Applied Statistics

Introduction to Probability

Cesar O. Aguilar
SUNY Geneseo

Portions of these notes were created from *Learning statistics with R* by Danielle Navarro, *Learning statistics with jamovi* by David Foxcroft, and *Introduction to Statistical Thinking* by Benjamin Yakir.

Introduction

- To a lot of people, descriptive statistics is all there is to statistics
- It's about calculating averages, collecting all the numbers, drawing pictures, and putting them all in a report somewhere
- In fact, descriptive statistics is one of the smallest parts of statistics and one of the least powerful
- The bigger and more useful part of statistics is that it provides information that lets you make **inferences** about the real-world
- **Inferential statistics** is about making decisions and predictions based on data, mathematical models, and reasoning
- One goal of inferential statistics is to generalize the results of surveys, sampling, or experiments to a larger group of individuals or a broader class of circumstances
- The theory of statistical inference is built on top of **probability theory**

What is probability theory?

- **Probability theory** is the branch of mathematics that deals with the analysis of random events
- For example, all of these questions are things you can answer using probability theory:
  1. What are the chances of a fair coin coming up heads 10 times in a row?
  2. If I roll two six sided dice, how likely is it that I'll roll two sixes?
  3. How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
  4. What are the chances that I'll win the lottery?
- In the questions above, we are given a **model** of a random process and we are interested in the **probability** of a certain event occurring as the random process unfolds in a particular way

What is probability theory?

- In the fair coin example, the underlying random process is flipping a fair coin
- The mathemtical model is that the probability of a head is 0.5 and the probability of a tail is 0.5
- In symbols, we write this as

  $$P(\text{heads}) = 0.5$$

  and

  $$P(\text{tails}) = 0.5$$
- The event we are interested in is the event that the coin comes up heads 10 times in a row

Probability vs Statistics

- In probability theory the model is known but the data are not
- This is in contrast to the situation in statistics where the data is known but the model is not
- In statistics, all we have is the data and it is from the data that we want to learn the truth about the world, that is, we want to learn the model of the random process that generated the data
- Statistical questions tend to look more like these:
  1. If my friend flips a coin 10 times and gets 10 heads, are they playing a trick on me?
  2. If five cards off the top of the deck are all hearts, how likely is it that the deck was shuffled?
  3. If the lottery commissioner's spouse wins the lottery, how likely is it that the lottery was rigged?
- In these questions, all we have is the data and we want to make inferences about the model

Probability vs Statistics

- In the problem: *If my friend flips a coin 10 times and gets 10 heads, are they playing a trick on me?*
- What we want to infer is whether or not we should conclude that what we saw was actually a fair coin being flipped 10 times in a row, or whether we should suspect that our friend is playing a trick on us
- If the coin is fair then the model we should adopt is one that says that the probability of heads is 0.5:

$$P(\text{heads}) = 0.5$$

- If the coin is not fair then the model we should adopt is one that says that the probability of heads is **not** 0.5:

$$P(\text{heads}) \neq 0.5$$

- The statistical inference problem is to figure out which of these probability models is right

What does probability mean?

- Say you want to bet on a sports game between Team A and Team B
- After thinking about it, you decide that there is an 80% probability of Team A winning and so you bet on Team A
- What do you mean by "**there is an 80% probability of Team A winning**"?
- The teams will play only once and the outcome will be either Team A wins or Team B wins

What does probability mean?

- There are three possible ways to think about the statement "**there is an 80% probability of Team A winning**"

    1. They're robot teams, so you can make them play over and over again, and if you did that, Team A would win 8 out of every 10 games on average

    2. For any given game, you would only agree that betting on this game is only "fair" if a $1 bet on Team B gives a $5 payoff (i.e. you get your $1 back plus a $4 reward for being correct), as would a $4 bet on Team A (i.e., your $4 bet plus a $1 reward)

    3. Your subjective "belief" or "confidence" in an Team A victory is four times as strong as my belief in a Team B victory

- These there are different statistical ideologies and depending on which one you subscribe to, you might say that some of those statements are meaningless or irrelevant

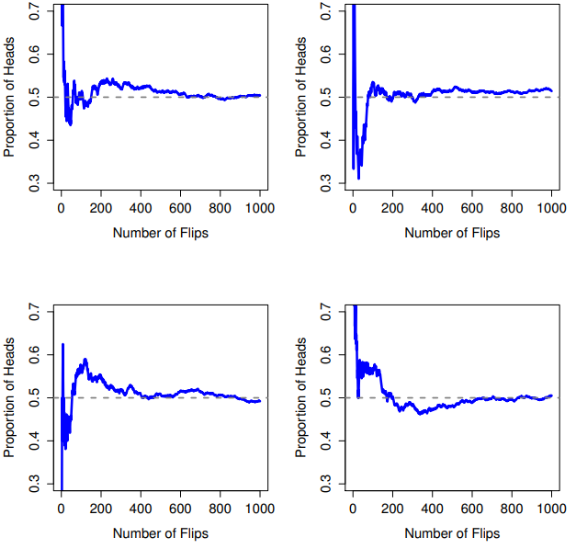
- We'll discuss two of the most common statistical ideologies

The frequentist view

- The first of the two major approaches to probability, and the more dominant one in statistics, is referred to as the **frequentist view**
- Most introductions to statistics are written from the frequentist perspective
- In the frequentist view, probability is defined as a **long-run frequency**
- Thus, the probabibility of an event is the long-run relative frequency with which the event would occur in a large number of independent trials of the event
- To say that the probability of a fair coin coming up heads is 0.5 means that if you flipped a fair coin over and over again, then in the long run, it would come up heads 50% of the time
- Or to say that there is an 80% probability of Team A winning means that if you played the game over and over again under equal conditions, then in the long run, Team A would win 80% of the time

Flipping a coin 1000 times, 4 times

As the number of flips N tends to infinity, $N \to \infty$, the fraction $N_H/N \to 0.5$



The frequentist view

- Two main advantages with the frequentist view is that it is **objective** and **unambiguous**
- The only way that probability statements can make sense is if they refer to (a sequence of) events that occur in the physical universe
- Or, to make a statement about probability, it must be possible to redescribe the statement in terms of a sequence of potentially observable events
- Any two people watching the same sequence of events must inevitably come up with the same probability statement
- The frequentist view has some disadvantages though: infinite sequences of events don't exist in the physical world!
- More seriously, the frequentist definition has a narrow scope: Not all events can be mapped into a hypothetical sequence of events
- "The probability of rain in Geneseo on 2 November 2048 is 60%"

The Bayesian view

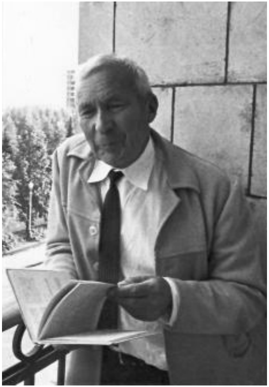
- The other main approach to probability is the **Bayesian view** which is often called the **subjectivist** view
- In this view, the probability of an event is the **degree of belief** that an intelligent and rational agent assigns to the truth of the event
- In this view, probabilities don't exist in the world but rather in the thoughts and assumptions of people and other intelligent beings
- This notion of "subjective probability" can be operationalised in terms of what bets an intelligent agent is willing to accept
- Suppose that I believe that there's a 60% probability of rain tomorrow
- If someone offers me a bet that if it rains tomorrow then I win $5, but if it doesn't rain I lose $5, then from my perspective, this is a pretty good bet
- On the other hand, if I think that the probability of rain is only 40% then it's a bad bet to take

The Bayesian view

- A main advantage with the Bayesian view is that it allows you to assign probabilities to any event you want to
- You don't need to be limited to those events that are repeatable
- An obvious disadvantage with this view (to many people) is that we aren't always purely objective
- Specifying a probability requires us to specify an entity that has the relevant degree of belief
- The Bayesian view allows everyone to have their own beliefs
- Two observers with different background knowledge can legitimately hold different beliefs about the same event
- I can believe the coin is fair and you don't have to, even though we're both rational
- To many people this is uncomfortable, it seems to make probability arbitrary

Basic probability theory

- Whether you subscribe to the frequentist or the Bayesian view, most people agree on the **rules** of probability theory
- The rules of modern probability theory are due to Soviet mathematician Andrey Kolmogorov in 1933



Source: https://commons.wikimedia.org/wiki/File:Andrej_Nikolajewitsch_Kolmogorov.jpg

Basic probability theory

- Suppose that you own 5 pairs of shoes, labelled X_1, X_2, X_3, X_4, and X_5
- Every day you pick one, and only one, pair of shoes to wear, depending on the weather, activities for the day, mood, etc.
- In the language of probability theory, we say that choosing a pair of shoes X is an **elementary event**
- The **set** of all possible elementary events is called the **sample space**:

$$S = \{X_1, X_2, X_3, X_4, X_5\}$$

- Each elementary event has an assigned probability of occuring, which is a number between 0 and 1
- The probability of choosing a pair of shoes X is denoted by $P(X)$
- The higher the probability $P(X)$, the more likely it is that the event X will occur

Basic probability theory

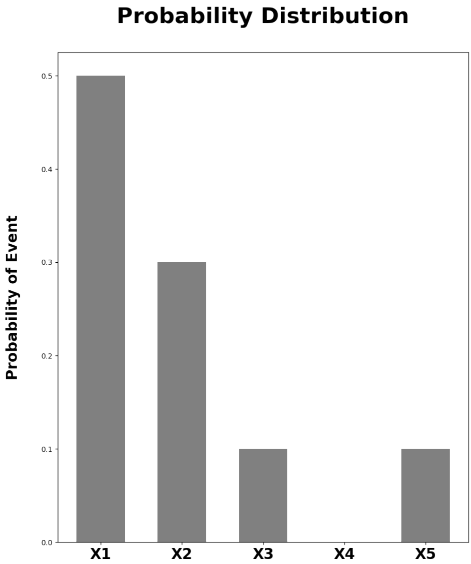
- If $P(X) = 0$ then the event X is **impossible** to occur
- If $P(X) = 1$ then the event X is **certain** to occur
- $P(X) = 0.5$ means that event X occurs half of the time
- Because at least one of the elementary events must occur, the sum of the probabilities of all elementary events must be equal to 1:

$$P(X_1) + P(X_2) + P(X_3) + P(X_4) + P(X_5) = 1$$

- These requirements define a **probability distribution**:

Event	Probability
X_1	$P(X_1) = .5$
X_2	$P(X_2) = .3$
X_3	$P(X_3) = .1$
X_4	$P(X_4) = 0$
X_5	$P(X_5) = .1$

Basic probability theory



Basic probability theory

- Once we have a sample space with elementary events and their probabilities, we can start to talk about **non-elementary events**
- For example, suppose that we want to know the probability of choosing a pair of running shoes among the shoes X_1, X_2, X_3, X_4, and X_5
- Denote by E the event that we choose running shoes
- If X_1 and X_2 are the only two pairs of running shoes, then

$$P(E) = P(X_1) + P(X_2) = 0.5 + 0.3 = 0.8$$

- From these simple rules, it's possible to construct some extremely powerful mathematical tools
- This is the basis of modern probability theory; fun stuff!
- In what follows, we'll discuss some probability distributions that arise frequently in statistics

The binomial distribution

- The **binomial distribution** is a probability distribution that describes the probability of a given number of **successes** in a fixed number of independent experiments or **trials**

- Examples where the binomial distribution is useful:
    - The probability of getting 3 heads in 20 coin tosses
    - The probability that 10 out of 80 lightbulbs in a box are defective
    - The probability that 5 out of 100 people in a room have brown eyes

- In each of these examples, there is a probability of success for each individual experiment or trial, denote the probability of each trial by θ

- Some authors use the notation *p* instead of θ; in any case $0 \leq \theta \leq 1$

- For the coin tosses example, the probability of a success (landing heads) in each individual coin toss is $\theta = 0.5$

- For the lightbulbs example, the probability of a success (a defective lightbulb) in each individual lightbulb might be something like $\theta = 0.01$

The binomial distribution

- In each example, there is a fixed number of trials N
- For the coin tosses example, $N = 20$
- In each example, we are interested in the probability of getting **exactly** k successes in N trials
- In the coin toss example, $k = 3$
- Let's backup and use X to denote the number of successes in N trials
- The possible values of X are $0, 1, 2, \ldots, N$; in the coin toss example we are interested in knowing the probability that $X = 3$
- For the binomial distribution, the probability that $X = k$ is denoted by

$$P(X = k \mid \theta, N)$$

- Because the numerical value of X varies randomly, we say that X is a **random variable**

The binomial distribution

- There is a formula for the probability that $X = k$:

$$P(X = k \mid \theta, N) = \binom{N}{k} \theta^k (1 - \theta)^{N-k}$$

- The notation $\binom{N}{k}$ is a binomial coefficient; it is defined as

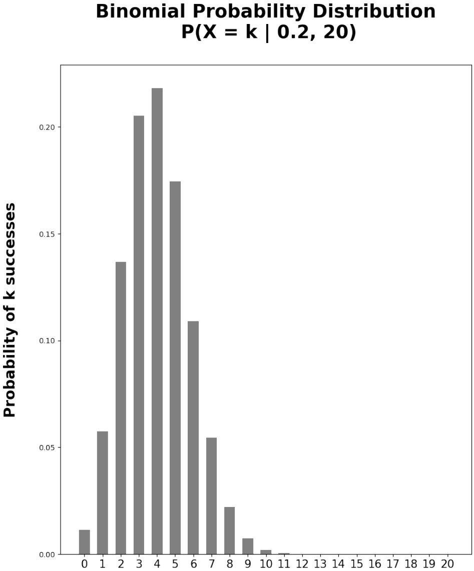
$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

- The probability distribution of X is called the **binomial distribution**, and θ and N are the **parameters** of the distribution
- You'll sometimes see the notation $X \sim \text{Binomial}(N, \theta)$
- The binomial distribution is an example of a **discrete distribution** because X is a discrete random variable
- Don't worry, you'll never have to calculate $P(X = k \mid \theta, N)$ by hand

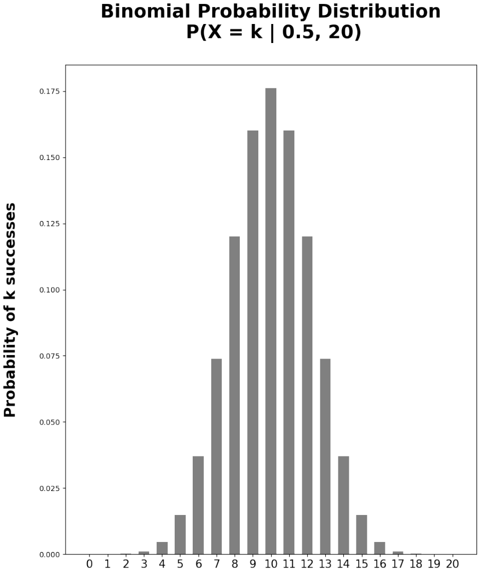
The binomial distribution: $\theta = 0.2$, $N = 20$

k	$P(X = k \mid \theta, N)$
0	0.01152922
1	0.05764608
2	0.13690943
3	0.20536414
4	0.21819940
5	0.17455952
6	0.10909970
7	0.05454985
8	0.02216088
9	0.00738696
10	0.00203141
11	0.00046168
12	0.00008657
13	0.00001332
14	0.00000166
15	0.00000017
16	0.00000001
17	0.00000000
18	0.00000000
19	0.00000000
20	0.00000000

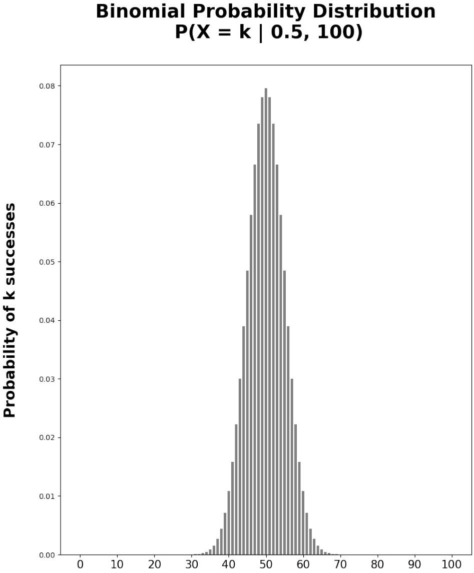
The binomial distribution: $\theta = 0.2,\ N = 20$



The binomial distribution: $\theta = 0.5$, $N = 20$



The binomial distribution: $\theta = 0.5$, $N = 100$



Example: The binomial distribution

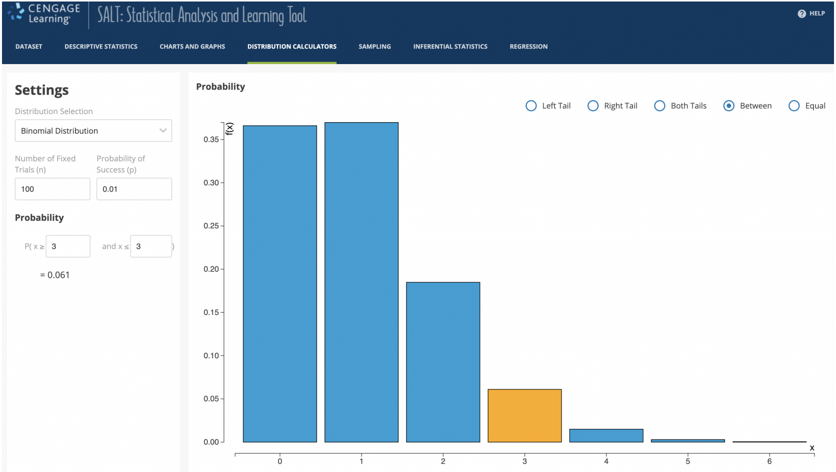
A company manufactures a product which is known to have a defect rate of 1%. Find the probability that out of 100 products:

1. exactly 3 are defective,
2. less than 3 are defective,
3. at least 3 are defective.

- Let X be the number of defective products out of 100. Then $X \sim \text{Binomial}(100, 0.01)$
- For 1., we are interested in $P(X = 3)$
- For 2., we are interested in $P(X < 3)$ which equals
$$P(0 \leq X \leq 2)$$
- For 3., we are interested in $P(X \geq 3)$ or
$$P(3 \leq X \leq 100)$$

Example: The binomial distribution

SALT: Statistical Analysis and Learning Tool



Settings

Distribution Selection

Binomial Distribution

Number of Fixed Trials (n): 100

Probability of Success (p): 0.01

Probability

P(x ≥ 3 and x ≤ 3)

= 0.061

Probability

○ Left Tail ○ Right Tail ○ Both Tails ● Between ○ Equal

Example: The binomial distribution

- For 1., we find that

$$P(X = 3) \approx 0.06$$

- For 2., we find that

$$P(X \leq 2) \approx 0.92$$

- For 3., we find that

$$P(X \geq 3) \approx 0.08$$

- Notice that because $X \leq 2$ and $X \geq 3$ are mutually exclusive events:

$$P(X \leq 2) + P(X \geq 3) = 1$$

- Therefore,

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.92 = 0.08$$

Example: The binomial distribution

In a community college training course, it is known that 10% of enrolled students will drop out before the end of the course. In a class of 40 students, what is the probability that that no more than 3 students will drop out?

- Let X be the number of students who drop out. Then $X \sim \text{Binomial}(40, 0.1)$
- We are interested in $P(X \leq 3)$
- In SALT, we can use the binomial distribution to calculate this probability and find that

$$P(X \leq 3) \approx 0.42$$

The normal distribution

- The **normal distribution** is the most important probability distribution in probability theory and statistics
- Many random variables in the real world are well approximated with a normal distribution such as SAT scores, shoe sizes, people's heights, IQ scores, etc.
- It is also called the "the bell curve" or the **Gaussian distribution** after German mathematician Carl Friedrich Gauss:

Source: https://commons.wikimedia.org/wiki/File:Carl_Friedrich_Gauss_1840_by_Jensen.jpg

The normal distribution

- Like with the binomial distribution, we are interested in the probability that a certain random variable X takes on a particular value x
- The normal distribution is used only for variables that are **continuous**, unlike the binomial distribution where the variable is **discrete**
- For example, if X is the height of a person, then X is a continuous variable and the normal distribution *might* be an appropriate model
- However, for **continuous** random variable X, computing the probability that X takes on a exact value is not useful
- This is in contrast to computing $P(X = x)$ for the binomial distribution
- **What is the problem?**: If X is a **continuous** random variable then

  $$P(X = x) = 0$$

  no matter what the x is
- Let's attempt to show why...

The normal distribution

- Suppose that X is a continous random variable and can take on any value between 0 and 10
- In a very small range of values, the probability that X equals any given value is the same regardless of the value
- This is similar to saying that the probability of a dice roll is the same regardless of the number (1/6 in this case)
- Let's say then that $P(X = x) = a$ for all x in a very small range of values
- Consider the following specific values that X can equal:

$$5,\ 5.1,\ 5.11,\ 5.111, \ldots$$

- The probability that X equals any one of these values is

$$P(\{X = 5\} \text{ or } \{X = 5.1\} \text{ or } \{X = 5.11\} \cdots) = a + a + a + \cdots$$

The normal distribution

- Now because the probability that X takes on some value between 0 and 10 is 1, we must have

$$a + a + a + a + a + \cdots <= 1$$

- Factor out the a:

$$a + a + a + a + a + \cdots = a \cdot (1 + 1 + 1 + 1 + 1 + \cdots)$$

- But now we have a problem because the following sum is unbounded:

$$1 + 1 + 1 + 1 + \cdots$$

- Thus if a is not zero then

$$a + a + a + a + \cdots > 1$$

  and this is a contradiction! **So we must have $a = 0$.**

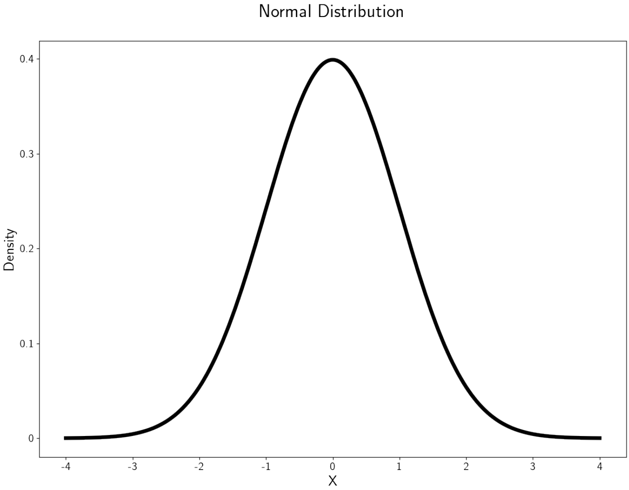
The normal distribution

- How then do we describe the normal distribution?
- We do this using what is called a **probability density function** (pdf)
- If X is a continous random variable, we use its pdf to compute the probability that X takes on a value in a chosen **range**
- We never talk about X taking on a specific value, only a range of values
- The normal distribution is described by two parameters:
  - the mean μ, and
  - the standard deviation σ
- The pdf for the normal distribution is given by:

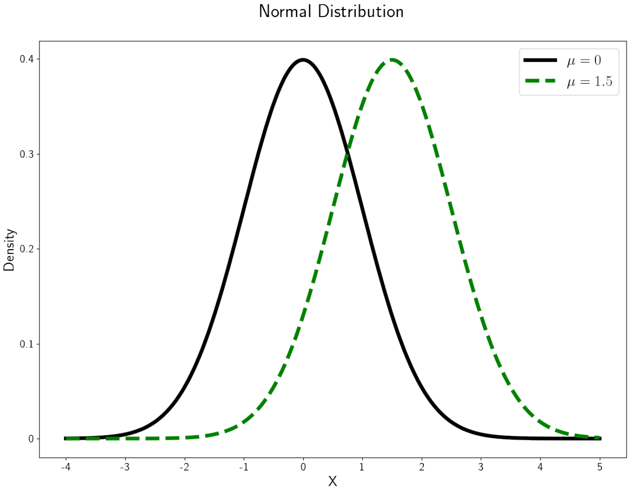
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- We also write that $X \sim N(\mu, \sigma)$

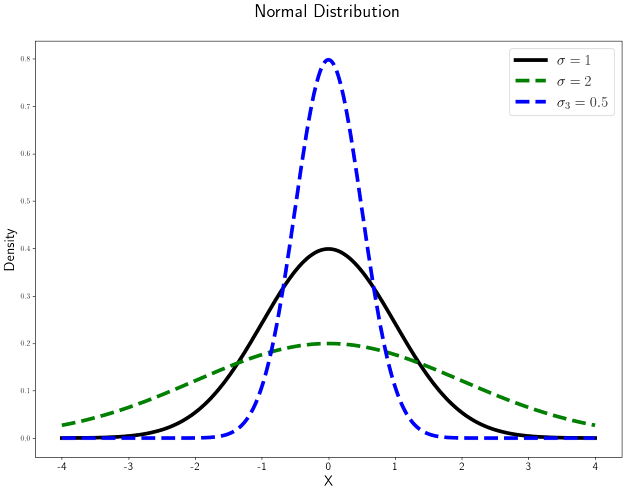
The normal distribution: $\mu = 0$, $\sigma = 1$



The normal distribution: $\mu = 0$ and $\mu = 1.5$, $\sigma = 1$



The normal distribution: $\mu = 0$, $\sigma = 1$ and $\sigma = 2$



The normal distribution

- How do we use the pdf to compute probabilities?
- To compute the probability that $a \leq X \leq b$:

$$P(a \leq X \leq b) = \int_a^b p(x)\, dx$$

- If you've taken calculus, you should recognize this as the area under $p(x)$ between a and b
- To compute the probability that $X \leq b$:

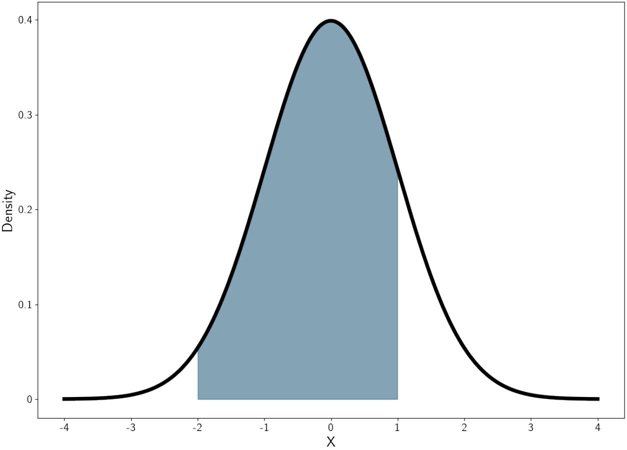
$$P(X \leq b) = \int_{-\infty}^b p(x)\, dx$$

- To compute the probability that $a \leq X$:

$$P(a \leq X) = \int_a^\infty p(x)\, dx$$

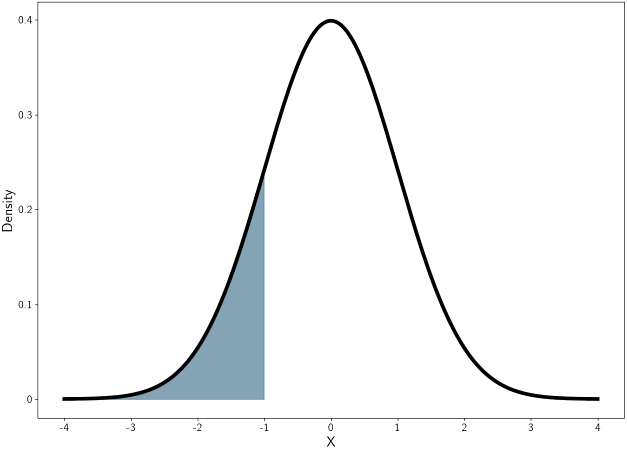
The normal distribution: $\mu = 0$, $\sigma = 1$

$$P(-2 \leq X \leq 1) = 0.8186$$



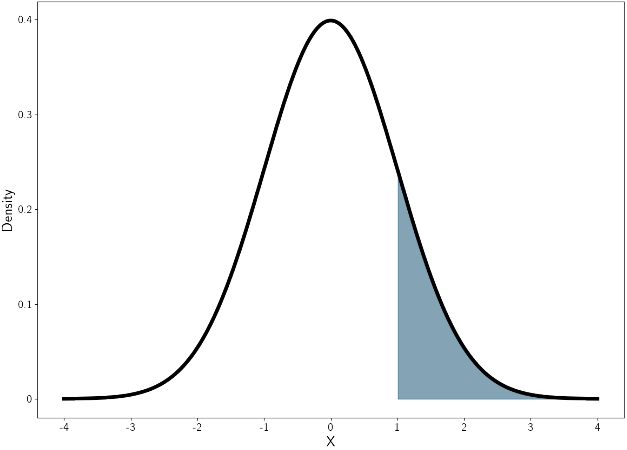
The normal distribution: $\mu = 0$, $\sigma = 1$

$$P(X \leq -1) = 0.1587$$



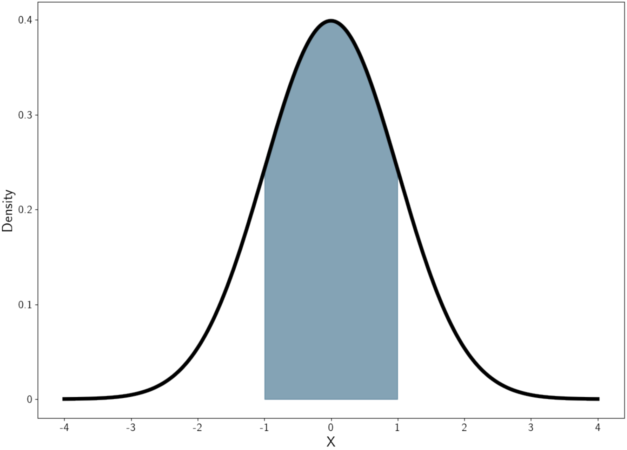
The normal distribution: $\mu = 0$, $\sigma = 1$

$$P(1 \leq X) = 0.1587$$



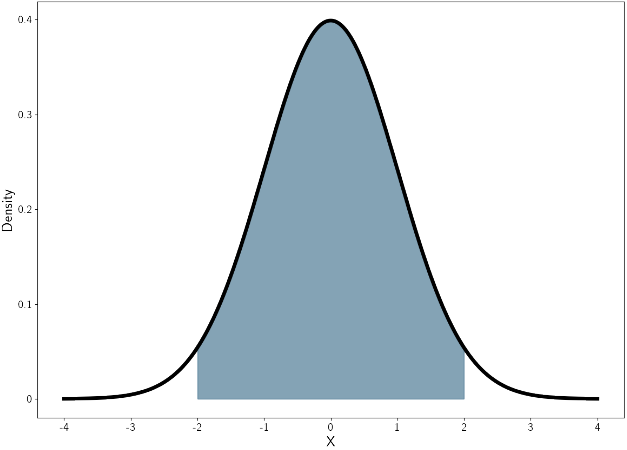
The normal distribution: $\mu = 0$, $\sigma = 1$

$$P(-\sigma \leq X \leq \sigma) = 0.6827$$



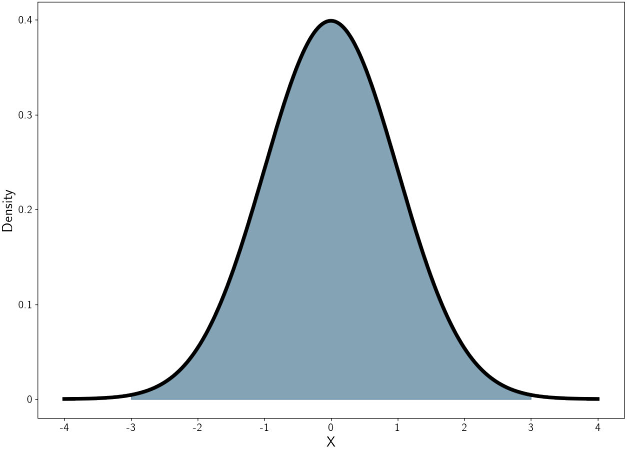
The normal distribution: $\mu = 0$, $\sigma = 1$

$$P(-2\sigma \leq X \leq 2\sigma) = 0.9545$$



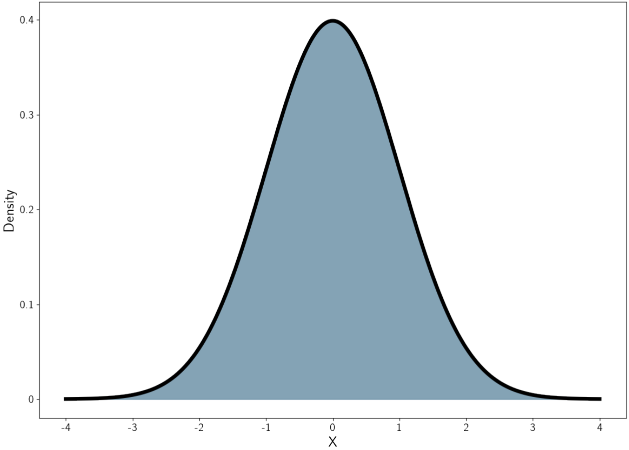
The normal distribution: $\mu = 0$, $\sigma = 1$

$$P(-3\sigma \leq X \leq 3\sigma) = 0.9973$$



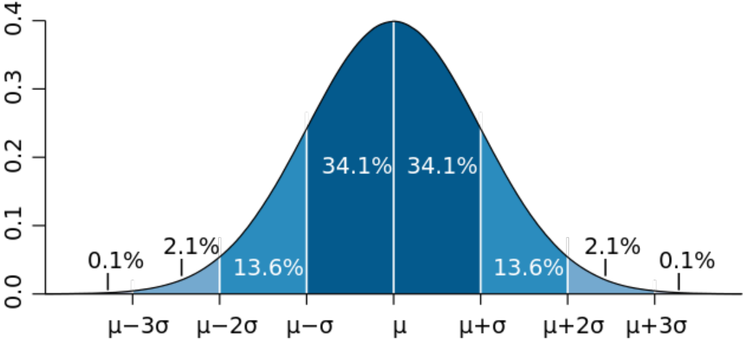
The normal distribution: $\mu = 0$, $\sigma = 1$

$$P(-\infty < X < \infty) = 1$$



The normal distribution

- For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%



Source: https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram_micro.svg

Example: The normal distribution

Suppose that the IQs of students at a college are normally distributed with $\mu = 100$ and $\sigma = 10$. What is the probability that a randomly selected student has an IQ

1. between 100 and 125
2. more than 140

- For 1. we have

$$P(100 \leq X \leq 125) \approx 0.49$$

- For 2. we have

$$P(X \geq 140) \approx 0.00003$$

Example: The normal distribution

A manufacturer produces lightbulbs whose lifetimes are normally distributed with $\mu = 1000$ hours and $\sigma = 300$ hours. What is the probability that a randomly selected lightbulb will last:

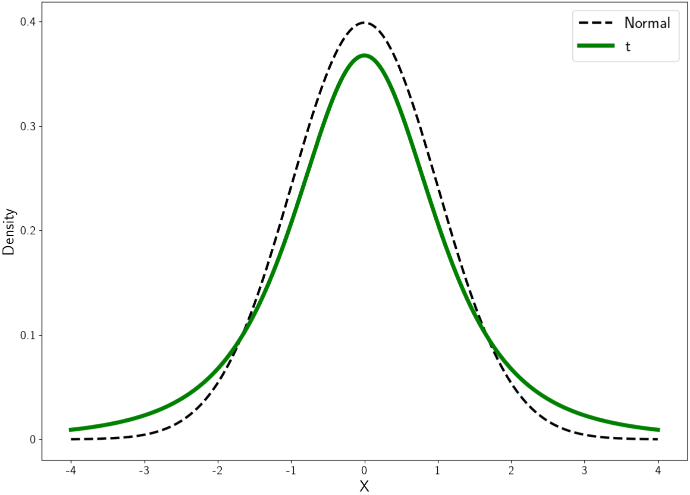
1. more than 1500 hours
2. less than 800 hours

- For 1. we have
$$P(X \geq 1500) \approx 0.047$$

- For 2. we have
$$P(X \leq 800) \approx 0.252$$

t-distribution

- The t-**distribution** is a continuous distribution that looks very similar to a normal distribution



t-distribution

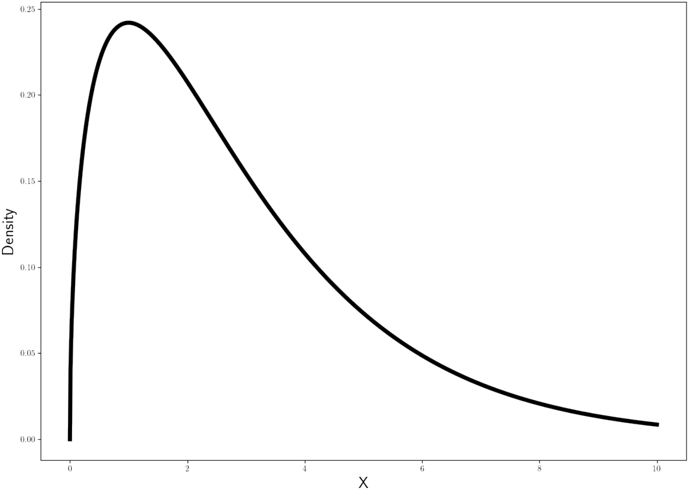
- The t-distribution tends to arise in situations where you think that the data actually follow a normal distribution, but you don't know the mean or standard deviation

- Specific problems where the t-distribution arises are:
  - when assessing the statistical significance of the difference between two sample means,
  - the construction of confidence intervals for the difference between two population means, and
  - in linear regression analysis

- The "tails" of the t-distribution extend further outwards than the tails of the normal distribution

χ^2-distribution

- chi-squared distribuition



χ^2-distribution

- The χ^2-distribution turns up in lots of different places, one being in categorical data analysis

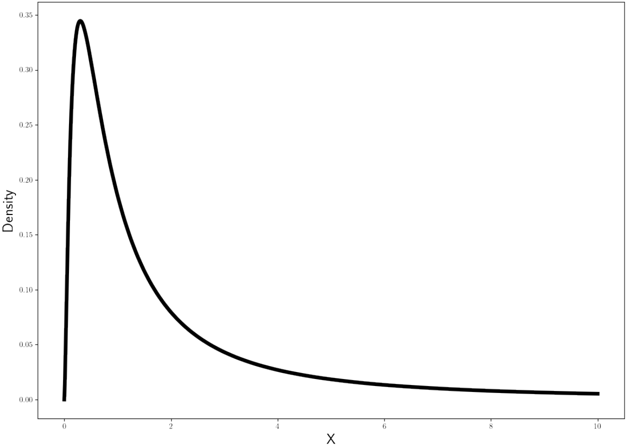
- If X_1, X_2, …, X_k are independent random variables with normal distribution then the sum of the squares of the X_i's has a χ^2 distribution:

$$Z = X_1^2 + X_2^2 + \cdots + X_k^2$$

- We'll see later that this fact turns out to be useful

F-distribution

- The F-distribution looks like the χ^2-distribution
- The F-distribution arises whenever you need to compare two χ^2-distributions to one another


Density
X