Statistical Thinking in Biology Research

Probability and Statistical Inference

Terry Neeman

Australian National University

30th July 2020

A few key ideas

- Probability: understanding possible outcomes under a set of "rules"
- Domain of probability: mathematics ("theoretical", "proof")
- Statistics: Given a set of outcomes, what can we *infer* about the possible rules?
- Domain of statistics: real world data ("pragmatic", "heuristic")

Probability and Statistics are two sides of the same subject.

Probability: measures of uncertainty

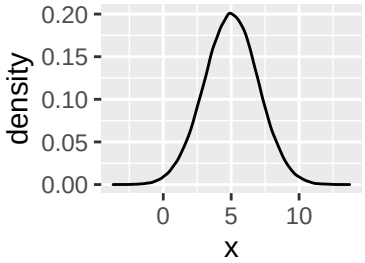
- Sample space: space of possible outcomes
- Distribution: relative frequencies (probabilities) of each outcome
- Summaries of distributions: average (expected) outcome, variation around average

Examples of common distributions in biological research

- Normal distribution
  - family of distributions
  - defined by two parameters: mean and standard deviation (variance)
  - many biological measures normally distributed, e.g. height, weight

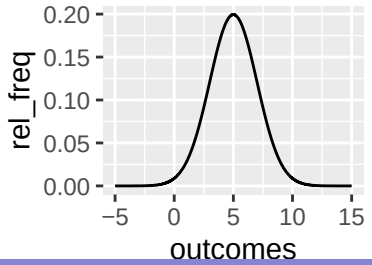
Sample from a normal distribution

```
library(tidyverse)
sample_normal <- tibble(x = rnorm(n=1e5, mean = 5, sd = 2))
ggplot(sample_normal, aes(x = x))+
   geom_density()
```



Plot the THEORETICAL Normal distribution

```
outcomes<-seq(-5,15, length.out = 1e4)
out_normal <- tibble(outcomes = outcomes,
                     rel_freq = dnorm(outcomes, mean = 5, sd = 2))
ggplot(out_normal, aes(x=outcomes,y = rel_freq))+
   geom_line()
```

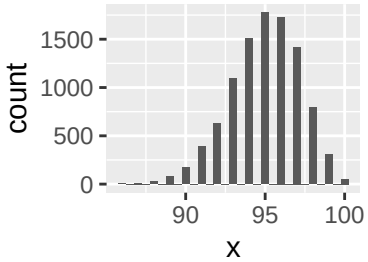


Examples of common distributions in biological research

- Binomial distribution
  - family of distributions
  - Describes potential outcomes: #successes out of n independent trials
  - defined by two parameters:
    - n = # of independent trials
    - p = probability of success in a trial

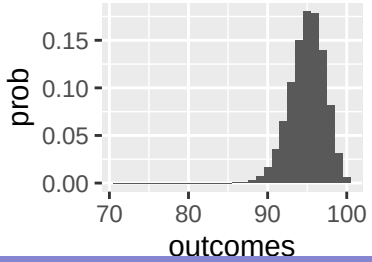
Sample from a binomial distribution

```
sample_binomial <- tibble(x = rbinom(1e4, size = 100, prob = 0.95))
ggplot(sample_binomial, aes(x = x))+
   geom_histogram(binwidth = 0.5)
```



Plot the THEORETICAL binomial distribution

```
outcomes<-seq(71,100, by=1)
outcomes_binomial <- tibble(outcomes = outcomes,
                            prob = dbinom(outcomes, size=100, prob=0.
ggplot(outcomes_binomial, aes(x=outcomes,y = prob))+
   geom_bar(stat="identity")
```



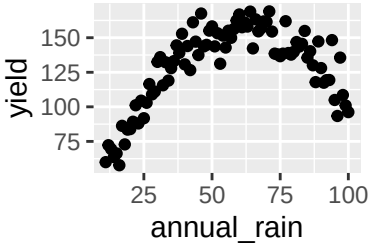
Sampling from a distribution: A data-generating machine

(Graphical description of data generating machine - rnorm) (AR(1) process)

Now let's look at this from the other end. Using data, can we build a machine that may have generated our data?

Machine 1: Precipitation -> Yield

```
set.seed(202073)
annual_rain<-seq(11,100, 1)
yield <- 2 + 5*annual_rain - 0.04* annual_rain^2 + rnorm(90,0,10)
yield_dat<-tibble(annual_rain = annual_rain, yield=yield)
ggplot(yield_dat, aes(annual_rain, yield))+geom_point()
```



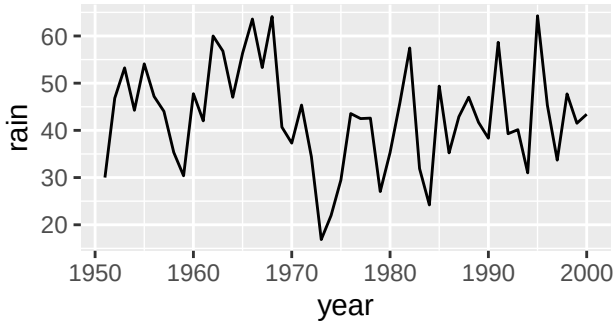
Machine 2: Precipitation -> Yield, AR process

```
set.seed(202073)
annual_rain<- rep(30,50); yield <-rep(150,50)
for (i in 2:50) {
     annual_rain[i]<- 0.3*annual_rain[i-1] + rnorm(1,30,10)
     yield[i]<- 5*annual_rain[i] - 0.04* annual_rain[i]^2 +
        0.2*yield[i-1] + rnorm(1,0,10)
}
yield_dat2<-tibble(year = 1951:2000, rain = annual_rain, yield=yield)
```

Machine 2: Precipitation -> Yield, AR process

Annual rain between 1951 and 2000

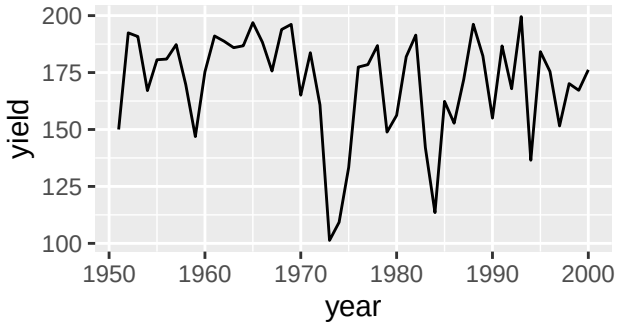
```
ggplot(yield_dat2, aes(year, rain))+geom_line()
```



Machine 2: Precipitation -> Yield, AR process

Crop yield between 1951 and 2000

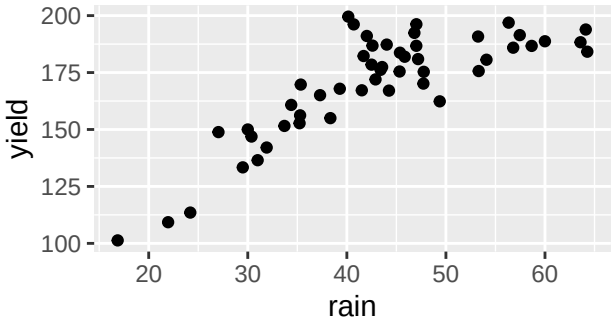
```
ggplot(yield_dat2, aes(year, yield))+geom_line()
```



Machine 2: Precipitation -> Yield, AR process

Crop yield vs Annual rain

```
ggplot(yield_dat2, aes(rain, yield))+geom_point()
```



Summary

- A probability distribution: a set of possible outcomes and associated probabilities
- Data generating process: set of rules for generating set of outcomes
- Probability: from rules to data
- Statistics: from data to rules

Statistics: re-constructing the rules, given the data

The Ultimate Challange!