Exam Revision

Class Quiz

Instructions

- The following quiz will give you a sense of the questions that will be asked in the final exam.
- As you will discover, there are no questions on R coding!
- You will compete in groups for an awesome prize!
- Each slide will present a multiple choice or short answer question
- You will have a limited amount of time to discuss the answer with your group.
- Your group must write the answer on the whiteboard before time is up.
- Each group will be scored and a running tally updated.
- The group with the highest score at the end of the quiz will be the winner.

Question 1

Suppose you take a random sample from the population and calculate the sample mean. Then you repeat this process again. The sample means will differ. This concept is know as

a. a biased sample.
b. natural variability.
c. induced variability.
d. sampling variability.

Question 2

You're taking a walk in your local neighbourhood. You walk past numerous trees planted on the side of the road. All the trees are the same species and were planted at the same time, but all the trees have different sizes, shapes and features. This variability is best described as

a. explainable variation.
b. natural variation.
c. induced variation.
d. sampling error.

Question 3

Hopefully you have learnt a lot about statistics in this course. Which of the following is NOT something I have taught?

a. Sample size matters.
b. Always think about the context behind the data.
c. With large samples, assumptions don't apply to statistical tests.
d. Leave the computation to computers.

Question 4

Which of the following descriptive statistics is said to be robust in the presence of outliers?

a. Median
b. Mean
c. Standard deviation
d. Minimum

Question 5

Which one of the following is a limitation of the box plot?

a. Does not show sample size.
b. Does not visually represent variability.
c. Does not show a measure of central tendency.
d. Difficult to detect skewness to the data.

Dr James Baglin (james.baglin@rmit.edu.au)


Question 6

You have a small sample (n = 20) of continuous data points and you want to visualise the distribution. Which one of the following plots would be the best choice?

a. histogram
b. bar chart
c. scatter plot
d. dot plot

Question 7

The middle 50% of an ordered dataset is know as

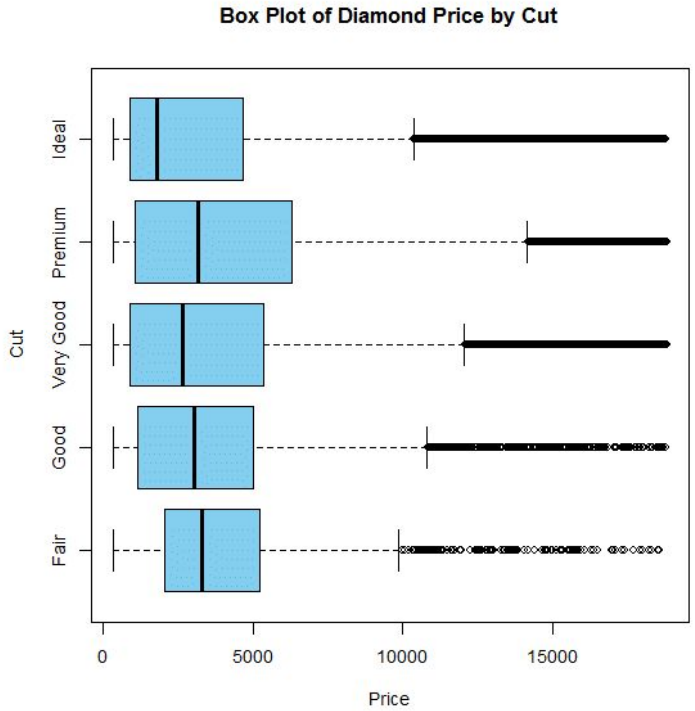
a. the interquartile range.
b. the range.
c. the median.
d. the first quartile.

Dr James Baglin (james.baglin@rmit.edu.au)


Question 8

The following box plot shows the distribution of price for different diamond cuts. Which of the following cuts shows the least skewness?

a. Premium
b. Very good
c. Good
d. Fair



Question 9

Which of the following two events are likely to be dependent?

a. A person wearing their lucky underwear and their sports team winning.
b. A person having a late night out (and a few too many) and being late to work the next day.
c. Two random people sharing the same birthday.
d. Flipping a coin and having it land heads and then flipping a coin again and having it land tails.

Question 10

Suppose you have three subjects that you need to study for exams. You want to study in a different order each day. For example, on Monday, you want to study statistics, maths, and then programming. On Tuesday, maths, programming and statistics... How many days can you study a different order of subject before you repeat a previous order.

a. 3
b. 4
c. 6
d. 9

Question 11

Suppose you work for a courier company. You're investigating the average number of successful, first time deliveries on a day-to-day basis for your couriers. Which one of the following probability distributions would be most suitable for modelling the average number of daily delivers per courier?

a. Normal distribution
b. Poisson distribution
c. Binomial distribution
d. *t* distribution

Question 12

Suppose the average human reaction time to a simple reaction time task is 250 ms with a standard deviation of 30 ms. Also assume reaction times are normally distributed. Which one of the following outcomes is most UNLIKELY?

a. The probability of a person's reaction time being less than 280 ms.
b. The probability of a person's reaction time being greater than 220 ms.
c. The probability of a person's reaction time being between 220 and 280 ms.
d. The probability of a person's reaction time being less than 220 ms.

Question 13

Investigators aim to conduct a survey of RMIT university students' study habits. Using university enrolment data, they divide the university into colleges, schools and student types (undergraduate, postgraduate). The investigators then randomly sample from these subgroups to ensure that each group is represented in the sample. This sampling method is best described as an example of

a. simple random sampling.
b. stratified sampling
c. cluster sampling
d. multi-stage sampling

Dr James Baglin (james.baglin@rmit.edu.au)


Question 14

Which one of the following is NOT an advantage of taking a sample over a census?

a. Samples are quicker.
b. If we are doing risky investigations, samples can minimise exposure to harm.
c. Samples are cheaper.
d. If the sample is large, it will represent the population.

Dr James Baglin (james.baglin@rmit.edu.au)


Question 15

Suppose you take 1000 random samples of size $n = 20$ from the population and using each sample, you calculate a mean and 90% *CI*. How many of the 90% *CIs* will you expect to capture the population mean?

a. 20
b. 100
c. 900
d. 1000

Question 16

An investigator comparing the average battery life of two competing brands concludes that brand A was significantly better than brand B according to the results of a two-sample *t*-test. What does the investigator mean by “significantly” better?

a. The investigator statistically proved that brand A had superior average battery life.
b. The investigator found that the difference between brand A and brand B didn’t occur by chance.
c. The investigator rejected the Null hypothesis that stated the mean battery life for the two brands was equal.
d. The investigator proved that brand B had superior average battery life.

Question 17

A *p*-value for a one-sample *t*-test can be defined as

a. the probability of observing a sample mean by chance.
b. the probability of observing a sample mean, or one more extreme, assuming the Null hypothesis is true.
c. the probability of the Null hypothesis being true.
d. the probability of the Alternate hypothesis being true.

Question 18

Consider the following formula for a 95% *CI* of the mean.

$$\bar{x} \pm z_{1-(0.05/2)} \frac{\sigma}{\sqrt{n}}$$

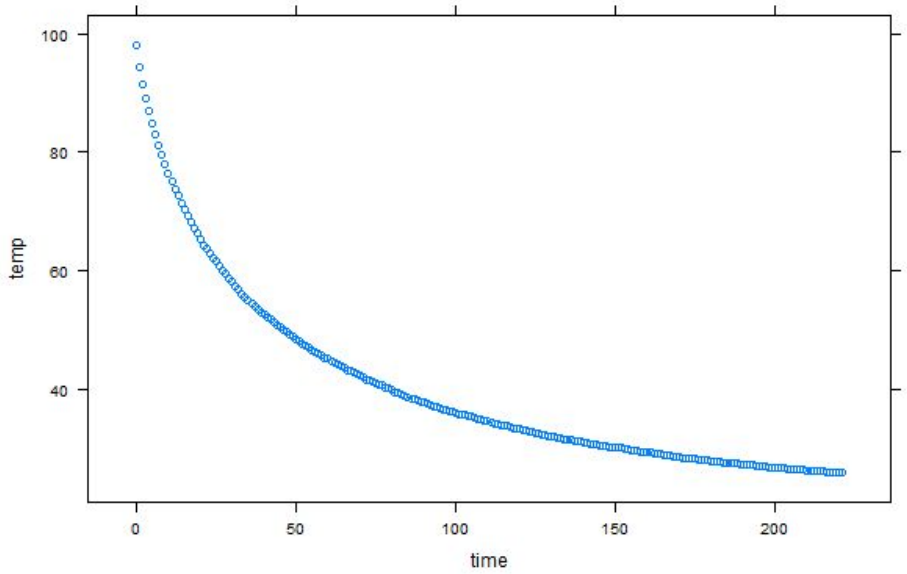
This formula makes a number of assumptions. Which one of the following is NOT an assumption?

a. The population standard deviation is known.
b. This interval will be a two-sided *CI*.
c. The sample size $n > 30$
d. x are normally distributed in the population

Question 19

The following plot shows the relationship between time (10 second intervals) and cooling water temperature. Correctly describe the nature of the relationship.

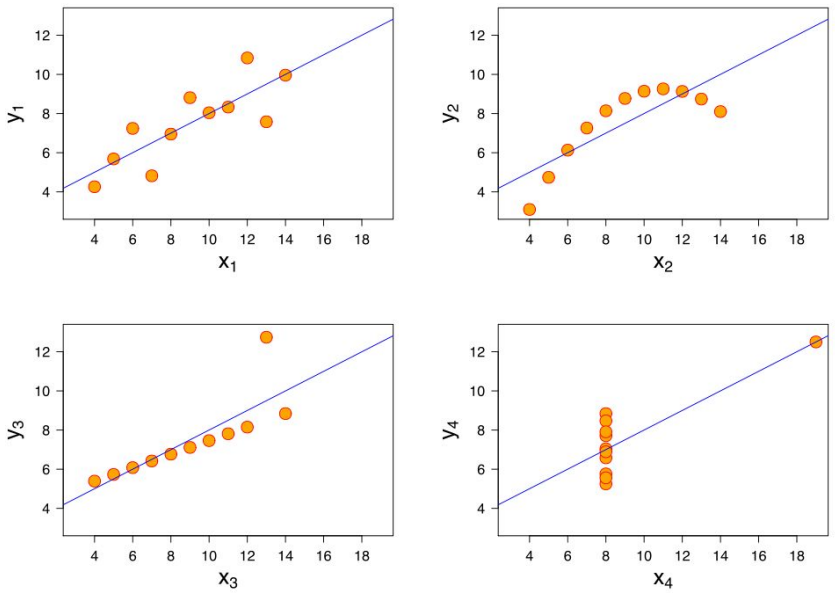
a. Negative, linear
b. Positive, linear
c. Negative, nonlinear
d. Positive, nonlinear



Question 20

The scatter plots to the right show four bivariate relationships. All plots share the same correlation of $r = .816$. These plots demonstrate

a. the limitations of using correlation when the data are not normally distributed.
b. the importance of running linear regression in conjunction with correlation.
c. the effect of small sample size on correlation.
d. the importance of visualising your data.



Question 21

For each of the following variables, determine its level of measurement:

a. Student ID number
b. Time (seconds)
c. Mass/weight
d. Calendar year (e.g. 2015)
e. Position of a team on a sporting ladder

(5 points)

Question 22

Use the following ABS data table (reported in ,000s) to answer the following questions. Demonstrate how you would calculate the following probabilities (no need to do the calculations!):

a. The probability accessing the internet every day in 2006-07.
b. The probability of accessing the at least weekly in 2008 - 2019
c. The probability of accessing the internet everyday daily for all years.

(3 points)

Households with internet access

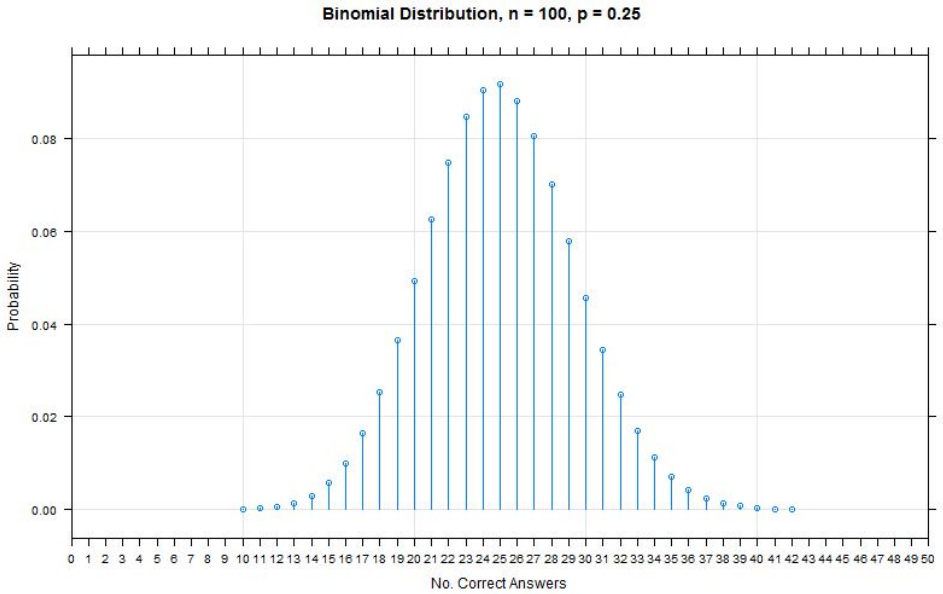
Frequency	2006–07	2007–08	2008–09	2010–11	2012–13	Total
Every day	3321	3703	4281	5157	5933	22395
At least weekly	1561	1540	1382	1339	1202	7024
At least monthly	185	176	132	148	129	770
At least yearly	21	26	25	20	23	115
Never	31	25	33	39	31	159
Don't know	19	20	26	21	26	112
Total	5138	5492	5878	6724	7343	30575

Question 23

Remember back to Week 4. A student was about to attempt a 100 question multiple choice exam without any preparation. Each multiple choice question has three distractors and one correct answer. Each question is independent so, the probability of guessing correctly is always 25%. Assuming the student guesses on every question, the exam score can be modelled using a binomial distribution shown to the right. Using this scenario and the plot provided, answer the following questions:

a. What is the expected score for the student?
b. What is the approximate probability the student will pass?
c. What is the approximate probability the student will score under 25%?
d. Is it more likely the student will score less than 17 or more than 35?

(4 points)

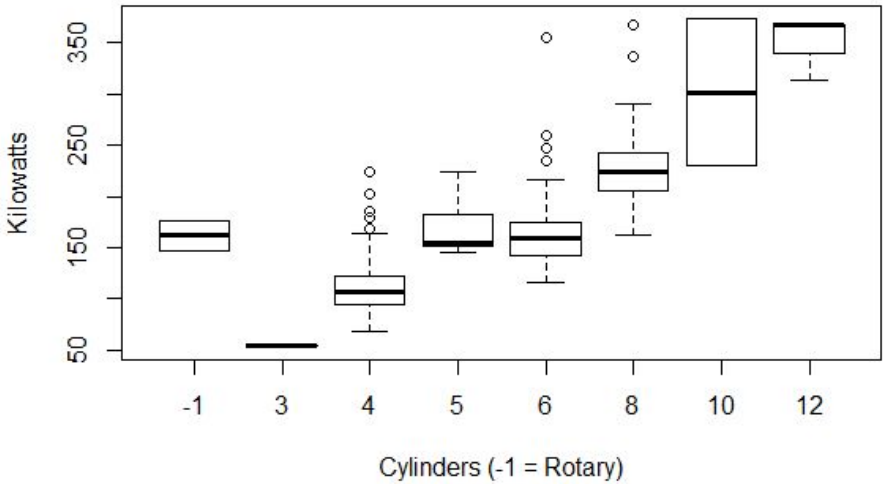


Question 24

The box plots to the right shows the distribution of car power (kilowatts) for cars with different numbers of cylinders (-1 = rotary). Using this plot, answer the following questions:

a. Cars with what number of cylinders have the highest median power?
b. Cars with what number of cylinders have the highest IQR?
c. Cars with what number of cylinders have the most number of suggested outliers?
d. Explain the unusual appearance of the box plot for 3 cylinder cars?
e. Why does the box plot for rotary engine cars have no whiskers?

(5 points)



Question 25

Is there are statistically significant difference between fathers' and sons' heights (inches)? Use the output to answer the following questions:

a. What was the estimated mean difference between fathers' and sons' heights?
b. Use the sample size to show how *df* was calculated?
c. What was the Null hypothesis for the test reported in the output?
d. Use the output to test the Null hypothesis and draw a conclusion.

(4 points)

```
#Fathers' Heights
 min Q1 median   Q3  max     mean       sd   n missing
  62 68     69 70.5 78.5 69.16817 2.299929 465       0
#Sons' Heights
 min   Q1 median Q3 max     mean       sd   n missing
  60 67.5   69.2 71  79 69.22882 2.631594 465       0
      Paired t-test

data:  Heritability_sons$father and Heritability_sons$height
t = -0.47822, df = 464, p-value = 0.6327
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -0.3098469  0.1885565
sample estimates:
mean of the differences
            -0.06064516
```


Question 26

The following statistical output reports the compares the means of 465 male and 433 female heights (inches). Using this output, answer the following questions:

a. Explain whether or not the analysis assumes homogeneity of variance.
b. State the Null and Alternate hypothesis for the hypothesis test reported.
c. Use the 95% *CI* to test the Null hypothesis.
d. Draw a conclusion in context.

(4 points)

```
  sex min   Q1 median   Q3  max mean  sd   n
1   F  56 62.5   64.0 65.5 70.5 64.11 2.37 433
2   M  60 67.5   69.2 71.0 79.0 69.22 2.63 465

     Welch Two Sample t-test

data:  height by sex
t = -30.662, df = 895.02, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -5.446293 -4.791018
sample estimates:
mean in group F mean in group M
       64.11016        69.22882
```

Question 27

You're on holidays and find yourself missing statistics and hypothesis testing. You decide to collect some data by tossing a coin 1000 times to determine if it was fair. You toss 510 heads and 490 tails. The results of the hypothesis test are reported to the right. Use this output to answer the following questions:

a. What type of statistical test has been performed?
b. What was the expected count for heads?
c. Show how df was calculated.
d. State the Null and Alternate hypothesis for the hypothesis test reported.
e. Interpret the hypothesis test and draw a conclusion in context.

(5 points)

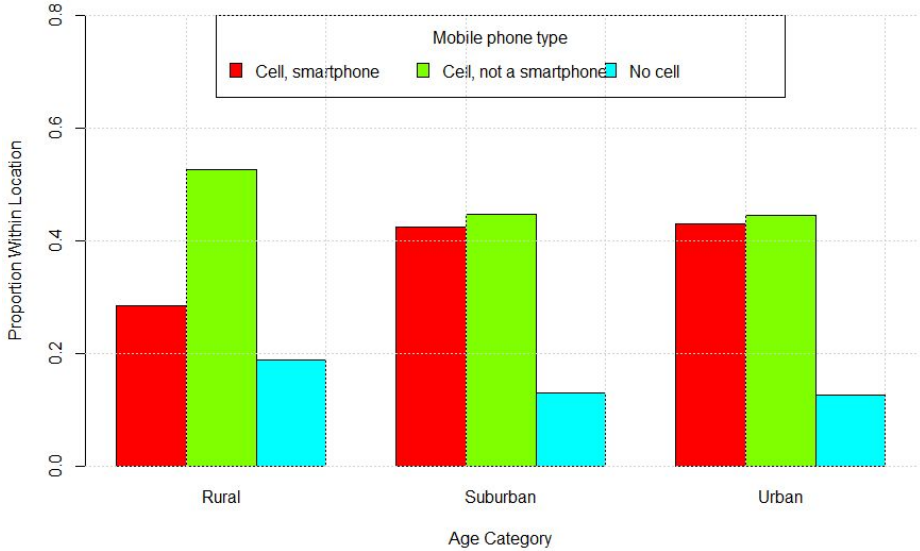
```
        Chi-squared test for given probabilities

data:  x
X-squared = 0.4, df = 1, p-value = 0.5271

  510      490
(500.00) (500.00)
 [0.20]   [0.20]
< 0.45>  <-0.45>

key:
     observed
     (expected)
     [contribution to X-squared]
     <residual>
```


Question 28

Investigators collect survey data looking at the association between a person's area of residence and the type of mobile phone they own. The data, clustered bar chart and results from an appropriate hypothesis test are reported to the right. Use this output to answer the following questions:

a. What type of statistical test has been performed? Why?
b. State the Null and Alternate hypothesis for the hypothesis test reported.
c. Interpret the hypothesis test and draw a conclusion in context.
d. Briefly explain the nature of the relationship between the variables.

(4 points)



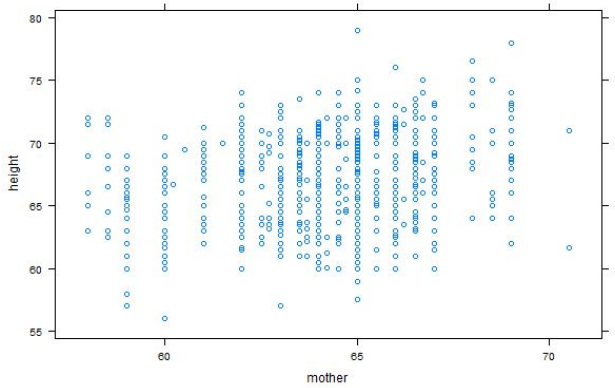

```
data:  x
X-squared = 44.925, df = 6, p-value = 4.844e-08
```

Question 29

The following output shows the relationship between a mother's height (inches) and their adult children. Using this output, answer the following questions:

a. State the assumptions that would need to be checked for the analysis reported in the output.
b. Interpret the slope of the model.
c. State the Null and Alternate hypothesis for the slope of the model.
d. Test the slope of the model and draw a conclusion in context.
e. Comment on the fit of the model.

(5 points)




```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.69077    3.25874  14.328  < 2e-16 ***
mother       0.31318    0.05082   6.163 1.08e-09 ***

Residual standard error: 3.511 on 896 degrees of freedom
Multiple R-squared:  0.04066,    Adjusted R-squared:
0.03959
F-statistic: 37.98 on 1 and 896 DF,  p-value: 1.079e-09
```

Question 30

The following two regression models show the relationship a father's height (inches) and their daughters or sons' adult heights. Using the output, determine which relationship is stronger and justify. (2 points)

Sons

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.5852     2.6632   13.36   <2e-16
father        0.4116     0.0384   10.72   <2e-16

Residual standard error: 2.109 on 431 degrees of freedom
Multiple R-squared:  0.2105,     Adjusted R-squared:  0.2086
F-statistic: 114.9 on 1 and 431 DF,  p-value: < 2.2e-16
```

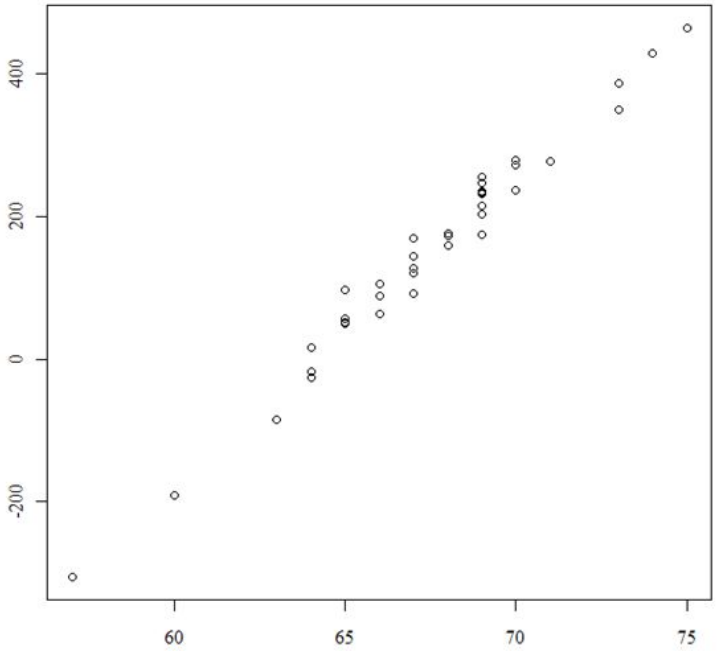
Daughters

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.25891    3.38663   11.30   <2e-16
father       0.44775    0.04894    9.15   <2e-16

Residual standard error: 2.424 on 463 degrees of freedom
Multiple R-squared:  0.1531,     Adjusted R-squared:  0.1513
F-statistic: 83.72 on 1 and 463 DF,  p-value: < 2.2e-16
```


Bonus points: Estimate that correlation!

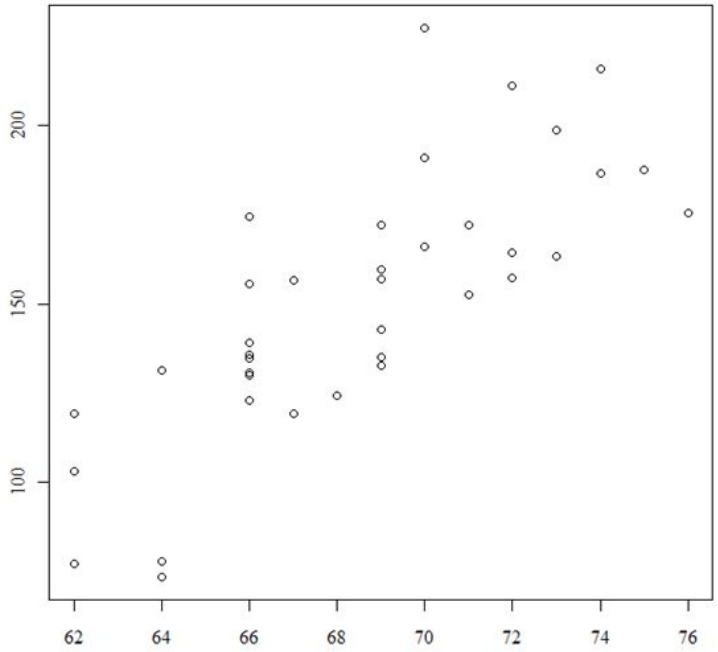
- If we have time and the scores are close….
- You will be shown a series of scatter plots of bivariate data.
- Estimate the Pearson correlation coefficient.
- If you get within ± 0.05 of the true value, you get a point.
- Ready?

Correlation 1

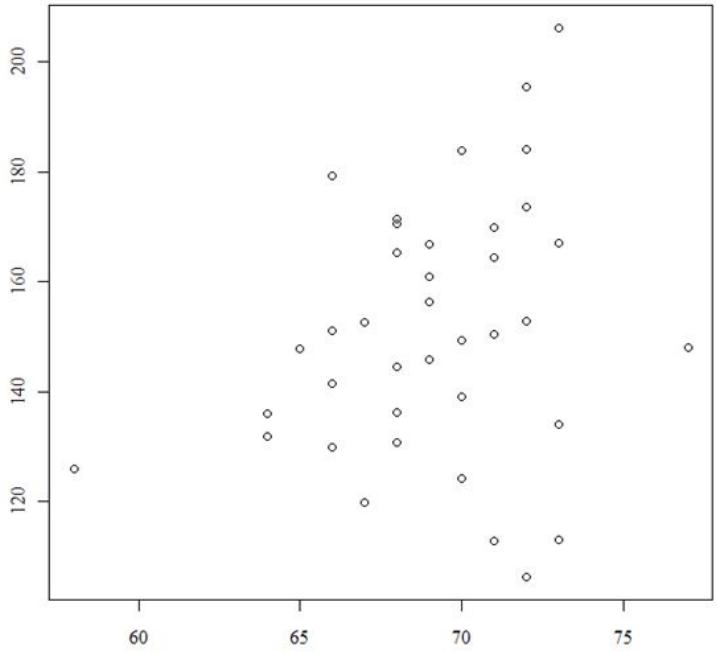


Dr James Baglin (james.baglin@rmit.edu.au)

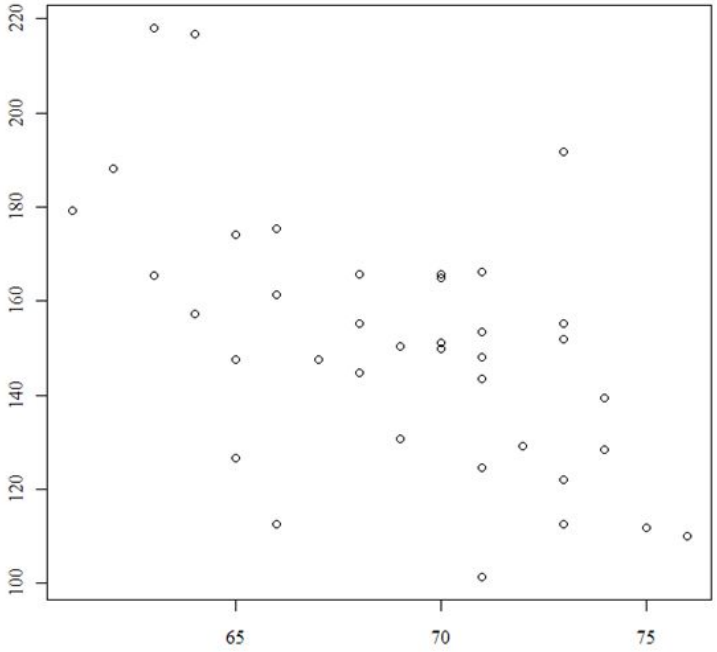
Correlation 2



Correlation 3



Correlation 4



Correlation 5