Part III — Statistics

Based on lectures by Brian
Notes taken by Dexter Chua

Lent 2017-2018

Contents

1 Representation and summary of data - location 4
- 1.1 Basic Concepts of Variable 4
- 1.2 Grouped data 4
- 1.3 Mean , mode and median 4
- 1.4 Linear interpolation 4
- 1.5 Coding 4

2 Representation and summary of data - measures of dispersion 5
- 2.1 Range and interquartile range 5
- 2.2 Percentiles split the data into 100 parts 5
- 2.3 Range and Interquartile range 5
- 2.4 Variance and standard deviation 5
- 2.5 Variance and standard deviation for grouped data 5
- 2.6 Coding 5

3 Representation of data 6
- 3.1 Stem and Leaf diagrams 6
- 3.2 Outlier 6
- 3.3 Box plot 6
- 3.4 Histogram 6
- 3.5 Skewness (Shape) 6
- 3.6 What!? 7

4 Probability 8
- 4.1 Classical Probability 8
- 4.2 Venn diagram and their rules 8
- 4.3 Conditional Probabilites 8
  - 4.3.1 Vann diagram 8
  - 4.3.2 Tree diagram 8
- 4.4 Special Events of Probabilites 8

5 Correlation 9
- 5.1 Correlation 9
- 5.2 Bivariate data 9
- 5.3 Product moment Correlation coefficient r 9
- 5.4 Coding 9

6 Regression 10
- 6.1 Linear 10
- 6.2 Coding 10
- 6.3 Interpolation and Extrapolation 10

7 Discrete random variables 11
- 7.1 Probability distribution 11

8 The normal distribution 12

9 Binomial distribution **13**
9.1 Basic Concept . . . 13
9.2 Mean and Variance . . . 13

10 Poisson distribution **14**
10.1 Basic Concepts . . . 14
10.2 Mean and Variance . . . 14
10.3 Approximate a Binomial with Poisson . . . 14

11 Continuous random variables **15**
11.1 Continous random variable . . . 15
11.2 Cumulative distribution function . . . 16
11.3 Mean and Variance . . . 17
11.4 Mode, median and quartiles . . . 17

12 Continuous uniform distribution **18**
12.1 Continuous uniform distribution . . . 18
12.2 Mean and Variance . . . 18
12.3 Choosing the right model . . . 18

13 Normal approximation **19**
13.1 Approximating binomial by normal . . . 19
13.2 Approximating Poisson by normal . . . 19
13.3 Choosing the appropriate approximation . . . 19

14 Population and samples **20**
14.1 The Concept of population and samples . . . 20

15 Hypothesis testing **21**
15.1 Concept of hypothesis testing . . . 21
15.2 One- and two-tailed tests . . . 22

16 Combination of random variables **23**

17 Sampling **24**

18 Estimation , confidence intervals and tests **25**
18.1 Estimation . . . 25
18.2 Confidence intervals and significance tests . . . 25
18.3 Combination of sampling distribution . . . 28

19 Goodness of fit and contingency tables **29**
19.1 Basic Concept . . . 29
19.2 Examples for all distributions . . . 29

20 Regression and correlation **35**

21 Quality of tests and estimators **39**

22 One-sample procedures **40**

23 Two-sample procedures **41**

1 Representation and summary of data - location

1.1 Basic Concepts of Variable

Definition (Quantitative variables and Qualitative variables)**.** Quantitative variable associated with numerical observation. Qualitative variables associated with non-numerical observations.

Definition (Continuous variable and discrete variable)**.** Continuous variable can take ant value in given range. Discrete can take only specific values in a given range.

1.2 Grouped data

Definition (Grouped data)**.** The groups are more commonly known as classes.

- class boundaries.
- mid-point of a class.
- class width.

Example. Example 5-6

Definition (Frequency and cumulative frequency)**.** Number of anything; example is how many sheeps. It is sometimes helpful to add a column to the table showing the running total of the frequencies. This is called the cumulative frequency

Definition (Ungrouped data)**.** Show all data

1.3 Mean , mode and median

Definition (Mode)**.** The mode is the value that occurs most often

Definition (Median)**.** n/2 term or 1 term above

Definition (Mean)**.**

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

1.4 Linear interpolation

Example. Example 14-15

1.5 Coding

Example. pick 1 example

2 Representation and summary of data - measures of dispersion

2.1 Range and interquartile range

The list of formula:

- Range = Upper value − Lowest value

Example. example 3

2.2 Percentiles split the data into 100 parts

Example. example 4

2.3 Range and Interquartile range

Example (Linear Interpolation)**.**

2.4 Variance and standard deviation

Definition (Variance)**.** Let f stand for the frequency, then $n = \sum f$ and

$$\text{Variance} = \frac{\sum f(x-\bar{x})^2}{\sum f} \text{ or } \frac{\sum fx^2}{\sum f} - (\frac{\sum fx^2}{\sum f})$$

2.5 Variance and standard deviation for grouped data

Definition.

Example. example 7-8

2.6 Coding

Example. example 9-11

3 Representation of data

3.1 Stem and Leaf diagrams

3.2 Outlier

Definition. An outlier is an extreme value that lies outside the overall pattern of the data.

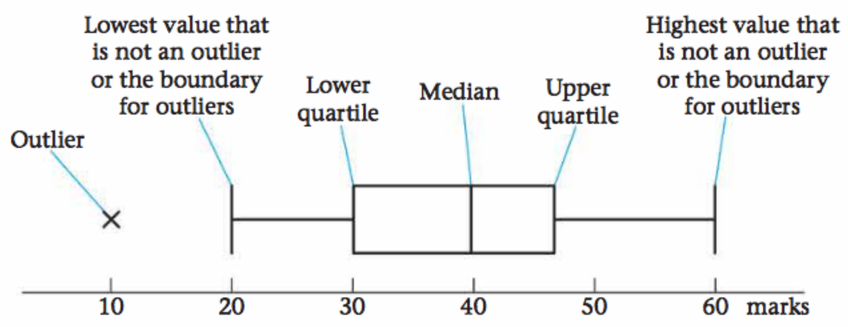
An outlier is any value, which is

greater than the upper quartile + 1.5 × interquartile range

OR

less than the lower quartile + 1.5 × interquartile range

3.3 Box plot



3.4 Histogram

Definition (Frequency density)**.**

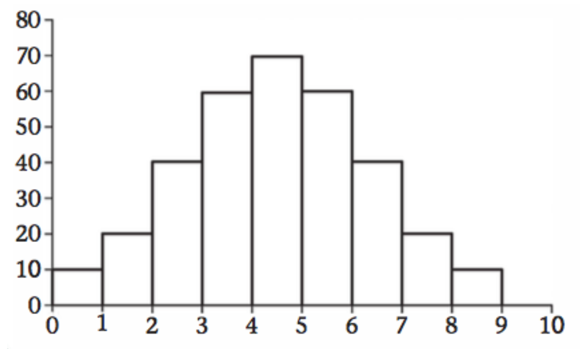
$$\text{frequency density} = \frac{\text{frequency}}{\text{class width}}$$

Example. 7

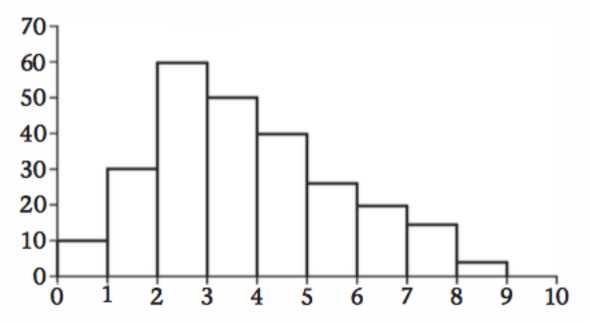
3.5 Skewness (Shape)

A distribution can be symmetrical , have positive skew or have negative skew

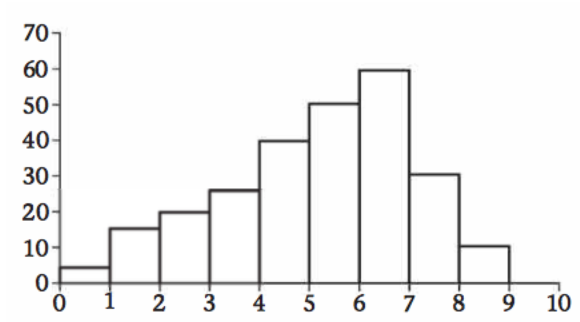
symmetrical $Q_2 - Q_1 = Q_3 - Q_2$ or mode=median=mean



positive :$Q_2 - Q_1 < Q_3 - Q_2$ or mode<median<mean



negative :$Q_2 - Q_1 > Q_3 - Q_2$ or mode>median>mean



Or you can calculate:

$$\frac{3(\text{mean} - \text{median})}{\text{SD}}$$

3.6 What!?

Example. example 10-12

4 Probability

4.1 Classical Probability

4.2 Venn diagram and their rules

Definition (Complementary Probability)**.**

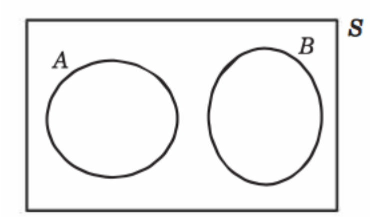
4.3 Conditional Probabilites

4.3.1 Vann diagram

4.3.2 Tree diagram

4.4 Special Events of Probabilites

Definition (Mutually exclusive)**.** When events have no outcomes in common, they are mutually exclusive.



There is no intersection of A and B, so $P(A \cap B) = 0$

We can use $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
result is

$$P(A \cup B) = P(A) + P(B)$$

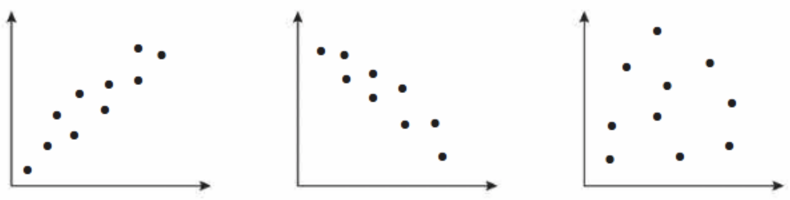
Definition (Independent events)**.** When one event has no effect on another, they are independent so $P(A|B) = P(A)$

by $\frac{P(A \cap B)}{P(B)} = P(A)$ we have:

$$P(A \cap B) = P(B) \times P(A)$$

5 Correlation

5.1 Correlation



5.2 Bivariate data

Recall this formula :

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

In correlation we write:

$$S_{xx} = \sum(x-\bar{x})^2$$

$$S_{yy} = \sum(y-\bar{y})^2$$

so

$$\text{Variance} = \frac{S_{xx}}{n}$$

Definition (Co-Variance)**.**

$$S_{xy} = \frac{\sum(x-\bar{x})(x-\bar{y})}{n}$$

5.3 Product moment Correlation coefficient r

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

The value of r varies between -1 and 1

If $r = 1$, positive linear correlation

If $r = -1$, nagative linear correlation

If $r = 0$, no linear correlation

limitation:

5.4 Coding

does not effect r

6 Regression

6.1 Linear

let $y = a + bx$ be a regression line
where

$$b = \frac{S_{xy}}{S_{xx}} \text{ and } a = \bar{y} - b\bar{x}$$

6.2 Coding

6.3 Interpolation and Extrapolation

7 Discrete random variables

7.1 Probability distribution

Definition (Mean / Expected value)**.**

$$E(X) = \sum xp(x)$$

when we find $E(X^n)$:

$$E(X^n) = \sum x^n p(x)$$

Definition (Variable)**.**

$$Var(X) = E(X^2) - (E(X))^2$$

The constant a and b affect on $E(X)$ and $Var(X)$

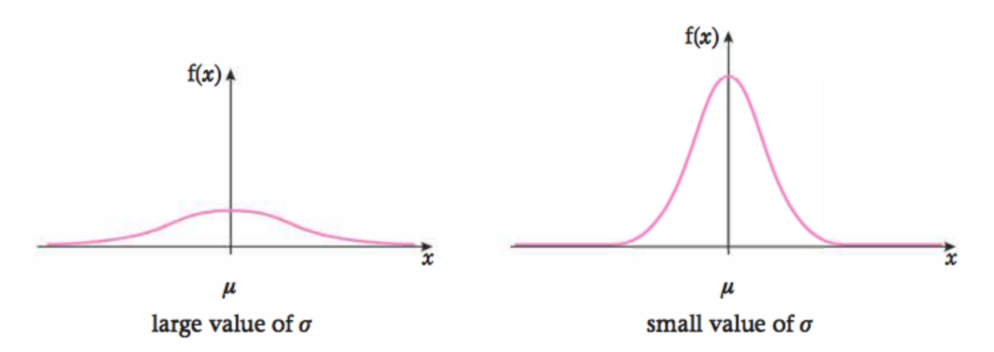
$$E(aX + b) = aE(x) + b$$

$$Var(aX + b) = a^2 Var(X)$$

Definition (Uniform distribution)**.** The distribution is uniform when all the probabilities is the same of all values.

8 The normal distribution

$Z \sim N(\mu, \sigma^2)$ represent the normal distribution.



The random variable X can be written as $X \sim N(\mu, \sigma^2)$

you can transformed X to Z by this formula

$$z = \frac{X - \mu}{\sigma}$$

Example. Example 8-9

9 Binomial distribution

9.1 Basic Concept

$X \sim B(n, p)$ represent the Binomial distribution, then

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

9.2 Mean and Variance

If $X \sim B(n, p)$ then

$$E(X) = \mu = np$$
$$Var(X) = \sigma^2 = np(1 - p)$$

Example. example 9-14

10 Poisson distribution

10.1 Basic Concepts

Recall the exponential function

$$e^x = 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + + \frac{x^r}{r!} + \cdots$$

If you let $x = \lambda$ and remember that $\lambda^0 = 1$ this gives

$$e^\lambda = \lambda^0 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots + + \frac{\lambda^r}{r!} + \cdots$$

Dividing by e^λ gives

$$\frac{e^\lambda}{e^\lambda} = \lambda^0 + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!} + \cdots + \frac{\lambda^r e^{-\lambda}}{r!} + \cdots$$

And the probability function is

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

We say that X has a Poisson distribution with parameter λ abd write

$$X \sim Po(\lambda)$$

10.2 Mean and Variance

$$Var(X) = E(X) = \mu = \sigma^2 = \lambda$$

Lemma. If mean and standard deviation square is same, we usually use Poisson distribution.

Example. example 5-6

10.3 Approximate a Binomial with Poisson

If $X \sim B(n, p)$ and

- n is large
- p is small

then X can be approximated by

$$Po(np)$$

Example. example 7-8 9-10

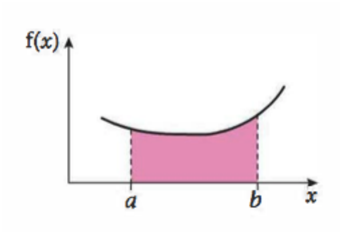
11 Continuous random variables

11.1 Continous random variable

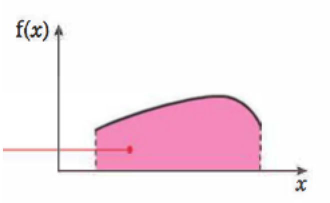
The Continuous random variables with p.d.f $f(x)$ satisfied the following properties:

(i) $f(x) \geq 0$ since we cannot have negative probabilities

(ii) $P(a < X < b) = \text{shaded area} = \int_a^b f(x)\,\mathrm{d}x$



(iii) $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$ since the area under the curve = 1.



Example. The random variable X has probability density function

$$f(x) = \begin{cases} kx(4-x) & 2 \leq x \leq 4 \\ 0 & \text{otherwise.} \end{cases}$$

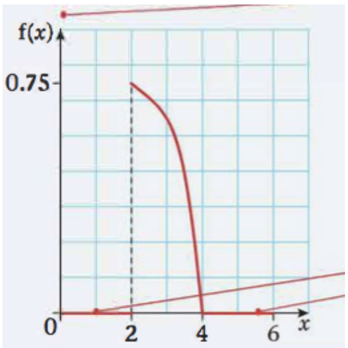
Find the value of k and sketch the p.d.f

Proof.

$$\int_2^4 k(4x - x^2)\,\mathrm{d}x = 1$$

$$k[2x^2 - \frac{x^3}{3}]_2^4 = 1$$

$$k = (\frac{3}{16})$$



□

Example. The random variable X has probability density function

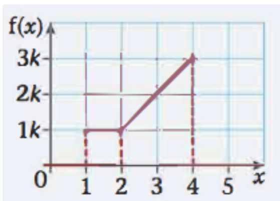
$$f(x) = \begin{cases} k & 1 < x < 2 \\ k(x-1) & 2 \le x \le 4 \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of k and sketch the p.d.f

Proof.

$$\int_1^2 k \, \mathrm{d}x + \int_2^4 k(x-1) \, \mathrm{d}x = 1$$

$$k = \frac{1}{5}$$

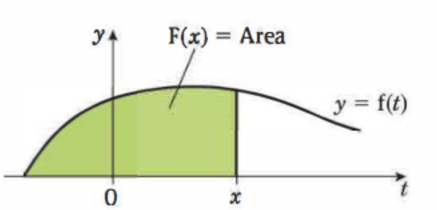


□

11.2 Cumulative distribution function

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t) \, \mathrm{d}t$$

where $F(x) = P(X \le x) = 1$



If X us a Continous random variable with c.d.f. $F(x)$ and p.d.f $f(x)$

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}t} F(x) \text{ and } F(x) = \int_{-\infty}^{x} f(t) \, \mathrm{d}t$$

Example. example 5-6

11.3 Mean and Variance

If X is a Continuous random variable with p.d.f $f(x)$

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x$$

$$\sigma^2 = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x) \, \mathrm{d}x - \mu^2$$

Remark. The range is the range of that function instead of negative infinity to infinity.

11.4 Mode, median and quartiles

The median m or Q_2 satisfies $F(m) = F(Q_2) = 0.5$
The lower quartile Q_1 satisfies $F(Q_1) = 0.25$
The lower quartile Q_1 satisfies $F(Q_3) = 0.75$

The mode is the x value at the highest point of the p.d.f $f(x)$

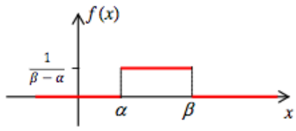
12 Continuous uniform distribution

12.1 Continuous uniform distribution

Definition

A continuous uniform distribution has **constant** probability density over a fixed interval.

Thus $f(x) = \dfrac{1}{\beta - \alpha}$ is the continuous uniform p.d.f. over the interval $[\alpha, \beta]$ and has a rectangular shape.



Median

By symmetry the median is $\dfrac{\alpha + \beta}{2}$

Mean and Variance

The expected mean is $\mathrm{E}[X] = \mu = \dfrac{\alpha + \beta}{2}$, which is the same as the median.

and the expected variance is $\mathrm{Var}[X] = \sigma^2 = \dfrac{(\beta - \alpha)^2}{12}$.

These formulae are proved in the appendix

12.2 Mean and Variance

Example. example 4-7

12.3 Choosing the right model

Example. example 8-10

13 Normal approximation

13.1 Approximating binomial by normal

If $X \sim B(n, p)$ and

- n is large
- p is close to 0.5

Then X can be approximated by

$$Y \sim N(np, np(1-p))$$

Example. $X \sim B(120, 0.25)$ approximated to $Y \sim N(30, (\sqrt{22.5})^2)$

Example. example 4

13.2 Approximating Poisson by normal

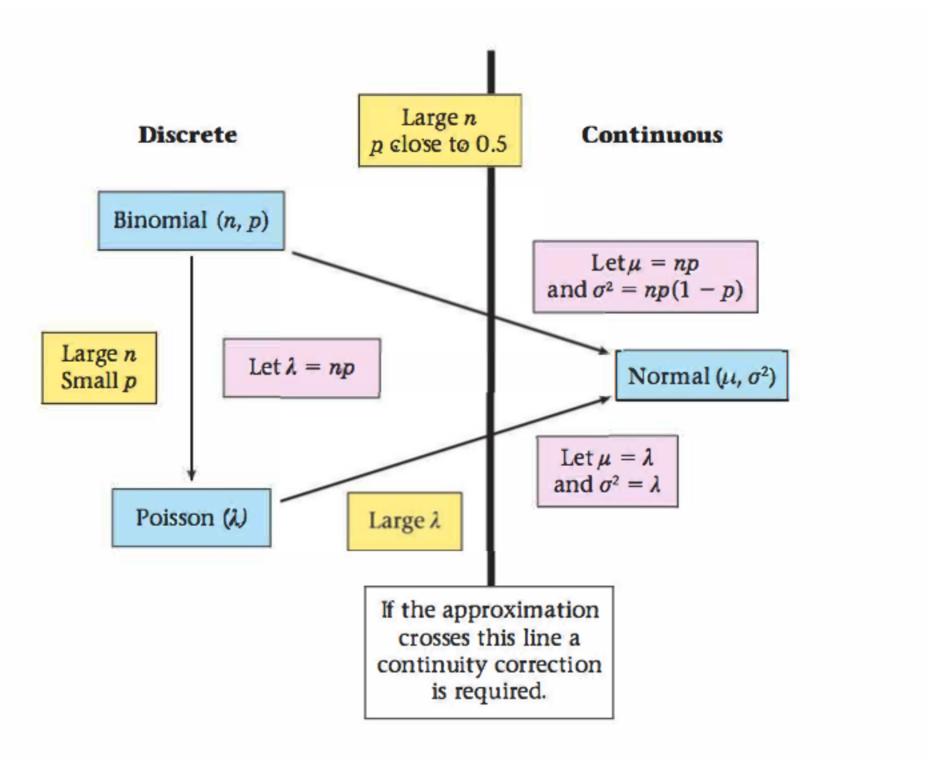
If λ is large

$$X \sim Po(\lambda) \text{to} Y \sim N(\lambda, (\sqrt{\lambda})^2)$$

Example. $X \sim Po(25)$ transformed to $Y \sim N(25, 5^2)$

Example. example 6

13.3 Choosing the appropriate approximation



Example. example 7

14 Population and samples

14.1 The Concept of population and samples

List of the possible samples and find their probabilities and distribution.

Example. example 5 6

15 Hypothesis testing

15.1 Concept of hypothesis testing

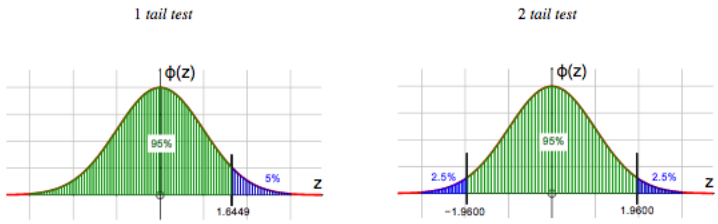
Definition (Null hypothesis H_0)**.** The hypothesis which is assumed to be correct unless shown otherwise.

Definition (Alternarive hypothesis H_1)**.** This is the conclusion that should be made if H_0 is rejected

Definition (Critical region)**.** The range of values which would lead you to reject the null hypothesis, H_0

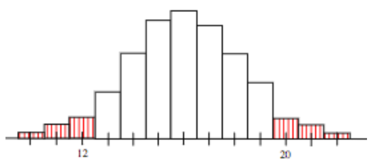
Definition (Significance level)**.** The actual significance level is the probability of rejecting H_0 when it is in fact true.

From your observed result (test statistic) you decide whether to reject or not to reject the null hypothesis H_0



The test statistic is significantat 5%, or that we reject H_0. Thus H_0could actually be true but we still reject it. Thus, the significance level, 5%, is the probability that we reject H_0 when it is in fact true, or the probability of incorrectly rejecting H_0.
When we reject the null hypothesis, H_0, we use the alternative hypothesis to write the conclusion.

The Poisson and Binomial distributions are discrete, and we look at probability histograms.



In the diagram, the critical region (shown by the shaded areas) is $X \leq 12$ or $X \geq 20$.

We include the whole bar around $X = 12$ and around $X = 20$

So $P(X \leq 12)$ is the area to the left of 12.5,
and $P(X \geq 20)$ is the area to the right of 19.5,

If $P(X \leq 12) = 0.0234$ and $P(X \geq 20) = 0.0217$, then
the actual signifaicance level is $0.0234 + 0.0217 = 0.0451 = 4.51\%$
Thus the probability of incorrectly rejecting H_0 is 0.0451.

15.2 One- and two-tailed tests

The One-tail test is

$$\begin{aligned} H_0 :& a = b \\ H_1 :& a > b \text{ or } a < b \end{aligned}$$

The Two-tail test is

$$\begin{aligned} H_0 :& a = b \\ H_1 :& a \neq b \end{aligned}$$

Example. 3

Example. 4-13

16 Combination of random variables

If X_1 and X_2 are independent normal random variables

$$X_1 \sim \mathbb{N}(\mu_1, \sigma_1^2) \text{ and } X_2 \sim \mathbb{N}(\mu_2, \sigma_2^2)$$

then $X_1 + X_2$ and $X_1 - X_2$ are also normal random variables

$$X_1 + X_2 \sim \mathbb{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \text{ and } X_1 - X_2 \sim \mathbb{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

17 Sampling

18 Estimation , confidence intervals and tests

18.1 Estimation

Definition (Biased and unbiased estimator)**.** If X (usually found from a sample) is used to estimate the value of a population parameter, t, then X is an unbiased estimator of t if $E[X] =$ the true value of the parameter t.

If an estimator, X, is biased, then the bias is the difference between $E[X]$ and the true value of the parameter t.

Definition (Unbiased estimators of μ and σ^2)**.**

18.2 Confidence intervals and significance tests

Definition (Sampling distribution of the mean)**.**

$$\mathbb{N}(\mu, \frac{\sigma^2}{n})$$

Theorem (Central limit theorem)**.** If $\{X_1, X_2, \cdots, X_n\}$ is a random sample of size n drawn any population with mean μ and variance σ^2 then the population of sample means.

(i) has expected mean μ

(ii) has expected variance $\frac{\sigma^2}{n}$

(iii) forms a normal distribution if n is 'large enough' i.e. $\bar{X} \sim \mathbb{N}(\mu, \frac{\sigma^2}{n})$

The standard error of the sample mean is $\frac{\sigma}{\sqrt{n}}$

Example.

Example:

A biscuit manufacturer makes packets of biscuits with a nominal weight of 250 grams. It is known that over a long period the variance of the weights of the packets of biscuits produced is 25 grams^2. A sample of 10 packets is taken and found to have a mean weight of 253·4 grams. Find 95% confidence limits for the mean weight of all packets produced by the machine.

Solution:

First assume that the machine is still producing packets with the same variance, 25.

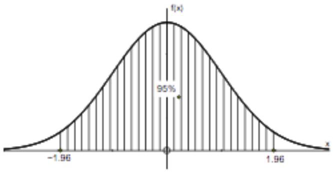
Suppose that the mean weight of all packets of biscuits is μ grams then the population of all packets has mean μ and standard deviation 5.

From the central limit theorem we can assume that the sample means form an approximately normal population with mean μ and standard error (standard deviation) $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{10}} = 1{\cdot}5811$

95% of the samples will have a mean in the region

$$-1{\cdot}96 < Z < 1{\cdot}96$$

We assume that the mean of this sample, 253·4, lies in this region



$\Rightarrow \quad -1{\cdot}9600 < \frac{253{\cdot}4-\mu}{1{\cdot}5811} < 1{\cdot}9600$

$\Rightarrow \quad -1{\cdot}9600 < \frac{253{\cdot}4-\mu}{1{\cdot}5811}$ and $\frac{253{\cdot}4-\mu}{1{\cdot}5811} < 1{\cdot}9600$

$\Rightarrow \quad \mu - 1{\cdot}9600 \times 1{\cdot}5811 < 253{\cdot}4$ and $253{\cdot}4 < \mu + 1{\cdot}9600 \times 1{\cdot}5811$

$\Leftrightarrow \quad \mu < 253{\cdot}4 + 1{\cdot}9600 \times 1{\cdot}5811$ and $253{\cdot}4 - 1{\cdot}9600 \times 1{\cdot}5811 < \mu$

$\Leftrightarrow \quad 253{\cdot}4 - 1{\cdot}9600 \times 1{\cdot}5811 < \mu < 253{\cdot}4 + 1{\cdot}9600 \times 1{\cdot}5811$

$\Leftrightarrow \quad 250{\cdot}3 < \mu < 256{\cdot}5$

This means that 95% of the samples will give an interval which contains the mean and we say that [250·3 g, 256·5 g] is a 95% *confidence interval* for μ.

This means that there is a 0·95 probability that **this interval contains the true mean**.

It *does not* mean that there is a probability of 0·95 that the true mean lies in this interval - the true mean is a fixed number, and either *does* or *does not* lie in the interval so the probability that the true mean lies in the interval is either 1 or 0.

In practice we go straight to the last line of the example:

95% confidence limits are $\mu \pm 1{\cdot}9600 \times \frac{\sigma}{\sqrt{n}}$ since $P(Z - 1{\cdot}9600 < z < 1{\cdot}9600) = 0{\cdot}95$

tables give $P(Z > 1{\cdot}9600) = 0{\cdot}025$

90% confidence limits are $\mu \pm 1{\cdot}6449 \times \frac{\sigma}{\sqrt{n}}$ since $P(Z - 1{\cdot}6449 < z < 1{\cdot}6449) = 0{\cdot}90$

tables give $P(Z > 1{\cdot}6449) = 0{\cdot}05$

Other confidence limits can be found using the Normal Distribution tables.

Example: A sample of 64 packets of cornflakes has a mean weight $\overline{X} = 510$ grams and a variance $S^2 = 36$ grams^2. Find 90% confidence limits for the mean weight of all packets.

(Note that the 'sample variance' is taken as the unbiased estimate of σ^2.)

Solution: We **assume** that the sample variance = the variance of the population of all packets

$\Rightarrow \quad S^2 = 36 = \sigma^2$.

Now find standard deviation (standard error) of the sampling distribution of the mean (population of sample means), standard error = $\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{64}} = 0 \cdot 75$

For 90% confidence limits $z = \pm 1{\cdot}6449$ (remember to use the 4 D.P. tables after the Normal Dist. tables), using the sample mean $\overline{X} = 510$ grams

$\Rightarrow$ 90% confidence limits are $510 \pm 1{\cdot}6449 \times 0.75 = 510 \pm 1{\cdot}234$

$\Rightarrow$ a 90% confidence interval is $[508{\cdot}8,\ 511{\cdot}2]$ to 4 S.F.

Note that we have assumed that the unbiased estimate, S^2 (=36), is the actual variance, σ^2, of the population.

This is a reasonable assumption as the number in the sample, 64, is large and the error introduced is therefore small.

Example.

Significance testing– variance of population known

Mean of normal distribution

Example:

A machine, when correctly set, is known to produce ball bearings with a mean weight of 84 grams with a standard deviation of 5 grams. The production manager decides to test whether the machine is working correctly and takes a sample of 120 ball bearings. The sample has mean weight 83.2 grams. Would you advise the production manager to alter the setting of his machine? Use a 5% significance level.

Solution:

1) H_0: $\mu = 84$ grams

2) H_1: $\mu \neq 84$ grams $\Rightarrow$ 2 tail test

(Note that the machine is not working correctly if the test result is too high *or* too low)

3) *5% Significance level*

4) *The Test*

We assume that the machine is still working with a standard deviation of $\sigma = 5\ g$.

From H_0, the mean weight of all ball bearings is assumed to be $\mu = 84\ g$.

These are the parameters for the population of **all** ball bearings.

We want to test a sample mean and therefore need the mean and standard deviation of the population of sample means (the sampling distribution of the sample mean, $\bar{X}$).

Expected mean of the sample means = $\mu = 84\ g$. and

expected standard deviation of the sample means = standard error = $\dfrac{\sigma}{\sqrt{n}} = \dfrac{5}{\sqrt{120}} = 0 \cdot 456435...$

.

We have an observed mean of 83·2

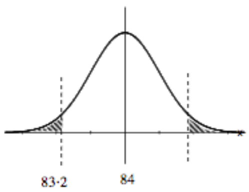
For a two-tailed test at 5%, we take 2·5% at each end

$P(\bar{X} < 83 \cdot 2) = \Phi\left(\frac{83 \cdot 2 - 84}{0 \cdot 456435...}\right) = \Phi(-1 \cdot 7527)$

$= (1 - \Phi(1 \cdot 75)) = 0 \cdot 0401$

$= 4 \cdot 01\% > 2 \cdot 5\%$

and so not significant at the 5% level.



5) *Conclusion*

Do not reject H_0 at the 5% level and advise the production manager that there is evidence that he should not change his setting, or that there is evidence that the machine is working correctly, etc.

Fortunately the formula for testing the difference between sample means

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

is in your formula booklet

18.3 Combination of sampling distribution

Same

19 Goodness of fit and contingency tables

19.1 Basic Concept

Definition (χ test)**.**

$$\chi^2 \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and expected frequencies

19.2 Examples for all distributions

Discrete uniform distribution

Example: A die is rolled 300 times and the frequency of each score recorded.

Score:	1	2	3	4	5	6
Frequency:	43	49	54	57	46	51

Test whether the die is fair at the 2·5% level of significance.

Solution: H_0: The die is fair, the probability of each score is $^1/_6$.

H_1: The die is not fair, the probability of each score is not $^1/_6$.

The expected frequencies are all $^1/_6 \times 300 = 50$ and we have

Score	Observed frequency	Expected frequency	$\frac{(O_i - E_i)^2}{E_i}$
1	43	50	0·98
2	49	50	0·02
3	54	50	0·32
4	57	50	0·98
5	46	50	0·72
6	51	50	0·02
Totals	**300**	**300**	**3·04**

$\Rightarrow$ $\chi^2 = 3{\cdot}04$

and ν = number of degrees of freedom $= n - 1 = 6 - 1 = 5$
since the total is a linear equation connecting the frequencies and is fixed.

From tables we see that $\chi_5^2(2{\cdot}5\%) = 12{\cdot}832 > 3{\cdot}04$, so our observed result is not significant.
We do not reject H_0 and conclude that the die is fair.

Continuous uniform distribution

This is very similar to the discrete uniform distribution – pay attention to the class boundaries and find the expected frequencies.

Binomial distribution

For H_0 *The Binomial distribution is a good fit*

we use the mean of the Observed frequencies to calculate the Expected frequencies, and so both O_i and E_i give the same mean and total: thus there are 2 linear equations connecting the frequencies and $\nu = n - 2$

but For H_0 *The Binomial distribution*, $B(30, 0{\cdot}3)$, *is a good fit*

the means using Oi and E_i will be different: thus there is only 1 linear equation, the total, connecting the frequencies and so $\nu = n - 1$.

Poisson distribution

For H_0 *The Poisson distribution is a good fit*

we use the mean of the Observed frequencies to calculate the Expected frequencies, and so both O_i and E_i give the same mean and total: thus there are 2 linear equations connecting the frequencies and $\nu = n - 2$

but For H_0 *The Poisson distribution*, $P_o(3)$, *is a good fit*

the means using Oi and E_i will be different: thus there is only 1 linear equation, the total, connecting the frequencies and so $\nu = n - 1$.

Example: A switchboard operator records the number of new calls in 69 consecutive one-minute periods in the table below.

number of calls	0	1	2	3	4	5	≥ 6
frequency	6	9	11	15	13	9	6

a) Say why you think that a Poisson distribution might be suitable.

b) Find the mean and variance of this distribution. Do these figures support the view that they might form a Poisson distribution?

c) Test the goodness of fit of a Poisson distribution at the 5% level.

Solution:

a) Telephone calls are likely to occur singly, randomly, independently and uniformly which are the conditions for a Poisson distribution.

b) Treating ≥ 6 as 7 we calculate the mean and variance

x	f	xf	x^2f
0	6	0	0
1	9	9	9
2	11	22	44
3	15	45	135
4	13	52	208
5	9	45	225
7	6	42	294
	69	215	915

$\Rightarrow$ mean = $^{215}/_{69} = 3{\cdot}12$
and variance = $^{915}/_{69} - (^{215}/_{69})^2 = 3{\cdot}55$.

From these figures we can see that the mean and variance are approximately equal: since the mean and variance of a Poisson distribution are equal this confirms the view that the distribution could be Poisson.

c) H_0: The Poisson distribution is a suitable model
H_1: The Poisson distribution is not a suitable model.

The Poisson probabilities can be calculated from $P(r) = \dfrac{\lambda^r e^{-\lambda}}{r!}$ where $\lambda = 3{\cdot}12$, and the expected frequencies by multiplying by $N = 69$.

Note that the probability for ≥ 6 is found by adding the other probabilities and subtracting from 1.

x	O	p	E	O (grouped)	E (grouped)	$\frac{(O-E)^2}{E}$
0	6	0·044337	3·059234			
1	9	0·138151	9·532395	15	12·59	0·461326
2	11	0·215235	14·8512	11	14·85	0·998148
3	15	0·223553	15·42515	15	15·43	0·011983
4	13	0·174145	12·01597	13	12·02	0·079900
5	9	0·108525	7·488214	9	7·49	0·304419
≥ 6	6	0·096056	6·627836	6	6·63	0·059864
	69		69		69.01	1.915641

The expected frequency for $x = 0$ is $3.06 < 5$ so it has been grouped with $x = 1$.

Thus we have $n = 6$ classes (after grouping) and $\nu = n - 2 = 4$

and $\chi_4^2(5\%) = 9.488$.

We have calculated $\chi^2 = 1.92 < 9.488$ which is not significant so we do not reject H_0 and conclude that the Poisson distribution is a suitable model.

The normal distribution

For H_0 *The Normal distribution is a good fit*

we use the mean and variance of the Observed frequencies to calculate the Expected frequencies, and so both O_i and E_i give the same mean, variance and total: thus there are 3 linear equations connecting the frequencies and $\nu = n - 3$

but For H_0 *The Normal distribution*, $N(14, 3^2)$, *is a good fit*

the means and variances using Oi and E_i will be different: thus there is only 1 linear equation, the total, connecting the frequencies and so $\nu = n - 1$.

Example: The sizes of men's shoes purchased from a shoe shop in one week are recorded below.

size of shoe	≤ 6	7	8	9	10	11	≥ 12
number of pairs	14	19	29	45	40	21	7

Is the manager's assumption that the normal distribution is a suitable model justified at the 5% level?

Solution: H_0: The normal distribution is a suitable model

H_1: The normal distribution is not a suitable model.

The total number of pairs, mean and standard deviation are calculated to be 175, 8.886 and 1.713 (taking ≤ 6 as 5 and ≥ 12 as 12)

Remembering that size 8 means from 7.5 to 8.5 we need to find the area between 7.5 and 8.5 and multiply by 175 to find the expected frequency for size 8, and similarly for other sizes.

x	$z = \frac{x-m}{s}$	$\Phi(z)$	class	area $= p$	$E = 175p$	O	$\frac{(O-E)^2}{E}$
6.5	-1.39	0.082	< 6.5	0.082	14.4	14	0.01
7.5	-0.81	0.209	6.5 to 7.5	$0.209 - 0.082 = 0.127$	22.2	19	0.46
8.5	-0.23	0.409	7.5 to 8.5	$0.409 - 0.209 = 0.200$	35.0	29	1.03
9.5	0.36	0.641	8.5 to 9.5	$0.641 - 0.409 = 0.232$	40.6	45	0.48
10.5	0.94	0.826	9.5 to 10.5	$0.826 - 0.641 = 0.185$	32.4	40	1.78
11.5	1.53	0.937	10.5 to 11.5	$0.937 - 0.826 = 0.111$	19.4	21	0.13
			> 11.5	$1 - 0.937 = 0.063$	11.0	7	1.45
							5.34

$n = 7$ classes & 3 linear equations connecting the frequencies $(N, m, s) \Rightarrow \nu = n - 3 = 4$.

$\chi_4^2(5\%) = 9.488$ and we have calculated $\chi^2 = 5.34 < 9.488$ and so we do not reject H_0 and therefore conclude that the normal distribution is a suitable model.

Contingency tables

For a 5×4 table in which the totals of each row and column are fixed the '?' cells represent the degrees of freedom since if we know the values of the ?s the frequencies in the other cells can now be calculated

	A	B	C	D	E	totals
W	?	?	?	?		✓
X	?	?	?	?		✓
Y	?	?	?	?		✓
Z						✓
totals	✓	✓	✓	✓	✓	✓

Thus there are $(5-1) \times (4-1) = 12$.

Generalising we can see that for an $m \times n$ table the number of degrees of freedom is $(m-1)(n-1)$.

Example: Natives of England, Africa and China were classified according to blood group giving the following table.

	O	A	B	AB
English	235	212	79	83
African	147	106	30	51
Chinese	162	135	52	43

Is there any evidence at the 5% level that there is a connection between blood group and nationality?

Solution: H_0: There is no connection between blood group and nationality.

H_1: There is a connection between blood group and nationality.

First redraw the table showing totals of each row and column

	O	A	B	AB	totals
English	235	212	79	83	609
African	147	106	30	51	334
Chinese	162	135	52	43	392
totals	544	453	161	177	1335

Now we need to calculate the expected frequency for English and group O. There are 609 English and 1335 people altogether so $^{609}/_{1335}$ of the people are English, and from H_0 we know that there is no connection between blood group and nationality, so there should be $^{609}/_{1335}$ of those with group O who are also English

$\Rightarrow$ expected frequency for English and group O is $\frac{609}{1335} \times 544 = \frac{609 \times 544}{1335} = 248.2$

this can become automatic if you notice that you just multiply the totals for the row and column concerned and divide by the total number

	O	A	B	AB	totals
English	$\frac{609\times544}{1335} = 248.2$	$\frac{609\times453}{1335} = 206.6$	$\frac{609\times161}{1335} = 73.4$	$\frac{609\times177}{1335} = 80.7$	608.9
African	$\frac{334\times544}{1335} = 136.1$	$\frac{334\times453}{1335} = 113.3$	$\frac{334\times161}{1335} = 40.3$	$\frac{334\times177}{1335} = 44.3$	334
Chinese	$\frac{392\times544}{1335} = 159.7$	$\frac{392\times453}{1335} = 133.0$	$\frac{392\times161}{1335} = 47.3$	$\frac{392\times177}{1335} = 52.0$	392
totals	544	452.9	161	177	1335

The value of χ^2 is calculated below

Observed frequency	*Expected frequency*	$\frac{(O-E)^2}{E}$
235	248.2	0.70
212	206.6	0.14
79	73.4	0.43
83	80.7	0.07
147	136.1	0.87
106	113.3	0.47
30	40.3	2.63
51	44.3	1.01
162	159.7	0.03
135	133.0	0.03
52	47.3	0.47
43	52.0	1.56
		8.41

We have $\nu = (4-1)(3-1) = 6$ degrees of freedom and $\chi^2_6(5\%) = 12.592$.

We have calculated $\chi^2 = 8.41 < 12.592$

$\Rightarrow$ do not reject H_0 and therefore conclude that there is no connection between nationality and blood group

20 Regression and correlation

Spearman's rank correlation coefficient

Ranking and equal ranks

Ranking is putting a list of figures in order and giving each one its position or ***rank***.

Equal numbers are given the average of the ranks they would have had if all had been different.

Example: Rank the following numbers: 45, 65, 76, 56, 34, 45, 23, 67, 65, 45, 81, 32.

Solution: First put in order and give ranks as if all were different: then give the average rank for those which are equal.

Numbers:	81	76	67	**65**	**65**	56	**45**	**45**	**45**	34	32	23
Actual rank	1	2	3	4=	4=	6	7=	7=	7=	10	11	12
Rank (if all different)	1	2	3	**4**	**5**	6	**7**	**8**	**9**	10	11	12
average for equal ranks				$\frac{4+5}{2}=4\frac{1}{2}$				$\frac{7+8+9}{3}=8$				
Modified rank	1	2	3	**4½**	**4½**	6	**8**	**8**	**8**	10	11	12

You must now calculate the PMCC, **not** Spearman, using the modified ranks.

Spearman's rank correlation coefficient

To compare two sets of rankings for the same n items, first find the difference, d, between each pair of ranks and then calculate Spearman's rank correlation coefficient

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

This is the same as the product moment correlation coefficient of the two sets of ranks and so we know that

$r_s = +1$ means rankings are in perfect agreement,

$r_s = -1$ means rankings are in exact reverse order,

$r_s = 0$ means that there is no correlation between the rankings.

Example: Ten varieties of coffee labelled A, B, C, ..., J were tasted by a man and a woman. Each ranked the coffees from best to worst as shown.

Man:	G	H	C	D	A	E	B	J	I	F
Woman:	C	B	H	G	J	D	I	E	F	A

Find Spearman's rank correlation coefficient.

Solution: Rank for each person, find d and then r_s.

Coffee	Man	Woman	d	d^2
A	5	10	-5	25
B	7	2	5	25
C	3	1	2	4
D	4	6	-2	4
E	6	8	-2	4
F	10	9	1	1
G	1	4	-3	9
H	2	3	-1	1
I	9	7	2	4
J	8	5	3	9
				86

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 86}{10\times 99} = 0.521212 = 0.521 \quad \text{to 3 S.F.}$$

Spearman or PMCC

Use of Spearman's rank correlation coefficient

(i) Use when one, or both, sets of data are **not** from a normal population.
(ii) Use when the data does not have to be measured on scales or in units (probably not normal).
(iii) Use when data is subjective – e.g. judgements in order of preference (not normal).
(iv) Can be used if the scatter graph indicates a non-linear relationship between the variables, since the PMCC is used to indicate **linear** correlation.
(v) Do **not** use for tied ranks (Spearman formula depends on non-tied ranks).

Use of Product moment correlation coefficient

(i) Use when ranks are tied – see above: modify the ranks and then use PMCC on the modified ranks.
(ii) Use when both sets of figures are normally distributed (this will not be the case when using ranks).
(iii) Use when the scatter diagram indicates a linear relationship between the variables – i.e. when the points lie close to a straight line.

Testing for zero correlation

N.B. the tables give figures for a ONE-TAIL test

Product moment correlation coefficient

PMCC tests to see if there is a **linear connection** between the variables. For strong correlation, the points on a scatter graph will lie close to a straight line.

Reminder: PMCC $= \rho = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

where $S_{xx} = \sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}$, $S_{yy} = \sum y_i^2 - \dfrac{\left(\sum y_i\right)^2}{n}$, $S_{xy} = \sum x_i y_i - \dfrac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$.

Example: The product moment correlation coefficient between 40 pairs of values is +0.52. Is there any evidence of correlation between the pairs at the 5% level?

Solution: H_0: There is no correlation between the pairs, $\rho = 0$.

H_1: There is correlation, positive or negative, between the pairs, $\rho \neq 0$, two-tail test

From tables for $n = 40$ which give **one**-tail figures, we must look at the 2.5% column and the critical values are ± 0.3120

The calculated figure is $0.52 > 0.3120$ and so is significant

$\Rightarrow$ we reject H_0 and conclude that there is some correlation (positive or negative) between the pairs.

Spearman's rank correlation coefficient

Spearman tests to see if there is a **connection** (or correlation) between the **ranks**.

Example: It is believed that a person who absorbs a drug well on one occasion will also absorb a drug well on another occasion. Tests on ten patients to find the percentage of drug absorbed gave the following value for Spearman' rank correlation coefficient, $r_s = 0.634$. Is there any evidence at the 5% level of a positive correlation between the two sets of results.

Solution: H_0: There is no correlation between the two sets of results, $\rho_s = 0$,

H_1: There is positive correlation between the two sets of results, $\rho_s > 0$, one-tail test.

From the tables for $n = 10$ and a one-tail test the critical value for 5% is 0.5364.

The calculated value is $0.634 > 0.5364$ which is significant

$\Rightarrow$ reject H_0; conclude that there is evidence of positive correlation between the two sets of results.

Note that this shows correlation between the **ranks** of the two sets of results.

Comparison between PMCC and Spearman

Example: A random sample of 8 students sat examinations in Geography and Statistics. The product moment correlation coefficient between their results was 0·572 and the Spearman rank correlation coefficient was 0·655.

(*a*) Test both of these values for positive correlation. Use a 5% level of significance.

(*b*) Comment on your results.

Solution:

(*a*) $H_0: \rho = 0$; $H_1: \rho > 0$

For the PMCC

the 5% Critical Value is **0·6215**

$0{\cdot}572 < \mathbf{0{\cdot}6215}$ $\Rightarrow$ not significant at %5

$\Rightarrow$ there is evidence that there is no positive correlation.

For Spearman's rank correlation coefficient

the 5% Critical Value is **0·6429**

$0{\cdot}655 > \mathbf{0{\cdot}6429}$ $\Rightarrow$ significant at 5%

$\Rightarrow$ there is evidence of positive correlation.

(*b*) From the PMCC there is not enough evidence to conclude that as Statistics marks increased Geography marks also increased
– i.e. conclude that the points on a scatter diagram do not lie close to a straight line.

From Spearman's rank correlation coefficient there is evidence that students **ranked** highly in Statistics were also **ranked** highly in Geography, or people with **high scores** in Statistics also had **high scores** in Geography

21 Quality of tests and estimators

22 One-sample procedures

23 Two-sample procedures