

18.443 Problem Set 10, due Wednesday, May 14, 2008

1. (a) For any real number m and any $\tau > 0$, if f is a probability density, show that $g(x) = \tau^{-1}f((x - m)/\tau)$ is also such a density. (Then m is called a location parameter and τ a scale parameter.)

(b) For any real x let

$$f_{m,\tau}(x) = \frac{\tau}{\pi(\tau^2 + (x - m)^2)}.$$

Show that this is a probability density. *Hint:* how does it relate to the standard Cauchy density?

(c) If X is such a random variable with $m = 0$, show that $E|X| = +\infty$.

Since the mean behaves badly for this distribution, let's consider the median.

(d) For any m and τ , what is the median of the distribution in part (b)?

2. Continuing with the distribution in Problem 1(b) and (d), it turns out that if X_1, \dots, X_n are i.i.d. with such a distribution then the sample mean \bar{X} doesn't converge to the median as n becomes large. (Actually \bar{X} has the same distribution for all n . We saw in PS9 problem 5 that the extreme observations $X_{(n)}$ and symmetrically $X_{(1)}$ each have medians of order n and hence they act as outliers, so that even when divided by n they have a non-negligible influence.)

Consider the sample median for large, odd values of n . Find its asymptotic distribution using the delta-method, as in the handout on sample medians. *Hint:* First suppose $m = 0$ and $\tau = 1$ (standard Cauchy distribution). The answer for general m and τ will then depend on them in an unsurprising way.

3. Suppose we have an unknown binomial probability p . In n independent trials with probability p of success we observe X successes.

(a) (Review) What is the MLE of p ? Is it unbiased? What is the 95% (two-sided) plug-in confidence interval for p ?

(b) Suppose we observe $X = n$ successes in the n trials. What does (a) give in this case (MLE and confidence interval)?

(c) What is the exact (one-sided) binomial 95% confidence interval for p if $X = n$?

(d) Now suppose we have a $U[0, 1]$ prior distribution. What is the Bayes estimate of p , for general X and for $X = n$?

(e) If $X = n$, for a $U[0, 1]$ prior, what is a 95% one-sided credible interval for p (an interval that has posterior probability 0.95)?

(f) Of the three intervals in (b), (c) and (e), which would be the last or worst choice of the three?

4. This relates to Example A on pp. 286-288 of Rice. Recall that a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ has density

$$f_{\alpha,\lambda}(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x} / \Gamma(\alpha)$$

for $x > 0$ and 0 for $x \leq 0$. The distribution has mean α/λ and variance α/λ^2 .

Assume that a distribution on the nonnegative integers is Poisson with unknown parameter $L > 0$. Suppose that before taking any observations, we want to specify a prior distribution for L with a mean of 200 and a standard deviation of 100.

- (a) Why is it not reasonable to make the prior distribution of L a normal distribution?
- (b) For the given prior mean and standard deviation, what should α and λ be?
- (c) Suppose we have n i.i.d. observations from the Poisson distribution with a sample mean equal to $\bar{X} = 273$. The posterior distribution will also be a gamma distribution, as Rice shows. This is what is meant by saying gamma distributions are “conjugate priors” for Poisson parameters. Find $\alpha = \alpha_n$ and $\lambda = \lambda_n$ as functions of n , which will also involve the observed \bar{X} .
- (d) Find the posterior mean and variance, also as functions of n .
- (e) For n large, the prior should have less and less influence on the posterior mean. So, it’s not surprising if the posterior mean becomes close to the non-Bayesian estimate \bar{X} (the maximum likelihood estimate of L). If \hat{L}_n is the Bayes estimate of L (the mean of its posterior distribution) find the expectation $E((\bar{X} - \hat{L}_n)^2)$ as a function of n , conditional on the observed value $\bar{X} = 273$.

5. A statistic X is measured in a medical test. Suppose that for a certain disease, D , the hypothesis H_0 is that the person being tested does not have D and in that case the distribution of X is $N(3, 1)$. The alternative hypothesis H_1 is that the person does have D , although they may not have any symptoms yet. Under H_1 , X has distribution $N(5, 1)$.

(a) Show that for any c , $X \geq c$ is a best critical region for testing H_0 vs. H_1 , in the sense that it has largest power for given size (recall the Neyman-Pearson lemma).

(b) Suppose it costs \$40 each time X is measured for one patient. Suppose this will be done for people in a “risk group” in which *a priori*, the probability of having D is 0.08. If the test is positive (H_0 is rejected), a test based on a different statistic costing \$1000 will be done which will give a correct result in essentially all cases. If the original \$40 test is negative (H_0 is not rejected) then no further test or treatment will be done. In that case, if a patient has D , assume a loss of \$1,000,000. How should c be chosen to minimize the expected loss?

(c) In a general population where the *a priori* probability of having D is 10^{-5} , is it cost-effective to do any such test procedure (for any c)? Hint: would it be, even if an initial \$40 test always gave the correct result?