Statistical Methods and Data Analysis I
Lecture 14: CDF. Covariance and Correlation.

Oleg Goldshmidt
oleg.goldshmidt@post.idc.ac.il

Arison School of Business
Interdisciplinary Center (IDC)
Herzliya, Israel

May 07, 2019

Cumulative Distribution Function (CDF)

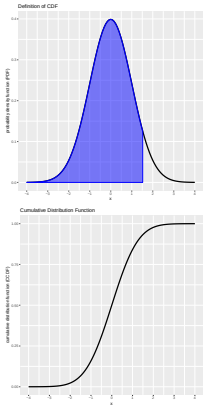
Definition: CDF of real-valued random variable X

$$F_X(x) = P(X \leq x)$$

Main properties:

1. $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to \infty} F_X(x) = 1$
2. *monotonically increases* (more precisely, *does not decrease*)
3. $P(a < X \leq b) = F_X(b) - F_X(a)$
4. if the *probability density function (PDF)* of X is $f_X(x)$ then

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx, \qquad f_x(x) = \frac{dF_X}{dx}$$



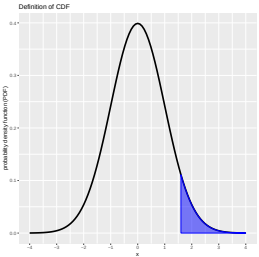
Tail Distribution Function (CCDF)

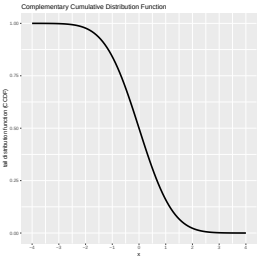
Definition: Complementary CDF (Tail Distribution)

$$\overline{F}_X(x) = P(X > x) = 1 - F_X(x)$$

Useful for:

- hypothesis testing,
- reliability or survival analysis.





Quantile Function

Definition: Quantile Function

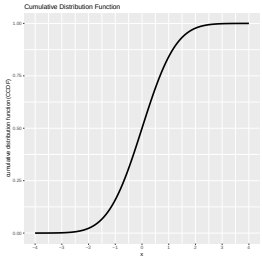
Quantile function $Q_X(p)$ is the threshold value x of a real-valued variable X such that the CDF

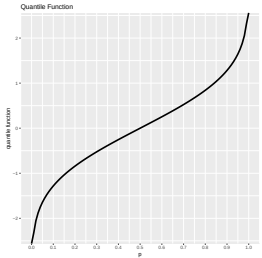
$$F_X(x) = P(X \leq x) = p$$

Interpretation: the *minimum* value of x for which $p \leq F_X(x)$.

Usually: (specifically when $F_X(x)$ is *continuous and monotonically increasing*)

$$Q_X = F_X^{-1}$$





CDF, Quantiles, and Friends in R

- Does it all look frightening? If it does, recall you have R...
- You are already familiar with `quantile()`...
- You also saw CDF in action, e.g., `pnorm()` — we used it precisely to compute the probability that the normally distributed variable is less than its argument...

```
dnorm(x, mean = 0, sd = 1) - probability density function (PDF)
pnorm(q, mean = 0, sd = 1) - cumulative distribution function (CDF)
qnorm(p, mean = 0, sd = 1) - quantile function
rnorm(n, mean = 0, sd = 1) - random generator

...

dunif(x, min = 0, max = 1) - probability density function (PDF)
punif(q, min = 0, max = 1) - cumulative distribution function (CDF)
qunif(p, min = 0, max = 1) - quantile function
runif(n, min = 0, max = 1) - random generator
```

Definition of Covariance

Definition: covariance of 2 random variables

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$$

Covariance is a measure of whether X and Y are likely to increase or decrease together.

- Simplifying the formula:

$$\begin{aligned} Cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - E[X]Y - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Covariance for Discrete Random Variables

- If the probability of $X = x_i$, $Y = y_i$ is p_i for $i = 1, \ldots, n$, then

$$Cov[X, Y] = \sum_{i=1}^{n} p_i (x_i - \bar{x}) (y_i - \bar{y})$$

- For equal probabilities $p_i = 1/n$:

$$Cov[X, Y] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})$$

- *Sample covariance*:

$$Cov[X, Y] = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})$$

Covariance: Properties

- Covariance with itself:

$$Cov[X,X] = E[(X - E[X])(X - E[X])] = E\left[(X - E[X])^2\right] = Var[X] = \sigma_X^2$$

- Symmetry:

$$Cov[Y,X] = Cov[X,Y]$$

- Linear combinations:

$$\begin{aligned} Cov[aX, bY] &= abCov[X,Y] \\ Cov[X + a, Y + b] &= Cov[X,Y] \\ Cov[aX + bU, cY + dV] &= acCov[X,Y] + bcCov[U,Y] + \\ &\quad adCov[X,V] + bdCov[U,V] \end{aligned}$$

Covariance of Independent Variables

- For independent variables

$$E[XY] = E[X]\,E[Y]$$

- From our formula for covariance:

$$Cov[X,Y] = E[XY] - E[X]\,E[Y] = 0$$

- The inverse is not necessarily true!
  - For instance, let X be *uniformly distributed* in $[-1,1]$, and $Y = X^2$
  - Clearly, X and Y are *not independent*!
  - However,

$$\begin{aligned} Cov[X,Y] &= Cov\left[X,X^2\right] = E\left[X \cdot X^2\right] - E[X] \cdot E\left[X^2\right] \\ &= E\left[X^3\right] - E[X] \cdot E\left[X^2\right] = 0 - 0 \cdot E\left[X^2\right] = 0 \end{aligned}$$

Variance and Standard Deviation of a Linear Combination

$$\begin{aligned} Var[aX+bY] &= E\left[(a(X-E[X])+b(Y-E[Y]))^2\right] \\ &= E\left[a^2(X-E[X])^2+2ab(X-E[X])(Y-E[Y])+b^2(Y-E[Y])^2\right] \\ &= a^2E\left[(X-E[X])^2\right]+2abE\left[(X-E[X])(Y-E[Y])\right]+b^2E\left[(Y-E[Y])^2\right] \\ &= a^2Var[X]+2abCov[X,Y]+b^2Var[Y] \end{aligned}$$

$$Var[X+Y]=Var[X]+2Cov[X,Y]+Var[Y]$$

$$Var[X-Y]=Var[X]-2Cov[X,Y]+Var[Y]$$

$$SD[aX+bY]=\sigma_{aX+bY}=\sqrt{a^2\sigma_X^2+2abCov[X,Y]+b^2\sigma_Y^2}$$

$$SD[X+Y]=\sigma_{X+Y}=\sqrt{\sigma_X^2+2Cov[X,Y]+\sigma_Y^2}$$

$$SD[X-Y]=\sigma_{X-Y}=\sqrt{\sigma_X^2-2Cov[X,Y]+\sigma_Y^2}$$

Correlation

Definition of correlation (coefficient)

Correlation is covariance normalized by the product of standard deviations:

$$Corr\,[X,Y] = \rho_{XY} = \frac{Cov\,[X,Y]}{\sigma_X \sigma_Y}$$

- *Sample correlation coefficient*:

$$\rho_{XY} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n-1}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Covariance and Correlation in R

- Afraid of the formulae? Of course, R gives you tools...

```
> library(tidyverse)
> mpg
# A tibble: 234 x 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     compact
 2 audi         a4           1.8  1999     4 manual(m5) f        21    29 p     compact
 3 audi         a4           2    2008     4 manual(m6) f        20    31 p     compact
 4 audi         a4           2    2008     4 auto(av)   f        21    30 p     compact
 5 audi         a4           2.8  1999     6 auto(l5)   f        16    26 p     compact
 6 audi         a4           2.8  1999     6 manual(m5) f        18    26 p     compact
 7 audi         a4           3.1  2008     6 auto(av)   f        18    27 p     compact
 8 audi         a4 quattro   1.8  1999     4 manual(m5) 4        18    26 p     compact
 9 audi         a4 quattro   1.8  1999     4 auto(l5)   4        16    25 p     compact
10 audi         a4 quattro   2    2008     4 manual(m6) 4        20    28 p     compact
#  with 224 more rows
> cov(mpg$cty,mpg$hwy)
[1] 24.22543
> cor(mpg$cty,mpg$hwy)
[1] 0.9559159
> cor(mpg$displ,mpg$hwy)
[1] -0.76602
```