2026 AH Statistics – Paper 1

1 a) H_0: there is NO association between county squad and number of times they represented their county

H_1: there is an association between county squad and number of times they represented their county

b) expected frequency $= \dfrac{\text{row total} \times \text{column total}}{132}$

$= \dfrac{(7+12+8+4+1) \times (8+6+3+6)}{132}$

$= \dfrac{32 \times 23}{132}$

$= 5.5757$

≈ 5.58 (2dp)

c) currently 8 of the 20 expected frequences are < 5
this is 40% of the expected frequencer, when there should be no more than 20%.

d) we have $X^2 = 13.942$
if we have 5% significance level, then $\chi^2_{9,0.95} = 16.919$
as $13.942 < 16.919$ we do not reject H_0
we donot have sufficient evidence to suggest that there is an association between county squad and number of times players represented their county.

e) i) there is a positive correlation between mass and height

ii) there is a greater gain in mass for the back players, for a given increase in height, compared to the forward players.

f) $H_0: \beta = 0$ two-tailed hypothesis test
$H_1: \beta \neq 0$

as p-value $< 0.00001, < 0.05$, we reject H_0
we have evidence to suggest that the model is useful for predicting a forward player's mass from their height.

1 g) fitted value $= \frac{101.2221 + 132.0617}{2}$

$= 116.642$.

or mass $= 40.75685 + 0.37567 \times 202$

$= 116.642$.

h) Prediction Interval: 95% of the time we would expect the mass of a single forward player of height 202cm to lie between 101.2221 kg and 132.0617 kg.

Confidence Interval: 95% of the time we would expect the mean mass of forward players of height 202cm to lie between 114.9216 kg and 118.3622 kg.

i) Figure 1 only contains data on 'back' players of height < 200 cm.
Hence the linear model is best suited for those heights
As $202 > 200$, we would be extrapolating the model, and therefore it would not be suitable to use to estimate the player's mass

2 a) the garages that did not appear on the internet would not be found online the researcher could have phoned them and asked them, as they had access to this alternative contact information.

b) a confidence interval is designed to capture a mean value, not a maximum value.

c) a two sample z-test would require knowledge of the population standard deviations of prices for each of the two groups of garages.

d) assumption: that the standard deviations of the populations from which the samples were taken were numerically equal.

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$= \frac{(18-1)\times 8.627261^2 + (34-1)\times 8.099399}{18+34-2}$$

$$= 68.6022$$

$$s = 8.28265$$

e) $m = 18$, $n = 34$, $W = 404$

So $E(W) = \frac{1}{2} \times 18 \times (18+34+1)$ $\quad V(W) = \frac{1}{12} \times 18 \times 34 \times (18+34+1)$

$= 477$ $\quad = 2703$

$$p\text{-value} = P(W \leq 404)$$

$$\approx P\left(Z \leq \frac{404.5 - 477}{\sqrt{2703}}\right) \quad \text{using continuity correction}$$

$$= P(Z \leq -1.39449)$$

$$= 0.081585$$

if the rounded z-value of just -1.39 is used, then $P(Z < -1.39) = 0.0823$.

2f) if we have a 10% level of significance, then the p-value of $0.0823 < 0.1$, but the p-value of $0.1019 > 0.1$. Hence they would lead to different conclusions.

2g) the Mann-Whitney test references medians, not a mean.
the conclusion is phrased as a two tailed test, but it was a one-tailed test that was performed. So it should be 'less than' not 'different to'

2h) i) the distributions did not look normally distributed, as they are not symmetrical particularly noticeable in the boxplot for the independent garages with very different length 'whiskers' and a median not equidistance between the lower and upper quartiles.

ii) the 95% confidence t-intervals would not be valid
the two-sample t-test would not be valid

2 i) all national chains were selected, rather than a sample of national chains. Hence we had the full population of national chains, but only a sample of independent garages.
We require both groups to have been randomly sampled for valid statistical inferences to be made

2026 AH Statistics – Paper 2.

1 a) lower fence $= LQ - 1.5 \times IQR$
$= 19.3 - 1.5 \times 1.7$
$= 16.75$

upper fence $= UQ + 1.5 \times IQR$
$= 21.0 + 1.5 \times 1.7$
$= 23.55$

as minimum $16.3 < 16.75$, there will be a possible outlier below the lower fence
as maximum $24.1 > 23.55$, there will be a possible outlier above the upper fence.

Hence there will be atleast two possible outliers.

b) $H_0 : \mu_W = \mu_A$

$H_1 : \mu_W \neq \mu_A$

two tailed. $\alpha = 0.1\% = 0.001$

assume H_0 to be true.

$\bar{x}_W = 20.2$ $\quad$ $\bar{x}_A = 19.8$

$s_W = 1.38$ $\quad$ $s_A = 1.15$

$n_W = 216$ $\quad$ $n_A = 87$

test statistic, $z = \dfrac{20.2 - 19.8}{\sqrt{\frac{1.38^2}{216} + \frac{1.15^2}{87}}}$

$z = 2.58103$

p-value $= 2 \times P(Z > 2.58103)$
$= 2 \times 0.004925$
$= 0.009851$
> 0.001

so we do not reject H_0

we do not have sufficient evidence that the mean handspans of white female pianists are different to those of asian female pianists

c) we assume that the two samples of pianists were randomly sampled from all such pianists.

2. $\Sigma x = 275$ $\Sigma y = 2468$ $S_{xx} = 1958.375$ $S_{xy} = -5029.5$ $S_{yy} = 18734.4$ $n = 40$

a) i) $b = \frac{S_{xy}}{S_{xx}} = -2.5682$

$\bar{x} = \frac{\Sigma x}{40} = 6.875$

$\bar{y} = \frac{\Sigma y}{40} = 61.7$

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

$= 79.3564$

hence regression line is $y = 79.3564 - 2.5682x$.

ii) fitted value, $y(10) = 79.3564 - 2.5682 \times 10$

$= 53.6744$

residual = observe - fitted

$= 58 - 53.6744$

$= 4.32563$.

b) The regression line is designed to calculate y from knowing x.

To calculate x from knowing y would require constructing a different regression line, formed by switching around all the x's and y's in the above calculation.

3. no. tickets sold = n.

X = number of the winning ticket

a) $X \sim U(n)$ a discrete uniform distribution.

b) $V(X) = \frac{n^2-1}{12}$

$2552 = \frac{n^2-1}{12}$

$n^2 = 1 + 12 \times 2552$

$n = \pm\sqrt{30625}$

as $n > 0$, $n = 175$.

c) Y = value of donation

$Y = 20 \times X$.

$V(Y) = V(20X)$

$= 20^2 V(X)$

$= 20^2 \times 2552$

$= 1020800$

$SD(Y) = 1010.35$.

4. H_0: population median difference = 0.

H_1: population median difference > 0

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

Assume H_0 to be true.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Difference :	8	−24	−10	−18	12	−8	−10
Rank :	1·5	7	3·5	6	5	1·5	3·5

$W_+ = 1{\cdot}5 + 5 = 6{\cdot}5$

$W_- = 7 + 3{\cdot}5 + 6 + 1{\cdot}5 + 3{\cdot}5 = 21{\cdot}5$

$W_+ + W_- = 28 = \frac{1}{2} \times 7 \times 8$ ✓

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

$= 6{\cdot}5$

5% critical value for $n = 7$ is 3

as $6{\cdot}5 > 3$, we do not reject H_0 at the 5% level

we do not have sufficient evidence to suggest that the population median difference is greater than zero.

Hence, we do not have evidence that, on average, the number of deliberate fires is more than 6 times the number of accidental fires.

5. a) target proportion = 0.15

sample size = 50

$$3 \text{ sigma limits} = 0.15 \pm 3\sqrt{\frac{0.15 \times 0.85}{50}}$$

$$= 0.301493, -0.001493$$

$$\approx 0.3015, -0.0015$$

the lower control limit is not required, as not only is it negative (and the limits should be between 0 and 1 inclusive), but also low proportions of rejected strawberries are not a concern.

b) between weeks 8 to 16 inclusive, there were more than 8 consecutive points falling on the same side of the centre line.

The farm would be pleased as it means that they are consistently having low proportions of rejected strawberries.

c) $p = 0.08$

if X = no. of rejected strawberries

then $X \sim B(n, 0.08)$

for a normal approximation, we need $np > 5$ and $nq > 5$

$n \times 0.08 > 5$ $\quad$ $n \times 0.92 > 5$

$n > \frac{5}{0.08}$ $\quad$ $n > \frac{5}{0.92}$

$n > 62.5$ $\quad$ $n > 5.43$

Hence, the smallest sample size would be 63.

6. a) we can see the shape of the distribution of the masses in the sample, which can provide an insight into the shape of the population distribution from which the sample was taken. We might also be able to spot any possible outliers.

b) i) $n = 15 \quad \bar{x} = 67 \quad s = 4$

X = mass of ice cream

we would need to assume that the distribution of X was normal, in order to use a t-test, as the sample size is only 15.

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

$H_0 : \mu = 70$

$H_1 : \mu < 70$

one tail test, $\alpha = 1\%$.

Assume H_0 is true

so $X \sim N(70, \sigma^2)$

so $\bar{X} \sim N\left(70, \frac{\sigma^2}{15}\right)$ where $\bar{X}$ = mean mass of ice cream, from sample of size 15

$$\frac{\bar{X} - 70}{\sqrt{\frac{\sigma^2}{15}}} \sim N(0, 1^2)$$

as we estimate σ^2 with s^2, then $N(0,1^2) \rightarrow t_{14}$

so $$\frac{\bar{X} - 70}{\sqrt{\frac{s^2}{15}}} \sim t_{14}$$

test statistic, $t = \dfrac{67 - 70}{\sqrt{\frac{4^2}{15}}} = -2.90474$

p-value $= P(t_{14} < -2.90474)$

$= 0.005767$

< 0.01

we reject H_0 at the 1% level

so we have evidence to suggest that the mean mass of ice cream is less than 70g.

7. a) X = no. of times kettle is boiled per hour.

$X \sim Po(2.5)$

Y = no. of times kettle is boiled in 3 hours

$Y \sim Po(7.5)$

$$P(Y=8) = \frac{e^{-7.5} \times 7.5^8}{8!}$$

$$= 0.137329.$$

b) The normal approximation will not be a good approximation as $7.5 < 10$.

c) W = no. times kettle boild in 8 hours

$W \sim Po(20)$

$$P(W>15) \approx P\left(Z > \frac{15.5-20}{\sqrt{20}}\right)$$ using continuity correction and normal approximation, as $20 > 10$.

$$= P(Z > -1.00623)$$

$$= 0.842848$$

As $0.842848 > 0.80$, the water heater is viable.

8.

14 month olds : $\frac{6}{31} \Rightarrow p_{14} = \frac{6}{31}$

18 month olds : $\frac{38}{55} \Rightarrow p_{18} = \frac{38}{55}$

pooled, $p = \frac{6+38}{31+55} = \frac{44}{86}$

$H_0 : p_{18} = p_{14}$

$H_1 : p_{18} > p_{14}$

one tail test, $\alpha = 1\%$.

Assume H_0 to be true

assume that the children's choice of what they offered was independent of the other children's choices, and of the researchers they were with.

$$\text{test statistic, } z = \frac{p_{18} - p_{14}}{\sqrt{pq\left(\frac{1}{n_{18}} + \frac{1}{n_{14}}\right)}}$$

$$= \frac{\frac{38}{55} - \frac{6}{31}}{\sqrt{\frac{44}{86} \cdot \frac{42}{86} \cdot \left(\frac{1}{55} + \frac{1}{31}\right)}}$$

$$= 4.43029$$

p-value $= P(Z > 4.43029)$

$= 0.000005$

$\ll 0.01$

so we reject H_0 at the 1% level

we have evidence to suggest that the proportion of 18-month olds who are more aware of other peoples' preferences is greater than the proportion of 14-month olds.

9.

24 cans in a multipack.

mass empty can, $C_i \sim N(18, 2)$

mass packaging, $K \sim N(196, 1)$

mass drink, $D_i \sim N(\mu, \sigma^2)$ assuming mass of drink is normally distributed

we shall also assume that all random variables are independent of one another

ie. mass of drink independent of mass of can and of mass of packaging, and that mass of can independent of mass of packaging

total mass of 24 filled cans with packaging, $T \sim N(8884, 121)$

$$\text{So } T = \sum_{i=1}^{24} C_i + \sum_{i=1}^{24} D_i + K$$

$$\text{So } E(T) = E(C_1 + \ldots + C_{24} + D_1 + \ldots + D_{24} + K)$$

$$E(T) = 24E(C_i) + 24E(D_i) + K.$$

$$8884 = 24 \times 18 + 24 \times \mu + 196.$$

$$\mu = \frac{8884 - 24 \times 18 - 196}{24}$$

$$\mu = 344.$$

$$\text{and } V(T) = V(C_1 + \ldots + C_{24} + D_1 + \ldots + D_{24} + K)$$

$$V(T) = 24V(C_i) + 24V(D_i) + K$$

$$121 = 24 \times 2 + 24\sigma^2 + 1$$

$$\sigma^2 = \frac{121 - 24 \times 2 - 1}{24}$$

$$\sigma^2 = 3.$$

So mass of drink in single can, $D_i \sim N(344, 3)$.

10.

the training provider would likely submit 20 names of people who had done well with their training, thereby being biased towards those who had a better experience. As a result, the training course may be rated higher in quality than it really was.

11.

a) let X = number of buses late in one week

$X \sim B(100, 0.043)$

$$P(X \geq 1) = 1 - P(X=0)$$
$$= 1 - {}^{100}C_0 (0.043)^0 (0.957)^{100}$$
$$= 1 - 0.012337$$
$$= 0.987663.$$

b) i) $$P(\text{train} \mid \text{late}) = \frac{P(\text{train} \cap \text{late})}{P(\text{late})}$$
$$= \frac{P(\text{train})P(\text{late} \mid \text{train})}{P(\text{train})P(\text{late} \mid \text{train}) + P(\text{bus})P(\text{late} \mid \text{bus})}$$
$$= \frac{\frac{90}{190} \times 0.09}{\frac{90}{190} \times 0.09 + \frac{100}{190} \times 0.043}$$
$$= 0.653226.$$

ii) we assume that the probability of taking bus or train is in proportion to their frequencies, and that no other factors sway them to take one mode of transport over the other.

c) 6 out of 75 trains were late

i) 95% C.I for proportion $= \frac{6}{75} \pm z_{0.975} \times \sqrt{\frac{\frac{6}{75} \times \frac{69}{75}}{75}}$
$$= \frac{6}{75} \pm 1.95996 \times 0.031326$$
$$= (0.018602, 0.141398)$$
$$\approx (0.0186, 0.1414)$$

ii) previously the proportion that were late was 0.09, which lies within this confidence interval, so the attempt to change the number of trains that were late seems not to have had any effect

iii) we need to check that $np > 5$ and $nq > 5$ in order to support the normal approximation to the binomial. We have $n \times p = 75 \times \frac{6}{75} = 6 > 5$

$n \times q = 75 \times \frac{69}{75} = 69 > 5.$

The interval is only approximate due to the normal approximation.

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

a) $P(\mu - 0.2\sigma < X < \mu + 0.2\sigma)$

$= P(-0.2\sigma < X - \mu < 0.2\sigma)$

$= P(-0.2 < \frac{X-\mu}{\sigma} < 0.2)$

$= P(-0.2 < Z < 0.2)$ where $Z \sim N(0, 1^2)$

$= 0.158519$

b) $n = 36$

i) $\bar{X} \sim N(\mu, \frac{\sigma^2}{36})$ if $\bar{X}$ were standardised, $Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{36}}}$

$= \frac{\bar{X}-\mu}{\frac{\sigma}{6}}$

$= \frac{6(\bar{X}-\mu)}{\sigma}$

ii) $P(\mu - 0.2\sigma < \bar{X} < \mu + 0.2\sigma)$

$= P(-0.2 < \frac{\bar{X}-\mu}{\sigma} < 0.2)$

$= P(-0.2 \times 6 < 6(\frac{\bar{X}-\mu}{\sigma}) < 0.2 \times 6)$

$= P(-1.2 < Z < 1.2)$

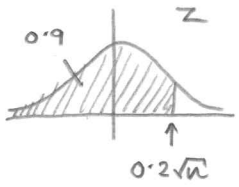
$= 0.769861$

c) if $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ then $P(\bar{X} < \mu + 0.2\sigma) \leq 0.9$

$P(\frac{\bar{X}-\mu}{\sigma} < 0.2) \leq 0.9$

$P(\sqrt{n}(\frac{\bar{X}-\mu}{\sigma}) < 0.2\sqrt{n}) \leq 0.9$

$P(Z < 0.2\sqrt{n}) \leq 0.9$



$0.2\sqrt{n} \leq \Phi^{-1}(0.9)$

$0.2\sqrt{n} \leq 1.28155$

$\sqrt{n} \leq 6.40776$

$n \leq 41.0594$

so we would need $n \leq 41$