MATH10282 Introduction to Statistics
Semester 2, 2019/2020
Examples 5, Solutions

1. $X \sim \mathrm{Bi}(100, 0.3)$

(i) $\mathrm{P}(X = 33)$, $\mathrm{P}(28 \leq X \leq 34)$ and $\mathrm{P}(X < 38)$

```
 > dbinom(x=33, size=100, prob=0.3)
[1] 0.0685392
> pbinom(q=34, size=100, prob=0.3)- pbinom(q=27, size=100, prob=0.3)
[1] 0.5407756
> pbinom(q=37, size=100, prob=0.3)
[1] 0.9469544
```

(ii) Probabilities of the same events calculated using the normal approximation, with continuity correction:

```
> stdev <- sqrt( 100*0.3*0.7 )
> pnorm(33.5, mean=30, sd=stdev)-pnorm(32.5, mean=30, sd=stdev)
[1] 0.0701851
> pnorm(34.5, mean=30, sd=stdev)- pnorm(27.5, mean=30, sd=stdev)
[1] 0.5442558
> pnorm(37.5, mean=30, sd=stdev)
[1] 0.9491465
```

The results are very similar, suggesting the normal approximation is a good one. This is to be expected, as n is large and p is not too small.

2. $X \sim \mathrm{Po}(10)$.

(i) To plot the p.m.f., first define a grid of x-values, then evaluate the p.m.f. on this grid. Finally use the `plot` command to create the graphic.

```
> xp<-seq(0, 25, 1)
> pxp<-dpois(xp, 10)
> plot(xp, pxp)
> plot(xp, pxp, main="Poisson(10) pmf")
```

To compute $\mathrm{P}(X < 15)$, $\mathrm{P}(X \geq 8)$ and $\mathrm{P}(6 \leq X \leq 16)$, first check the help for the function `ppois`. This reveals that the function computes $\mathrm{P}(X \leq q)$. Thus we do the following:

```
> help(ppois)
> ppois(14, 10)
[1] 0.9165415
> 1-ppois(7, 10)
[1] 0.7797794
> ppois(16, 10)-ppois(5, 10)
[1] 0.9058724
```

(ii) To evaluate the percentiles, use the quantile function.

```
> qpois(0.25, 10)
[1] 8
> qpois(0.50, 10)
```

```
[1] 10
> qpois(0.75, 10)
[1] 12
```

3. $X \sim \text{Exp}(0.2)$.

(i) To calculate and plot the p.d.f:

```
> xe<-seq(0, 25, 0.2)
> dxe<-dexp(xe, 0.2)
> plot(xe, dxe, type="l")
> plot(xe, dxe, type="l", main="Ex(0.2) pdf")
```

To compute $\text{P}(X < 12)$, $\text{P}(X > 3)$ and $\text{P}(4 < X < 20)$:

```
> pexp(12, 0.2)
[1] 0.909282
> 1-pexp(3, 0.2)
[1] 0.5488116
> pexp(20, 0.2)-pexp(4, 0.2)
[1] 0.4310133
```

(ii) To find the percentiles use the quantile function:

```
> qexp(c(0.2, 0.5, 0.8), 0.2)
[1] 1.115718 3.465736 8.047190
```

4. $X \sim N(20, 7^2)$

(i) To plot the p.d.f.:

```
> xn<-seq(0, 40, 0.2)
> pxn<-dnorm(xn, 20 ,7)
> plot(xn, pxn, type="l", main="Normal pdf with mean=20, sd=7")
```

To calculate $\text{P}(X < 17)$, $\text{P}(X > 25)$ and $\text{P}(13 < X < 27)$:

```
> pnorm(17, 20, 7)
[1] 0.3341176
> 1-pnorm(25, 20, 7)
[1] 0.2375253
> pnorm(27, 20, 7)-pnorm(13, 20, 7)
[1] 0.6826895
```

(ii) To find the percentiles:

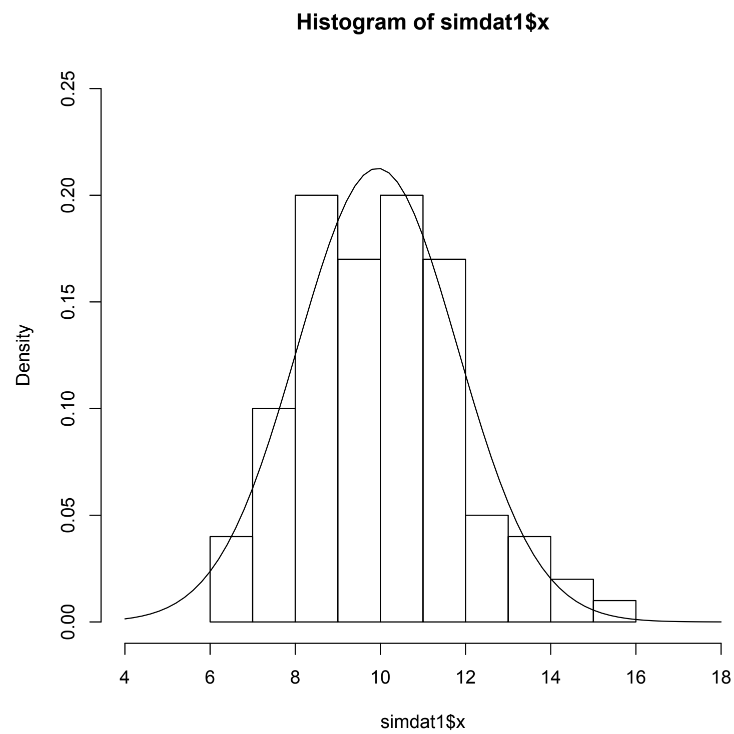
```
 > qnorm(c(0.05, 0.10, 0.90, 0.95), 20, 7)
[1]  8.486025 11.029139 28.970861 31.513975
```

(iii) If $X \sim N(\mu, \sigma^2)$, then $Q(p) = \mu + \sigma\Phi^{-1}(p)$, where $\Phi^{-1}(p)$ is the p quantile of a standard normal distribution. Hence the following code gives the same results as above:

```
> 20+7*qnorm(c(0.05, 0.10, 0.90, 0.95), 0, 1)
[1]  8.486025 11.029139 28.970861 31.513975
```

5. (i) For the `simdat` data, a normal density can be superimposed on a histogram as follows:

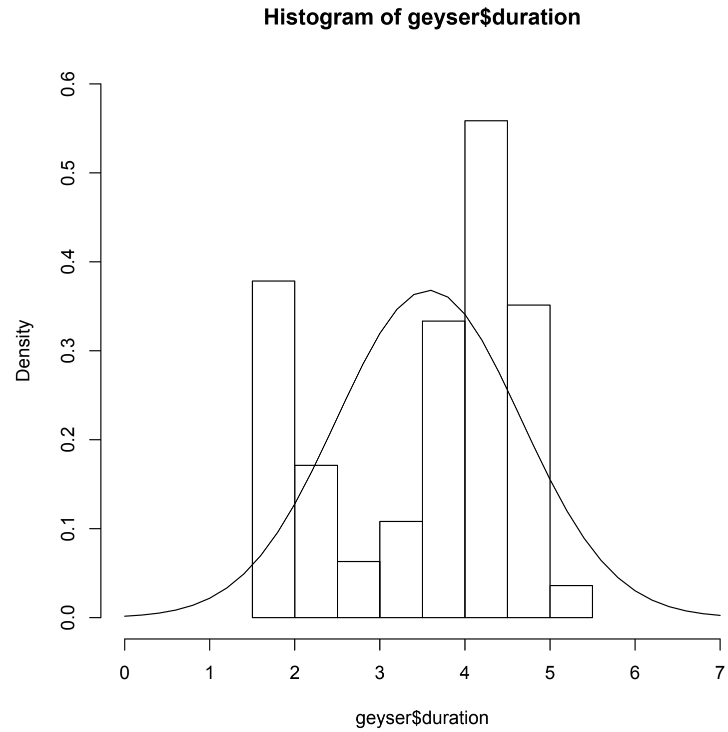
```
> simdat1<-read.table(file="https://minerva.it.manchester.ac.uk/~saralees/simdat1.txt",
                      header=T)
> hist(simdat1$x, freq=F, xlim=c(4, 18), ylim=c(0, 0.25))
> xs<-seq(from=4, to=18, by=0.2)
> yxs<-dnorm(xs, mean=mean(simdat1$x), sd=sd(simdat1$x))
> lines(xs, yxs)
```



The normal distribution looks like a reasonably good fit. There are possibly slightly fewer observations in the left tail than might be expected under the normal model.

(ii) For the `geyser` data, it is equally straightforward.

```
> geyser<-read.table(file="https://minerva.it.manchester.ac.uk/~saralees/geyser.txt",
                          header=T)
> names(geyser)
[1] "day"      "duration" "interval"
> hist(geyser$duration, freq=F, xlim=c(0, 7), ylim=c(0, 0.6))
> xg<-seq(0, 7, 0.2)
> yxg<-dnorm(xg, mean=mean(geyser$duration), sd=sd(geyser$duration))
> lines(xg, yxg)
```

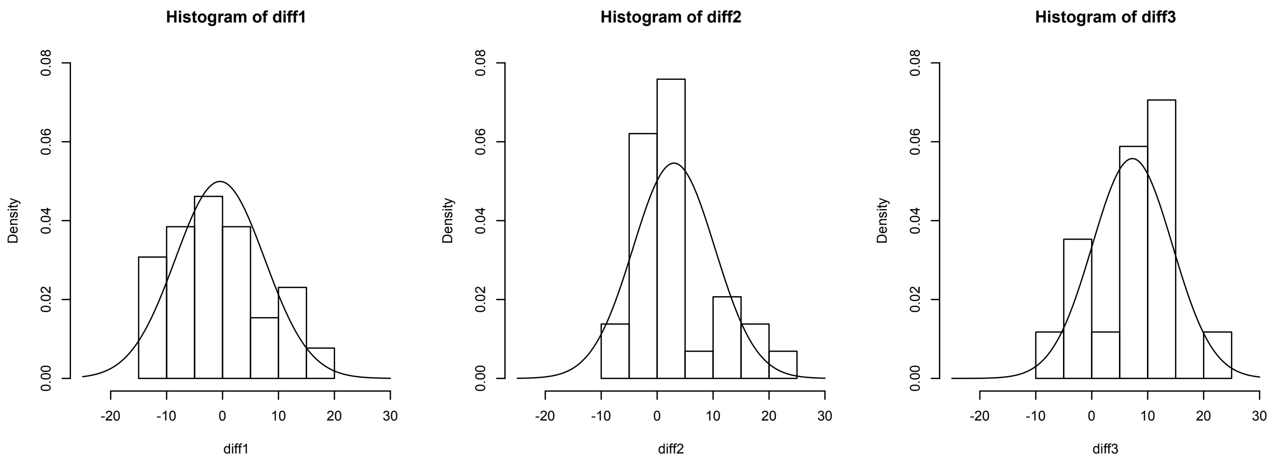


The normal distribution looks like a poor fit. The empirical distribution appears somewhat bimodal, i.e. there are two modes.

(iii) For the `anorexia` data we investigate normality within the three groups.

```
> anorexia<-read.table(file="https://minerva.it.manchester.ac.uk/~saralees/anorexia.txt",
                       header=T)
> names(anorexia)
[1] "case"      "prewt"     "postwt"    "treatment"

> diff<-anorexia$postwt-anorexia$prewt
> diff1<-diff[anorexia$treatment==1]
> diff2<-diff[anorexia$treatment==2]
> diff3<-diff[anorexia$treatment==3]

> hist(diff1, freq=F, xlim=c(-25, 30), ylim=c(0, 0.08))
> xd1<-seq(-25, 30, 0.2)
> yxd1<-dnorm(xd1, mean=mean(diff1), sd=sd(diff1))
> lines(xd1, yxd1)

> hist(diff2, freq=F, xlim=c(-25, 30), ylim=c(0, 0.08))
> xd2<-seq(-25, 30, 0.2)
> yxd2<-dnorm(xd2, mean=mean(diff2), sd=sd(diff2))
> lines(xd2, yxd2)

> hist(diff3, freq=F, xlim=c(-25, 30), ylim=c(0, 0.08))
> xd3<-seq(-25, 30, 0.2)
> yxd3<-dnorm(xd3, mean=mean(diff3), sd=sd(diff3))
```

```
> lines(xd3, yxd3)
```

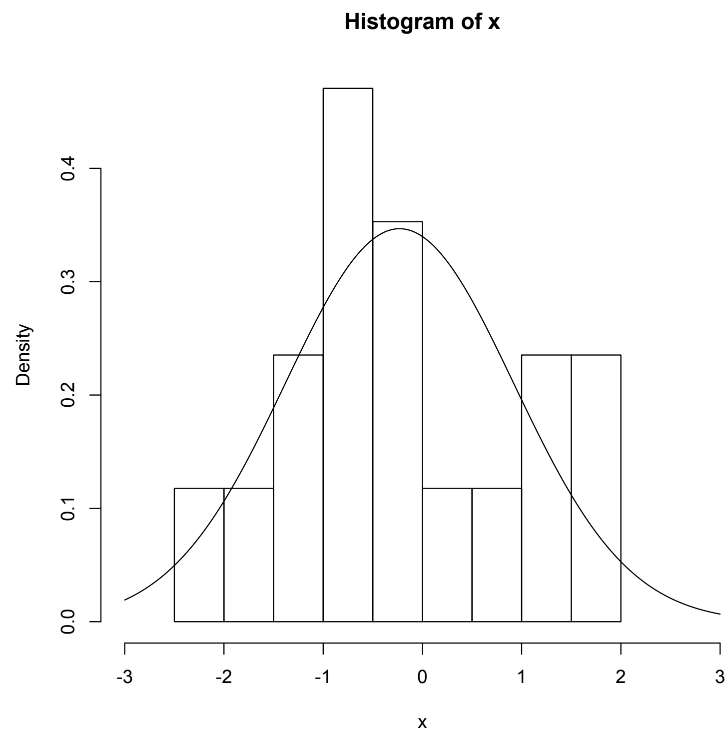


It is not clear whether the fits are reasonable. There is possibly some suggestion of bimodality within some of the groups. However, the sample size within each group is not particularly large:

```
> table(anorexia$treatment)
 1  2  3
26 29 17
```

Compare the above with a plot of the histogram and normal p.d.f. for a random sample of 17 observations simulated from $N(0,1)$. Even though the simulated data really are normally distributed in this case, the p.d.f. often does not look like a particularly good fit. This shows that it is difficult to verify normality for a small sample by looking at such plots. One can only really spot when the normality assumption is violated very strongly.

```
> x <- rnorm(17)
> xs <- seq(from=-3,to=3,length=200)
> ys <- dnorm(xs)
> hist(x,xlim=c(-3,3),freq=F)
> lines(xs,ys)
```


Histogram of x
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x