2019 AH Stats.

1 a) X = no. breakdowns per day $\quad X \sim Po(3)$

$$P(X < 3) = P(X \leq 2)$$

$$= e^{-3} + \frac{3e^{-3}}{1!} + \frac{3^2 e^{-3}}{2!}$$

$$= e^{-3}\left(1 + 3 + \frac{9}{2}\right)$$

$$= 0.42319...$$

$$\approx 0.4232 \text{ (4dp)}$$

or by poissCdf(3, 0, 2)

b) let Y = no. breakdowns per week $\quad Y \sim Po(21)$

$$P(Y = 25) = \frac{21^{25} e^{-21}}{25!}$$

$$= 0.055546...$$

$$\approx 0.0555 \text{ (4dp)}$$

or by poissPdf(21, 25)

2. a) $P(\text{left handed} \mid \text{colour blind}) = \frac{10}{10+80}$

$= \frac{1}{9}$

b) $P(\text{left handed}) = \frac{130+10}{1000} = \frac{7}{50}$

if $P(\text{left handed} \mid \text{colour blind}) = P(\text{left handed})$ then the two events are independent

$LHS = \frac{1}{9}$ $\quad RHS = \frac{7}{50}$

$\neq \frac{1}{9}$

So LHS ≠ RHS, so colour blind is not independent of left handed ness.

c)

f_o	M	F	
No	450	506	956
Yes.	40	4	44
	490	510	1000

f_e	M	F	
No	468.44	487.56	← $\frac{956 \times 510}{1000}$
Yes.	21.56	22.44	

H_0: colour blindness and gender are not associated

H_1: they are associated

Assume H_0 to be true.

$\alpha = 5\%$ one-tail test

So $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

$= 32.3478$

$P(\chi_1^2 > 32.3478) = 1.28 \times 10^{-8}$

< 0.05

or critical value at 5% on χ_1^2 is 3.841

So $32.3478 > 3.841$

So we are in critical region.

So we have evidence to reject H_0

We conclude that colour blindness and gender are associated.

We conjecture that you are more likely to be colour blind if you are male

3. a) the population is all children's books

we have a sample of all 100 books in a local bookshop, that has <u>not</u> been randomly selected

we therefore have used <u>convenience sampling</u>.

a disadvantage of this technique is that the results of the analysis may not reflect the population of all children's books, as the single local bookshop may not have a representative sample of all children's books. A consequence of this is that their estimated proportion of books with a female central character may not be an accurate estimate of the true proportion in the population.

b) The books would first be placed in an order, and in a large book shop that would typically be by surname of author

We would count the total number of books in the shop, call it N.

Now work out 4% of N to get the number of books we need, call it B.

We would out how many times B goes into N, call it T.

We therefore need to select every T^{th} book from the shelves.

We also generate a starting number from 1 to (T-1) using random number generator.

We then start at that random number and take every T^{th} book after it until we have a total of B books.

eg. say N = 3000, so B = 4% of 3000 = 120

now $\frac{3000}{120} = 25$, which is T

generate random number between 1 and 25, say 13.

∴ select 13^{th} book, 38^{th} book, 63^{rd} book, 88^{th} book, etc

until 120 books selected

2019 AH Stats.

Q4. Claim : first scoring more likely to win match.

Two teams over one season ∴ sample of size 2.

H_0: team no more likely to win $(p=0.5)$

H_1: team more likely to win $(p>0.5)$

Combining two teams' performances, out of 64 games they collectively won $22+15=37$ when they scored first

So we are performing a proportion test to establish if $\frac{37}{64}$ provides evidence against H_0 being true.

let X = no. games won, when team scored first

$X \sim B(64, p)$

under H_0, we have $p=0.5$, so $X \sim B(64, 0.5)$

p-value $= P(X \geq 37)$

$= 0.130218$ from binomcdf(64, 0.5, 37, 64)

> 0.05

Hence we do not have sufficient evidence to reject H_0, and so proportion of wins $= 0.5$

So we conclude that the fan's claim that scoring first increases your chances of winning is not supported by the evidence.

OR

we perform proportion test via normal approximation....

if $X \sim B(64, p)$

let $Y \sim N(64p, 64pq)$ where Y is approx to X

so $\frac{Y}{64} \sim N\left(p, \frac{pq}{64}\right)$

we have $\hat{p} = \frac{37}{64}$, so test statistic, $z = \dfrac{\frac{37}{64} - p}{\sqrt{\frac{\frac{37}{64} \times \frac{27}{64}}{64}}}$ where $p=0.5$ under H_0

$= 1.26$

now $1.26 < 1.64$, so we are not in the 5% critical region

so we conclude that this is no evidence for the fan's claim.

5. let $X \sim U(7, 22)$ (cts uniform)

$n = 25$

a) let $\bar{X}$ = sample mean of sample of size 25

now $E(X) = \frac{7+22}{2} = \frac{29}{2}$

$V(X) = \frac{(22-7)^2}{12} = \frac{15^2}{12} = \frac{225}{12}$.

so, by CLT, $\bar{X} \approx N\left(\frac{29}{2}, \frac{225/12}{25}\right)$

$\Rightarrow \bar{X} \approx N(14.5, 3/4)$

so $P(\bar{X} > 16.7) \doteq P\left(Z > \frac{16.7 - 14.5}{\sqrt{3/4}}\right)$

$= P(Z > 2.5403)$

$= 0.005537$ by normCdf (2.5403, 9E99)

≈ 0.00554 (3sf).

b) the CLT is justified here as the sample size is greater than 20.

6. a) species is categorical data

mean energy is continuous data

so a bar chart would be suitable, eg.

b) we have $\Sigma x = 25995$

$\Sigma x^2 = 68\,456\,947$

$n = 10$

Assuming that the energy produced is normally distributed, we calculate the sample mean and sample standard deviation

So $\bar{x} = \frac{1}{n}\Sigma x$

$= \frac{1}{10} \times 25995$

$= 2599.5$

$s_{n-1} = \sqrt{\frac{S_{xx}}{n-1}}$

$= \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$

$= \sqrt{\frac{68\,456\,947 - \frac{25995^2}{10}}{9}}$

$= \sqrt{\frac{1765889}{18}}$

$= 313.217...$

So 95% CI for mean is $\bar{x} \pm t_{9,0.975} \times \sqrt{\frac{313.217^2}{10}}$

$= 2599.5 \pm 2.262 \times \sqrt{\frac{313.217^2}{10}}$

$= (2375.44, 2823.56)$

(and check with tInterval energy, 1, 0.95 on TI-Nspire, with energy being the list of the 10 values)

c) we seek the tree type whose mean energy is in the interval $(2375.44, 2832.56)$

This includes Birch (2660) and Maple (2820)

So the 10 samples could have come from either tree.

d) 99% interval is $\bar{x} \pm t_{9,0.995} \times \sqrt{\frac{s_{n-1}^2}{10}}$

$= 2599.5 \pm 3.250 \times \sqrt{\frac{313.217^2}{10}}$

$= (2277.61, 2921.39)$

This interval is now wider and includes Pine, Birch, Maple, Elm

Thus her conclusion now presents a longer list of possible trees.

7. Sample size = 6.

a) $\bar{\bar{X}}$ = mean of $\bar{X}$ values

$= \frac{1}{8}(29.1 + 30.3 + \ldots + 29.2)$

$= 29.55$

so target value = 29.55.

$\bar{R}$ = mean of R values:

$= \frac{1}{8}(3.0 + 2.0 + \ldots + 1.8)$

$= 2.25$

So $\hat{\sigma} = \frac{\bar{R}}{d} = \frac{2.25}{2.534} = 0.887924$

So 3-sigma limits are $29.55 \pm 3 \times \frac{0.887924}{\sqrt{6}}$

$= (28.4625, 30.6375)$

b) now $n = 9$

$\bar{\bar{X}} = 29.92$

$\bar{R} = 2.35$

3-sigma upper = 30.71

$30.71 = 29.92 + 3 \times \frac{2.35/d}{\sqrt{9}}$

$30.71 = 29.92 + \frac{2.35}{d}$

$0.79 = \frac{2.35}{d}$

$d = \frac{2.35}{0.79}$

$d = 2.97468$

$d \approx 2.975$. (3dp)

8 a) there appears to be a positive correlation between gestation and IQ score.
It is not unreasonable to view a positive linear correlation to be present.

b) $S_{gg} = 531.5676$ $S_{gc} = 555.0811$ $S_{cc} = 1731.2973$

coeff of determination $= r^2$

$$= \frac{S_{gc}^2}{S_{gg} \times S_{cc}}$$

$$= \frac{555.0811^2}{531.5676 \times 1731.2973}$$

$$= 0.334798$$

$$\approx 0.335 \quad (3sf)$$

This measures the percent of variability explained by the linear regression model.
Here, it explains 33.5% of the variability.

c) Hypothesis test on ρ:

we have $r = \sqrt{0.335} = 0.578617$ and $n = 37$

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

We assume H_0 to be true

$\alpha = 1\%$ two tail test.

test statistic, $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 4.19709$

$p\text{-value} = 2 \times P(t_{35} > 4.19709)$

$= 2 \times 0.000088$

$= 0.000176$

< 0.01

OR

$t_{35, 0.995} = 2.724$

as $4.19709 > 2.724$, we are in critical region

So we have evidence to reject H_0 and population correlation coefficient is non zero.
we conclude that we have evidence of a linear association between gestation and IQ. We have assumed that the sample of 37 children are independent.

d) we should construct a residual plot to verify that the residuals are plausibly normally distributed with mean 0 and a constant variance.

9.

1. median = 65

n = 22 scripts.

We will assume that the parent distribution of exam scores is distributed symmetrically.

H_0 : new median = 65

H_1 : new median < 65

Assume H_0 to be true

$\alpha = 5\%$, one tail test

score	86	80	78	73	69	65	62	61	59	58	54	51	49	47	43	40	38	37	35	32	29	29
score-65	21	15	13	8	4	0	-3	-4	-6	-7	-11	-14	-16	-18	-22	-25	-27	-28	-30	-33	-36	-36
|score-65|	21	15	13	8	4		3	4	6	7	11	14	16	18	22	25	27	28	30	33	36	36
rank	13	10	8	6	2.5		1	2.5	4	5	7	9	11	12	14	15	16	17	18	19	20.5	20.5

So $W_+ = 13+10+8+6+2.5 = 39.5$

we have sample of 21 samples (due to removing 0) > 20 so Normal Approximation required.

let W = rank sum $E(W) = \frac{1}{4} \times 21 \times 22 = \frac{231}{2} = 115.5$ $\quad V(W) = \frac{1}{24} \times 21 \times 22 \times (2 \times 21 + 1) = \frac{3311}{4}$

we have observed value of 39.5 and $W \approx N\left(115.5, \frac{3311}{4}\right)$

so p-value $= P(W \leq 39.5)$

$\approx P\left(Z \leq \frac{40 - 115.5}{\sqrt{\frac{3311}{4}}}\right)$ using continuity correction

$= P(Z \leq -2.6242)$

$= 0.004343...$ from normCdf(-9E99, -2.6242)

< 0.05

so we have evidence to reject H_0

we conclude that the median score of the new cohort is less than 65.

10. W = wingspan $W \sim N(50, 4^2)$

$n = 25 \quad \bar{x} = 48.3$

a) $H_0: \mu = 50$

$H_1: \mu < 50$

Assume H_0 to be true.

$\alpha = 1\%$ and 5%

so $\bar{W} \sim N\left(50, \frac{4^2}{25}\right)$

$p\text{-value} = P(\bar{W} < 48.3)$

$= P\left(Z < \frac{48.3 - 50}{\sqrt{4^2/25}}\right)$

$= P(Z < -2.125)$

$= 0.016793$ from normCdf(-9E99, -2.125)

So at 1% we would not reject H_0

but at 5% we would reject H_0

So there is evidence of a decrease in wingspan at the 5% level, but not at the 1% level.

b) at 1% level, the critical value from $N(0,1^2)$ is -2.32635

so we reject H_0 if $\frac{\bar{x} - 50}{\sqrt{4^2/25}} < -2.32635$

$\Rightarrow \quad \bar{x} < 50 - 2.32635\sqrt{\frac{4^2}{25}}$

$\bar{x} < 48.1389$

at 5% level, we use -1.64485, giving $\bar{x} < 50 - 1.64485\sqrt{\frac{4^2}{25}}$

$\bar{x} < 48.6841$

so, to summarise, at 1%, $b = 48.14$ (2dp) cm

at 5%, $b = 48.68$ (2dp) cm.

if $\bar{x}$ is below either of these values of b, then you have evidence at that significance level that the wingspan has decreased.

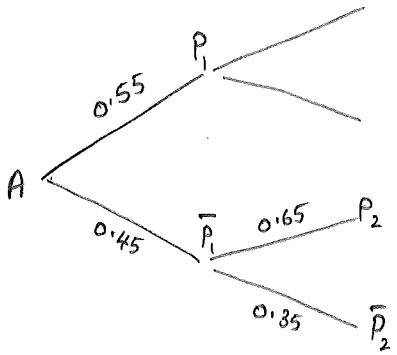
if $\bar{x}$ is above the values of b, then you do not have evidence.

c) with population variance unknown, you would seek to estimate the population variance from the sample variance.

so $\hat{\sigma}^2 = s^2 = \sqrt{\frac{S_{xx}}{n-1}}$

a consequence of this would be that the t_{24} distribution would be used instead of the $N(0, 1^2)$ distribution, but otherwise the process would be the same.

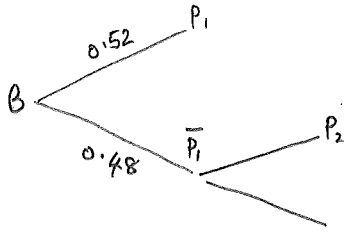
11.



a) $P(\text{pass after 1 or 2 attempts}) = P(\text{Pass}_1) + P(\overline{\text{Pass}_1})\,P(\text{Pass}_2 \mid \overline{\text{Pass}_1})$

$= 0.55 + 0.45 \times 0.65$

$= \underline{\underline{0.8425}}$. or 84.3% (approximately)



$P(\overline{P_1} \cap P_2) = 0.35$

b) i) $P(\text{pass after 1 or 2 attempts}) = P(\text{Pass}_1) + P(\overline{P_1} \cap P_2)$

$= 0.52 + 0.35$

$= 0.87$

> 0.8425

So Instructor B's <u>claim is correct</u>.

ii) $P(P_2 \mid \overline{P_1}) = \dfrac{P(P_2 \cap \overline{P_1})}{P(\overline{P_1})}$

$= \dfrac{0.35}{0.48}$

$= 0.729167$

≈ 0.7292 (4dp)

So <u>73% are expected to pass at their second attempt</u>.

c) $P(\text{instructor A}) = \frac{1}{3}$

$P(\text{instructor B}) = \frac{2}{3}$.

$P(\text{instructor B} \mid \text{failed after 2 attempts}) = ?$

we have that $P(\text{failed after 2 attempts} \mid \text{instructor A}) = 0.45 \times 0.35$
$= 0.1575$

and $P(\text{failed after 2 attempts} \mid \text{instructor B}) = 0.48 \times (1 - 0.7292)$
$= 0.13$

So $P(\text{instructor B} \mid \text{failed after 2 attempts})$

$$= \frac{P(\text{instructor B and failed after 2 attempts})}{P(\text{failed after 2 attempts})}$$

$$= \frac{P(\text{instructor B})\, P(\text{failed after 2 attempts} \mid \text{instructor B})}{P(\text{failed after 2 attempts})}$$

$$= \frac{\frac{2}{3} \times 0.13}{\frac{2}{3} \times 0.13 + \frac{1}{3} \times 0.1575}$$

$$= 0.622754$$

$$\approx 0.6228 \quad (4dp)$$

12.

unbiased octahedral dice : 1 to 8

if 1,1,1 then win \$100

if 1,1,X
1,X,1
X,1,1
} then win \$10

if 1,X,X
X,1,X
X,X,1
} then win \$1

$P(\text{win \$100}) = \left(\frac{1}{8}\right)^3 = \frac{1}{512}$

$P(\text{win \$10}) = 3 \times \left(\frac{1}{8}\right)^2 \times \left(\frac{7}{8}\right) = \frac{21}{512}$

$P(\text{win \$1}) = 3 \times \left(\frac{1}{8}\right) \times \left(\frac{7}{8}\right)^2 = \frac{147}{512}$

so if X = profit for one game

If pay \$1 and win \$100, then profit is \$99

so

x	-1	0	9	99
$P(X=x)$	$\frac{343}{512}$	$\frac{147}{512}$	$\frac{21}{512}$	$\frac{1}{512}$

or, in decimals,

x	-1	0	9	99
$P(X=x)$	0.6699	0.2871	0.0410	0.0020

so $E(X) = \sum x P(X=x)$

$= -1 \times 0.6699 + 0 + 9 \times 0.0410 + 99 \times 0.0020$

$= -0.1029$ as required.

$E(X^2) = 1 \times 0.6699 + 0 + 9^2 \times 0.0410 + 99^2 \times 0.0020$

$= 23.5929$

so $V(X) = E(X^2) - E^2(X)$

$= 23.5929 - (-0.1029)^2$

$= 23.5823$

so $SD(X) = \sqrt{23.5823}$

$= 4.8562$ (4dp) as required.

12 cont

Y = profit for one game

$E(Y) = -0.06$

$SD(Y) = 20 \quad \Rightarrow V(Y) = 20^2 = 400$

let Y_i = profit for game i

so $E(Y_i) = -0.06$ and $V(Y_i) = 400$

b) i) plays low stakes 60 times and high stakes 45 times

so let $T = X_1 + \ldots + X_{60} + Y_1 + \ldots + Y_{45}$ stand for total winnings.

$$\begin{aligned} E(T) &= E(X_1 + \ldots + X_{60} + Y_1 + \ldots + Y_{45}) \\ &= E(X_1) + \ldots + E(X_{60}) + E(Y_1) + \ldots + E(Y_{45}) \\ &= 60 \times (-0.1029) + 45 \times (-0.06) \\ &= -8.874 \end{aligned}$$

$$\begin{aligned} V(T) &= V(X_1 + \ldots + X_{60} + Y_1 + \ldots + Y_{45}) \\ &= V(X_1) + \ldots + V(X_{60}) + V(Y_1) + \ldots + V(Y_{45}) \quad \text{assuming all games are independant for this step} \\ &= 60 \times V(X_i) + 45 \times V(Y_i) \\ &= 60 \times 4.8562^2 + 45 \times 400 \\ &= 19415 \end{aligned}$$

so $SD(T) = \sqrt{19415}$

$= 139.338\ldots$

$= 139.34$ (2dp)

so mean winnings = -8.874 dollars, with standard deviation 139.34 dollars.

ii) The player can therefore expect to lose \$8.87 in the 105 games that they play

We'd expect their winnings to within a couple of standard deviations of the mean

i.e. $-8.874 \pm 2 \times 139.34$

$\approx (-287.55, 269.80)$