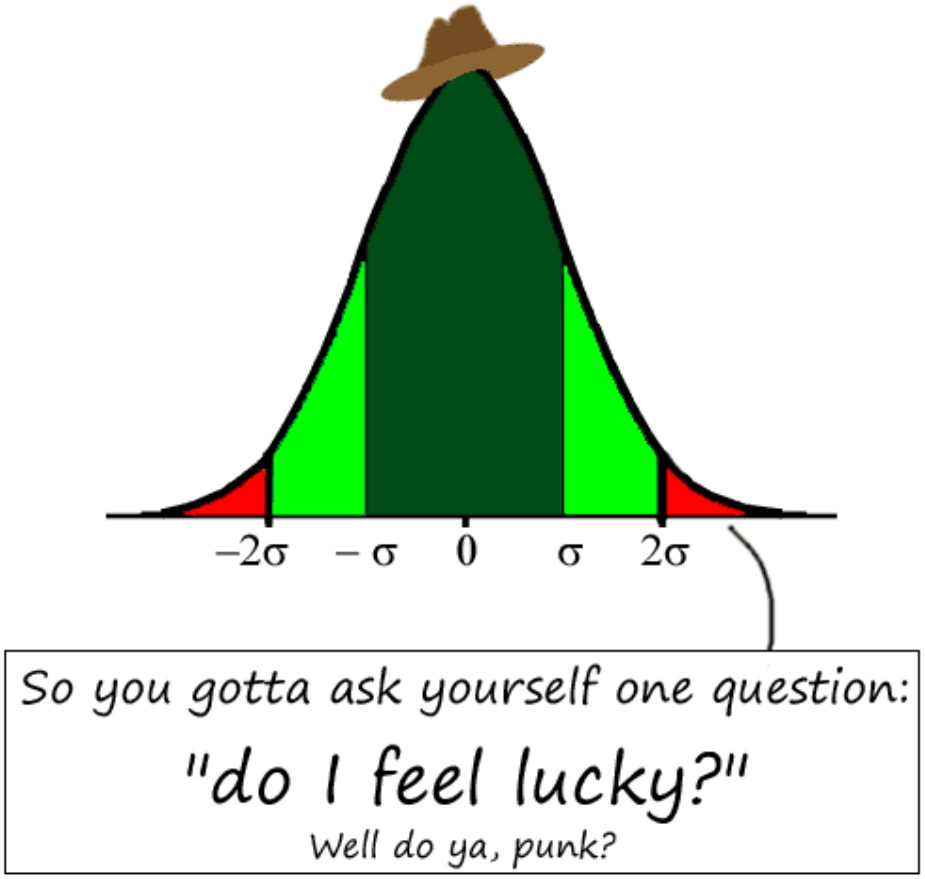
Probability distributions

A Field Guide to Probability theory & distributions for biologists



Thursday
October 2nd
2014

Goals for Today

- Re-familiarize ourselves with conditional probabilities, and the the language of probability.
- Clarify Probability distributions:
  - Probability MASS functions (discrete)
  - Probability DENSITY functions (continuous).
  - Simple calculations using these distributions.

Readings

- Chapter 4 of Bolker (Ecological examples)

  OR
- Chapters 6 &7 of Seefeld (Genomics examples)
- VERY useful reference as well:
- Crawley (The R Book) Chapter 7
- Vasishth & Broe: The foundations of statistics: A simulation-based approach

Some basic tools from probability theory to keep in mind

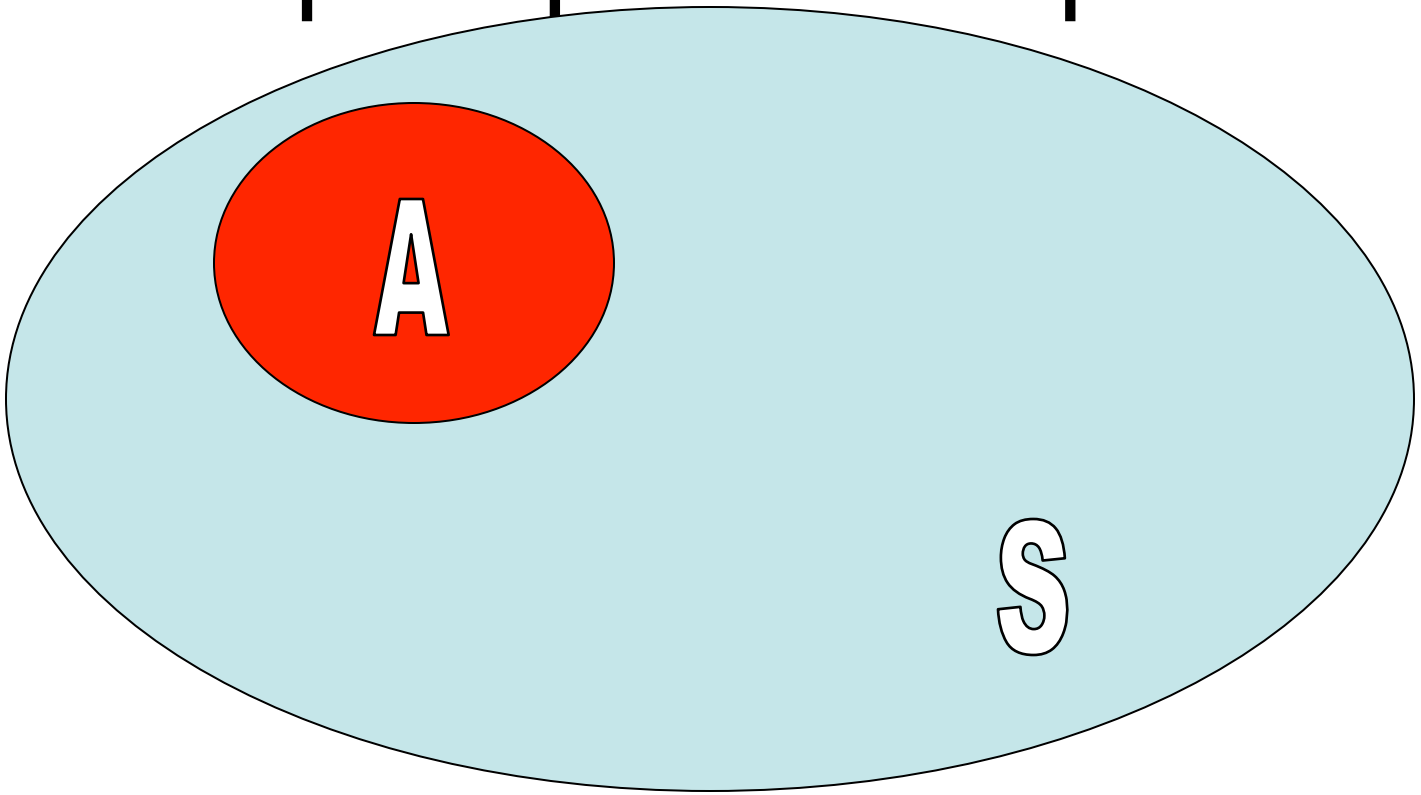
Sample spaces….

What is the sample space of a 6 sided die?

A coin?

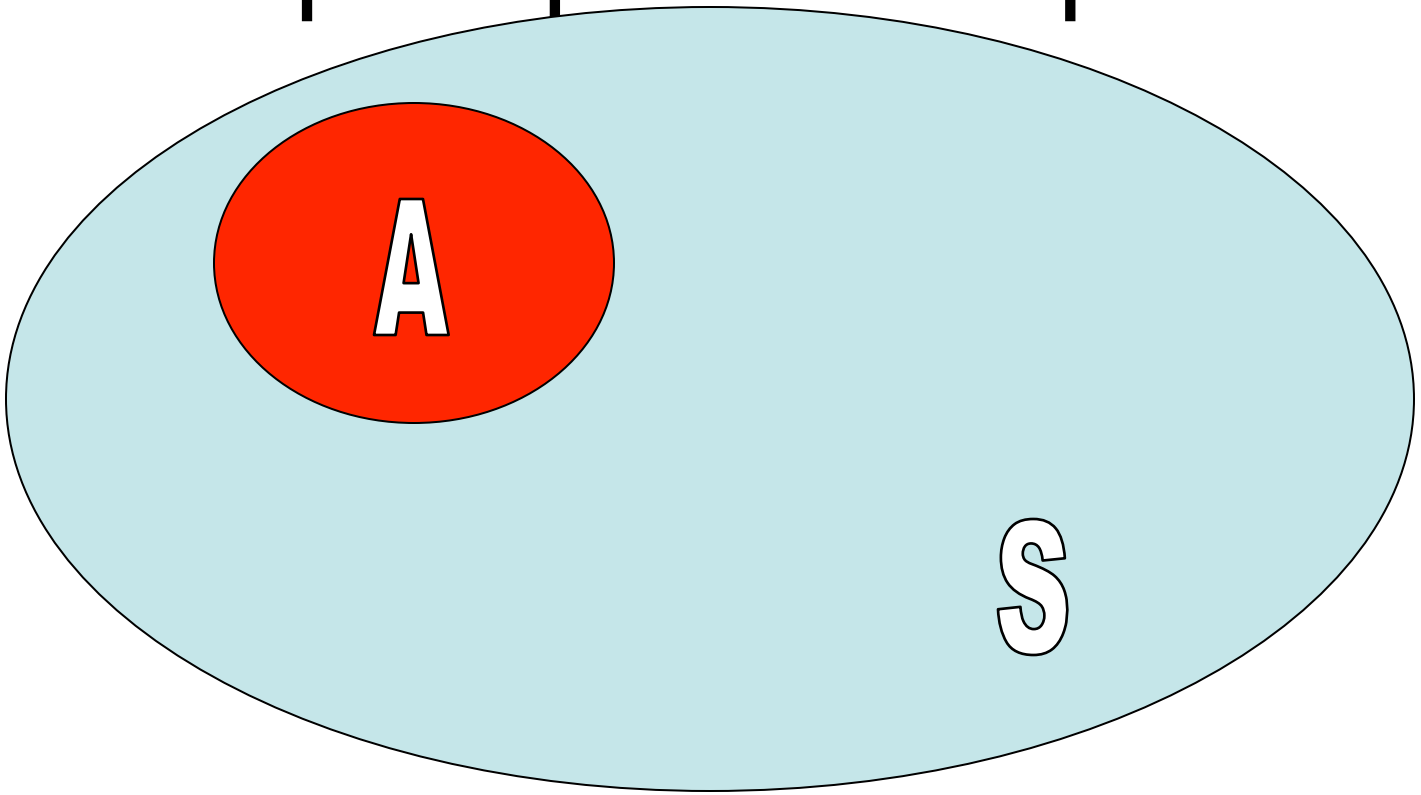
[P] of Lake Michigan?

Sample space and probability



What is the probability of A?

Sample space and probability



Pr(A) = Area of A/ Area of S

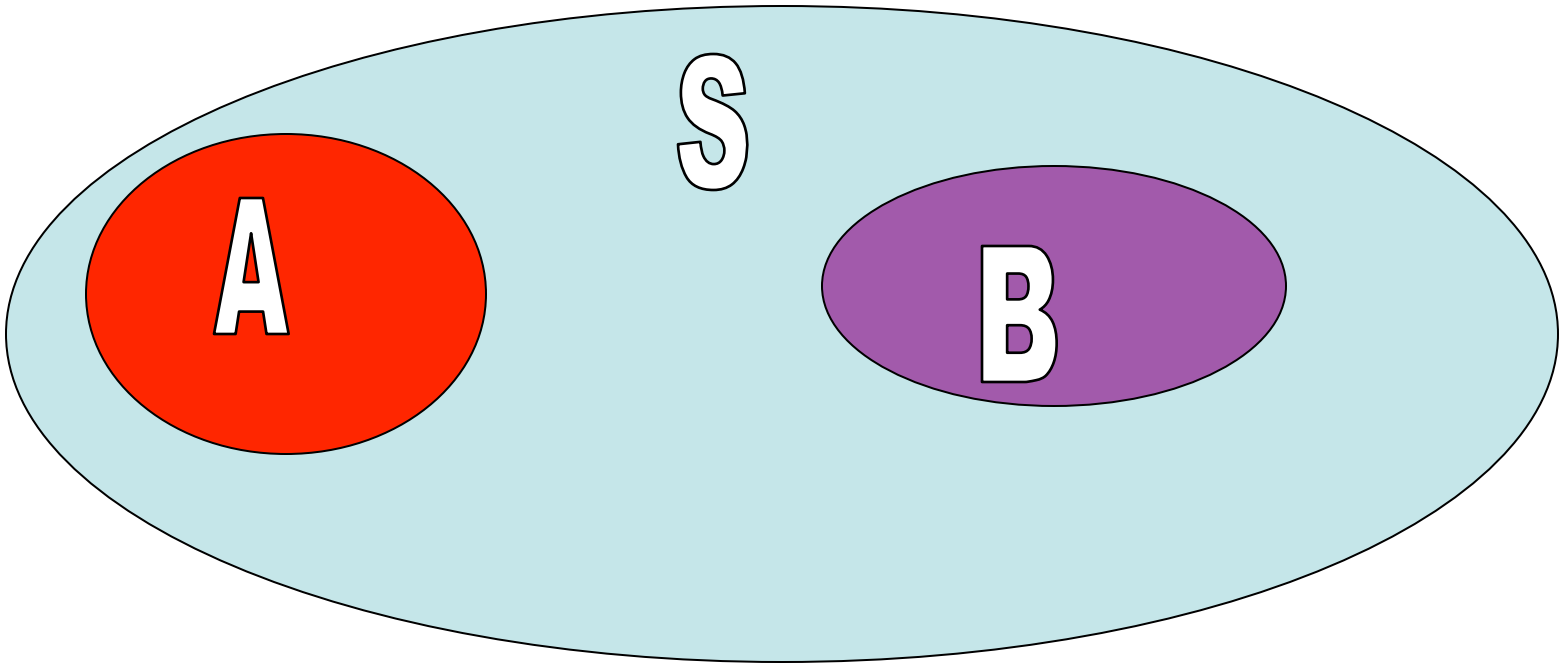
Some basic tools from probability theory to keep in mind

Or statement - the Union ($\cup$).

For two **mutually exclusive** events, the probability of getting A or B

$(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$

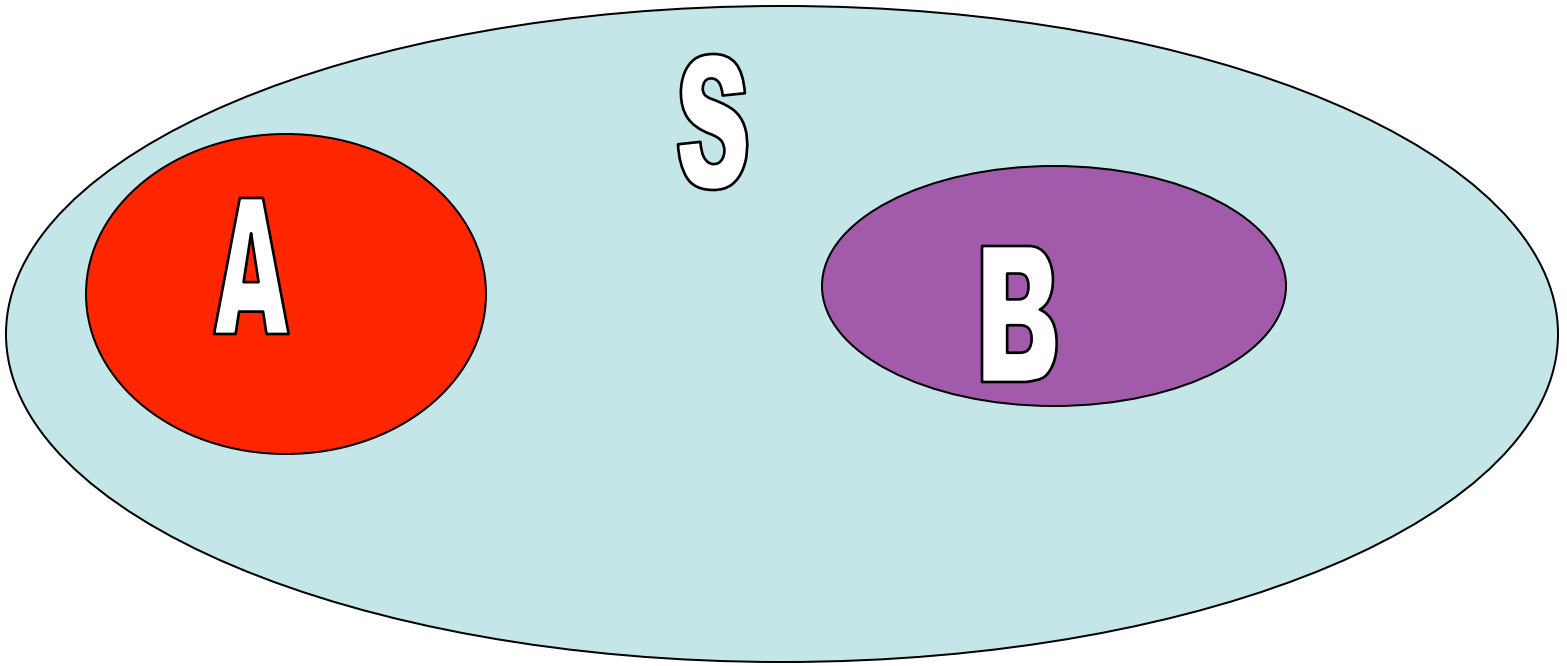
Two independent events



What is the probability of A? of B?

What is the probability of A or B?

Two independent events



Pr(A) = Area of A/ Area of S

Pr(B) = Area of B/ Area of S

Pr(A **OR** B) = (Area of A + Area of B)/ Area of S = Pr(A) + Pr(B)

The probability of a shared event (joint probability)

The **joint** probability of two events A and B, occurring (assuming they are **independent**) is the product of their probabilities (their **intersection**).

$P(A \& B) = P(A,B) = P(A \cap B) = P(A) * P(B)$

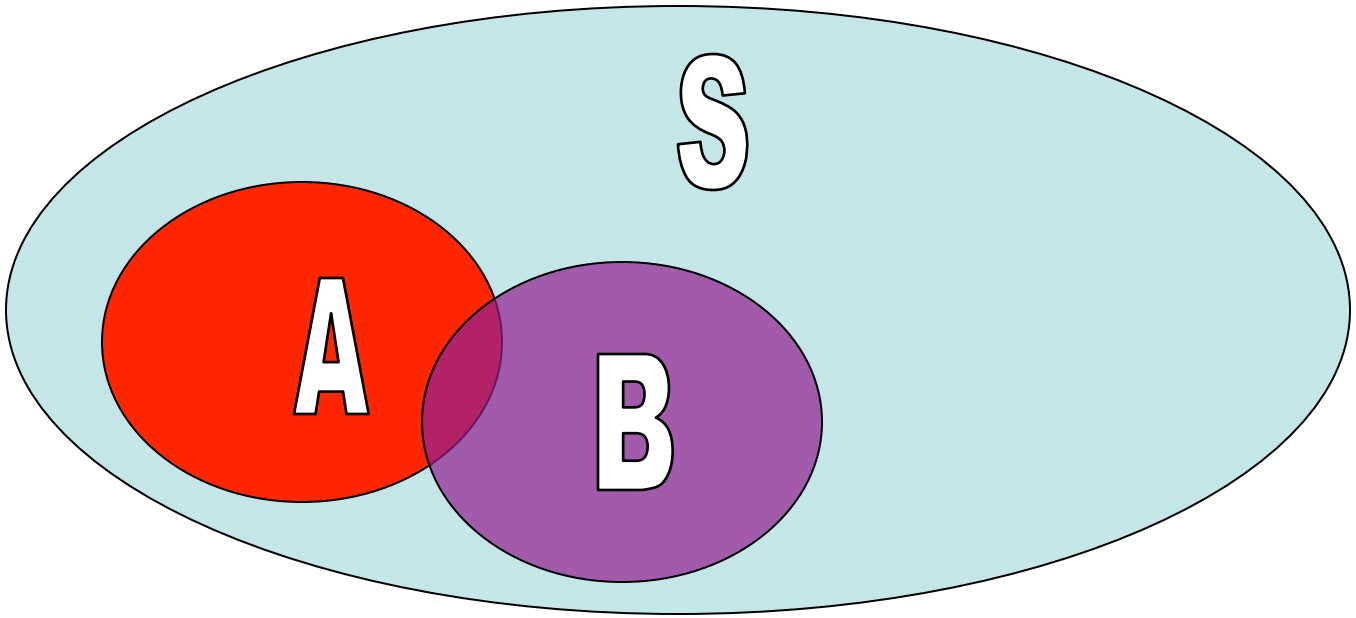
ASSUMES A and B are independent.

The probability of a shared event

Useful identity

$P(A \cap B) = P(A) * P(B) = P(B \cap A)$

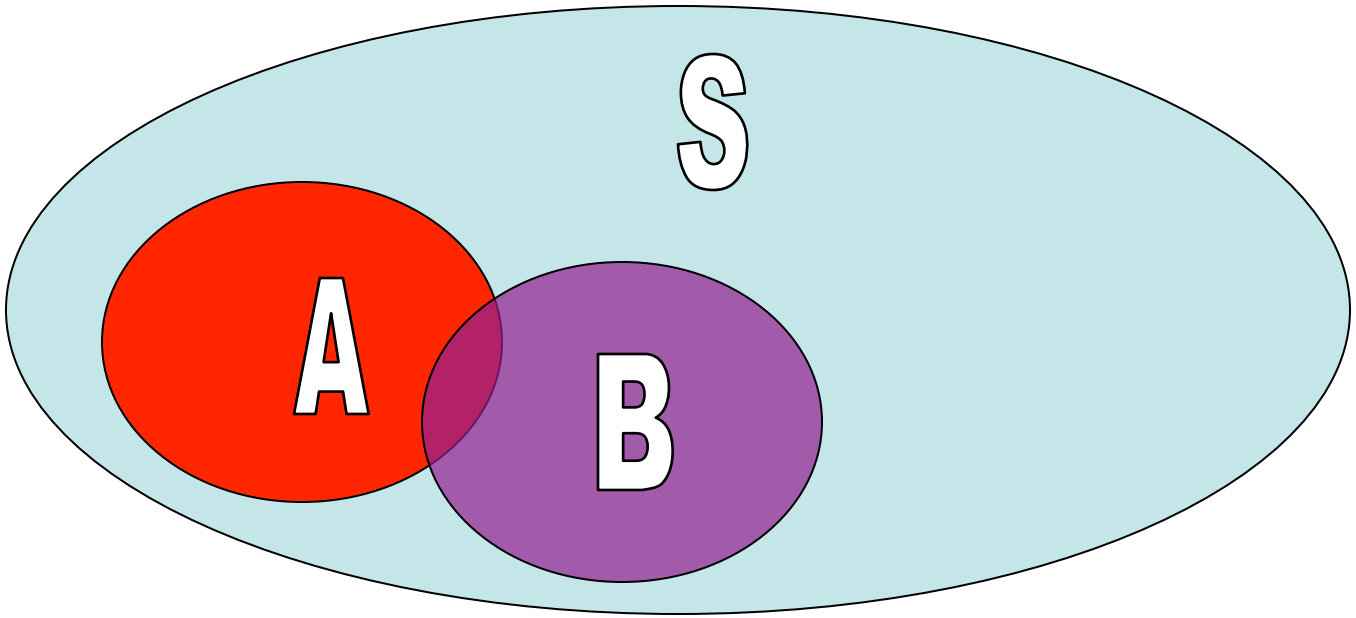
Two events (not independent)



What is the probability of A? of B?

What is the probability of A or B?

Two events (not independent)



Pr{A} = Area of A/ Area of S Pr{B} = Area of B/ Area of S

Pr(A **OR** B) = (Area of A + Area of B - Area of AB)/ Area of S
= Pr(A) + Pr(B) - Pr(A & B)

Pr(A & B) = Pr(A,B) = Joint probability of A and B

Calculating probabilities of combined events

If events A and B are not mutually exclusive. What is

P(A or B) = P (A $\cup$ B) =

Calculating probabilities of combined events

If events A and B are not mutually exclusive. What is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

$$P(A) + P(B) - P(A)*P(B)$$

Conditional probabilities

- If we are calculating the probability of an event, and we have information about the outcome of the event (for instance the distribution of the sample space…..), we should include this information somehow.

- These “updated” estimates are **conditional probabilities**.

Conditional probabilities

- Conditional probabilities are written as:

$P(A|B)$

Translation: P(A) Given outcome B

Or

P(A) conditional on B.

Probability that A occurred given B occurred.

= Area common to A and B/ Area of B

Conditional probabilities

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

Thus the conditional probability of A given B is equal to the intersection of A and B divided by the probability of B.

Conditional probabilities

This can be usefully re-arranged:

$$P(A,B) = P(A|B)*P(B)$$

But the following must also be true…

$$P(B,A) = P(B|A)*P(A) = P(A|B)*P(B)$$

Bayes Rule

Sometimes we know the conditional probabilities, without any information on the joint probabilities….

P (B|A)*P(A) = P(A|B)*P(B)

We can re-arrange this to get Bayes Rule

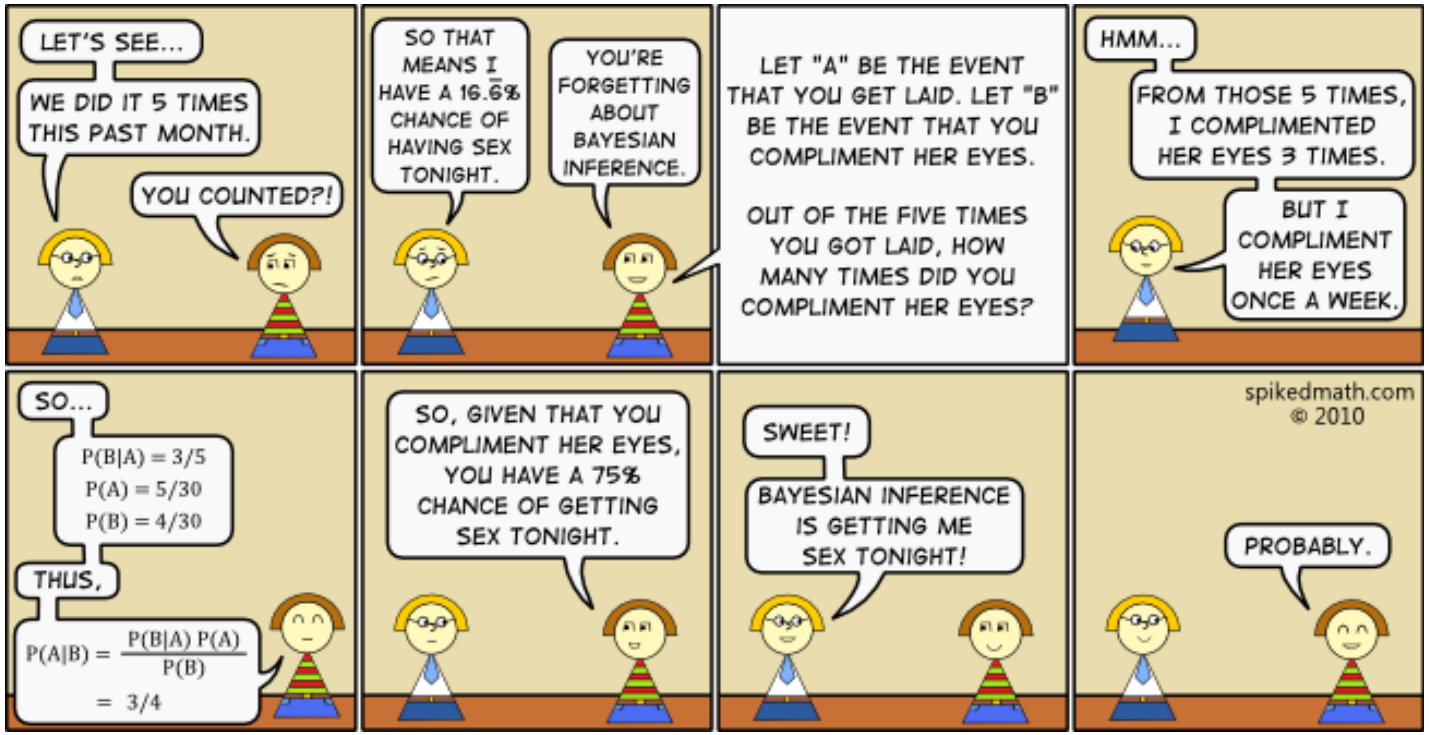
posterior

prior model

$$P(B \mid A) = \frac{P(A \mid B) * P(B)}{P(A)}$$

probability of data

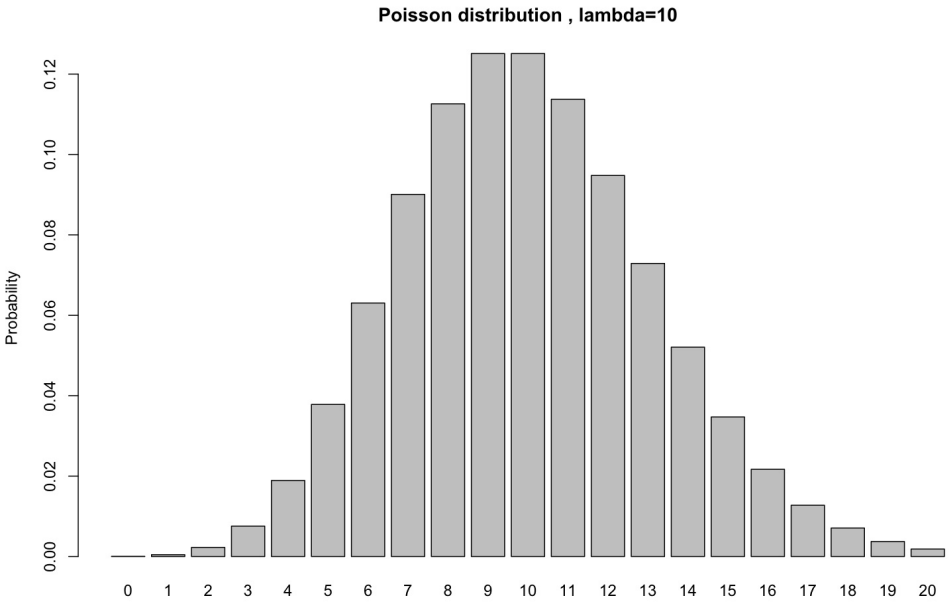
Informative?



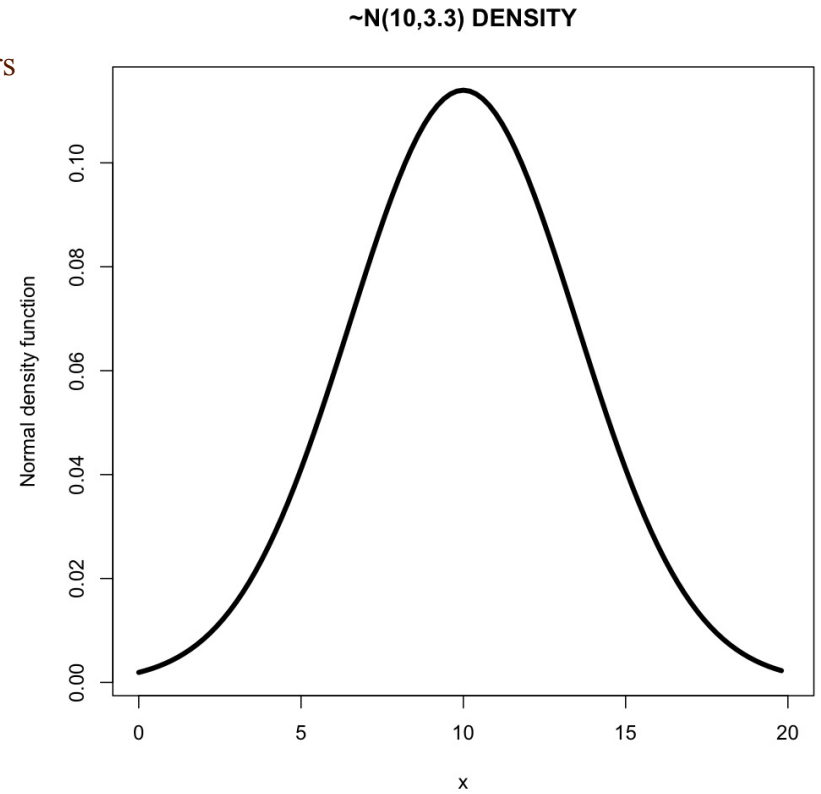
if you understand the previous slides, you’re fine. no extra reading necessary

Probability Density vs. Mass function

discrete distribution; only 0 (boundary condition) or positive integers

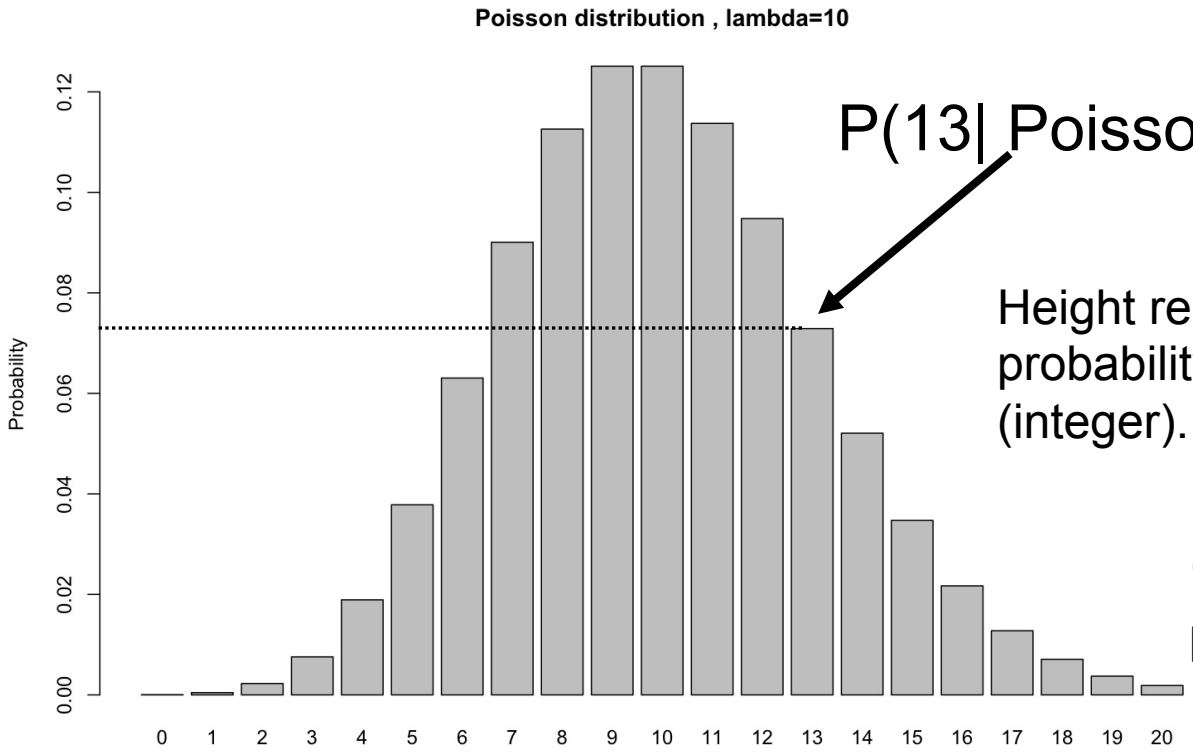


Probability Mass function for a discrete variable.



Probability Density function for a continuous variable.

Probability Mass function



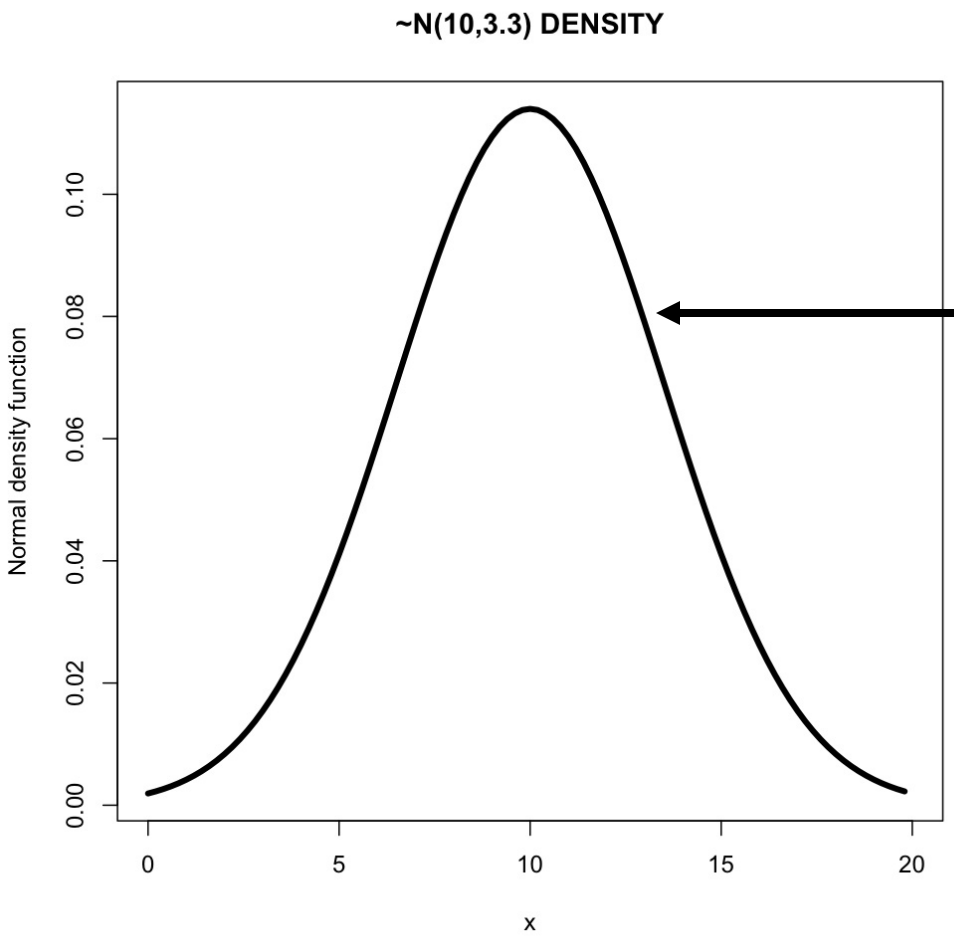
P(13| Poisson (λ=10)) =0.073

Height represents the probability at that point (integer).

"Area" of the box has no particular meaning.

P(integer) ≥ 0
P(non-integers) = 0.

Probability Density function



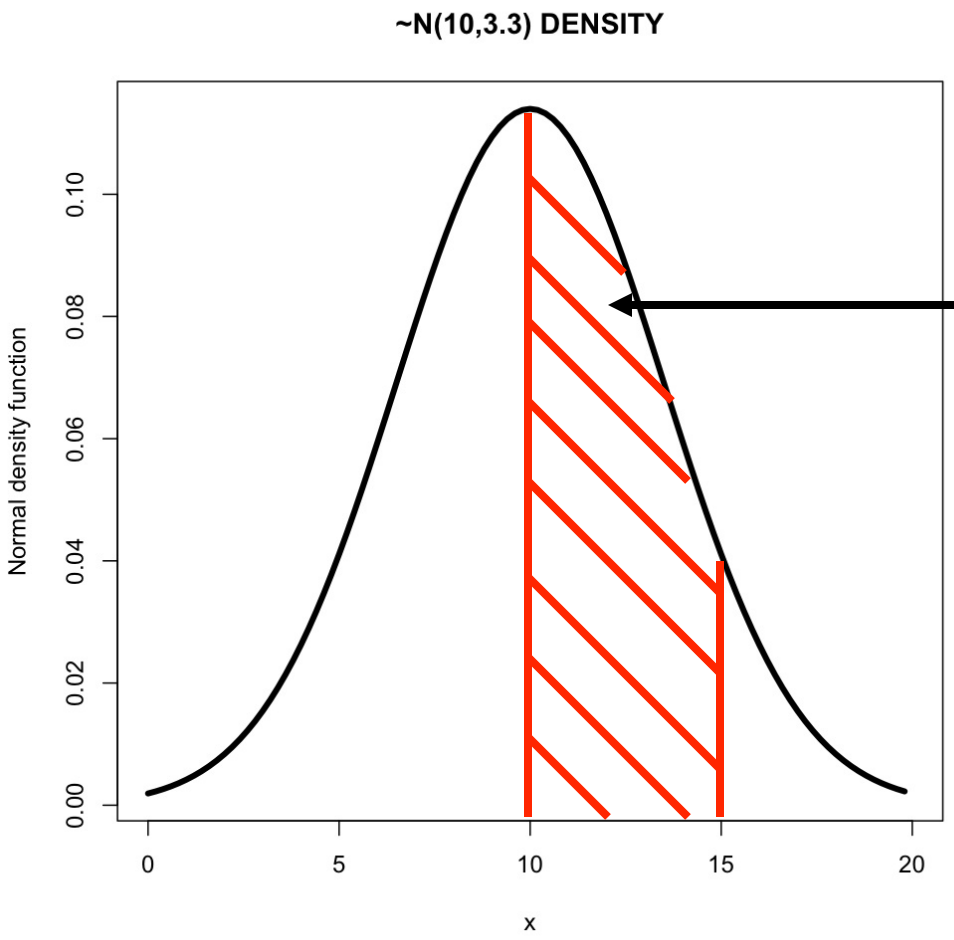
Height at x= 13 is 0.0799

This is not the probability at x=13, but the density.
i.e. f(13) = 0.0799, where f(x) is the normal distribution.

P(x=13| N(mean=10,sd=3.3)) = 0

WHY? it's a continuous distribution, the probability of landing on a specific number is infinismal

Probability Density function

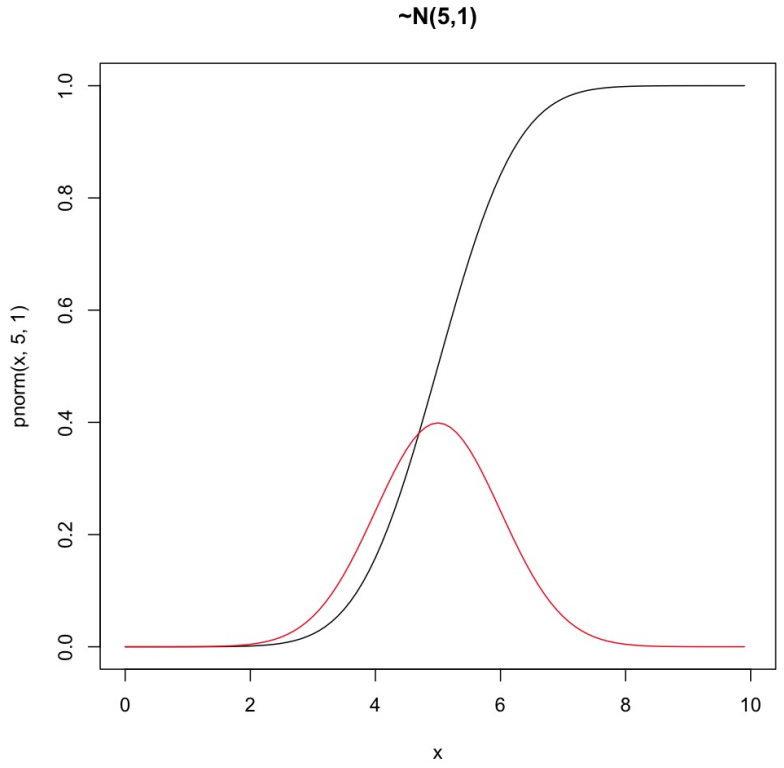


We can define the probability in the interval $10 \leq x \leq 15$

However, we can define an interval then make this calculation.

$P(10 \leq x \leq 15 | N(10,3.3)) = 0.435$

Clarifications on continuous distributions.



$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(a \le X \le b) = \int_{a}^{b} f(x)dx$$

$$P(X = x) = 0$$

AREA UNDER CURVE OF PDF =1

(The integral of the normal)

All of this seems mighty abstract…

- But, if you remember from our first few lectures. I often said things like
- “Probability of the data given the hypothesis….”
- Now we have some tools (or at least symbols) to translate this.

What likelihood is all about…

P(data | hypothesis) = P(D|H)

If we want to know what the probability of our data is (the data which we have collected), we need to have a context for that… our hypothesis.

What do we mean by hypothesis?

In parametric statistics (MLE or Bayesian), this mean some kind of model:

$$Y \sim N(\beta_0 + \beta_1 X, \sigma) = \theta$$

$$P(D|H) = P(Y| N(\beta_0 + \beta_1 X, \sigma)) = P(Y| \theta)$$

What do we mean by hypothesis?

$P(Y| N(\beta_0 + \beta_1 X, \sigma)) = P(Y| \theta)$

Usually our theta (θ) includes actual parameter values (in this case for β_0, β_1 and σ).

This is what we are really trying to do.

An example

- Let's say we have a small data set of the plant heights (cm) in a population (7,4,3,7,7).

- A previous population we examined looked like plant heights were distributed (approximately) normal with a mean of 5cm, and a S.D.= 1.

What is the probability of a plant of height 7cm coming from the population with mean 5 cm and sd =1 ?

What do we need to assume?

P(7 | ~N(5,1))

Thus our data in this case is the “7” and our hypothesis is that the data is ~ N(5,1).

Now we feed this into the normal distribution.

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

P(7 | ~N(5,1))

Thus our data in this case is the “7” and out hypothesis is that the data is ~ N(5,1).

Now we feed this into the normal distribution.

$$\frac{1}{\sqrt{2\pi 1^2}}\exp\left(-\frac{(7-5)^2}{2*1^2}\right)\approx 0.054$$

It is worth mentioning this is not the probability (since we are looking at the interval), but it is proportional to it. As I mentioned in general we will be interested in relative probabilities.

Probability of the data given the hypothesis

- So we have figured out the probability of a single data point, given an arbitrary hypothesis. So what?
- We could extend this to the whole data set (how?).
- We can also use some optimization criteria (such as Maximum Likelihood) to find an estimate of “best fit” (given the criteria).

In Bayesian statistics we will soon learn that we are not interested in P(D|H), but in P(H|D)

$$P(H \mid D) = \frac{P(D \mid H) * P(H)}{P(D)}$$

One more useful bit….

- Multiplying probabilities is a pain. Remember that if you log transform, you can add.

$P(A \cap B) = P(A)*P(B)$

$Log(P(A \cap B)) = Log(P(A)) + Log(P(B))$

Probability distributions: what they are **NOT** … the "true" distributions for real data.

- Geary (1947) - normality is a myth; there never was and never will be a normal distribution.
- This extends for all distributions
- They are an artifice; a mathematical convenience.
- We do not know the actual form of the distributions of the data. We use known probability distributions to approximate what we observe &/or predict.

So why do we use them? It's all about shape and scale!

- Because they provide a usable framework for framing our questions, and allowing for parametric methods; i.e likelihood and Bayesian.
- Even if we do not know its actual distribution, it is clear frequency data is generally going to be better fit by a binomial than a normal distribution. Why?

Why will it be a better fit?

- The binomial is **bounded** by zero and 1
- Other distributions (gamma, poisson, etc) have a lower boundary at zero.

- This provides a convenient framework for the relationship between means and variance as one approaches the boundary condition.

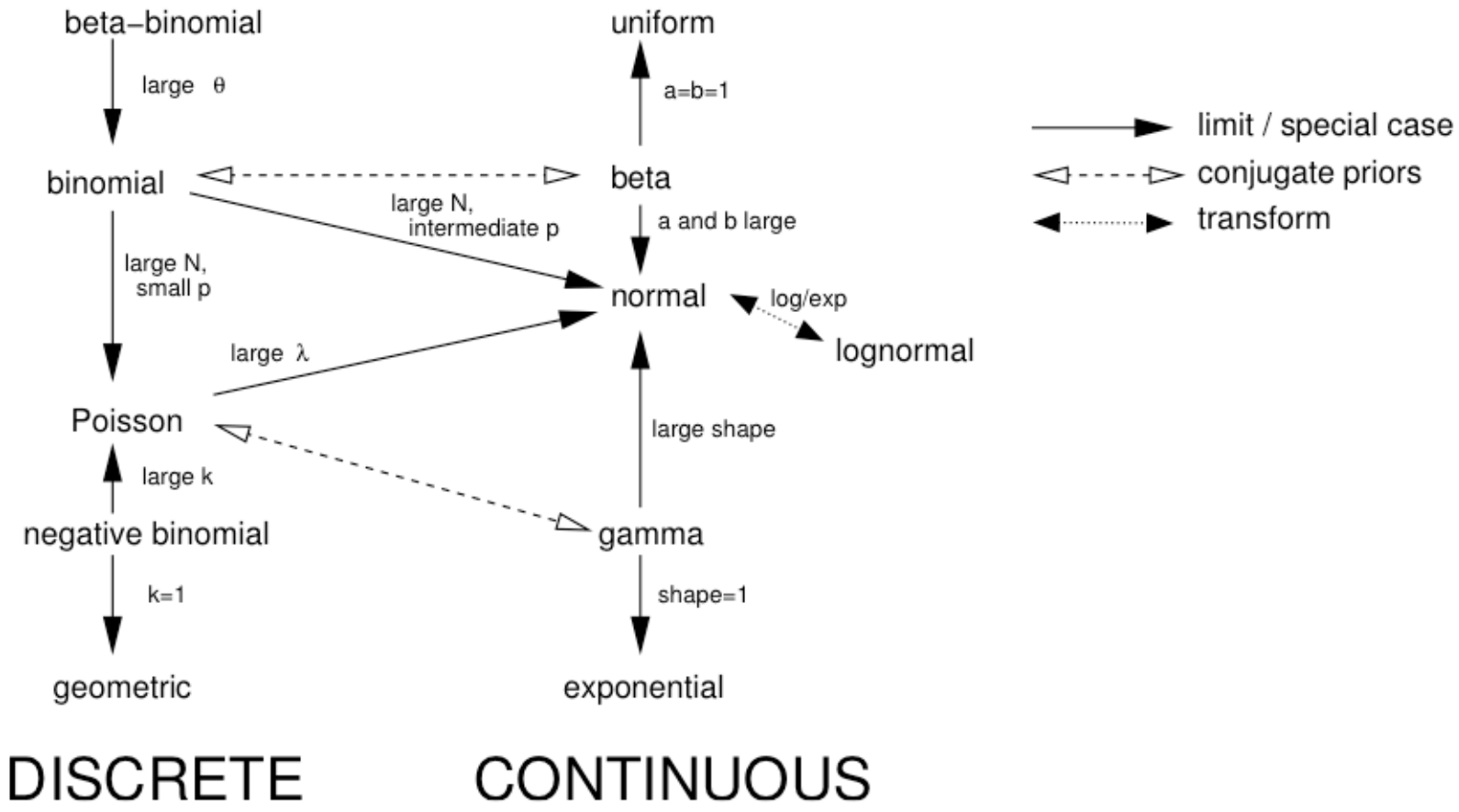
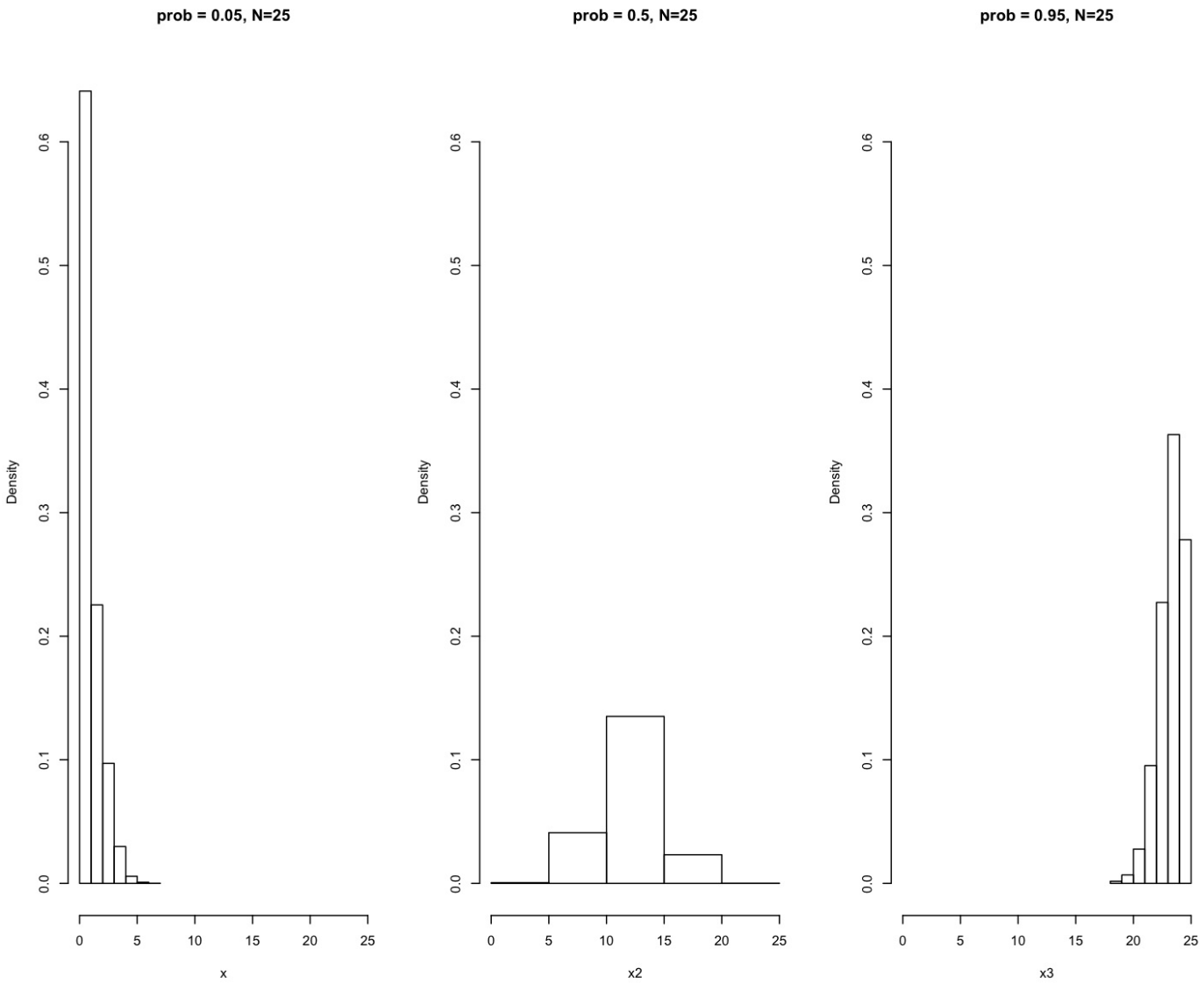


Figure 4.17 Relationships among probability distributions.

Bolker CH4 page 137

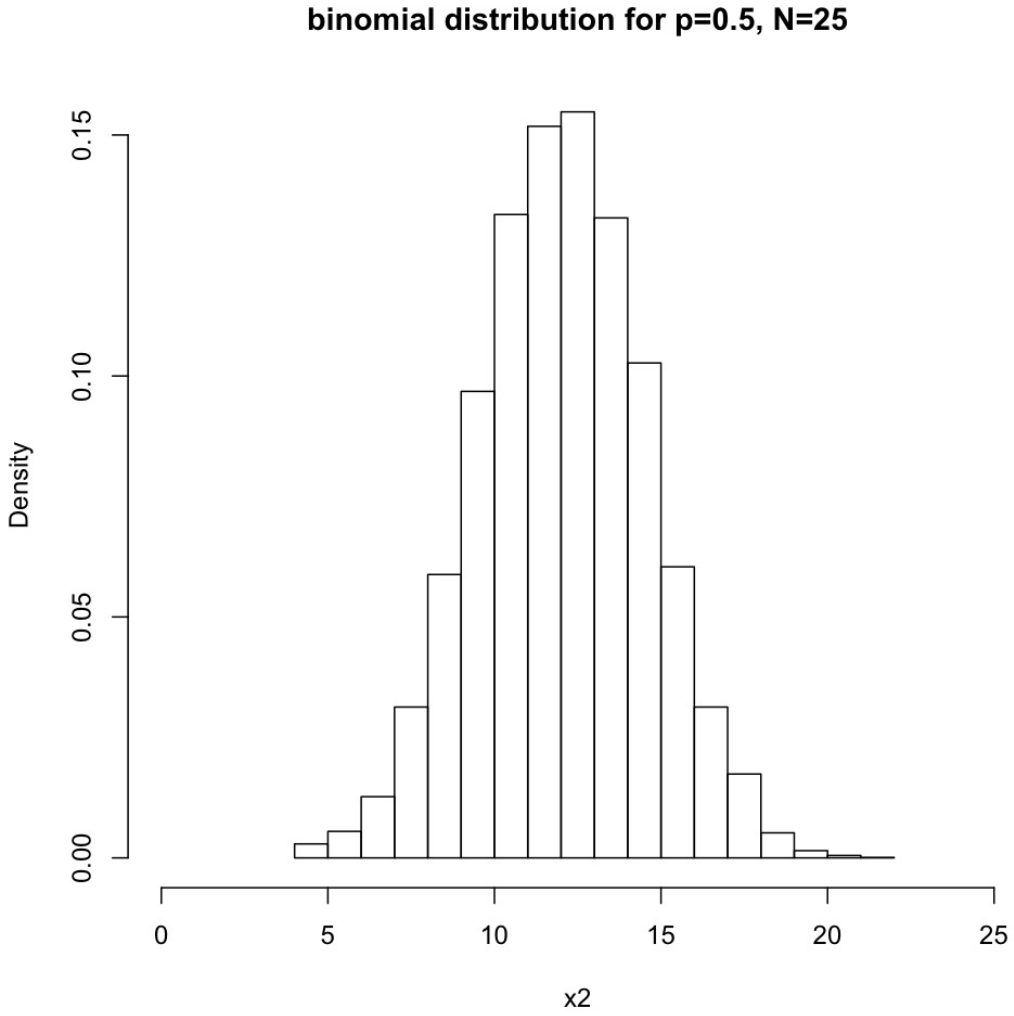
Binomial distributions



continued

- If one is examining proportions $x \sim bi(p=0.5)$ then the binomial may look sort of normal.

Binomial, p=0.5



Looks a lot like a normal distribution in the centre of its range.

continued

- Same with poisson, away from zero it looks normal
- i.e. Number of sex comb teeth (discrete) is well fit by a normal distribution (continuous).

Those pesky boundary conditions

- Why not just use a Normal distribution, where anything less than 0 goes to zero
- i.e x <- rnorm(1000, 1, 1);

  y <- ifelse(x<0, 0, x)
- This produces a ***zero inflated*** normal similar to a binomial or a…
- However, it is not not easy to work with mathematically.

The multitude of probability distributions allow us to to choose those that match our data or theoretical expectations in terms of shape location, scale.

Fitting a distribution is an art and science of utmost importance in probability modeling. The idea is you want a distribution to fit your data model "just right" without a fit that is "overfit" (*or underfit*). Over fitting models is sometimes a problem in modern data mining methods because the models fit can be too specific to a particular data set to be of broader use.

Seefeld 2007

Some things to consider

- There are three types of parameters for probability distributions; shape, scale and location.
- Some distributions have only 2 (normal or 1 (Poisson) parameters.
- While considering over and under dispersion also consider *Occam's Razor*. The simplest model that fits your data is the one to start with.

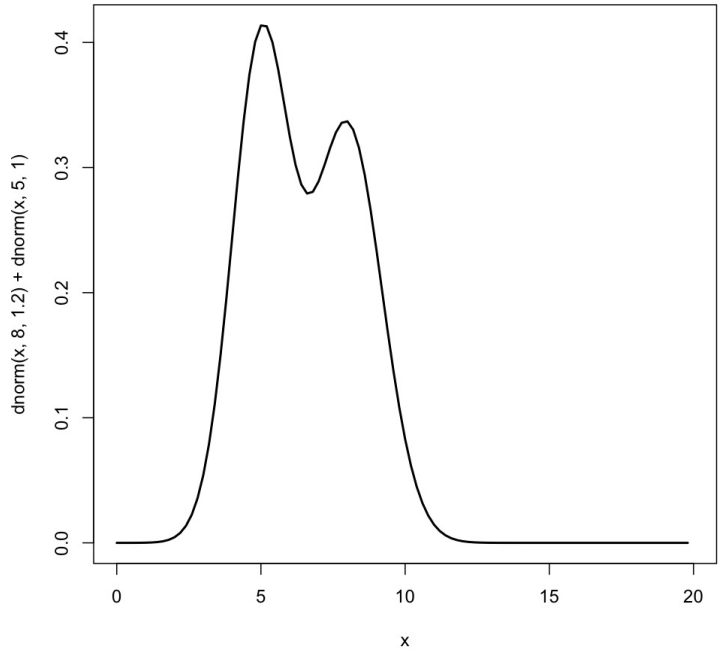
Parameters for probability distributions

- Location = mean for a normal. Scale = Standard Deviation. (NO shape, that is why it is symmetrical).

- The fact that location=mean and scale = Sd is NOT TRUE for all probability distributions.

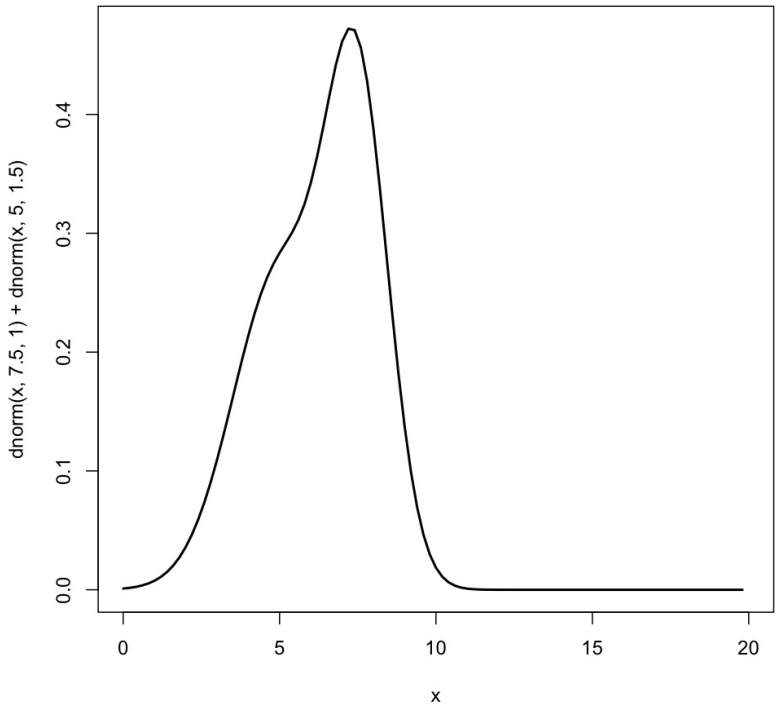
In statistical models we are interested both in modeling the data as a whole, and the distribution of the error in particular.

- If your data looked like this, what kind of distribution might you use?



In statistical models we are interested both in modeling the data as a whole, and the distribution of the error in particular.

- If your data looked like this, what kind of distribution might you use?



Both examples were the result of a mixture of two normal distributions

You could use a mixture of two normals with 4 parameters (mean and variance for each peak). However Occam's Razor suggests this may not be required as only the location/mean differs.

An important point is that you may not need to fit your whole data set to a particular probability distribution, per se. Instead remember that there may be other model parameters (I.e. a mean for each peak) to consider, with a single underlying error distribution.

Keep in my mind..

- It is not the distribution of the whole data set that will necessarily determine what distribution to use to model it. Instead we are often more concerned with the distribution of the residual variation once we have accounted for all the parameters that we are estimating.