

Biostatistics

Math 322 — Spring 2009

Homework 4 — Suggested solution

Problem 1 (Problems 6.7 – 6.9 in the book)

To read in the data, we can do

```
> lvef <-  
+ read.table("http://www.math.wustl.edu/~hjelle/m322/m322_090213table1.txt",  
+           header = TRUE)
```

1. The standard deviation of LVEF for these patients is

```
> sd(lvef$LVEF)  
[1] 0.07991627
```

2. The standard error of the mean for LVEF is

```
> sd(lvef$LVEF) / sqrt(length(lvef$LVEF))  
[1] 0.01537989
```

3. To draw the 50 subsamples one could use the same methods we used in class on February 2nd (generate.samples and estimate.mean available in the file 090202_functions.R on the web page). Here, let us use a different method based on the replicate function.

```
> samples <- replicate(50, sample(lvef$LVEF, 10))  
> means <- colMeans(samples)  
> means  
[1] 0.214 0.217 0.212 0.221 0.216 0.253 0.215 0.242 0.203 0.235 0.207  
[12] 0.236 0.245 0.203 0.226 0.228 0.205 0.235 0.206 0.241 0.246 0.187  
[23] 0.229 0.226 0.231 0.225 0.204 0.191 0.220 0.213 0.200 0.213 0.279  
[34] 0.208 0.199 0.243 0.199 0.206 0.219 0.191 0.264 0.190 0.197 0.218  
[45] 0.244 0.197 0.191 0.256 0.238 0.188
```

There are different methods for checking the normality of data that you may have seen in earlier statistics courses (including QQ-plots available in R through qqnorm). Here we will just compare a histogram of the means with the normal curve having the same mean and standard deviation as our data.

```
> h <- hist(means, plot = FALSE)  
> ylim <- range(0, h$density, dnorm(mean(means), mean(means), sd(means)))  
> hist(means, freq = FALSE, ylim = ylim, col = gray(.7))  
> curve(dnorm(x, mean(means), sd(means)), add = TRUE)
```

The histogram is included in Figure 1. From the plot we see that the sample means are decently close to normally distributed. In general, sample size 10 is a little small for good approximations using the central limit theorem, and one should be cautious. However, how good the approximation is depends heavily on the distribution of the original data. If the original data are already normally distributed, then also the sample means will be normally distributed.

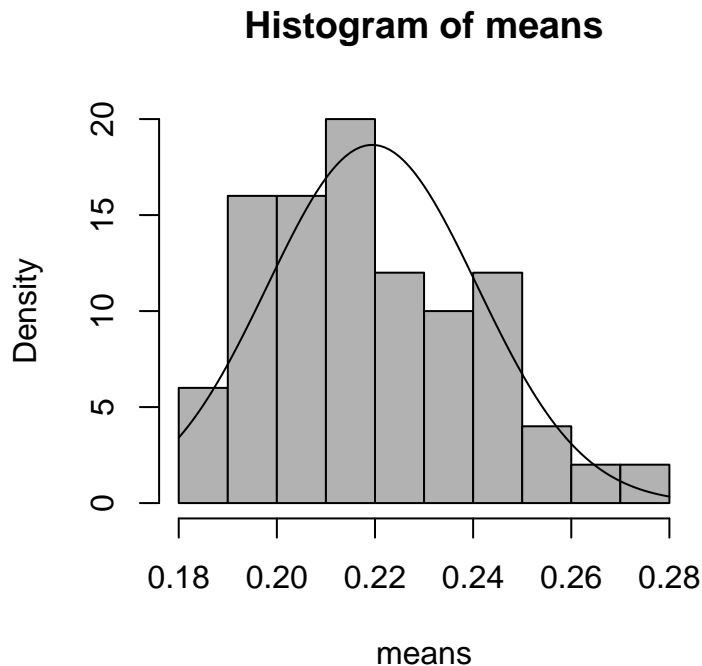


Figure 1: Histogram of the 50 sample means in problem 1c) compared with a normal curve.

Problem 2 (Problems 6.14 – 6.16 in the book)

To calculate the confidence intervals we need to know the sample mean \bar{x} , the sample standard deviation s , and the sample size n . We can calculate these by hand, or we can input the data into R and do the calculations there.

```
> wbc <- scan()
1: 8 5 12 4 11 6 8 7 7 12 7 3 11 14 11 9 6 6 5 6 10 14 4 5 5
26:
Read 25 items
> mean(wbc)
[1] 7.84
> sd(wbc)
[1] 3.2104
```

That is, $\bar{x} = 7.84$, $s = 3.2104$, and $n = 25$.

1. The 95% confidence interval is calculated as

$$\bar{x} \pm t_{n-1, .975} \frac{s}{\sqrt{n}} = 7.84 \pm 2.0639 \frac{3.2104}{\sqrt{25}} = (6.51, 9.17).$$

Using R,

```
> mean(wbc) + qt(c(.025, .975), 24) * sd(wbc) / sqrt(length(wbc))
[1] 6.514812 9.165188
```

or since we have the sample observations, we can use `t.test`,

```
> t.test(wbc)$conf.int
[1] 6.514812 9.165188
attr(,"conf.level")
[1] 0.95
```

2. The 90% confidence interval is calculated similarly,

$$\bar{x} \pm t_{n-1, .95} \frac{s}{\sqrt{n}} = 7.84 \pm 1.7109 \frac{3.2104}{\sqrt{25}} = (6.74, 8.94).$$

With R,

```
> t.test(wbc, conf.level = 0.90)$conf.int
[1] 6.741477 8.938523
attr(,"conf.level")
[1] 0.9
```

3. The 90% confidence interval is shorter than and contained inside the 95% confidence interval.

Problem 3 (Problems 6.31 – 6.32 in the book)

- a) We calculate a 95% confidence interval for the mean DBP for children in the community as

$$\bar{x} \pm t_{n-1, .975} \frac{s}{\sqrt{n}} = 56.2 \pm 2.045 \frac{7.9}{\sqrt{30}} = (53.3, 59.1).$$

Since the confidence interval does not contain the national average, 64.2 mm Hg, we conclude that there is a difference in DBP for the children in the community compared to the national average. Using R,

```
> 56.2 + qt(c(.025, .975), 29) * 7.9 / sqrt(30)
[1] 53.25009 59.14991
```

- b) For a 95% confidence interval for the standard deviation we need to use the χ^2 -distribution. A 95% confidence interval for the variance σ^2 is given by

$$\left(\frac{(n-1)s^2}{\chi_{n-1, .975}^2}, \frac{(n-1)s^2}{\chi_{n-1, .025}^2} \right) = \left(\frac{29 \cdot 7.9^2}{45.72}, \frac{29 \cdot 7.9^2}{16.05} \right) = (39.58, 112.79).$$

Then, a 95% confidence interval for the standard deviation σ is

$$(\sqrt{39.58}, \sqrt{112.79}) = (6.3, 10.6).$$

Using R,

```
> sqrt(29 * 7.9^2 / qchisq(c(.975, .025), 29))
[1] 6.291615 10.620090
```

Problem 4 (Problems 6.38 – 6.40 in the book)

- a)
1. The best point estimate p of the probability of a guinea pig having an enlarged liver, is the proportion $p = \frac{15}{40} = \frac{3}{8} = .375$.
 2. We now assume that the normal approximation to the binomial distribution is valid. That is, we assume that the proportion of guinea pigs with enlarged livers is normally distributed with mean $\mu = .375$ and variance $\sigma^2 = \frac{.375(1-.375)}{40} = .00586$. Then a 95% confidence interval is given by

$$.375 \pm z_{.975} \sqrt{.00586} = (.225, .525).$$

Using R,

```
> .375 + qnorm(c(.025, .975)) * sqrt(.375 * (1-.375) / 40)
[1] 0.2249715 0.5250285
```

- b) To find an exact 95% confidence interval for p we have to find $p_1 < p < p_2$, where p_1 and p_2 are such that

$$Pr(X \geq 15 \mid p = p_1) = .025 \quad \text{and} \quad Pr(X \leq 15 \mid p = p_2) = .025.$$

The easiest way to do this is the `binom.test` function in R,

```
> binom.test(15, 40)$conf.int
[1] 0.2272627 0.5419852
attr(,"conf.level")
[1] 0.95
```

Alternatively, we can use R to continuously “guess” values of p_1 and p_2 , until we arrive at a good approximation. Note that

$$Pr(X \geq 15 \mid p = p_1) = 1 - Pr(X < 15 \mid p = p_1) = 1 - Pr(X \leq 14 \mid p = p_1),$$

so the R-expressions to use are

```
> 1 - pbinom(14, 40, p.1) # Find p.1 so that this equals 0.025
> pbinom(15, 40, p.2) # Find p.2 so that this equals 0.025
```

Carrying this out, we will arrive at the interval $(.227, .542)$.

Problem 5 (Problems 6.58 – 6.62 in the book)

To draw 6 random samples of size 5 from the data in Table 6.2, we first draw 6 samples of size 5 from the integers between 0 and 999, inclusive. We then use these integers as indexes into the data, and enter the 30 values needed.

```
> samp.ind <- replicate(6, sample(0:999, 5))
> samp.ind
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]  310  901   19  991  636  557
[2,]  340  917   22   78   86  275
[3,]  860  667   50  928  556  206
[4,]  132  354  362  992  148  365
[5,]  817  566  688   92  784  201
> samp.data <- cbind(c(122, 110, 78, 79, 106), # Using Table 6.2, page 172
+                   c(170, 98, 121, 134, 72),
+                   c(121, 121, 124, 125, 113),
+                   c(153, 81, 68, 97, 137),
+                   c(100, 109, 100, 89, 138),
+                   c(128, 121, 124, 127, 118))
> samp.data
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]  122  170  121  153  100  128
[2,]  110   98  121   81  109  121
[3,]   78  121  124   68  100  124
[4,]   79  134  125   97   89  127
[5,]  106   72  113  137  138  118
```

Note that since these weights are sampled at random, your numbers will be different.

- a) 1. The mean birthweight for the first sample is

$$\frac{1}{5}(122 + 110 + 78 + 79 + 106) = 99.0.$$

The mean birthweights of the other samples are calculated similarly. They are 119.0, 120.8, 107.2, 107.2, and 123.6, respectively. The mean of all 30 observations is also the mean of the 6 sample means. It is $\frac{1}{6}(99.0 + 119.0 + 120.8 + 107.2 + 107.2 + 123.6) = 112.8$. Using R,

```
> colMeans(samp.data)
[1] 99.0 119.0 120.8 107.2 107.2 123.6
> mean(samp.data)
[1] 112.8
```

2. The variation based on the sample of 6 means is

$$\frac{1}{5}((99.0 - 112.8)^2 + (119.0 - 112.8)^2 + (120.8 - 112.8)^2 + (107.2 - 112.8)^2 + (107.2 - 112.8)^2 + (123.6 - 112.8)^2) = 94.448.$$

The standard deviation is then

$$\sqrt{94.448} = 9.72.$$

This quantity is also called the standard error of the mean (SEM).

```
> sd(colMeans(samp.data))
[1] 9.718436
```

- b) 1. The third weight sampled in each of the 6 samples is 78, 121, 124, 68, 100, 124, respectively. Their mean is 102.5.

```
> samp.data[3,]
[1] 78 121 124 68 100 124
```

The variation of these 6 numbers is

$$\frac{1}{5}((78 - 102.5)^2 + (121 - 102.5)^2 + (124 - 102.5)^2 + (68 - 102.5)^2 + (100 - 102.5)^2 + (124 - 102.5)^2) = 612.7$$

and the standard deviation is

$$\sqrt{612.7} = 24.8.$$

With R,

```
> sd(samp.data[3,])
[1] 24.75278
```

2. If we let σ denote the standard deviation of the birthweights, then the standard error of the mean should be approximately $\frac{\sigma}{\sqrt{5}}$. Looking at the 6 third samples is similar to looking at 6 individual sample points, so their standard deviation should be approximately σ . The ratio between the two standard deviations should therefore be approximately $\sqrt{5}$.
3. Our results are not too far off, although the standard deviation of the third samples is a little bigger than what we would expect from the standard error of the mean.

```
> sqrt(5) * sd(colMeans(samp.data))
[1] 21.73108
```

Problem 6 (Problem 6.70 in the book)

This problem is very open ended, and many different kinds of analysis are possible. The following is just one example.

```
> # Read in the data
> sexrat <- read.csv("sexrat.txt")
> attach(sexrat)
>
> # Some basic statistics
> number.of.children <- sum(Nm_chld * Num_fam)
> number.of.children
[1] 175313
> number.of.boys <- sum((Sx_1 == "M") * Num_fam,
+                       (Sx_2 == "M") * Num_fam,
+                       (Sx_3 == "M") * Num_fam,
+                       (Sx_4 == "M") * Num_fam,
+                       (Sx_5 == "M") * Num_fam)
> number.of.girls <- number.of.children - number.of.boys
> c(number.of.boys, number.of.girls)
[1] 88775 86538
> c(number.of.boys, number.of.girls) / number.of.children
[1] 0.50638 0.49362
> ratio.M <- number.of.boys / number.of.children
```

We will now work under the assumption that the gender of child is not predictable by the gender of its previous born siblings. In particular, we will consider each birth a Bernoulli trial with 50.6% of having a boy, and look at how probable the observed results are.

To assess whether the sex of the second birth is predictable from the sex of the first birth, we can do something like the following.

```
> # Sum up number of families with different combinations
> number.of.M <- sum((Sx_1 == "M") * Num_fam)
> number.of.F <- sum((Sx_1 == "F") * Num_fam)
> number.of.MM <- sum((Sx_2[Sx_1 == "M"] == "M") * Num_fam[Sx_1 == "M"])
> number.of.FM <- sum((Sx_2[Sx_1 == "F"] == "M") * Num_fam[Sx_1 == "F"])
>
> # Check confidence intervals for ratio of second born males
> binom.test(number.of.MM, number.of.M, p = ratio.M)$conf.int
[1] 0.5121837 0.5243338
attr(,"conf.level")
[1] 0.95
> binom.test(number.of.FM, number.of.F, p = ratio.M)$conf.int
[1] 0.4929954 0.5052551
attr(,"conf.level")
[1] 0.95
```

Note that none of the confidence intervals contain the population mean of 50.6% male births, so there seems to be significant evidence that a second birth is more likely to have the same sex as the first birth than should be expected. This significance can also be seen from the p -values, both of which are below 5%.

```
> binom.test(number.of.MM, number.of.M, p = ratio.M)$p.value
[1] 0.0001226621
> binom.test(number.of.FM, number.of.F, p = ratio.M)$p.value
[1] 0.02001633
```

Similar analyses can be carried out also for later births.

Problem 7 (Problem 6.114 in the book)

We assume the number of injuries on artificial turf are Poisson distributed. Our best estimator for the parameter λ (number of injuries per 1000 games) is $45 \cdot \frac{1000}{10112} = 4.45$. To find a confidence interval for λ , we first find a 95% confidence interval for $\mu = \lambda \cdot \frac{10112}{1000}$ (number of injuries per 10112 games).

The simplest way to find this confidence interval is to use the table in Table 8 in the appendix of the book (page 835), with x , the observed number of injuries equal to 45. We can then read of the confidence interval (32.82, 60.21) for μ . We then convert this to a 95% confidence interval for λ by

$$\left(32.82 \cdot \frac{1000}{10112}, 60.21 \cdot \frac{1000}{10112}\right) = (3.25, 5.95).$$

Alternatively, we can use R to “guess” for the limits as in Problem 4. We then want to find 1.1 and 1.2 such that

```
> 1 - ppois(44, 10112 / 1000 * 1.1) # Find 1.1 so that this equals 0.025
> ppois(45, 10112 / 1000 * 1.2) # Find 1.2 so that this equals 0.025
```