Worksheet 3

Once again load the fastR2 package. If it is not installed, run the following commands.

```
install.packages("devtools")
```

```
install.packages("fastR2",repos = "http://cran.us.r-project.org")
```

The following command loads the fastR2 package.

```
require(fastR2)
```

Confidence Intervals

As is well known, a confidence interval is an estimator and changes with the sample. Here we see some examples of how to obtain a confidence interval - 1. in the case of a normal r.v. and 2. a large sample from a binomial r.v. We also show how someone might produce them through simulation. Particularly, we can check, in a simulation, that a 95% confidence interval will contain the observed value of the estimate ~95% of the time. Finally, we also see how we may visualize them.

Normal Random Variable

A sample from a binomially distributed random variable can be obtained as follows

```
asample <- rnorm(100,22,5)
```

We can then find out the observed mean and standard deviation of that sample

```
m <- mean(asample)
s <- sd(asample)
```

Now, to find a 95% confidence interval, we first find out that x value that gives 97.5% mass. For this purpose, we use the normal approximation with variance unknown, hence we use the t-distribution.

```
cutoff <- qt(0.975, 99)
```

Now the confidence interval is

```
l <- m - cutoff*s/sqrt(100)
u <- m + cutoff*s/sqrt(100)
```

Approximately Normal Random Variable

A sample from a binomially distributed random variable can be obtained as follows

```
asample <- rbinom(100,50,0.4)
```

We can then find out the observed mean and standard deviation of that sample

```
m <- mean(asample)
s <- sd(asample)
```

Now, to find a 95% confidence interval, we first find out that x value that gives 97.5% mass. For this purpose, we use the normal approximation with variance unknown, hence we use the t-distribution.

```
cutoff <- qt(0.975, 99)
```

Now the confidence interval is

```
l <- m - cutoff*s/sqrt(100)
u <- m + cutoff*s/sqrt(100)
```

> **Exercise 1**: Create a sample from a normal random variable. Find an exact 95% confidence interval for the variance of that sample assuming that the mean and variance are not known.

Simulation

In the simulation, we will create an R function that computes a confidence interval for a given sample. After that, we find the proportion of times when the observed estimate is found in that confidence interval.

```
confint <- function(x, level)
{
  l <- length(x)
  m <- mean(x)
  s <- sd(x)
  cutoff <- qt((1+level)/2,l-1)
  lower <- m - cutoff*s/sqrt(l)
  upper <- m + cutoff*s/sqrt(l)
  belongs <- (20 >= lower && 20 <= upper)
  print(c(lower,upper, belongs,m))
}
```

Replicate the confidence intervals 5000 times.

Check how many of the observed values belong to the confidence interval

```
sum(intervals[,3])/5000
```

```
## [1] 0.9552
```

> **Exercise 2**: Perform a similar simulation for variance of a normally distributed random variable.

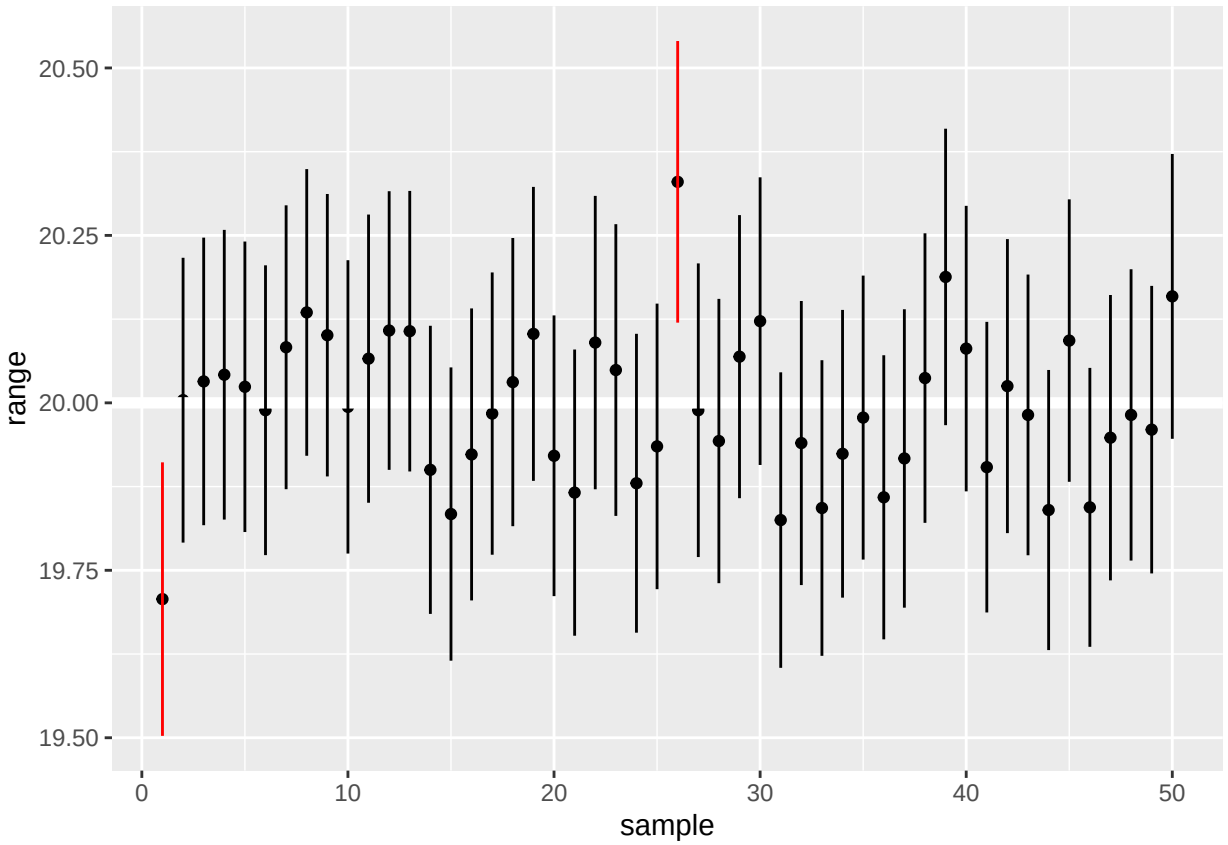
Visualization

```
lower = intervals[,1]
upper = intervals[,2]
mean = intervals[,4]

my_data <- data.frame(lower1 = lower[1:50], upper1 = upper[1:50], mean = mean[1:50], x = 1:50)

my_data <- my_data %>%
  mutate(mycolor = ifelse((upper1 < 20)|(lower1 > 20), "red", "black"))

gf_point(mean ~ x, data = my_data) %>%
  gf_hline( yintercept = 20, color = "white", linewidth = 2) %>%
  gf_segment(lower1 + upper1 ~ x + x, color = my_data$mycolor, data = my_data) %>%
  gf_labs(y="range", x="sample")
```



Exercise 3: Create a similar visualization for variance of a normally distributed random variable.

Hypothesis Tests

We will look at how to perform a hypothesis test : 1. an exact binomial test, 2. an exact t test, 3. hypothesis testing under approximate normality.

An exact binomial test

In an experiment, we find the mean of a binomially distributed r.v. to be 25.2 in a sample of size 200. Suppose that through theoretical means or from prior experience, we knew that p=0.2. This forms the null hypothesis. Therefore, the alternate hypothesis is that p is not 0.2. Do we have evidence to reject null hypothesis? We compute the p-value of the observed data as follows.

```
pbinom(25.2,200,0.2)  #This is the probability of getting 25.2 or less
```

```
## [1] 0.003628754
```

Now, we must find out all those x values for which the probability would be lower than or equal to that for x=25.

```
probs <- dbinom(0:200,200,0.2)
pval <- sum(probs[probs <= dbinom(25,200,0.2)])
if (pval < 0.05)
{
  print("Reject Null Hypothesis")
}
```

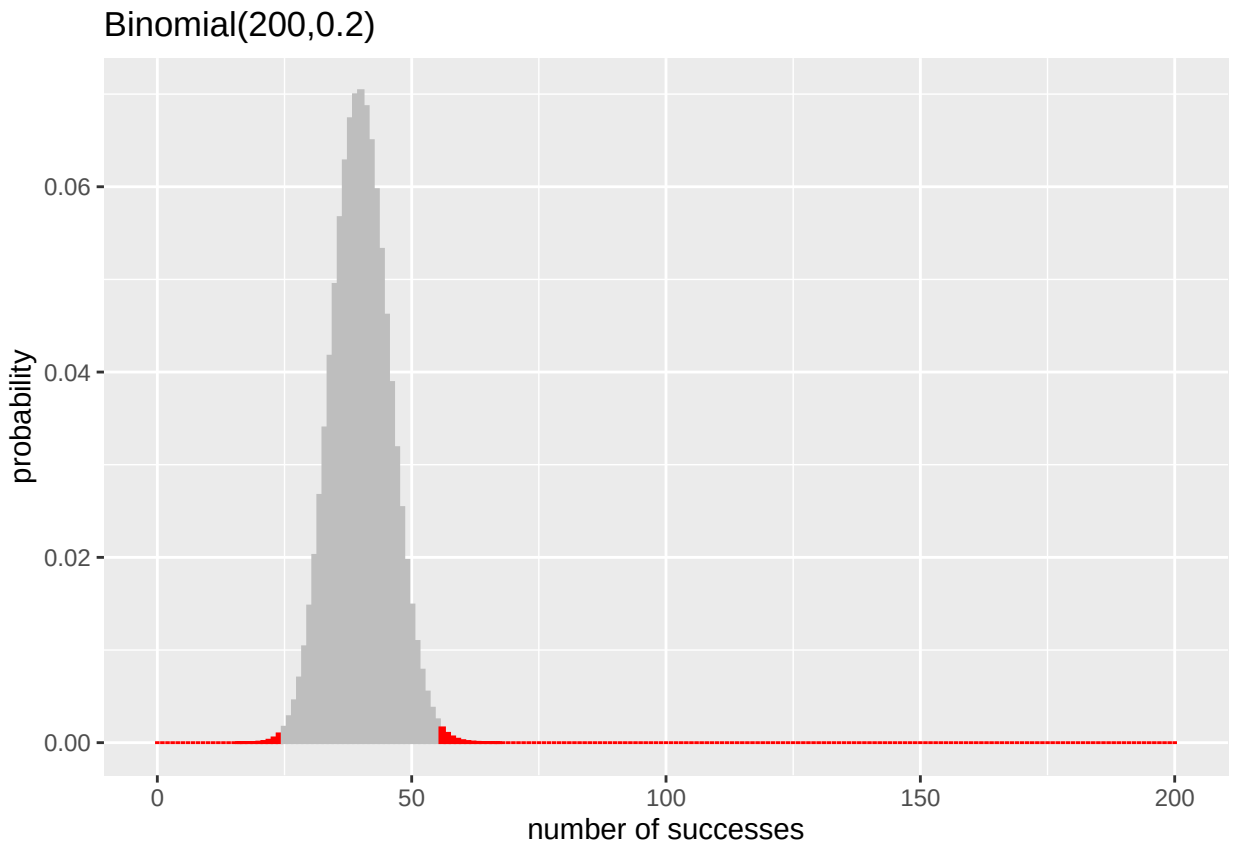
```
## [1] "Reject Null Hypothesis"
```

Let us visualize the reject region

```
probsf <- data.frame(probs, x = 0:200)

probsf <- probsf %>%
  mutate(mycolor = ifelse((probs < dbinom(25,200,0.2)), "red", "gray"))

gf_col(probs ~ x, fill = probsf$mycolor, colour = probsf$mycolor, data = probsf) %>%
  gf_labs(title = "Binomial(200,0.2)", x="number of successes", y="probability")
```



Exercise 4: In an experiment, we find the mean of a binomially distributed r.v. to be 50 in a sample of size 100. Suppose that through theoretical means or from prior experience, we knew that p=0.4. This forms the null hypothesis. Therefore, the alternate hypothesis is that p is not 0.4. Do we have evidence to reject null hypothesis?

An exact z test

In an experiment, we find the mean of a normally distributed r.v. to be 289 in a sample of size 50. We also know that the standard deviation is 100. Suppose that through theoretical means or from prior experience, we believed that the true mean is 252. This forms the null hypothesis. Therefore, the alternate hypothesis is that the mean is not 252. Do we have evidence to reject null hypothesis at 5% significance level? We compute the p-value of the observed data as follows.

```
qnorm(0.975)
```

```
## [1] 1.959964
```

Calculate a t-statistic

```
val <- sqrt(50)*(289-252)/100

val
```

```
## [1] 2.616295
```

```
if (val > qnorm(0.975))
{
  print("Reject Null Hypothesis")
}
```

```
## [1] "Reject Null Hypothesis"
```

```
pvalue <- 2*(1-pnorm(val))

pvalue
```

```
## [1] 0.00888897
```

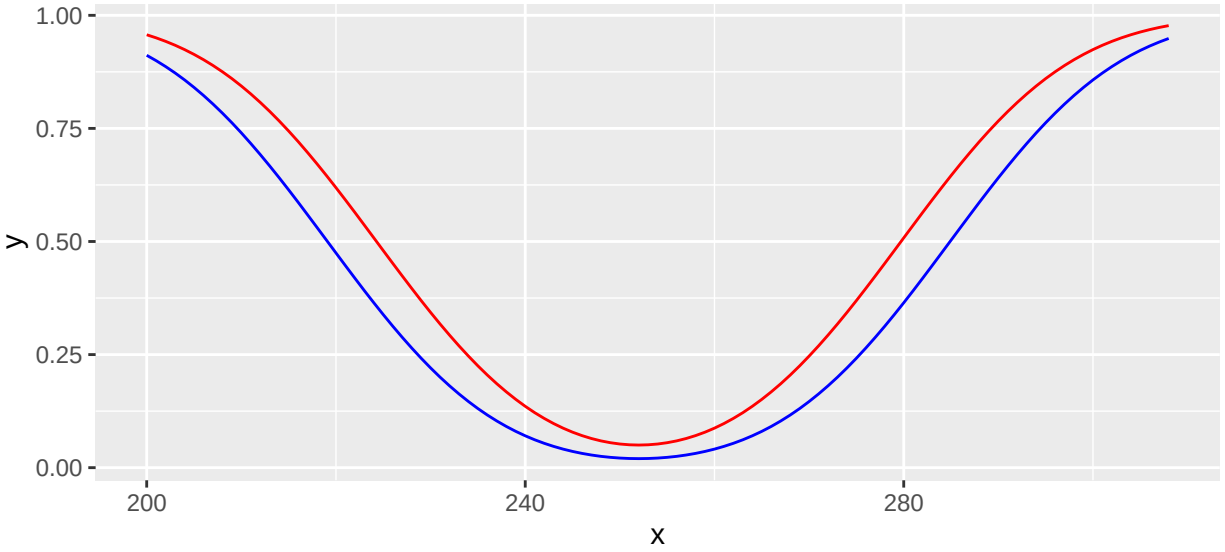
Let us compute the power of this test. Power is the probability of the complement of the event that null hypothesis is false but it was accepted. This is easiest to do by simulation. We simulate a large number of samples of size 50 according to alternative hypothesis. Then, we compute the test statistic for them. We find the proportion of the test-statistics that are above the threshold according to the null-hypothesis.

```
1-pnorm(qnorm(0.975),val,1)
```

```
## [1] 0.7441944
```

Let us plot power as a function of the alternative hypothesis

```
x = seq(200,308,by=1)
val = sqrt(50)*(x-252)/100
y = 1-pnorm(qnorm(0.975),val,1)+pnorm(qnorm(0.025),val,1)
z = 1-pnorm(qnorm(0.99),val,1)+pnorm(qnorm(0.01),val,1)
gf_line(y ~ x, color = "red") %>%
  gf_line(z ~ x, color = "blue")
```



Exercise 5: In an experiment, we find the mean of a normally distributed r.v. to be 500 in a sample of size 20. Suppose that through theoretical means or from prior experience, we knew that the mean is 480. This forms the null hypothesis. Therefore, the alternate hypothesis is that

the mean is not 480. Repeat the steps above while assuming that the standard deviation is 35.

References

1. Randall Pruim, Foundations and Applications of Statistics, An Introduction Using R, Second Edition, American Mathematical Society, 2018