Introduction to Probability and Statistics - 18.05

Test 3 - solutions

1. (a) Defininition of mean squared error: let T be an estimator for a parameter θ. The mean squared error of T is

$$MSE(T) = E[(T - \theta)^2]$$

   (b) The function $g(x) = 1/x$ is strictly convex (in the relevant range, i.e. $x > 0$). Also we know that $\bar{X}_n$ is an unbiased estimator for the expectation of the distribution which is $1/\lambda$, which means that $1/E[\bar{X}_n] = \lambda$. So by Jensen's inequality:

$$E[1/\bar{X}_n] > 1/E[\bar{X}_n] = \lambda$$

   Therefore the estimator is positively biased.

2. (a) The likelihood function is the probability that we get x_1 in the first sample, x_2 in the second and so on. By the fact that the dataset realizes a geometric distribution, we get:

$$L(p) = (1-p)^{x_1-1}p\cdot(1-p)^{x_2-1}p\cdots(1-p)^{x_n-1}p = p^n\cdot(1-p)^{\sum x_i}\cdot(1-p)^{-n}$$

   (b) Taking the logarithm of the likelihood function we get the log-likelihood function:

$$\ell(p) = n\ln p + (\sum x_i)\ln(1-p) - n\ln(1-p)$$

   (c) To find the maximum likelihood estimate, we need to find the p that maximizes $\ell(p)$. We derive:

$$\frac{d}{dp}\ell(p) = \frac{n}{p} - \frac{\sum x_i}{1-p} + \frac{n}{1-p}$$

   equating to zero and solving the equation in the unknown p, we get that the maximum likelihood estimate for p is $\frac{1}{\bar{x}_n}$.

3. (a) By the fact that we set a sharp threshold between keeping H_0 and rejecting H_1, we can define a setting which is just beyond that threshold (i.e. we should reject H_0). In this case, since we just change the parameters by a small amount, the distribution is almost as the one under H_0, but now keeping H_0 (which happens with probability almost $1 - \alpha$) is an error (type II to be precise).

   (b) Under the null hypothesis the probability that a sample is positive is exactly 1/2, so the number of positive samples has a binomial distribution with parameters 10 and 1/2, and its expectation is 5. So values around 5 support H_0. Under H_1, on the other hand, we expect more than 5 to be positive (because more of the mass is in the positive side since $\delta_1 > -\delta_2$). So values towards 10 support H_1. To find the critical region with respect to significance level 0.05 we need to find a number c such that the probability that a binomial r.v. Y with parameters 10 and 1/2 is more than c is at most 0.05 (but close to it as possible). Calculating the binomial tails (i.e. $\binom{10}{k}(1/2)^{10}$) we see that $Pr[Y \geq 9] = 0.01074$ while $\Pr[Y \geq 8] = 0.05469$. So 8 is too low, therefore the critical region is $\{9, 10\}$.

4. (a) Under the studentized procedure, a 95% confidence interval is given by the formula:

$$(\bar{x}_n - t_{n-1,0.025}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,0.025}\frac{s_n}{\sqrt{n}}$$

   In other words, the mean of the dataset is exactly the middle of the interval which is 4.7.

   (b) A 99% confidence interval is given by,

$$(\bar{x}_n - t_{n-1,0.005}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,0.005}\frac{s_n}{\sqrt{n}})$$

   We have $n = 16$ and $x_n = 4.7$. From the table we get that $t_{n-1,0.005} = 2.947$. We need to compute s_n, this is given to us by the equation:

$$\bar{x}_n - t_{n-1,0.025}\frac{s_n}{\sqrt{n}} = 1.6$$

Plugging in the number and solving we get $s_n = 1.4547$. This gives us a 99% confidence interval: $(0.413, 8.987)$.