

Math 466/566 - Homework 4

1. We want to test a hypothesis involving a population proportion. The unknown population proportion is p . The null hypothesis is $p = 1/2$ and the alternative hypothesis is $p > 1/2$. We want the level of the test to be 0.01, and we want the power of the test to be at least 0.9 when $p > 0.55$. Determine how large a sample we need and specify what the test is, i.e., when we accept the null hypothesis and when we reject it. Note that we worked out the power function for this setting of a one sided alternative in class. (It is also worked out in the book.) If you want to make the problem a little more challenging you can instead take the alternative hypothesis to be that $p \neq 1/2$.

466 Solution: The statistic we use is

$$Z = \frac{f - 0.5}{\sqrt{0.5(1 - 0.5)/n}} \quad (1)$$

For a level of 0.01 with a one-sided test, we use $P(Z > 2.33) = 0.01$ to see that the test should be that we reject the null hypothesis if $Z > 2.33$.

Over the range $p \geq 0.55$, the power will be smallest at $p = 0.55$. So we want the power at $p = 0.55$ to be 0.9. When $p = 0.55$, Z is not a standard normal RV anymore. The standard normal now is

$$Z' = \frac{f - 0.55}{\sqrt{0.55(1 - 0.55)/n}} \quad (2)$$

Note that $\sqrt{0.5(1 - 0.5)} = 0.5$ while $\sqrt{0.55(1 - 0.55)} = 0.4975$. So it is a good approximation to take Z' to be

$$Z' = \frac{f - 0.55}{\sqrt{0.5(1 - 0.5)/n}} \quad (3)$$

We reject the null hypothesis if $Z > 2.33$. In terms of Z' this corresponds to

$$P(Z > 2.33) = P(Z' > 2.33 + (0.5 - 0.55)2\sqrt{n}) = P(Z' > 2.33 - 0.1\sqrt{n}) \quad (4)$$

(5)

We want this to equal 0.9. Using tables of software we find $P(Z' > -1.28) = 0.9$. So

$$2.33 - 0.1\sqrt{n} = -1.28 \quad (6)$$

Solving for n we get n must be approximately 1300.

566 Solution: The 566 version had an alternative hypothesis of $p \neq 1/2$. So to get a level of 0.01, you need $c = 2.58$, i.e., we reject the null hypothesis if $|Z| > 2.58$. An analysis similar to the above than shows to get the desired power we must take n to be about 1490.

2. Problem 1 in chapter 5 in the book.

Solution: Follow the book's suggestion. The joint density is $f(x_1, x_2, \dots, x_n) = 1/n!$ when $x_1 < x_2 < x_3 < \dots < x_n$, and = 0 otherwise. We can compute the mean of $U_{(i)}$ using this joint density:

$$\begin{aligned} E[U_{(i)}] &= \int f(x_1, x_2, \dots, x_n) x_j dx_1 dx_2 \dots dx_n \\ &= \frac{1}{n!} \left(\int_0^1 \dots \left(\int_0^{x_5} \left(\int_0^{x_4} \left(\int_0^{x_3} \left(\int_0^{x_2} x_j dx_1 \right) dx_2 \right) dx_3 \right) dx_4 \right) \dots dx_n \right) \end{aligned}$$

First do the integral without the x_j :

$$\begin{aligned} &\frac{1}{n!} \left(\int_0^1 \dots \left(\int_0^{x_5} \left(\int_0^{x_4} \left(\int_0^{x_3} \left(\int_0^{x_2} dx_1 \right) dx_2 \right) dx_3 \right) dx_4 \right) \dots dx_n \right) \\ &\frac{1}{n!} \left(\int_0^1 \dots \left(\int_0^{x_5} \left(\int_0^{x_4} \left(\int_0^{x_3} x_2 dx_2 \right) dx_3 \right) dx_4 \right) \dots dx_n \right) \\ &\frac{1}{n!} \left(\int_0^1 \dots \left(\int_0^{x_5} \left(\int_0^{x_4} \frac{1}{2} x_3^2 dx_3 \right) dx_4 \right) \dots dx_n \right) \\ &\frac{1}{n!} \left(\int_0^1 \dots \left(\int_0^{x_5} \frac{1}{3!} x_4^3 dx_4 \right) \dots dx_n \right) \\ &\frac{1}{n!} \left(\int_0^1 \dots \left(\int_0^{x_6} \frac{1}{4!} x_5^4 dx_5 \right) \dots dx_n \right) = \dots = \frac{n!}{n!} \end{aligned} \tag{7}$$

Now consider what happens with the x_j in there. It will go along for the ride until we do the x_j integral. Then instead of $\int_0^{x_{j+1}} x_j^{j-1} dx_j$ we will have $\int_0^{x_{j+1}} x_j^j dx_j$. So we will skip over the factor of $1/j$ and generate a factor of $1/(j+1)$. So in the end we get

$$\frac{n!}{1 \cdot 2 \cdot 3 \cdots (j-1) \cdot (j+1) \cdots (n+1)} = \frac{j}{n+1}$$

Grading note: Any attempt at this problem got the full 5 points.

3. Problem 2 in chapter 5 in the book.

Solution: The Cauchy distribution does not have moments of any order. In particular the variance is infinite, so the theoretical variance of the sample mean is infinite. The variance of the sample median is approximately $1/(4f(m)^2n)$. Here $f(m) = c(0)/s = 1/(\pi s)$. So

$$\sigma_M^2 = \frac{100\pi^2}{4 \times 49} = 5.03$$

So the standard deviation of the sample median is 2.24. Theoretically, the sample median should be a better estimator of the median than the sample mean. For the data we find the sample mean is 89.49 while the sample median is 100.05. The sample median is indeed much closer to the population median of 100.

4. Problem 3 in chapter 5 in the book. If you are using R you will find the following functions useful. `pnorm()` computes the c.d.f. of a normal distribution. `sort()` puts a sample in increasing order. `(1:n-0.5)/n` will create an array with the numbers $(i - 0.5)/n$ for $i = 1, 2, \dots, n$.

Solution: The null hypothesis is that the population is normal with the given mean and variance. The alternative hypothesis is that it is not. As stated in class, $P(D > 1.36/\sqrt{n}) = 0.05$. So if we use a significance level of 5%, we will reject the null hypothesis if $D > 1.36/\sqrt{n} = 0.23$. Using R or your favorite software, you should find for the data set that the KS statistic $D = 0.1803$. So we accept the null hypothesis that the data is normal with the given mean and variance.

5. Suppose X (the population) is a continuous random variable with probability density function

$$f(x) = \frac{\lambda}{2} \exp(-\lambda|x - \mu|), \quad -\infty < x < \infty \quad (8)$$

where λ and μ are unknown parameters. The mean of X is μ and the density is symmetric about $x = \mu$, so the median is also μ . We are given a random sample X_1, X_2, \dots, X_n with n large. Determine whether the sample mean or the sample median is a better estimator for μ . By “better” I mean “has smaller variance.”

Solution: First we compute the population variance.

$$\begin{aligned} \text{var}(X) &= \frac{\lambda}{2} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-\lambda|x-\mu|} dx = \frac{\lambda}{2} \int_{-\infty}^{\infty} x^2 e^{-\lambda|x|} dx \\ &= \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx = \frac{2}{\lambda^2} \end{aligned} \tag{9}$$

Thus the variance of the sample mean is σ^2/n which is

$$\text{var}(\bar{X}_n) = \frac{2}{\lambda^2 n}$$

For the sample median the variance is

$$\text{var}(M_n) = \frac{1}{4f^2(\mu)n} = \frac{1}{\lambda^2 n} \tag{10}$$

So the sample median is the better estimator.

6. The random variable X is uniformly distributed on the interval $[0, \theta]$. (This is the population.) θ is an unknown parameter. We have a random sample X_1 of size 1. We want to use it to estimate the unknown parameter θ . Consider estimators of the form $T = cX_1$ where c is a constant.

(a) Find the value of c which makes this an unbiased estimator.

Solution: We want $E[T] = \theta$. $E[T] = cE[X_1] = c\theta/2$. So $c = 2$.

(b) Find the value of c which minimizes the mean square error. This is the risk when we take the loss function to be $(T - \theta)^2$.

Solution: $\text{var}(X_1) = \theta^2/12$, and the bias is $c\theta/2 - \theta$. So the MSE is

$$\frac{c^2\theta^2}{12} + \left(\frac{1}{2}c\theta - \theta\right)^2$$

Note that we can factor out a θ^2 , which implies that the minimizing c will not depend on θ . Minimize this as a function of c and you should get $c = 3/2$.

(c, 566 only) Find the value of c which makes this an unbiased estimator.

Solution: $E[X_{(n)}] = n/(n+1)$, so $c = (n+1)/n$.

(d, 566 only) Find the value of c which minimizes the mean square error. This is the risk when we take the loss function to be $(T - \theta)^2$.

Solution: Compute the mean square error as above using the formula from the book for the variance of $X_{(n)}$. You should get

$$c = \frac{(n+1)(n+2)}{1+n(n+2)}$$