Probability

Felipe José Bravo Márquez

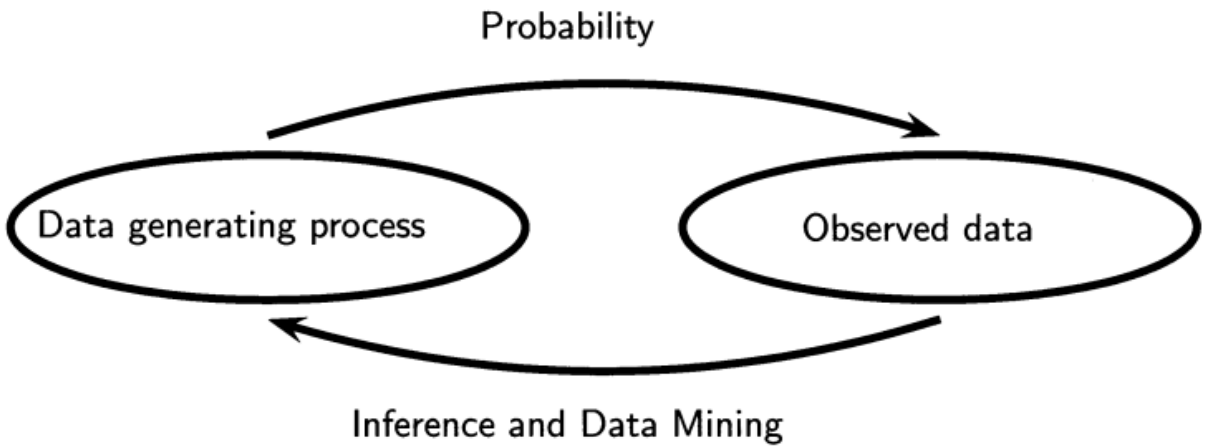
April 5, 2021

Probability and Statistics

- Probability is the language of **uncertainty** that is also the basis for statistical inference [Poldrack, 2019].
- It forms an important part of the foundation for statistics, because it provides us with the mathematical tools to describe uncertain events.
- The study of probability arose in part due to interest in understanding games of chance, like cards or dice.
- These games provide useful examples of many statistical concepts, because when we repeat these games the likelihood of different outcomes remains (mostly) the same.

Probability and Statistics

- The problem studied in probabilities is: given a data generating process, which are the properties of the outputs?
- The problem studied in statistical inference, data mining and machine learning is: given the outputs, what can we say about the process that generates the observed data?



[1] Figure taken from [Wasserman, 2013]

What is Probability?

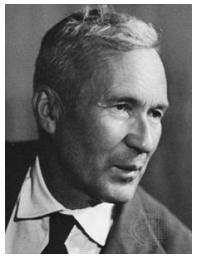
- We think of probability as a number that describes the likelihood of some event occurring, which ranges from zero (impossibility) to one (certainty).
- Probabilities can also be expressed in percentages: when the weather forecast predicts a twenty percent chance of rain today.
- In each case, these numbers are expressing how likely that particular event is, ranging from absolutely impossible to absolutely certain.

Probability Concepts

- A **random experiment** in the act of measuring a process whose output is uncertain.
- Examples: flipping a coin, rolling a 6-sided die, or trying a new route to work to see if it's faster than the old route.
- The set with all possible outputs of a random experiment is the **sample space** Ω (it can be discrete or continuous).
- For a coin flip $\Omega = \{\text{heads, tails}\}$, for the 6-sided die $\Omega = \{1, 2, 3, 4, 5, 6\}$, and for the amount of time it takes to get to work Ω is all possible real numbers greater than zero.
- An **event** $E \subseteq \Omega$ corresponds to a subset of those outputs.
- For example, $E = \{2, 4, 6\}$ is the event of observing an even number when rolling a die.

Probability

- Now we can outline the formal features of a probability, which were first defined by the Russian mathematician Andrei Kolmogorov.



- A probability $\mathbb{P}$ is a real-valued function defined over Ω that satisfies the following properties:

Properties

1. For any event $E \subseteq \Omega$, $0 \leq \mathbb{P}(E) \leq 1$.
2. The probability of the sample space is 1: $\mathbb{P}(\Omega) = 1$
3. Let $E_1, E_2, \ldots, E_k \in \Omega$ be disjoint sets

$$\mathbb{P}(\bigcup_{i=1}^{k} E_i) = \sum_{i}^{k} P(E_i)$$

Probabilities cannot be negative or greater than 1.

Interpretation of Probabilities

The are two common interpretations of probabilities: frequencies and degrees of beliefs (or Bayesian probabilities).

Frequentist probability

- In the frequency interpretation, $\mathbb{P}(E)$ is the long run proportion (limiting frequency) of times that E is true in repetitions.[a]
- For example, if we say that the probability of heads is 1/2, we mean that if we flip the coin many times then the proportion of times we get heads tends to 1/2 as the number of tosses increases.
- When the sample space Ω is finite, we can say that

$$\mathbb{P}(E) = \frac{\text{Favorable cases}}{\text{total cases}} = \frac{|E|}{|\Omega|}$$

[a] https://en.wikipedia.org/wiki/Frequentist_probability

Interpretation of Probabilities

Probability as a degree of belief

- The degree-of-belief interpretation (a.k.a Bayesian interpretation or Subjective interpretation) is that $\mathbb{P}(E)$ measures an observer's strength of belief that E is true.
- If I were to ask you "How likely is it that the US will return to the moon by 2026", you can provide an answer to this question based on your knowledge and beliefs.
- Even though there are no relevant frequencies to compute a frequentist probability.

- In either interpretation, we require that properties 1 to 3 hold.
- The difference in interpretation will not matter much until we deal with statistical inference.
- There, the differing interpretations lead to two schools of inference: the frequentist and the Bayesian schools.

The rule of subtraction

- The probability of some event E not happening is one minus the probability of the event happening:

$$\mathbb{P}(\neg E) = 1 - \mathbb{P}(E)$$

- For example, if the probability of rolling a one in a single die throw is 1/6, then the probability of rolling anything other than a one is 5/6.

Combinatorial methods

- There are a few facts from counting theory that are useful for calculating probabilities.
- Given n objects, the number of ways of ordering these objects is $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$.
- For convenience, we define $0! = 1$.
- We also define

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

read “n choose k”, which is the number of distinct ways of choosing k objects from n.

Combinatorial methods

- For example, if we have a class of 20 people and we want to select a committee of 3 students, then there are

$$\binom{20}{3} = \frac{20!}{3!17!} = 1140$$

possible committees.

- We note the following properties:

$$\binom{n}{0} = \binom{n}{n} = 1$$

and

$$\binom{n}{k} = \binom{n}{n-k}.$$

Random Variable

- A **random variable** is a mapping (or function)

$$X : \Omega \to \mathbb{R}$$

  which assigns a real value $X(e)$ to any event of Ω.
- Example: We flip a fair coin 10 times. The outcome of each toss is a head H or a tail T.
- Let $X(e)$ be the number of heads in the sequence of outcomes.
  - If $e = HHTHHTHHTT$, then $X(e) = 6$
- A random variable can in many cases be the identity function. Example, if we roll a 6-sided die once and $X(e)$ is the resulting die value, then $X(e) = e$ for $e \in \{1, 2, 3, 4, 5, 6\}$.
- It is important to understand that we can easily have different random variables from a same sample space. Example: we roll a die two times, $X(e)$ is the sum of the resulting two rolls and $Y(e)$ is the product of these two numbers.
- In most cases we work directly with the random variable and the sample space becomes implicit.

Example

- We flip a coin 2 times. Let X be the number of tails obtained.
- The random variable and its distribution is summarized as:

e	$\mathbb{P}(e)$	$X(e)$
HH	1/4	0
HT	1/4	1
TC	1/4	1
TT	1/4	2

x	$\mathbb{P}(X = x)$
0	1/4
1	1/2
2	1/4

R.V Definitions

- Let X be a R.V , we define **cumulative distribution function** (CDF) or $F_X : \mathbb{R} \to [0, 1]$ as:

$$F_X(x) = \mathbb{P}(X \leq x)$$

Discrete Random Variables

- A R.V X is **discrete** if it maps the outputs to a countable set.
- We define the **probability function** or **probability mass function** of a discrete R.V X as $f_X(x) = \mathbb{P}(X = x)$.
- Then $f_X(x) \geq 0 \ \forall x \in \mathbb{R}$, and $\sum_i f_X(x_i) = 1$
- The CDF of X is related to f_X as follows:

$$F_X = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

R.V Definitions II

Continuous Random Variable

- A R.V X is continuous if:
- there exists a function f_X such that $f_X(x) \geq 0 \; \forall x$, $\int_{-\infty}^{\infty} f_X(x)dX = 1$

$$\int_{-\infty}^{\infty} f_X(x)dX = 1$$

- For all $a \geq b$:

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$$

- The function f_X is called the probability density function (PDF).
- The PDF is related to the CDF as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

- Then $f_X(x) = F'_X(x)$ at all points x where F_X is differentiable.
- For continuous distributions the probability that X takes a particular value x is always zero.

Some Properties

1. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$
2. $\mathbb{P}(X > x) = 1 - F(x)$
3. If X is continuous:

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) \end{aligned}$$

Quantiles

- Let X be a R.V with CDF F. The inverse CDF or quantile function is defined as

$$F^{-1}(q) = inf\{x : \ F(x) > q\}$$

- For $q \in [0, 1]$ if F is strictly increasing and continuous, $F^{-1}(q)$ is the only real value such that $F(x) = q$.
- Then $F^{-1}(1/4)$ is the first quartile, $F^{-1}(1/2)$ the median (or second quartile) and $F^{-1}(3/4)$ the third quartile.

Some distributions

	Probability Function	Parameters
Binomial	$f_x = \binom{n}{x} p^x (1-p)^{n-x}$	n, p
Normal	$f_x = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	μ, σ
Poisson	$f_x = \frac{1}{x!}\lambda^x \exp^{-\lambda}$	λ
Exponential	$f_x = \lambda \exp^{-\lambda x}$	λ
Gamma	$f_x = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp^{-\lambda x}$	λ, α
Chi-square	$f_x = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(\frac{k}{2}-1)} \exp^{-x/2}$	k

Warning

We defined random variables to be mappings from a sample space Ω to $\mathbb{R}$ but we did not mention the sample space in any of the distributions above. The sample space often becomes implicit but it is really there in the background.

Binomial Distribution

- The binomial distribution is a discrete distribution that provides a way to compute the probability of some number of successes out of a number of trials.
- In each trial there is either success or failure and nothing in between (known as "Bernoulli trials") given some known probability of success on each trial.
- Let n be the number of trials, x the number of successes, and p the probability of a success, the probability mass function of the Binomial distribution is as follows:

$$f_x(n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

- The binomial coefficient $\binom{n}{x}$ describes the number of different ways that one can choose x items out of n total items.

Binomial Distribution

- Example: on Jan 20 2018, the basketball player Steph Curry hit only 2 out of 4 free throws in a game against the Houston Rockets.
- We know that Curry's overall probability of hitting free throws across the entire season was 0.91.
- What is probability that he would hit only 50% of his free throws in a game?

$$f_2(4, 0.91) = \binom{4}{2} 0.91^2 (1 - 0.91)^{4-2} = 0.040$$

In R:

```
 > dbinom(x=2,size=4,p=0.91)
[1] 0.04024566
```

Normal Distribution

- X has a Normal or Gaussian distribution of parameters μ and σ, $X \sim N(\mu, \sigma^2)$ if

$$f_x = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- Where $\mu \in \mathbb{R}$ is the "center" or the "mean" of the distribution and $\sigma > 0$ is the "standard deviation".
- When $\mu = 0$ and $\sigma = 1$ we have a **Standard Normal Distribution** denoted by Z.
- We refer to the PDF by $\phi(z)$ and to the CDF of a Standard Normal by $\Phi(z)$.
- The values of $\Phi(z)$, $\mathbb{P}(Z \leq z)$ are tabulated.

Useful Properties

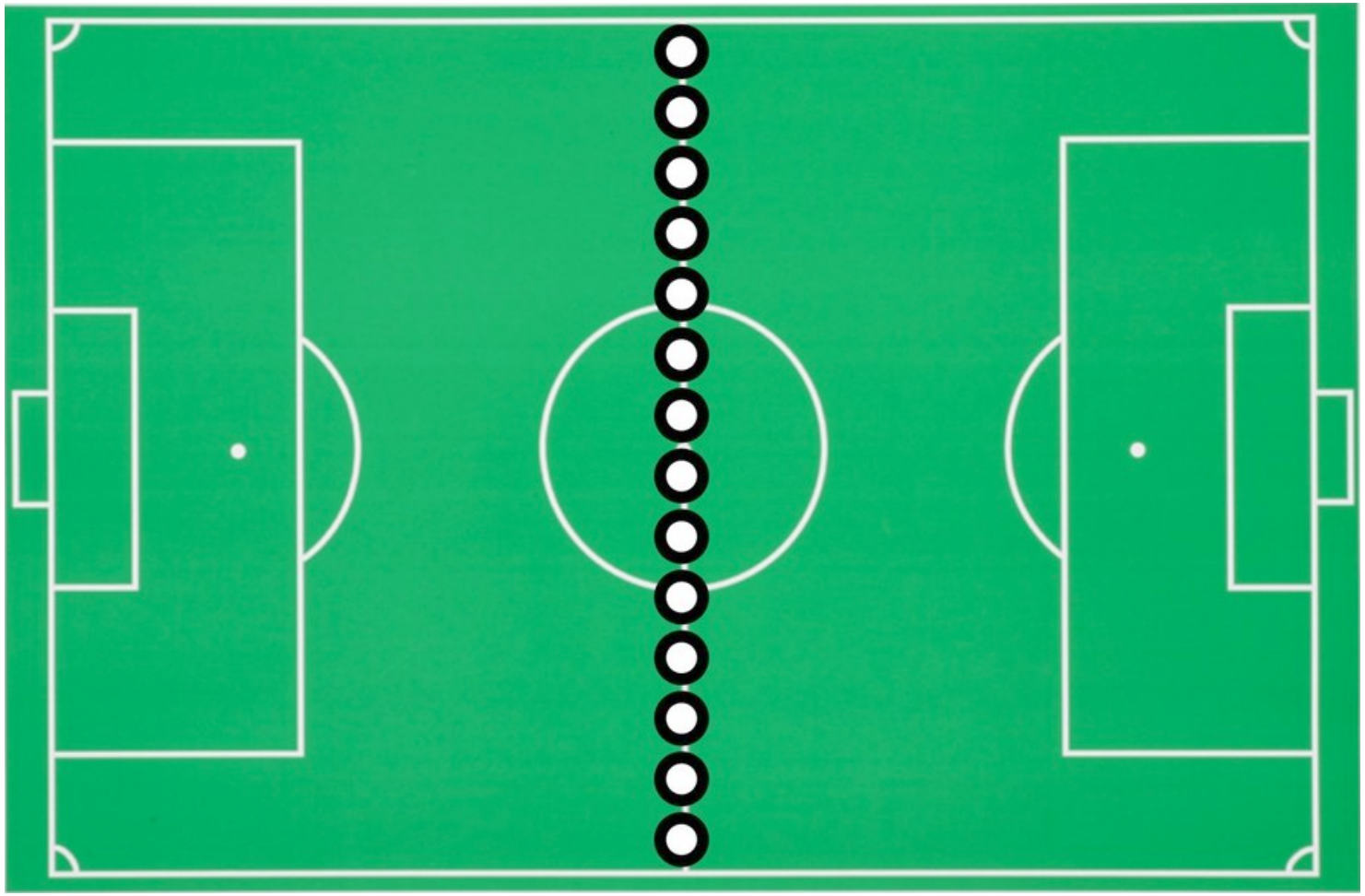
1. If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$
2. If $Z \sim N(0, 1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$
3. Let $X_i \sim N(\mu_i, \sigma_i^2)$,$i = 1, \ldots, n$ be independent R.Vs:

$$\sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$$

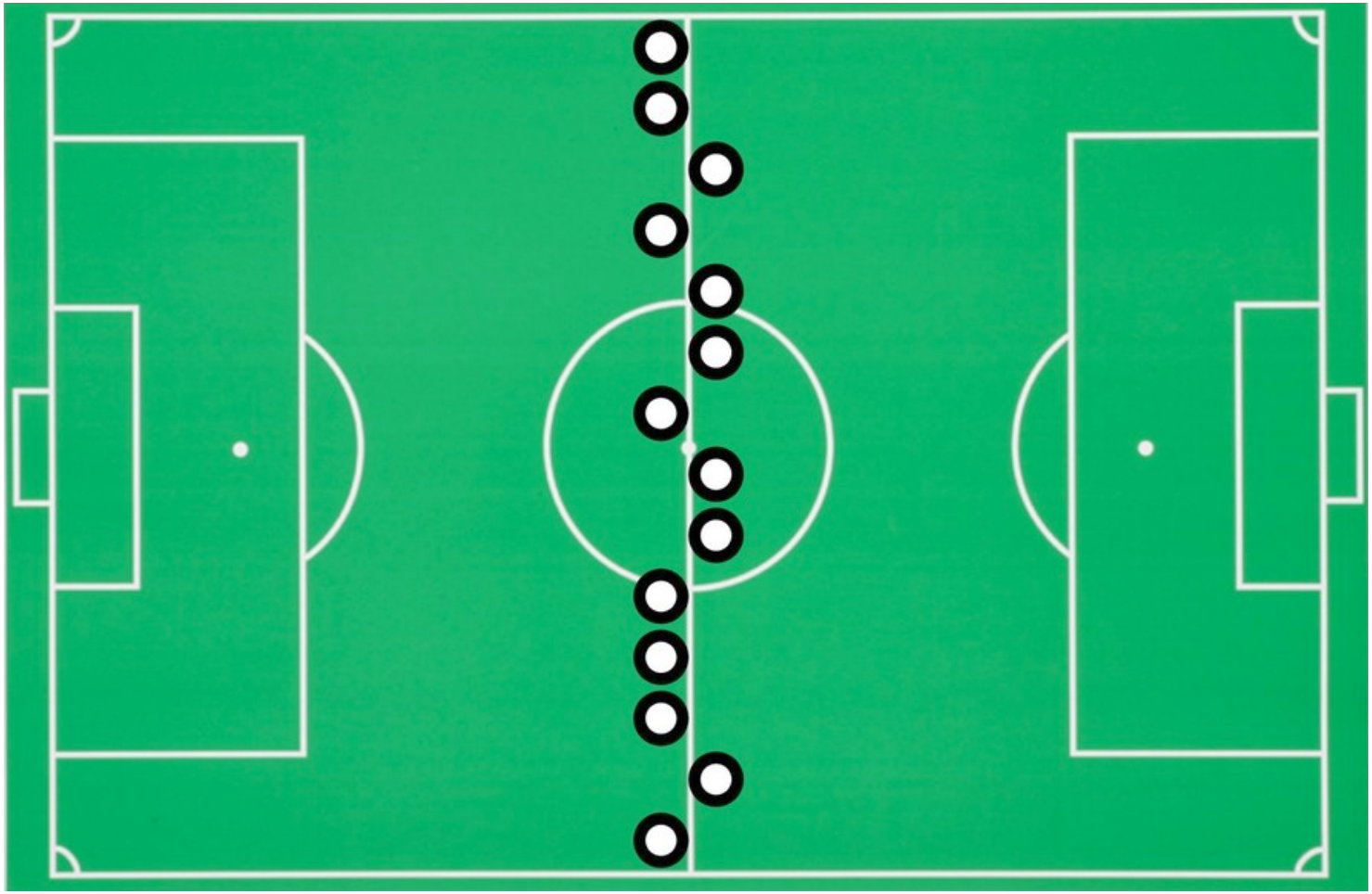
Normal Distribution

- Suppose you and a thousand of your closest friends line up on the halfway line of a soccer field (football pitch).
- Each of you has a coin in your hand. At the sound of the whistle, you begin flipping the coins.
- Each time a coin comes up heads, that person moves one step towards the left-hand goal. Each time a coin comes up tails, that person moves one step towards the right-hand goal.
- Each person flips the coin 16 times, follows the implied moves, and then stands still.
- Now we measure the distance of each person from the halfway line.
- Can you predict what proportion of the thousand people who are standing on the halfway line? How about the proportion 5 yards left of the line?
- It's hard to say where any individual person will end up, but you can say with great confidence what the collection of positions will be.
- The distances will be distributed in approximately normal, or Gaussian, fashion.
- This is true even though the underlying distribution is binomial.
- It does this because there are so many more possible ways to realize a sequence of left-right steps that sums to zero.
- There are slightly fewer ways to realize a sequence that ends up one step left or right of zero, and so on, with the number of possible sequences declining in the characteristic bell curve of the normal distribution.

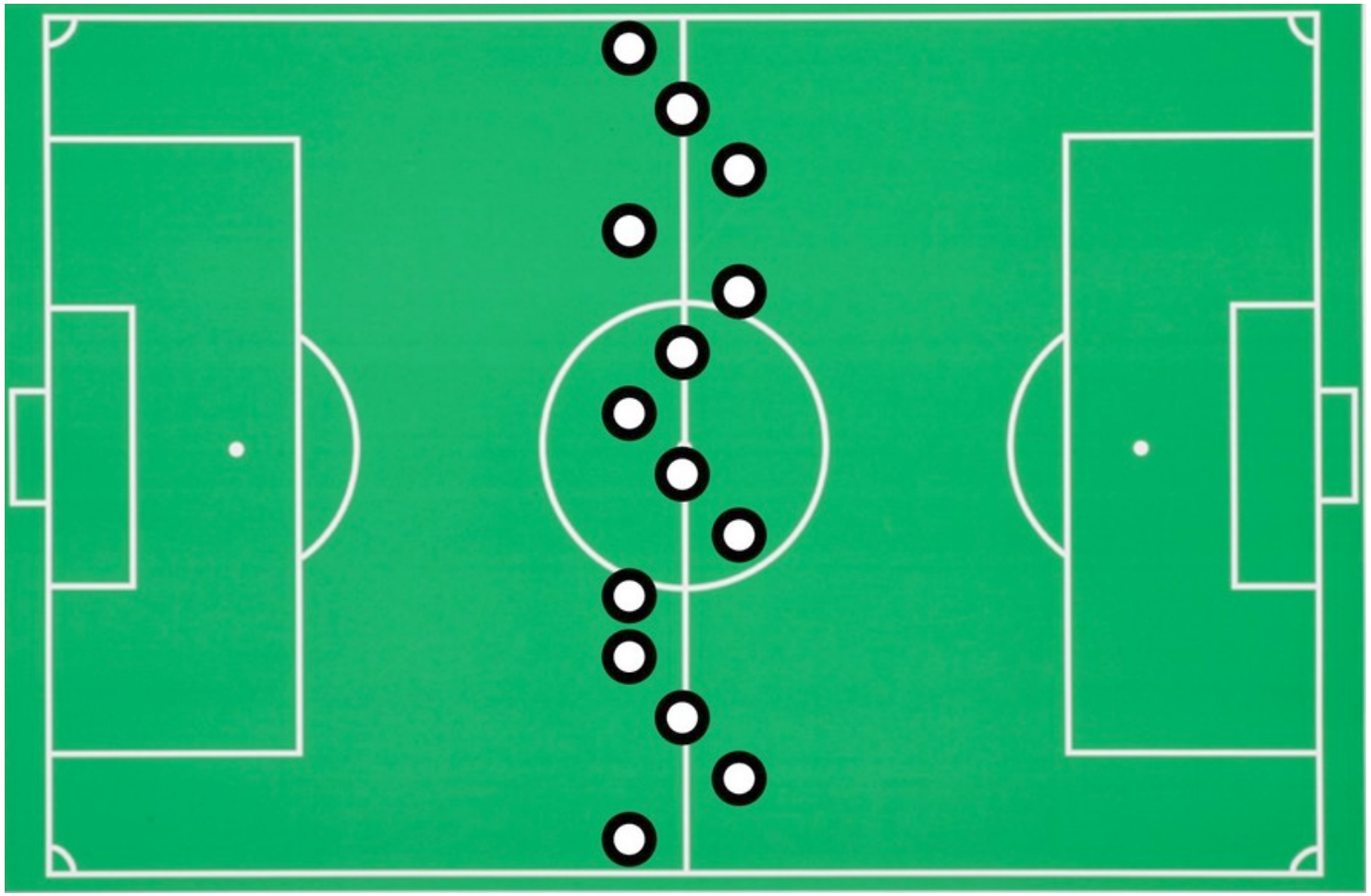
Soccer 1



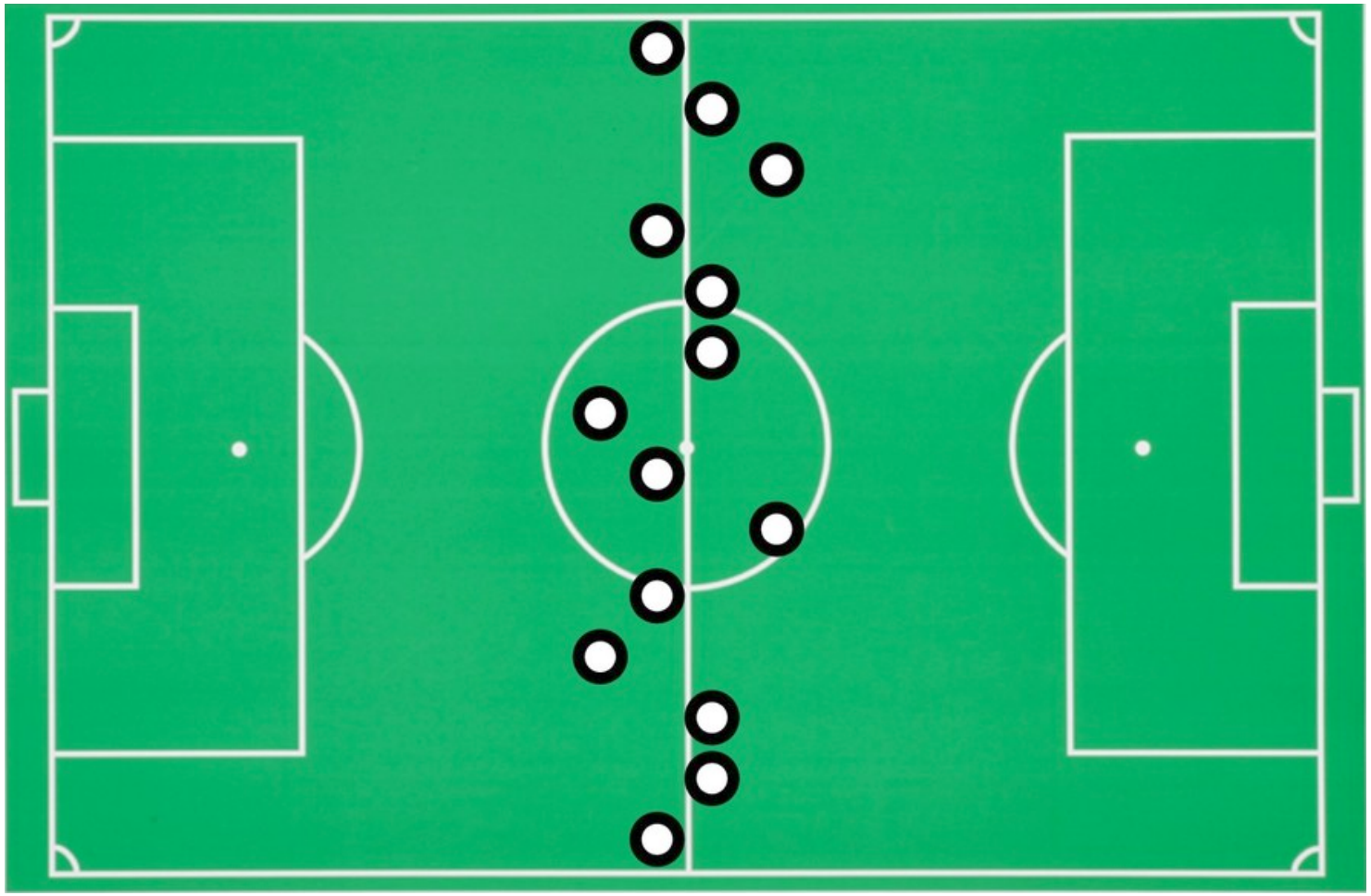
Soccer 2



Soccer 3



Soccer 4



Example Normal

- In R we can access the PDF, CDF, quantile function and random number generation of the distributions.
- For a Normal distribution the R commands are:

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Example

Let $X \sim N(3, 5)$, calculate $\mathbb{P}(X > 1)$

$\mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}(Z < \frac{1-3}{\sqrt{5}}) = 1 - \Phi(-0.8944) = 0.81$

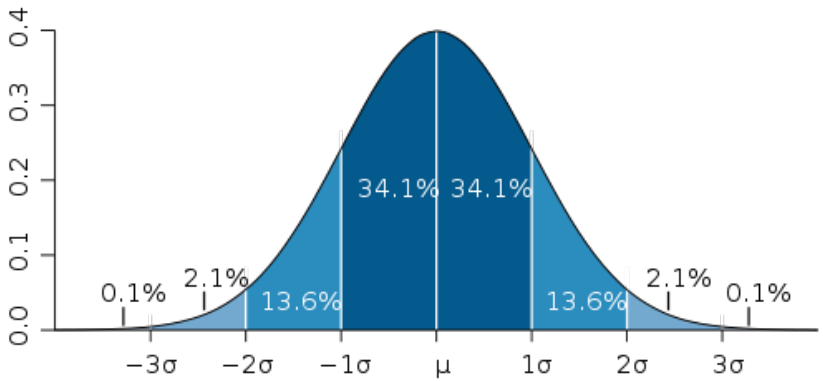
In R:

```
 > 1-pnorm(q=(1-3)/sqrt(5))
[1] 0.8144533
```

Or directly:

```
> 1-pnorm(q=1,mean=3,sd=sqrt(5))
[1] 0.8144533
```

The 68-95-99.7 rule of a Normal Distribution



Let X be a R.V $simN(\mu, \sigma^2)$.

- $\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6827$
- $\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545$
- $\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$

In R for $X \sim N(0, 1)$:

```
> pnorm(1)-pnorm(-1)
[1] 0.6826895
> pnorm(2)-pnorm(-2)
[1] 0.9544997
> pnorm(3)-pnorm(-3)
[1] 0.9973002
```

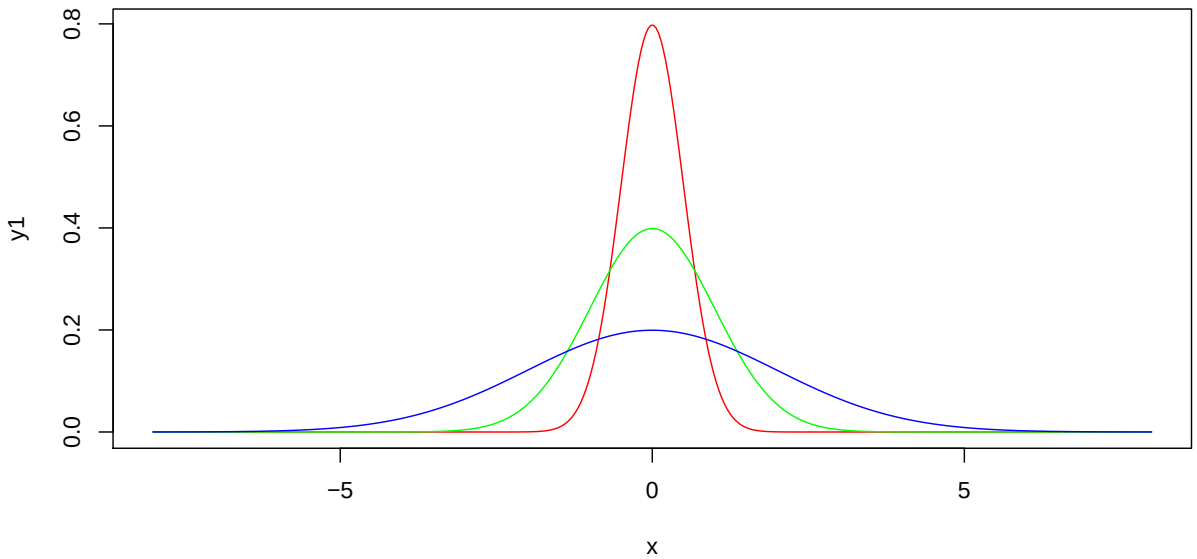
Symmetry of the Normal Distribution

- The PDF of a normal is symmetric around μ.
- Then $\phi(z) = \phi(-z)$
- $\Phi(z) = 1 - \Phi(-z)$

```
> dnorm(1)
[1] 0.2419707
> dnorm(-1)
[1] 0.2419707
> pnorm(0.95)
[1] 0.8289439
> 1-pnorm(-0.95)
[1] 0.8289439
```

Plotting the PDF of Normals with different variance in R

```
x=seq(-8,8,length=400)
y1=dnorm(x,mean=0,sd=0.5)
y2=dnorm(x,mean=0,sd=1)
y3=dnorm(x,mean=0,sd=2)
plot(y1~x,type="l",col="red")
lines(y2~x,type="l",col="green")
lines(y3~x,type="l",col="blue")
```



Joint and Conditional Probabilities

- The notion of probability function (mass or density) can be **extended** to more than one R.V.
- Let X and Y be two R.Vs, $\mathbb{P}(X, Y)$ represents the joint probability function.
- X and Y are independent of each other, if

$$\mathbb{P}(X, Y) = \mathbb{P}(X) \times \mathbb{P}(Y)$$

- The **conditional probability** for Y given X is defined as:

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)}$$

- If X and Y are independent $\mathbb{P}(Y|X) = \mathbb{P}(Y)$

Joint and Conditional Probabilities (2)

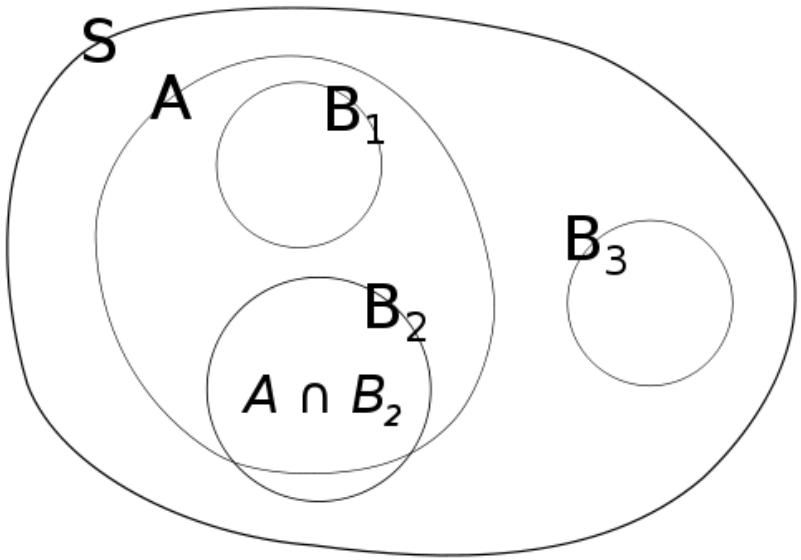


Figure: Source: en.wikipedia.org/wiki/Conditional_probability

- Let S be the sample space, A and B_n events.
- The probabilities are proportional to the area.
- $\mathbb{P}(A) \sim 0.33$, $\mathbb{P}(A|B_1) = 1$
- $\mathbb{P}(A|B_2) \sim 0.85$ y $\mathbb{P}(A|B_3) = 0$

Bayes' Theorem and Total Probabilities

- The conditional probability $\mathbb{P}(Y|X)$ and $\mathbb{P}(X|Y)$ can be expressed as a function of each other using Bayes' theorem.

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}$$

- $P(Y|X)$ can be interpreted as the fraction of times Y occurs when X is known to occur.
- Then let $\{Y_1, Y_2, \ldots, Y_k\}$ be a set of mutually exclusive events of the sample space of a R.V X, the denominator of Bayes' theorem can be expressed as:

$$\mathbb{P}(X) = \sum_{i=1}^{k} \mathbb{P}(X, Y_i) = \sum_{i=1}^{k} \mathbb{P}(X|Y_i)\mathbb{P}(Y_i)$$

Example

- I split my emails into three categories: A_1="“spam”, A_2="“low priority”, A_3="“high priority”.’
- We know that $\mathbb{P}(A_1) = 0.7$, $\mathbb{P}(A_2) = 0.2$ and $\mathbb{P}(A_3) = 0.1$, clearly $0.7 + 0.2 + 0.1 = 1$.
- Let B be the event that the mail contains the word “free”.
- We know that $\mathbb{P}(B|A_1) = 0.9$ $\mathbb{P}(B|A_2) = 0.01$ y $\mathbb{P}(B|A_3) = 0.01$ clearly $0.9 + 0.01 + 0.01 \neq 1$
- What is the probability that an email with the word “free” in it is “spam”?
- Using Bayes and Total Probabilities:

$$\mathbb{P}(A_1|B) = \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = 0.995$$

Expectation

- Let X be a R.V, we define its **expectation** or **first-order moment** as:

$$\mathbb{E}(X) = \begin{cases} \sum_x (x \times f(x)) & \text{If } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x \times f(x)) dx & \text{If } X \text{ is continuous} \end{cases}$$

- The expectation is the weighted average of all the possible values that a random variable can take.
- For the case of tossing a coin twice with X the number of heads:

$$\begin{aligned} \mathbb{E}(X) &= (0 \times f(0)) + (1 \times f(1)) + (2 \times f(2)) \\ &= (0 \times (1/4)) + (1 \times (1/2)) + (2 \times (1/4)) = 1 \end{aligned}$$

- Let the random variables $X_1, X_2, \ldots, X_n$ and the constants $a_1, a_2, \ldots, a_n$,

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i)$$

Variance

- The variance measures the "dispersion" of a distribution.
- Lex X be a R.V of mean μ, we define the variance of X denoted as σ^2, σ_X^2 or $\mathbb{V}(X)$ as:

$$\mathbb{V}(X) = \mathbb{E}(X-\mu)^2 = \begin{cases} \sum_{i=1}^{n} f_x(x_i)(x_i-\mu)^2 & \text{If } X \text{ is discrete} \\ \int (x-\mu)^2 f_X(x)dx & \text{If } X \text{ is continuous} \end{cases}$$

- The **standard deviation** σ is defined as $\sqrt{\mathbb{V}(X)}$

Properties

- $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2) - \mu^2$
- If a and b are constants, then $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$
- If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{V}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathbb{V}(X_i)$$

Law of the Large Numbers

Weak Form

- Let $X_1, X_2, \ldots X_n$ be IID random variables of mean μ and variance σ^2.
- The mean $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$ onverges in probability to μ, $\overline{X_n} \xrightarrow{P} \mu$
- This is equivalent to saying that for all $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}(|\overline{X_n} - \mu| < \epsilon) = 1$$

- Then the distribution of $\overline{X_n}$ becomes centered around μ as n grows.

Example

- Let be the experiment of flipping a coin where the probability of heads is p.
- For a Bernoulli distributed R.V $E(X) = p$.
- Let be $\overline{X_n}$ the fraction of heads after n tosses.
- The law of large numbers tells us that $\overline{X_n}$ converges in probability to p.
- This does not imply that $\overline{X_n}$ is numerically equal to p.
- But if n in large enough, the distribution of $\overline{X_n}$ will be centered around p.

Central Limit Theorem

- While the law of large numbers tells us that $\overline{X_n}$ approaches μ as n grows.
- This is not sufficient to say anything about the distribution of $\overline{X_n}$.

Central Limit Theorem (CLT)

- Let X_1, X_n be IID random variables of mean μ and variance σ^2.
- Let $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$

$$Z_n \equiv \frac{\overline{X_n} - \mu}{\sqrt{\mathbb{V}(\overline{X_n})}} = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow Z$$

  where $Z \sim N(0,1)$
- This is equivalent to:

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Central Limit Theorem (2)

- The theorem allows us to approximate the distribution of $\overline{X_n}$ to a Gaussian distribution when n is large.
- Even if we do not know the distribution of X_i, we can approximate the distribution of its mean.

Alternative notations showing that Z_n converges to a Normal

$$
\begin{aligned}
Z_n &\approx N(0,1) \\
\overline{X_n} &\approx N\left(\mu, \frac{\sigma^2}{n}\right) \\
\overline{X_n} - \mu &\approx N\left(0, \frac{\sigma^2}{n}\right) \\
\sqrt{n}(\overline{X_n} - \mu) &\approx N(0, \sigma^2) \\
\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} &\approx N(0,1)
\end{aligned}
$$

Central Limit Theorem (3)

- Suppose that the number of errors of a computer program follows a Poisson distribution with parameter $\lambda = 5$
- If $X \sim Poisson(\lambda)$, $\mathbb{E}(X) = \lambda$ and $\mathbb{V}(X) = \lambda$.
- If we have 125 independent programs $X_1, \ldots, X_{125}$ we would like to approximate $\mathbb{P}(\overline{X_n} < 5.5)$
- Using the CLT we have that

$$\begin{aligned} \mathbb{P}(\overline{X_n} < 5.5) &= \mathbb{P}\left(\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{5.5 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\ &\approx \mathbb{P}\left(Z < \frac{5.5 - 5}{\frac{\sqrt{5}}{\sqrt{125}}}\right) = \mathbb{P}(Z < 2.5) = 0.9938 \end{aligned}$$

References I


Poldrack, R. A. (2019).
Statistical Thinking for the 21st Century.


Wasserman, L. (2013).
All of statistics: a concise course in statistical inference.
Springer Science & Business Media.