8135

THE UNIVERSITY OF SYDNEY

FACULTIES OF ARTS AND SCIENCE

**STAT 2012: Statistical Tests (Normal)**

Semester 2, 2011 Time allowed: Two hours

---

**Solution to examination**

1. One sample:

   (a) (3 marks) The signs of $d_i = x_i - y_i$ are

   $$-0.2\ (-)\ 1.5\ (+)\ 1.4\ (+)\ 0.7\ (+)\ -0.6\ (-)\ 0.5\ (+)\ 0.0\ (\text{drop})$$

   The sign test for the reduction in COHb after smoking low-tar cigarettes is

   1. **Hypotheses:** $H_0:\ \mu_d = 0$ against $H_1:\ \mu_d > 0$.
   2. **Test statistic:** $X = 4$ $(m = 6)$.
   3. **Assumption:** Distribution of $d_i = x_i - y_i$ is symmetric. Then $X \sim \mathcal{B}(6, 0.5)$ under $H_0$.
   4. ***P*-value:** $\Pr(X \geq 4) = \Pr(X \leq 2) = 0.3438$ (bin. table; $n$=6, $p$=0.5, $x$=2)
   5. **Decision:** Since the $p$-value$> 0.05$, the data are consistent with $H_0$. The reduction in mean blood COHb level after smoking *low-tar* cigarettes is not significant.

   (b) (2 marks) $z$-test: the rejection region is

   $$\bar{d} \ \geq \ \mu_d + z_{0.05}\frac{\sigma}{\sqrt{n}} = 0 + 1.645\frac{1}{\sqrt{7}} = 0.6217$$

   Hence $k = 0.6217$. Since $\bar{d} = 0.4714$ lies outside the rejection region, $H_0$ is accepted.

(c) (3 marks) $z$-test: the type II error under $H_1$ when $\mu = 1$ is

$$\begin{aligned}
\beta(1) &= \Pr(\text{type II error}) \\
&= \Pr(\text{Accept } H_0 \mid H_0 \text{ is false and } \mu = 1) \\
&= \Pr(\bar{d} < 0.6217 |\ \bar{X} \sim \mathcal{N}(1, 1/7)) \\
&= \Pr\left(z < \frac{0.6217 - 1}{\sqrt{1/7}}\right) \\
&= \Pr(z < -1.001) = 1 - 0.8413 = 0.1584
\end{aligned}$$

(d) (2 marks) $z$-test: the sample size is

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 1 \Rightarrow n = (2z_{1-\alpha/2}\sigma)^2 = [2(1.96)]^2 = 15.36583528$$

Take $n = 16$.

(e) (3 marks) The *two-sided* 95% confidence interval for the variance of $d_i$ is

$$\left(s^2\,\frac{n-1}{\chi^2_{n-1,1-\alpha/2}},\ s^2\,\frac{n-1}{\chi^2_{n-1,\alpha/2}}\right) = \left(0.6324\,\frac{7-1}{14.449},\ 0.6324\,\frac{7-1}{1.237}\right) = (0.2626,\ 3.0665)$$

Since the CI includes 1, the variance does not differ significantly from 1.

2. Two or three samples:

(a) (8 marks) 2-sample Wilcoxon rank sum test:

(i) We have $n_x = 7$, $n_y = 6$ and $N = n_x + n_y = 13$. The ranks in the combined sample are

A: -0.2 (2) 1.5 (8.5) 1.4 (7) 0.7 (5) -0.6 (1) 0.5 (4) 0.0 (3)

B: 1.5 (8.5) 2.7 (13) 2.6 (12) 1.3 (6) 1.8 (11) 1.7 (10)

Since there are ties, normal approximation should be used and so there is NO need to just base on sample of lower size. Hence

$$\begin{aligned}
W &= 8.5 + 13 + 12 + 6 + 11 + 10 = 60.5 \text{ for B} \quad \text{or } 91 - 60.5 = 30.5 \text{ for A} \\
E(W) &= \frac{1}{2}n_y(n_x + n_y + 1) = \frac{1}{2}6(7 + 6 + 1) = 42 \text{ for B} \quad \text{or } \frac{1}{2}7(14) = 49 \text{ for A}
\end{aligned}$$

$$\begin{aligned}
\sum_i r_i^2 &= 2^2 + 8.5^2 + 7^2 + 5^2 + 1^2 + 4^2 + 3^2 + 8.5^2 + 13^2 + 12^2 + 6^2 + 11^2 + 10^2 = 818.5 \\
Var(W) &= \frac{n_x \times n_y}{N(N-1)} \left( \sum_i r_i^2 - \frac{1}{4} N(N+1)^2 \right) = \frac{7(6)}{13(12)} \left( 818.5 - \frac{1}{4} 13 \times 14^2 \right) = 48.8654 \\
p\text{-value} &= \Pr(W \geq 60.5) = \Pr\left( Z \geq \frac{60.5 - 42}{\sqrt{48.8654}} \right) = \Pr(Z \geq 2.6465) = 1 - 0.9959 = 0.0041 \\
&\overset{or}{=} \Pr(W \leq 30.5) = \Pr\left( Z \leq \frac{30.5 - 49}{\sqrt{48.8654}} \right) = \Pr(Z \leq -2.6465) = 0.0041 \text{ for B}
\end{aligned}$$

(ii) Condition on the observed ranks for brand B, the cases with $W \geq 60.5$ are :

| Rank of $Y$ | $W$ | Times |
|---|---|---|
| 8.5, 8.5, 10, 11, 12, 13 | 63 | once |
| 7, 8.5, 10, 11, 12, 13 | 61.5 | twice |
| 6, 8.5, 10, 11, 12, 13 | 60.5 | twice |

Under $H_0$,

there are $\begin{pmatrix} N \\ n_y \end{pmatrix} = \begin{pmatrix} 13 \\ 6 \end{pmatrix} = \frac{13 \times 12 \times 11 \times 10 \times 9 \times 8}{6 \times 5 \times 4 \times 3 \times 2} = 1716$ possible cases.

Thus $\Pr(W \geq 60.5) = \frac{5}{1716} = 0.002913753$

(b) (5 marks) Three sample KW test:

| Brand | Reduction in COHb | | | | | | Mean $\bar{r}_i$ |
|---|---|---|---|---|---|---|---|
| B | 1.5 | 2.7 | 2.6 | 1.3 | 1.8 | 1.7 | |
| Rank | 4 | 13 | 12 | 3 | 7 | 6 | 7.5 |
| C | 1.6 | 2.4 | 0.7 | 1.2 | 2.0 | | |
| Rank | 5 | 10 | 1 | 2 | 9 | | 5.4 |
| D | 3.2 | 2.5 | 3.0 | 1.9 | 3.6 | | |
| Rank | 15 | 11 | 14 | 8 | 16 | | 12.8 |

We have $N = 16$, $\bar{r} = \frac{1}{2}(1 + 16) = 8.5$. There are no ties. The KW test for the equality of mean reduction in COHb level across the three brands of low-tar cigarettes is

1. **Hypotheses:** $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_1$ : Not all the $\mu_j$'s are equal.

2. **Test statistic:**

$$\begin{aligned} k_0 &= \frac{12}{N(N+1)}\left[n_1 \times \bar{r}_1^2 + n_2 \times \bar{r}_2^2 + n_3 \times r_3^2\right] - 3(N+1) \\ &= \frac{12}{16(17)}\left[6 \times 7.5^2 + 5 \times 5.4^2 + 5 \times 12.8^2\right] - 3(17) = 6.4632 \end{aligned}$$

3. **Assumption:** Same distribution of $Y_{ij}$ in each group $i$. We have $k_0 \sim \chi^2_{g-1}$ under $H_0$.

4. **P-value:** $p$-value $= \Pr(\chi^2_2 \geq k_0) = \Pr(\chi^2_2 \geq 6.4632) < 0.05$ (0.03949 from R).

5. **Decision:** Since $p$-value $< 0.05$, we reject $H_0$. There is strong evidence in the data against $H_0$. The mean reduction in COHb for the three brands of low-tar cigarettes are not all equal.

3. Two-way data without replicates:

The number of data $n = 15$, the number of blocks $r = 5$ and the number of treatments $c = 3$ .

(a) (6 marks)

$$\begin{aligned} CM &= n\bar{y}^2 = 15(25.467^2) = 9728.267 \\ SST_o &= \sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}^2 - n\overline{y}^2 = 10104 - 9728.267 = 375.733 \\ SST &= r\sum_{j=1}^{c} \overline{y}_{\cdot j}^2 - n\overline{y}^2 = 5(26^2 + 28.6^2 + 21.8^2) - 9728.267 = 117.733 \\ SSB &= c\sum_{i=1}^{r} \overline{y}_{i\cdot}^2 - n\overline{y}^2 = 3(24^2 + 23.667^2 + 20.667^2 + 29^2 + 30^2) - 9728.267 \\ &= 184.4 \\ SSR &= SST_o - SST - SSB = 375.733 - 117.733 - 184.4 = 73.6 \end{aligned}$$

The ANOVA table for two-way data without replicate is

**ANOVA table**

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Treatments (City) | 2 | 117.733 | $\frac{117.733}{2} = 58.867$ | $\frac{58.867}{9.2} = 6.3986$ |
| Blocks (Size) | 4 | 184.4 | $\frac{184.4}{4} = 46.1$ | $\frac{46.1}{9.2} = 5.0109$ |
| Residuals | 8 | 73.6 | $\frac{73.6}{8} = 9.2$ | |
| Total | 14 | 375.733 | | |

(b) (4 marks) This is a *randomized block design*. The two-way ANOVA test for city effects is

1. **Hypothesis:** $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs $H_1$ : Not all $\beta_j$ are the same.
2. **Test statistic:**

$$f_{t0} = \frac{SST/(c-1)}{SSR/(r-1)(c-1)} = \frac{117.733/2}{73.6/8} = 6.3986$$

3. **Assumption:** $Y_{ij} \sim \mathcal{N}(\mu + \alpha_i + \beta_j, \sigma^2)$ and $Y_{ij}$ are independent.
4. ***P*-value:**

$$p\text{-value} = \Pr(F_{2,8} \geq 6.3986) < 0.025 \quad \text{(R: 0.0219; Table: } F_{2,8,0.975} = 6.06).$$

5. **Decision:** Since the $p$-value for city effect $< 0.05$, we reject $H_0$. There is strong evidence in the data that the number of hours of television watching differs across the three cities.

(c) (3 marks) For a *completely randomized design*,

$$\begin{aligned} SSR_{new} &= SSB + SSR = 184.4 + 73.6 = 258 \\ f_{t01} &= \frac{SST/(c-1)}{SSR_{new}/(n-c)} = \frac{117.733/2}{258/12} = 2.7380 \\ p - \text{value} &= \Pr(F_{2,12} \geq 2.7380) > 0.1 \quad \text{(R: 0.1048; Table: } F_{2,12,0.900} = 2.81). \end{aligned}$$

Hence we accept $H_0$ that there is no difference in the numbers of hours of television watching across the three cities.

The insignificant city effect is due to the inflated MSR from 9.2 to 21.5 when the age effect is not accounted for in the one-way ANOVA test.

4. Regression analysis:

(a) (4 marks) Given

$$\sum_{i=1}^{n} x_i = 300, \quad \sum_{i=1}^{n} y_i = 1694, \quad n = 14,$$
$$\sum_{i=1}^{n} x_i^2 = 6572, \quad \sum_{i=1}^{n} y_i^2 = 207492, \quad \sum_{i=1}^{n} x_i y_i = 36763,$$

$$\begin{aligned}
S_{xx} &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2 = 6572 - \frac{1}{14}\,300^2 = 143.42857, \\
S_{xy} &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big) = 36763 - \frac{1}{14}(300)(1694) = 463, \\
\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{463}{143.42857} = 3.228088 \\
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = \frac{1694}{14} - 3.228088\,\frac{300}{14} = 51.826693.
\end{aligned}$$

Hence the fitted least squares line is

$$\hat{y} = \hat{\alpha} + \hat{\beta}\,x = 51.826693 + 3.228088\,x.$$

(b) (4 marks) The test for the regression model in (a) is

1. **Hypotheses:** $H_0$: $\beta = 0$ vs $H_1$ : $\beta \neq 0$.
2. **Test statistic:** $t_0 = \dfrac{\hat{\beta}}{\sqrt{\dfrac{s^2}{S_{xx}}}} = \dfrac{3.228088}{\sqrt{\dfrac{85.282952}{143.42857}}} = 4.186316346$, where

$$\begin{aligned}
S_{yy} &= \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 207492 - \frac{1694^2}{14} = 2518, \\
SSR &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 2518 - \frac{463^2}{143.42857} = 1023.395418 \\
s^2 &= \frac{SSR}{n-2} = \frac{1023.395418}{12} = 85.282952
\end{aligned}$$

3. **Assumption:** $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$. $Y_i$ are independent.
4. **P-value:** $p$-value $= 2\Pr(t_{12} > 4.186316346) < 0.002$ ($t_{12,0.999} = 3.930$)
5. **Decision:** Since $p$-value $< 0.05$, we reject $H_0$. There is strong evidence in the data that a linear relationship exists between $y$, weight of female at age 30, and $x$, weight at age 1.

(c) (2 marks) The coefficient of determination

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{463^2}{143.42857(2518)} = 0.593568142 = 59.4\%.$$

The model fit is just satisfactory.

(d) (6 marks) The predicted weight at age 30 when the weight at age 1 is $x_0 = 12$ lb:

$$\begin{aligned} \widehat{y}|\, x_0 = 12 &= \hat{\alpha} + \hat{\beta}x_0 = 51.826693 + 3.228088\,(12) = 90.56375. \\ \text{s.e.}(\widehat{y}|\, x_0 = 12) &= \sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \\ &= \sqrt{85.282952\left(1 + \frac{1}{14} + \frac{(12 - 21.42857)^2}{143.42857}\right)} = 12.00973. \end{aligned}$$

The 95% Prediction Interval for the weight of female at age 30 in lb when the weight at age 1 is $x_0 = 12$ lb:

$$\begin{aligned} &\left[(\hat{\alpha} + \hat{\beta}x_0) - t_{\alpha/2,n-2}\; s_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\; (\hat{\alpha} + \hat{\beta}x_0) + t_{\alpha/2,n-2}\; s_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right] \\ = \;& (90.56375 \;-\; 2.179 \times 12.00973,\; 90.56375 \;+\; 2.179 \times 12.00973) \\ = \;& (64.3968,\; 116.7307). \end{aligned}$$

The fitted regression model may not predict the weight of female at age 30 accurately because $x = 12$ lies outside the data range and the model fit is not good enough.

(e) (2 marks) Since when $x = \bar{x}$, $\hat{y} = \hat{\alpha} + \hat{\beta}\bar{x} = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}\bar{x} = \bar{y}$. Hence the regression line always passes through the point of sample means $(\bar{x},\ \bar{y})$.

5. Chi-square GOF test for normality:

(a) (5 marks) The expected frequencies:

| Interval | $O_i$ | Probability of event $\pi_i$ | $E_i = 120\pi_i$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|
| Less than 80 | 28 | $\Pr(Z < -1) = 0.1587$ | 120(.1587)=19.044 | $\frac{(28-19.044)^2}{19.044} = 4.212$ |
| 80 to 100 | 41 | $\Pr(-1 < Z < 0) = 0.3413$ | 120(.3413)=40.956 | $\frac{(41-40.956)^2}{40.956} = 0.000$ |
| 100 to 120 | 35 | $\Pr(0 < Z < 1) = 0.3413$ | 120(.3413)=40.956 | $\frac{(35-40.956)^2}{40.956} = 0.866$ |
| More than 120 | 16 | $\Pr(Z > 1) = 0.1587$ | 120(.1587)=19.044 | $\frac{(16-19.044)^2}{19.044} = 0.487$ |
| Sum | 120 | 1.0000 | 120 | 5.565 |

(b) (3 marks) The Chi-square test is

1. **Hypothesis:** $H_0$: $y_i \sim N(\mu, \sigma^2)$ vs $H_1$: $y_i$ do not follow $N(\mu, \sigma^2)$
2. **Test statistic:** $\chi_0^2 = \sum_{i=1}^{4} \frac{(y_i - 120\pi_i)^2}{120\pi_i} = 5.564571$
3. **P-value:** $\Pr(\chi_1^2 \geq 5.565) < 0.05 \quad (\chi_{1,0.95}^2 = 3.841)$
4. **Conclusion:** Reject $H_0$. There is strong evidence in the data against $H_0$. The data do not follow $N(\mu, \sigma^2)$.