June 2009 - Part B (Statistics)

(Q4) a) i) $n = 20 \rightarrow m = \frac{1}{2}\left(x_{(10)} + x_{(11)}\right)$

$= \frac{1}{2}(2.03 + 2.05) = 2.04 \ (\mu)$

$q_1 = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = \frac{1}{2}(1.27 + 1.40) = 1.335 \ (\mu)$

$q_3 = \frac{1}{2}\left(x_{(15)} + x_{(16)}\right) = \frac{1}{2}(2.24 + 2.31) = 2.275 \ (\mu)$

↳ 5-number summary = $\{1.06, 1.335, 2.04, 2.275, 2.64\}$

ii) $iqr = q_3 - q_1 = 0.94$

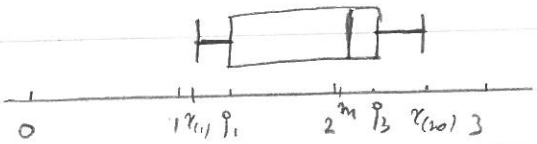
→ admissible range = $[q_1 - 1.5 \times iqr, \ q_3 + 1.5 \times iqr]$

$= [1.335 - 1.5 \times 0.94, \ 2.275 + 1.5 \times 0.94]$

$= [-0.075, \ 3.685]$

→ no outliers

iii)



→ fairly symmetric, no outliers

iv) Assuming the normality of the observations, we can build the two-sided t-confidence interval for μ:

$\left[\bar{x} \pm t_{n-1,\,1-\alpha/2}\frac{s}{\sqrt{n}}\right] = \left[1.8810 \pm 1.729 \times \frac{0.5239}{\sqrt{20}}\right]$

$t_{19,\,0.975} = 1.729$ (table) $= [1.6785; \ 2.0835]$

a) we need a random sample (no real way of checking this assumption) and the population must be normal (here $n=20$, probably too small to rely on the CLT). We can check this by plotting a quantile plot of the observations.

b) n components, each working independently of each other

→ X, the number of working components out of n, is binomially distributed:

$$X \sim Bin(n, 0.9)$$

→ long-run proportion of time that a 3-out-of-5 system will function is

$$P(X \geq 3) \quad \text{for } X \sim Bin(5, 0.9)$$

$$= 1 - P(X \leq 2)$$

$$= 1 - 0.009 \quad \text{(table)}$$

$$= 0.991$$

(Q5) i) $* = 64 + 2 = 66$

that's the total number of degrees of freedom, which is $n-1$ (n being the total number of observations)

$$\rightarrow n = 67$$

ii) Define μ_1, μ_2, μ_3 the true heat rates for the three types of turbines

The ANOVA tests the null hypothesis

$H_0: \mu_1 = \mu_2 = \mu_3$ against H_a: not all means equal

We need the 3 samples to be random and independent samples. We need the observations to come from normal populations. We need the variance of the observations to be the same in each group.

iii) We observed a test statistic $f_0 = 15.74$
We know that the test statistic follows the $F_{k-1, n-k}$ distribution

→ reject H_0 if $f > f_{k-1, n-k; 1-\alpha} = f_{2, 64; 0.95} \overset{(table)}{=} 3.15$

⇒ REJECT H_0

The associated p-value is $P(X > 15.74)$ for $X \sim F_{2,64}$

According to the table, this probability is < 0.005 (by far)

In plain language: there is a significant difference between the mean heat rates for the different types of turbines.

b) topic not addressed in 2011

c) i) $X \sim N(0,1)$; $Y \sim N(0,1)$, independent

$\Rightarrow X + Y \sim N(0, 2)$

- 0: sum of the means
- 2: sum of the variances
- sum of N-r.v. remains a N-r.v.

ii) $P(X+Y<1) = P\left(Z < \frac{1-0}{\sqrt{2}}\right) = P(Z < 0.7071) \overset{(table)}{\simeq} 0.76$

June 2010 – Part B (Statistics)

(Q4) a) i) 5-number summary : {3, 8, 12, 17, 23}
(from the boxplot)

ii) the boxplot is symmetric without outlier.
the qqplot does not show any "strong" departure from the normality (the 10 dots are reasonably close to the straight line)
→ normality assumption is reasonable.

iii) Normal population → t-confidence interval

$$\left[\bar{x} \pm t_{n-1,\,1-\alpha/2}\ \frac{s}{\sqrt{n}}\right] : \left[12.2 \pm 3.250 \times \frac{6.8}{\sqrt{10}}\right]$$

$t_{9;\,0.995} = 3.250$ (table) $= [5.2114,\ 19.1886]$

iv) · random sample → no way of checking that
· normal population → previous qqplot shows that it is reasonable.

b) Chit-Chat

c) Let π be the true proportion of damaged bumper cars.

Out of 96 cars tested, we estimate π by $\hat{p} = \frac{12}{96} = 0.125$

We know that, by the CLT,

$$\sqrt{n}\ \frac{(\hat{p} - \pi)}{\sqrt{\hat{p}(1-\hat{p})}} \overset{\cdot}{\sim} N(0,1)$$

$$\rightarrow \text{C.I. for } \pi = \left[\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$1-\alpha = 0.98 \rightarrow 1-\alpha/2 = 0.99$

$z_{0.99} = 2.33$ (table)

$$\rightarrow \left[0.125 \pm 2.33 \times \sqrt{\frac{0.125 \times 0.875}{96}} \right] = [0.046; 0.204]$$

ii) $n\pi(1-\pi) > 5$ for the CLT to be valid

$\rightarrow 96 \times 0.125 \times 0.875 = 10.5$

$\rightarrow$ n is certainly large enough for CLT to be valid.

Q5 i) $A = \frac{75081}{3} = 25027$

$B = \frac{235424}{16} = 14714$

(Check: $\frac{A}{B}$ must be equal to $f = 1.7$)

$$\frac{25027}{14714} = 1.7 \quad \checkmark$$

ii) Denote $\mu_1, \mu_2, \mu_3, \mu_4$ the true mean number of km for each type of plugs.

ANOVA test: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

against: H_a: not all the means are equal

Assumptions:
- random samples
- independent samples
- normal population for each group
- same variance in each group

iii) $\alpha = 0.01$ Observed value of the test statistic $f = 1.7$

We know that the test statistic follows the $F_{k-1, n-k}$ distribution

↳ reject H_0 if $f_0 > f_{k-1, n-k, 1-\alpha} = f_{3, 16, 0.95}$

$\triangleq 5.40$ (table)

$\Rightarrow$ no reject of H_0

Associated p-value = $P(X > 1.7)$ for $X \sim F_{3,16}$

$p > 0.05$ (table)

$\Rightarrow$ Conclusion: there is no significant difference in the mean number of driving km for the different types of plugs.

b) X = opening altitude

$\sim N(200, 35)$

i) $P(X < 100) = P\left(Z < \frac{100 - 200}{35}\right)$

$= P(Z < -2.857)$

$= 0.0021$ (table)

ii) Y = # of parachutes damaged out of 5

$\sim Bin(5; 0.0021)$

$P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1 - 0.0021)^5$

$= 0.0105$

c) i) $X \sim P(2) \rightarrow P(X=1) = e^{-2} \times 2 = 0.2707$

ii) $Y \sim Exp(1) \rightarrow P(Y<1) = 1 - e^{-1} = 0.6321$

iii) $P((X=1) \cup (Y<1))$

$= P(X=1) + P(Y<1) - P((X=1) \cap (Y<1))$

$= P(X=1) + P(Y<1) - P(X=1) \times P(Y<1)$ ← ind.

$= 0.2707 + 0.6321 - 0.2707 \times 0.6321$

$= 0.7317$

(Q6) a) i) $H_0: \beta_1 = 0$ $\quad H_a: \beta_1 \neq 0$

ii) (1) test statistic: $\sqrt{s_{xx}} \frac{\hat{b}_1}{s} = -7.80$

(follows the t_{n-2} distribution)

Here $n = 30 \rightarrow t_{28}$

(2) $p = 2 \times P(T > |-7.801|) = 2 \times P(T > 7.80)$

for $T \sim t_{28}$

Hence p-value = 0 (output)

(3) → we reject H_0

→ ppv has a significant effect on the response Ratio.

b) $\hat{\rho} = -\sqrt{r^2} = -\sqrt{0.685} = -0.828$

(negative as the slope is negative)

c) the change in the mean of Ratio for a unit increase in ppv is the slope β_1

↳ CI for $\beta_1 = \left[\hat{b}_1 \pm t_{n-2,\, 1-\alpha/2} \frac{s}{\sqrt{s_{xx}}} \right]$

$t_{28;\, 0.975} = 2.048$ (table)

$\hat{b}_1 = -0.00001484$

$\frac{s}{\sqrt{s_{xx}}} = 0.0000019$ } (output)

$\rightarrow \left[-0.00001484 \pm 2.048 \times 0.0000019 \right]$

$= \left[-1.873, -1.095 \right] \times 10^{-5}$

d) i) that's the interval that we are 95 % confident to find the true straight line in, at ppv = 750
(position of $\mu_{Y|X=750}$)

ii) the prediction interval is the interval that we are 95 % confident to find the next observation in, if ppv is set to 750
(position of Y_{n+1})

iii) the best estimation of $\mu_{Y|X=x} = \hat{y}(x)$

the best prediction of the next value of Y when $X = x$ is also $\hat{y}(x)$

↳ both intervals are centred at this value $\hat{y}(750) = 0.988$

e) (1) the error terms ε_i are independent
(2) they are from a normal population
(3) they have common variance

(1) → ok with plot of residuals against observation order
(no trend)

(2) → ok with qq-plot of the residuals

(3) → ok with plot of residuals against fitted values
(constant variability around 0,
no fan-like shape etc)