Statistics 3

Revision Notes

June 2016

Statistics 3

1 Combinations of random variables

Expected mean and variance for $X \pm Y$

Reminder

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

$\mathrm{E}[aX] = a\mathrm{E}[X]$ and $\mathrm{Var}[aX] = a^2\,\mathrm{Var}[X]$

$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$ and $\mathrm{E}[X - Y] = \mathrm{E}[X] - \mathrm{E}[Y]$

and for two **independent** random variables

$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$ and $\mathrm{Var}[X - Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Combining independent normal random variables Y

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

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

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

$X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

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

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

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

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

Solution: $\mathrm{E}[X_1] = 21$, $\mathrm{Var}[X_1] = 12$ and $\mathrm{E}[X_2] = 9$, $\mathrm{Var}[X_2] = 6$

$\Rightarrow$ $\mathrm{E}[2X_2] = 2\mathrm{E}[X_2] = 2 \times 9 = 18$

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

$\Rightarrow$ $\mathrm{E}[X_1 - 2X_2] = \mathrm{E}[X_1] - \mathrm{E}[2X_2] = 21 - 18 = 3$

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

$\Rightarrow$ the expected mean and standard deviation of $X_1 - 2X_2$ are 3 and $\sqrt{36} = 6$. 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 0.06kg.

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. $X \sim N(0{\cdot}1, 0{\cdot}02^2)$ and $Y \sim N(1, 0{\cdot}06^2)$.

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 $E[W] = 12\,E[X] + 12\,E[Y] = 12 \times 0{\cdot}1 + 12 \times 1 = 13{\cdot}2$

and, assuming independence,

$Var[W] = 12\,Var[X] + 12\,Var[Y] = 12 \times 0{\cdot}02^2 + 12 \times 0{\cdot}06^2 = 0{\cdot}048.$

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

$$1 - \Phi\left(\frac{13{\cdot}5 - 13{\cdot}2}{\sqrt{0{\cdot}048}}\right) = 1 - \Phi(1{\cdot}37) = 0{\cdot}0853 \text{ to 3 S.F. Answer.}$$

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^{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^{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:

	Boys	Girls
Upper sixth	30	40
Lower sixth	30	30
Fifth form	70	60
Fourth form	60	70
Third form	50	60

Solution: As 50 is $^1/_{10}$ of the total population, $^1/_{10}$ of each stratum should be selected in the sample. Thus the sample would comprise

	Boys	Girls
Upper sixth	3	4
Lower sixth	3	3
Fifth form	7	6
Fourth form	6	7
Third form	5	6

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, $\mu = \mathrm{E}[X]$.

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, μ. Show that Q_2 is a biased estimator of μ, and find the bias.

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

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

Solution:

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

(*b*)

Sample	Probability		median	mean
(2,2,2)	$\left(\frac{2}{5}\right)^3$	$= \frac{8}{125}$	2	2
(2,2,5), (2,5,2), (5,2,2)	$3 \times \left(\frac{2}{5}\right)^2 \times \frac{3}{5}$	$= \frac{36}{125}$	2	3
(2,5,5), (5,2,5), (5,5,2)	$3 \times \frac{2}{5} \times \left(\frac{3}{5}\right)^2$	$= \frac{54}{125}$	5	4
(5,5,5)	$\left(\frac{3}{5}\right)^3$	$= \frac{27}{125}$	5	5

Sampling distribution

(i) Median

$Q_2 = x_i$	p_i	$x_i p_i$
2	$\frac{8+36}{125}$	$\frac{88}{125}$
5	$\frac{54+27}{125}$	$\frac{405}{125}$
		$\frac{493}{125}$

$\Rightarrow \quad \mathrm{E}[Q_2] = \frac{493}{125} = 3{\cdot}944$

(ii) Mean

$\overline{X} = x_i$	p_i	$x_i p_i$
2	$\frac{8}{125}$	$\frac{16}{125}$
3	$\frac{36}{125}$	$\frac{108}{125}$
4	$\frac{54}{125}$	$\frac{216}{125}$
5	$\frac{27}{125}$	$\frac{135}{125}$
		$\frac{475}{125}$

$\mathrm{E}[\overline{X}] = \frac{475}{125} = 3{\cdot}8$

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] = the true value of the parameter 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 μ is 3·8, and in part (b) we have shown that E[Q_2] = 3·944

$\Rightarrow$ Q_2 is a biased estimator of μ,
and the bias is |E[Q_2] – true value of μ| = |3·944 – 3·8| = 0·144

(ii) The mean $\overline{X}$ is used as the estimate of the mean.
From part (a) we know that the true value of the mean μ is 3·8, and in part (b) we have shown that E[$\overline{X}$] = 3·8

$\Rightarrow$ E[$\overline{X}$] = the true value of the mean

$\Rightarrow$ $\overline{X}$ is an unbiased estimator of μ,

(d) If we now use kQ_2 as an **unbiased** estimator of the mean value of all the coins.

E[kQ_2] = k E[Q_2] = $\frac{493}{125}k$

But the true mean $\mu = \frac{475}{125}$

If kQ_2 is an unbiased estimator of μ, E[kQ_2] = true value of μ

$\Rightarrow$ $\frac{493}{125}k = \frac{475}{125}$ $\Rightarrow$ $k = \frac{475}{493}$.

Example: A sample of size 3 is drawn from a binomial distribution B(10, 0·25) and the mean, $\overline{X}$, is calculated.

The probability of success, p, is estimated by $\hat{p} = \frac{1}{10}\overline{X}$. Show that $\hat{p}$ is an unbiased estimator of p.

Solution: E[X] = np = 10 × 0·25 = 2·5

For a sample $\{X_1, X_2, X_3\}$, $\overline{X} = \frac{1}{3}(X_1 + X_2 + X_3)$

$\Rightarrow$ $\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\frac{1}{3}(X_1 + X_2 + X_3)\right] = \frac{1}{3}(\mathrm{E}[X_1] + \mathrm{E}[X_2] + \mathrm{E}[X_3])$

$\Rightarrow$ $\mathrm{E}\left[\overline{X}\right] = \frac{1}{3} \times 3 \times 2 \cdot 5 = 2 \cdot 5$ since $\mathrm{E}[X_i] = \mathrm{E}[X] = 2 \cdot 5$, for i = 1, 2, 3

$\Rightarrow$ $\mathrm{E}[\hat{p}] = \mathrm{E}\left[\frac{1}{10}\overline{X}\right] = \frac{1}{10}\mathrm{E}\left[\overline{X}\right] = \frac{1}{10} \times 2 \cdot 5 = 0 \cdot 25$, which is the true value of p

$\Rightarrow$ $\hat{p} = \frac{1}{10}\overline{X}$ is an unbiased estimator of p.

Example: A sample of size 4 is drawn from a continuous uniform distribution, U[3, β]. The mean of the sample, $\overline{X}$, is calculated.

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

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

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

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

$\Rightarrow \mathrm{E}\left[\overline{X}\right] = \frac{1}{4} \times 4 \times \mathrm{E}[X] = \mathrm{E}[X]$ since $\mathrm{E}[X_i] = \mathrm{E}[X]$, for i = 1, 2, 3, 4

$\Rightarrow \mathrm{E}\left[\overline{X}\right] = \frac{1}{2}(3 + \beta)$

$\Rightarrow \mathrm{E}[\hat{\beta}] = \mathrm{E}[2\overline{X} - 3] = 2\mathrm{E}\left[\overline{X}\right] - 3 = 2 \times \frac{1}{2}(3 + \beta) - 3 = \beta,$

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

$\Rightarrow \hat{\beta} = 2\overline{X} - 3$ is an unbiased estimator of β.

There is more on biased and unbiased estimators in the Appendix.

Unbiased estimators of μ and σ^2

Estimating μ and σ^2 from a sample

We usually do not know the mean, μ, and the variance, σ^2, of a population.

To estimate these values we take a sample $\{X_1, X_2, X_3, \ldots, X_n\}$ 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 $\mathrm{E}\left[\overline{X}\right]$ = the true value of μ

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

It can be shown that $\mathrm{E}\left[(sd)_x{}^2\right] = \frac{n-1}{n}\sigma^2$ ..... **I**

$\Rightarrow$ $(sd)_x{}^2$ is a *biased* estimator of σ^2.

I $\Rightarrow$ $\mathrm{E}\left[\frac{n}{n-1}(sd)_x{}^2\right] = \frac{n}{n-1} \times \frac{n-1}{n}\sigma^2 = \sigma^2$, 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, σ^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 σ^2.

Also, the term *Sample Variance* is used to denote the unbiased estimate of σ^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 σ^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:

X	$X-\bar{X}$	$(X-\bar{X})^2$
56	1·8	3·24
53	-1·2	1·44
57	2·8	7·84
51	-3·2	10·24
54	-0·2	0·04
271		22·8

$\Rightarrow \quad \bar{X} = \frac{1}{n}\sum X = \frac{271}{5} = 54{\cdot}2$

$\Rightarrow \quad (sd)_x{}^2 = \frac{1}{n}\sum(X-\bar{X})^2 = \frac{22{\cdot}8}{5} = 4\cdot 56$

$\Rightarrow \quad \hat{\sigma}^2 = \frac{n}{n-1}(sd)_x{}^2 = \frac{5}{4} \times 4\cdot 56 = 5\cdot 7$

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, $\bar{X}$, $= \frac{14228}{14} = 1016{\cdot}286\ldots$.

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

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

$\sum X^2 = 14460232$

$\Rightarrow \quad (sd)_x{}^2 = \frac{1446032}{14} - \left(\frac{14228}{14}\right)^2 = 37{\cdot}06122\ldots$

$\Rightarrow \quad S^2 = s_x{}^2 = \frac{n}{n-1}(sd)_x{}^2 = \frac{14}{14-1} \times 37\cdot 06122\ldots = 39\cdot 91209\ldots$

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.

ΣX = 3797 grams, and ΣX^2 = 973692.

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

Solution: $\mu = \bar{X} = \frac{3797}{15} = 253 \cdot 133 \ldots = 252 \cdot 1$ grams.

$$(sd)_x{}^2 = \frac{1}{n}\sum X^2 - \bar{X}^2 = \frac{973692}{15} - \left(\frac{3797}{15}\right)^2 = 836{\cdot}3156\ldots$$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{15}{14} \times 836 \cdot 3156 \ldots = 896 \cdot 0524 \ldots = 896 \cdot 1 \text{ grams}^2.$$

Answer Unbiased estimators are $\hat{\mu}$ =252·1 *g*, and $\hat{\sigma}^2$ = 896·1 g^2.

Example: The lengths of 10 rods are measured, and the sample has mean, $\bar{X}$ = 26·7 *cm* and variance s^2 = 76·9 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.

$$\text{For } n = 10, \ \bar{X}_{10} = \frac{1}{10}\sum_{i=1}^{10} X_i = 26 \cdot 7 \quad \Rightarrow \quad \sum_{i=1}^{10} X_i = 267$$

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

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

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

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

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

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

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

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

For 11 rods, sample mean is 27 *cm*, and sample variance is 70·2 cm^2.

4 Confidence intervals and significance tests

Sampling distribution of the mean

X is a random variable draw from a population with mean μ and standard deviation σ.

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

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

and the expected mean of the population of sample means is

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

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

$$= \mu.$$

Also the expected variance of the population of sample means is

$$\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]$$ assuming that all the X_i are independent

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

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

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 μ and variance σ^2,

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

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 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. $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

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 grams^2.

(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, $X \sim N(23{\cdot}4, 6^2)$

$\Rightarrow$ $\mathrm{P}(X > 25) = \Phi\left(\frac{25-23{\cdot}4}{6}\right) = \Phi(0\cdot27) = 0\cdot6064$

(ii) $\mu = 23{\cdot}4$, $\sigma^2 = 36$

The sample mean $\overline{X} \sim N\left(23{\cdot}4, \left(\frac{6}{\sqrt{50}}\right)^2\right)$

$\Rightarrow$ standard error is $\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0{\cdot}848528137$

$\bar{X} \sim N(23\cdot4,\ 0\cdot8485\ldots^2)$

$\Rightarrow$ $P(\bar{X} > 25) = 1 - \Phi\left(\frac{25-23{\cdot}4}{0{\cdot}8485\ldots}\right)$

$= 1 - \Phi(1{\cdot}89) = 1 - 0{\cdot}9706$ (from Normal tables)

$= 0{\cdot}0294.$

(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:

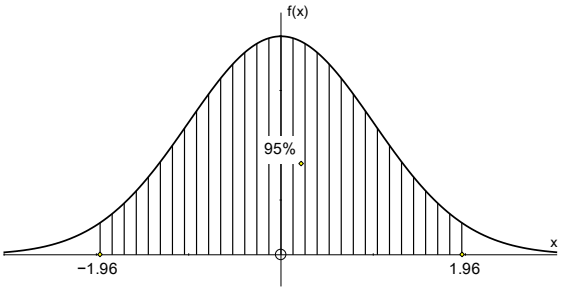
First assume that the machine is still producing packets with the same variance, 25.

Suppose that the mean weight of all packets of biscuits is 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

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



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

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

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

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

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

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

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

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 [508·8, 511·2] 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.

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 = $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{120}} = 0\cdot456435...$

.

We have an observed mean of 83·2

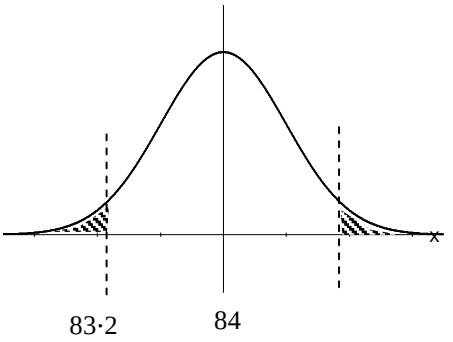
For a two-tailed test at 5%, we take 2·5% at each end

$P(\bar{X} < 83\cdot2) = \Phi\left(\frac{83\cdot2-84}{0\cdot456435...}\right) = \Phi(-1\cdot7527)$

$= (1 - \Phi(1\cdot75)) = 0\cdot0401$

$= 4\cdot01\% > 2\cdot5\%$

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

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

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

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

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

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 $\overline{X}_A = 123{\cdot}1$ and $\overline{X}_B = 124{\cdot}4$ 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 μ_A and μ_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,

$\overline{X}_B - \overline{X}_A = 124{\cdot}4 - 123{\cdot}1 = 1{\cdot}3$,

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 $\overline{X}_B - \overline{X}_A$.

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

expected variance $\text{Var}[\overline{X}_B] = \frac{{\sigma_B}^2}{n_B} = \frac{3}{16}$

and secondly, for the population of sample means for machine *A*

expected variance Var[$\overline{X}_A$] = $\frac{\sigma_A^2}{n_A} = \frac{4}{25}$

and so for the population of differences of sample means

expected mean = E[$\overline{X}_B - \overline{X}_A$] = $\mu_A - \mu_B = 0$ (from H_0)

and Var[$\overline{X}_B - \overline{X}_A$] = Var[$\overline{X}_B$] + Var[$\overline{X}_A$]

= $\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}$ = $^3/_{16} + ^4/_{25}$ = 0·3475.

The observed difference, the test statistic, is 124·4 – 123·1 = 1·3

and the standard error is $\sqrt{0 \cdot 3475)}$

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

$\Rightarrow$ $P(\text{difference} > 1 \cdot 3) = 1 - \Phi\left(\frac{1 \cdot 3 - 0}{\sqrt{0 \cdot 3475)}}\right) = 1 - \Phi(2 \cdot 2053)$

$= 1 - \Phi(2 \cdot 20)$ = 1– 0·9861 = 0·0134

= 1·34% < 5%

$\Rightarrow$ significant at 5% level so reject H_0 and conclude that there is evidence that machine *B* is producing bars of chocolate with a heavier mean weight than machine *A*.

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

$Z = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$ is in your formula booklet.

Significance testing – variance of population NOT known, large sample

When the variance of the population, σ^2, is not known and when the sample is **large,** we assume that the variance of the sample (meaning the unbiased estimate of σ^2), S^2, is the variance of the population, σ^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 $0{\cdot}33^2$ cm^2. Advise the production manager, using a 5% level of significance.

Important assumption

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

Solution:

H_0: $\mu = 25{\cdot}4$.

H_1: $\mu \neq 25{\cdot}4$ two-tail test, 2·5% in each tail

We assume that population variance σ^2 = the sample variance $S^2 = 0{\cdot}33^2$

$\Rightarrow \sigma = 0{\cdot}33$

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

expected mean = 25·4 from hypothesis

and standard error = $\frac{\sigma}{\sqrt{n}} = \frac{0 \cdot 33}{\sqrt{80}} = 0{\cdot}036895121$.

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

$$\Phi\left(\frac{25{\cdot}31 - 25{\cdot}4}{0{\cdot}036895121}\right) = \Phi(-2 \cdot 4393) = 1 - \Phi(2 \cdot 44) = 0{\cdot}0073 < 2{\cdot}5\%$$

$\Rightarrow$ reject H_0 and conclude that there is evidence that 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 75^2 N^2, whereas 65 cables produced by machine *B* have a mean breaking strain of 1757 N and a variance of 63^2 N^2. Is there any evidence, at the 10% level, to suggest that machine *B* is producing stronger cables than machine *A*?

Solution:

Let μ_A and μ_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, σ_A^2 = the sample variance, $S_A^2 = 75^2$

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

For Machine *B*

We assume that the population variance, σ_B^2 = the sample variance, $S_B^2 = 63^2$

$\Rightarrow$ variance of sample means $\text{Var}[\overline{X}_B] = \dfrac{\sigma_B^2}{n} = \dfrac{63^2}{65} = 61 \cdot 0615....$

For differences in sample means $\overline{X}_B - \overline{X}_A$

Expected mean = 0 from hypothesis

Expected variance is $\text{Var}[\overline{X}_B - \overline{X}_A] = \text{Var}[\overline{X}_B] + \text{Var}[\overline{X}_A]$

$= 140 \cdot 625 + 61.0615\ldots = 201 \cdot 6865\ldots$

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

We have an observed difference in means,

test statistic, $\overline{X}_B - \overline{X}_A = 1757 - 1728 = 29$

and for a 1-tail test that *B* is stronger

we need the area to the right of 29 mean is treated as continuous, so do not use 28.5

$= 1 - \Phi\left(\dfrac{29 - 0}{14 \cdot 2016...}\right) = 1 - \Phi(2 \cdot 04) = 0 \cdot 0207 < 10\%$

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, χ^2 test

General points

- The χ^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, ν, 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.

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 $\chi^2_5(2{\cdot}5\%) = 12{\cdot}832 > 3{\cdot}04$, 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, 0·3)*, 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 = $^{215}/_{69}$ = 3·12
and variance = $^{915}/_{69} - (^{215}/_{69})^2$ = 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 $\lambda = 3{\cdot}12$, 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 $\chi^2_4(5\%) = 9.488$.

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 $(N, m, s) \Rightarrow \nu = n - 3 = 4$.

$\chi^2_4(5\%) = 9.488$ 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	$\frac{609\times 544}{1335} = 248.2$	$\frac{609\times 453}{1335} = 206.6$	$\frac{609\times 161}{1335} = 73.4$	$\frac{609\times 177}{1335} = 80.7$	608.9
African	$\frac{334\times 544}{1335} = 136.1$	$\frac{334\times 453}{1335} = 113.3$	$\frac{334\times 161}{1335} = 40.3$	$\frac{334\times 177}{1335} = 44.3$	334
Chinese	$\frac{392\times 544}{1335} = 159.7$	$\frac{392\times 453}{1335} = 133.0$	$\frac{392\times 161}{1335} = 47.3$	$\frac{392\times 177}{1335} = 52.0$	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.

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.

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**

0·572 < **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

7 Appendix

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, $\sum_{i=1}^{n} r_{ij} = q_j$

$E[X + Y] = E[X] + E[Y]$

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

$$= \sum_i x_i \sum_j r_{ij} + \sum_j y_j \sum_i r_{ij}$$

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

$$\Rightarrow \quad \mathbf{E}[\boldsymbol{X+Y}] = \mathbf{E}[\boldsymbol{X}] + \mathbf{E}[\boldsymbol{Y}]$$

Var[X + Y] = Var[X] + Var[Y]

In this case we take X and Y to be **independent**,

$\Rightarrow \quad r_{ij} = \mathrm{P}(x_i \textbf{ and } y_j) = \mathrm{P}(x_i) \times \mathrm{P}(y_j) = p_i \times q_j.$

Also notice that $\sum p_i = \sum q_j = 1$

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

$$= \sum_i \sum_j (x_i + y_j)^2 r_{ij} - (\mathrm{E}[X] + \mathrm{E}[Y])^2$$

$$= \sum_i \sum_j {x_i}^2 p_i q_j + 2\sum_i \sum_j x_i y_j p_i q_j + \sum_i \sum_j {y_j}^2 p_i q_j - ((\mathrm{E}[X])^2 + 2\mathrm{E}[X]\mathrm{E}[Y] + (\mathrm{E}[Y])^2)$$

$$= \sum_i {x_i}^2 p_i \sum_j q_j + 2\sum_i x_i p_i \sum_j y_j q_j + \sum_j {y_j}^2 q_j \sum_i p_i - (\mathrm{E}[X])^2 - 2\mathrm{E}[X]\mathrm{E}[Y] - (\mathrm{E}[Y])^2$$

$$= \sum_i {x_i}^2 p_i + 2\mathrm{E}[X]\mathrm{E}[Y] + \sum_j {y_j}^2 q_j - (\mathrm{E}[X])^2 - 2\mathrm{E}[X]\mathrm{E}[Y] - (\mathrm{E}[Y])^2$$

$$= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 + \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2$$

$\Rightarrow$ Var[X + Y] = Var[X] + Var[Y], **if X and Y are independent.**

Unbiased & biased estimators

Unbiased estimators

An estimator $\hat{\lambda}$ for a parameter λ 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

estimate $\hat{g} = \frac{i}{20} \times 468.$

We now find the expected value of this estimate

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

$\Leftrightarrow$ 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 λ 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 $E[\hat{n}] = \sum_{0}^{30} \frac{1500}{i} \times p_i$

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 μ and variance σ^2, then

$E[X] = \mu$, 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 $\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n)$.

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

Unbiased estimate of the mean

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

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

$\Rightarrow$ $E[\bar{X}] = \mu$, 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) $\text{Var}[X] = \text{E}[X^2] - (\text{E}[X])^2 = \text{E}[X^2] - \mu^2$

$\Rightarrow \quad \text{E}[X^2] = \text{Var}[X] + \mu^2 = \sigma^2 + \mu^2$ **I**

(ii) $\text{Var}[\bar{X}] = \text{E}[\bar{X}^2] - (\text{E}[\bar{X}])^2 = \text{E}[\bar{X}^2] - \mu^2$

$\Rightarrow \quad \text{E}[\bar{X}^2] = \text{Var}[\bar{X}] + \mu^2$

$= \text{Var}\left[\frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n)\right] + \mu^2$

$= \frac{1}{n^2}\text{Var}[X_1 + X_2 + X_3 + \ldots + X_n] + \mu^2$

$= \frac{1}{n^2}(\text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3] + \cdots + \text{Var}[X_n]) + \mu^2$

$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \sigma^2 + \ldots + \sigma^2) + \mu^2$

$\Rightarrow \quad \text{E}[\bar{X}^2] = \frac{1}{n}\sigma^2 + \mu^2$ **II**

Proof

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

$\text{Variance} = (\text{s.d.})^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2$

$\Rightarrow \quad \text{E}[(\text{s.d.})^2] = \text{E}\left[\frac{1}{n}\sum X_i^2 - \bar{X}^2\right]$

$= \text{E}\left[\frac{1}{n}\sum X_i^2\right] - \text{E}[\bar{X}^2]$

$= \frac{1}{n}\text{E}\left[\sum X_i^2\right] - \text{E}[\bar{X}^2]$

$= \frac{1}{n}\sum \text{E}\left[X_i^2\right] - \text{E}[\bar{X}^2]$

$= \frac{1}{n}\sum(\sigma^2 + \mu^2) - \text{E}[\bar{X}^2]$ since $\text{E}[X_i^2] = (\sigma^2 + \mu^2)$ from **I**

$= \frac{1}{n}\left(n(\sigma^2 + \mu^2) - \left(\frac{1}{n}\sigma^2 + \mu^2\right)\right)$ since $\text{E}[\bar{X}^2] = \left(\frac{1}{n}\sigma^2 + \mu^2\right)$ from **II**

$\Rightarrow \quad \text{E}[(\text{s.d.})^2] = (\sigma^2 + \mu^2) - \left(\frac{1}{n}\sigma^2 + \mu^2\right) = \frac{n-1}{n}\sigma^2$

Thus $\text{E}[(\text{s.d.})^2]$ is **not** equal to the true value, and so $(\text{s.d.})^2$ is a *biased* estimator of σ^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 σ^2.

Bias

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

(*a*) Find the mean, μ, and variance, σ^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 $\mathrm{E}[\bar{X}]$, and $\mathrm{Var}[\bar{X}]$.

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

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

Solution:

(*a*) $\mu = \sum x_i p_i = 0 \times 0{\cdot}6 + 1 \times 0{\cdot}4 = 0{\cdot}4$

$\sigma^2 = \sum x_i^2 p_i - \mu^2 = (0^2 \times 0{\cdot}6 + 1^2 \times 0{\cdot}4) - 0{\cdot}4^2 = 0{\cdot}24$

(*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	$0{\cdot}6^3$	$3 \times 0{\cdot}6^2 \times 0{\cdot}4$	$3 \times 0{\cdot}6 \times 0{\cdot}4^2$	$0{\cdot}4^3$
	0·216	0·432	0·288	0·064

(*d*) $\mathrm{E}[\bar{X}] = 0 \times 0{\cdot}216 + \frac{1}{3} \times 0{\cdot}432 + \frac{2}{3} \times 0{\cdot}288 + 1 \times 0{\cdot}064 = 0{\cdot}4$

$\mathrm{Var}[\bar{X}] = \left(0^2 \times 0{\cdot}216 + \left(\frac{1}{3}\right)^2 \times 0{\cdot}432 + \left(\frac{2}{3}\right)^2 \times 0{\cdot}288 + 1^2 \times 0{\cdot}064\right) - 0{\cdot}4^2$

$\Rightarrow \quad \mathrm{Var}[\bar{X}] = 0{\cdot}08$

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

M	0	1
p	$0{\cdot}6^3 + 3 \times 0{\cdot}6^2 \times 0{\cdot}4$	$3 \times 0{\cdot}6 \times 0{\cdot}4^2 + 0{\cdot}4^3$
	0·648	0·352

(*f*) $\mathrm{E}[M] = 0 \times 0{\cdot}648 + 1 \times 0{\cdot}352 = 0{\cdot}352$

$\mathrm{Var}[M] = \left(0^2 \times 0{\cdot}648 + 1^2 \times 0{\cdot}352\right) - 0{\cdot}352^2$

$\Rightarrow \quad \mathrm{Var}[M] = 0{\cdot}228096$

Thus the sample mean is an *unbiased estimator* of the mean of the population

since $\mathrm{E}[\bar{X}] = 0{\cdot}4 = \mu$, the true value

but the sample mode is a *biased estimator* of the mode of the population

$\mathrm{E}[M] = 0{\cdot}352$, but the true value of the mode of the population is 0.

We say that the bias is $\mathrm{E}[M]$ – (the true value) = $0{\cdot}352 - 0 = 0{\cdot}352$

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

bias = $\mathrm{E}[\hat{\lambda}] - \lambda$

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

bias = $\mathrm{E}[M]$ – true value

= $0{\cdot}352 - 0 = 0{\cdot}352$

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 $G(t) = p_0 + p_1 t + p_2 t^2 + \ ... \ + p_n t^n$ is a probability generating function, p.g.f.,

if (i) $\sum_1^n p_i = 1$, and (ii) $p_i \geq 0 \ \forall \ i$

$P(X = i)$ = the coefficient of t^i. The probability *generating* function can be thought of as a probability *labelling* 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 $E[X] = \sum x_i p_i = 0 \times p_0 + 1 \times p_1 + 2 \times p_2 + \ ... \ + n \times p_n = \sum i p_i$

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

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

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

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

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

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

$\Rightarrow \quad G''(1) = \sum_1^n i(i-1)p_i = \sum_1^n i^2 p_i - \sum_1^n i p_i$

$\Rightarrow \quad \sum_1^n i^2 p_i = G''(1) + \sum_1^n i p_i$

$\Rightarrow \quad \text{Var}[X] = \sum_1^n i^2 p_i - \left(\sum_1^n i p_i\right)^2$

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

Thus for a probability generating function $G(t) = p_0 + p_1 t + p_2 t^2 + \ ... \ + p_n t^n$,

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

Mean and variance of a Binomial distribution

If $X \sim \mathrm{B}(n, p)$ then $\mathrm{P}(X = i) = {}^{n}\mathrm{C}_i\, p^i q^{n-i}$, 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 $\mathrm{B}(n, p)$ is $\mathrm{G}(t) = (q + pt)^n$.

$\Rightarrow$ $\mathrm{G}'(t) = np(q + pt)^{n-1}$, and $\mathrm{G}''(t) = n(n-1)p^2 (q + pt)^{n-2}$

$\Rightarrow$ $\mu = \mathrm{E}[X] = \mathrm{G}'(1) = np$ since $p + q = 1$

and $\sigma^2 = \mathrm{Var}[X] = \mathrm{G}''(1) + \mathrm{G}'(1) - (\mathrm{G}'(1))^2$

$= n(n-1)p^2 + np - (np)^2 = n^2p^2 - np^2 + np - n^2p^2$

$\Rightarrow$ $\sigma^2 = \mathrm{Var}[X] = np(1-p)$ or npq.

Mean and variance of a Poisson distribution

If $X \sim \mathrm{Po}(\lambda)$ then, in a given interval, $\mathrm{P}(X = i) = \frac{\lambda^i e^{-\lambda}}{i!}$, where λ is the mean number of occurrences in an interval of the same length, $i = 0, 1, 2, 3, \ldots$

$$\Longrightarrow \quad \mathrm{G}(t) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} t^i = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} t^i = e^{-\lambda} e^{\lambda t}$$

$\Rightarrow$ $\mathrm{G}'(t) = \lambda e^{-\lambda} e^{\lambda t}$ and $\mathrm{G}''(t) = \lambda^2 e^{-\lambda} e^{\lambda t}$

$\Rightarrow$ $\mu = \mathrm{E}[X] = \mathrm{G}'(1) = \lambda\, e^{-\lambda} e^{\lambda} = \lambda$

$\Rightarrow$ $\mu = \mathrm{E}[X] = \lambda$

and $\sigma^2 = \mathrm{Var}[X] = \mathrm{G}''(1) + \mathrm{G}'(1) - (\mathrm{G}'(1))^2$

$= \lambda^2 + \lambda - \lambda^2$ since $e^{-\lambda} e^{\lambda} = 1$

$\Rightarrow$ $\sigma^2 = \mathrm{Var}[X] = \lambda$

Index