

PubH 7401: Elements of Biostatistical Inference I
Homework 7, due November 6

1. In R, type `> boxplot(rnorm(1000))` `[Enter]`. This produces a boxplot of 1000 *iid* $N(0, 1)$ draws. Now hit `[↑]` `[Enter]` several times (e.g. say 10) in a row. This will let you get a feel of how variable a truly normal random sample can be. Now repeat this with sample sizes of 100 and 10. Do not include any plots, but describe what you see, particularly in terms of boxplot symmetry and outlying observations.
2. Repeat the above experiment by plotting histograms instead of boxplots. For sample size 10, you'd type `hist(rnorm(10))` `[Enter]` `[↑]` `[Enter]` `[↑]` `[Enter]`, etc.
3. Let's examine the Law of Large numbers, which states for any *iid* sequence of random variables X_1, X_2, X_3, \dots , that $\bar{X}_n \xrightarrow{P} E(X_i)$. That is, with high probability the sample mean will be close to the population mean when the sample size n is large. We will consider the population density

$$f(x) = \frac{0.6 \exp(-0.5(x - 5)^2)}{\sqrt{2\pi}} + \frac{0.4 \exp(-0.5(x - 10)^2/4)}{\sqrt{8\pi}}.$$

The R code `LLN1.txt` generates 10 \bar{X}_n 's for each of $n = 1, 10, 100, 1000, 10000, 100000$ and plots them versus $i = 1, 2, 3, 4, 5, 6$. The code also plots the above bimodal density. For these data $E(X_i) = 7$ exactly.

- (a) Download `LLN1.txt` and run it from the R command line, e.g.

```
> source("c:/biostat/LLN1.txt").
```

Include the plot with your homework. You can also just copy the code and paste it into R and it will run it. You may have to hit a final `[Enter]` for the last command.

- (b) As n increases from 1 to 10 to 100 to 1000 to 10000 to 100000, what happens to the 10 \bar{x}_n draws in terms of estimating the population mean $E(X_i) = 7$?
- (c) Repeat the experiment in part (a). Now you have a different set of 10 \bar{X}_n 's for each sample size. How do the plots differ? Remember: when collecting actual data, you only get to see *one* of the 60 dots on the plot at one sample size and you don't know $E(X_i)$. By increasing n you ensure that with high probability that \bar{X}_n is close to $E(X_i)$ and therefore a "good" estimator.

4. Now let's consider the population from the previous problem and examine the Central Limit Theorem (CLT). For these data $\text{var}(X_i) = \sigma^2 = 8.20$. The CLT says

$$\bar{X}_n \overset{\circ}{\sim} N(\mu, \sigma^2/n) = N(7, 8.2/n).$$

You will have R generate $m = 10000$ independent \bar{X}_n 's from *iid* samples of sizes $n = 1, 2, 3, 5, 10, 20$ and plot three things on the same graph: (a) the original population density $f(x)$, (b) a nonparametric histogram estimate of the sampling distribution of \bar{X}_n , call it $\hat{f}_{\bar{X}_n}(x)$, based on the $m = 10000$ samples, and (c) the $N(7, 8.2/n)$ density that the CLT says should approximate the true density of the sample mean $\bar{X}_n(x)$. The idea is, that as n gets large, the CLT “kicks in” and the histogram of the $m = 10000$ \bar{X}_n 's will look like the approximating $N(7, 8.2/n)$ density.

Code to perform these steps is posted online, called `CLT1.txt`. You can run the code from the command prompt or just cut and paste it into R. You will primarily be changing `n=2` to `n=1`, `n=5`, etc.

Run the code with $n = 1, n = 2, n = 3, n = 5, n = 10$, and $n = 20$. Describe what you see. Include plots of $n = 2$ and $n = 10$. At what sample size does the CLT “kick in?”

5. Let $X_1, \dots, X_n \overset{iid}{\sim} \text{beta}(\alpha, 1)$. Here $E(X_i) = \alpha/(\alpha + 1)$. The density for each X_i is $f(x|\alpha) = \alpha x^{\alpha-1}$ on the range $R = (0, 1)$.

(a) Show that the MOM is $\hat{\alpha} = \bar{X}_n/(1 - \bar{X}_n)$.

(b) Show that the MLE is $\hat{\alpha} = -n/\sum_{i=1}^n \log X_i$.

(c) For sample sizes of $n = 5, 25, 125, 625$, obtain histograms of $m = 10000$ MOM's obtained from $\text{beta}(5, 1)$ data.

(d) Repeat for MLE's. Code to do parts (c) and (d) are in `MLE2a.txt` and `MLE2b.txt`. Include the plots in your homework.

(e) Is there much difference between the distribution of the MLE and the MOM based on the histograms? What happens as the sample size n increases?

6. Consider the family of $\text{beta}(\alpha, 1)$ densities from Problem 5. For $X_i \sim \text{beta}(\alpha, 1)$, $E(X_i) = \alpha/(\alpha + 1)$ and $\text{var}(X_i) = \alpha/[(\alpha + 1)^2(\alpha + 2)]$.

(a) Let $g(x) = 1/(1 - x)$. Find $g'(x)$.

(b) What is the approximate distribution of the MOM $\hat{\alpha}$? Use the delta method and part (a).

(c) The log pdf is $\log f(X_i|\alpha) = \log(\alpha) + (\alpha - 1) \log X_i$. Find $\frac{d^2}{d\alpha^2} [\log(\alpha) + (\alpha - 1) \log X_i]$. What is the information $I(\alpha)$?

(d) What is the approximate distribution of the MLE $\hat{\alpha}$? Hint: this is very similar to the result obtained for exponential data on pp. 8–11 in `ch8c08.pdf`.

7. Chapter 5, Problem 18 (p. 190).

8. Chapter 8, Problem 7a,b. Show all work. The likelihood for the model $X_1, \dots, X_n \overset{iid}{\sim} \text{geom}(\pi)$ is

$$\mathcal{L}(\pi) = \prod_{i=1}^n p(x_i|\pi) = \prod_{i=1}^n \pi(1 - \pi)^{x_i-1} = \pi^n (1 - \pi)^{n\bar{x}_n - n}.$$