2021 AH Stats - Paper 1

1 a)
- title each side - which is cats, which is dogs
- legend eg. 1|6 = 1·6
- left side ordered in increasing order away from stem

b) cats upper fence $= Q3 + 1{\cdot}5\,(Q3 - Q1)$

$= 6{\cdot}5 + 1{\cdot}5\,(6{\cdot}5 - 3{\cdot}7)$

$= 10{\cdot}7$

no. data points above 10·7 on right side of stem and leaf $= \underline{\underline{5}}$

c) i) two sample proportion test using $\frac{121}{150}$ and $\frac{145}{150}$ as test statistics

ii) $H_0: p_{cats} = p_{dogs}$

$H_1: p_{cats} \neq p_{dogs}$.

d) assumptions: shape and spread of drawing times were the same for the two populations of times, from which the samples were taken.

the samples were independent and randomly selected from the population

e) $m = 121$ cats

$n = 145$ dogs

$W_{cats} = 12048$

$E(W) = \frac{1}{2}m(m+n+1)$

$= \frac{1}{2} \times 121 \times (121 + 145 + 1)$

$= 16153{\cdot}5$

$V(W) = \frac{1}{12} \times mn(m+n+1)$

$= \frac{1}{12} \times 121 \times 145 \times (121 + 145 + 1)$

$= 390376{\cdot}25$

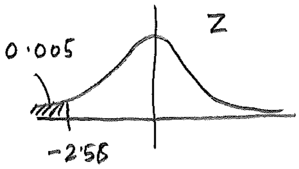
test statistic, $z = \dfrac{12048{\cdot}5 - 16153{\cdot}5}{\sqrt{390376{\cdot}25}}$ with continuity correction as we want $P(W \leq 12048)$

$= -6{\cdot}57069$

$\approx -6{\cdot}57$

1f) H_0 : median time to draw a cat = median time to draw a dog

H_1 : median time to draw a cate ≠ median time to draw a dog.



2-tailed test

⇒ significante level = 2×0.005

= 0.01

= 1%

we reject Ho and
we have evidence to suggest that the median time to draw a doodle of a cat is different to that for a dog.

1g) assumption : the variances of the two populations are equal

we would look at the standard deviations of the samples, 2.307 and 2.655.

2. a) the relationship has a positive correlation, but it does not appear to be a linear relationship.

b) each model has a high value of r (>0.89 in both cases) and both have very small p-values under the assumption that the correlation is 0

hence, both models would seem to be appropriate

c) $R^2 = (0.8971518)^2$

$= 0.804881$

$\approx 80.5\%$ (3 sf)

a least squares regression line would explain 80.5% of the variation in the data, when estimating $\sqrt{cost}$ from length.

d) Both residual plots have a random scatter of points around 0, with constant variation. Hence, they support the assumptions that $E(\varepsilon_i)=0$ and $V(\varepsilon_i)=\sigma$, a constant

Hence both models still remain valid

e) estimated $\sqrt{cost}$ = midpoint of interval

$= \frac{1}{2}(592.0915 + 661.634)$

$= 626.863$

$\Rightarrow$ estimated cost $= (626.863)^2$

$= £392\,957.$

95% CI for $\sqrt{cost}$ $= (592.0915, 661.634)$

95% CI for cost $= (350572, 437760)$

$= (£350\,572, £437\,760)$

f) i) the models were fitting cost to length, and not length to cost

ii) Peter should fit a new model of length based on cost, and see if a linear relationship could exist after any appropriate transformations.

2021 - AH Stats - Paper 2.

1.

a)

s	2	4	6	8	10
$P(S=s)$	$\frac{2}{k}$	$\frac{4}{k}$	$\frac{6}{k}$	$\frac{8}{k}$	$\frac{10}{k}$

$$\sum P(S=s) = \frac{2+4+6+8+10}{k} = \frac{30}{k}$$

as $\sum P(S=s) = 1$, then $k = 30$.

So

s	2	4	6	8	10
$P(S=s)$	$\frac{2}{30}$	$\frac{4}{30}$	$\frac{6}{30}$	$\frac{8}{30}$	$\frac{10}{30}$

b) $E(S) = \frac{1}{30}(4+16+36+64+100)$

$= \frac{1}{30} \times 220$

$= \frac{22}{3}.$

$E(S^2) = \frac{1}{30}\left(2^2 \times 2 + 4^2 \times 4 + 6^2 \times 6 + 8^2 \times 8 + 10^2 \times 10\right)$

$= \frac{1}{30}\left(1800\right)$

$= 60$

So $V(S) = E(S^2) - E^2(S)$

$= 60 - \left(\frac{22}{3}\right)^2$

$= \frac{56}{9}.$

2 F = female

T = teacher

A = admin

a) $$P(F \cap T) = \frac{18}{(18+12+7+3+5+5)}$$

$$= \frac{18}{50}$$

$$= \underline{\underline{\frac{9}{25}}}$$

$$P(F \cup \bar{A}) = P(F) + P(\bar{A}) - P(F \cap \bar{A})$$

$$= \frac{18+7+5}{50} + \frac{18+12+5+5}{50} - \frac{18+5}{50}$$

$$= \frac{30}{50} + \frac{40}{50} - \frac{23}{50}$$

$$= \underline{\underline{\frac{47}{50}}}$$

b) i) $P(T) = \frac{18+12}{50} = \frac{30}{50}$ $\qquad P(D|T) = 0.8$

$P(A) = \frac{10}{50}$ $\qquad P(D|A) = 0.5$ $\qquad P(\bar{D}|A) = 0.5$

$P(O) = \frac{10}{50}$ $\qquad P(D|O) = 0.3$

let D = drive to school

$$P(D) = P(D|T)P(T) + P(D|A)P(A) + P(D|O)P(O)$$

$$= 0.8 \times \frac{30}{50} + 0.5 \times \frac{10}{50} + 0.3 \times \frac{10}{50}$$

$$= \frac{32}{50}$$

$$= \underline{\underline{0.64}}$$

ii) $$P(A|\bar{D}) = \frac{P(A \cap \bar{D})}{P(\bar{D})}$$

$$= \frac{P(\bar{D} \cap A)}{P(\bar{D})}$$

$$= \frac{P(\bar{D}|A)P(A)}{P(\bar{D})}$$

$$= \frac{0.5 \times \frac{10}{50}}{1 - 0.64}$$

$$= 0.27777\ldots$$

$$= \underline{\underline{0.2778 \text{ (4dp)}}}$$

3.

Assume that distribution of daily mite counts is symmetrical.

H_0 : median mite count = 7

H_1 : median mite count > 7

Assume H_0 to be true

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

mites	6	8	11	13	6	14	11	9	6	7	11	8	6	14
mite − 7	−1	1	4	6	−1	7	4	2	−1	0	4	1	−1	7
|mite − 7|	1	1	4	6	1	7	4	2	1		4	1	1	7
rank.	1	2	8	11	3	12	9	7	4		10	5	6	13
	↓	↓	↓		↓	↓	↓		↓		↓	↓	↓	↓
	3·5	3·5	9		3·5	12·5	9		3·5		9	3·5	3·5	12·5

$W_- = 3{\cdot}5 + 3{\cdot}5 + 3{\cdot}5 + 3{\cdot}5 = 14$

W_+ is $> W_-$

$n = 13.$

we want $P(W_- \leq 14)$

from tables for $n=13$, $P(W \leq 21) = 0{\cdot}05$
$P(W \leq 17) = 0{\cdot}025$
$P(W \leq 12) = 0{\cdot}01$

Hence $14 < 21$, so we are in the 5% critical region

We have endence to reject H_0 and conclude that the median mite count is greater than 7, and thus the colony's health is at risk

4. $W \sim Po(5)$

$F \sim Po(2.3)$

$M \sim Po(1.2)$

a) i) $P(W > 10) = 1 - P(W \le 10)$

$= 1 - 0.986305$ from poissCdf (5, 0, 10)

$= 0.013695$

≈ 0.0137 (4dp)

ii) let $X = F + M$

$X \sim Po(2.3 + 1.2)$

$X \sim Po(3.5)$

$P(X=2) = 0.184959$ from poissPdf (3.5, 2)

$= 0.1850$ (4dp)

b) 6 weeks = $6 \times 7 = 42$ days

let Y = total captures over 6 weeks

$Y = W_1 + \dots + W_{42} + F_1 + \dots + F_{42} + M_1 + \dots + M_{42}$ (ie. 126 random variables)

$Y \sim Po(5 \times 42 + 2.3 \times 42 + 1.2 \times 42)$ as all W_i, F_i and M_i are iid.

$Y \sim Po(357)$

let W = normal approximation for Y

$W \sim N(357, 357)$

$P(Y < 340) = P(W < 339.5)$ by cc.

$= P\left(Z < \dfrac{339.5 - 357}{\sqrt{357}}\right)$

$= P(Z < -0.926198)$

$= 0.177171$ by normCdf (−9E99, −0.926)

$= 0.1772$ (4dp)

5. $H_0 : \mu_A = \mu_B$

$H_1 : \mu_A \neq \mu_B$

Assume H_0 to be true

$\alpha = 5\%$ 2-tail test

$\bar{x}_A = 54$	$s_A = 5$	$n_A = 11$
$\bar{x}_B = 47$	$s_B = 11$	$n_B = 15$

$$\text{pooled } s^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

$$= \frac{10 \times 5^2 + 14 \times 11^2}{11 + 15 - 2}$$

$$= 81$$

$$\Rightarrow \quad s = 9$$

$$\text{test statistic, } t = \frac{\bar{x}_A - \bar{x}_B - (0)}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{54 - 47 - 0}{9\sqrt{\frac{1}{11} + \frac{1}{15}}} = 1.95934$$

$P(t_{24} > 1.95934) = 0.030893$

p-value $= 2 \times 0.030893 = 0.061787 > 0.05$.

So we do not reject H_0 at the 5% level of significance, and we conclude that we do not have evidence that the mean warm up times are different.

6. a) stratified random sampling

b) convenience sampling

results would be unreliable as it would be biased and not be representative of the intended population

c) $\bar{x} = 409$

$\sigma = 130$

$n = 25$

so $\bar{X} \sim N(\mu, \frac{130^2}{25})$

so 95% CI is $\bar{x} \pm Z_{0.975} \times \sqrt{\frac{130^2}{25}}$

$= 409 \pm 1.96 \times \sqrt{\frac{130^2}{25}}$

$= (358{\cdot}041, 459{\cdot}959)$

$\approx (358{\cdot}0, 460{\cdot}0)$ minutes.

d) the previous population mean of 458 mins is captured by the 95% CI, but not by the 90% CI

Hence, by presenting the 90% CI only, the student representative would be able to argue that mean screen time had decreased, and thus claim the reward.

7. $X \sim U(78, 83)$

$\bar{X} = \frac{1}{75}(X_1 + \dots + X_{75})$

a) $$P(80.45 < X < 80.83) = \frac{80.83 - 80.45}{83 - 78}$$

$$= \frac{0.38}{5}$$

$$= 0.076.$$

b) by CLT, $\bar{X} \approx N(\mu, \frac{\sigma^2}{75})$ $\quad \mu = \frac{78+83}{2}$ $\quad \sigma^2 = \frac{(83-78)^2}{12}$

$\mu = 80.5$ $\quad \sigma^2 = \frac{25}{12}$

so $\bar{X} \approx N(80.5, \frac{1}{36})$ by CLT, as $n > 20$

c) $$P(80.45 < \bar{X} < 80.83) = P\left(\frac{80.45 - 80.5}{\sqrt{\frac{1}{36}}} < Z < \frac{80.83 - 80.5}{\sqrt{\frac{1}{36}}}\right)$$

$$= P(-0.3 < Z < 1.98)$$

$$= 0.59406$$

$$\approx 0.5941 \text{ (4dp)}$$

8. $X =$ time taken

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

$n = 50, \bar{x} = 16 \cdot 1$

$H_0 : \mu = 15$

$H_1 : \mu > 15$

assume H_0 to be true.

$\alpha = 5\%$. 1 tail test

$X \sim N(15, 2^2)$

$\bar{X} \sim N(15, \frac{2^2}{50})$

$$P(\bar{X} \geq 16 \cdot 1) = P(Z > \frac{16 \cdot 1 - 15}{\sqrt{\frac{2^2}{50}}})$$

$$= P(Z > 3 \cdot 88909)$$

$$= 0 \cdot 00005$$

$$< 0 \cdot 05$$

So we have evidence at the 5% level (and also 1% level) to reject H_0 and conclude that the mean time to solve the problem is more than 15 mins.

possible explanations – candidates not good enough

– expectations to complete the task were not reasonable ie. the task is too hard to solve in 15 mins.

9 $E(A) = 2.5$ $V(A) = 4^2$

$E(B) = 1$ $V(B) = 5^2$

a) $C = A - B$

$$E(C) = E(A-B) = E(A) - E(B) = 2.5 - 1 = \underline{\underline{1.5}}$$

$$V(C) = V(A-B) = V(A) + V(B) = 16 + 25 = \underline{\underline{41.}}$$

b) C = extra monthly profit the company makes from policy A over policy B, for each £10 premium.

c) T = total profit = $A_1 + \dots + A_{33} + B_1 + \dots + B_{26}$

$$V(T) = V(A_1 + \dots + A_{33}) + V(B_1 + \dots + B_{26})$$
$$= 33 \times V(A) + 26 \times V(B)$$
$$= 33 \times 4^2 + 26 \times 5^2$$
$$= 1178$$
$$SD(T) = \underline{\underline{34.32}}$$

∴ standard deviation of total monthly profit is $\underline{\underline{£34.32}}$.

10.

a) $X \sim B(50, p)$ $\quad$ X = no. shops selling produce past sell-by date

Y = nomal approx to X

$Y \sim N(50p, 50pq)$

$\frac{Y}{n}$ = proportion of shops selling produce past sell-by date

$\frac{Y}{50} \sim N(p, \frac{pq}{50})$

so 95% CI for $p = \hat{p} \pm z_{0.975} \times \sqrt{\frac{\hat{p}\hat{q}}{50}}$

$= \frac{13}{50} \pm 1.95996 \times \sqrt{\frac{\frac{13}{50} \times \frac{37}{50}}{50}}$

$= (0.138419, 0.381581)$

$= \underline{(0.1384, 0.3816)}$ to 4dp.

b) width of 0.04 ⇒ ± 0.02

so $z_{0.975} \sqrt{\frac{\frac{13}{50} \times \frac{37}{50}}{n}} = 0.02$

$\frac{\frac{481}{2500}}{n} = \left(\frac{0.02}{1.96}\right)^2$

$n = \frac{\frac{481}{2500}}{\left(\frac{0.02}{1.96}\right)^2}$

$n = 1847.74$

So we would need a sample size of at least <u>1848 shops</u>

11. a) the distribution of scores appear normally distributed

b) we perform a non-paired two sample t-test

this requires the assumption that the population standard deviations are equal (here, the sample standard deviations are 10.08 and 10.49, so it's plausible for this assumption to be true)

we could also perform a non-paired two sample z-test

this requires us to assume the population standard deviations to be well estimated by the sample standard deviations, which is plausible given the large sample sizes.

both of the above tests additionally require the sample for group B to be independent from group C, so that the variance of their difference can be calculated

given that the t-test will lead to a distribution of t_{128} which is very close to Z, we proceed with the z-test as it does not require the assumption of equal variaces.

$H_0: \mu_B = \mu_C$

$H_1: \mu_B \neq \mu_C$

Assume H_0 to be true

$\alpha = 5\%$ 2-tail test

$$\text{test statistic, } z = \frac{\bar{x}_B - \bar{x}_C - (0)}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_C^2}{n_C}}}$$

$$= \frac{55.4 - 51.8 - (0)}{\sqrt{\frac{10.08^2}{70} + \frac{10.49^2}{60}}}$$

$$= 1.9861$$

$$\text{p-value} = 2 \times P(Z > 1.9861)$$

$$= 2 \times 0.023511$$

$$= 0.047022$$

$$< 0.05$$

(as an aside, the t-test statistic has p-value of 0.048472 calculated on TI-Nspire)

So we have evidence to reject H_0 and conclude that mean score for group B is different to mean score for group C.

11 c)

m_B = maximum mean score for B

$H_0: \mu_A = \mu_B$
$H_1: \mu_A > \mu_B.$

test statistic would be $z = \dfrac{\bar{x}_A - m_B}{\sqrt{\dfrac{s_A^2}{120} + \dfrac{s_B^2}{70}}}$

$$= \frac{56 - m_B}{\sqrt{\dfrac{10.71^2}{120} + \dfrac{10.08^2}{70}}}$$

and we want this to be 10% significant ⇒ critical value of $z_{0.90} = 1.28155$

$$\Rightarrow \quad 1.28155 = \frac{56 - m_B}{\sqrt{2.40739}}$$

$$\Rightarrow \quad m_B = 56 - 1.28155 \times \sqrt{2.40739}$$

$$\Rightarrow \quad m_B = 54.0116$$

Hence maximum mean score for group B is $\underline{54.01}$ (2dp)

12.

X = weights of cereal bags

$X \sim N(500, 5.73^2)$

$n = 5$

$\bar{X} \sim N\left(500, \frac{5.73^2}{5}\right)$ where $\bar{X}$ = mean weight of 5 bags.

a) 1σ limits. $= 500 \pm 1 \times \sqrt{\frac{5.73^2}{5}}$

$= 497.437, 502.563$

b) i) $P(\bar{X} > 502.563) = P(Z > 1) = 0.158655$

let Y = no. of samples beyond 1σ limit

$Y \sim B(3, 0.158655)$

$P(Y \geq 2) = 0.067527$ from binomCdf(3, 0.158 655, 2, 3)

so P(at least 2 beyond same 1σ limit) = P(above + below)

$= 2 \times P(Y \geq 2)$

$= 2 \times 0.067527$

$= 0.135055$

$= 0.1351$ (4dp)

ii) the probability from (a)(i) is too high, giving rise to too many false alarms.

13. a) explanations

- one lecturer was better than the other
- one group of students was more academically able than the other, possibly due to the entrance requirements of their course.

b) improvement - combining categories to reduce occurence of low expected frequencies.

i.e. group D + E together

f_o	A	B	C	D+E
Psy	13	11	6	6
Bio	4	4	6	9

f_e	A	B	C	D+E
Psy	10·4	9·2	7·3	9·2
Bio	6·6	5·8	4·7	5·8

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 6{\cdot}06123$$

H_0 : no association between grade and course

H_1 : association between grade and course

$$df = (4-1) \times (2-1) = 3$$

$$P(\chi^2_3 > 6{\cdot}06123) = 0{\cdot}108669$$

$$> 0{\cdot}10$$

so we don't have evidence to reject H_0 and conclude that we don't have evidence of an association between grade and course.

c) the greatest contribution to X^2 was of 1·69963 by the Biology Grade D+E group

(obtained by scrutinising the stat. CompMatrix output after performing χ^2 test on Ti-Nspire)