Mathematics 376 – Probability and Statistics II
Solutions for Final Examination
May 8, 2010

General Note: A "random sample" always consists of independent measurements from an indicated distribution.

I. Let Y_1, Y_2, Y_3, Y_4, Y_5 be a random sample from a normal distribution with mean $\mu = 4$ and standard deviation $\sigma = 10$.

A) What is the distribution of $\overline{Y} = \frac{1}{5}(Y_1 + Y_2 + Y_3 + Y_4 + Y_5)$? Why?

Solution: Any linear combination of normal random variables has a normal distribution by results from last semester. The expected value of $\overline{Y}$ is

$$E(\overline{Y}) = \frac{1}{5}(E(Y_1) + \cdots + E(Y_5)) = \frac{1}{5}(4 + \cdots + 4) = 4.$$

By the independence assumption, the variance is

$$V(\overline{Y}) = \frac{1}{25}(V(Y_1) + \cdots + V(Y_5)) = \frac{1}{25}(100 + \cdots + 100) = 20.$$

Therefore $\overline{Y} \sim N(4, 2\sqrt{5})$.

B) What is the distribution of $U = \frac{(Y_1-4)^2+(Y_2-4)^2+(Y_3-4)^2}{100}$? Why?

Solution: Each $\frac{Y_i-4}{10}$ has a standard normal distribution, so $\left(\frac{Y_i-4}{10}\right)^2$ has a χ^2 distribution with one degree of freedom. Then U is the sum of three independent χ^2-distributed random variables, so U also has a χ^2 distribution with 3 degrees of freedom.

C) What is the distribution of $V = \frac{\sqrt{3}(Y_4-4)}{10\sqrt{U}}$, where U is as in part B? Why?

Solution: We can rearrange the formula for V like this:

$$V = \frac{\frac{Y_4-4}{10}}{\sqrt{\frac{U}{3}}}$$

Since Y_4 and U are independent, this has the form of a random variable with a t-distribution with 3 degrees of freedom.

D) How would you determine the PDF for the sample maximum $Y_{(5)}$? (Note: The CDF for a normal random variable is not an elementary function; just give a "recipe" for how it might be computed.)

Solution: The PDF for each Y_i is

$$f(y) = \frac{1}{\sqrt{200\pi}} e^{-(y-4)^2/200}$$

The CDF is

$$F(y) = \int_{-\infty}^{y} f(t)\ dt.$$

(This is not an elementary function.) The PDF for the sample minimum is then

$$f_{(5)}(y) = 5(F(y))^4 f(y).$$

II. A random variable Y is said to have a *log-normal* distribution with parameters μ, σ if its pdf has the form

$$f(y) = \begin{cases} \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-(\ln(y)-\mu)^2/(2\sigma^2)} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

A) Compute $E(\ln(Y))$ for a log-normal random variable. (Hint: Set up the integral, then change variables.)

Solution: By the definition,

$$E(\ln(Y)) = \int_0^\infty \ln(y) \cdot \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-(\ln(y)-\mu)^2/(2\sigma^2)}\ dy$$

To evaluate the integral, let $u = \ln(y)$ then $du = \frac{1}{y}\ dy$. The limits of integration become $-\infty$ to ∞ in the new variable, so

$$\begin{aligned} E(\ln(Y)) &= \int_{-\infty}^\infty u \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(u-\mu)^2/(2\sigma^2)}\ du \\ &= \mu, \end{aligned}$$

since the integral in terms of u just computes the expected value of a $N(\mu, \sigma)$ random variable.

B) Let $Y_1, \ldots, Y_n$ be a random sample from a log-normal distribution with unknown μ and known $\sigma = 1$. Find the maximum-likelihood estimator for μ.

Solution: The likelihood is

$$\begin{aligned} L(y_1, \ldots, y_n) &= \prod_{i=1}^n \frac{1}{y_i\sqrt{2\pi}} e^{-(\ln(y_i)-\mu)^2/2} \\ &= \frac{1}{y_1 \cdot \cdots \cdot y_n} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n (\ln(y_i)-\mu)^2/2} \end{aligned}$$

The log-likelihood function is

$$
\begin{aligned}
\ln(L) &= -\sum_{i=1}^{n} \ln(y_i) - n\ln(\sqrt{2\pi}) - \sum_{i=1}^{n}(\ln(y_i) - \mu)^2/2 \\
\Rightarrow \frac{d}{d\mu}\ln(L) &= \sum_{i=1}^{n}(\ln(y_i) - \mu) \\
&= \sum_{i=1}^{n} \ln(y_i) - n\mu
\end{aligned}
$$

Setting this equal to 0 and solving for μ, we see there is a unique critical point at

$$\widehat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} \ln(y_i)$$

This is a maximum because the second derivative of $\ln(L)$ with respect to μ is

$$\frac{d^2\ln(L)}{d\mu} = -n < 0.$$

C) Is your estimator from part B biased or unbiased? Why?

Solution: The estimator is unbiased because of the result from part A:

$$
\begin{aligned}
E(\widehat{\mu}_{ML}) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\ln(Y_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(\ln(Y_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \mu.
\end{aligned}
$$

III. Solid copper produced by melting powdered ore is tested for "porosity" (the volume fraction due to air bubbles). A sample of $n_1 = 40$ porosity measurements made in one lab has $\overline{y}_1 = .25$ and $s_1^2 = .001$. A second set of $n_2 = 50$ measurements is made using identical ore in a second lab, yielding $\overline{y}_2 = .17$ and $s_2^2 = .002$. Find a 95% confidence interval for $\mu_1 - \mu_2$, the difference of the population mean porosity measurements from the two labs.

Solution: We can use the large-sample formulas here because the numbers of samples in each group is > 30. The confidence interval is

$$
\begin{aligned}
\mu_1 - \mu_2 &= \overline{y_1} - \overline{y_2} \pm z_{.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
&= .25 - .17 \pm 1.96\sqrt{\frac{.001}{40} + \frac{.002}{50}} \\
&= .08 \pm .0158
\end{aligned}
$$

The confidence interval contains only numbers greater than zero, so there is evidence to say that $\mu_1 > \mu_2$.

IV. The Rockwell hardness index for steel is determined by pressing a diamond point into the metal with a specified force and measuring the depth of penetration. Out of a random sample of 50 ingots of a certain grade of steel from one manufacturer, the Rockwell hardness index was greater than 62 in 31 of the ingots.

A) The manufacturer claims that at least 75% of the ingots of this type of it produces steel will have Rockwell hardness index greater than 62. Is there sufficient evidence to refute this claim? Use a test at the $\alpha = .01$ level.

Solution: This will be a large-sample (z-) test of $H_0 : p = .75$ vs. $H_a : p < .75$. The rejection region is $RR = \{z \mid z < -2.33\}$. The test statistic is computed from $\frac{31}{50} = .62$.

$$z = \frac{.62 - .75}{\sqrt{\frac{(.75)(.25)}{50}}} = -2.12.$$

This is not negative enough to reject H_0 at $\alpha = .01$.

B) Using the rejection region you found in part A, compute the Type II error probability β of your test if it is actually true that 65% of the ingots have hardness index greater than 62.

Solution: The Type II error probability β is the chance that we do not reject H_0 when $p = .65$. This will happen when

$$\frac{(Y/50) - .75}{\sqrt{\frac{(.75)(.25)}{50}}} > -2.33$$

or

$$(Y/50) > .75 - 2.33\sqrt{\frac{(.75)(.25)}{50}} = 0.6073.$$

If p is actually .65, the probability that this happens is

$$P\left(\frac{(Y/50) - .65}{\sqrt{\frac{(.65)(.35)}{50}}} > \frac{.6073 - .65}{\sqrt{\frac{(.65)(.35)}{50}}}\right)$$

which we estimate as

$$P(Z > -.63) = 1 - P(Z > .63) = 1 - .2643 = .7357.$$

V. Consider the following measurements of the weights of yields of two breeds of apple trees (in kilograms):

Breed 1 :	80.6	80.3	81.5	80.7	80.4	
Breed 2 :	79.5	79.9	81.0	79.4	79.2	81.4

Assume the measurements come from normal populations.

A) Estimate the *variances* σ_1^2 and σ_2^2 of the yields of the two breeds. Is there evidence to suspect that $\sigma_2^2 > \sigma_1^2$? Explain.

Solution: The population variances are estimated by the two sample variances:

$$s_1^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - \overline{x})^2 = .225$$

$$s_2^2 = \frac{1}{5}\sum_{i=1}^{6}(y_i - \overline{y})^2 \doteq .8387.$$

We can test $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_a : \sigma_2^2 > \sigma_1^2$ using a upper-tail F-test. The test statistic is

$$F = \frac{s_2^2}{s_1^2} = \frac{.8387}{.225} \doteq 3.727.$$

From the F-table with 5 numerator and 4 denominator degrees of freedom, we see $F_{.1} = 4.05$. This means the p-value is $> .1$, which is usually too large to consider rejecting H_0.

B) Test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_a : \mu_1 \neq \mu_2$. Estimate the p-value of your test and state your conclusion clearly and succinctly.

Solution: Since we do not reject $\sigma_1^2 = \sigma_2^2$, we will use the small sample (t-) test for equality of means. The pooled estimator for the variance is:

$$S_p^2 = \frac{4(.225) + 5(.8387)}{9} \doteq .5659.$$

$$t = \frac{80.7 - 80.07}{\sqrt{.5659}\sqrt{\frac{1}{5} + \frac{1}{6}}} \doteq 1.383.$$

From the t-table with 9 degrees of freedom, we see that this is almost exactly equal to $t_{.1}$. Hence for the two-tail test, the p-value would be $p = .2$. There is not sufficient evidence to conclude that the two population means are different.

VI. The following table gives measurements of the amount of sodium chloride that dissolved in 100 grams of water at various temperatures in a chemistry experiment.

x (degrees C)	y (grams)
0	7.3
15	13.0
30	23.3
45	30.7
60	39.7
75	47.7

A) Find the equation of the least squares regression line for this data set.

Solution: The main steps in the calculation:

$$\begin{aligned}
\overline{x} &= 37.5 \\
\overline{y} &= 26.95 \\
S_{xx} &= 3937.5 \\
S_{xy} &= 2171.25 \\
\widehat{\beta_1} &= \frac{S_{xy}}{S_{xx}} \doteq .5514 \\
\widehat{\beta_0} &= \overline{y} - \widehat{\beta_1}\overline{x} \doteq 6.271
\end{aligned}$$

The least squares regression line is

$$y = .5514x + 6.271.$$

B) Is there sufficient evidence to say that $\beta_1 > .45$? Explain, using the p-value of an appropriate test.

Solution We test $H_0 : \beta_1 = .45$ vs. $H_1 : \beta_1 > .45$ with an upper tail t-test. The test statistic is

$$t = \frac{\widehat{\beta_1} - .45}{S\sqrt{c_{11}}}$$

We compute the components of this as follows:

$$\begin{aligned}
S^2 &= \frac{1}{n-2}(S_{yy} - \widehat{\beta_1}S_{xy}) = (1201.235 - (.5514)(2171.25))/4 \doteq .9864 \\
c_{11} &= \frac{1}{S_{xx}} \doteq .000254
\end{aligned}$$

So then

$$t = \frac{.5514 - .45}{\sqrt{.9864}\sqrt{.000254}} \doteq 6.406$$

From the t-table with 4 degrees of freedom, we see that $p < .005$. So there is strong evidence to reject H_0 and conclude that $\beta_1 > .45$.

VII. Does a "statistically significant" result where we reject some H_0 mean that H_0 is *far from being true*? Answer intuitively first. Then answer the following: Suppose each individual we draw from a population has either property A or property B (but not both). Let p be the proportion of the population that has property A. We want to test $H_0 : p = 1/2$ versus $H_a : p > 1/2$ with a large-sample test. Suppose that tests are done with $n = 100, 1000, 10000, 100000$. What must the observed proportion of sampled individuals with property A be in order to reject H_0 at the $\alpha = .05$ level in each case? How does this square with what you said at first?

Solution: We reject H_0 when

$$\frac{Y}{n} > .5 + 1.645\sqrt{\frac{(.5)(.5)}{n}}$$

With the given values of n this yields

n	$.5 + 1.645\sqrt{\frac{(.5)(.5)}{n}}$
100	.58225
1000	.526
10000	.508225
100000	.5026

Note what this is saying. For a large n like $n = 100000$, we reject H_0 for any observed fraction $> .5026$. This is not very far from $p = .5$(!) *Comment:* Observations like this have been used, in fact, to raise the question whether the theory of hypothesis testing as we have developed it is really appropriate for detecting meaningful differences in real world data(!)

Extra Credit. Recall that in the "Big Theorem" in the multiple regression case, we said that in the entries $c_{ij}\sigma^2$ of the covariance matrix of the least squares estimators, the c_{ij} were the entries of the matrix $(X^tX)^{-1}$. Show this is true by direct computation for the X from a simple linear model of the form $Y = \beta_0 + \beta_1 x + \epsilon$.

Solution: For the simple linear model, the X matrix has 1's in the first column and the $x_1, \ldots, x_n$ in the second. Hence

$$X^tX = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

For a 2×2 matrix, recall from Linear Algebra that there is a comparatively simple formula for the inverse matrix: If the determinant $ad - bc \neq 0$, then

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Applying that here,

$$\begin{aligned}(X^tX)^{-1} &= \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sum x_i^2}{nS_{xx}} & \frac{-\sum x_i}{nS_{xx}} \\ \frac{-\sum x_i}{nS_{xx}} & \frac{1}{S_{xx}} \end{pmatrix}\end{aligned}$$

This matches our formulas

$$c_{00} = \frac{\sum x_i^2}{nS_{xx}}$$

$$c_{01} = \frac{-\sum x_i}{nS_{xx}}$$

$$c_{11} = \frac{1}{S_{xx}}$$

(Recall that the rows and columns of these matrices are indexed starting from 0.)