GG501

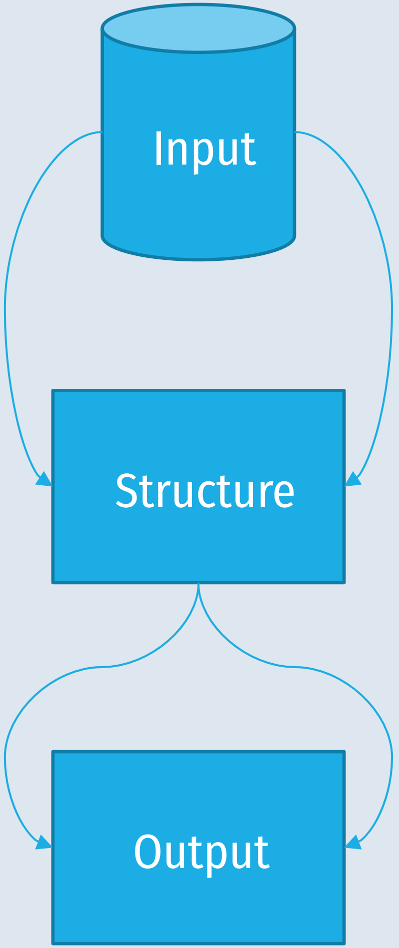
6. Model data

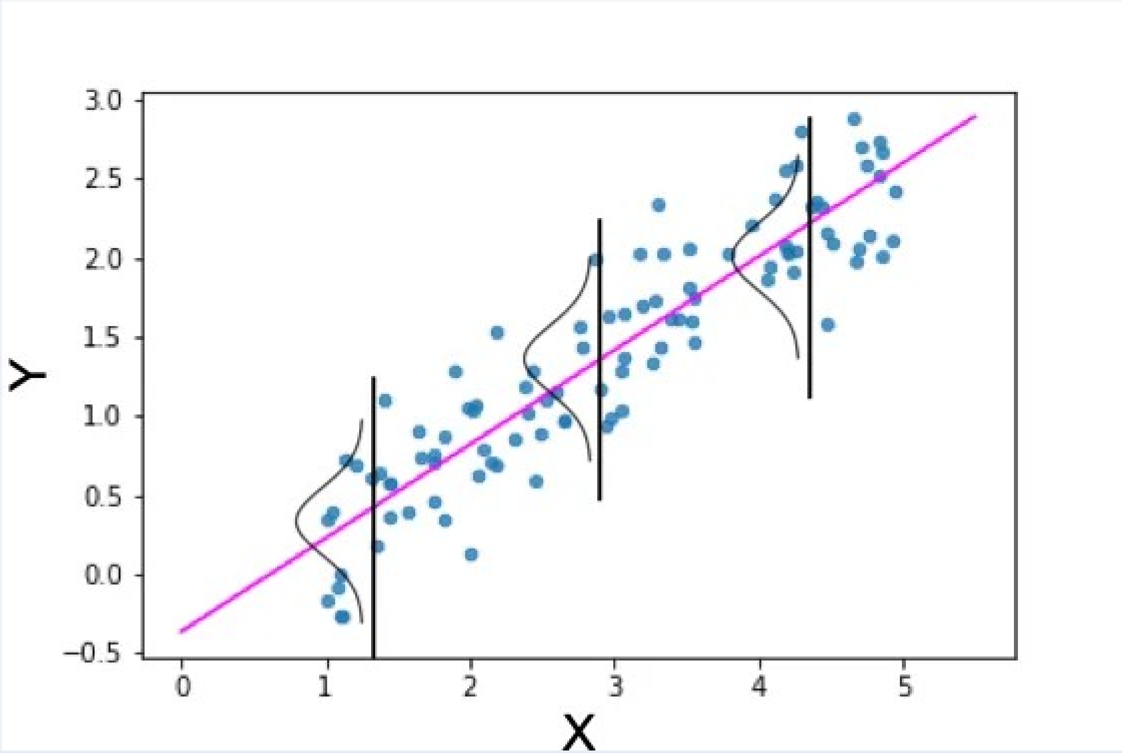
Models in R

- What do we mean by ‘model’?

Many variety of models

- Each take some input data
- Attempt to generalize about the underling data-generating-process
- Can be used for a variety of purposes
  - description
  - explanation
  - prediction





https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

Statistical vs probability model

- Statistical model
  - Describes one or more variables & their relationship
- Probability model
  - Describes outcome of random *event*
  - Sometimes called a random variable

Describing randomness

- Random event/variable
  - Sample space - what are the possible outcomes?
- Probability model
  - Assign probability to each member of sample space
    - For a coin toss this is 0.5 for heads & 0.5 for tails
- Purely random
  - Probability model contains all the information
    - *No explanatory variables needed to account for variation*

Settings for probability models

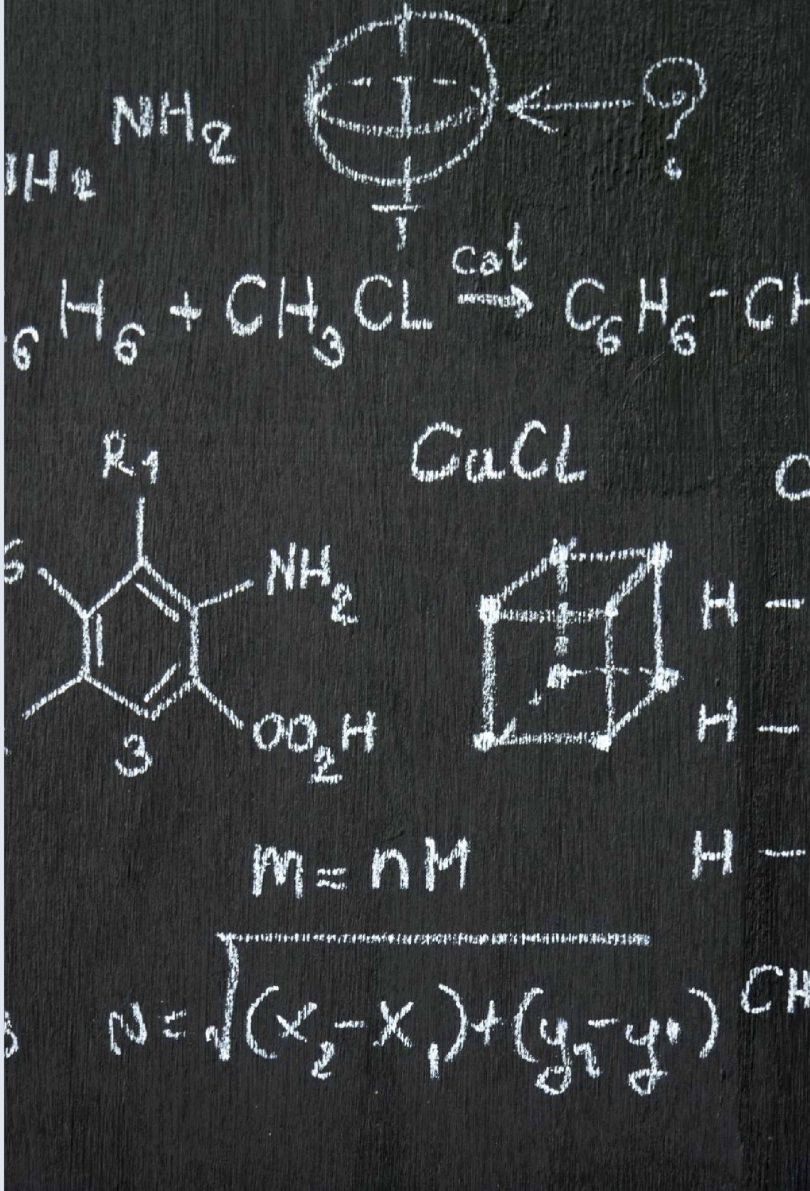
- How is the event configured/measured/represented?
- Examples
  - Number of radioactive particles per minute
  - Student test score on standardized test
  - Number of blood vessels in microscope slide
  - Number of people who support a candidate in a random sample
- Must pick form of model that fits setting
  - Combines expert knowledge & probability calculus

Discrete vs Continuous

- Helpful to distinguish between two kinds of sample spaces
  - Discrete numbers - outcome of rolling a die
  - Continuous numbers - any value in a range
- Possible to assign a probability to each outcome for discrete numbers
- Possible to assign probability to range of outcomes for continuous numbers

Probability density

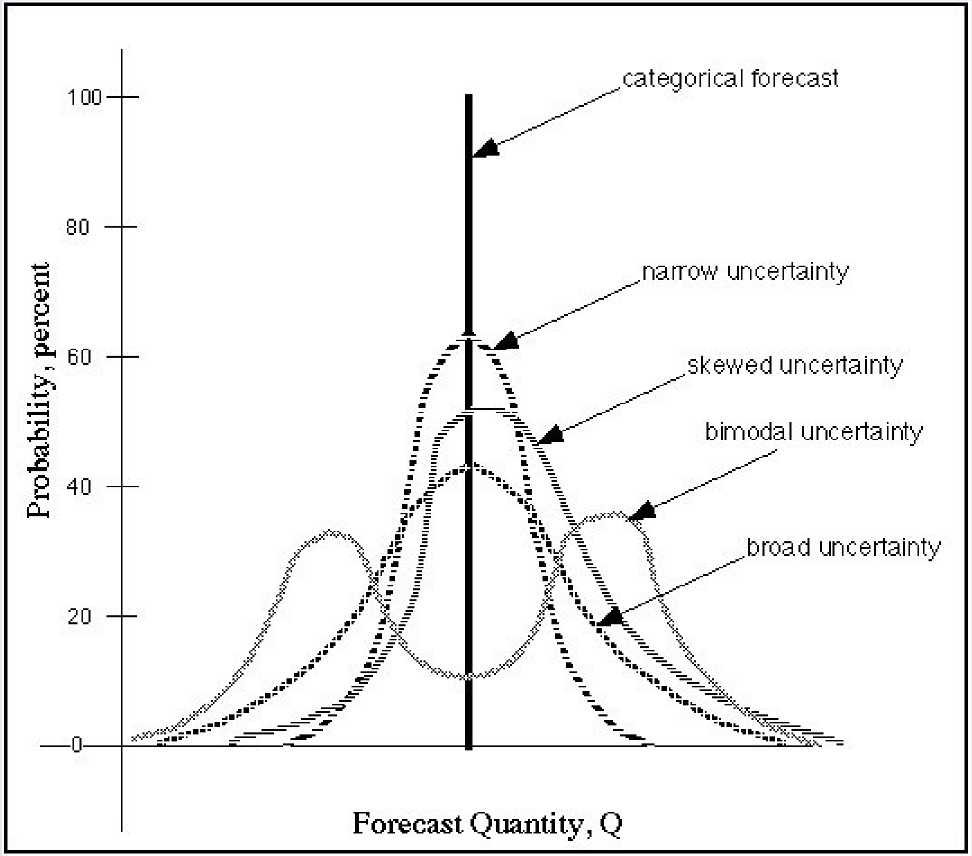
- Can assign probability density to each outcome by dividing probability by extent of range
  - Usually treat this probability density as function of value of random value: *p(x)*
- Often use probabilities & probability densities in a similar way
  - For discrete sample space, assigned probabilities over all the members of space must add to 1
  - For continuous sample space, integral over assigned probability over possible values must be 1



Multiple views of probability

- Frequentist view of probability
  - Describe how often outcomes occur
    - Example - 100 coin flips should lead to 50 heads
  - Based on large number of possible trials
- • Subjectivist view of probability
  - Encodes modeller's assumptions/beliefs
  - Assess degree of belief
  - Probability is assigned to a hypothesis

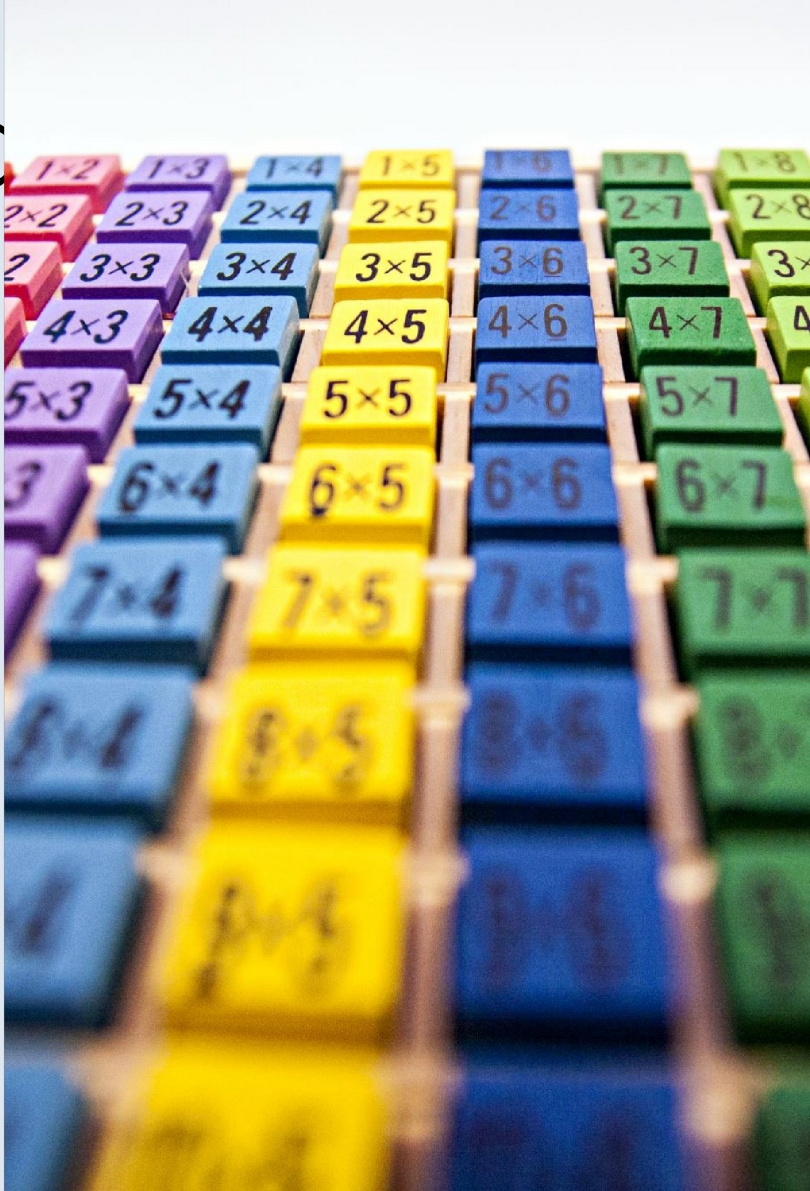
“It will snow today …”



http://www.hpc.ncep.noaa.gov/wwd/winter_wx.shtml

Standard probability moc

- Small set of standard probability models apply to wide range of settings
- Don't need to derive them!
- Each model has parameters that need to be adjusted
  - Parameters are similar to coefficients from regression models



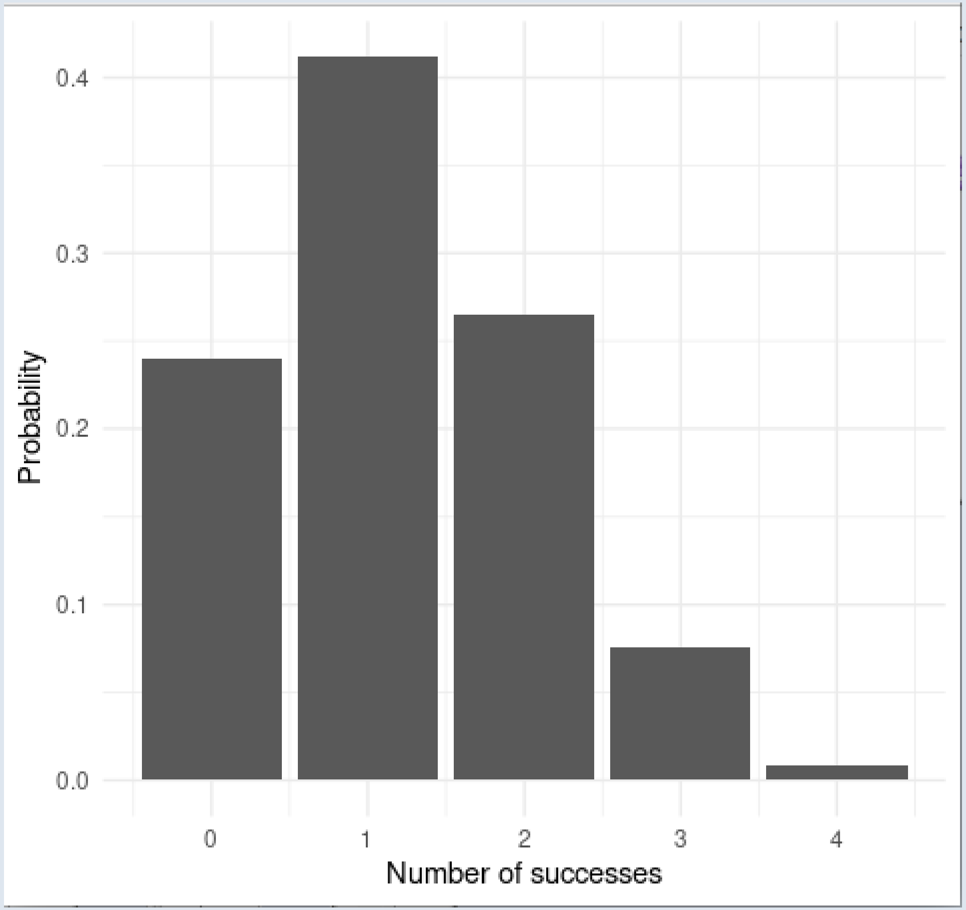
Discrete

- Equal probabilities
  - Examples - die toss, coin flip, distributions of ranks of any continuous variable
  - Parameter
    - size - how many possibilities

Discrete

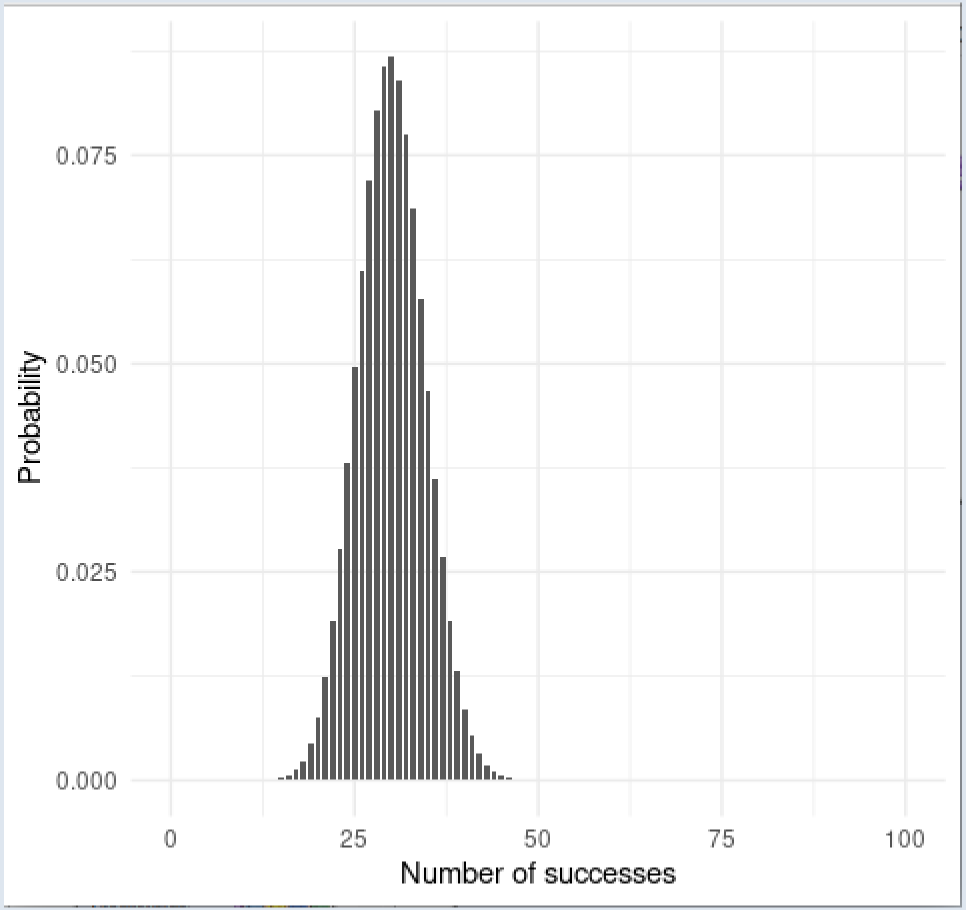
- Binomial
  - Example - trials of coin flip where outcome is count of “successes” or “heads” or “1s”
  - Parameters
    - size - number of trials
    - prob - probability of success on each trial

- Binomial
  - n=4, p=0.30




```
dbinom(0:4, size=4, p=0.3) %>% as_tibble() %>%
ggplot(aes(x=0:4, y=value)) + geom_bar(stat="identity") +
labs(x="Number of successes", y="Probability")
```

- Binomial
    - n=100, p=0.30

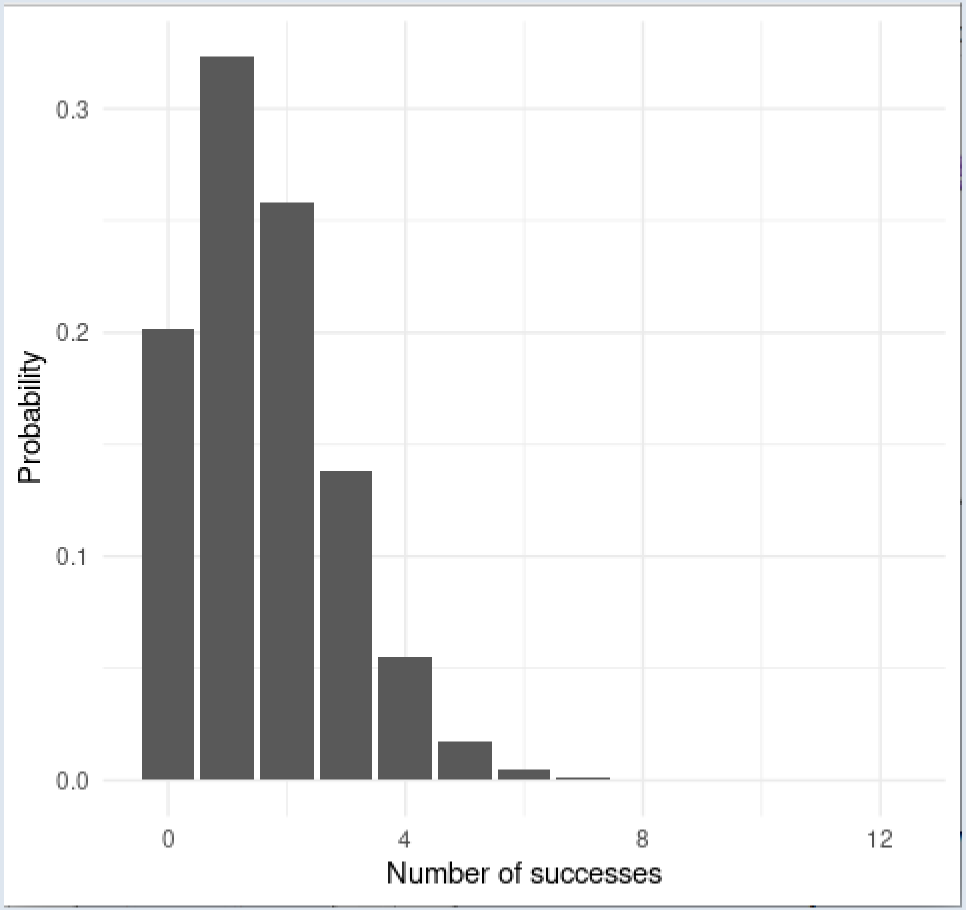



```
dbinom(0:100, size=100, p=0.3) %>% as_tibble() %>%
ggplot(aes(x=0:100, y=value)) + geom_bar(stat="identity")
+ labs(x="Number of successes", y="Probability")
```

Discrete

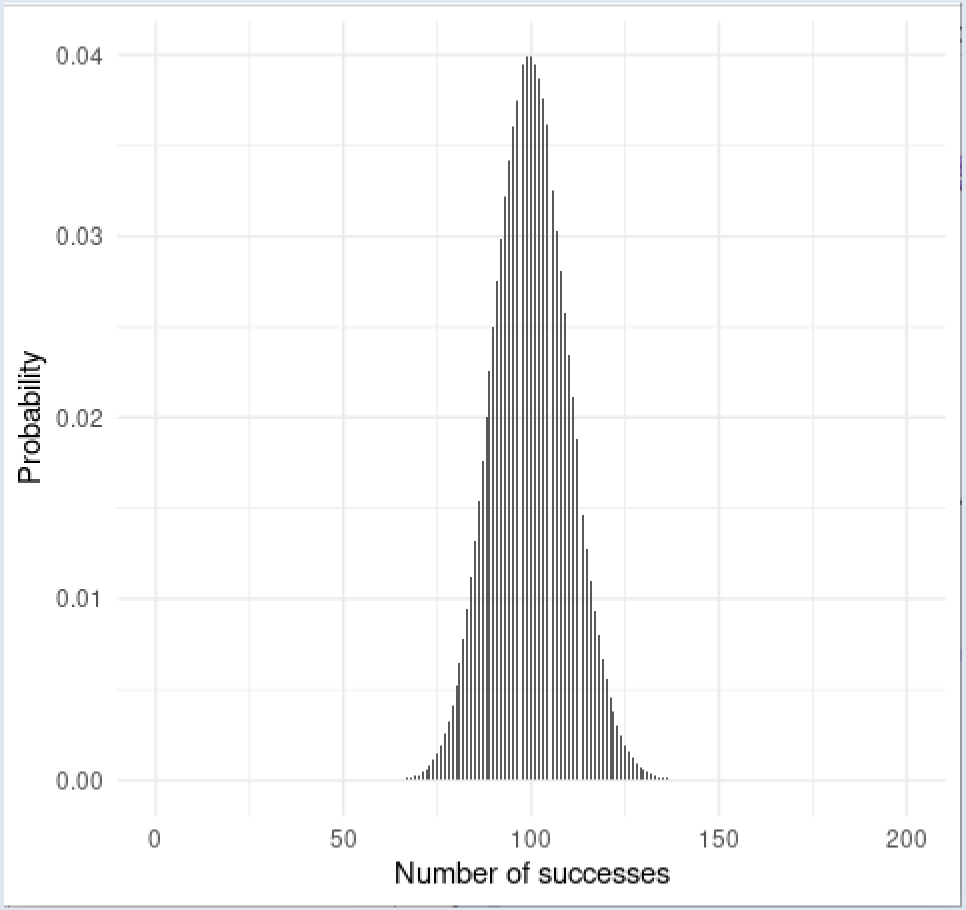
- Poisson
  - Number of events that happen in a given time
  - Example - number of cars passing by a point, number of shooting stars in minute, number of snowflakes that land on a glove in a minute
  - Parameter
    - lambda - mean number of events

- Poisson
  - lambda(rate)=0.1.6



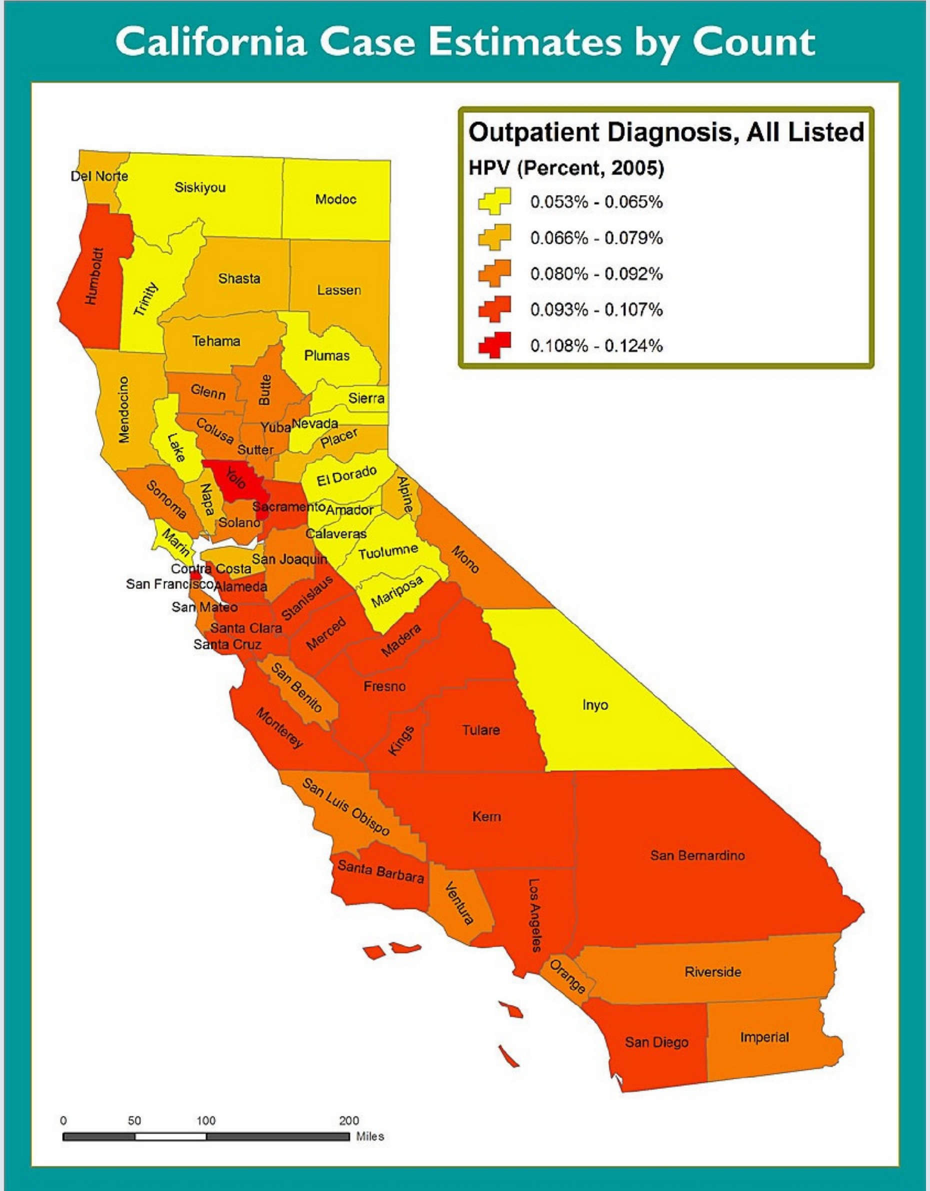

```
dpois(x=0:12,lambda=1.6) %>% as_tibble() %>%
ggplot(aes(x=0:12, y=value)) + geom_bar(stat="identity")
+ labs(x="Number of successes", y="Probability")
```

- Poisson
  - lambda(rate)=0.1.6




```
dpois(x=0:200,lambda=100) %>% as_tibble() %>%
ggplot(aes(x=0:200, y=value)) + geom_bar(stat="identity")
+ labs(x="Number of successes", y="Probability")
```

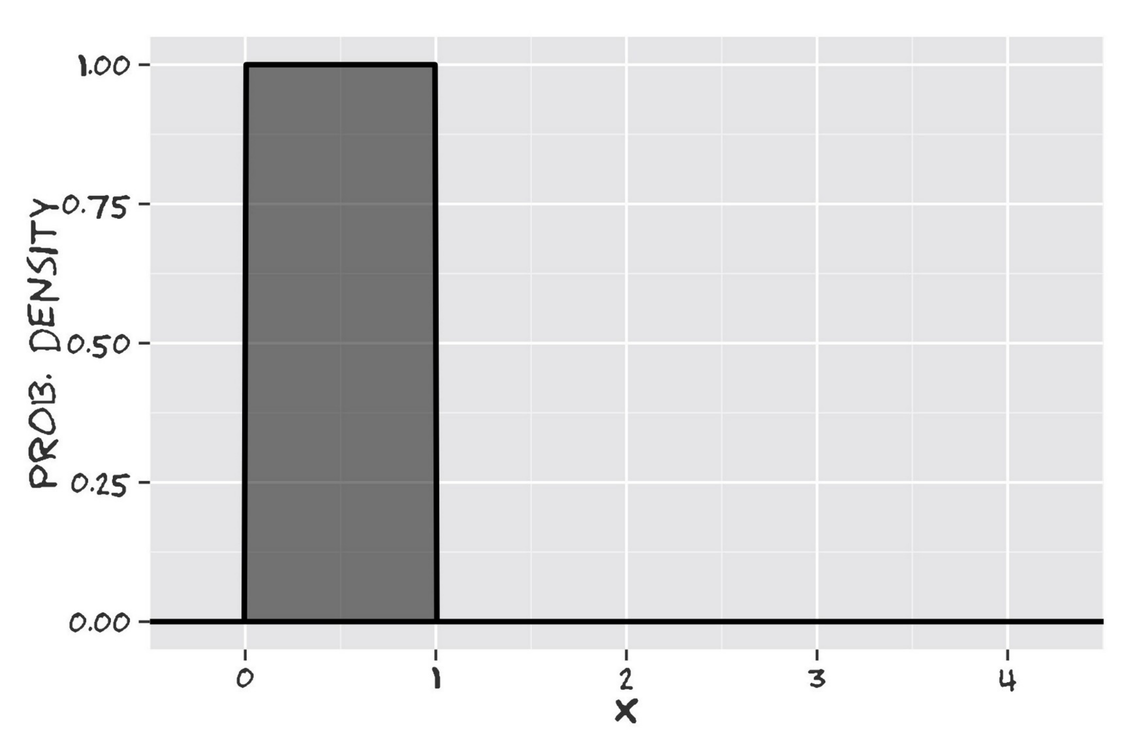
Map of California human papilloma virus (HPV) cases by outpatient diagnosis in each county



Continuous

- Uniform
  - Parameters: Max and min
  - Models: spinners (angles in 2-d, but not higher), p-values under the Null Hypothesis

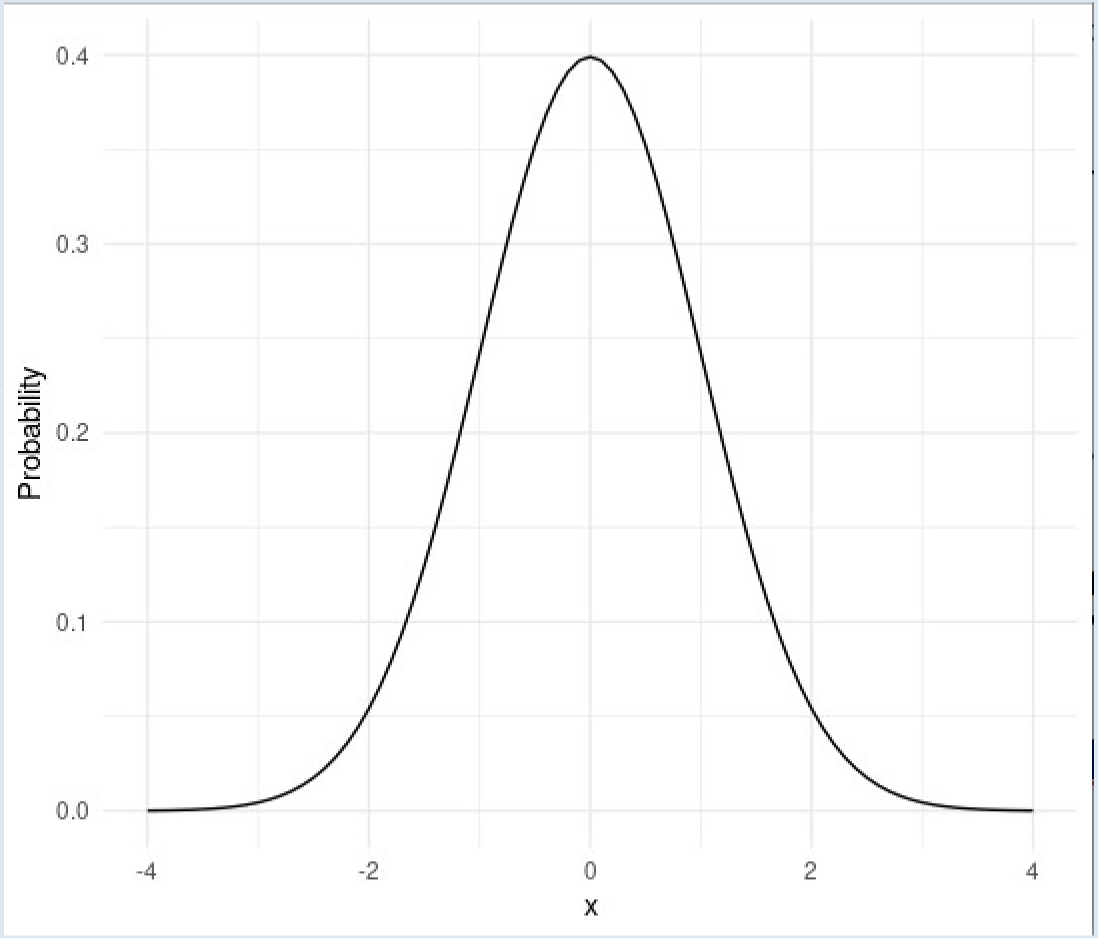
- Uniform
  - min=0, max=1



Continuous

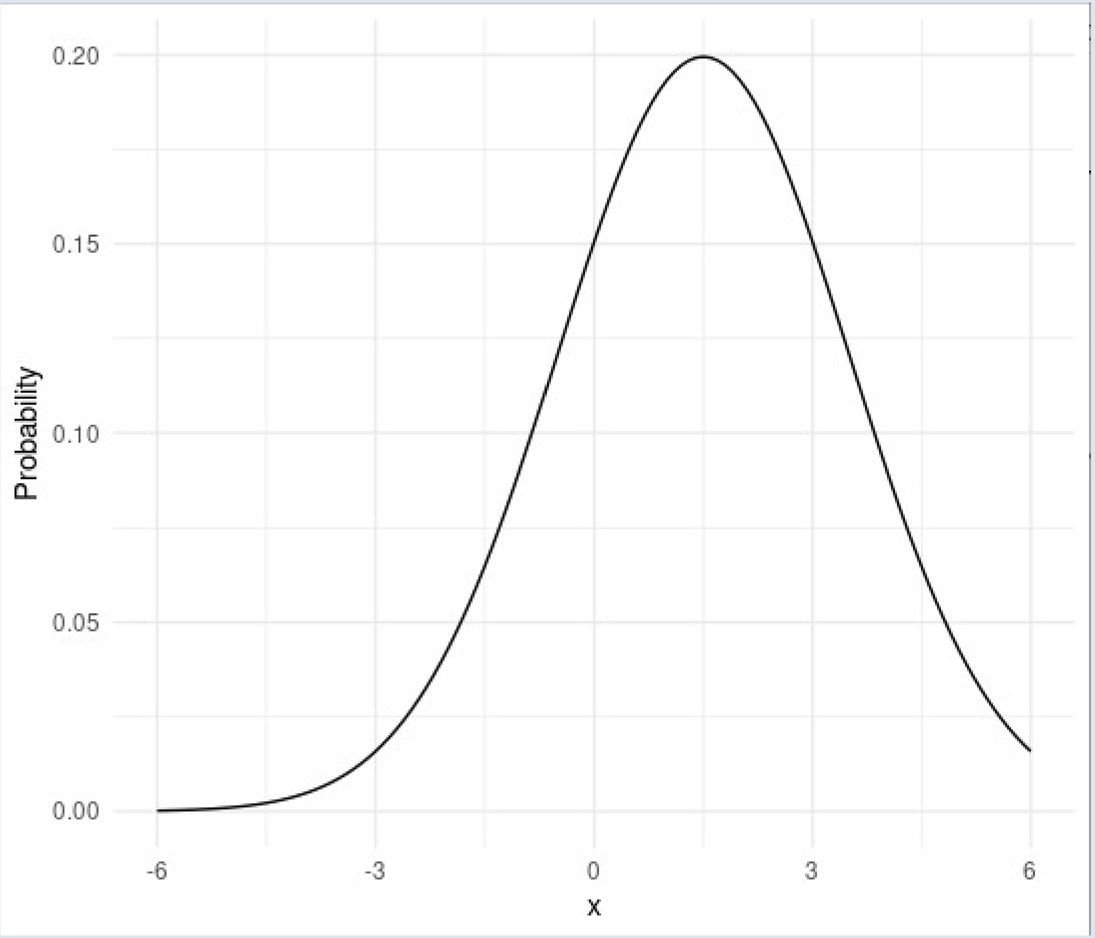
- Normal (gaussian)
  - Parameters: mean and sd
  - Models: your general purpose model

- Normal
  - mean=0, sd=1




```
dnorm(x=seq(-4,4,0.1), mean=0, sd=1) %>% as_tibble() %>%
ggplot(aes(x=seq(-4,4,0.1), y=value)) + geom_line() +
labs(x="x", y="Probability")
```

- Normal
  - mean=1.5, sd=2




```
dnorm(x=seq(-6,6,0.1), mean=0, sd=1) %>% as_tibble() %>%
ggplot(aes(x=seq(-6,6,0.1), y=value)) + geom_line() +
labs(x="x", y="Probability")
```

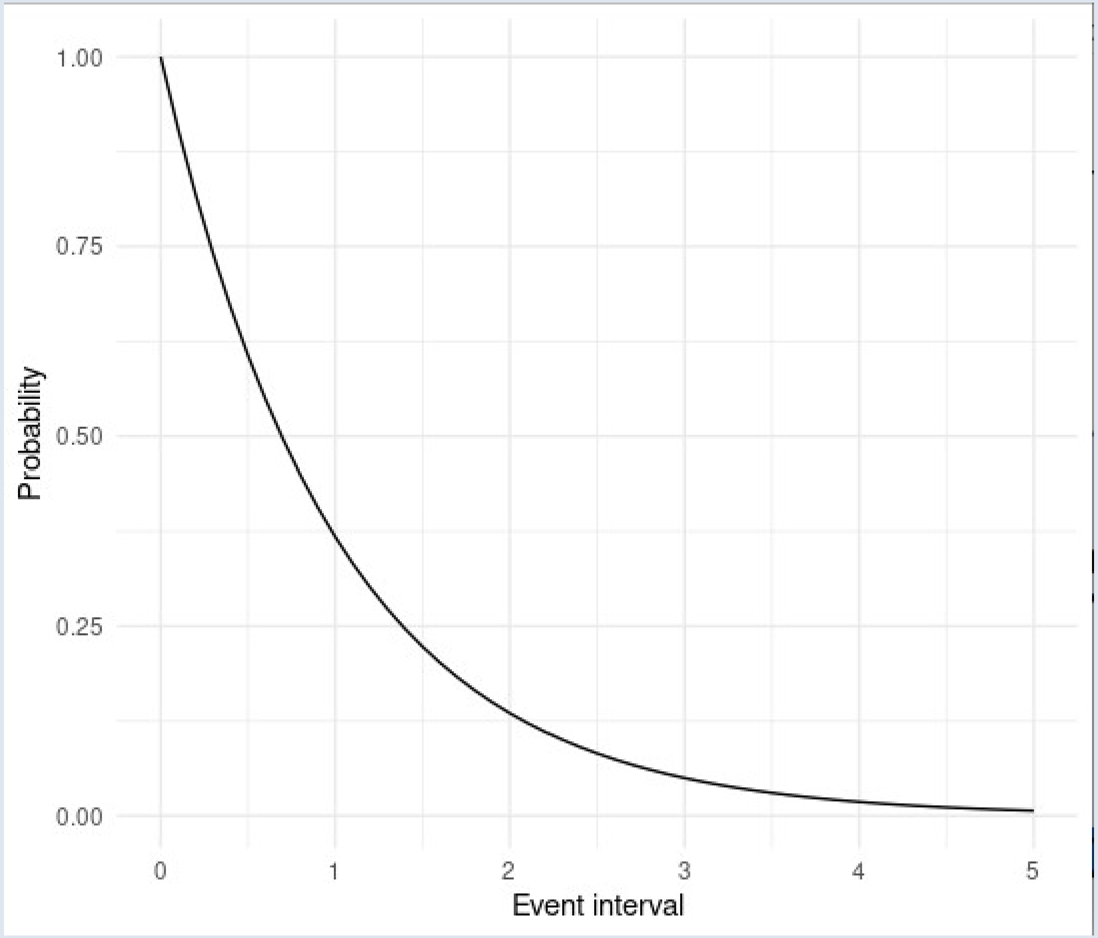
What can we do with probability models?

- Percentiles: what is the range of values within some probability range?
  - e.g., 90th percentile: value of the random variable such that 90% of the time values will be equal or smaller
- Quantiles: What is the percentile of a given outcome?
  - e.g., how unusual is this observation given the underlying probability model?

Continuous

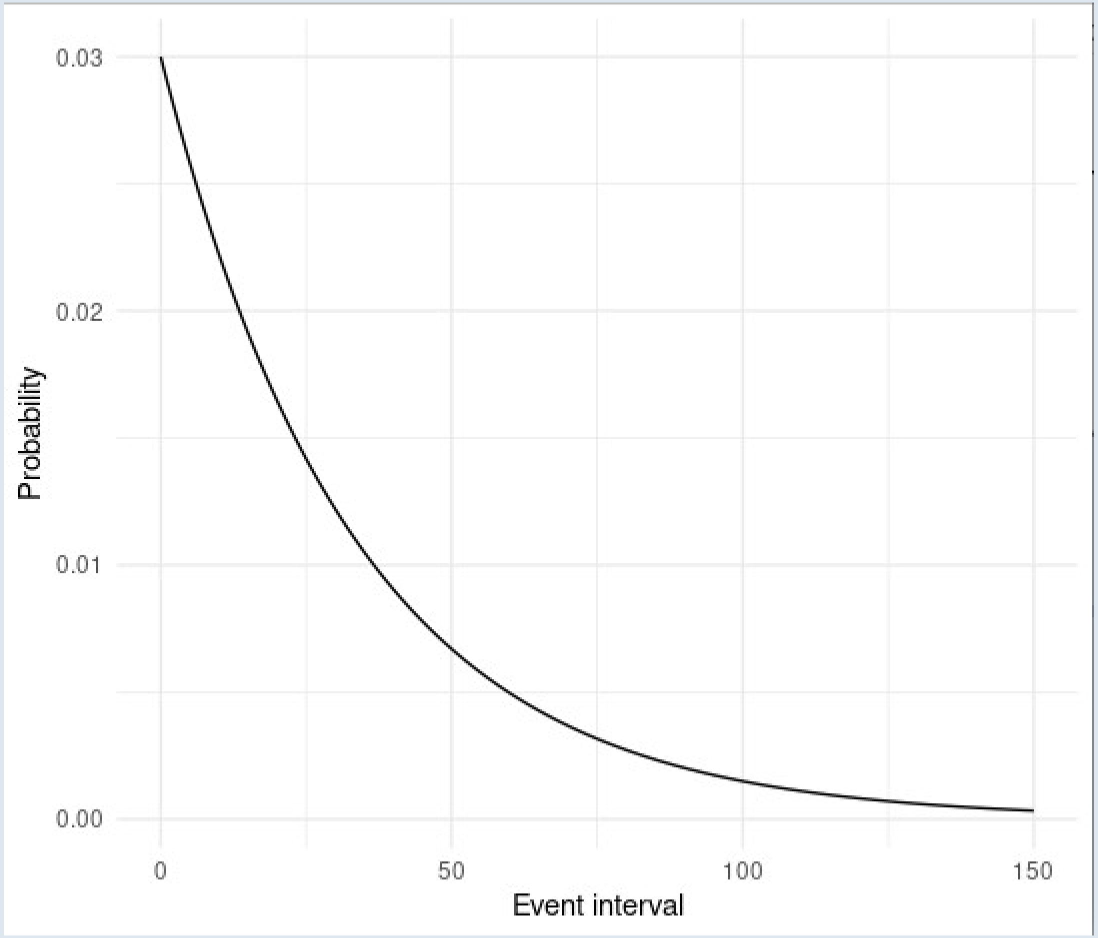
- Exponential
  - Times between random events (earthquakes, storms, floods)
  - Parameters: rate (mean time is 1/rate)

- Exponential
  - rate=1




```
dexp(seq(0,5,0.1), rate=1)%>% as_tibble() %>%
ggplot(aes(x=seq(0,5,0.1), y=value)) + geom_line() +
labs(x="Event interval", y="Probability")
```

- Exponential
  - rate=1

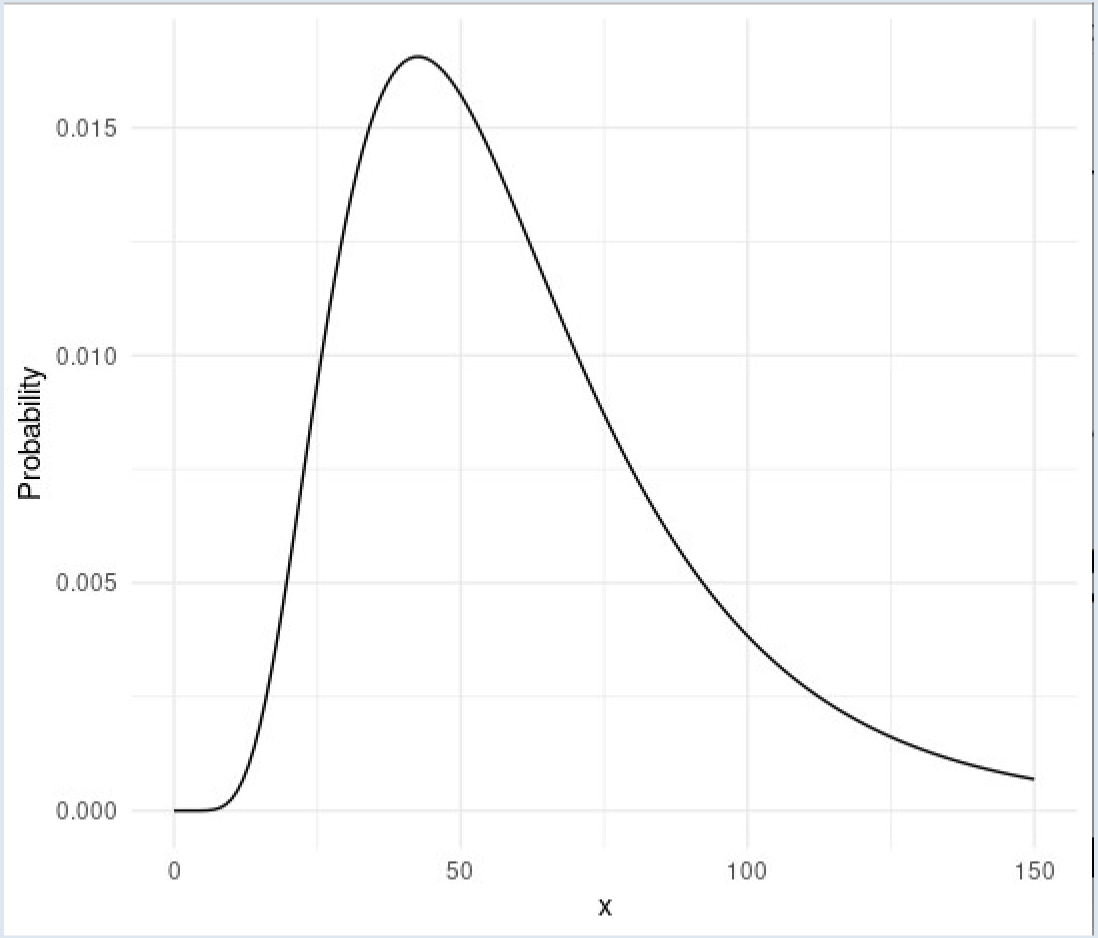



```
dexp(seq(0,150,1), rate=0.03)%>% as_tibble() %>%
ggplot(aes(x=seq(0,150,1), y=value)) + geom_line() +
labs(x="Event interval", y="Probability")
```

Continuous

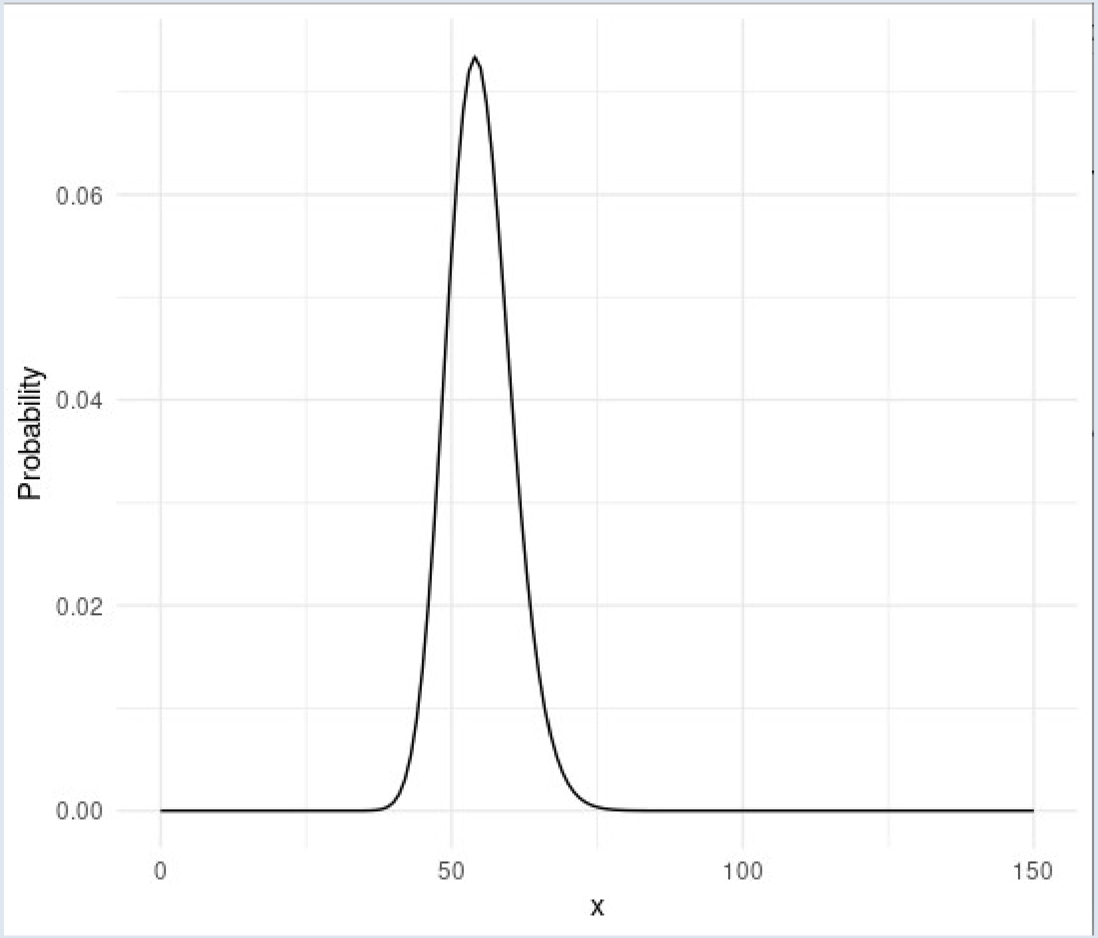
- Lognormal
  - Models: useful when skew is important
  - Parameters: mean, sd (of the log of the values)

- lognormal
  - meanlog=4, sdlog=0.5




```
dlnorm(seq(0,150,1), meanlog=4, sdlog=0.5) %>%
as_tibble() %>% ggplot(aes(x=seq(0,150,1), y=value)) +
geom_line() + labs(x="x", y="Probability")
```

- lognormal
  - meanlog=4, sdlog=0.1



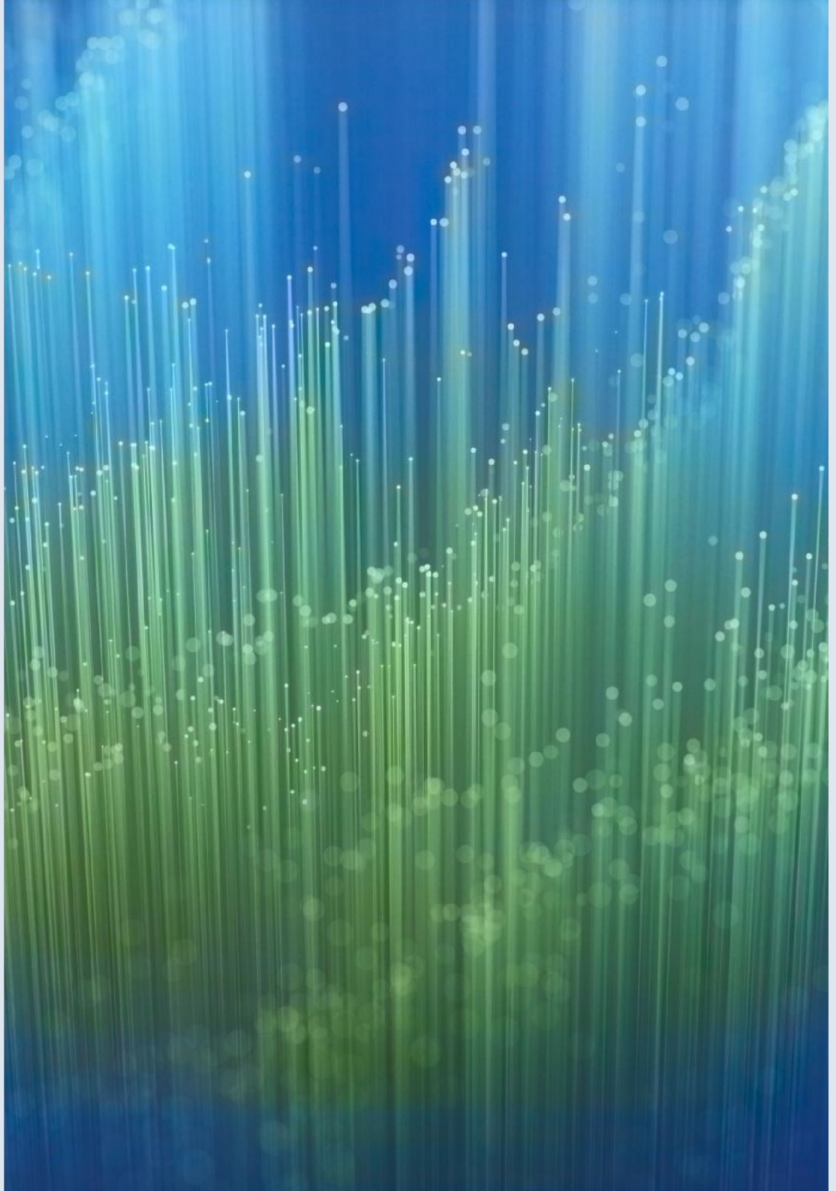

```
dlnorm(seq(0,150,1), meanlog=4, sdlog=0.1) %>%
as_tibble() %>% ggplot(aes(x=seq(0,150,1), y=value)) +
geom_line() + labs(x="x", y="Probability")
```

Why the normal distribution?

- Many chance phenomena are at least approximately described by a normal probability density function
- Example
  - Collect 1000 snowflakes & weigh them, would find distribution of weights accurately described by a normal curve
  - Measure the strength of bones in wildebeests; likely to find they are normally distributed

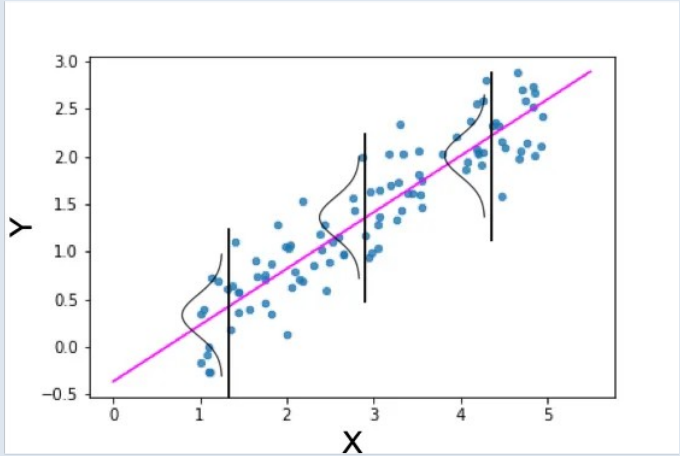
Normal nature

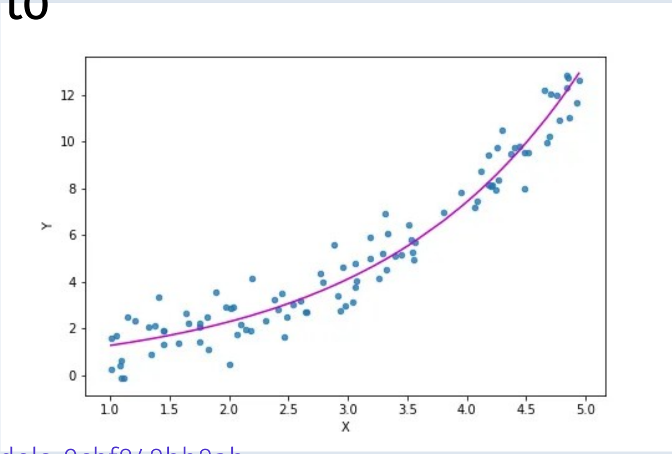
- Random biological or physical processes affected by *large* number of random processes with *individually* small effects…
- Sum of all these random components creates random variable that converges on normal distribution
- Regardless of the underlying distribution of processes causing the small effects!

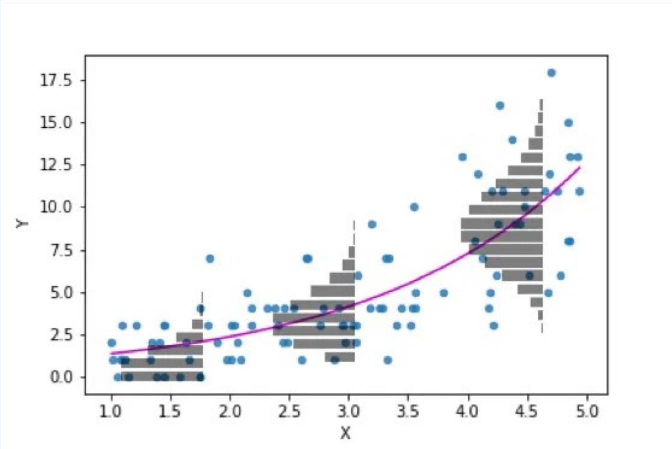


What does this have to do with modelling?

- The relationships in models are based on sample data – therefore have randomness
- They could change if repeated
- Need to use randomness to evaluate precision of models
- We can use visualization to better understand the characteristics of a given model







https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab

Models are important

- Each take some data
- Attempt to generalize about the underling data-generating-process
- Visualization of models is key to understanding what you can do with information embedded within them
  - regression trees
  - Kaplan-Meier plots, ROCs, marginal/conditional effects