Ma 322: Biostatistics
Homework Assignment 7

Prof. Wickerhauser

Read Chapter 13, "Foundations of Statistical Inference," pages 217–239 of our text.

1. Plot the F densities with every pair of numerator, denominator degrees of freedom chosen from the list 3, 10, 50, over the interval $[0, 4]$. Arrange the graphs into a 3×3 grid. (Hint: modify the code on page 227 of our text.)

   Solution: Following the hint, use the following `R` code:

```
x <- seq(0,4,by=.005); m <- c(3,10,50); n <- c(3,10,50);
par(mfrow=c(3,3));
for (i in 1:3) {
 for (j in 1:3) {
 plot(x,df(x,m[i],n[j]),type='l',ylab="f(x)",cex=.6)
 title(paste(paste("dof =",m[i]),n[j],sep=","))
 }
}
```

   That yields the graph in Figure HW7,Ex1 below. □

2. This problem will illustrate the Central Limit Theorem. Let X be a random variable taking real values $x \in [-1, 0] \cup [1, 2]$ with uniform pdf

$$f(x) = \begin{cases} 1/2, & \text{if } -1 \le x \le 0 \text{ or } 1 \le x \le 2; \\ 0, & \text{otherwise.} \end{cases}$$

   (a) Generate N samples from this pdf using `runif(N)+sample(c(-1,1),N,replace=TRUE)`. Do this with $N = 500$ and plot the histogram to see how little this pdf resembles the bell-shaped curve e^{-x^2} of the normal density.

   (b) What is the exact mean μ of X? (Hint: do not use R or Calculus.)

   (c) What is the exact variance σ^2 of X? (Hint: use Calculus.)

   (d) Fix $n = 3$ and $m = 200$. Generate m vectors $\{X_i : i = 1, \ldots, m\}$ of n random samples $X_i(1), \ldots, X_i(n)$ of X and form m normalized averages

$$\bar{X}_i \stackrel{\text{def}}{=} \frac{S_i - n\mu}{\sigma\sqrt{n}}, \qquad i = 1, \ldots, m,$$

   where $S_i = \sum_{k=1}^{n} X_i(k)$, and μ and σ are from parts b and c. Plot the histogram of $\bar{X}_i$ and the quantile-quantile plot `qqnorm()` against the normal pdf.

   (e) Repeat part d with $n = 50$ and $m = 200$.

   Solution: (a) Use the given `R` code:

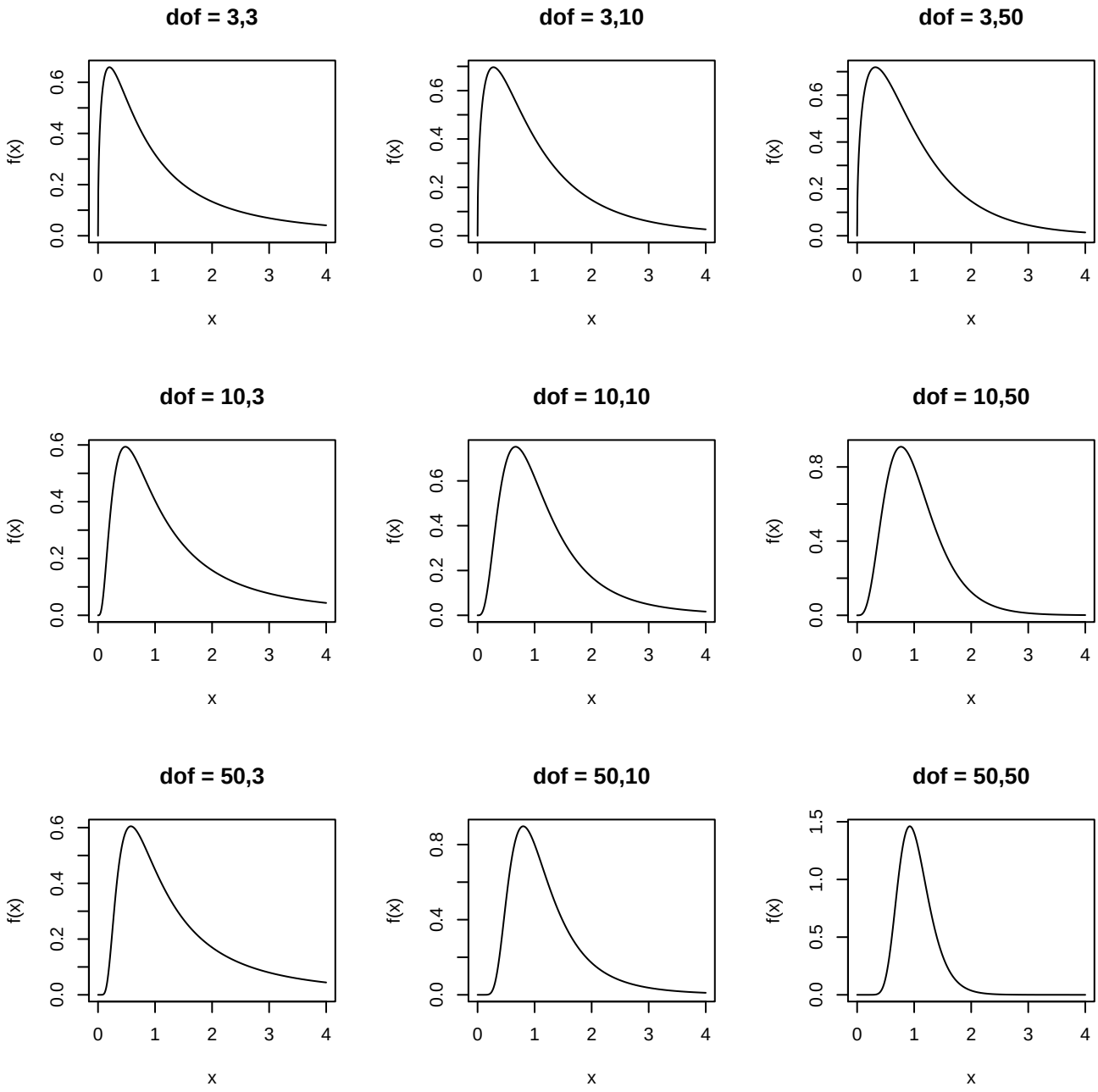

Figure 1: HW7,Ex1: Plots of F densities with various degrees of freedom.

```
N <- 500; X <- runif(N)+sample(c(-1,1),N,replace=TRUE); hist(X);
```

The results are shown in Figure HW7,Ex2a.

(b) Compute the mean μ by

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-1}^{2} x f(x)\,dx = 1/2,$$

which is evident since f is symmetric about $1/2$ and vanishes outside $[-1, 2]$. Alternatively, use Macsyma:

```
f:1/2;   mu: integrate(x*f,x,-1,0) + integrate(x*f,x,1,2);
```

(c) Compute the variance σ^2 by

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{-1}^{0} (x-\frac{1}{2})^2(\frac{1}{2})\,dx + \int_{1}^{2} (x-\frac{1}{2})^2(\frac{1}{2})\,dx = \frac{13}{12},$$

by elementary methods or with Macsyma:

```
f:1/2; mu: integrate(x*f,x,-1,0) + integrate(x*f,x,1,2);
sigma2: integrate((x-mu)*(x-mu)*f,x,-1,0) + integrate((x-mu)*(x-mu)*f,x,1,2);
```

This shows that the pdf has finite variance.

(d,e) Use the following `R` code:

```
xbars <- function(n,m) {
  mu <- 1/2; sigma <- sqrt(13/12); xbar <- vector(mode="numeric",length=m);
  for (i in 1:m ) {
    X <- runif(n)+sample(c(-1,1),n,replace=TRUE);
    xbar[i] <-(sum(X)-n*mu)/(sigma*sqrt(n)); }
  xbars <- xbar;
}
par(mfrow=c(2,2));
xbar <- xbars(n=3, m=200);  hist(xbar);  qqnorm(xbar);
xbar <- xbars(n=50, m=200);  hist(xbar);  qqnorm(xbar);
```

The results are shown in Figure HW7,Ex2de. □

3. Alleles A and a are present in a population in unknown proportions p and $1-p$. Assuming a Hardy-Weinberg equilibrium distribution of the resulting diploid genotypes, find the maximum likelihood estimator for p given the following experimental results:

Genotype Count Data for One Allele

Genotype	Count Data	Variable
AA	314	n_{AA}
Aa	531	n_{Aa}
aa	289	n_{aa}

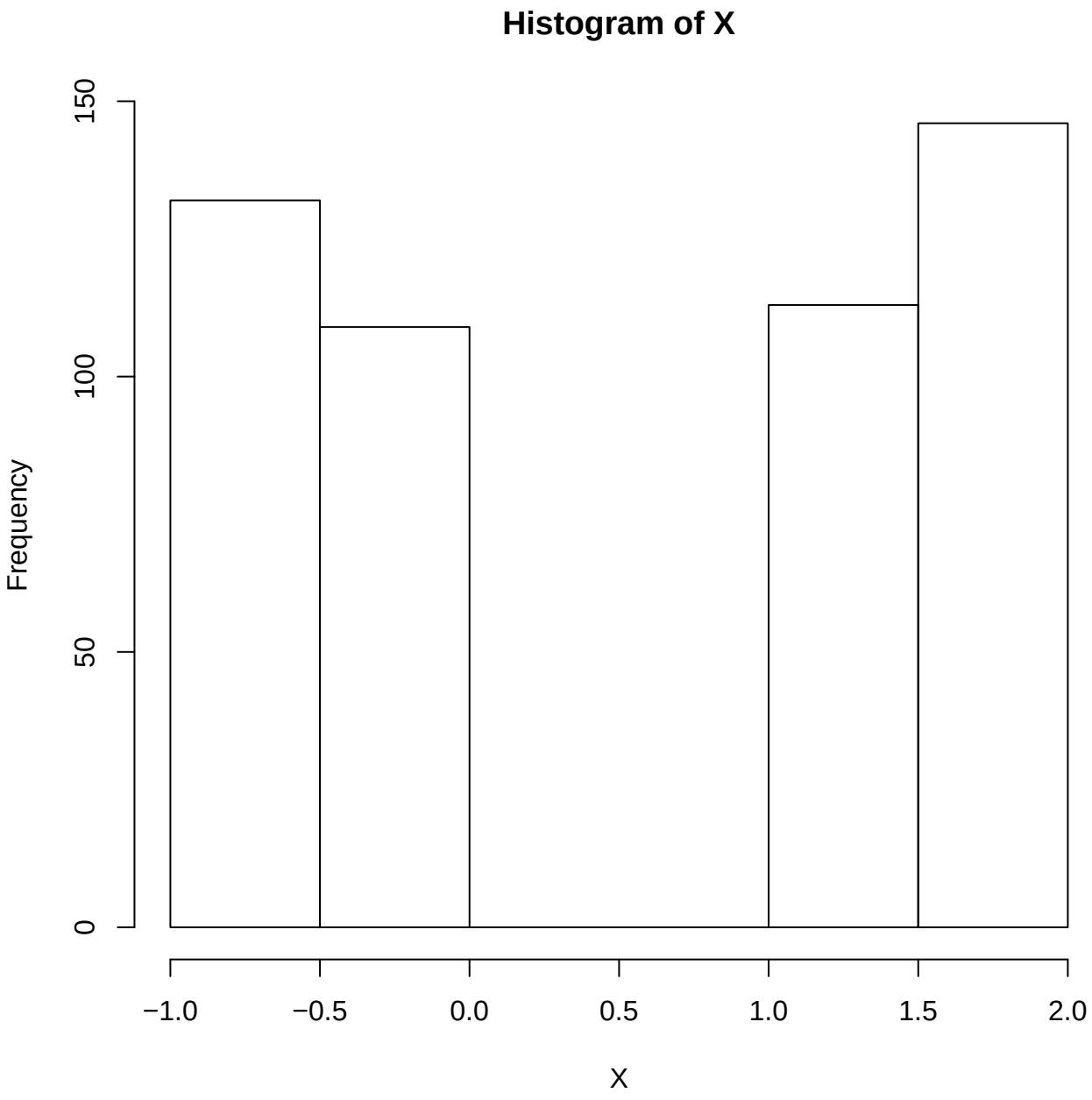

Figure 2: HW7,Ex2a: Histogram of samples from a certain non-normal PDF.

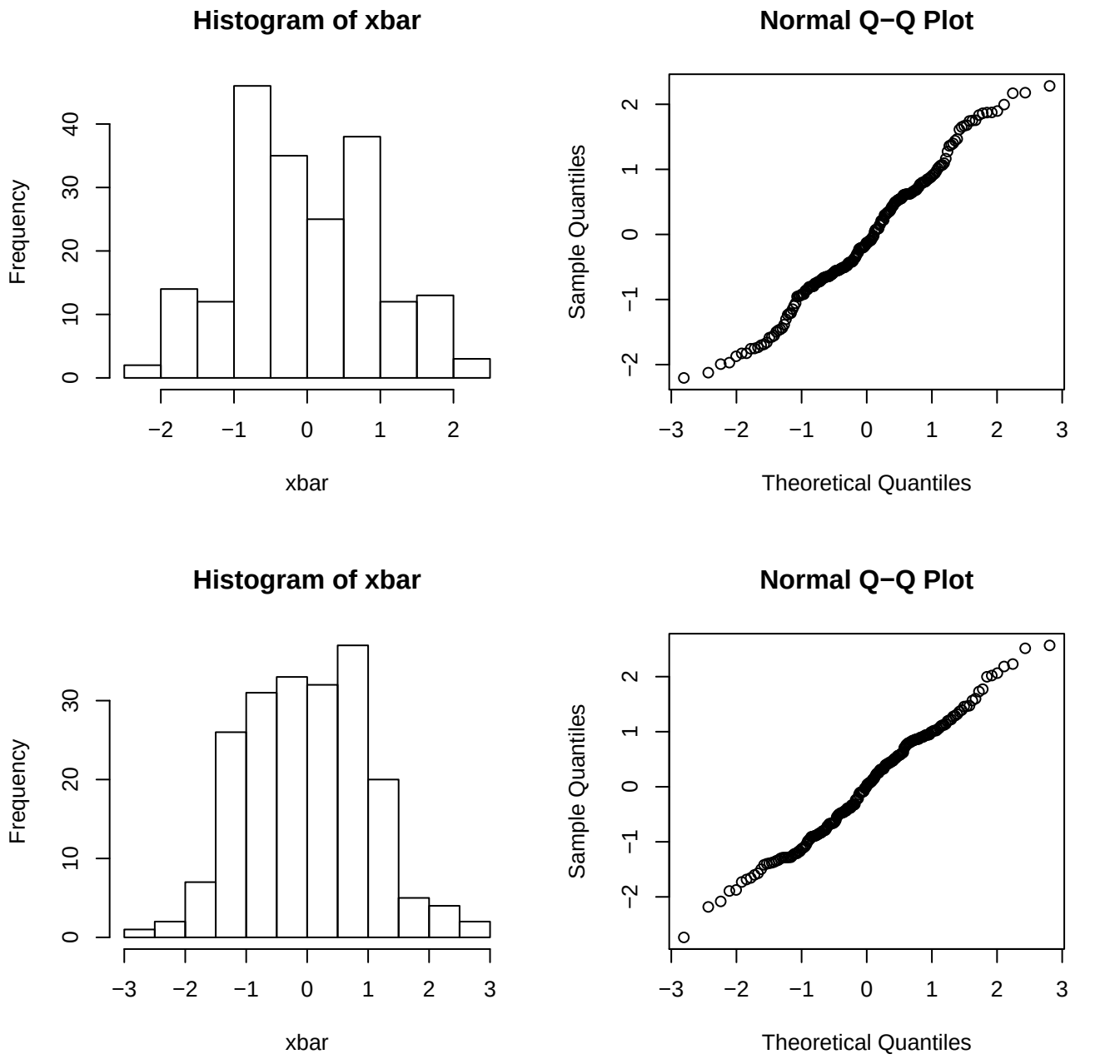

Figure 3: HW7,Ex2de: the 3-means and 50-means from a non-normal PDF.

Solution: As on page 231 of our text, the likelihood function satisfies

$$L_n(p) \propto (p)^{2n_{AA}}(2p(1-p))^{n_{Aa}}(1-p)^{2n_{aa}} = (p)^{2\times 314}(2p(1-p))^{531}(1-p)^{2\times 289},$$

where $n = (n_{AA}, n_{Aa}, n_{aa}) = (314, 531, 289)$ is the count data from the experiment.

Setting $g(p) \stackrel{\text{def}}{=} (\log L_n(p))' = 0$ and solving for p gives the desired estimator:

$$0 = g(p) = \frac{2n_{AA}}{p} + \frac{n_{Aa}}{p} - \frac{n_{Aa}}{1-p} - \frac{2n_{aa}}{1-p},$$

as derived on page 231 of our text. Solving for p gives

$$p = \frac{n_{Aa} + 2n_{AA}}{2n_{Aa} + 2n_{AA} + 2n_{aa}} = \frac{531 + 2\times 314}{2\times 531 + 2\times 314 + 2\times 289} = \frac{1159}{2268} \approx 0.511023$$

This may be calculated using the Macsyma commands

```
g: 2*nAA/p + nAa/p - nAa/(1-p) - 2*naa/(1-p);
solve(g=0,p);
subst([nAA=314,nAa=531,naa=289],%);
```

Alternatively, we may plot the log-likelihood function to determine its maximum within some reasonably fine grid. This may be done with the `R` commands

```
pdf("hw7ex4.pdf");  par(mfrow=c(2,3));  # send nice output to a PDF file
G<-function(p,nAA=314,nAa=531,naa=289) {
2*nAA*log(p) + nAa*log(p*(1-p)) + 2*naa*log(1-p);}
p<-seq(0.4,0.6,by=0.001);   plot(p,G(p));
p<-seq(0.49,0.53,by=0.0001);   plot(p,G(p));
p<-seq(0.510,0.512,by=0.00001);   plot(p,G(p));
```

We may also find the zero of the derivative of the log-likelihood function graphically:

```
g<-function(p,nAA=314,nAa=531,naa=289) {
  2*nAA/p + nAa/p - nAa/(1-p) - 2*naa/(1-p);}
p<-seq(0.4,0.6,by=0.001);   plot(p,g(p)); abline(0,0);
p<-seq(0.49,0.53,by=0.0001);   plot(p,g(p)); abline(0,0);
p<-seq(0.510,0.512,by=0.00001);   plot(p,g(p)); abline(0,0);
dev.off();  # close the PDF file with 2 rows of 3 graphs
```

These six plots are displayed in Figure HW7,Ex4. □

4. Following are some samples from a population with unknown (but finite) mean μ and standard deviation σ:

```
6.92 11.9 8.94 3.18 10.3 9.90 9.22 5.61 6.73 6.66 9.86 5.50 8.53 5.46 4.95
```

(a) Compute an estimate for σ.

(b) Compute an estimate for μ.

(c) Find the median of the samples.

(d) Find the quartile deviation of the samples.

Solution: Use `x<-scan()` with cut and paste to load the data.

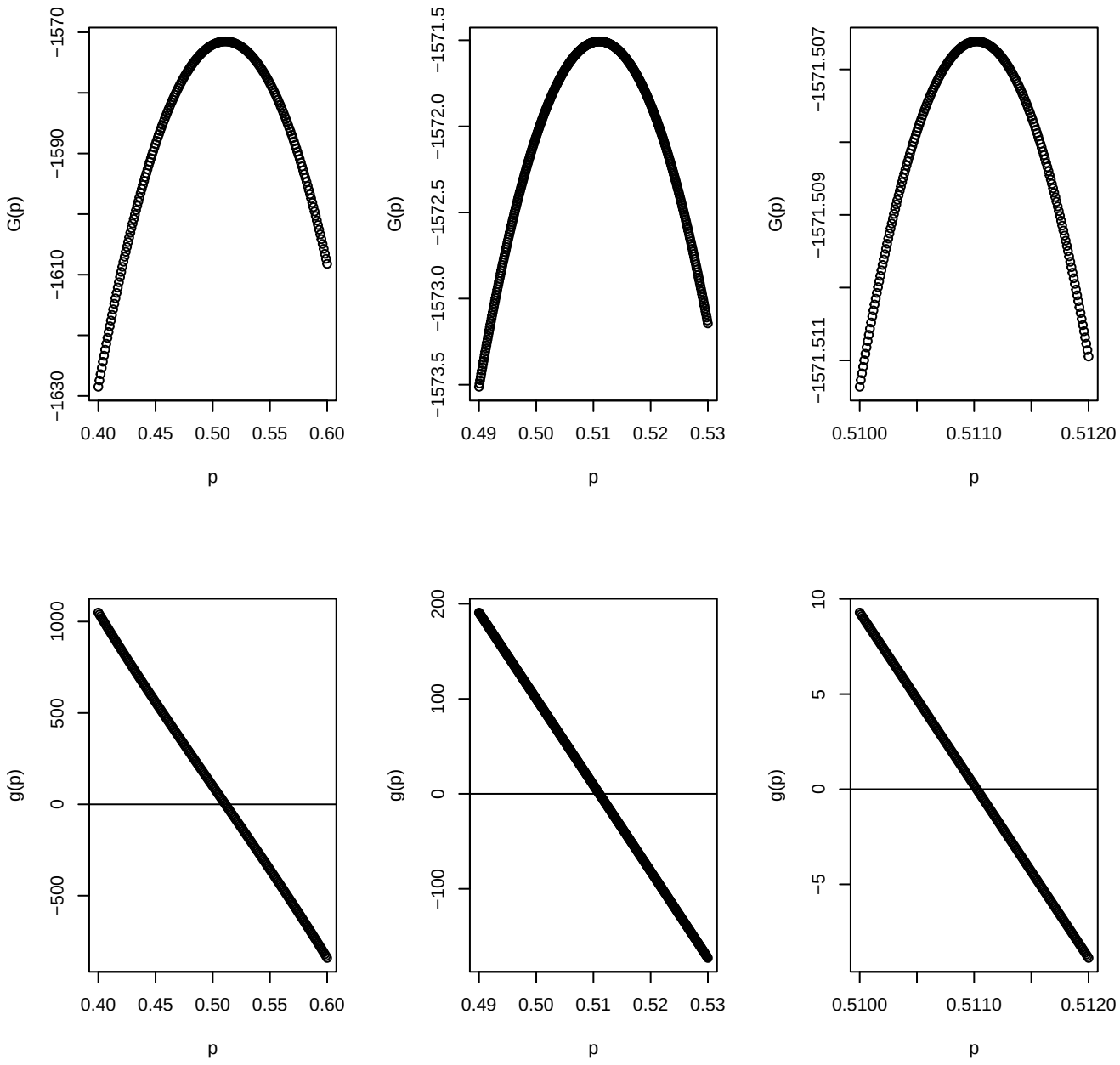

Figure 4: HW7,Ex4: Plots of the log-likelihood function $G(p)$ and its derivative $g(p) = G'(p)$ near the maximum-likelihood $p = 0.511023$.

```
x<-scan()
```

That yields the following values:

(a) `sd(x) = 2.435838`.

(b) `mean(x) = 7.577333`.

(c) `median(x)` gives 6.92; `summary(x)[3]` also gives 6.92.

(d) `(summary(x)[5]-summary(x)[2])/2` gives 1.9925. □

5. This problem will illustrate nonparametric bootstrap estimation of sample variability. First, let `MYSID` be your student ID number and generate a 200 sample data set as follows:

```
set.seed(MYSID); data<- c(rnorm(90,mean=3,sd=2), rexp(110,rate=1));
```

(a) Plot the histogram of `data`.

(b) Find the mean and standard deviation of `data`.

(b') Estimate the "standard error" of a 200-sample mean by $s/\sqrt{200}$ using the standard deviation from part b.

(c) Find the median and the 1st and 3rd quartile values of `data`.

Now apply the bootstrap method: generate 100 replications of 200 samples of `data`, with replacement, and calculate their means and medians.

(d) Calculate the mean and standard deviation of the 100 bootstrap means.

(d') Which is bigger, the bootstrap standard deviation of the means, or the "standard error" from part b'?

(e) Calculate the median and the 1st and 3rd quartile values of the 100 bootstrap medians.

(e') Compute the ratio of the differences between the 3rd and 1st quartiles for the bootstrap medians and the original data.

Solution: Here is the output from an experiment with `MYSID=12345`:

```
MYSID <- 12345;
set.seed(MYSID); data<- c(rnorm(90,mean=3,sd=2), rexp(110,rate=1));
hist(data);  mean(data); sd(data); sd(data)/sqrt(200);
summary(data); xstar<-matrix(0,100,200);
for(i in 1:100) xstar[i,]<-sample(data,200,replace=TRUE);
xmedians<-rep(0,100); for(i in 1:100)xmedians[i]<-median(xstar[i,]);
xmeans<-rowSums(xstar)/200;   sd(xmeans);  summary(xmedians);
```

That produces the following output:

(a) See Figure HW7,Ex5a below.

(b) mean = 2.109065, sd = 2.077493.

(b') sd/sqrt(200) = 0.1469010.

(c) 1st Quartile q1 = 0.3853; Median = 1.5940; 3rd Quartile q3 = 3.6940.

(d) Bootstrap means: mean = 2.097153; sd = 0.1482708.

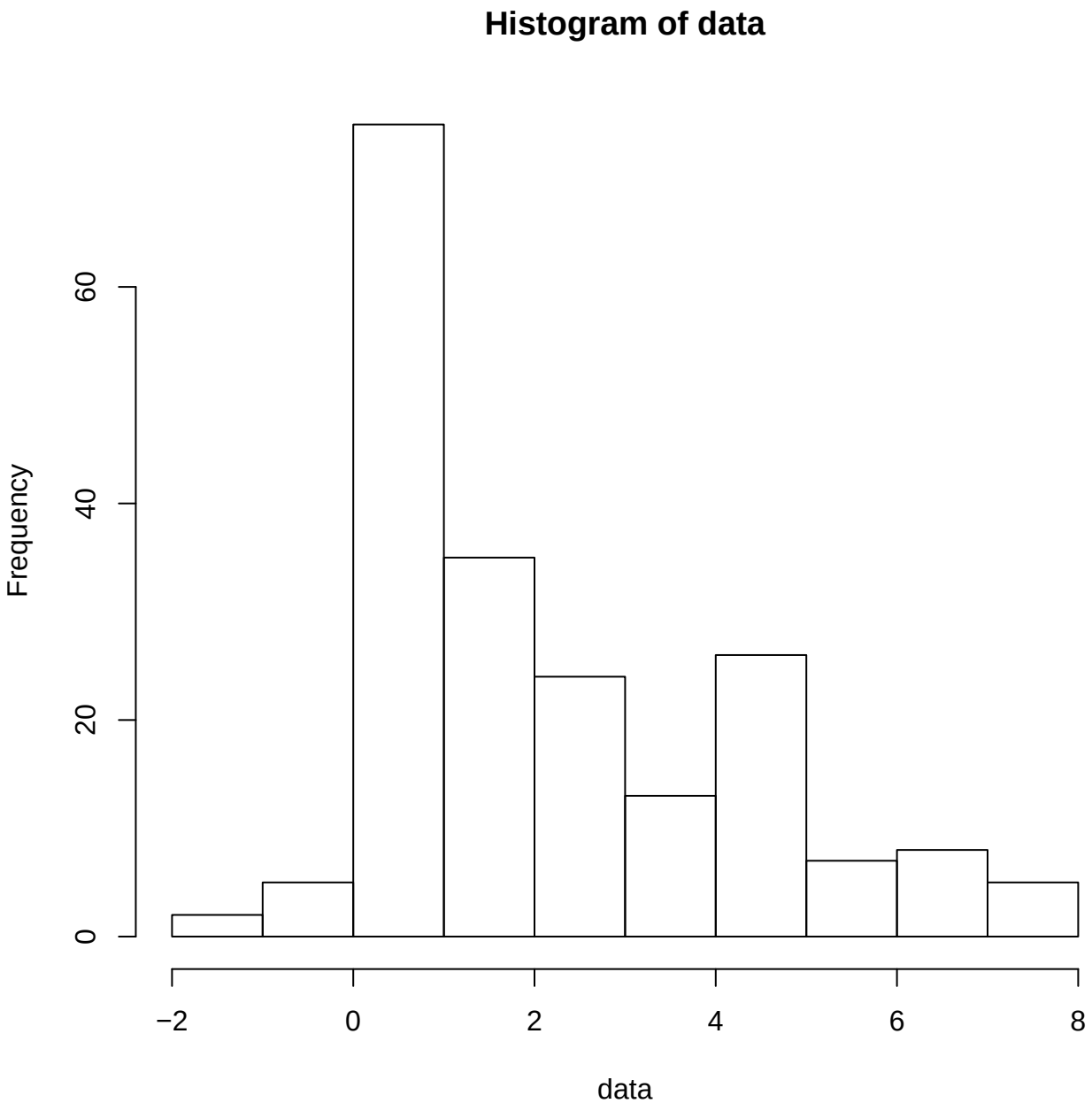

Figure 5: HW7,Ex5a: Histogram of another non-normal PDF.

(d') The standard deviation of the bootstrap means is slightly bigger than the "standard error" computed in part b'.

(e) Bootstrap medians: median = 1.5970; 1st quartile qb1 = 1.3500; 3rd quartile qb3 = 1.6570.

(e') The ratio of interquartile differences

$$\frac{q_3 - q_1}{qb_3 - qb_1} = (3.6940 - 0.3853)/(1.6570 - 1.3500) = 10.77752$$

shows an improvement similar to the $\sqrt{100}$ reduction that we would expect by taking the mean of 100 samples. □