

- 1 a)
- no two killer sounds occur at the same time
 - the killer sounds occurrences are all independent of each other
 - the killer sounds occur at a constant mean rate.

b) ensure that the vertical axis scale was the same for all three charts.

c) mean rate, $\bar{x} = \frac{\sum x_i \cdot O_i}{\sum O_i}$

$$= \frac{1 \times 3 + 2 \times 6 + 3 \times 6 + 4 \times 5 + 5 \times 5 + 6 \times 2 + 7 \times 1}{3 + 6 + 6 + 5 + 5 + 2 + 1}$$
$$= \frac{97}{28}$$
$$\approx \underline{3.4643} \text{ (4 dp)}$$

d) E_i when $x_i = 5+$ is $28 - (0.85 + 2.96 + 5.18 + 6.04 + 5.29)$

$$= 28 - 20.32$$
$$= \underline{7.68}$$

or $28 \times P(X \geq 5)$ where $X \sim Po(3.5)$

$$= 28 \times [1 - P(X \leq 4)]$$
$$= 28 \times (1 - 0.7254) \quad \text{by } \text{poisscdf}(3.5, 0, 4)$$
$$= 28 \times 0.2746$$
$$= \underline{7.6888}$$

e) $df = \text{categories} - \text{constraints}$

$$= 6 - 1 - 1$$
$$= \underline{4}$$

← as we estimated the mean rate parameter from the observed data.

- f) i) • no expected frequency should be less than 1, and $0.85 < 1$
 • also 2 out of the 6 expected frequencies are less than 5, which equates to 33% of the categories, and this number should not be more than 20%.

ii) table 2 should be

x_i	O_i	E_i
1-	3	3.81
2	6	5.18
3	6	6.04
4	5	5.29
5+	8	7.69 ← (from part (d))

$$\begin{aligned} \chi^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(3 - 3.81)^2}{3.81} + \dots + \frac{(8 - 7.69)^2}{7.69} \\ &= \underline{0.3307} \text{ (4 dp)} \end{aligned}$$

- iii) H_0 : number of filler sounds in 30 seconds $\sim P_0(3.5)$
 H_1 : number of filler sounds in 30 seconds is not $\sim P_0(3.5)$

$\alpha = 10\%$

$\chi^2_{3, 0.90} = 6.251$ as $df = 5 - 2 = 3$

as $0.3307 < 6.251$, we do not reject H_0

we do not have evidence that suggests that the number of filler sounds in 30 seconds is not distributed as $P_0(3.5)$

- g) common assumption is that the distributions of the populations of filler sounds in 30 seconds would be normally distributed.

2 a) number of athletes = $df + 2$
 $= 293 + 2$
 $= \underline{\underline{295}}$

b) as $p\text{-value} < 0.0001 < 0.05$, we reject H_0
 we have evidence to suggest that the true correlation is not equal to zero
 this means we have evidence that the sprint times are correlated to the hurdle times.

c) we want $E(\varepsilon_i) = 0$ for all x_i
 this appears to be satisfied as the residuals are randomly scattered around zero.
 we want $V(\varepsilon_i) = \sigma^2$ for all x_i
 this appears to be satisfied as the residuals have no change to their spread, regardless of the fitted value.

d) sprint = $a + 0.9665$ hurdles
 $24.1366 = a + 0.9665 \times 13.09$
 $a = 24.1366 - 12.6514$
 $a = \underline{\underline{11.4851}}$

Prediction interval is symmetrical around fitted value

\Rightarrow upper bound = $24.1366 + (24.1366 - 22.5224)$
 $= 24.1366 + 1.6142$
 $= \underline{\underline{25.7508}}$

e) the residuals need to be assumed to be normally distributed
 $\varepsilon_i \sim N(0, \sigma^2)$ for some σ^2 .

f) if she were to run her sprint 100 times, we would expect that her time for 99 of those sprints to lie within the interval of 22.5224 seconds and 25.7508 seconds.

g) the model in Output 2 predicts the sprint time from a hurdle time, not vice versa.

In order to predict hurdle time from sprint time, we would need to recalculate the regression line with x -data and y -data swapped around.

h) confidence interval predicts the mean sprint time from multiple competitions.

KJT's time of 23.08 seconds is a single, individual occurrence, not the average of several sprints.

(but 23.08 was in the prediction interval, from Output 2).

1. a)

x	1	2	3	4	5	6	
$P(X=x)$	$\frac{1}{60}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{19}{60}$	
$xP(X=x)$	$\frac{1}{60}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	$\frac{114}{60}$	$\rightarrow \text{sum} = \frac{17}{4}$
$x^2P(X=x)$	$\frac{1}{60}$	$\frac{4}{6}$	$\frac{9}{6}$	$\frac{16}{6}$	$\frac{25}{6}$	$\frac{684}{60}$	$\rightarrow \text{sum} = \frac{245}{12}$

$$E(X) = \sum xP(X=x) = \underline{\underline{\frac{17}{4}}}$$

$$E(X^2) = \sum x^2P(X=x) = \frac{245}{12}$$

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{245}{12} - \left[\frac{17}{4}\right]^2 \\ &= \underline{\underline{\frac{113}{48}}} \end{aligned}$$

b) $Y \sim U(8)$

$$\therefore E(Y) = \frac{8+1}{2} = \frac{9}{2}$$

$$V(Y) = \frac{8^2-1}{12} = \frac{63}{12}$$

$$V(X-Y) = V(X) + V(Y)$$

$$= \frac{113}{48} + \frac{63}{12}$$

$$= \frac{365}{48}$$

$$\begin{aligned} \therefore SD(X-Y) &= \sqrt{\frac{365}{48}} \\ &= \underline{\underline{2.7576}} \text{ (4 dp)} \end{aligned}$$

we require X and Y to be independent.

2.

a) Upper Fence = $Q_3 + 1.5 \times IQR$ Lower Fence = $Q_1 - 1.5 \times IQR$
 $= 22 + 1.5 \times 15$ $= 7 - 1.5 \times 15$
 $= 44.5$ $= -15.5$

as $1 > -15.5$, the lowest value is not an outlier

as $46 > 44.5$, the highest value is a possible outlier

b)

O_i	< 20 km	> 20 km	E_i	< 20	> 20
Retail	36	26	R	42.78	19.22
Food	33	5	F	26.22	11.78

H_0 : there is no association between type of outlet visited and distance travelled

H_1 : there is an association between type of outlet visited and distance travelled

Assume H_0 to be true

$\alpha = 5\%$ one tailed test

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 9.12165$$

$$p\text{-value} = 0.002526$$

as $0.002526 < 0.05$ we reject H_0

we have evidence to suggest that there is an association between the type of outlet visited, and the distance travelled.

(comparing the observed and expected frequencies, we conjecture that customers travel further for Retail, and less far for food shops)

or $\chi^2_{1,0.95} = 3.841$

as $9.12 > 3.841$, we reject H_0 .

3. $X = \text{no. of stoppages per month}$

$$X \sim P_0(14)$$

$$P(X > 20) = ?$$

let Y be a normal approximation to X

$$\Rightarrow Y \sim N(14, 14) \quad (\text{as } \lambda > 10, \text{ this approximation is valid})$$

$$\text{so } P(X > 20) \approx P(Y > 20.5) \quad \text{using continuity correction.}$$

$$= P\left(Z > \frac{20.5 - 14}{\sqrt{14}}\right)$$

$$= P(Z > 1.7372)$$

$$= 0.041176$$

by normcdf(1.7372, 9E99)

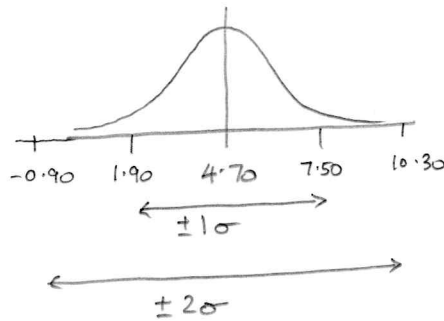
$$\approx \underline{\underline{0.0412}} \quad (4\text{dp})$$

4.

a) mean = \$4.70

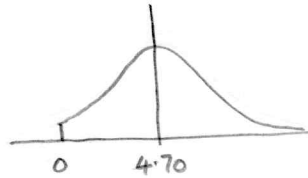
st. dev = \$2.80

if it were normally distributed, then it would be symmetrical about the mean



the value that is 2 standard deviations below the mean is a negative amount of money which is not a valid value.

Hence the tip-per-customer's distribution would be truncated at zero, possibly looking like this:



b) i) let \bar{X} be mean tip-per-customer

$$\bar{X} \approx N(4.70, \frac{2.80^2}{50})$$

$$P(\bar{X} > 5.50) = P(Z > \frac{5.50 - 4.70}{\sqrt{\frac{2.80^2}{50}}})$$

$$= P(Z > 2.02031)$$

$$= \underline{\underline{0.0217}} \quad (4 \text{ dp})$$

from normCdf(2.02031, 9E99)

- ii) • The distribution of tips-per-customer was not normally distributed
- The sample size of 50 was greater than 20.

5.

x_i	40.1	40.3	39.5	40.6	40.8	39.9	40.0	41.3	38.3
η	40	40	40	40	40	40	40	40	40
$x_i - \eta$	0.1	0.3	-0.5	0.6	0.8	-0.1	0	1.3	-1.7
$ x_i - \eta $	0.1	0.3	0.5	0.6	0.8	0.1	0	1.3	1.7
rank	1 ↓ 1.5	3	4	5	6	2 ↓ 1.5	/	7	8

$$H_0: \eta = 40$$

$$H_1: \eta \neq 40$$

Assume H_0 is true

$\alpha = 5\%$ two tailed test

$$W_+ = 1.5 + 3 + 5 + 6 + 7 = 22.5 \quad 21.5$$

$$W_- = 4 + 1.5 + 8 = 13.5$$

$$W = \min(W_+, W_-)$$

$$W = 13.5$$

critical value for $n = 8$ is 3

as $13.5 > 3$, we are not in critical region

we do not have evidence to reject H_0

so we do not have evidence to suggest that the population median is different to 40.

$$\left. \begin{array}{l} \text{check} \\ 22.5 + 13.5 = 36 \\ \frac{1}{2} \times 8 \times 9 = 36 \end{array} \right\} \checkmark \textcircled{U}$$

6.

$$\bar{x}_1 = 3.9868$$

$$\bar{x}_2 = 4.2510$$

$$s_1 = 1.3396$$

$$s_2 = 1.2342$$

$$n_1 = 990$$

$$n_2 = 735$$

let X_1 = tidal range for 2000-2009

X_2 = tidal range for 2010-2018

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

we are given that $\sigma_1^2 = s_1^2$ and $\sigma_2^2 = s_2^2$

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 < \mu_2$$

Assume H_0 to be true and that $s_1 \approx \sigma_1$, $s_2 \approx \sigma_2$

$$\alpha = 0.5\% = 0.005 \quad \text{2-tailed test}$$

$$X_1 \sim N(\mu_1, \sigma_1^2) \quad X_2 \sim N(\mu_2, \sigma_2^2)$$

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

$$\text{test statistic, } Z = \frac{\bar{x}_1 - \bar{x}_2 - (0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{3.9868 - 4.2510}{\sqrt{\frac{1.3396^2}{990} + \frac{1.2342^2}{735}}} = -4.2387$$

$$p\text{-value} = P(Z < -4.2387)$$

$$= 0.000011$$

from norm Cdf(-999, -4.2387)

$$< 0.005$$

so we reject H_0

we have evidence to suggest that the mean monthly extreme

tidal range has increased between 2000-2009 and 2010-2018.

7. a) $X = \text{no. of fours when rolling fair die, 10 times.}$

so we have $n = 10$

$$P(\text{get a 4}) = \frac{1}{6} \quad \therefore \underline{\underline{X \sim B(10, \frac{1}{6})}}$$

we have 10 independent trials with a fixed probability of success of obtaining a four.

b) $L = \text{leftover string}$

L can be any value between 0 and 8.0cm, with equal likelihood.

Hence L is a continuous uniform distribution

$$\underline{\underline{L \sim U(0, 8.0)}}$$

8 a). number all members of the population with a unique integer between 1 and N inclusive

- use a random number generator to give n distinct integers between 1 and N inclusive
- select members of the population whose number corresponds to the randomly generated values, and they are the random sample of size n from a population of size N .

b) centre A : 14 out of 29 passed $\Rightarrow \hat{p}_A = \frac{14}{29}$

centre B : 21 out of 30 passed $\Rightarrow \hat{p}_B = \frac{21}{30}$

we shall use a z-test for a difference in population proportions.

let X = no. passed at centre A

let Y = no. passed at centre B

$$X \sim B(29, p_A)$$

$$Y \sim B(30, p_B)$$

$$H_0: p_A = p_B$$

$$H_1: p_A < p_B$$

Assume H_0 to be true. $\alpha = 5\%$. one-tailed test

test requires approximation of both binomial distributions to normal distributions

so $np > 5$ and $nq > 5$ need to be satisfied for both centre's distributions

$$\text{Centre A: } 29 \times p_A \approx 29 \times \hat{p}_A = 29 \times \frac{14}{29} = 14 > 5$$

$$29 \times q_A \approx 29 \times \hat{q}_A = 29 \times \frac{15}{29} = 15 > 5 \quad \checkmark$$

$$\text{Centre B: } 30 \times p_B \approx 30 \times \hat{p}_B = 30 \times \frac{21}{30} = 21 > 5$$

$$30 \times q_B \approx 30 \times \hat{q}_B = 30 \times \frac{9}{30} = 9 > 5 \quad \checkmark$$

$$\text{pooled proportion, } p = \frac{14 + 21}{29 + 30} = \frac{35}{59}$$

$$\text{test statistic, } z = \frac{\frac{14}{29} - \frac{21}{30}}{\sqrt{\frac{35}{59} \times \frac{24}{59} \times \left(\frac{1}{29} + \frac{1}{30}\right)}} = -1.6982$$

$$p\text{-value} = P(Z < -1.6982)$$

$$= 0.044735$$

from normCDF(-9E99, -1.6982)

$$< 0.05$$

so we reject H_0

we have evidence to suggest that the pass rate at Centre A is lower than the pass rate at Centre B.

c) samples were taken in different months, so we are not comparing like with like.

9.

$$X \sim N(0.1, 0.25^2)$$

X in inches

$$1 \text{ inch} = 2.54 \text{ cm}$$

a) $Y = 2.54X$

$$\begin{aligned} E(Y) &= E(2.54X) \\ &= 2.54 E(X) \\ &= 2.54 \times 0.1 \\ &= \underline{\underline{0.254}} \end{aligned}$$

$$\begin{aligned} V(Y) &= V(2.54X) \\ &= 2.54^2 V(X) \\ &= 2.54^2 \times 0.25^2 \\ &= \underline{\underline{0.403225}} \end{aligned}$$

b) within 1 cm $\Rightarrow -1 < Y < 1$ where $Y \sim N(0.254, 0.403225)$

so $P(-1 < Y < 1)$

$$= P\left(\frac{-1 - 0.254}{\sqrt{0.403225}} < Z < \frac{1 - 0.254}{\sqrt{0.403225}}\right)$$

$$= P(-1.9748 < Z < 1.1748)$$

$$= 0.855818$$

from normCdf(-1.9748, 1.1748)

let W = no. of planks that are within 1cm

$$W \sim B(80, 0.855818)$$

$$\therefore E(W) = 80 \times 0.855818$$

$$= 68.4654$$

$$\approx 68.5$$

Hence we expect 68 or 69 planks to be within 1cm of the required length.

10. X = mass of powdered milk.

$$\sum x_i = 27.9 \quad \sum x_i^2 = 130.2 \quad n = 6. \quad X \sim N(\mu, \sigma^2)$$

a) assume that each spoonful's mass is independent of all other spoonfuls' masses.

$$H_0: \mu = 4.5$$

$$H_1: \mu \neq 4.5$$

$\alpha = 10\%$, two tailed test

Assume H_0 to be true.

$$\text{so } X \sim N(4.5, \sigma^2)$$

$$\Rightarrow \bar{X} \sim N(4.5, \frac{\sigma^2}{6})$$

$$\frac{\bar{X} - 4.5}{\sqrt{\frac{\sigma^2}{6}}} \sim N(0, 1)$$

$$\Rightarrow \frac{\bar{X} - 4.5}{\sqrt{\frac{s^2}{6}}} \sim t_5 \quad \text{as we estimate } \sigma \text{ by } s$$

$$\text{now sample mean, } \bar{x} = \frac{1}{n} \sum x = \frac{1}{6} \times 27.9 = 4.65$$

$$\begin{aligned} \text{sample standard deviation, } s_{n-1} &= \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} \\ &= \sqrt{\frac{130.2 - \frac{(27.9)^2}{6}}{5}} \\ &= 0.304959 \end{aligned}$$

$$\text{test statistic, } t = \frac{4.65 - 4.5}{\sqrt{\frac{0.304959^2}{6}}} = 1.20483$$

$$p\text{-value} = 2 \times P(t_5 > 1.20483)$$

$$= 2 \times 0.14109$$

$$= 0.28218$$

$$> 0.10$$

$$\text{from } t\text{CDF}(1.20483, 999, 5)$$

so we do not reject H_0

we do not have evidence to suggest that the mean mass per spoonful is different from 4.5 grams.

b) they would perform a z-test, as they would not be estimating σ from the sample.

11. a) i) $p = 0.92$

$$p - 2\sigma = 0.821$$

$$0.92 - 2\sigma = 0.821$$

$$2\sigma = 0.92 - 0.821$$

$$2\sigma = 0.099$$

$$\sigma = 0.0495$$

$$\therefore \text{Lower } 1\sigma \text{ limit} = 0.92 - 0.0495$$

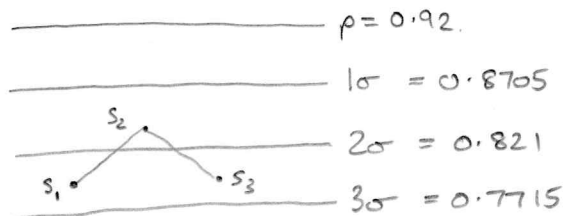
$$= \underline{\underline{0.8705}}$$

$$\text{Lower } 3\sigma \text{ limit} = 0.92 - 3 \times 0.0495$$

$$= \underline{\underline{0.7715}}$$

ii) having a high proportion of success is not a cause for concern.

b)



so we have 2 out of 3 samples beyond the 2 sigma limit

ii) process is considered out of statistical control.

12. a) $n=15$
 $\bar{x} = 139.5 \text{ cm}$
 $s = 0.7 \text{ cm}$

let $X = \text{bounce height}$

- assume that X is normally distributed
- assume that bounce heights of the sample are all independent of each other.

so $X \sim N(\mu, \sigma^2)$

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1^2)$

$\frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{14}$ as we estimate σ from s .

so 98% confidence interval is $\bar{x} \pm t_{14, 0.99} \times \sqrt{\frac{s^2}{n}}$

$\Rightarrow 139.5 \pm 2.62 \times \sqrt{\frac{0.7^2}{15}}$

$\Rightarrow 139.5 \pm 0.474349$

$\Rightarrow (139.026, 139.974)$

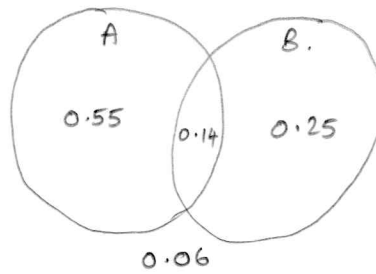
- b) as 141 cm is outwith, and above the 98% confidence interval, the balls do not bounce high enough, and thus they would not be recommended to be used.

13.

$$P(A) = 0.69$$

$$P(A \cap \bar{B}) = 0.55$$

$$P(\bar{A} \cap \bar{B}) = 0.06$$



$$\begin{aligned} \text{a) } P(\bar{A} \cap B) &= 1 - 0.55 - 0.14 - 0.06 \\ &= \underline{\underline{0.25}} \end{aligned}$$

$$\begin{aligned} \text{b) } P((A \cap B) \cup (\bar{A} \cap \bar{B})) &= 0.14 + 0.06 \\ &= \underline{\underline{0.20}} \end{aligned}$$

$$\begin{aligned} \text{c) } P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{0.14}{0.14 + 0.25} \\ &= \underline{\underline{0.358974}} \end{aligned}$$