2024 AH Statistics Paper 1

1a) there is a mismatch between the intended population of UK films, and the sample that was from Worldwide films. ie. the sample is not representative of the UK population.

1b) i) The sample took a fixed number of films per year, and per decade. It was therefore not a proportional sample that took into account the numbers of films released in each year. Stratified sampling requires a proportional sample to be taken from each strata/year/decade.

ii) For each year that was sampled, count up the total number of films that were released. Call this total count to be N.

Now obtain 1% of N by dividing it by 100, and rounding to the nearest integer. Call this M, and it's the number of films to be sampled from each year.

Number all films in a year from 1 to N, and use a random number generator to give M unique values between 1 and N. These are the numbered films to be sampled.

Repeat for each year, and combine the samples together to obtain a 1% stratified sample from the population.

1c) i) H_0 : no association between decade and age rating of film

H_1 : there is an association between decade and age rating of film.

c) ii) Check that no expected frequency is less than 1, and at least 80% of the calculated expected frequencies would be greater than or equal to 5.

c) iii) One would either combine rows or columns. At present it is a 4 row by 5 column table. If, say, combining the first two columns satisfied the check, then we would have a 4 row × 4 column table $\Rightarrow (4-1)\times(4-1) = 9$ degrees of freedom.

1 d) i) $\hat{p}_1 = \frac{37}{150}$ $n_1 = 150$

$\hat{p}_2 = \frac{5}{50}$ $n_2 = 50$

pooled proportion, $p = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

$= \dfrac{37 + 5}{150 + 50}$

$= \dfrac{42}{200}$.

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{\frac{37}{150} - \frac{5}{50}}{\sqrt{\frac{42}{200} \times \frac{158}{200}\left(\frac{1}{150} + \frac{1}{50}\right)}}$$

$= 2 \cdot 20508$

to have obtained $z = -2 \cdot 20508$, they must have had $p_1 = \frac{5}{50}$, $p_2 = \frac{37}{150}$

$n_1 = 50$, $n_2 = 150$

which leads to the negative value of the above calculation.

1d) ii) A proportion test is based upon a normal approximation to a binomial distribution. This therefore requires $n_ip_i > 5$ and $n_iq_i > 5$. were $B(n_i, p_i)$ was the starting distribution. for each of the two samples.

If this check is not satisfied, then the nomal approximation is not a good one, and it jeopardises the relliability of the test. Hence, the final conclusion of the test may not be valid, or robust.

1 e)

"numbers of films in every age category have changed"

- not true, as the report did not look at number of films that were released.

"increase in number of 'family friendly' films"

- not true, as family friendly was defined earlier as U, PG and 12 films, and not just '12' films.

(indeed the U and PG films seem to have decreased over the decades).

2 a) • neither distribution for standard or new treatment appears to have a normal distribution of viral load.

• a two sample z-test would require knowing the population variances, which we do not have.

2 b) i) H_0: population median viral load of standard treatment = population median viral load of new treatment

H_1: population median viral load of standard treatment ≠ population median viral load of new treatment

(or $H_0: \eta_{standard} = \eta_{new}$
$H_1: \eta_{standard} \neq \eta_{new}$)

b) ii) all 29 numbers were placed in ascending value order, keeping track of which numbers stand for 'standard' and which for 'new'

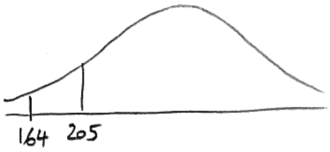
rank all numbers from 1 to 29, dealing with any tied ranks that may exist.

sum the ranks for the 'new' group, as it is the smallest sized group (of 14)

also rank all numbers from 29 to 1 (the reverse of previously) and recalculate the sum of ranks for the 'new' group.

Of the two rank sums for new that are calculated, we the smallest, that should be equal to 205.

b) iii) $\alpha = 5\%$.
two tail test
$m = 14$, $n = 15$
$\Rightarrow$ critical value from tables = 164.



as $205 > 164$, we do not reject Ho

we have evidence to suggest that there is no difference in the median viral load of the two treatments, after 3 months

2 c) missing value is $\underline{\underline{6{\cdot}5}}$, as it is the same as the other difference value of 20.

2 d) $W_- = 3+3+3+9 = 18$

$W_+ = 6{\cdot}5 + 12 + 6{\cdot}5 + 15 + 3 + 9 + 12 + 12 + 3 + 14 + 9$
$= 102$ } or $W_+ = \frac{1}{2} \times 15 \times (15+1) - 18$

$\therefore$ W = minimum of 18 and 102
$\underline{\underline{= 18}}$

2 e) i) $n = 15$
$W = 18$

two-tailed test.

from tables, critical value of 19 $\leftrightarrow$ $p = 0{\cdot}02$
" " " 15 $\leftrightarrow$ $p = 0{\cdot}01$

hence $\underline{0{\cdot}01 < p < 0{\cdot}02}$.

ii) as p-value < 5%, we have evidence to reject H_0

so we have evidence to suggest that there is a difference in median viral loads between 3 months and 6 months.

2024 AH Statistics – Paper 2

1. $11 \quad 27 \quad 45 \quad \underset{LQ}{\boxed{63}} \quad 65 \quad 70 \quad 77 \,|\, 79 \quad 87 \quad 88 \quad \underset{UQ}{\boxed{90}} \quad 95 \quad 102 \quad 130$

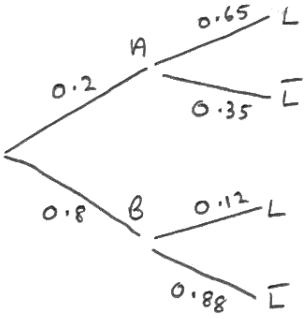
$IQR = 90 - 63$
$= 27$

lower fence $= LQ - 1{\cdot}5 \times IQR$
$= 63 - 1{\cdot}5 \times 27$
$= 22{\cdot}5$

Upper Fence $= UQ + 1{\cdot}5 \times IQR$
$= 90 + 1{\cdot}5 \times 27$
$= 130{\cdot}5$.

as $11 < 22{\cdot}5$, it is a possible outlier, There are no data points above the upper fence.

2.



L = late for work

a) $P(L) = P(A \cap L) + P(B \cap L)$
$= P(A)P(L|A) + P(B)P(L|B)$
$= 0{\cdot}2 \times 0{\cdot}65 + 0{\cdot}8 \times 0{\cdot}12$.
$= 0{\cdot}226$.

b) $P(B|L) = \dfrac{P(B \cap L)}{P(L)}$
$= \dfrac{0{\cdot}8 \times 0{\cdot}12}{0{\cdot}226}$
$= 0{\cdot}424779\ldots$
$\approx 0{\cdot}4248$ (4dp)

3. a) X = no. bins emptied on first lift

$X \sim B(12, 0.88)$

$P(X=9) = {}^{12}C_9 \, (0.88)^9 \, (0.12)^3$

$= 0.120312\ldots$

$= \underline{\underline{0.1203}}$ (4dp) or via binompdf(12, 0.88, 9)

b) Y = no. bins emptied

$Y \sim B(48, 0.88)$

$\begin{cases} 48 \times 0.12 = 5.76 > 5 \ \checkmark \\ 48 \times 0.88 = 42.24 > 5 \ \checkmark ☺ \end{cases}$

use normal approximation.

$B(48, 0.88) \approx N(48 \times 0.88, \ 48 \times 0.88 \times 0.12)$

$= N(42.24, \ 5.0688)$

So 75% of 48 = 36

$P(Y > 36) \approx P\left(Z > \dfrac{36.5 - 42.24}{\sqrt{5.0688}}\right)$ using continuity correction

$= P(Z > -2.54953)$

$= 0.994606$ from normcdf(−2.54953, 9E99)

$\approx \underline{\underline{0.9946}}$ (4dp)

4.

grade	A	B	C	D	E	total
f_o	185	170	197	163	155	870.
f_e	174	174	174	174	174	

$\frac{870}{5} = 174.$

H_0: number of students at each grade is uniformally distributed, $U(5)$

H_1: number of students at each grade is not uniformally distributed, $U(5)$

one-tail test

$\alpha = 10\%$

assume Ho to be true.

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

$$= 6.5977$$

we have $5-1=4$ degrees of freedom (5 categories, 1 constraint being the total)

$P(\chi^2_4 > 6.5977) = 0.158738$

$\therefore$ p-value $= 0.158738 > 0.10$

$\therefore$ we do not reject Ho

So we do not have evidence to suggest that the distribution of grades is not unifomally distributed as $U(5)$.

5 a) the relationship between errors and attempts is non-linear with a negative association. (It appears to be inversely proportional).

b) i) $b = \frac{S_{xy}}{S_{xx}}$ now $S_{xx} = 158 \cdot 89$

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

and in this context $S_{x\frac{1}{y}} = \sum x_i \frac{1}{y_i} - \frac{\sum x_i \sum \frac{1}{y_i}}{n}$

$$= 34 \cdot 03 - \frac{55 \times 3 \cdot 45}{9}$$

$$= 12 \cdot 9467$$

so $b = \frac{12 \cdot 9467}{158 \cdot 89} = 0 \cdot 081482$

$$a = \bar{y} - b\bar{x}$$

$$= \frac{1}{9} \times 3 \cdot 45 - 0 \cdot 081482 \times \frac{1}{9} \times 55$$

$$= -0 \cdot 114612$$

so least squares regression equation, $\frac{1}{y} = -0 \cdot 114612 + 0 \cdot 081482x$.

b) ii) single rat → prediction interval

$x = 7 \Rightarrow \frac{1}{y} = -0 \cdot 114612 + 0 \cdot 081482 \times 7$

$= 0 \cdot 455762$

$t_{9-2, 0 \cdot 975} = 2 \cdot 36462$

so 95% PI for $\frac{1}{y} = 0 \cdot 455762 \pm 2 \cdot 36462 \times 0 \cdot 078 \times \sqrt{1 + \frac{1}{9} + \frac{\left(7 - \frac{55}{9}\right)^2}{158 \cdot 89}}$

$= (0 \cdot 26091, 0 \cdot 650614)$

$\Rightarrow y = \left(\frac{1}{0 \cdot 650614}, \frac{1}{0 \cdot 26091} \right)$

$\Rightarrow y = (1 \cdot 53701, 3 \cdot 83275)$

so we would expect a rat on its 7th attempt to make 2 or 3 errors

(as $1 \cdot 53701 < n < 3 \cdot 83275$, $n \in \mathbb{Z}$. $\Rightarrow n = 2, 3$)

6.

a) X = no. of crackers that work

$X \sim B(20, p)$

sample proportion, $\hat{p} = \frac{14}{20}$.

let Y = normal approx to X $\qquad np > 5,\ nq > 5$ when $n = 20,\ \hat{p} = \frac{14}{20},\ \hat{q} = \frac{6}{20}$ ✓ ☺.

$Y \sim N(20p, 20pq)$

let $\frac{Y}{20}$ = proportion of crackers that work

so $\frac{Y}{20} \sim N\left(p, \frac{pq}{20}\right)$

so 99% confidence interval $= \hat{p} \pm Z_{0.995}\sqrt{\frac{\hat{p}\hat{q}}{20}}$

$= \frac{14}{20} \pm 2.57583\sqrt{\frac{\frac{14}{20} \times \frac{6}{20}}{20}}$

$= (0.436056, 0.963944)$

$\approx (0.4361, 0.9639)$ (4dp)

assumptions made:

- the chance of any cracker working is independent of any other cracker working (from binomial model)
- the likelihood of a cracker working is fixed, and does not change (from binomial model)

b) as 0.75 lies within the interval $(0.4361, 0.9639)$ our sample of 14 out of 20 supports the belief that 75% of crackers do work properly.

7. $X \sim U(8) \quad \Rightarrow E(X) = \frac{9}{2} \quad V(X) = \frac{63}{12}.$

$Y = 3X - 2.$

$$\begin{aligned} E(Y) &= 3E(X) - 2 \\ &= 3 \times \frac{9}{2} - 2 \\ &= \frac{27}{2} - 2 \\ &= \frac{23}{2} \\ &= \underline{\underline{11.5}}. \end{aligned}$$

$$\begin{aligned} V(Y) &= 3^2 V(X) \\ &= 9 \times \frac{63}{12} \\ &= \frac{189}{4} \\ &= 47.25. \end{aligned}$$

$$\begin{aligned} \Rightarrow SD(Y) &= \sqrt{47.25} \\ &= 6.87386\ldots \\ &\approx \underline{\underline{6.8739}} \ (4dp) \end{aligned}$$

8. we will conduct a two sample t-test for a difference in population means:

- two unpaired data sets
- population variances not known
- small sample sizes.

so $H_0 : \mu_1 = \mu_2$ where μ_1 = population mean height in Area 1

$H_1 : \mu_1 > \mu_2$ μ_2 = population mean height in Area 2.

let $\alpha = 5\%$.

one-tailed test

assume H_0 to be true.

test statistic, $$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$ where $$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s = \sqrt{\frac{8 \times 3.42^2 + 10 \times 3.51^2}{18}}$$

$$s = \sqrt{12.0429}$$

$$s = 3.47029...$$

$$t = \frac{38.7 - 36.9 - (0)}{3.47029 \times \sqrt{\frac{1}{9} + \frac{1}{11}}}$$

$$t = 1.15401$$

$$\text{p-value} = P(t_{18} > 1.15401)$$
$$= 0.131794$$
$$> 0.05$$

so we do not reject H_0

so we do not have evidence to suggest that the mean heights of plants in Area 1 is greater than those in Area 2.

Underlying assumptions: • heights of plants in each area are normally distributed

• variance of heights of plants in each area are equal.

9. a) mean number of failures $= \dfrac{6000}{2400}$

$= 2 \cdot 5$

so if X = no. of failures per 6000 mile race

$X \sim Po\,(2 \cdot 5)$

assumptions are - component failures are independent of each other

- the mean failure rate remains constant throughout the race.

b) $P(X=0) = \dfrac{2 \cdot 5^0 \times e^{-2 \cdot 5}}{0!}$

$= 0.082085$

≈ 0.0821 (4dp)

c) we want $P(X \leq n) \geqslant 0.90$ where n = number of failures.

refer to Table 2 on Page 10 of data booklet ....

for $\lambda = 2 \cdot 5$ $\quad P(X \leq 4) = 0.8912 < 0.90$

$P(X \leq 5) = 0.9580 > 0.90$

$\therefore$ minimum number of spares $= 5$.

10. $X \sim N(\mu, 2.9^2)$

so $\bar{x} \pm Z_{0.95}\sqrt{\frac{2.9^2}{n}} < \bar{x} \pm 0.7$ ← from $\frac{1.4}{2}$ where '<' means "narrower interval"

$$Z_{0.95}\sqrt{\frac{2.9^2}{n}} < 0.7$$

$$\sqrt{\frac{2.9^2}{n}} < \frac{0.7}{Z_{0.95}}$$

$$\frac{2.9^2}{n} < \left(\frac{0.7}{Z_{0.95}}\right)^2$$

$$\frac{n}{2.9^2} > \left(\frac{Z_{0.95}}{0.7}\right)^2$$

$$n > 2.9^2 \times \left(\frac{Z_{0.95}}{0.7}\right)^2$$

$$n > 2.9^2 \times \left(\frac{1.64485}{0.7}\right)^2$$

$$n > 46.436$$

so minimum sample size $\underline{\underline{= 47}}$ (as n increases, interval narrows).

11. a) i) discrete data : age, pulse rate

ii) test on medians would likely be a Mann-Whitney test

⇒ assumption would be that the heights of the two groups would be distributed with same shape and spread.

b) χ^2 test for association requires categorical data

categorical data : activity level, smoker

c) Pearson's product moment correlation coefficient, to help determine if there is a linear association.

12. a) jar mass, $J \sim N(69, 6)$

honey mass, $H \sim N(453, 16)$

T = total mass of 48 jars of honey

$T = J_1 + \dots + J_{48} + H_1 + \dots + H_{48}$

$E(T) = 48E(J) + 48E(H)$

$= 48 \times 69 + 48 \times 453$

$= 25056.$

$V(T) = V(J_1 + \dots J_{48}) + V(H_1 + \dots H_{48})$

$= V(J_1) + \dots V(J_{48}) + V(H_1) + \dots V(H_{48})$ all r.v.'s independent

$= 48 \times V(J) + 48 \times V(H)$

$= 48 \times 6 + 48 \times 16$

$= 1056.$

so $T \sim N(25056, 1056)$

$P(T > 25000) = P\left(Z > \frac{25000 - 25056}{\sqrt{1056}}\right)$

$= P(Z > -1.72328)$

$= 0.957581$ from normcdf$(-1.72328, 9E99)$

$= 0.9576.$

assumption required: masses of each jar and each jar's honey content are all independent of one another (to support the calulation of $V(T)$).

12 b) $\bar{x} = 527.5$

$n = 10$

let X = mass of jar and honey

$X \sim N(\mu, 5^2)$

$H_0 : \mu = 522$

$H_1 : \mu \neq 522$

$\alpha = 1\%$

two tailed test

Assume H_0 to be true

so $X \sim N(522, 5^2)$

so $\bar{X} \sim N\left(522, \frac{5^2}{10}\right)$ where $\bar{X}$ = mean of sample of size 10.

$$\frac{\bar{X} - 522}{\sqrt{5^2/10}} \sim N(0, 1^2)$$

test statistic, $z = \frac{\bar{x} - 522}{\sqrt{5^2/10}}$

$$= \frac{527.5 - 522}{\sqrt{5^2/10}}$$

$$= 3.47851$$

p-value $= 2 \times P(Z > 3.47851)$

$= 2 \times 0.000252$

$= 0.000504$

< 0.01

Hence we reject H_0

We have evidence to suggest that the mean mass of jars of honey from the beekeeper is not equal to 522g.

Assumption required : the standard deviation of the mass of honey jars in the sample remains at 5g. (to be aligned with the Committee's model)

13. $n = 5$, per hour.

historical $\mu = 102g$.

$\sigma = 0.13$

a) $X =$ sugar content per litre bottle

$X \sim N(102, 0.13^2)$

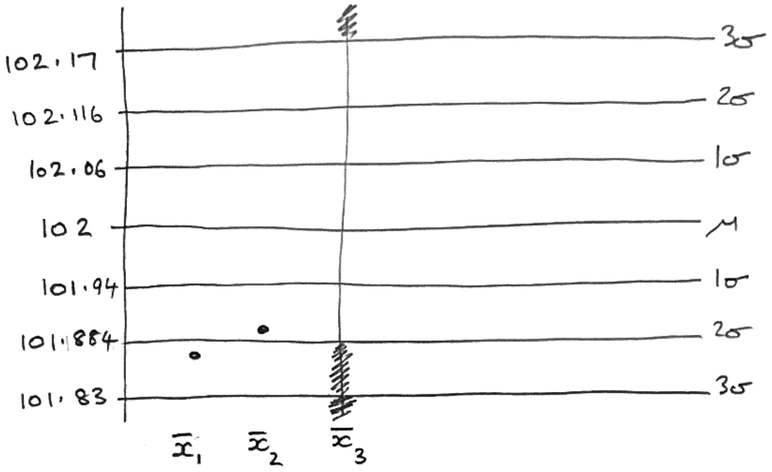
$\bar{X} \sim N(102, \frac{0.13^2}{5})$ where $\bar{X} =$ mean sugar content, from sample of size 5.

so 2σ limits are $102 \pm 2\sqrt{\frac{0.13^2}{5}}$

$= (101.884, 102.116)$

b) $\bar{x}_1 = 101.86$

$\bar{x}_2 = 101.89$.



3rd sample cannot be less that 101.884, else 2 out of 3 results are beyond same 2σ limit.

It must not lie beyond the upper 3σ limit

$\therefore \quad 101.884 < \bar{x}_3 < 102.17$ for the process to remain in control.

14. a) i) the sample mean has an approximate normal distribution.

ii) distribution mean is equal to the population mean

distibution variance is equal to the population variance divided by the sample size.

b) i) birth weights are already known to be normally distributed, so the distribution is not unkown.

ii) it was not a random sample, but rather a non-random sample (convenience sample) as they used the first 25 babies that were delivered at their nearest matemity hospital.