MATH10282 Introduction to Statistics
Semester 2, 2019/2020
Examples 7, Solutions

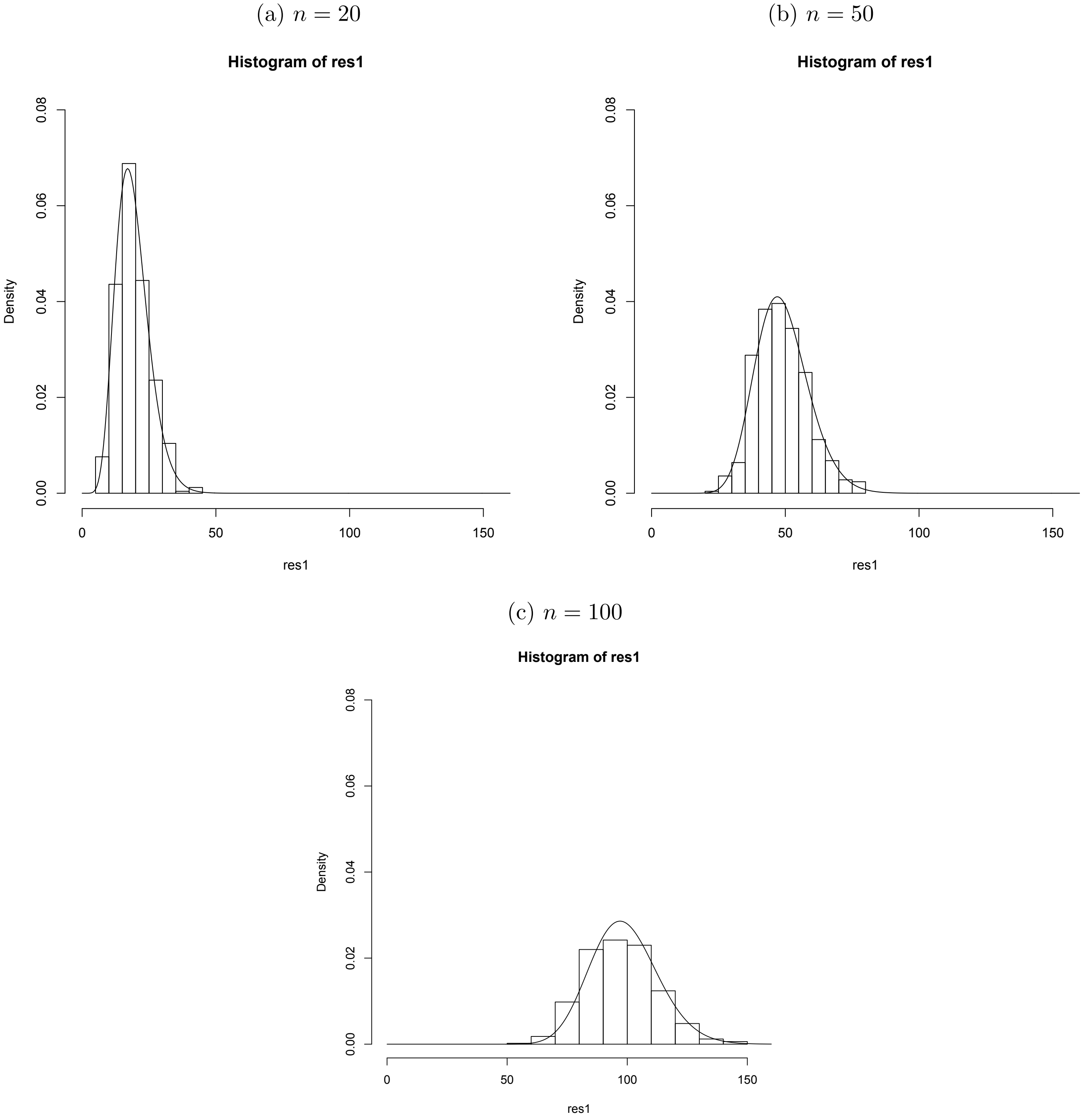
Sample variance

For $n = 50$, the code is given on the Example Sheet. For different values of n, we need to change: (i) the value of `n` that is supplied to the simulation function `newf.1`, (ii) the plotting points for the density function, and (iii) the degrees of freedom for the χ^2 density.

The following code is more flexible, if a different value of `n` is used on the first line, then then a correct plot will still be produced. We plot all of the histograms on the same axes (via `xlim` and `ylim`) to facilitate comparison.

```
n <- 50
res1 <- newf.1(n)
hist(res1, freq=F, xlim=c(0,160))
xv <- seq(from=0, to=160, by=0.5)
cv <- dchisq(xv, df=n-1)
lines(xv, cv)
```

This results in the plots overleaf, which indicate that the theoretical $\chi^2(n-1)$ distribution fits the simulated distributions very well. Note that your plots may be slightly different, as simulation produces a different random sample each time.



(a) $n = 20$

(b) $n = 50$

(c) $n = 100$

Figure: simulated sampling distributions of $(n-1)S^2/\sigma^2$, normally distributed data

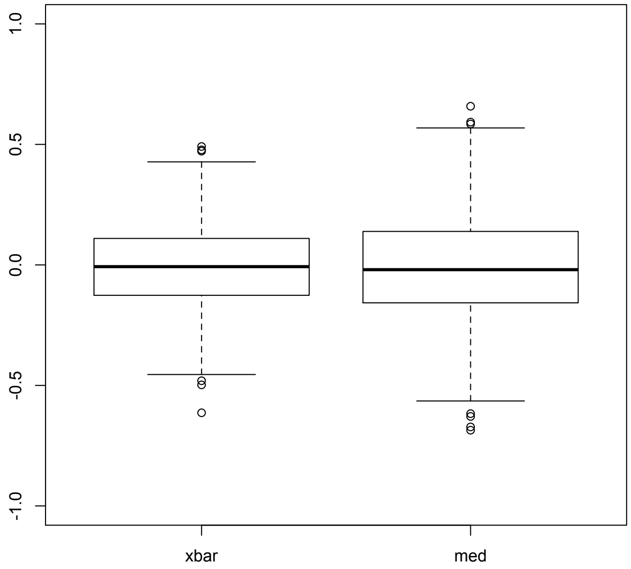
Sample mean and median

The plots are as below. We have put all of these on the same axes to facilitate comparison, using `ylim` as follows in the boxplot command:

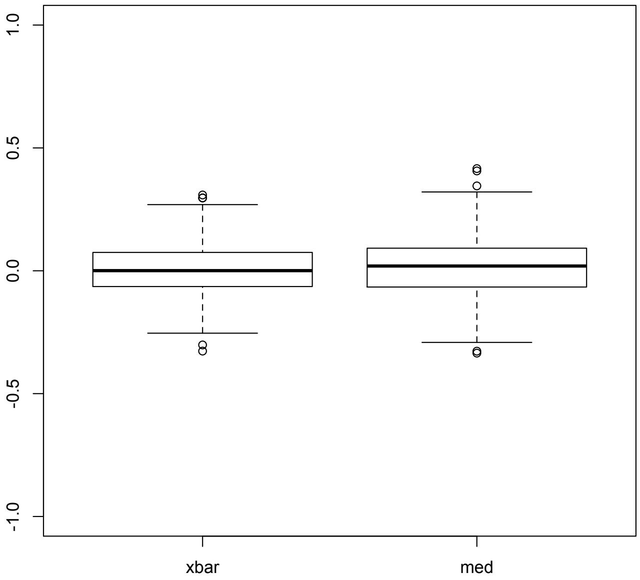
```
boxplot(res2, ylim=c(-1,1))
```

We see that the variance of the sampling distribution of $\bar{X}$ is lower than that of $\hat{Q}(0.5)$. Thus, for normally distributed data, we prefer to use the sample mean. However, in general the sample median is more robust to departures from the assumption of normality. The variance of the sampling distribution of both estimators decreases as n increases.

(a) $n = 30$



(b) $n = 100$



(c) $n = 200$

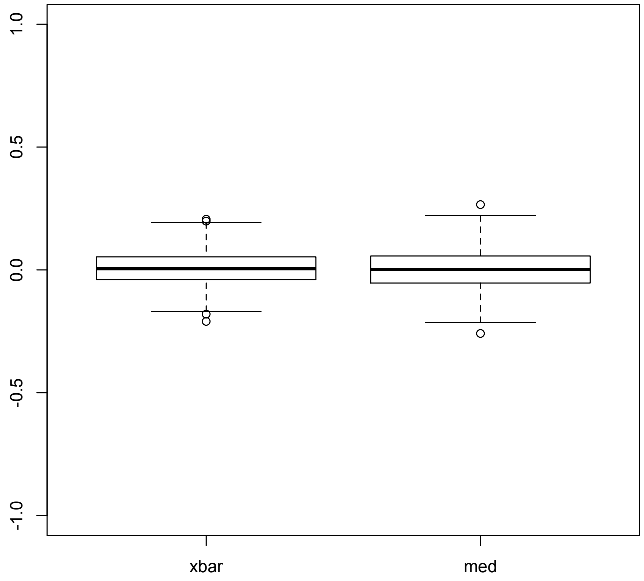


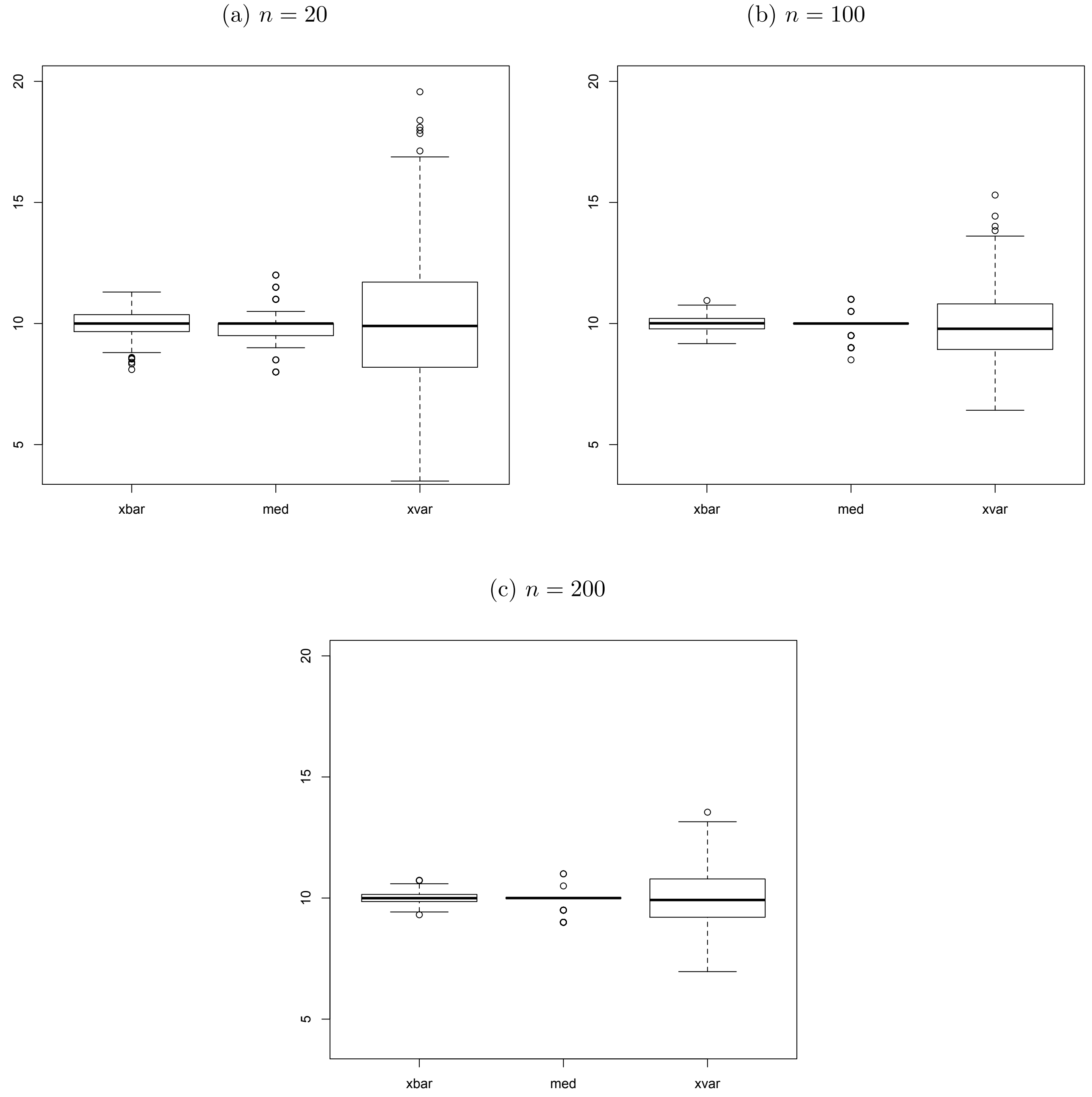
Figure: simulated sampling distributions of $\bar{X}$ and $\hat{Q}(0.5)$, normally distributed data

Sample mean, median and variance for Poisson data

To facilitate comparison across different sample sizes, we put all box plots on the same axes using `ylim`, for example as follows:

```
res3 <- newf.3(n=30); boxplot(res3, ylim=c(4,20))
```

The plots are as below:



(a) $n = 20$

(b) $n = 100$

(c) $n = 200$

Figure: simulated distributions of three estimators of λ, Poisson distributed data

The first feature we notice is that in fact the distribution of the median is discrete. As the data are integer valued, the median can only take values which are an integer, or an integer plus 0.5. This is an undesirable property of an estimator, since in principle λ may not be an integer (or an integer plus 0.5). Thus the sample median is not a good estimator in this context.

The variance of the sampling distribution of $\bar{X}$ is much lower than that of S^2, and so $\bar{X}$ is preferable overall. The variance of the sampling distribution of both $\bar{X}$ and S^2 decreases as n increases from 100 to 200, but perhaps not as much as might be anticipated. Thus it is a matter of judgement whether the additional cost of gathering 200 rather than 100 observations is worthwhile in terms of the improvement in the statistical inference.

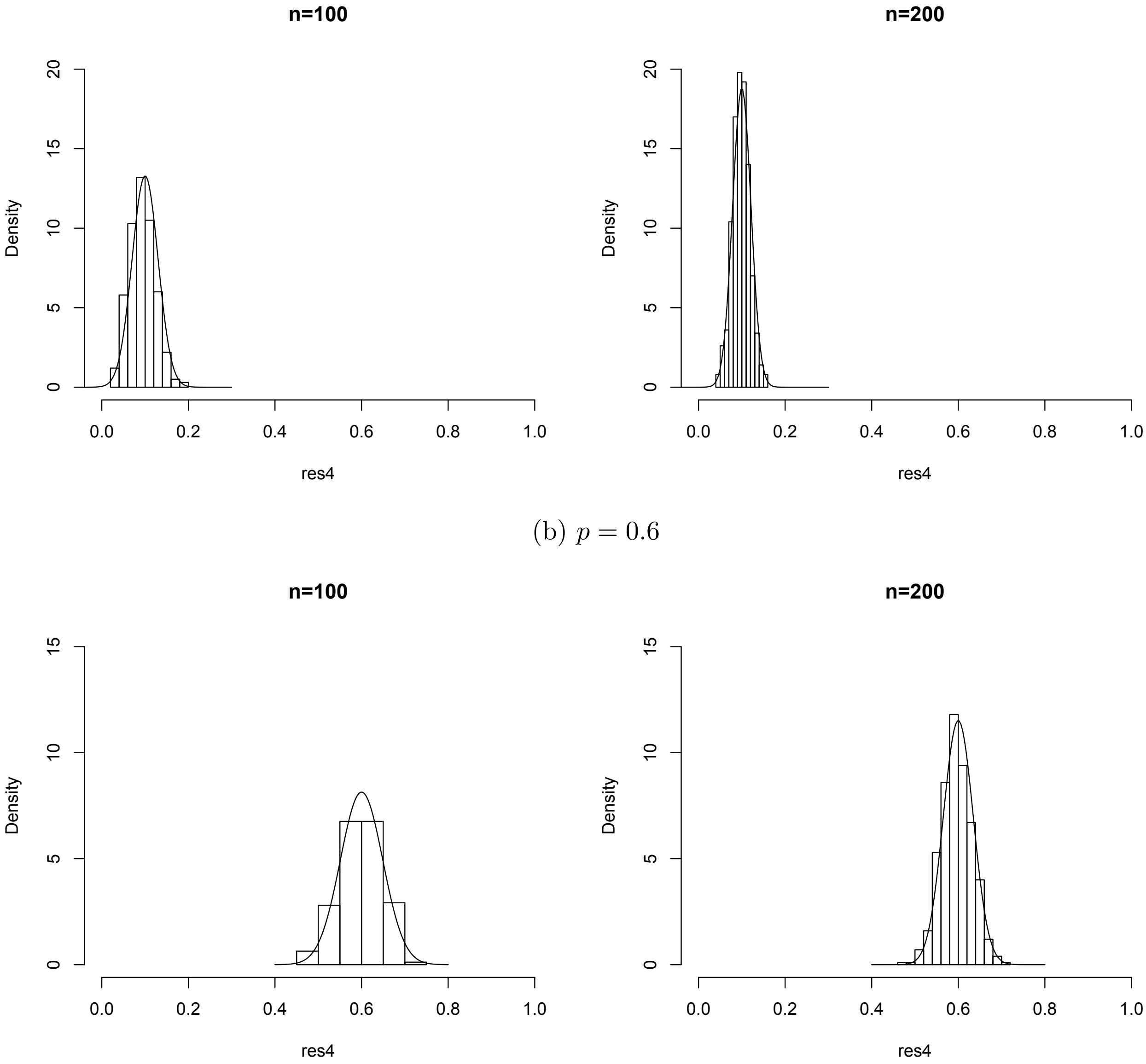
Sample proportion, Binomial data

To facilitate comparison, we plot all histograms on the same axes via `xlim` and `ylim`, e.g.:

```
n <- 100
p <- 0.6

res4 <- newf.4(n, p)
summary(res4)
var(res4)
sd(res4)
hist(res4, freq=F, xlim=c(0,1), ylim=c(0,15))
xv <- seq(from=p-0.2, to=p+0.2, length=100)
dv <- dnorm(xv, mean=p, sd=sqrt(p*(1-p)/n))
lines(xv, dv)
```

The plots are shown on the next page. The fit of the normal approximation is better when $n = 200$, and when $p = 0.6$. The variance of the sampling distribution reduces as n increases.



(a) $p = 0.1$

(b) $p = 0.6$

Figure: simulated distribution of the sample proportion, Binomially distributed data