2023 AH Statistics – Paper 1

a) bakery items have a greater median fat content than non-bakery items.

the two types of items have a similar spread as their interquartile ranges are very similar

b) outliers are data points that are above the upper fence, or below the lower fence. (upper fence = $Q3 + 1.5 \times IQR$
lower fence = $Q1 - 1.5 \times IQR$)

c) Bakery – calories has $IQR = Q3 - Q1$
$= 430 - 320$
$= 110$

lower fence $= 320 - 1.5 \times 110$
$= 320 - 165$
$= 155$

The two data points are below 150, which is below 155, and so they are outliers.

d) The bakery items may have had their calorie content miscalculated or mislabelled, or they could have been recorded incorrectly by the data analysts.

e) t-tests require estimating population standard deviation from sample standard deviation. Hence the sample standard deviations were missing from output 1.

f) Output 2: H_0: mean fat in bakery items = mean fat in non-bakery items
i.e. $\mu_{bakery\,fat} = \mu_{non\text{-}bakery\,fat}$

Output 3: H_0: mean calories in bakery items = mean calories in non-bakery items.
i.e. $\mu_{bakery\,calories} = \mu_{non\text{-}bakery\,calories}$.

g) p-value $= 2 \times P(t_{75} > 1.0496)$
$= 2 \times P(Z > 1.05)$
$= 2 \times (1 - 0.8531)$ from table 3
$= 2 \times 0.1469$
$= 0.2938$

h) p-value = 0.006931

< 0.05

So we have evidence to reject H_0

So mean bakery calories are not equal to mean non-bakery calories.

So there would be an impact on your mean calorie impact.

i) we have performed cluster sampling

a single cluster (coffee shop chain) was randomly selected from all coffee shop chains in the UK.

all coffee shop chains must have been listed, given a unique number from 1 to n, and then one of these numbers was randomly selected.

simple random sampling was then conducted on the bakery and non-bakery items in the selected coffee shop chain.

Hence, this is two-stage cluster sampling.

<u>alternatively</u> (if we only consider the selection of the coffee shop chain, and not the bakery goods)

we have performed simple random sampling

all coffee shop chains were listed and given a number

one number was randomly selected to give the single coffee shop chain, which was then used to source the bakery/non-bakery items.

2 a) i) the scatterplot shows a positively correlated linear relationship between crop yield and crop density.

ii) a residual = observed value - fitted value

where fitted value is calculated from the equation of the regression line.

a residual measures the error in the observed data that is not explained by the linear model.

iii) the residuals do have a mean of 0 ($E(\varepsilon_i)=0$) and they do have a constant variance ($V(\varepsilon_i)=\sigma^2$) but there is a clear "U-shaped" pattern that is not randomly scattered.

this non-random scatter indicates that a transformation of data is required to obtain a better fitting model.

b) i) we want $\sqrt{y} = a + bx$ where $b = \frac{S_{x\sqrt{y}}}{S_{xx}}$ and $a = \overline{\sqrt{y}} - b\bar{x}$

so $b = \frac{78.8165}{45.5} = 1.73223$

$a = \frac{370.2569}{13} - 1.73223 \times \frac{65}{13}$

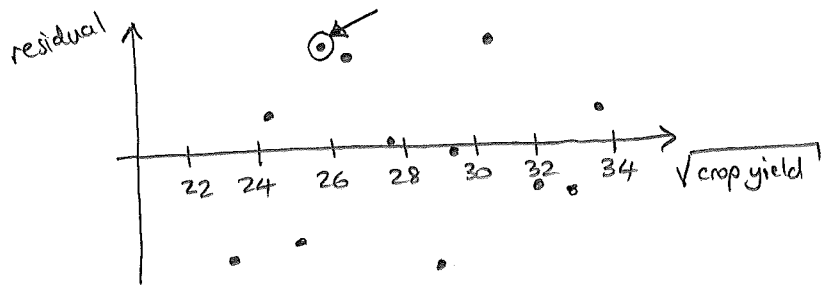
$a = 19.8201$

$\sqrt{y} = 19.8201 + 1.73223x$

ii) if $x = 3.5$, then $\sqrt{y} = 19.8201 + 1.73223 \times 3.5$
$= 25.883$

observed $\sqrt{\text{crop yield}}$ when $x = 3.5$ was 26.0

$\therefore$ residual = observed - fitted
$= 26 - 25.883$
$= 0.117046$



c) i) the original data set covered crop densities between 2 and 8. So using values greater than 8 would be extrapolating the model and the results may not be reliable.

ii) the aim was to find the crop density that would maximise the crop yield.
however a linear model is always increasing, and will not have a "maximum turning point"
Hence no maximum can be found by this method.

d) The linear model for $\sqrt{\text{crop yield}}$ and crop density is a good fit for crop densities between 2 and 8 plants/m^2.

2023 AH Statistics – Paper 2.

1. X = heights of boys aged 5 years

$X \sim N(109, 7^2)$

a) $P(X > 111) = P\left(Z > \frac{111-109}{7}\right)$

$= P(Z > \frac{2}{7})$

$= 0.3875$ (4dp) from normCdf $(\frac{2}{7}, 9E99)$

b) $\bar{X}$ = sample mean height of 25 boys

$\bar{X} \sim N(109, \frac{7^2}{25})$

$P(\bar{X} > 111) = P\left(Z > \frac{111-109}{\sqrt{\frac{7^2}{25}}}\right)$

$= P(Z > \frac{10}{7})$

$= 0.0766$ (4dp) from normCdf $(\frac{10}{7}, 9E99)$

c) $\bar{X}$ has a smaller standard deviation than X, so the distribution of $\bar{X}$ is more "tightly packed" around the mean of 109cm.

2. assume that the distribution of steps is symmetrical

steps	320	310	321	304	298	328	296	307	314	295
median	300	300	300	300	300	300	300	300	300	300
steps − median	20	10	21	4	−2	28	−4	7	14	−5
ranks	8	6	9	2·5	1	10	2·5	5	7	4

$W_- = 1+2{\cdot}5+4 = 7{\cdot}5$

$W_+ = 8+6+9+2{\cdot}5+10+5+7 = 47{\cdot}5$

$= 55 = \frac{1}{2}\times 10\times 11$ check ✓

H_0: median steps = 300

H_1: median steps > 300

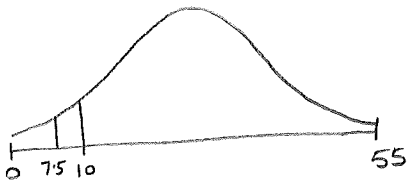
Assume H_0 is true

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

let $W = \min(W_-, W_+) = 7{\cdot}5$

we have $n = 10$

5% critical value = 10.



as $W = 7{\cdot}5 < 10$

we are in the critical region

so we reject H_0

and so we have evidence to suggest that the median number of steps recorded is greater than 300

so the mobile phone does appear to over count the steps.

3.

a) X = no. of donors of type B^-

$X \sim B(20, 0.020)$

$P(X \geq 2) = \underline{0.0599}$ (4dp) from binomCdf(20, 0.02, 2, 20)

b) Y = no. donors of O^+ or O^-

$Y \sim B(50, 0.409 + 0.095)$

$Y \sim B(50, 0.504)$

approximate Y with a normal distribution $np = 50 \times 0.504 = 25.2 > 5$

$nq = 50 \times 0.496 = 24.8 > 5$ ✓ allowed

so W = normal approx to Y

$W \sim N(50 \times 0.504, 50 \times 0.504 \times 0.496)$

$W \sim N(25.2, 12.4992)$

$P(Y \leq 30) = P(W \leq 30.5)$ by continuity correction

$= P\left(Z \leq \frac{30.5 - 25.2}{\sqrt{12.4992}}\right)$

$= P(Z \leq 1.49911)$

$= \underline{0.9331}$ (4dp) from normCdf(-9E99, 1.49911)

4.

observed	78	90	152
expected	80	80	160

H_0 : data follows specified ratio

H_1 : data does not follow specified ratio

Assume H_0 to be true.

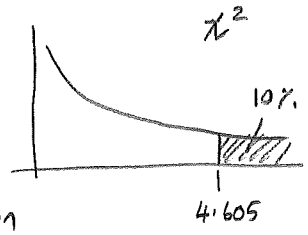
$\alpha = 10\%$ one-tail test

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(78-80)^2}{80} + \frac{(90-80)^2}{80} + \frac{(152-160)^2}{160}$$

$$= \frac{4}{80} + \frac{100}{80} + \frac{64}{160}$$

$$= 1.7$$

we have degrees of freedom, $\nu = 3 - 1 = 2$.



critical value, $\chi^2_{2,0.90} = 4.605$ from tables.

as $X^2 = 1.7 < 4.605$, we are not in critical region

so we do not reject H_0

we have evidence to suggest that the data follow the specified ratio.

hence the sample supports the genetic theory.

5.

x	0	1	2	3	4
$P(X=x)$	p	p	$2p$	$5p$	

a) i) $P(X=4) = 1 - P(X \leq 3)$

$= 1 - p - p - 2p - 5p$

$= \underline{1 - 9p}$.

$E(X) = \sum x_i P(X = x_i)$

$= 0 \times p + 1 \times p + 2 \times 2p + 3 \times 5p + 4 \times (1 - 9p)$

$= 0 + p + 4p + 15p + 4 - 36p$

$= \underline{4 - 16p}$ as required

ii) $3 = 4 - 16p$.

$16p = 1$

$\underline{p = \frac{1}{16}}$.

$E(X^2) = \sum x_i^2 P(X = x_i)$

$= 0 + p + 4 \times 2p + 9 \times 5p + 16 \times (1 - 9p)$

$= p + 8p + 45p + 16 - 144p$.

$= 16 - 90p$

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

$= (16 - 90p) - 3^2$

$= 7 - 90p$

$= 7 - 90 \times \frac{1}{16}$

$= \frac{11}{8}$

$= \underline{1.375}$.

b) $K = 2Y - X + 3$

$Y \sim Po(1)$ so $E(Y) = V(Y) = 1$

and $E(X) = 3$ $V(X) = \frac{11}{8}$

$E(K) = 2E(Y) - E(X) + 3$

$= 2 \times 1 - 3 + 3$

$= \underline{2}$.

$V(K) = 4V(Y) + V(X) + 0$

$= 4 \times 1 + \frac{11}{8}$

$= \frac{43}{8}$

$SD(K) = \sqrt{\frac{43}{8}}$

$= \underline{2.3184}$ (4dp)

6. X = length of baby at birth in cm

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

$n = 75$

$\Sigma x = 3840 \Rightarrow \bar{x} = \frac{3840}{75}$

$\Sigma x^2 = 198240 \Rightarrow s_{n-1} = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$

$= \sqrt{\frac{198240 - \frac{3840^2}{75}}{74}}$

$= 4.69617$

$H_0: \mu = 50$

$H_1: \mu > 50$

Assume H_0 to be true

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

so $X \sim N(50, \sigma^2)$

so $\bar{X} \sim N(50, \frac{\sigma^2}{75})$

we are bold to assume $\sigma^2 \approx s_{n-1}^2$

so $\bar{X} \sim N(50, \frac{4.69617^2}{75})$

p-value $= P(\bar{X} > \frac{3840}{75})$

$= P\left(Z > \frac{\frac{3840}{75} - 50}{\sqrt{\frac{4.69617^2}{75}}}\right)$

$= P(Z > 2.21293)$

$= 0.013451$ from normCdf(2.21293, 9E99)

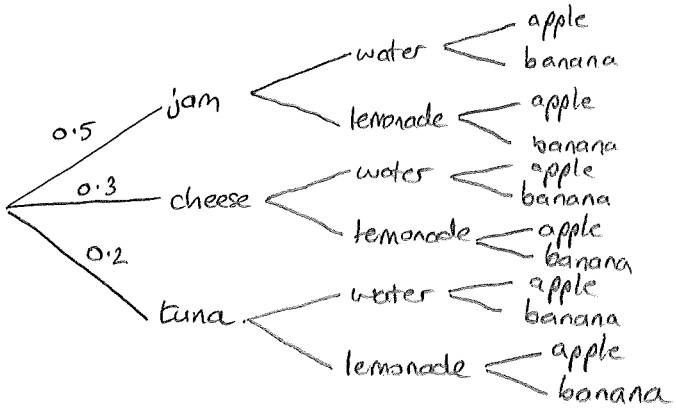
> 0.01

so we do not reject H_0

so we do not have evidence to suggest that the mean baby length is greater than 50cm.

This suggests that the midwife's theory is not true

7.



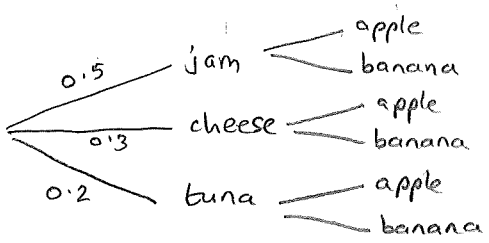
$P(\text{tuna} \cap \text{water}) = 0.035$

a) $P(\text{water} \mid \text{tuna}) = \dfrac{P(\text{water} \cap \text{tuna})}{P(\text{tuna})}$

$= \dfrac{0.035}{0.2}$

$= 0.175$

b) $P(\text{cheese} \cap \text{banana}) = 0.12.$ $\quad P(\text{jam} \cap \text{apple}) = ?$



now $P(\text{cheese} \cap \text{banana}) = P(\text{cheese}) \times P(\text{banana})$ as independent

$0.12 = 0.3 \times P(\text{banana})$

$P(\text{banana}) = \dfrac{0.12}{0.3}$

$= 0.4$

$\Rightarrow P(\text{apple}) = 0.6$

$\Rightarrow P(\text{jam} \cap \text{apple}) = P(\text{jam}) \times P(\text{apple})$ as independent

$= 0.5 \times 0.6$

$= 0.3.$

8. $n = 25$

pmcc, $r = 0.652$

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

assume H_0 to be true

$\alpha = 0.1\%$, two tailed test

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

$= \frac{0.652 \times \sqrt{25-2}}{\sqrt{1-0.652^2}}$

$= 4.12398$

p-value $= 2 \times P(t_{23} > 4.12398)$

$= 2 \times 0.000207$

$= 0.000413$

< 0.001

So we reject H_0.

So we have evidence to suggest that $\rho \neq 0$

which means at the 0.1% level there is evidence of a linear association between bodymass and plasma volume in healthy women.

9. a) paired data, unknown σ^2 $\Rightarrow$ t-test for paired data.

let X = difference in distances

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

$H_0 : \mu = 0$

$H_1 : \mu > 0$

assume H_0 to be true

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

$X \sim N(0, \sigma^2)$

$\bar{X} \sim N(0, \frac{\sigma^2}{12})$ as $n = 12$

$\frac{\bar{X} - 0}{\sqrt{\frac{\sigma^2}{12}}} \sim N(0, 1^2)$

we estimate σ^2 which s_{n-1}^2, so we use t_{11}

$\frac{\bar{X} - 0}{\sqrt{\frac{s_{n-1}^2}{12}}} \sim t_{11}$

test statistic, $t = \frac{\bar{x} - 0}{\sqrt{\frac{s_{n-1}^2}{12}}}$

$= \frac{0.145 - 0}{\sqrt{\frac{0.927^2}{12}}}$

$= 1.6816$

p-value $= P(t_{11} > 1.6816)$

$= 0.060392$ from tcdf(1.6816, 9E99, 11)

> 0.05

so we do not reject H_0

so we do not have evidence to suggest that the mean difference in distances is greater than zero

this means that it appears that runners do not run further when wearing a fitness tracker.

b) i) the assumption of normality is questionable as the distribution seems not to be symmetrical, with a negative skew.

ii) a Wilcoxon Rank Sum test for paired data could be used, but it too requires symmetrical differences, which makes it unsuitable.

10. 2010 → 2017, proportion = 0·624

2018, $\hat{p} = \frac{23312}{37878}$

let X = number of homeless veterans in sheltered accomodation in 2018

$X \sim B(37878, p)$

$H_0: p = 0.624$

$H_1: p \neq 0.624$

assume H_0 to be true

$\alpha = 0.5\%$ two tail test

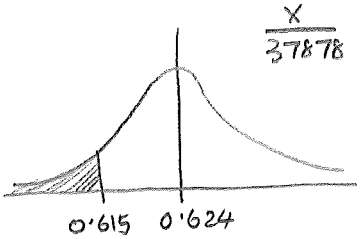
so $X \sim B(37878, 0.624)$

$\Rightarrow X \approx N(37878 \times 0.624, 37878 \times 0.624 \times (1-0.624))$ valid as $np > 5$, $nq > 5$

let $\frac{X}{37878}$ = proportion of veterans in sheltered accomodation

$\frac{X}{37878} \sim N\left(0.624, \frac{0.624 \times (1-0.624)}{37878}\right)$

$\frac{X}{37878} \sim N(0.624, 0.002489^2)$



test statistic, $p = \frac{23312}{37878} = 0.61545$

$p\text{-value} = 2 \times P\left(\frac{X}{37878} < 0.61545\right)$

$= 2 \times P\left(Z < \frac{0.61545 - 0.624}{0.002489}\right)$

$= 2 \times P(Z < -3.43553)$

$= 2 \times 0.000296$ from normCdf(-9E99, −3.43553)

$= 0.000591$

< 0.005

So we reject H_0 at the 0·5% level

we have evidence to suggest that the proportion of homeless veterans in sheltered accommodation is different from the previous 8 years

(we conjecture that the proportion has decreased)

check using pure binomial

$p\text{-value} = 2 \times P(X \leq 23312)$

$= 2 \times 0.000307$ from binomCdf(37878, 0.624, 0, 23312)

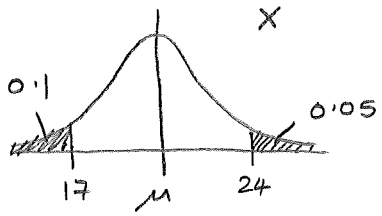
$= 0.000613$

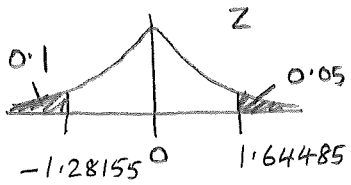
< 0.005

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

$P(X > 24) = 0.05$

$P(X < 17) = 0.1$





so $\frac{X-\mu}{\sigma} = Z$ gives two equations: $\frac{17-\mu}{\sigma} = -1.28155$

and $\frac{24-\mu}{\sigma} = 1.64485$

Solve simultaneously using LinSolve $\begin{cases} 17-m = -1.28155s \\ 24-m = 1.64485s \end{cases}, \{m, s\}$

gives $\mu = 20.0655$

$\sigma = 2.39201$

12. $n = 100$

$\hat{p} = 0.55$

a) 99% CI is $\hat{p} \pm z_{0.995} \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$

$= 0.55 \pm 2.57583 \times \sqrt{\frac{0.55 \times 0.45}{100}}$

$= (0.421854, 0.678146)$

$\approx (42\%, 68\%)$

it is an approximate interval as we have approximated a binomial distribution with a normal distribution when using proportions.

b) what is n so that lower bound of interval is greater than 50%.

lower bound is $\hat{p} - z_{0.995}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

so $\hat{p} - z_{0.995}\sqrt{\frac{\hat{p}\hat{q}}{n}} > 0.50$

$0.55 - z_{0.995}\sqrt{\frac{0.55 \times 0.45}{n}} > 0.50$

$0.05 > z_{0.995}\sqrt{\frac{0.2475}{n}}$

$\frac{0.05}{z_{0.995}} > \sqrt{\frac{0.2475}{n}}$

$\left(\frac{0.05}{z_{0.995}}\right)^2 > \frac{0.2475}{n}$

$n > \frac{0.2475}{\left(\frac{0.05}{z_{0.995}}\right)^2}$

$n > 656.855$

so, the smallest sample size would be $n = 657$