Probability and Statistics

The theory from probability can be used to help analyse the data obtained from statistical studies. This forms the last part of Strand 1.

The key idea here is that of a probability distribution and the corresponding probability histogram or curve. The most important example is the standard normal probability distribution. The values of the probabilities, i.e. the areas under this curve, are given on pages 36 and 37 of the *Formulae and Tables*. You should practise calculating areas in right tails, left tails and between two given values of z.

By converting to standard units, we can also calculate probabilities for any normal variable. Associated ideas are the empirical rule and using standard units (z scores) to determine relative standing.

This theory from probability can be used to determine the reliability of statistics derived from sample data. The margin of error of a sample proportion can be used to give a confidence interval for a population proportion.

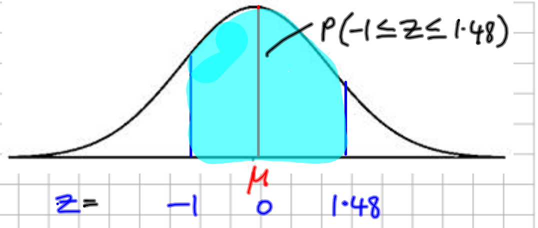
Another application is the use of the margin of error in making a decision about whether an obtained sample proportion is consistent with an assumption about a population proportion. This is called hypothesis testing.

The final application is to use the theory associated with the Central Limit Theorem to deal with the distribution of sample means and hence determine the probability that the mean of an individual sample differs from the mean of the population by a specified amount.

None of the calculations in this section is in any way challenging. Rather the difficulty lies in understanding the concepts, and being able to decide what to do in any given situation.

1. Standard normal tables

e.g. if Z is a standard normal variable, calculate $P(-1 \le Z \le 1\cdot48)$



① Look up $P(z \le 1\cdot48)$

② Look up $P(z \le 1)$

(negative values not in tables)

③ $P(z \le -1) = 1 - P(z \le 1)$

④ $P(-1 \le z \le 1\cdot48)$
$= P(z \le 1\cdot48) - P(z \le -1)$

$P(z \le 1\cdot48) = 0\cdot9306$

$P(z \le 1) = 0\cdot8413$

$P(z \le -1) = 1 - 0\cdot8413 = 0\cdot1587$

$P(-1 \le z \le 1\cdot48) = 0\cdot9306 - 0\cdot1587$
$= 0\cdot7719$

2. Other normal distributions

e.g. A certain brand of chocolate bar has weight which is normally distributed with mean 150 g and standard deviation 5 g.

(i) What is the probability that a bar of chocolate chosen at random has a weight less than 140 g?

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

$\mu = 150\text{ g}$
$\sigma = 5\text{ g}$
$x = 140\text{ g}$

z value?

$$z = \frac{x-\mu}{\sigma} = \frac{140-150}{5} = -2$$

Related positive value?
$z = -2$ is related to $z = 2$

$P(z \le 2) = 0{\cdot}9772$

$P(z \le -2) = 1 - P(z \le 2)$
$= 1 - 0{\cdot}9772$
$= 0{\cdot}0228$

Answer: $P(\text{Weight} \le 140\text{g}) = 2{\cdot}28\%$

(ii) In a delivery of 2000 of these bars, how many would we expect to weigh less than 140 g?

Expected no. =
Probability × trials

Expected no.?

$= (2000)(0{\cdot}0228) = 45{\cdot}6$

≈ 46 bars

3. Empirical rule

e.g. X is a continuous random variable which is normally distributed with mean $5{\cdot}6$ and standard deviation $0{\cdot}8$.
Calculate $P(4{\cdot}8 \le X \le 7{\cdot}2)$, i.e. the probability that X lies between $4{\cdot}8$ and $7{\cdot}2$, using the empirical rule.

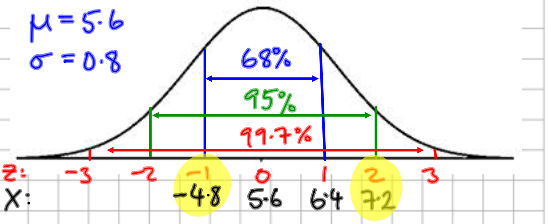
Empirical Rule:

$P(-\sigma \le x \le \sigma) = 68\%$
$P(-2\sigma \le x \le 2\sigma) = 95\%$
$P(-3\sigma \le x \le 3\sigma) = 99{\cdot}7\%$

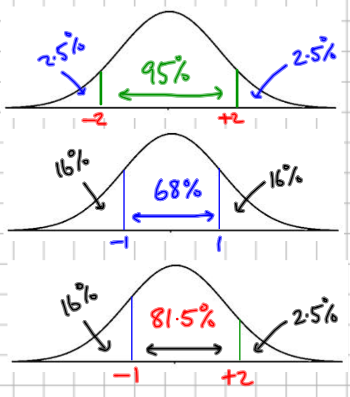
$X = 7{\cdot}2 = 5{\cdot}6 + 2(0{\cdot}8)$
$\Leftrightarrow z = 2$
$X = -4{\cdot}8 = 56 - 0{\cdot}8$
$\Leftrightarrow z = -1$

see diagram

$\mu = 5{\cdot}6$
$\sigma = 0{\cdot}8$



$P(4{\cdot}8 \le x \le 7{\cdot}2) = P(-1 \le z \le 2)$



$(16 + 2{\cdot}5)\% = 18{\cdot}5\%$
$(100 - 18{\cdot}5)\% = 81{\cdot}5\%$

ANSWER: $P(4{\cdot}8 \le x \le 7{\cdot}2) = 81{\cdot}5\%$

4. Relative standing

e.g. Paul plays an online game against many other players. On Monday, he scores 21050 and is told that for his group of players the mean and standard deviation are 19550 and 780 respectively. On Tuesday, he plays again against the same group of players. This time he scores 20440, and is informed that the mean and standard deviation are 19400 and 580 respectively. On which day did Paul perform better relative to the other players in his group? Give a reason.

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

*percentile rankings are not required here you may just compare z-scores.

Answer :

On which day did Paul get better z-score?

Monday :

$X = 21050$
$\mu = 19550$
$\sigma = 780$

$$z = \frac{21050-19550}{780} \approx 1.92$$

* Percentile Ranking $= P(z \leq 1.92) = 97.26\%$

Tuesday :

$X = 20440$
$\mu = 19400$
$\sigma = 580$

$$z = \frac{20440-19400}{580} \approx 1.79$$

*Percentile Ranking $= P(z \leq 1.79) = 96.33\%$

On Monday Paul had higher z-score ($\Rightarrow$ percentile ranking).

5. Margin of error

e.g. In a by-election, a random sample of 400 voters suggests that 38% will vote for candidate A.

(i) What is the margin of error?

$$E = \frac{1}{\sqrt{n}}$$

(ii) Give the 95% confidence interval for the true proportion of voters who intend to vote for candidate A.

95% confidence interval

$$\hat{p} - E \leq p \leq \hat{p} + E$$

$\hat{p}$ = sample proportion
$\hat{p}$ = 38%

p = actual population proportion

Margin of error

$$E = \frac{1}{\sqrt{400}} = \frac{1}{20} = 5\%$$

95% confidence interval

$$38\% - 5\% \leq p \leq 38\% + 5\%$$

$$33\% \leq p \leq 43\%$$

6. Hypothesis testing

e.g. The manufacturers of a new treatment claim that it cures hiccups within 20 seconds for 90% of people. An independent body decides to check this claim. The new treatment is tested on a random sample of 4000 hiccups sufferers, and cures the hiccups of 3540 of these.

(i) State the hypothesis that should be made by the independent body.

(ii) On what basis should the independent body reject or not reject the hypothesis?

(iii) Determine if the independent body should have concern about the claim made by the manufacturers.

95% Confidence interval

$\hat{p} - E \le p \le \hat{p} + E$

(i) null Hypothesis

H_0 = hiccups cured within 20 seconds for 90% of people

ie.. $p \le 90\%$

alternative Hypothesis

H_1 = hiccups not cured within 20 seconds for 90% of people.

ie.. $p \geq 90\%$

(ii) Accept or reject depending on pass/fail the hypothesis test

(iii) Hypothesis test.

Margin of error (E)

$$E = \frac{1}{\sqrt{4000}} \approx 1{\cdot}58\%$$

Populalation Sample $(\hat{p})$

$$\hat{p} = \frac{3540}{4000} = 88{\cdot}5\%$$

95% Confidence interval

$$(88{\cdot}5 - 1{\cdot}58)\% \le p \le (88{\cdot}5 + 1{\cdot}58)\%$$

$$86{\cdot}92\% \le p \le 90{\cdot}08\%$$

ie.. $p \le 90{\cdot}08\%$

Result

H_0 is accepted. The claim is valid

7. Central Limit Theorem

e.g. A trial consists of tossing a fair coin 4 times. Let X be the number of heads obtained.

(i) Complete the probability distribution table below.

$x \in X$	0	1	2	3	4
$P(x)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

(ii) Represent this distribution on a histogram.

Normal distribution :

$P(H) = \frac{1}{2} = p$, $P(T) = \frac{1}{2} = q$

Bernoulli distribution

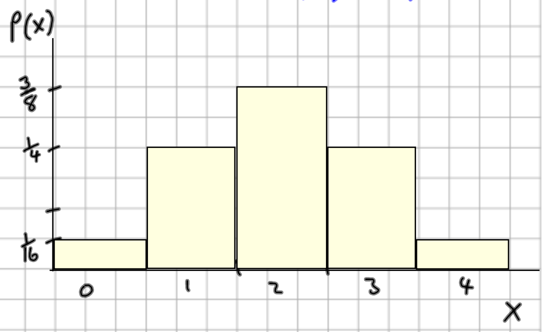
$$P(X=r) = \binom{n}{r} p^r q^{n-r}$$

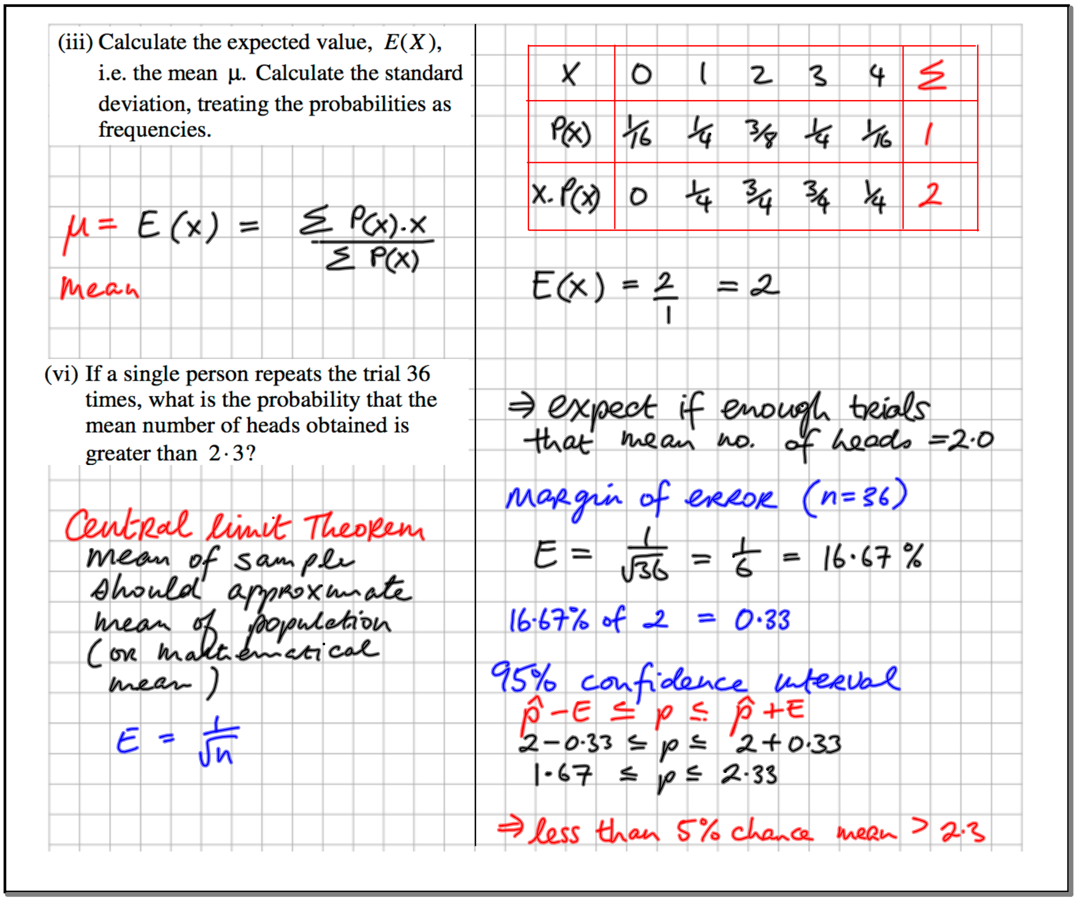
$$P(1\text{ Head}) = \binom{4}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

$$P(2\text{ Heads}) = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \frac{3}{8}$$

$$P(3\text{ Heads}) = \binom{4}{3}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

$$P(4\text{ Heads}) = \binom{4}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 = \frac{1}{16}$$





(iii) Calculate the expected value, $E(X)$, i.e. the mean μ. Calculate the standard deviation, treating the probabilities as frequencies.

$$\mu = E(x) = \frac{\sum P(x).x}{\sum P(x)}$$

Mean

X	0	1	2	3	4	$\sum$
$P(x)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$	1
$x.P(x)$	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{4}$	2

$$E(x) = \frac{2}{1} = 2$$

(vi) If a single person repeats the trial 36 times, what is the probability that the mean number of heads obtained is greater than $2 \cdot 3$?

Central limit Theorem

mean of sample should approximate mean of population (or mathematical mean)

$$E = \frac{1}{\sqrt{n}}$$

⇒ expect if enough trials that mean no. of heads $= 2.0$

Margin of error $(n=36)$

$$E = \frac{1}{\sqrt{36}} = \frac{1}{6} = 16.67\%$$

16.67% of $2 = 0.33$

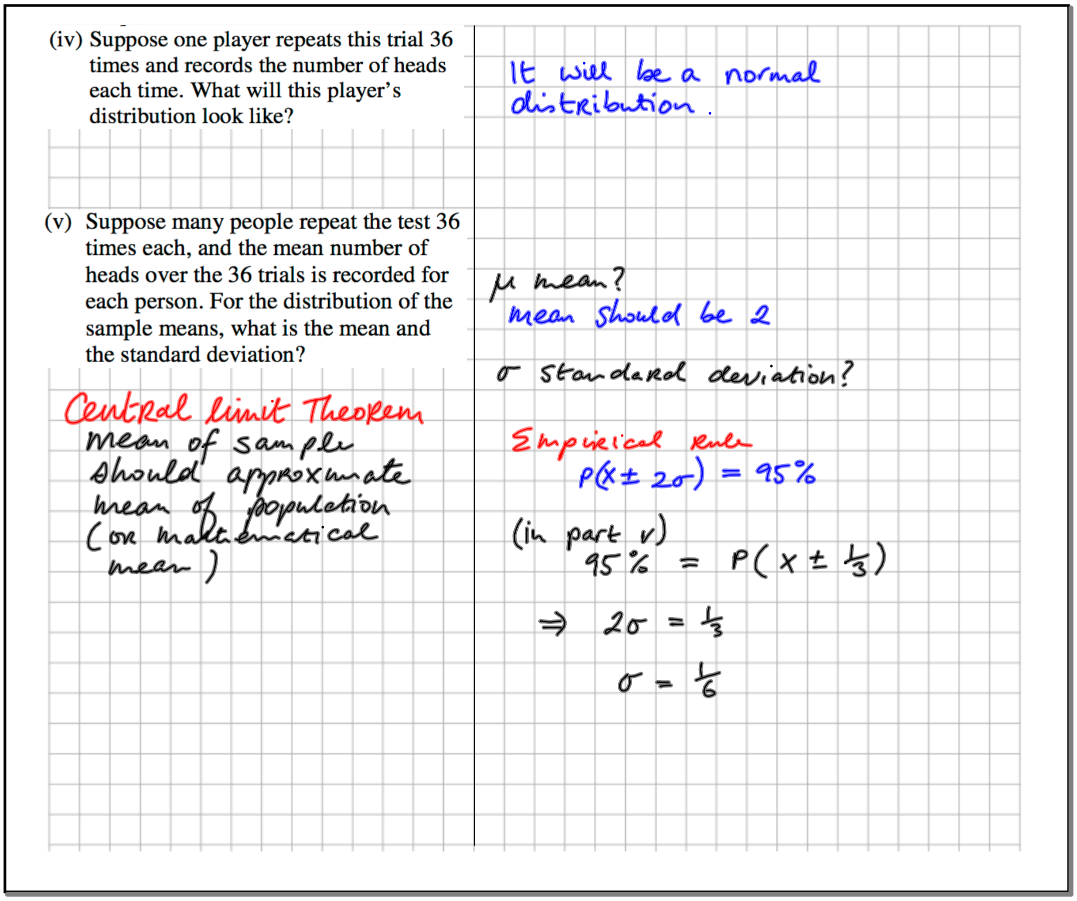
95% confidence interval

$$\hat{p} - E \le p \le \hat{p} + E$$

$$2 - 0.33 \le p \le 2 + 0.33$$

$$1.67 \le p \le 2.33$$

⇒ less than 5% chance mean > 2.3



(iv) Suppose one player repeats this trial 36 times and records the number of heads each time. What will this player's distribution look like?

It will be a normal distribution.

(v) Suppose many people repeat the test 36 times each, and the mean number of heads over the 36 trials is recorded for each person. For the distribution of the sample means, what is the mean and the standard deviation?

Central limit Theorem

mean of sample should approximate mean of population (or mathematical mean)

μ mean?

mean should be 2

σ standard deviation?

Empirical rule

$$P(x \pm 2\sigma) = 95\%$$

(in part v)

$$95\% = P(x \pm \tfrac{1}{3})$$

$$\Rightarrow 2\sigma = \tfrac{1}{3}$$

$$\sigma = \tfrac{1}{6}$$