Mathematics 376 – Probability and Statistics 2
Midterm Examination 2 – Solutions
April 21, 2006

I. (20) Let $Y_1, \ldots, Y_n$ be independent samples from the distribution with pdf containing the unknown parameter θ:

$$f(y|\theta) = \begin{cases} \frac{1}{\theta} y^{1/\theta - 1} & \text{if } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Determine the maximum likelihood estimator for θ using the Y_i.

Solution: The log-likelihood function is:

$$\begin{aligned} \ln(L) &= \ln\left(\frac{1}{\theta^n}(y_1 \ldots y_n)^{1/\theta - 1}\right) \\ &= -n\ln(\theta) + (1/\theta - 1)\sum_i \ln(y_i) \end{aligned}$$

So

$$\begin{aligned} 0 = \frac{d}{d\theta}\ln(L) &= \frac{-n}{\theta} - \frac{1}{\theta^2}\sum_i \ln(y_i) \\ \Rightarrow \widehat{\theta} &= \frac{-\sum_i \ln(y_i)}{n} \end{aligned}$$

(It can be checked that this is a maximum for the likelihood by the second derivative test.)

II. A machine shop manufactures toggle levers. Let d be the proportion of defectives in the shop's output. A random sample of size $n = 150$ toggle levers produced 6 defectives.

A) (10) Find a 95% confidence interval for d using this information.

Solution: Since $n = 150$ we use the large sample formula, and $z_{.05/2} = 1.96$. The confidence interval is

$$\begin{aligned} d &= \frac{6}{150} \pm 1.96\sqrt{\frac{(6/150)(144/150)}{150}} \\ &= .04 \pm .03136 \end{aligned}$$

(Note that this interval contains only positive numbers.)

B) (10) Test the hypothesis $H_0 : d = .1$ versus $H_a : d \neq .1$ using this data. Take $\alpha = .02$ (probability of Type I error). Also give the p-value of your test.

Solution: Again, since $n = 150$, we use a z-test. The test statistic is

$$z = \frac{.04 - .10}{\sqrt{\frac{(.1)(.9)}{150}}} = -2.45$$

The rejection region for the two-tailed test (H_a is $d \neq .1$) is

$$RR = \{z : |z| > z_{.01} = 2.33\}$$

So we reject H_0 at the $\alpha = .02$ level. The p-value is $p = 2(.0071) = .0142$ from the standard normal table with $z = 2.45$.

C) (10) For the test in part B, what is β (probability of Type II error) if the true value of d is .03?

Solution: We find β as follows from $z = \frac{Y/n - .01}{\sqrt{\frac{(.1)(.9)}{150}}}$ as above, using the fact that d is really .03 (so Y/n is approximately normal with mean .03 and variance $\frac{(.03)(.97)}{150}$):

$$\begin{aligned}
\beta &= P(TS \notin RR \mid d = .03) \\
&= P(-2.33 < TS < 2.33 \mid d = .03) \\
&= P(.0429 < Y/n < .1571) \\
&= P\left(\frac{.0429 - .03}{\sqrt{\frac{(.03)(.97)}{150}}} < z < \frac{.1571 + .03}{\sqrt{\frac{(.03)(.97)}{150}}}\right) \\
&= P(.9281 < z < 9.144) \\
&= P(z > .9281) - P(z > 9.144) \\
&\doteq .1762 - 0 \\
&= .1762
\end{aligned}$$

III. Consider the following measurements of the heat-producing capacity of the oil produced by two fields (in billions of calories per ton):

Field 1 :	8.26	8.13	8.35	8.07	8.34	
Field 2 :	7.95	7.89	7.90	8.14	7.92	7.84

A) (10) Estimate the *variance* σ^2 of the heat producing capacity of the oil from each field.

Solution: Use the usual unbiased estimator formula for the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

This gives

$$s_1^2 \doteq .01575 \qquad s_2^2 \doteq .01092$$

B) (10) Test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative $H_a : \sigma_1^2 \neq \sigma_2^2$ at the $\alpha = .05$ level. State your conclusion clearly and succinctly.

Solution: To test equality of variances, we use an F-test. The test statistic is $f = s_1^2/s_2^2 \doteq 1.44$. The rejection region for a test with $\alpha = .05$ is

$$\{f < f_{.975}(4,5)\} \cup \{f > f_{.025}(4,5)\}$$

From the F-table we have $f_{.025}(4,5) = 7.39$ and $f_{.975}(4,5) = 1/f_{.025}(5,4) = 1/9.36 = .1068$. 1.44 is not in the rejection region, so there is not enough evidence here to indicate that the variances are different.

C) (15) Construct a two-sided 95% confidence interval for the difference of the population mean heat producing capacities $\mu_1 - \mu_2$.

Solution: (Under assumption variances equal) we use the pooled estimator

$$s_p^2 = \frac{(4)(.01575) + (5)(.01092)}{9} = .01307.$$

Then the confidence interval is

$$\begin{aligned}\overline{Y_1} - \overline{Y_2} \pm t_{.025}(9)\sqrt{s_p^2(1/5 + 1/6)} &= 8.23 - 7.94 \pm (2.262)\sqrt{(.01307)(1/5 + 1/6)} \\ &= .29 \pm .1566\end{aligned}$$

IV. (15) Let (x_i, y_i), $i = 1, \ldots, n$ be a collection of data points. Using the matrix formulation, derive the normal equations for the least squares estimators for the coefficients β_0, β_1 in the model $Y = \beta_0 + \beta_1 x + \varepsilon$ fitting the data.

Solution: We have

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

So the normal equations are

$$X^t X \beta = X^t Y$$

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \beta = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

Extra Credit (10) In the situation of question III part C above, suppose you had $n = 50 = m$ measurements from each oil field instead of $n_1 = 5$ and $n_2 = 6$. What, if anything, would change in the method you would use? Explain.

Solution: With 50 measurements from each oil field, you could use the large sample formula. The $t_{\alpha/2}$ would be replaced by $z_{\alpha/2}$. Moreover, for the large sample test, it is not necessary to assume that the individual variances are equal, so you do not need to use the pooled estimator for the variance:

$$\overline{Y_1} - \overline{Y_2} \pm z_{.025}\sqrt{s_1^2/50 + s_2^2/50}$$