ECO 602
Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

Probability Distributions 2

Today's Agenda

1. Revisit distributions, discrete distributions
2. A note on distribution function plots
3. Group activity/quiz
4. Continuous distributions
5. How to choose a distribution

Probability key concepts

The sum of the probabilities of **all possible events** is 1.0

The probability of a **specific event** is usually less than 1.0

Independent events: the value of one observation gives us **no information** about the value of another observation.

Probability key concepts

Independent events: the probability of observing a **specific series** of events is equal to the **product** of the **individual events**.

The set of all possible events of a stochastic process is the **sample space**.

What is the sample space of a single coin flip?

What is the sample space of two independent coin flips?

Sampling and sample spaces

- Sample spaces can be discrete or continuous.
- A sample space can be finite or infinite
  - Finite, discrete example?
  - Infinite, discrete example?
  - Finite, continuous example?

Probability Distributions

A **distribution** associates a probability with every possible **event** in the **sample space**.

Theoretical distributions have well-defined functions.

Empirical distributions are calculated from data.

We usually want to **infer** a **theoretical distribution** of a **population** using an **empirical distribution** calculated from data.

Discrete distributions

Sample space is discrete – events cannot take on intermediate values.

For example, in a series of tosses of a coin, it is never possible to observe 1.34 heads.

But unintuitively, the sample space can still be **infinite**!

Binomial distribution

Describes a set of n independent Bernoulli trials.

Each trial has the same success probability.

Two parameters:

- p = probability of success
- n = number of trials

Repeated coin flips

Binomial examples from reading

Parameters:

n = number of trials

p = probability of success

Probability functions:

mass, cumulative mass

empirical, quantile

Binomial examples from reading

Brown creeper experiment:

- 10 sites
- Success: observing a bird at the site
- Failure: not observing a bird at the site
- Binary outcome, multiple trials.
- Binomial is a good candidate model.
- What could go wrong?

Probability Distributions… discrete

Example: *Binomial distribution*

Probability mass function (pmf):

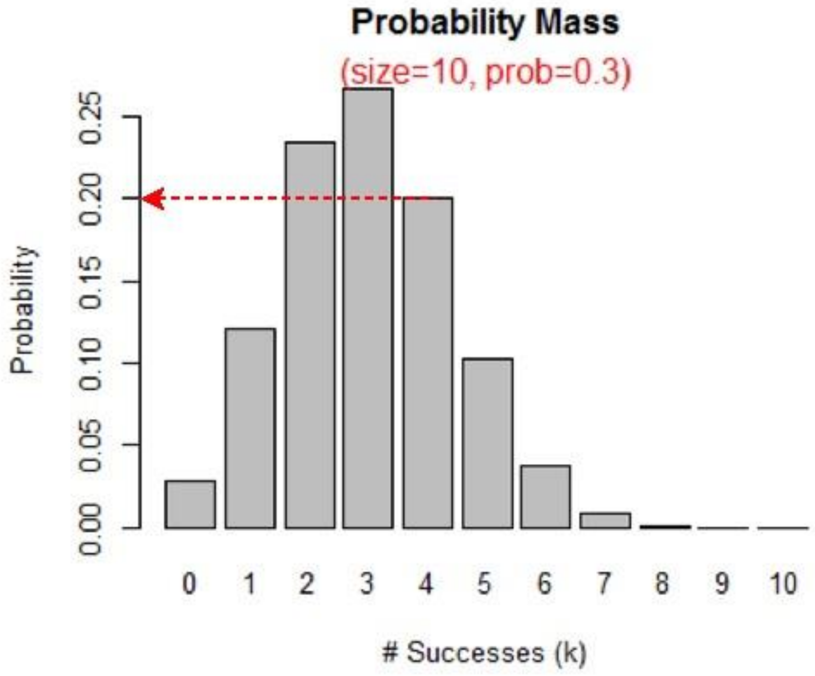
f(x) = Prob(X=x)

Binomial pmf:

$$\binom{N}{x} p^x (1 - p)^{N-x}$$

N = trial size
p = per trial prob(success)
x = #successes (k)

dbinom(x=4,size=10,prob=0.3)
= 0.2



Probability Distributions… discrete

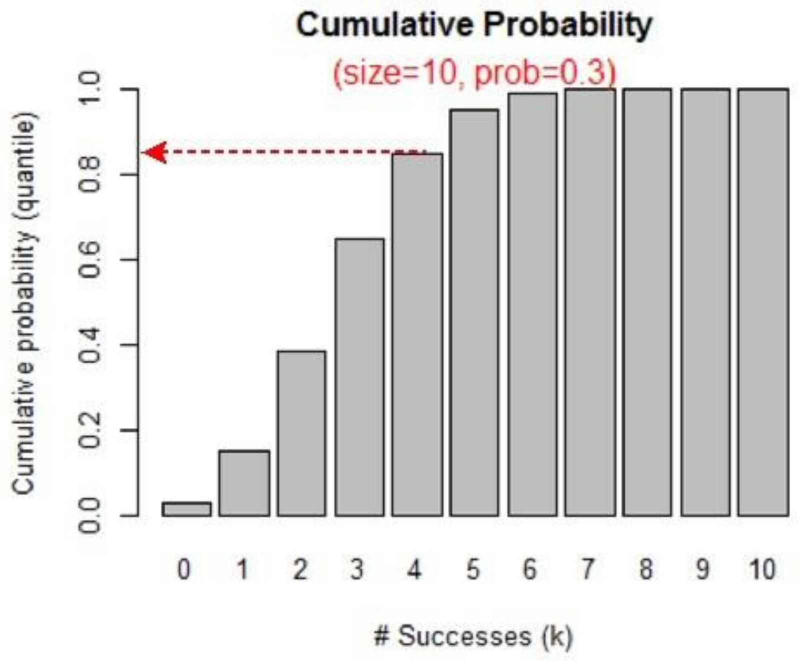
Example: *Binomial distribution*

Cumulative probability distribution:

$$f(x) = \mathrm{Prob}(X \leq x)$$

- Denotes probability of x being less than or equal to any particular value (basis for *p*-values)

pbinom(x=4,size=10,prob=0.3) = 0.85

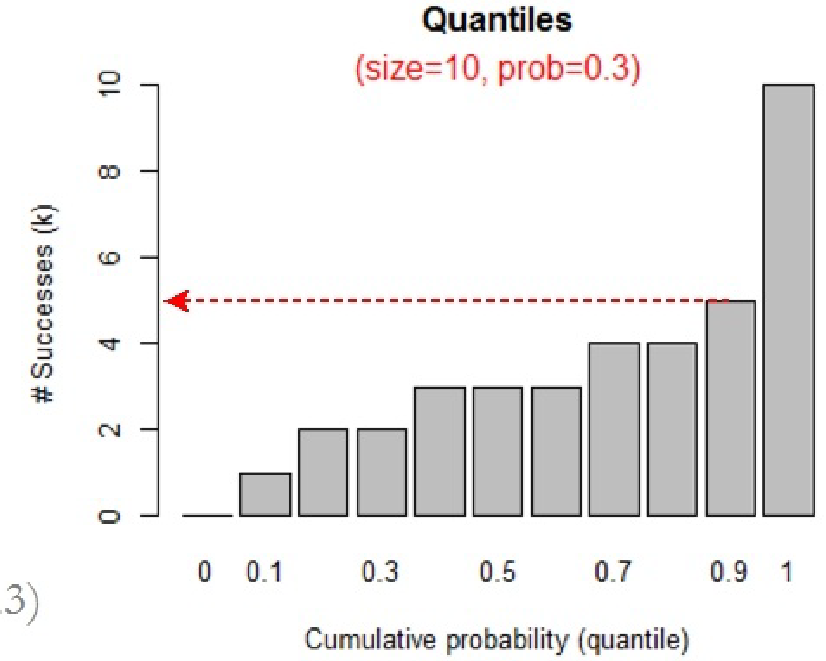


Probability Distributions... discrete

Example: *Binomial distribution*

Quantile distribution:

- Denotes value of x for any given quantile of the cumulative probability distribution; i.e., it is the opposite of the cumulative probability distribution



qbinom(p=.9,size=10,prob=0.3) = 5

Quantile functions are confusing*
*to me

I find the concept of quantile functions much more confusing than probability mass or cumulative probability functions!

We'll go over the concept several times.

Quantile functions

The reading says they are the ‘opposite’ of the cumulative mass function.

You can think of it as an inverse function to the cumulative mass function.

If you have a headache at this point, it’s ok.

Probability Distributions… discrete

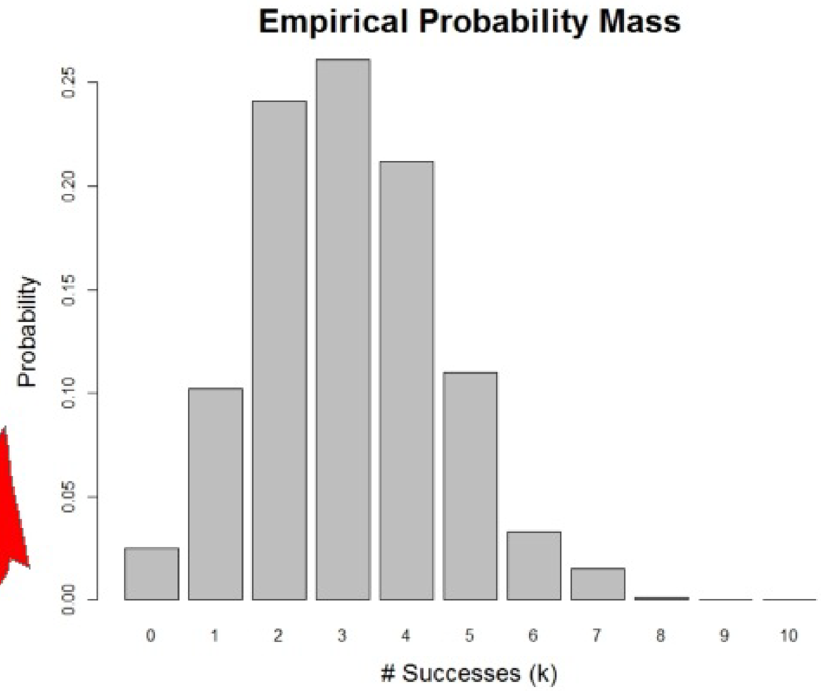
Example: *Binomial distribution*

<u>Example:</u>

Size(#trials) = 10
prob(present) = 0.3

Sample	Trial outcome	*k*
Sample 1	0 1 0 0 0 1 0 1 1 0	4
Sample 2	0 0 0 0 0 0 1 0 0 0	1
Sample 3	0 1 1 0 0 0 0 0 1 0	3
etc…		

Note, divide frequencies by total frequency to convert to a probability



Histograms and mass functions

1. Did you notice a similarity between histograms, probability mass functions, and empirical mass functions?
2. Hint: probability mass functions are just a type of normalized histogram.

In-class activity/quiz: sample of 30 fish from 1 lake

Think broadly for these questions, don't try to give technical answers, but rather focus on the conceptual and philosophical aspects:

Q1: What can we learn about our particular sample of fish? What can't we learn?

Q2: What can we learn about the population of fish in the lake? What can't we learn?

In-class activity/quiz: sample of 30 fish from 1 lake

Q3:

- What if we could repeat our sampling process 10 times?

In-class activity/quiz: sample of 30 fish from 1 lake

Q4: What if we cannot repeat our sampling procedure, and we know nothing about the lake?

Q5: What if we cannot repeat our sampling procedure, but we already have some information about the fish in this lake?

Continuous Distributions

Normal distribution

1. Two parameters: mean, standard deviation
2. Reading example: fish in streams
   1. Mean length = 10cm
   2. Standard deviation of length = 2cm

Probability Density

1. Density at a point is 0
2. Density of a range can be > 0
    1. Density of a range is a definite integral

Probability Distributions… continuous

Example: *Normal distribution*

Probability density function (pdf):

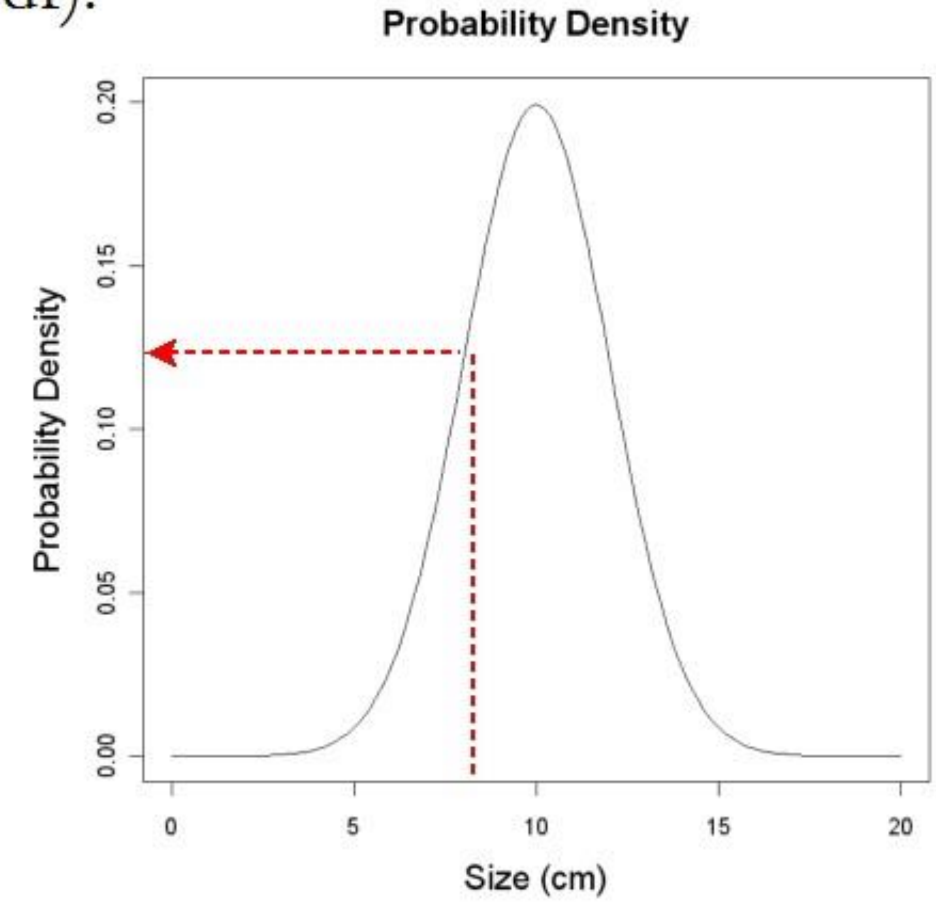
$$f(x) = \frac{prob\big(x \le X \le (x + \Delta x)\big)}{\Delta x}$$

Normal pdf:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x - \mu)^2}{2\sigma^2} \right)$$

μ = mean
σ = standard deviation
x = value of random variable

dnorm(x=8,mean=10,sd=2)
= 0.12



Cumulative Distribution Function

1. Indefinite integral of a density function
2. Mean value is on x-axis
3. Cumulative probability, i.e. quantile is on y-axis

Probability Distributions... continuous

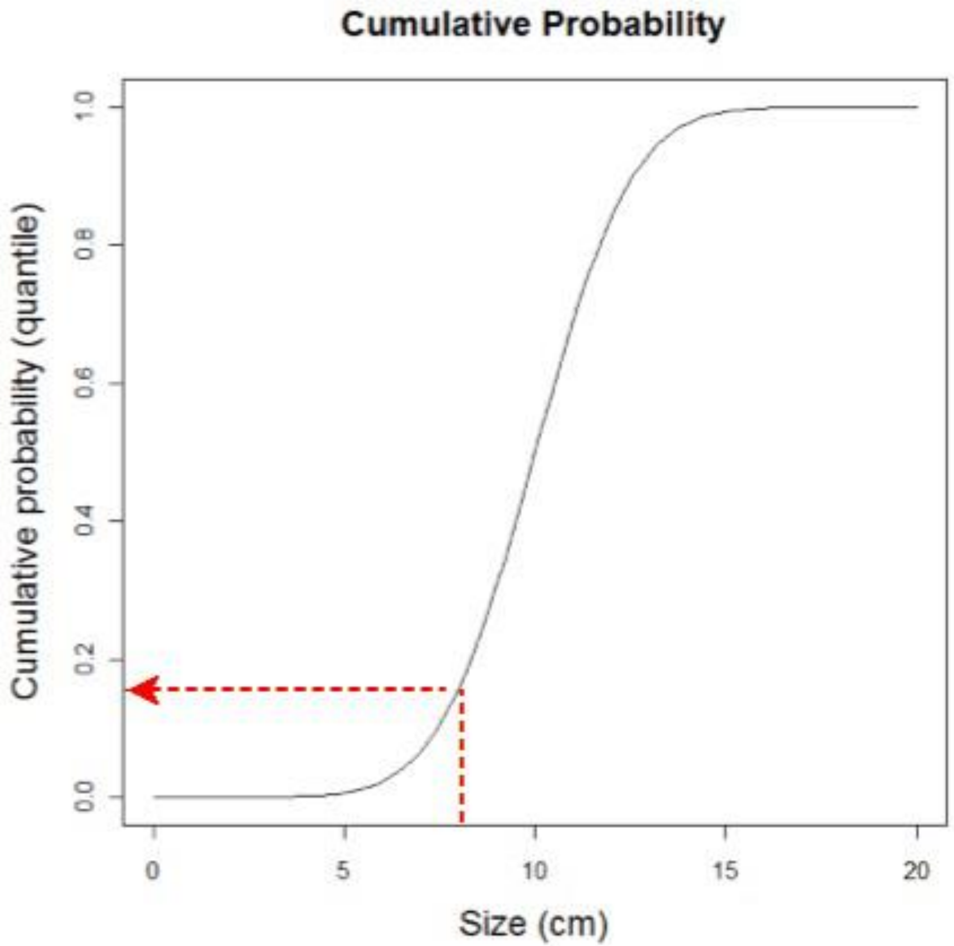
Example: *Normal distribution*

Cumulative Probability Distribution:

$$f(x) = \text{Prob}(X \leq x)$$

- Denotes probability of x being less than or equal to any particular value (basis for *p*-values)

pnorm(x=8,mean=10,sd=2) = 0.16



Quantile distribution function

Function inverse of the cumulative distribution function

Percentiles are on the x-axis

Mean values are on the y-axis

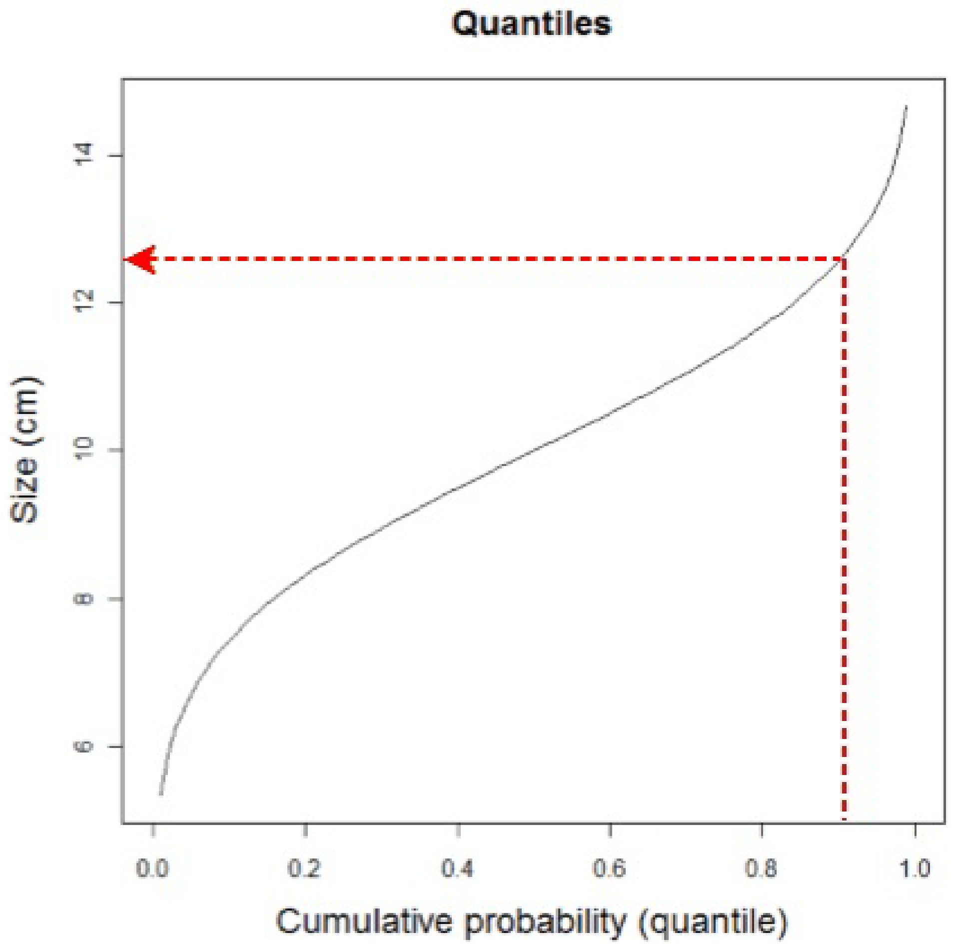
Probability Distributions... continuous

Example: *Normal distribution*

Quantile Distribution:

- Denotes value of x for any given quantile of the cumulative probability distribution; i.e., it is the opposite of the cumulative probability distribution

qnorm(p=.9,mean=10,sd=2)
= 12.56



Our example distributions

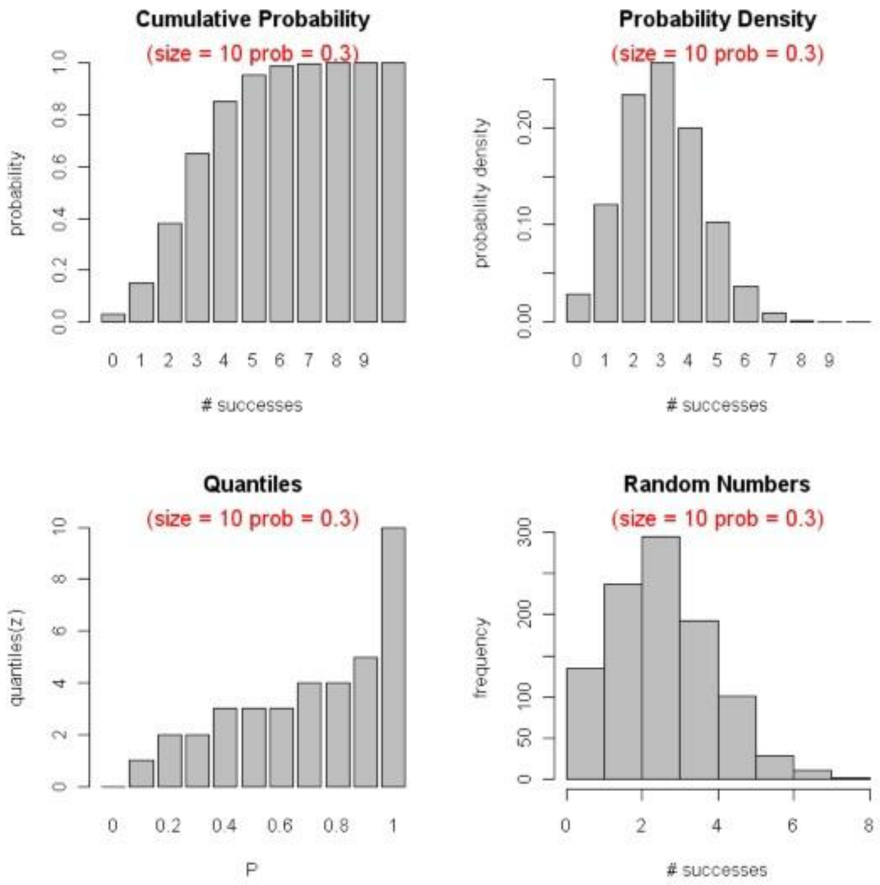
Discrete: Binomial distribution

Continuous: Normal distribution

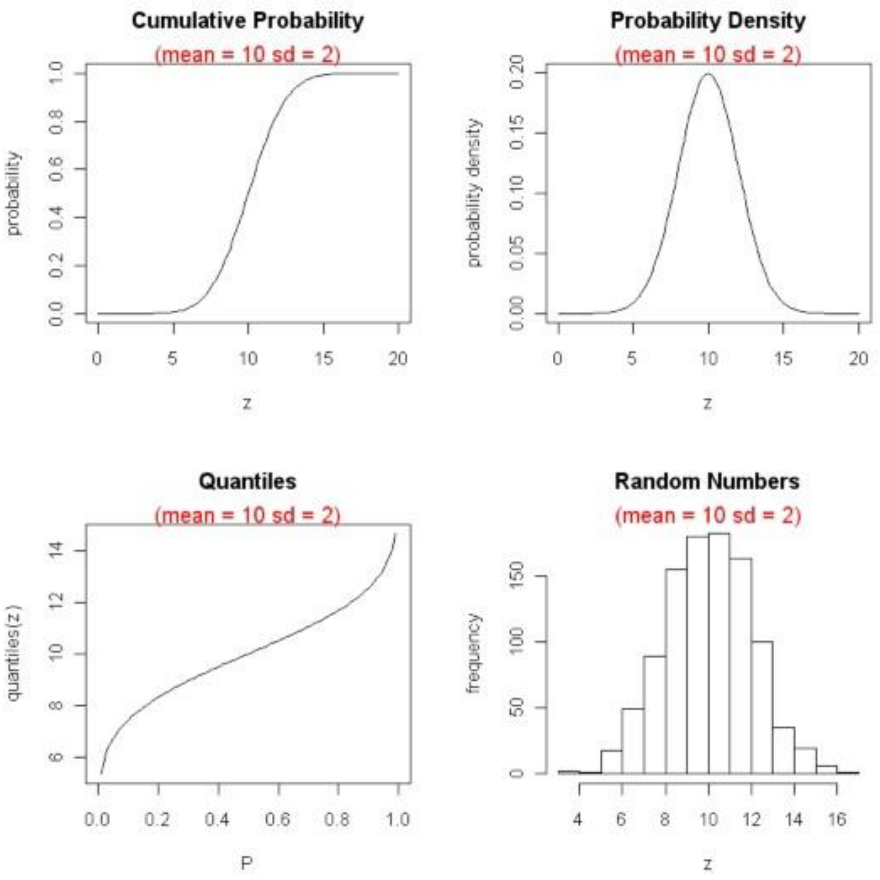
Probability Distributions… review

Discrete versus Continuous

Discrete distributions
(binomial)



Continuous distributions
(normal)



How to choose a distribution?

1. Experimental design: mechanistic
   - What is the possible sample space?

2. Matching data: phenomenological
   - What distribution fits the data best?

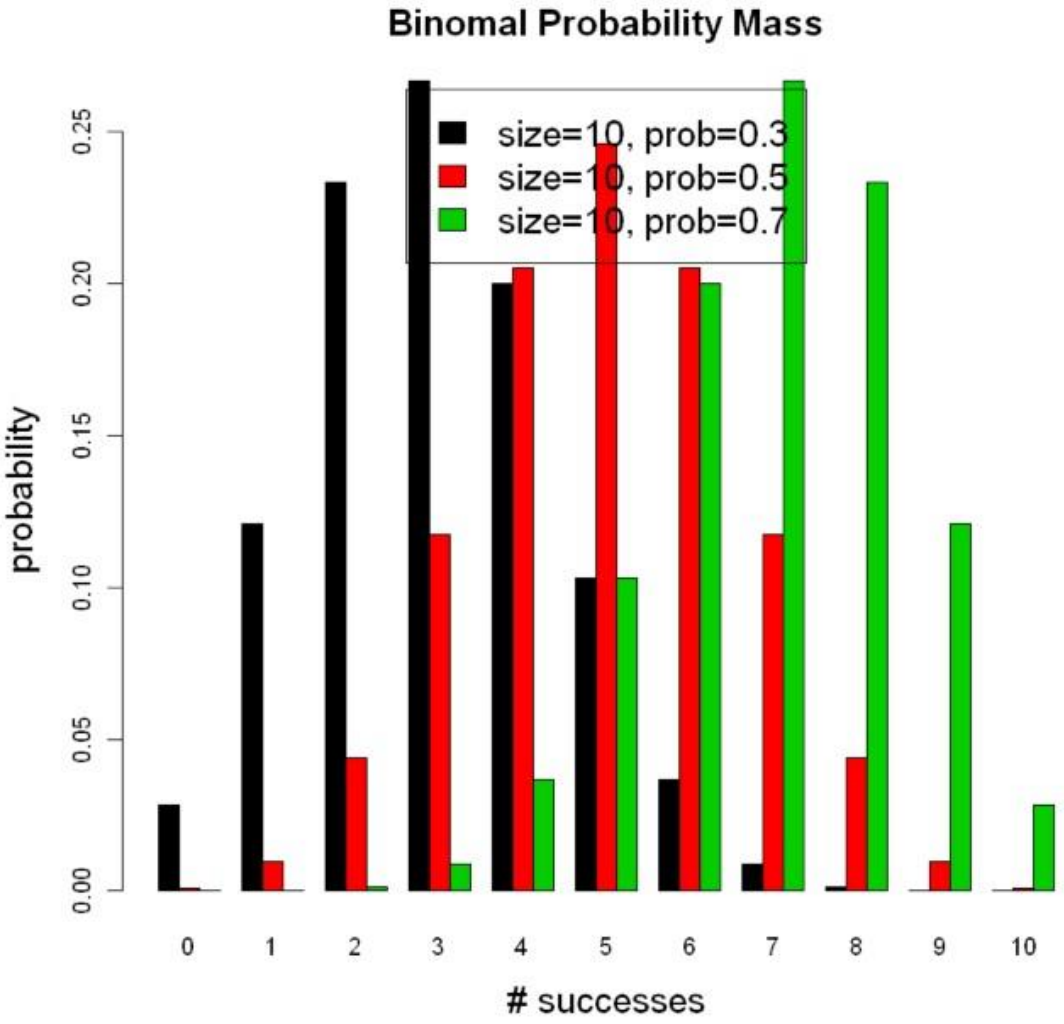
Selected distributions

Probability Distributions… bestiary of functions

Discrete Distributions

Binomial Distribution:

- Gives number of successes (*k*) in a sample with fixed number of trails (*size*) and equal probability of success (*prob*) in every trial
- Used when *k* has an upper limit; when *size* is large and *prob* is small, approaches Poisson
- Lots of examples (very common)

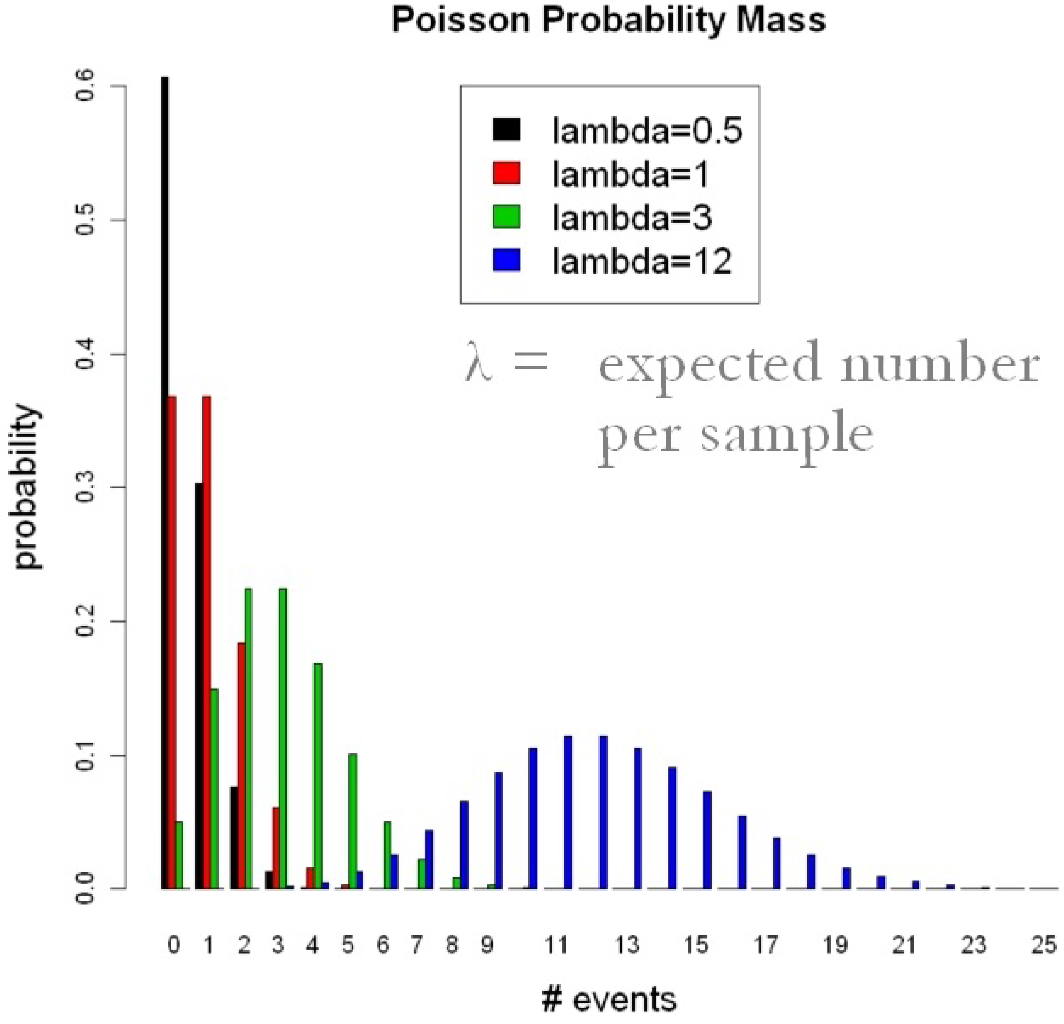


Probability Distributions… bestiary of functions

Discrete Distributions

Poisson Distribution:

- Gives number of events in a given unit of sampling effort if each event is independent
- Used when you expect the number of events to be effectively unlimited
- Used only for count data (very common)



$\lambda =$ expected number per sample

Other discrete distributions you might encounter

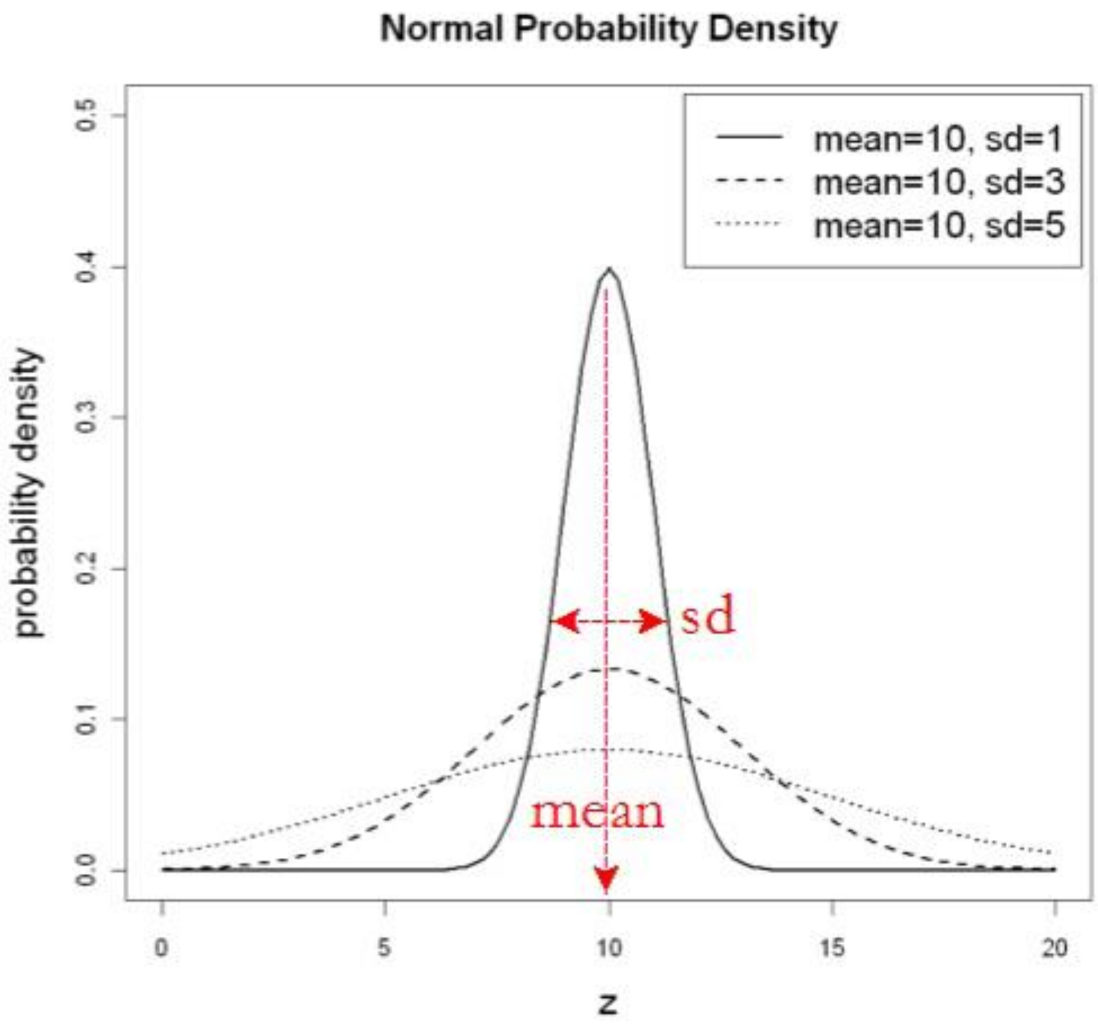
1. Geometric and negative binomial
   1. Binary events
   2. Counts of failures before first success
   3. Survival analyses

Probability Distributions… bestiary of functions

Continuous Distributions

Normal Distribution:

- Distribution of the *sum* of many independent samples from the same distribution; ubiquity lies in the Central Limit Theorum
- Mean and variance are independent, so used when variance is constant
- Used with continuous, unimodal and symmetric distributions (everywhere), and the basis for most classical methods

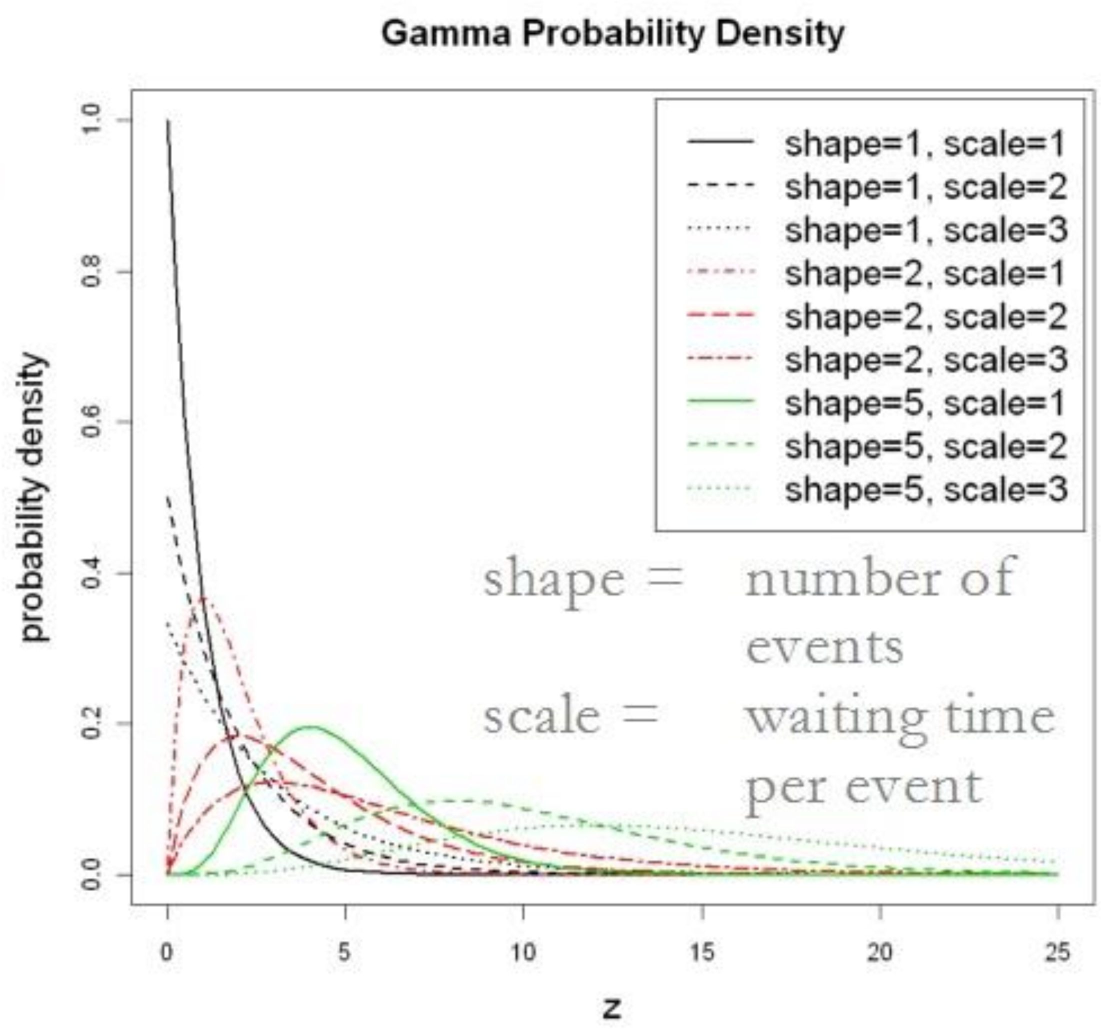


Probability Distributions… bestiary of functions

Continuous Distributions

Gamma Distribution:

- Distribution of waiting times until a certain number of events (*shape*) takes place given an average waiting time per event (*scale*)
- Continuous counterpart of negative binomial
- Used phenomenologically with continuous, positive data having too much variance (overdispersed normal) and right skew

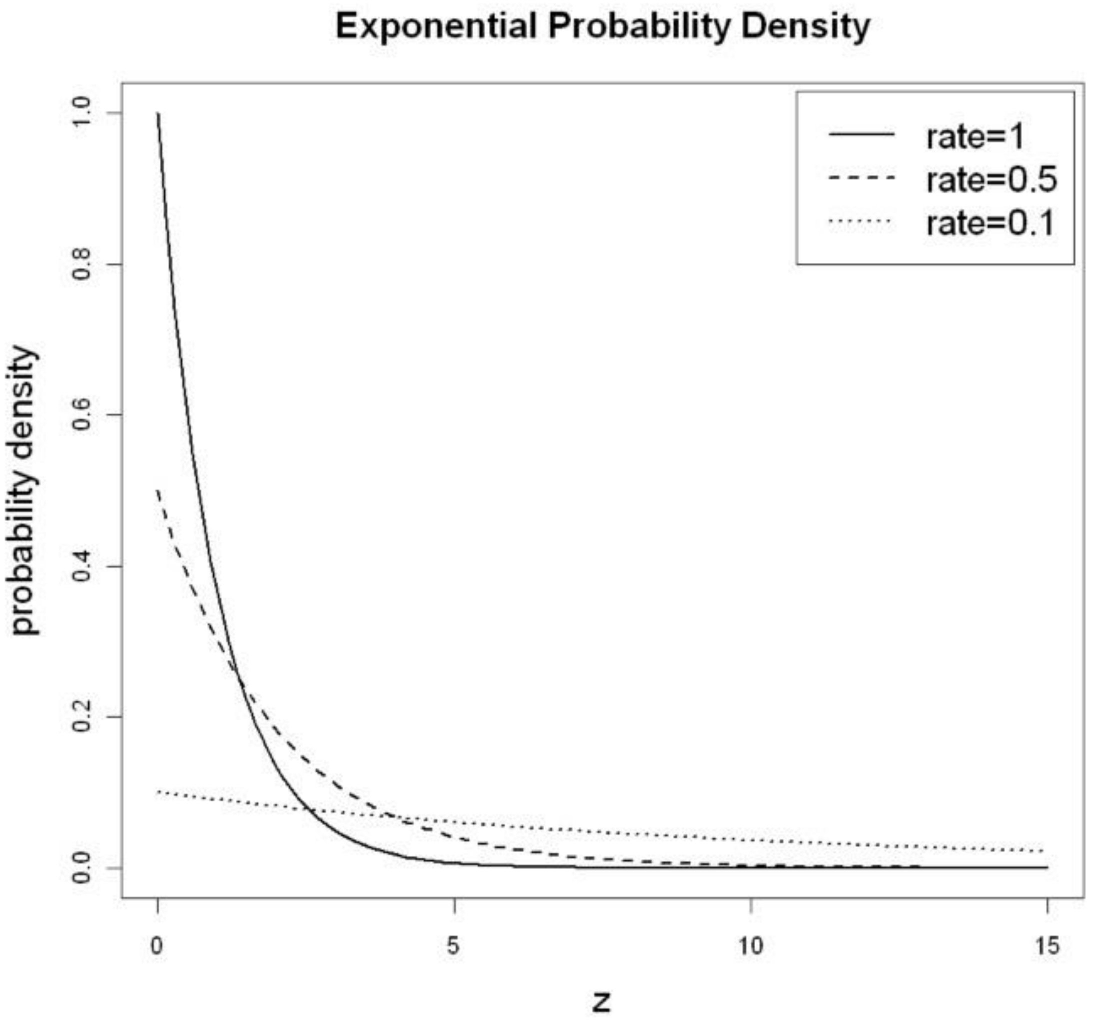


Probability Distributions… bestiary of functions

Continuous Distributions

Exponential Distribution:

- Distribution of waiting times for a single event to happen given a constant probability per unit time (*rate*) that it will happen
- Continuous counterpart of geometric and special case of Gamma (*shape*=1)
- Used phenomenologically and mechanistically (very common)

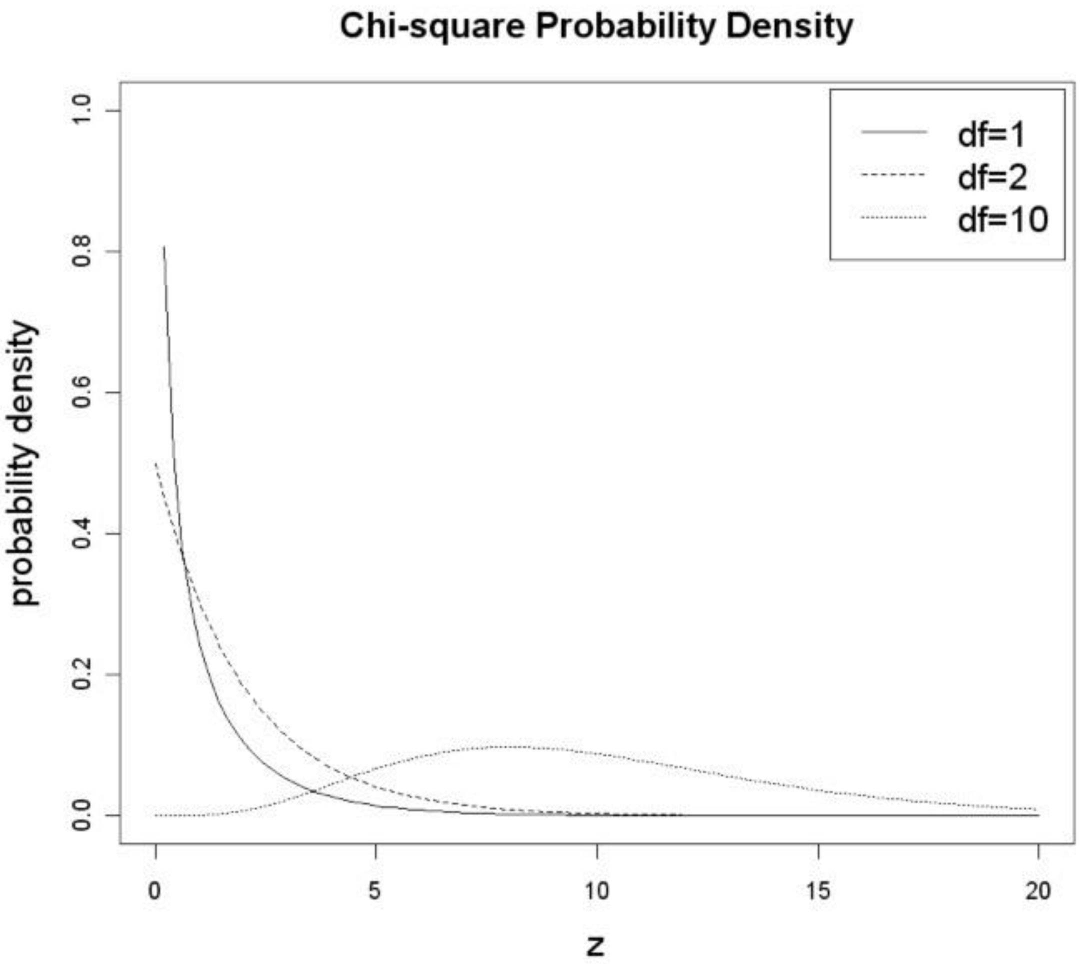


Probability Distributions… bestiary of functions

Continuous Distributions

Chi-square Distribution:

- Distribution of the sum of squares of *n* (degrees of freedom) normals each with variance one
- Famous for its use in contingency table analysis
- Important because Likelihood ratio statistics are distributed approximately chi-square

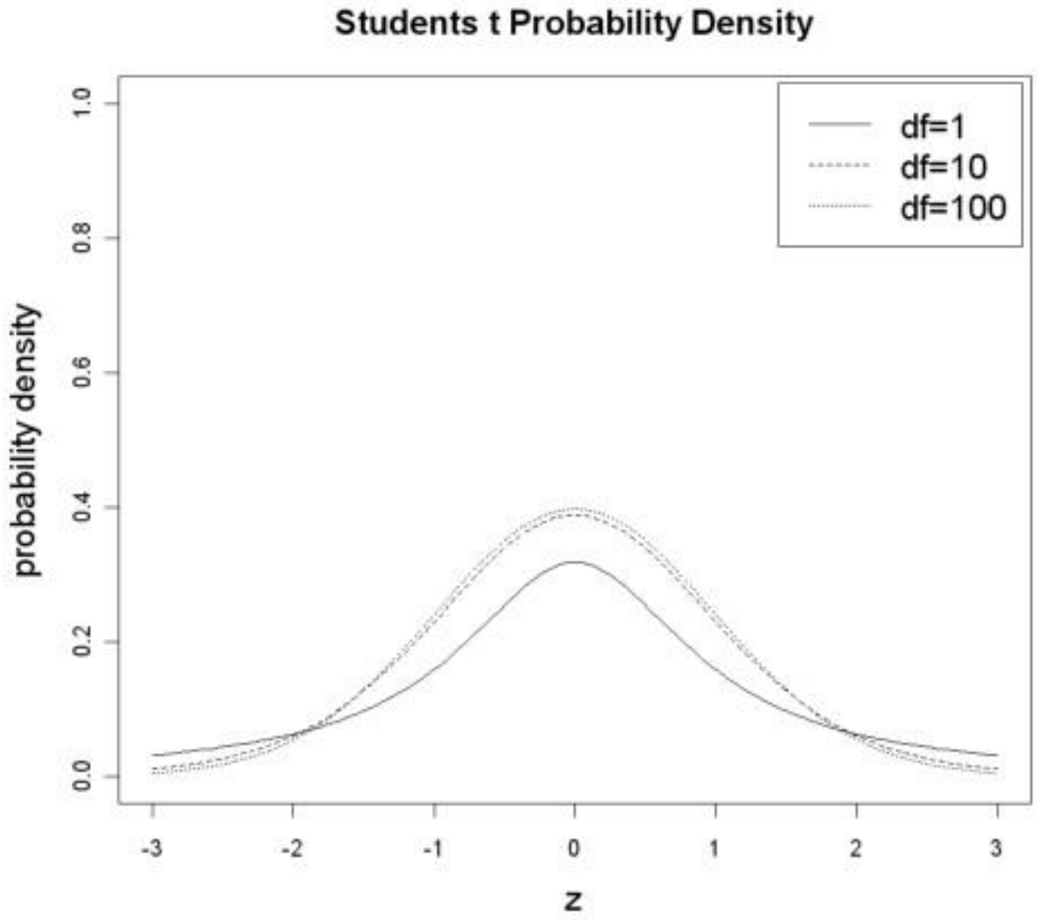


Probability Distributions… bestiary of functions

Continuous Distributions

Student's *t* Distribution:

- Distribution of a random variable that is the ratio of the difference between a sample statistic and its population value to the standard deviation of the distribution of the sample statistic (standard error)
- Famous for its use in testing for difference between the means of two normally distributed samples and for whether a parameter estimate differs from zero.



Choosing a distribution

Can be complicated.

But there are a small number of distributions that are used very frequently.

Sometimes software can help choose an optimal distribution

For next time:

1. Finish discrete distributions
2. Continuous distributions