ASDS Statistics, YSU, Fall 2020

Lecture 03

Michael Poghosyan

19 Sep 2020

Contents

- BarPlot, PieChart, LineGraph, Frequency Polygon
- Empirical CDF

Last Lecture Recap

- Can you classify Variable by Types?

Last Lecture Recap

- Can you classify Variable by Types?
- Give the Definition of the Frequency and the Relative Frequency

Visualizing Frequency and Relative Frequency Tables

Now, having the Frequency or the Relative Frequency Tables, we can visualize the Dataset by using a BarPlot (BarChart), PieChart, Line Graph or a Frequency Polygon.

Frequency Tables, Example

Now, consider the *iris* dataset in **R**:

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Frequency Tables, Example

Now, consider the *iris* dataset in **R**:

```r
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Frequency Tables, Example, Cont'd

To get the *Species* Variable of the iris Dataset, we use

```
iris$Species
```

Frequency Tables, Example, Cont'd

To get the *Species* Variable of the iris Dataset, we use

```
iris$Species
```

And to calculate the Frequency of each of the Species, we use

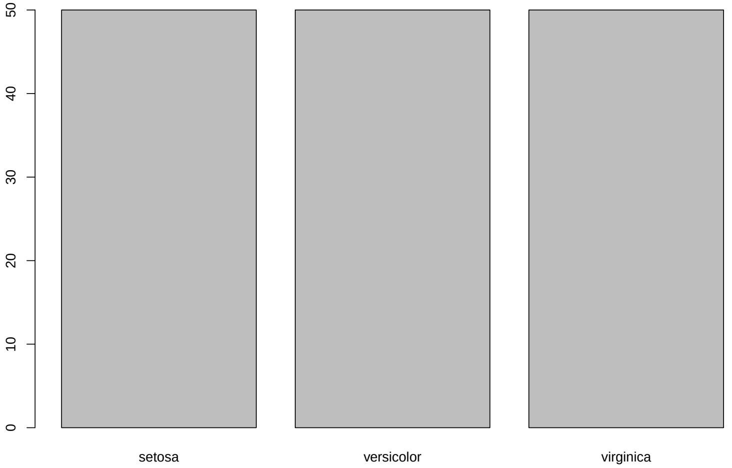
```
table(iris$Species)
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

BarPlot

Now, let us visualize our Frequency Table by using a BarPlot:

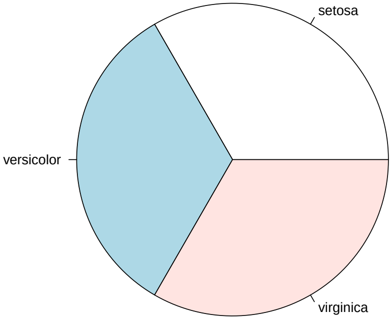
```
barplot(table(iris$Species))
```



PieChart

Also, we can visualize the same Frequency Table (or, in fact, the Relative Frequency Table) using a PieChart:

```
pie(table(iris$Species))
```



BarPlot

Another standard Dataset, *mtcars*, again about cars ⌣̈:

```
head(mtcars, 3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear c
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4
```

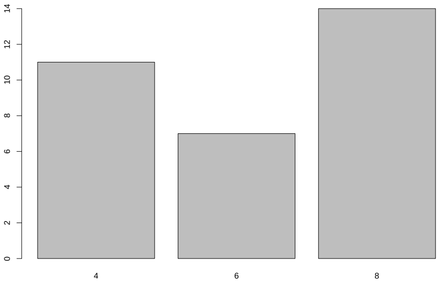
BarPlot

Another standard Dataset, *mtcars*, again about cars ☺:

```
head(mtcars, 3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear c
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4
```

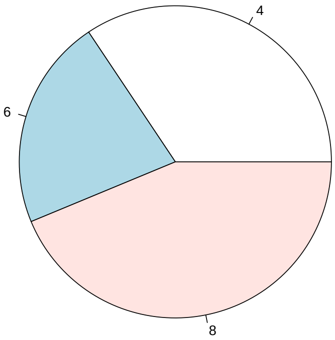
```
barplot(table(mtcars$cyl))
```



mtcars CYL with PieChart

The same, but with PieChart:

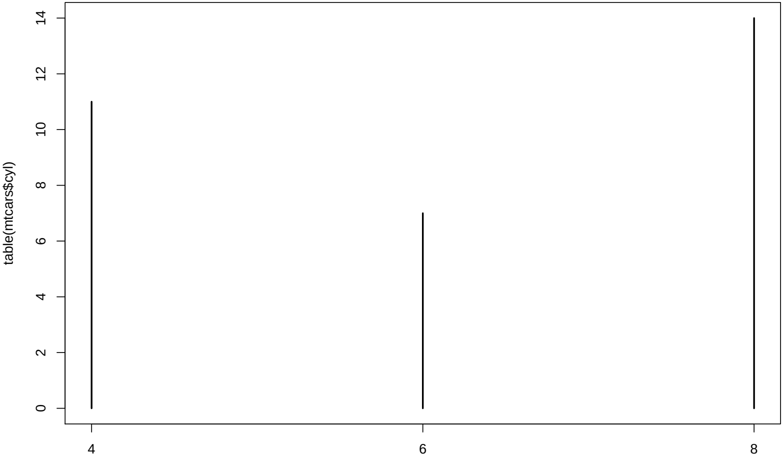
```
pie(table(mtcars$cyl))
```



LineGraph and Barplot

Now, with the Line Graph:

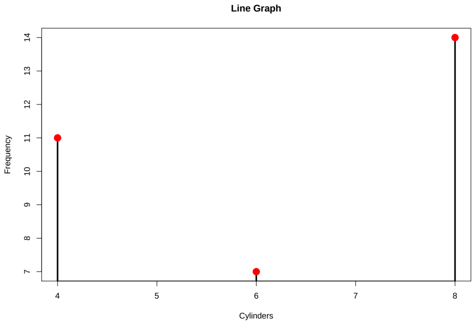
```
plot(table(mtcars$cyl))
```



LineGraph and Barplot

More sophisticated (titiz) version:

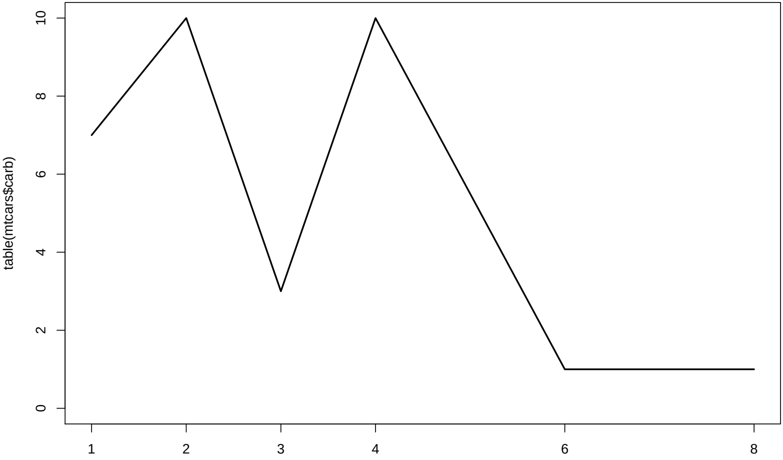
```
x <- mtcars$cyl; y <- as.data.frame(table(x))
a <- as.numeric(as.character(y$x)); b <- y$Freq
plot(a,b,type="h", lwd=3,  xlab = "Cylinders",
     ylab = "Frequency", main = "Line Graph")
points(a,b, pch=16, cex=2, col="red")
```



The Frequency Polygon

Again, same cars, but now the *carb* Variable Frequencies:

```
plot(table(mtcars$carb), type = "l")
```



Supplements

If our Dataset has more complex structure, say, we have categories, and categories can be separated by some groups, then we can use **Stacked** or **Grouped BarPlots** to visualize the Dataset.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights. And we assume *Height* is our r.v., and we have some observations from that r.v.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights. And we assume *Height* is our r.v., and we have some observations from that r.v.

From the Probability course, we know two complete characteristics of a Random Variable:

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights. And we assume *Height* is our r.v., and we have some observations from that r.v.

From the Probability course, we know two complete characteristics of a Random Variable: the **CDF and PD(M)F**.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights. And we assume *Height* is our r.v., and we have some observations from that r.v.

From the Probability course, we know two complete characteristics of a Random Variable: the **CDF and PD(M)F**. So to describe our Data Distribution, we can try to describe the CDF and/or PD(M)F behind the Data.

Empirical CDF

First let's estimate the CDF. We will estimate CDF by the Empirical CDF:

Definition: The **Empirical Distribution Function, ECDF** or the **Cumulative Histogram** $ecdf(x)$ of our data $x_1, ..., x_n$ is defined by

$$ecdf(x) = \frac{\text{number of elements in our dataset} \leq x}{\text{the total number of elements in our dataset}} =$$

$$= \frac{\text{number of elements in our dataset} \leq x}{n}, \qquad \forall x \in \mathbb{R}.$$

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

- Analytical Part - on the board

To do the graphical part, we

- Sort our Dataset from the lowest to the largest values

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

- Analytical Part - on the board

To do the graphical part, we

- Sort our Dataset from the lowest to the largest values
- Plot the Data points on the OX axis

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

- Analytical Part - on the board

To do the graphical part, we

- Sort our Dataset from the lowest to the largest values
- Plot the Data points on the OX axis
- ECDF is 0 for values of x less that the smallest Datapoint, and is 1 for values of x bigger than the largest Datapoint

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

- Analytical Part - on the board

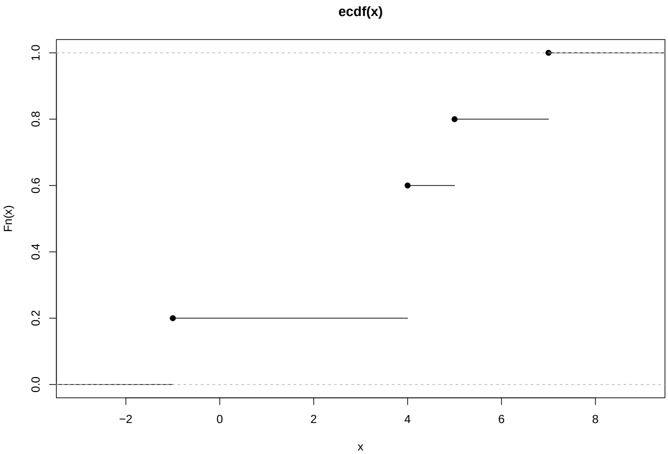
To do the graphical part, we

- Sort our Dataset from the lowest to the largest values
- Plot the Data points on the OX axis
- ECDF is 0 for values of x less that the smallest Datapoint, and is 1 for values of x bigger than the largest Datapoint
- For each Data point, calculate the Relative Frequency of that Datapoint (the number of times it occurs in our Dataset over the total number of Datapoints). At that Datapoint, do a Jump of the size of the Relative Frequency, and draw a horizontal line up to the next Datapoint.

Example

Now, using **R**:

```
x <-c(-1, 4, 7, 5, 4)
f <- ecdf(x)
plot(f)
```



Note: It is easy to see that the ECDF satisfies all properties of a CDF.

Note: It is easy to see that the ECDF satisfies all properties of a CDF.

Note: It is easy to see that the ECDF for a Dataset

$$-1, 4, 7, 5, 4$$

coincides with the CDF of a r.v.

X	-1	4	5	7
$\mathbb{P}(X = x)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, ..., x_n$ comes from the Distribution with the CDF $F(x)$,

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, ..., x_n$ comes from the Distribution with the CDF $F(x)$,and if we will denote by $F_n(x)$ the ECDF constructed for $x_1, ..., x_n$, then

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, ..., x_n$ comes from the Distribution with the CDF $F(x)$,and if we will denote by $F_n(x)$ the ECDF constructed for $x_1, ..., x_n$, then

$$F_n(x) \to F(x) \quad \text{uniformly on } \mathbb{R}.$$

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, ..., x_n$ comes from the Distribution with the CDF $F(x)$,and if we will denote by $F_n(x)$ the ECDF constructed for $x_1, ..., x_n$, then

$$F_n(x) \to F(x) \quad \text{uniformly on } \mathbb{R}.$$

This Theorem says that if you will have enough datapoints from a Distribution, you can approximate the unknown CDF of your Distribution pretty well by using the ECDF.

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, ..., x_n$ comes from the Distribution with the CDF $F(x)$,and if we will denote by $F_n(x)$ the ECDF constructed for $x_1, ..., x_n$, then

$$F_n(x) \to F(x) \quad \text{uniformly on } \mathbb{R}.$$

This Theorem says that if you will have enough datapoints from a Distribution, you can approximate the unknown CDF of your Distribution pretty well by using the ECDF.

Above, we need to be more precise about in which sense the convergence holds.

Glivenko-Cantelli Theorem

In fact, the following Theorem Holds:

Theorem (Glivenko, Cantelli): If $X_1, ..., X_n$ are IID r.v.s from the Distribution with the CDF $F(x)$, and $F_n(x)$ is the ECDF constructed by using $X_1, ..., X_n$, then

$$\sup_x |F_n(x) - F(x)| \to 0 \qquad a.s.$$

Estimation of the CDF through ECDF

Let us check this theorem using **R**:

```
plot(pnorm, lwd = 3, col = 'red', xlim = c(-3,3),
     ylim = c(0,1), ylab = "ecdf and CDF")
n <- 30 ; x <- rnorm(n) #Taking a sample of size n from N(0,1)
f <- ecdf(x) #f will be the ECDF of our data x
par(new = TRUE) #this is to keep the previous graph
plot(f, xlim = c(-3,3), ylim = c(0,1), ylab = "ecdf and CDF")
```