BIOMED SCI 552:

STATISTICAL THINKING

LECTURE 3: PROBABILITY PART 2

QUESTIONS FROM TUESDAY?

A PROBABILITY GAME

- Guess the number of plastic disks I have in this cup

LET'S DO SOME SAMPLING

- Let's draw five of the disks out of the bag, and I'll let you all revise your guesses

TWO DIFFERENT STATISTICAL APPROACHES

- $\frac{1}{N} \times \frac{1}{N-1} \times \frac{1}{N-2} \times \frac{1}{N-3} \times \frac{1}{N-4} \times 120$
- We want the minimum value of *N*, as this is the *most likely* answer given our data
- That's the maximum observed value we see
- But…
- That number will *at best* be right, and in all other cases, always be an underestimate

ANOTHER WAY

- $N = Max(Observed) + \frac{Max+k}{k}$, where k is the number of observations (i.e. 5)

- How on earth did we get this?

ANOTHER WAY

- $N = Max(Observed) + \frac{Max+k}{k}$, where k is the number of observations (i.e. 5)

- How on earth did we get this?
- $\frac{Max+k}{k}$ is approximately the average gap between the draws, and we're assuming (if everything is fair), that there is reasonably an average sized gap between the highest one we drew and the highest possible number

WHERE DID THIS COME FROM?

- Like many statistical examples, this one comes from WW2
- "How did this bit of statistics come about?" tends to be either biology or from a war – both circumstances with lots of uncertainty and incomplete information
- In this case, trying to estimate the number of tanks produced by the Germans

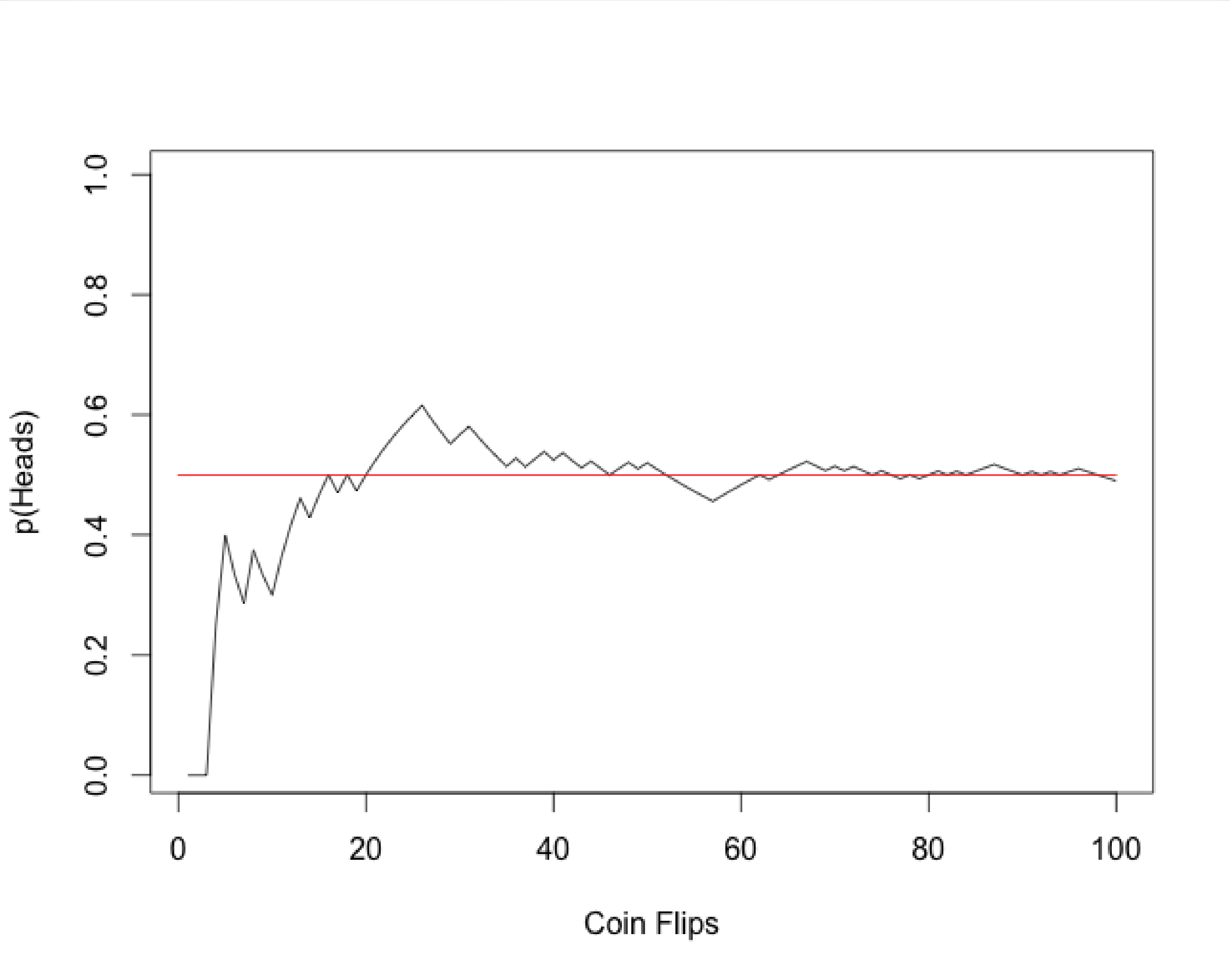
Date	Estimated Monthly Production		Monthly Production Speer Ministry
	Serial Number Estimate	Munitions Record 10 Aug. 42	
June, 1940	169	1000	122
June, 1941	244	1550	271
August, 1942	327	1550	342

ARE WE THINKING STOCHASTICALLY?

LET'S DO THIS A FEW MORE TIMES…

THE LAW OF LARGE NUMBERS

- As sample size grows, the mean of a stochastic process approaches the mean of the whole population
- Formally:
- $\lim_{n \to \infty} P\left(\left|\overline{X}_N - \mu\right| > \varepsilon\right) = 0$
- Where $\overline{X}_N$ is the sample average, μ is the expected value (aka the population average), and e is some threshold
- Some people in finance use “The Law of Large Numbers” to mean something completely different, and they are *wrong*


p(Heads)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Coin Flips

WHAT ARE WE ASSUMING?

SERIAL NUMBERS ARE SEQUENTIAL

- There's no attempt to obfuscate serial numbers
- This is not hard to do – there are functions called hashes that are heavily used in cryptography to take a known input and turn it into an encrypted output

WolfsbergPanther+1: e1a3
WolfsbergPanther+2: e0e3
WolfsbergPanther+273: 9d1e

WHAT ABOUT TIME?

WHAT ABOUT TIME?

- This is an issue that comes up *a lot* in biomedicine, and is one of the reasons survival analysis is such a big deal
- Obviously, a tank that's been fighting since 1939 is much more likely to show up in your sample of destroyed tanks than one built in 1943, which is in turn much more likely than one built in 1945
- Similarly, a worker who has been working in a factory for 10 years is far more likely to have had an occupation-related injury than a new hire
  - Or are they?

TYPES OF VARIABLES

- Continuous vs. Discrete
- Nominal vs. Ordinal

DISCRETE VARIABLES

- These are variables that can take a finite number of values
- Examples?

- For the most part these are what we've been working with so far, because they are easy to think about in terms of paths to events, etc.
- The field of math that concerns itself with these is aptly named Discrete Math

NOMINAL AND ORDINAL VARIABLES

- Nominal variables have no specific ordered value
  - Species: Oak, Ash, Aspen, etc.
  - Genotypes
- Ordinal variables have a clear order
  - Very bad, bad, neutral, good, very good
  - A, B, C, D, F
  - Low, Middle, High Income

CONTINUOUS VARIABLES

- Continuous variables are those that have potentially infinite values
- Examples?

- There are a lot of these in biomedicine
- There are also a lot of variables that are somewhat ambiguous
    - Examples?

STATISTICAL VARIATION

- We've talked a lot about statistical variation, but we can now start to think about quantifying it – a process known as estimation
- Estimation is a process of taking data and estimating an unknown quantity from it

Statistic:	**Parameter:**
Mean is $\bar{x}$	Mean is μ
SD is s	SD is σ

- We use *probability distributions* to quantify and model variation

PROBABILITY DISTRIBUTIONS

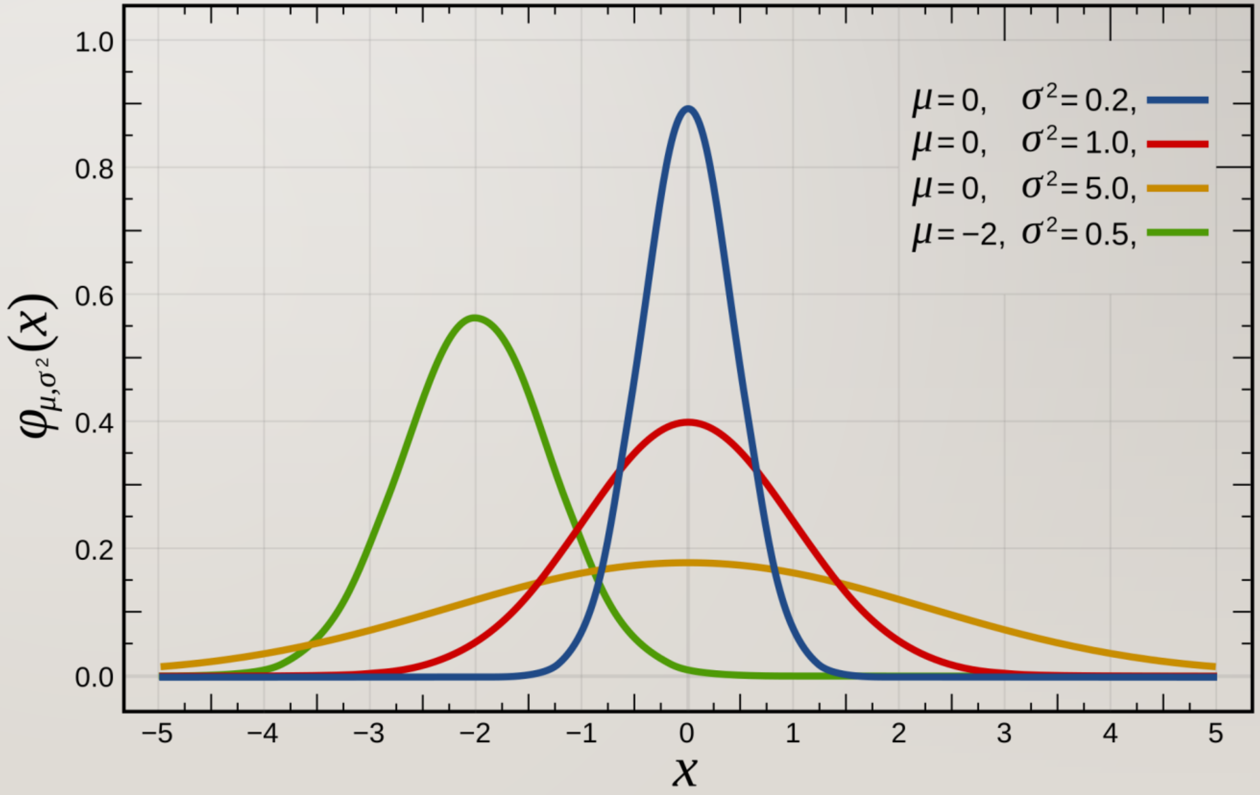
- Probability Distributions are *a model*
  - We are asserting the world works according to a particular description
  - We can assess how well that assertion works, but it's still an assertion
  - They can be represented by a list of possible outcomes (as we have been doing) or an equation
  - The former is used (as we have been) for discrete variables, especially in small numbers
  - The latter is used for continuous variables

PROBABILITY DISTRIBUTIONS

- Why are these useful?
  - They are helpful for describing the variability in a population, and visualizing it
  - We can start asking questions *about* the distribution, as well as it's relationship with our data

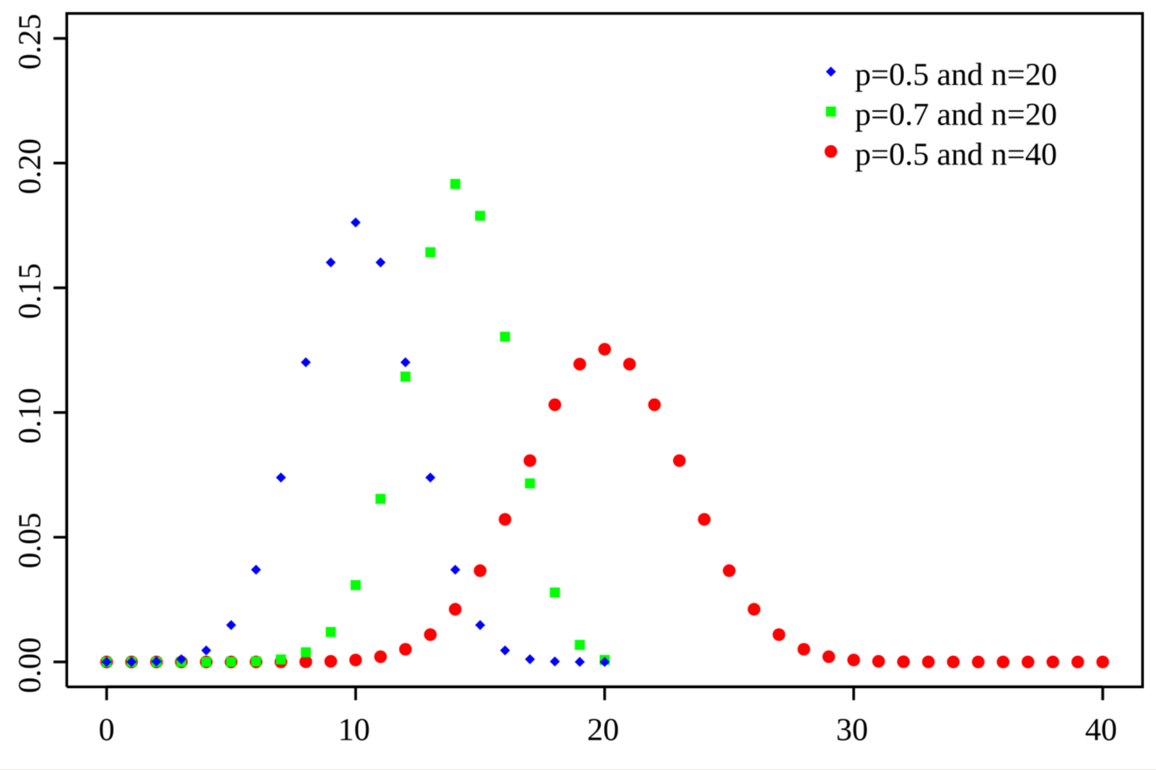
NORMAL DISTRIBUTION

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



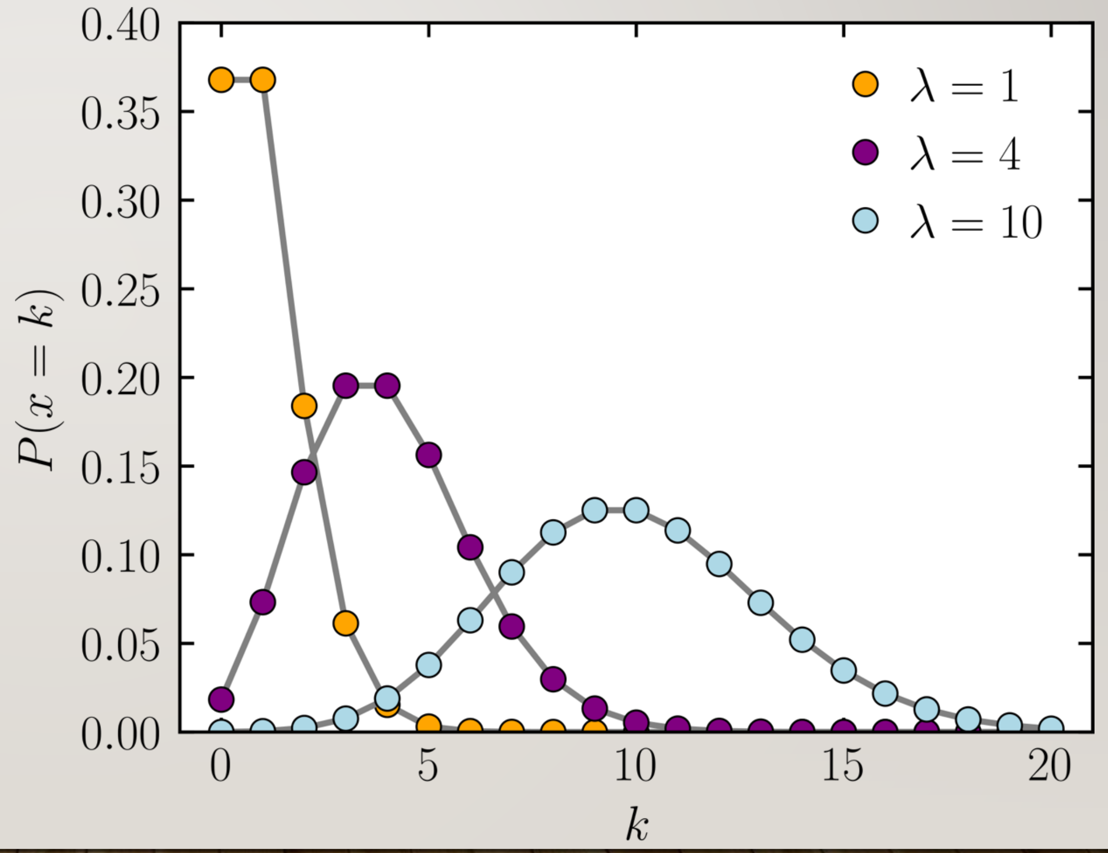
BINOMIAL DISTRIBUTION

$$f(s) = \frac{n!}{s!(n-s)!} p^s (1-p)^{n-s}$$



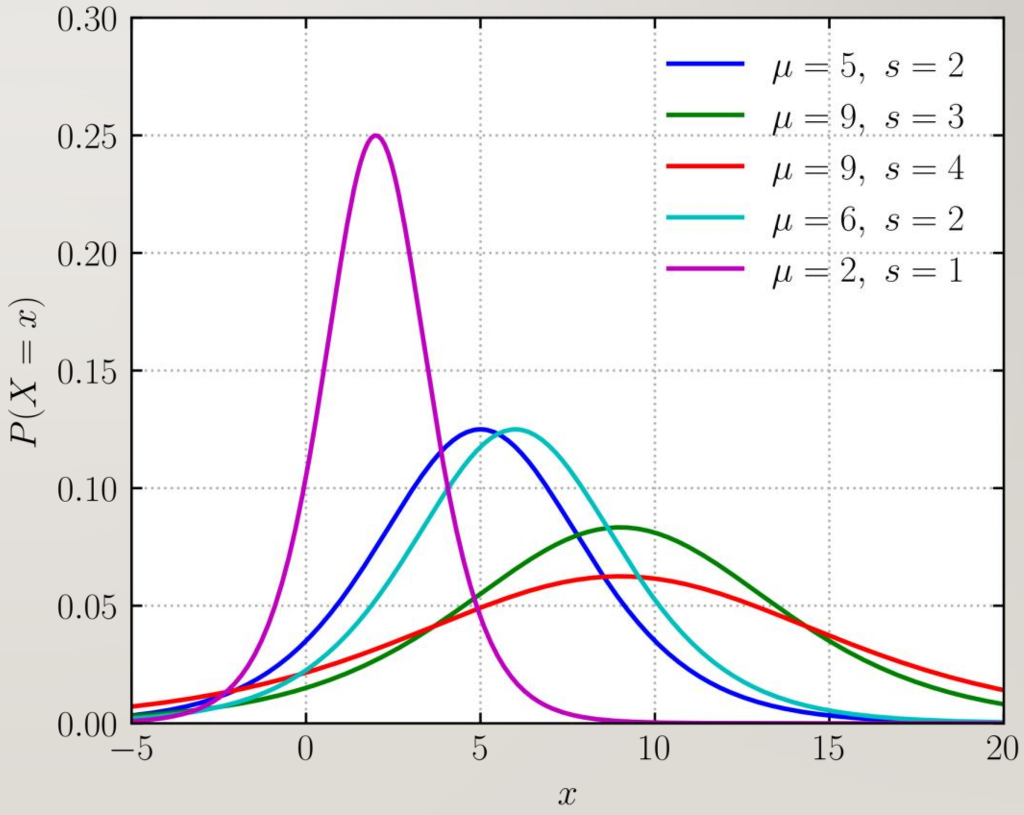
POISSON DISTRIBUTION

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



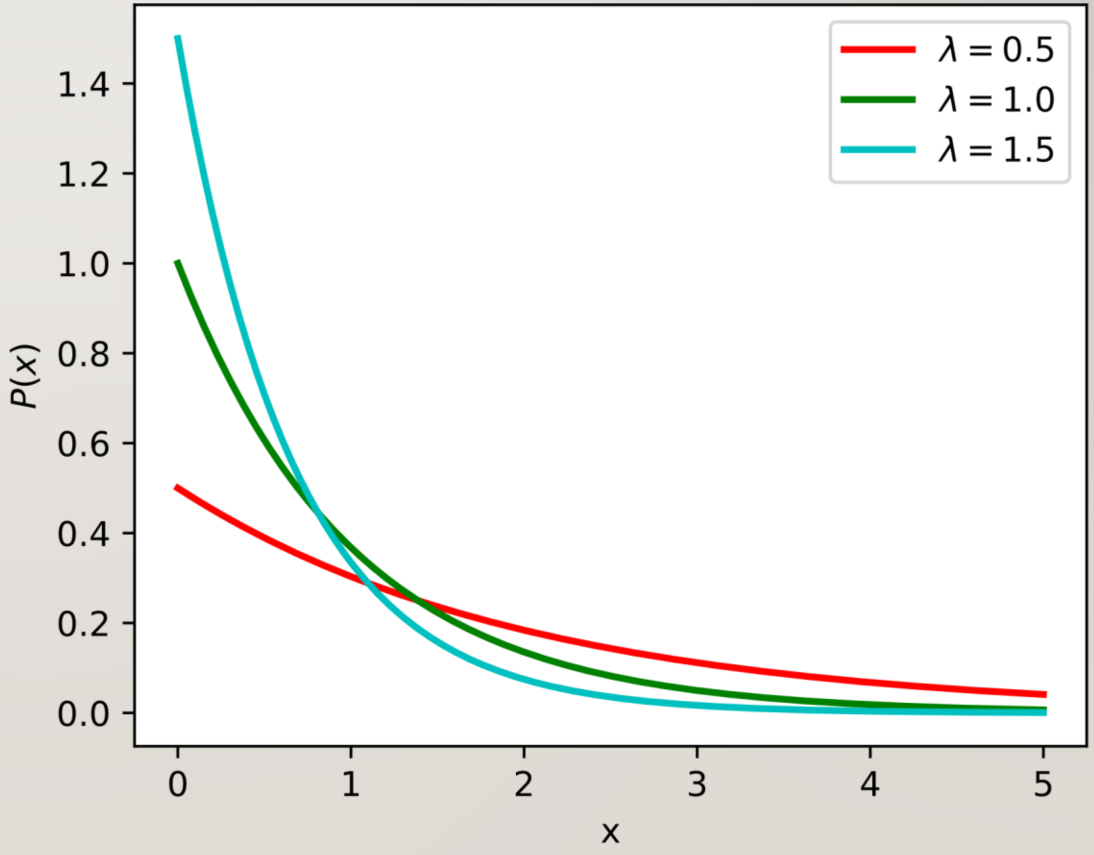
LOGISTIC DISTRIBUTION

$$f(x) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-\frac{x-\mu}{s}})^2}$$



EXPONENTIAL DISTRIBUTION

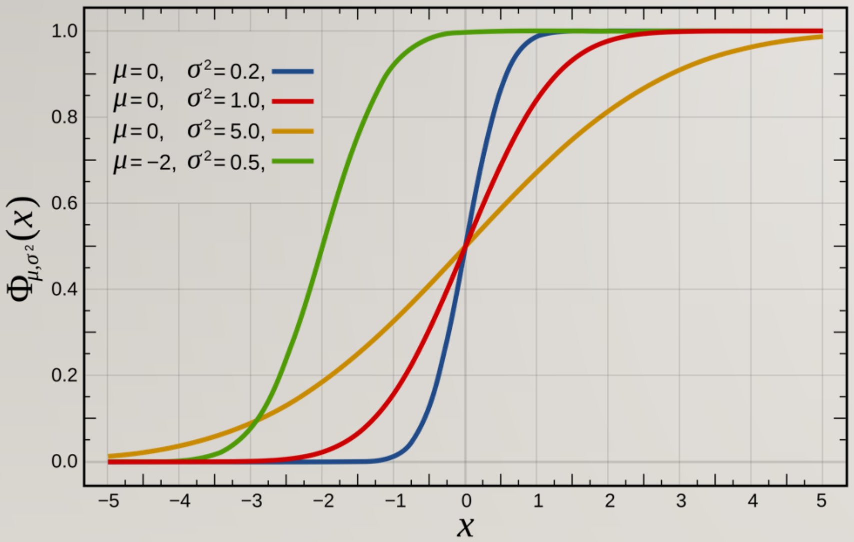
$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$$



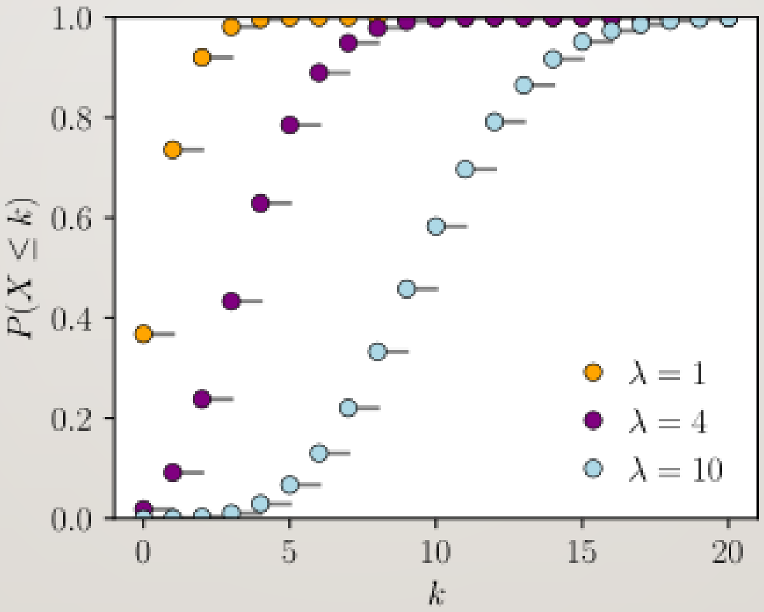
HOW TO WRITE THESE

- Often, a particular variable is described by it’s probability distribution
  - This is given as Variable ~ Distribution(Parameters)
  - Example: Height ~ Normal(3,10)
  - This means that Height has a mean of 3 with a variance of 10
  - What does Time ~ Exponential(7) mean?
  - What does this tell you about an exponential distribution?

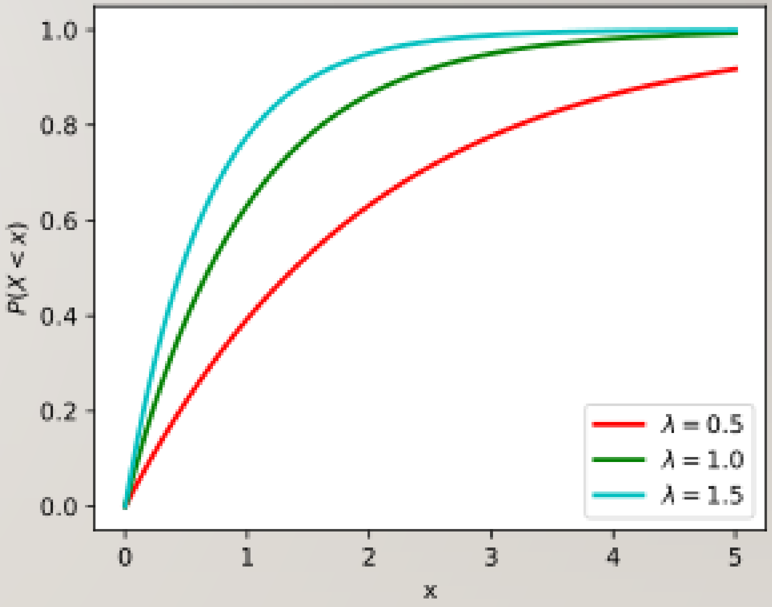
CUMULATIVE DISTRIBUTION FUNCTIONS



Normal



Poisson



Exponential