Statistical Thinking in Biology Research

Probability and Statistical Inference

Terry Neeman

Australian National University

12th August 2022

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: simulating data from a set of rules

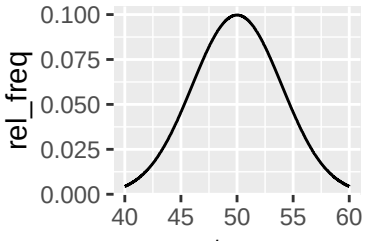
- Sample space: space of possible outcomes
- Distribution: relative frequencies (probabilities) of each outcome
- Summaries of distributions: average (expected) outcome, variation around average

Examples of common distributions in biological research

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

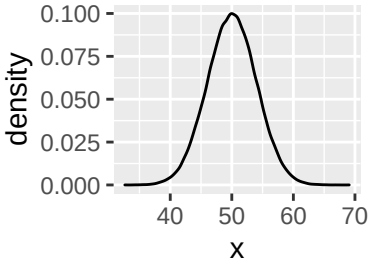
Relative frequencies in the Normal distribution

```
library(tidyverse)
outcome<-seq(40,60, length.out = 1e4)
out_normal <- tibble(outcome = outcome,
                     rel_freq = dnorm(outcome, mean = 50, sd = 4))
ggplot(out_normal, aes(x=outcome,y = rel_freq))+
   geom_line()
```



Sample from a normal distribution

```
sample_normal <- tibble(x = rnorm(n=1e5, mean = 50, sd = 4))
ggplot(sample_normal, aes(x = x))+
   geom_density()
```

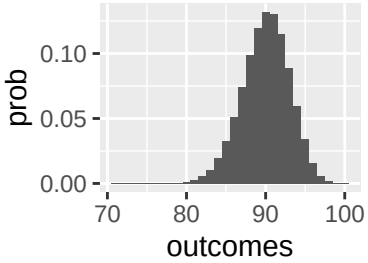


Examples of common distributions in biological research

- 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 space $0, 1, ..., n$

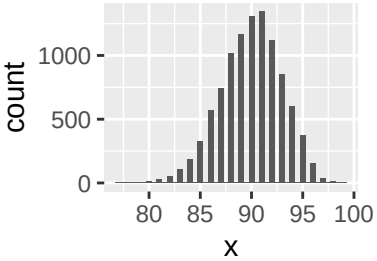
Probabilities of possible outcomes for 100 flips of a very biased coin

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

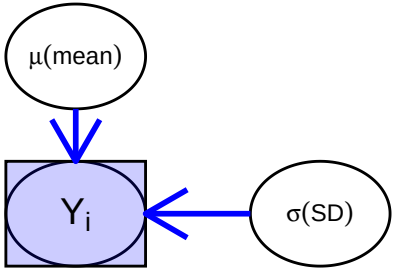


Simulating outcomes from a coin-flipping machine

```
sample_binomial <- tibble(x = rbinom(1e4, size = 100, prob = 0.9))
ggplot(sample_binomial, aes(x = x))+
   geom_histogram(binwidth = 0.5)
```



Sampling from a distribution: A data-generating machine



Now let's generate data that looks more like real biological data. Data generating process = Model

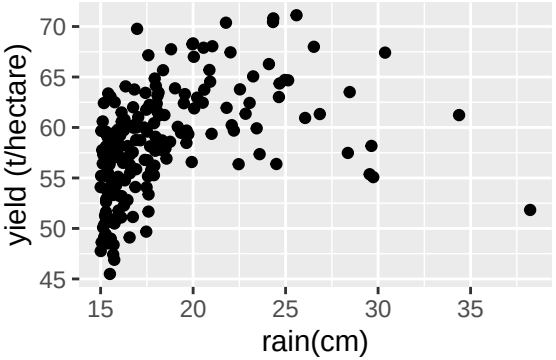
Model 1: Precipitation -> Yield

```
library(Pareto)
set.seed(2021906)
annual_rain<-rPareto(190,15,alpha=6)
yield <- 2 + 5*annual_rain - 0.1* annual_rain^2 + rnorm(190,0,4)
yield_dat<-tibble(annual_rain = annual_rain, yield=yield)
head(yield_dat)
```

```
## # A tibble: 6 x 2
##   annual_rain yield
##         <dbl> <dbl>
## 1        18.0  62.4
## 2        17.5  59.6
## 3        15.4  59.2
```

Model 1: Precipitation -> Yield

```
ggplot(yield_dat, aes(annual_rain, yield))+geom_point()+
   xlab("rain(cm)")+ylab("yield (t/hectare)")
```



“Data generating Machine” for Model 1

-
-
-

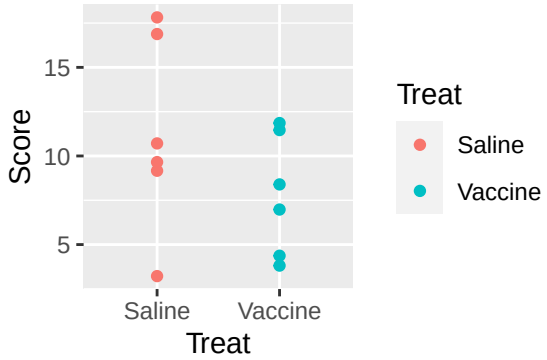
Model 2: Simulate Mouse Challenge Experiment

```
set.seed(220621)
challenge <- tibble(MouseID = 1:12,
            Treat = rep(c("Saline", "Vaccine"), each = 6),
            Score = rep(c(10,6),each = 6) + rnorm(12,0,sd = 4))
head(challenge)
```

```
## # A tibble: 6 x 3
##   MouseID Treat  Score
##     <int> <chr>  <dbl>
## 1       1 Saline 17.8
## 2       2 Saline  3.22
## 3       3 Saline  9.17
## 4       4 Saline 10.7
```

Model 2: Simulate Mouse Challenge Experiment

```
ggplot(challenge, aes(x = Treat, y = Score, col =Treat))+
   geom_point()
```



“Data generating Machine” for Model 2

Model 3: Mouse Challenge Experiment with Variant

Challenge with WT virus or variant

```
set.seed(362154)
challenge2 <- tibble(MouseID = 1:24,
            Challenge = rep(c("WT", "var1"), each = 12),
            Treat = rep(rep(c("Saline","Vaccine"),each = 6),2),
          Score = rep(c(10,6,14,7),each = 6) + rnorm(24,0,sd = 4))
```

Model 3: Mouse Challenge Experiment with Variant

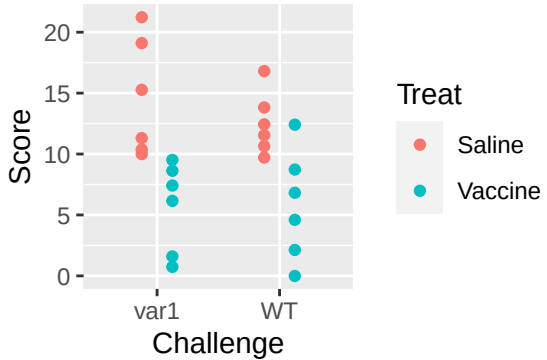
Challenge with WT virus or variant

```
head(challenge2)
```

```
## # A tibble: 6 x 4
##   MouseID Challenge Treat  Score
##     <int> <chr>     <chr>  <dbl>
## 1       1 WT        Saline 11.6
## 2       2 WT        Saline 13.8
## 3       3 WT        Saline 12.4
## 4       4 WT        Saline 16.8
## 5       5 WT        Saline 10.6
## 6       6 WT        Saline  9.71
```

Model 3: Mouse Challenge with Variant

```
ggplot(challenge2, aes(x = Challenge, y = Score, col =Treat))+
   geom_point(position = position_dodge(width = 0.5))
```



“Data generating Machine” for Model 3

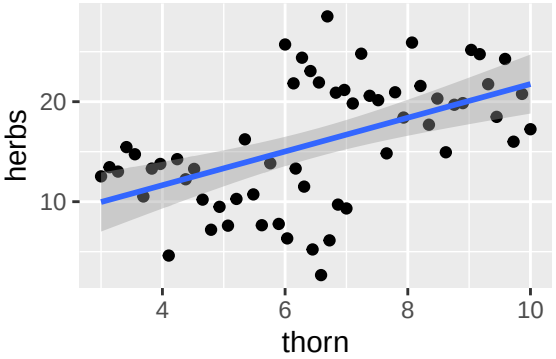
Model 4: Recreate some thorn density data

```
site <- rep(c("Site1", "Site2"), each = 30)
thorn<- c(seq(3,7,length.out = 30), seq(6, 10, length.out = 30))
set.seed(65432)
herbs <- rep(c(20,35), each = 30) - 1.8*thorn+ rnorm(60,0,3)
herb_data <- tibble(site = site, thorn = thorn, herbs = herbs)
head(herb_data)
```

```
## # A tibble: 6 x 3
##   site  thorn herbs
##   <chr> <dbl> <dbl>
## 1 Site1  3     12.5
## 2 Site1  3.14  13.4
## 3 Site1  3.28  13.0
```

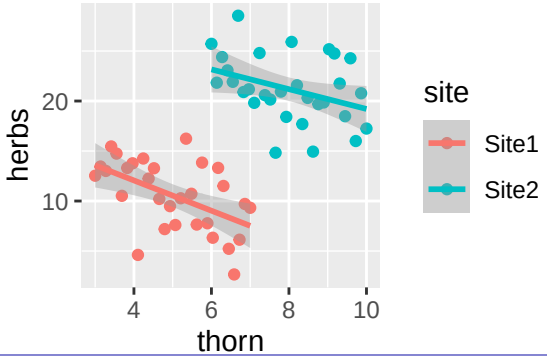
Model 4: Thorn density data

```
ggplot(herb_data, aes(thorn, herbs))+geom_point()+
   geom_smooth(method = "lm")
```



Model 4: Recreate some thorn density data

```
ggplot(herb_data, aes(thorn, herbs, col = site))+
   geom_point()+
   geom_smooth(method = "lm")
```



“Data generating Machine” for Model 4

-
-
-
-
-
-

Summary

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

Statistics: re-constructing the model, given the data

The Ultimate Challenge!