**Variance of random variables**

In addition to knowing what we would expect Y to be on average, we may want to know something about the “variability” of Y from μ = E(Y). In the transaction time example, we want to numerically quantify the expected deviation from the mean for Y. We already did this in terms of E[|Y-30|] = E[|Y-μ|]. There is another way more often used for doing this in terms of the squared deviation, E[(Y-μ)2].

Let Y be a random variable with PDF f(y) and mean μ. The variance of Y is

σ2 = E[(Y-μ)2] = 

when Y is discrete, and

σ2 = E[(Y-μ)2] = 

when Y is continuous. The positive square root of the variance, σ, is called the standard deviation of Y.

Notes:

* The variance is the expected average squared deviation of Y from μ.
* This an example of using g(Y) = (Y-μ)2 when finding expected values earlier.
* E[(Y-μ)2] is equivalently written as E{[Y-E(Y)]2} so that there is an expectation within an expectation. Once E(Y) is found, it is a constant value as has been shown in the last section.
* Common notation that is often used here includes: Var(Y) =  = σ2 = E[(Y-μ)2]. The Var() part is just a nice way to replace the E[ ] part.
* The subscript Y on σ2 is often helpful to use when there is more than one random variable under consideration. Thus,  denotes the variance of Y.
* The reason for considering the positive square root of σ2 is so that we can use the original units of Y.
* σ2 ≥ 0 and σ ≥ 0 for all possible values of Y

The variance of a random variable Y can also be found with

σ2 = Var(Y) = E(Y2) – μ2 = E(Y2) – [E(Y)]2

proof:

E[(Y-μ)2] = E(Y2 - 2Yμ + μ2)

= E(Y2) - 2μE(Y) + E(μ2)

= E(Y2) - 2μ2 + μ2

= E(Y2) - μ2

Example: Let’s play Plinko! (plinko.xlsx, plinko.R)

Let X be a random variable denoting the amount won for one chip. Find the variance and standard deviation for X.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Drop Plinko chip in column above:** | | | | |
|  | **$100** | **$500** | **$1,000** | **$0** | **$10,000** |
| **x** | **f(x)** | **f(x)** | **f(x)** | **f(x)** | **f(x)** |
| **$100** | 0.1667 | 0.1339 | 0.0822 | 0.0359 | 0.0176 |
| **$500** | 0.3571 | 0.3080 | 0.2204 | 0.1287 | 0.0882 |
| **$1,000** | 0.2976 | 0.2991 | 0.2862 | 0.2545 | 0.2353 |
| **$0** | 0.1429 | 0.1920 | 0.2796 | 0.3713 | 0.4118 |
| **$10,000** | 0.0357 | 0.0670 | 0.1316 | 0.2096 | 0.2471 |
|  |  |  |  |  |  |
| **** | $850.00 | $1,136.16 | $1,720.39 | $2,418.26 | $2,751.76 |
| **** | $1,799.31 | $2,404.79 | $3,246.57 | $3,923.92 | $4,170.28 |
|  |  |  |  |  |  |
| **** | -$2,748.61 | -$3,673.42 | -$4,772.75 | -$5,429.57 | -$5,588.79 |
| **** | $4,448.61 | $5,945.74 | $8,213.54 | $10,266.10 | $11,092.32 |
|  |  |  |  |  |  |
| **** | -$4,547.92 | -$6,078.21 | -$8,019.33 | -$9,353.49 | -$9,759.06 |
| **** | $6,247.92 | $8,350.54 | $11,460.12 | $14,190.01 | $15,262.59 |

When dropping the chip above the $100 reservoir/slot,

σ2 = E[(X-μ)2] = 

= (100-850)2×0.1667 + (500-850)2×0.3571 +

(1000-850)2×0.2976 + (0-850)2×0.1429 +

(10000-850)2×0.0357

= 3,237,500 dollars2

where T = {100, 500, 1000, 0, 10000}.

To put this into units of just dollars, we take the positive square root to find = $1,799.31.

What does this value really mean?

“Rule of thumb” for the # of standard deviations all possible observations (data) lies from its mean: 2 or 3. This is an ad-hoc interpretation of Chebyshev’s Rule and the Empirical Rule (discussed later). This rule of thumb is discussed now to help you understand standard deviations.

When someone drops one chip from the top of the Plinko board, I would generally expect the amount of money they will win to be between μ - 2σ and μ + 2σ. A more conservative expected amount would be μ - 3σ and μ + 3σ.

Examine the application of the rule of thumb in the table and how it makes sense when the chip is dropped above the $100 spot.

R:

> x <- c(100, 500, 1000, 0, 10000)

> fx10000 <- c(0.0176, 0.0882, 0.2353, 0.4118, 0.2471)

> mu <- sum(x\*fx10000)

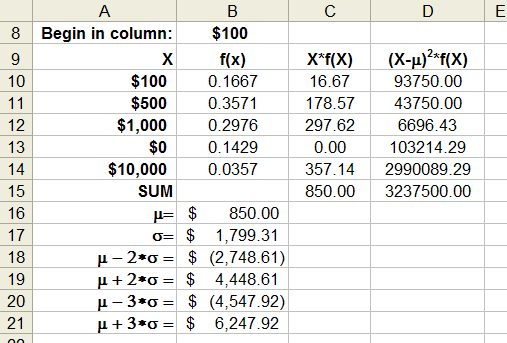
> sigmasq <- sum((x-mu)^2 \* fx10000)

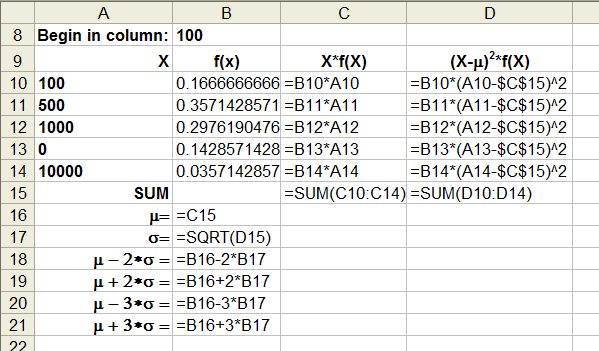
> data.frame(mu, sigmasq, sigma = sqrt(sigmasq))

mu sigmasq sigma

1 2752.16 17393141 4170.509

Excel (negative values are represented by a “( )” instead of a negative sign):

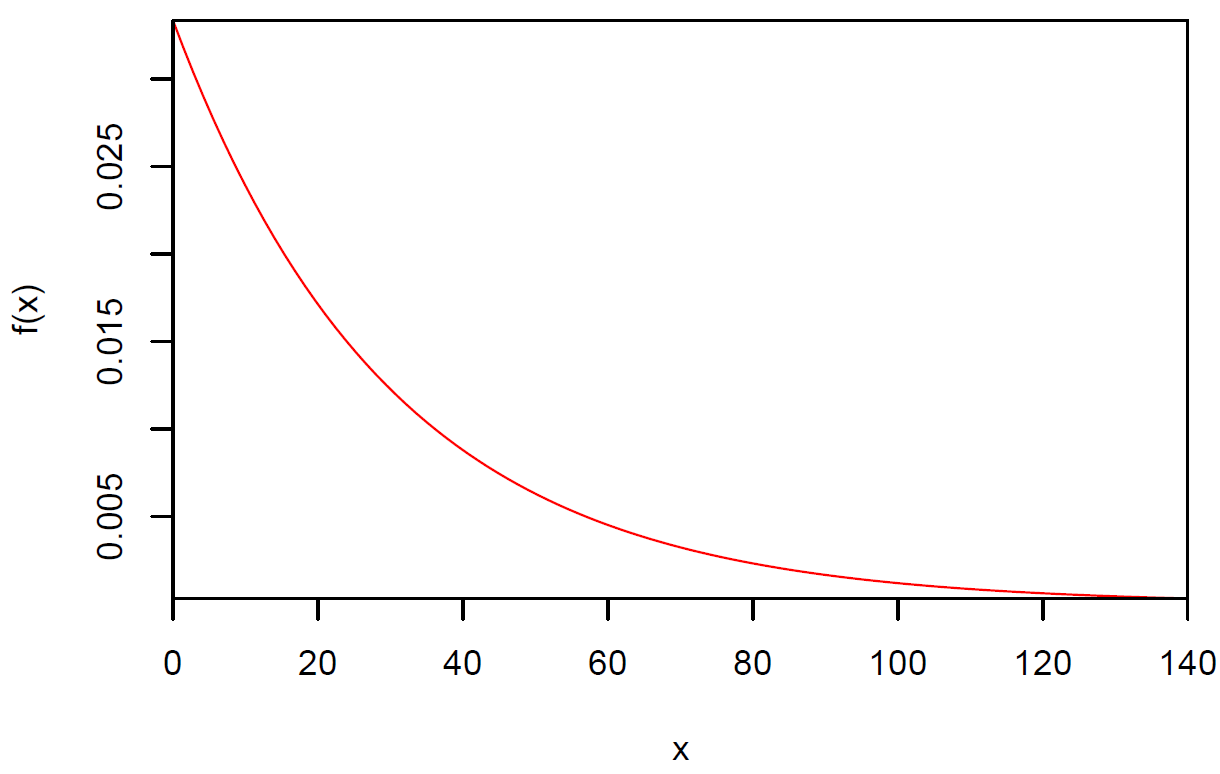




Example: Transaction time (TransExpect.R, TransExpect.ipynb)

The number of seconds between transactions (e.g., purchases) on a website can be represented by a PDF. Let X be a random variables representing the seconds. Suppose the PDF for X is





Find the variance and standard deviation for the time between transactions.

σ2 = Var(X) = E[(X-μ)2] =  =  where μ= 30.

Instead of doing this integral, it is often a little easier to work with σ2 = E(X2) - μ2. We found previously that E(X2) = 1800. Thus, 2 = 1800 – 302 = 900!

R:

> # E[(X-mu)^2]

> pdf.var <- function(x, mu) {

(x-mu)^2 \* 1/30\*exp(-x/30)

}

> integrate(f = pdf.var, lower = 0, upper = Inf, mu = mu)

900 with absolute error < 0.022

> integrate(f = pdf.var, lower = 0, upper = Inf, mu = mu,

rel.tol = 0.00000001)

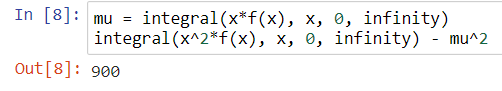
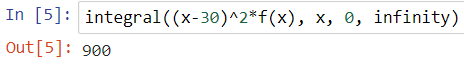
900 with absolute error < 1.1e-06

> integrate(f = pdf.xsq, lower = 0, upper = Inf, rel.tol =

0.00000001)$value - mu^2

[1] 900

Sage:



Notice that σ =  = 30. Generally for random variables, the mean and the standard deviation will not be the same!!! This just happens to be a special PDF called the “Exponential PDF” where this will always occur. Using the rule of thumb,

μ - 2σ = 30 - 2×30 = -30 and μ+2σ = 30 + 2×30 = 90

and

μ - 3σ = 30 - 3×30 = -60 and μ+3σ = 30 + 3×30 = 120

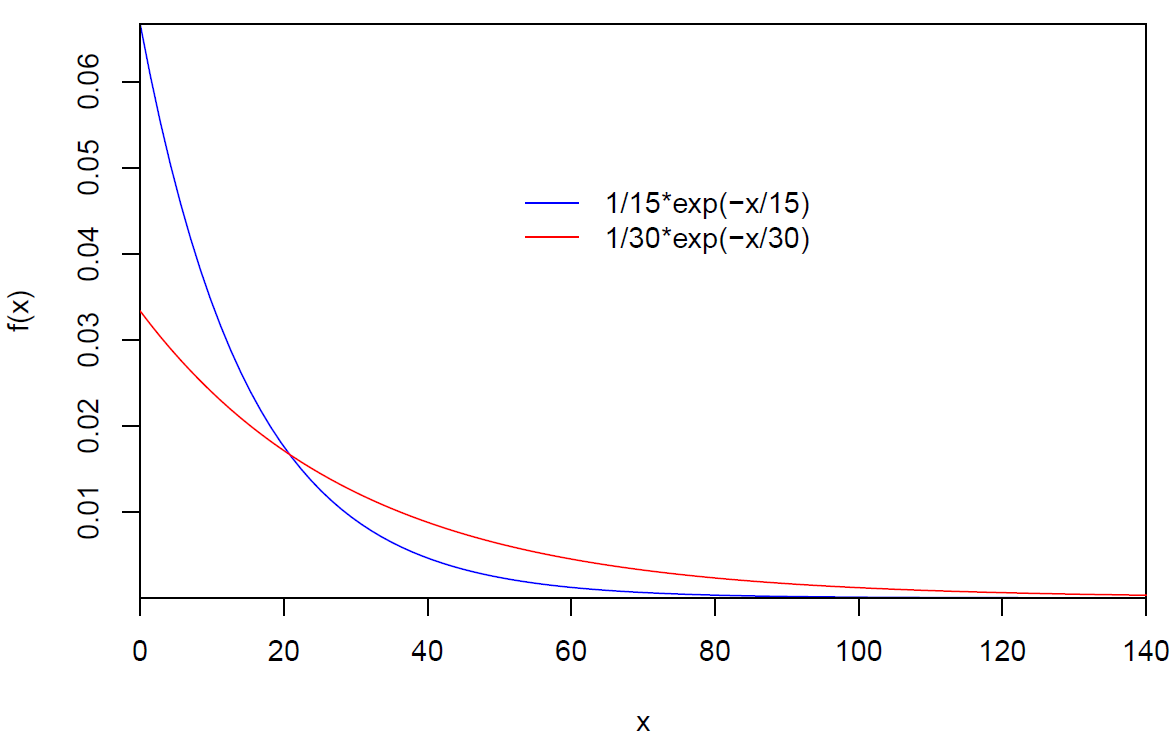
Thus, one would expect all of the times to be between -30 and 90. A more conservative range would be -60 and 120. Of course, the negative values do not make sense for this problem. However, examine where the upper bound values fall on the PDF plot. Make sure you understand why these upper values make sense in terms of the plot!

Compare σ = 30 to E(|X-μ|) = 22.07 found earlier.

Suppose the PDF was changed to



In this case, one can show that E(X) = 15, Var(X) = σ2 = 225, and σ = 15. The PDF is shown below.



Notice the σ = 30 plot has a PDF more spread out than the σ = 15 plot. This is because the standard deviation (and variance) is larger for the σ = 30 plot!

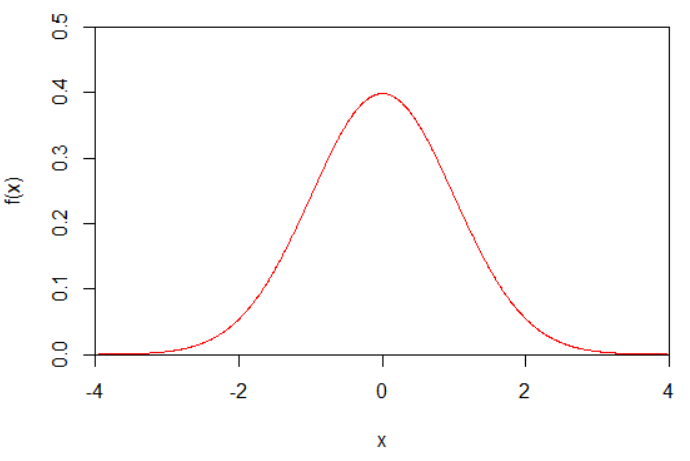
Relate the two PDFs and their associated μ and σ values to the time between transactions.

Other common g(Y) functions used with E[g(Y)]:

* The skewness for a random variable Y is defined to be

E[(Y-μ)3] / E[(Y-μ)2]3/2 = E[(Y-μ)3] / σ3

This quantity measures the lack of symmetry in a PDF. For example, the PDFs on the previous page are very skewed (non-symmetric). The PDF shown below is symmetric.



R code to plot the above will be discussed later in the course.

Note that E[(Y-μ)2]3/2 ≠ E[(Y-μ)3] ≠ E[(Y-μ)]3.

* The kurtosis of a random variable Y is defined to be

E[(Y-μ)4] / E[(Y-μ)2]2 = E[(Y-μ)4] / σ4

This quantity measures the amount of peakedness or flatness of a PDF. For example, the red PDF below has a higher peak than the blue PDF which is more flat.



Variance for functions of random variables

Let Y be a random variable with PDF f(y). The variance of the random variable g(Y) is



when Y is discrete, and



when Y is continuous. We will discuss easier ways to express these variances shortly.