SQA AH Statistics – 2018

1. A = no. wakes per night $A \sim Po(3)$

B = no. wakes per night $B \sim Po(2)$

a) $P(B=3) = \dfrac{e^{-2}2^3}{3!} = 0.180447$ or by poissPdf(2,3)

$\simeq 0.1804$ (4dp)

b) P(neither baby wakes) $= P(A=0 \cap B=0)$

$= P(A=0)\,P(B=0)$ assuming independence

$= \dfrac{e^{-3}3^0}{0!} \times \dfrac{e^{-2}2^0}{0!}$

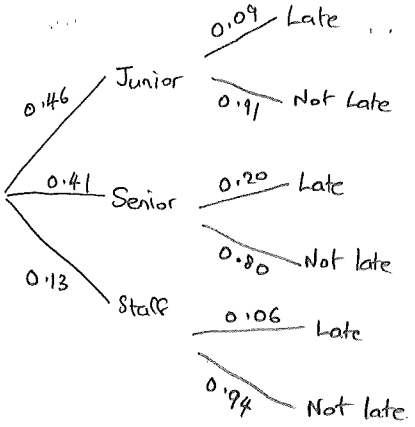
$= e^{-3} \times e^{-2}$

$= e^{-5}$

$= 0.006738$

$\simeq 0.0067$ (4dp) or by poissPdf(2,0) × poissPdf(3,0)

2



a) i) P(Junior and Not Late)

= P (Junior) × P (Not Late | Junior)

= 0·46 × 0·91

= 0·4186

ii) P (Late)

= P (Junior ∩ Late) + P (Senior ∩ Late) + P(Staff ∩ Late)

= 0·46×0·09 + 0·41×0·20 + 0·13×0·06

= 0·0414 + 0·082 + 0·0078

= 0·1312.

b) $P(\text{Senior} \mid \text{Late}) = \frac{P(\text{Senior} \cap \text{Late})}{P(\text{Late})}$

$= \frac{0.41 \times 0.20}{0.1312}$

$= 0.625.$

c) We would sample 10% of the population through Stratified Random Sampling.

We would randomly sample 10% of the Junior Pupils, 10% of the Senior Pupils and 10% of the Staff. When sampling from each of these strata, we would use simple random sampling.

3. $S_{xx} = 10005.4$

$S_{xy} = 2265.43$

$S_{yy} = 623.429$

$n = 7$.

a) $H_0 : \beta = 0$

$H_1 : \beta \neq 0$

Assume H_0 to be true

$\alpha = 5\%$.

test statistic, $t = \dfrac{b\sqrt{S_{xx}}}{s}$

here $b = \dfrac{S_{xy}}{S_{xx}} = 0.226421$ and $s^2 = \dfrac{SSR}{n-2}$ where $SSR = S_{yy} - \dfrac{S_{xy}^2}{S_{xx}}$

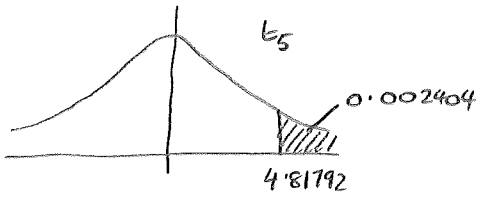
$\Rightarrow SSR = 110.489$

so $s^2 = \dfrac{110.489}{7-2} = 22.0977$

$\Rightarrow s = 4.70082$

$\therefore t = \dfrac{0.226421\sqrt{10005.4}}{4.70082}$

$t = 4.81792$



$p\text{-value} = 2 \times P(t_5 > 4.81792)$

$= 2 \times 0.002404$

$= 0.004808$

< 0.05.

from tCdf(4.81792, 9E99, 5)

Hence we have evidence to reject H_0

We conclude that the slope parameter is non-zero

b) We would also calculate the correlation coefficient (or the Coefficient of Determination) and see if it was sufficiently high.

A high correlation coefficient would mean that the linear relation fits the data closely, meaning that it would give reliable predictions of y from x.

4. X = life expectancy of bulb.

$X \sim N(10000, 250^2)$

a) $P(X > 10100) = P\left(Z > \frac{10100 - 10000}{250}\right)$ where $Z \sim N(0, 1^2)$

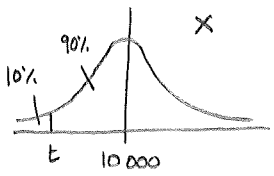
$= P\left(Z > \frac{100}{250}\right)$

$= P(Z > 0.4)$

$= 0.344578$ from normCdf (0.4, 9E99)

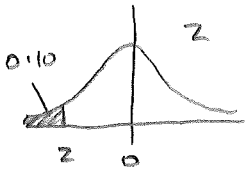
≈ 0.3446 (4dp)

b)



we want t, such that $P(X > t) = 0.90$

$\Rightarrow P(X < t) = 0.10$



$\Phi^{-1}(0.10) = -1.28155$ by invNorm (0.1)

So $P(Z < -1.28155) = 0.10$

So $\frac{X - 10000}{250} = -1.28155$

$X = 10000 - 1.28155 \times 250$

$X = 9679.61$

So the packaging should claim that 90% of bulbs exceed 9680 hrs (3sf)

c) B_i = weight of one bulb $B_i \sim N(24, 1)$

X_i = weight of one box $X_i \sim N(5, 0.5)$

C = weight of crate $C \sim N(75, 7)$

T = total weight of crate + 100 bulbs in boxes

$T = C + B_1 + \ldots + B_{100} + X_1 + \ldots + X_{100}$

$E(T) = E(C) + 100E(B_i) + 100E(X_i)$

$= 75 + 100 \times 24 + 100 \times 5$

$= 2975$

$V(T) = V(C) + 100\,V(B_i) + 100\,V(X_i)$

$= 7 + 100 \times 1 + 100 \times 0.5$

$= 157$

we need to assume weights of all boxes, bulbs and the crate are all independent in order to work out $V(T)$

so $T \sim N(2975, 157)$

$P(T < 3000) = P\left(Z < \frac{3000 - 2975}{\sqrt{157}}\right)$

$= P(Z < 1.99522)$

$= 0.9767$ (4sf) from normCdf (-9E99, 1.99522).

5. a) systematic sampling.

b) depending on the location picked, there may have been a non-representative collection of pebbles in that place. For example, smaller pebbles would be in the shallower water, or on the inside of bends where the water flows slower. And larger pebbles are in deeper water, or where the water flows fastest, say on the outside of bends.

c) old method gave mean 115·3 mm, st.dev 21·6

new method of $n=100$ gives mean 119·4mm.

let X = size of pebble

we assume $V(X) = 21{\cdot}6^2$, and $E(X) = \mu$.

let $\bar{X}$ = mean size of pebble, from sample of size 100

so, by Central Limit Theorem, $\bar{X} \approx N\left(\mu, \frac{21{\cdot}6^2}{100}\right)$

so 90% CI for μ is sample mean $\pm z_{0.95} \sqrt{\frac{21{\cdot}6^2}{100}}$

$= 119{\cdot}4 \pm 1{\cdot}64485 \sqrt{\frac{21{\cdot}6^2}{100}}$

$= 119{\cdot}4 \pm 3{\cdot}55288$

$= (115{\cdot}847, 122{\cdot}953)$

$\approx (115{\cdot}8, 123{\cdot}0)$ (1dp)

This confidence interval suggests that the true mean is between 115·8 and 123·0 This interval does not contain the mean of 115·3mm from the older method, so this either means that the pebbles in the stream are now not as small as they use to be, or that the old method underestimated the pebble size.

6. $m = 4$, $n = 7$

$W_M = 14$.

$P(\text{sum of ranks} \leq 14) = \dfrac{\text{no. ranks} \leq 14}{\text{no. possible ranks}}$

no. possible ranks $= {}^{11}C_4 = 330$.

rank total	positions	1	2	3	4	5	6	7	8	9	10	11
10		1	2	3	4							
11		1	2	3		5						
12		1	2		4	5						
13		1		3	4	5						
14			2	3	4	5						
12		1	2	3			6					
13		1	2		4		6					
14		1		3	4		6					
14		1	2			5	6					
13		1	2	3				7				
14		1	2		4			7				
14		1	2	3					8			

Total of 12 ways of obtaining rank sum of 14 or less

$\therefore P(\text{sum of ranks} \leq 14) = \dfrac{12}{330}$

$= \dfrac{2}{55}$.

as required.

7. Packets of 6.

10% fail to grow.

120 packets tested.

X = number of failed bulbs in a packet

$X \sim B(6, p)$

H_0: Data fits Bin (6, 0.1)

H_1: Data does not fit Bin (6, 0.1)

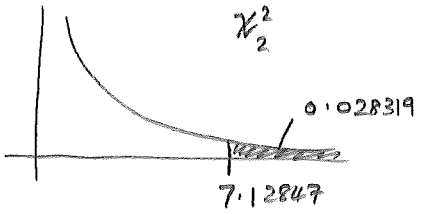
x	0	1	2	3	4	5	6
f_o, observed	59	38	19	3	1	0	0
f_e, expected	63.8	42.5	11.8	1.7	0.14	0.006	0.00012.

combine, to ensure $f_e > 5$

↓

x	0	1	2+
f_o	59	38	23.
f_e	63.77	42.51	13.7118

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 7.12847$$



degrees freedom = 3 − 1 = 2

p-value = $P(\chi^2_2 > 7.12847)$

= 0.028319

< 0.05

Hence we have evidence to reject H_0, and conclude that the data does not fit a Binomial Distribution with $p = 0.1$.

b) Binomial model requires trials to be independent, and with fixed probability. In this case, that equates to not having a packet with all-good bulbs, or all-bad bulbs. It's likely that bad bulbs come in batches, rendering the requirement of independence to be void.

8. 11.9% accidents involve young driver, 2008-12

In 2013, 100 accidents had 18 young drivers.

a) let X = number of accidents with young driver

$X \sim Bin(100, p)$

approx X to normal to give $Y \sim N(100p, 100pq)$ [valid if $100p > 5$ and $100(1-p) > 5$]

let $\frac{Y}{100}$ = proportion of accidents with young driver

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

$H_0: p = 0.119$

$H_1: p > 0.119$

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

so $\frac{Y}{100} \sim N\left(0.119, \frac{0.119 \times 0.881}{100}\right)$

$p\text{-value} = P\left(\frac{Y}{100} \geq \frac{18}{100}\right)$

$= P\left(\frac{Y}{100} \geq 0.18\right)$

$= P\left(Z \geq \frac{0.18 - 0.119}{\sqrt{\frac{0.119 \times 0.881}{100}}}\right)$

$= P(Z \geq 1.88395)$

$= 0.029786$ from normCdf(1.88395, 9E99)

< 0.05

So we have evidence to reject H_0

We conclude that there has been an increase in the proportion of accidents that involve a young driver, compared to 2008-12.

b) now, $n = 40$ instead of 100 (as in part a)

proportion involving young driver is $\frac{d}{40}$.

z-value to be in lowest 5% tail is $z_{0.05} = -1.64485$ from invNorm(0.05)

so $\frac{\frac{d}{40} - 0.119}{\sqrt{\frac{0.119 \times 0.881}{40}}} = -1.64485$

$\Rightarrow \frac{d}{40} = 0.119 - 1.64485\sqrt{\frac{0.119 \times 0.881}{40}}$

$\Rightarrow d = 40 \times 0.034791$

cont/ $\Rightarrow$ $d = 1.39164$

So if $d = 1$, then proportion of drivers is less than 2008-2012 figures.

Interestingly, if there are 40 accidents in four-months, then over a year there will be 120 accidents > 100 accidents in 2013.

Hence, this location is not getting any safer, regardless of the age of the drivers!

9. batches of 5 readings

previous variance = 0.0576.

let X = pH value of one reading $\quad E(X)=\mu \quad V(X)=0.0576$

and $\bar{X}$ = mean pH value over 5 readings $\quad E(\bar{X})=\mu \quad V(\bar{X})=\frac{0.0576}{5}$

from 10 batches, sample mean $= \frac{5.488+\ldots+5.491}{10}$

$= \frac{54.3}{10}$

$= \underline{\underline{5.43}}$

and thus 1 sigma limits are $5.43 \pm 1 \times \sqrt{\frac{0.0576}{5}}$

$= 5.43 \pm 0.107331$

$= \underline{\underline{5.323, 5.537}} \quad (3dp)$

b) let 5th reading be x

$\therefore$ mean of five readings $= \frac{5.481+5.392+5.606+5.463+x}{5}$

$= \frac{21.942+x}{5}$

we know that $x > 5.43$

For process to be out of control, if the 21st batch has a mean that's > the upper 1-sigma limit, then 4 of the last 5 batches will fall beyond the same 1σ limit (WECO Rule #3)

So $\frac{21.942+x}{5} > 5.537$

$21.942+x > 27.685$

$x > 5.743$

Hence, smallest possible value of 5th individual reading that would indicate process is out of control is $\underline{\underline{5.743}}$.

10. a) i)

x	7	8	9	10	11	12	13	14	15	16	17
f	1	2	2	3	4	5	5	4	2	0	2
$P(X=x)$	$\frac{1}{30}$	$\frac{2}{30}$	$\frac{2}{30}$	$\frac{3}{30}$	$\frac{4}{30}$	$\frac{5}{30}$	$\frac{5}{30}$	$\frac{4}{30}$	$\frac{2}{30}$	$\frac{0}{30}$	$\frac{2}{30}$.

ii) X = no. of turns taken to win

$E(X) = 12$

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

$= E(X^2) - 12^2$

$= \sum x^2 P(X=x) - 12^2$

$= \left(7^2 \times \frac{1}{30} + 8^2 \times \frac{2}{30} + \ldots + 17^2 \times \frac{2}{30}\right) - 12^2$

$= 150 - 12^2$

$= 6.$ as required.

b) i) we have $E(X) = 12$, $V(X) = 6$, so approx X with $N(12, 6)$

reason: normal distribution seems appropriate due to "bell shape" of bars.

ii) $P(X = 10) \approx P(9.5 < Y < 10.5)$ where $Y \sim N(12, 6)$, by continuity correction

$= P\left(\frac{9.5-12}{\sqrt{6}} < Z < \frac{10.5-12}{\sqrt{6}}\right)$

$= P(-1.02 < Z < -0.612)$

$= 0.116429$ by normCdf$(-1.02, -0.612)$

≈ 0.1164 (4dp)

[note, this compares with the observed "probability" of $\frac{3}{30} = \frac{1}{10} = 0.1$]

iii) The statistician could work out all the other theoretical probabilities for values 7 to 17 inclusive, scale them up to give expected frequencies (by multiplying by 30) and then do a χ^2 Goodness of Fit test.

11. a) i) outlier is point furthest from the linear relation, at the point (60, 21) (roughly)

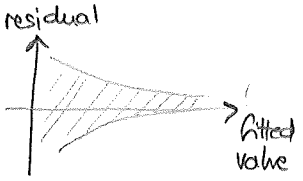
This equates to the Czech Republic.

ii) "correlation does not imply causation"

There may be an underlying factor to both these criteria that drives them both, rather than one directly affecting the other.

iii) There is a wide scatter of points for lower values of x (Welfare Generosity) and a tighter spread of points for larger values.

This would likely lead to a residual plot that looks like:



which would indicate that a transformation of data would be required.

b) $b = \frac{S_{xy}}{S_{xx}}$

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

$= 141677.3 - \frac{2208 \times 1054.8}{18}$

$= 12288.5$

so $b = \frac{S_{xy}}{S_{xx}} = \frac{12288.5}{105904} = 0.116034$

and $\bar{y} = \frac{1}{18} \times 1054.8 = 58.6$

$\bar{x} = \frac{1}{18} \times 2208 = 122.667$

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

$= 58.6 - 0.116034 \times 122.667$

$= 44.3665$

so $\underline{y_i = 44.3665 + 0.116034\, x_i}$

residual = observed − fitted

$= 42.6 - (44.3665 + 0.116034 \times 59)$

$= 42.6 - 51.2125$

$= -8.61248$

$= \underline{-8.61}$ (2dp)

12. a) i)

t-test for difference in population means.

we have $n_a = 10$ $\bar{x}_a = 12.5$ $s_a^2 = 1.53$

and $n_b = 8$ $\bar{x}_b = 14.0875$ $s_b^2 = 10.8241$

we assume - parent populations of weights are each normally distributed.
- parent populations of weight have the same standard deviations.

let A = weight of fish at location A $\quad A \sim N(\mu_a, \sigma_a^2)$
B = weight of fish at location B $\quad B \sim N(\mu_b, \sigma_b^2)$

$H_0 : \mu_a = \mu_b$

$H_1 : \mu_a \neq \mu_b$

We assume H_0 to be true. Two tailed test. $\alpha = 5\%$.

we pool samples to estimate $s^2 = \dfrac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}$

$= \dfrac{9 \times 1.53 + 7 \times 10.8241}{16}$

$= 5.59805$ (exactly, using raw data on Nspire and tTest_2Samp command)

test statistic, $t = \dfrac{\bar{x}_a - \bar{x}_b - (0)}{s\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}$

$= \dfrac{12.5 - 14.0875}{\sqrt{5.59805}\sqrt{\frac{1}{10} + \frac{1}{8}}}$

$= -1.4145$ (exactly, using tTest_2Samp)

p-value $= 2 \times P(t_{16} < -1.4145)$

$= 2 \times 0.08819$ from tCdf(-9E99, -1.4145, 16)

$= 0.176379$

> 0.05

Hence, we have no evidence to reject H_0, and we conclude that we do not have evidence that the mean weights of salmon in the two locations are different.

ii) we have $s_a^2 = 1.53$ and $s_b^2 = 10.8241$

These are very different in size, so the assumption of the parent populations of weights having equal standard deviations is in great doubt.

b) we have $V(A) = V(B) = 2.25$

we can perform a two sample z-test where $A \sim N(\mu_a, 2.25)$

$B \sim N(\mu_b, 2.25)$

$H_0: \mu_a = \mu_b$

$H_1: \mu_a \neq \mu_b$

Assume H_0 to be true. 2 sample test. $\alpha = 5\%$

test statistic, $z = \dfrac{\bar{x}_a - \bar{x}_b - (0)}{\sqrt{\frac{2.25}{10} + \frac{2.25}{8}}} = \dfrac{12.5 - 14.0875}{\sqrt{\frac{2.25}{10} + \frac{2.25}{8}}}$

$= -2.23116$

(and checked using zTest_2Samp command on TI-Nspire)

so p-value $= 2 \times P(Z < -2.23116)$

$= 2 \times 0.012835$ from normCdf(-9E99, -2.23116)

$= 0.02567$

< 0.05

So, we have evidence to reject H_0, and conclude that the mean weights of salmon in the two locations are different.

c) assumption for t-test: normality and equal st. deviations

assumption for z-test: known population variance.

so, a non-parametric test would avoid these.

Hence, use a Mann-Whitney rank sum test for non-paired data.

this has the underlying assumption that the populations have the same shape and spread only, with no requirement of normality

However, this assumption is questionable from the combined dot plot.....