INTRO TO PROBABILITY & STATISTICS

Asst. Prof. Gwendolyn Eadie

David A. Dunlap Dept. of Astronomy & Astrophysics / Dept. of Statistical Sciences

University of Toronto, Toronto, Ontario, Canada

Nov. 29, 2021

WHAT IS STATISTICS?

Statistics

Study of uncertainty, estimating and summarizing properties about data, learning from data, analysing data, collecting data …

It can be used to infer facts, make predictions, and make recommendations, based on data.

Nearly all fields that collect data use statistics: astronomy, biology, chemisty, environmental science, forestry, geophysics, law, politics, sports analytics, etc.

What's the recurring theme?

Data!

Broad categories of data

Quantitative

- Numerical
- e.g., height, age, time since an event, blood pressure, brightness of a star, etc.

Categorical

- Can be grouped into a category, type, or quality
- e.g., letter grade, month of birthday, type of galaxy, type of treatment, etc.

Ordinal

- Have a natural order
- Differences between two values may not be meaningful

How we visualize and summarize data depends on the type of data we have

Terminology: a "population" versus a "sample"

Population

- The true, underlying distribution for some quantity
- E.g., the distribution of heights of people all over the world

Sample

- A sample drawn from some distribution
- E.g., randomly select 100 people from around the world and measure their heights
- Will not be exactly like the population because of randomness

Statistics can be used to try to understand the underlying population, when all you have is a sample

A DISTRIBUTION…

Tells you the frequency or relative frequency of each possible value/event, or of some data that was collected

Could be empirical or analytic

Can be useful for modelling a population of objects

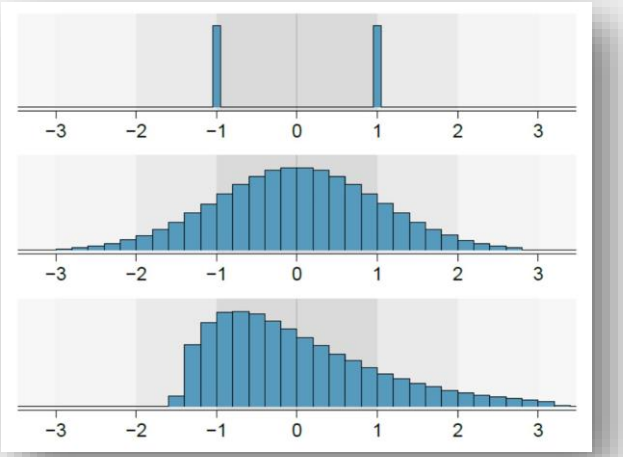
Is often a foundation of statistical reasoning

Can be continuous or discrete

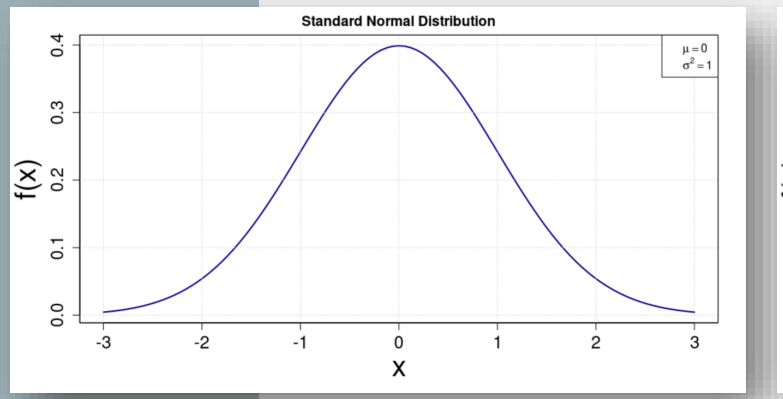
That is analytic has parameters that define its shape

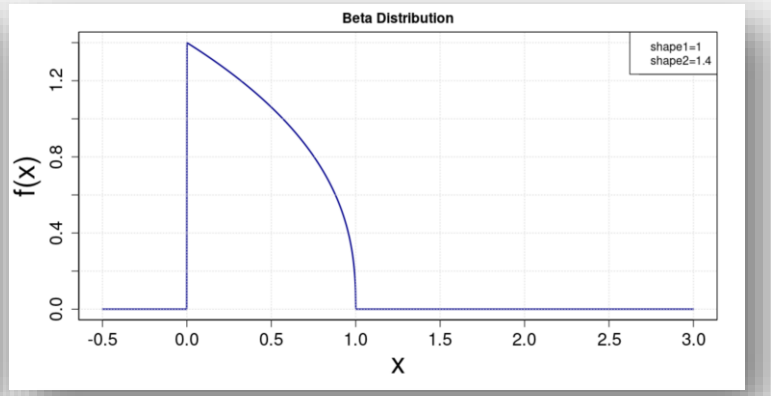
Can be univariate or multivariate

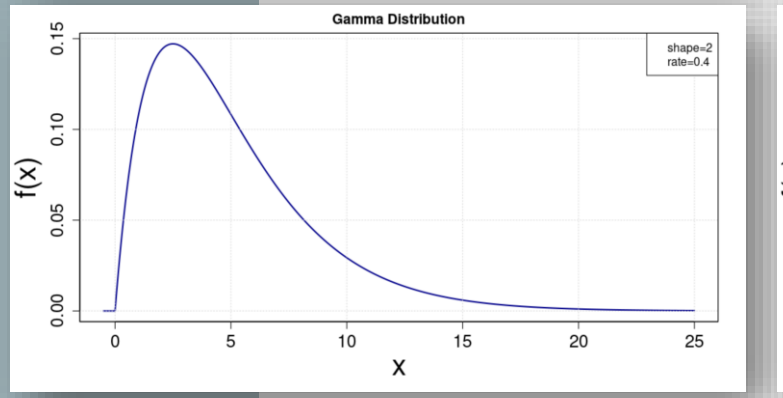
Example histograms (figure from Open Intro Statistics 4th ed.)

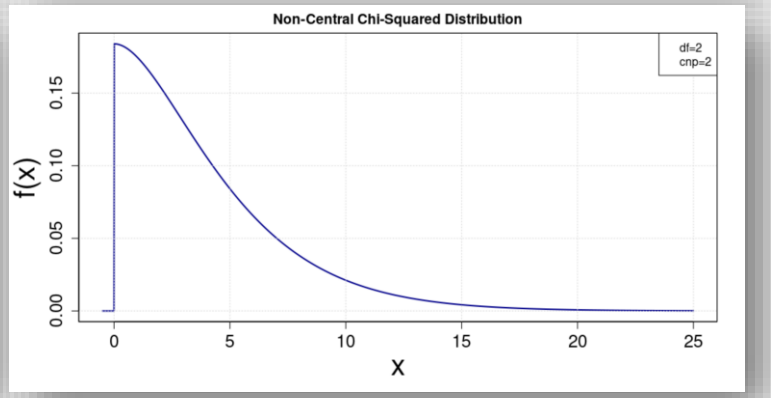


Some analytic probability distributions (plotted by me)









Before we dive into distributions…

We first need the basics of ***probability*** and an understanding of a statistical concept called ***random variables***…

… and to understand these two things, we need to understand ***sample spaces*** and ***events***

INTRODUCTION TO SAMPLE SPACE AND EVENTS

SAMPLE SPACE & VENN DIAGRAMS

- Can be useful for checking your understanding and for visualization of problems

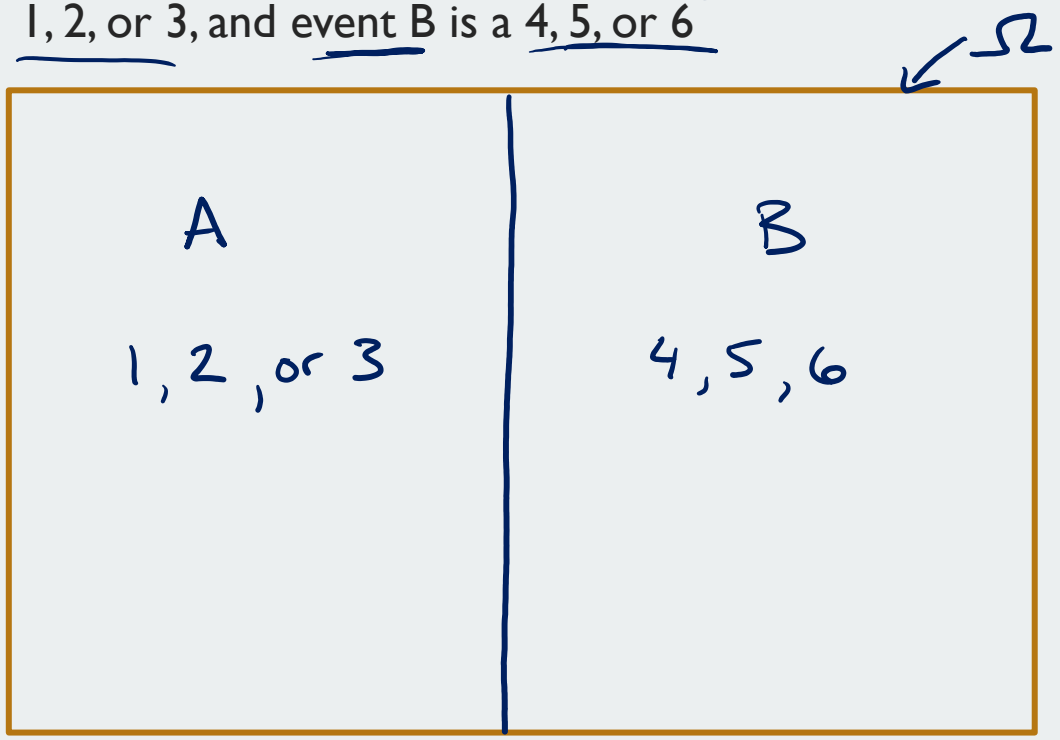
Sample Space

Contains all the possible outcomes or events

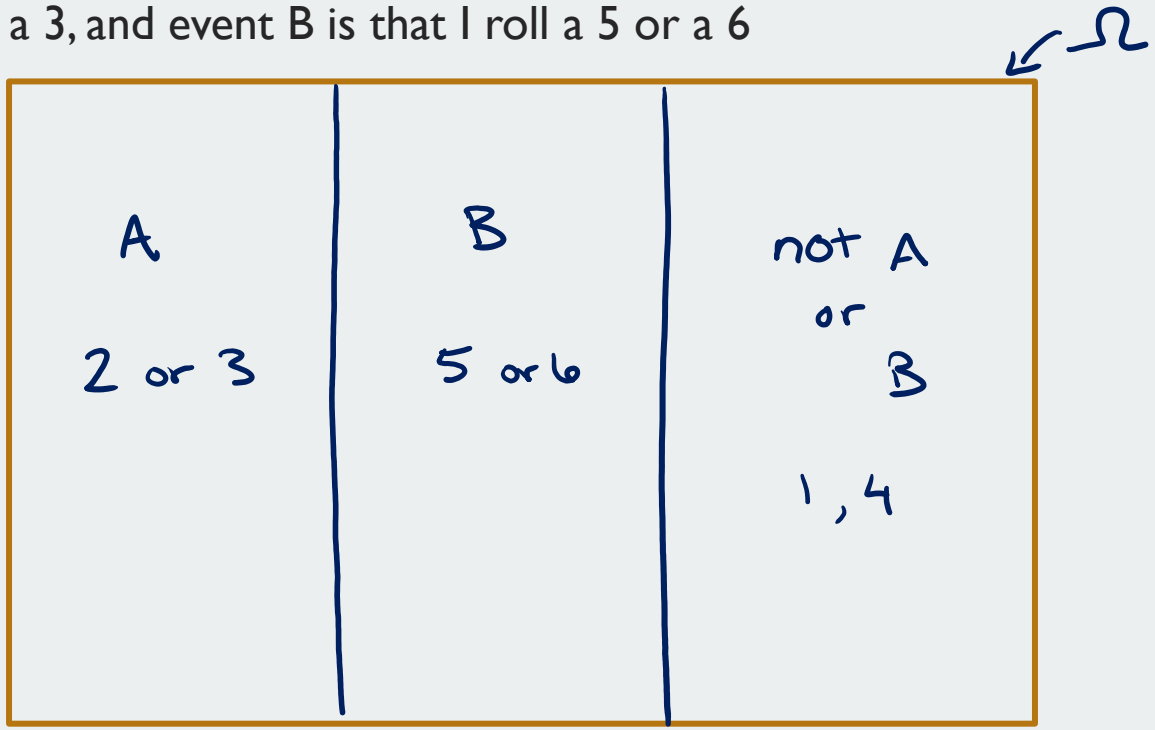
A: 2 or 3

B: 5 or 6

Example 1: Gwen rolls a die. Event A is that I roll a 1, 2, or 3, and event B is a 4, 5, or 6



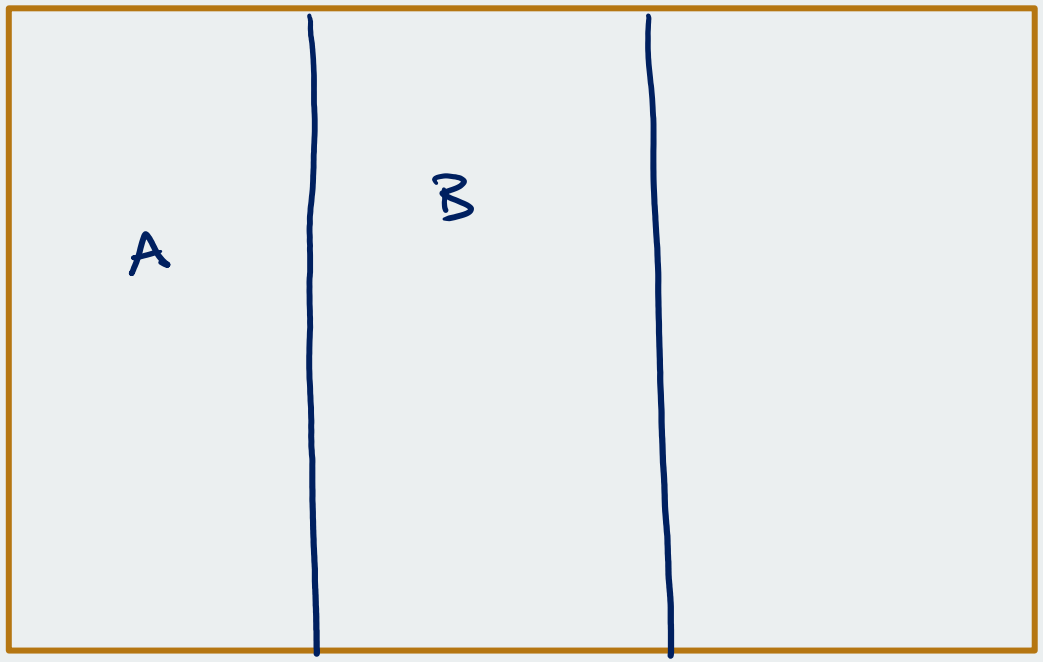
Example 2: Gwen rolls a die. Event A is that I roll a 2 or a 3, and event B is that I roll a 5 or a 6



Events

Disjoint/Mutually Exclusive

- Two events are disjoint/mutually exclusive if they have no outcomes in common



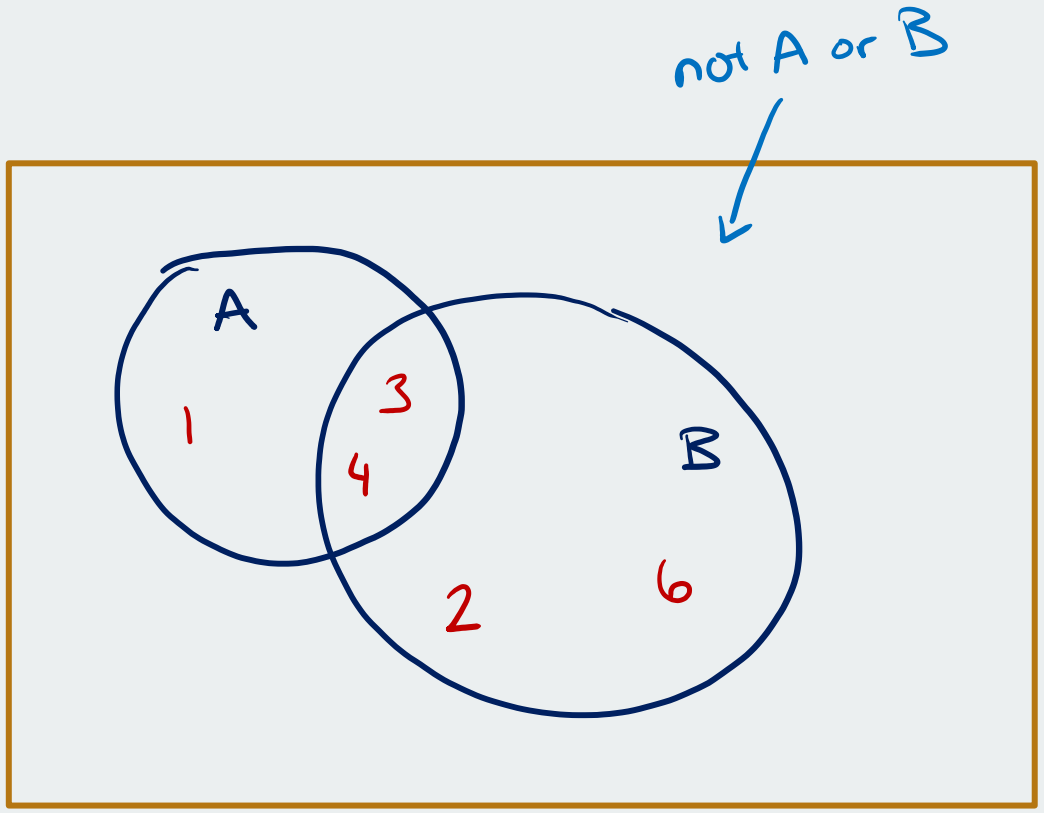
eg. A: 2, 3
B: 5, 6

A and B are disjoint because they don't share any outcomes/events

Events

Not disjoint or not mutually exclusive

- Some outcomes are common between events



eg. A: 1, 3, or 4

B: 2, 3, 4, or 6

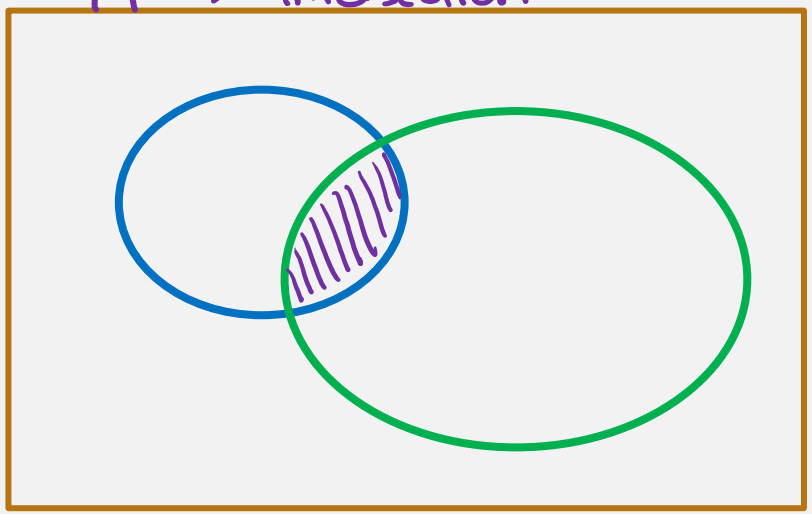
A and B are not disjoint

Sample Space

Contains all the possible outcomes or events

- Let's go through some more examples of events, and how we can visualize them in these types of diagrams

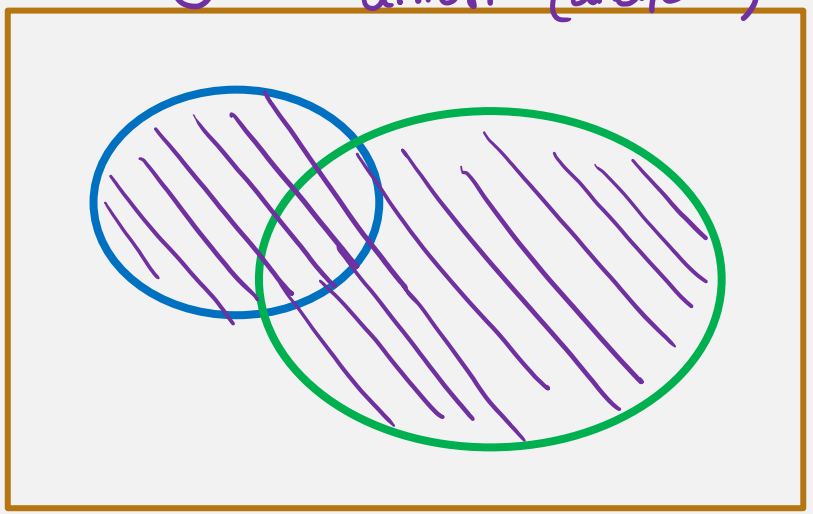
∩ → "intersection"



A *and* B

$A \cap B$

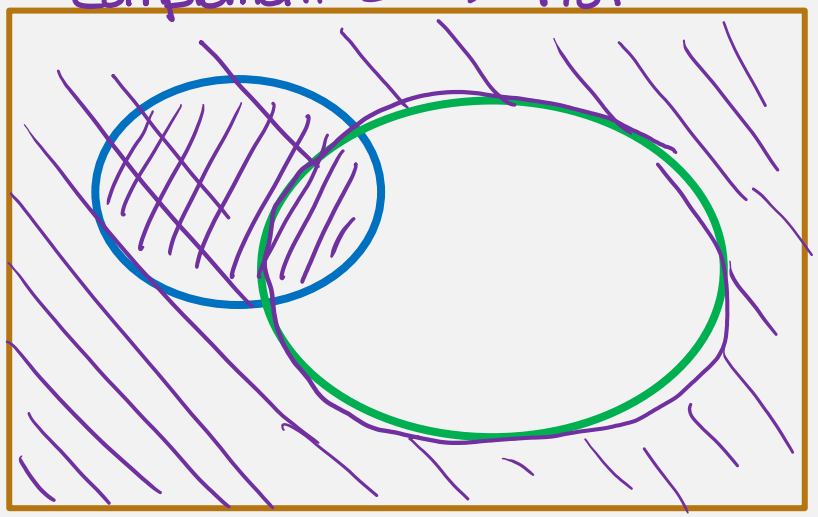
∪ → "union" ("and/or")



A *or* B

$A \cup B$

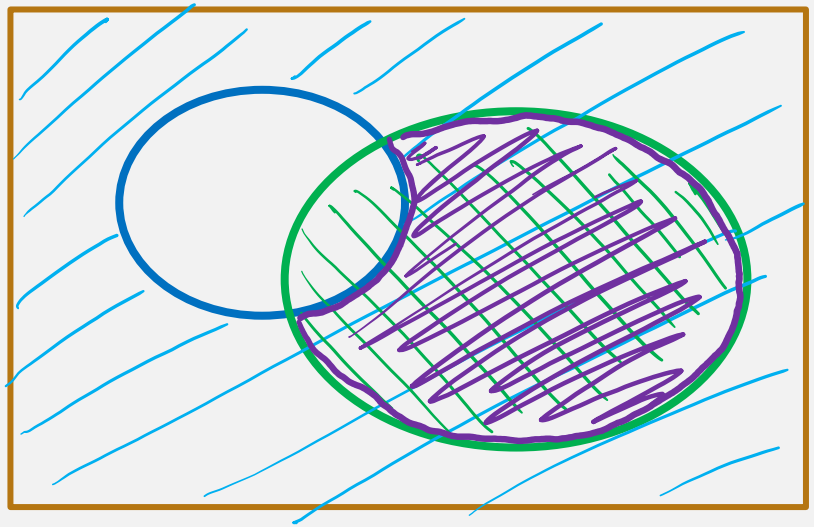
Complement ^{c} → "not"



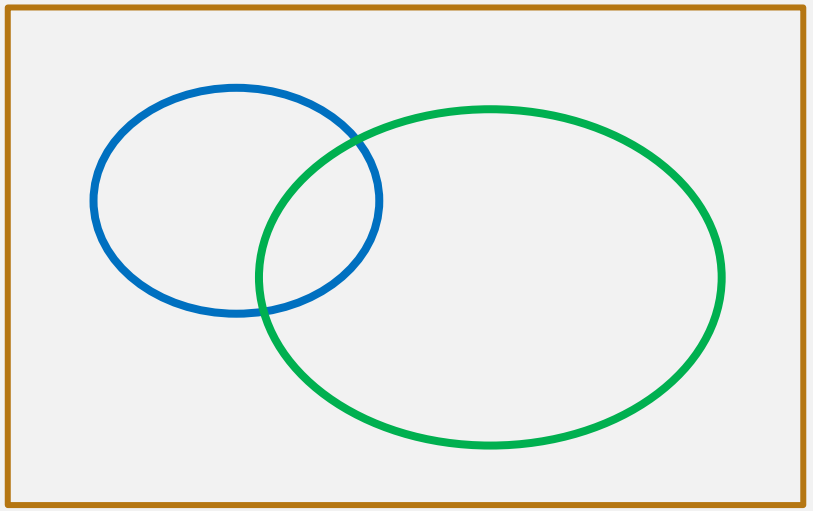
A or B complement

$A \cup B^c$

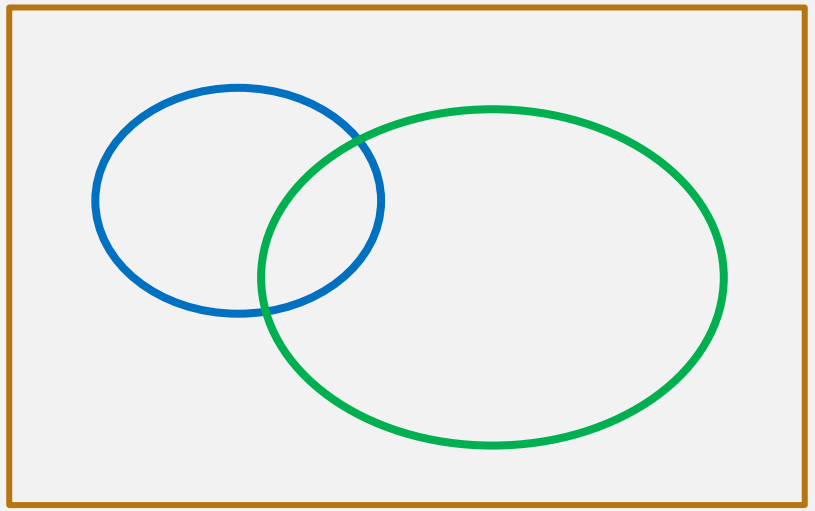
Try these last three on your own! ☺



A complement and B $A^c \cap B$



Complement of (A and B) $(A \cap B)^c$



A complement or B $A^c \cup B$

RANDOM VARIABLES

Random Variables

Assigns a *numerical* value to an outcome/event from an experiment

- Notation:

$$X, Y, W, \ldots etc.$$

- Examples of random variables:
  - Birth weight of a baby
  - Height of a person → continuous random variable
  - How long you wait for the bus
  - The winning score for a basketball game → discrete random variable

Random Variables

Can be *continuous* or *discrete*

- What are some examples of *continuous random variables* in astronomy?
  - flux density of a star
  - mass of an exoplanet
  - age of a cluster
- What are some examples of *discrete random variables* in astronomy?
  - number of exoplanets in a system
  - number of stars in galaxy
  - number of photon counts

CHECK-IN QUESTION

- The number of photons that hit a CCD in an hour is an example of a continuous random variable. True or False?

A. TRUE

B. FALSE

A (data) *sample* is a realization of a random variable

- Imagine we measured the brightnesses of 25 randomly selected stars from the sky

  →These data are realizations x of a random variable X that represents the brightness of a star

- To perform (parametric) *statistical inference* on our data x, we assume how the random variable X is distributed

- Once we have a statistical model for how X is distributed, then we can say things like *"the probability of observing a star with magnitude less than 17 is …"*

PROBABILITY

PROBABILITY

- Formally, probability is given by a number between 0 and 1
- Probabilities of all events must add to 1
- Intuitively, we tend to think about probability by counting
  - For example, rolling a six-sided die can result in 6 different outcomes. If the die is fair, then rolling e.g., a two, is $1/6$, and e.g., a two or a three is $2/6$
- In our Sample Space diagrams, we were sketching the possible ***events***.
- The relative areas of the Sample Space are supposed to represent the ***probabilities***
  - *NOTE: You can think of the total sample space area as adding to* 1

Probability and Notation

Description	Notation	Notes
• Probability of Ω	$P(\Omega) = 1$	probability of all events is 1
• probability of A	$P(A)$	" " event A happening regardless of anything else
• probability of A *and* B	$P(A \cap B)$, $P(A, B)$	joint probability of A and B
• probability of A *or* B	$P(A \cup B)$	prob. of A and/or B happening
• probability of A *given* B	$P(A \mid B)$ ↑ "given"	conditional probability — prob. of A given that B happened
• probability of *not* A	$P(A^c) = 1 - P(A)$	prob. of A not happening

Conditional Probability

Description

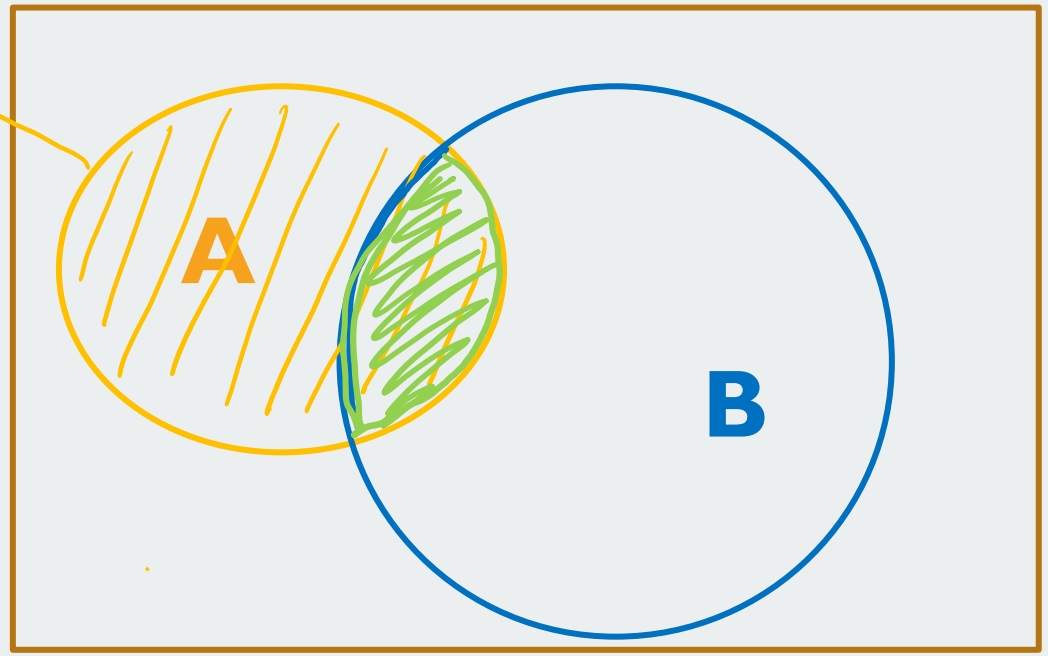
- The probability of an event, given some other event
- $P(B|A) = \frac{P(A,B)}{P(A)}$

→ $P(A,B) = P(B|A)P(A)$

P(B|A) → ratio of green area to the yellow area

- In our sample space diagrams, a conditional probability is the *relative* area

- The *condition* restricts the space we are referring to.
- It refers to a *subset* of the sample space.



Basic Rules of Probability

Addition Rules

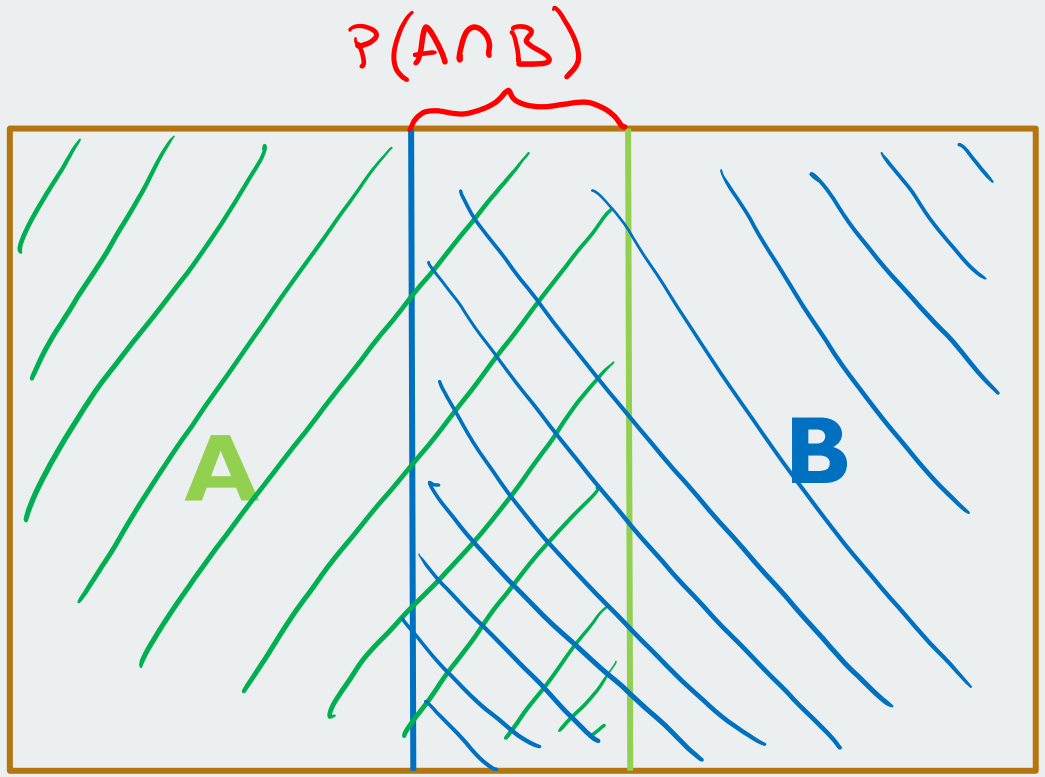
- The probability of all possible events in the sample space must add up to 1
  → $P(\Omega) = 1$
- Probability of events A and B:

  → $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

  $P(A \cap B)$

  area of $A \cup B$ = [area of A] + [area of B] − [area of A and B]

- Probability of the complement of an event A is
  → $P(A^c) = 1 - P(A)$



Basic Rules of Probability

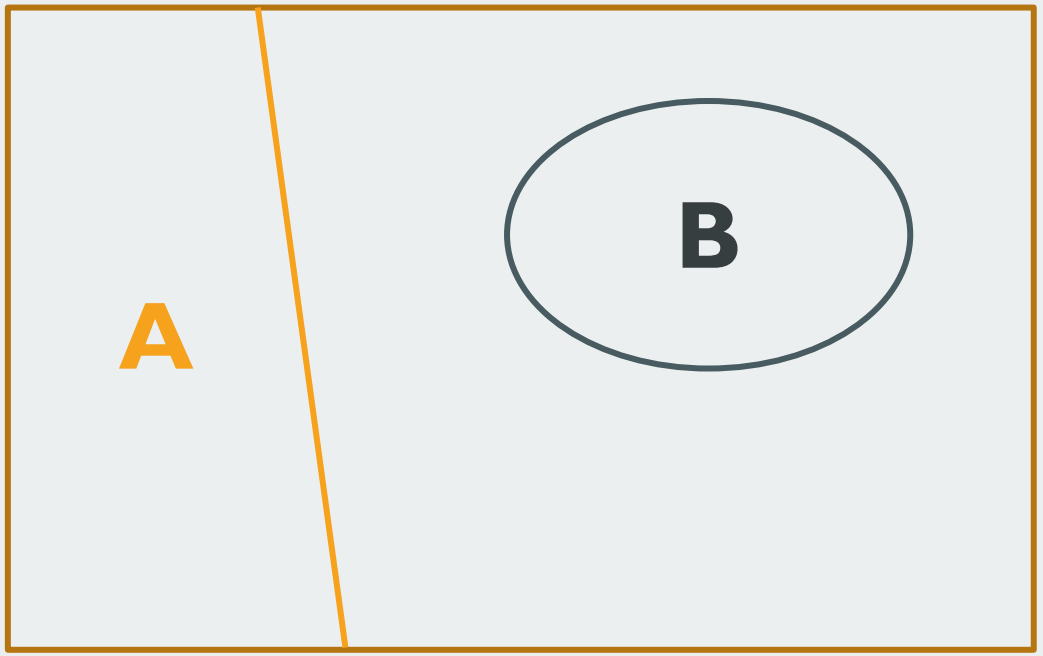
Disjoint or Mutually Exclusive

- A and B do not share any outcomes
- In a sample space diagram, the events do not overlap
- We say that A and B are *disjoint* or *mutually exclusive*

- If P(A) = 0.3 and P(B) = 0.25, and A and B are disjoint, then how does the addition rule simplify?

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.3 + 0.25 - 0$$
$$= 0.55$$

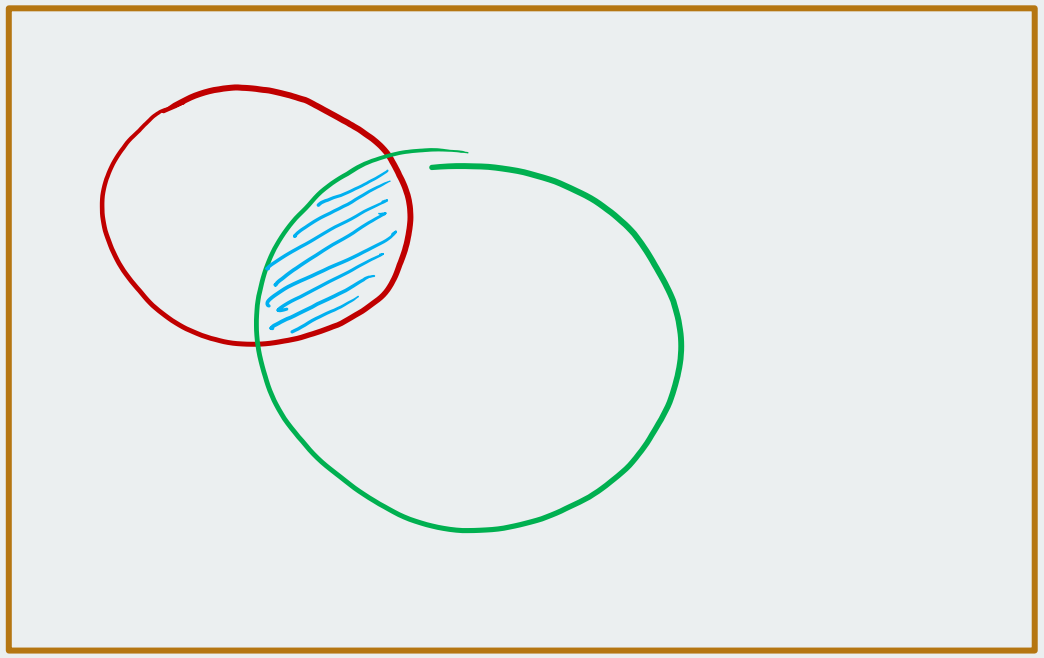
A and B don't span the sample space



Basic Rules of Probability

Independence

- Events A and B are independent if $P(A, B) = P(A)P(B)$
- Also, if A and B are independent, then $P(A|B) = P(A)$, and $P(B|A) = P(B)$



$P(A) = 0.25$

$P(B) = 0.30$

$P(A \cap B) = P(A)P(B)$

$= (0.25)(0.30)$

$= 0.075$

CHECK-IN QUESTION

- TRUE or FALSE. If events A_1 and A_2 are independent, then they are mutually exclusive.
  - True
  - False

Why?

If events A_1 and A_2 are ~~independent~~ disjoint, then they don't share any outcomes or events in common

→ $P(A_1 \cap A_2) = 0$

But, if A_1 and A_2 are independent, then by definition…

→ $P(A_1, A_2) = P(A_1)P(A_2) \neq 0$

contradictory

Law of Total Probability

If $A_1, A_2, \ldots, A_k$ are a partition of Ω, then for any event B:

- $P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots$

→ In other words, add up all the ways B can happen

BAYES' THEOREM

BAYES' THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- A theorem relating conditional probabilities
- Use when you want to invert a conditional probability
- In data analysis: useful for inferring model parameters from data (more on this tomorrow!)

Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Conditional Rules:

$P(A|B)\,P(B) = P(A,B)$ $\longleftrightarrow$ equal $\longleftrightarrow$ $P(B,A) = P(B|A)\,P(A)$

$P(A|B)P(B) = P(B|A)P(A)$

$\rightarrow P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

normalization

How else can we write p(B)?

- P(B) is the probability that event B happens
  - Need to think about all the possible ways B could have happened!

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

CHECK-IN QUESTION

In a galaxy, 60% of the stellar systems have an Earth-like planet, and all these systems *also* have a Jupiter-like planet. *However*, only half of the systems without an Earth-like planet have a Jupiter-like planet.

What's the probability that a stellar system with a Jupiter-like planet also has an Earth-like planet?

Mathematically, we're being asked to find P(Earth-like|Jupiter-like). We can write this out using Bayes' theorem:

$$P(E|J) = \frac{P(J|E)P(E)}{P(J)} = \frac{(1)(0.6)}{0.8} = 0.75$$

We are told that:

$P(J|E) = 1$

$P(E) = 0.6$

$P(E^c) = 1 - P(E)$

$P(J) = P(J|E)P(E) + P(J|E^c)P(E^c) = (1)(0.6) + (0.5)(0.4) = 0.8$

BAYES' THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- A theorem relating conditional probabilities
- Use when you want to invert a conditional probability
- In data analysis: useful for inferring model parameters from data (more on this tomorrow!) → first need to know more about probability *distributions*

DISTRIBUTIONS

A DISTRIBUTION…

Tells you the frequency or relative frequency of each possible value/event, or of some data that was collected

Could be empirical or analytic

Can be useful for modelling a population of objects

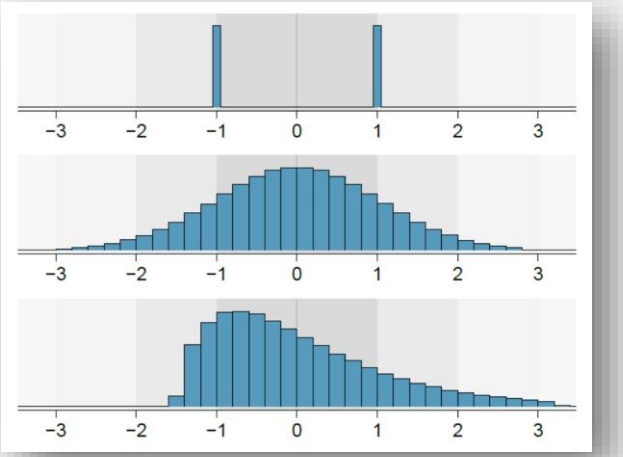
Is often a foundation of statistical reasoning

Can be continuous or discrete

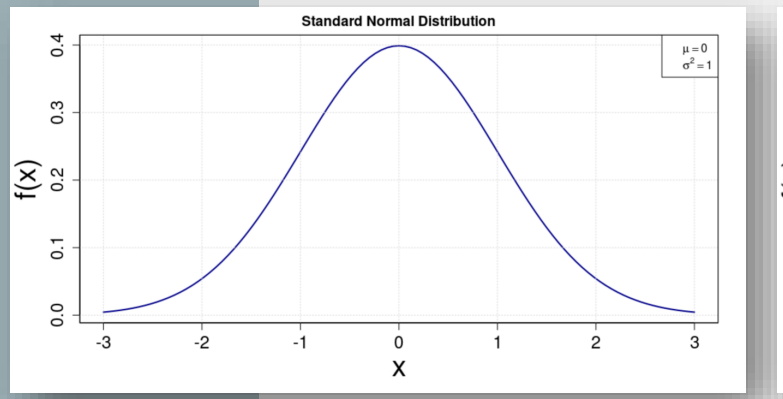
That is analytic has parameters that define its shape

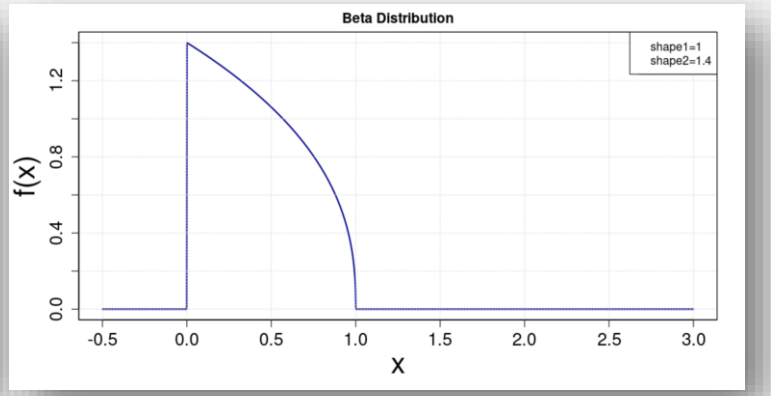
Can be univariate or multivariate

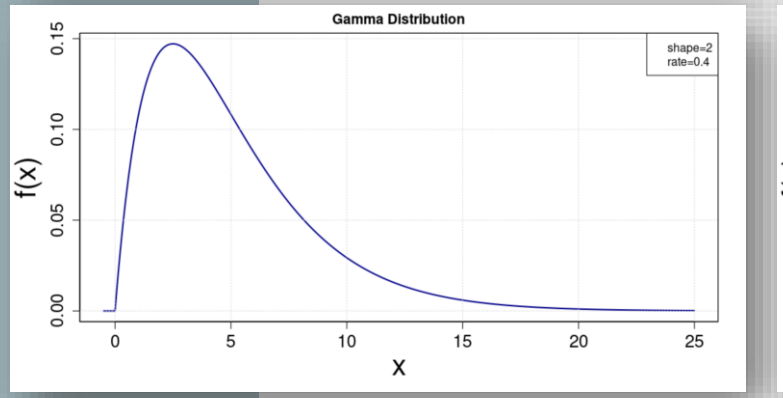
Example histograms (figure from Open Intro Statistics 4th ed.)

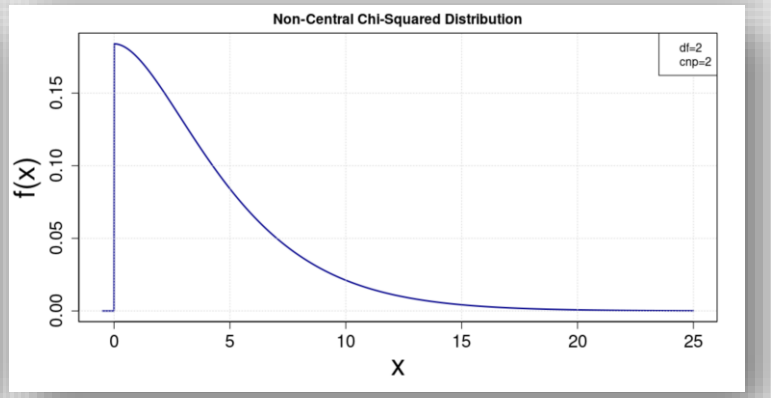


Some analytic probability distributions (plotted by me)





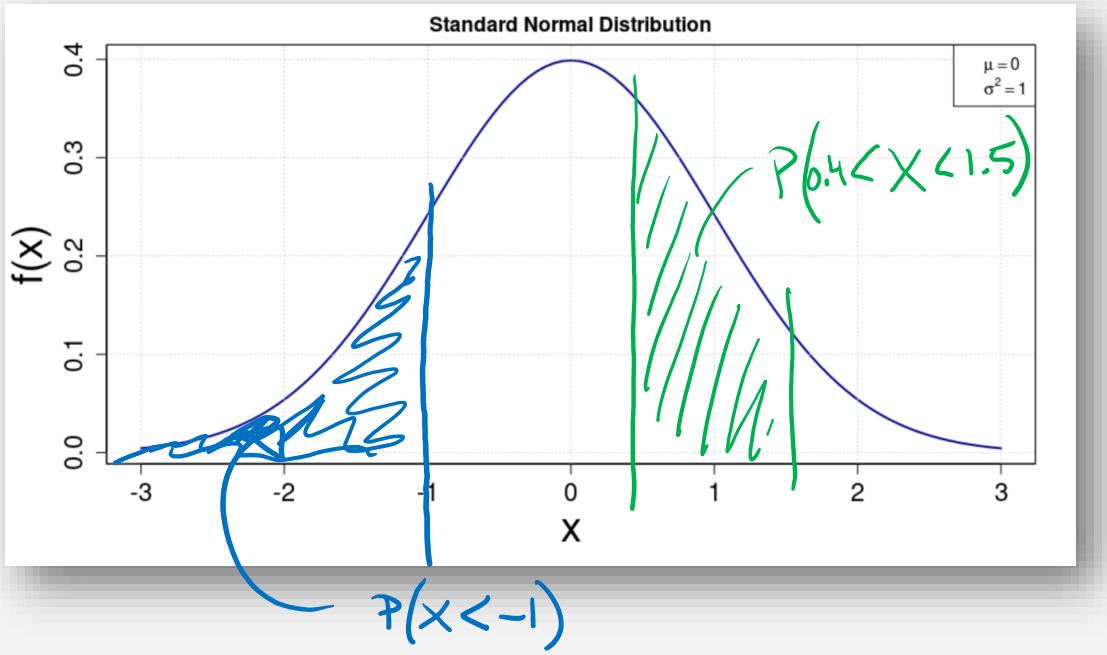




PROBABILITY DISTRIBUTIONS

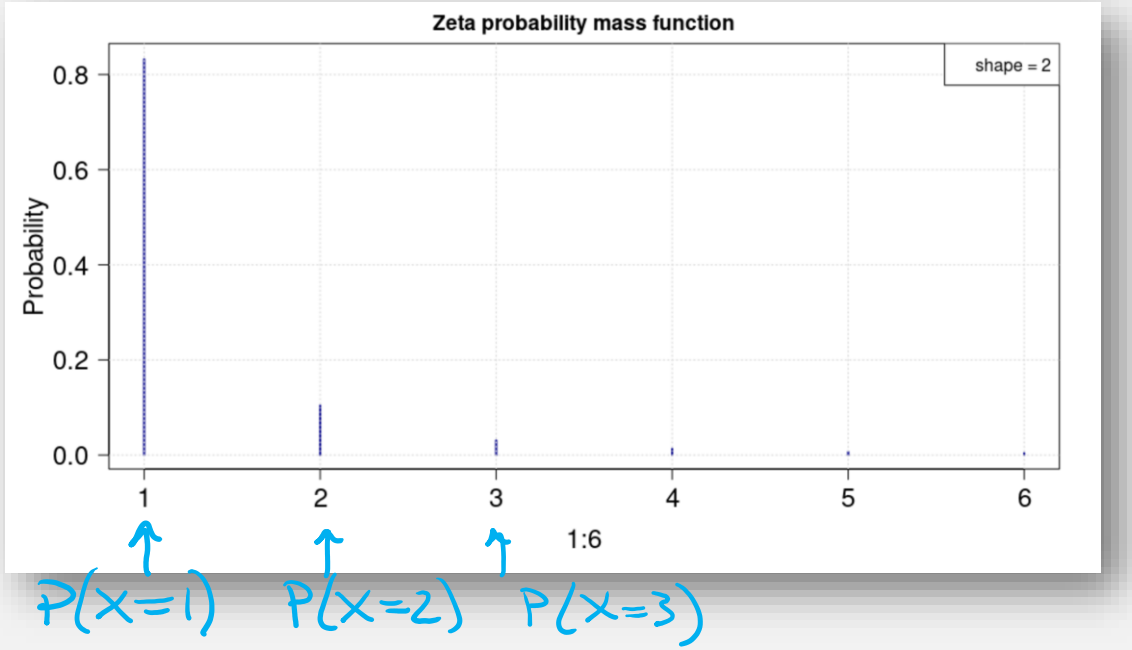
Continuous quantities

probability density function (pdf)

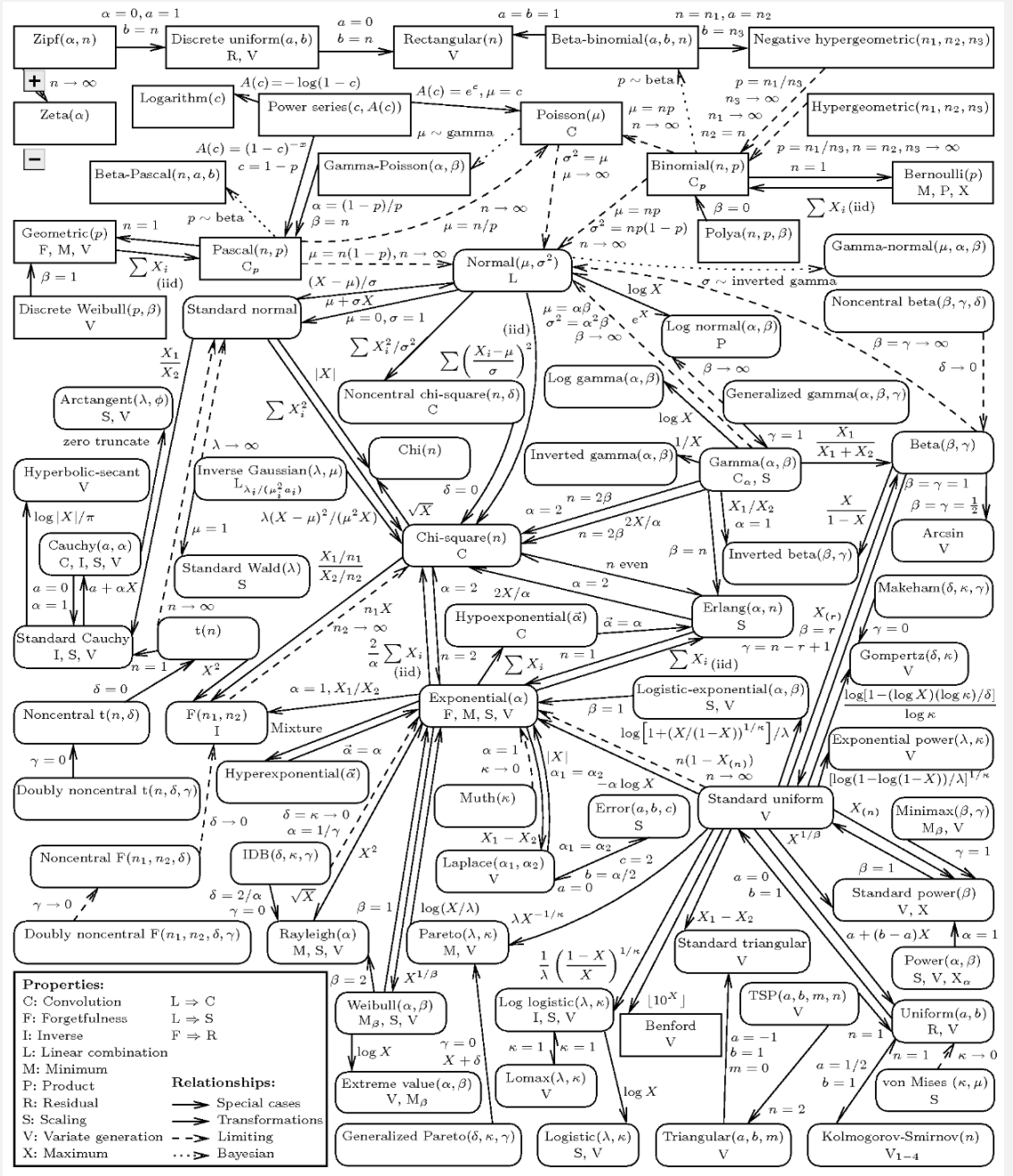


Discrete quantities

Probability mass function (pmf)



THERE ARE MANY UNIVARIATE DISTRIBUTIONS!



http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

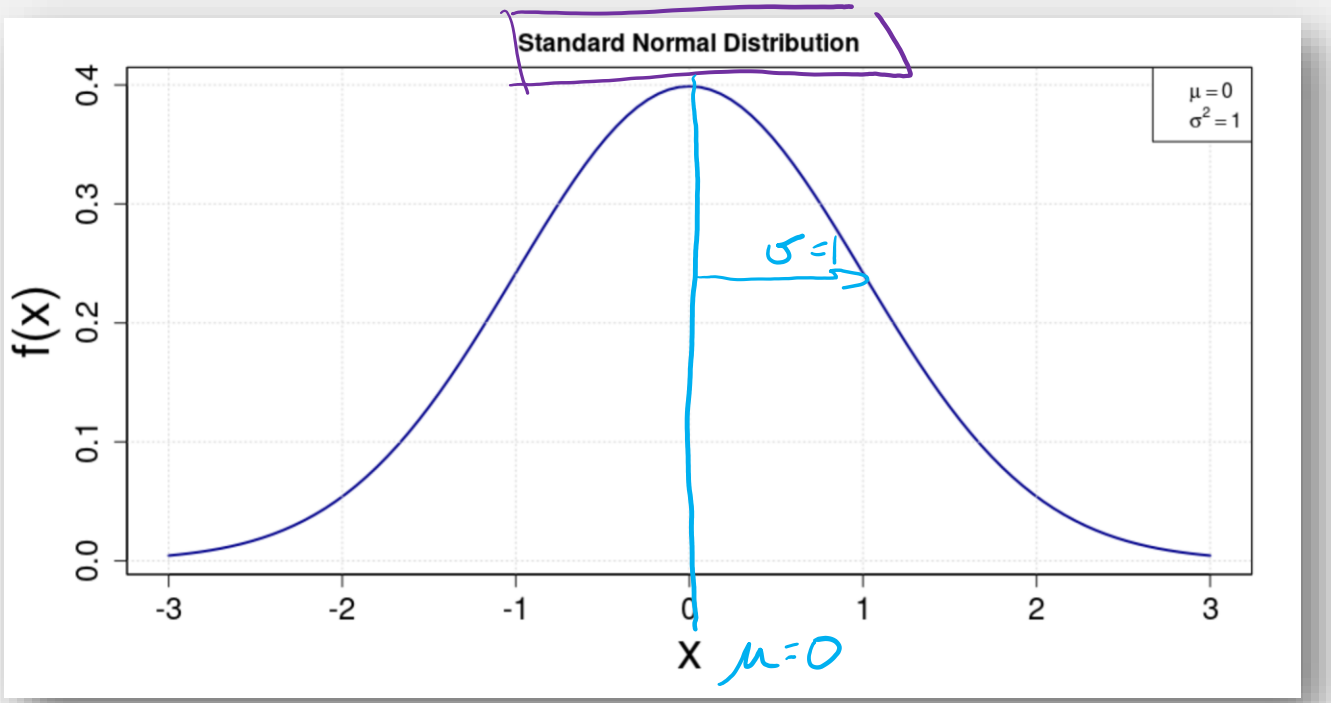
THE NORMAL DISTRIBUTION

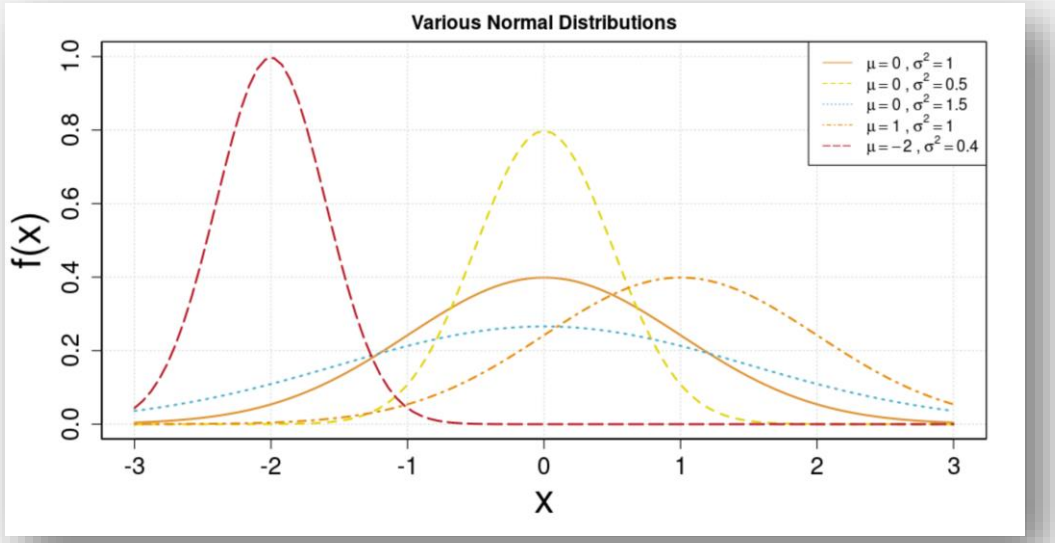
$$N(\mu, \sigma^2)$$

mean

variance

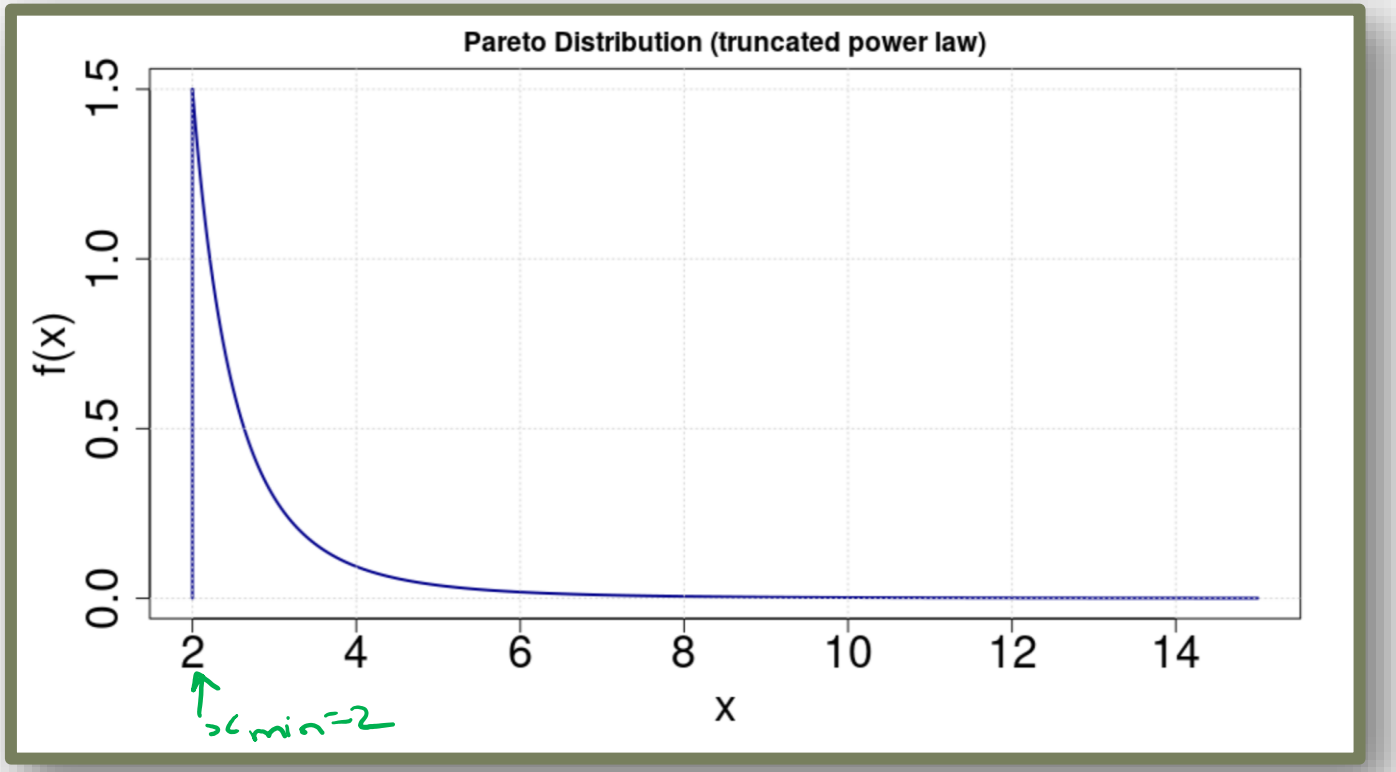
The mean and variance entirely define the Normal → if you know these you can plot it





pdf

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



Example:

Pareto distribution (truncated power-law)

Probability distribution function (pdf)

$$f(x) = \frac{\alpha x_{\min}^{\alpha}}{x^{\alpha+1}}$$

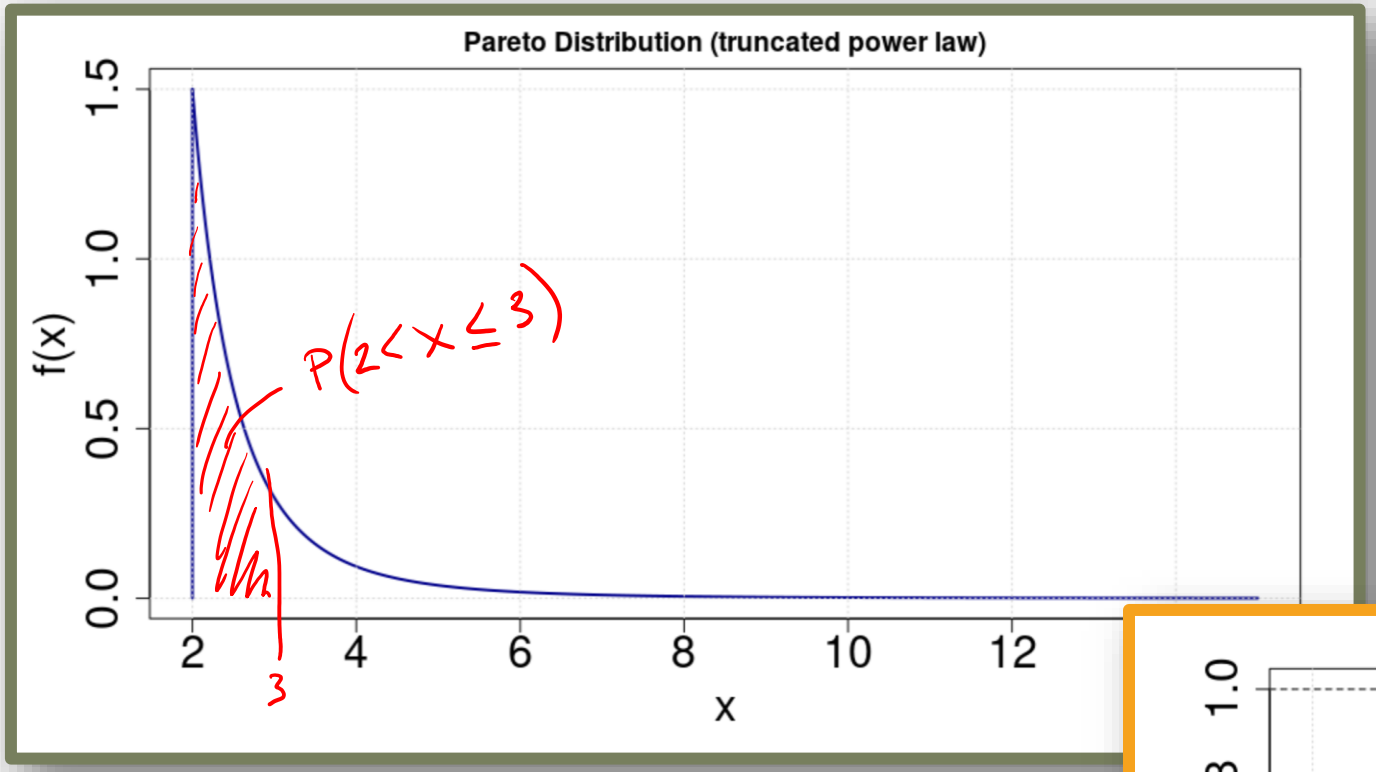
two parameters:

x_{min}

α

CDF:

$$F(x) = \int f(x)\,dx$$

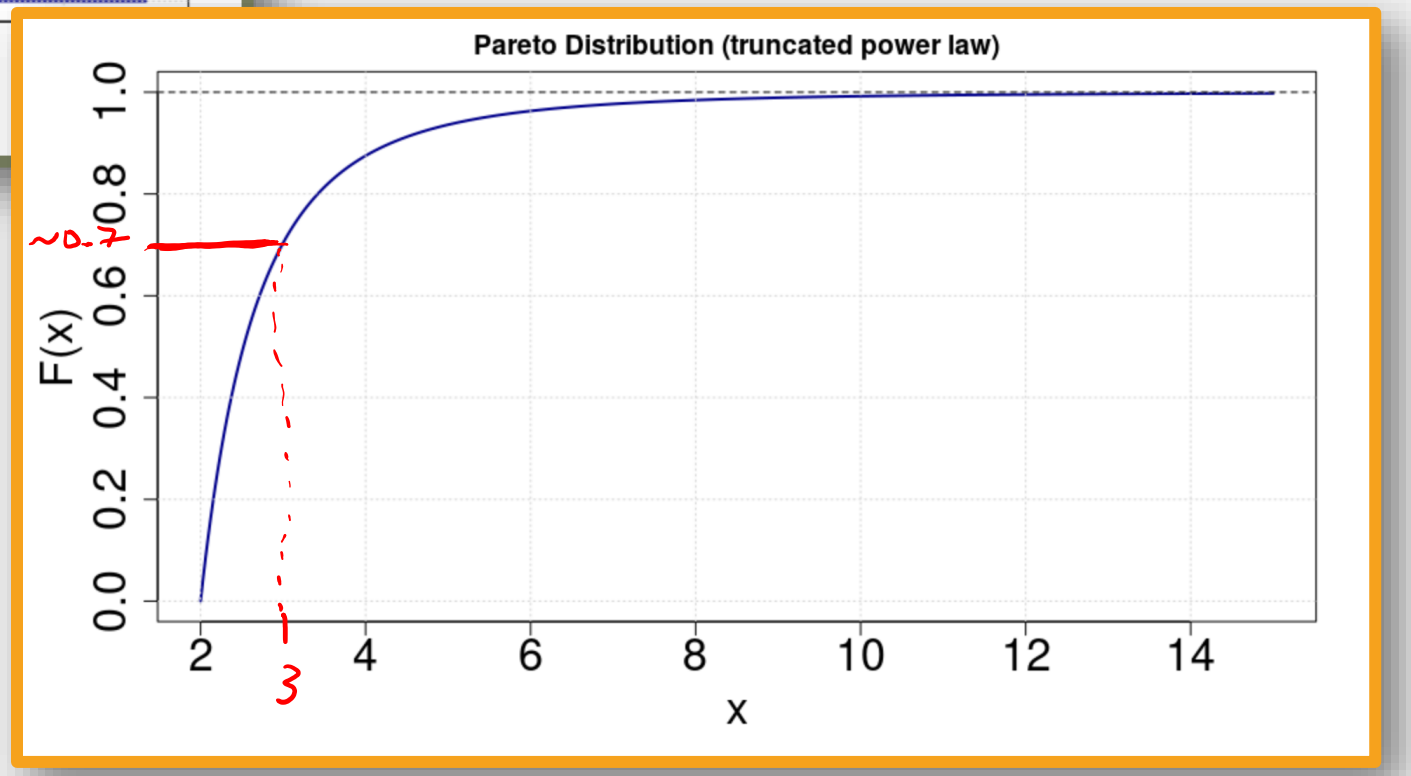


Example:

Pareto distribution (truncated power-law)

$$F(x) = P(X \leq x) = 1 - \left(\frac{x_{\min}}{x}\right)^{\alpha}$$

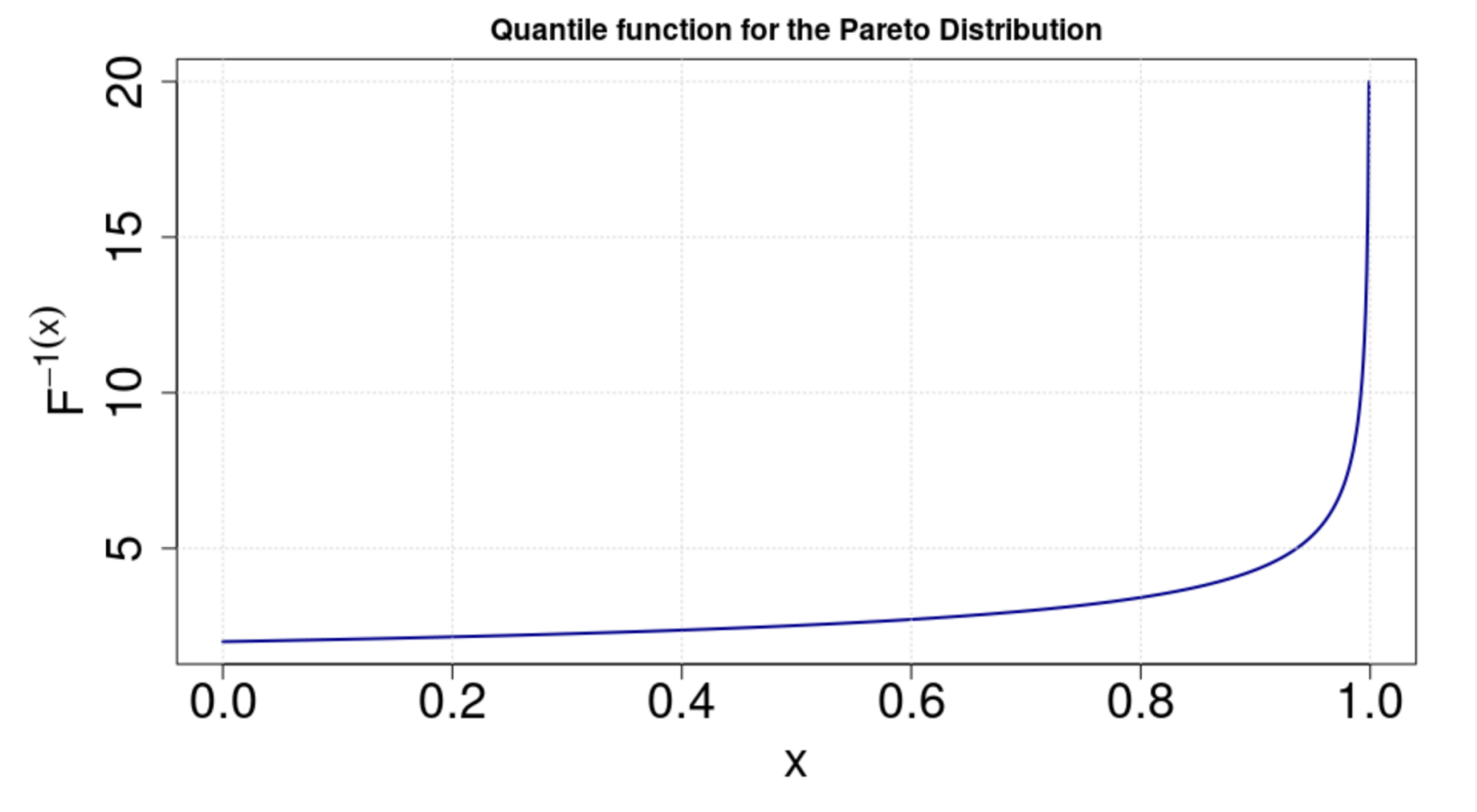
CUMULATIVE DISTRIBUTION FUNCTION (CDF)



Probability distribution function (pdf)

$$f(x) = \frac{\alpha x_{\min}^{\alpha}}{x^{\alpha+1}}$$

QUANTILE FUNCTION



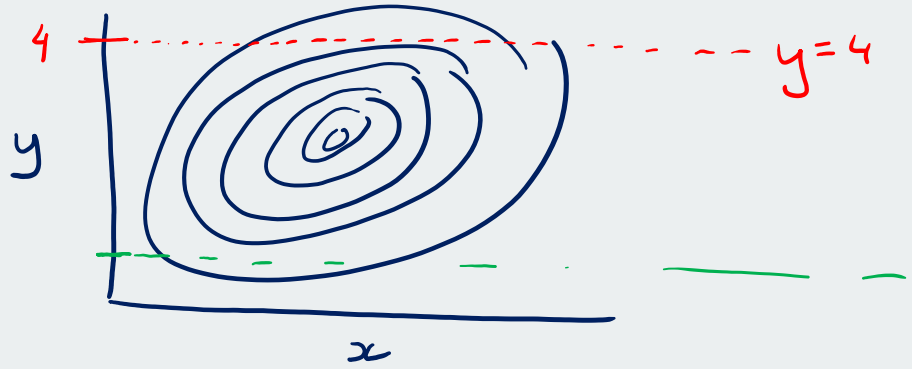
This is the inverse of the cumulative distribution function (cdf)

JOINT, CONDITIONAL, AND MARGINAL DISTRIBUTIONS

Joint, Conditional, and Marginal Distributions

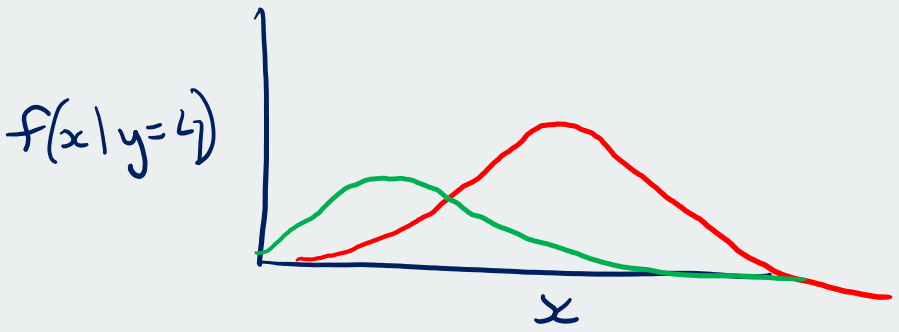
Joint Distribution

- The 2-d (or more!) distribution of two or more things



Conditional Distribution

- The distribution of a variable given an event or value for another variable



Marginal Distribution

- The distribution of a variable *regardless* of the values of the other variables

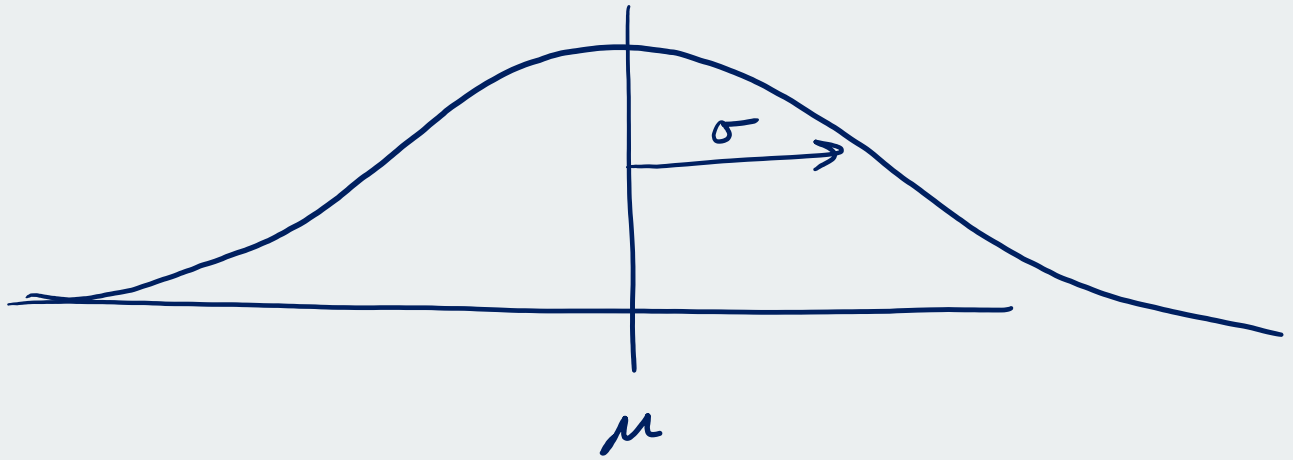
QUICK NOTE THE NORMAL DISTRIBUTION AND THE EMPIRICAL RULE

Leading up to the Empirical Rule

Normal or Gaussian distribution is defined
by a **mean** and a **variance**

$N(\mu, \sigma^2)$

$N(\mu, \sigma)$



Empirical Rule – "68-95-99 rule"

Normal or Gaussian distribution is defined by a **mean** and a **variance.**

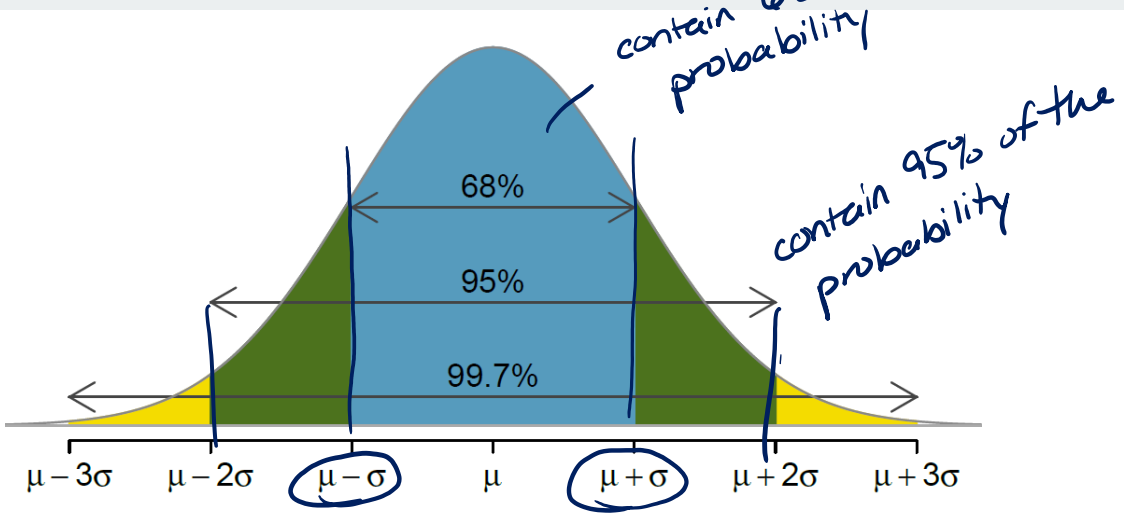
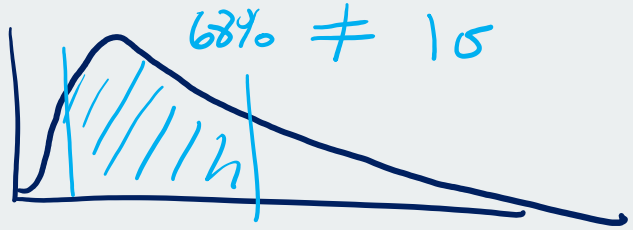


Figure 4.7: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

IMPORTANT NOTE:

68% is equivalent to 1σ (1 standard deviation) only in the case of Normal/Gaussian distributions!

The 68% quantile is not necessarily equal to 1σ in non-Gaussian distributions



DATA: REALIZATIONS OF RANDOM VARIABLES

Recall: a "population" versus a "sample"

Population

- The true, underlying distribution for some quantity
- E.g., the distribution of heights of people all over the world

Sample

- A sample drawn from some distribution
- E.g., randomly select 100 people from around the world and measure their heights
- Will not be exactly like the population because of randomness

Statistics can be used to try to understand the underlying population, when all you have is a sample

A (data) *sample* is a realization of a random variable

- Imagine we measured the heights of 25 randomly selected people from all over the world
  - These data are realizations x of a random variable X that represents the height of a person

- To perform (parametric) *statistical inference* on our data x, we assume how the random variable X is distributed

$X \sim N(\mu, \sigma)$

- For example, we could assume that the *population* of heights follows a normal distribution. Then we would write

"distributed as"

$$X \sim N(\mu, \sigma)$$

$X \sim U(0,1)$ (uniform)

$X \sim \text{Binom}(n, p)$

and perform statistical inference to infer the model parameters μ and σ

Randomness matters!

- We usually don't know how X is distributed!
- Our *random* data sample x is just that – a sample!
- Randomness can trick us! Humans like to look for patterns
- Things get even trickier when our data samples suffer from selection bias, observation bias, etc.
- We should look at and summarize our data in different ways. Be skeptical.
  - → Exploratory Data Analysis

EMPIRICALLY SUMMARIZING DATA

We have some data --- how can we summarize it?

1. Minimum
2. Maximum
3. Median
4. First quartile

   1/4 of the way through the ordered data
5. Third quartile

   3/4 of the way through the ordered data

Terminology difference:
quartile and
quantile/percentile

quartiles → "quarters"

quantile → any percentage

10%, 20%, ...

Box Plot

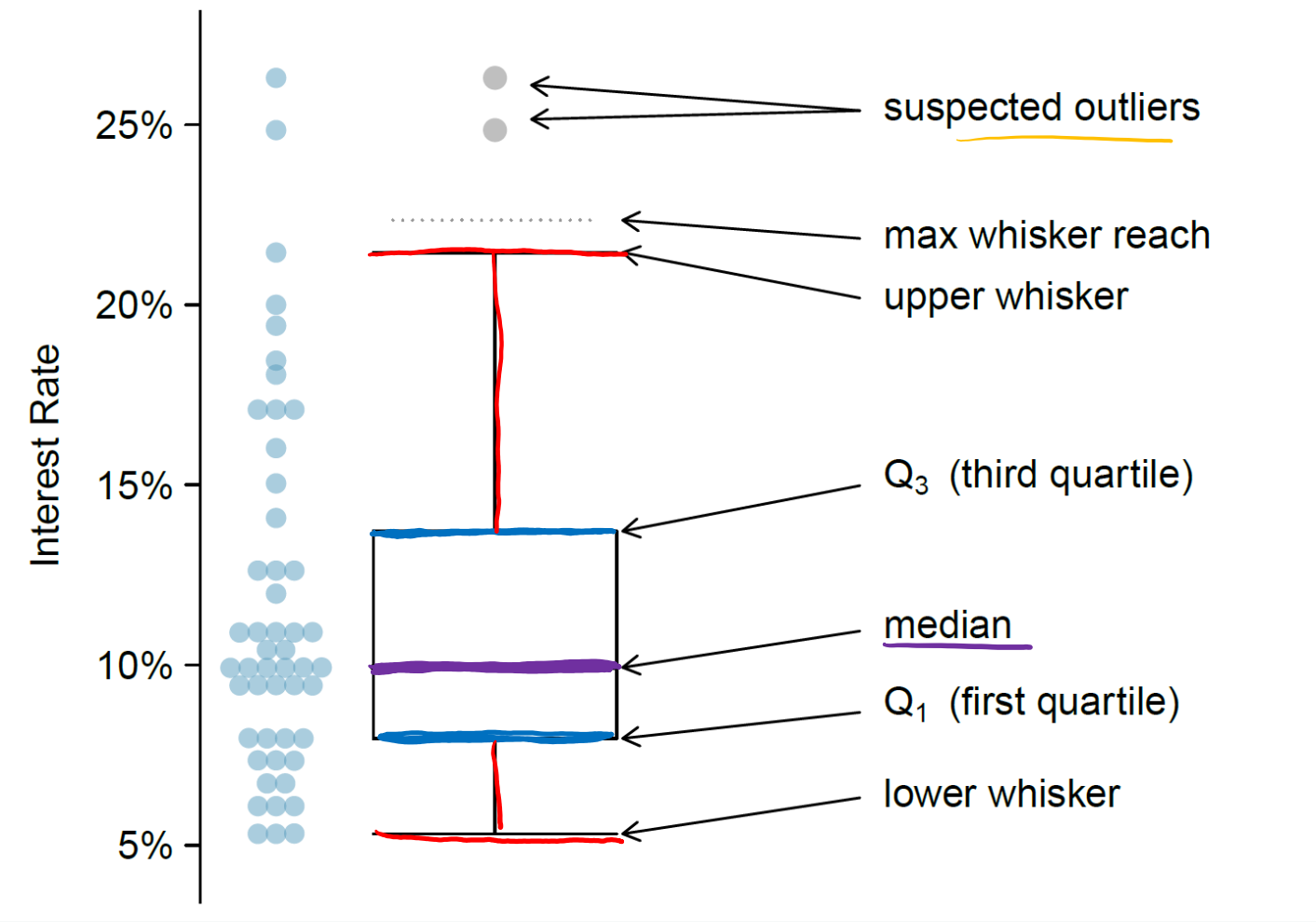
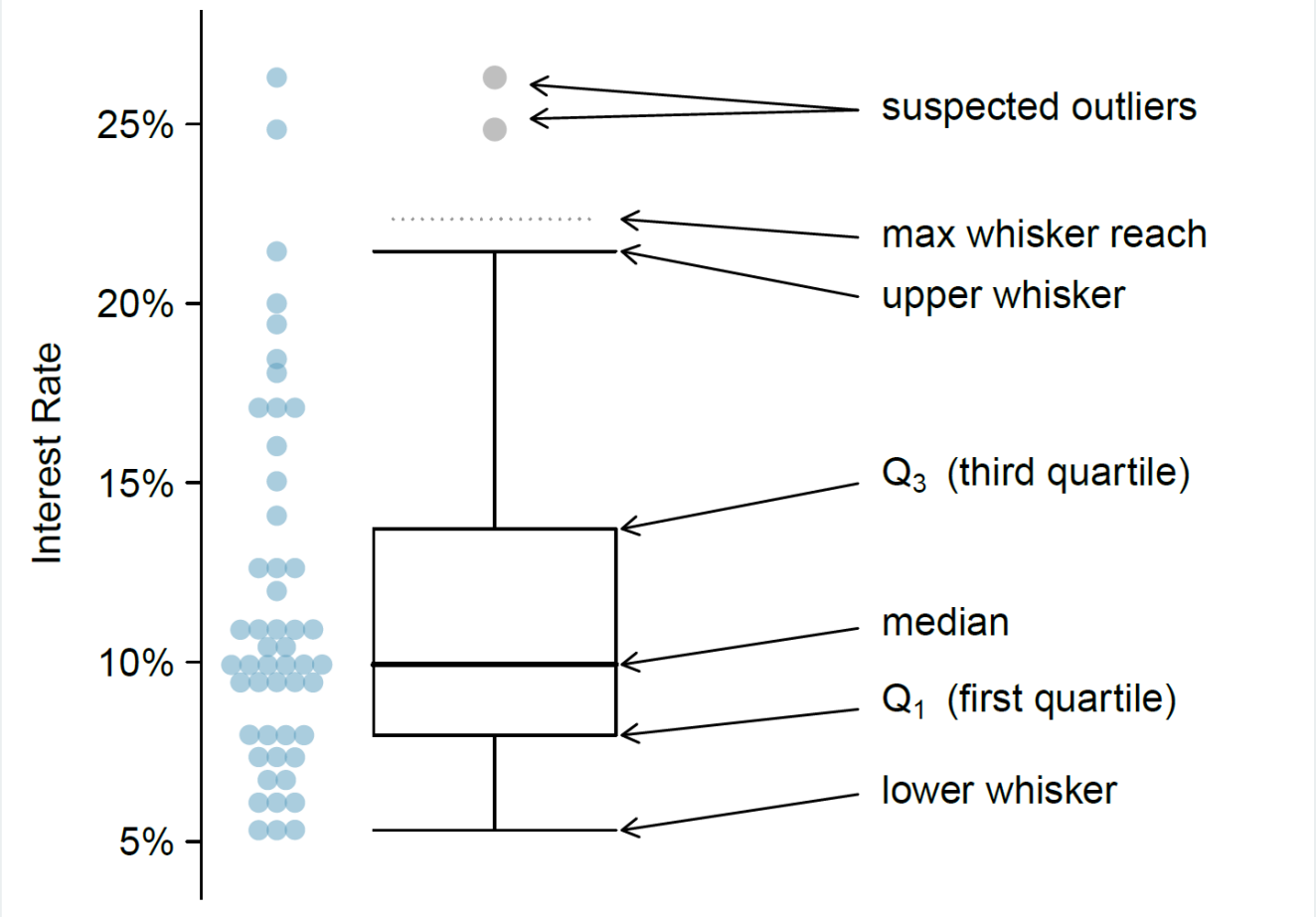


Figure 2.10, OpenIntro (4th ed.)

- Building a box plot:
  - Find the median first
  - Draw a rectangle that shows the interquartile range (IQR)
  - Extend the whiskers out to the furthest data point that is still within 1.5xIQR
  - Show the individual points that are outside the whiskers

Inter Quartile Range (IQR)

- *75% quartile – 25% quartile*
- Tells us how spread out the middle half of the data are



Mean vs. Trimmed Mean

- The mean is the average
- *The sample mean is:*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- The mean is not a robust statistic, because it is not "resistant" to extreme observations

- The trimmed mean is the average after removing the k largest and smallest values.

Sample Variance

- The average squared distance from the mean
- *The sample variance is:*

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

- What does the squaring do?
  - Makes all differences positive
  - Makes large differences relatively much larger

Variance

- *Distributions don't have to look the same to have the same variance*
- What other ways could you describe these distributions to differentiate them?

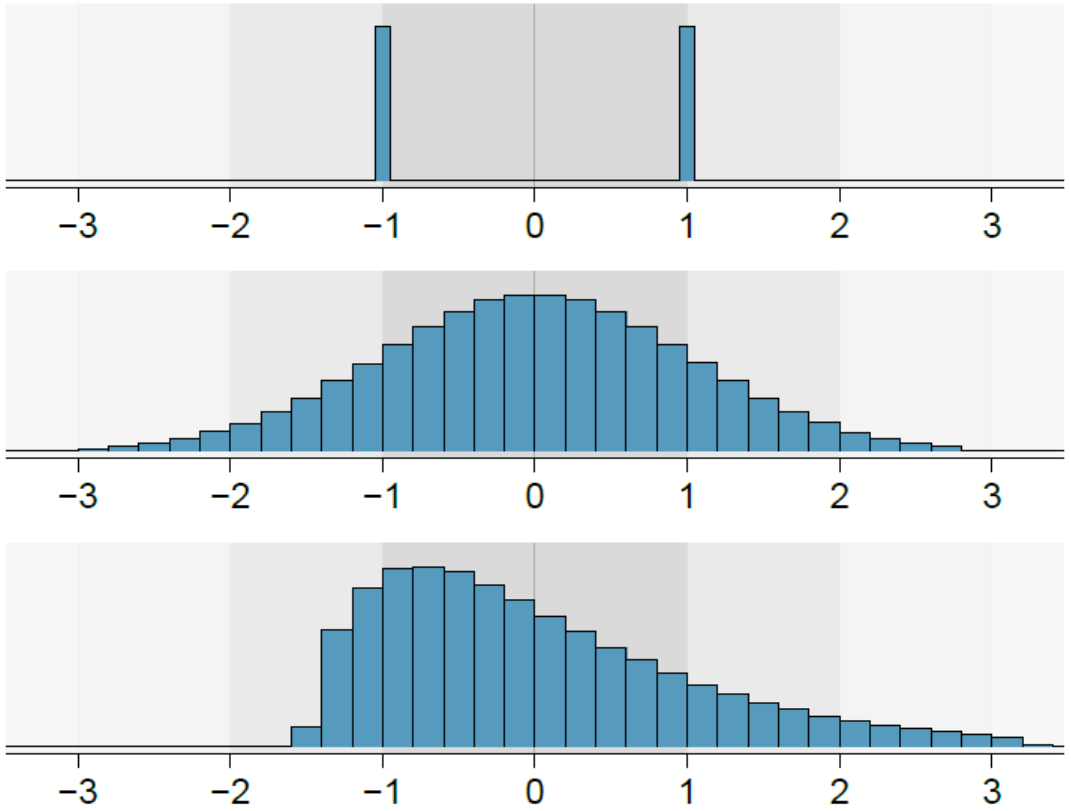


Figure 2.9: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

Standard Deviation

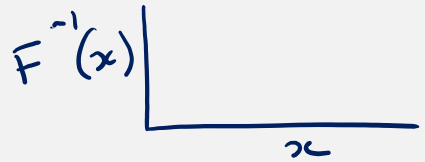
- *Standard deviation is the square root of the variance*

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

- Sometimes this is more intuitive to think about
- This is the *sample* standard deviation

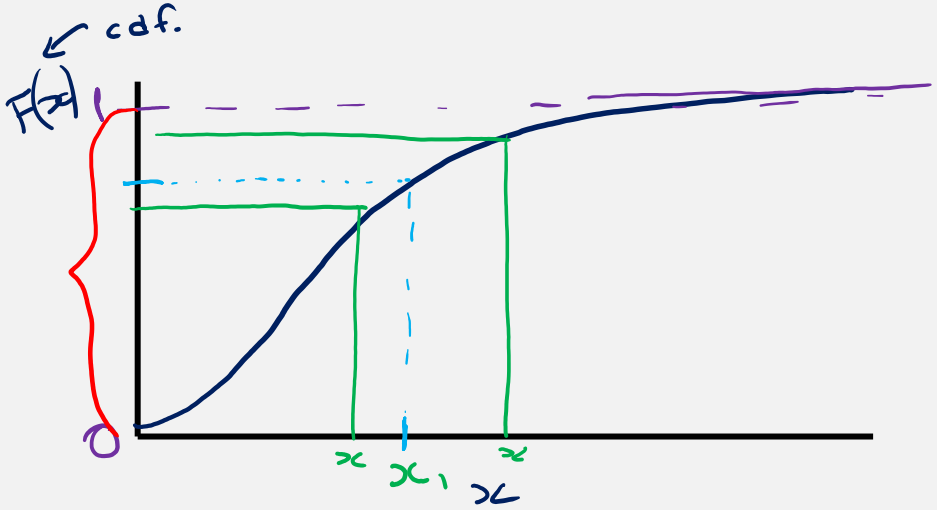
SAMPLING MOCK DATA FROM A DISTRIBUTION

TWO BASIC APPROACHES FOR SAMPLING FROM A DISTRIBUTION



Inverse cdf Method, $F^{-1}(x)$

- First choice if the inverse cdf is tractable

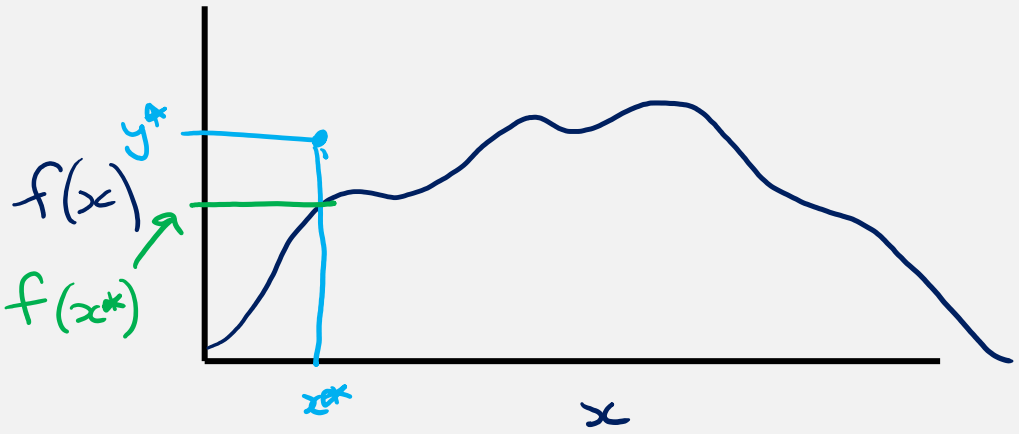


(1) draw a # between 0 and 1
(2) Pass that # to $F^{-1}(x)$

x will follow the distribution

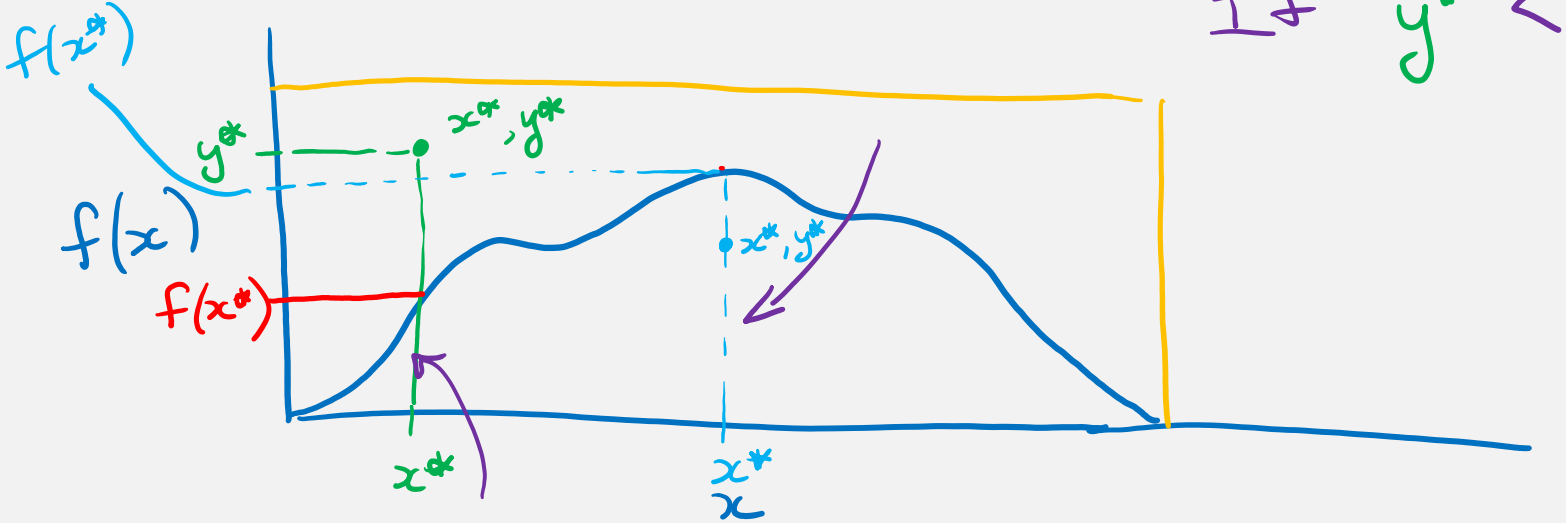
Accept/Reject Algorithm

- Useful when you can't write down the inverse cdf



(1) draw a random pair x^* and y^*
(2) compare y^* to $f(x^*)$ and keep x^* if $y^* \leq f(x^*)$
(3) repeat

JUPYTER NOTEBOOK EXERCISE



If $y^* < f(x^*)$ then accept x^*

→ in this drawing blue x^* would be accepted

→ green x^* would be rejected