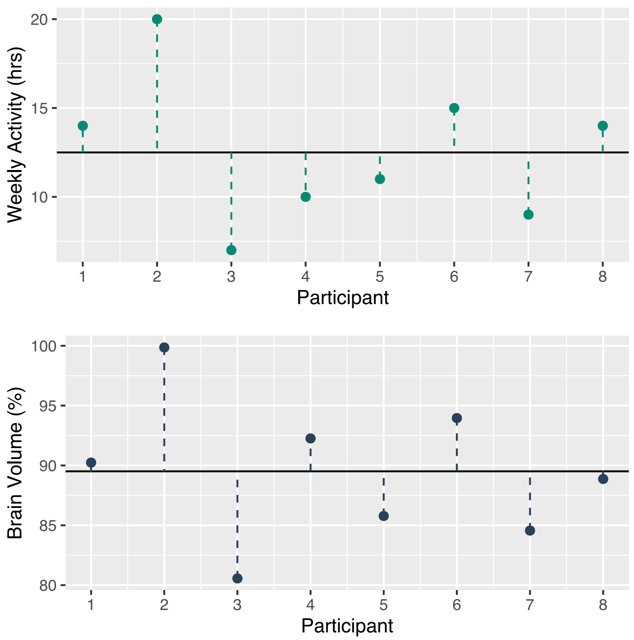
Correlation

- A measure of the relationship between two continuous variables
- Does a linear relationship exist between *x* and *y* ?
- Specifically, do two variables covary?
  - A change in one equates to a change in the other
- Does y vary with x ?
- Equivalent to asking "does y differ from its mean in the same way x does?"



Correlation

- A measure of the relationship between two continuous variables
- Does a linear relationship exist between *x* and *y* ?
- Specifically, do two variables covary?
    - A change in one equates to a change in the other
- Does y vary with x ?
- Equivalent to asking "does y differ from its mean in the same way x does?"
- It's likely the variables are related if observations differ proportionally from their means



Covariance

Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{\sum (x - \bar{x})(x - \bar{x})}{n}$$

Covariance

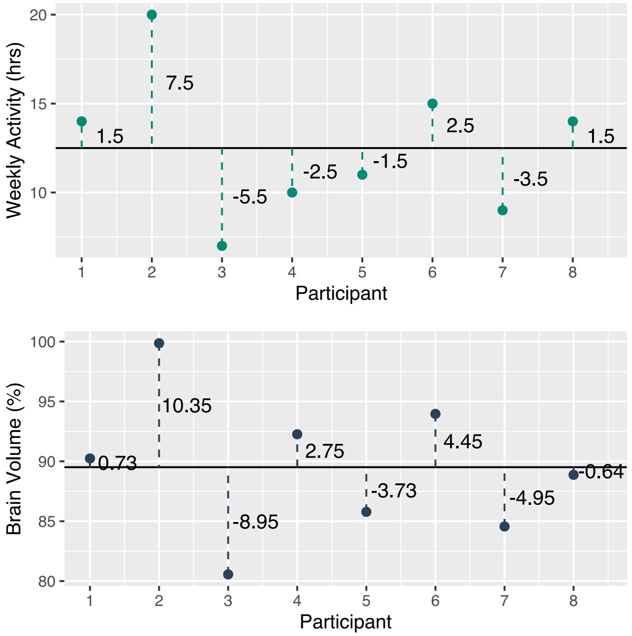
Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{\sum (x - \bar{x})(x - \bar{x})}{n}$$

Covariance

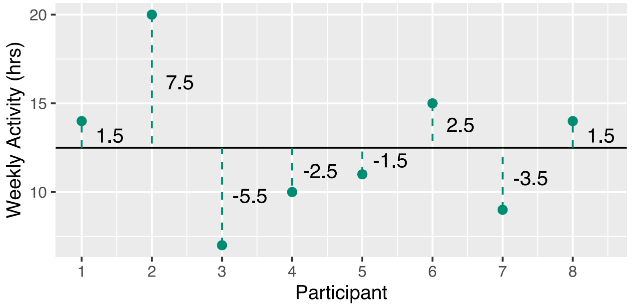
$$\operatorname{cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}$$

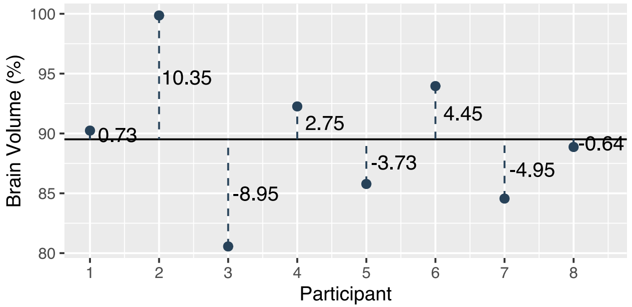
Covariance



$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1.5	0.73	1.095
7.5	10.35	77.625
-5.5	-8.95	49.225
-2.5	2.75	-6.875
-1.5	-3.73	5.595
2.5	4.45	11.125
-3.5	-4.95	17.325
1.5	-0.64	-0.96
		154.16

Covariance





$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1.5	0.73	1.095
7.5	10.35	77.625
-5.5	-8.95	49.225
-2.5	2.75	-6.875
-1.5	-3.73	5.595
2.5	4.45	11.125
-3.5	-4.95	17.325
1.5	-0.64	-0.96
		154.16

$$\text{cov}(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n} = \frac{154.16}{8} = 30.83$$

The Problem With Covariance

Miles

$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
-0.99	-0.1	0.1
3.22	1.78	5.73
2.46	0.97	2.38
-2.65	-1.31	3.47
-2.04	-1.34	2.73
		14.41

$$\text{cov}(x, y) = \frac{14.41}{5} \simeq 2.88$$

Kilometres

$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
-1.6	-0.16	0.25
5.19	2.86	14.84
3.96	1.56	6.16
-4.27	-2.11	8.99
-3.28	-2.15	7.06
		37.3

$$\text{cov}(x, y) = \frac{37.3}{5} \simeq 7.46$$

Correlation Coefficient

- The standardised version of covariance is the **correlation coefficient**, r

$$r = \frac{\text{covariance}(x, y)}{\text{standard deviation}(x) \cdot \text{standard deviation}(y)}$$

Correlation Coefficient

- The standardised version of covariance is the **correlation coefficient**, r

$$r = \frac{\text{covariance}(x, y)}{\text{standard deviation}(x) \cdot \text{standard deviation}(y)}$$

$$r = \frac{\frac{\sum (x-\bar{x})(y-\bar{y})}{N}}{\sqrt{\frac{\sum (x-\bar{x})^2}{N}}\sqrt{\frac{\sum (y-\bar{y})^2}{N}}}$$

Correlation Coefficient

- The standardised version of covariance is the correlation coefficient, r

$$r = \frac{\text{covariance}(x, y)}{\text{standard deviation}(x) \cdot \text{standard deviation}(y)}$$

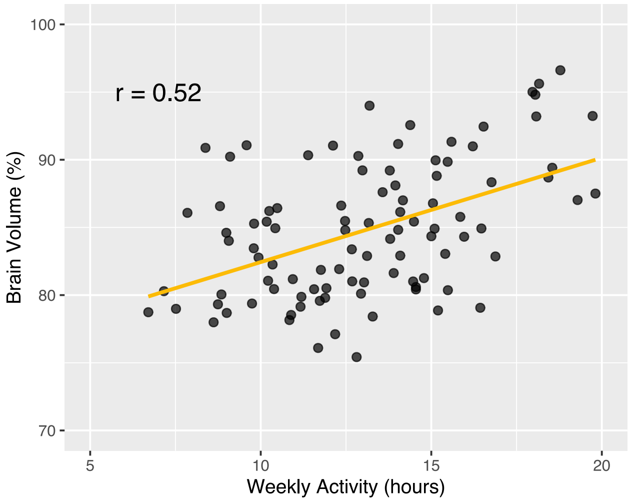
$$r = \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{N}}{\sqrt{\frac{\sum(x-\bar{x})^2}{N}}\sqrt{\frac{\sum(y-\bar{y})^2}{N}}}$$

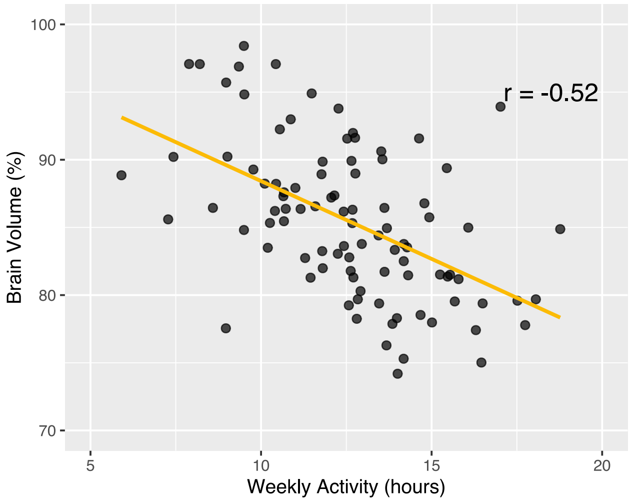
$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

Interpeting r

$-1 \leq r \leq 1$ (± 1 = perfect fit; 0 = no fit; sign shows direction of slope)

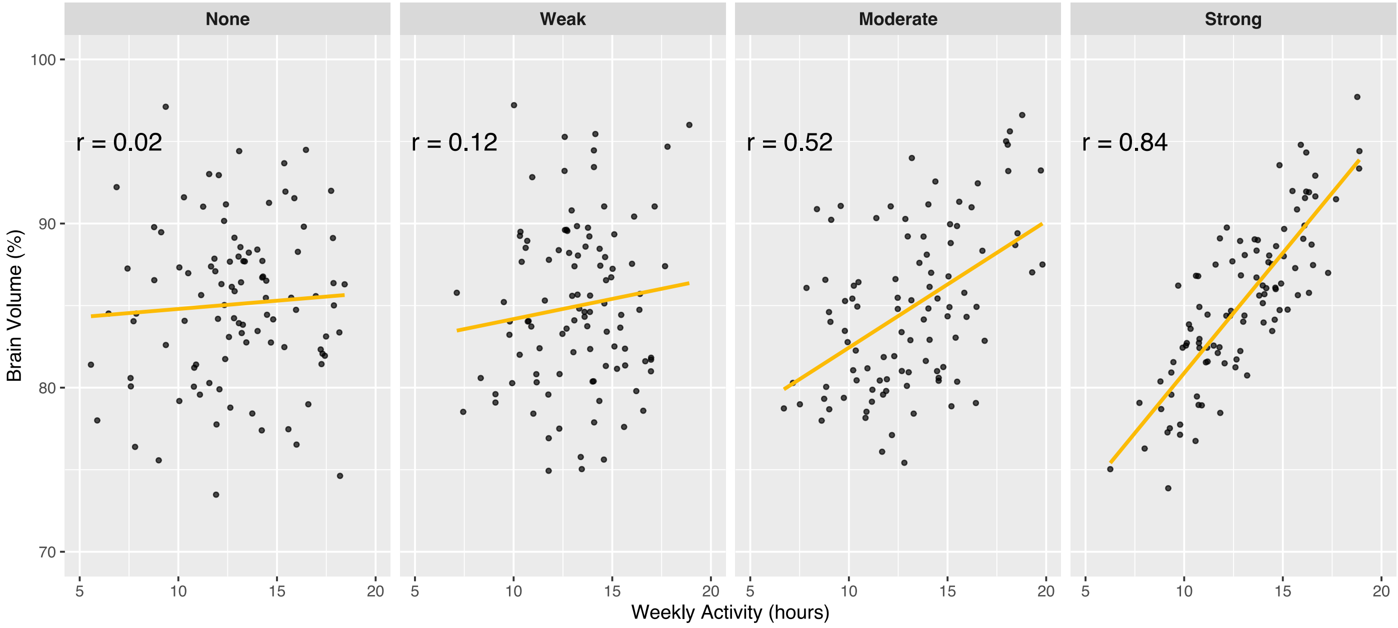
The sign of r gives you information about the direction of the relationship





Interpreting r

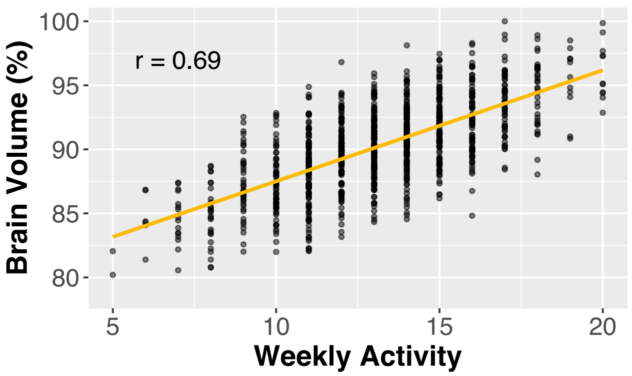
The magnitude of r gives you information about the strength of the relationship



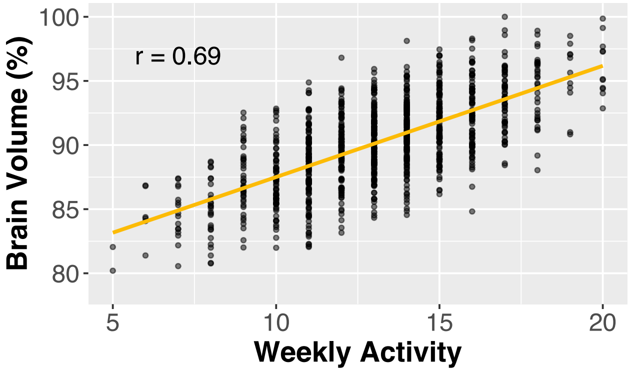
Part 2

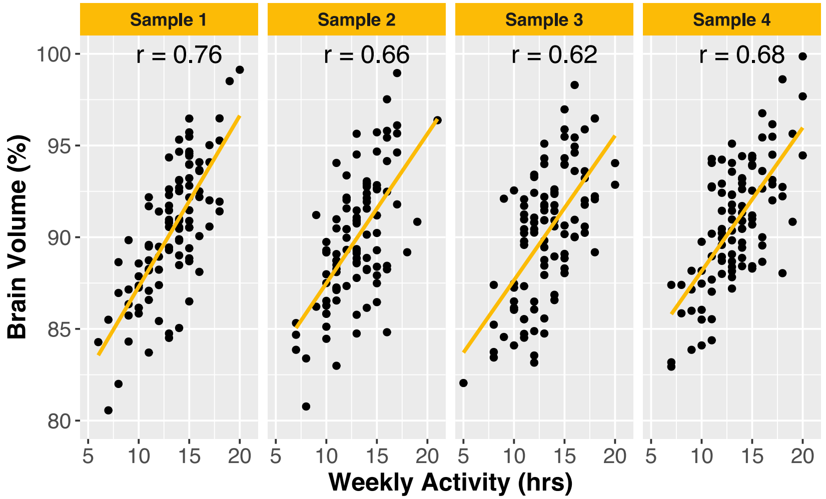
Hypothesis Testing with Correlation

Sampling from the Population

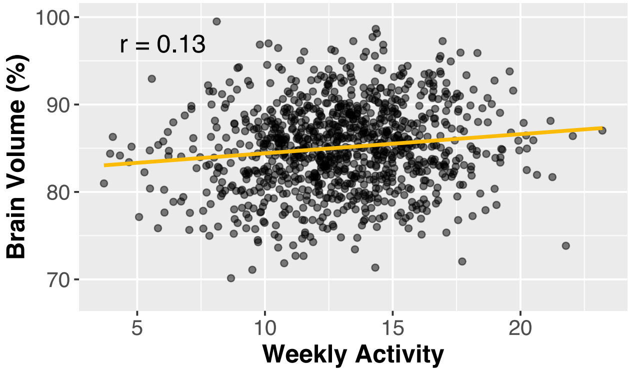


Sampling from the Population

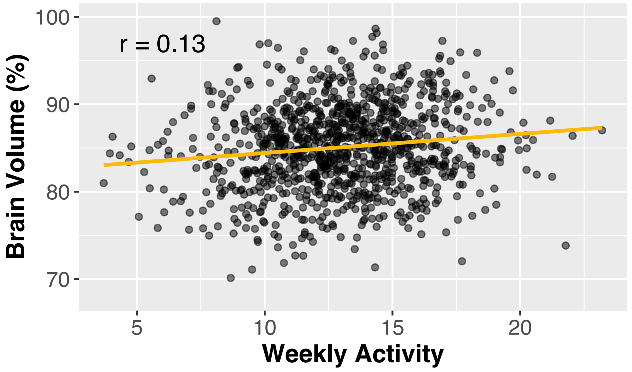


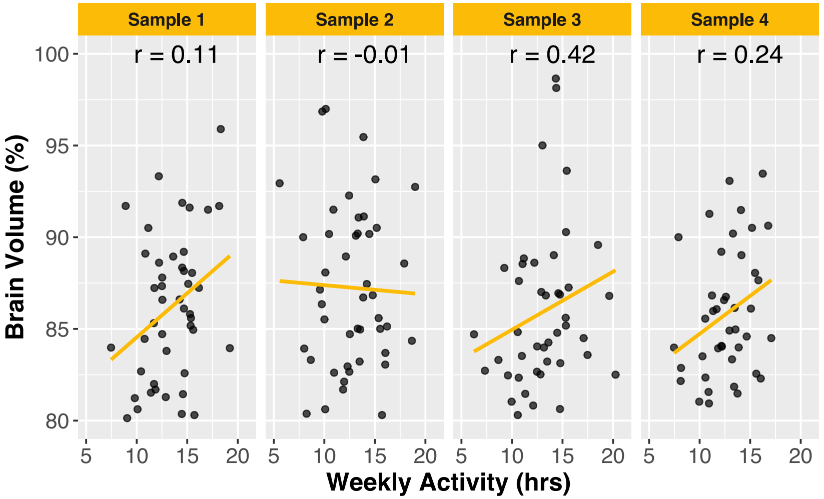


Sampling from the Population



Sampling from the Population





Hypothesis Testing

- Does a linear relationship exist between *x* and *y* ?
- $H_0 : r_{population} = 0$

Hypothesis Testing

- Does a linear relationship exist between *x* and *y* ?
- $H_0 : r_{population} = 0$
- **Two-tailed**
  - $H_1 : r_{population} \neq 0$
  - There is a significant relationship between brain volume and weekly activity.
  - As brain volume changes, weekly activity changes.

Hypothesis Testing

- Does a linear relationship exist between *x* and *y* ?
- $H_0 : r_{population} = 0$
- Two-tailed
  - $H_1 : r_{population} \neq 0$
  - There is a significant relationship between brain volume and weekly activity.
  - As brain volume changes, weekly activity changes.
- One-tailed
  - $H_1 : r_{population} > 0$ OR $r_{population} < 0$
  - As weekly activity increases, brain volume increases.
  - As weekly activity increases, brain volume decreases.

Significance of a Correlation

- We want to know whether a correlation is **significant**
  - i.e., whether the probability of finding it by chance is low enough

- Cardinal rule in NHST: compare everything to chance
- Let's investigate by examining the range of r values we expect from random data

Random Correlations

- **Step 1:** Pick two random sets of numbers

Random Correlations

- **Step 1:** Pick two random sets of numbers

```
x <- runif(10, min=0, max=100)
y <- runif(10, min=0, max=100)
head(cbind(x,y))
```

```
##             x      y
## [1,]  1.223 14.537
## [2,] 13.186  7.402
## [3,] 13.800 45.028
## [4,] 55.523 50.858
## [5,] 19.738 36.407
## [6,] 29.011 82.642
```

Random Correlations

- **Step 1:** Pick two random sets of numbers

```
x <- runif(10, min=0, max=100)
y <- runif(10, min=0, max=100)
head(cbind(x,y))
```

```
##           x      y
## [1,]  1.223 14.537
## [2,] 13.186  7.402
## [3,] 13.800 45.028
## [4,] 55.523 50.858
## [5,] 19.738 36.407
## [6,] 29.011 82.642
```

- **Step 2:** Run a correlation

```
cor(x,y)
```

```
## [1] 0.6615
```

Random Correlations

- **Step 1:** Pick two random sets of numbers

```
x <- runif(10, min=0, max=100)
y <- runif(10, min=0, max=100)
head(cbind(x,y))
```

```
##            x      y
## [1,]  1.223 14.537
## [2,] 13.186  7.402
## [3,] 13.800 45.028
## [4,] 55.523 50.858
## [5,] 19.738 36.407
## [6,] 29.011 82.642
```

- **Step 2:** Run a correlation

```
cor(x,y)
```

```
## [1] 0.6615
```

- **Step 3:** Repeat. A lot.

Random Correlations

- **Step 3:** Repeat. A lot.

```
randomCor <- function(size) {
  x <- runif(size, min=0, max=100)
  y <- runif(size, min=0, max=100)
  cor(x,y) # calculate r
}

# then we can use the usual trick:
rs <- data.frame(corrDat =
                   replicate(1000, randomCor(5)))
head(rs)
```

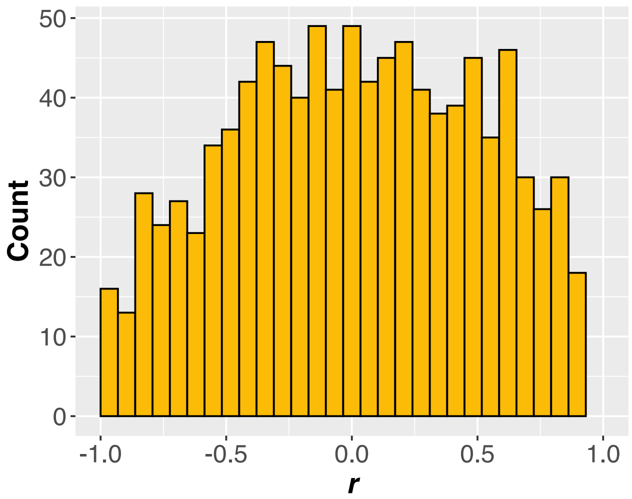
```
##    corrDat
## 1 -0.42851
## 2 -0.05464
## 3 -0.70928
## 4  0.80346
## 5 -0.34746
## 6 -0.08687
```

Random Correlations

- **Step 3:** Repeat. A lot.

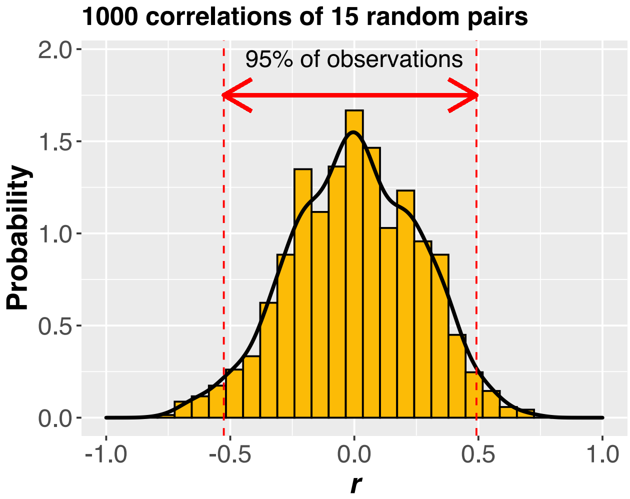
```
randomCor <- function(size) {
  x <- runif(size, min=0, max=100)
  y <- runif(size, min=0, max=100)
  cor(x,y) # calculate r
}

# then we can use the usual trick:
rs <- data.frame(corrDat =
                   replicate(1000, randomCor(5)))
head(rs)
```

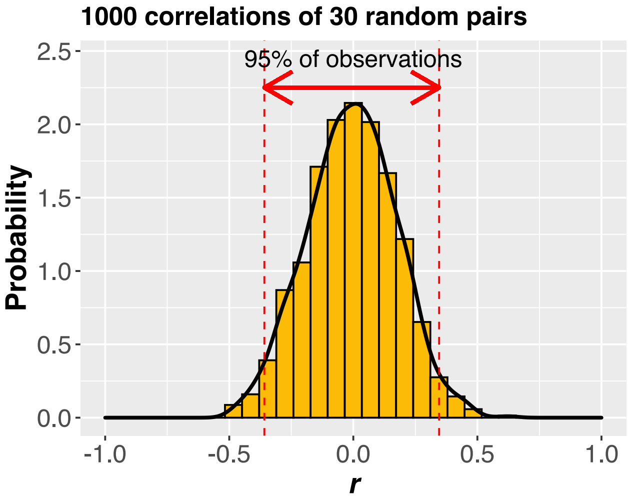
```
##    corrDat
## 1 -0.42851
## 2 -0.05464
## 3 -0.70928
## 4  0.80346
## 5 -0.34746
## 6 -0.08687
```



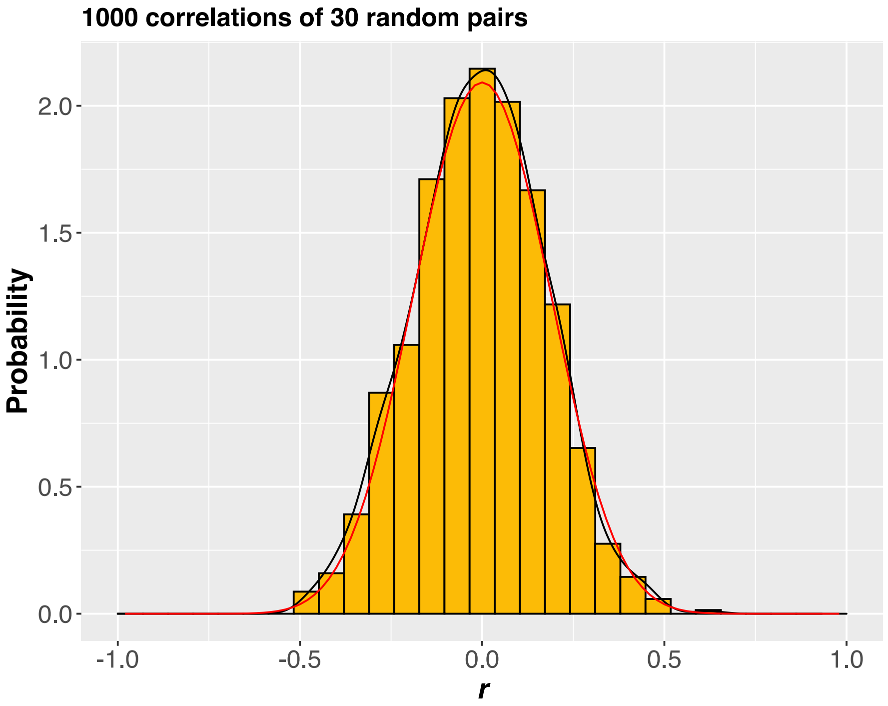
Random Correlations

- Extreme scores are less common, so the distribution narrows as more observations are added.





The *t* distribution



- The distribution of random rs is the t distribution, with $n-2$ df
- This formula computes the corresponding t statistic for the observed r value

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

- Allows you to calculate the probability of getting a value **equal to or more extreme than** r for sample size n by chance

Correlation in R

- In R, you can get the correlation value alone:

```
cor(bvAl$weekly_actv, bvAl$brain_vol)
```

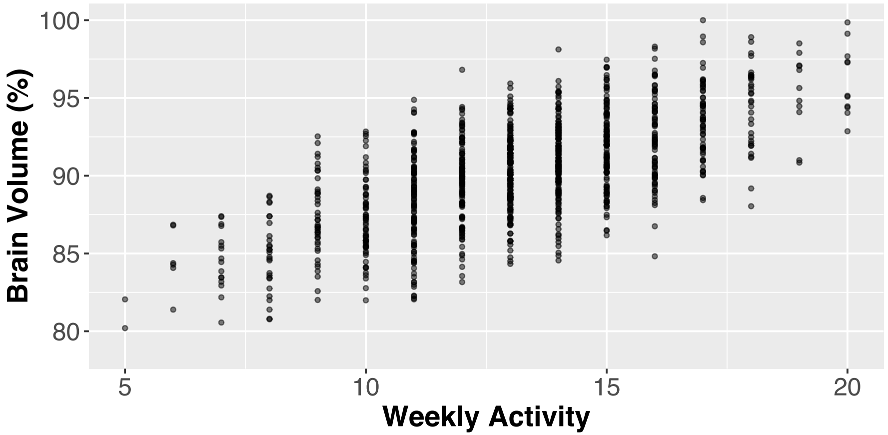
```
## [1] 0.6874
```

- ...or you can get the full results from a t -test of your correlation:

```
cor.test(bvAl$weekly_actv, bvAl$brain_vol)
```

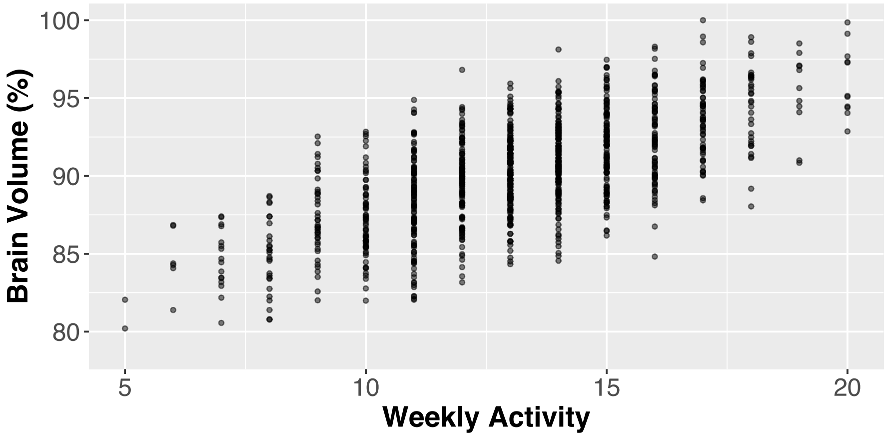
```
##
##     Pearson's product-moment correlation
##
## data:  bvAl$weekly_actv and bvAl$brain_vol
## t = 30, df = 998, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6532 0.7188
## sample estimates:
##    cor
## 0.6874
```

Reporting Correlation Results



"There was a positive relationship between weekly activity level and brain volume, $r(998) = 0.69, p < .001$."

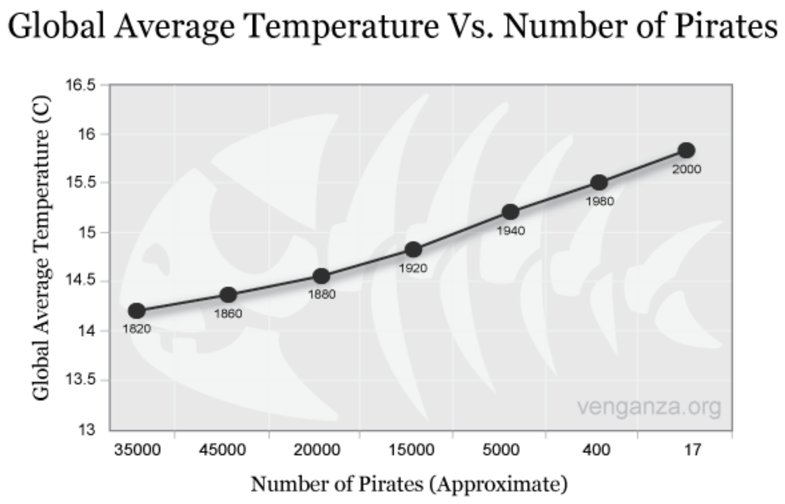
Reporting Correlation Results



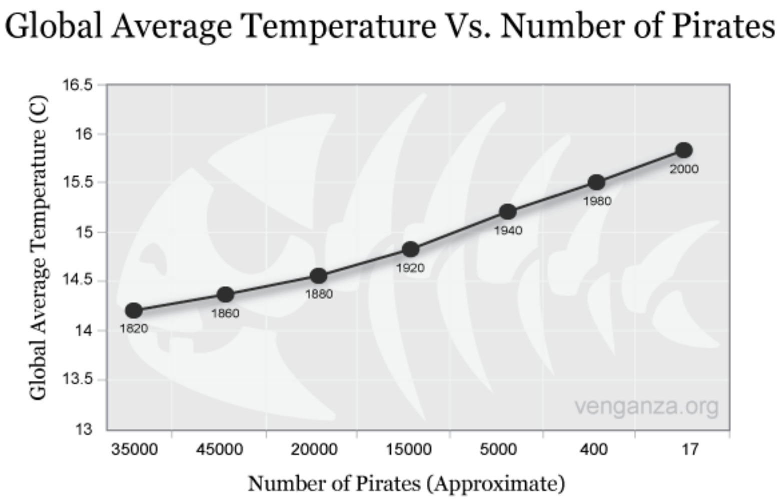
"There was a positive relationship between weekly activity level and brain volume, $r(998) = 0.69, p < .001$."

- Note the lack of causal language!
  - CANNOT SAY "An increase in weekly activity *leads to* an increase in brain volume."

Pirates and Global Warming

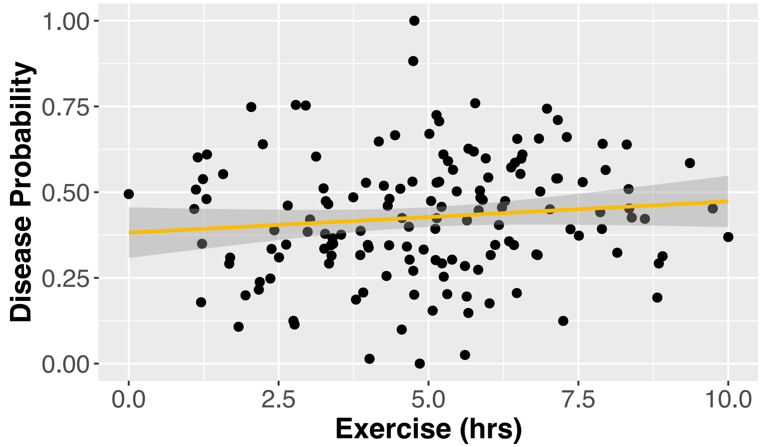


Pirates and Global Warming



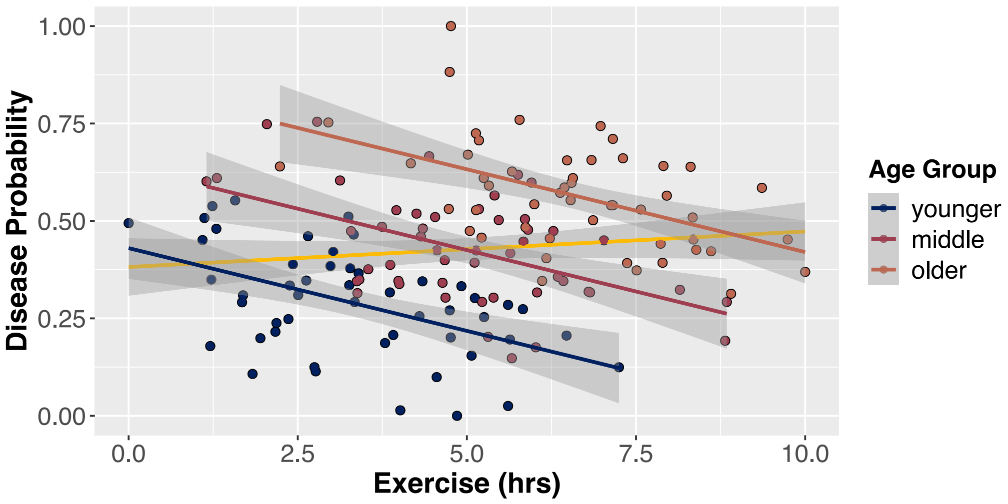
- Clear *negative* correlation between number of pirates and mean global temperature
- We need pirates to combat global warming

Simpson's Paradox



- The more hours of exercise, the greater the risk of disease

Simpson's Paradox



- Age groups mixed together
- An example of a *mediating variable*

Interpreting Correlation

- Correlation does not imply causation
- Correlation simply suggests that two variables are related
  - There may be mediating variables
- Interpretation of that relationship is key
- Never rely on statistics such as r without
  - Looking at your data
  - Thinking about the real world

Part 3

Putting it all Together

Has Statistics Got You Frazzled?

- We've bandied a lot of terms around in quite a short time
- We've tended to introduce them by example
- Time to step back...

What is NHST all about?

Null Hypothesis Statistical Testing

- Two premises

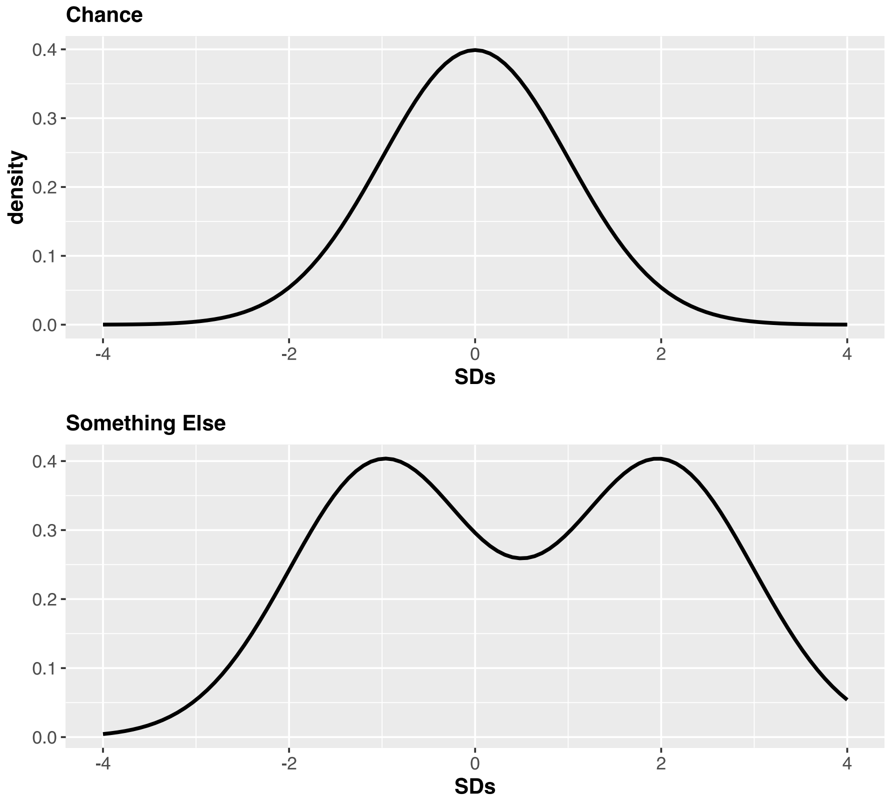
    1. Much of the variation in the universe is due to *chance*

    2. We can't *prove* a hypothesis that something else is the cause

Chance

- When we say *chance*, what we really mean is "stuff we didn't measure"
- We believe that "pure" chance conforms approximately to predictable patterns (like the normal and t distributions)
    - If our data isn't in a predicted pattern, perhaps we haven't captured all of the non-chance elements

Patterns attributable to



Proof

- Can't prove a hypothesis to be true
- "The sun will rise tomorrow"

Proof

- Can't prove a hypothesis to be true
- "The sun will rise tomorrow"
- *Just takes one counterexample*

Chance and Proof

If the likelihood that the pattern of data we've observed would be found *by chance* is low enough, propose an alternative explanation

- Work from summaries of the data (e.g., $\bar{x}$, σ)
- Use these to approximate chance (e.g., t distribution)

Chance and Proof

If the likelihood that the pattern of data we've observed would be found *by chance* is low enough, propose an alternative explanation

- Work from summaries of the data (e.g., $\bar{x}$, σ)
- Use these to approximate chance (e.g., t distribution)
  - Catch: we can't estimate the probability of an exact value (this is an example of the measurement problem)
  - Estimate the probability of finding the measured difference *or more*

Alpha and Beta

- We need an agreed "standard" for proposing an alternative explanation
  - Typically in psychology, we set α to 0.05
  - "If the probability of finding this difference or more under chance is α (e.g., 5%) or less, propose an alternative"

Alpha and Beta

- We need an agreed "standard" for proposing an alternative explanation
  - Typically in psychology, we set α to 0.05
  - "If the probability of finding this difference or more under chance is α (e.g., 5%) or less, propose an alternative"
- We also need to understand the quality of evidence we're providing
  - Can be measured using β
  - **power** = 1 - β
  - Psychologists typically aim for $\beta = 0.20$ (i.e., a power level of 80%)
  - "Given that an effect truly exists in a population, what is the probability of finding $p < \alpha$ in a sample (of size n etc.)?"

The Rest is Just Nuts and Bolts

- Type of measurement
- Relevant laws of chance
- Suitable estimated distribution (normal, t, χ^2, etc.)
- Suitable summary statistic (z, t, χ^2, r, etc.)
- Use statistic and distribution to calculate p and compare to α
- Rinse, repeat

End