Math Stat, solutions to HW5

Book problems.

12. By Proposition 7.5, the confidence interval is

$$\sqrt{\frac{(n-1)s^2}{x_2}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{x_1}} \quad (q)$$

where x_1 and x_2 are from Table 4 with DF$= n-1 = 29$ and columns $(1-q)/2$ and $(1+q)/2$. With $q = 90$ we get $x_1 = 17.71$ and $x_2 = 42.56$ and confidence interval $15.2 \le \sigma \le 23.6 \quad (0.90)$. With $q = 0.95$, we get $x_1 = 16.05$ and $x_2 = 45.72$ and confidence interval $14.7 \le \sigma \le 24.8 \quad (0.95)$.

13. The length of the confidence interval is $L = 2ts/\sqrt{n}$, which we note is a random variable because of s. We want to determine n such that $P(L \le \varepsilon\sigma) = q$ (larger n will obviously give a higher probability). Let

$$Y = \frac{(n-1)s^2}{\sigma^2}$$

so that $Y \sim \chi^2_{n-1}$ by Proposition 7.4. We get

$$P(L \le \varepsilon\sigma) = P\left(\frac{2ts}{\sqrt{n}} \le \varepsilon\sigma\right) = P\left(\frac{s}{\sigma} \le \frac{\varepsilon\sqrt{n}}{2t}\right)$$

$$= P\left(\frac{s^2}{\sigma^2} \le \frac{n\varepsilon^2}{4t^2}\right) = P\left(Y \le \frac{n(n-1)\varepsilon^2}{4t^2}\right) = F_{\chi^2_{n-1}}\left(\frac{n(n-1)\varepsilon^2}{4t^2}\right)$$

Practice problems.

1(a) The obvious estimator of $a\mu_X + b\mu_Y$ is $a\bar{X} + b\bar{Y}$ where

$$a\bar{X} + b\bar{Y} \sim N\left(a\mu_X + b\mu_Y, \sigma^2(\frac{a^2}{n} + \frac{b^2}{m})\right)$$

where we estimate σ^2 by s_p^2, as usual. Then

$$T = \frac{a\bar{X} + b\bar{Y} - (a\mu_X + b\mu_Y)}{s_p\sqrt{\frac{a^2}{n} + \frac{b^2}{m}}}$$

has a t_{n+m-2} distribution and for given q we can find t such that $P(-t \leq T \leq t) = q$ in the usual way. We get the confidence interval

$$a\mu_X + b\mu_Y = a\bar{X} + b\bar{Y} \pm t s_p \sqrt{\frac{a^2}{n} + \frac{b^2}{m}} \; (q)$$

(b) Let X values be Norwegian salmon and let Y values be Canadian salmon. To have Norwegian salmon "twice as large" means that $\mu_1 = 2\mu_2$, that is, $\mu_1 - 2\mu_2 = 0$ so we use the confidence interval in **(a)** with $a = 1$ and $b = -2$. We have $n = 4, m = 3, \bar{X} = 39.8, \bar{Y} = 18.6, s_X^2 = 10.2$, and $s_Y^2 = 16.8$. The pooled sample variance is

$$s_p = \frac{3 \cdot 10.2 + 2 \cdot 16.8}{5} = 12.8$$

which gives $s_p = 3.6$. With $q = 0.95$, we get $t = 2.57$ and the confidence interval becomes

$$\mu_1 - 2\mu_2 = 39.8 - 2 \cdot 18.6 \pm 3.6 \cdot 2.57 \cdot \sqrt{\frac{1}{4} + \frac{4}{3}} = 2.6 \pm 11.6 \;\; (0.95)$$

and as 0 is in the interval, there is no support for either claim.

5. Solve $(4 \cdot 9 + (m-1) \cdot 4)/(3 + m) = 6$ to get $m = 7$.

6. The estimators are

$$\widehat{b} = \frac{S_{xY} - S_x S_Y/n}{S_{xx} - S_x^2/n} = \frac{460.92 - 58.4 \cdot 46.8/6}{573.8 - 58.4^2/6} = 1.0$$

and

$$\widehat{a} = \bar{Y} - \widehat{b}\bar{x} = 7.8 - 1.0 \cdot 9.7 = -1.9$$

which gives the estimated regression line $y = -1.9 + x$. For the confidence intervals, we also need to estimate the variance σ^2:

$$s^2 = \frac{1}{n-2} \sum_{k=1}^{n} (Y_k - \widehat{a} - \widehat{b}x_k)^2 = \frac{1.15}{4} = 0.29$$

which gives $s = \sqrt{0.29} = 0.54$. With $q = 0.95$, we get $(1+q)/2 = 0.975$ and as DF$= n - 2 = 4$, we get $t = 2.78$ and the confidence intervals

$$a = \widehat{a} \pm \frac{ts}{\sqrt{n - \frac{S_x^2}{S_{xx}}}} = -1.9 \pm \frac{2.78 \cdot 0.54}{\sqrt{6 - \frac{58.4^2}{573.8}}} = -1.9 \pm 6.33 \quad (0.95)$$

$$b = \widehat{b} \pm \frac{ts}{\sqrt{S_{xx} - \frac{S_x^2}{n}}} = 1.0 \pm \frac{2.78 \cdot 0.54}{\sqrt{573.8 - \frac{58.4^2}{6}}} = 1.0 \pm 0.28 \quad (0.95)$$

7. $\widehat{p} = 523/1000 = 0.523$ and ME $= 1.96\sqrt{0.523(1-0.523)/1000} \approx 0.03$. The confidence interval is $\widehat{p}\pm$ME which ranges from 49.3% to 55.3% and as this includes 50%, there is no support for his claim.

8. The 95% confidence interval is

$$p_2 - p_1 = \widehat{p}_2 - \widehat{p}_1 \pm 1.96\sqrt{\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2} + \frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}} \quad (95\%)$$

where $\widehat{p}_2 = 523/1000 = 0.523$ (second poll) and $\widehat{p}_1 = 410/900 = 0.456$ which gives $p_2 - p_1 = 0.067 \pm 0.045(95\%)$. As this interval is entirely above 0, there is support for his claim.

9(a) We have $\widehat{p} = 0.505$ and with $q = 0.95$ we get $z = 1.96$ and the interval 0.505 ± 0.031 (95%).

(b) With $q = 0.50$, we get $(1+q)/2 = 0.75$ and $z = 0.67$ which gives the interval 0.505 ± 0.01 (50%).

(c) The confidence interval must be at most 0.505 ± 0.005 which gives the equation $0.005 = z\dfrac{0.505 \cdot 0.495}{\sqrt{1000}}$ which gives $z = 0.32$. As $\Phi(0.32) = 0.63 = (1+q)/2$, we get $q = 0.26$, a 26% confidence level.

Turn-in problems.

1. We can write T as

$$T = \frac{Z}{\sqrt{U/r}}$$

where $Z \sim N(0,1)$, $U \sim \chi_r^2$, and Z and U are independent. Hence

$$T^2 = \frac{Z^2}{U/r} = \frac{Z^2/1}{U/r}$$

where $Z^2 \sim \chi_1^2$. Hence, T^2 has an F distribution with $m = 1$ and $n = r$.

2(a) $\widehat{p} = 20/100 = 0.20$ which gives the confidence interval

$$p = 0.20 \pm 1.96\sqrt{\frac{0.20 \cdot 0.80}{100}} = 0.20 \pm 0.08 \quad (0.95)$$

(b) The equation to solve for n is ME $= 0.03$ in $p = \widehat{p} \pm$ME. The "worst-case" method amounts to setting $\widehat{p} = 0.5$ and as the confidence level $q = 0.99$ gives $z = 2.58$, we get the equation $2.58\sqrt{0.5(1-0.5)/n} = 0.03$ which gives $n = 1850$ (rounded up).

Alternatively, use the estimate from **(a)** to set $\widehat{p} = 0.20$ which gives the equation $1.96\sqrt{0.20(1-0.20)/n} = 0.03$ which gives $n = 1184$ (rounded up).

3(a) The 2006 poll has $n = 600$ and $\widehat{p} = 312/600 = 0.52$ and the 95% confidence interval for the true proportion p of supporters is

$$p = \widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0.52 \pm 0.04 \quad (95\%)$$

which is the interval $[48\%, 56\%]$. As it is entirely below the election result of 57%, we conclude that the decreased support is statistically significant.

(b) The 2005 poll had a proportion of supporters equal to $235/413 = 0.57$ which is the same as the election result. However, we need to take into account the uncertainty of this poll and cannot repeat the argument from **(a)**. Instead, compare the true proportions p_1 and p_2 at the times of the two polls and compute the confidence interval for the difference $p_1 - p_2$. We have $n_1 = 413, \widehat{p}_1 = 0.57, n_2 = 600$ and $\widehat{p}_2 = 0.52$ which gives the 95% confidence interval

$$p_1 - p_2 = \widehat{p}_1 - \widehat{p}_2 \pm 1.96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} = 0.05 \pm 0.06 \quad (95\%)$$

and as the interval includes 0, we cannot conclude that the actual support has gone down.

4(a) Since $Y_k \sim N(bx_k, \sigma^2)$, the likelihood function is

$$L(b) = \prod_{k=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_k - bx_k)^2} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(Y_k - bx_k)^2}$$

We can now take the logarithm, differentiate with respect to b and set the derivative equal to 0, or simply note that maximum of $L(b)$ is attained when the sum in the exponent is minimized. Thus, let

$$S(b) = \sum_{k=1}^{n}(Y_k - bx_k)^2$$

which we differentiate and set equal to 0:

$$\frac{d}{db}S(b) = 2\sum_{k=1}^{n}(Y_k - bx_k) \cdot (-x_k) = -2(S_{xY} - bS_{xx}) = 0$$

which gives the MLE

$$\widehat{b} = \frac{S_{xY}}{S_{xx}}$$

(b) For Hubble's data we have $S_{xY} = 12519$ and $S_{xx} = 29.5$ which gives $\widehat{b} = 12519/29.5 \approx 424$ and the estimated regression line $y = 424x$.

5(a) $H_0 : p = 0.05 \quad vs. \quad H_A : p > 0.05$

(b) $H_0 : p = 0.40 \quad vs. \quad H_A : p > 0.40$

(c) $H_0 : p = 0.5 \quad vs. \quad H_A :\neq 0.5$