MATH 476 Statistics

Fred J. Hickernell **Spring 2010**
Test **Monday, March 1**

Instructions:

i. This test consists of FOUR *questions. Answer all of them.*

ii. The time allowed for this test is 75 minutes.

iii. This test is closed book, but you may use 2 double-sided letter-size sheets of notes.

iv. Calculators, even of the programmable variety, are allowed. Computers using JMP or MATLAB, are also allowed. No internet access, web browsing, email, chat, etc. is allowed.

v. Show all your work to justify your answers. Answers without adequate justification will not receive credit.

1. (25 marks)
Jill is an IIT student studying the failure rates of computer hard drives. She observes, X, the number of hard drive failures during t months in a data center containing n hard drives. Note that X follows a Poisson process with parameter $nt\lambda$, i.e., $E(X) = \text{var}(X) = nt\lambda$. Here, λ denotes the average number of failures for a single hard drive in a single month.

   a) Assuming n and t to be known, find an unbiased estimate of λ based on the observation X.

   Answer: Let $Y = X/(nt)$. Then $E(Y) = E[X/(nt)] = nt\lambda/(nt) = \lambda$, so Y is an unbiased *estimate of λ.*

   b) Is your estimate of λ in part a) a *consistent* estimate as $nt \to \infty$? Why or why not?

   Answer: Note that

$$\text{var}(Y) = \text{var}\left(\frac{X}{nt}\right) = \frac{\text{var}(X)}{(nt)^2} = \frac{nt\lambda}{(nt)^2} = \frac{\lambda}{nt}$$

   Since $E(Y) = \lambda$ and $\lim_{nt\to\infty} \text{var}(Y) = 0$, Y is a consistent estimator of λ.

2. (30 marks)
Jill wants to construct a *confidence interval* for λ based on her observation of X. One may write

$$X = \sum_{i=1}^{n} \sum_{j=1}^{t} Z_{ij}.$$

where Z_{ij} denotes the number of failures for a single hard drive i for a single month j, and the Z_{ij} are i.i.d. Poisson random variables.

   a) Assuming that nt is large, use the Central Limit Theorem find the approximate distribution of X.

Answer: Note that Y derived in the previous question is just the sample average. By the Central Limit Theorem:

$$\frac{Y-\lambda}{\sqrt{\lambda/(nt)}} \approx N(0,1), \qquad Y \approx N(\lambda, \lambda/(nt)) \qquad X = ntY \approx N(nt\lambda, nt\lambda),$$

b) Use this approximation to construct a 95% *confidence interval* for λ based on X.

Answer: From the above one has

$$\begin{aligned}
95\% &\approx \text{Prob}\left[\left|\frac{Y-\lambda}{\sqrt{\lambda/(nt)}}\right| \le 1.96\right] && \textit{by the Central Limit Theorem} \\
&= \text{Prob}\left[Y - 1.96\sqrt{\lambda/(nt)} \le \lambda \le Y + 1.96\sqrt{\lambda/(nt)}\right] && \textit{expanding the inequality} \\
&\approx \text{Prob}\left[Y - 1.96\sqrt{Y/(nt)} \le \lambda \le Y + 1.96\sqrt{Y/(nt)}\right] && \textit{since } \lambda \approx Y \\
&\approx \text{Prob}\left[\frac{X - 1.96\sqrt{X}}{nt} \le \lambda \le \frac{X + 1.96\sqrt{X}}{nt}\right] && \textit{since } X = ntY.
\end{aligned}$$

Thus, the approximate 95% confidence interval is $(X \pm 1.96\sqrt{X})/(nt)$. *For*

c) Jill observes 203 failures in a lab with 1000 hard drives over the course of 12 months. What is the approximate 95% confidence interval for λ, the average number of failures per month for a single hard drive?

Answer:

$$\frac{X \pm 1.96\sqrt{X}}{nt} = \frac{203 \pm 1.96\sqrt{203}}{12000} = 0.0169 \pm 0.0023.$$

3. (25 marks)
Jill takes a professional exam and wonders whether IIT or UIUC engineers do better on average. Compute a 95% confidence interval for the difference of the means of the two populations of engineers based on the exam scores below. Is there compelling evidence that one school's engineers score better than the other school's engineers?

IIT	76	82	45	93	98	72	65	76	95	88	85	79	83	72	89	82	65	74
UIUC	75	61	89	73	92	68	54	73	95	84	71	68	57	69	74	81	60	70

Answer: In JMP put all test scores in a column labeled `Scores`*. In the second column, labeled* `School` *identify each score by its university. Then choose* `Analyze/Fit Y By X` *and put* `Scores` *into* `Y` *and* `University` *into* `X` *and submit. Finally choose* `Means/ANOVA/Pooled t` *from the red triangle menu to get a 95% confidence interval for the difference of the means of* $[2.425, -14.091]$. *Since this confidence interval contains zero, there is no compelling evidence that one school's engineers are better than the other school's engineers.*

4. (20 marks)
It is time to apply for summer jobs. Jill and her friends keep track of the number of resumes that each of them must send out before receiving the first job offer, $X_1, X_2, \ldots$. The X_i may be modeled by the geometric distribution, which has probability mass function

$$f(x) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, 3, \ldots.$$

Here θ denotes the probability of a resume leading to a job interview. Find the *maximum likelihood estimator* for θ in terms of the data $x_1, \ldots, x_n$.

Answer: The likelihood function is

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \left[\theta(1-\theta)^{x_i-1}\right] = \theta^n(1-\theta)^{x_1+\cdots+x_n-n} = \theta^n(1-\theta)^{n(\bar{x}-1)}.$$

Differentiating the log likelihood function and setting it to zero gives

$$\begin{aligned}\log\left(L(x_1, \ldots, x_n|\theta)\right) &= n\log(\theta) + n(\bar{x}-1)\log(1-\theta),\\ 0 = \frac{\mathrm{d}\log\left(L(x_1, \ldots, x_n|\theta)\right)}{\mathrm{d}\theta} &= \frac{n}{\theta} + \frac{-n(\bar{x}-1)}{1-\theta},\\ (1-\theta) = (\bar{x}-1)\theta, \qquad \theta\bar{x} &= 1, \qquad \theta_{\mathrm{MLE}} = \frac{1}{\bar{x}}.\end{aligned}$$