Probability!

Probability - The fraction of observations of an event given multiple repeated independent observations.

A Feeding Trial Example

Let's say you've offered 50 budworms a leaf to eat. 45 eat. P(eats) $= \frac{45}{50} = 0.9$

Now you offer 50 others a treated leaf. 10 eat. P(eats) $= \frac{10}{50} = 0.2$

Probability of NOT doing something

What is the probability of not eating if you are fed a treated leaf?

P(! eats) $= 1 - \frac{10}{50} = 0.8$

P(!A) = 1-P(A)

Probability of Exclusive Events

What if we offered our budworms both a treated and untreated leaf? 20 eat the control, 5 eat the treated leaf.

P(eats) $= \frac{20}{50} + \frac{5}{50} = 0.5$

P(A or B) = P(A) + P(B)

Two Events

We offer our budworms a leaf. 45 eat it. Then we offern them seconds. 20 of the original 45 each the second leaf.

P(eats twice) $= \frac{20}{50} = 0.4$

$= \frac{45}{50} * \frac{20}{45}$

P(A and B) = P(A)P(B)

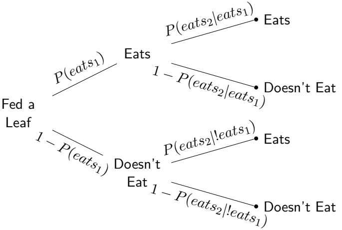
Two Conditional Events

If we are interested in the probability of eating twice - i.e. the probability of eating a second time *given* that a budworm ate once, we phrase that somewhat differently.

$P(eats_2|eats_1)$

So, P(A and B) = P(A)P(B$\|A$)

Probability Tree



$P(eats_2) = P(eats_2|eats_1) + P(eats_2|!eats_1)$

Bayes Theorem

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

So...
$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Rev. T. Bayes

GMO Collateral Damage

Let's say you you had a rare but extremely harmful budworm munching ravenously through your fields. You've developed a really effective GMO tobacco leaf to help stop it. It has a 75% kill rate. And, miraculously, it only has a 15% kill rate of non-budworms. Given that the budworms make up about 10% of the insects in a field, what's porportion of dead insects WON'T be budworms?

$P(!W|D) = 1 - P(W|D)$

$P(W|D) = \frac{P(D|W)P(W)}{P(D)}$
$= \frac{P(D|W)P(W)}{P(D|W)P(W)+P(D|!W)P(!T)} = \frac{0.75*0.1}{0.75*0.1+0.15*0.9} = 0.357$

$1 - 0.357 = 0.643$ - the majority of the dead!

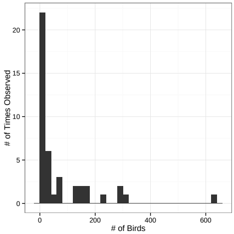
Why are we talking about this??

As we test hypotheses in a *frequentist* framework, we'll be asking about the probability of observing data given that a hypothesis is true - P(Data $||Hypothesis$).

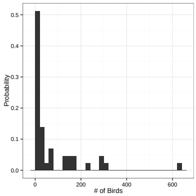
Distributions!

(when a point probabilty just ain't enough)

Frequency Distributions Make Intuitive Sense

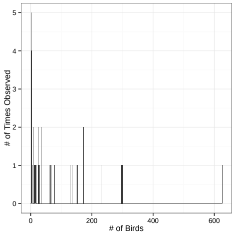


Frequencies Can be Turned Into Probabilities



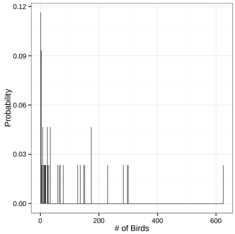
Just divide by total # of observations
But - we have binned observations...

Frequencies of Individual Observations



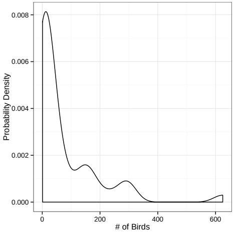
Can we turn these into probabilities?

Probabilities of Individual Measurements



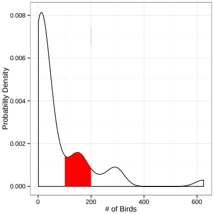
Many probabilities small, and what about the gaps?

Continuous Probability Distributions



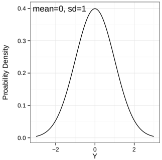
Any individual observation has a *probability density*.

Probability as Integral Under the Curve



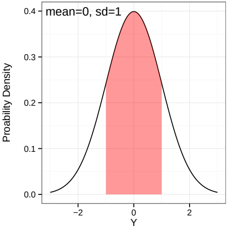
We obtain probabilities of observations between a range of values by integrating the distribution over selected values.

The Normal Distribution

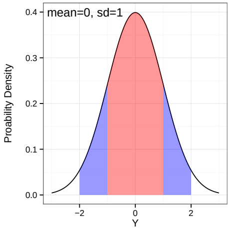


- Defined by it's mean and standard deviation.
- Y ~N(μ, σ)
- Single mode
- Symmetric

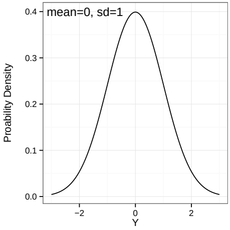
67% of Values within 1 SD



95% of Values within 2 (1.96) SD

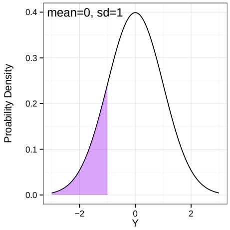


How to Get A Probability Density in R



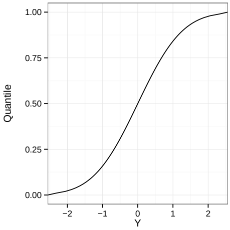

```
dnorm(Y, mean = 0, sd = 1)
```

The Probability of a Value or More Extreme Value




```
pnorm(Y, mean = 0, sd = 1)
```

The Cummulative Distribution/Quantile Function



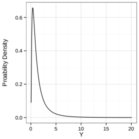

```
qnorm(p, mean = 0, sd = 1)
```

The Cummulative Distribution/Quantile Function

pnrm an dqnorm are the inverse of one another

```
pnorm(-1)

## [1] 0.1587

qnorm(pnorm(-1))

## [1] -1

qnorm(0.025)

## [1] -1.96
```

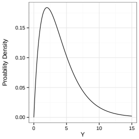
The Lognormal Distribution



- An exponentiated normal
- Defined by the mean and standard deviation of its log.
- Y ~LN(μ_{log}, σ_{log})
- Generated by multiplicative processes

```
dlnorm(Y, meanlog = 0, sdlog = 1)
```

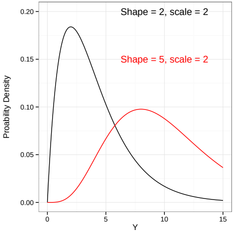
The Gamma Distribution



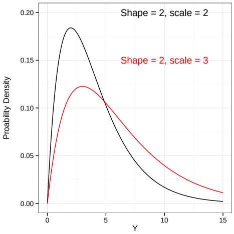
- Defined by number of events(shape) average time to an event (scale)
- Can also use rate (1/scale)
- Y ~G(shape, scale)
- Think of time spent waiting for a bus to arrive

```
dgamma(Y, shape = 2, scale = 2)
```

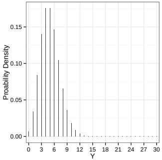
Waiting for more events



Longer average time per event



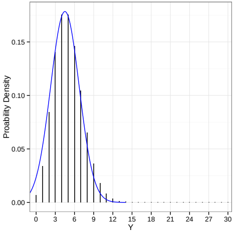
The Poisson Distribution



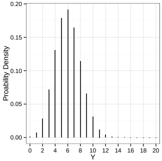
- Defined by λ - the mean and variance
- Y ~ P(lambda)

```
dpois(Y, lambda = 5)
```

When Lambda is Large, Approximately Normal



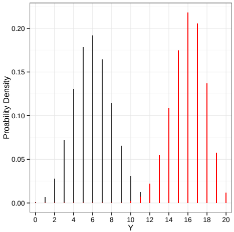
The Binomial Distribution



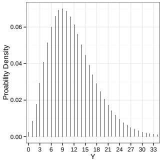
- Results from multiple coin flips
- Defined by size (# of flips) and prob (probability of heads)
- Y ~ B(size, prob)
- bounded by 0 and size

```
dpois(Y, size, prob)
```

Increasing Probability Shifts Distribution



The Negative Binomial Distribution



- Distribution of number of failures before n number of successes in k trials
- Or mean # of counts, μ, with an overdispersion parameter, size
- Y ~ NB(μ, size)

```
dnbin(Y, mu, size)
```

Exercise

- Explore the distributions we have discussed
- Examine how changing parameters shifts the output of probability function
- Compare curves generated using density functions (e.g., dnorm) and large number of random draws (e.g. from rnorm)
- Overlay these in plots if you can (hist, lines, etc.)
- Challenge: graphically show integration under the different types of distribution curves (?polygon or ?geom_ribbon)