

## Homework 4

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

**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  $X$ .
- (b) Plot the density for  $(\mu, \sigma^2) = (0, 1)$ ,  $(\mu, \sigma^2) = (3, 1)$ , and  $(\mu, \sigma^2) = (0, 10)$ .
- (c) Show that the mode of  $f_X(x)$  is at  $x = \exp(\mu - \sigma^2)$ .
- (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  $x_{.75}$ .

**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

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

**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  $i = 1, \dots, n$ ,

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

**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.
- (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.]

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

**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 log-normally 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 lecture to estimate  $\mu$  and  $\sigma$ , as follows:
- i* Compute the quantiles needed from the normal density (see hint below) and plot  $\log(\mathbf{X})$  versus the normal quantiles. Label the x-axis with 'standard normal quantile' and the y-axis with ' $\log(\mathbf{X})$ '.
  - ii* Fit an OLS regression of  $\log(\mathbf{X})$  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 lecture 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, `qnorm((1:13 - 1/2)/13)`. The function `lm` fits OLS regression (see S-PLUS notes).

- (b) *Maximum Likelihood.* The function `ensorReg` 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 <- censorReg(censor(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).