

## Homework 4 Outline of Solution

**Problem 1.** A study to determine whether two toxins were carcinogenic was performed on laboratory rats. The results are given below:

		<i>Toxin</i>		
		#1 ( $A_1$ )	#2 ( $A_2$ )	
<i>Occurrence</i>	No ( $B_1$ )	24	18	42
<i>of tumor</i>	Yes ( $B_2$ )	16	2	18
		40	20	60

(a) Test the independence of A and B using Pearson's  $X^2$  statistic.

**Solution.** Denote by  $n_{ij}$  the counts in the table ( $i$  indexing rows and  $j$  indexing columns). Under the null-model of  $A$  and  $B$  (rows and columns) being independent, the expected table counts (conditional on the marginal totals) are given by  $E_{ij} \equiv n_i \cdot n_{\cdot j} / n_{\dots}$ . Pearson's  $X^2$  statistic is defined as

$$X^2 \equiv \sum_{i,j} (n_{ij} - E_{ij})^2 / E_{ij}.$$

S-PLUS has a function that does the Pearson's test (`chisq.test`). First, import the data into S-PLUS:

```
> prob.1 <- matrix(c(24,16,18,2),2,2)
> dimnames(prob.1) <- list(c('Tumor: no ', 'Tumor: yes'),
+                           c('Tox 1', 'Tox 2'))
> prob.1
      Tox 1 Tox 2
Tumor: no   24   18
Tumor: yes  16    2
>
```

Then, use the `chisq.test` function:

```
> chisq.test(prob.1)
```

Pearson's chi-square test with Yates' continuity correction

```
data:  prob.1
X-square = 4.375, df = 1, p-value = 0.0365
```

(Yates' continuity correction replaces  $(n_{ij} - E_{ij})$  with  $(n_{ij} - E_{ij} - 1/2)$ ).  
Without the correction:

```
> chisq.test(prob.1,correct=F)
```

Pearson's chi-square test without Yates' continuity correction

```
data: prob.1
X-square = 5.7143, df = 1, p-value = 0.0168
```

In both cases we reject the null-model (null-hypotheses) of independence of rows and columns (tumor and toxic) at the 5% significance level, but not at 1% level.

The expected counts under the null-model are easily computed:

```
> row.marg <- apply(prob.1,1,sum)
> row.marg
  Tumor: no  Tumor: yes
           42         18
> col.marg <- apply(prob.1,2,sum)
> col.marg
  Tox 1 Tox 2
       40   20
> E <- (row.marg %*% t(col.marg)) / sum(prob.1)
> E
      Tox 1 Tox 2
[1,]    28   14
[2,]    12    6
>
```

And so are  $X^2$  and the p-value:

```
> X2 <- sum((prob.1 - E)^2/E)
> X2
[1] 5.714286
> 1 - pchisq(X2,1)
[1] 0.01682741
>
```

(b) Compute the estimated relative risk  $\hat{\rho}$  of Toxin #1 relative to Toxin #2.

**Solution.** The estimated relative risk of Toxin #1 relative to Toxin #2 is given by

$$\hat{\rho} = \frac{n_{21}/n_{.1}}{n_{22}/n_{.2}}.$$

For the data given in the table,  $\hat{\rho} = 4$ .

- (c) Compute the estimated odds ratio  $\hat{\omega}$  of Toxin #1 relative to Toxin #2.

**Solution.** The estimated odds ratio  $\hat{\omega}$  of Toxin #1 relative to Toxin #2 is given by

$$\hat{\omega} = \frac{n_{21}/n_{11}}{n_{22}/n_{12}}.$$

For the data given in the table,  $\hat{\omega} = 6$ .

**Problem 2.** Consider the two-parameter lognormal random variable  $X = e^Y$ , where  $Y \sim N(\mu, \sigma^2)$ .

- (a) Derive the density  $f_X(x)$  of  $X$ .

**Solution.** We have

$$F_X(x) = \Pr(X \leq x) = \Pr(Y \leq \log(x)) = F_Y(\log(x)).$$

Taking the derivative gives  $f_X(\cdot)$ , the density, as

$$\begin{aligned} f_X(x) &= \frac{\partial F_X(x)}{\partial x} = \frac{\partial F_Y(\log(x))}{\partial x} = f_Y(\log(x))(1/x) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right). \end{aligned}$$

- (b) Plot the density for  $(\mu, \sigma^2) = (0, 1)$ ,  $(\mu, \sigma^2) = (3, 1)$ , and  $(\mu, \sigma^2) = (0, 10)$ .

**Solution.**

Fig. 1 shows the three different densities.

- (c) Show that the mode of  $f_X(x)$  is at  $x = \exp(\mu - \sigma^2)$ .

**Solution.** To find the mode of  $f_X(x)$  we take the derivative and solve for the  $x$  that makes the derivative zero. Note first that finding the mode of  $f_X(x)$  is equivalent to finding the mode of  $\log(f_X(x))$  (and it is easier in this case). We have,

$$\frac{\partial \{\log(f_X(x))\}}{\partial x} = -\frac{1}{x} \left[ \frac{\log(x) - \mu}{\sigma^2} + 1 \right],$$

which is equal to zero if and only if,  $(\log(x) - \mu)/\sigma^2 + 1 = 0$ , or  $x = \exp(\mu - \sigma^2)$ . Hence, the mode is at  $\exp(\mu - \sigma^2)$ .

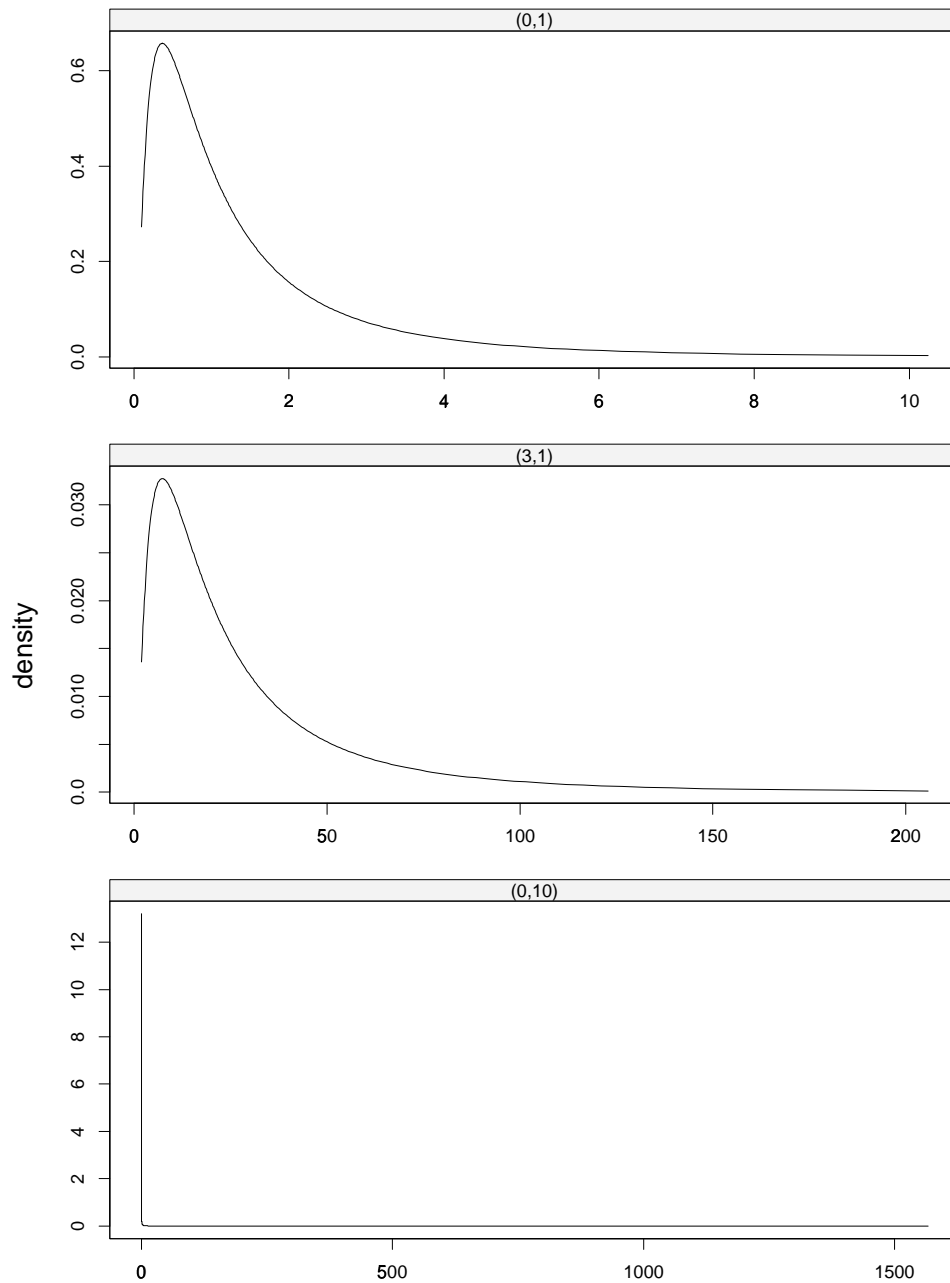


Figure 1: The lognormal density for three different values of  $\mu$  and  $\sigma^2$ :  $(\mu, \sigma^2) = (0, 1)$ ,  $(\mu, \sigma^2) = (3, 1)$ , and  $(\mu, \sigma^2) = (0, 10)$ . Note the different scale. Each density was plotted in the (1%, 99%) percentile range.

- (d) If  $u_\alpha$  is the  $\alpha$ -th quantile of the standard normal distribution (i.e.,  $\Phi(u_\alpha) = \alpha$ ), show that the  $\alpha$ -th quantile of  $X$  is  $x_\alpha = \exp(\mu + u_\alpha\sigma)$ ,  $0 < \alpha < 1$ . Evaluate  $x_{.25}$ ,  $x_{.5}$ , and  $x_{.75}$ .

**Solution.** The  $\alpha$ -th quantile of the distribution of  $X$  is defined as that  $x_\alpha$  such that  $\Pr(X \leq x_\alpha) = \alpha$ . Now,

$$\Pr(X \leq x) = \Pr(Y \leq \log(x)) = \Pr\left(\frac{Y - \mu}{\sigma} \leq \frac{\log(x) - \mu}{\sigma}\right),$$

where  $(Y - \mu)/\sigma$  has a standard normal distribution ( $N(0, 1)$ ). Then  $\Pr(X \leq x_\alpha) = \alpha$  if and only if  $(\log(x_\alpha) - \mu)/\sigma = u_\alpha$ , where  $u_\alpha$  is the  $\alpha$ -th quantile of a standard normal. Hence,

$$x_\alpha = \exp(\mu + u_\alpha\sigma)$$

Now;  $u_{1/4} \approx -0.675$ ,  $u_{1/2} = 0$ , and  $u_{3/4} \approx 0.675$ . Therefore;  $x_{1/4} \approx \exp(\mu - 0.675\sigma)$ ,  $x_{1/2} = \exp(\mu)$ , and  $x_{3/4} = \exp(\mu + 0.675\sigma)$ .

**Problem 3.** Let  $X_1, \dots, X_n$  be i.i.d. according to the lognormal distribution, as defined in 2. Show that the maximum likelihood estimators  $\hat{\mu}$  and  $\hat{\sigma}$ , are given by

$$\hat{\mu} = (\sum \log x_i)/n, \quad \hat{\sigma} = \{\sum(\log x_i - \hat{\mu})^2/n\}^{1/2}.$$

**Solution.**

Let  $Y_i \equiv \log(X_i)$ , then  $Y_i \sim N(\mu, \sigma^2)$ , independently. The maximum likelihood estimators  $\hat{\mu}$  and  $\hat{\sigma}$  are then given by  $\hat{\mu} = \bar{y}$  and  $\hat{\sigma} = \{\sum(y_i - \hat{\mu})^2/n\}^{1/2}$ . Plugging in for  $y_i = \log(x_i)$  gives the required result.

**Problem 4.** Consider the problem of left-censoring in the case where the uncensored observations come from a location-scale family whose density is  $f((x - \mu)/\sigma)/\sigma$  and whose cdf is  $F((x - \mu)/\sigma)$ , where  $dF(z)/dz = f(z)$ .

- (a) There are  $n_1$  censored data, censored at  $L_1, \dots, L_{n_1}$ , respectively. The remaining  $n_2$  observations  $X_{n_1+1}, \dots, X_{n_1+n_2}$  are uncensored. Show that the log likelihood of the left-censored data,  $L_1, \dots, L_{n_1}, X_{n_1+1}, \dots, X_{n_1+n_2}$ , is:

$$\ell(\mu, \sigma) = \sum_{i=1}^n \ell_i(\mu, \sigma),$$

with  $n = n_1 + n_2$ , where for  $i = 1, \dots, n$ ,

$$\ell_i(\mu, \sigma) = \begin{cases} \log f\left(\frac{x_i - \mu}{\sigma}\right) - \log \sigma; & \text{if } x_i \text{ is uncensored} \\ \log F\left(\frac{L_i - \mu}{\sigma}\right); & \text{if left-censoring occurs at } L_i. \end{cases}$$

**Solution.**

From class, we have that

$$\ell(\mu, \sigma) = \sum_{i=1}^n \ell_i(\mu, \sigma),$$

with

$$\ell_i(\mu, \sigma) = \begin{cases} \log f_X(x; \mu, \sigma), & \text{if } x_i \text{ is uncensored} \\ \log F_X(L_i; \mu, \sigma), & \text{if left-censoring occurs at } L_i. \end{cases}$$

Now,  $\log f_X(x; \mu, \sigma) = f((x - \mu)/\sigma) - \log \sigma$  and  $F_X(L_i; \mu, \sigma) = F((L_i - \mu)/\sigma)$ . Plugging this in gives the desired result.

**Problem 5.** Consider the censored data:

ND, ND, 4.5, 5.3, 5.3, 5.6, 7.7, 8.5, 9.5, 9.6, 11.0, 17.0, 17.0,

where ND (non detect) denotes a measurement  $< 4.5$ . The value  $L_1 = L_2 = 4.5$  is known as the LOD (limit of detection). Assume that, before censoring, the data were lognormally distributed, as defined in 2.

- (a) Ignore the NDs and find maximum likelihood estimates of  $\mu$  and  $\sigma$  based on the 11 remaining measurements.

**Solution.** Using S-PLUS:

```
> X <- c(NA,NA,4.5,5.3,5.3,5.6,7.7,8.5,9.5,9.6,11,17,17)
> x <- log(X[-(1:2)])
> mean(x)
[1] 2.120084
> sqrt(var(x,unbiased=F))
[1] 0.4353482
```

That is,  $\hat{\mu}_a \approx 2.120$  and  $\hat{\sigma}_a \approx 0.435$ .

- (b) After log-transforming the data, find estimates of  $\mu$  and  $\sigma$  based on the median and the interquartile range.

[Hint: From a sample of size 13, the lower quartile is the 4th obs., the median is the 7th obs., and the upper quartile is the 10th obs. The interquartile range is the difference between the upper quartile and the lower quartile.]

**Solution.**

The median of the data is 7.7; hence  $\hat{\mu}_b = \log 7.7 \approx 2.041$ . The interquartile range is  $\log 9.6 - \log 5.3 \approx 0.594$ . From problem 2, we have that  $\text{IQR} = u_{3/4} - u_{1/4} \approx (\mu + 0.675\sigma) - (\mu - 0.675\sigma) = 1.35\sigma$ , for a  $N(\mu, \sigma^2)$  distributed random variable. Our estimate of  $\sigma$  is therefore,  $\hat{\sigma}_b = \text{IQR}/(u_{3/4} - u_{1/4}) \approx 0.594/1.35 \approx 0.440$ .

- (c) Replace the two NDs with the LOD, 4.5, and obtain maximum likelihood estimates of  $\mu$  and  $\sigma$ , assuming now that the 13 observations are lognormally distributed.

**Solution.**

The MLEs of  $\mu$  and  $\sigma$  were derived in problem 2. Using S-PLUS:

```
> x <- c(NA,NA,4.5,5.3,5.3,5.6,7.7,8.5,9.5,9.6,11,17,17)
> x <- log(x)
> x[1:2] <- log(4.5)
> mean(x)
[1] 2.025314
> sqrt(var(x,unbiased=F))
[1] 0.4580039
```

That is,  $\hat{\mu} \approx 2.025$  and  $\hat{\sigma} \approx 0.458$ .

- (d) Comment on each of the methods you have used in (a), (b), and (c), to deal with the presence of NDs.

**Solution.**

Method (a) under-estimates both  $\mu$  and  $\sigma$ , method (c) over-estimates  $\mu$  but under-estimates  $\sigma$ . Method (b) is unbiased, but uses the data inefficiently, but is the preferred one of these three methods.

**Problem 6.** *Limit of Detection (S-PLUS).* Consider the censored data given in problem 5, made by an instrument with LOD of 4.5. The original observations are assumed independent and identically lognormally distributed, that is,  $\log(X_i) \sim N(\mu, \sigma^2)$ , for  $i = 1, \dots, 13$ .

The data can be typed into S-PLUS by

```
> X <- c(NA,NA,4.5,5.3,5.3,5.6,7.7,8.5,9.5,9.6,11,17,17)
```

where we have used NA (Not Available) in S-PLUS to denote ND observations (NA is the symbol used in S-PLUS to denote 'not available', or 'missing' observations — in our case, those are NDs).

(a) *Graphical Method.* Use the graphical method given in class to estimate  $\mu$  and  $\sigma$ , as follows:

*i* Compute the quantiles needed from the normal density (see hint below) and plot  $\log(X)$  versus the normal quantiles. Label the x-axis with 'standard normal quantile' and the y-axis with ' $\log(X)$ '.

*ii* Fit an OLS regression of  $\log(X)$  versus the normal quantiles (see hint below) to get estimates of  $\mu$  and  $\sigma$  (remember to ignore ND observations). Add the fitted line to the plot by `abline(mu,sigma)`, where `mu` and `sigma` are the estimated values (intercept and slope) from the OLS regression.

*iii* Carry out steps 4 and 5 as given in class notes and derive the refined estimates of  $\mu$  and  $\sigma$ . Add another line to the plot, using the refined estimates, with line type 2 (`lty=2`) in the `abline` function.

How much is gained (how different are  $\mu$  and  $\sigma$ ) by carrying out steps 4 and 5?

Hint: The function `qnorm` gives quantiles from the standard normal distribution; for example, use `qnorm((1:13 - 1/2)/13)`. The function `lm` fits OLS regression (see S-PLUS notes).

(b) *Maximum Likelihood.* The function `sensorReg` in S-PLUS deals with censored data via maximum likelihood. In our case, observations are either *left* censored, with LOD as censoring value, or observed. To get MLEs of  $\mu$  and  $\sigma$ , do the following in S-PLUS:

```
> cens.code <- ifelse(is.na(X), 0, 1) ## 1=obs, 0=cens
> X[1:2] <- 4.5 ## put the upper bound (LOD) in place of NA
> fit.mle <- sensorReg(sensor(X,cens.code,type='left') ~ 1,
                      distribution='lognormal')
> fit.mle
```

Add this new line to the plot by:

```
> abline(a=fit.mle$coef,b=fit.mle$scale,lty=3)
```

and add the following legend:

```
> legend(-0.87,2.83,c('Q-Q','OLS','MLE'),lty=1:3)
```

How different are the MLEs from the final estimates in part (a)?

Print out and return the final plot along with the commands used in part (a).

**Solution.** The following code does the job:

```
> ### PART A:
>
> Y <- log(X)
> Q <- qnorm((1:13-0.5)/13)
> plot(Q, Y, xlab='standard normal quantile', ylab='log(X)')
> fit <- lm(Y ~ Q, na.action=na.omit)
> abline(fit,lwd=1)
> coef(fit)
(Intercept)          Q
  1.970102  0.5560098
> Y.hat <- coef(fit)[1] + coef(fit)[2] * Q
> Y.hat[3:13] <- Y[3:13] ## use the data
> new.est <- c('mean'=mean(Y.hat),'sd'=sqrt(var(Y.hat)))
> new.est
      mean      sd
1.970102 0.5585292
> abline(new.est,lty=2)
>
> ### PART B:
>
> cens.code <- ifelse(is.na(X), 0, 1) ## 1=obs, 0=cens
> X[1:2] <- 4.5 ## put the upper bound (LOD) in place of NA
> fit.mle <- censorReg(censor(X,cens.code,type='left') ~ 1,
+                      distribution='lognormal')
> fit.mle
Call:
censorReg(formula = censor(X, cens.code, type = "left") ~ 1, distribution =
"lognormal")

Distribution: Lognormal

Coefficients:
(Intercept)
  1.981478

Dispersion (scale) = 0.5243186
Log-likelihood: -33.91867
```

```
Observations: 13 Total; 2 Censored
Parameters Estimated: 2
> abline(a=fit.mle$coef,b=fit.mle$scale,lty=3)
> legend(-0.87,2.83,c('Q-Q','OLS','MLE'),lty=1:3)
>
```

The MLE method gives slightly smaller mean and standard deviation, but very little is gained by the extra step in the graphical method

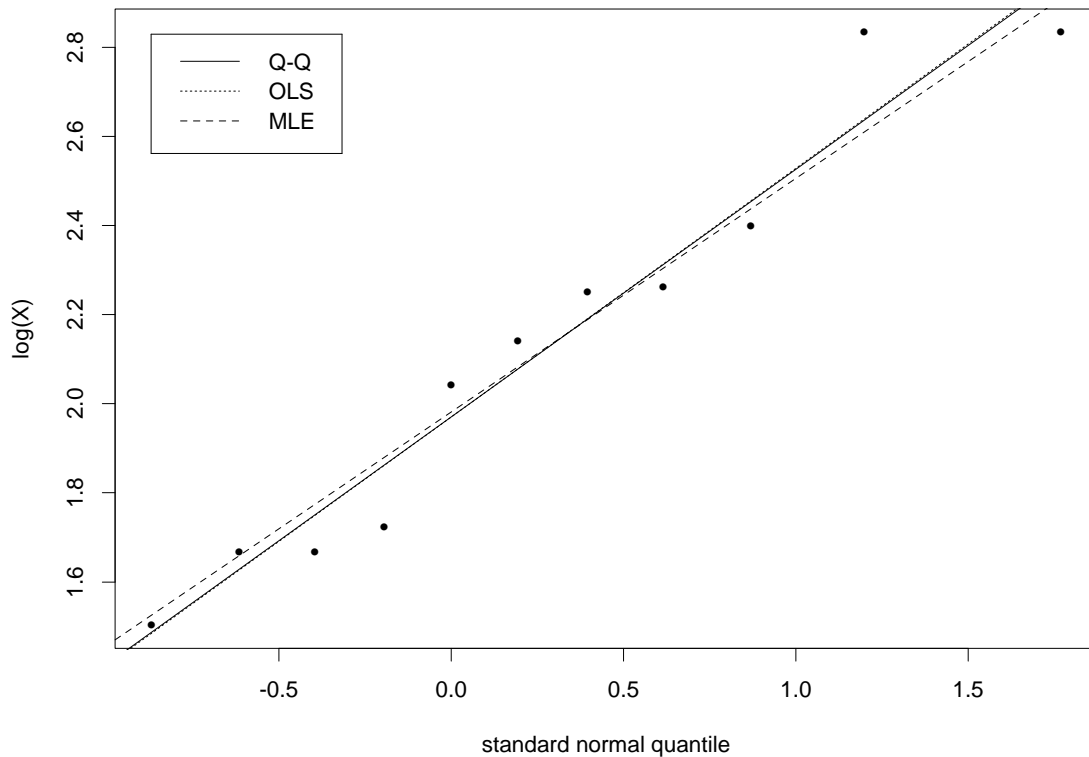


Figure 2: The plot created by the commands in problem 6, showing the three different estimates of  $\mu$  and  $\sigma$ .