Solution

Practice for Concepts in CASI Chapters 4 and 5

Finish Example from end of class on Poisson MLE and Cramer Rao Bound.

Then, try these questions with those around you. Important points from chapters 4 and 5 are covered.

1. Binomial Info

Find the Fisher information for a single observation drawn from a Binomial(n, p) distribution. Remember that the pmf takes the form: $P(X=x)=\binom{n}{x}(p)^x(1-p)^{n-x}, x=0,\ldots n, 0<p<1.$

$$\ell(x|p) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p)$$

$$\ell'(x|p) = \frac{x}{p} + \frac{n-x}{1-p}\cdot(-1) = xp^{-1} - (n-x)(1-p)^{-1}$$

$$\ell''(x|p) = -xp^{-2} + (n-x)(1-p)^{-2}(-1) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}$$

$$-E(\ell''(x|p)) = -E\left[\frac{-x}{p^2} - \frac{n-x}{(1-p)^2}\right] = \frac{E(x)}{p^2} + \frac{n-E(x)}{(1-p)^2}$$

$$= \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p} + \frac{n(1-p)}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = \frac{n-pn+pn}{p(1-p)}$$

$$= \frac{n}{p(1-p)} = I(p) \quad \text{(Fisher Info)} \Rightarrow \text{Var bound is } \frac{p(1-p)}{n}$$

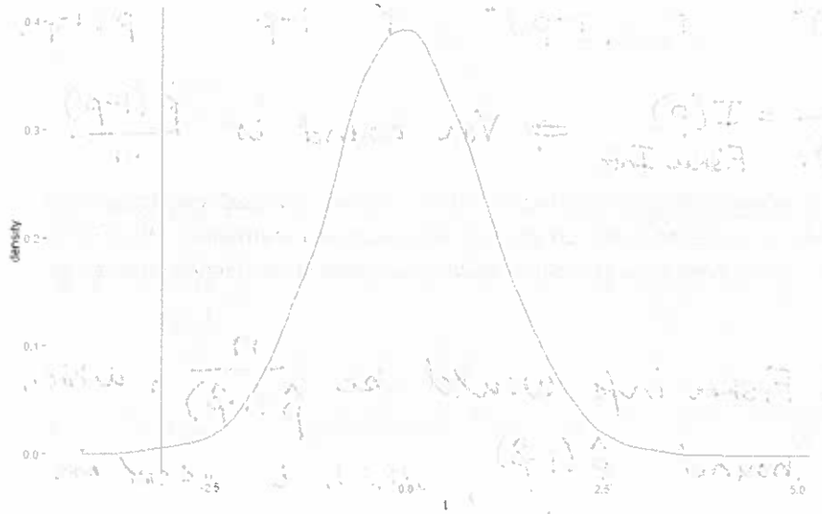
Now, the computation above gave the theoretical (expected) information. The text suggests considering the observed Fisher information (evaluate based on the MLE for unknown parameters). What is the observed Fisher information here? How does this value relate to our typical confidence intervals for a proportion?

The observed Fisher info would be $\frac{n}{\hat{p}(1-\hat{p})}$, which would make the bound $\frac{\hat{p}(1-\hat{p})}{n}$, which is used in spread component for making the CI for p. ($\hat{p}=\frac{x}{n}$ can be derived as MLE for p.) $\hat{p} \pm z^{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

2. Randomization/Permutation Procedures

The Leukemia data set for gene 136 (with correct labels) was used in the analysis below. The mean expression values for ALL and AML are being compared. Use the output to address the questions that follow.

```
leukemia_big <- read.csv("http://web.stanford.edu/~hastie/CASI_files/DATA/leu
kemia_big.csv")
gene136 <- t(leukemia_big[136, ]) #says pictures from row 136
type <- c(rep("ALL", 20), rep("AML", 14), rep("ALL", 27), rep("AML", 11))
leukemia <- data.frame(gene136, type)
leukemia <- rename(leukemia, gene136 = X136)


t.test(gene136 ~ type, data = leukemia)

        Welch Two Sample t-test

data:  gene136 by type
t = -3.1323, df = 54.667, p-value = 0.002786
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.32386725 -0.07112012
sample estimates:
mean in group ALL mean in group AML
        0.7524794         0.9499731

set.seed(500) #make the results reproducible
ttest <- do(10000) * (t.test(gene136 ~ shuffle(type), data = leukemia)$statis
tic)
ttest <- as.data.frame(ttest)

gf_dens(~ t, data = ttest) %>% gf_vline(xintercept = -3.1323, color = "red")
```




```
pdata(~ t, -3.1323, data = ttest, lower.tail = TRUE)
[1] 0.0017
```

a. Explain what the t-test results on the original data suggest. (Assuming appropriate conditions hold).

$t = -3.1323$ and the corresponding small p-value suggest a sign. diff in mean expression levels for ALL vs. AML for gene 136.

b. There were concerns about the original t-test analysis. In your own words, explain what distribution is being created in the densityplot, and how it can be used to perform an appropriate analysis.

10000 t stats are computed in a setting where group labels have been scrambled. This effectively is generating an empirical sampling distribution for the t stats when there is no diff in mean expression levels by group (ALL vs. AML), ie. when the null hypothesis is true.

c. What do you conclude from the randomization/permutation test results?

The observed t stat is far in the lower tail of the empirical null sampling dist. This suggests that the result would be unusual if H_0 was true. Hence, we reject H_0, and conclude there is a significant diff. in mean exp. levels for gene 136 for ALL vs. AML.

3. **Quiz Retake**

nice math!

a. What explains the classical preference for parametric models? Mathematical tractability

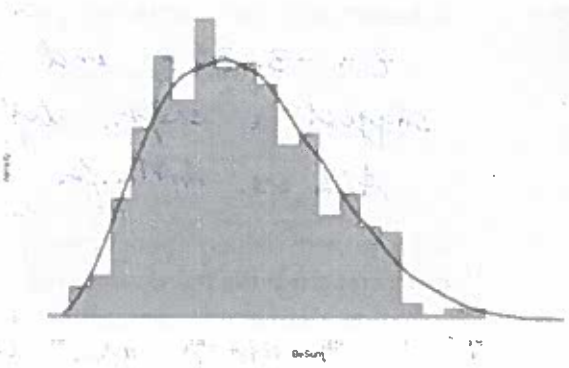
b. Name two distributions that are part of an exponential family.

Poisson, Beta, Gamma, Normal, Exp., Chisq.

c. Name one of the two multivariate distributions examined in depth in the chapter.

Multivariate normal, Multinomial

4. Creating Betas - The textbook states that there are many relationships among the distributions listed in table 5.1. They give an example of creating a Beta from two independent Gamma RVs with the same second parameter. This is illustrated below:



```
set.seed(495)
Gam1 <- rgamma(500, 3, 5)
Gam2 <- rgamma(500, 6, 5)
BeSum <- Gam1/(Gam1+Gam2)

BeMatch <- rbeta(50000, 3, 6)

gf_dhistogram(~ BeSum) %>% gf_dens(~ BeMatch, color = "blue", size = 2)
```

Verify, using appropriate variable transformation techniques, that if X and Y are independent Gamma RVs with the same second parameter, then W = X/(X+Y) follows a Beta distribution with parameters inherited from X and Y. The statement from the text is provided for your reference (CASI page 54). Hint: Jacobians.

Gamma $\text{Gam}(\nu, \sigma)$	$\frac{x^{\nu-1}e^{-x/\sigma}}{\sigma^{\nu}\Gamma(\nu)}$	$x \geq 0$	$\nu > 0$ $\sigma > 0$	$\sigma\nu$ $\sigma^2\nu$
Beta $\text{Be}(\nu_1, \nu_2)$	$\frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1}(1-x)^{\nu_2-1}$	$0 \leq x \leq 1$	$\nu_1 > 0$ $\nu_2 > 0$	$\nu_1/(\nu_1+\nu_2)$ $\frac{\nu_1\nu_2}{(\nu_1+\nu_2)^2(\nu_1+\nu_2+1)}$

Relationships abound among the table's families. For instance, independent gamma variables $\text{Gam}(\nu_1, \sigma)$ and $\text{Gam}(\nu_2, \sigma)$ yield a beta variate according to

$$\text{Be}(\nu_1, \nu_2) \sim \frac{\text{Gam}(\nu_1, \sigma)}{\text{Gam}(\nu_1, \sigma) + \text{Gam}(\nu_2, \sigma)} \tag{5.3}$$

$W = \frac{X}{X+Y}$ $\quad X, Y \perp$ Gamma

r_X, σ and r_Y, σ for parameters

$R = X+Y \rightarrow$ should be Gamma$(r_X + r_Y, \sigma)$ could verify via mgfs.

Inverses $\quad X = RW \quad Y = R - RW = R(1-W)$ $\quad R > 0$, $0 < W < 1$

$$J = \begin{vmatrix} d/dw & d/dR \\ R & W \\ -R & 1-W \end{vmatrix} = R(1-W) + RW = R \qquad |R| = R$$

$f(w, r) = f_{X,Y}(RW, R(1-W)) \cdot R$ $\quad$ use product of two Gamma densities $\cdot R$

$$= \frac{(RW)^{r_X-1} e^{-(RW)/\sigma}}{\sigma^{r_X}\Gamma(r_X)} \cdot \frac{(R(1-W))^{r_Y-1} e^{-(R(1-W)/\sigma)}}{\sigma^{r_Y}\Gamma(r_Y)} \cdot R$$

Combine common terms.

Second page for #4. Hint: The sum X+Y should have a Gamma distribution. This is verifiable by mgfs, but you can take it as fact. You should be able to use W and the sum in the Jacobian transformation process fairly easily, and then integrate out the sum from the resulting joint distribution. For the integration, the trick is to recognize it as a distribution you know, but with pieces missing. (Integrate without integrating.)

$$f(w,r) = \frac{R^{r_x+r_y-1}\, w^{r_x-1}(1-w)^{r_y-1}\, e^{-\frac{1}{\sigma}(Rw+R(1-w))}}{\sigma^{r_x+r_y}\,\Gamma(r_x)\Gamma(r_y)}$$

R

for $R>0$, $0<w<1$

Need to integrate out R.

convenient "1"

looks Gamma (r_x+r_y, σ)

$$f_w(w) = \frac{w^{r_x-1}(1-w)^{r_y-1}}{\Gamma(r_x)\Gamma(r_y)} \cdot \frac{\Gamma(r_x+r_y)}{1} \int_0^\infty \frac{1}{\Gamma(r_x+r_y)} \frac{R^{r_x+r_y-1}}{\sigma^{r_x+r_y}} e^{-R/\sigma}\, dR$$

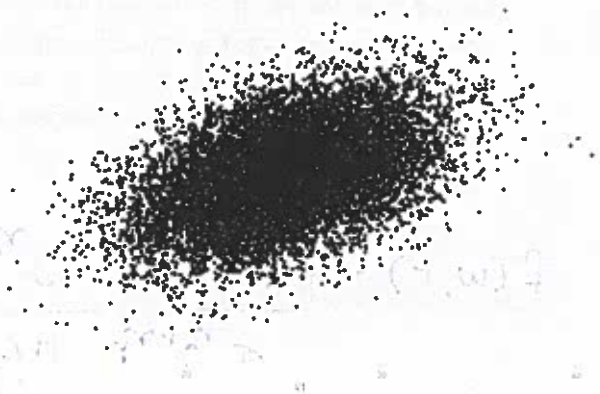
integrates to 1

$$= \frac{\Gamma(r_x+r_y)}{\Gamma(r_x)\Gamma(r_y)}\, w^{r_x-1}(1-w)^{r_y-1}, \quad 0<w<1$$

$W \sim \text{Beta}(r_x, r_y)$ as expected

5. Multivariate Normals

Multivariate normals can be simulated in R using the mvtnorm package. You need a mean vector and the covariance matrix. An example in 2-D is provided.



```
set.seed(52)
sigma <- matrix(c(16, 9, 9, 25), ncol=2)
NormData <- as.data.frame(rmvnorm(10000,
mean = c(25, 55), sigma = sigma))
gf_point(V2 ~ V1, data = NormData)
with(NormData, cor(V1, V2))
[1] 0.4474522
```

a. Find the theoretical correlation between V1 and V2 based on the provided covariance matrix. How well does it match the observed correlation? (Be sure you understand what each entry is in the covariance matrix, ask if unsure!)

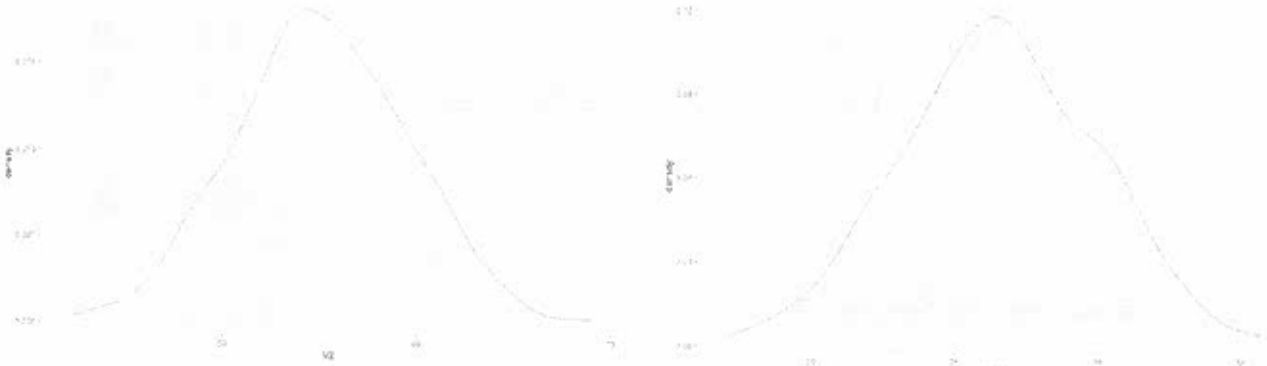
$$\Sigma = \begin{bmatrix} 16 & 9 \\ 9 & 25 \end{bmatrix}$$

$$Cor(X,Y) = \frac{Cov(X,Y)}{SD(X)\,SD(Y)} = \frac{9}{4 \cdot 5} = \frac{9}{20} = 0.45$$

0.45 is very close to the observed 0.4475

b. The following code and plots illustrate a neat property of the Multivariate normal. What property is being shown?

```
rounded <- round(NormData, 0)
rounded %>% filter(V1 == 25) %>% gf_dens(~ V2)
rounded %>% filter(V2 == 60) %>% gf_dens(~ V1)
```



If X_1 and X_2 (or if there are more) are bivariate (multi) normal, then the dists $X_2 | X_1$ and $X_1 | X_2$ are also normal.

6. Multinomial Distributions

The second multivariate distribution covered in the chapter is the multinomial, a generalization of the binomial. There are multiple functions available in R to simulate from this distribution, depending on what you want. You can obtain a summary of the counts from a random draw of a certain number of objects from the distribution, or get the category numbers as data to use.

```
rmultinom(1, 40, c(0.1, 0.2, 0.3, 0.4)) #gives count summaries
     [,1]
[1,]    4
[2,]    9
[3,]    9
[4,]   18
```

A related function in the Hmisc package gives you a random draw of the category numbers as a vector.

```
m <- rMultinom(rbind(c(0.1, 0.2, 0.3, 0.4)), 40)
m #shows all forty entries are 1:4
t(apply(m, 1, table)/40)
     1     2    3     4
[1,] 0.05 0.225 0.35 0.375
```

a. Explain why X1 = number of observations in category 1 and X2 = number of observations in category 2 have a negative covariance/correlation in a multinomial setting.

The total # of objects, n, is fixed in this setting. If a category count increases for 1 category, then somewhere there must be a ↓ (perhaps in more than 1 category). This induces a negative correlation/covariance structure.

b. What distribution has neat relationships to the multinomial as described in the text?

The Poisson is focused on.

They do mention power of multinomial for discrete settings

7. Exponential Families

The text presents exponential families via a formula relating any two densities in the family via a renormalized exponential tilt. There are other ways to recognize that a density is in an exponential family.

Suppose X is a random variable with density given by: $f(x \mid \theta) = a(\theta)b(x)\exp[c(\theta)d(x)]$, where a() and c() are functions only of the parameter theta, and b() and d() are functions of x (the data).

For example, for the Bernoulli distribution, some re-writing enables us to see that:

Then X is an an Exp family.

$$f(x \mid p) = p^x(1-p)^{1-x} = (1-p)\left(\frac{p}{1-p}\right)^x = (1-p)\exp\left[x\log\left(\frac{p}{1-p}\right)\right].$$

Here, $a(p) = 1-p, b(x) = 1, c(p) = \log\left[\frac{p}{1-p}\right], d(x) = x$. Hint: $a = \exp(\log(a))$.

If you have a random sample of observations from a distribution in this family, then the sample mean (or sum of the sample observations) is a sufficient statistic for the parameter(s).

Verify that the Poisson density can be written in this form.

Poisson
$\text{Poi}(\mu)$ $\quad \frac{e^{-\mu}\mu^x}{x!}$

$$f(x|\mu) = \frac{e^{-\mu}\mu^x}{x!} = e^{-\mu}\frac{1}{x!}e^{x\log\mu}$$

$a(\theta) = a(\mu) = e^{-\mu}$ $\qquad c(\mu) = \log\mu$

$b(x) = \frac{1}{x!}$ $\qquad d(x) = x$

The Poisson satisfies the form, as requested.

Not all distributions we have seen exist in exponential families.

An example of a distribution in a *curved* exponential family is the Cauchy distribution.