Mathematics 376 – Probability and Statistics II
Midterm Exam I – Solutions
March 1, 2006

I. Assume that the calorie content of chocolate Power Bars is normally distributed. A random sample of 10 chocolate Power Bars yields calorie content values of:

$$210, 220, 234, 227, 229, 240, 221, 228, 225, 219$$

A) (10) Estimate the population mean calorie content of the chocolate Power Bars.

Solution: We estimate μ with the sample mean $\overline{Y} \doteq 225.3$

B) (10) The sample variance of the calorie content values above is $S^2 = 70.68$ cal. Find a 99% confidence interval for the population mean calorie content, μ.

Solution: Since $n = 10$, we use the small sample formula (using percentage points of the t-distribution with $n - 1 = 9$ degrees of freedom):

$$\begin{aligned} \mu &= 225.3 \pm t_{.005} \frac{\sqrt{70.68}}{\sqrt{8}} \\ &= 225.3 \pm (3.25) \frac{8.407}{\sqrt{10}} \\ &= 225.3 \pm 8.64 \end{aligned}$$

C) (10) Find a 95% confidence interval for the calorie content variance σ^2.

Solution: The 95% confidence interval is found using the percentage points of the χ^2 distribution with $n - 1 = 9$ degrees of freedom:

$$\begin{aligned} \sigma^2 &\in \left[\frac{9(70.68)}{\chi^2_{.025}}, \frac{9(70.68)}{\chi^2_{.975}} \right] \\ &= \left[\frac{9(70.68)}{19.0228}, \frac{9(70.68)}{2.70039} \right] \\ &= [33.44, 235.57] \end{aligned}$$

(Note: The fact that this interval is so large comes from the fact that the number of samples is quite small. It is difficult to pin down the population variance from this little data.)

II. A researcher found that 107 out of 200 rutabaga seeds germinated when they were planted in soil maintained at $5°C$, while 123 out of 200 seeds germinated when they were planted in soil maintained at $15°C$.

A) (20) Construct a 95% confidence interval for the difference in the proportions of seeds that germinate at the two temperatures.

Solution: Since $n = 200$ for the two samples, we use the large-sample formula for the difference of proportions (with the percentage points of the standard normal distribution). The point estimators for the two proportions are $\hat{p}_{5^\circ} = 107/200 = .535$ and $\hat{p}_{15^\circ} = 123/200 = .615$. Then

$$p_{5^\circ} - p_{15^\circ} = .535 - .615 \pm (1.96)\sqrt{\frac{(.535)(.465)}{200} + (.615)(.385)/200} \doteq -0.08 \pm .0966$$

B) (10) Can we rely on the results of this study to conclude that a greater proportion of rutabaga seeds will germinate at the higher temperature in general? Why or why not?

Solution: No – the data does not support that conclusion because the confidence interval contains zero and positive numbers as well as negative numbers. The fact that $\hat{p}_{5^\circ} < \hat{p}_{15^\circ}$ might be a result of random variation.

III. (15) Let $Y_1, \ldots, Y_n$ denote a random sample from a population whose density is given by

$$f(y) = \begin{cases} 3\beta^3 y^{-4} & \text{if } y \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

where β is unknown. Consider the estimator $\widehat{\beta} = \min(Y_1, \ldots, Y_n)$. Is $\widehat{\beta}$ biased or unbiased?

Solution: We need to compute the expected value of $\widehat{\beta} = Y_{(1)}$, so we need the pdf for the sample minimum in this case. The general formula is $f_{(1)}(y) = n(1-F(y))^{n-1} f(y)$, where $F(y)$ is the cdf for the distribution.

$$F(y) = \int 3\beta^3 y^{-4}\, dy = -\beta^3 y^{-3} + c$$

(for $y \geq \beta$, 0 otherwise). The constant c must be chosen to make this a valid cdf. Since $\lim_{y \to +\infty} F(y) = 1$, we see $c = 1$. Then

$$f_{(1)}(y) = n(\beta^3 y^{-3})^{n-1} \cdot 3\beta^3 y^{-4}$$

and

$$\begin{aligned} E(Y_{(1)}) &= \int_\beta^\infty y \cdot n(\beta^3 y^{-3})^{n-1} \cdot 3\beta^3 y^{-4}\, dy \\ &= 3n\beta^{3n} \int_\beta^\infty y^{-3n}\, dy \\ &= \frac{3n}{3n-1}\beta \end{aligned}$$

Since we do not get exactly β, this is a *biased* estimator.

IV. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be two independent random samples of equal sizes from a normal distribution with *known* mean μ, but *unknown* variance σ^2. Let Σ_X^2 and Σ_Y^2 be the statistics:

$$\Sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \qquad \Sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$$

A) (10) What is the distribution of $n\Sigma_X^2/\sigma^2$? Explain.

Solution: We see

$$n\Sigma_X^2/\sigma^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

is the sum of squares of n independent standard normals. Hence this has a χ^2 distribution with n degrees of freedom. (It's *not* $n-1$ degrees of freedom here because we know the exact population μ and we use that rather than the sample mean $\overline{X}$.)

B) (5) Use your answer from part A to show that Σ_X^2 is an unbiased estimator for σ^2.

Solution: Recall that $E(Z^2) = V(Z) + (E(Z))^2 = 1$. So a χ^2 random variable with n degrees of freedom has expected value n. Hence

$$\frac{n}{\sigma^2}E(\Sigma_X^2) = E\left(n\Sigma_X^2/\sigma^2\right) = n \Rightarrow E(\Sigma_X^2) = \sigma^2,$$

so this is an unbiased estimator for σ^2. (Once again, this is different from the case we discussed in class because we have μ rather than $\overline{X}$.)

C) (10) What is the distribution of Σ_X^2/Σ_Y^2? Explain.

Solution: Rewriting the quotient as

$$\Sigma_X^2/\Sigma_Y^2 = \frac{\frac{n\Sigma_X^2}{\sigma^2}/n}{\frac{n\Sigma_Y^2}{\sigma^2}/n}$$

we see that Σ_X^2/Σ_Y^2 has an $F(n,n)$ distribution.

Extra Credit (10) The *median* of a sample is the "middle" value, that is the value with the property that half the sample values are greater than or equal to the median and half are less than or equal to the median. Give formulas for computing the sample median in terms of the *order statistics* of the sample. (Note: the cases n even and n odd are different.)

Solution: When n is odd, the middle value is the order statistic $Y_{((n+1)/2)}$. When n is even, we take the average of the two order statistics on either side of the middle point:

$$\frac{1}{2}(Y_{(n/2)} + Y_{(n/2+1)})$$