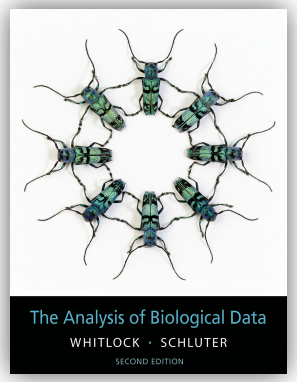
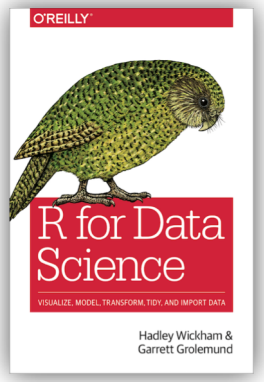
Data Science in Bioinformatics

week.03.remember.your.PROBABILITY.class

Palle Villesen & Thomas Bataillon





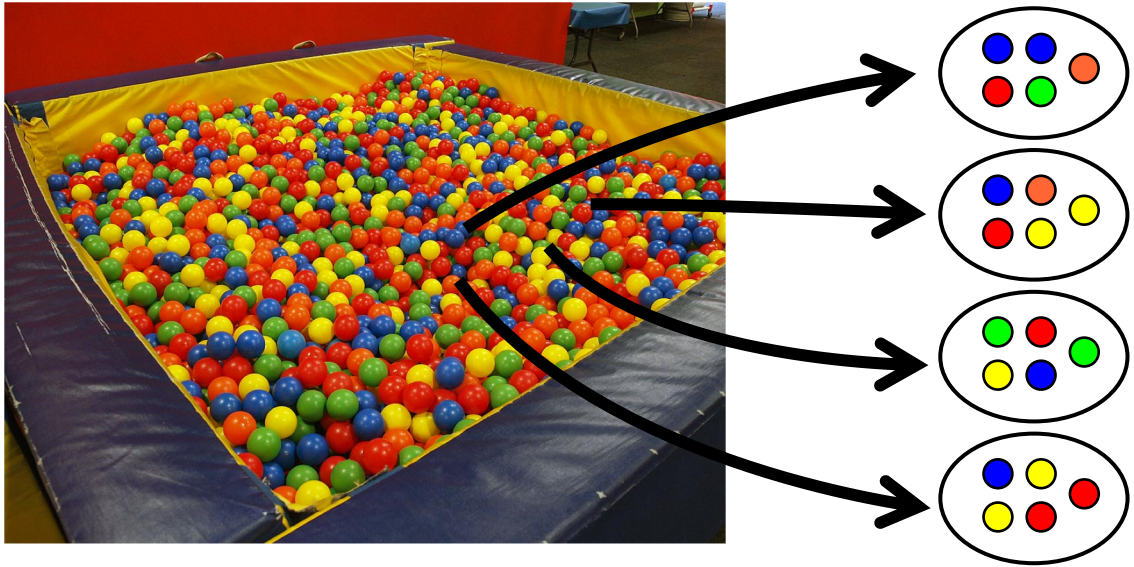
Population VS sample

UNKNOWN

- Virtually infinite
- Parameters
- Probability distribution

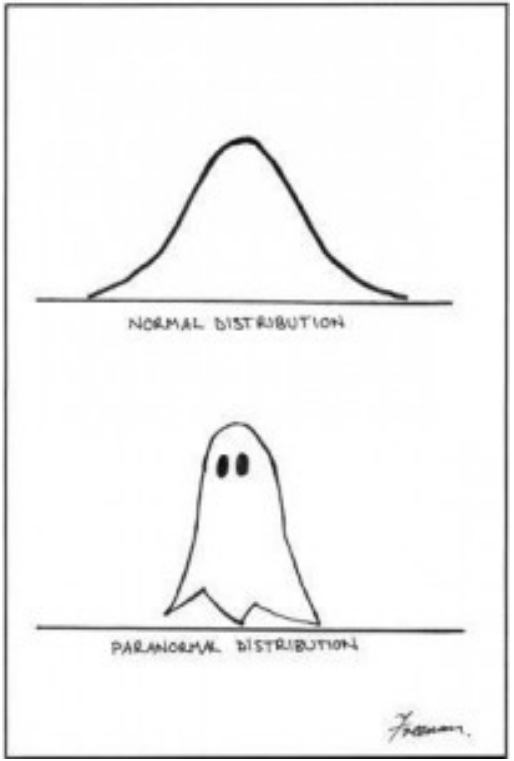
KNOWN

- n obs: x_1, x_2, …x_n
- Parameter estimates
- Sampling distribution



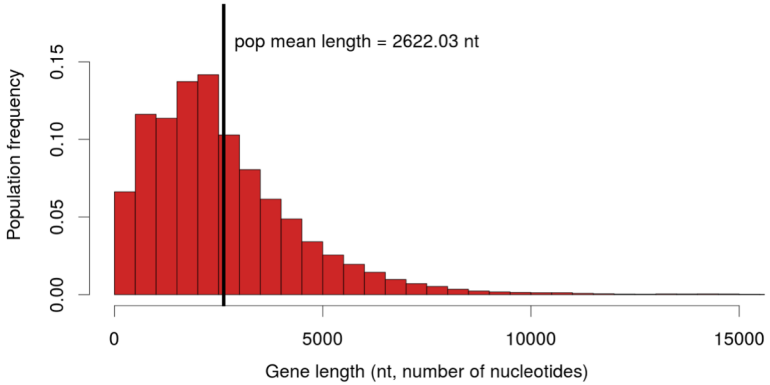
Outline for week 03

- Any questions from last week ?
- Probabilities : why do we care
- Measuring uncertainty
- Basic probabilities
  - Random events & Random variables
  - Probability distributions
    - Discrete
    - Continuous
    - “pseudo continuous “
    - in R …
  - Histograms : Probability vs Empirical (= DATA) distribution

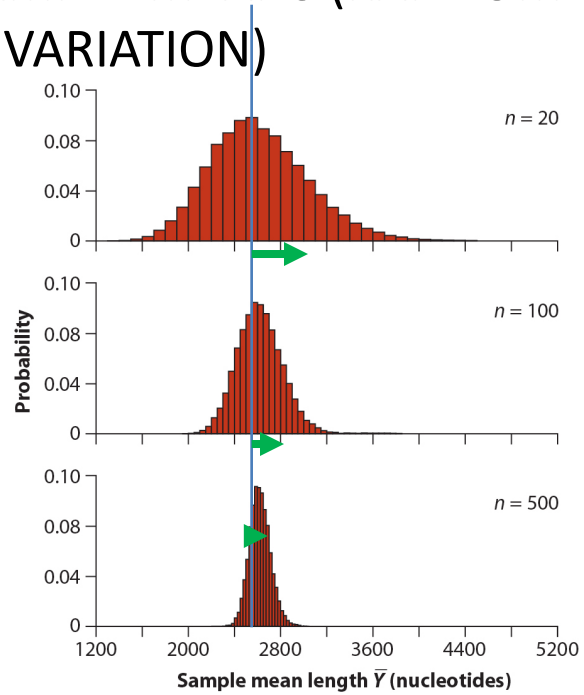


Population parameters versus sampling distributions of sample statistics

POPULATION MEAN, SD (FIXED)



SAMPLE MEANS (RANDOM VARIATION)



SE = SD / Sqrt(n)

What is a probability ? ?

To define a Probability we need a ”thought experiment” (a random trial) where you can oberve **stable repeated** outcomes :

Examples

- Flipping a coin (2 posible outcomes)
- Rolling a dice (6 possible outcomes)
- Drawing a card from a well mixed deck
- Genotype an individual at a SNP position of the genome (AA AT TT)
- Measuring an individual from a population or an experiment

What is a probability ?

To define the **Probability of an event**, we need a ”thought experiment” (a random trial) where you can potentially oberve many **repeated** outcomes :

Examples

- Flipping a coin (2 outcomes), Rolling a dice (6 outcomes .. Or 10 20 if you are into roleplaying games)
- Drawing a card rom a well mixed deck (Poker)
- Genotype an individual at a SNP position of the genome (AA AT TT)
- Measuring an individual from a population or an experiment

The probability of an event is the proportion of the time you will see that outcome matching that event (say AA) if you observe many outcomes

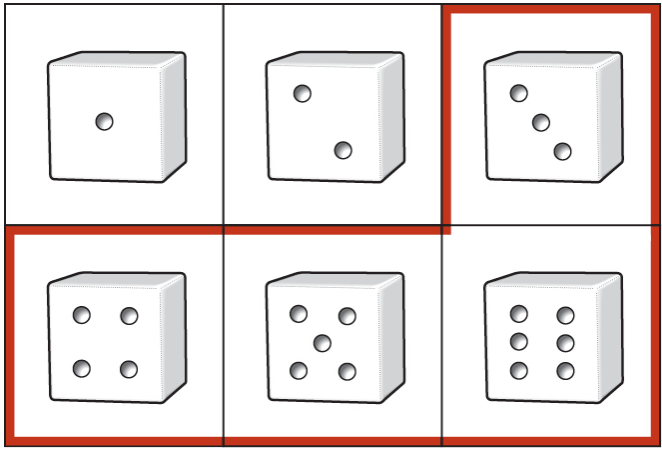
Example I roll 100 times a dice and I observe 23 ”6” so the probability is 23/100 (note it is not exactly 1/6)
If you had the patience to roll it a million times the probability will **converge** to 1/6

Defining the space of possibles:

The **sample space** Ω is the **set** of all possible outcomes in an experiment.

$\Omega = \{ "1", "2", "3", "4", "5", "6" \}$

We are interested in specific "events" that can be a very specific outcome (roll and get a "6") $\mathcal{E}_1$={6}
several outcomes (("roll and get higher than 2")
$\mathcal{E}_2$ ={"3","4"5","6"}



Total number of outcomes in S and E resp is n(S) and n($\mathcal{E}$)
IF all individual (elementary outcomes) are equiprobable then
P($\mathcal{E}$)= n($\mathcal{E}$) / n(Ω)

A useful thing: random variables (r.v.)

Random variables ”record” the outcome of probability event when the outcome is a number

Example:

Let X be a random variable that records the throw of a dice

The **sample space Ω** is the **set** of all possible outcomes in an experiment.

$\Omega = \{ \text{”1”}, \text{”2”}, \text{”3”}, \text{”4”}, \text{”5”}, \text{”6”} \}$

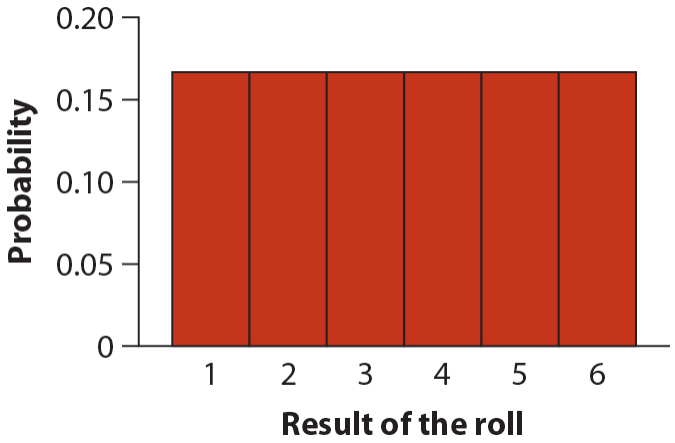
X={1, 2, 3, 4, 5 or 6}

Here if it s a ”fair dice” we can say that all events are equiprobable

So using our notation : P(X=1)=P(X=2)= …=P(X=6)

And we also know that all these events are mutually exclusive so their probabilities sum to 1

A discrete probability distribution



X records the result of the roll of a fair dice

X={1,2,3,4,5,6}

One of the possible event must happen → sum of probs =1

NOTE Here all events are

equiprobable and this need not be the case

Outlook: sometimes you can have a potentially infinite number of events

Example: number of random events in a time interval

Random variables always have an associated probability distribution

P(X= …) or in a more detailed way

P(X=1) P(X=2) … etc.

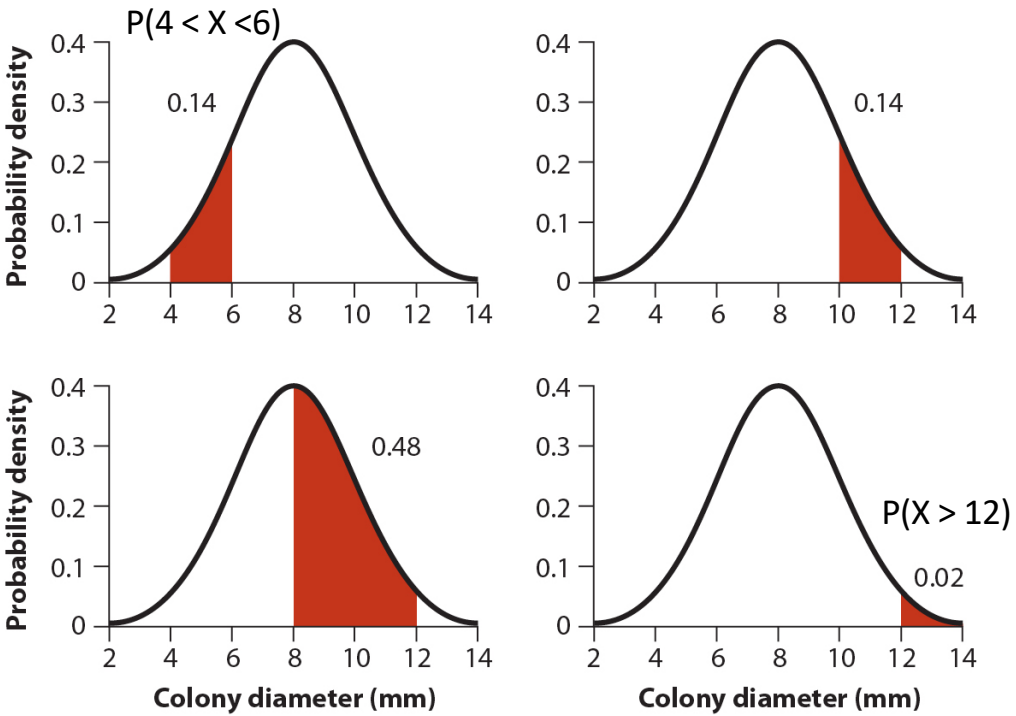
Note here we used random variables that ”map” to discrete event (i.e. Dice rolled or a coin flip) but we can also use them for continuous probability distributions

Random variables have a ”mean” (Expectation) and a variance

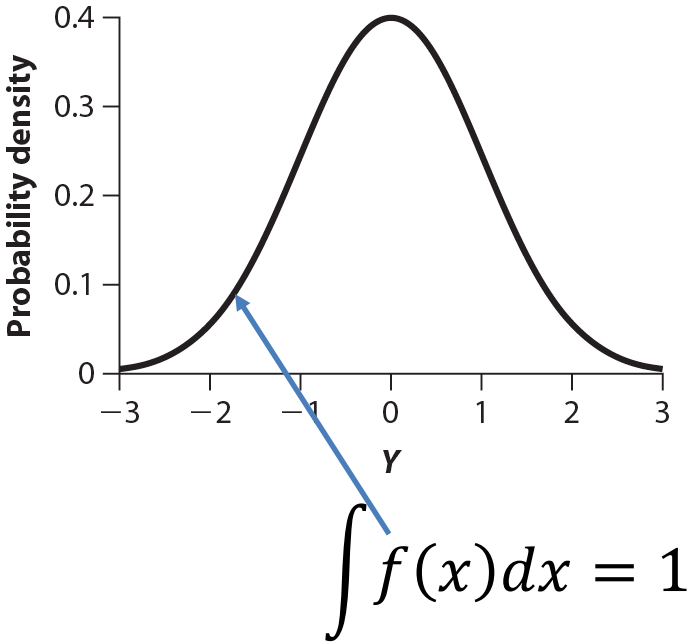
$E[X] = 1\ P(X=1) + 2\ P(X=2) + \ldots + 6\ P(X=6)$ (WEIGHTED MEAN)

$V[X] = E(X\ X) - E(X)\ E(X) = E[(\ X - E(X)\ )^2]$ (~ like variance in sample)

X = colony diameter,



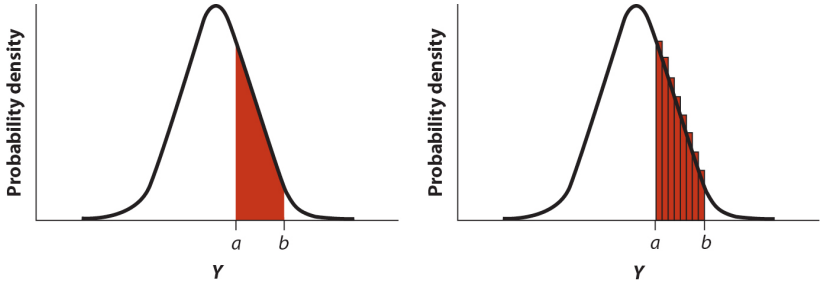
Continuous random variables



$$\int f(x)dx = 1$$

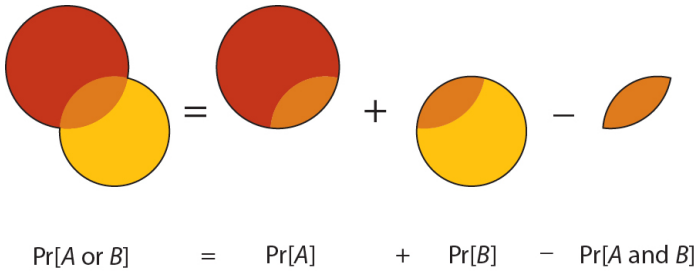
- Sometimes the events we record are not discrete but **continuous**
- **Examples: height measurements, gene expression measurements, enzymatic activity**
- The probability distribution becomes a density function and we replace sum of all events by **integrals**
- **The area under the curve is 1.**

Dealing with continuous random variables



- The probability of a specific point is zero !
- The probability of a random variable to be in an interval [a,b] is given by the area under the curveof that interval
- $P(a < Y < b) = \int_a^b f(x)dx$

Either this OR that

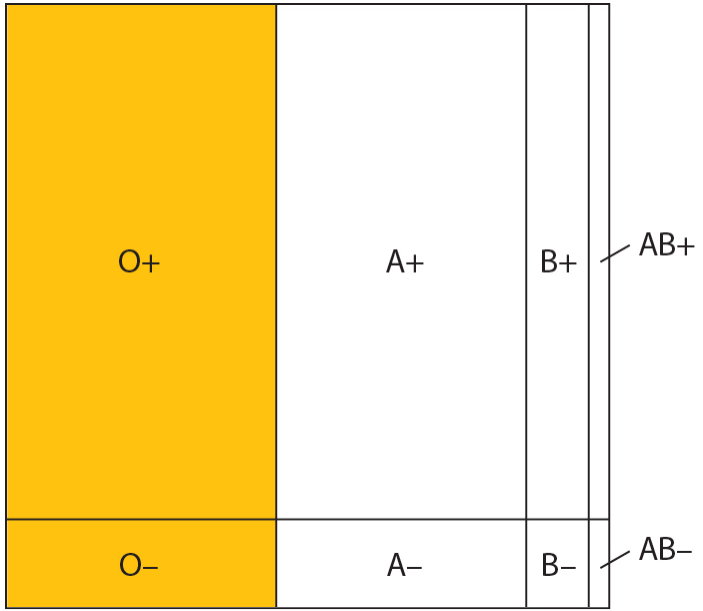


P(A OR B)

= P(A) + P(B) – P (A AND B)

IF A AND B are mutually exclusive

Either this OR that event



E= ” being Type O”

E ={ ”0+”, ”O-”}

Being O+ and O- are mutually exclusive events

Probability distribution can be defined also for sums of random variables

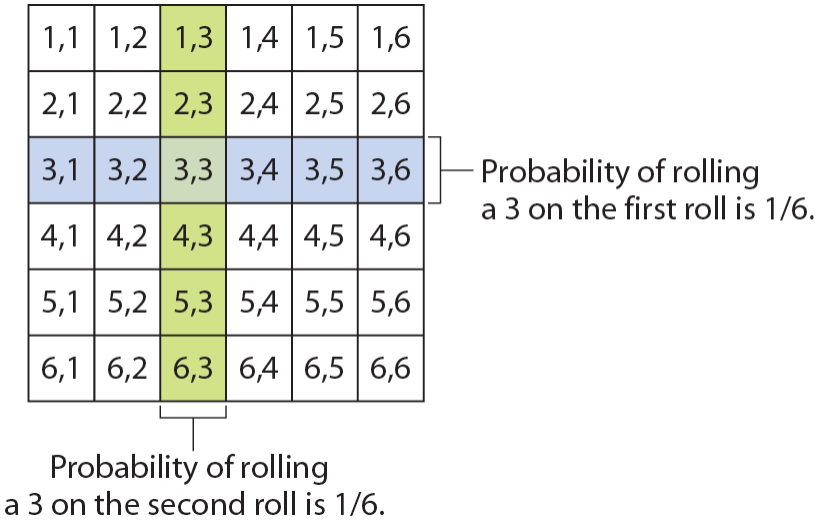
- X1 records the 1st throw of a dice
- X2 records the 2nd throw of dice
- S=X1+X2 records the sum of two dice

Ω ={2,3,4,…,12}

- How can we calculate P(S) the probability distibution of that sum ?

THIS & THAT :
Independence and the multiplication rule

A more complicated venn diagram depicting all individual outcomes (Ω)



Dice rolls are **independent**

X1 and X2 are r.v. recording the outcome of each roll

X1 and X2 and independent

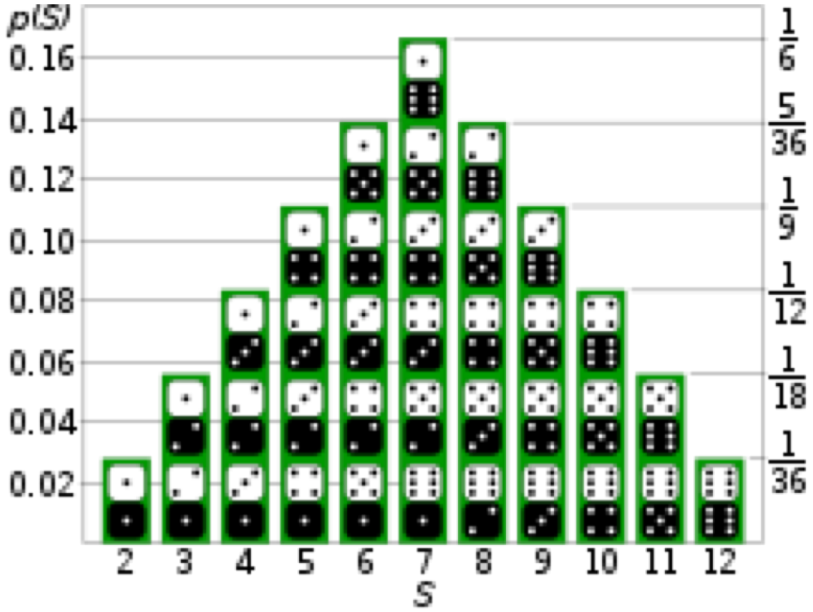
→ Each ”micro” outcome (X1, X2) has probability 1/36 because

$P(X1=3) = 1/6$

$P(X2=3) = 1/6$

$P(X1=3 \ \& \ X2=3) = 1/6 \times 1/6 = 1/36$

Adding up individual mutually exclusive events



1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

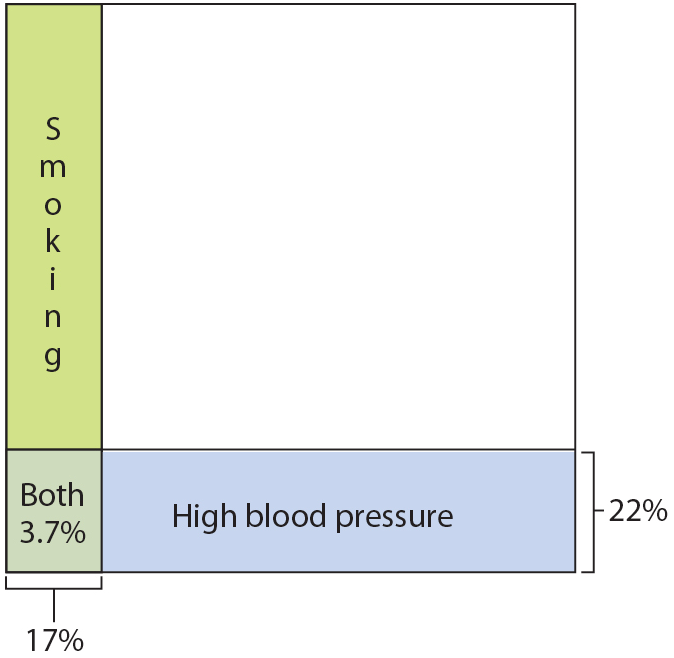
Probability of rolling a 3 on the first roll is 1/6.

Probability of rolling a 3 on the second roll is 1/6.

Modeling a sample of *n* observations with r.v.

- Imagine the sum of n throws of a dice … $S_n = X_1 + X_2 + X_3 \ldots + X_n$
- What can we say about Sn ?
  - Sn={6, 2…, 6n}
  - P(Sn) (hard)?
  - E(Sn) and V(Sn) easier (See exercise / Problem in TE)
- One strategy is to enumerate all the individual (mutually exclusive) events , and sum them up
  - Tedious BUT Guaranteed to work
  - Some math person might have done it before (see binomial distribution ch7)

Independent events → multiplication rule

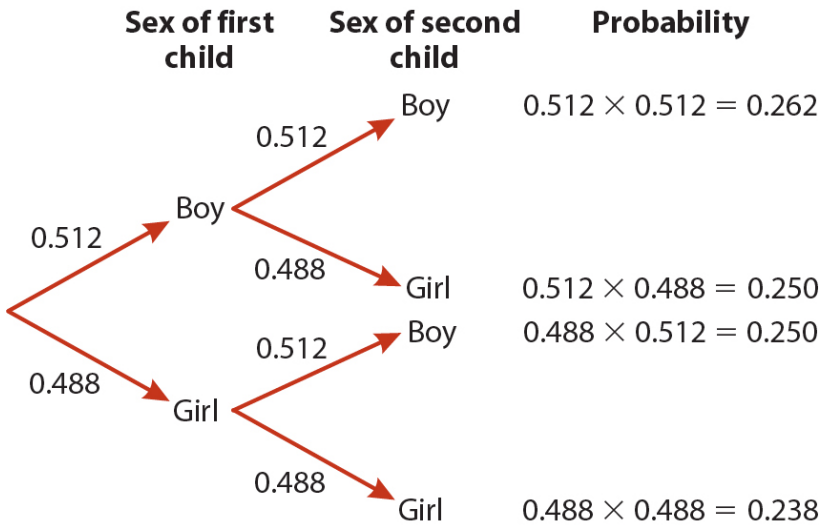


P(Smoking & HBP) =

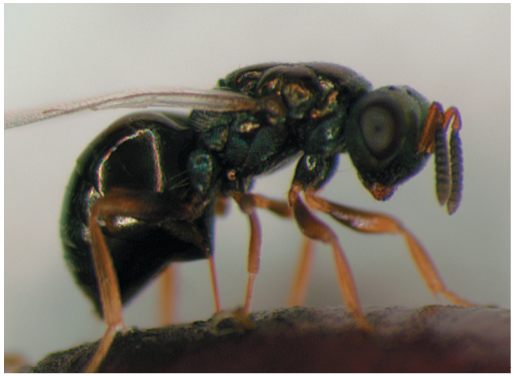
P(Smoking) x P(HBP)

Probability trees are good for calculating the probability of sequence of events

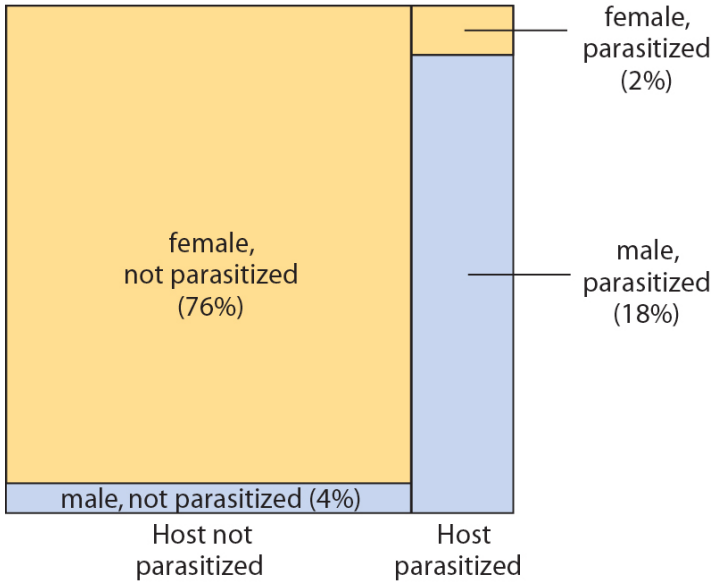
First child second child

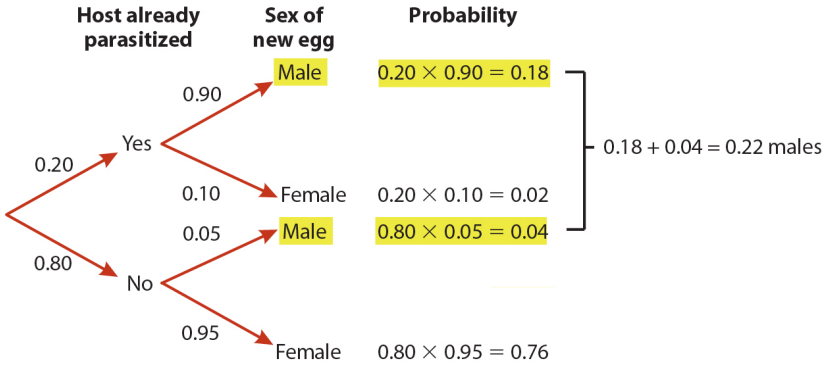


What are wasps doing ?

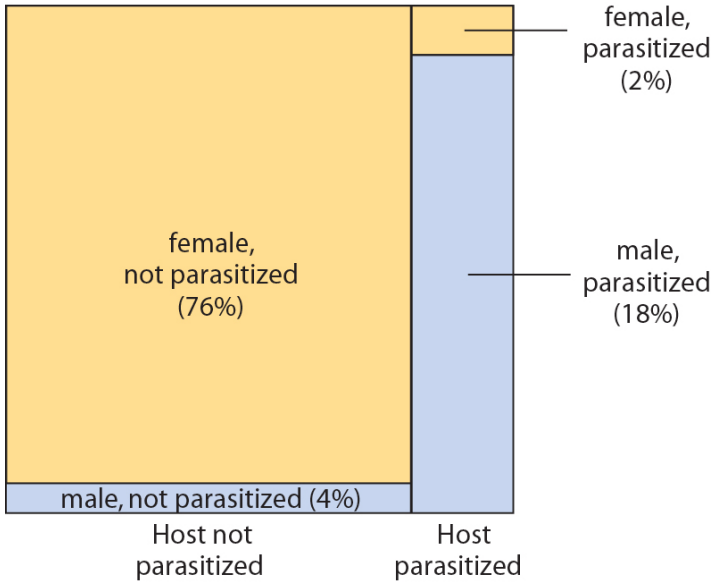


Events can be dependent upon each other





Dependent events → multiplication rules do not work anymore !!!



P(Egg is male) = 0.22

P(Host is parasitized) = 0.2

P(Hist is parasitized & egg is male) **is NOT 0.22 x 0.2** (=0.044)

(here it is 18%)

This week‘s focus

Lecture

- Probability calculations
- Random variables
  - E [] Expectation
  - V[] Variance
  - … higher moment

TE

Same .. !

Pencil and paper probability

Using R to

Sample

Calculate probabilities

Vizualize probabilities

Let’s do it ”in R”…Using the binomial distribution

Random sample (n=100) from a distribution

rbinom(n = 100, size = 6,prob = 1/6)

rnorm(n=100, mean = 1, sd=2)

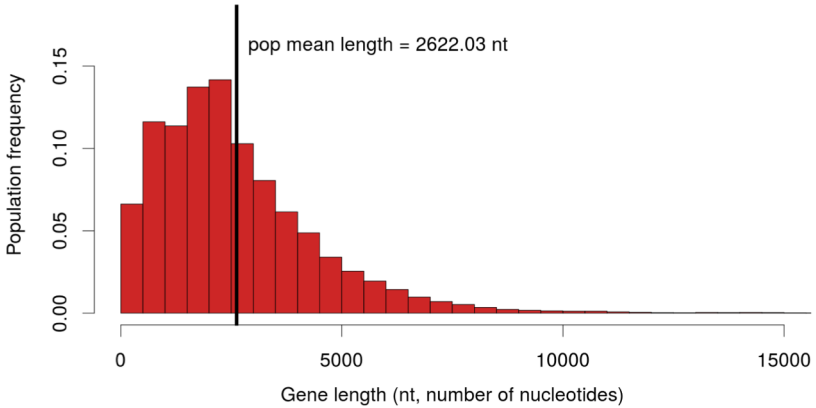
Calculate probabilities of a certain range

pbinom() OR dbinom()

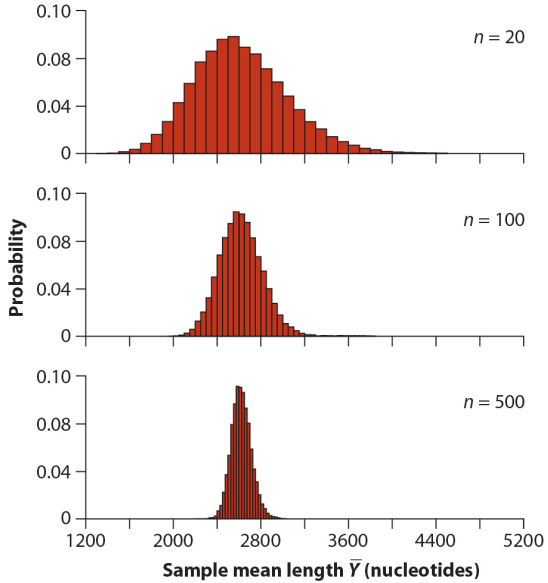
pnorm()

We learned about probabilities… but how do we use that …back to statistics ?

Population (probability distribution)



Sample probability distribution



Displaying data

Exercise

- Make a small dataset with 3 outliers:
  `x = c(70:90,1000,1100,1200)`
- Calculate mean and median
  `mean(x)`
  `median(x)`
- Calculate sd, iqr and mad
  `sd(x)`
  `quantile(x)`
  `mad(x)`
- Conclusions?

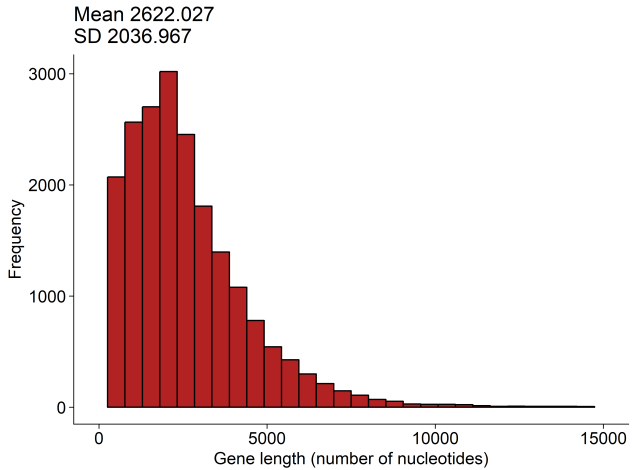
Estimating with uncertainty

Keywords

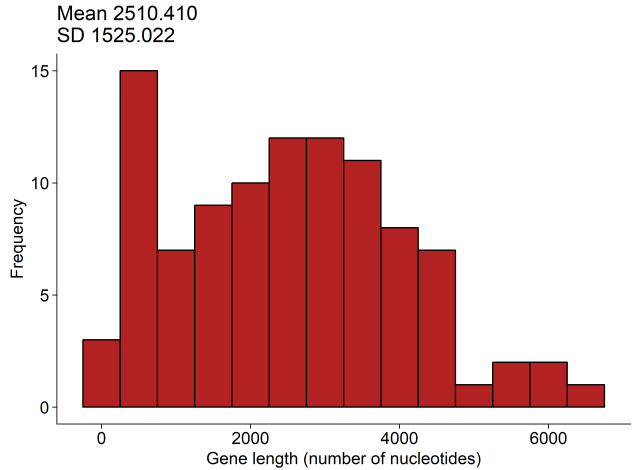
- We want to say something about the **population**
- But we only have a **sample**
- So the **sample estimate** is different from the true value because of **sampling error**
- **The sampling distribution** is the distribution of estimate from different samples
- **The standard error** is the standard deviation of the sampling distribution
- **Confidence intervals on the estimate**
  - The 2SE rule of thumb
  - Bootstrap (chapter 19)

All genes

```
df <- read_csv(file =
"chap04e1HumanGeneLengths.csv")
```



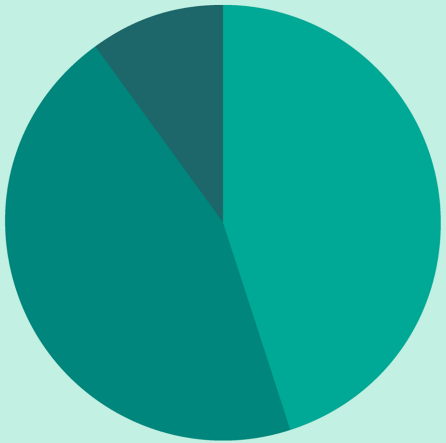

```
set.seed(0)
dfsub <- df %>% sample_n(size = 100)
```



- `df <- read_csv(file="chap04e1HumanGeneLengths.csv")`
- Replicate figure 4.1-3
  - the sampling distribution of the mean
- Replicate figure 4.1-4
  - the sampling distribution of the mean for n=20, n=100 and n=500
- Calculate standard error from your samples (n=20, n=100, n=500)
- Compare with table 4.2-1

```
set.seed(0)
r <- data.frame() %>% tbl_df()
for (i in 1:10000) {
  dfsub = df %>% sample_n(size = 100)
  r <- rbind(r, data.frame(n = nrow(dfsub), gene.mean = mean(dfsub$geneLength)))
}
```

GROUP WORK



- Splitting up tasks
- Scolding and accusing others of not doing their part
- Actual work

TRUTH FACTS