Statistics 3

1 Combinations of random variables .................................................................................................................3
   Expected mean and variance for \( X \pm Y \) ........................................................................................................3
   Reminder ..........................................................................................................................................................3
   Combining independent normal random variables \( Y \) ..............................................................................3

2 Sampling .........................................................................................................................................................4
   Methods of collecting data ............................................................................................................................4
   Taking a census .................................................................................................................................................4
   Sampling ..........................................................................................................................................................4
   Simple random sampling ..............................................................................................................................5
   Using random number tables ..........................................................................................................................5
   Systematic sampling .......................................................................................................................................5
   Stratified sampling ..........................................................................................................................................5
   Sampling with and without replacement .........................................................................................................6
   Quota sampling ............................................................................................................................................6
   Primary data ..................................................................................................................................................7
   Secondary data ...............................................................................................................................................7

3 Biased & unbiased estimators ...........................................................................................................................8
   Unbiased estimators of \( \mu \) and \( \sigma^2 \) ........................................................................................................11
   Estimating \( \mu \) and \( \sigma^2 \) from a sample ....................................................................................................11

4 Confidence intervals and significance tests ...................................................................................................14
   Sampling distribution of the mean ..................................................................................................................14
   Central limit theorem and standard error ......................................................................................................15
   Confidence intervals .......................................................................................................................................16
   Central Limit Theorem Example ....................................................................................................................16
   Significance testing—variance of population known ....................................................................................18
   Mean of normal distribution ...........................................................................................................................18
   Difference between means of normal distributions .......................................................................................19
   Significance testing—variance of population NOT known, large sample ..................................................21
   Mean of normal distribution ...........................................................................................................................21
   Important assumption .....................................................................................................................................21
   Difference between means .............................................................................................................................22
5  Goodness of fit, $\chi^2$ test ..................................................................................................................... 23
   General points........................................................................................................................................... 23
   Discrete uniform distribution................................................................................................................... 23
   Continuous uniform distribution .............................................................................................................. 24
   Binomial distribution ............................................................................................................................... 24
   Poisson distribution ................................................................................................................................. 24
   The normal distribution ........................................................................................................................... 26
   Contingency tables................................................................................................................................... 27
6  Regression and correlation................................................................................................................ 29
   Spearman’s rank correlation coefficient ................................................................................................. 29
   Ranking and equal ranks .......................................................................................................................... 29
   Spearman’s rank correlation coefficient .................................................................................................. 29
   Spearman or PMCC ................................................................................................................................... 30
   Testing for zero correlation ..................................................................................................................... 31
   Product moment correlation coefficient .................................................................................................. 31
   Spearman’s rank correlation coefficient .................................................................................................. 31
   Comparison between PMCC and Spearman............................................................................................ 32
7  Appendix............................................................................................................................................. 33
   Combining random variables .................................................................................................................. 33
   $E[X + Y] = E[X] + E[Y]$ ............................................................................................................................ 33
   $Var[X + Y] = Var[X] + Var[Y]$ ................................................................................................................ 34
   Unbiased & biased estimators .................................................................................................................. 34
   Unbiased estimators .................................................................................................................................. 34
   Biased Estimators ..................................................................................................................................... 35
   Unbiased estimates of population mean and variance ............................................................................ 35
   Unbiased estimate of the mean .................................................................................................................. 35
   Unbiased estimate of the variance of the population ............................................................................. 36
   Bias............................................................................................................................................................ 37
   Probability generating functions............................................................................................................. 39
   Expected mean and variance for a p.g.f. .................................................................................................. 39
   Mean and variance of a Binomial distribution ......................................................................................... 40
   Mean and variance of a Poisson distribution ......................................................................................... 40
Index............................................................................................................................................................ 41
1 Combinations of random variables

Expected mean and variance for $X \pm Y$

Reminder

For any two random variables $X$ and $Y$

\[ E[aX] = aE[X] \quad \text{and} \quad \text{Var}[aX] = a^2 \text{Var}[X] \]

\[ E[X + Y] = E[X] + E[Y] \quad \text{and} \quad E[X - Y] = E[X] - E[Y] \]

and for two independent random variables

\[ \text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] \quad \text{and} \quad \text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y]. \]

Combining independent normal random variables $Y$

If $X_1$ and $X_2$ are independent normal random variables

\[ X_1 \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad X_2 \sim N(\mu_2, \sigma_2^2) \]

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

\[ X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \]

and \[ X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \]

Example: $X_1$ and $X_2$ are independent normal random variables

$X_1 \sim N(21, 12)$ and $X_2 \sim N(9, 6)$.

Find the expected mean and standard deviation of $X_1 - 2X_2$.

Solution:

\[ E[X_1] = 21, \quad \text{Var}[X_1] = 12 \quad \text{and} \quad E[X_2] = 9, \quad \text{Var}[X_2] = 6 \]

\[ \Rightarrow \quad E[2X_2] = 2E[X_2] = 2 \times 9 = 18 \]

and \[ \text{Var}[2X_2] = 2^2 \times \text{Var}[X_2] = 4 \times 6 = 24 \]

\[ \Rightarrow \quad E[X_1 - 2X_2] = E[X_1] - E[2X_2] = 21 - 18 = 3 \]

and \[ \text{Var}[X_1 - 2X_2] = \text{Var}[X_1] + \text{Var}[2X_2] = 12 + 24 = 36 \]

\[ \Rightarrow \quad \text{the expected mean and standard deviation of } X_1 - 2X_2 \text{ are } 3 \quad \text{and} \quad \sqrt{36} = 6. \quad \text{Answer} \]
Example: The weights of empty coffee jars are normally distributed with mean 0.1 kg and standard deviation 0.02 kg. The weight of coffee in the jars is normally distributed with mean 1 kg and standard deviation 0.06 kg.

Find the distribution of 12 full jars of coffee.

What is the probability that 12 full jars weigh more than 13.5 kg?

Solution: Let $X_1, X_2, \ldots, X_{12}$ be the weights of 12 empty jars and $Y_1, Y_2, \ldots, Y_{12}$ be the weights of coffee in the jars. $X \sim N(0.1, 0.02^2)$ and $Y \sim N(1, 0.06^2)$.

Let $W$ be the total weight of 12 full jars then $W = X_1 + X_2 + \ldots + X_{12} + Y_1 + Y_2 + \ldots + Y_{12}$.

Then $E[W] = 12 E[X] + 12 E[Y] = 12 \times 0.1 + 12 \times 1 = 13.2$

and, assuming independence,

$\text{Var}[W] = 12 \text{Var}[X] + 12 \text{Var}[Y] = 12 \times 0.02^2 + 12 \times 0.06^2 = 0.048$.

As we are combining normal distributions the distribution for 12 full jars is $N(13.2, 0.048)$. Answer

The probability that 12 full jars weigh more than 13.5 kg is

$$1 - \Phi \left( \frac{13.5 - 13.2}{\sqrt{0.048}} \right) = 1 - \Phi(1.37) = 0.0853 \text{ to 3 s.f. Answer}.$$ 

2 Sampling

Methods of collecting data

Taking a census

A census involves observing every member of a population and is used if

the size of the population is small

or if extreme accuracy is required.

Advantages

it should give a completely accurate result, a full picture.

Disadvantages

very time consuming and expensive

it cannot be used when testing process destroys article being tested

information is difficult to process because there is so much of it.

Sampling

Sampling involves observing or testing a part of the population. It is cheaper but does not give such a full picture. The size of the sample depends on the accuracy desired (for a varied population a large sample will be required to give a reasonable accuracy).
Simple random sampling
Every member of the population must have an equal chance of being selected.

Using random number tables
To take a simple random sample of size $n$ from a population of $N$ sampling units first make a list and give each member of the population a number. Then use random number tables to select the sample.
We ignore any numbers which do not refer to a member of the population – for example using three figure random numbers for a population numbered from 001 to 659 we would ignore numbers from 660 to 999.
Also we ignore the second occurrence of the same number.

Advantages
- the numbers are truly random and free from bias
- it is easy to use
- each member has a known equal chance of selection

Disadvantages
- it is not suitable when the sample size is large.

Lottery sampling
A sampling frame is needed – identifying each member of the population. The name or number of each member is written on a ticket (all the same size, colour and shape), and the tickets are all put in a container which is then shaken. Tickets are then drawn without replacement.

Advantages
- the tickets are drawn at random.
- it is easy to use.
- each ticket has a known chance of selection (considered as constant as long as the sample size is much smaller than the total number of tickets).

Disadvantages
- it is not suitable for a large sample
- a sampling frame is needed.

Systematic sampling
First make an ordered list, and divide into equal groups each of size 50 (or??).
Second select every 50$^{\text{th}}$ (or ??) member from the list.
In order to make sure that the first on the list is not automatically selected random number tables must be used to select the member in the first group, then select every 50$^{\text{th}}$ (or ??) after that.

Used when the population is too large for simple random number sampling.

Advantages
- simple to use
- suitable for large samples

Disadvantages
- only random if the ordered list is truly random.
- it can introduce bias
Stratified sampling

First divide the population into exclusive (distinct) groups or strata and then select a sample so that the proportion of each stratum in the sample equals the proportion of that stratum in the population.

Example: How would you take a stratified sample of 50 children from a school of 500 pupils divided as follows:

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper sixth</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Lower sixth</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Fifth form</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>Fourth form</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>Third form</td>
<td>50</td>
<td>60</td>
</tr>
</tbody>
</table>

Solution: As 50 is \(\frac{1}{10}\) of the total population, \(\frac{1}{10}\) of each stratum should be selected in the sample. Thus the sample would comprise:

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper sixth</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Lower sixth</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Fifth form</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Fourth form</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Third form</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

and simple random number sampling would be used within each stratum.

Used when

the sample is large
the population divides naturally into mutually exclusive groups.

Advantages

it can give more accurate estimates (or a more representative picture) than simple random number sampling when there are clear strata present.
It reflects the population structure.

Disadvantages

within the strata the problems are the same as for any simple random sample
if the strata are not clearly defined they may overlap.

Sampling with and without replacement

Simple random sampling is sampling without replacement in which each member of population can be selected at most once.

In sampling with replacement each member of the population can be selected more than once: this is called unrestricted random sampling.
Quota sampling

This is a non-random method.

First decide on groups into which the population is divided and a number from each group to be interviewed to form *quotas*.

Then go out and interview and enter each result into the relevant quota. If someone refuses to answer or belongs to a quota which is already full then ignore that persons reply and continue interviewing until all quotas are full.

**Used when** it is not possible to use random methods - for example when the whole population is not known (homeless in a big city).

**Advantages**
- can be done quickly as a representative sample can be obtained with a small sample size
- costs are kept to a minimum
- administration is fairly easy.

**Disadvantages**
- it is not possible to estimate the sampling errors (as it is not a random process)
- interviewer may not put into correct quota
- non-responses are not recorded
- it can introduce interviewer bias

Primary data

Primary data is data collected by or on behalf of the person who is going to use the data.

**Advantages**
- collection method is known
- accuracy is known
- exact data needed are collected

**Disadvantages**
- costly in time and effort

Secondary data

Secondary data is data *not* collected by or on behalf of the person who is going to use it. The data are second-hand – e.g. government census statistics.

**Advantages**
- cheap to obtain
- large quantity available (e.g. internet)
- much has been collected year on year and can be used to plot trends

**Disadvantages**
- collection method may not be known
- accuracy may not be known
- it can be in a form which is difficult to handle
- bias is not always recognised.
3 Biased & unbiased estimators

Example: A bag contains a large number of coins, of which \( \frac{2}{5} \) are 2p coins and \( \frac{3}{5} \) are 5p coins.

**(a)** \( X \) is the value of a single coin draw from the bag. Find the expected mean of all coins in the bag, \( \mu = E[X] \).

Samples of size 3 are now drawn from the bag.

**(b)** Find the sampling distribution of and the expected value of (i) the median, and (ii) the mean.

**(c)** (i) The median, \( Q_2 \), is used as an estimator of the mean of all the coins, \( \mu \). Show that \( Q_2 \) is a biased estimator of \( \mu \), and find the bias.

(ii) The mean, \( \bar{X} \), is used as an estimator of the mean of all the coins, \( \mu \). Show that \( \bar{X} \) is an unbiased estimator of \( \mu \).

**(d)** \( kQ_2 \) is now used as an unbiased estimator of the mean of all the coins. Find the value of \( k \).

Solution:

**(a)** \( \mu = E[X] = \sum x_i p_i = \frac{2}{5} \times 2 + \frac{3}{5} \times 5 = 3 \cdot 8 \)

**(b)**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Probability</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2,2,2)</td>
<td>( \left( \frac{2}{5} \right)^3 ) = ( \frac{8}{125} )</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>(2,2,5), (2,5,2), (5,2,2)</td>
<td>( 3 \times \left( \frac{2}{5} \right)^2 \times \frac{3}{5} = \frac{36}{125} )</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>(2,5,5), (5,2,5), (5,5,2)</td>
<td>( 3 \times \frac{2}{5} \times \left( \frac{3}{5} \right)^2 = \frac{54}{125} )</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>(5,5,5)</td>
<td>( \left( \frac{3}{5} \right)^3 ) = ( \frac{27}{125} )</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Sampling distribution

(i) Median

\( Q_2 = x_i \quad p_i \quad x_i p_i \)


<table>
<thead>
<tr>
<th>Sample</th>
<th>Probability</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( \frac{8+36}{125} ) = ( \frac{88}{125} )</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>( \frac{54+27}{125} ) = ( \frac{405}{125} )</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

\( \Rightarrow E[Q_2] = \frac{493}{125} = 3.944 \)

(ii) Mean

\( \bar{X} = x_i \quad p_i \quad x_i p_i \)


<table>
<thead>
<tr>
<th>Sample</th>
<th>Probability</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( \frac{8}{125} ) = ( \frac{16}{125} )</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>( \frac{36}{125} ) = ( \frac{108}{125} )</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>( \frac{54}{125} ) = ( \frac{216}{125} )</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>( \frac{27}{125} ) = ( \frac{135}{125} )</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

\( \Rightarrow E[\bar{X}] = \frac{475}{125} = 3.8 \)
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 $E[X] = t$.

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

(c) (i) The median $Q_2$ is used as the estimate of the mean.
From part (a) we know that the true value of the mean $\mu$ is 3.8, and in part (b) we have shown that $E[Q_2] = 3.944$
$\Rightarrow$ $Q_2$ is a biased estimator of $\mu$
and the bias is $|E[Q_2] - \text{true value of } \mu| = |3.944 - 3.8| = 0.144$

(ii) The mean $\bar{X}$ is used as the estimate of the mean.
From part (a) we know that the true value of the mean $\mu$ is 3.8, and in part (b) we have shown that $E[\bar{X}] = 3.8$
$\Rightarrow$ $E[\bar{X}] = \text{the true value of the mean}$
$\Rightarrow \bar{X}$ is an unbiased estimator of $\mu$,

(d) If we now use $kQ_2$ as an unbiased estimator of the mean value of all the coins.
$E[kQ_2] = kE[Q_2] = \frac{493}{125}k$
But the true mean $\mu = \frac{475}{125}$
If $kQ_2$ is an unbiased estimator of $\mu$, $E[kQ_2] = \text{true value of } \mu$
$\Rightarrow \frac{493}{125}k = \frac{475}{125} \Rightarrow k = \frac{475}{493}$.

Example: A sample of size 3 is drawn from a binomial distribution B(10, 0.25) and the mean, $\bar{X}$, is calculated.
The probability of success, $p$, is estimated by $\hat{p} = \frac{1}{10} \bar{X}$. Show that $\hat{p}$ is an unbiased estimator of $p$.

Solution: $E[X] = np = 10 \times 0.25 = 2.5$
For a sample $\{X_1, X_2, X_3\}$, $\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3)$
$\Rightarrow E[\bar{X}] = E\left[\frac{1}{3}(X_1 + X_2 + X_3)\right] = \frac{1}{3}(E[X_1] + E[X_2] + E[X_3])$
$\Rightarrow E[\bar{X}] = \frac{1}{3} \times 3 \times 2 \cdot 5 = 2 \cdot 5$ since $E[X_i] = E[X] = 2 \cdot 5$, for $i = 1, 2, 3$
$\Rightarrow E[\hat{p}] = E\left[\frac{1}{10} \bar{X}\right] = \frac{1}{10} E[\bar{X}] = \frac{1}{10} \times 2 \cdot 5 = 0 \cdot 25$, which is the true value of $p$
$\Rightarrow \hat{p} = \frac{1}{10} \bar{X}$ is an unbiased estimator of $p$. 
Example: A sample of size 4 is drawn from a continuous uniform distribution, U[3, \( \beta \)]. The mean of the sample, \( \bar{X} \), is calculated.

The upper limit, \( \beta \), is estimated by \( \hat{\beta} = 2\bar{X} - 3 \). Show that \( \hat{\beta} \) is an unbiased estimator of \( \beta \).

Solution: \( E[X] = \frac{1}{2} (3 + \beta) \)

For a sample \( \{X_1, X_2, X_3, X_4\} \), \( \bar{X} = \frac{1}{4} (X_1 + X_2 + X_3 + X_4) \)

\( \Rightarrow E[\bar{X}] = E\left[ \frac{1}{4} (X_1 + X_2 + X_3 + X_4) \right] = \frac{1}{4} (E[X_1] + E[X_2] + E[X_3] + E[X_4]) \)

\( \Rightarrow E[\bar{X}] = \frac{1}{4} \times 4 \times E[X] = E[X] \) since \( E[X_i] = E[X] \), for \( i = 1, 2, 3, 4 \)

\( \Rightarrow E[\bar{X}] = \frac{1}{2} (3 + \beta) \)

\( \Rightarrow E[\hat{\beta}] = E[2\bar{X} - 3] = 2E[\bar{X}] - 3 = 2 \times \frac{1}{2} (3 + \beta) - 3 = \beta, \)

\( \Rightarrow E[\hat{\beta}] = \beta \) which is the true value of the (unknown) upper limit, \( \beta \).

\( \Rightarrow \hat{\beta} = 2\bar{X} - 3 \) is an unbiased estimator of \( \beta \).

There is more on biased and unbiased estimators in the Appendix.
Unbiased estimators of $\mu$ and $\sigma^2$

Estimating $\mu$ and $\sigma^2$ from a sample

We usually do not know the mean, $\mu$, and the variance, $\sigma^2$, of a population. To estimate these values we take a sample $\{X_1, X_2, X_3, ..., X_n\}$ of size $n$ and calculate

- the sample mean, $\bar{X} = \frac{1}{n}\sum X_i$, and
- $(sd)_x^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum(X - \bar{X})^2$

these can be compared with the formulae for population variance from the S1 module.

It can be shown that $E[\bar{X}] = \mu$, the true value of $\mu$.

$\Rightarrow \bar{X} = \hat{\mu}$ is an unbiased estimator of the population mean $\mu$.

It can be shown that $E[(sd)_x^2] = \frac{n-1}{n} \sigma^2$ \hspace{1cm} \text{I}

$\Rightarrow (sd)_x^2$ is a biased estimator of $\sigma^2$.

\text{I} \Rightarrow E\left[\frac{n}{n-1}(sd)_x^2\right] = \frac{n}{n-1} \times \frac{n-1}{n} \sigma^2 = \sigma^2$, the true value of the variance

$\Rightarrow \frac{n}{n-1}(sd)_x^2 = \hat{\sigma}^2$ is an unbiased estimator of the population variance, $\sigma^2$.

(Proofs of these results are given in the Appendix.)

Note: the Edexcel course uses both the letters $S^2$ and $s_x^2$ to mean the unbiased estimate of $\sigma^2$.

Also, the term Sample Variance is used to denote the unbiased estimate of $\sigma^2$, the variance of the population.

In these notes I shall always think of the variance, $(sd)_x^2$, as $\frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum(X - \bar{X})^2$

To find $S^2$ or $s_x^2$, the unbiased estimator for $\sigma^2$:

Calculate $(sd)_x^2$, and then multiply by $\frac{n}{n-1}$.
Example:
The weights of a sample of five chocolate bars produced by a machine were 56, 53, 57, 51 and 54 grams. Find unbiased estimators for the weight of all chocolate bars produced by that machine.

Solution:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$X - \bar{X}$</th>
<th>$(X - \bar{X})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>1.8</td>
<td>3.24</td>
</tr>
<tr>
<td>53</td>
<td>-1.2</td>
<td>1.44</td>
</tr>
<tr>
<td>57</td>
<td>2.8</td>
<td>7.84</td>
</tr>
<tr>
<td>51</td>
<td>-3.2</td>
<td>10.24</td>
</tr>
<tr>
<td>54</td>
<td>-0.2</td>
<td>0.04</td>
</tr>
</tbody>
</table>

$\Rightarrow \quad \bar{X} = \frac{1}{n} \sum X = \frac{271}{5} = 54.2$

$\Rightarrow \quad (sd)_x^2 = \frac{1}{n} \sum (X - \bar{X})^2 = \frac{22.8}{5} = 4.56$

$\Rightarrow \quad \sigma^2 = \frac{n}{n-1} (sd)_x^2 = \frac{5}{4} \times 4.56 = 5.7$

Answer

Unbiased estimators for the mean and variance of all chocolate bars are 54.2 grams and 5.7 grams$^2$.

Example: The volume of water in each of a sample of 14 litre bottles of water from a day’s production is taken. The results are shown below, in $ml$.

1023, 1019, 1004, 1011, 1023, 1014, 1017, 1020, 1020, 1010, 1025, 1007, 1016, 1019

Find unbiased estimates for the mean and variance of all bottles produced on that day.

Solution: First find the sample mean, $\bar{X} = \frac{14228}{14} = 1016.286$. …

(finding $X - \bar{X}$ each time) would give unpleasant arithmetic,

so use $(sd)_x^2 = \frac{1}{n} \sum X^2 - \bar{X}^2$

$\sum X^2 = 14460232$

$\Rightarrow \quad (sd)_x^2 = \frac{1446032}{14} - \left(\frac{14228}{14}\right)^2 = 37.06122…$

$\Rightarrow \quad S^2 = s_x^2 = \frac{n}{n-1} (sd)_x^2 = \frac{14}{14-1} \times 37.06122… = 39.91209…$

Answer

Unbiased estimators for the mean and variance of the whole day’s production are 1016.3 $ml$ and 39.91 $ml^2$. 
Example: The weights of a sample of 15 packets of biscuits are recorded and give the following results.

\[ \sum X = 3797 \text{ grams}, \quad \sum X^2 = 973692. \]

Find unbiased estimators for the mean and variance of all biscuits produced by this process.

Solution: \[ \bar{X} = \frac{3797}{15} = 253 \cdot 133 \ldots = 252 \cdot 1 \text{ grams}. \]

\[
\begin{align*}
(sd)_X^2 &= \frac{1}{n} \sum X^2 - \bar{X}^2 = \frac{973692}{15} - \left(\frac{3797}{15}\right)^2 = 836 \cdot 3156 \\
\Rightarrow \sigma^2 &= \frac{15}{14} \times 836 \cdot 3156 \ldots = 896 \cdot 0524 \ldots = 896 \cdot 1 \text{ grams}^2.
\end{align*}
\]

Answer Unbiased estimators are \( \hat{\mu} = 252 \cdot 1 \text{ g}, \) and \( \hat{\sigma}^2 = 896 \cdot 1 \text{ g}^2. \)

Example: The lengths of 10 rods are measured, and the sample has mean, \( \bar{X} = 26 \cdot 7 \text{ cm} \) and variance \( s^2 = 76 \cdot 9 \text{ cm}^2. \) An eleventh rod has length 30 cm.

Find (a) the mean and (b) the variance of the sample of 11 rods.

Solution: (a) With the sample mean there are no complications.

For \( n = 10, \) \[ \bar{X}_{10} = \frac{1}{10} \sum_{i=1}^{10} X_i = 26 \cdot 7 \Rightarrow \sum_{i=1}^{10} X_i = 267 \]

For \( n = 11, \) \[ \sum_{i=1}^{11} X_i = 267 + 30 = 297 \Rightarrow \bar{X}_{11} = \frac{1}{11} \sum_{i=1}^{11} X_i = \frac{297}{11} = 27 \text{ cm} \]

(b) WARNING: The question refers to the variance of the sample, which means the unbiased estimate of the variance of the population.

\[ s_{10}^2 = 76 \cdot 9 = \frac{10}{(10-1)} \times (sd)_{10}^2 \Rightarrow (sd)_{10}^2 = \frac{9}{10} \times 76 \cdot 9 = 69 \cdot 21 \]

\[ \Rightarrow (sd)_{10}^2 = \frac{1}{10} \sum_{i=1}^{10} X_i^2 - \bar{X}_{10}^2 = 69 \cdot 21 \]

\[ \Rightarrow \sum_{i=1}^{10} X_i^2 = 692 \cdot 1 + 10 \times 26 \cdot 7^2 = 7821 \]

with extra rod \( \Rightarrow \sum_{i=1}^{11} X_i^2 = 7821 + 30^2 = 8721 \)

\[ \Rightarrow (sd)_{11}^2 = \frac{1}{11} \sum_{i=1}^{11} X_i^2 - \bar{X}_{11}^2 = \frac{8721}{11} - 27^2 = \frac{702}{11} \]

\[ \Rightarrow s_{11}^2 = \frac{11}{(11-1)} (sd)_{11}^2 = \frac{11}{10} \times \frac{702}{11} = 70 \cdot 2 \text{ cm}^2 \]

For 11 rods, sample mean is 27 cm, and sample variance is 70 \cdot 2 \text{ cm}^2.
4 Confidence intervals and significance tests

Sampling distribution of the mean

$X$ is a random variable draw from a population with mean $\mu$ and standard deviation $\sigma$.

If $\{X_1, X_2, \ldots, X_n\}$ is a random sample of size $n$ with mean $\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$

then $E[X_i] = \mu$, and $\text{Var}[X_i] = \sigma^2$, for $i = 1, 2, 3, \ldots, n$

and the expected mean of the population of sample means is

$$E[\bar{X}] = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]$$

$$= \frac{1}{n}(E[X_1] + E[X_2] + \ldots + E[X_n]) = \frac{1}{n}(\mu + \mu + \ldots + \mu)$$

$$= \mu$$

Also the expected variance of the population of sample means is

$$\text{Var}[\bar{X}] = \text{Var}\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]$$

assuming that all the $X_i$ are independent

$$= \frac{1}{n^2}(\text{Var}[X_1] + \text{Var}[X_2] + \ldots + \text{Var}[X_n]) = \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

This means that if very many samples were taken and the mean of each sample calculated then the mean of these means would be $\mu$ and the variance of these means would be $\frac{\sigma^2}{n}$.

It can also be shown that the sample means form a Normal distribution (provided that $n$ is ‘large enough’).

We can then say that for samples drawn from a population with mean $\mu$ and variance $\sigma^2$,

the sampling distribution of the mean is $N(\mu, \frac{\sigma^2}{n})$. 


Central limit theorem and standard error

The central limit theorem states that

If \( \{X_1, X_2, \ldots, X_n\} \) is a random sample of size \( n \) drawn from any population with mean \( \mu \) and variance \( \sigma^2 \) then the population of sample means

(i) has expected mean \( \mu \)

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

(iii) forms a normal distribution if \( n \) is ‘large enough’.

i.e. \( \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \).

The central limit theorem is used for sampling when the sample size is ‘large’ (> about 50) as the population of sample means is then approximately normal whatever the distribution of the original population.

The standard error of the sample mean is \( \frac{\sigma}{\sqrt{n}} \).

Example: A sample of size 50 is taken from a population of eggs with mean 23.4 grams and variance 36 grams².

(i) Find the probability that a single egg weighs more than 25 grams.

(ii) Find the probability that the sample mean is larger than 25.

(iii) What assumptions did you make?

Solution:

(i) The weight of a single egg, \( X \sim N(23.4, 6^2) \)

\[ \Rightarrow P(X > 25) = \Phi \left( \frac{25 - 23.4}{6} \right) = \Phi(0.27) = 0.6064 \]

(ii) \( \mu = 23.4, \ \sigma^2 = 36 \)

The sample mean \( \bar{X} \sim N\left(23.4, \frac{6}{\sqrt{50}}\right) ^2 \)

\[ \Rightarrow \text{standard error is} \quad \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.848528137 \]

\[ \bar{X} \sim N(23.4, 0.8485 ^2) \]

\[ \Rightarrow P(\bar{X} > 25) = 1 - \Phi \left( \frac{25 - 23.4}{0.8485} \right) \]

\[ = 1 - \Phi(1.89) = 1 - 0.9706 \quad \text{(from Normal tables)} \]

\[ = 0.0294. \]

(iii) We have assumed the Central Limit Theorem: in particular that the sample means form a normal distribution.
Confidence intervals

Central Limit Theorem 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 $\mu$ grams then the population of all packets has mean $\mu$ and standard deviation 5.

From the central limit theorem we can assume that the sample means form an approximately normal population with mean $\mu$ and standard error (standard deviation) $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{10}} = 1.5811$.

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

$$-1.96 < Z < 1.96$$

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

$$\Rightarrow -1.9600 < \frac{253.4 - \mu}{1.5811} < 1.9600$$
$$\Rightarrow -1.9600 < \frac{253.4 - \mu}{1.5811} \text{ and } \frac{253.4 - \mu}{1.5811} < 1.9600$$
$$\Rightarrow \mu - 1.9600 \times 1.5811 < 253.4 \text{ and } 253.4 < \mu + 1.9600 \times 1.5811$$
$$\Rightarrow \mu < 253.4 + 1.9600 \times 1.5811 \text{ and } 253.4 - 1.9600 \times 1.5811 < \mu$$
$$\Rightarrow 253.4 - 1.9600 \times 1.5811 < \mu < 253.4 + 1.9600 \times 1.5811$$
$$\Rightarrow 250.3 < \mu < 256.5$$

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 $\mu$.

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 \( \mu \pm 1.9600 \times \frac{\sigma}{\sqrt{n}} \) since \( P(-1.9600 < z < 1.9600) = 0.95 \) tables give \( P(z > 1.9600) = 0.025 \)

90% confidence limits are \( \mu \pm 1.6449 \times \frac{\sigma}{\sqrt{n}} \) since \( P(-1.6449 < z < 1.6449) = 0.90 \) tables give \( P(z > 1.6449) = 0.05 \)

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

**Example:** A sample of 64 packets of cornflakes has a mean weight \( \bar{X} = 510 \) 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 \( \sigma^2 \).)*

**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),

\[ \text{standard error} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{64}} = 0.75 \]

For 90% confidence limits \( z = \pm 1.6449 \) (remember to use the 4 D.P. tables after the Normal Dist. tables), using the sample mean \( \bar{X} = 510 \) grams

\( \Rightarrow \quad 90\% \text{ confidence limits are} \quad 510 \pm 1.6449 \times 0.75 = 510 \pm 1.234 \)

\( \Rightarrow \quad \text{a 90\% confidence interval is} \quad [508.8, \ 511.2] \quad \text{to 4 S.F.} \)

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

This is a reasonable assumption as the number in the sample, 64, is large and the error introduced is therefore small.
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) \( H_0: \mu = 84 \text{ grams} \)

2) \( H_1: \mu \neq 84 \text{ grams} \) \( \Rightarrow \) 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 \( \sigma = 5 \text{ g} \).

From \( H_0 \), the mean weight of all ball bearings is assumed to be \( \mu = 84 \text{ g} \).

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 = \( \mu = 84 \text{ g} \) and

expected standard deviation of the sample means = standard error = \( \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{120}} = 0.456435... \).

We have an observed mean of 83.2

For a two-tailed test at 5%, we take 2.5% at each end

\[
P(\bar{X} < 83.2) = \Phi\left( \frac{83.2 - 84}{0.456435...} \right) = \Phi(-1.7527)
\]

\[
= (1 - \Phi(1.75)) = 0.0401
\]

\[
= 4.01\% > 2.5\%
\]

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.
Difference between means of normal distributions

Suppose that $X$ and $Y$ are two independent random variables from different normal distributions –

$X \sim \text{N}(\mu_x, \sigma_x^2)$ and $Y \sim \text{N}(\mu_y, \sigma_y^2)$.

If samples of sizes $n_x$ and $n_y$ are drawn from these populations then the distributions of the sample means, $\bar{X}$ and $\bar{Y}$ will be normal

$\bar{X} \sim \text{N}\left(\mu_x, \frac{\sigma_x^2}{n_x}\right)$ and $\bar{Y} \sim \text{N}\left(\mu_y, \frac{\sigma_y^2}{n_y}\right)$

$\mathbb{E}[\bar{X} - \bar{Y}] = \mathbb{E}[\bar{X}] - \mathbb{E}[\bar{Y}]$ and $\text{Var}[\bar{X} - \bar{Y}] = \text{Var}[\bar{X}] + \text{Var}[\bar{Y}]$

$\Rightarrow$ the differences of the sample means, $\bar{X} - \bar{Y}$, will be normal

$(\bar{X} - \bar{Y}) \sim \text{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$

Example: The weights of chocolate bars produced by two machines, $A$ and $B$, are known to be normally distributed with variances $\sigma_A^2 = 4$ and $\sigma_B^2 = 3$ grams$^2$. Samples are taken from each machine of sizes $n_A = 25$ and $n_B = 16$ which have means $\bar{X}_A = 123.1$ and $\bar{X}_B = 124.4$ grams. Is there any evidence at the 5% significance level that the bars produced by machine $B$ are heavier than the bars produced by machine $A$?

Solution:

Suppose that the mean weights for all bars from the two machines are $\mu_A$ and $\mu_B$

$H_0$: $\mu_A = \mu_B$

$H_1$: $\mu_B > \mu_A$ one-tail test at 5% level

The test statistic is the observed difference between sample means,

$\bar{X}_B - \bar{X}_A = 124.4 - 123.1 = 1.3$,

and we must find the variance of this population of differences of sample means (the sampling distribution of differences of sample means).

Consider the population of differences of sample means $\bar{X}_B - \bar{X}_A$.

Firstly, for the population of sample means for machine $B$

expected variance $\text{Var}[\bar{X}_B] = \frac{\sigma_B^2}{n_B} = \frac{3}{16}$
and secondly, for the population of sample means for machine \( A \)

expected variance \( \text{Var}[\bar{X}_A] = \frac{\sigma_A^2}{n_A} = \frac{4}{25} \)

and so for the population of differences of sample means

expected mean \( = \text{E}[\bar{X}_B - \bar{X}_A] = \mu_A - \mu_B = 0 \) \( \text{(from H}_0\text{)} \)

and \( \text{Var}[\bar{X}_B - \bar{X}_A] = \text{Var}[\bar{X}_B] + \text{Var}[\bar{X}_A] \)

\[ = \frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A} = \frac{3}{16} + \frac{4}{25} = 0.3475. \]

The observed difference, the test statistic, is \( 124.4 - 123.1 = 1.3 \)

and the standard error is \( \sqrt{0.3475} \)

The Central Limit Theorem tells us that we have a Normal distribution

\[ \Rightarrow \quad P(\text{difference} > 1.3) = 1 - \Phi\left( \frac{1.3 - 0}{\sqrt{0.3475}} \right) = 1 - \Phi(2.2053) \]

\[ = 1 - \Phi(2.20) = 1 - 0.9861 = 0.0134 \]

\[ = 1.34\% < 5\% \]

\[ \Rightarrow \quad \text{significant at 5\% level so reject } \text{H}_0 \text{ and conclude that there is evidence that machine } B \text{ is producing bars of chocolate with a heavier mean weight than machine } A. \]

Fortunately (!!) the formula for testing the difference between sample means

\[ Z = \frac{\bar{X}_B - \bar{X}_A - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \]

is in your formula booklet.
Significance testing – variance of population NOT known, large sample

When the variance of the population, $\sigma^2$, is not known and when the sample is large, we assume that the variance of the sample (meaning the unbiased estimate of $\sigma^2$), $S^2$, is the variance of the population, $\sigma^2$. As the sample is large, the error introduced is small.

**Mean of normal distribution**

*Example:* A machine usually produces steel rods with a mean length of 25.4 cm. The production manager wants to test 80 rods to see whether the machine is working correctly. The sample has mean 25.31 cm and variance 0.332 cm$^2$. Advise the production manager, using a 5% level of significance.

**Important assumption**

*The sample variance, $S^2$, is taken as, $\sigma^2$, the unbiased estimate of the variance of the population, $\sigma^2$, and we then assume that the population variance equals the unbiased estimate.*

**Solution:**

$$H_0: \mu = 25.4.$$  
$$H_1: \mu \neq 25.4 \quad \text{two-tail test, 2.5\% in each tail}$$

We assume that population variance $\sigma^2 = \text{the sample variance} \ S^2 = 0.33^2$

$\Rightarrow \sigma = 0.33$

For the population of sample means (the sampling distribution of the sample means)

expected mean = 25.4 \quad \text{from hypothesis}

and standard error $= \frac{\sigma}{\sqrt{n}} = \frac{0.33}{\sqrt{80}} = 0.036895121$

The observed sample mean is 25.31 and for a two-tail test at 5% we consider

$$\Phi \left( \frac{25.31 - 25.4}{0.036895121} \right) = \Phi(-2.4393) = 1 - \Phi(2.44) = 0.0073 < 2.5\%$$

$\Rightarrow$ reject $H_0$ and conclude that there is evidence that the machine is not producing rods of mean length 25.4 cm.
Difference between means

Example:

A firm has two machines, \( A \) and \( B \), which make steel cable. 40 cables produced by machine \( A \) have a mean breaking strain of 1728 N and variance of 752 N^2, whereas 65 cables produced by machine \( B \) have a mean breaking strain of 1757 N and a variance of 632 N^2. Is there any evidence, at the 10% level, to suggest that machine \( B \) is producing stronger cables than machine \( A \)?

Solution:

Let \( \mu_A \) and \( \mu_B \) be the mean breaking strengths of all cables produced by machines \( A \) and \( B \).

1) \( H_0: \mu_A = \mu_B \)

2) \( H_1: \mu_B > \mu_A \)

1 tail test

3) Significance Level 10%.

4) The Test

For Machine \( A \)

We assume that the population variance, \( \sigma_A^2 = \) the sample variance, \( S_A^2 = 75^2 \)

\[ \Rightarrow \text{variance of sample means} \ Var[ \bar{X}_A ] = \frac{\sigma_A^2}{n} = \frac{75^2}{40} = 140 \cdot 625. \]

For Machine \( B \)

We assume that the population variance, \( \sigma_B^2 = \) the sample variance, \( S_B^2 = 63^2 \)

\[ \Rightarrow \text{variance of sample means} \ Var[ \bar{X}_B ] = \frac{\sigma_B^2}{n} = \frac{63^2}{65} = 61.0615\ldots. \]

For differences in sample means \( \bar{X}_B - \bar{X}_A \)

Expected mean = 0 \hspace{1cm} \text{from hypothesis}

Expected variance is \( \text{Var}[ \bar{X}_B - \bar{X}_A ] = \text{Var}[ \bar{X}_B ] + \text{Var}[ \bar{X}_A ] \)

\[ = 140 \cdot 625 + 61.0615\ldots = 201 \cdot 6865\ldots \]

\[ \Rightarrow \text{standard deviation or standard error} = \sqrt{201 \cdot 6865\ldots} = 14 \cdot 2016\ldots. \]

We have an observed difference in means, test statistic, \( \bar{X}_B - \bar{X}_A = 1757 - 1728 = 29 \)

and for a 1-tail test that \( B \) is stronger we need the area to the right of 29 \hspace{1cm} \text{mean is treated as continuous, so do not use 28.5}

\[ = 1 - \Phi\left(\frac{29 - 0}{14 \cdot 2016\ldots}\right) = 1 - \Phi(2 \cdot 04) = 0.0207 < 10\% \]

which is significant at 10%.

5) Conclusion

Reject \( H_0 \) at the 10% level and conclude that there is evidence that machine \( B \) produces cables with a greater mean strength than machine \( A \).
5 Goodness of fit, $\chi^2$ test

General points

- The $\chi^2$ test can only be used to test two lists of frequencies – the observed and the expected frequencies calculated from the hypothesis.
- The expected frequencies do not need to be integers (give 2 D.P.)
- $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, where $O_i$ and $E_i$ are the observed and expected frequencies.
- If the expected frequency for a class is less than 5, then you must group this class with the next class (or two ...).
- The number of degrees of freedom, $\nu$, is the number of cells (after grouping if necessary) minus the number of linear equations connecting the frequencies.

Discrete uniform distribution

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

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>2</td>
<td>49</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>57</td>
</tr>
<tr>
<td>5</td>
<td>46</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
</tr>
</tbody>
</table>

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 $\frac{1}{6}$.

$H_1$: The die is not fair, the probability of each score is not $\frac{1}{6}$.

The expected frequencies are all $\frac{1}{6} \times 300 = 50$ and we have

<table>
<thead>
<tr>
<th>Score</th>
<th>Observed frequency</th>
<th>Expected frequency</th>
<th>$\frac{(O_i - E_i)^2}{E_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43</td>
<td>50</td>
<td>0.98</td>
</tr>
<tr>
<td>2</td>
<td>49</td>
<td>50</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
<td>50</td>
<td>0.32</td>
</tr>
<tr>
<td>4</td>
<td>57</td>
<td>50</td>
<td>0.98</td>
</tr>
<tr>
<td>5</td>
<td>46</td>
<td>50</td>
<td>0.72</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
<td>50</td>
<td>0.02</td>
</tr>
</tbody>
</table>

$\Rightarrow \chi^2 = 3.04$

and $\nu = \text{number of degrees of freedom} = n - 1 = 6 - 1 = 5$

since the total is a linear equation connecting the frequencies and is fixed.

From tables we see that $\chi^2_5(2.5\%) = 12.832 > 3.04$, 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, \( B(30, 0.3) \), is a good fit*

the means using \( O_i \) 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 \( O_i \) 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.

<table>
<thead>
<tr>
<th>number of calls</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>≥ 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>6</td>
<td>9</td>
<td>11</td>
<td>15</td>
<td>13</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

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 \( \geq 6 \) as 7 we calculate the mean and variance

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f )</th>
<th>( xf )</th>
<th>( x^2 f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>22</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>45</td>
<td>135</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>52</td>
<td>208</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>45</td>
<td>225</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>42</td>
<td>294</td>
</tr>
</tbody>
</table>

\[ \begin{align*}
69 & \quad 215 \\
915 &
\end{align*} \]
⇒ mean = \( \frac{215}{69} = 3.12 \)
and variance = \( \frac{915}{69} - \left( \frac{215}{69} \right)^2 = 3.55 \).

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 \( P(r) = \frac{\lambda^r e^{-\lambda}}{r!} \) where \( \lambda = 3.12 \), and the expected frequencies by multiplying by \( N = 69 \).

Note that the probability for \( r \geq 6 \) is found by adding the other probabilities and subtracting from 1.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( O )</th>
<th>( p )</th>
<th>( E )</th>
<th>( O )_\text{,(grouped)}</th>
<th>( E )_\text{,(grouped)}</th>
<th>( \frac{(O - E)^2}{E} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>0.044337</td>
<td>3.059234</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>0.138151</td>
<td>9.532395</td>
<td>15</td>
<td>12.59</td>
<td>0.461326</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>0.215235</td>
<td>14.8512</td>
<td>11</td>
<td>14.85</td>
<td>0.998148</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>0.223553</td>
<td>15.42515</td>
<td>15</td>
<td>15.43</td>
<td>0.011983</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>0.174145</td>
<td>12.01597</td>
<td>13</td>
<td>12.02</td>
<td>0.079900</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0.108525</td>
<td>7.488214</td>
<td>9</td>
<td>7.49</td>
<td>0.304419</td>
</tr>
<tr>
<td>( \geq 6 )</td>
<td>6</td>
<td>0.096056</td>
<td>6.627836</td>
<td>6</td>
<td>6.63</td>
<td>0.059864</td>
</tr>
<tr>
<td>( 69 )</td>
<td>( 69 )</td>
<td>( 69.01 )</td>
<td>( 1.915641 )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 \( \chi^2 (5\%) = 9.488 \).

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 $O_i$ 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.

<table>
<thead>
<tr>
<th>size of shoe</th>
<th>≤ 6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>≥ 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of pairs</td>
<td>14</td>
<td>19</td>
<td>29</td>
<td>45</td>
<td>40</td>
<td>21</td>
<td>7</td>
</tr>
</tbody>
</table>

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.

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

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

$\chi^2 (5\%) = 9.488$ 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 $5 \times 4$ 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.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>totals</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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

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.

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>A</th>
<th>B</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>235</td>
<td>212</td>
<td>79</td>
<td>83</td>
</tr>
<tr>
<td>African</td>
<td>147</td>
<td>106</td>
<td>30</td>
<td>51</td>
</tr>
<tr>
<td>Chinese</td>
<td>162</td>
<td>135</td>
<td>52</td>
<td>43</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>235</td>
<td>212</td>
<td>79</td>
<td>83</td>
<td>609</td>
</tr>
<tr>
<td>African</td>
<td>147</td>
<td>106</td>
<td>30</td>
<td>51</td>
<td>334</td>
</tr>
<tr>
<td>Chinese</td>
<td>162</td>
<td>135</td>
<td>52</td>
<td>43</td>
<td>392</td>
</tr>
<tr>
<td>totals</td>
<td>544</td>
<td>453</td>
<td>161</td>
<td>177</td>
<td>1335</td>
</tr>
</tbody>
</table>
Now we need to calculate the expected frequency for English and group O. There are 609 English and 1335 people altogether so \( \frac{609}{1335} \) of the people are English, and from H₀ we know that there is no connection between blood group and nationality, so there should be \( \frac{609}{1335} \) of those with group O who are also English.

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

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.

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>248.2</td>
<td>206.6</td>
<td>73.4</td>
<td>80.7</td>
<td>608.9</td>
</tr>
<tr>
<td>African</td>
<td>136.1</td>
<td>113.3</td>
<td>40.3</td>
<td>44.3</td>
<td>334</td>
</tr>
<tr>
<td>Chinese</td>
<td>159.7</td>
<td>133.0</td>
<td>47.3</td>
<td>52.0</td>
<td>392</td>
</tr>
<tr>
<td>totals</td>
<td>544</td>
<td>452.9</td>
<td>161</td>
<td>177</td>
<td>1335</td>
</tr>
</tbody>
</table>

The value of \( \chi^2 \) is calculated below.

<table>
<thead>
<tr>
<th>Observed frequency</th>
<th>Expected frequency</th>
<th>((O - E)^2 / E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>235</td>
<td>248.2</td>
<td>0.70</td>
</tr>
<tr>
<td>212</td>
<td>206.6</td>
<td>0.14</td>
</tr>
<tr>
<td>79</td>
<td>73.4</td>
<td>0.43</td>
</tr>
<tr>
<td>83</td>
<td>80.7</td>
<td>0.07</td>
</tr>
<tr>
<td>147</td>
<td>136.1</td>
<td>0.87</td>
</tr>
<tr>
<td>106</td>
<td>113.3</td>
<td>0.47</td>
</tr>
<tr>
<td>30</td>
<td>40.3</td>
<td>2.63</td>
</tr>
<tr>
<td>51</td>
<td>44.3</td>
<td>1.01</td>
</tr>
<tr>
<td>162</td>
<td>159.7</td>
<td>0.03</td>
</tr>
<tr>
<td>135</td>
<td>133.0</td>
<td>0.03</td>
</tr>
<tr>
<td>52</td>
<td>47.3</td>
<td>0.47</td>
</tr>
<tr>
<td>43</td>
<td>52.0</td>
<td>1.56</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 8.41 < 12.592 \]

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

We have calculated \( \chi^2 = 8.41 < 12.592 \)

\( \Rightarrow \) do not reject H₀ and therefore conclude that there is no connection between nationality and blood group.
6 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.

<table>
<thead>
<tr>
<th>Numbers:</th>
<th>81</th>
<th>76</th>
<th>67</th>
<th>65</th>
<th>65</th>
<th>56</th>
<th>45</th>
<th>45</th>
<th>45</th>
<th>34</th>
<th>32</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4=</td>
<td>4=</td>
<td>6</td>
<td>7=</td>
<td>7=</td>
<td>7=</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Rank (if all different)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>average for equal ranks</td>
<td>4½</td>
<td>4½</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Modified rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4½</td>
<td>4½</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

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 \( r_s \).

<table>
<thead>
<tr>
<th>Coffee</th>
<th>Man</th>
<th>Woman</th>
<th>( d )</th>
<th>( d^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>10</td>
<td>-5</td>
<td>25</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>6</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>8</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>4</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>H</td>
<td>2</td>
<td>3</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>J</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

\[
r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 86}{10 \times 99} = 0.521212 = 0.521 \quad \text{to 3 s.f.}
\]

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: \( \text{PMCC} = \rho = \frac{S_{yy}}{\sqrt{S_{xx}S_{yy}}} \)

where 
\[
S_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}, \quad S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}, \quad S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}.
\]

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 \( \pm 0.3120 \)

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’s 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 \( 0.6215 \)

\( 0.572 < 0.6215 \) \( \Rightarrow \) 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 \( 0.6429 \)

\( 0.655 > 0.6429 \) \( \Rightarrow \) 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
Combining random variables

Let \( X \) and \( Y \) be random variables with probability distributions
\[ X \sim \{X_1, X_2, X_3, \ldots, X_n\} \] with probabilities \((p_1, p_2, p_3, \ldots, p_n)\), and
\[ Y \sim \{Y_1, Y_2, Y_3, \ldots, Y_m\} \] with probabilities \((q_1, q_2, q_3, \ldots, q_m)\),
Then the random variable \( X + Y \) is all possible combinations \( x_i + y_j \) as \( i \) varies from 1 to \( n \) and \( j \) varies from 1 to \( m \).

Let \( P(x_i + y_j) = r_{ij} \).

Notice that
\[ \sum_{j=1}^{m} r_{ij} = r_{i1} + r_{i2} + r_{i3} + \ldots + r_{im} = p_i \]
and, similarly,
\[ \sum_{i=1}^{n} r_{ij} = q_j \]

\[ \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \]

\[ \mathbb{E}[X + Y] = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i + y_j) r_{ij} = \sum_{i} \sum_{j} x_i r_{ij} + \sum_{j} \sum_{i} y_j r_{ij} \]

\[ = \sum_{i} x_i \sum_{j} r_{ij} + \sum_{j} y_j \sum_{i} r_{ij} \]

\[ = \sum_{i=1}^{n} x_i p_i + \sum_{j=1}^{m} y_j q_j \]

\[ \Rightarrow \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \]
\[ \text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] \]

In this case we take \( X \) and \( Y \) to be independent,

\[ \Rightarrow r_{ij} = P(x_i \text{ and } y_j) = P(x_i) \times P(y_j) = p_i \times q_j. \]

Also notice that \( \sum p_i = \sum q_j = 1 \)

\[ \text{Var}[X + Y] = E[(X + Y)^2] - (E[X + Y])^2 \]

\[ = \sum_{i} \sum_{j} (x_i + y_j)^2 r_{ij} - (E[X] + E[Y])^2 \]

\[ = \sum_{i} \sum_{j} x_i^2 p_i q_j + 2 \sum_{i} \sum_{j} x_i y_j p_i q_j + \sum_{i} \sum_{j} y_j^2 p_i q_j - ((E[X])^2 + 2E[X]E[Y] + (E[Y])^2) \]

\[ = \sum_{i} x_i^2 p_i \sum_{j} q_j + 2 \sum_{i} x_i p_i \sum_{j} y_j q_j + \sum_{j} y_j^2 q_j \sum_{i} p_i - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2 \]

\[ = E[X^2] - (E[X])^2 + E[Y^2] - (E[Y])^2 \]

\[ \Rightarrow \text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y], \quad \text{if } X \text{ and } Y \text{ are independent.} \]

**Unbiased & biased estimators**

**Unbiased estimators**

An estimator \( \hat{\lambda} \) for a parameter \( \lambda \) is said to be unbiased if \( E[\hat{\lambda}] = \lambda \).

*Example:*

A bag has 468 beads of two colours, white and green. 20 beads are taken at random and the number, \( i \), of green beads in the sample is counted.

To estimate the true number of green beads, \( g \), in the bag, we calculate

\[ \hat{g} = \frac{i}{20} \times 468. \]

If \( g \) is the true number of green beads in the bag

then the probability of drawing a green bead in a single trial is \( p = \frac{g}{468} \),

and drawing \( n = 20 \) beads with replacement gives a Binomial distribution \( B(n, p) \).

Thus \( \mu = E[i] = np = 20 \times \frac{g}{468} \)

We do not actually know the number of green beads, and want to estimate this number after taking one sample

\[ \hat{g} = \frac{i}{20} \times 468. \]
We now find the expected value of this estimate

\[ E[\hat{g}] = E\left[ \frac{i}{20} \times 468 \right] = \frac{468}{20} \times E[i] = \frac{468}{20} \times 20 \times \frac{g}{468} = g, \] the true number

\[ \Rightarrow \] the expected value of the estimator, \( \hat{g} \), is equal to the true value, \( g \)

\[ \Rightarrow \] the estimator, \( \hat{g} \), is unbiased.

**Biased Estimators**

An estimator \( \hat{\lambda} \) for a parameter \( \lambda \) is said to be biased if \( E[\hat{\lambda}] \neq \lambda \).

**Example**

A naturalist wishes to estimate the number of squirrels in a wood. He first catches 50 squirrels, marks them and then releases them. Later he catches 30 squirrels and counts the number, \( i \), which have been marked.

The true number in the population, \( n \), is then estimated as \( \hat{n} \) from the equation

\[ \frac{50}{\hat{n}} = \frac{i}{30} \Rightarrow \hat{n} = \frac{1500}{i}. \]

Now \( E[\hat{n}] = \sum_{i=0}^{30} \frac{1500}{i} \times p_i \)

i.e. it is possible that \( i = 0 \), in which case the estimate \( \hat{n} \) is infinite when \( i = 0 \),

\[ \Rightarrow E[\hat{n}] \] is also infinite and so cannot be equal to its true value

\[ \Rightarrow \] in this case the estimator \( \hat{n} = \frac{1500}{i} \) is biased.

**Unbiased estimates of population mean and variance**

Let \( X \) be a random variable drawn from a population with mean \( \mu \) and variance \( \sigma^2 \), then

\[ E[X] = \mu, \text{ and } \text{Var}[X] = \sigma^2. \]

A random sample, \( X_1, X_2, X_3, \ldots, X_n \), of size \( n \) is taken from the population.

The sample mean is \( \bar{X} = \frac{1}{n} (X_1 + X_2 + X_3 + \ldots + X_n) \).

\[ \Rightarrow E[X_i] = \mu, \text{ and } \text{Var}[X_i] = \sigma^2 \text{ for } i = 1, 2, 3, \ldots, n. \]

**Unbiased estimate of the mean**

\[ E[\bar{X}] = E \left[ \frac{1}{n} (X_1 + X_2 + X_3 + \ldots + X_n) \right] = \frac{1}{n} (E[X_1] + E[X_2] + E[X_3] + \ldots + E[X_n]) \]

\[ = \frac{1}{n} (\mu + \mu + \ldots + \mu) = \mu \]

\[ \Rightarrow E[\bar{X}] = \mu, \text{ the true value of the mean} \]

\[ \Rightarrow E[\bar{X}] \] is an unbiased estimate of the mean of the population.
Unbiased estimate of the variance of the population

Preliminary results

(i) \[ \text{Var}[X] = E[X^2] - (E[X])^2 = E[X^2] - \mu^2 \]
\[ \Rightarrow E[X^2] = \text{Var}[X] + \mu^2 = \sigma^2 + \mu^2 \quad \text{I} \]

(ii) \[ \text{Var}[\bar{X}] = E[\bar{X}^2] - (E[\bar{X}])^2 = E[\bar{X}^2] - \mu^2 \]
\[ \Rightarrow E[\bar{X}^2] = \text{Var}[\bar{X}] + \mu^2 \]
\[ = \text{Var}\left[ \frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n) \right] + \mu^2 \]
\[ = \frac{1}{n^2} \text{Var}[X_1 + X_2 + X_3 + \ldots + X_n] + \mu^2 \]
\[ = \frac{1}{n^2} (\text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3] + \cdots + \text{Var}[X_n]) + \mu^2 \]
\[ = \frac{1}{n^2} (\sigma^2 + \sigma^2 + \sigma^2 + \ldots + \sigma^2) + \mu^2 \]
\[ \Rightarrow E[\bar{X}^2] = \frac{1}{n} \sigma^2 + \mu^2 \quad \text{II} \]

Proof

The variance of \( X_1, X_2, X_3, \ldots, X_n \) is defined to be

\[ \text{Variance} = (\text{s.d.})^2 = \frac{1}{n} \sum X_i^2 - \bar{X}^2 \]
\[ \Rightarrow E[(\text{s.d.})^2] = E \left[ \frac{1}{n} \sum X_i^2 - \bar{X}^2 \right] \]
\[ = E \left[ \frac{1}{n} \sum X_i^2 \right] - E[\bar{X}^2] \]
\[ = \frac{1}{n} E[\sum X_i^2] - E[\bar{X}^2] \]
\[ = \frac{1}{n} \sum E[X_i^2] - E[\bar{X}^2] \]
\[ = \frac{1}{n} \sum (\sigma^2 + \mu^2) - E[\bar{X}^2] \]
\[ = \frac{1}{n} \left( n(\sigma^2 + \mu^2) - \left( \frac{1}{n} \sigma^2 + \mu^2 \right) \right) \]
\[ = \frac{n-1}{n} \sigma^2 \quad \text{since } E[X_i^2] = (\sigma^2 + \mu^2) \text{ from I} \]
\[ = \frac{n-1}{n} \sigma^2 \quad \text{since } E[\bar{X}^2] = \left( \frac{1}{n} \sigma^2 + \mu^2 \right) \text{ from II} \]
\[ \Rightarrow E[(\text{s.d.})^2] = (\sigma^2 + \mu^2) - \left( \frac{1}{n} \sigma^2 + \mu^2 \right) = \frac{n-1}{n} \sigma^2 \]

Thus \( E[(\text{s.d.})^2] \) is not equal to the true value, and so \( (\text{s.d.})^2 \) is a biased estimator of \( \sigma^2 \),
but multiplying both sides by \( \frac{n}{n-1} \), we can see that

\( \frac{n}{n-1} (\text{s.d.})^2 \) is an unbiased estimator of \( \sigma^2 \).
Bias

Example: A large bag contains counters: 60% have the number 0, and 40% have the number 1.

(a) Find the mean, \( \mu \), and variance, \( \sigma^2 \).

A simple random sample of size 3 is drawn.

(b) List all possible samples.

(c) Find the sampling distribution for the mean \( \bar{X} = \frac{X_1 + X_2 + X_3}{3} \).

(d) Use your answers to part (c) to find \( E[\bar{X}] \), and \( \text{Var}[\bar{X}] \).

(e) Find the sampling distribution for the mode \( M \).

(f) Use your answers to part (e) to find \( E[M] \), and \( \text{Var}[M] \).

Solution:

(a) \( \mu = \sum x_i p_i = 0 \times 0.6 + 1 \times 0.4 = 0.4 \)

\( \sigma^2 = \sum x_i^2 p_i - \mu^2 = (0^2 \times 0.6 + 1^2 \times 0.4) - 0.4^2 = 0.24 \)

(b) Possible samples are

(0, 0, 0)  (1, 0, 0)  (1, 1, 0)  (1, 1, 1)
(0, 1, 0)  (1, 0, 1)  (0, 0, 1)  (0, 1, 1)

(c) From (c) we can find the sampling distribution of the mean

| \( \bar{X} \) | 0 | \( \frac{1}{3} \) | \( \frac{2}{3} \) | 1 |
| \( p \)    | 0.6^3 | 3 \times 0.6^2 \times 0.4 | 3 \times 0.6 \times 0.4^2 | 0.4^3 |
|           | 0.216 | 0.432 | 0.288 | 0.064 |

(d) \( E[\bar{X}] = 0 \times 0.216 + \frac{1}{3} \times 0.432 + \frac{2}{3} \times 0.288 + 1 \times 0.064 = 0.4 \)

\( \text{Var}[\bar{X}] = (0^2 \times 0.216 + \left(\frac{1}{3}\right)^2 \times 0.432 + \left(\frac{2}{3}\right)^2 \times 0.288 + 1^2 \times 0.064) - 0.4^2 \)

\( \Rightarrow \text{Var}[\bar{X}] = 0.08 \)

(e) From (c) we can find the sampling distribution of the mode \( M \)

| \( M \) | 0 | 1 |
| \( p \) | 0.6^3 + 3 \times 0.6^2 \times 0.4 | 3 \times 0.6 \times 0.4^2 + 0.4^3 |
|        | 0.648 | 0.352 |

(f) \( E[M] = 0 \times 0.648 + 1 \times 0.352 = 0.352 \)

\( \text{Var}[M] = (0^2 \times 0.648 + 1^2 \times 0.352) - 0.352^2 \)

\( \Rightarrow \text{Var}[M] = 0.228096 \)

Thus the sample mean is an unbiased estimator of the mean of the population since \( E[\bar{X}] = 0.4 = \mu \), the true value.
but the sample mode is a *biased estimator* of the mode of the population

\[ E[M] = 0.352, \text{ but the true value of the mode of the population is } 0. \]

We say that the bias is \( E[M] - (\text{the true value}) = 0.352 - 0 = 0.352 \)

In general, if \( \hat{\lambda} \) is a biased estimator of the parameter \( \lambda \) then the bias is defined to be

\[
\text{bias} = E[\hat{\lambda}] - \lambda
\]

In the above example, the bias in estimating the mode from the sample is

\[
\text{bias} = E[M] - \text{true value} \\
= 0.352 - 0 = 0.352
\]
Probability generating functions

Probability functions are a neat idea, and are useful for finding the expected mean and variance for distributions which have a probability generating function which is easy to differentiate.

If $X$ is a random variable on the set $[1, n]$, then $G(t) = p_0 + p_1 t + p_2 t^2 + \ldots + p_n t^n$ is a probability generating function, p.g.f.,

if (i) $\sum_{1}^{n} p_i = 1$, and (ii) $p_i \geq 0 \forall i$

$P(X = i) = \text{the coefficient of } t^i$. The probability generating function can be thought of as a probability labeling function, where $t^i$ acts as a label for the probability that $X = i$.

Expected mean and variance for a p.g.f.

We know that $E[X] = \sum x_i p_i = 0 \times p_0 + 1 \times p_1 + 2 \times p_2 + \ldots + n \times p_n = \sum i p_i$

and that $\text{Var}[X] = E[X^2] - (E[X])^2$

$= 0 \times p_0 + 1^2 \times p_1 + 2^2 \times p_2 + \ldots + n^2 \times p_n - (\sum i p_i)^2$

Notice that $G'(t) = 0 \times p_0 + 1 \times p_1 + 2 \times p_2 t + 3 \times p_3 t^2 + \ldots + n \times p_n t^{n-1}$

$\Rightarrow$ $G'(1) = 0 \times p_0 + 1 \times p_1 + 2 \times p_2 + 3 \times p_3 + \ldots + n \times p_n$

$\Rightarrow$ Expected mean $= E[X] = G'(1)$

and $G''(t) = 0 \times (-1) \times p_0 + 1 \times 0 \times p_1 + 2 \times 1 \times p_2 + 3 \times 2 \times p_3 t + \ldots + n(n - 1) \times p_n t^{n-2}$

$\Rightarrow$ $G''(1) = \sum_{1}^{n} i(i-1)p_i = \sum_{1}^{n} i^2 p_i - \sum_{1}^{n} i p_i$

$\Rightarrow$ $\sum_{1}^{n} i^2 p_i = G''(1) + \sum_{1}^{n} i p_i$

$\Rightarrow$ $\text{Var}[X] = \sum_{1}^{n} i^2 p_i - \left(\sum_{1}^{n} i p_i\right)^2$

$\Rightarrow$ $\text{Var}[X] = G''(1) + G'(1) - (G'(1))^2$

Thus for a probability generating function $G(t) = p_0 + p_1 t + p_2 t^2 + \ldots + p_n t^n$, $E[X] = G'(1)$ and $\text{Var}[X] = G''(1) + G'(1) - (G'(1))^2$. 

S3 JUNE 2016 SDB 39
Mean and variance of a Binomial distribution

If \( X \sim B(n, p) \) then \( P(X = i) = \binom{n}{i} p^i q^{n-i} \), where \( p + q = 1 \).
These probabilities are the coefficients of \( t^i \) in the expansion of \( (q + pt)^n \)
\( \Rightarrow \) the p.g.f. for the binomial distribution \( B(n, p) \) is \( G(t) = (q + pt)^n \).
\( \Rightarrow \) \( G'(t) = np(q + pt)^{n-1} \), and \( G''(t) = n(n-1)p^2 (q + pt)^{n-2} \)
\( \Rightarrow \) \( \mu = E[X] = G'(1) = np \) since \( p + q = 1 \)
\( \Rightarrow \) \( \sigma^2 = \text{Var}[X] = G''(1) + G'(1) - (G'(1))^2 \)
\( = n(n-1)p^2 + np - (np)^2 = n^2p^2 - np - n^2p^2 \)
\( \Rightarrow \) \( \sigma^2 = \text{Var}[X] = np(1 - p) \) or \( npq \).

Mean and variance of a Poisson distribution

If \( X \sim P_0(\lambda) \) then, in a given interval, \( P(X = i) = \frac{\lambda^i e^{-\lambda}}{i!} \), where \( \lambda \) is the mean number of occurrences in an interval of the same length, \( i = 0, 1, 2, 3, \ldots \)
\( \Rightarrow \) \( G(t) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} t^i = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i t^i}{i!} = e^{-\lambda} e^{\lambda t} \)
\( \Rightarrow \) \( G'(t) = \lambda e^{-\lambda} e^{\lambda t} \) and \( G''(t) = \lambda^2 e^{-\lambda} e^{\lambda t} \)
\( \Rightarrow \) \( \mu = E[X] = G'(1) = \lambda e^{-\lambda} e^\lambda = \lambda \)
\( \Rightarrow \) \( \mu = E[X] = \lambda \)
\( \Rightarrow \) \( \sigma^2 = \text{Var}[X] = G''(1) + G'(1) - (G'(1))^2 \)
\( = \lambda^2 + \lambda - \lambda^2 \) since \( e^{-\lambda} e^\lambda = 1 \)
\( \Rightarrow \) \( \sigma^2 = \text{Var}[X] = \lambda \)
Index

$\chi^2$ test
  binomial dist, 24
  continuous uniform dist., 24
  degrees of freedom, 23
  discrete uniform dist., 23
  general points, 23
  normal dist, 26
  Poisson dist, 24

Bias, 9, 37
  Biased estimators, 8
    bias, 38
    examples, 35
  Binomial distribution p.g.f.
    expected mean and variance, 40
  Census, 4
  Central limit theorem, 15
  Combinations of random variables
    expected mean of $X \pm Y$, 3
    expected variance of $X \pm Y$, 3
    independent normal variables, 3
    $\text{Var}[X+Y]$, 34
    $\mathbb{E}[X + Y]$, 33
  Confidence intervals, 16
  Contingency tables
    $\chi^2$ test, 27
    degrees of freedom, 27
  Data
    primary data, 7
    secondary data, 7
  Estimators
    population mean, 11
    population variance, 11
  Lottery sampling, 5
  PMCC
    comparison with Spearman, 32
    when to use, 30
  Poisson distribution p.g.f.
    expected mean and variance, 40
  Probability generating functions, 39
    expected mean and variance, 39
  Random number tables, 5
  Ranks
    equal ranks, 29
  Sample variance
    estimator of population variance, 11
  Sampling, 4
    quota sampling, 7
    sample means, 14
    simple random sampling, 5
    stratified sampling, 6
    systematic sampling, 5
    with and without replacement, 6
  Significance test
    zero correlation, 31
  Significance test – variance of population known
    difference between means, 19
    mean of normal distribution, 18
  Significance test – variance of population NOT known
    difference between means, 22
    mean of normal distribution, 21
  Spearman
    comparison with PMCC, 32
  Spearman’s rank correlation coefficient, 29
    when to use, 30

Standard error, 15
  Unbiased estimators, 8
    examples, 34
    of the population mean, 35
    of the population variance, 36
  variance of $(X + Y)$, 33