STAT 513 fa 2018 Final Exam

Karl B. Gregory

Do not open this test until told to do so; no calculators allowed; no notes allowed; no books allowed; show your work so that partial credit may be given.

pmf/pdf	$\mathcal{X}$	$M_X(t)$	$\mathbb{E}X$	$\operatorname{Var} X$
$p_X(x;p) = p^x(1-p)^{1-x}$,	$x = 0, 1$	$pe^t + (1-p)$	p	$p(1-p)$
$p_X(x;n,p) = \binom{n}{x}p^x(1-p)^{n-x}$,	$x = 0, 1, \ldots, n$	$[pe^t + (1-p)]^n$	np	$np(1-p)$
$p_X(x;p) = (1-p)^{x-1}p$,	$x = 1, 2, \ldots$	$\frac{pe^t}{1-(1-p)e^t}$	p^{-1}	$(1-p)p^{-2}$
$p_X(x;p,r) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$,	$x = r, r+1, \ldots$	$\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$	rp^{-1}	$r(1-p)p^{-2}$
$p_X(x;\lambda) = e^{-\lambda}\lambda^x/x!$	$x = 0, 1, \ldots$	$e^{\lambda(e^t-1)}$	λ	λ
$p_X(x;N,M,K) = \binom{M}{x}\binom{N-M}{K-x}/\binom{N}{K}$	$x = 0, 1, \ldots, K$	¡complicadísimo!	$\frac{KM}{N}$	$\frac{KM}{N}\frac{(N-K)(N-M)}{N(N-1)}$
$p_X(x;K) = \frac{1}{K}$	$x = 1, \ldots, K$	$\frac{1}{K}\sum_{x=1}^K e^{tx}$	$\frac{K+1}{2}$	$\frac{(K+1)(K-1)}{12}$
$p_X(x;x_1,\ldots,x_n) = \frac{1}{n}$	$x = x_1, \ldots, x_n$	$\frac{1}{n}\sum_{i=1}^n e^{tx_i}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$	$\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$
$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$-\infty < x < \infty$	$e^{\mu t + \sigma^2 t^2/2}$	μ	σ^2
$f_X(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)$	$0 < x < \infty$	$(1-\beta t)^{-\alpha}$	$\alpha\beta$	$\alpha\beta^2$
$f_X(x;\lambda) = \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)$	$0 < x < \infty$	$(1-\lambda t)^{-1}$	λ	λ^2
$f_X(x;\nu) = \frac{1}{\Gamma(\nu/2)2^{\nu/2}}x^{\nu/2-1}\exp\left(-\frac{x}{2}\right)$	$0 < x < \infty$	$(1-2t)^{-\nu/2}$	ν	2ν
$f_X(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$0 < x < 1$	$1 + \sum_{k=1}^\infty \frac{t^k}{k!}\left(\prod_{r=0}^k \frac{\alpha+r}{\alpha+\beta+r}\right)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Let $X_1, \ldots, X_n$ be a random sample with likelihood function $L(\theta; X_1, \ldots, X_n)$. Then for hypotheses of the form H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$ the likelihood ratio takes the form

$$\mathrm{LR}(X_1, \ldots, X_n) = \frac{L(\theta_0; X_1, \ldots, X_n)}{L(\hat{\theta}; X_1, \ldots, X_n)},$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ.

1. Given a set of outcomes from n independent Bernoulli trials with unknown success probability p, Agresti and Coull in [1] suggested the estimator

$$\tilde{p} = \frac{Y+2}{n+4}$$

for p, where Y is the number of success in the n trials. This can be interpreted as the proportion of successes after adding two successes and two failures to the observed set of outcomes. This question is designed to make you understand why this is a reasonable thing to do.

Suppose we have the Bayesian hierarchical model

$$\begin{aligned} Y|p &\sim \text{Binomial}(n,p) \\ p &\sim \text{Beta}(\alpha,\beta). \end{aligned}$$

(a) Find the posterior distribution of $p|Y$.

Solution: We have

$$p|Y \sim \text{Beta}(Y+\alpha, n-Y+\beta).$$

(b) Give a formula for the posterior mean of $p|Y$.

Solution: We have

$$\mathbb{E}[p|Y] = \frac{Y+\alpha}{n+\alpha+\beta}.$$

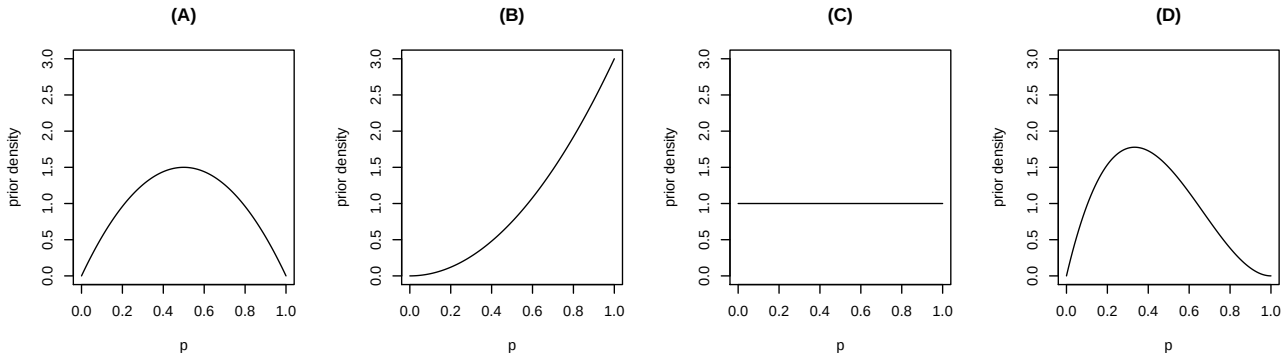
(c) Describe how you would construct a 95% credible interval for p after observing a value of Y.

Solution: One way is to set the bounds of the interval equal to the lower and upper $\alpha/2$ quantiles of the posterior distribution, in this case those of the Beta$(Y+\alpha, n-Y+\alpha)$ distribution.

(d) Find choices of α and β such that the posterior mean of $p|Y$ is equal to the estimator $\tilde{p}$.

Solution: If we choose $\alpha = 2$ and $\beta = 2$, the posterior mean is given by $(Y+2)/(n+4)$.

(e) The pdfs of four Beta distributions are plotted below. Select the distribution which, if used as a prior for p, results in a posterior mean of $p|Y$ which is given by adding *one* success and *one* failure to the observed set of outcomes.



Solution: If we choose $\alpha = 1$ and $\beta = 1$ we get a posterior mean equal to $(Y + 1)/(n + 2)$. For $\alpha = 1$ and $\beta = 1$ the Beta(α, β) distribution is the same as the Uniform$(0, 1)$ distribution, which has the flat density appearing in plot (C).

(f) Suppose two researchers have different prior beliefs about p such that they make different choices of the prior parameters α and β. Then, suppose the two researchers are given the same data set and each of them computes the posterior mean of p. Consider the following scenarios:

1. The data set contains 10 observations.
2. The data set contains 100 observations.

In which scenario will the researchers' posterior means be closer to one another?

Solution: When $n = 100$ the posterior means computed by the two researchers will be closer to one another than when $n = 10$, because as the sample size grows, the posterior mean approaches the data mean, and the researchers have been given the same data.

2. Suppose $Y_1, \ldots, Y_n$ are independent random variables having pdf

$$f(y;\beta) = \frac{y}{\beta^2} \exp\left(-\frac{y}{\beta}\right) \mathbf{1}(y > 0).$$

(a) Write down the likelihood function $L(\beta; Y_1, \ldots, Y_n)$ of the parameter β based on the data $Y_1, \ldots, Y_n$.

Solution: We have

$$\begin{aligned} L(\beta; Y_1, \ldots, Y_n) &= \prod_{i=1}^{n} \frac{Y_i}{\beta^2} \exp\left(-\frac{Y_i}{\beta}\right) \\ &= \frac{1}{\beta^{2n}} \prod_{i=1}^{n} Y_i \exp\left(-\frac{\sum_{i=1}^n Y_i}{\beta}\right). \end{aligned}$$

(b) Give the log-likelihood function $\ell(\beta; Y_1, \ldots, Y_n)$.

Solution: We have

$$\ell(\beta; Y_1, \ldots, Y_n) = -2n \log \beta + \sum_{i=1}^{n} \log Y_i - \frac{\sum_{i=1}^n Y_i}{\beta}.$$

(c) Find the maximum likelihood estimator of β.

Solution: We write

$$\frac{\partial}{\partial \beta} \ell(\beta; Y_1, \ldots, Y_n) = -\frac{2n}{\beta} + \frac{\sum_{i=1}^n Y_i}{\beta^2} = 0,$$

to which the solution is

$$\hat{\beta} = \frac{1}{2n} \sum_{i=1}^{n} Y_i = \frac{1}{2} \bar{Y}_n.$$

(d) Suppose it is of interest to test

$$H_0\colon\ \beta = \beta_0 \text{ versus } H_1\colon\ \beta \neq \beta_0$$

for some β_0. Write down the likelihood ratio corresponding to these hypotheses and simplify it so that it is a function of $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$, β_0, and n.

Solution: The likelihood ratio is given by

$$\begin{aligned}
\mathrm{LR}(Y_1,\ldots,Y_n) &= \frac{L(\beta_0;Y_1,\ldots,Y_n)}{L(\hat{\beta};Y_1,\ldots,Y_n)} \\
&= \frac{\beta_0^{-2n}\prod_{i=1}^n Y_i \exp\left[-\beta_0^{-1}\sum_{i=1}^n Y_i\right]}{\hat{\beta}^{-2n}\prod_{i=1}^n Y_i \exp\left[-\hat{\beta}^{-1}\sum_{i=1}^n Y_i\right]} \\
&= \left[\frac{\hat{\beta}}{\beta_0}\right]^{2n} \exp\left[-\sum_{i=1}^n Y_i\left(\frac{1}{\beta_0}-\frac{1}{\hat{\beta}}\right)\right] \\
&= \left[\frac{\bar{Y}_n/2}{\beta_0}\right]^{2n} \exp\left[-n\bar{Y}_n\left(\frac{1}{\beta_0}-\frac{1}{\bar{Y}_n/2}\right)\right] \\
&= \left[\frac{\bar{Y}_n/2}{\beta_0}\right]^{2n} \exp\left[-\frac{n\bar{Y}_n}{\beta_0}+2n\right] \\
&= \left[\frac{\bar{Y}_n}{2\beta_0}\right]^{2n} \exp\left[-\frac{\bar{Y}_n}{2\beta_0}\right]^{2n} \exp(2n).
\end{aligned}$$

(e) For any $\alpha \in (0,1)$, give a test which will have size approximately equal to α when n is large. Your answer should be of the form, "Reject H_0 if and only if ..."

Solution: We have

$$-2\log \mathrm{LR}(Y_1,\ldots,Y_n) = -4n\left[\log\left(\frac{\bar{Y}_n}{2\beta_0}\right) - \frac{\bar{Y}_n}{2\beta_0} + 1\right],$$

so that the asymptotic likelihood ratio test is given by

$$\text{Reject } H_0 \text{ iff } -4n\left[\log\left(\frac{\bar{Y}_n}{2\beta_0}\right) - \frac{\bar{Y}_n}{2\beta_0} + 1\right] > \chi^2_{1,\alpha}.$$

3. Consider the 2×2 table below, which contains counts from a fictional study in which it was recorded whether subjects in a control and a treatment group experienced an adverse event (like a headache or nausea, etc.).

	Adverse event	No adverse event
Control	18	382
Treatment	2	98

(a) Compute the table of expected values under the hypothesis that there is no association between the control and treatment groups and experiencing an adverse event (no calculators allowed; you shouldn't need one).

Solution: The table of expected values is

	Adverse event	No adverse event
Control	16	384
Treatment	4	96

(b) Which of the tests that we learned in class do you recommend for testing the null hypothesis of no association? Explain your answer.

Solution: Since one of the expected cell counts is less than 5, we should not use Pearson's chi-squared test or the asymptotic likelihood ratio test. This leaves Fisher's exact test.

(c) Let p_1 and p_2 be the probabilities of experiencing an adverse event in the control and treatment groups, respectively, and let Y_1 and Y_2 be the observed numbers of subjects experiencing adverse events in the control and treatment groups, respectively, and suppose we adopt the Bayesian setup

$$Y_1|p_1 \sim \text{Binomial}(400, p_1), \quad p_1 \sim \text{Beta}(2,2)$$
$$Y_2|p_2 \sim \text{Binomial}(100, p_2), \quad p_2 \sim \text{Beta}(2,2),$$

where p_1 and p_2 are independent and $Y_1|p_1$ and $Y_2|p_2$ are independent.

Under this Bayesian setup, give an expression for the posterior mean of $p_2 - p_1|Y_1, Y_2$, and give its value based on observing $Y_1 = 18$ and $Y_2 = 2$ under the sample sizes $n_1 = 400$ and $n_2 = 100$.

Solution: We have

$$\mathbb{E}[p_2 - p_1|Y_1, Y_2] = \mathbb{E}[p_2|Y_2] - \mathbb{E}[p_1|Y_1] = \frac{Y_2 + 2}{n_2 + 4} - \frac{Y_1 + 2}{n_1 + 4}$$

Plugging in $Y_1 = 18$, $Y_2 = 2$, $n_1 = 400$, and $n_2 = 100$ gives

$$\frac{4}{104} - \frac{20}{404}.$$

4. The following data is a subset of a larger data set from [2] containing the waiting times of patients waiting to receive a liver transplant between 1990 and 1999. The data below are waiting times for females with blood type 'O' who are 40 years old or younger.

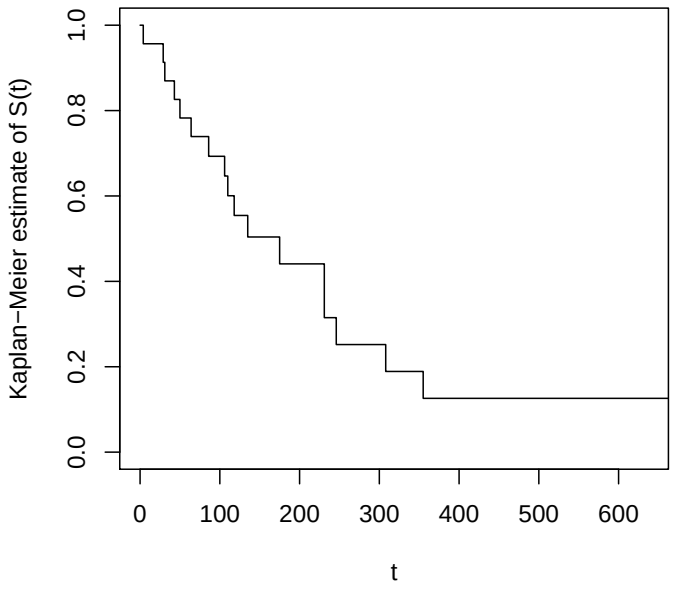
```
4    29    31    43    50    64    83+   86   106   110   118   129+
135  144+ 146+ 175   231   231 246   308   355   406+ 637+
```

A '+' marker indicates a censoring time (death, withdrawal from the study, etc.) prior to the patient's receiving a transplant.

(a) A researcher who wishes to estimate the mean waiting time for patients in this group to receive a liver transplant proposes taking the mean of the above 23 numbers, ignoring the censoring information. Give a critique of this approach, explaining your reasoning.

Solution: If the researcher simply averages these numbers, the result will tend to underestimate the true mean. This is because those observations marked with a plus sign are less than or equal to the true event times.

(b) The Kaplan-Meier estimate of the survival function based on these data is plotted below.



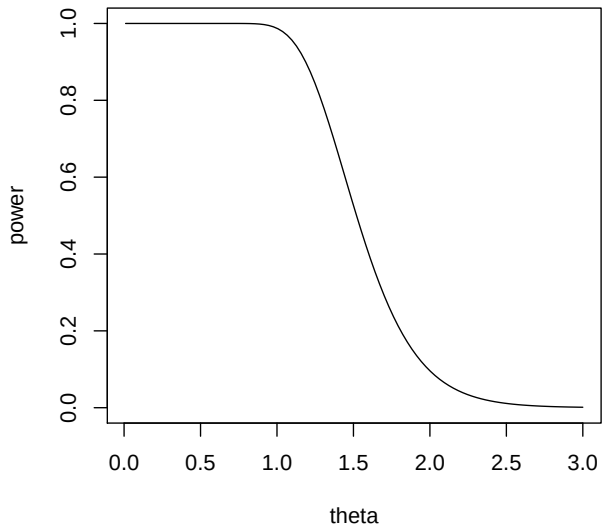
According to this estimate of the survival function, what is the probability that a patient from this group will receive a liver transplant in *fewer than* 300 months?

Solution: The probability that a patient in this group waits for more than 300 months is estimated to be about 0.25 (This is the height of the curve at 300), so the probability of waiting for fewer than 300 months is estimated to be 0.75.

5. Below is a plot of the power curve of a test of the hypotheses

$$H_0\colon\ \theta \geq 2 \text{ versus } H_1\colon\ \theta < 2$$

for some parameter θ.



Use the plot to find the following as accurately as you can:

(a) The size of the test.

(b) The probability of making an incorrect decision when $\theta = 1.5$.

(c) The probability of making a correct decision when $\theta = 2.5$.

6. Let $Y_1, \ldots, Y_n$ be a random sample from the Normal(μ, σ^2) distribution, where σ^2 is known.

(a) Given a real number μ_0 and a value of $\alpha \in (0, 1)$, give a size-α test of the hypotheses

$$H_0\colon\ \mu \leq \mu_0 \text{ versus } H_1\colon\ \mu > \mu_0.$$

Solution: A size-α test is

$$\text{Reject } H_0 \text{ iff } \frac{\bar{Y}_n - \mu_0}{\sigma/\sqrt{n}} > z_\alpha.$$

(b) Give an expression for the power function $\gamma(\mu)$ of your test.

Solution: The power is given by

$$\begin{aligned}
\gamma(\mu) &= P_\mu\left(\frac{\bar{Y}_n - \mu_0}{\sigma/\sqrt{n}} > z_\alpha\right) \\
&= P_\mu\left(\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} > z_\alpha\right) \\
&= P_\mu\left(Z > z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right), \quad Z \sim \text{Normal}(0,1) \\
&= 1 - \Phi\left(z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

(c) For any value $\mu_1 > \mu_0$, interpret the value of $\gamma(\mu_1)$.

Solution: The value of $\gamma(\mu_1)$ is the probability of rejecting H_0 when the true mean is equal to μ_1. Since $\mu_1 > \mu_0$, this represents the probability of correctly rejecting H_0.

(d) For any value $\mu_1 > \mu_0$, describe in detail how the value of $\gamma(\mu_1)$ is affected by

i. the sample size n.

Solution: If the sample size n is increased, $\gamma(\mu_1)$ will increase.

ii. the value of σ^2.

Solution: Larger values of σ^2 make $\gamma(\mu_1)$ smaller.

iii. the choice of α.

Solution: A smaller choice of α will make $\gamma(\mu_1)$ smaller.

References

[1] Alan Agresti and Brent A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

[2] W. Ray Kim, Terry M. Therneau, Joanne T. Benson, Walter K. Kremers, Charles B. Rosen, Gregory J. Gores, and E. Rolland Dickson. Deaths on the liver transplant waiting list: an analysis of competing risks. *Hepatology*, 43(2):345–351, 2006.