Probability

Felipe José Bravo Márquez

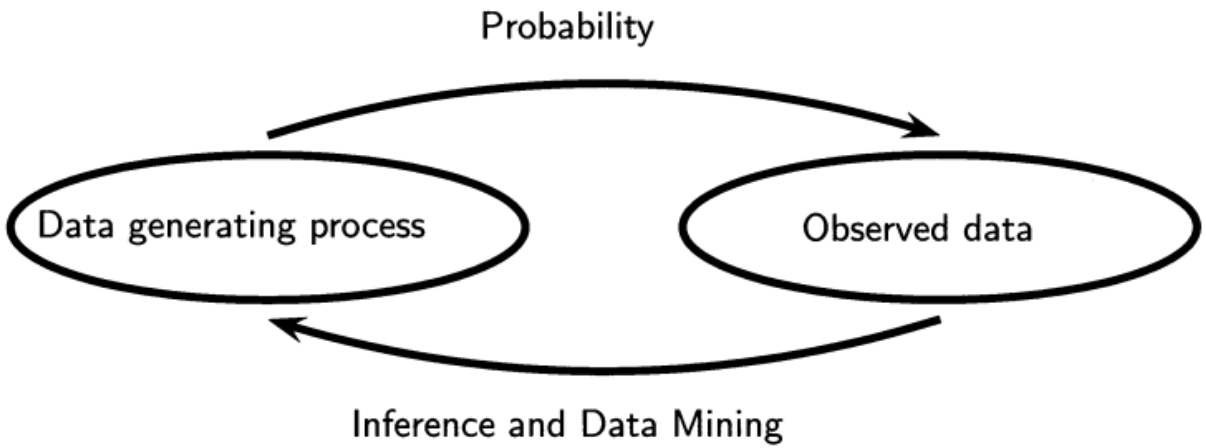
August 18, 2021

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 probability 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?

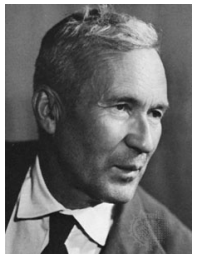
- We think of probability as a number that describes the likelihood of some event occurring, which ranges from zero (impossibility) to one (certainty).
- Probabilities can also be expressed in percentages: when the weather forecast predicts a twenty percent chance of rain today.
- In each case, these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

- In R:

```
> factorial(20)/(factorial(3)*factorial(17))
[1] 1140
> choose(20,3)
[1] 1140
```

- We note the following properties:

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

and

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

```
> choose(20,17)
[1] 1140
```

Conditional Probabilities

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

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

- $\mathbb{P}(Y|X)$ can be interpreted as the fraction of times Y occurs when X is known to occur.
- If X and Y are independent $\mathbb{P}(Y|X) = \mathbb{P}(Y)$

Example

- In the experiment of rolling a fair die, let G be the event of getting and outcome greater than 2 ($G = \{3, 4, 5, 6\}$) and O the event of getting an odd number ($O = \{1, 3, 5\}$).
- What is the value of $\mathbb{P}(G|O)$?

Conditional Probabilities

- Using the definition that $\mathbb{P}(X) = \frac{\text{Favorable cases}}{\text{total cases}}$: $\mathbb{P}(G) = 4/6$, $\mathbb{P}(O) = 3/6$, and $\mathbb{P}(G|O) = 2/3$.
- Notice that once we know O, the number of favourable cases for G was reduced to $\{3, 5\}$ and the total number of cases to $\{1, 3, 5\}$.
- Now, according to the definition above $\mathbb{P}(G|O) = \frac{\mathbb{P}(G,O)}{\mathbb{P}(O)}$.
- Where $\mathbb{P}(G, O) = 2/6$ (favourable cases correspond to the intersection between G and O $\{3, 5\}$).
- So, $\mathbb{P}(G|O) = \frac{\mathbb{P}(G,O)}{\mathbb{P}(O)} = \frac{2/6}{3/6} = 2/3$.

Warning

- In general it is not the case that $\mathbb{P}(Y|X) = \mathbb{P}(X|Y)$.
- People get this confused all the time.
- For example, the probability of spots given you have measles is 1 but the probability that you have measles given that you have spots is not 1.
- In this case, the difference between $\mathbb{P}(Y|X)$ and $\mathbb{P}(X|Y)$ is obvious but there are cases where it is less obvious.

Conditional Probabilities: Example

- A medical test for a disease D has outcomes “positive” and “negative”, the probabilities are:

	D	¬ D
positive	0.009	0.099
negative	0.001	0.891

- From the definition of conditional probability

$$\mathbb{P}(\text{positive}|D) = \frac{\mathbb{P}(\text{positive}, D)}{\mathbb{P}(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

```
> pos.d <- 0.009/(0.009+0.001)
> pos.d
[1] 0.9
```

and

$$\mathbb{P}(\text{negative}|\neg D) = \frac{\mathbb{P}(\text{negative}, \neg D)}{\mathbb{P}(\neg D)} = \frac{0.891}{0.891 + 0.0991} \approx 0.9$$

```
> neg.notd <- 0.891/(0.891+0.0991)
> neg.notd
[1] 0.8999091
```

Conditional Probabilities: Example

- Apparently, the test is fairly accurate.
- Sick people yield a positive 90 percent of the time and healthy people yield a negative about 90 percent of the time.
- Suppose you go for a test and get a positive.
- What is the probability you have the disease? Most people answer 0.90.
- The correct answer is

$$\mathbb{P}(D|\text{positive}) = \frac{\mathbb{P}(D, \text{positive})}{\mathbb{P}(\text{positive})} = \frac{0.009}{0.009 + 0.099} \approx 0.08$$

```
> d.pos <- 0.009/(0.009+0.099)
> d.pos
[1] 0.08333333
```

- The lesson here is that you need to compute the answer numerically.
- Don't trust your intuition.

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.

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

- 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:

$$\mathbb{P}(A_1|B) = \frac{\mathbb{P}(B|A_1) \times \mathbb{P}(A_1)}{\mathbb{P}(B)} = \frac{0.9 \times 0.7}{\mathbb{P}(B)} = \frac{0.63}{\mathbb{P}(B)}$$

Example

- Using Total Probabilities:

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B|A_1) \times \mathbb{P}(A_1) + \mathbb{P}(B|A_2) \times \mathbb{P}(A_2) + \mathbb{P}(B|A_3) \times \mathbb{P}(A_3) \\ &= 0.9 \times 0.7 + 0.01 \times 0.2 + 0.01 \times 0.1 = 0.633\end{aligned}$$

Finally,

$$\mathbb{P}(A_1|B) = \frac{0.63}{0.633} = 0.995$$

- In R:

```
> a1 <-0.7
> a2 <- 0.2
> a3 <-0.1
> b.a1 <- 0.9
> b.a2<-0.01
> b.a3<-0.01
> b<-b.a1*a1+b.a2*a2+b.a3*a3
> a1.b<-b.a1*a1/b
> a1.b
[1] 0.9952607
```

Random Variable

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

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

  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 $X(e) = 6$

Random Variable

- 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 $e \in \{1, 2, 3, 4, 5, 6\}$.
- In these cases, the notion that a random variable is a function can be confusing (it looks more like a set), but we can always reconstruct the mapping function as the identity function.
- 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.
- For the event $e = \{4, 5\}$, $X(e) = 9$ and $Y(e) = 20$.

Example

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

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

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

R.V Definitions

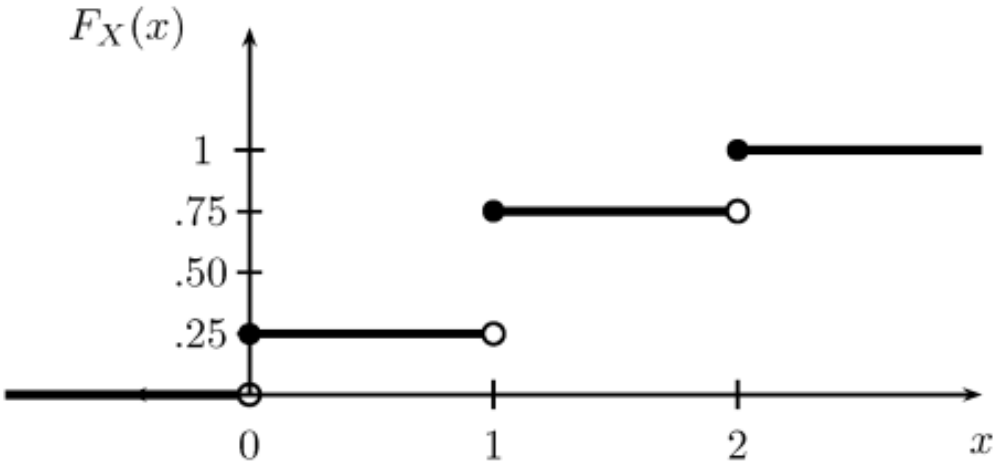
- Let X be a R.V , we define **cumulative distribution function** (CDF) or $F_X : \mathbb{R} \to [0, 1]$ as:

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

- For the previous example of flipping a fair coin twice and counting the number of heads, the CDF is as follows:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2. \end{cases}$$

R.V Definitions



- CDF's can be very confusing.
- Notice that the function is right continuous, non-decreasing, and that it is defined for all x, even though the random variable only takes values 0,1, and 2.
- Notation: when X is a random variable; x denotes a particular value of the random variable.

R.V Definitions

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** PMF 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:

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

The PMF for the previous example is:

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise.} \end{cases}$$

R.V Definitions

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.

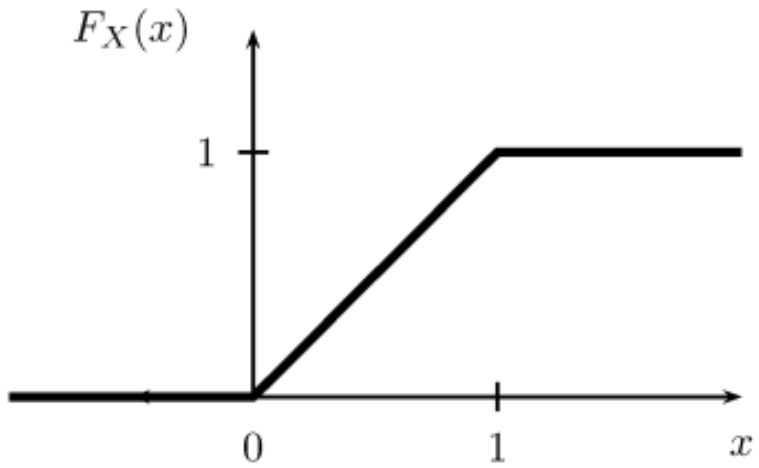
R.V Definitions

- Example: Suppose that X has PDF

$$f_X(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Clearly, $f_X(x) \geq 0$ and $\int f_X(x) = 1$.
- A random variable with this density is said to have a Uniform (0,1) distribution.
- This is meant to capture the idea of choosing a point at random between 0 and 1.
- The CDF is given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1. \end{cases}$$



R.V Definitions

- Continuous random variables can lead to confusion.
- First, note that if X is continuous then $\mathbb{P}(X = x) = 0$ for every x.
- Don't try to think of $f(x)$ as $\mathbb{P}(X = x)$.
- This only holds for discrete random variables.
- We get probabilities from a PDF by integrating.

R.V Definitions

- A PDF can be bigger than 1 (unlike a mass function).
- For example, if $f(x) = 5$ for $x \in [0, 1/5]$ and 0 otherwise.
- Then $f(x) \geq 0$ and $\int f(x)dx = 1$, so this is a well-defined PDF even though $f(x) = 5$ in some places.
- A PDF can also be interpreted as the rate of change in the CDF.
- So where cumulative probability is increasing rapidly, density can easily exceed 1.
- But if we calculate the area under the density function, it will never exceed 1.

R.V Definitions

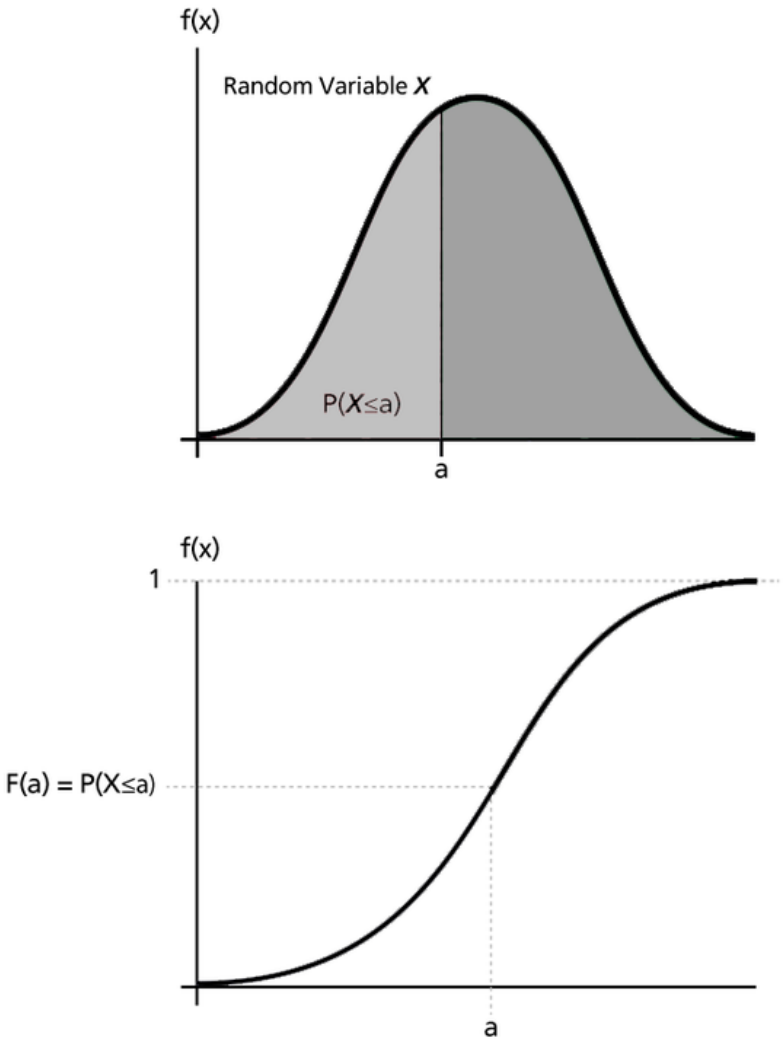


Figure: Source: `http://reliawiki.org/index.php/Basic_Statistical_Background`

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.

Quantiles

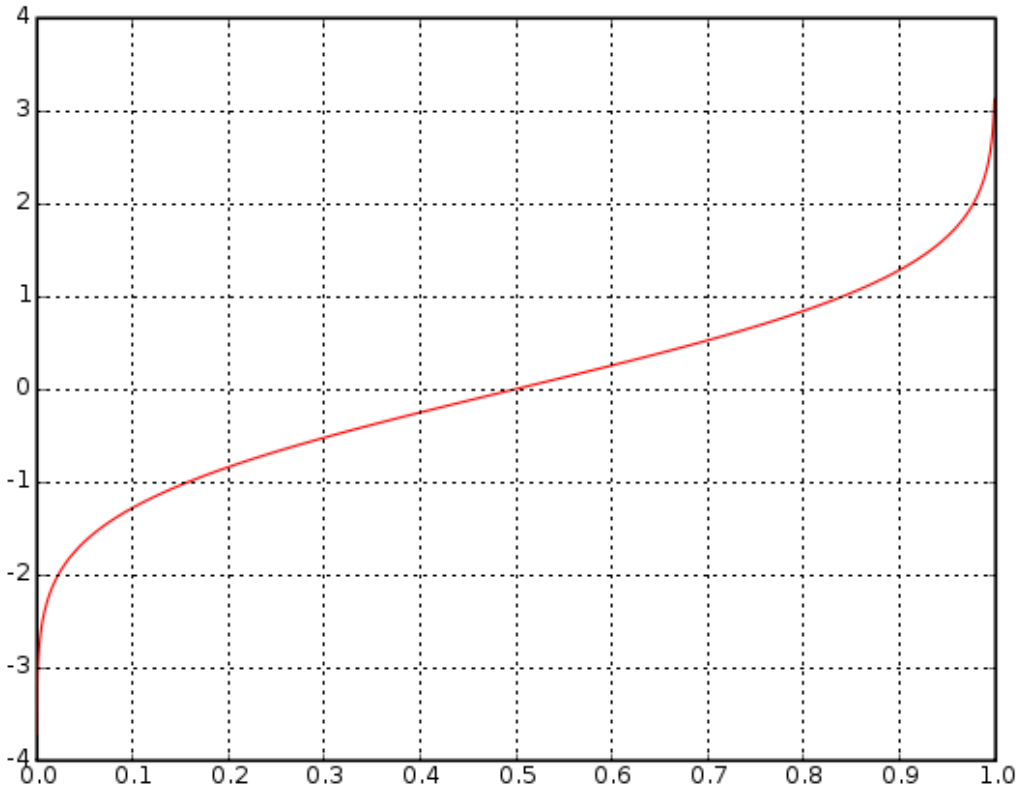


Figure: The quantile function of the Normal distribution (probit)[2]

[2] https://en.wikipedia.org/wiki/Quantile_function

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?

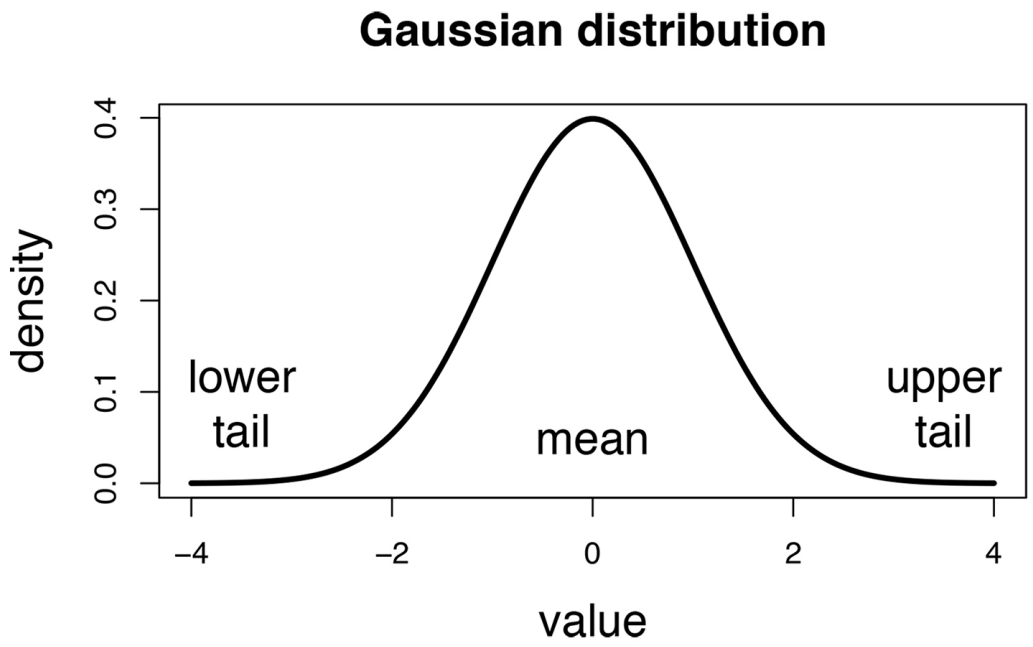
$$f_2(4, 0.91) = \binom{4}{2} 0.91^2 (1 - 0.91)^{4-2} = 0.040$$

In R:

```
> choose(4,2)*0.91^2*(1-0.91)^2
[1] 0.04024566
> # more compactly
> dbinom(x=2,size=4,p=0.91)
[1] 0.04024566
```

The Normal Distribution

- The normal or Gaussian distribution is extremely important in statistics, in part because it shows up all the time in nature.
- It is controlled by two parameters, a mean μ and a standard deviation σ.



- As we will see in the following example [McElreath, 2020], the normal distribution is observed whenever many small, independent variations are summed together to produce a value.

[3]Figure source: `http://sfonline.barnard.edu/wp-content/uploads/2015/12/gaussian-distribution.jpg`

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?

Normal Distribution

- 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

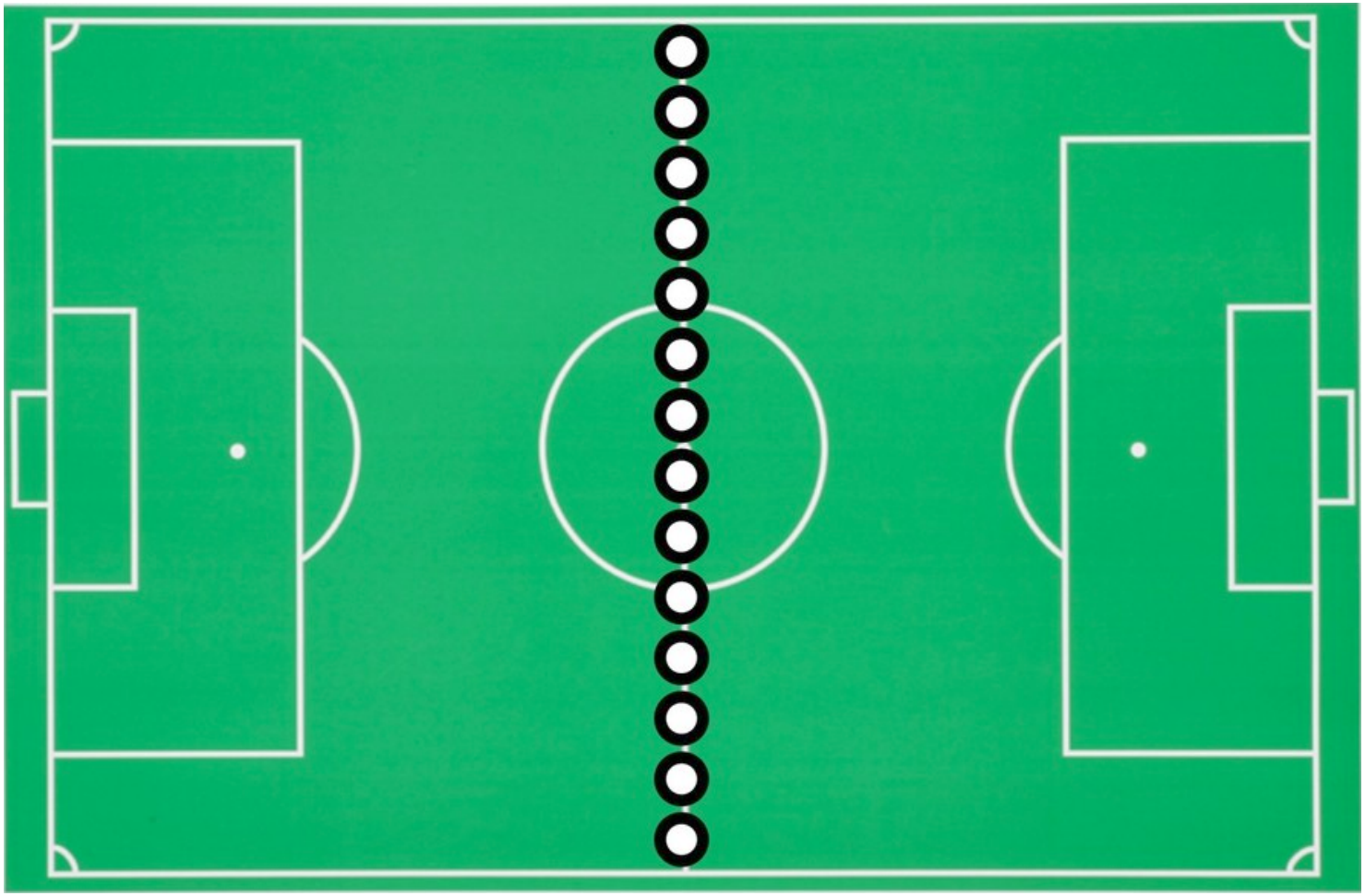
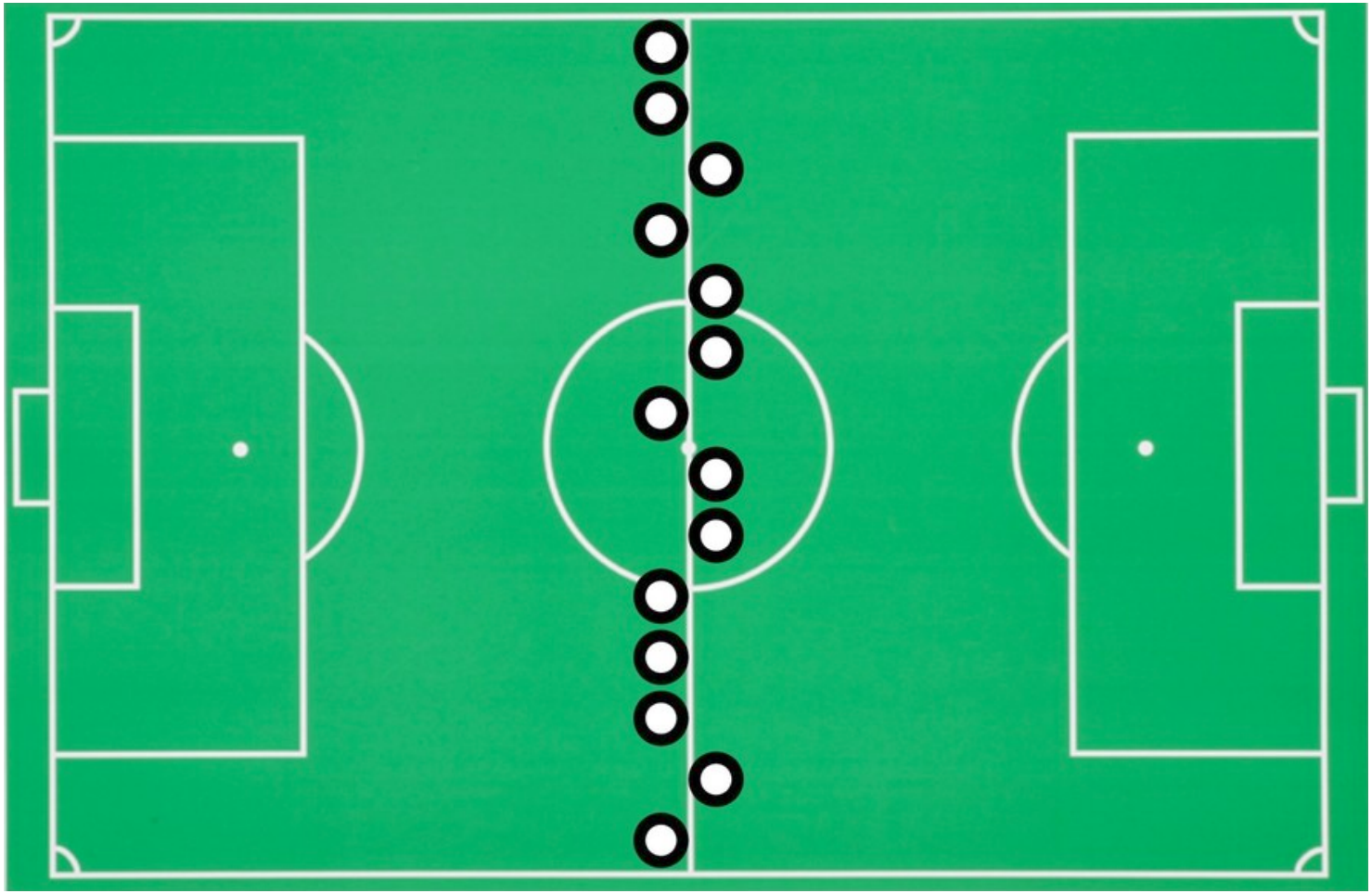
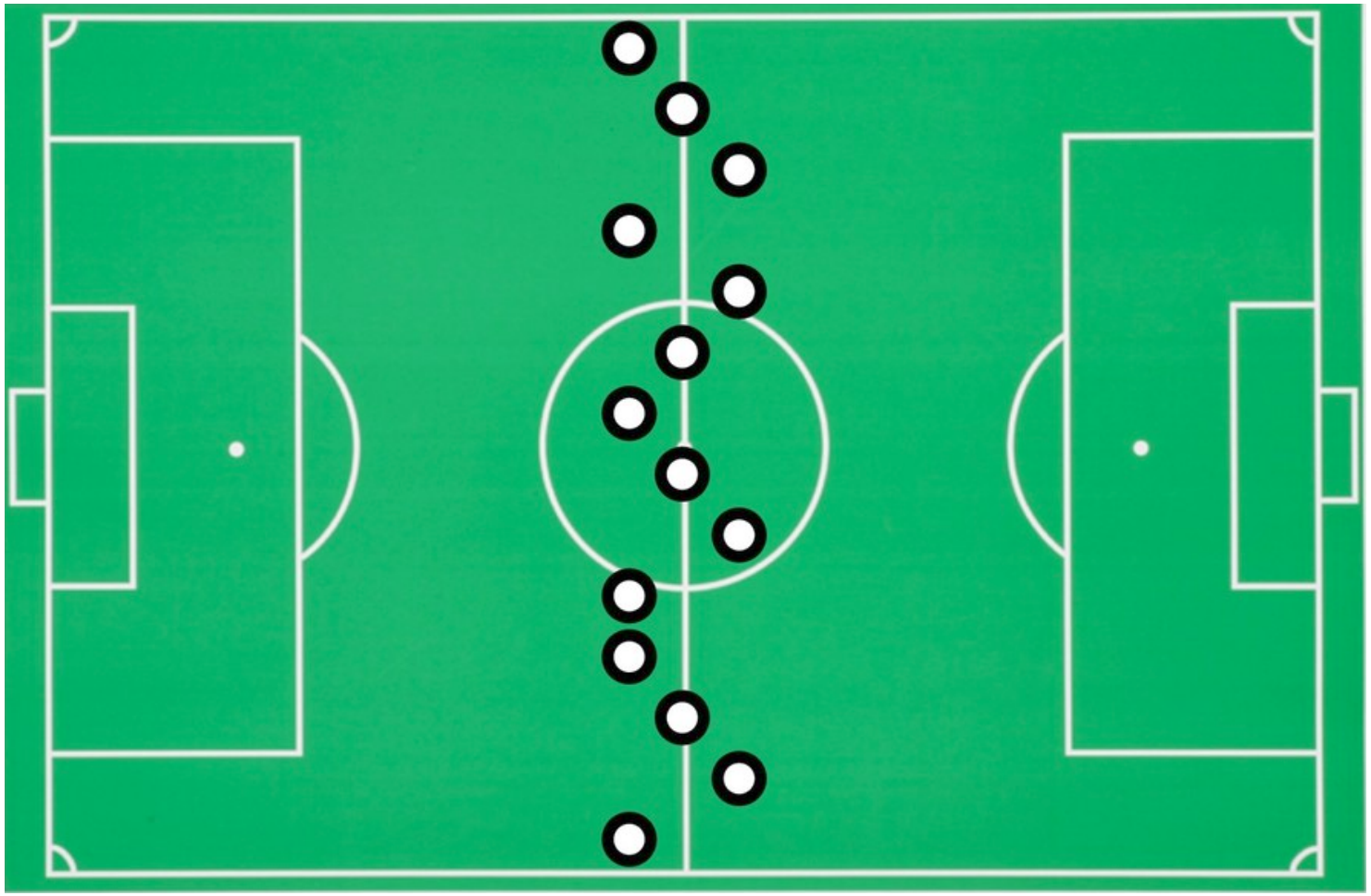


Figure: Source:
https://github.com/rmcelreath/stat_rethinking_2020

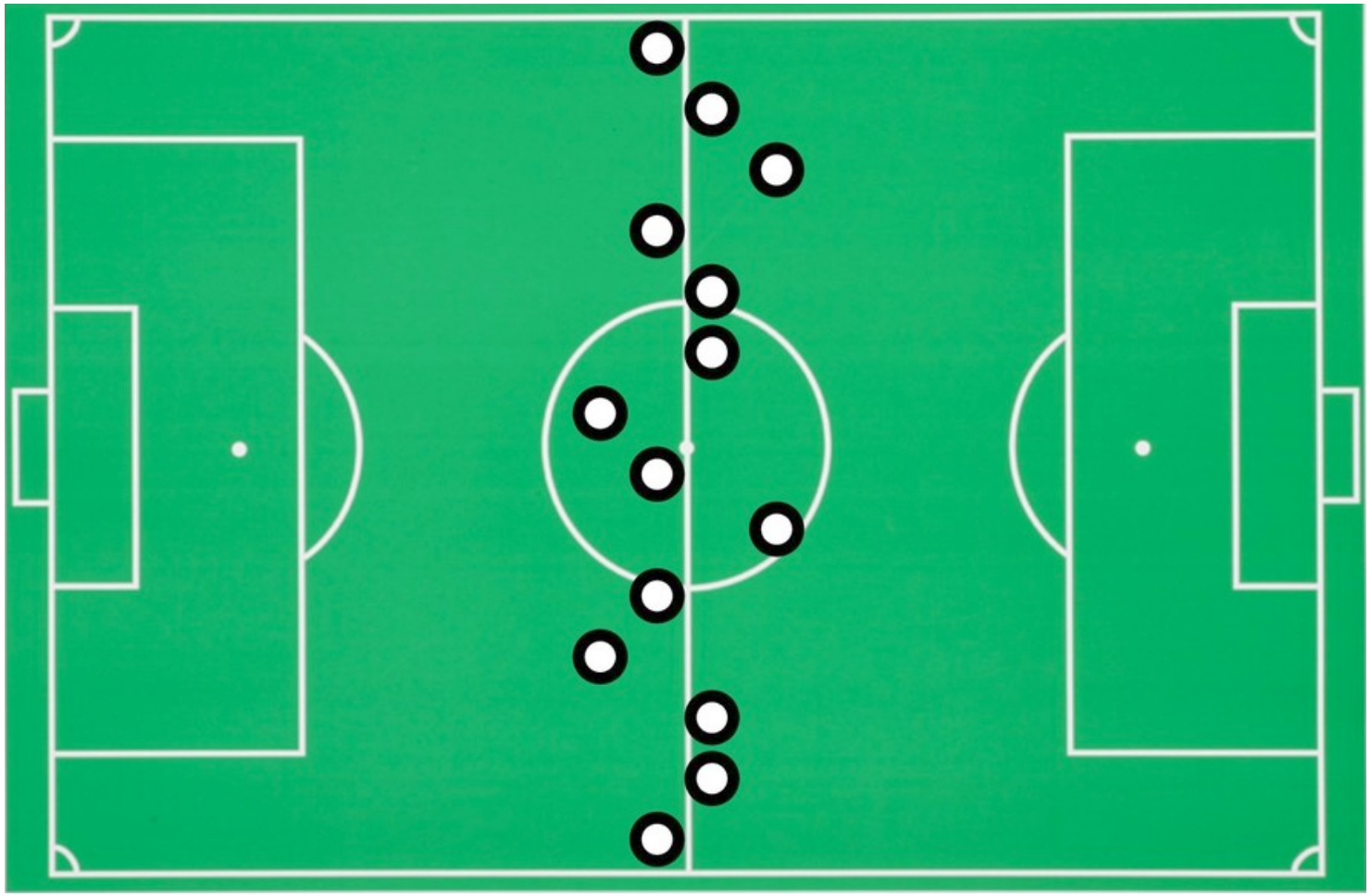
Soccer 2



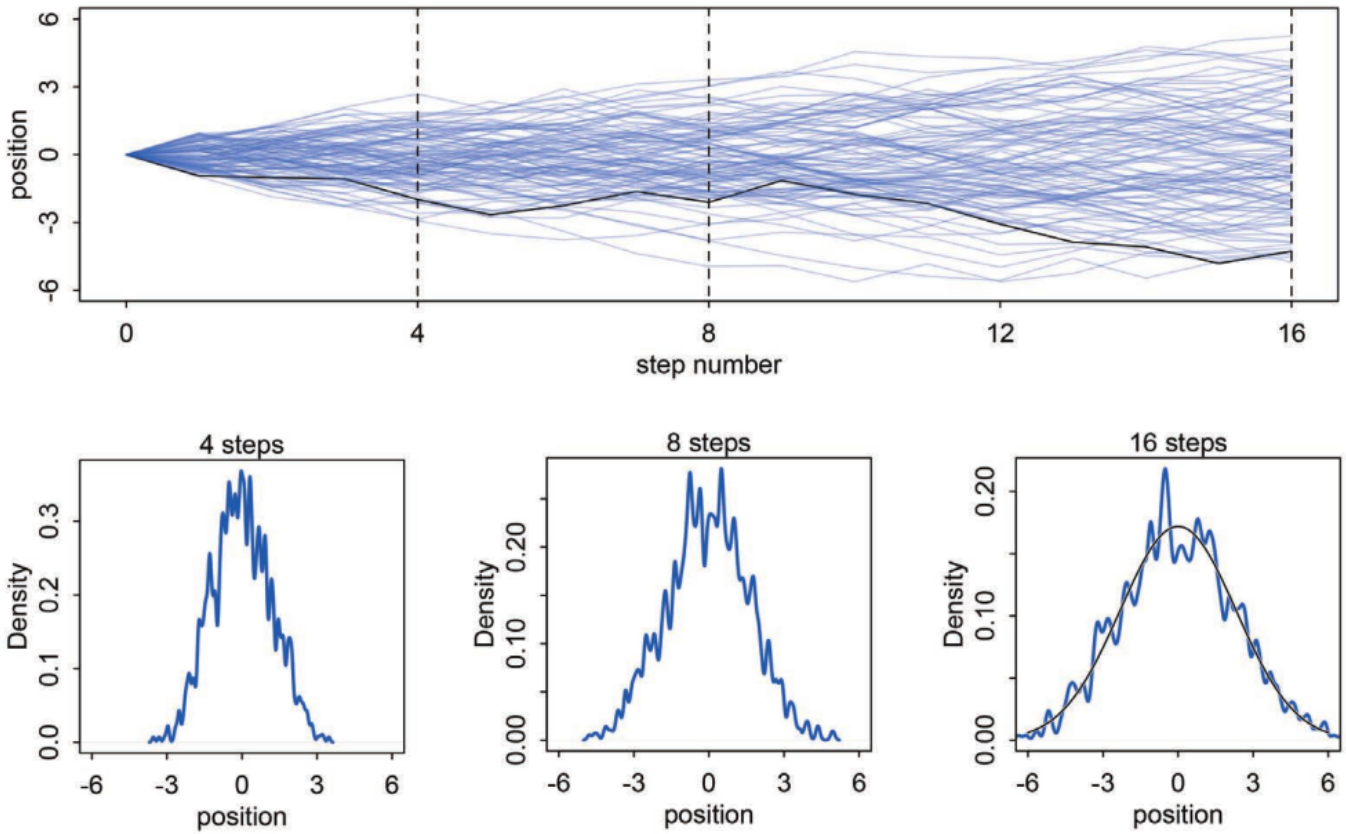
Soccer 3



Soccer 4



Soccer 5



- Random walks on the soccer field converge to a normal distribution.
- The more steps are taken, the closer the match between the real empirical distribution of positions and the ideal normal distribution, superimposed in the last plot in the bottom panel.

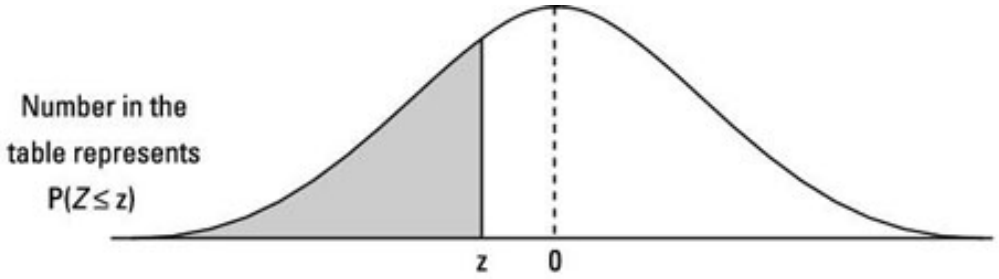
The Normal Distribution

- 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".
- The mean shifts the distribution along the x axis.
- The standard deviation affects the shape such that the larger the σ, the wider the shape.
- 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.

The Normal Distribution



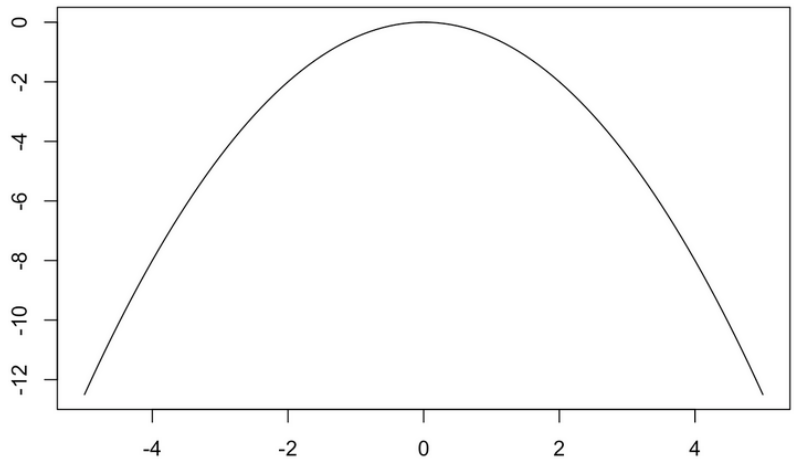
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367

The Normal Distribution

- Let's try to understand the main components of $\phi(z)$ as showed in [Quirk, 2020].

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}x^2}$$

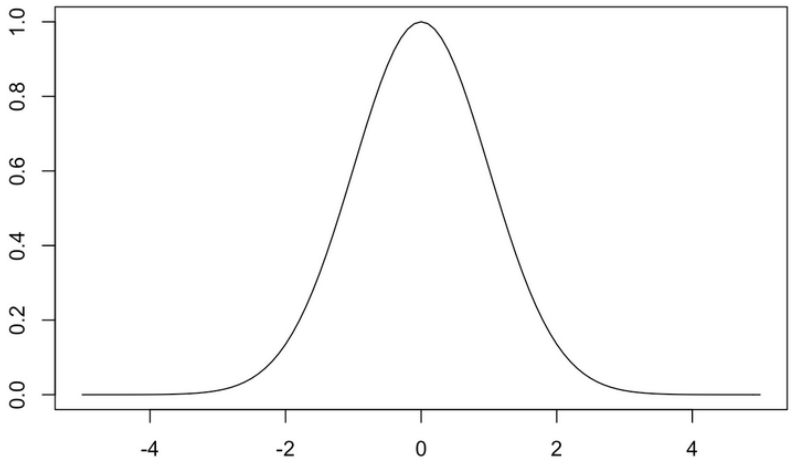
- Here's a plot of just the exponent $-\frac{1}{2}x^2$.



- As you can see, this is a simple parabola.

The Normal Distribution

- When we raise exp to the power of $-\frac{1}{2}x^2$ we get the following plot.



- Taking the exponential of a negative parabola is what gives us the bell curve.
- As is, the area under the curve is $\sqrt{2\pi}$, so the constant $\frac{1}{\sqrt{2\pi}}$ is included to make it equal to 1.
- Finally, the $\frac{1}{2}$ in the exponent ensures that the variance is equal to 1.

The Normal Distribution

- In order to generalize to any normal distribution we simply include the parameters μ and σ

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

- Notice that σ scales the constant, but that the real action happens in the exponent.
- The density's highest point is shifted by μ, then divided by σ.

Useful Properties

1. If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$.[a]
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:

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

[a] This value is also called **Z-score**.

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 Normal Distribution

- If follows from property 1 that if $X \sim N(\mu, \sigma^2)$, then:

$$\mathbb{P}(a < X < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \quad (1)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \quad (2)$$

Example

Let $X \sim N(15, 5)$, calculate $\mathbb{P}(13 < X < 18)$:

$\mathbb{P}(13 < X < 18) = \Phi\left(\frac{18-15}{5}\right) - \Phi\left(\frac{13-15}{5}\right) = \Phi(0.6) - \Phi(-0.4) = 0.381$

In R:

```
> pnorm(0.6)-pnorm(-0.4)
[1] 0.3811686
```

Or directly:

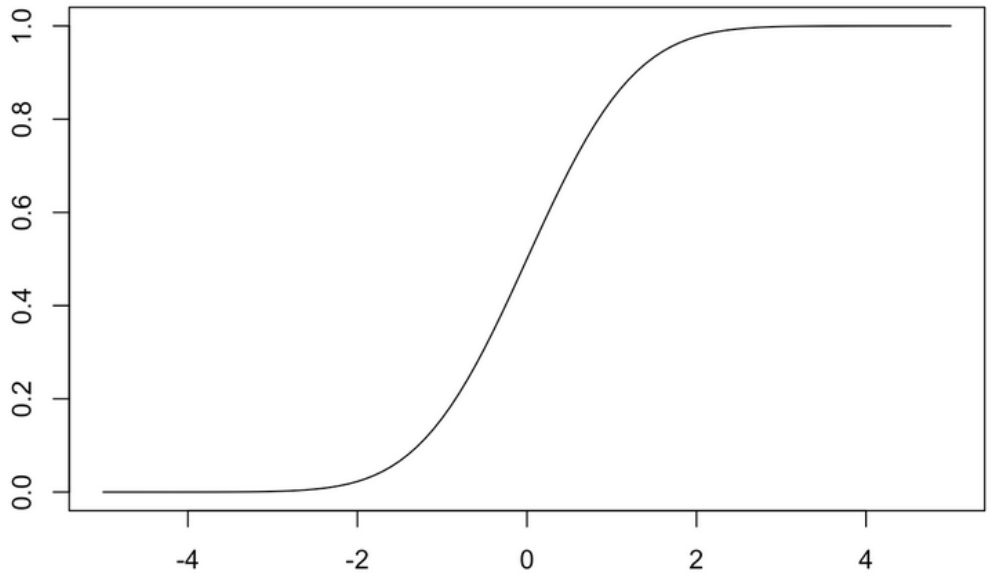
```
> pnorm(18,mean=15,sd=5)-pnorm(13,mean=15,sd=5)
[1] 0.3811686
```

The Normal Distribution's CDF

- As discussed above, when dealing with probability over continuous values we are primarily interested in probability over a range.
- For some purposes, it is convenient to think of a distribution in terms of total probability of an event occurring in the range of $-\infty$ to z.
- For the standard normal distribution we get the following CDF:

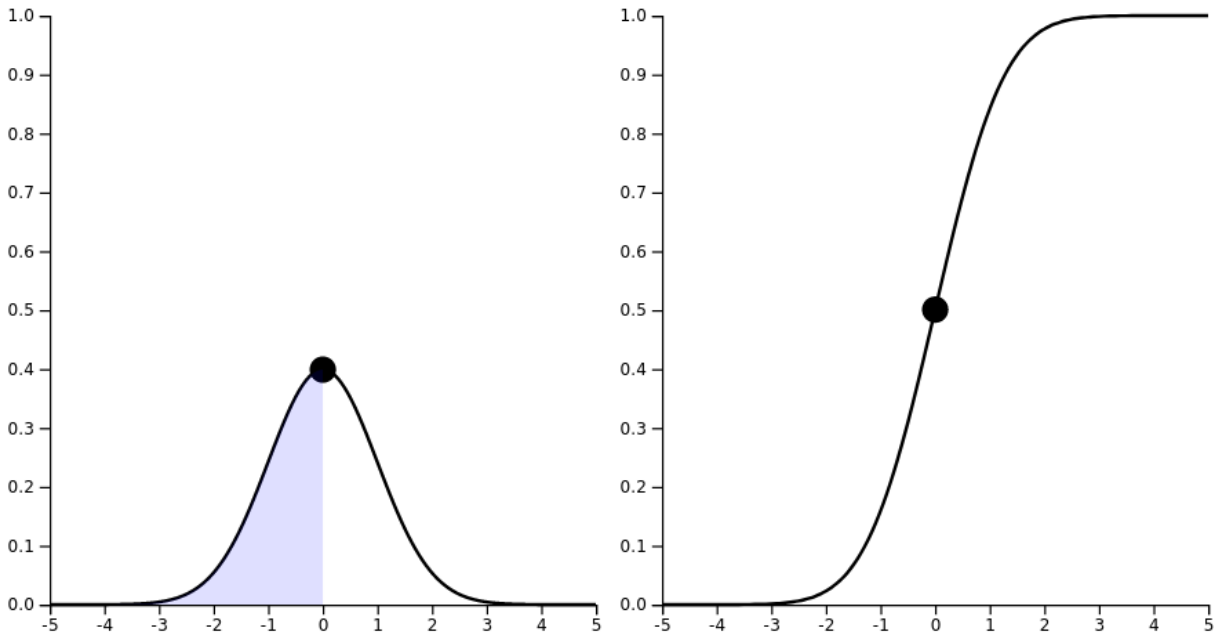
$$\Phi(z) = \mathbb{P}(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp^{\frac{-t^2}{2}} dt$$

- Here is a plot of the standard normal CDF.

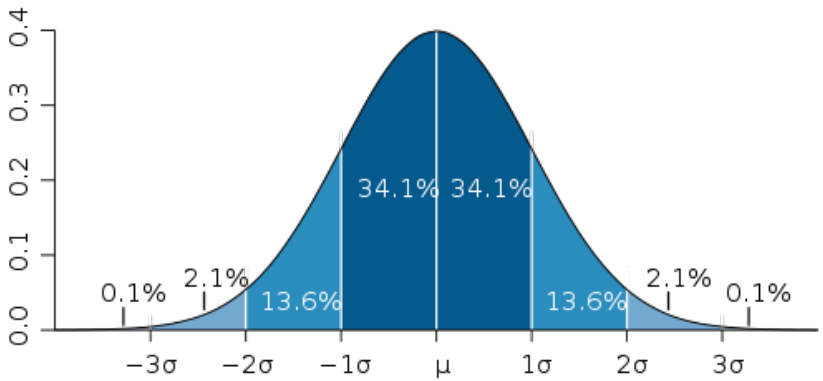


The Normal Distribution's CDF

- The y axis shows cumulative probability so the function is always increasing.
- The CDF is simply expressing the area under the curve (i.e. the integral from $-\infty$ to z) of the PDF.
- With $z = 0$, we are at the mean of the standard normal, so values are equally likely to be less than or greater than z.
- This means that the CDF at $z = 0$ should be 0.5 as we have accumulated half of the available probability.



The 68-95-99.7 rule of a Normal Distribution



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

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

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
```

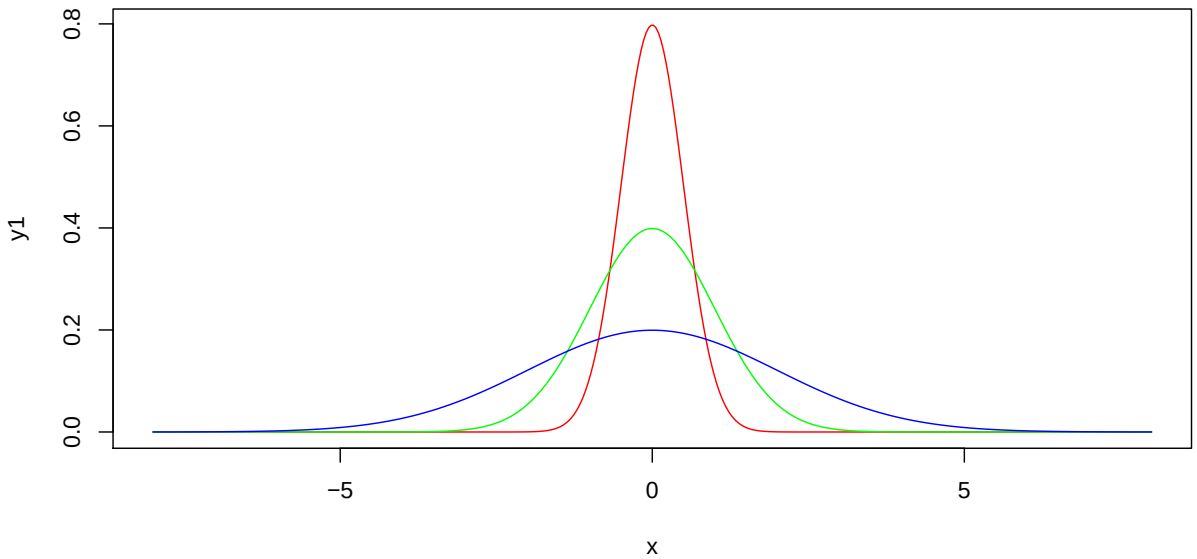
Symmetry of the Normal Distribution

- 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 Probabilities

- The notion of probability function (mass or density) can be **extended** to more than one R.V.
- Given a pair of discrete random variables X and Y, define the joint mass function by $f(x, y) = \mathbb{P}(X = x \text{ and } Y = y)$ or $\mathbb{P}(X = x, Y = y)$.
- We write f as $f_{X,Y}$ when we want to be more explicit.
- Example: Here is a bivariate distribution for two random variables X and Y each taking values 0 or 1:

	$Y = 0$	$Y = 1$	
$X = 0$	1/9	2/9	1/3
$X = 1$	2/9	4/9	2/3
	1/3	2/3	1

- Thus, $f(1, 1) = \mathbb{P}(X = 1, Y = 1) = 4/9$.

Joint Probabilities

Continuous Case

In the continuous case, we call a function $f(x, y)$ a PDF for the random variables (X, Y) if

1. $f(x, y) \geq 0$ for all (x, y),
2. $\int_{-\infty}^{\infty} f(x, y)dxdy = 1$ and,
3. for any set $A \subset \mathbb{R} \times \mathbb{R}, \mathbb{P}((X, Y) \in A) = \int \int_A f(x, y)dxdy$.

In the discrete or continuous case we define the joint CDF as $F_{X,Y} = \mathbb{P}(X \leq x, Y \leq y)$.

Independent Random Variables

- Two random variables X and Y are independent if, for every A and B,

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

- For the discrete case that means that $\mathbb{P}(X, Y) = \mathbb{P}(X) \times \mathbb{P}(Y)$ and for the continuous case we have that $f_{X,Y}(x, y) = f_X(x) \times f_Y(y)$ for all values x and y.

Expectation

- Let X be a R.V, we define its **expectation**, or **mean**, 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.
- Think of $\mathbb{E}(X)$ as the average $\sum_{i=1}^{n} X_i/n$ of a large number of IID draws $X_1, \ldots, X_n$. [4]
- 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 $a_1, a_2, \ldots, a_n$,

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

[4] This is a theorem called the law of large numbers to be discussed next.

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

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

Expectation and Variance of Well-known Distributions

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ,σ^2)	μ	σ^2
Exponential(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
t_ν	0 (if $\nu>1$)	$\nu/(\nu-2)$ (if $\nu>2$)
χ^2_p	p	$2p$
Multinomial(n,p)	np	see below
Multivariate Normal(μ,Σ)	μ	Σ

Source: [Wasserman, 2013].

Law of the Large Numbers

Weak Form

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

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

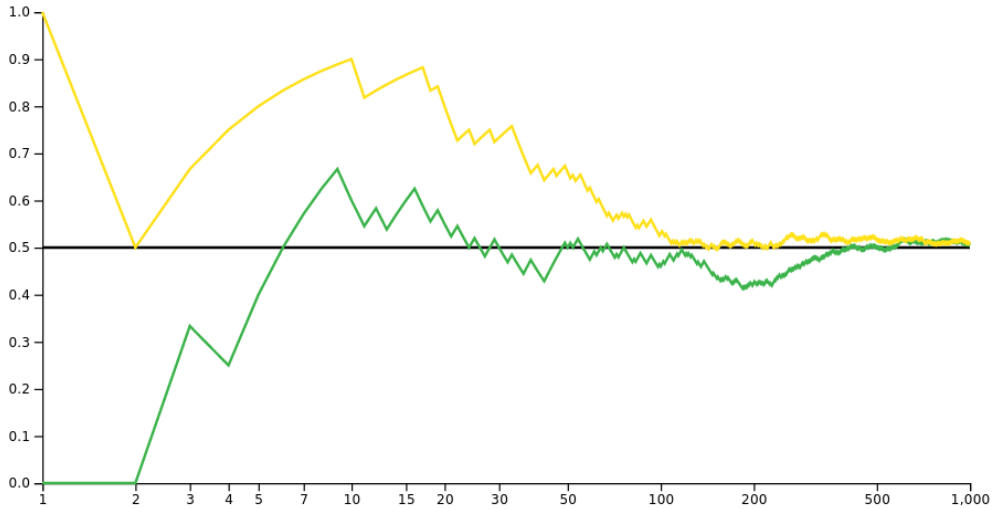
- Then the distribution of $\overline{X_n}$ more concentrated around μ as n grows.

Example

- Let be the experiment of flipping a coin where the probability of heads is p.
- For a Bernoulli distributed R.V $\mathbb{E}(X) = p$.
- 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.

The Law of Large Numbers

- Let's visualize executing the coin tossing experiment twice with $n = 1,000$ and $p = 0.5$ in the following plot.



- On the x-axis we show the flip number and along the y-axis, we show the cumulative mean, or our average number of heads up to flip x.
- Early on there is a lot of variation in mean scores.
- However, as you move towards 1,000 flips, the runs will converge towards the expected probability of 0.5.
- The more independent random events you have, the less variability there will be around the theoretically expected results.

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:

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

- 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.

Central Limit Theorem (2)

- The previous example about the soccer field was a clear manifestation of the CLT.
- In that case we were summing up various independent random variables coming from a non-normal (Binomial) distribution.
- The CLT also holds for the sum of random variables, because the sum is just $\overline{X_n}$ multiplied by a constant.

Alternative notations showing that Z_n converges to a Normal

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

Why is the Central Limit Theorem Important?

- For experimentalists, the CLT is an extremely important concept.
- For many practical questions, we cannot get measurements for an entire population of interest.
- So we have to select a sample from which to draw conclusions.
- How can we be confident that the conclusions we draw about the sample generalize across the population?
- The central limit theorem allows us to make claims about the distribution of our **sample means**.
- This will be a fundamental idea for statistical inference.

Example: Central Limit Theorem

- Suppose that the number of errors X of a computer program in a week follows a Poisson distribution with parameter $\lambda = 5$.
- If $X \sim \textit{Poisson}(\lambda)$, $\mathbb{E}(X) = \lambda$ and $\mathbb{V}(X) = \lambda$.
- This means that, on average, a computer program makes 5 errors in a week.
- If we have 125 independent programs $X_1, \ldots, X_{125}$ we would like to approximate the probability that the average number of errors for all these programs during a week is less than 5.5: $\mathbb{P}(\overline{X_n} < 5.5)$.

Example: Central Limit Theorem

- Using the CLT we have that:

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

- In R:

```
> n <- 125
> sigma <- sqrt(5)
> mu <-5
> pnorm(5.5,mean = 5,sd =sigma/sqrt(n))
[1] 0.9937903
> # alternatively
> pnorm(2.5)
[1] 0.9937903
```

Conclusions

- We have visited the main concepts of probability, the language uncertainty.
- Random variables, PDFs, PMFs, CDFs, are fundamental building blocks for statistical inference.
- The law of large numbers and the central limit theorem will allow us to make probabilistic statements about the population even if we only work with samples.

References I


McElreath, R. (2020).
Statistical rethinking: A Bayesian course with examples in R and Stan.
CRC press.


Poldrack, R. A. (2019).
Statistical thinking for the 21st century.
https://statsthinking21.org/.


Quirk, C. (2020).
Let's explore statistics.
https://bookdown.org/cquirk/LetsExploreStatistics/.


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