Stat 342 Homework Fall 2014

Assignment 1 Due 9/5/14

Section 1.1 of the Course Outline

1. Consider a probability model for the random pair (y,x) with joint density

$$f(y,x)=\begin{cases}\frac{1}{x}\exp\left(-\frac{y}{x}\right) & \text{for } 0<x<1 \text{ and } y>0\\ 0 & \text{otherwise}\end{cases}$$

a) Find the marginal distribution of x and the conditional density of $y\,|\,x$.

b) Find both the optimal squared error loss and optimal absolute error loss predictors of y based on x.

c) Evaluate the risk of the optimal squared error loss predictor.

2. Suppose that the random pair (y,x) has a joint distribution on $\{0,1\}\times(0,\infty)$ that may be specified as follows:

$P[y=1]=.7$ and $P[y=0]=.3$
conditional on the value of y, the variable x is Exponential with mean $y+1$

A "density" for (y,x) (that one adds over x and integrates over y) is then

$$f(y,x)=\begin{cases}.7I[y=0]\exp(-x)+.3I[y=1]\left(\frac{1}{2}\exp\left(-\frac{x}{2}\right)\right) & \text{if } (y,x)\in\{0,1\}\times(0,\infty)\\ 0 & \text{otherwise}\end{cases}$$

a) Evaluate $P[x>1]$.

b) Find the conditional probability that $y=1$ given the value of x, $P[y=1\,|\,x]$.

c) Find the 0-1 loss optimal predictor of y based on x.

d) Evaluate the risk of your predictor in c).

3. Below is a table giving a joint pmf for the random pair (y, x).

	$x=1$	$x=2$	$x=3$
$y=3$	.05	.2	.1
$y=2$	.1	.1	.11
$y=1$	.15	.1	.09

a) Find the SEL optimal predictor of y based on x. (You need to evaluate all of $\hat{y}^{\text{opt}}(1), \hat{y}^{\text{opt}}(2),$ and $\hat{y}^{\text{opt}}(3)$.) Evaluate the risk of this predictor.

b) Find the 0-1 loss predictor of y based on x. (Again, you need to evaluate all of $\hat{y}^{\text{opt}}(1), \hat{y}^{\text{opt}}(2),$ and $\hat{y}^{\text{opt}}(3)$.) Evaluate the risk of this predictor.

Section 1.2 of the Course Outline

4. Consider the toy statistical model where the entries of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are iid $\mathrm{N}(\mu, 1)$.

a) Write out the likelihood function for inference about μ.

b) Find the maximum likelihood estimator of μ (the function of $\mathbf{x}$ that maximizes the likelihood function, $\hat{\mu}^{\text{MLE}}(\mathbf{x})$). (Operationally, you will probably find it easiest to maximize the logarithm of the likelihood function.)

c) Find the SEL risk function $R(\mu)$ for the maximum likelihood estimator of μ.

d) A second possible estimator of μ is $.5\hat{\mu}^{\text{MLE}}(\mathbf{x})$. Find the risk function of this second estimator of μ and compare it to your answer to c). Is one of the two estimators always (for all μ) better than the other? If not, for which values of (the unknown) μ is this second estimator preferable?

5. Suppose that $x \sim \mathrm{Bi}(5, p)$.

a) Plot on the same set of axes, the 6 different possible log-likelihood functions

$$\ln\left(p^x(1-p)^{5-x}\right)$$

(the $\binom{5}{x}$ is inconsequential).

b) In light of the plot in a) for the $x=0$ and $x=5$ cases and some calculus, what is the maximum likelihood estimator of p, $\hat{p}^{\mathrm{MLE}}(x)$? What is the risk function of this estimator?

c) Two more possible estimators of p are

$$\hat{p}_2(x)=\frac{\frac{1}{2}+\frac{x}{5}}{2} \text{ and } \hat{p}_3(x)=\frac{1}{2}$$

Find the SEL risk functions for these estimators and plot them together with the risk function from b) on the same set of axes. Is any one of these uniformly/always smaller than the others?

6. Consider the two binomial pmf's

$$f(x\,|\,0)=\binom{5}{x}(.25)^x(.75)^{5-x} \text{ and } f(x\,|\,1)=\binom{5}{x}(.75)^x(.25)^{5-x}$$

defined on $\{0,1,\ldots,5\}$ and decision about $\theta\in\{0,1\}$. Find the 0-1 loss risk function for the decision function

$$a(x)=I[x>2.5]$$

(You must find the two values $R(0)$ and $R(1)$.)

Section 1.3 of the Course Outline

7. For $x\sim\mathrm{N}(\mu,1)$ and a $\mathrm{N}\left(0,(10)^2\right)$ prior distribution for μ, what is the SEL Bayes optimal estimator of μ. Hint: What is the form of the conditional distribution of μ given x?

8. For $x\sim\mathrm{Bi}(5,p)$ and a $\mathrm{Uniform}(0,1)$ prior distribution for p, what is the SEL Bayes optimal estimator of p? Hint: What is the form of the conditional distribution of p given x?

Assignment 2 Due 9/19/14

Section 2.1 of the Course Outline

9. In class we derived the distribution of $\bar{U}=\frac{1}{2}(U_1+U_2)$ for U_1 and U_2 iid with pmf

u	0	1	2
$f(u)$	.3	.4	.3

a) Find the distribution of $\bar{U}=\frac{1}{4}(U_1+U_2+U_3+U_4)=\frac{1}{2}\left(\frac{1}{2}(U_1+U_2)+\frac{1}{2}(U_3+U_4)\right)$ for iid random variables U_1,U_2,U_3,U_4 with this marginal distribution.

b) Compare the mean and variance for U_1 to the means and variances for both the $n=2$ and $n=4$ versions of $\bar{U}$.

10. Suppose that U_1 and U_2 are iid discrete random variables, each uniformly distributed on the set $\{0,1,2,3,4,5,6\}$. Find the pmf of the random variable $V=U_1U_2$. (Determine the set of possible values and corresponding probabilities and put them into a table specifying this pmf.)

11. Suppose that (U,V) is uniformly distributed on the unit square. That is, suppose that the pair has joint pdf

$$f(u,v)=I[0<u<1 \text{ and } 0<v<1]$$

Consider the random variable $T=U+V$.

a) Find the cdf of T. (You can find different "formulas" for $F(t)$ depending upon how t compares to the values $0,1,$ and 2. Once you identify the set of (u,v) with total less than or equal to t, simple geometry can be used to find the value of $F(t)$.)

b) Use a) and find the pdf of T.

12. Suppose that U and V are iid $\text{Exp}(1)$, i.e. have joint pdf

$$f(u,v) = I[0 < u < 1 \text{ and } 0 < v < 1]\exp(-(u+v))$$

Find the cdf and pdf for the product of U and V. (The first will require computation of a double integral over an appropriate (u,v) region.)

13. Find a function h so that for $U \sim \mathrm{U}(0,1)$ the random variable $h(U)$ has pdf

$$f(v) = \frac{3}{8}v^2 I[0 < v < 2]$$

14. Consider the discrete distribution of Problem 9. Find a function h so that for $V \sim \mathrm{U}(0,1)$ the random variable $h(V)$ has that simple distribution. (There are many solutions here. h should have only 3 possible values, and it suffices to break its domain into 3 intervals corresponding to those possible values.)

15. Find the moment generating function of the $\mathrm{U}(0,\theta)$ distribution.

16. Argue using moment generating functions that the sum of $n = 5$ iid $\mathrm{Exp}(1)$ random variables has a gamma distribution.

17. Suppose that the entries of $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are iid $\mathrm{N}(\mu_x, \sigma^2)$, the entries of $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ are iid $\mathrm{N}(\mu_y, \sigma^2)$, and $\mathbf{x}$ and $\mathbf{y}$ are independent. Let $\overline{x}$ and s_x^2 are the sample mean and variance of the x_i's and $\overline{y}$ and s_y^2 are the sample mean and variance of the y_i's. Define

$$s_{\text{pooled}}^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{(n-1)+(m-1)}$$

a) Argue carefully that there is a multiple of s_{pooled}^2 that has a χ^2 distribution. (What is the multiplier and what are the degrees of freedom?)

b) Identify a function of the random variables $\overline{x}-\overline{y}$ and s^2_{pooled} and the difference $\mu_x - \mu_y$ that has a t distribution. What are appropriate degrees of freedom?

c) Suppose that $\overline{x}=7, s_x=1, n=10, \overline{y}=5, s_y=.8,$ and $m=12$. Use your answer to b) and find 90% two-sided confidence limits for $\mu_x - \mu_y$.

Assignment 3 Due 9/26/14

18. In Example 18 the claim is made that the sum of two independent Poisson random variables is Poisson. Prove this.

Section 2.2 of the Course Outline

19. Use the fact that for large λ a Poisson distribution with mean λ is approximately normal and find limits based on a Poisson variable X that (at least for large λ) function as approximate 90% confidence limits for λ. What interval is produced if a Poisson observation is $X=500$?

20. Suppose that $x_1, x_2, x_3, \ldots$ are iid $\mathrm{U}(0,\theta)$. Consider the random variable

$$m_n = \max\{x_1, x_2, \ldots, x_n\}$$

the largest of the first n of these observations. Notice that $m_n \le t$ if and only if all of the first n observations are less than or equal to t. (So, for $t<\theta$ the $P[m_n \le t] = P[x_1 \le t]^n$.) Make from m_n the random variable

$$y_n = n(\theta - m_n)$$

a) Find for $y>0$ the limit as $n \to \infty$ of

$$P[y_n \le y]$$

What is the approximate distribution of y_n?

b) What is an approximate distribution for y_n/θ? (What is the limit of $P\left[\frac{y_n}{\theta} \le t\right]$ for positive t?)

c) Use your answer to b) and find a large n approximately 95% upper confidence bound for θ that is a multiple of m_n.

21. As in the previous problem, suppose that $x_1, x_2, x_3, \ldots$ are iid $\mathrm{U}(0,\theta)$. For $\bar{x}_n$ the sample mean of the first n of these variables, find (using the CLT and a delta method argument) an approximate distribution for $\bar{x}_n^2$.

Section 2.3 of the Course Outline

22. Use `R` simulations to

a) Find approximate answers for all of Problem 11.

b) Find approximate answers for all of Problem 12.

(In both cases use at least 10,000 simulated values of the variables in question.)

Assignment 4 Due 10/17/14

23. Below is some `BUGS` code that can be used to do the simulations in Problem 22a). You can read about `BUGS` syntax in the user manual available through the help menu in either `WinBUGS` or `OpenBUGS`. Run this code in either `WinBUGS` or `OpenBUGS` with at least 100,000 iterations and get estimated densities for $U, V,$ and T and estimated values for the cdf of T at the values $t = 0, .2, .4, \ldots, 1.8, 2.0$.

```
model {
U~dunif(0,1)
V~dunif(0,1)
T<-U+V
I.2<-step(.2-T)
I.4<-step(.4-T)
I.6<-step(.6-T)
I.8<-step(.8-T)
I1.0<-step(1.0-T)
I1.2<-step(1.2-T)
I1.4<-step(1.4-T)
I1.6<-step(1.6-T)
I1.8<-step(1.8-T)
I2.0<-step(2.0-T)
}
```

24. Modify the code in Problem 23 and get estimated densities for U, V, and T for Problem 22b) using either WinBUGS or OpenBUGS with at least 100,000 iterations.

25. Suppose that $x \sim \text{Bi}(5, p)$ and *a priori* $p \sim \text{U}(0,1)$.

a) What is the posterior distribution of $p \mid x$?

b) Use WinBUGS or OpenBUGS with at least 100,000 iterations to approximate the posterior density and the posterior mean of p for the observation $x = 2$. Also find a corresponding 95% credible interval for p. Here is some relevant BUGS code:

```
model {
X~dbin(p,5)
p~dunif(0,1)
}
list(X=2)
```

Section 2.4 of the Course Outline

26. Below is some R code for implementing a non-parametric bootstrap for studying the distribution of

$$T_n = \text{the sample median of } n \text{ observations}$$

(supposing that one is making iid draws from some distribution). Apply this code to the sample of $n = 20$ numbers

$$0.18, 0.15, 0.14, 0.44, 2.89, 1.23, 0.54, 0.96, 0.15, 1.39, 0.76, 1.24, 4.42, 1.05, 1.04, 1.88, 0.65, 0.34, 0.59, 2.36$$

(that actually comprise a rounded sample of size $n = 20$ generated from an $\text{Exp}(1)$ distribution.

Use that code to estimate $\text{E}T_{20}, \sqrt{\text{Var}T_{20}}$, and the first and 9th deciles of the distribution of T_{20}.

```
#Set the seed on the random number generator to get the same results for
repeat runs
set.seed(0)

#Create some data input "data"
n<-20
observed<-c(round(rexp(n),digits=2))
observed

#Ready a matrix with B rows and n columns for B bootstrap samples of size n
B<-10000
Boot<-matrix(c(rep(0,B*n)),nrow=B,byrow=TRUE)

#Create the matrix of bootstrapped samples
```

```
for (i in 1:B) {
  Boot[i,]<-sample(observed,replace=TRUE,n)
}

#Ready a vector for bootstrapped values of T
Tstar<-c(rep(0,B))

#Make a vector of bootstrapped values of T
for (i in 1:B) {
  Tstar[i]<-median(Boot[i,])
}

#Get some summaries of the bootstrap distrbution of Tstar
summary(Tstar)
hist(Tstar)
quantile(Tstar,probs=seq(0,1,.025))
```

27. The mean of the observations listed in Problem 26 is $\bar{x} = 1.12$. If one assumes that the values in Problem 26 are a rounded sample from some exponential distribution, the exponential distribution with mean 1.12 is a plausible choice. (If one acknowledges the rounding, 1.12 is not necessarily quite the MLE of the exponential mean, but we'll ignore this fine point.) Thus a parametric bootstrap replaces sampling with replacement from the observed values with sampling from a rounded version of the exponential distribution with mean 1.12 (and thus "rate" $1/1.12$). In the code above, one can simply replace

```
sample(observed,replace=TRUE,n)
```

with

```
round(rexp(n,rate=1/1.12),2)
```

a) Redo Problem 26 using this parametric bootstrap.

b) Replace 1.12 above with 1.00 and thereby use simulation to find the true values of the characteristics of the distribution of the sample median of $n = 20$ rounded exponential variables that are being estimated in Problem 26 and above in part b).

Section 3.1 of the Course Outline

28. Suppose that $x_1, x_2, \ldots, x_n$ are iid $\text{Beta}(\alpha, \beta)$. Identify a two-dimensional sufficient statistic for the parameter vector (α, β).

29. Suppose that $x_1, x_2, \ldots, x_n$ are iid $\mathrm{N}\left(\mu, \sigma^2\right)$.

a) Argue carefully that the statistic

$$\mathbf{T} = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2 \right)$$

is sufficient for the parameter vector $\left(\mu, \sigma^2\right)$.

b) Then show that $\mathbf{S} = \left(\bar{x}, s^2\right)$ is a 1-1 onto function of $\mathbf{T}$ (is equivalent to $\mathbf{T}$) and is thus also sufficient.

30. For $x_1, x_2, \ldots, x_n$ iid $\text{Poisson}(\lambda)$ argue carefully that $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$ is minimal sufficient for λ. (Showing sufficiency is easy. Showing minimal sufficiency is somewhat harder. You can argue that if $T(\mathbf{x}) \neq T(\mathbf{x}')$ then $\Lambda_{\mathbf{x}}(\lambda, 1) \neq \Lambda_{\mathbf{x}'}(\lambda, 1)$ as functions of λ. That means that you have to show that for some particular λ the values of the likelihood ratios differ.)

31. Consider the statistical model for x with $\theta \in \{1, 2, 3\}$ and pmfs $f(x \mid \theta)$ given in the table below. Identify a minimal sufficient statistic for this model, say T . (Give values of T for each possible value of x.)

	x									
	1	2	3	4	5	6	7	8	9	10
$\theta = 3$	.05	.05	.10	.05	.10	.20	.05	.20	.10	.10
$\theta = 2$	.025	.05	.075	.10	.05	.10	.15	.10	.30	.05
$\theta = 1$	.05	.10	.15	.05	.10	.15	.05	.10	.10	.15

Section 3.2 of the Course Outline

32. Find the Fisher information in a single $\text{Bernoulli}(p)$ random variable about the parameter p in two different ways. First use the definition of Fisher information, then use Proposition 22 of the typed outline.

33. Find the Fisher information in a single Exponential variable with mean μ about the parameter μ. First use the definition of Fisher information, then use Proposition 22 of the typed outline.

34. Suppose that x and y are independent random variables, $x \sim \text{Poisson}(\mu)$ and y Exponential with mean μ.

a) What is the Fisher information in the pair (x, y) about μ?

b) If $T(x, y)$ is unbiased for μ, what is the minimum possible value for $\text{Var}_\mu T(x, y)$?

For a fixed number $\alpha \in (0,1)$ let $\hat{\mu}_\alpha(x, y) = \alpha x + (1-\alpha) y$.

c) Show that each $\hat{\mu}_\alpha$ is unbiased for μ.

d) Plot the SEL risk functions for both $\hat{\mu}_{\frac{1}{2}}$ and $\hat{\mu}_{\frac{2}{3}}$ over the range $0 < \mu < 4$. On the same set of axes plot, the Cramèr-Rao lower bound for the variance of an unbiased estimator of μ (this is a function of μ). For which values of μ do the two linear combinations of x and y achieve this lower bound?

<u>Assignment 5 (Not to be Collected, but to be *Covered on Exam 2*)</u>

35. Argue carefully that for $x_1, x_2, \ldots, x_n$ iid $\text{Bernoulli}(p)$ variables, $\hat{p}_n = \overline{x}_n$ is unbiased for p. Then show that the variance of this estimator achieves the Cramèr-Rao lower bound for the variance of an unbiased estimator of p for all values of p.

36. Argue carefully that for $x_1, x_2, \ldots, x_n$ iid Exponential variables with mean μ, $\overline{x}_n$ is unbiased for μ. Then show that the variance of this estimator achieves the Cramèr-Rao lower bound for the variance of an unbiased estimator of μ for all values of μ.

Section 3.3 of the Course Outline

37. Consider the context of Problem 35.

a) Argue that $\hat{p}_n = \overline{x}_n$ is the maximum likelihood estimator of p .

b) Argue two ways that for any $\varepsilon > 0$

$$P_{p_0}\left[\left|\hat{p}_n - p_0\right| > \varepsilon\right] \to 0 \text{ as } n \to \infty$$

First use Proposition 27 and then use Theorem 13 from the typed outline.

c) Argue two ways that for large n, valid but unusable approximately 95% confidence limits for p are

$$\hat{p}_n \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

First use Proposition 28 and then use Theorem 14 from the typed outline.

d) Use Corollary 29 and find a valid usable replacement for the limits of part c). (Use the "expected information" fix.)

e) Use Corollary 30 and find a valid usable replacement for the limits of part c). (Use the "observed information" fix.)

38. Suppose that $x_1, x_2, \ldots, x_n$ are iid $\text{Poisson}(\lambda)$.

a) Argue that $\hat{\lambda}_n = \overline{x}_n$ is the maximum likelihood estimator of λ .

b) Argue two ways that for any $\varepsilon > 0$

$$P_{\lambda_0}\left[\left|\hat{\lambda}_n - \lambda_0\right| > \varepsilon\right] \to 0 \text{ as } n \to \infty$$

First use Proposition 27 and then use Theorem 13 from the typed outline.

c) Argue two ways that for large n, valid but unusable approximately 95% confidence limits for λ are

$$\hat{\lambda}_n \pm 1.96\sqrt{\frac{\lambda}{n}}$$

First use Proposition 28 and then use Theorem 14 from the typed outline.

d) Use Corollary 29 and find a valid usable replacement for the limits of part c). (Use the "expected information" fix.)

e) Use Corollary 30 and find a valid usable replacement for the limits of part c). (Use the "observed information" fix.)

39. Suppose that $x_1, x_2,, \ldots, x_n$ are iid with marginal pmf $f(x \mid p)$ specified below.

x	0	2	3
$f(x \mid p)$	$(1-p)^3 + 3(1-p)^2\, p$	$3(1-p)\, p^2$	p^3

Let

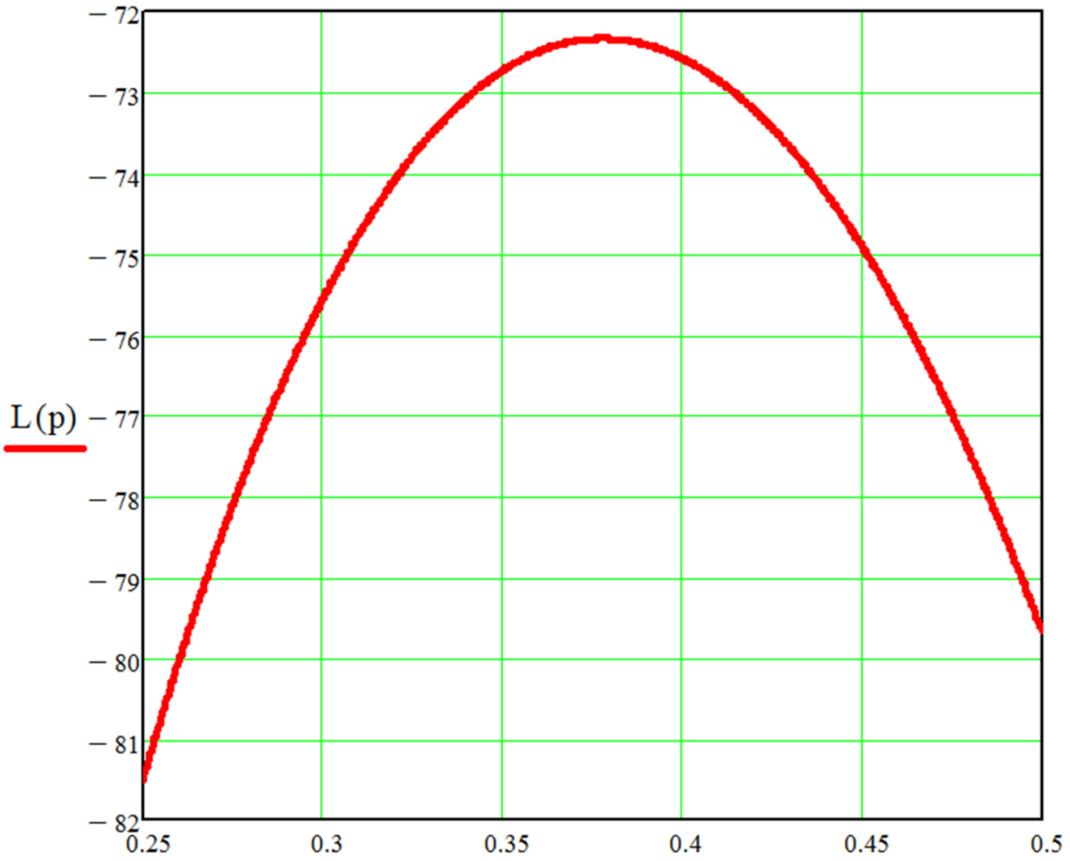
n_0 = the number of x_i taking the value 0
n_2 = the number of x_i taking the value 2
n_3 = the number of x_i taking the value 3

a) Give a formula for the log-likelihood function based on the x_i, $L(p)$.

b) A sample of size $n = 100$ produces $n_0 = 64, n_2 = 29,$ and $n_3 = 7$ and a log-likelihood that is plotted below. Further, some numerical analysis can be done to show that

$$L(.378) \approx -72.331, L'(.378) \approx 0 \text{ and } L''(.378) \approx -1011$$

Use Corollary 30 (the "observed information" fix) and give approximate 95% confidence limits for p based on this sample.



Section 3.4 of the Course Outline

40. Consider a statistical model where a single discrete observation x has probability mass function $f\left(x \mid \theta\right)$ indicated in the table below.

				x			
		0	1	2	3	4	5
	5	.1	0	.2	.4	.2	.1
	4	.1	.3	.2	.2	.1	.1
θ	3	.1	0	.2	.4	.2	.1
	2	.2	.1	.4	0	0	.3
	1	.2	.1	.4	.2	.1	0

a) Initially consider only the possibilities that $\theta = 4$ and $\theta = 1$. Identify a minimum Bayes risk test of $\mathrm{H}_0 : \theta = 4$ vs $\mathrm{H}_\mathrm{a} : \theta = 1$ for a prior distribution with $g(4) = .6$ and $g(1) = .4$. What are the size and power for this test?

b) Now consider testing $\mathrm{H}_0 : \theta = 4$ vs $\mathrm{H}_\mathrm{a} : \theta \neq 4$. Find values for the generalized log likelihood ratio statistic

$$\lambda(x) = \ln \max_{\theta} \Lambda_x(\theta, 4)$$

potentially useful in this simple versus composite testing problem. Consider the test that decides in favor of $\mathrm{H}_\mathrm{a} : \theta \neq 4$ exactly when $\lambda(x) > \ln(3/2)$. For which x values does this test reject H_0? What is the size of this test?

41. For $x_1, x_2, \ldots, x_n$ iid Exponential variables with mean μ, consider testing $\mathrm{H}_0 : \mu = 1$ vs $\mathrm{H}_\mathrm{a} : \mu \neq 1$.

a) Find the generalized log likelihood ratio statistic

$$\lambda_n(\mathbf{x}) = \ln \max_{\mu} \Lambda_{\mathbf{x}}(\mu, 1)$$

and note that it is a function of the sample mean. Argue that as a function of $\bar{x}$ it is decreasing to the left of $\bar{x} = 1$ and increasing to the right of $\bar{x} = 1$.

b) On the basis part a), for $n = 100$ give two equations in $\bar{x}$ that can be solved numerically to produce $d_{100}^{\mathrm{L}} < 1$ and $d_{100}^{\mathrm{U}} > 1$ so that a test rejecting H_0 if $\bar{x} < d_{100}^{\mathrm{L}}$ or $\bar{x} > d_{100}^{\mathrm{U}}$ is a likelihood ratio test of size approximately .05.

c) Show how you could use the Central Limit Theorem to evaluate the (Type II) error probability for this test for the possibility that $\mu = 2$. (Give a formula for this in terms of d_{100}^{L} and d_{100}^{U}.)

d) Suppose that $n = 100$ and $\bar{x}_{100} = 3.2$. Plot the log-likelihood and show graphically how you can make an approximately 95% confidence interval for μ. What is this interval?

42. Return to Problem 39 and the plot of the likelihood provided for a particular sample of size $n = 100$. Read off from that plot an approximate 95% confidence interval for p as the set of all p's with log-likelihood within .5 times the upper 5% point of the χ_1^2 distribution of the maximum log-likelihood.

Assignment 6 Due 11/14/14

Sections 4.1 and 4.2 of the Course Outline

43. Vardeman will send you an R file providing code that will allow you to study simple 2-class classification problems with $\mathbf{x} \in [0,1]^2$. As provided, the code is set up to study a case where

$$f(\mathbf{x}\,|\,0) = I\left[\mathbf{x} \in [0,1]^2\right], g(0) = .5, f(\mathbf{x}\,|\,1) = (x_1 + x_2)\, I\left[\mathbf{x} \in [0,1]^2\right], \text{ and } g(1) = .5$$

You can make simple modifications to change all of these, but for this problem, begin with this set-up. (You should thoroughly understand the code and be able to change parameters it uses and modify it to handle other problems.)

a) Find the form of the Bayes optimal classifier in this problem and use 2-variable calculus or geometry to find the error rate for this classifier. (Note that NO data-based approximation to the Bayes optimal classifier can have a real error rate lower than this value.)

b) Generate (using the code) a training set of size $N = 100$ for this problem. Make a plot of the regions in $[0,1]^2$ where the 5 nearest neighbor classifier classifies to 0 and to 1. How does this classifier compare to the optimal classifier? What is the $K = 5$ fold cross validation error rate for this classifier? Now make a similar plot and evaluate the corresponding cross-validation error rate for the 9 nearest neighbor classifier.

c) Redo part b) for a training set of size $N = 400$.

d) For the size $N = 400$ training set, find the k for a nearest neighbor classifier with the best $K = 10$ cross-validation error rate.

e) Redo part d) for the bootstrap estimate of error rate.

f) Common practice in predictive analytics is to look for the least complex classifier with predicted error rate "fairly close" to the minimum. What does this heuristic suggest is a good k in the present problem with $N = 400$?

44. Redo Problem 43 using

$$f\left(\mathbf{x}\,|\,1\right)=\left(x_1-x_2+1\right)I\left[\mathbf{x}\in\left[0,1\right]^2\right]$$

Assignment 7 Due 11/21/14

45. Return to the situation of Problem 43 and generate training samples of sizes 400, 4000, and 40000.

a) For each training set size, use cross-validation to identify a good k for use in a *k*nn classifier. Compare the resulting classifiers to the Bayes/optimal classifier in terms of appearance of a plot of $\hat{y}\left(x_1,x_2\right)$ (white for 0 and black for 1) at the points on the 51×51 grid used in the first set of classification code Vardeman distributed.

b) Create a test set of 100,000 pairs $\left(y,\mathbf{x}\right)$ (separate from those used to make the classifiers). Compute test error rates based on this set for the 3 classifiers of a). How do these compare to the theoretically optimal error rate you computed in 43a)?

46. (**Extra Credit, not required but worth an extra "point" if done completely/well**.) In Section 4.1 of the typed outline, the suggestion is made that one *might* try estimating $f\left(\mathbf{x}\,|\,0\right)$ and $f\left(\mathbf{x}\,|\,1\right)$ and with $\hat{g}\left(0\right)$ the fraction of $y=0$ cases in the training set, consider the classifier

$$I\left[\hat{g}\left(0\right)\hat{f}\left(\mathbf{x}\,|\,0\right)<\left(1-\hat{g}\left(0\right)\right)\hat{f}\left(\mathbf{x}\,|\,1\right)\right] \qquad (*)$$

In the context of Problem 43 with bivariate continuous $\mathbf{x}=\left(x_1,x_2\right)$, a way to estimate densities is for a "bandwidth parameter" $\sqrt{\lambda}$, to use

$$\hat{f}\left(\left(x_1,x_2\right)|\,0\right)=\frac{1}{N_0}\sum_{\substack{i\text{ s.t.}\\ y_i=0}}\frac{1}{2\pi\lambda}\exp\left(-\frac{1}{2\lambda}\left(\left(x_1-x_{1i}\right)^2+\left(x_2-x_{2i}\right)^2\right)\right)$$

and

$$\hat{f}\left(\left(x_1,x_2\right)|\,1\right)=\frac{1}{N_1}\sum_{\substack{i\text{ s.t.}\\ y_i=1}}\frac{1}{2\pi\lambda}\exp\left(-\frac{1}{2\lambda}\left(\left(x_1-x_{1i}\right)^2+\left(x_2-x_{2i}\right)^2\right)\right)$$

where N_0 and N_1 are the counts of $y=0$ and $y=1$ cases in the training set. For the $n=400$ training set of Problem 43, treat λ as a complexity parameter and compare several cases of λ in terms of the training error rates for the classifier (*) and appearance of a plot of $\hat{y}\left(x_1,x_2\right)$ (white for 0 and black for 1) at the points on the 51×51 grid used in the first set of classification code Vardeman distributed. (I'm guessing that a reasonable place to start looking for a good bandwidth parameter is around $\sqrt{\lambda}=.05$.)

Sections 4.3 and 4.4 of the Course Outline

47. Return to the situation of Problem 43. (Unless specifically indicated to the contrary, use the $N=400$ training set size below.)

a) Fit classification trees using the `tree()` function with both default parameter settings and those provided in Vardeman's code. Which "full" tree is simpler? Using the more complex full tree, use cross validation with the cost-complexity pruning idea to find a good sub-tree of the full tree. What is the cross-validation error rate for the chosen complexity/weight/number-of-final-nodes?

b) Run the `randomForest` code provided by Vardeman. Does the random forest predictor seem to behave more sensibly if it is built on a much larger training set than 400? (Also run the code for $n=4000$.) Compare error rates and plots of how the classifiers split up $\left[0,1\right]^2$ into classification regions.

c) Fit and compare (based on training error rates and plots of how the classifiers break up $\left[0,1\right]^2$ into classification regions) good logistic regression-based classifiers produced using `glmnet()` with $\alpha=0,.5,1$. (Use `lambda.1se`.)

d) How would you modify the classifiers you have identified in c) if you were given the information that actually, the prevalence of $y=0$ case in the universe is not as represented in the training set but is actually much more like $g\left(0\right)=.01$. (If the training set had been made up to be representative of the universe, one would have expected only about 4 instances of $y=0$.) (See the outline discussion of case-control studies.) Show how these classifiers split up $\left[0,1\right]^2$ into classification regions as compared to the ones from which they are derived.

e) Choosing between a number of possible cost parameters on the basis of cross-validation, find a good support vector classifier for the problem. What is its training error rate? Make a plot of how it breaks up $[0,1]^2$ into 2 classification regions.

f) Combine the three classifiers you identify in c) in 2 different ways. First, with equal weights average fitted probabilities that $y=0$ and use a classifier built on the average probability. Then simply use equal weight majority voting between the classifiers to define a new one. (These are the two ideas of the "Ensemble" video.) Compare these two classifiers to each other and to the three classifiers from which they were made in terms of a test set error rate like that made in Problem 45b).

48. Redo Problem 47 using

$$f(\mathbf{x}|1) = (x_1 - x_2 + 1) I\left[\mathbf{x} \in [0,1]^2\right]$$

In a) look for a choice of parameters that gives a large initial tree.

<u>Assignment 8 Not to be Collected but Covered on Exam 3 12/5/14</u>

Section 5.1 of the Course Outline

49. (Simple linear regression in terms of the normal linear model) Below is a small set of fake (x,y) data.

x	-1	0	1	2	3
y	4	4	3	3	2

Consider the (SLR) normal linear model

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

where the ε_i are independent $\text{Normal}\left(0,\sigma^2\right)$ random variables. This is written in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

for

$$\mathbf{Y}=\begin{pmatrix}y_1\\y_2\\y_3\\y_4\\y_5\end{pmatrix},\ \mathbf{X}=\begin{pmatrix}1&4\\1&4\\1&3\\1&3\\1&2\end{pmatrix},\ \boldsymbol{\beta}=\begin{pmatrix}\beta_1\\\beta_2\end{pmatrix},\ \text{and }\boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\end{pmatrix}$$

Using matrix calculations (either "by hand" or aided by the matrix calculation facility of R) do all of the following.

a) Find $\hat{\boldsymbol{\beta}}^{\text{OLS}}=\hat{\boldsymbol{\beta}}^{\text{MLE}}$. Based on this fit to the "training data," what is $\hat{y}^{\text{OLS}}(1.5)$? What is the function $\hat{y}^{\text{OLS}}(x)$?

b) Find SSE and the value of the MLE for σ^2.

c) Make 95% two-sided confidence limits for σ in the normal linear model.

d) Give a 90% two-sided confidence interval for the increase in average y that accompanies a unit increase in x. (See again the "non-matrix" form of the model.)

e) Give a 90% two-sided confidence interval for the average value of y when $x=1.5$.

f) Give a 90% two-sided prediction interval for the next value of y when $x=1.5$.

g) Give a 90% two-sided prediction interval for the sample mean of the next 5 values of y when $x=1.5$.

h) Give a 90% two-sided confidence interval for the difference in mean y at $x=-.5$ and mean y at $x=1.5$.

i) Give a 90% two-sided prediction interval for the difference between a new y at $x=-.5$ and a different new y at $x=1.5$.

50. (The one-way normal model in terms of the general normal linear model) In an ISU engineering research project, so called "tilt table tests" were done in order to determine the angles at which certain vehicles experience lift-off of the "high side" set of wheels and begin to roll over on their sides. So called "tilt table ratios" (which are the tangents of angles at which lift-off occurred) were measured for 4 different vans with the following results.

Van #1	Van #2	Van #3	Van #4
1.096, 1.093, 1.090, 1.093	.962, .970, .967, .966	1.010, 1.024, 1.021, 1.020,1.022	1.002, 1.001, 1.002, 1.004

(Notice that Van #3 was tested *5* times while the others were tested 4 times each.) Vans #1 and #2 were minivans and Vans #3 and #4 were full size vans.

We'll consider analysis of these data using a "one-way normal model" for

$$y_{ij} = j\text{th tilt table ratio for van } i$$

of the form

$$y_{ij} = \beta_j + \varepsilon_{ij}$$

for ε_{ij} iid $\mathrm{N}\left(0,\sigma^2\right)$. With

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{14} \\ y_{21} \\ \vdots \\ y_{24} \\ y_{31} \\ \vdots \\ y_{35} \\ y_{41} \\ \vdots \\ y_{44} \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} \underset{4\times1}{\mathbf{1}} & \underset{4\times1}{\mathbf{0}} & \underset{4\times1}{\mathbf{0}} & \underset{4\times1}{\mathbf{0}} \\ \underset{4\times1}{\mathbf{0}} & \underset{4\times1}{\mathbf{1}} & \underset{4\times1}{\mathbf{0}} & \underset{4\times1}{\mathbf{0}} \\ \underset{5\times1}{\mathbf{0}} & \underset{5\times1}{\mathbf{0}} & \underset{5\times1}{\mathbf{1}} & \underset{5\times1}{\mathbf{0}} \\ \underset{4\times1}{\mathbf{0}} & \underset{4\times1}{\mathbf{0}} & \underset{4\times1}{\mathbf{0}} & \underset{4\times1}{\mathbf{1}} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}, \ \text{and} \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{24} \\ \varepsilon_{31} \\ \vdots \\ \varepsilon_{35} \\ \varepsilon_{41} \\ \vdots \\ \varepsilon_{44} \end{pmatrix}$$

this can be written in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Using matrix calculations (either "by hand" or aided by the matrix calculation facility of R) do all of the following.

a) Find $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \hat{\boldsymbol{\beta}}^{\mathrm{MLE}}$.

b) Find SSE and the value of the MLE for σ^2.

c) Make 95% two-sided confidence limits for σ in the normal linear model.

d) Give 95% two-sided confidence limits for β_1 (the mean tilt table ratio for Van #1).

e) Give 95% two-sided confidence limits for β_3 (the mean tilt table ratio for Van #3).

f) Give 95% two-sided prediction limits for the next tilt table ratio for Van #3.

g) Give 95% two-sided confidence limits for $\beta_1 - \beta_3$ (the difference in mean tilt table ratios for Van #1 and Van #3).

h) It might be of interest to compare the average of the tilt table ratios for the minivans to that of the full size vans. Accordingly, give a 95% two-sided confidence interval for the quantity

$$\frac{1}{2}\left(\beta_1+\beta_2\right)-\frac{1}{2}\left(\beta_3+\beta_4\right)$$

<u>Assignment 9 Not to be Collected but Covered on the Final Exam</u>

Section 5 of the Course Outline (Practical SEL Prediction)

51. This question concerns the analysis of a set of home sale price data obtained from the Ames City Assessor's Office. Data on sales May 2002 through June 2003 of $1\frac{1}{2}$ and 2 story homes built 1945 and before, with (above grade) size of 2500 sq ft or less and lot size 20,000 sq ft or less, located in Low- and Medium-Density Residential zoning areas. (The data are in an Excel® spreadsheet on the Stat 342 Web page. These need to be loaded into R for analysis.) $n = 88$ different homes fitting this description were sold in Ames during this period. (2 were actually sold twice, but only the second sales prices of these were included in our data set.) For each home, the value of the response variable

$Price =$ recorded sales price of the home

and the values of 14 potential explanatory variables were obtained. These variables are

Size	the floor area of the home above grade in sq ft,
Land	the area of the lot the home occupies in sq ft,
Bedrooms	a count of the number in the home
Central Air	a **dummy** variable that is 1 if the home has central air conditioning and is 0 if it does not,
Fireplace	a count of the number in the home,
Full Bath	a count of the number of full bathrooms above grade,

Half Bath a count of the number of half bathrooms above grade,

Basement the floor area of the home's basement (including both finished and unfinished parts) in sq ft,

Finished Bsmnt the area of any finished part of the home's basement in sq ft,

Bsmnt Bath a **dummy** variable that is 1 if there is a bathroom of any sort (full or half) in the home's basement and is 0 otherwise,

Garage a **dummy** variable that is 1 if the home has a garage of any sort and is 0 otherwise,

Multiple Car a **dummy** variable that is 1 if the home has a garage that holds more than one vehicle and is 0 otherwise,

Style (2 *Story*) a **dummy** variable that is 1 if the home is a 2 story (or a $2\frac{1}{2}$ story) home and is 0 otherwise, and

Zone (*Town Center*) a **dummy** variable that is 1 if the home is in an area zoned as "Urban Core Medium Density" and 0 otherwise.

a) In preparation for analysis, standardize all variables that are not dummy variables (those we'll leave in raw form), making a data frame with 15 columns. Say clearly how one goes from a particular new set of home characteristics to a corresponding set of predictors. Then say clearly how a prediction for the standardized price to a prediction for the dollar price.

b) Find predictors for standardized price of all the following forms:

- OLS
- Lasso (choose λ by cross-validation)
- Ridge (choose λ by cross-validation)
- Elastic Net with $\alpha = .5$ (choose λ by cross-validation)
- Nearest Neighbor (based on the $k \le 4$ predictors that have the largest coefficients in the linear predictors you identify)
- Single Regression Tree (choose the tree by cost-complexity pruning of full trees)
- Random Forest (use default parameters)
- N-W Kernel Smoother (based on the $k \le 4$ predictors that have the largest coefficients in the linear predictors you identify) (look for a good bandwidth)
- Local Linear Regression Smoother (based on the $k \le 4$ predictors that have the largest coefficients in the linear predictors you identify) (look for a good bandwidth)
- Neural Network (use one hidden layer with 8 nodes and cross-validation in the fitting)

Find the sets of predictions all these methods produce for the training set. Compare these sets of predictions by making scatterplots for all pairs of prediction types and computing all pairs of correlations between predictions. Which two methods give the least similar predictions? (If these both have good cross-validation errors, they would become good candidates for combining into a single "stacked" predictor.)