

Ma 322: Biostatistics

Solutions to Homework Assignment 2

Prof. Wickerhauser

Due Friday, February 3rd, 2012

Read Chapter 7, pages 80–107, of our e-text to review some basic probability density functions and their properties, concentrating especially on the normal pdf. Consult Chapters 1-5 as needed to find function names and syntax to solve the computation problems below.

- a. Using Venn diagrams, depict five subsets A, B, C, D, E satisfying $A \subset B \subset C$, $B \cap D \neq \emptyset$, $A \cap D = \emptyset$, $C \cap E = \emptyset$, and $D \cap E \neq \emptyset$.
b. Is $C \cap D = \emptyset$?

Solution: a. See the figure below.

b. No, the conditions imply that $C \cap D \neq \emptyset$, since $B \subset C$ and $B \cap D \neq \emptyset$. \square

- A standard set of 52 playing cards is divided into 4 suits of 13 ranks each: ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, jack, queen, and king. The suits are called clubs, diamonds, hearts and spades, with clubs and spades being “black” and hearts and diamonds being “red.”
 - Taking 1 card at random, what is the probability of drawing a king of spades? a black king? a face card (jack, queen, or king)?
 - Taking 2 cards at random without replacement, what is the probability of drawing a pair of kings? a pair of clubs? a pair of black cards? a pair of cards of different ranks and suits?
 - Taking 5 cards at random without replacement, what is the probability of drawing a “full house,” namely 3 cards of one rank and 2 cards of a second rank?

Solution: (a) With 1 card:

$$P(\text{king of spades}) = 1/52 \approx 0.0192;$$

$$P(\text{black king}) = 2/52 \approx 0.0385;$$

$$P(\text{face card}) = 12/52 \approx 0.231.$$

(b) With 2 cards:

$$P(\text{pair of kings}) = \binom{4}{2} / \binom{52}{2} = \frac{4}{52} \times \frac{3}{51} \approx 0.00452, \text{ or } \text{choose}(4, 2) / \text{choose}(52, 2) \text{ in R.}$$

$$P(\text{pair of clubs}) = \binom{13}{2} / \binom{52}{2} = \frac{13}{52} \times \frac{12}{51} \approx 0.0588, \text{ or } \text{choose}(13, 2) / \text{choose}(52, 2) \text{ in R.}$$

$$P(\text{pair of black cards}) = \binom{26}{2} / \binom{52}{2} = \frac{26}{52} \times \frac{25}{51} \approx 0.245, \text{ or } \text{choose}(26, 2) / \text{choose}(52, 2) \text{ in R.}$$

$$P(\text{pair of cards of different ranks and suits}) = \frac{52}{52} \times \frac{36}{51} \approx 0.706.$$

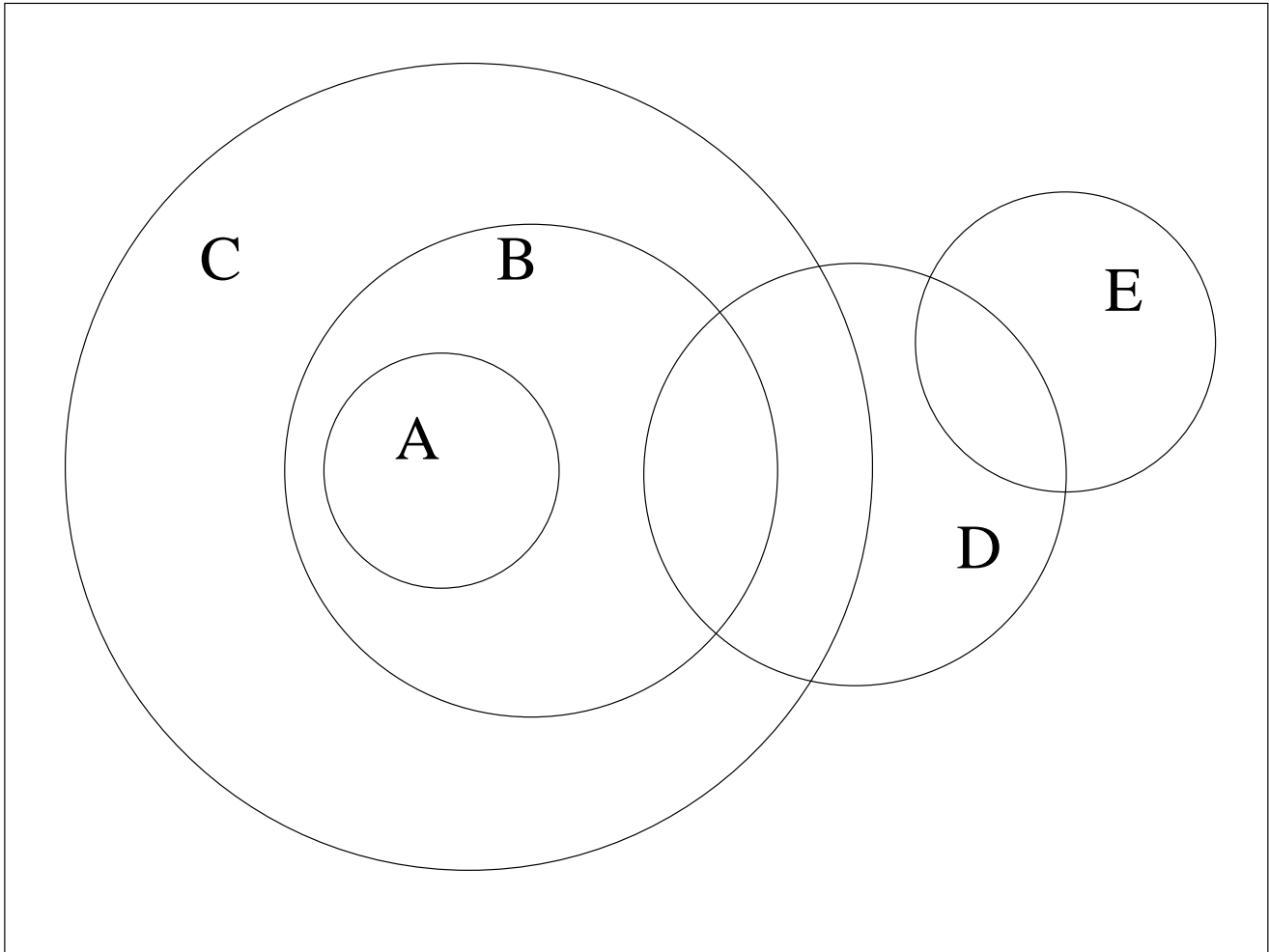


Figure 1: Venn Diagram

(c) With 5 cards:

$$P(\text{full house}) = \frac{\binom{13}{1} \binom{4}{3} \binom{12}{1} \binom{4}{2}}{\binom{52}{5}} = \frac{(13 \times 12)(4)(6)(5!)}{52 \times 51 \times 50 \times 49 \times 48} \approx 0.00144$$

The choice sequence is first rank (of 13), three suits at that rank (of 4), second rank (of the remaining 12), two suits at that second rank (of 4). Divide by the number of 5-card hands (of 52 cards). Compute the value with the R commands:

```
choose(13,2)*choose(4,3)*choose(12,1)*choose(4,2)/choose(52,5).
```

□

3. Suppose that x is a normally distributed random variable with mean 21.8 and standard deviation 1.34.

(a) What is the probability that x is greater than 20? greater than 23? less than 20?

(b) What is the probability that x lies between 20.8 and 22.8? between 22.8 and 24.8? between 24.8 and 26.8?

(c) Use R commands to generate 10 independent random values of x . Compute the mean and standard deviation of those samples.

(d) What is the theoretical standard deviation of the mean of 10 independent random samples of x ? What is the probability that a random set of 10 independent samples has a mean greater than 22? What is the probability that a random set of 1000 samples has a mean greater than 22?

Solution: (a) The R commands to compute this are:

```
> pnorm(20, mean=21.8, sd=1.34, lower.tail=FALSE)
0.91041
> pnorm(23, mean=21.8, sd=1.34, lower.tail=FALSE)
0.1852540
> pnorm(20, mean=21.8, sd=1.34, lower.tail=TRUE)
0.08959008
```

(b) Use the `diff()` function and exploit R's behavior on vectors:

```
> diff(pnorm(c(20.8,22.8), mean=21.8, sd=1.34))
0.5444949
> diff(pnorm(c(22.8,24.8), mean=21.8, sd=1.34))
0.2151683
> diff(pnorm(c(24.8,26.8), mean=21.8, sd=1.34))
0.01248904
```

Note that the result is the same for `lower.tail=TRUE` (the default) as it is for `lower.tail=FALSE`.

(c)

```
> x<-rnorm(10, mean=21.8, sd=1.34); x
25.65681 21.45387 20.51436 20.22023 21.92318 20.69021 20.87604 21.47374
22.67400 22.34372
> mean(x)
21.78262
> sd(x)
1.577861
```

Your answer may vary depending on the random number seed when you called `rnorm()`.

(d) In theory, the standard deviation of the mean of n samples from this population is $1.34/\sqrt{n}$. Compute the probabilities using this for `sd` in `pnorm()`:

```
> msd <- 1.34/sqrt(10); msd
0.4237452
> pnorm(22, mean=21.8, sd=msd, lower.tail=FALSE)
0.3184699
> msd <- 1.34/sqrt(1000); msd
0.04237452
> pnorm(22, mean=21.8, sd=msd, lower.tail=FALSE)
1.180282e-06
```

Evidently the 1000-sample mean is likely to be very close to the population mean. □

4. Present the following data on a graph that shows the mean, 95% confidence interval, and range for each month:

Table of caloric intakes (kcal/g of body weight) of squirrels.

Month	No. of data	Mean	Std. error	Range
January	14	0.662	0.040	0.451–0.903
February	13	0.544	0.039	0.385–0.793
March	17	0.487	0.028	0.412–0.589

Solution: First, compute the 95% confidence intervals $\bar{X} \pm \delta$ using the given standard error $\sigma_{\bar{X}}$ instead of $s_{\bar{X}}$ in the formula $\delta = t_{0.05(2),\nu} s_{\bar{X}}$ with $\nu = n - 1$:

95% confidence intervals.					
Month	n	ν	$\sigma_{\bar{X}}$	$t_{0.05(2),\nu}$	δ
January	14	13	0.040	2.160	0.0864
February	13	12	0.039	2.179	0.0850
March	17	16	0.028	2.120	0.0594

Get δ using the t -distribution quantile function in R:

```
quantiles <- c(0.025, 0.975) # 95% lies between 2.5% and 97.5% marks
xbar<-0.662 # January mean value
sxbar<- 0.040 # January standard error of the mean
xbar+ sxbar*qt(quantiles,13) # January 95% confidence interval
```

Then, use R to produce a box plot depicting the mean within the 95% confidence interval within the range:

```
quantiles <- c(0.025, 0.5, 0.975) # middle 0.5 marks the mean
xbar<-0.662; sxbar<-0.040; minv<-0.451; maxv<-0.903; xdf <- 13 # January
jan<-c(minv, xbar+sxbar*qt(quantiles,xd), maxv)
xbar<-0.544; sxbar<-0.039; minv<-0.385; maxv<-0.793; xdf <- 12 # February
feb<-c(minv, xbar+sxbar*qt(quantiles,xd), maxv)
xbar<-0.487; sxbar<-0.028; minv<-0.412; maxv<-0.589; xdf <- 17 # March
mar<-c(minv, xbar+sxbar*qt(quantiles,xd), maxv)
pdf(file="boxplot.pdf") # redirect graphics output to a PDF file
boxplot(list(jan,feb,mar),range=0)
dev.off() # close the PDF file
```

The resulting plot is shown in the figure below. □

5. A random sample of size 19 from a normal population has mean 17.92 cm and variance 8.2261 cm².
- Calculate the 95% confidence interval of the population mean μ .
 - How large a sample would be needed to estimate μ to within 1.00 cm with 95% confidence?
 - How large a sample would be needed to estimate μ to within 2.00 cm with 95% confidence?
 - How large a sample would be needed to estimate μ to within 2.00 cm with 99% confidence?

Solution: (a) First note that $s_{\bar{X}} = \sqrt{s^2/n} = \sqrt{8.2261/19} = 0.6580$ for the $n = 19$ samples with the given variance. Using $\alpha(2) = 0.05$ and $\nu = n - 1 = 18$, find $t_{\alpha(2),\nu}$ with the `qt()` function, then calculate $\delta = t_{\alpha(2),\nu} s_{\bar{X}}$ in R as follows:

```
> ss<-8.2261; n<-19; xm<-17.92; nu<-n-1; sx<-sqrt(ss/n); sx
0.6579914
> alpha<-0.05; t<-qt(alpha/2, df=nu, lower.tail=FALSE); t
2.100922
```

```
> delta <- t*sx; delta
1.382389
> xm+c( -delta, delta )
16.53761 19.30239
```

This last is the 95% (or $1 - \alpha$) confidence interval 17.92 ± 1.38 cm.

(b) Set $\delta = 1.00$, $\alpha = 0.05$, and solve for n in the equation

$$n = \frac{s^2 t_{\alpha(2),n-1}^2}{\delta^2} \quad (1)$$

The R commands are:

```
power.t.test(delta=1, sd=sqrt(8.2261), sig.level=0.05, power=0.5, type="one.sample")
```

The resulting output indicates $n = 33.5554$. The number of samples to use is the least integer greater than or equal to this, or $n = 34$.

NOTE: not including argument n , or equivalently including $n=NULL$ in `power.t.test()`, means “compute n ” when the four other arguments `delta`, `sig.level`, `sd`, `power` are provided. This n solves

$$n = \frac{2s^2}{\delta^2} \left(t_{\alpha(2),n-1}^2 + t_{\beta(1),n-1}^2 \right).$$

Putting `power=0.5` makes $\beta = 0.5$, so that $t_{\beta(1),n-1}^2 = 0$ for all n . Then putting `sd=sqrt(s^2/2)` makes this function solve Equation 1.

(c) Set $\delta = 2.00$, $\alpha = 0.05$, and solve for n in Equation 1 in part (a) above, using the following R commands:

```
power.t.test(delta=2, sd=sqrt(8.2261), sig.level=0.05, power=0.5,type="one.sample")
```

The resulting output indicates $n = 9.93135$.

(c) Set $\delta = 2.00$, $\alpha = 0.01$, and solve for n in Equation 1 in part (a) above with the following R commands:

```
power.t.test(delta=2, sd=sqrt(8.2261), sig.level=0.01, power=0.5,type="one.sample")
```

The resulting output indicates $n = 16.9727$. □

6. Generate 1000 samples from each of the following probability spaces (populations) and plot the resulting histogram.

(a) X is standard normal; bins are of the form $\frac{1}{5}[k, k + 1)$ where k is an integer.

(b) X is χ^2 on 10 degrees of freedom; bins are of the form $[k, k + 1)$ where k is an integer.

Solution: (a)

```
> x<-rnorm(1000); range(x)
-3.115844 3.725171
> bins<-seq(-4,4,by=0.2); hist(x,breaks=bins)
> pdf("hista.pdf"); hist(x,breaks=bins); dev.off()
```

(b)

```
> y<-rchisq(1000,df=10); range(y)
1.444421 30.110476
> bins<-seq(0,31,by=1); hist(y, breaks=bins)
> pdf("histb.pdf"); hist(y, breaks=bins); dev.off()
```

Your results will vary depending upon the value of the random number seed.

□

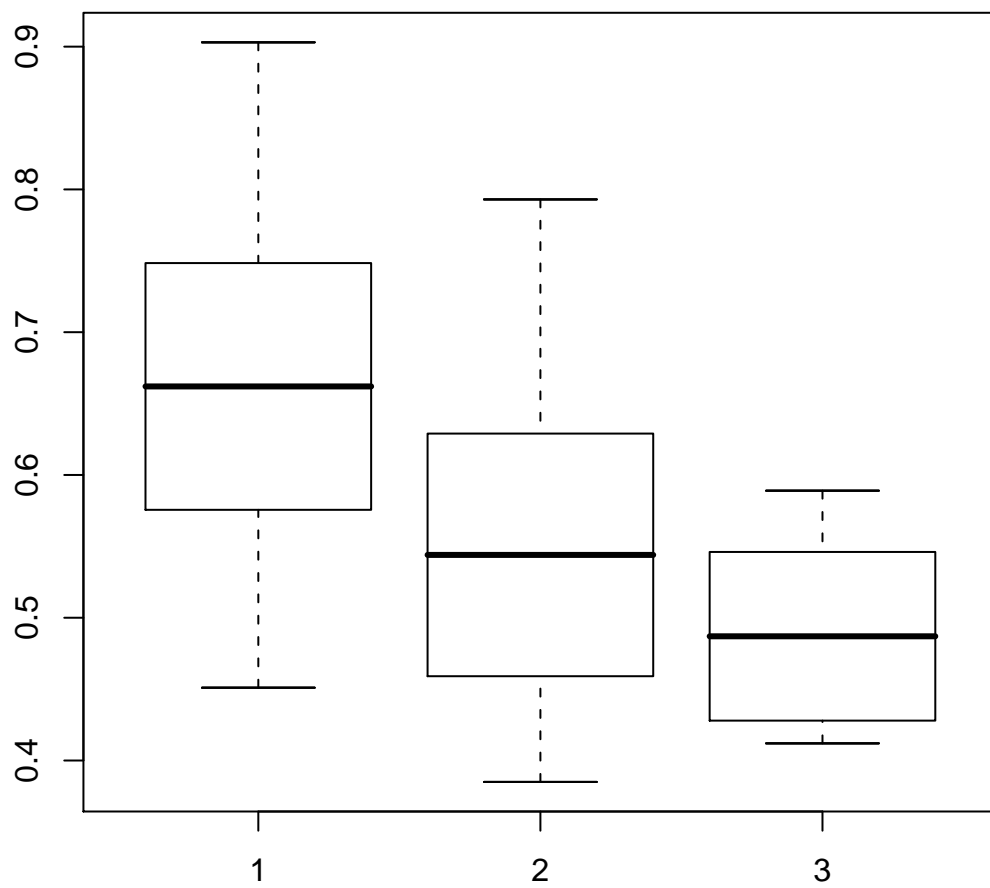


Figure 2: Box Plots for Caloric Intakes Data

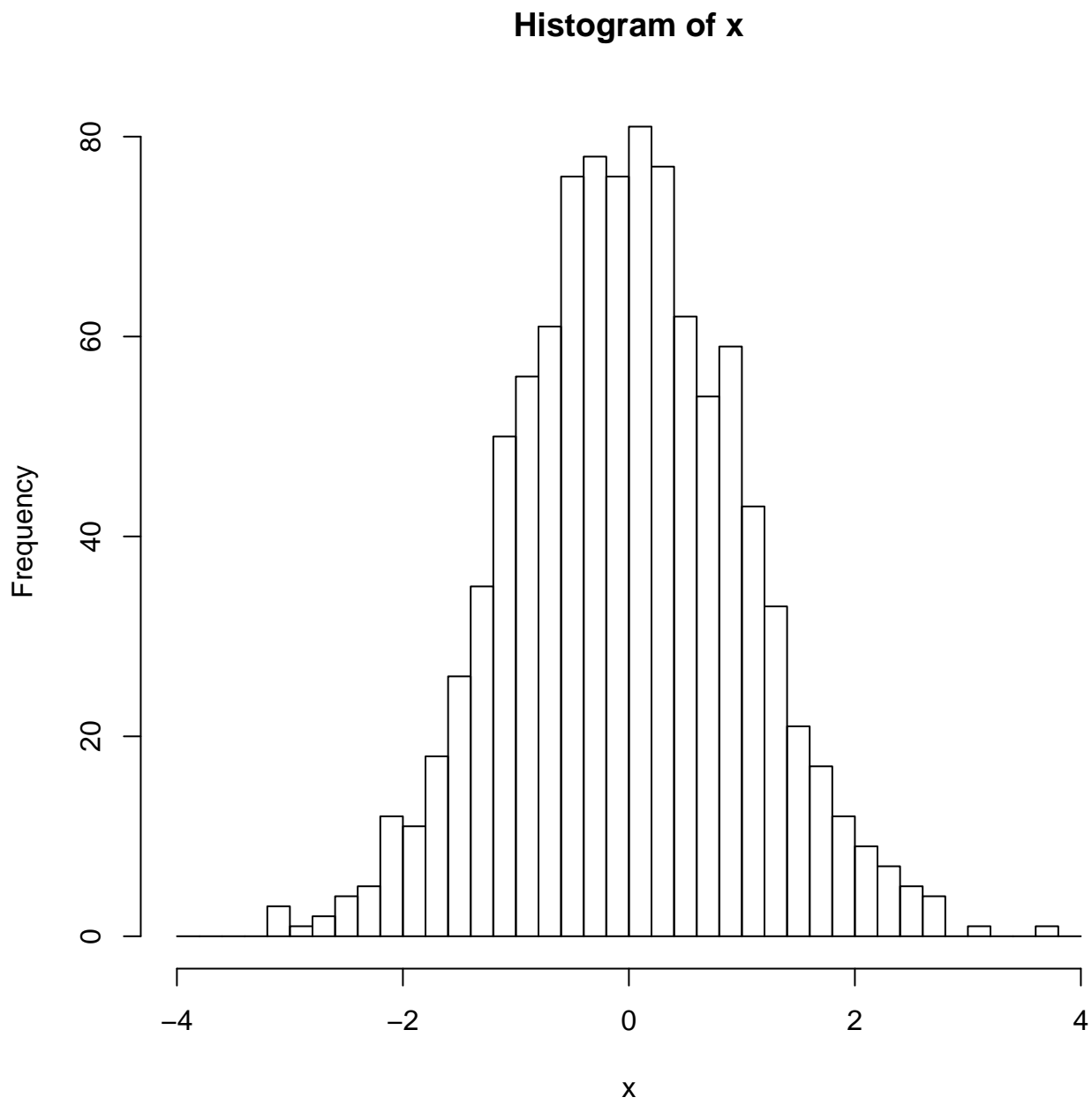


Figure 3: Standard Normal Simulation

Histogram of y

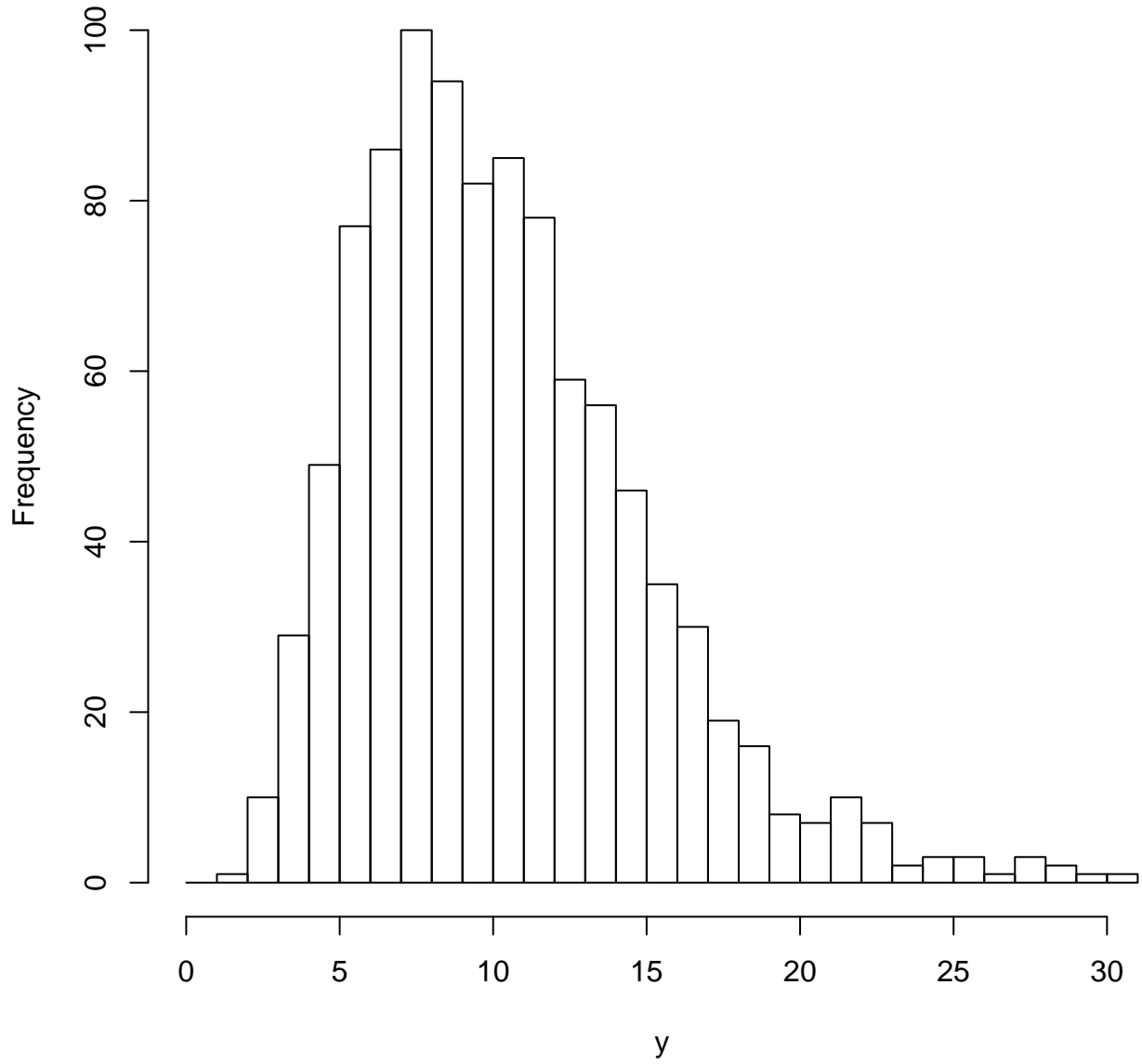


Figure 4: Chi Squared Simulation