2016 AH Statistics

1a)

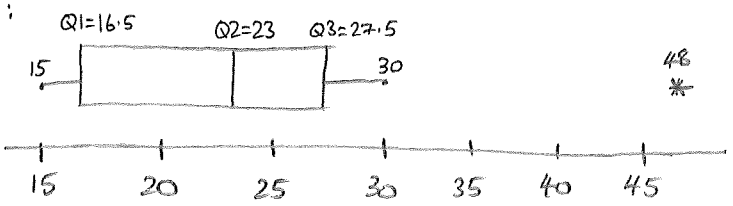
1	5 5 6 7
2	6 8 3 5 0 2 7
3	0
4	8

→

1	5 5 6 7
2	0 2 3 5 6 7 8
3	0
4	8

2|3 = 23

drawn as a boxplot:



b) IQR = Q3 − Q1
= 27·5 − 16·5
= 11

so upper fence = Q3 + 1·5 × IQR
= 27·5 + 1·5 × 11
= 44

So 48 is considered an outlier as 48 > 44.

2.

a) of the 100 patients, some will have been treated by the surgeons who've done 100's, and some by those who have done few.

As there are more patients who have been treated by the surgeons who have done 100's, then it is more likely that those surgeons will have their patients selected for the sample.

Hence, you would get a larger proportion of the surgeons who have done 100's patients.

Also, if you are looking at "quality of life", those surgeons who do 100's of operations, may be lots of small ailments, whereas those who do fewer operations may be much more technically demanding or serious. The patients' quality of life is likely connected to the complexity of their knee ailment.

b) You would do stratified random sampling.

You would divide all the patients up into 10 strata (one for each surgeon)

You would take a random sample from each of the 10 strata, whose size would be proportional to the number of patients in that strata.

This would guarantee that patients from each and every surgeon would be included in the sample of 100

c) Another cause of variation would be the seriousness of the knee ailment, which may itself be connected with the patients' age and health.

If we focus on their after-operation-care, then we need to look at the nurse provision, ward space and possible hospital food on offer!

3. X = weight of one honey jar

$n = 12$

sample mean, $\bar{x} = 147.8$

sample st.dev, $s_{n-1} = 2.379$.

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

$H_0: \mu = 150$

$H_1: \mu < 150$

Assume H_0 to be true

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

hence $X \sim N(150, \sigma^2)$

$\bar{X} \sim N\left(150, \frac{\sigma^2}{12}\right)$ where $\bar{X}$ = mean weight of 12 jars

$$\frac{\bar{X} - 150}{\sqrt{\frac{\sigma^2}{12}}} \sim N(0, 1^2)$$

$$\frac{\bar{X} - 150}{\sqrt{\frac{s^2}{12}}} \sim t_{11}$$ as we estimate σ^2 with s^2_{n-1}

so $t = \frac{147.8 - 150}{\sqrt{\frac{2.379^2}{12}}} = -3.20346$

$P(t_{11} < -3.20346) = 0.004201$

so p-value $= 0.004201 < 0.01$

Hence we reject H_0, and we have evidence to suggest that the mean weight of jars of honey is less than 150g.

b) It would be inappropriate to use the z-test as we have had to estimate the parent population's standard deviation from the small sample of 12 values. A z-test would have required us to know the parent population's standard deviation without the need for estimation.

4. a) i) the data has a positively skewed distribution, with most values between 5.2 years and 7.7 years, with possible outliers stretching to 12.8 and 13.7 years.

ii) mean $= \frac{1}{20} \times \Sigma x$

$= \frac{142}{20}$

$= 7.1$

$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$

$= \frac{1120.16 - \frac{(142)^2}{20}}{19}$

$= 5.89263$

$\Rightarrow s_{n-1} = 2.42747...$

$= 2.427$ (3dp)

b) let X = gannet age, from second biologist

from the first sample of 20, we estimate that $E(X) = 7.1$

$V(X) = 5.89263$.

So, by CLT, we can estimate that $\bar{X} \approx N\left(7.1, \frac{2.427^2}{20}\right)$

where $\bar{X}$ = mean age of gannets from second biologist

Hence $P(\bar{X} > 8.1) = P\left(Z > \frac{8.1 - 7.1}{\sqrt{\frac{2.427^2}{20}}}\right)$

$= P(Z > 1.8423)$

$= 0.032716$ from normcdf(1.8423, 9E99)

≈ 0.0327 (4dp)

5.

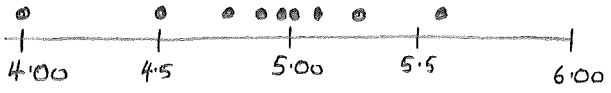
let X = yield from Gardener's Delight.

$E(X) = 4.75$.

$n = 9$.

a) i) we shall assume yields are distributed normally.

a quick dot plot shows that this assumption is plausible:



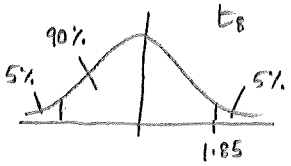
so $X \sim N(\mu, \sigma^2)$

$\Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{9}\right)$

we have to estimate σ^2 from $s_{n-1}^2 = 0.461844^2$, and estimate μ from $\bar{x} = 4.90$

as we estimate σ, we use a t_8 distribution

$t_{8,0.95} = 1.85955$



$\therefore$ 90% CI is $4.90 \pm 1.85955\sqrt{\frac{0.461844^2}{9}}$

$= (4.61373, 5.18627)$

$\approx (4.61, 5.19)$ to 2 dp.

ii) As this confidence interval contains 4.75, we can conclude that watering only once a week, rather than once a day, makes no difference to the yield. Had 4.75 been outwith the CI, we would have had evidence that less frequent watering would indeed have had an effect on yields.

b) The amount of sunlight, that each plant had, would have to be equitable, as well as any temperature fluctuations.

6. a) Months 2 to 10 inclusive are above the centre line, which is at least 8 consecutive points, falling foul of WECO rule #4.

b) X = data point is above the line

$X \sim B(14, 0.5)$ [$p = 0.5$, as the distribution should be symmetrical if process is in control]

$P(X = 12) = 0.005554$

$P(12 \text{ out of } 14 \text{ either above, or below line}) = 2 \times P(X = 12)$

$= 0.011108$

≈ 0.0111 (4dp)

c) change between months 14 and 15.

d) so $\hat{p} + 3\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.04$

so $3\sqrt{\frac{0.025 \times 0.975}{n}} = 0.015$

$\sqrt{\frac{0.025 \times 0.975}{n}} = 0.005$

$\frac{0.025 \times 0.975}{n} = 0.000025$

$n = \frac{0.025 \times 0.975}{0.000025}$

$n = 975$.

e) if $X \sim B(n, p)$ then you can approximate X with a Normal <u>if</u> $np > 5$, $nq > 5$

Here $n = 975$ and $p = 0.025$ $\Rightarrow$ $np = 24.375 > 5$

$nq = 950.625 > 5$

Thus conditions for a good approximation are satisfied.

so $X \sim B(975, 0.025)$ where X = no. of difficult recoveries per month

if Y is nomal approximation to X, then $Y \sim N(975 \times 0.025, 975 \times 0.025 \times 0.975)$

let $\frac{Y}{975}$ = proportion of difficult recoveries per month

$\frac{Y}{975} \sim N\left(0.025, \frac{0.025 \times 0.975}{975}\right)$

$\frac{Y}{975} \sim N(0.025, 0.000025)$

$P\left(\frac{Y}{975} > 0.03\right) = P\left(Z > \frac{0.03 - 0.025}{\sqrt{0.000025}}\right)$

$= P(Z > 1)$

$= 0.158655$

$\approx 15.9\%$

(oh! this is obvious from the control chart that 0.03 is one standard deviation above the mean of 0.025. Should have spotted that sooner)

7.

	male	female	
virus	40	10	50
non-virus	50	50	100
	90	60	150.

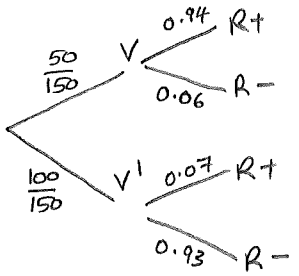
a) $P(\text{individual has virus}) = \frac{50}{150} = \frac{1}{3}$.

b) $P(\text{all 3 have disease}) = \frac{50}{150} \times \frac{49}{149} \times \frac{48}{148}$

$= \frac{196}{5513}$

≈ 0.035552

≈ 0.0356 (4dp)



V = has virus
R+ = react positively
R- = react negatively.

c) i) $P(R+) = P(R+|V)P(V) + P(R+|V')P(V')$

$= 0.94 \times \frac{50}{150} + 0.07 \times \frac{100}{150}$

$= 0.36$.

ii) $P(V|R+) = \frac{P(V \cap R+)}{P(R+)}$

$= \frac{\frac{50}{150} \times 0.94}{0.36}$.

$= 0.87037$

≈ 0.8704 (4dp)

8. a) $n_1 = 100$, $p_1 = 0.75$ } treated

$n_2 = 100$, $p_2 = 0.65$ } control

we set $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$= \frac{75 + 65}{200}$

$= \frac{140}{200}$

$H_0 : p_1 = p_2$

$H_1 : p_1 > p_2$

Assume H_0 to be true

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

test statistic, $z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$= \frac{0.75 - 0.65}{\sqrt{\frac{140}{200} \times \frac{60}{200}\left(\frac{1}{100} + \frac{1}{100}\right)}}$

$= 1.54303$

$P(Z > 1.54303) = 0.061411$

p-value $= 0.061411 > 0.05$

Hence we are not in 5% critical region, and we do not have sufficient evidence to reject H_0 and we conclude that the recovery rate proportions are equal, meaning that there is no evidence that the new drug has a higher recovery rate.

b) a new drug may do better, or worse, than an existing drug

Hence looking for a decrease as well as an increase in recovery rates would detect such a shortcoming of a new drug.

9.

X = no. injuries in 1 week

$X \sim Po(4)$

a) – injuries within any single week are independent of one another
– injuries from one week to the next happen at a constant rate

b) if $X \sim Po(4)$ then $E(X) = 4$
$V(X) = 4 \Rightarrow$ St. dev $= 2$

so $P(X > \mu + 2\sigma)$
$= P(X > 4 + 2\times 2)$
$= P(X > 8)$
$= P(X \geq 9)$
$= 1 - P(X \leq 8)$
$= 1 - 0.978637$
$= 0.021363$
$\approx \underline{0.0214}$ (4dp)

c) let $T = X_1 + X_2 + \ldots + X_{38}$ where $X_i \sim Po(4)$

so $T \sim Po(4\times 38)$
$T \sim Po(152)$ as weeks are all independent

so $P(T < 140)$ is what we want

approx T with $S \sim N(152, 152)$

so $P(T < 140) = P(S < 139.5)$ by continuity correction
$= P\left(Z < \frac{139.5 - 152}{\sqrt{152}}\right)$
$= P(Z < -1.01388)$
$= 0.155319$ by normcdf$(-9E99, -1.01388)$
$= \underline{0.1553}$ (4dp)

d) Sporting injuries often can arise from two players colliding, thereby injuring two players, which means that each injury is no longer independent of any other.

Also contact sports (in which most injuries occur) are not played in every week of the year, so there will be more injuries during, say, the rugby season than during the summer months. This compromises the assumption that the rate of injuries is constant over the session.

10. $E(X) = \frac{7}{4}$

$V(X) = \frac{3}{16}$

a)

x	1	2
$P(X=x)$	a	b.

where $a+b=1$

$E(X) = a+2b.$

$$\left.\begin{aligned} a+b &= 1 \quad (1) \\ a+2b &= \tfrac{7}{4} \quad (2) \\ \hline b &= \tfrac{3}{4} \quad (2)-(1). \end{aligned}\right\}$$

$\Rightarrow a = \frac{1}{4}$

So

x	1	2
$P(X=x)$	$\frac{1}{4}$	$\frac{3}{4}$

b) $E(Y) = \sum y P(Y=y)$

$= 1 \times \frac{2}{5} + 2 \times \frac{3}{5}$

$= \frac{1}{5}(2+6)$

$= \frac{8}{5}.$

$E(Y^2) = \sum y^2 P(Y=y)$

$= 1 \times \frac{2}{5} + 4 \times \frac{3}{5}$

$= \frac{1}{5}(2+12)$

$= \frac{14}{5}$

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

$= \frac{14}{5} - \left(\frac{8}{5}\right)^2$

$= \frac{70}{25} - \frac{64}{25}$

$= \frac{6}{25}.$

c) $E(3X-Y)$

$= E(3X) - E(Y)$

$= 3E(X) - E(Y)$

$= 3 \cdot \frac{7}{4} - \frac{8}{5}$

$= \frac{21}{4} - \frac{8}{5}$

$= \frac{105-32}{20}$

$= \frac{73}{20}.$

$(= 3.65)$

$V(3X-Y)$

$= 3^2 Var(X) + Var(Y)$

$= 9 \times \frac{3}{16} + \frac{6}{25}$

$= \frac{27}{16} + \frac{6}{25}$

$= \frac{771}{400}$

$(= 1.9275)$

11. $n=12$.

a) $H_0: \mu_{cross} = \mu_{SELF}$

$H_1: \mu_{cross} > \mu_{SELF}$

Assume H_0 to be true.

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

assuming difference normally distributed $\Rightarrow$ t-test as we will have to estimate st. deviation of difference.

So let X = difference (cross - self)

positive values of X support H_1 and go against H_0.

sample mean, $\bar{x} = 1.175$

$s_{n-1} = 4.27171$

test statistic, $t = \frac{\bar{x} - 0}{\sqrt{\frac{s_{n-1}^2}{12}}} = \frac{1.175}{\sqrt{\frac{4.27171^2}{12}}} = 0.952855$

p-value $= P(t_{11} > 0.952855)$

$= 0.180566$ from tCdf(0.952855, 9E99, 11)

> 0.05

Hence we have insufficient evidence to reject H_0 and we conclude that we do not have evidence that the mean height of cross-polinated is more than self-polinated.

b) We would do the Wilcoxon Signed Rank test for paired data.

This assumes that the distribution of differences is symmetrical

Diff	6.1	-8.4	1	0	2	2.9	3.5	5.1	1.8	3.1	3.0	-6.0
|Diff|	6.1	8.4	1		2	2.9	3.5	5.1	1.8	3.1	3	6.
rank	10	11	1		3	4	7	8	2	6	5	9

$W_- = 11+9 = 20$

$W_+ = 10+1+\ldots+6+5 = 46.$

let $W = \min(W_-, W_+) = 20$

H_0: median difference $= 0$ where difference = CROSS-SELF.

H_1: median difference > 0.

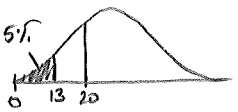
we want $P(W \leq 20)$ for $n = 11$

From tables; $P(W \leq 13) = 0.05$

$P(W \leq 10) = 0.025$

$P(W \leq 7) = 0.01$

Hence $W = 20$ is outside the 5% critical region, and we again have no reason to reject H_0, as in part (a)



Both tests give the same conclusion, as they both took into account the relative sizes of the difference in heights.

c) The paired study allowed you control over many other variables in their growing conditions thereby allowing more meaningful comparisons to be made between the two plants.

12. a) there appears to be a positive correlation between the two measures.
It is possible that a linear relationship exists.

b) $n = 70$
$\Sigma x = 12.0345$
$\Sigma y = 179.9185$
$S_{xx} = 58.3111$
$S_{yy} = 26.1816$
$S_{xy} = 15.6348$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{15.6348}{\sqrt{58.3111 \times 26.1816}} = 0.400146$$
$$\approx 0.400 \text{ (3dp)}$$

This measures the strength of a linear relationship between the two variables.
Its value is quite low, so it is not a strong linear relationship.

The coefficient of determination, $r^2 = 0.16$, which means that any linear relationship would only account for 16% of the variability in the data.

c) $H_0 : \rho = 0$
$H_1 : \rho \neq 0$
assume H_0 to be true
$\alpha = 5\%$
$r = 0.400146$
test statistic, $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.400146 \times \sqrt{68}}{\sqrt{1-0.400146^2}} = 3.60051$

p-value $= 2 \times P(t_{68} > 3.60051)$
$= 2 \times 0.000299$
$= 0.000599$
< 0.05

or $2 \times P(Z > 3.60051)$
$= 2 \times 0.000159$
$= 0.000318$
< 0.05

Hence we are well inside a 5% critical region and we have evidence to reject H_0 and conclude the linear correlation coefficient is significantly different from zero.

d) In part (c) the low value of r was statistically significant due to the large sample size, $n = 70$. A smaller sample with a similar value of r would not have done so well under the test.

Final conclusion: there is an association between Attractiveness and Performance. In order to better establish if it's a linear association, it would be wise to calculate a least squares regression line and examine a residual plot. This would add weight to the nature of the association.

e) Even if you did fit a linear regression line to the data, the low level of correlation would mean that it would not be reliable for prediction purposes, and thus not useful.