**Chebyshev’s Rule**

From earlier in the course:

“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.

The purpose of this section to give a more precise definition and explanation of this powerful rule.

Chebyshev’s Rule: The probability that any random variable Y will assume a value within k standard deviations of the mean is at least 1-1/k2. Thus,

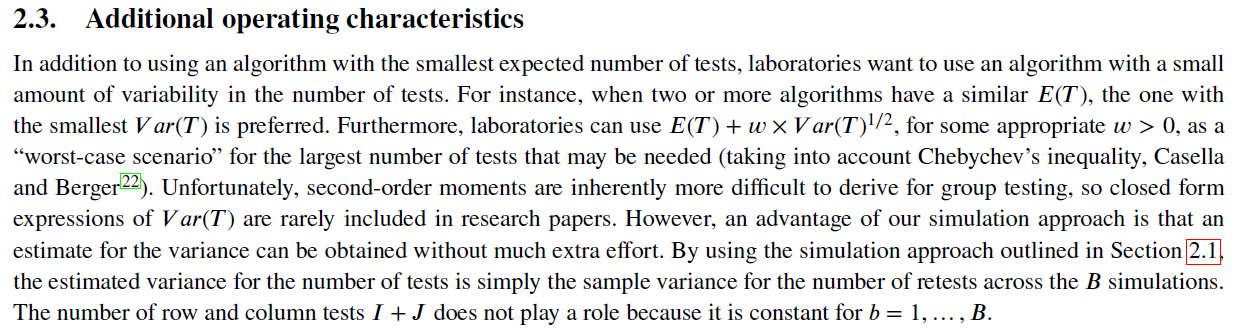
P(μ - kσ < Y < μ + kσ) ≥ 1 - 

Notes:

* There is no mention of the PDF for Y here!!! Thus, this holds for ANY PDF!!!
* Below are some values of k and lower bounds for the probability which are often of interest.

| **k** | **1 -** |
| --- | --- |
| 1 | 0 |
| 2 | 0.75 |
| 3 |  |
| 4 | 0.94 |
| 5 | 0.96 |

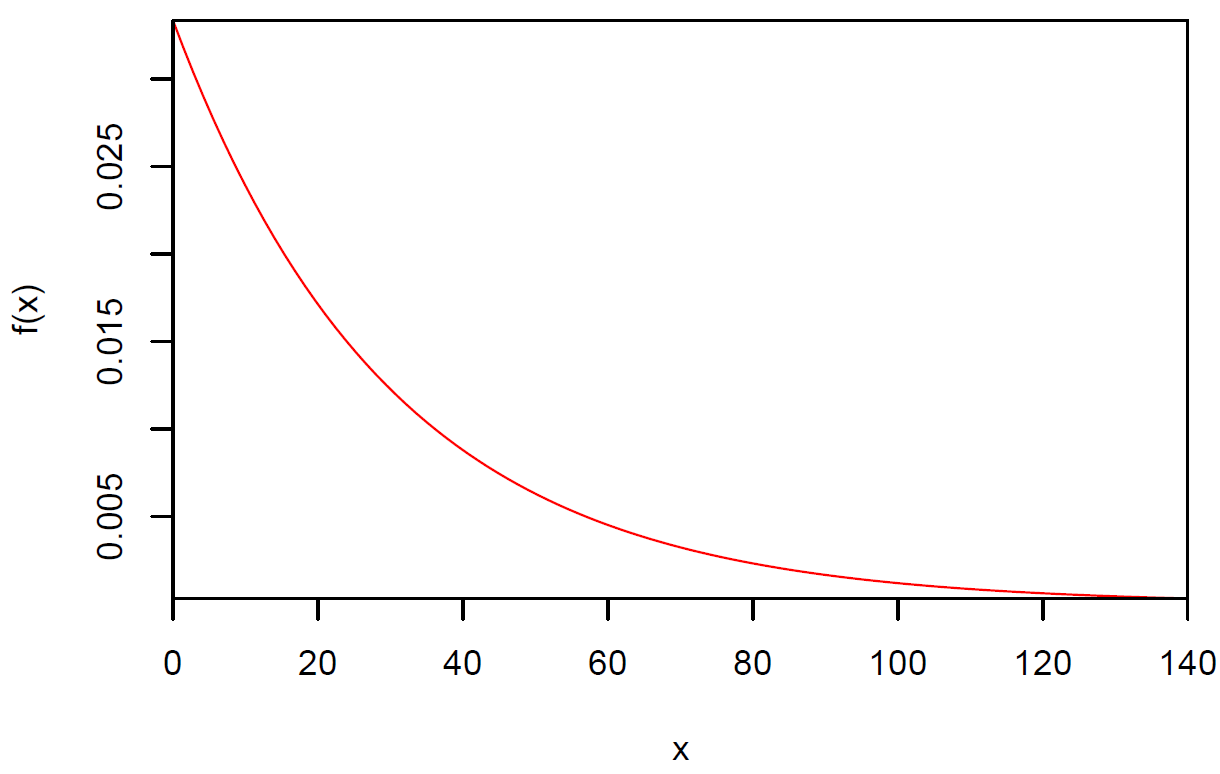
* While the probabilities for 2 or 3 may be lower than you would expect based on the rule of thumb, remember that the above table gives lower bounds only. For a given PDF, the probabilities could be higher!
* A proof is available on p. 122 of Casella and Berger (2002).
* Side note: In a paper that I recently revised for a journal, I used Chebyshev’s Rule!



Example: Transaction time (TransExpect.R)

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 P(μ - kσ < X < μ + kσ). Note that we could insert the values of μ and σ into the probability. I will wait until the end to do that.



Since μ - kσ could be less than 0 (outside of the possible values of X), we should use the following expression for the probability:

If μ - kσ > 0 then

P(μ - kσ < X < μ + kσ) = -e-(μ+kσ)/30 + e-(μ-kσ)/30

If μ - kσ < 0 then

P(0 < X < μ+kσ) = -e-(μ+kσ)/30 + e-0/30 = -e-(μ+kσ)/30 + 1

Earlier, we found that μ = 30 and σ = 30. Then

P(30-30k < X < 30+30k) can be found for various values of k. Note that for k ≥ 1, 30-30k ≤ 0, so we can find P(0 < X < 30+30k) instead.

> k <- 1:5

> cheby.prob <- round(1 - 1/k^2,4)

> pdf <- function(x) {

1/30\*exp(-x/30)

}

> # P(X > 30)

> k1 <- integrate(f = pdf, lower = 0, upper = 30 + 30)$value

> k2 <- integrate(f = pdf, lower = 0, upper = 30 +

30\*2)$value

> k3 <- integrate(f = pdf, lower = 0, upper = 30 +

30\*3)$value

> k4 <- integrate(f = pdf, lower = 0, upper = 30 +

30\*4)$value

> k5 <- integrate(f = pdf, lower = 0, upper = 30 +

30\*5)$value

> data.frame(k, cheby.prob, prob = round(c(k1, k2, k3, k4,

k5),4))

k cheby.prob prob

1 1 0.0000 0.8647

2 2 0.7500 0.9502

3 3 0.8889 0.9817

4 4 0.9375 0.9933

5 5 0.9600 0.9975

Would you expect most times between transactions to be within 2 (or 3) standard deviations from the mean?