Math 218 Mathematical Statistics

Final Exam

May 2009

Problem 1. [24; 6 points each part] Suppose that you're interested in determining whether a particular dangerous substance, like mercury or a pesticide, is entering the food chain and concentrating further along the food chain. You might sample the amounts of chemical residues in two different species—one predator species and one prey species—to see if the prey species has a lower concentration of the substance and the predator species that eats the prey species has a higher concentration of the substance.

a. Suppose that you are designing this experiment. Answer each of these questions with a sentence, or two at most: (1) Would this be a comparative or a noncomparative study? (2) What confounding factors might there be? (3). What can you do to assure that the sampling is random?

It's a comparative study. There are lots of compounding factors–age of captured animals, high or low tolerance levels for the substance, ability to break down the substance, the predator might get the the substance from some other source, these predators might be found primarily in places where the prey does or doesn't have access to the substance but the prey is sampled from a larger population. Ideally, perfect random sampling where each animal of a species has the same probability of being chosen can't be done. But choosing the animals from a variety of locations and times will go a long way toward making the sampling more random.

b. Determine what the null hypothesis H_0 and the alternative hypothesis H_1 should be. Should this be a one-sided test or a two-sided test; why? Explain why why you chose what you did to be H_0 instead of H_1.

The alternative hypothesis, H_1, should be what we expect to find, that the mean μ_1 of the substance per unit weight of the predator animals is greater than μ_2 of that of the prey animal. The null hypothesis is that it isn't, $H_0 : \mu_1 \leq \mu_2$, or $H_0 : \mu_1 = \mu_2$. This is a one-sided test.

c. The experiment will involve measuring the substance levels in n_1 animals of the predator species and n_2 animals in the prey species. Suppose that for economic reasons you can only have small samples of the two populations. What is the test statistic and the rejection region for the null hypothesis H_0 you gave in part b?

There's a T-test that does this for small populations. We have to assume that both populations have (approximately) normal distributions to use it, but for this test we don't need to assume they have the same variance.

$$T = \frac{X_1 - X_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

We reject $H - 0$ if $t > t_{\nu,\alpha}$ where

$$\nu = \frac{(w_1 + w_2)^2}{w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)}$$

and $w_1 = s_1^2/n_1$ and $w_2 = s_2^2/n_2$.

Perhaps a better experiment would be a matched pairs design. Predators and prey could each be captured from a several different locations, those from the same location being matched. As you can see from part d, however, that's not the kind of experiment we're considering.

d. Suppose that you test $n_1 = 10$ predators and $n_2 = 13$ prey and you find the mean concentration of the substance in the predator sample is $\overline{x}_1 = .041$ while that in the prey sample is $\overline{x}_2 = .026$, and the standard deviation for the predator sample is $s_1 = .017$ while that for the prey is $s_2 = .006$. Is H_0 rejected at the $\alpha = .05$ significance level?

$$\begin{aligned}
t &= \frac{x_1 - x_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \\
&= \frac{.041 - .026}{\sqrt{.017^2/10 + .006^2/13}} \\
&= \frac{.015}{.0056} = 2.68 \\
w_1 &= \frac{s_1^2}{n_1} = 0.000289 \\
w_2 &= \frac{s_2^2}{n_2} = 0.00000277 \\
\nu &= \frac{(w_1 + w_2)^2}{w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)} \\
&= \frac{1.003 \cdot 10^{-9}}{9.344 \cdot 10^{-11}} = 10.7 \\
t_{\nu,\alpha} &= t_{10.7,0.05} \approx t_{11,0.05} = 1.8
\end{aligned}$$

Since $t = 2.68$ is greater than $t_{\nu,\alpha} = 1.8$, reject H_0 and conclude that the predators have a higher level of the substance than the prey.

Problem 2. [24; 8 points each part] Suppose that the median sales prices for new single-family homes in a particular locale are as follows

year	median sales price in thousands
1992	207
1993	223
1994	235
1995	249
1996	264
1997	278
1998	295
1999	312

a. Given this data, find the least squares line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. (You may also want to graph the data and visually compare it to this line, but that's not required.)

If you graph the data, you'll see it's almost a straight line.

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{619.5}{42} = 14.75\\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x}\\ &= 257.875 - 14.75 \cdot 1995.5 = -29175\end{aligned}$$

Thus, the least squares line is $y = -29175 + 14.75x$.

b. Compute the error sum of squares (SSE), the total sum of squares (SST), and the regression sum of squares (SSR). Also find the coefficient of determination r^2 and the sample correlation coefficient r.

$$\begin{aligned}\text{SSE} &= 19.25\\ \text{SST} &= 9157\\ \text{SSR} &= \text{SST} - \text{SSE} = 9138\\ r^2 &= \frac{\text{SSR}}{\text{SST}} = 0.998\\ r &= \sqrt{0.997} = 0.997\end{aligned}$$

c. Test the significance of the linear relationship between year and median sales price using a significance level of $\alpha = .05$. (Refer to the section on analysis of variance for simple linear regression.)

You can use an F-test or a T-test. The statistic

$$F = \frac{\text{MSR}}{\text{MSE}} = T^2$$

has an F-distribution with 1 and $n-2$ degrees of freedom, and its square root T has a T-distribution with $n-2$ degrees of freedom. You can use either one to conclude that we can reject the hypothesis H_0 that $\beta_1 = 0$ (that there is no linear relationship between year and median sales price) at the $\alpha = .05$ significance level and conclude that there is, in fact, a linear relationship.

Note that the value of r was so extremely high in this problem that you might wonder if the data was really made up.

Problem 3. [24; 6 points each part] Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with mean λ. This distribution has the density function $f(x) = \frac{1}{x!}\lambda^x e^{-\lambda}$, for $x = 0, 1, 2, \ldots$. Furthermore, the mean of this Poisson distribution equals λ, and its variance also equals λ.

a. Write down the likelihood function

$$L(\lambda | x_1, \ldots, x_n) = f(x_1, \ldots, x_n | \lambda) = \prod_{i=1}^{n} f(x_i | \lambda),$$

then find its logarithm and simplify that.

You can do this using $\prod$ product notation and $\sum$ sum notation, but I'll write it up without those notations.

$$\begin{aligned}L(\lambda|x_1,\ldots,x_n) &= f(x_1,\ldots,x_n|\lambda)\\ &= f(x_1|\lambda) f(x_2|\lambda) \cdots f(x_n|\lambda)\\ &= \frac{\lambda^{x_1}e^{-\lambda}}{x_1!}\frac{\lambda^{x_2}e^{-\lambda}}{x_2!}\cdots\frac{\lambda^{x_n}e^{-\lambda}}{x_n!}\\ \ln L(\lambda|x_1,\ldots,x_n) &= (x_1 \ln\lambda - \lambda - \ln(x_1!))\\ &\quad + (x_2 \ln\lambda - \lambda - \ln(x_2!))\\ &\quad + \cdots\\ &\quad + (x_n \ln\lambda - \lambda - \ln(x_n!))\\ &= (x_1 + x_2 + \cdots + x_n)\ln\lambda - n\lambda\\ &\quad - (\ln(x_1!) + \cdots + \ln(x_n!))\end{aligned}$$

b. Compute the derivative of the logarithm of the likelihood function.

$$\frac{d}{d\lambda} L(\lambda|x_1,\ldots,x_n) = \frac{x_1 + x_2 + \cdots + x_n}{\lambda} - n$$

c. Determine the maximum likelihood estimator $\hat{\lambda}$ for λ by finding the critical point of the logarithm of the likelihood function.

The critical points will be where the derivative $\frac{d\lambda}{dL}$ equals 0. That's where

$$\frac{x_1 + x_2 + \cdots + x_n}{\lambda} - n = 0,$$

so when $\lambda = \frac{x_1 + x_2 + \cdots + x_n}{n} = \overline{x}$. Thus, the maximum likelihood estimator is $\hat{\lambda} = \overline{X}$, the sample mean.

d. Determine the mean and variance of $\hat{\lambda}$.

The mean $E(\overline{X})$ of the sample mean $\overline{X}$ equals the mean of the population distribution, and the population distribution is the Poisson distribution with mean $\mu = \lambda$. The variance

$\mathrm{Var}(\overline{X})$ of the sample mean $\overline{X}$ is $1/n$ times the variance of the population distribution. Since the variance of this Poisson distribution is λ, therefore the variance of $\overline{X}$ is λ/n.

Problem 4. [10] In your own words, briefly explain the difference between a 95% confidence interval as used by a classical statistician and a 95% probability interval as used by a Bayesian statistician.

The main point to make about the confidence interval of a classical statistician is that the probability that the parameter θ lies in the interval $[l, r]$ is either 0 or 1, not 0.95, but when many confidence intervals are estimated at the 95% confidence level, in the long run about 95% of the confidence intervals will, in fact, contain θ.

The main point to make about the probability interval of a Bayesian statistician is that the probability that the parameter θ lies in the interval $[l, r]$ actually is 0.95, but that depends on the assumption that the the prior distribution for the parameter θ was valid.

Problem 5. [18; 9 points each part] A publication for college students is planning to survey students to estimate the average amount students spend on living expenses, excluding rent, per month. The publication expects to sample a large number of students, at least $n = 40$, and has done surveys like this in the past and has the ability to select random samples of students.

a. Describe a process, which may depend on n, that will give a 95% confidence interval for the averge living expenses.

Since n is large, we may conclude, by the central limit theorem, that the sample mean $\overline{X}$ is approximately normally distributed with mean μ and variance σ^2/n where μ and σ^2 are the mean and variance of the population distribution. Furthermore, since n is large, the sample variance s^2 is a good approximation for the population variance σ^2. Therefore the endpoints of a 95% confidence interval can be taken to be

$$\overline{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $z_{\alpha/2} = z_{.025} = 1.96$.

b. Suppose that the publication wants the margin of error (half the length of the confidence interval) to no more than \$20. Explain how you could determine how large n has to be to achieve this margin of error.

What you want is for $1.96 \dfrac{\sigma}{\sqrt{n}} = 20$, that is,

$$n = \left(\frac{1.96}{20} \sigma \right)^2 \approx 0.01\sigma^2$$

The value of σ is approximated by s, and that's provided by the survey, but n has to be known before executing the survey. That's a problem.

This is a problem for almost every survey, and various solutions are used. Here's one. Survey 40 people first to get an estimate of σ, then use that to determine n to see how many more people to survey. Another solution is to have a prior estimate of the range of the average living expenses. In terms of classical statistics, one fourth of that range is about σ. In terms of Bayesian statistics, that information can be incorporated as a prior probability distribution. Either way, however, there's still the problem of estimating that range.