STAT 703/J703 – Spring 2007 - Take Home Exam 3

Due by7:30pm Thursday, May 3rd (I will check my e-mail, mail box and the fax machine at 9:00am on Friday for them)

Answer 10 of the following 11 questions (I will grade your best 10, there is one more to choose from this time.) Show all of your work for credit. There are no "trick" questions.

---

1) One famous (and old) data set concerns the number of deaths per year per group of soldiers due to being kicked by pack animals. In total, 10 groups of soldiers were monitored for 20 years giving a total of 200 effective years. This data was summarized in the following table:

No. of Deaths	0	1	2	3	4	5 or more
No. Group-Years	109	65	22	3	1	0

Consider being asked to test whether this data follows a Poisson distribution by first finding the usual MLE of lambda assuming the data follows a Poisson distribution and then using that with the likelihood ratio test for the multinomial distribution (don't actually do the test!) What is the "obvious" number of degrees of freedom to use for the resulting test based on the above set-up? Explain how you could easily (and reasonably) want to use a different degrees of freedom from this same observed data set?

2) Consider the choice between using the two-sample t-test (based on the sample mean and standard deviation) and the Mann-Whitney-Wilcoxon rank sum test (based on ranking the observations). In a few sentences, explain what effect a single large outlier in one of the samples would have on the t-test and what effect it would have on the Mann-Whitney-Wilcoxon rank sum test, and why. Be sure to indicate for each test whether it would be more, less, or equally likely to reject the null hypothesis as compared to that same test when the outlier was a more reasonable value.

3) Two procedures for testing whether variables are correlated (H_0: variables are uncorrelated) are based on Pearson's correlation coefficient *r* and Kendall's τ. The asymptotic relative efficiency of τ relative to *r* for three distributions is given in the following table.

Distribution	Uniform	Normal	Double Exponential
eff(τ,*r*)	1.000	0.912	1.266

Briefly summarize what the value of 0.912 is telling us when the variables we are observing are approximately normally distributed? (Which procedure is better? By how much in terms of sample size? And what is a major limitation of this interpretation?)

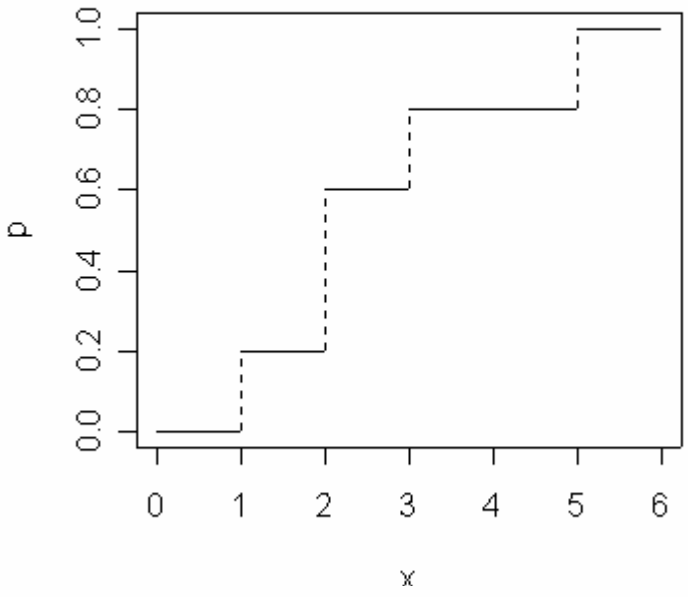
4) The Rayleigh distribution has pdf: $f(x|\theta) = \dfrac{x\exp\left(-\dfrac{x^2}{2\theta^2}\right)}{\theta^2}$ for $x > 0$. It's method of moments estimator is $\hat{\theta}_{mom} = \bar{x}\sqrt{\dfrac{2}{\pi}}$, the mle was $\hat{\theta}_{mle} = \sqrt{\dfrac{\sum_{i=1}^{n} x_i^2}{2n}}$, and it's information function is $I(\theta) = \dfrac{4}{\theta^2}$. What do we know about the variance of any unbiased estimator of θ from a sample of size *n*?

5) The pdf of the beta distribution is $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $0<x<1$, and the beta with parameters α and β has mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$.

Show that the beta distribution is a member of the exponential family of distributions, and find the natural parameters and corresponding sufficient statistics.

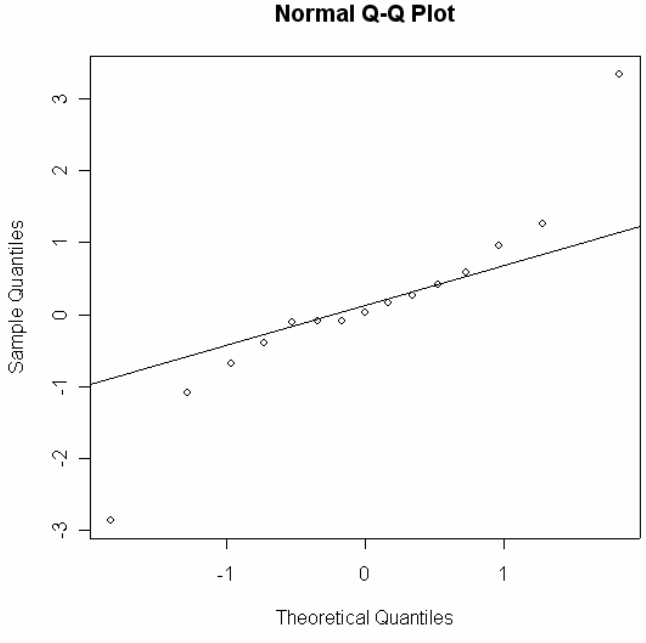
6) The Bernoulli (binomial with n=1) is a member of the exponential family with parameter $p/(1-p)$ (or equivalently p) and statistic $\sum_{i=1}^{m} x_i$ for a sample of size m. Prove that about $\bar{x}$ is the UMVUE for p. (You may assume needed regularity conditions are met—find the right theorem and this will only be a few lines long.)

7) The plot shown at the right is the EDF for a data set. Determine the relative frequency distribution and the median. Give one reason to suspect that this EDF is for data from a discrete random variable and not a continuous one.



8) A random sample of size 15 is supposed to have been generated from a standard normal distribution. It results in the q-q plot shown at the right and a Kolmogorov-Smirnov p-value of 0.6298.

Which of the two "supports" normality and which does not? Which do you "trust more" and why?



9) Consider the observed data set 5.1 5.8 7.2 9.1 10.5 12.4 14.7 .
Use the nonparametric bootstrap (B=100) to estimate the standard deviation of the median of this data set. Include all of the computer code used.

Questions 10-11 are based on the Bayesian estimation of the *p* parameter for a single binomial random variable X. The observation X is binomial with parameters *n* and *p*, and *p* has a prior that is beta with parameters *a* and *b*.

10) Show that the posterior distribution of *p* given *x* follows a beta distribution with parameters $a+x$ and $b+n-x$.

11) Consider the baseball player who has 1 hit in 25 at bats and using a prior that is beta with $a=76.57$ and $b=211.26$. First, verify that this prior does indeed have mean 0.266 and sd 0.026 like the example in class. Second, find the posterior mean estimate of this batters *p*. (The important formulae are given in problem 5.)