MOEC0021
Empirical Methods
University of Zurich HS 2019
PROF: Greg Crawford / TA: E. Dicarlo, M.R. Greco, S. Bagagli, A. Jenni Handout 1

Exercise 1

Suggested Solutions

1 Theory

1. Suppose you *knew* the process generating the data in a population of interest was of the form

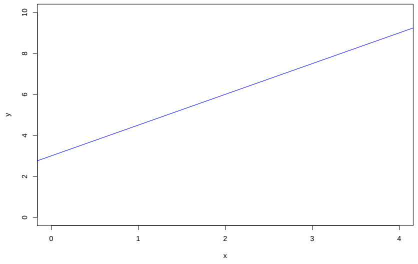
$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

with $\beta_1 = 3$ and $\beta_2 = 1.5$.

(a) Write down the population regression function. Draw a picture of $E(Y_i|X_i)$, the non-random part of the PRF.

Simply

$$Y_i = 3 + 1.5X_i + \epsilon_i$$



(b) You are given the following 4 observations drawn independently from this population:

Construct a table with the following values: the mean of X, $\bar{X}$; the mean of Y, $\bar{Y}$; X in mean-deviation form, $x_i = (X_i - \bar{X})$; Y in mean-deviation form, $y_i = (Y_i - \bar{Y})$, and the cross-product $x_i y_i$, and square of x_i, x_i^2, both in mean-deviation form.

	X	Y
1	1	4
2	4	10
3	3	9
4	2	7

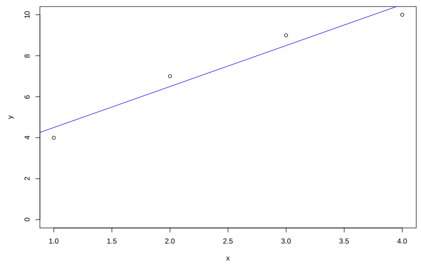
X_i	$\bar{X}$	x_i	Y_i	$\bar{Y}$	y_i	x_i^2	x_iy_i	ϵ_i	e_i	y_i^2
1	2.5	-1.5	4	7.5	-3.5	2.25	5.25	-0.5	-0.5	12.25
4	2.5	+1.5	10	7.5	+2.5	2.25	3.75	1.0	-0.5	6.25
3	2.5	+0.5	9	7.5	+1.5	0.25	0.75	1.5	0.5	2.25
2	2.5	-0.5	7	7.5	-0.5	0.25	0.25	1.0	0.5	0.25
10	10	0	30	30	0	5	10	3.0	0	21

(c) Calculate the OLS estimates of β_1 and β_2.

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{10.0}{5.0} = 2$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 7.5 - 2 * 2.5 = 2.5$$

(d) Plot the 4 points and the OLS line. Note this is the non-residual part of your Sample Regression Function.



```
## (Intercept)           x
##         2.5         2.0
```

(e) How does the OLS line compare to the line you drew from your Population Regression Function?

Our estimated intercept is "too low" and our estimated slope is "too high". The difference between the population regression function and sample regression arises from sampling. For every sample n from the population there are n different sample regressions. (Hence the need for hypothesis tests regarding the relation between estimated values and true underlying parameters). The SRF will almost never coincided with the PRF. As a researcher, you must have faith that your estimating technique ("the estimator") provides good estimates of the underlying the true population regression line.

(f) Does your sample regression function cross the population regression function? Suppose you were to select another sample from the same population. Is it possible that the two (sample) lines would *not* cross? Why or why not?

The sample regression function and the popuation regression function will *always* cross. If they don't have the same slope, they cross. The only case when this does not happen is when they have the same slope (β_2), but a different intercept. This implies that the sample covariance of x and y and the sample variance of x are the same of the population one. However, the population and sample mean of x and y are not the same. This can only happen if the sample is a linear additive transformation of the population (which means this is not really a "sample").

Must two sample regression functions cross? It can be shown that the $cov(\hat{\beta}_1, \hat{\beta}_2) < 0$. In other words, if a sample over estimates the slope coefficient it will under estimate the intercept coefficient. Therefore, the two lines will cross. Try this by running a model with and without a constant; you should be able to predict whether your $\hat{\beta}_2$ will be larger or smaller *ex ante*. Following the same reasoning above, the two lines won't cross in the unlikely case that you that the two sample you extract from the population are exactly a linear additive transformation of each other, but then your sample is not a random one.

(g) Calculate the error, ϵ_i, for the data points in your sample. Also calculate the residuals, e_i, for these data points. Do the errors sum to zero? Do the residuals? Do your answers differ? If so, explain why in your own words.

The errors do *not* sum to zero. The residuals do. This is expected. The errors are random variables with an expectation of zero, but that doesn't mean they will be zero either individually or on average (for example, even if men's average height is 5'10", there is no guarantee in any sample that the resulting sample average is 5'10"). By contrast, the residuals sum to zero as a process of calculating the OLS estimators. In any regression, they will always sum to zero.

(h) Show that $\sum_{i=1}^{4}(X_i - \bar{X}) = 0$. Is this an idiosyncratic feature of this sample or would you expect it to hold in every sample?

This always holds. It's a simple feature of how means are calculated.

(i) Show that $\sum x_i y_i = \sum x_i Y_i$. Also show that this equals $\sum X_i y_i$. Is this an idiosyncratic feature of this sample or would you expect it to hold in every sample?

See the table below, esp. the last row with the sum of the elements in each column. This also always holds and is a result of the fact that the sum of two variables in mean-deviation form equals the sum of the same two variables, one in mean-deviation form and one in "regular" form (or "Sum Little*Little = Sum Little*Big").

X_i	$\bar{X}$	x_i	Y_i	$\bar{Y}$	y_i	$x_i y_i$	$x_i Y_i$	$X_i y_i$
1	2.5	-1.5	4	7.5	-3.5	5.25	-6.0	-3.5
4	2.5	+1.5	10	7.5	+2.5	3.75	15.0	10.0
3	2.5	+0.5	9	7.5	+1.5	0.75	4.5	4.5
2	2.5	-0.5	7	7.5	-0.5	0.25	-3.5	1.0
10	10	0	30	30	0	10	10	10

2 Empirical Application

1. In this empirical exercise, we will illustrate the impact of sample size on the variance of the sample mean using what are called "Monte carlo methods". In monte carlo methods, you *create your own data* and then evaluate the properties of functions of that data. While the concepts at play in this question are (fairly) easy, it is not necessarily as easy to program the computer to have it do exactly what you want it to. Thus this question is about having you develop some of your programming skills.

   In this question, we will work with data that are drawn from an *exponential* distribution. If you are not familiar with the exponential distribution, look it up on Wikipedia or Wolfram MathWorld. If a random variable, x_i, is distributed as an exponential, we denote this, $x_i \sim \exp(\lambda)$, where λ is the parameter governing the shape of the distribution. For $x_i \sim \exp(\lambda)$, you can show (or look up) that $E(x_i) = \frac{1}{\lambda}$ and $V(x_i) = \frac{1}{\lambda^2}$. For the rest of this question, we will assume $x_i \sim \exp(1)$.

   This question asks you to draw many samples of data from the distribution of x_i. Each sample is distinguished by its number of observations, which we denote (as usual) N. But in each question below, I will ask you to draw samples of size N many times. We will call these different samples *replications* and index them by the letter $r = 1, \ldots, R$. Thus the i^{th} draw of x from the r^{th} replication can be denoted x_i^r. And the sample average of the N values of x_i^r in the r^{th} replication can be denoted $\bar{x}^r$. We can also take the sample average and variance across the R replications of $\bar{x}^r$, which we will denote $\bar{x}$ (note *no* r) and $s_{\bar{x}}$. $s_{\bar{x}}$ is our estimate of the variance of the sample mean from a sample of size N discussed extensively in lecture.

   (a) Let $N = 1$ and $R = 200$. Calculate $\bar{x}^r$ for $r = 1, \ldots, 200$ show them in a histogram. Also calculate the across-replication average, $\bar{x}$, and sample variance, s_{barx}.

   (b) Let $N = 5$ and $R = 200$. Calculate $\bar{x}^r$ for $r = 1, \ldots, 200$ show them in a histogram. Also calculate the across-replication average, $\bar{x}$, and sample variance, s_{barx}.

   (c) Let $N = 20$ and $R = 200$. Calculate $\bar{x}^r$ for $r = 1, \ldots, 200$ show them in a histogram. Also calculate the across-replication average, $\bar{x}$, and sample variance, s_{barx}.

   (d) Let $N = 1,000$ and $R = 200$. Calculate $\bar{x}^r$ for $r = 1, \ldots, 200$ show them in a histogram. Also calculate the across-replication average, $\bar{x}$, and sample variance, s_{barx}.

   (e) Based on your answers to the previous parts of this question,

      i. For each of $N = 1$, $N = 5$, $N = 20$, and $N = 1,000$: Does the distribution of $\bar{x}^r$ look more like an exponential distribution or a normal distribution?

      ii. Is your estimate of $\bar{x}$ close to $E(x_i) = 1$ in each experiment? If not, why not?

      iii. Is your estimate of $s_{\bar{x}}$ close to $V(x_i) = 1$ in each experiment? If not, why not?

Empirical Application

Suggested Solutions

Emanuele, Matteo

October 2019

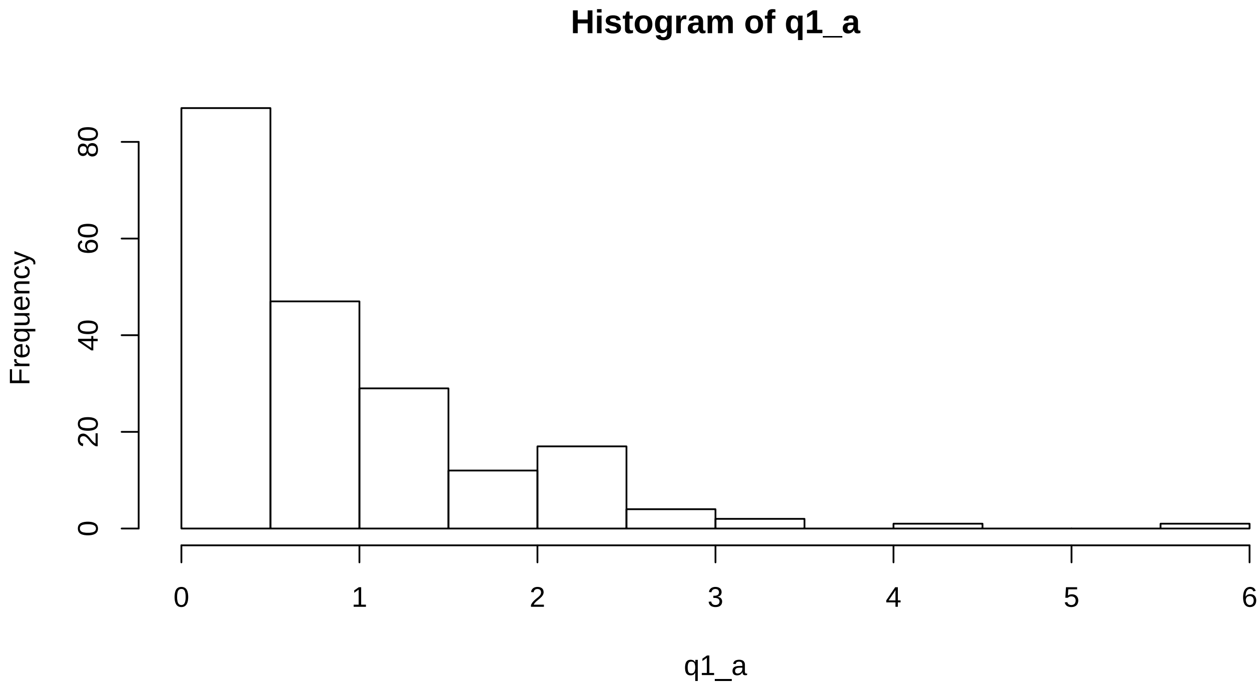
1.a

```r
# Do the replication
set.seed(12321) # This helps with replicability
q1_a <- replicate(n=200, rexp(1, rate=1))

# Store the mean
mean1_a <- mean(q1_a)

# Store the variance
var1_a <- var(q1_a)
```

```r
# Histogram and print mean and variance
hist(q1_a)
```



```r
print(paste0("Mean=",round(mean1_a, digits=4)))
```

```
## [1] "Mean=0.871"
```

```r
print(paste0("Variance=", round(var1_a, digits=4)))
```

```
## [1] "Variance=0.7114"
```

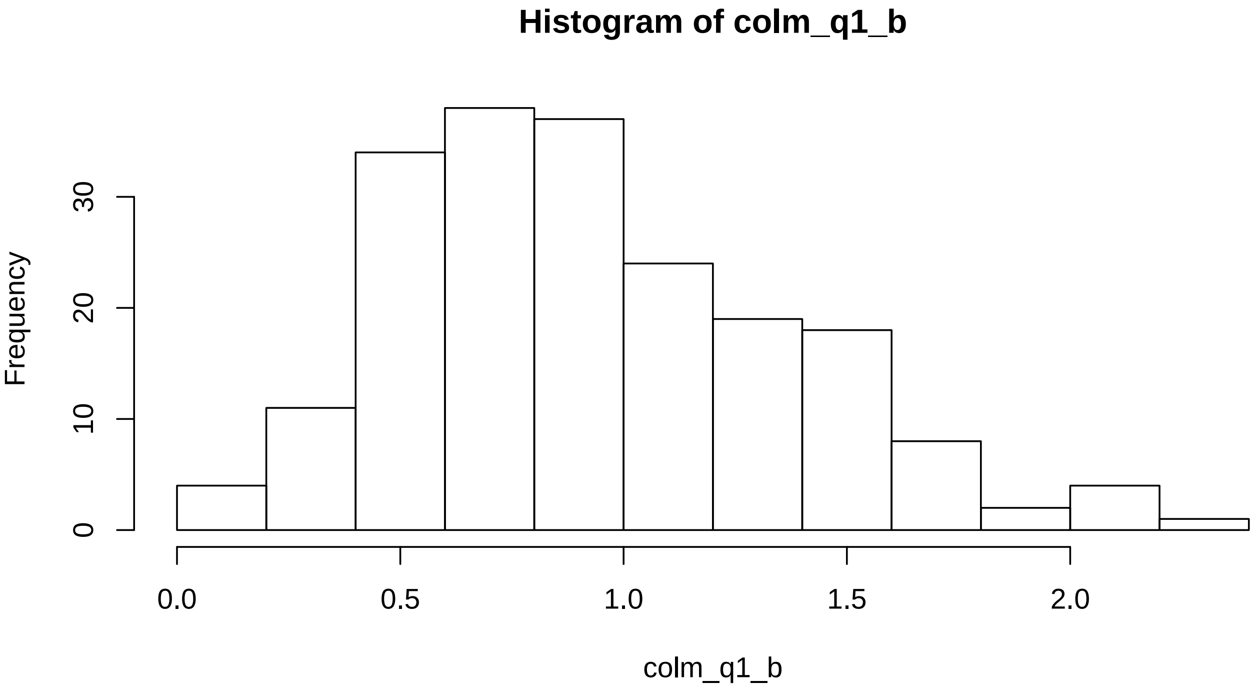
1.b

```r
# Do the replication
set.seed(12321)
q1_b <- replicate(n=200, rexp(5, rate=1))
# This time you also need the sample mean
colm_q1_b <- colMeans(q1_b)

# Store the mean
mean1_b <- mean(colm_q1_b)

# Store the variance
var1_b <- var(colm_q1_b)
```

```r
# Histogram and print mean and variance
hist(colm_q1_b)
```



```r
print(paste0("Mean=",round(mean1_b, digits=4)))
```

```
## [1] "Mean=0.9367"
```

```r
print(paste0("Variance=", round(var1_b, digits=4)))
```

```
## [1] "Variance=0.1871"
```

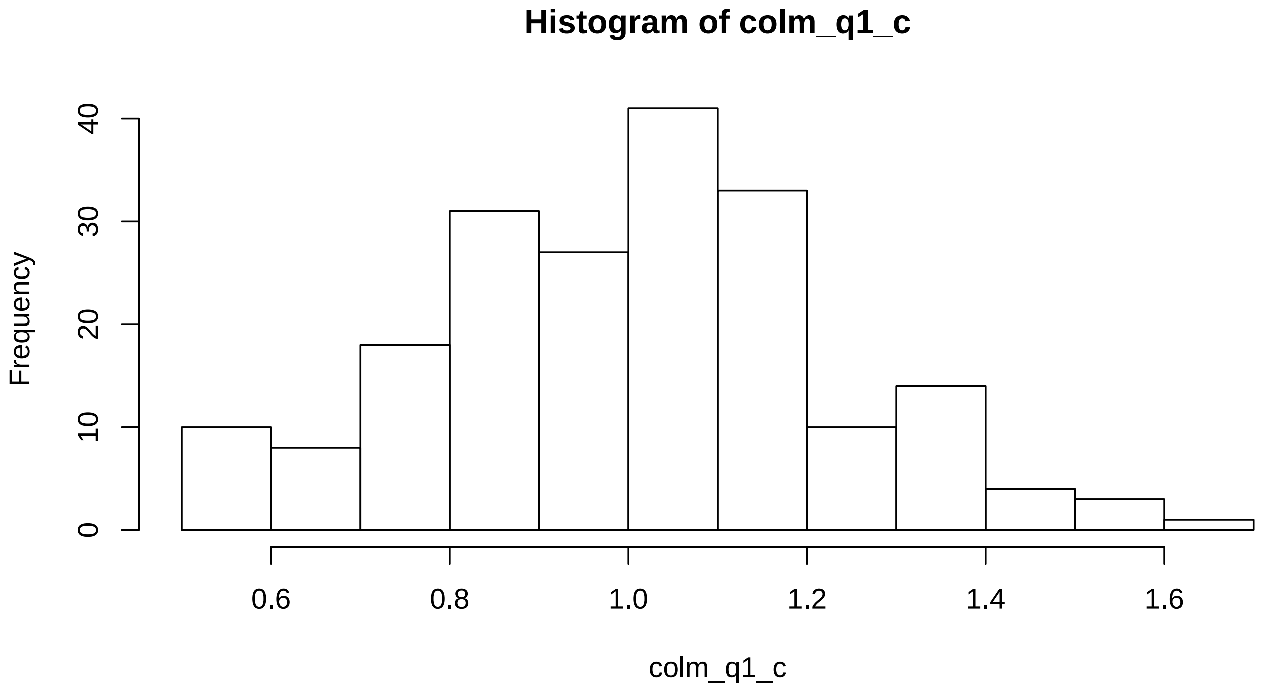
1.c

```r
# Do the replication
set.seed(12321)
q1_c <- replicate(n=200, rexp(20, rate=1))
colm_q1_c <- colMeans(q1_c)
```

```
# Store the mean
mean1_c <- mean(colm_q1_c)

# Store the variance
var1_c <- var(colm_q1_c)
```

```
# Histogram and print mean and variance
hist(colm_q1_c)
```



```
print(paste0("Mean=",round(mean1_c, digits=4)))
```

```
## [1] "Mean=1.0049"
```

```
print(paste0("Variance=", round(var1_c, digits=4)))
```

```
## [1] "Variance=0.0497"
```

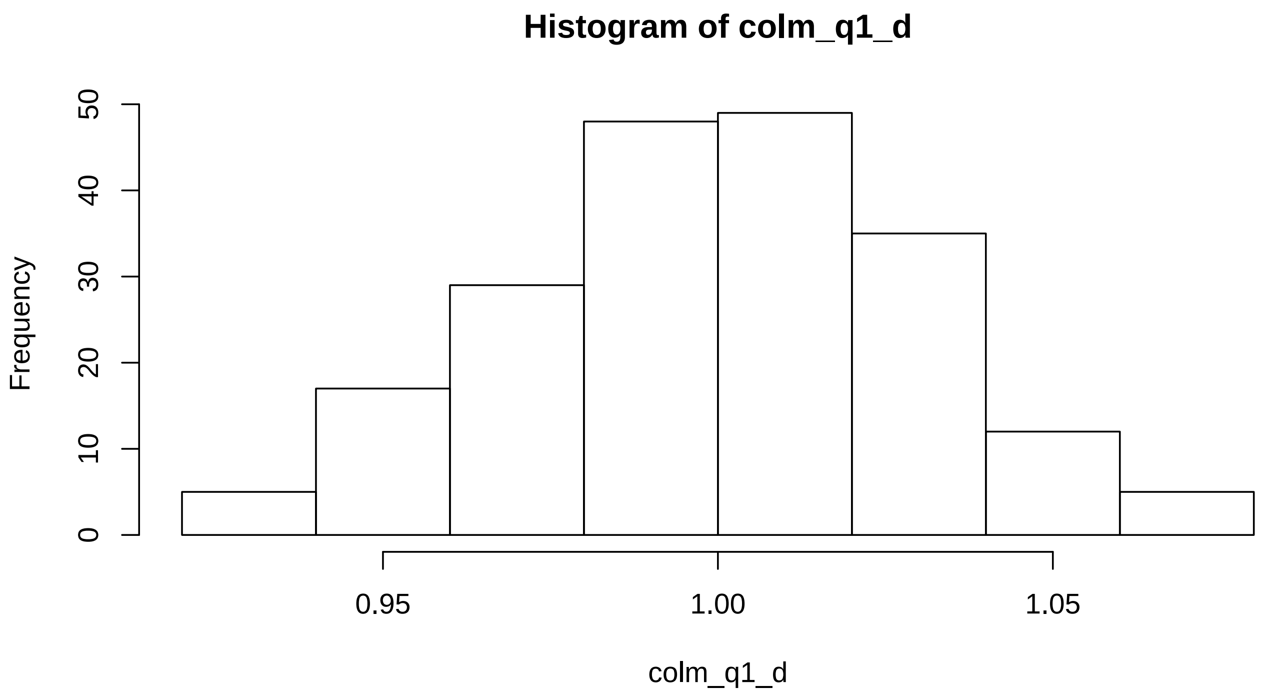
1.d

```
# Do the replication
set.seed(12321)
q1_d <- replicate(n=200, rexp(1000, rate=1))
colm_q1_d <- colMeans(q1_d)

# Store the mean
mean1_d <- mean(colm_q1_d)

# Store the variance
var1_d <- var(colm_q1_d)
```

```
# Histogram and print mean and variance
hist(colm_q1_d)
```



```
print(paste0("Mean=",round(mean1_d, digits=4)))
```

```
## [1] "Mean=0.9998"
```

```
print(paste0("Variance=", round(var1_d, digits=4)))
```

```
## [1] "Variance=9e-04"
```

1.e

Based on your answers to the previous parts of this question,

1. For each of $N = 1$, $N = 5$, $N = 20$, and $N = 1,000$: Does the distribution of $\bar{x}^r$ look more like an exponential distribution or a normal distribution? The answer is straightforward. This is consistent with which theoretical result? CLT
2. Is your estimate of $\bar{x}$ close to $E(x_i) = 1$ in each experiment? If not, why not? Yes, why? LLN
3. Is your estimate of $s_{\bar{x}}$ close to $V(x_i) = 1$ in each experiment? If not, why not? No, why? Again LLN