Basic statistical review

Chi-square test, t-test, correlations

Recap

Basic statistical review

Chi-square test, t-test, correlations

Lung cancer epidemiology

- 1922: 617 deaths from lung cancer in the UK
- 1947: 9287 deaths from lung cancer in the UK

He's one of the busiest men in town. While his door may say *Office Hours 2 to 4*, he's actually on call 24 hours a day.
The doctor is a scientist, a diplomat, and a friendly sympathetic human being all in one, no matter how long and hard his schedule.
According to a recent Nationwide survey:
MORE DOCTORS SMOKE CAMELS THAN ANY OTHER CIGARETTE
DOCTORS in every branch of medicine—113,597 in all—were queried in this nationwide study of cigarette preference. Three leading research organizations made the survey. The gist of the query was—What cigarette do you smoke, Doctor?
The brand named most was Camel!
The rich, full flavor and cool mildness of Camel's superb blend of costlier tobaccos seem to have the same appeal to the smoking tastes of doctors as to millions of other smokers. If you are a Camel smoker, this preference among doctors will hardly surprise you. If you're not—well, try Camels now.
CAMEL
Your "T-Zone" Will Tell You...
T for Taste...
T for Throat...
that's your proving ground for any cigarette. See if Camels don't suit your "T-Zone" to a "T."
CAMELS

Doll and Hill's study

BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

Doll and Hill’s study

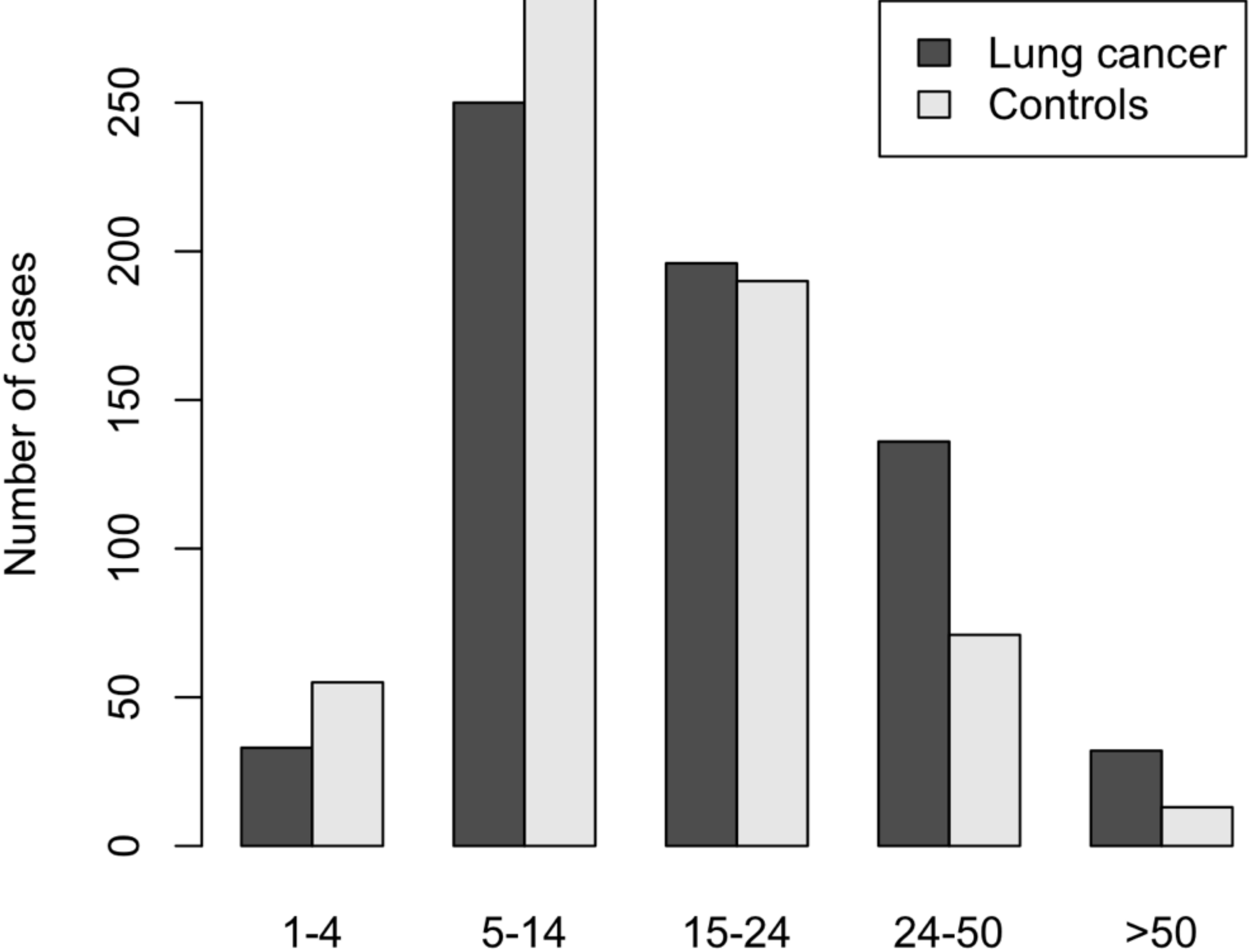
- April 1948 to October 1949 2370 cases reported
- 150 over 75 and 80 incorrectly diagnosed
- 408 could not be interviewed for various reasons

Doll and Hill's study

- 1723 patients with carcinoma interviewed
- 743 general medical or surgical patients interviewed as controls for lung cancer patients
- Some carcinoma patients later proved to be misdiagnosed

between lung cancer patients and controls

Disease Group	No. Smoking Daily				
	1 Cig.–*	5 Cigs.–	15 Cigs.–	25 Cigs.–	50 Cigs. +
Males:					
Lung-carcinoma patients (647)	33 (5·1%)	250 (38·6%)	196 (30·3%)	136 (21·0%)	32 (5·0%)
Control patients with diseases other than cancer (622)..	55 (8·8%)	293 (47·1%)	190 (30·5%)	71 (11·4%)	13 (2·1%)



Lung cancer
Controls
Number of cases
0
50
100
150
200
250
1-4
5-14
15-24
24-50
>50

Contingency table

Number of cigs	1-4	5-14	15-24	25-49	50+
Lung cancer	33	250	196	136	32
Control	55	293	190	71	13

Contingency table

Number of cigs	1-4	5-14	15-24	25-49	50+	Total
Lung cancer	33	250	196	136	32	647
Control	55	293	190	71	13	622
Total	88	543	386	207	45	1269

Expected value

For each cell in the table, the expected value is:

$$\frac{\text{Column total X Row total}}{\text{Grand total}}$$

Contingency table

Number of cigs	1-4	5-14	15-24	25-49	50+	Total
Lung cancer	33	250	196	136	32	647
Control	55	293	190	71	13	622
Total	88	543	386	207	45	1269

Expected value for highlighted cell is:

$\underline{88 \times 647}$ 44.87

Deviation from expected

$$33 - 44.87 = -11.87$$

We can calculate this for each cell in the table and get an indication of the extent by which the observed values deviate from the expected ones

Deviation from expected

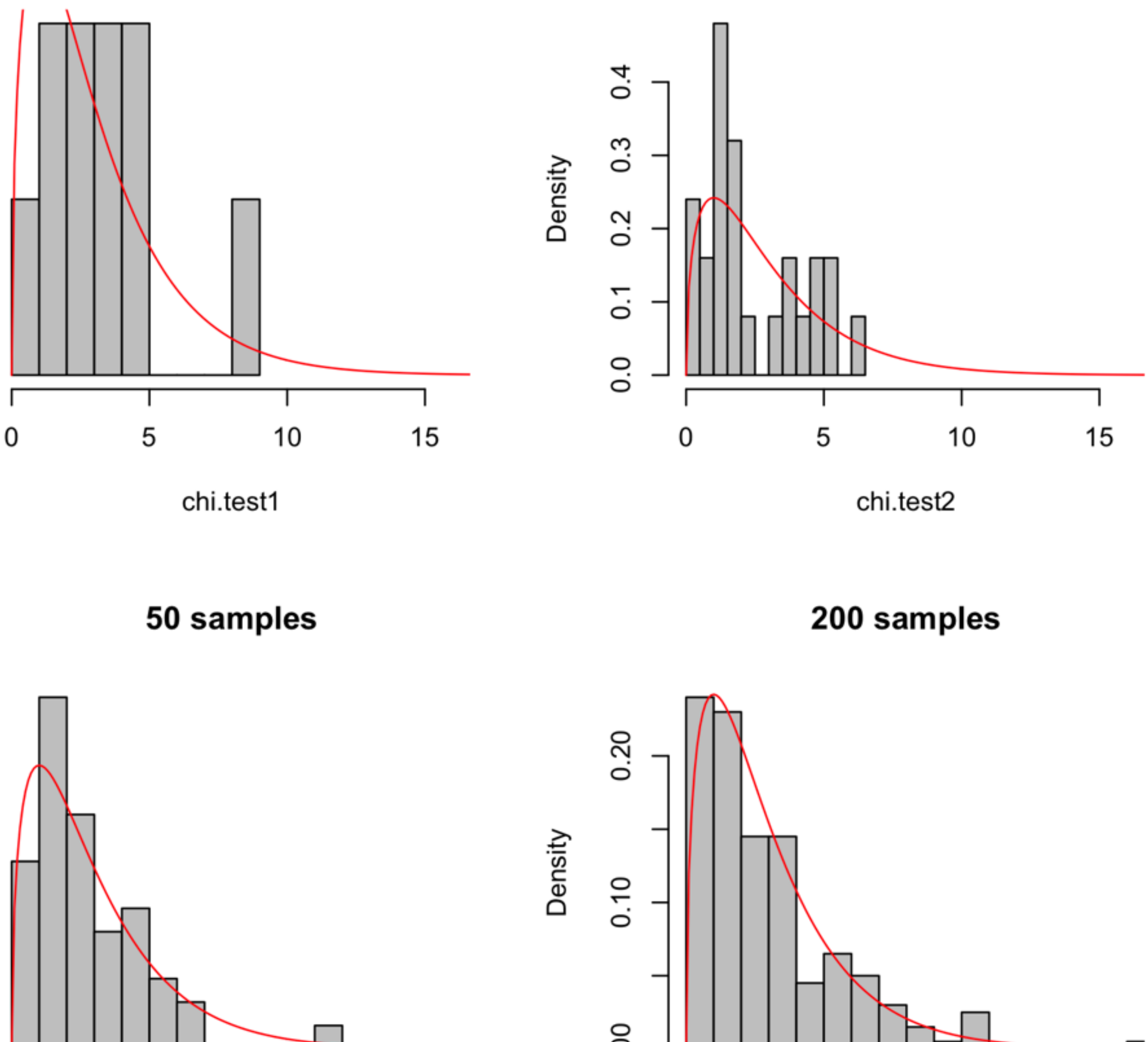
Number of cigs	1-4	5-14	15-24	25-49	50+
Lung cancer	-11.87	-26.84	-0.8	30.46	9.06
Control	11.87	26.84	0.8	-30.46	-9.06

Deviation from the expected

For each cell, we calculate:

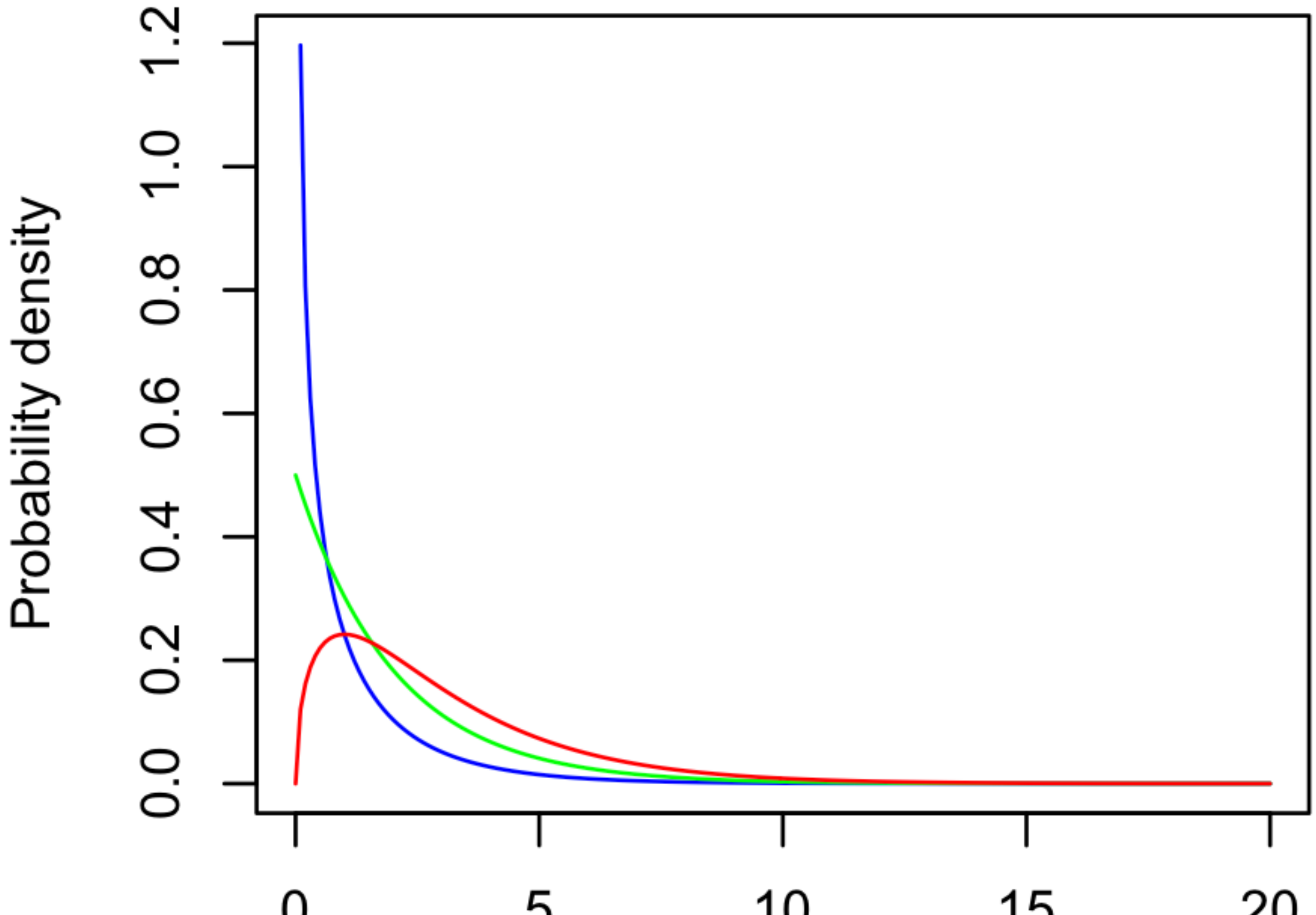
$$\frac{(\text{observed-expected})^2}{\text{expected}}$$

The sum of these is 36.95

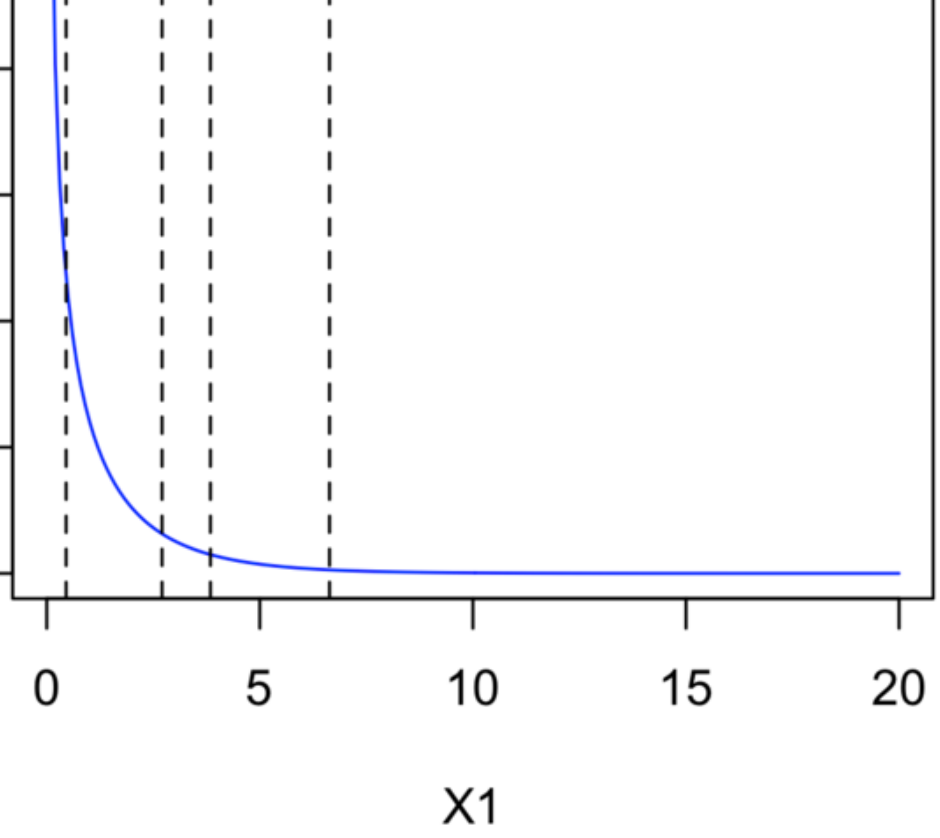


200 random draws from standard normal distribution, squared and summed and compared to Chi-square probability distribution

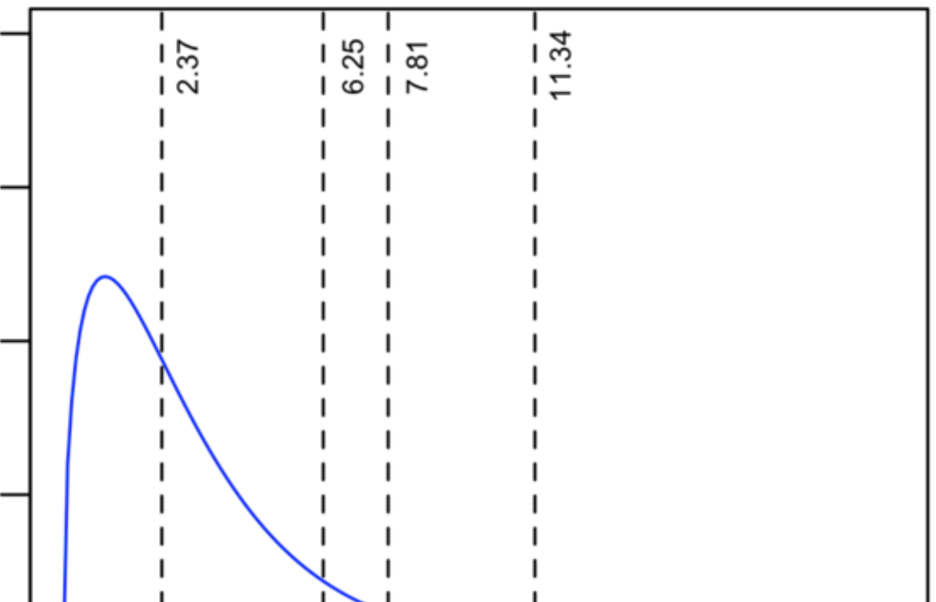
Chi-squared distribution



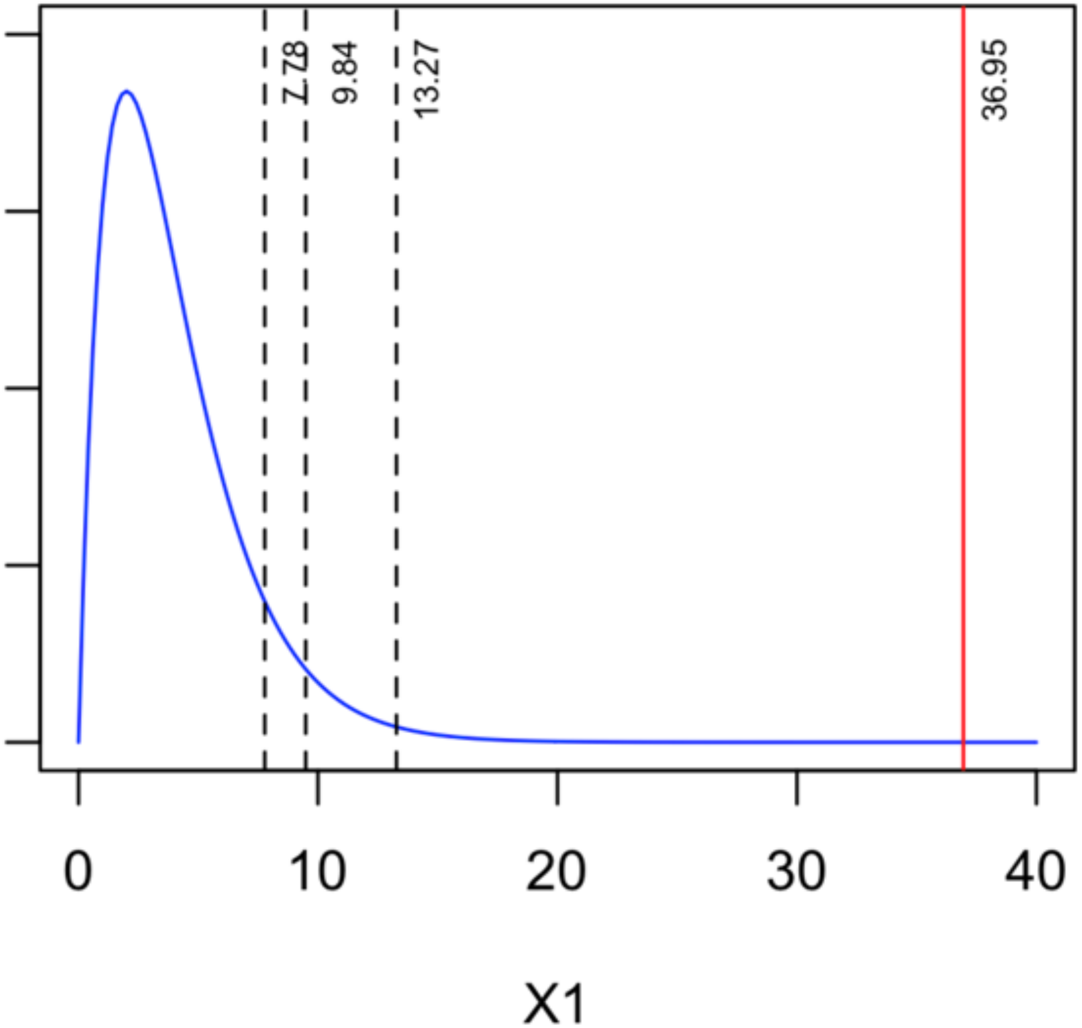
Blue = 1df
Green = 2 df
Red = 3 df



Probability density function for Chi square
distribution with 1 df. 50%, 90%, 95% and
quantiles are marked.



Probability density function for Chi square
distribution with 3 df. 50%, 90%, 95% and
quantiles are marked.



Probability density function for Chi squa
distribution with 4 df. 90%, 95% and 99
quantiles are marked. The test statistic
our lung cancer analysis is in red.

BLE V.—*Most Recent Amount of Tobacco* Consumed Regula[r] by Smokers Before the Onset of Present Illness; Lung-carcino[ma] Patients and Control Patients with Diseases Other Than Canc[er]*

[D]isease Group	No. Smoking Daily					Probabilit[y] Test
	1 Cig.–*	5 Cigs.–	15 Cigs.–	25 Cigs.–	50 Cigs. +	
[M]ales:						
[Lu]ng-carcinoma [p]atients (647)	33 (5·1%)	250 (38·6%)	196 (30·3%)	136 (21·0%)	32 (5·0%)	$\chi^2 = 36{\cdot}95$ n=4; $P < 0{\cdot}001$
[Co]ntrol patients [w]ith diseases [o]ther than [c]ancer (622) . .	55 (8·8%)	293 (47·1%)	190 (30·5%)	71 (11·4%)	13 (2·1%)	

in R:

use `chisq.test`

```
> lungs<-
matrix(data=c(33,250,196,136,32,55,293,190,71,13),
byrow=TRUE, nrow=2)

> chisq.test(lungs)

	Pearson's Chi-squared test

data:  lungs
X-squared = 36.953, df = 4, p-value = 1.842e-07
```

Basic statistical review

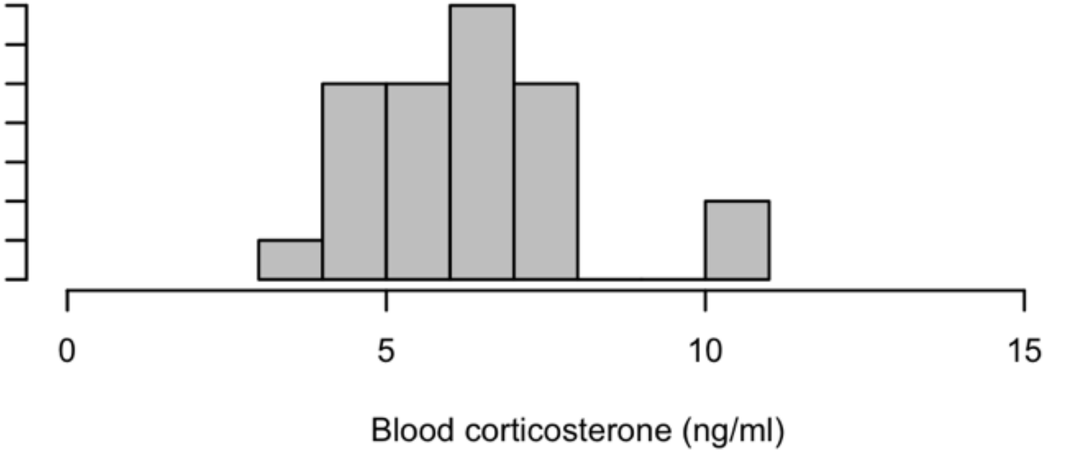
Chi-square test, **t-test**, correlations

A test of corticosterone levels in rutting stags

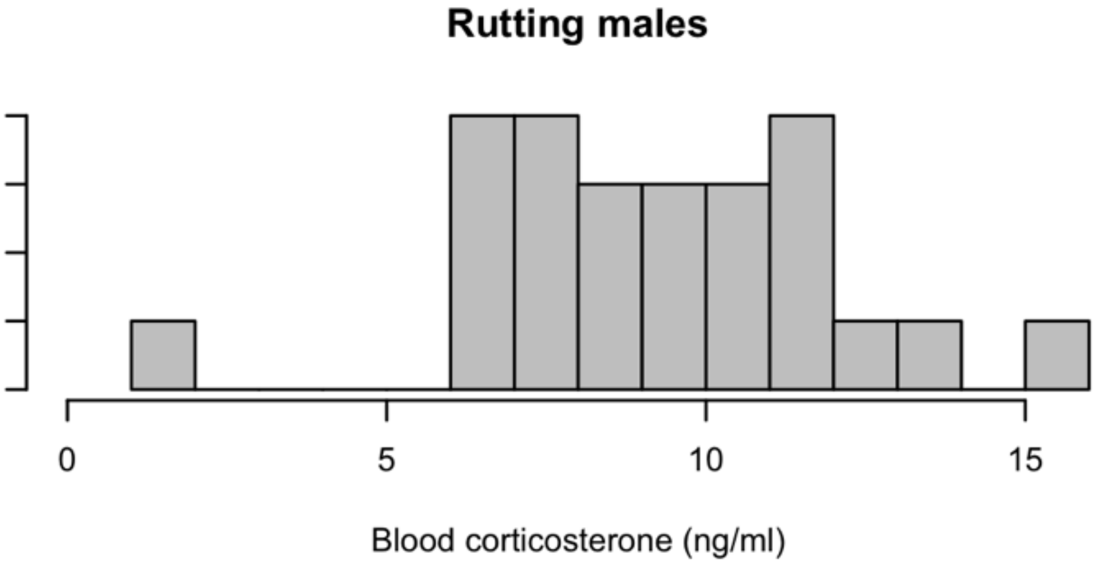
Study to investigate whether rutting causes elevated corticosterone levels in stags, and what levels of corticosterone are found in does

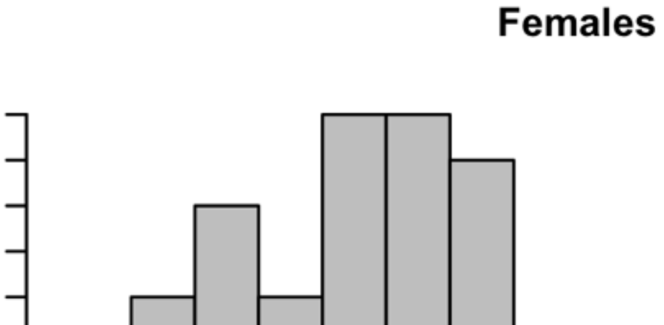
Stags darted and blood samples taken before and during the rut, from the same 25 individuals

Does darted and blood samples taken during the rut



Initial data exploration





Sample	Mean (ng.ml)	Varianc
Non-rutting males	6.38	2.78
Rutting males	9.43	8.42
Females	4.49	2.82

Initial data exploration

All samples are roughly normally distributed

Rutting males>non-rutting males>females

Is the difference in means between rutting and non-rutting males statistically significant?

Testing the differences between males

We have sampled each male twice, before and during the rut

Therefore, we can express the change in corticosterone titre for each male as the second measurement minus the first

Testing the differences between males

```
er$Males.in.rut-deer$Males
  0.123  3.815 -0.403  8.179  6.266  5.849
3  4.492  4.894  2.734  0.733  7.589  5.03
9
 -1.338 -3.132 -5.721  5.125  1.555  9.207
1  0.390  0.746  0.862  8.742

an(deer$Males.in.rut-deer$Males)
3.09616

rt(var(deer$Males.in.rut-deer$Males))
```

Statistical testing recap

- We're trying to calculate the probability that our value for the mean could arise by random error when sampling from a population with a mean of zero
- Null Hypothesis:
- There is no difference between samples. The differences between samples are drawn from a population with a mean of zero
- Alternative hypothesis:
- There is a difference between samples. The differences between samples are drawn from a population with a mean not equal to zero

Calculating t

- We can do our statistical test if we calculate a value called "t", which is defined as the difference in means divided by the standard error
- The mean difference between our two samples is 3.10, s=3.85 so

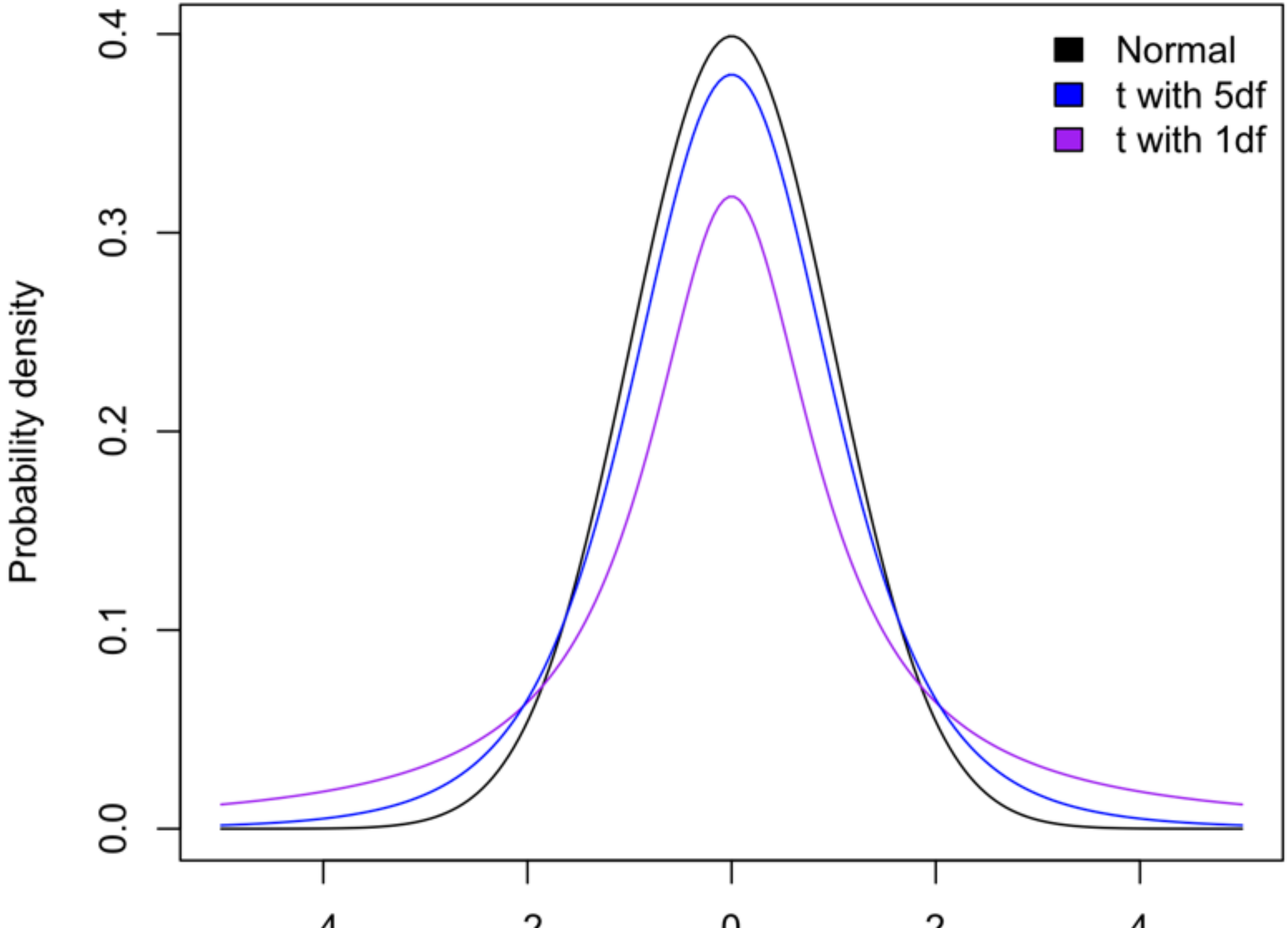
$$t = \frac{X - \mu}{s/\sqrt{N}} \qquad t = \frac{3.10 - 0}{3.85/\sqrt{25}} \qquad t = 3.10/0.77 = 4.02$$

Testing our mean

Using the estimate of the standard deviation causes problems because it will lead to systematic ***underestimation*** of σ

This is solved by comparing our value of t with ***Student's t distribution***, which takes account of this

Student's t-distribution



Testing our means

Our value of is t 4.02 with 24 df

The critical value of t for a 2-tailed test at 24 df is 2.064

Therefore we **reject** H_0 and **accept** H_1

Rutting stags have significantly higher corticosterone titres than the same stags sampled before the rut

in R:

use t.test

```
> t.test(deer$Males.in.rut,deer$Males, paired=TRUE)

	Paired t-test

data:  deer$Males.in.rut and deer$Males
t = 4.0203, df = 24, p-value = 0.0005005
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 1.506704 4.685616
sample estimates:
mean of the differences
               3.09616
```

Basic statistical review

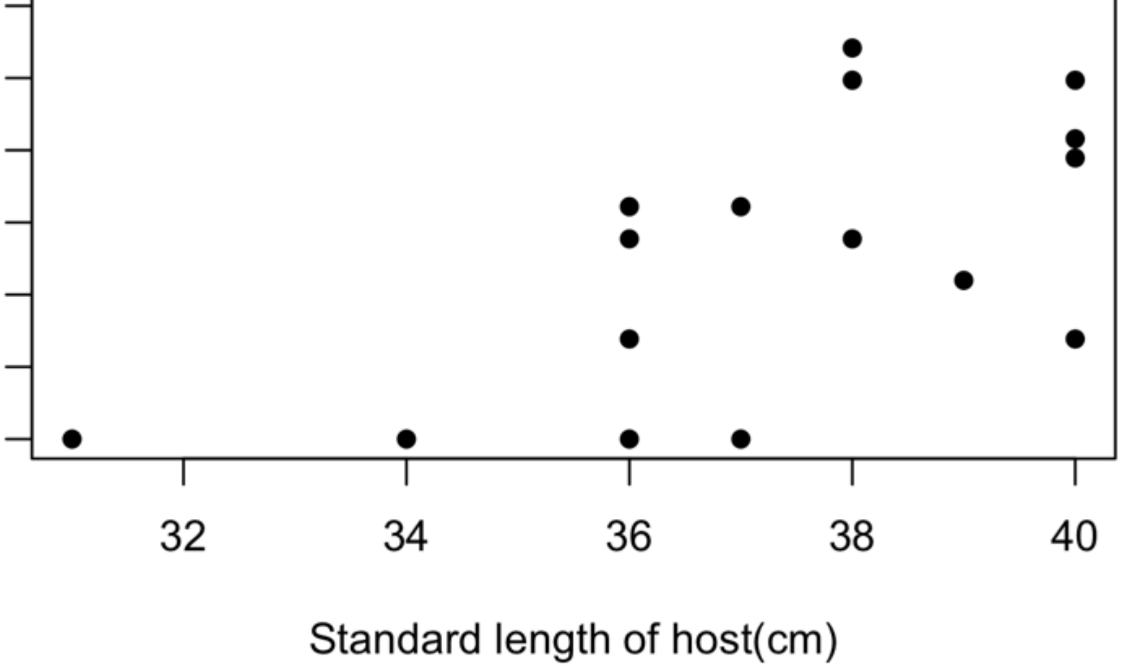
Chi-square test, t-test, **correlations**

2 types of statistical test

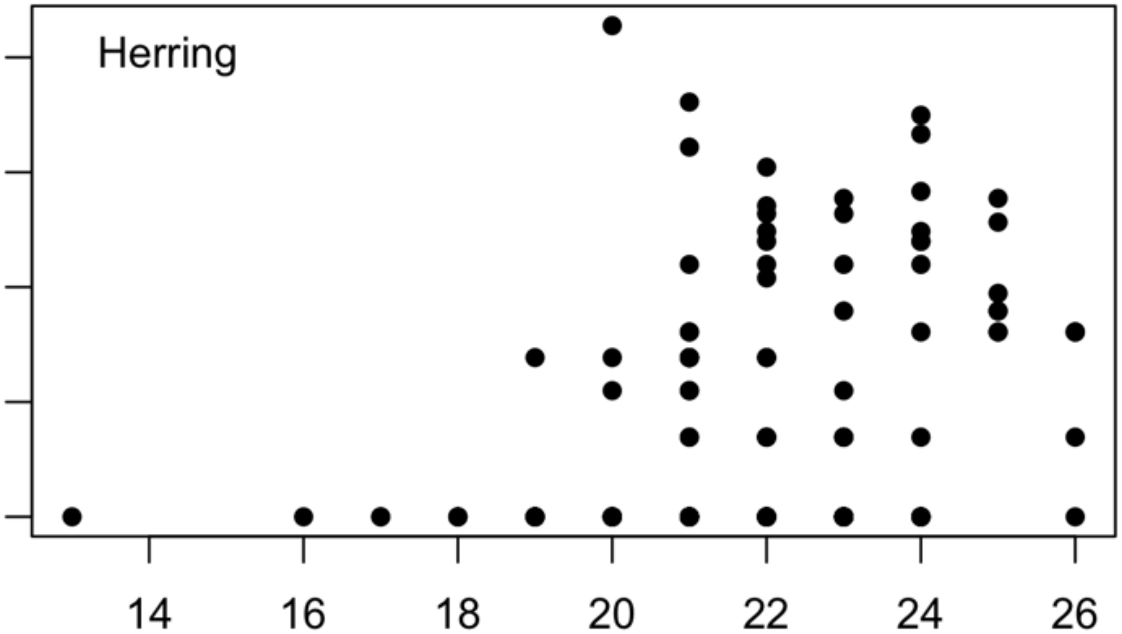
- Comparisons
- Relationships

Correlation

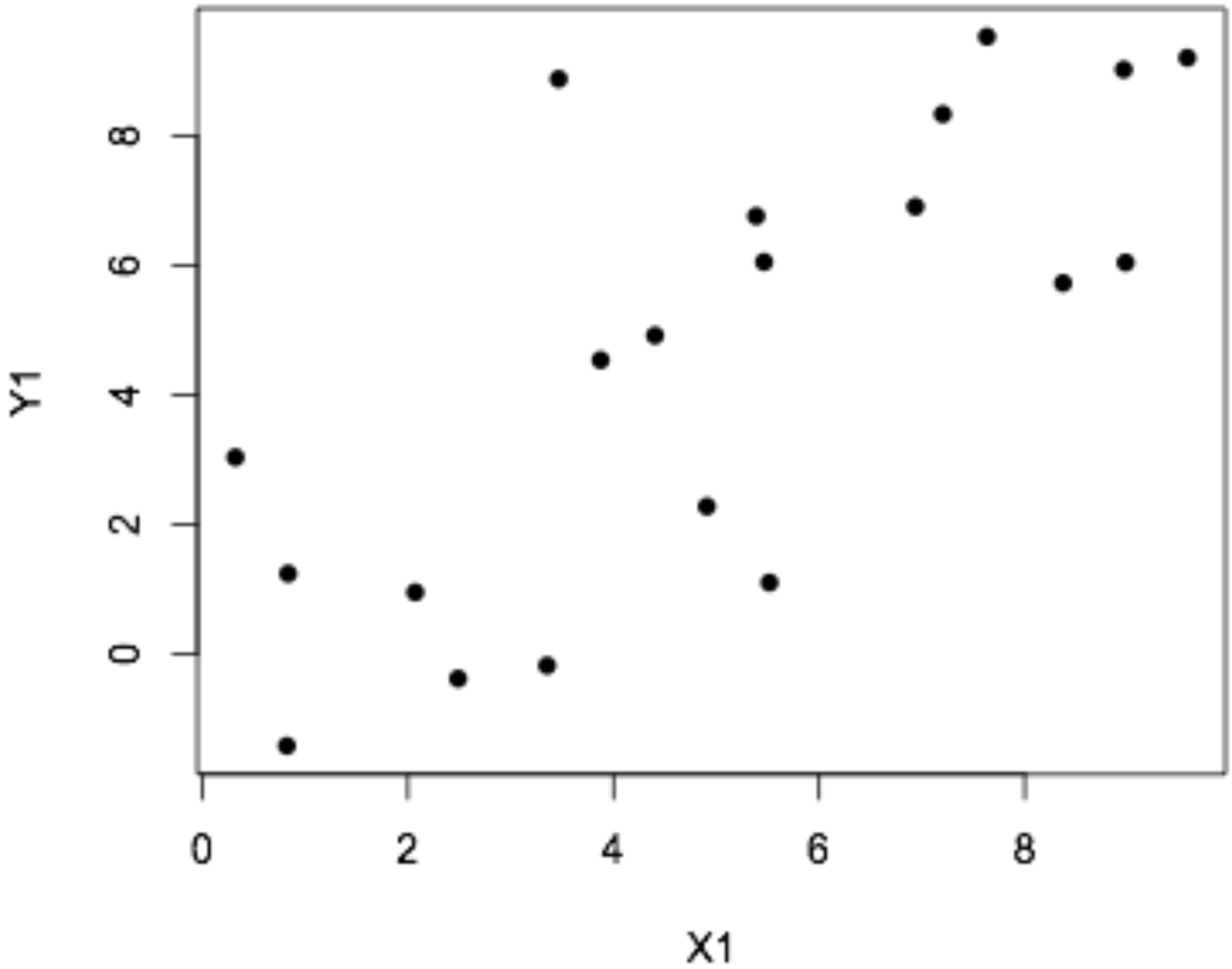
- Relationships or associations between continuous variables
- Can be positive or negative
- Shows the strength and significance of th relationship between 2 variables



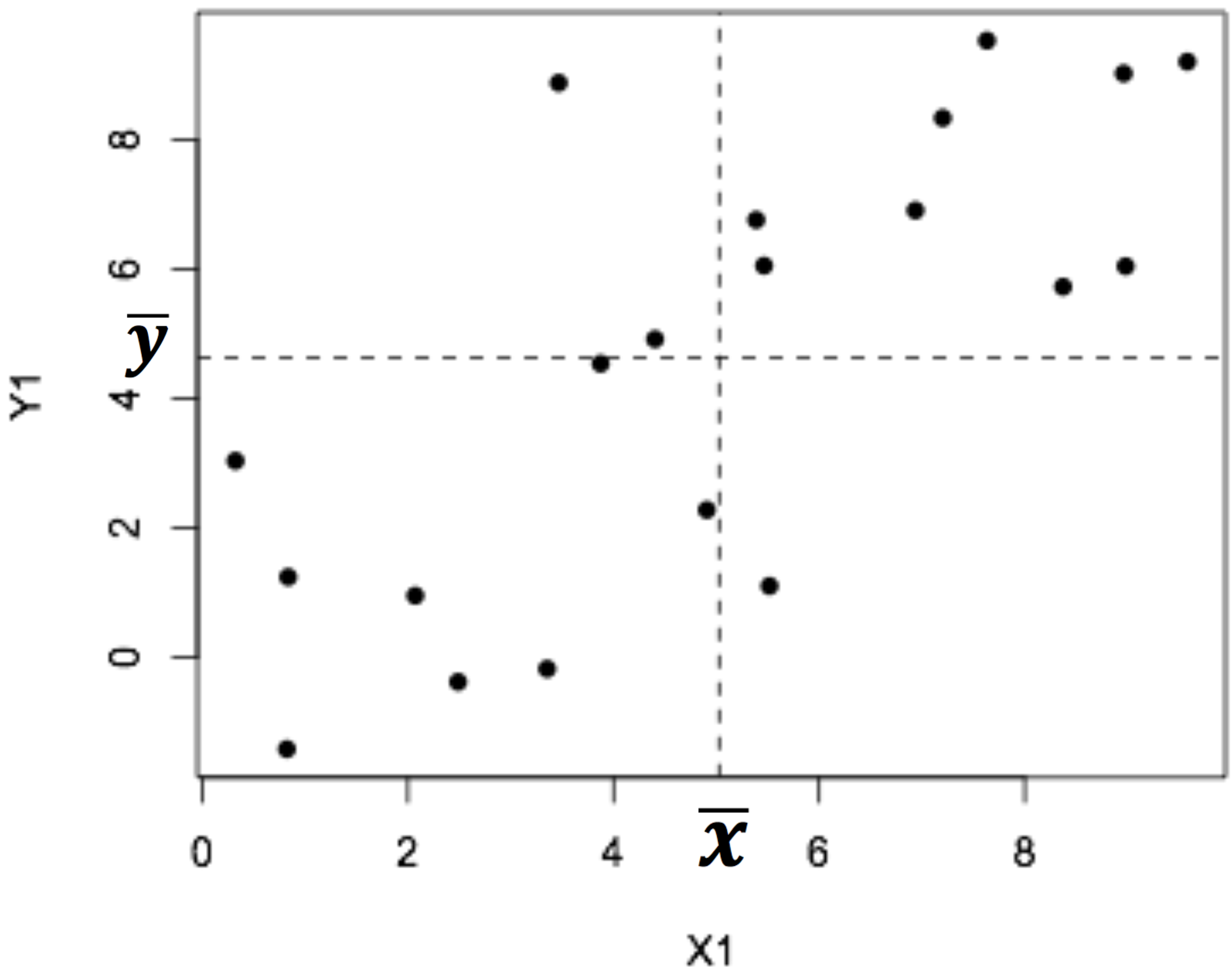
Data from SBS205 pract



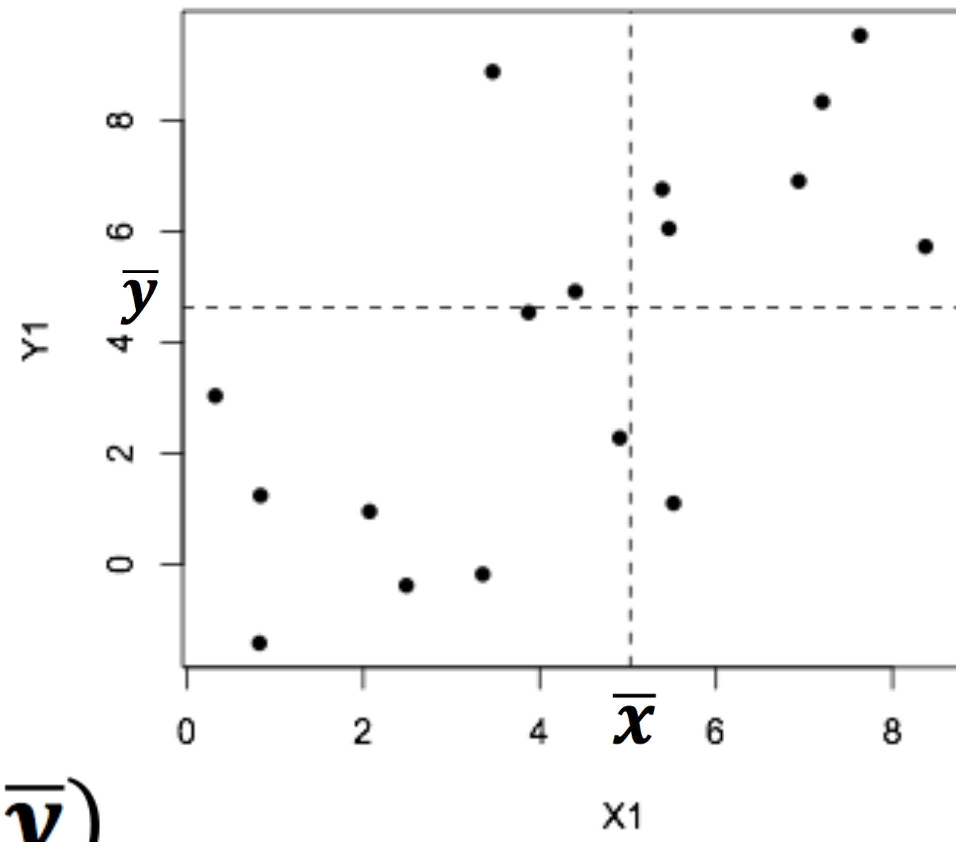
Calculating a correlation coefficient



Calculating a correlation coefficient



Calculating a correlation coefficient

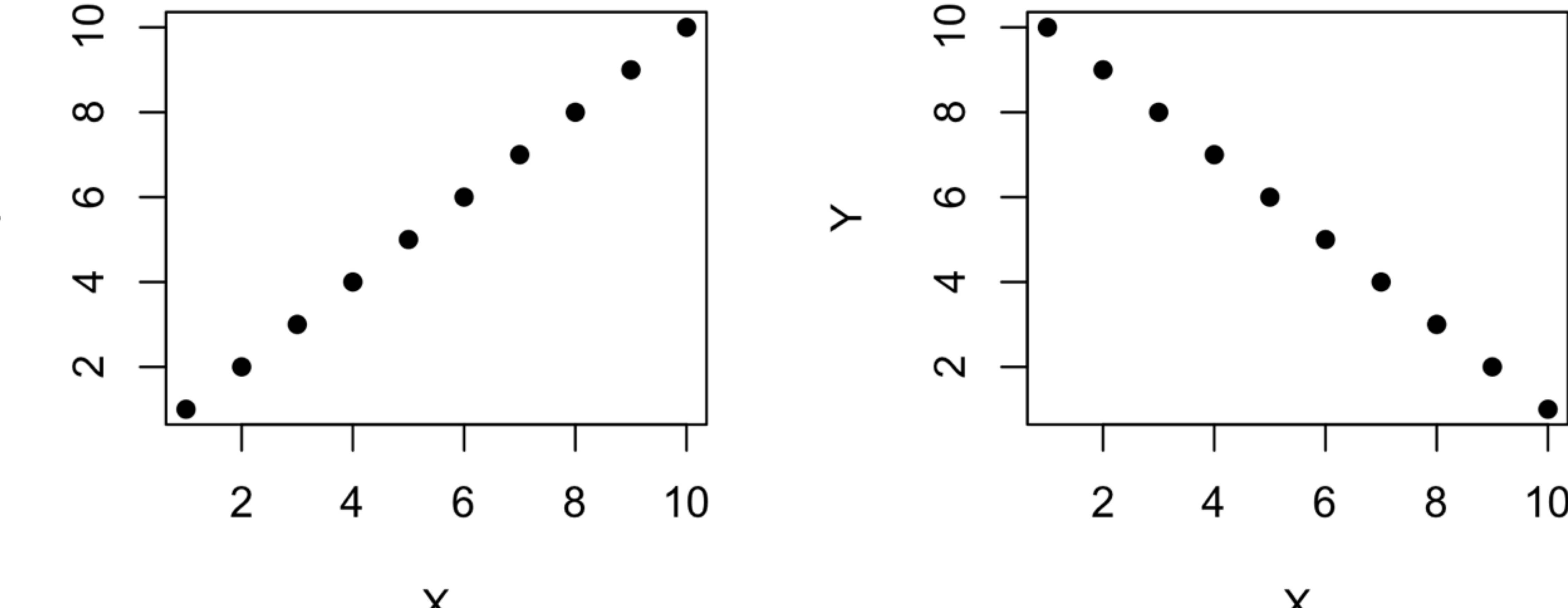


$$variance = \sum \frac{(x - \overline{x})(y - \overline{y})}{n - 1}$$

$$r = \sum \frac{(x - \overline{x})(y - \overline{y}) / _{n-1}}{s_x s_y}$$

Correlation coefficients

- ***r*** falls between +1 and -1



Correlation coefficients: statistical significance

Ho: the two variables are unrelated

Calculate a *t*-statistic and test at n-2df:

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

in R:

use cor.test

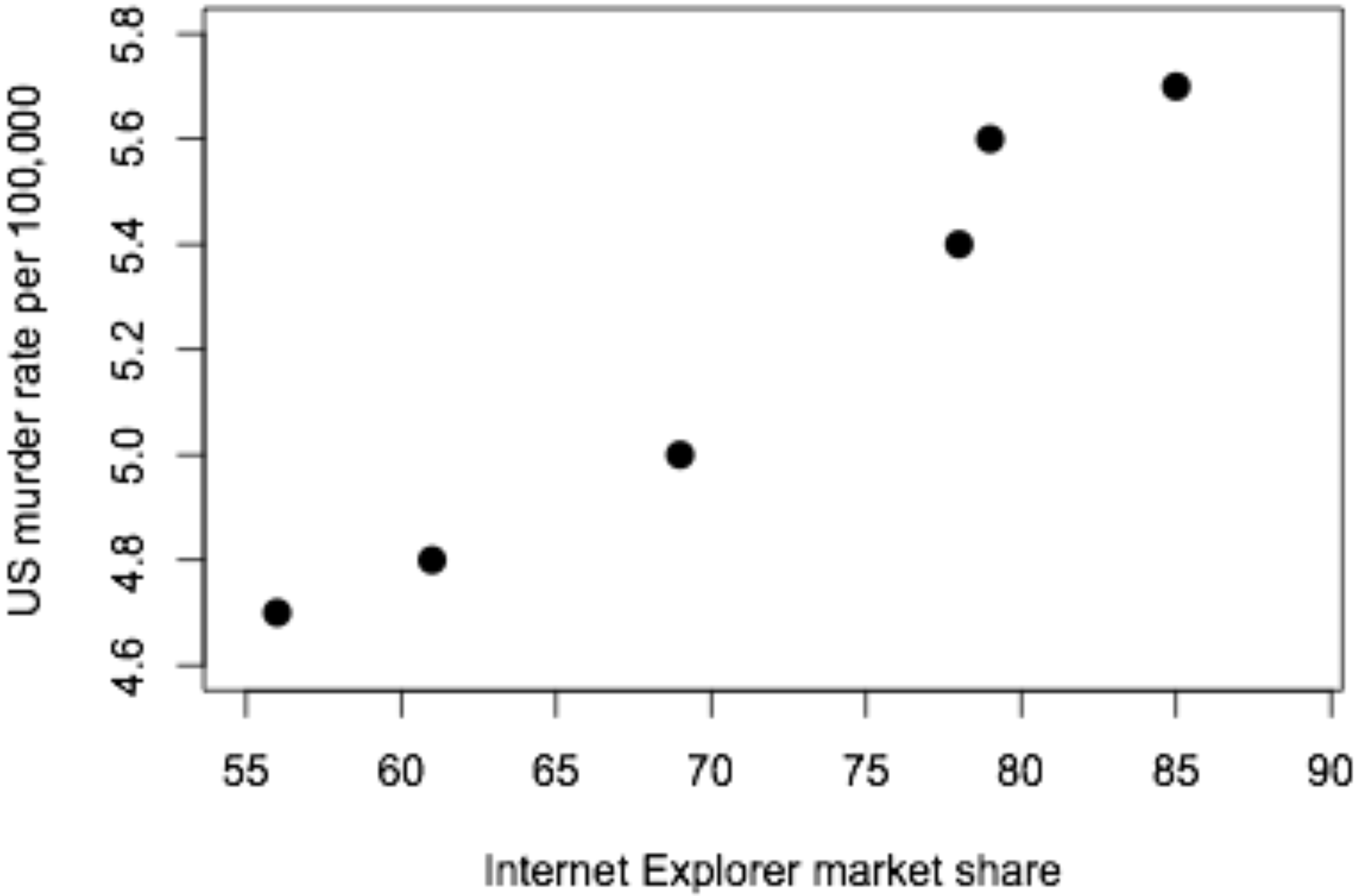
```
Z1<-runif(20,0,10)
Z2<-Z1+rnorm(20,0,2.5)
cor.test(Z1,Z2)

 Pearson's product-moment correlation

ata:  Z1 and Z2
 = 2.7642, df = 18, p-value = 0.01278
lternative hypothesis: true correlation is not equal to 0
5 percent confidence interval:
0.1362879 0.7960970
ample estimates:
     cor
.5458861
```

Correlation coefficients: summary

- r varies between +1 and -1. The closer to or -1, the stronger the correlation
- We can calculate a p-value associated with to allow us to test for a statistically significant correlation
- Coefficient of determination, r^2, is an estimate of the % variability in one variable explained by the other variable

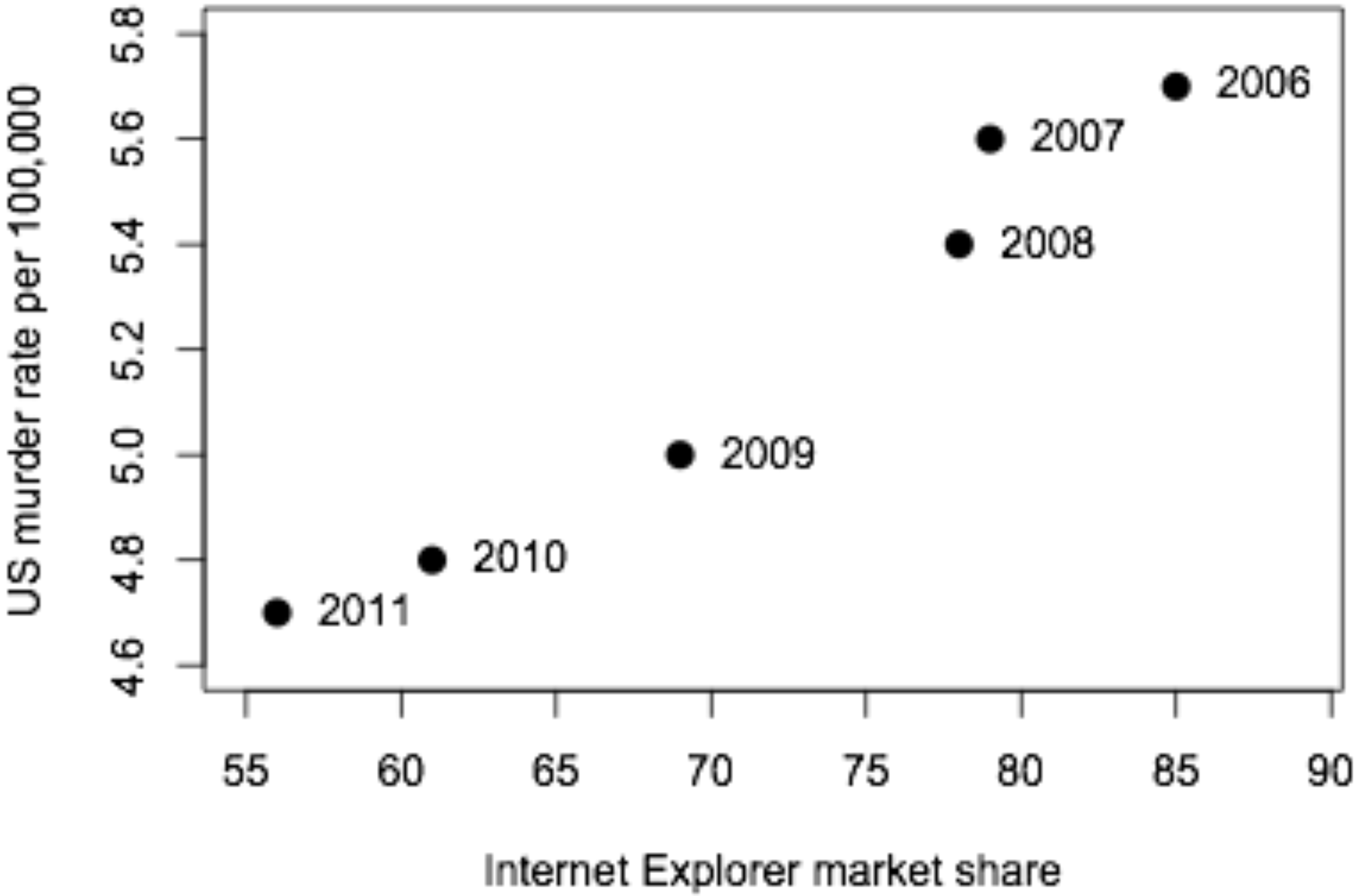
Note: correlation does NOT mean causality

- If two variables are strongly and significantly correlated it does not mean one is the cause of the other



US murder rate per 100,000
4.6
4.8
5.0
5.2
5.4
5.6
5.8
55
60
65
70
75
80
85
90
Internet Explorer market share

```
> cor.test(IE,murder)

  Pearson's product-moment correlati

data:  IE and murder
t = 10.1718, df = 4, p-value =
0.0005261
alternative hypothesis: true
correlation is not equal to 0
95 percent confidence interval:
 0.8329100 0.9980292
sample estimates:
     cor
0.981212
```



A third variable is correlated with both of our variables

ews Front Page
World
UK
England
orthern Ireland
Scotland
Wales
Business
Politics
Health
Education
Science/Nature

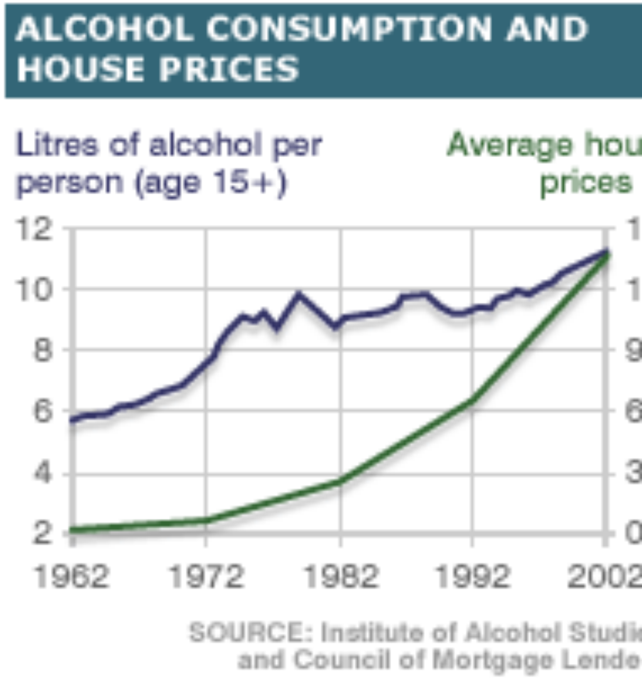
Last Updated: Tuesday, 4 October 2005, 09:43 GMT 10:43 UK

E-mail this to a friend
Printable version

'Binge drinking? Blame house prices'

By Tom Geoghegan
BBC News Magazine

The debate about binge drinking has focused mainly on licensing hours and discounted drinks. But could house prices be to blame?



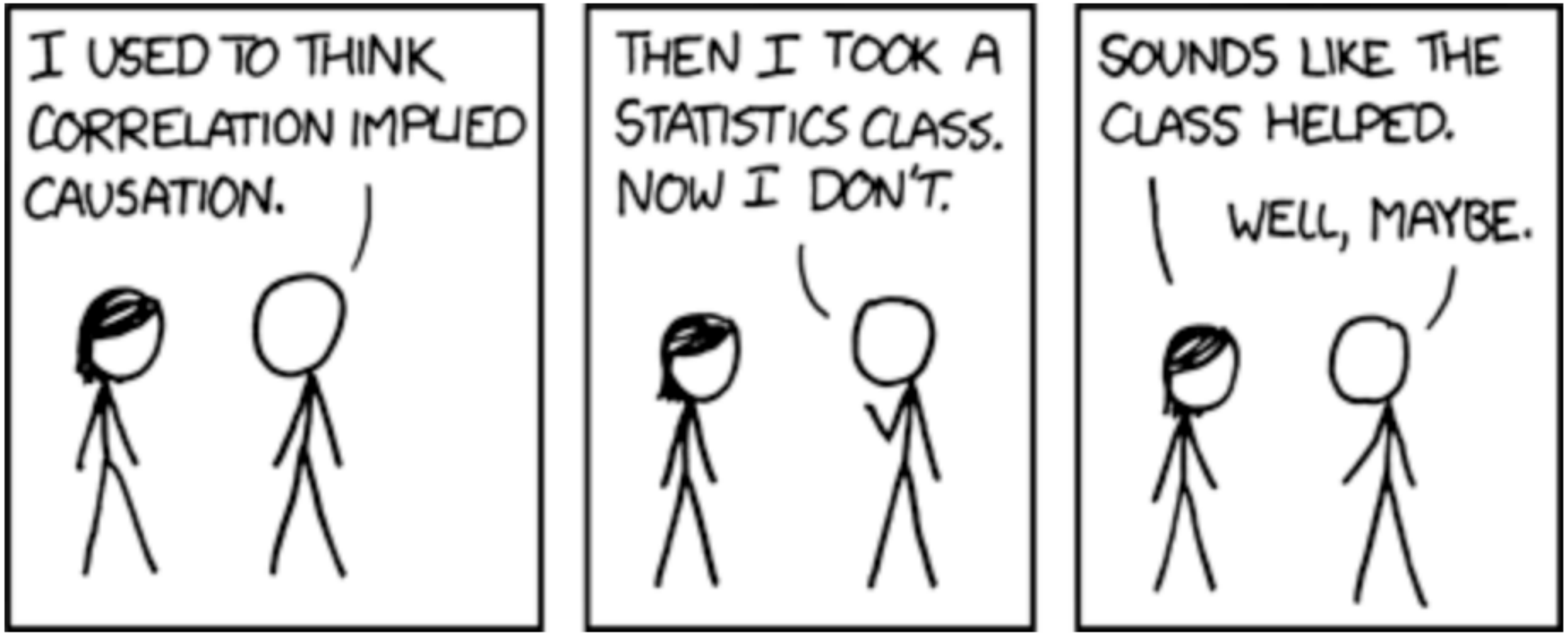
drew McNeil, director of the
stitute of Alcohol Studies,
inks the major factor is the
fordability of alcohol, which
s increased over the years.
hen excise duty has
creased, consumption has
llen, he says.

The drinking culture is highly visible

/e're not advocating
ohibition by price but we're
ying it would be useful to deal properly with heavily
scounted promotions by the on-street trade and
permarkets," he says. "And secondly if excise duties were

e, Professor Cooper spent an
rs on the streets of Manchester.
as £75 and most of the revellers he
nts or in rented shared

be as likely to spend that kind of
because they have a mortgage and
ture and the garden. Plus they are



http://xkcd.com/552/