Math 218 Mathematical Statistics

First Test Answers

February 2016

Scale. 89–106 A, 73–88 B, 65–72 C. Median 87.

1. [15; 3 points each part] Data variables come in various types. Some are categorical (also called qualitative) while others are numerical. Those that are qualitative are either nominal or ordinal. Those that are numerical are either continuous or discrete. For each of the following examples, decide whether it is nominal, ordinal, continuous, or discrete.

a. Grade of meat: prime, choice, good.

There is a clear order here. Prime costs more than choice, choice more than good. So this is an ordinal scale.

b. Birthweight of newborns in grams.

If the number of grams has to be a whole number, then this is discrete, but if fractions are used, then continuous.

c. Car model (e.g. Ford Escort, or Mitsubishi Galant).

Nominal. The names themselves are what matter.

d. Time until a lightbulb burns out.

Continuous.

e. Days absent from class.

Discrete.

2. [16; 4 points each part] A study was done to evaluate the benefits of an exercise program to reduce mortality among patients who have survived a heart attack. Mortality among volunteers who enrolled in an exercise program was compared with that among controls selected from medical records of eligible patients who chose not to join the exercise program. Reasons given for not exercising included physician disapproval, shift work, and lack of interest. The study results showed that the patients in the exercise program had half as many deaths as those in the control group.

a. What is the response variable in this study?

Deaths, or mortality rate.

b. What is the explanatory variable in this study?

Participation in the program.

c. Name at least one confounding variable.

A confounding variable is an uncontrolled variable whose effects cannot be separated from the predictor variable. For example, willingness to exercise. Certainly those who aren't willing to exercise are less likely to volunteer. They may also be more likely to die soon even after surviving a heart attack. A more noticeable confounding variable is whether the physician recommended not enrolling in the program.

d. Use the statement above to give at least one explanation why the chosen control group is not appropriate and how it could influence study results.

The first one mentioned is that patients self selected to be or not to be in the program. Self selection nearly always introduces bias. Others can be tied to willingness to exercise, or physician recommendation not to participate.

3. [15; 3 points each part] True or false. Write the whole word "true" or the whole word "false."

a. A normal plot of data indicates that the data is normally distributed if the plot approximates a straight line.

True. That's the whole purpose of a normal plot.

b. Whereas treatment factors are controlled in an experiment, nuisance factors, also called noise factors, are all the other factors that might affect the response variable.

True. This defines when a factor is a nuisance factor.

c. Student's T-distribution is the primary distribution used to study Poisson processes.

False. They're unrelated.

d. To standardize a random variable, subtract its mean and divide by its standard deviation, $Y = (X - \mu_X)/\sigma_X$, for then $\mu_Y = 0$ and $\sigma_Y = 1$.

True.

e. Simple random sampling is where the population can be divided into homogeneous subpopulations and a small sample is drawn from each subpopulation resulting in a sample that is representative of the population.

False. This is stratified random sampling.

4. [20; 5 points each part] We have used the term "variance" in several ways.

a. Define the variance σ^2 of a random variable X.

The variance σ^2 is a number defined as $\sigma^2 = \mathrm{Var}(X) = E((X - \mu_X)^2$. It's a parameter, that is, a constant.

b. Define the variance S^2 of a sample $X_1, X_2, \ldots, X_n$.

The sample variance S^2 is a statistic defined as $S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})$ or $S^2 = \frac{1}{n}\sum(X_i - \overline{X})$. It's a statistic, that is, a function of the sample $X_1, X_2, \ldots, X_n$.

c. How is S^2 related to σ^2?

S^2 is an estimator of σ^2. When the $\frac{1}{n-1}$ is used, it's unbiased, too, that is, $E(S^2) = \sigma^2$, but when $\frac{1}{n}$ is used, the MSE is smaller. Also, as $n \to \infty$, $S^2 \to \sigma^2$.

d. How is the variance $\mathrm{Var}(\overline{X})$ of the sample mean related to either σ^2 or to S^2?

$\mathrm{Var}(\overline{X})$ is a number, specifically, $\mathrm{Var}(\overline{X}) = \frac{1}{n}\sigma^2$. That's because

$$\begin{aligned}\mathrm{Var}(\overline{X}) &= \mathrm{Var}(\frac{1}{n}\sum X_i) \\ &= \frac{1}{n^2}\sum \mathrm{Var}\, X_i \\ &= \frac{1}{n^2}\sum \sigma^2 = \frac{1}{n^2}(n\sigma^2 = \frac{1}{n}\sigma^2\end{aligned}$$

5. [20; 4 points each part] On maximum likelihood functions. Consider the family of exponential distributions parametrized by θ, a positive real number. The density function is

$$f(x\,|\,\theta) = \theta e^{-\theta x} \qquad \text{for } x \text{ positive.}$$

a. What is the joint density function $f(x_1, x_2, \ldots, x_n|\theta)$ for a random sample $X_1, X_2, \ldots, X_n$ from that distribution? (Assume all the x_i's are positive.)

$$f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_i \theta e^{-\theta x_i},$$

which, if you want to simplify it now, is

$$\theta^n\, e^{-\theta \sum x_i}.$$

b. Write down the likelihood function $L(\theta|x_1, x_2, \ldots, x_n)$?

$L(\theta|x_1, x_2, \ldots, x_n)$ is $f(x_1, x_2, \ldots, x_n|\theta)$, which is part a.

c. Express the log of the likelihood function, and simplify it.

$$\ln L(\theta|\mathbf{x}) = \ln \theta^n e^{-\theta \sum x_i} = n \ln \theta - \theta \sum x_i.$$

d. Take $\frac{d}{d\theta}$ of the log of the likelihood function.

$$\frac{d}{d\theta} \ln L(\theta|\mathbf{x}) = \frac{n}{\theta} - \sum x_i.$$

e. Use the work you've done so far to determine is the maximum likelihood estimator $\hat{\theta}$ for θ, that is, find the value of θ maximizes this likelihood for given $x_1, x_2, \ldots, x_n$?

$\frac{d}{d\theta} \ln L(\theta|\mathbf{x})$ is 0 when

$$\frac{n}{\theta} = \sum x_i,$$

that is, when

$$\theta = \frac{n}{\sum x_i} = \frac{1}{\overline{x}}.$$

Thus, $\hat{\theta}$ is $1/\overline{x}$.

6. [20] On estimators. Recall the geometric distribution with parameter p. Recall that when independent Bernoulli trials are repeated, each with probability p of success, then the number of trials X it takes to get the first success has a geometric distribution. The probability mass function for the geometric distribution is $f(x) = q^{x-1}p$, for $x = 1, 2, \ldots$. The mean of this geometric distribution is $\mu = 1/p$, and its variance is $\sigma^2 = (1-p)/p^2$.

a. [4] Given a random sample $X_1, X_2, \ldots, X_n$ from a geometric distribution with an unknown parameter p, let our estimator $\hat{\mu}$ for μ be the sample mean $\hat{\mu} = \overline{x}$. Determine the expectation $E(\hat{\mu})$ of this estimator.

It is always the case that $E(\overline{X}) = E(\frac{1}{n}\sum X_i) = \frac{1}{n}\sum \mu = \mu$, so $E(\overline{X}) = 1/p$.

b. [4] Is this $\hat{\mu}$ an unbiased estimator for μ?

Yes, since $E(\hat{\mu}) - \mu = 0$.

c. [4] Determine the variance $\mathrm{Var}(\hat{\mu})$ of this estimator.

$$\begin{aligned}\mathrm{Var}(\hat{\mu}) &= \mathrm{Var}\left(\frac{1}{n}\sum X_i\right) \\ &= \frac{1}{n^2}\sum \mathrm{Var}(X_i) \\ &= \frac{1}{n^2}\sum \sigma^2 \\ &= \frac{\sigma^2}{n} = \frac{q}{np^2}\end{aligned}$$

d. [8] Determine the mean squared error $\mathrm{MSE}(\hat{\mu})$ of this estimator.

The MSE is always the sum of the variance plus the bias^2, but the bias is 0 in this case, so it's just the variance which was found in part c.