STA 312f10 Assignment 1

Do this review assignment in preparation for the quiz on Friday, Sept. 17th. The problems are practice for the quiz, and are not to be handed in.

The first part of this assignment is based on material that you probably know already. However, the notation used in Statistics can be an obstacle for some students, so we will review the following basic rules.

- The distributive law: $a(b+c) = ab + ac$. You may see this in a form like
$$\theta \sum_{i=1}^n x_i = \sum_{i=1}^n \theta x_i$$
- Power of a product is the product of powers: $(ab)^c = a^c\, b^c$. You may see this in a form like
$$\left(\prod_{i=1}^n x_i\right)^\alpha = \prod_{i=1}^n x_i^\alpha$$
- Multiplication is addition of exponents: $a^b a^c = a^{b+c}$. You may see this in a form like
$$\prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n \exp(-\theta \sum_{i=1}^n x_i)$$
- Powering is multiplication of exponents: $(a^b)^c = a^{bc}$. You may see this in a form like
$$(e^{\mu t + \frac{1}{2}\sigma^2 t^2})^n = e^{n\mu t + \frac{1}{2}n\sigma^2 t^2}$$
- Log of a product is sum of logs: $\ln(ab) = \ln(a) + \ln(b)$. You may see this in a form like
$$\ln \prod_{i=1}^n x_i = \sum_{i=1}^n \ln x_i$$
- Log of a power is the exponent times the log: $\ln(a^b) = b\, \ln(a)$. You may see this in a form like
$$\ln(\theta^n) = n \ln \theta$$
- The log is the inverse of the exponential function: $\ln(e^a) = a$. You may see this in a form like
$$\ln\left(\theta^n \exp(-\theta \sum_{i=1}^n x_i)\right) = n \ln \theta - \theta \sum_{i=1}^n x_i$$

Choose the correct answer.

1. $\prod_{i=1}^{n} e^{x_i} =$
   - (a) $\exp(\prod_{i=1}^{n} x_i)$
   - (b) e^{nx_i}
   - (c) $\exp(\sum_{i=1}^{n} x_i)$
2. $\prod_{i=1}^{n} \lambda e^{-\lambda x_i} =$
   - (a) $\lambda e^{-\lambda^n x_i}$
   - (b) $\lambda^n e^{-\lambda n x_i}$
   - (c) $\lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i)$
   - (d) $\lambda^n \exp(-n\lambda \sum_{i=1}^{n} x_i)$
   - (e) $\lambda^n \exp(-\lambda^n \sum_{i=1}^{n} x_i)$
3. $\prod_{i=1}^{n} a_i^b =$
   - (a) na^b
   - (b) a^{nb}
   - (c) $(\prod_{i=1}^{n} a_i)^b$
4. $\prod_{i=1}^{n} a^{b_i} =$
   - (a) na^{b_i}
   - (b) a^{nb_i}
   - (c) $\sum_{i=1}^{n} a^{b_i}$
   - (d) $a^{\prod_{i=1}^{n} b_i}$
   - (e) $a^{\sum_{i=1}^{n} b_i}$
5. $\left(e^{\lambda(e^t-1)}\right)^n =$
   - (a) $ne^{\lambda(e^t-1)}$
   - (b) $e^{n\lambda(e^t-1)}$
   - (c) $e^{\lambda(e^{nt}-1)}$
   - (d) $e^{n\lambda(e^t-n)}$
6. $\left(\prod_{i=1}^{n} e^{-\lambda x_i}\right)^2 =$
   - (a) $e^{-2n\lambda x_i}$
   - (b) $e^{-2\lambda \sum_{i=1}^{n} x_i}$
   - (c) $2e^{-\lambda \sum_{i=1}^{n} x_i}$

7. True, or False?

   (a) $\sum_{i=1}^n \frac{1}{x_i} = \frac{1}{\sum_{i=1}^n x_i}$

   (b) $\prod_{i=1}^n \frac{1}{x_i} = \frac{1}{\prod_{i=1}^n x_i}$

   (c) $\frac{a}{b+c} = \frac{a}{b} + \frac{a}{c}$

   (d) $\ln(a+b) = \ln(a) + \ln(b)$

   (e) $e^{a+b} = e^a + e^b$

   (f) $e^{a+b} = e^a e^b$

   (g) $e^{ab} = e^a e^b$

   (h) $\prod_{i=1}^n (x_i + y_i) = \prod_{i=1}^n x_i + \prod_{i=1}^n y_i$

   (i) $\ln(\prod_{i=1}^n a_i^b) = b \sum_{i=1}^n \ln(a_i)$

   (j) $\sum_{i=1}^n \prod_{j=1}^n a_j = n \prod_{j=1}^n a_j$

   (k) $\sum_{i=1}^n \prod_{j=1}^n a_i = \sum_{i=1}^n a_i^n$

   (l) $\sum_{i=1}^n \prod_{j=1}^n a_{i,j} = \prod_{j=1}^n \sum_{i=1}^n a_{i,j}$

8. Simplify as much as possible.

   (a) $\ln \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$

   (b) $\ln \prod_{i=1}^n \binom{m}{x_i} \theta^{x}(1-\theta)^{m-x_i}$

   (c) $\ln \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

   (d) $\ln \prod_{i=1}^n \theta(1-\theta)^{x_i-1}$

   (e) $\ln \prod_{i=1}^n \frac{1}{\theta} e^{-x_i/\theta}$

   (f) $\ln \prod_{i=1}^n \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x_i/\beta} x_i^{\alpha-1}$

   (g) $\ln \prod_{i=1}^n \frac{1}{2^{\nu/2}\Gamma(\nu/2)} e^{-x_i/2} x_i^{\nu/2-1}$

   (h) $\ln \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

   (i) $\prod_{i=1}^n \frac{1}{\beta-\alpha} I(\alpha \leq x_i \leq \beta)$ (Express in terms of the minimum and maximum y_1 and y_n.)

9. For each of the following distributions, derive a general expression for the Maximum Likelihood Estimator (MLE). You don't have to do the second derivative test. Then use the data to calculate a numerical estimate.

   (a) $p(x) = \theta(1-\theta)^x$ for $x = 0, 1, \ldots$, where $0 < \theta < 1$. Data: 4, 0, 1, 0, 1, 3, 2, 16, 3, 0, 4, 3, 6, 16, 0, 0, 1, 1, 6, 10. Answer: 0.2061856

   (b) $f(x) = \frac{\alpha}{x^{\alpha+1}}$ for $x > 1$, where $\alpha > 0$. Data: 1.37, 2.89, 1.52, 1.77, 1.04, 2.71, 1.19, 1.13, 15.66, 1.43 Answer: 1.469102

   (c) $f(x) = \frac{\tau}{\sqrt{2\pi}} e^{-\frac{\tau^2 x^2}{2}}$, for x real, where $\tau > 0$. Data: 1.45, 0.47, -3.33, 0.82, -1.59, -0.37, -1.56, -0.20 Answer: 0.6451059

   (d) $f(x) = \frac{1}{\theta} e^{-x/\theta}$ for $x > 0$, where $\theta > 0$. Data: 0.28, 1.72, 0.08, 1.22, 1.86, 0.62, 2.44, 2.48, 2.96 Answer: 1.517778

10. Let $X_1, \ldots, X_N$ be a random sample from a Poisson distribution with parameter λ. Notice that the formula for the Poisson probability function is not being supplied. You need to remember it.

    (a) Derive a general formula for the maximum likelihood estimator (MLE) of λ. Show your work, but don't bother with the second derivative test.

    (b) Find the MLE $\widehat{\lambda}$ based on the following data: `7 7 6 4 2 5 2 3 7 2`. Answer: 4.5

11. Let $X_1, \ldots, X_{N_1}$ be a random sample from a Poisson distribution with parameter λ_1, and let $Y_1, \ldots, Y_{N_2}$ be a random sample from another Poisson distribution with parameter λ_2. The two random samples are independent.

    (a) Derive a general formula for the maximum likelihood estimator (MLE) of the pair (λ_1, λ_2). Show your work, but don't bother with anything like a second derivative test.

    (b) Find the MLE of (λ_1, λ_2) based on the following data:
    X: `7 0 3 2 7 5 4 4 4 7 7 4 5 6 5`,
    Y: `7 8 8 4 5 11 11 5 9 4`.
    Answer: (4.67,7.2).

    (c) For the data given above, find the MLE of (λ_1, λ_2) under the restriction that $\lambda_1 = \lambda_2$. Answer: (5.68,5.68).

12. This problem establishes a result that will be used later in our course. It requires you to remember that the sum of two independent Poisson random variables also has a Poisson distribution. So, let the number of girls born on one day in Toronto have a Poisson distribution with parameter λ_1, and let let the number of boys have a Poisson distribution with parameter λ_2. These two random variables are independent, and the whole thing is based upon the reasonable assumption that boys and girls are being born according to independent Poisson processes – see class notes.

    Anyway, *given* that n babies were born on a particular day, what is the probability distribution of the number of girls born on that day? Show all your work.

    Here are two hints. First, the word "given" is a clue that you are being asked for a conditional probability. The second hint is that you should start by specifying the possible values of the number of girls born, given that a total of n babies were born. This is the *support* of the distribution for which you are being asked.

13. A fast food chain is considering a change in the blend of coffee beans they use to make their coffee. To determine whether their customers prefer the new blend, the company plans to select a random sample of N coffee-drinking customers and ask them to taste coffee made with the new blend and with the old blend, in cups marked "A" and "B." Half the time the new blend will be in cup A, and half the time it will be in cup B. Let θ denote the true probability that a customer will prefer the new blend.

    (a) Derive a general formula for the maximum likelihood estimator (MLE) of θ. Show your work, but don't bother with the second derivative test.

    (b) Suppose $N = 100$ consumers participate in the taste test, and 60 of them prefer the new blend. What is the MLE $\widehat{\theta}$. The answer is a number.

    (c) Using the Central Limit Theorem, derive a general $(1-\alpha)100\%$ confidence interval for θ. "Derive" means give the details. Start by stating the Central Limit Theorem, and saying what μ and σ are in terms of this problem. Of course it's okay to use $\widehat{\sigma}$ since σ is unknown. Don't bother with any correction for continuity (if you happen to know what that is).

    (d) Using the 60 out of 100 result from above, give the upper and lower 95% confidence limits. Your answer is a pair of numbers. Answer: $(0.50398, 0.69602) \approx (0.5, 0.7)$.

14. Nothing is perfect, and that definitely applies to medical tests. Suppose a blood test is used to detect thyroid disease. The *prevalence* of the disease is the probability that a randomly chosen person actually has the disease. Even with a perfect test, this could never be known exactly without testing the whole population. The *sensitivity* of the test is a conditional probability. It is the probability that a person who actually has the disease will test positive. The *specificity* of the test is another conditional probability. It is the probability that a person who does not have the disease will test negative.

    Suppose that the sensitivity of the test is 95%, the specificity is 90%, and the underlying rate of the disease is one percent. So it's a pretty good test for a rare condition. What percent of people in the population will test positive for the disease? The answer is a single number. Show your work. My answer is 0.1085. The moral of the story is that if you confuse prevalence with the probability of testing positive (a very natural mistake), this good test can make you think the prevalence of the disease is nearly eleven times as great as it actually is.

    A good way to think about this problem is to let X be a Bernoulli random variable, with $X = 1$ indicating Disease and $X = 0$ indicating No Disease. You can never observe the value of X for any patient, but you *can* observe another Bernoulli random variable Y, with $Y = 1$ indicating Test Positive for Disease and $Y = 0$ indicating Test Negative. $P(X = 1) =$ Prevalence; X and Y are linked by the sensitivity and specificity. $P(Y = 1|X = 1) =$ Sensitivity, and so on. With this setup, you should be able to solve the problem.

    Here is a final comment. You can think of Y as an imperfect *measurement* of X, with errors of measurement coming from two sources: sensitivity not equal to one and specificity not equal to one. The kind of measurement error in this problem is called "classification error," because a patient could be misclassified as having the disease when he or she does not (false positive), or as not having the disease when he or she actually does (false negative). Classification error is very common in practice — everywhere, not just in medicine. It can produce very misleading results when it is ignored.