Lab 6 - Homework, Part C2

Nikita Grabher-Meyer

11/2/2020

PART C (coding)

Part 2

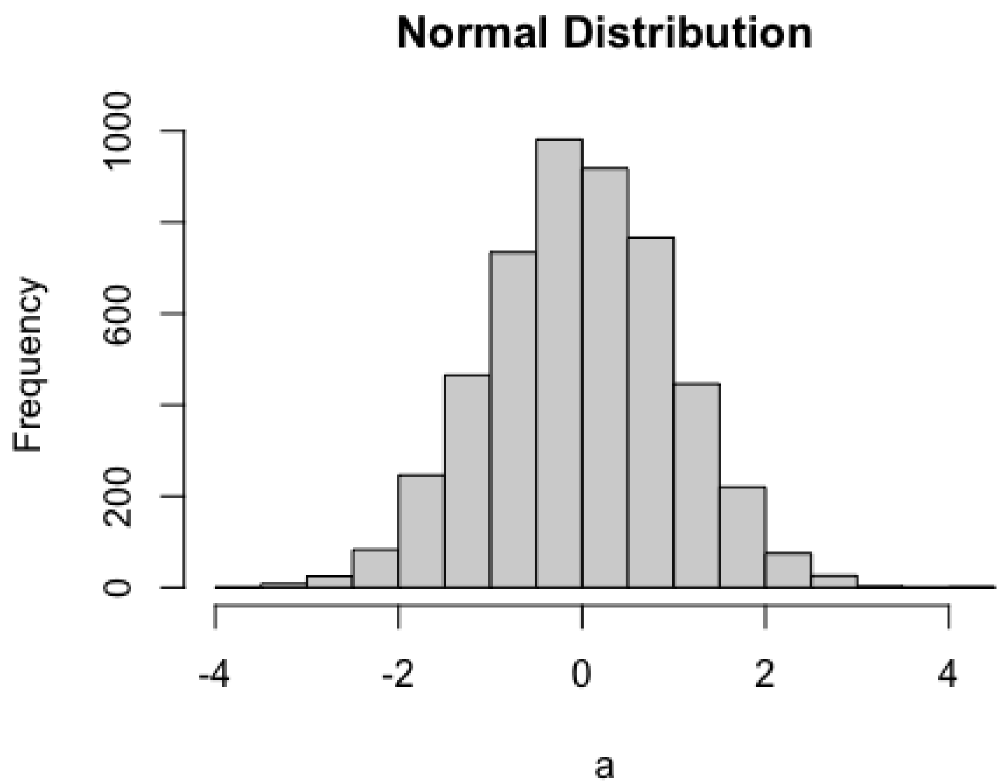
Choose at least 3 distributions and try to simulate data (e.g. 5000 draws) from a random variable that follows this distribution

Normal distribution

```
a <- rnorm(5000)
a[1:20]
```

```
##  [1] -1.33848985  0.28189075 -0.11710750 -0.32896624  0.72546409 -1.160577
35
##  [7] -0.45382283  1.65168427 -1.18073029  0.33555488 -1.25131628  0.146488
10
## [13]  0.11733128  0.04955321  0.96881637  0.36203229 -1.28443476  0.469270
16
## [19] -1.03710100  1.48940570
```

```
hist(a, main = "Normal Distribution")
```



Binomial distribution: number of successes in a set of pass/fail trials with success estimated at probability p The function takes three arguments: number of observations you want to see, number of trials per observation, probability of success for each trial.

```
b <- rbinom(5000,2,.4)
b[1:20]

##  [1] 1 2 1 1 0 1 1 2 0 1 2 2 0 2 1 0 0 1 1 1
```

Poisson distribution: number of expected events for a process given we know the average rate at which events occur during a given unit of time The function takes three arguments: Number of observations you want to see,The estimated rate of events for the distribution; this is expressed as average events per period.

```
c <- +rpois(5000,2)
c[1:20]

##  [1] 4 2 3 2 0 2 3 4 5 2 4 4 0 1 1 3 3 1 1 3
```

Draw two samples from each variable and report the sample means. Are they equal?

Normal distribution

```
s=500 #sample size

samplea1 <- sample(a,s)
samplea2 <- sample(a,s)

mean(samplea1)

## [1] 0.02571229

mean(samplea2)

## [1] -0.09240917
```

Binomial distribution

```
s=500 #sample size

sampleb1 <- sample(b,s)
sampleb2 <- sample(b,s)

mean(sampleb1)

## [1] 0.842

mean(sampleb2)

## [1] 0.77
```

Poisson distribution

```
s=500 #sample size

samplec1 <- sample(c,s)
samplec2 <- sample(c,s)

mean(samplec1)

## [1] 1.896

mean(samplec2)

## [1] 2.032
```

The sample means (no matter which distribution is considered) are not equal.

Follow the steps for illustrating the Central Limit Theorem using a poisson, a uniform and a normal distribution.Compare the normal, poisson and the uniform, what do you notice for 100 draws?

Poisson distribution

```
poisson10 <- replicate (10000, {
  a <- rpois(10,5)
```

```
  mean(a)
})

poisson100 <- replicate (10000, {
  a <- rpois(100,5)
  mean(a)
})
```

Uniform distribution

```
uniform10 <- replicate (10000, {
  b <- rpois(10,5)
  mean(b)
})

uniform100 <- replicate (10000, {
  b <- rpois(100,5)
  mean(b)
})
```

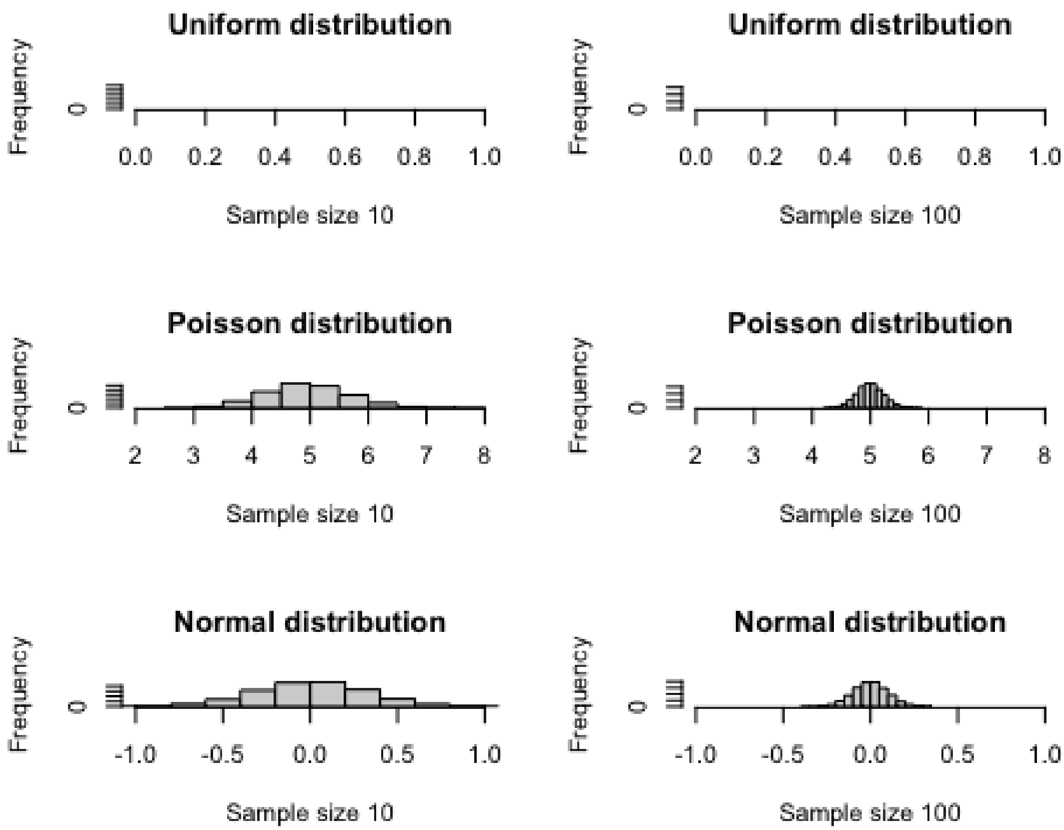
Normal distribution

```
normal10 <- replicate (10000, {
  c <- rnorm(10)
  mean(c)
})

normal100 <- replicate (10000, {
  c <- rnorm(100)
  mean(c)
})
```

Compare the three distributions

```
par(mfrow=c(3,2))
hist(uniform10, xlim=c(0,1), main="Uniform distribution", xlab="Sample size 1
0")
hist(uniform100, xlim=c(0,1), main="Uniform distribution", xlab="Sample size
100")
hist(poisson10, xlim=c(2,8), main="Poisson distribution", xlab="Sample size 1
0")
hist(poisson100, xlim=c(2,8), main="Poisson distribution", xlab="Sample size
100")
hist(normal10, xlim=c(-1,1), main="Normal distribution", xlab="Sample size 10
")
hist(normal100, xlim=c(-1,1), main="Normal distribution", xlab="Sample size 1
00")
```



The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population, then the distribution of the sample means will be approximately normally distributed (regardless of the population distribution).

Testing

Load the data

```
setwd("/Users/nikitagrabher-meyer/Desktop/PHD/Econometrics/Labs/Lab 6, Homewo
rk")

library(data.table)
library(ggplot2)
library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary St
atistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
load("dt_wages.RData")
dt.wages <- data.table(dt.wages)
```

Summaries

```r
stargazer(dt.wages, type = "text")

##
## ===============================================================
## Statistic  N   Mean   St. Dev.  Min   Pctl(25) Pctl(75)  Max
## ---------------------------------------------------------------
## wage      526  5.896   3.693   0.530   3.330    6.880   24.980
## educ      526 12.563   2.769     0       12       14      18
## exper     526 17.017   13.572    1       5        26      51
## tenure    526  5.105   7.224     0       0        7       44
## nonwhite  526  0.103   0.304     0       0        0       1
## female    526  0.479   0.500     0       0        1       1
## married   526  0.608   0.489     0       0        1       1
## numdep    526  1.044   1.262     0       0        2       6
## smsa      526  0.722   0.448     0       0        1       1
## northcen  526  0.251   0.434     0       0       0.8      1
## south     526  0.356   0.479     0       0        1       1
## west      526  0.169   0.375     0       0        0       1
## construc  526  0.046   0.209     0       0        0       1
## ndurman   526  0.114   0.318     0       0        0       1
## trcommpu  526  0.044   0.205     0       0        0       1
## trade     526  0.287   0.453     0       0        1       1
## services  526  0.101   0.301     0       0        0       1
## profserv  526  0.259   0.438     0       0        1       1
## profocc   526  0.367   0.482     0       0        1       1
## clerocc   526  0.167   0.374     0       0        0       1
## servocc   526  0.141   0.348     0       0        0       1
## lwage     526  1.623   0.532   -0.635  1.203    1.929   3.218
## expersq   526 473.435 616.045    1       25       676    2,601
## tenursq   526 78.150  199.435    0       0        49     1,936
## ---------------------------------------------------------------
```

Create an estimator for the average wage by group

```r
wage <- dt.wages[,list(avg_wage=mean(wage))]

wage.nowh <- dt.wages[,list(avg_wage=mean(wage)), by=nonwhite]
wage.fem <- dt.wages[,list(avg_wage=mean(wage)), by=female]

wage.nowh

##    nonwhite avg_wage
## 1:        0 5.944174
## 2:        1 5.475926

wage.fem
```

```
##    female avg_wage
## 1:      1 4.587659
## 2:      0 7.099489
```

Confidence intervals How to calculate the 95% confidence intervals for the population wage?

```
dt.wages[ , list (avg_wage=mean(wage) , sd_wage=sd(wage))]
```

```
##    avg_wage  sd_wage
## 1: 5.896103 3.693086
```

Create a function to calculate the 95% confidence interval

```
conf.int <- function(X){
     n <- length(X)
     error <- qt(0.975, df=n-1) * sd(X) / sqrt(n)
     mean.X <- mean(X)
     return(list(lower = mean.X - error, upper = mean.X +error))
}
```

Apply the "conf.int" function to the wage variable

```
dt.wages[,conf.int(wage)]
```

```
##       lower    upper
## 1: 5.579768 6.212437
```

Confidence intervals by group

```
dt.wages[,conf.int(wage), by=nonwhite]
```

```
##    nonwhite    lower    upper
## 1:        0 5.605032 6.283316
## 2:        1 4.614661 6.337191
```

```
dt.wages[,conf.int(wage), by=female]
```

```
##    female    lower    upper
## 1:      1 4.273855 4.901462
## 2:      0 6.604626 7.594352
```

Hypothesis testing, using the function t-test

```
dt.wages[female==1, t.test(wage,mu=5)]
```

```
## 
##  One Sample t-test
## 
## data:  wage
## t = -2.5879, df = 251, p-value = 0.01022
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.273855 4.901462
```

```
## sample estimates:
## mean of x
##  4.587659
```

Are wages of men different from wages of women?

```
t.test(dt.wages[female==1,wage],dt.wages[female==0,wage])

##
##  Welch Two Sample t-test
##
## data:  dt.wages[female == 1, wage] and dt.wages[female == 0, wage]
## t = -8.44, df = 456.33, p-value = 4.243e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.096690 -1.926971
## sample estimates:
## mean of x mean of y
##  4.587659  7.099489
```

Are wages of whites different from wages of non-whites?

```
t.test(dt.wages[nonwhite==1,wage],dt.wages[nonwhite==0,wage])

##
##  Welch Two Sample t-test
##
## data:  dt.wages[nonwhite == 1, wage] and dt.wages[nonwhite == 0, wage]
## t = -1.0118, df = 71.298, p-value = 0.3151
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3909497  0.4544541
## sample estimates:
## mean of x mean of y
##  5.475926  5.944174
```