

# **Week 3:**

# **Probability and distributions**

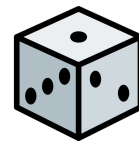
Phase 1

# Key Ideas

- Statistics is particularly useful when making inferences about data.
- We can never be 100% certain about our inferences. It's all about quantifying the certainty of an observation using **probability**.
- In general, observations follow a **probability distribution**, i.e. they exhibit different probabilities.

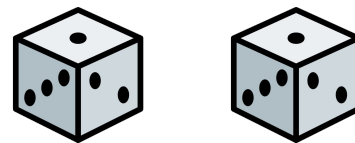
# Examples

Imagine you have one dice:



- Chance of getting 1 when rolling it?  $1/6$
- Chance of getting a 1 or 2 in the next roll?  $2/6$
- Chance of getting either 1, 2, 3, 4, 5, or 6 on the next roll?  $1$  (100%)
- Chance of not rolling a 2?  $5/6$

Now you have two dice:



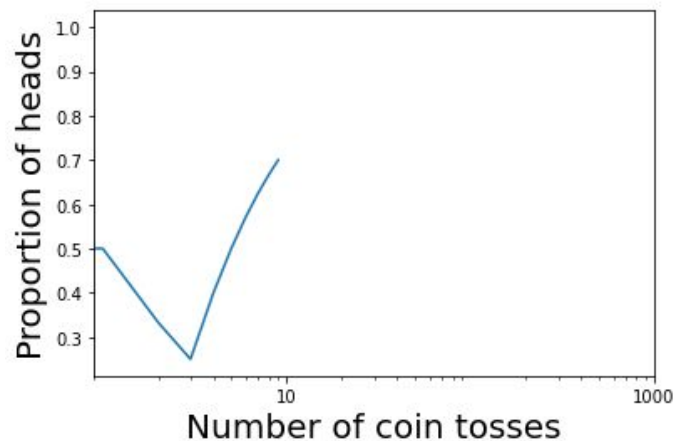
- What is the chance of getting two 1s?  $1/6 \times 1/6 = 1/36$

# Probability: a formal definition

The probability of an outcome is the **proportion** of times the outcome would occur if we observed the **random** process an **infinite** number of times.

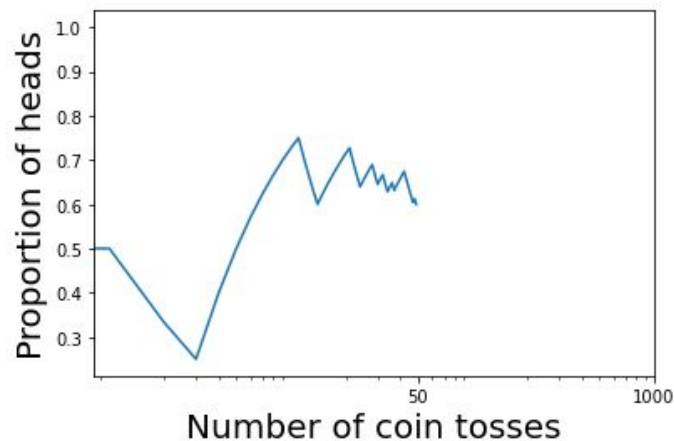
# Probability

Example: Probability of heads after N coin tosses...



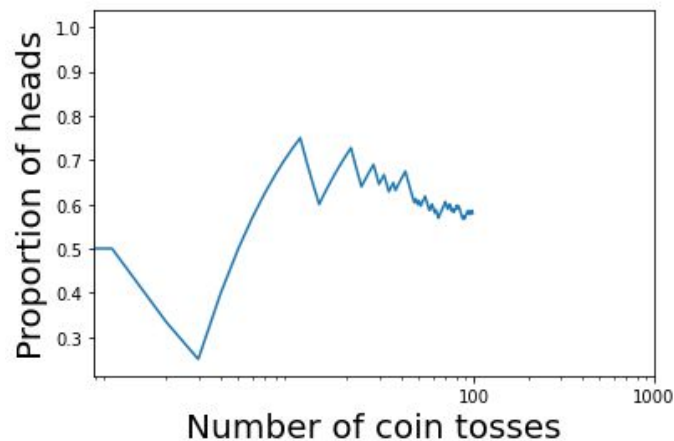
# Probability

Example: Probability of heads after N coin tosses...



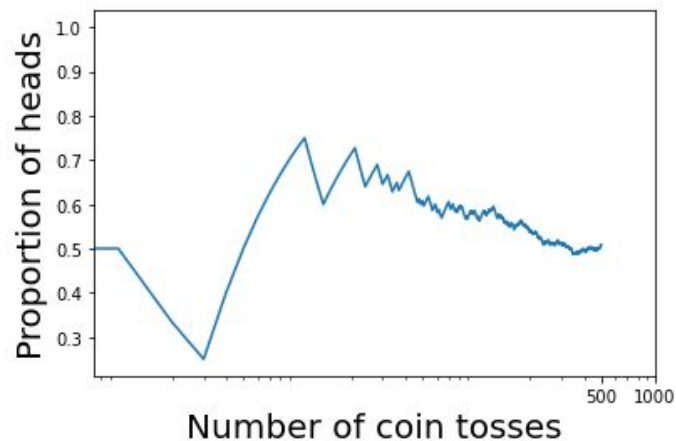
# Probability

Example: Probability of heads after N coin tosses...



# Probability

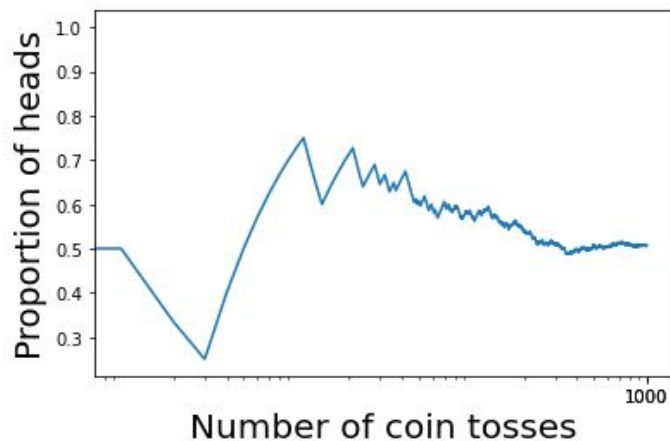
Example: Probability of heads after  $N$  coin tosses...





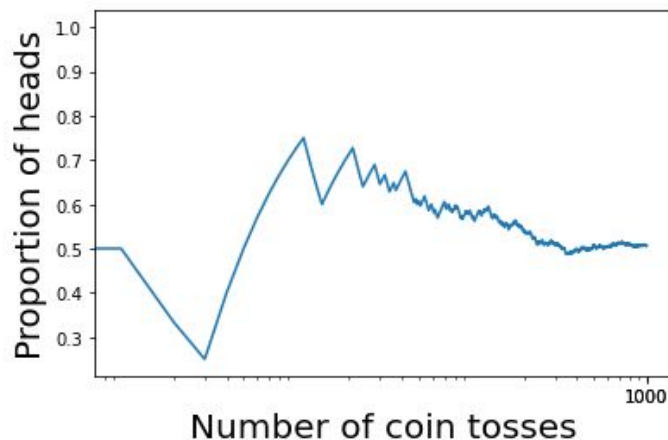
# Probability

Example: Probability of heads after  $N$  coin tosses...



# Probability

Example: Probability of heads after  $N$  coin tosses... After a sufficiently **large** number of times, the probability **converges** to the real value (here 0.5).



# Probability models

A random phenomenon can be mathematically represented in a **probability model**, which has the following ingredients:

- Sample space: The list of all possible outcomes in a random phenomenon.  
e.g. : {Blue jeans, green jeans, grey jeans, black suit, blue track}
- Event: An outcome or a set of outcomes of a random phenomenon.  
e.g. {Blue jeans}
- Each event gets assigned a probability.

# Probability distributions: Sample space

The sample space  $S$  is the set of **all possible outcomes** of a random phenomenon.

# Probability distributions: Sample space

The sample space  $S$  is the set of **all possible outcomes** of a random phenomenon.

**Example:** What is the sample space of a dice?

# Probability distributions: Sample space

The sample space  $S$  is the set of **all possible outcomes** of a random phenomenon.

**Example:** What is the sample space of a dice?

$\{1, 2, 3, 4, 5, 6\}$

# Probability distributions: Sample space

The sample space  $S$  is the set of **all possible outcomes** of a random phenomenon.

**Important:** The possible outcomes (i.e. the sample space) depend upon how our "experiment" is.

# Probability distributions: Sample space

The sample space  $S$  is the set of **all possible outcomes** of a random phenomenon.

**Important:** The possible outcomes (i.e. the sample space) depend upon how our "experiment" is.

**Question:** what is the sample space of tossing 2 coins?



# Probability distributions: Sample space

The sample space  $S$  is the set of **all possible outcomes** of a random phenomenon.

**Important:** The possible outcomes (i.e. the sample space) depend upon how our "experiment" is.

**Question:** what is the sample space of tossing 2 coins?

$\{HH, HT, TH, TT\}$

# Probability vs Statistics

**Question:** How different are Probability and Statistics?

# Probability vs Statistics

- What are the chances of a fair coin coming up heads 10 times in a row?
- If I roll two six sided dice, how likely is it that I'll roll two sixes?
- How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
- What are the chances that I'll win the lottery?
- If my friend flips a coin 10 times and gets 10 heads, are they playing a trick on me?
- If five cards off the top of the deck are all hearts, how likely is it that the deck was shuffled?
- If the lottery commissioner's spouse wins the lottery, how likely is it that the lottery was rigged?

# Probability vs Statistics

- Probabilistic questions assume a **known** model of the world (e.g.  $P(\text{heads}) = 0.5$ )
- We use this model to perform calculations.
- Here, the model is known, but the data are not.
- We do not know the truth about the world.
- We only have the data.
- We want to use the data to learn (infer) the truth about the world.

# Probability vs Statistics



1. The world generates our data.
2. Statistical models and data generate probabilities.
3. With probabilities, we can make predictions about the world.

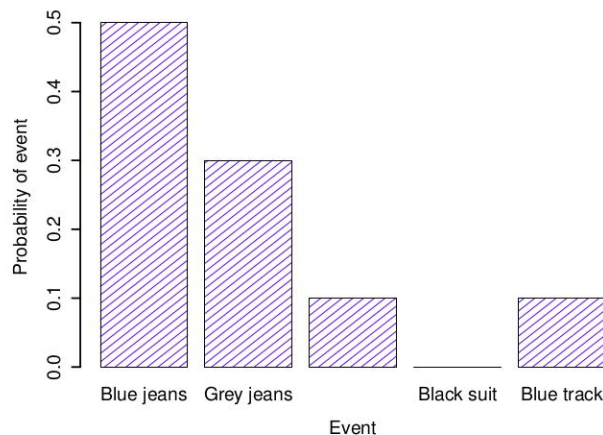
# Probability vs Statistics



Statistical inference is to figure out **which** probability models are right. Although not the same, both are deeply **connected** to one another.

# Probability distributions

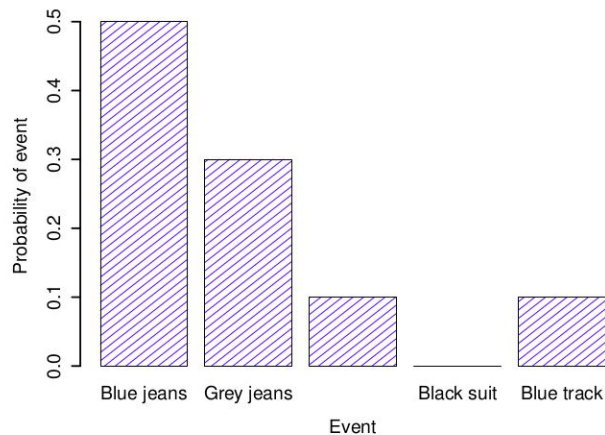
- A **probability distribution**  $P(X)$  is a mathematical function that provides the probabilities of an event  $X$  from all different possible **outcomes (the sample space)**.



Which pants?	Label	Probability
Blue jeans	$X_1$	$P(X_1) = .5$
Grey jeans	$X_2$	$P(X_2) = .3$
Black jeans	$X_3$	$P(X_3) = .1$
Black suit	$X_4$	$P(X_4) = 0$
Blue tracksuit	$X_5$	$P(X_5) = .1$

# Probability distributions

- A **probability distribution**  $P(X)$  is a mathematical function that provides the probabilities of an event  $X$  from all different possible **outcomes (the sample space)**.



Which pants?	Label	Probability
Blue jeans	$X_1$	$P(X_1) = .5$
Grey jeans	$X_2$	$P(X_2) = .3$
Black jeans	$X_3$	$P(X_3) = .1$
Black suit	$X_4$	$P(X_4) = 0$
Blue tracksuit	$X_5$	$P(X_5) = .1$

Form of a density function  $\longrightarrow$  Their probabilities must sum 1.



# Probability distributions

- A **probability distribution**  $P(X)$  is a mathematical function that provides the probabilities of an event  $X$  from all different possible **outcomes (the sample space)**.

# Probability distributions

- A **probability distribution**  $P(X)$  is a mathematical function that provides the probabilities of an event  $X$  from all different possible **outcomes (the sample space)**.
- As any function, it will in general depend on some underlying parameters  $\theta_i$

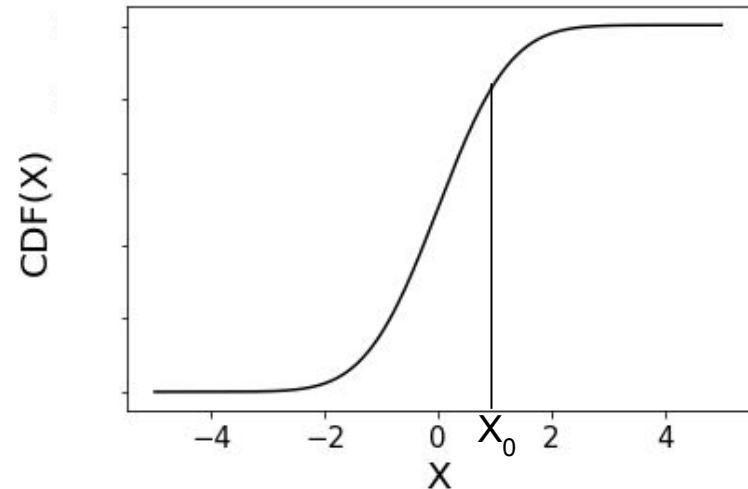
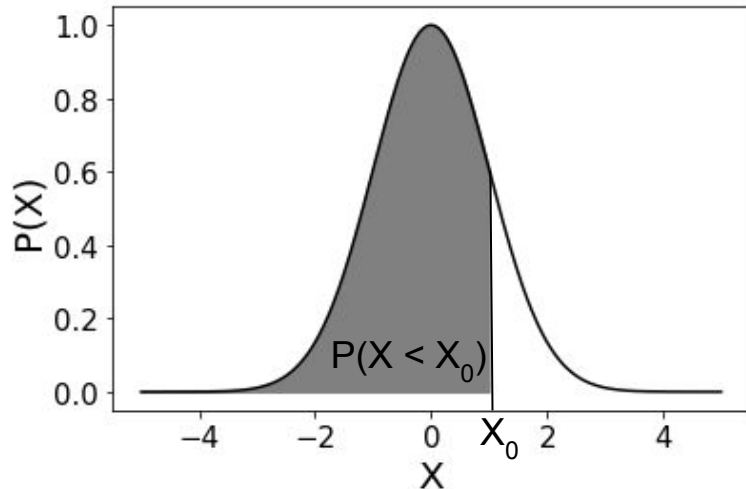
# Probability distributions

- A **probability distribution**  $P(X)$  is a mathematical function that provides the probabilities of an event  $X$  from all different possible **outcomes (the sample space)**.
- As any function, it will in general depend on some underlying parameters  $\theta_i$ .
- Keep this in mind for the future: Probability distributions can be **discrete**, if the outcomes take only a set of finite values, or **continuous**, if it takes an infinite set of outcomes.

# Probability distributions: some derived functions

## ➤ Cumulative distribution function CDF(x):

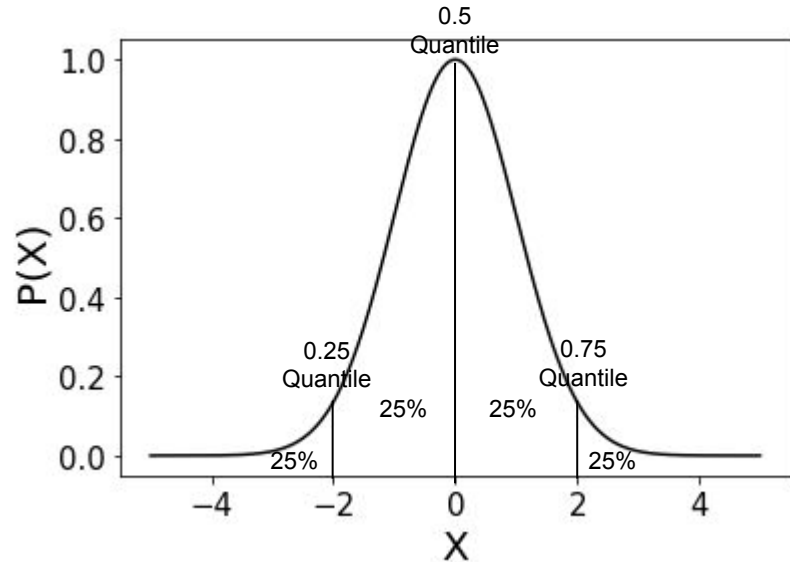
It tells you the **probability** of obtaining an outcome smaller than or equal to  $X_0$ , that is,  $P(X \leq X_0)$ .



# Probability distributions: some derived functions

## ➤ Quantile function $Q(p)$ :

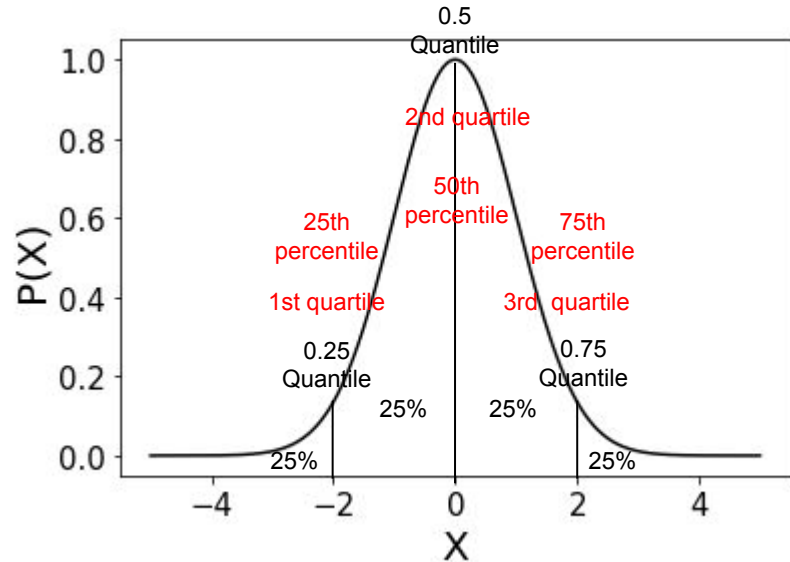
It's the inverse of the CDF, and tells you the particular  $X_0$  for which the probability is less than or equal to a given value  $p$ . It is often given in terms of percentages (e.g. 25%, 75%, etc)



# Probability distributions: some derived functions

## ➤ Quantile function $Q(p)$ :

It's the inverse of the CDF, and tells you the particular  $X_0$  for which the probability is less than or equal to a given value  $p$ . It is often given in terms of percentages (e.g. 25%, 75%, etc)



# Binomial distribution

- It's a **discrete** distribution, that models the probability that **positive** events occur in a given sample of **repeated** independent experiments.

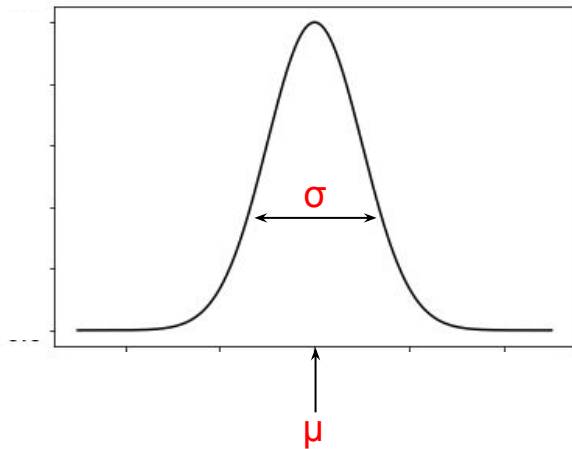
## Example:

- If I toss a coin 5 times, what's the probability of getting 3 heads?
  - If I roll a dice 10 times, what's the probability of getting 4 fives?
- Therefore, it will depend on the number of independent occurrences  $N$ , and the probability of the positive occurrence  $p$ .

$$\theta_i = \{p, N\}$$

# Gaussian distribution

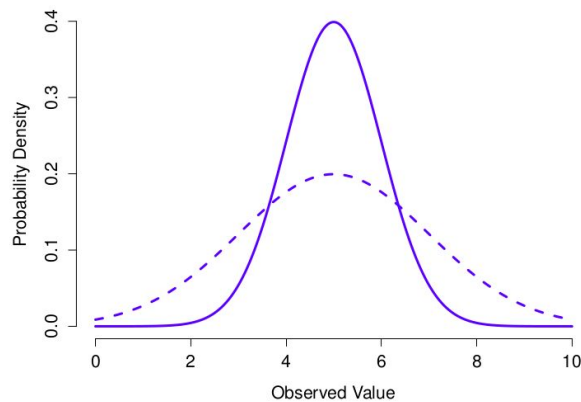
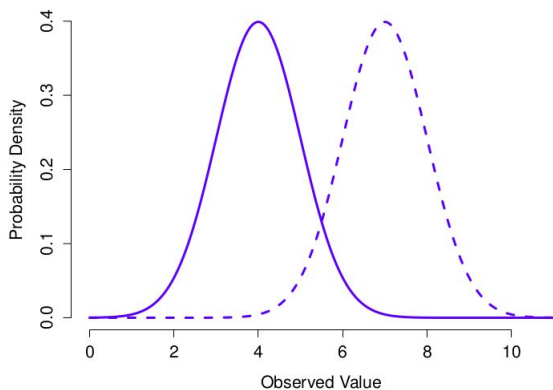
- It's a **continuous** distribution and probably the most important one in statistics. (Why? We'll see this in the future...)
- A gaussian distribution is described by two parameters: the mean of the distribution  $\mu$ , and the standard deviation of the distribution  $\sigma$ .  $\theta_i = \{\mu, \sigma\}$





# Gaussian distribution

- It's a **continuous** distribution and probably the most important one in statistics. (Why? We'll see this in the future...)
- A gaussian distribution is described by two parameters: the mean of the distribution  $\mu$ , and the standard deviation of the distribution  $\sigma$ .  $\theta_i = \{\mu, \sigma\}$

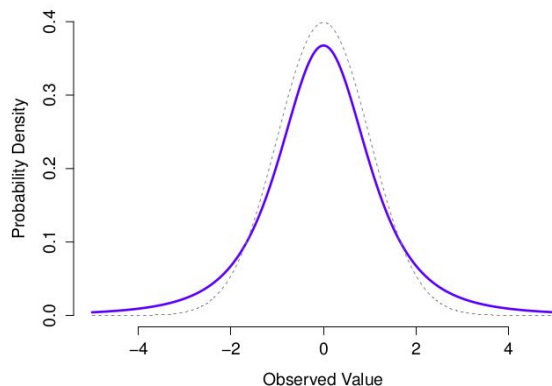


# Gaussian distribution

- It's a **continuous** distribution and probably the most important one in statistics. (Why? We'll see this in the future...)
- A gaussian distribution is described by two parameters: the mean of the distribution  $\mu$ , and the standard deviation of the distribution  $\sigma$ .  $\theta_i = \{\mu, \sigma\}$
- The special case  $\mu=0$  and  $\sigma=1$  has its own name that you'll often see: **The normal distribution.**

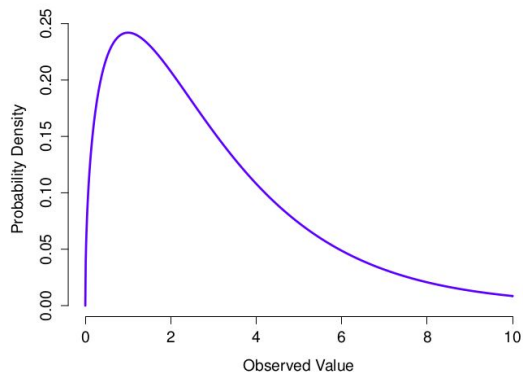
## Other types: Student's $t$ -distribution

- The Student's  $t$ -distribution is like a normal distribution but with **heavier tails**.  
 $\theta_i = \{df\}$  ( $df$  stands for degrees of freedom).
- It is a very important distribution in statistics, particularly for things like assessing differences between the **means of two samples** and constructing **confidence intervals** (we'll see this in the future).



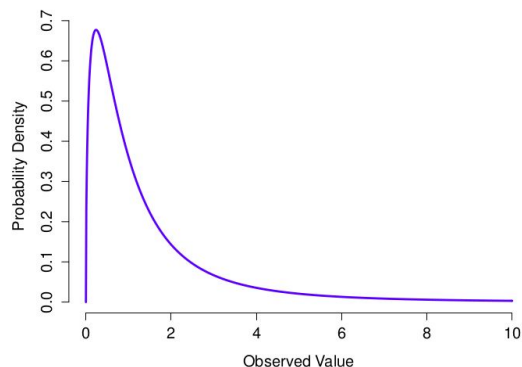
## Other types: $\chi^2$ distribution

- It is the distribution of the **sum of squares** (keep this in mind for the next slide) of independent variables **normally** distributed  $\theta_i = \{df\}$ .
- It is a widely used distribution in statistics, e.g. testing differences in **proportions** between groups, or **goodness fit** of the data.



# Other types: F-distribution

- It is related to the  $\chi^2$  distribution; specifically as the **ratio** between two  $\chi^2$  statistics.  $\theta_i = \{df_1, df_2\}$
- It usually arises as the ratio between variances (aha, here you have the “sum of squares” that I mentioned before). This ratio is common in testing **mean differences** across groups (ANOVA test).



# Recap

- We can use probability models to quantify the **certainty** of our observations.

# Recap

- We can use probability models to quantify the **certainty** of our observations.
- Statistics develops on these probability models to quantify how **likely** our world is given a **certain testable condition** (This is basically the famous **p-value!!!**)

# Recap

- We can use probability models to quantify the **certainty** of our observations.
- Statistics develops on these probability models to quantify how **likely** our world is given a **certain testable condition** (This is basically the famous **p-value!!!**)
- There exists **many** probability distributions, each suitable for **specific situations/problems**, so you'll need to choose wisely.

