

Name: -----

Computing in Statistics, 22S:166
Midterm 2, Fall 2009

Download the file called "mid2ans.txt" from the Handouts section of the course web page. Copy and paste your R code and output into the appropriate place to answer each exam question. You do NOT have to deal with L^AT_EX or Sweave for this exam. Upload the completed "mid2ans.txt" into ICON to submit your exam.

1. Parametric bootstrap

A dataset called `OECD.txt` is available under Datasets on the course web page. Read it into an R data frame. Also look at the documentation in `OECD.info`.

We will treat the 29 countries in this dataset as a random sample from the population of all countries in the world, even though that is a questionable assumption. We are interested in estimating the median number of doctors per 1000 people in all countries in the world. We will use a parametric bootstrap to get a confidence interval for this median.

- (a) Use R to get a point estimate of the population median, based on the data in this sample. Paste your R code and output into the answer file.

```
> OECD <- read.table("http://www.stat.uiowa.edu/ftp/kcowles/datasets/OECD.txt", header=TRUE)
> attach(OECD)
> median(docs)
[1] 2.8
```

- (b) Assume that the distribution of the `docs` variable in the population of all countries is normal. Use the `boot` package to carry out a parametric bootstrap to obtain an estimate of bias and standard error of your point estimate. Paste your R code for defining any needed functions and running the bootstrap into the answer file, along with the output.

```
ran.gen.normal <- function( d, p ) {
  rnorm( length(d), mean = p$xbar, sd = p$sd)
}

>library(boot)
>boot.out <- boot( data = docs, statistic = median, R = 1999, sim = "parametric",ran.gen=ran.gen.normal)
>boot.out
```

PARAMETRIC BOOTSTRAP

Call:

```
boot(data = docs, statistic = median, R = 1999, sim = "parametric",
      ran.gen = ran.gen.normal, mle = list(xbar = mean(docs), sd = sd(docs)))
```

Bootstrap Statistics :

	original	bias	std. error
t1*	2.8	-0.04815826	0.2183897

- (c) Produce a 90% confidence interval for the population median, using the percentile method and your bootstrap results. Paste your R code and output into the answer file.

```
> ci <- boot.ci( boot.out, type="perc", conf = .90 )
> ci
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1999 bootstrap replicates
```

CALL :

```
boot.ci(boot.out = boot.out, conf = 0.9, type = "perc")
```

Intervals :

```
Level      Percentile
90%      ( 2.396, 3.101 )
```

Calculations and Intervals on Original Scale

2. Simulation study

Recall the following facts:

- A Poisson random variable with parameter λ has both mean and variance equal to λ .
- The maximum likelihood estimator of a population variance based on i.i.d. observations y_1, y_2, \dots, y_n drawn from the population is

$$\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

This estimator is known to be biased.

- The R function called `var` computes the unbiased estimator of variance (that is, it puts $(n - 1)$ in the denominator instead of n).

- (a) Write an R function to compute the maximum likelihood estimator of variance given a numeric vector. Paste the R code into the answer file.

```
> varmle <- function(x) {
n <- length(x)
(n-1) * var(x) / n
}
```

or

```
> varmle2 <- function(x) {
n <- length(x)
sum( (x - mean(x))^2 ) /n
}
```

- (b) Apply your function to a random sample of 10 observations drawn from a Poisson distribution with parameter 4. Paste the R code and output into the answer file.

```
> mypoissons <- rpois( 10, 4 )
> varmle(mypoissons)
[1] 6.6
> varmle2(mypoissons)
[1] 6.6
> var(mypoissons)      # just for comparison; not required
[1] 7.333333
```

- (c) Carry out a simulation study to evaluate the bias of the maximum likelihood estimator of variance when the population distribution is Poisson(4) and the sample size is 10. Test your procedure with a very small number S of replicate samples. Perform your final run with $S = 5000$. Paste your R code, and the output as a single-number estimate of bias, into the answer file.

```
> poismat <- matrix( rpois( S * n, 4 ), nrow = S )
> mlevarout <- apply( poismat, 1, varmle )
> mean(mlevarout) - 4
[1] -0.408306
```

- (d) Write a sentence or two telling whether your results were what you expected.

I would expect the bias to be negative, since dividing by n instead of $(n-1)$ will give a smaller value. Specifically in this example, I would expect the bias to be approximately $-(1/n) * \text{truth} = -0.1 * 4 = -0.4$. Thus the result of my simulation study (estimated bias = -0.408) is

very much what I expected.

3. A law firm secretary wishes to store information about the firm's lawyers and clients in a database. Each lawyer in the firm may have more than one client, and each client may work with more than one lawyer. Below are all the variables that the secretary wishes to store. In addition, the database must make it possible to identify all clients of each lawyer, and all lawyers working with each individual client. In the answer file, list all the tables that would be needed to store this information in 3rd normal form. For each table, list all its fields, and designate any primary and foreign keys.

```
lawyer first name
lawyer last name
lawyer office number
lawyer office phone
lawyer pager number
client first name
client last name
client address
client telephone number
client date of first contact
```

This data has two entities (lawyers and clients), and a many-to-many relationship between them (1 lawyer can serve more than 1 client, and the same client can be served by more than 1 lawyer). Thus we need 3 tables: lawyer table, client table, and linking table, as follows.

Lawyer table

```
lawyer i.d. (primary key)
lawyer first name
lawyer last name
lawyer office number
lawyer office phone
lawyer pager number
```

Client table

```
client i.d. (primary key)
client first name
client last name
client address
client telephone number
```

client date of first contact

Linking table

service i.d. (primary key)

lawyer i.d. (foreign key)

client i.d. (foreign key)

(Note: you could add possible other fields describing the case.

Also you could have used the combination

lawyer i.d. and client i.d.

as the primary key in this table.)