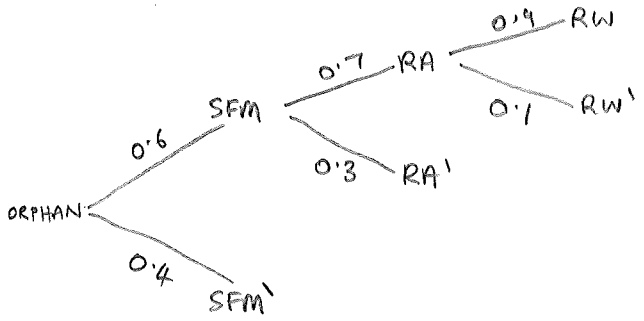
2017 SQA AH Statistics

1.



a) i) $P(\text{return to wild}) = P(SFM \cap RA \cap RW)$

$= 0.6 \times 0.7 \times 0.9$

$= 0.378$

ii) $P(\text{not returned to wild}) = P(SFM \cap RA \cap RW')$

$= 0.6 \times 0.7 \times 0.1$

$= 0.042$

b) $P(SFM' \mid RA') = \dfrac{P(SFM' \cap RA')}{P(RA')}$

$= \dfrac{P(SFM')}{1 - P(RA)}$

$= \dfrac{0.4}{1 - 0.6 \times 0.7}$

$= \dfrac{0.4}{0.58}$

$= 0.68965\ldots$

≈ 0.690 (3sf)

2.

x	20	30	40	50	60
$P(X=x)$	c	c	c	c	c

a) as $\Sigma P(X=x) = 1 \Rightarrow 5c = 1 \Rightarrow c = 0.2$

x	20	30	40	50	60
$P(X=x)$	0.2	0.2	0.2	0.2	0.2

b) i) $E(X) = \Sigma x P(X=x)$

$= 20 \times 0.2 + \ldots + 60 \times 0.2$

$= (20 + 30 + \ldots + 60) \times 0.2$

$= 200 \times 0.2$

$= 40.$

$E(X^2) = \Sigma x^2 P(X=x)$

$= (20^2 + 30^2 + \ldots + 60^2) \times 0.2$

$= 9000 \times 0.2$

$= 1800$

$Var(X) = E(X^2) - [E(X)]^2$

$= 1800 - 40^2$

$= 200.$

So $E(X) = 40$, $Var(X) = 200$.

ii) $P(X > \mu) = P(X > 40)$

$= P(X = 50) + P(X = 60)$

$= 0.2 + 0.2$

$= 0.4$

3. a) i) not possible to obtain a sample of exam marks as...

- the results may be confidential
- only fail/pass/merit/distinction may be recorded centrally rather than any numerical scores.

ii) obtain 15% of pianists from 20 centres.

Due to travel cost restrictions, we want to avoid visiting all 20 centres.

Hence a form of cluster sampling may suit

So consider the 20 centres to each be a cluster

Say we have sufficient expenses to visit 4 centres.

Use simple random sampling to select the 4 centres from the 20

(allocate each centre a number from 1 to 20, and then randomly generate four integers between 1 and 20 inclusive).

With these four centres identified, visit each one and conduct either a simple random sample of 15% of their pianists scores, or perform stratified sampling,

For stratified sampling, treat each Grade of Pianist (grade 1 to 8) as a strata, and select by simple random sample a proportional number of pianists from each strata to obtain 15% of the total

3 b) let X_p = piano score

$n_p = 16 \quad \bar{x}_p = 77 \quad s_p = 10$

let X_v = violin score

$n_v = 14 \quad \bar{x}_v = 82 \quad s_p = 8$

as we are told to perform a t-test, we assume that the parent population standard deviations are the same, so that we pool the sample standard deviations.

we also assume that X_p and X_v are distributed normally.

$$\text{So } s^2 = \frac{(n_p - 1)s_p^2 + (n_v - 1)s_v^2}{n_p + n_v - 2}$$

$$= \frac{15 \times 10^2 + 13 \times 8^2}{16 + 14 - 2}$$

$$= 83.2857$$

$$s = 9.1261$$

so if $X_p \sim N(\mu_p, \sigma_p^2)$ and $X_v \sim N(\mu_v, \sigma_v^2)$

then $H_0: \mu_p = \mu_v$

$H_1: \mu_p < \mu_v$

Assume H_0 to be true. $\alpha = 5\%$ 1 tail test

$$\text{test statistic, } t = \frac{\bar{x}_p - \bar{x}_v}{s\sqrt{\frac{1}{n_p} + \frac{1}{n_v}}}$$

$$= \frac{77 - 82}{9.12\sqrt{\frac{1}{16} + \frac{1}{14}}}$$

$$= -1.49709$$

$$p\text{-value} = P(t_{28} < -1.49709)$$

$$= 0.072779 \qquad \text{from tCdf}(-9E99, -1.497, 28)$$

$$> 0.05$$

Hence we do not have sufficient evidence at the 5% level to reject H_0 and thus we conclude that the mean piano and violin marks are the same.

4. a) 22 27 27 (29) 29 30 30 | 30 30 32 (32) 33 34 35

Q1 — median 30. — Q3

So $IQR = Q_3 - Q_1$
$= 32 - 29$
$= 3$

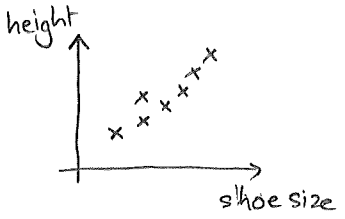
upper fence $= Q_3 + 1.5\,IQR$
$= 32 + 1.5 \times 3$
$= 36.5$

lower fence $= Q1 - 1.5\,IQR$
$= 29 - 1.5 \times 3$
$= 24.5$

as $22 < 24.5$ we have one outlier at the lower end.

b) i) the plot is inappropriate as it does not incorporate the paired nature of the data.

ii) a scatter plot would be better eg.



If the distributions of the separate data sets were assumed to be normal, then bivariate analysis could be performed to establish if a linear relationship existed.

A hypothesis test on the correlation coefficient, r, would be done

If successful, then a linear equation could be calculated and residual plots examined.

It may be that a non-linear relation exists, and thus the data may have to be transformed in some way to find a linear relationship.

5. weekly sample size, $n = 5$

39 weeks' samples.

For first 20 weeks, $\bar{x} = -6.1$, $s = 9.0561$

a) i) X = SCC measurement

if X has $E(X) = \mu$, $Var(X) = \sigma^2$

then $\bar{X}$ has $E(\bar{X}) = \mu$ $Var(\bar{X}) = \frac{\sigma^2}{5}$

so high alarm limit $= \bar{x} + 6\sqrt{\frac{s^2}{5}}$

$= -6.1 + 6 \times \sqrt{\frac{9.0561^2}{5}}$

$= -6.1 + 6 \times 4.05001$

$= 18.2001$

≈ 18.2 (1dp)

ii) First value exceeds 18.2 in week 25.

b) i) If process in control, $P(\text{above centre line}) = P(\text{below centre line}) = \frac{1}{2}$

$\therefore P(9 \text{ points all on one side}) = P(9 \text{ points above}) + P(9 \text{ points below})$

$= \left(\frac{1}{2}\right)^9 + \left(\frac{1}{2}\right)^9$

$= 0.003906$

ii) 9 points in a row (all above) in week 33

2 out of 3 beyond same 2sigma limit in week 16

6. 130 skiers.

S = weight of skier $\quad S \sim N(75, 16)$

E = weight of equipment $\quad E \sim N(10, 2)$

we will assume independence of weights of skiers, weights of equipment and weights of equipment of each skier.

let T_i = weight of a skier$_i$ with their equipment

$T_i = S_i + E_i$

$T_i \sim N(75+10, 16+2)$

$T_i \sim N(85, 18)$

let total weight of 130 skiers, $TW = T_1 + T_2 + \ldots + T_{130}$

$$E(TW) = 130 \times E(T_i)$$
$$= 130 \times 85$$
$$= 11050$$

$$Var(TW) = Var(T_1) + Var(T_2) + \ldots + Var(T_{130})$$
$$= 130 \times 18$$
$$= 2340$$

so $TW \sim N(11050, 2340)$

$$P(TW > 11200) = P\left(Z > \frac{11200 - 11050}{\sqrt{2340}}\right)$$

$$= P(Z > 3.10087)$$

$$= 0.000965$$

$$\approx 0.0010 \quad (4dp)$$

7.

18 samples gave $\bar{x} = 7.46$, $s_{n-1} = 1.46$

X = no. of trees per hectare.

a) let us assume X distributed normally

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

so $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{18}\right)$

as we're estimating σ with s, we use a t_{17} distribution.

so 95% CI for μ is $\bar{x} \pm t_{17, 0.975}\sqrt{\frac{s^2}{18}}$

$7.46 \pm 2.10982\sqrt{\frac{1.46^2}{18}}$

7.46 ± 0.726041

$(6.73396, 8.18604)$

$(6.73, 8.19)$ (3sf)

b) in 2000, $\bar{x} = 5.87$

i) as 5.87 is outside the CI from the 1930's, we could reasonably conclude that there has been a decline in the number of large trees.

Even if we treated 5.87 as the centre of a confidence interval constructed in a similar way to before, $5.87 + 0.726$ is still not within $(6.73, 8.19)$

Here the population mean is very unlikely to be the same now as it was then.

ii) I would be cautious about stating this analysis validates the biologists' concerns as there may be many other factors at work - beyond water shortage - that are leading to a decline in numbers of large trees.

Timber felling could easily have the same effect.

Also large trees take a long time to grow, and the forrest may not have matured enough to have enough trees of sufficient girth to be counted, after any areas had been felled, or burnt due to forest fires.

8. a) $W \sim Po(4)$

$E(W) = 4$

$Var(W) = 4 \Rightarrow$ st. dev $= 2$

so $P(4-2 \leq W \leq 4+2)$

$= P(2 \leq W \leq 6)$

$= 0.797748$ from poissCdf (4, 2, 6)

$\approx \underline{0.7977}$ (4dp)

b) $X \sim N(4, 2^2)$

so $P(4-2 \leq X \leq 4+2)$

$= P(2 \leq X \leq 6)$

$= P\left(\frac{2-4}{2} \leq Z \leq \frac{6-4}{2}\right)$

$= P(-1 \leq Z \leq 1)$

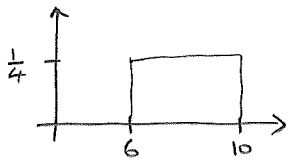
$= 0.682628$ from normCdf (-1, 1)

$\approx \underline{0.6826}$ (4dp)

c) $Y \sim U(6, 10)$

so $E(Y) = \frac{6+10}{2} = 8$

$Var(Y) = \frac{(10-6)^2}{12} = \frac{4^2}{12} = \frac{4}{3}$.



$\Rightarrow$ st. dev $= \sqrt{\frac{4}{3}}$

$\Rightarrow P\left(8 - \sqrt{\frac{4}{3}} \leq Y \leq 8 + \sqrt{\frac{4}{3}}\right)$

$= \frac{1}{4}\left[\left(8+\sqrt{\frac{4}{3}}\right) - \left(8-\sqrt{\frac{4}{3}}\right)\right]$

$= \frac{1}{4} \cdot 2 \cdot \sqrt{\frac{4}{3}}$

$= \frac{1}{2} \cdot \frac{2}{\sqrt{3}}$

$= \frac{1}{\sqrt{3}}$

$\approx \underline{0.5774}$ (4dp)

9. X = lead concentration in a sample

a) problem: the centre of a 50x50m square may not be an area of open soil. Indeed it may be under one of the village homes.

solution: consider looking at smaller squares of land, say 10x10m squares. There would be 25 of these smaller squares in the 50x50 region. Then perform random sampling without replacement of these 25 regions until a square was located whose centre was open soil, and not blocked by any dwelling or similar obstruction.

b) $E(X) = 165.6$

$Var(X) = 23.1^2$

$n = 25$.

if $\bar{X}$ = mean concentration, then by Central Limit Theorem, $\bar{X}$ is approximately normally distributed, $\bar{X} \sim N\left(165.6, \frac{23.1^2}{25}\right)$, as the sample size is larger than 20.

c) let Y = contamination at village A

Assume $Var(Y) = 23.1^2$, as before

let $\bar{Y} \sim N\left(\mu, \frac{23.1^2}{25}\right)$ where $\bar{Y}$ = mean of 25 samples

$\bar{y} = 174.5$

$H_0: \mu = 165.6$

$H_1: \mu > 165.6$

Assume H_0 to be true

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

so $\bar{Y} \sim N\left(165.6, \frac{23.1^2}{25}\right)$

$p\text{-value} = P(\bar{Y} > 174.5)$

$= P\left(Z > \frac{174.5 - 165.6}{\sqrt{\frac{23.1^2}{25}}}\right)$

$= P(Z > 1.92641)$

$= 0.027027$ from normCdf(1.92,9E99)

< 0.05

So we have evidence to reject H_0 and conclude that the soil in Village A has a mean concentration that is higher than 165.6 mg/kg

10. $P(\text{recapture}) = 0.2$

a) X = no. of rodents previously captured

$X \sim B(20, 0.2)$

$P(X=3) = \binom{20}{3} 0.2^3 0.8^{17}$

$= 0.205364$ from binomPdf(20, 0.2, 3)

$\approx \underline{0.2054}$ (4dp)

b) Y = no. of rodents previously captured, at second site

$Y \sim B(45, 0.2)$

approximate Y to normal distribution, as $45 \times 0.2 = 9 > 5$
$45 \times 0.8 = 36 > 5$ ✓ approx is acceptable as $np > 5, nq > 5$

let W = normal approx to Y

$W \sim N(45 \times 0.2, 45 \times 0.2 \times 0.8)$

$W \sim N(9, 7.2)$

so $P(5 \leq Y \leq 10) = P(4.5 \leq W \leq 10.5)$ by c.c.

$= P\left(\frac{4.5-9}{\sqrt{7.2}} \leq Z \leq \frac{10.5-9}{\sqrt{7.2}}\right)$ where $Z \sim N(0, 1^2)$

$= P(-1.67705 \leq Z \leq 0.559017)$

$= 0.665159$ from normCdf(-1.677, 0.559)

$\approx \underline{0.6652}$ (4dp)

c)

obs	M	F
R	58	51
NR	255	182

exp	M	F
R	62.4853	46.5147
NR	250.515	186.485

H_0: no association between gender and recapture status
H_1: there is an association
Assume H_0 to be true
$\alpha = 5\%$ 1-tail test

$X^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 0.942675$

$\nu = (2-1) \times (2-1) = 1$

so $P(\chi_1^2 > 0.942675) = 0.331591 > 0.05$

Hence we do not have evidence to reject H_0, and we conclude that there is no association between gender and recapture status.

11. a) X = no. of residents in favour (company A)

$X \sim B(100, \hat{p})$

approximate X with normal distribution, Y → if $\hat{p} = 0.61$, then $n\hat{p} > 5$ and $n\hat{q} > 5$ so acceptable approximation

so $Y \sim N(100\hat{p}, 100\hat{p}\hat{q})$

let $\frac{Y}{100}$ = proportion of residents in favour

so $\frac{Y}{100} \sim N\left(\hat{p}, \frac{\hat{p}\hat{q}}{100}\right)$

$\therefore$ 99% CI is $\hat{p} \pm z_{0.995}\sqrt{\frac{\hat{p}\hat{q}}{100}}$

$= 0.61 \pm 2.57583\sqrt{\frac{0.61 \times 0.39}{100}}$

$= 0.61 \pm 0.125636$

$= (0.484364, 0.735636)$

$\approx (48.4\%, 73.6\%)$ (3sf)

It is an approximate interval due to the Normal approximation to the binomial step at the outset.

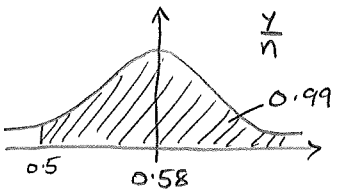
b) if X = no. of residents in favour (company B)

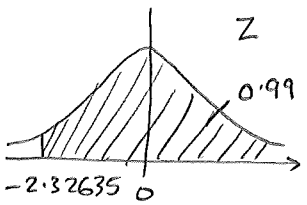
$X \sim B(n, \hat{p})$

so $\frac{Y}{n} \sim N\left(\hat{p}, \frac{\hat{p}\hat{q}}{n}\right)$

now they are 99% confidence that the true proportion is over 50% (the majority)

so $P\left(\frac{Y}{n} > 0.5\right) = 0.99$ where $\frac{Y}{n} \sim N\left(0.58, \frac{0.58 \times 0.42}{n}\right)$





↑ from invNorm(0.01)

so $\frac{0.5 - 0.58}{\sqrt{\frac{0.58 \times 0.42}{n}}} = -2.32635$

$\frac{0.5 - 0.58}{-2.32635} = \sqrt{\frac{0.58 \times 0.42}{n}}$

$(0.034389)^2 = \frac{0.58 \times 0.42}{n}$

$n = \frac{0.58 \times 0.42}{(0.034389)^2}$

$n = 205.99$

Hence company B must have asked at least 206 residents.

12. a) Mann-Whitney is appropriate as the two distributions of percentage blue bell cover have similar shapes to their distributions and the data is not paired.

Also, the distributions do not appear to be symmetrical/normally distributed (ie they are skewed) and thus t-tests and z-tests are not appropriate.

b) $W_1 = 480$.

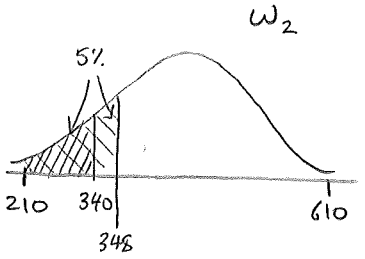
i) H_0: median strip 1 = median strip 2

H_1: median strip 1 ≠ median strip 2

Assume H_0 to be true

$\alpha = 10\%$, 2-tail test

$W_1 = 480$ from $n = 20$



Total rank sum $= \frac{1}{2} \times 40 \times 41 = 820$

Hence $W_2 = 820 - 480 = 340$.

Minimum rank sum $= \frac{1}{2} \times 20 \times 21 = 210$

Maximum rank sum $= \frac{1}{2} \times 40 \times 41 - 210 = 610$

we want to know $P(W_2 \leq 340)$ where $n = m = 20$.

From tables, we know that $P(W \leq 348) = 0.05$

so $P(W_2 \leq 340) < 0.05$

so $2P(W_2 \leq 340) < 0.10$

Hence, we have evidence to reject H_0 at the 10% level

We conclude that there is likely to be a difference in the population's median percentage bluebell cover, between the two strips.

ii) At the 5% level, we would have likely not rejected H_0 (we assume here that $P(W_2 \leq 340)$ is just under 0.05, so that doubling it makes it exceed 5%)

If we'd not rejected, then we'd have concluded that the medians were likely to be the same and thus that cutting in spring or autumn makes little difference in the number of bluebells.

iii) Precision might be improved by placing more than 20 quadrats in each strip as well as ensuring none of the placements overlapped the location of a previous quadrat.

13. a) The scatterplot suggests

- there is a non-linear correlation between area & number of species
- the data for Florida is such an outlier that its status as being treated as an island is highly questionable.

b) $\Sigma x = 17.375 \quad S_{xx} = 26.2676 \quad S_{yy} = 2.6895 \quad S_{xy} = 7.5234 \quad n = 13.$

$y = 0.9202 + 0.2864x$

$x = \log_{10}$ Area

$y = \log_{10}$ (Number of Species)

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{7.5234}{\sqrt{26.26 \times 2.6395}} = 0.903531$$

so coeff of determination, $r^2 = 0.816369$

Hence the least squares regression line accounts for 81.6% of the variation present in the transformed data.

This high value suggests that the linear model of the transformed data would generate reliable predictions.

c) $x = 2.5441$

so $\hat{y} = 0.9202 + 0.2864 \times 2.5441$

$= 1.64883$

As we have an individual case, we shall construct a prediction interval around this $\hat{y}$ value.

$$\Rightarrow \quad \hat{y} \pm t_{11, 0.95} \times s \times \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}$$

$$1.64883 \pm 1.79588 \times s \times \sqrt{1 + \frac{1}{13} + \frac{(2.5441 - \frac{17.375}{13})^2}{26.2676}}$$

we need $s = \sqrt{\frac{SSR}{n-2}}$ where $SSR = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

$= 0.484695$

so $s = \sqrt{\frac{0.484695}{11}}$

$\Rightarrow s = 0.209912$

$\therefore$ 90% PI is $1.64883 \pm 1.79588 \times 0.209912 \sqrt{1 + \frac{1}{13} + 0.055513}$

$\Rightarrow 1.64883 \pm 0.401165$

$\Rightarrow (1.24766, 2.05)$

So min. species : $\log_{10}(\text{min}) = 1.24766$

$$\Rightarrow \quad \text{min} = 10^{1.24766}$$

$$\Rightarrow \quad \text{min} = 17.6874$$

similarly, $\text{max} = 10^{2.05}$

$$= 112.201$$

Hence a 90% prediction interval for the number of species is 17.6 to 112.2

or roughly 18 to 112.