Chapter 4: Probability Distributions

Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

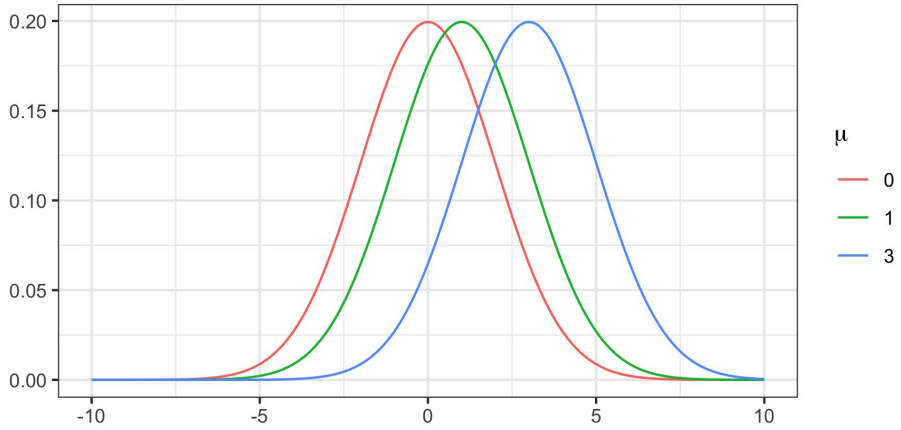
Icahn
School of
Medicine at
Mount
Sinai

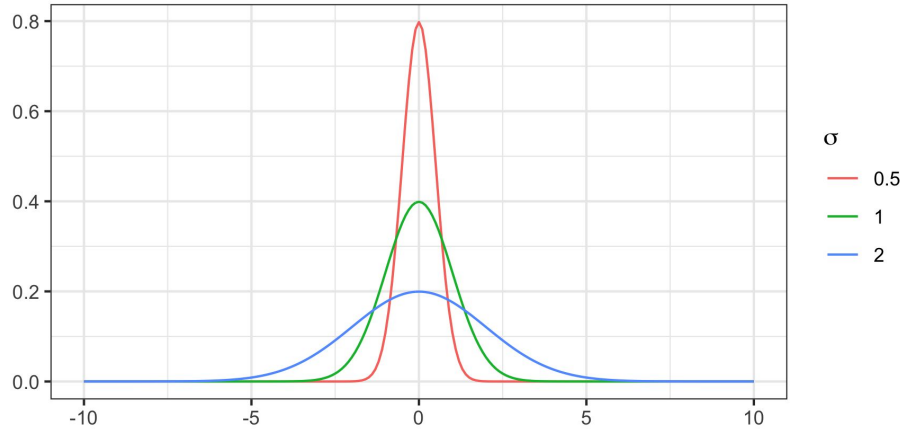
How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 15 min, max) while following along in the notes

Question 4.1

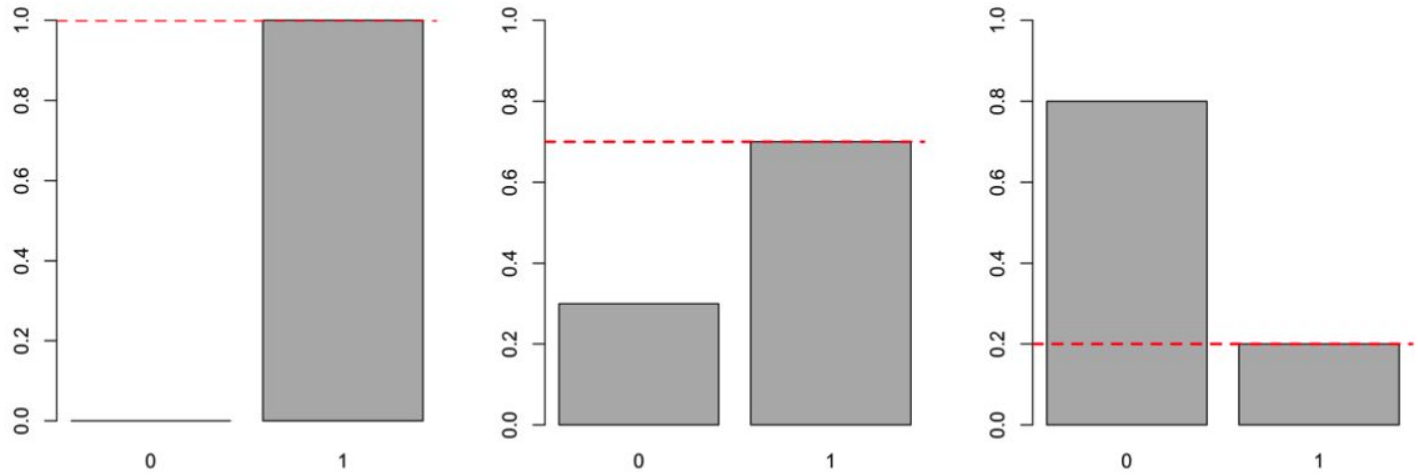
List 5 random variables from medicine or biology that should follow normal distributions.





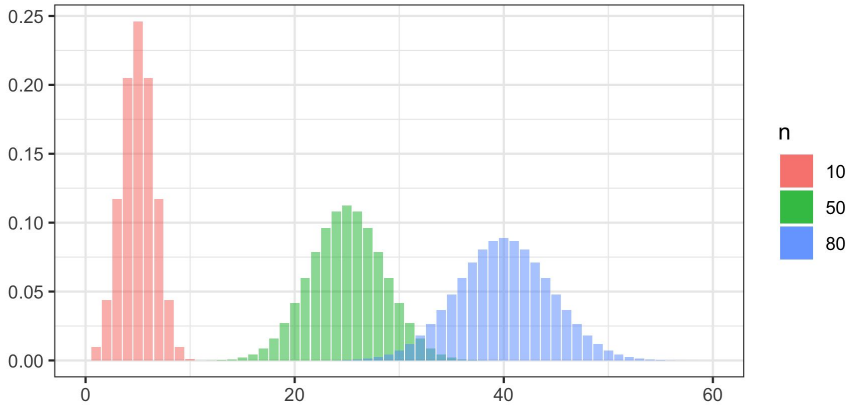
Question 4.2

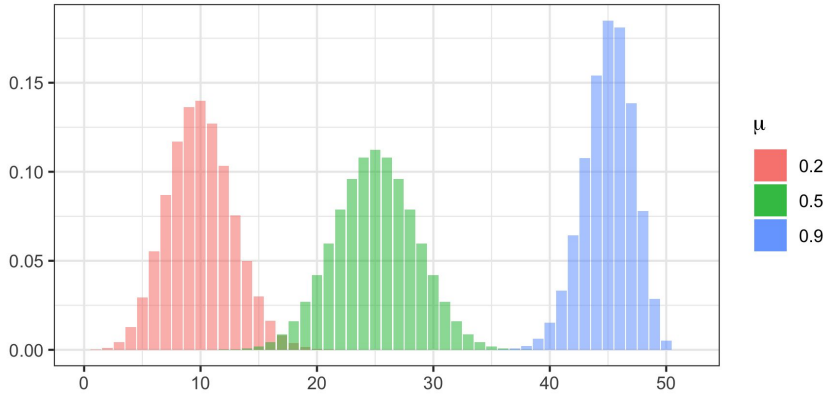
List 5 random variables from medicine or biology that should follow Bernoulli distributions.



Question 4.3

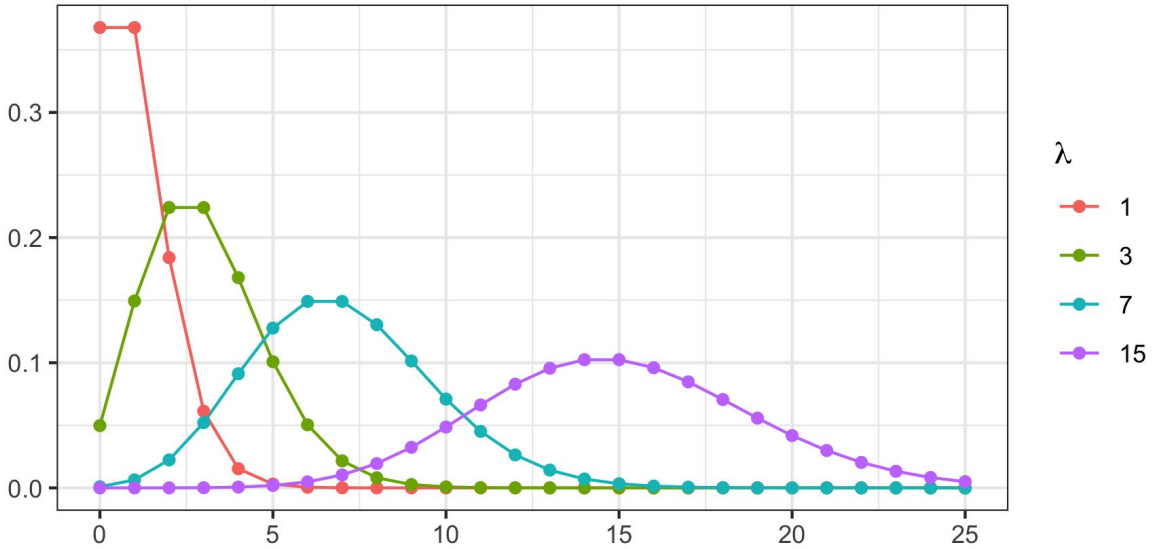
List 5 random variables from medicine or biology that should follow binomial distributions.





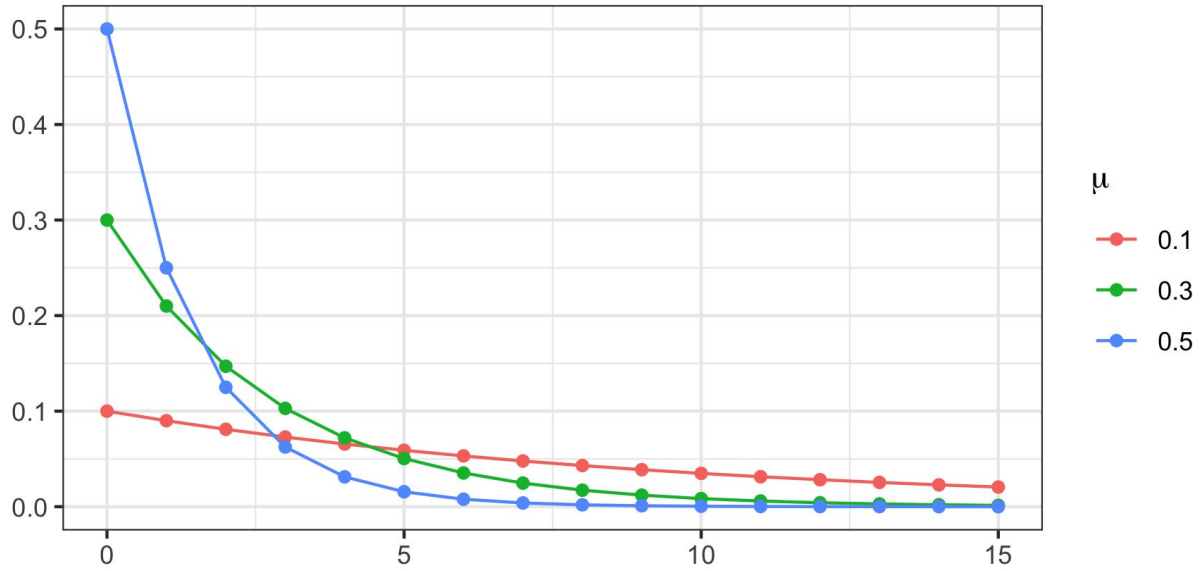
Question 4.4

List 5 random variables from medicine or biology that should follow Poisson distributions.



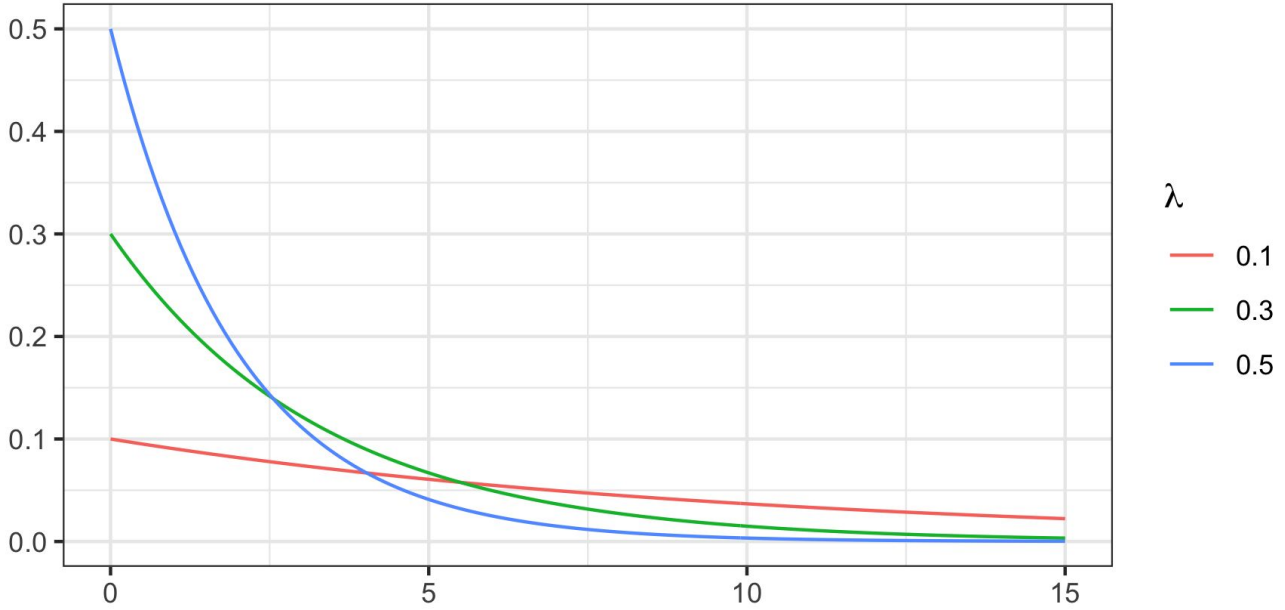
Question 4.5

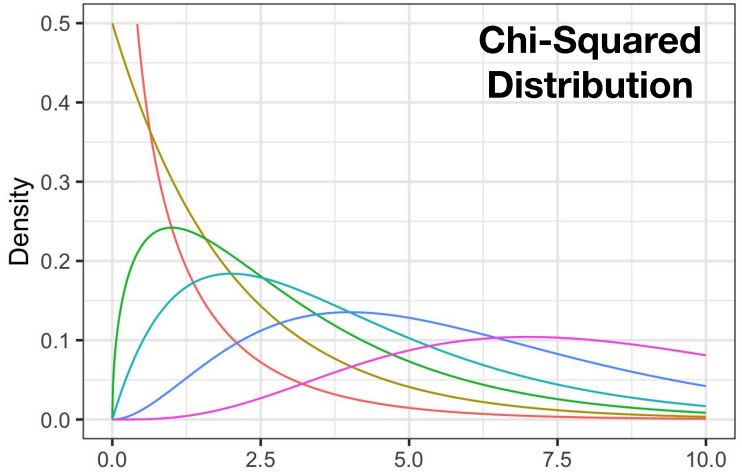
List 5 random variables from medicine or biology that should follow geometric distributions.

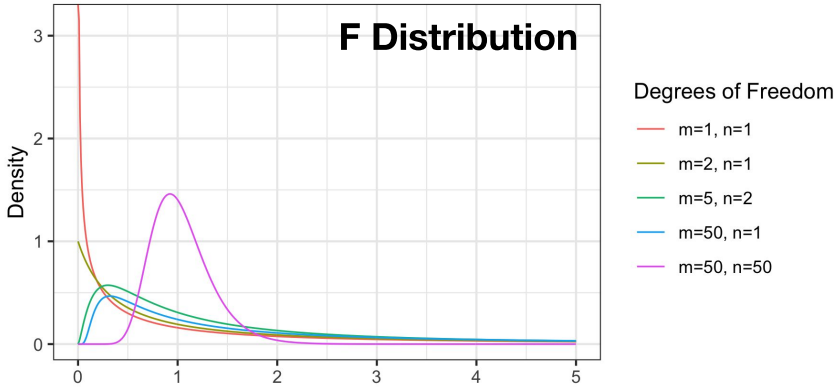


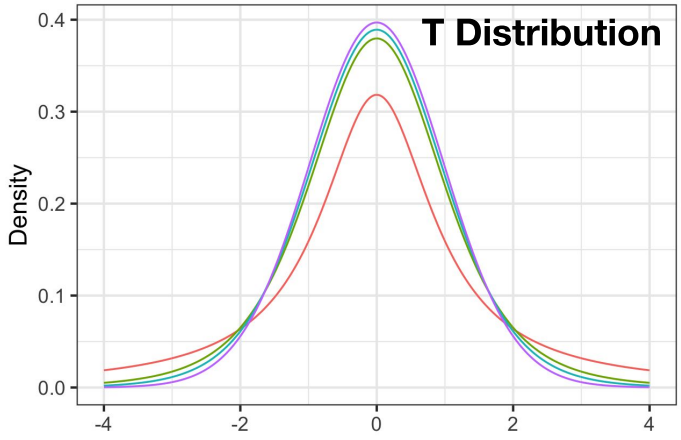
Question 4.6

List 5 random variables from medicine or biology that should follow exponential distributions.




Chi-Squared Distribution
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.0
2.5
5.0
7.5
10.0
Degrees of Freedom
1
2
3
4
6
9


F Distribution
Density
0
1
2
3
0
1
2
3
4
5
Degrees of Freedom
m=1, n=1
m=2, n=1
m=5, n=2
m=50, n=1
m=50, n=50


T Distribution
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Degrees of Freedom
1
5
10
50

Question 4.7

For each of the following experimental conditions, which distribution (from those listed above) provides the best model for how the data $x^{(1)}, \ldots, x^{(n)}$ are generated?

(a) You are observing several patients' skin in a clinical study to see how long it takes them to develop a rash. You take a picture each day. Let $x^{(i)}$ be the number of days of *no rash* before the rash occurs.

Patient ID (i)	$x^{(i)}$
1	4
2	1
3	0
4	2
5	2
6	4
7	3
8	1
9	0
10	1

(b) Same situation as above except that instead of taking a picture each day, the patient texts you at the moment he/she observes a rash. The data look like this, where $x^{(i)}$ is the time (in days) at which patient i develops a rash:

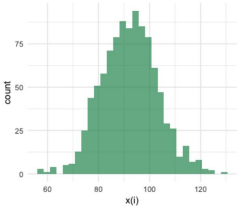
Patient ID (i)	$x^{(i)}$
1	2.25
2	3.43
3	0.68
4	0.04
5	3.78
6	5.65
7	2.88
8	3.88
9	2.83

(c) Imagine you are Ladislaus Bortkiewicz, and you are modeling the number of persons killed by mule or horse kicks in the Prussian army per year. You have data from the late 1800s over the course of 20 years. Let $x^{(i)}$ be the number of people killed in year i.

Year (i)	$x^{(i)}$	Year (i)	$x^{(i)}$
1	8	11	9
2	10	12	7
3	5	13	10
4	3	14	12
5	10	15	8
6	8	16	7
7	7	17	8
8	2	18	8
9	6	19	10
10	11	20	7

(d) Every year, 10 scientists go to the same geographic area (same Lyme prevalence) and they each collect 40 ticks. They test each tick for Lyme disease and record the number of ticks that have Lyme. Let $x^{(i)}$ be the number of ticks with Lyme in the ith scientist's bunch.

Scientist ID (i)	$x^{(i)}$
1	8
2	9
3	14
4	15
5	12
6	7
7	6
8	8
9	8
10	14



(e) You have waist circumference data on 1045 men aged 70 and above (see Dey's 2002 paper in the Journal of the American Geriatric Society). It looks like this: