2022 AH Statistics – Paper 1

1 a) $IQR = Q3 - Q1$
$= 13 - 5$
$= 8$

upper fence $= Q3 + 1.5\,IQR$
$= 13 + 1.5 \times 8$
$= 13 + 12$
$= 25$

lower fence $= Q1 - 1.5 \times IQR$
$= 5 - 1.5 \times 8$
$= 5 - 12$
$= -7$

As the maximum number of weeks for 1980's is 21 which is less than 25, and the minimum is 2 with is more than -7, there are no outliers.

b) each year (1980, 1981, 1982, ..., 1989) is a strata

a simple random sample of 2% of all of the songs in each of these years would have been taken

Combining all of these samples together would give the sample for the decade.

c) location: both the means and the medians are increasing as time passes.
this suggests that songs are in the charts for longer periods of time.

spread: the standard deviations are increasing as time passes
this suggests that there is less consistency in how long a song is in the Top 40 charts.

sample size: the sample sizes are decreasing as time passes
this suggests that fewer songs are in the charts

all in all, as time passes, we have fewer songs staying in the charts for longer, but that it is less predictable how long they will be in the charts for.

d) i) let X = number of weeks a song is in the charts in the 1990's.

$E(X) = \mu$
$V(X) = \sigma^2$

we shall assume that X is normally distributed

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

so $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ where $\bar{X}$ = sample mean number of weeks.

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

we estimate σ^2 with s^2, so we use t_{32}

so $\frac{\bar{X} - \bar{x}}{\sqrt{\frac{s^2}{33}}} \sim t_{32}$

so 95% CI for mean is $\bar{x} \pm t_{32, 0.975} \sqrt{\frac{s^2}{33}}$

$= 12.700 \pm 2.03693 \sqrt{\frac{7.038^2}{33}}$

$= (10.2044, 15.1956)$

$\approx (10.20, 15.20)$

ii) this interval would be expected to capture the true value of the population mean number of weeks roughly 95% of the time, if the sampling process was repeated many times.

e) as p-value $= 0.4042 > 0.05$, we do not have evidence to reject H_0

So, we would conclude that the mean number of weeks that a song is in the charts in the 2010's is the same as that for the 2000s.

f) in 2000s, we have that $s_{n-1} = 8.991$
in 2010s, we have that $s_{n-1} = 12.713$
I would challenge the validity of the assumption as 12.713 is quite different in value to 8.991 (in fact, it's 41% larger!)

2. a) sampling method : convenience sampling

disadvantage : may not be representative of population, due to non responses.

b) offering an incentive, such as entering their name in a prize draw, might generate a higher response rate.

c) the 'teachers' group is most affected as there were only 60 of them, compared to the pupils groups which were all about four times larger. Hence one teacher accounts for a greater proportion than one pupil

d) i) expected wearers $= \dfrac{79+66+64+55+54+39+44}{7}$

$= 57.2857...$

$\approx \underline{\underline{57.3}}.$

ii) the introduction was centred on pupils, and so the group of teachers should not have been included.

e) $H_0 : p_{S1} = p_{S5}$

$H_1 : p_{S1} \neq p_{S5}$

so $\hat{p}_{S1} = \dfrac{79}{224}$, $\hat{p}_{S5} = \dfrac{54}{206}$

$n_{S1} = 224$ $n_{S5} = 206$

pooled $p = \dfrac{79+54}{224+206} = \dfrac{133}{430}$

test statistic, $z = \dfrac{\hat{p}_{S1} - \hat{p}_{S5}}{\sqrt{pq\left(\frac{1}{n_{S1}} + \frac{1}{n_{S5}}\right)}}$

$= \dfrac{\frac{79}{224} - \frac{54}{206}}{\sqrt{\frac{133}{430} \cdot \frac{297}{430}\left(\frac{1}{224} + \frac{1}{206}\right)}}$

$= 2.02928$

$\approx \underline{\underline{2.03}}.$

e) (cont) p-value $= 2 \times P(Z > 2.03)$ from normCdf(2.03, 9E99)

$= 2 \times 0.021215$

$= 0.04243$

≈ 0.042.

f) the choice of secondary school was not randomly selected from the wider population of secondary schools.

Furthermore, only one school does not encompass the full diversity of young people.

2022 Alt Statistics – Paper 2.

1. H_0: no association between infection and sex

H_1: there is an association between infection and sex.

assume H_0 to be true

$\alpha = 1\%$, one-tailed test

observed	male	female	
infected	76	129	205
not infected	399	332	731
	475	461	936

expected	male	female
infected	104.03	100.967
not infected	370.967	360.033

$\frac{475 \times 731}{936}$ etc.

now $$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$
$$= 19.6383$$

all expected frequencies are > 5 ☺

so no merging of rows or columns required.

$$df = (\text{rows} - 1) \times (\text{cols} - 1)$$
$$= (2-1) \times (2-1)$$
$$= 1$$

$$p\text{-value} = P(\chi^2_1 > 19.6383)$$
$$= 0.000009$$
$$\ll 0.01$$

as p-value $0.000009 < 0.01$ we have evidence to reject H_0 and conclude that there is an association between the prevalence of infection and sex of the fish.

2.

X = no. bees leaving hive A per minute $\quad X \sim Po(2.3)$

Y = no. bees leaving hive B per minute $\quad Y \sim Po(1.7)$

a) $\underline{\underline{X \sim Po(2.3)}}$

b) $P(X=0) = \dfrac{e^{-2.3} \times 2.3^0}{0!}$

$= 0.100259$

$\underline{\underline{= 0.1003}}$ (4dp)

c) $P(X=2 \text{ and } Y=2) = P(X=2) \times P(Y=2)$ as X and Y are independent

$= \dfrac{e^{-2.3} \times 2.3^2}{2!} \times \dfrac{e^{-1.7} \times 1.7^2}{2!}$

$= 0.265185 \times 0.263978$

$= 0.070003$

$\underline{\underline{\approx 0.0700}}$

d) let $W = X + Y$

$W \sim Po(2.3 + 1.7)$

$W \sim Po(4)$

$P(W > 5) = 1 - P(W \leq 5)$

$= 1 - 0.78513$ from poissCdf(4,5)

$= 0.21487$

$\underline{\underline{\approx 0.2149}}$.

3.

	0	0	2	4	4
0	·	0	2	4	4
0	0	·	2	4	4
2	2	2	·	6	6
4	4	4	6	·	8
4	4	4	6	8	·

T = total of two cards, without replacement

So

t	0	2	4	6	8
$P(T=t)$	$\frac{2}{20}$	$\frac{4}{20}$	$\frac{8}{20}$	$\frac{4}{20}$	$\frac{2}{20}$
$=$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

$$E(T) = \sum t P(T=t)$$

$$= 0\times\frac{1}{10} + 2\times\frac{2}{10} + 4\times\frac{4}{10} + 6\times\frac{2}{10} + 8\times\frac{1}{10}$$

$$= \frac{1}{10}(0+4+16+12+8)$$

$$= \frac{40}{10}$$

$$= 4.$$

$$E(T^2) = \sum t^2 P(T=t)$$

$$= \frac{1}{10}(0^2\times 1 + 2^2\times 2 + 4^2\times 4 + 6^2\times 2 + 8^2\times 1)$$

$$= \frac{1}{10}(0 + 8 + 64 + 72 + 64)$$

$$= \frac{208}{10}$$

$$= 20.8$$

$$V(T) = E(T^2) - E^2(T)$$

$$= 20.8 - 4^2$$

$$= 4.8.$$

4. a) let X = no. tests that are passed

$X \sim B(104, 0.44)$

$P(X=52) = {}^{104}C_{52} \times 0.44^{52} \times 0.56^{52}$

$= 0.036713$ from binomPdf(104, 0.44, 52)

$\approx \underline{\underline{0.0367}}$

b) let Y = normal approximation to X

$Y \sim N(104 \times 0.44, 104 \times 0.44 \times 0.56)$

$Y \sim N(45.76, 25.6256)$

so $P(40 \leq X \leq 50) = P(39.5 \leq Y \leq 50.5)$ by continuity correction.

$= P\left(\frac{39.5 - 45.76}{\sqrt{25.6256}} \leq Z \leq \frac{50.5 - 45.76}{\sqrt{25.6256}}\right)$

$= P(-1.23662 \leq Z \leq 0.936357)$

$= 0.717342$ from normCdf(-1.23662, 0.936357)

$\approx \underline{\underline{0.7173}}$

5, a) the data is paired data, and a t-test for a difference in population means is for non-paired data.

b) we shall assume that the distribution of each of the sets of scores are normally distributed, so that their differences are also normally distributed.

French	67	83	71	59	49	89	42	55	77
German	64	82	71	62	42	85	39	50	75
F-G.	3	1	0	-3	7	4	3	5	2.

let $D = F - G$, so $D \sim N(\mu, \sigma^2)$

so $H_0 : \mu_{difference} = 0$

$H_1 : \mu_{difference} \neq 0$

assume H_0 to be true

$\alpha = 5\%$, two tailed test

$\bar{x}_D = \dfrac{3+1+\ldots+5+2}{9} = 2.44$, $n=9$, $s_{n-1} = 2.92024$

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

$\bar{D} \sim N(\mu, \frac{\sigma^2}{9})$

$$\frac{\bar{D} - \mu}{\sqrt{\frac{\sigma^2}{9}}} \sim N(0, 1^2)$$

we estimate σ^2 with s_{n-1}^2, so we use t_8 distribution

$$\frac{\bar{D} - \mu}{\sqrt{\frac{s_{n-1}^2}{9}}} \sim t_8$$

test statistic, $t = \dfrac{2.44 - 0}{\sqrt{\frac{2.92024^2}{9}}} = 2.51121$

p-value $= 2 \times P(t_8 > 2.51121)$

$= 2 \times 0.018151$

$= 0.036302$

< 0.05

so we have evidence to reject H_0, and conclude that the mean difference between French and German marks is non-zero

6. a) the residual plot should have $E(\varepsilon_i)=0$ and $V(\varepsilon_i)=\sigma^2$

This plot seems not to have the expected residual to be zero

and it also has a non-constant variance shown by the parabolic pattern of points.

b) we have $\sum x = 3740$, $n=85$ $\Rightarrow \bar{x} = \frac{3740}{85}$

$\sum w = 101.2529$, $n=85$ $\Rightarrow \bar{y} = \frac{101.2529}{85}$

so $b = \frac{S_{xw}}{S_{xx}} = \frac{-715.456}{51170} = -0.013982$

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

$= 1.80642$.

so $\underline{\underline{w = 1.80642 - 0.013982x}}$

in 1927, $x = 1927-1840 = 87$

so $w = 1.80642 - 0.013982 \times 87$

$= 0.589987$

$\Rightarrow \log_{10} y = 0.589987$

$\Rightarrow y = 10^{0.589987}$

$\Rightarrow y = 3.89034$

so Percentage of men with sideburns $= \underline{\underline{3.9\%}}$

7.

a) $X \sim B(n, p)$ so $E(X) = np$, $V(X) = npq$

$$E\left(\frac{X}{n}\right) = E\left(\frac{1}{n}X\right) = \frac{1}{n}E(X) = \frac{1}{n} \times np = p.$$

$$V\left(\frac{X}{n}\right) = V\left(\frac{1}{n}X\right) = \left(\frac{1}{n}\right)^2 V(X) = \frac{1}{n^2} \times npq = \frac{pq}{n}$$

b) let $\hat{p} = \frac{14}{50} = 0.28$

so if $X \sim B(50, p)$

then approx X to normal, $X \approx N(50p, 50pq)$

this is valid if $50p > 5$ and $50q > 5$

we estimate p, with $\hat{p} = 0.28$, so $50\hat{p} = 14 > 5$, $50\hat{q} = 36 > 5$ ✓

Hence normal approximation is valid.

so $\frac{X}{50}$ is proportion of successes, $\frac{X}{50} \approx N\left(\frac{50p}{50}, \frac{50pq}{50^2}\right)$

$$\frac{X}{50} \approx N\left(p, \frac{pq}{50}\right)$$

so 99% CI for p $= \hat{p} \pm z_{0.995}\sqrt{\frac{\hat{p}\hat{q}}{50}}$

$$= 0.28 \pm 2.57583\sqrt{\frac{0.28 \times 0.72}{50}}$$

$$= (0.11644, 0.44356)$$

$$\approx (0.1164, 0.4436)$$

8.

a) $P(\text{spin a 4 and then goldfish})$

$= P(\text{spin a 4}) \times P(\text{goldfish} \mid \text{card number 4})$

$= \frac{1}{5} \times \frac{5}{8}$

$= \frac{1}{8}.$

b) i) $P(\text{shark} \mid \text{spin a 1}) = \frac{2}{5}$

ii) $P(\text{lose game}) = P(\text{reveal a shark})$

$= \sum_{i=1}^{5} P(\text{spin an } i \text{ and pick a shark})$

$= \sum_{i=1}^{5} P(\text{spin an } i)\, P(\text{shark} \mid \text{card number } i)$

$= \frac{1}{5} \times \frac{2}{5} + \frac{1}{5} \times \frac{0}{7} + \frac{1}{5} \times \frac{0}{6} + \frac{1}{5} \times \frac{3}{8} + \frac{1}{5} \times \frac{0}{6}$

$= \frac{2}{25} + 0 + 0 + \frac{3}{40} + 0$

$= \frac{31}{200}$

iii) $P(\text{spin a 1} \mid \text{lost game}) = \dfrac{P(\text{spin a 1 and lost game})}{P(\text{lost game})}$

$= \dfrac{2/25}{31/200}$

$= \frac{16}{31}.$

9 a) the distribution of the sample mean is approximately normal, regardless of the population distribution.

so if X has $E(X) = \mu$ and $V(X) = \sigma^2$ (and X is not normally distributed)

then $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$ where n is the sample size

b) let X = width of batten so $E(X) = \mu$, $V(X) = \sigma^2$

$n = 45$

$\bar{x} = 52.6$ $s^2 = 103.25$

$H_0 : \mu = 50$

$H_1 : \mu > 50$

assume H_0 to be true

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

we have been told to do a z-test, which requires knowing σ^2

we shall assume that σ^2 is best approximated by s^2, given that the sample is large.

this is the "further assumption".

so $E(X) = 50$, $V(X) = 103.25$

by CLT, $\bar{X} \approx N(50, \frac{103.25}{45})$

$$\frac{\bar{X} - 50}{\sqrt{\frac{103.25}{45}}} \approx N(0, 1^2)$$

test statistic, $z = \dfrac{52.6 - 50}{\sqrt{\frac{103.25}{45}}} = 1.71646$

p-value $= P(Z > 1.71646)$

$= 0.043039$ from normCdf(1.71646, 9E99)

< 0.05

so we have evidence to reject H_0 and conclude that the mean batten width is greater than 50mm.

10. a) $H_0: \rho = 0$

$H_1: \rho \neq 0$.

assume H_0 to be true

$\alpha = 5\%$, two tail test.

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

here $n = 6$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{46.29}{\sqrt{278.61 \times 10.95}} = 0.838073$$

So $t = 3.07235$.

we have $df = 4$

So P-value $= 2 \times P(t_4 > 3.07235)$

$= 2 \times 0.018604$

$= 0.037208$

< 0.05

So we have evidence to reject H_0 and conclude that exposure index and number of death are linearly associated.

underlying assumption to this test: number of deaths in each town is normally distributed.

and/or: data for each town is independent of all other towns.

b) linear correlation does not imply causation.

There may be another reason that explains the deaths.

11. a) i)

adults		juveniles
3	0	7 9
5 3	1	1 7 3 9 1
9 3 7 2	2	9 8
5 5	3	
0	4	

$\Rightarrow$ ordered

adults		juveniles
3	0	7 9
5 3	1	1 1 3 7 9
9 7 3 2	2	8 9
5 5	3	
0	4	

where $2|8 = 2.8$

Diagram shows that the adults appear to have more widely spread and longer reaction times.

ii) so $m = n = 10$.

H_0 : $\text{median}_{\text{juvenile}} = \text{median}_{\text{adult}}$

H_1 : $\text{median}_{\text{juvenile}} \neq \text{median}_{\text{adult}}$

assume H_0 to be true

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

$W_{\text{juvenile}} = 89$

from tables, we have $P(W \leq 78) = 0.025$

so, as $89 > 78$ we are not in the critical region. So we do not have evidence to reject H_0 and conclude that there is not a difference between the median reaction times of adult and juvenile foxes.

b) if $A \sim N(2.5, 0.5)$
$J \sim N(2.0, 0.3)$

$$P(A > J) = P(A - J > 0)$$
$$= P(D > 0) \quad \text{where } D = A - J, \; D \sim N(0.5, 0.8)$$
$$= P\left(Z > \frac{0 - 0.5}{\sqrt{0.8}}\right)$$
$$= P(Z > -0.559017)$$
$$= 0.711925 \quad \text{from normCdf}(-0.559, 9E99)$$
$$\simeq \underline{\underline{0.7119}}.$$