Introduction to Probability and Statistics

Question: If you flip a quarter, what are the chances that the coin lands with the "heads" side facing up?

Introduction

In this notebook, we will use a combination of logic, experiment, and computing to get an overview of (almost) all of the concepts from probability and statistics that will we cover in more detail as the course progresses. That is, we are going to get a glimpse of the main ideas of both probability and statistics (statistical inference).

Data science and machine learning rely heavily on both probability and statistics. This article provides some perspective on probability and statistics for machine learning. In addition, the following articles from Wikipedia are worth skimming:

- Probability is the measure of the likelihood that an event will occur.
- Statistics is a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation.
- Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution.

One of the main takeaways that I would like for you to get from this introduction is that statistics (statistical inference) relies on probability. Thus, it is not really possible to understand statistics without probabilty.

Coin Flipping and a Probability Model

Consider the flip of a *fair*, two-sided coin. Flipping the coin is an example of an (random) **experiment**.

Suppose that we call one side of the coin heads and the other tails. After flipping the coin, the (only) possible **outcomes** are heads (H) or tails (T).

In this case, we say that the "**sample space**" (set of all possible outcomes) for the coin flipping experiment is $\Omega = \{H, T\}$. For reasons that you will see later, we call the subsets of the sample space Ω the **events** of the experiment. Thus the events corresponding to the coin flipping sample space Ω are: $\emptyset$, $\{H\}$, $\{T\}$, and $\{H, T\} = \Omega$. (Recall that a set with n elements has 2^n possible subsets.)

We call each run of an experiment (*e.g.*, flip a coin) a **trial**.

By a fair coin, we mean that the chances of the coin landing with heads facing up when flipped is the same as it landing with tails facing up. Thus, it is commonly said that the coin has a 50% chance of turning up heads (and also a 50% chance of turning up tails).

Another way to say this is that the probability that the coin lands heads up is 1/2 (so the probability of landing a tails is also 1/2).

In notation,

$P(\{H\}) = \frac{1}{2}$

and

$P(\{T\}) = \frac{1}{2}$.

(For technical reasons we also write that $P(\emptyset) = 0$, and $P(\Omega) = 1$). Note that we will always take probability values to be nonnegative numbers less than or equal to 1.

Our goal is to construct a robust mathematical model (*i.e.*, a probability model) for the coin tossing experiment (and more generally any random experiment). One initial annoyance is that our outcomes are described by the elements of the set Ω and these are not numerical quantities which are much easier to work with in practice (especially on a computer). There is a very powerful idea that will allow us to convert general questions about probabilities into statments about numberical quantities.

For example, we can convert the coin flipping experiment into a mathematical function $X : \Omega \to \{0, 1\}$ that takes on numerical values by defining

$X(\{H\}) = 1$,

and

$X(\{T\}) = 0$.

That is, we are simply counting the number of heads after one flip of a coin. A function like this is called a **random variable**. Since we are now working with numerical values, we can exploit this to develop a mathematical model for coin flipping, because the probability that the random variable X will equal 1 is $\frac{1}{2}$ and the probability that X will equal 0 is also $\frac{1}{2}$. That is,

$P(X = 1) = \frac{1}{2}$, and $P(X = 0) = \frac{1}{2}$

We can go even further and write an expression for this, if $x \in \{0, 1\}$,

$p(x) = P(X = x) = \left(\frac{1}{2}\right)^{x} \left(1 - \frac{1}{2}\right)^{1-x}$

A function like p is called a **probability mass function** (pmf). There is also a realted function called the **cummulative distribution function** (CDF) denoted $F(x)$ and defined as

$F(x) = \sum_{s \leq x} p(s)$ for $x \in \{0, 1\}$.

Thus,

$F(0) = p(0) = \frac{1}{2}$,

and

$F(1) = p(0) + p(1) = \frac{1}{2} + \frac{1}{2} = 1$.

These are implemented in R by the functions dbinom and pbinom respectively. For example,

```r
dbinom(1,size=1,prob=0.5)
```

```
## [1] 0.5
```

and

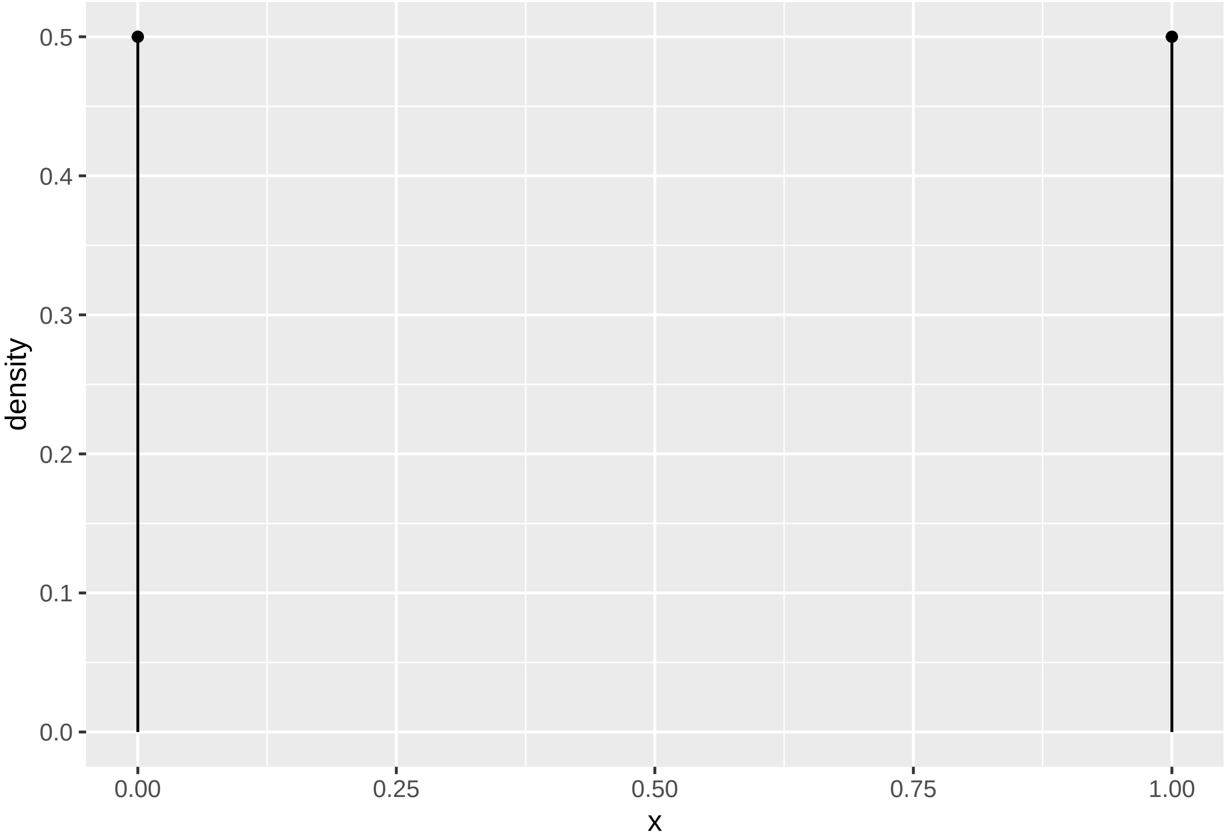
```r
pbinom(1,size=1,prob=0.5)
```

```
## [1] 1
```

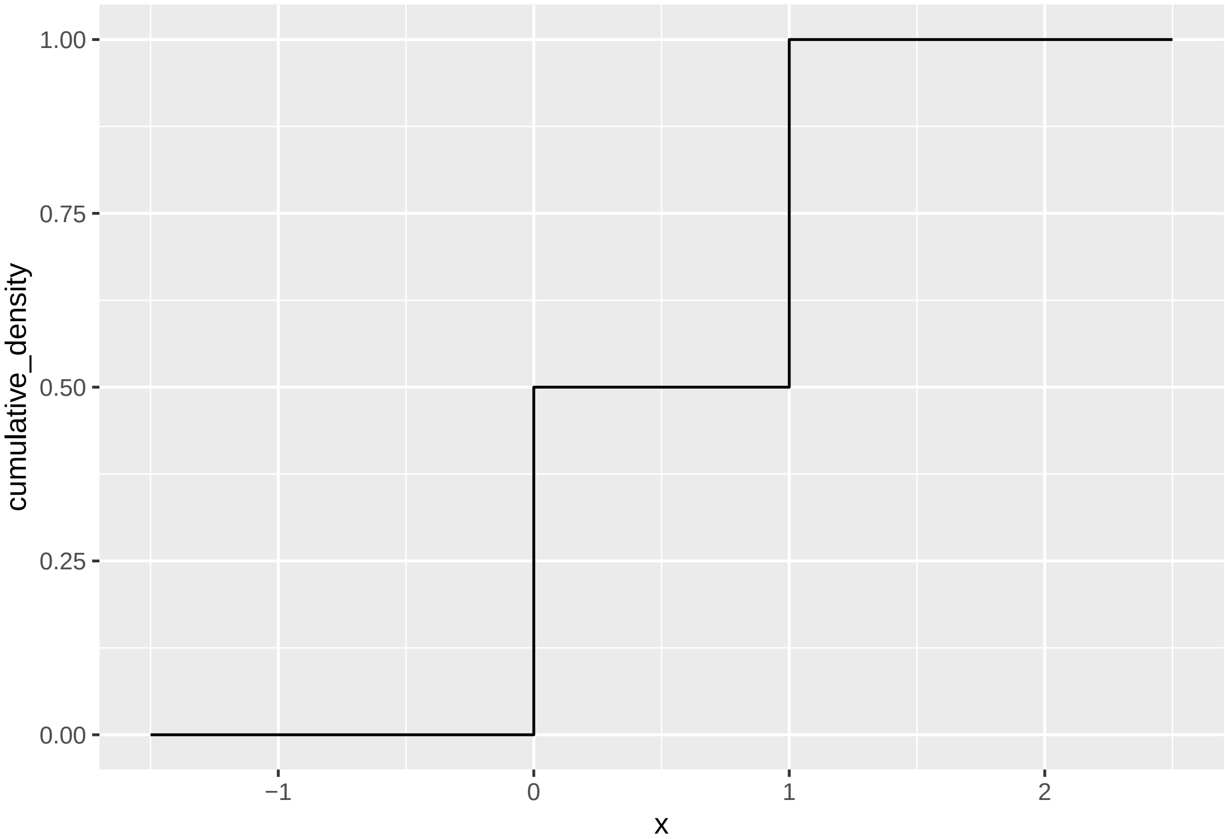
Exercise Try varying the first number in the input to the functions in the last two lines of code.

In addition, we can visualize these functions in R.

```
gf_dist("binom",params=list(size=1,prob=0.5),kind="density")
```



```
gf_dist("binom",params=list(size=1,prob=0.5),kind="cdf")
```



The benefit of all of this is that now it is easy to abstract to the situation where we want to compute the probability of getting a certain number x of heads after flipping the coin n times, that is, after n trials.

Exercise What would be an appropriate sample space for the experiment of flipping a coin n times?

Now for the situation where we want to compute the probability of getting a certain number x of heads after flipping the coin n times, the appropriate probability mass function would be

$p(x) = \left(\begin{array}{c} n \\ x \end{array}\right) \left(\frac{1}{2}\right)^x \left(1-\frac{1}{2}\right)^{n-x}$,

where $x \in \{0, 1, 2, 3, \ldots, n\}$, and $\left(\begin{array}{c} n \\ x \end{array}\right) = \frac{x!}{x!(n-x)!}$.

Note: This is derived by simple counting.

Again, this function is implemented in R. For example, we can compute the probability of getting 5 heads out of 10 flips:

```
dbinom(5,size=10,prob=0.5)
```

```
## [1] 0.2460938
```

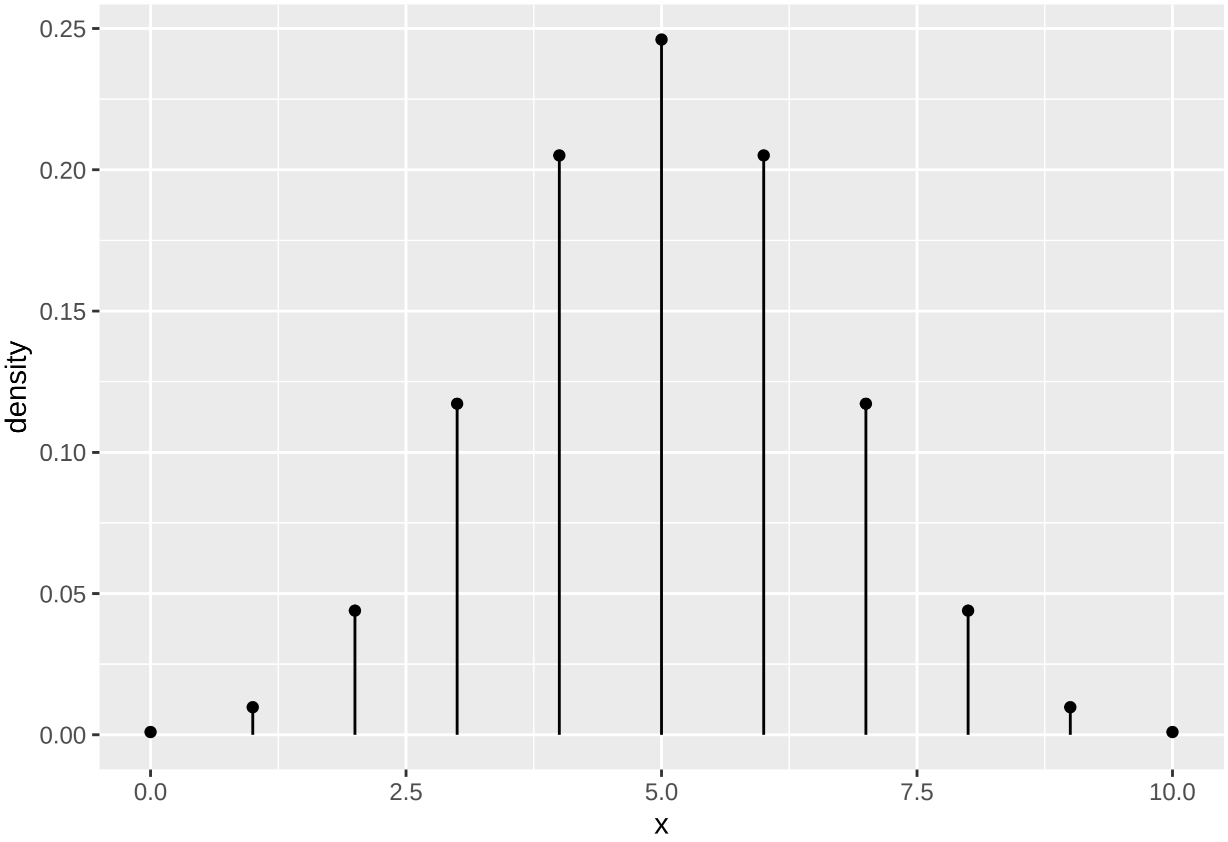
and also the probability of getting at most 5 heads out of ten flips:

```
pbinom(5,size=10,prob=0.5)
```

```
## [1] 0.6230469
```

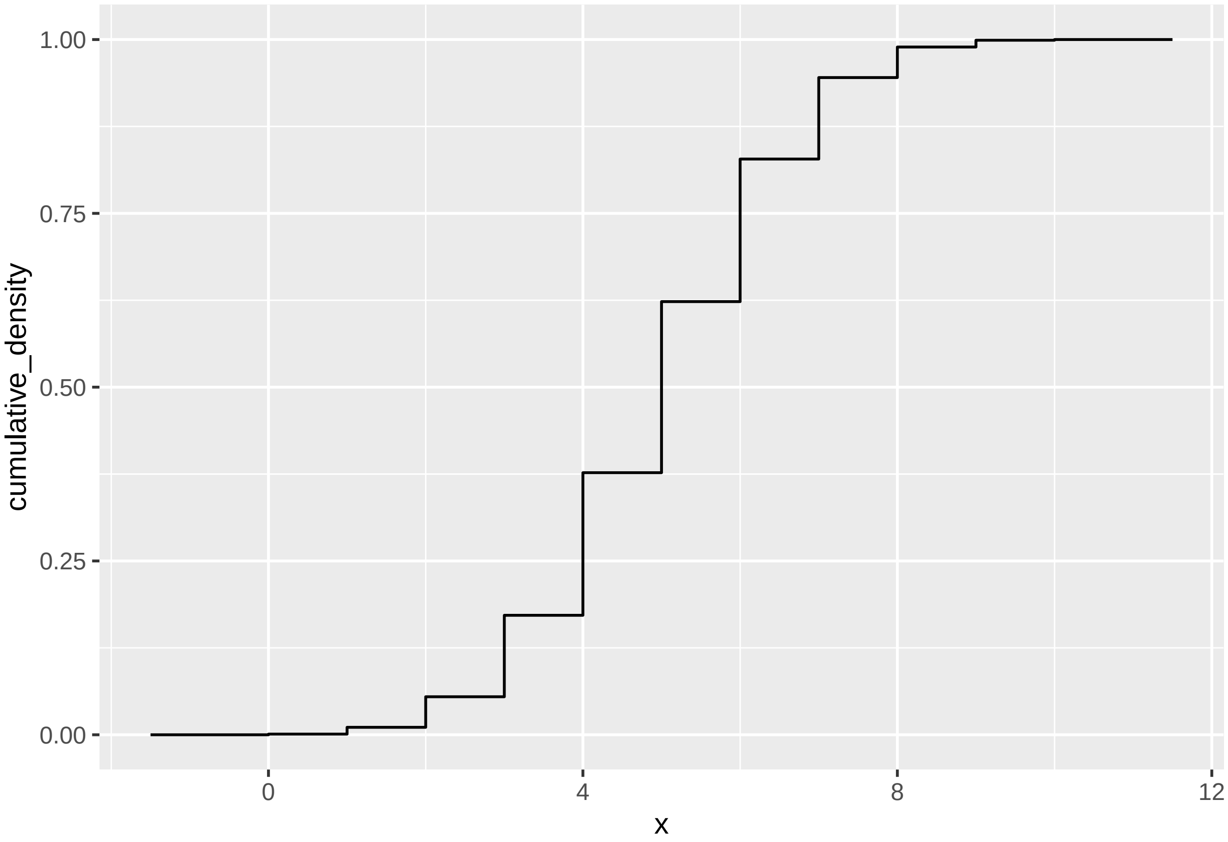
Of course we can visualize the probability mass function for getting x heads out of say 10 flips:

```
gf_dist("binom",params = list(size=10,prob=0.5),kind="density")
```



While the distribution function is a little more complicated to write down, it is easy to visualize:

```
gf_dist("binom",params = list(size=10,prob=0.5),kind="cdf")
```



Exercise Try changing the values for n in the code that is used to create the previous figures.

What we have developed here is a mathematical model, or *probability model* for modeling the random experiment of flipping a two-sided fair coin some set number of times. However, we still have not really addressed exactly what it is we mean by a **probability**. We just sort of asserted that the probability of getting heads after one flip of a (fair) coin is $\frac{1}{2}$.

What we can do is use our mathematical model to simulate the experiment of coin flipping in order to get some sense for what we might mean by a probability. That is, we can simulate coin flipping. This is done in R as follows:

```
rbinom(10,size=1,prob=0.5)
```

```
##  [1] 0 1 0 0 0 0 1 1 1 0
```

Everytime we run this function we get a different number of 0's and 1's.

Let's add up the number of 1's and divide by the number of flips:

```
n <- 10
sum(rbinom(n,size=1,prob=0.5))/n
```

```
## [1] 0.3
```

Notice that the values change each time we flip the coin.

Observe what happens as we take n increasingly larger:

trials	probability
10	0.700000
100	0.470000

trials	probability
1000	0.515000
10000	0.487600
100000	0.500440
1000000	0.498933

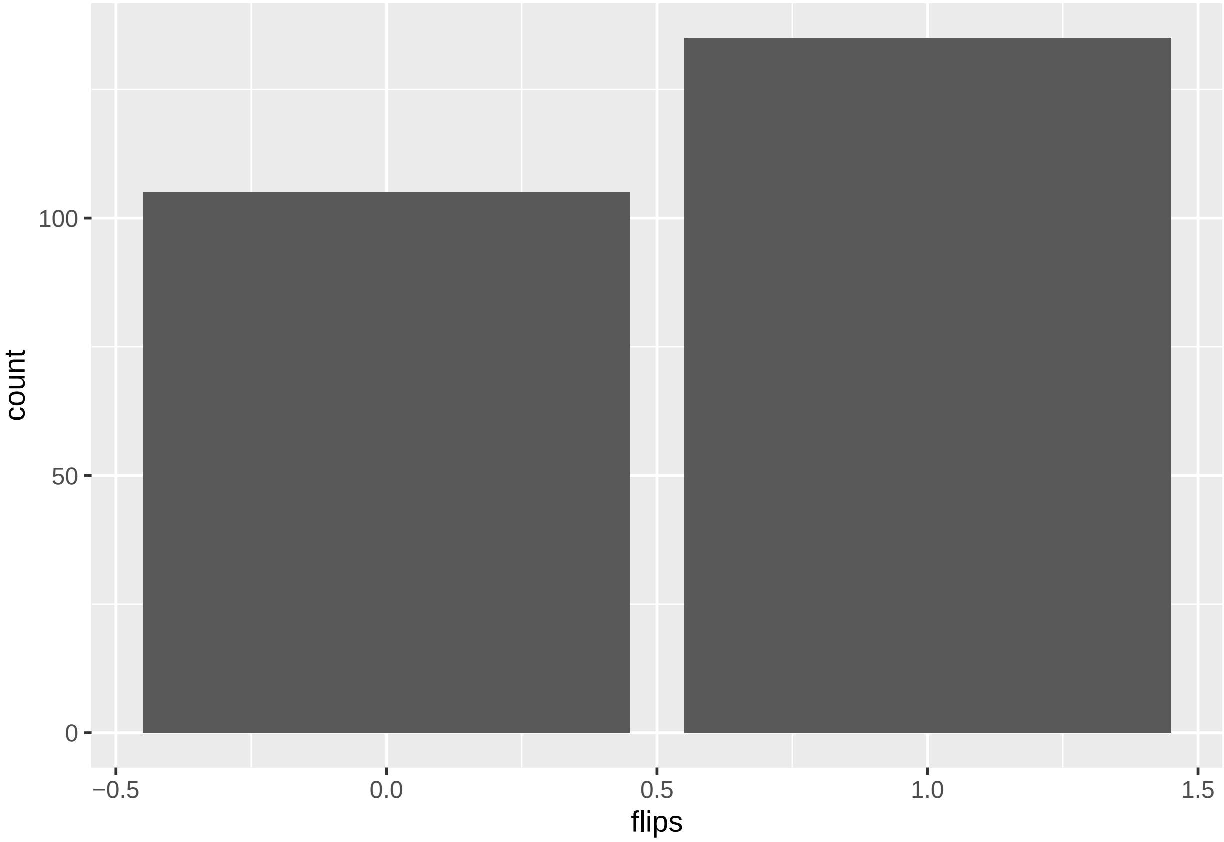
As n gets very large there is little variation of the result from 0.5. What this captures is the idea that in the long run (as $n \to \infty$) we expect that flipping a coin will produce as many heads as it will tails.

Comparison with the Real World

Now an actual experiment:

Everyone flip a quarter 20 times and report the number of heads that you get. We can record the results. (Why is flipping 12 quarters 20 times each equivalent to flipping 1 quarter 240 times?)

```
n_trials <- 240
n_heads <- 135
n_tails <- (n_trials-n_heads)
df <- data.frame(flips=c(rep(0,n_tails),rep(1,n_heads)))
df %>% ggplot(aes(x=flips)) + geom_bar()
```



```
n_heads/n_trials
```

```
## [1] 0.5625
```

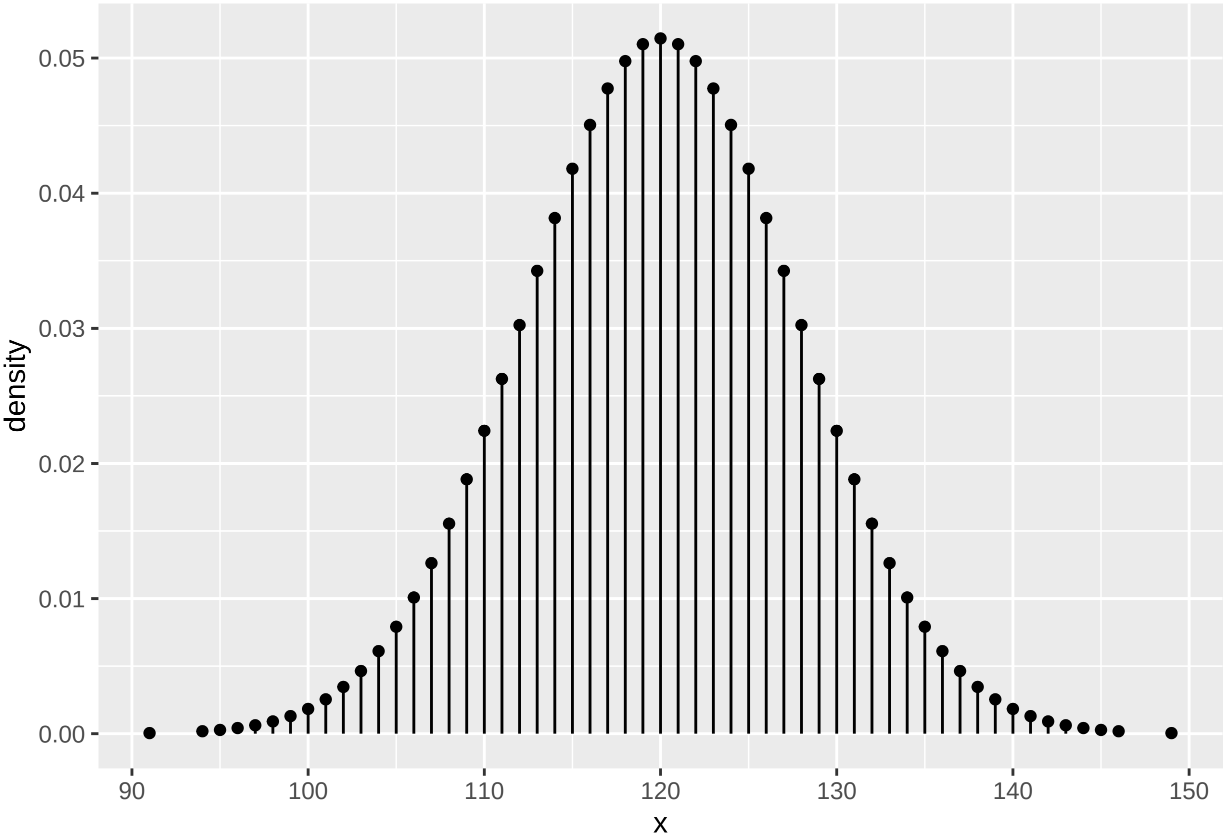
What is the probability of this data (assuming that the coin is fair)?

```
dbinom(n_heads,size=240,prob=0.5)
```

```
## [1] 0.007913272
```

Let's look at the mass function for the case of $n = 240$:

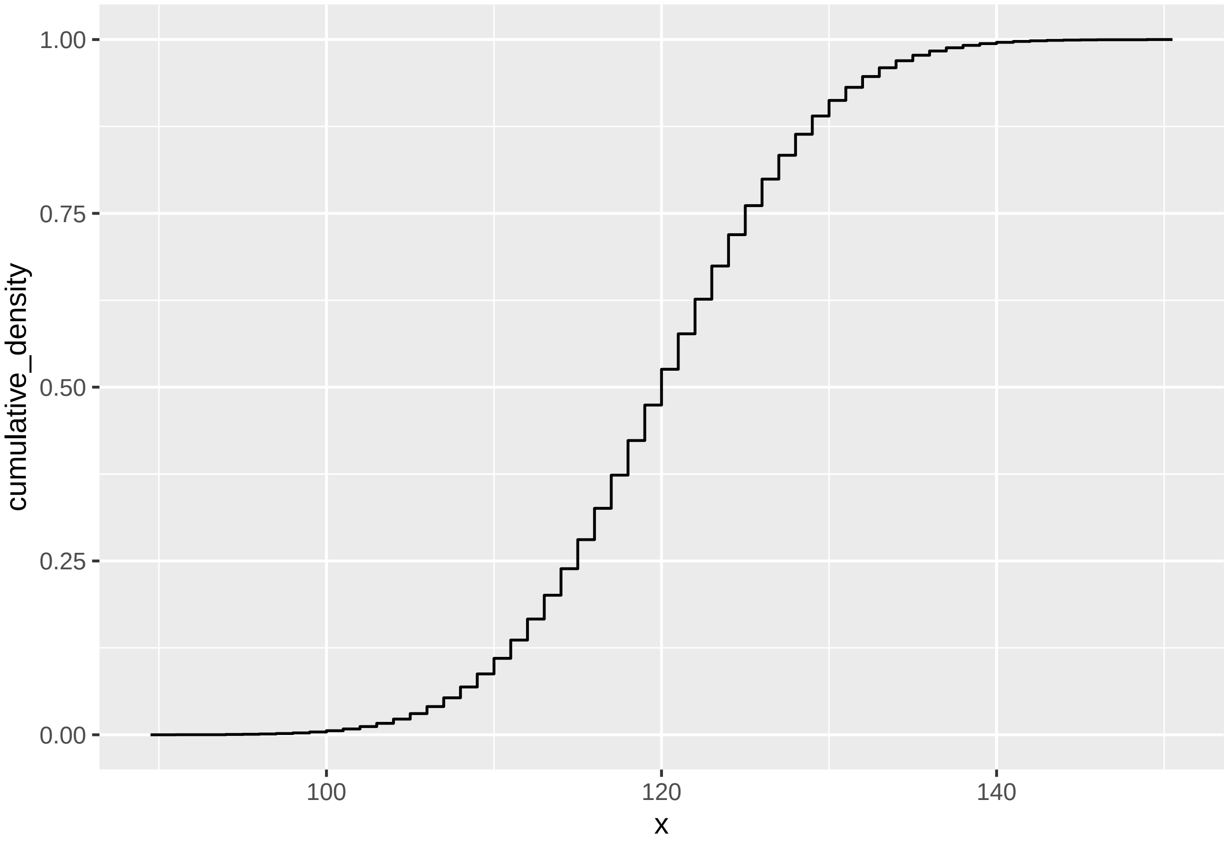
```
gf_dist("binom",params=list(size=n_trials,prob=0.5),kind="density")
```



This shows the (theoretical) **sampling distribution** for the probability of getting x heads out of n trials.

Based on this, is the probability of getting n_{heads} heads out of n_{trials} flips high or low? The distribution function may also be useful to look at:

```
gf_dist("binom",params=list(size=n_trials,prob=0.5),kind="cdf")
```



How come we do not get exaclty half heads and half tails in the real experiment? There are two possibilities:

1) Randomness
2) The coins are not really fair

How do we determine which of these two possibilities is the likely culprit? Well, if the coins are not fair, then the probability of getting heads (in a single flip) is

$p(x) = P(X = x) = q^x \left(1 - q\right)^{1-x}$

where the **parameter** q is some value (between 0 and 1) that is not equal to $\frac{1}{2}$.

Exercise Think about why this might be.

On the other hand, if the coins are fair then we expect that the probability of getting heads after one flip is $\frac{1}{2}$. So assume that the coin is fair and ask "Does the sample data that we have collected support this hypothesis or not?" This is where we transition from probability to (inferential) **statistics**.

From Probability to Statistics

We begin by forming a **hypothesis**, if the coins are fair then we expect that the probability of getting heads after one flip is $\frac{1}{2}$. Does the sample data that we have collected support this hypothesis or not? That is what we will address. Formally, we state a **null hypothesis**:

- H_0 - the coin is fair, *i.e.* $q = \frac{1}{2}$

to which there corresponds a mutually exclusive **alternative hypothesis**

- H_a - the coin is not fair, *i.e.* $q \neq \frac{1}{2}$

What we have to do is to "test" the null-hypothesis. This is done as follows:

1) Assume the null-hypothesis is true.
2) Compute the probability of getting a result that is at least as extreme as your data.
3) Decided on a cut-off probability.
4) If the probability of getting a restult that is at least as extreme as your data is less than the cut-off value, then you conclude that the data provides sufficient evidence to reject the null hypothesis. Otherwise, you do not reject the null hypothesis.

Here we go:

Assuming that the coin(s) is (are) fair, what is the probability of getting a result that is at least as extreme as getting n_{heads} heads out of n_{trials} tosses? This can be calculated in R as follows:

```
(p <- 2*(1 - pbinom(n_heads-1,n_trials,0.5)))
```

```
## [1] 0.06098896
```

Note that we multiply by two because this is a symmetric distribution and we want to conduct a **two-sided** test.

What this says is that under the null hypothesis ($q = \frac{1}{2}$) about 0.06% of the time we expect to get a result that is at least as extreme as getting n_{heads} heads. While this probability is not high, it is also not extremely low. Thus, we fail to reject the null hypothesis.

R will do all of this "hypothesis testing" for us automatically:

```
(b_test<-binom.test(n_heads,n=n_trials,p=0.5,alternative="two.sided"))
```

```
##
##
##
## data:  n_heads out of n_trials
## number of successes = 135, number of trials = 240, p-value = 0.06099
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4972115 0.6262254
## sample estimates:
## probability of success
##                0.5625
```

In addition to the probability value 0.060989 (*i.e.* the p-value), we also get a **confidence interval** (or at least an estimation for one) for the parameter q. In this case, we get a 95% confidence interval. Again from Wikipedia:

> Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 95%. Thus,tThe (95%) confidence interval represents values for the population parameter (q) for which the difference between the parameter and the observed estimate is not statistically significant at the 5% level. A 95% confidence interval does not mean that for a given realized interval there is a 95% probability that the population parameter lies within the interval (*i.e.*, a 95% probability that the interval covers the population parameter).

Further Motivation

Consider the mpg data set cotained in the ggplot2 package:

```
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

This data contains information for various automobiles. Among the data are values for the class and highway mileage for each vehicle. Suppose we want to know if compact cars really get better highway driving gas mileage than midsize cars. First, we compute the **mean** and **standard deviation** (and also **standard error**) for the highway gas mileage per class:

```
mpg_stats <- mpg %>%
  group_by(class) %>%
  summarise(hwy_mean=mean(hwy,na.rm=T),hwy_sd=sd(hwy,na.rm=T),hwy_se=plotrix::std.error(hwy,na.rm=T))
```

Here is the result:

```
mpg_stats
```

```
## # A tibble: 7 x 4
##   class      hwy_mean hwy_sd hwy_se
##   <chr>         <dbl>  <dbl>  <dbl>
## 1 2seater        24.8   1.30  0.583
## 2 compact        28.3   3.78  0.552
## 3 midsize        27.3   2.14  0.334
## 4 minivan        22.4   2.06  0.622
## 5 pickup         16.9   2.27  0.396
## 6 subcompact     28.1   5.38  0.909
## 7 suv            18.1   2.98  0.378
```

Is the difference in the mean highway gas mileage between compact and midsize cars significant? These leads to a hypothesis that we can test.

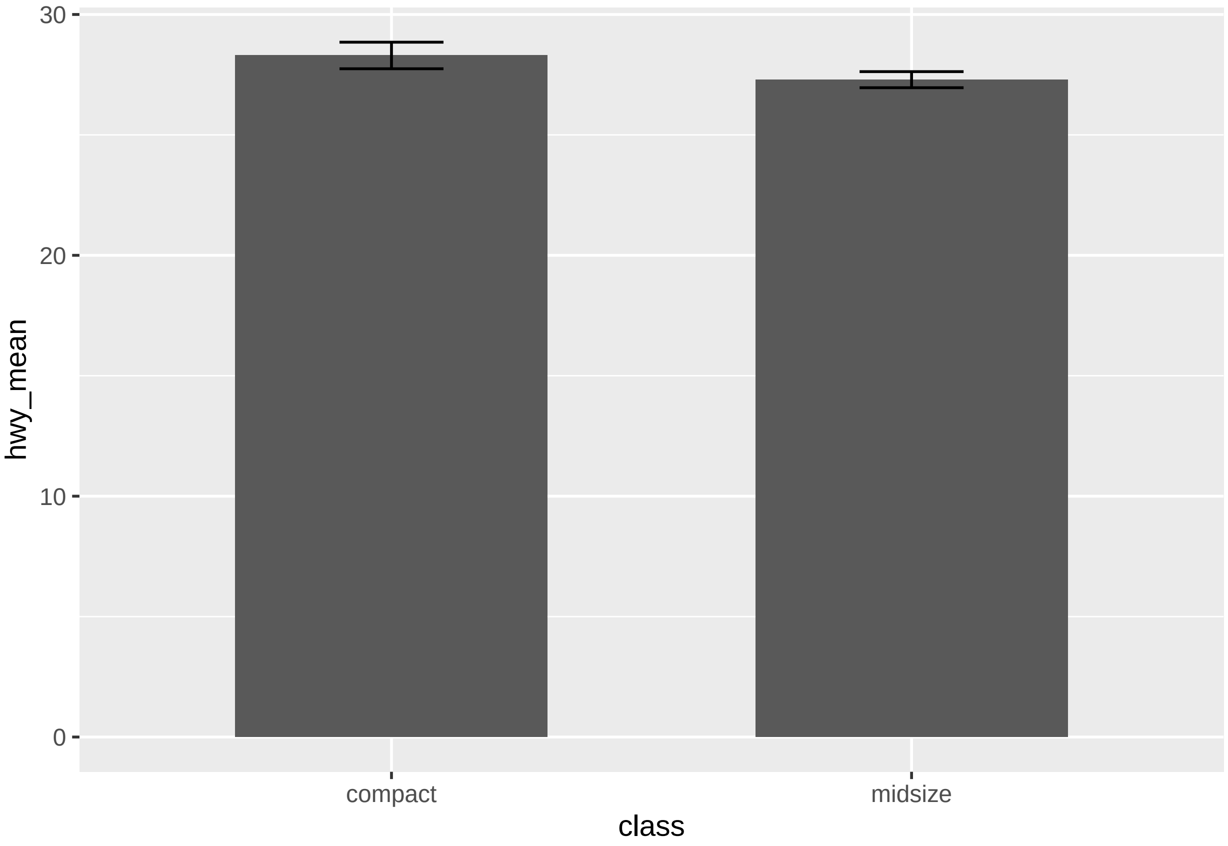
Exercise State a null hypothesis for this scenario.

The following code carries out the hypothesis test. Later we will examine the details of this type of test.

```
class_1 <- "midsize"
class_2 <- "compact"
```

We first look at the mean highway gas mileage in each class and also include the standard error:

```
mpg_stats %>% filter(class %in% c(class_1,class_2)) %>%
  ggplot(aes(x=class,y=hwy_mean)) +
  geom_bar(stat = "identity",width=0.6) +
  geom_errorbar(mapping=aes(ymin=hwy_mean-hwy_se,ymax=hwy_mean+hwy_se),width=0.2)
```



Is this difference significant? Let's see the test:

```
df_testing <- mpg %>% filter(class %in% c(class_1,class_2)) %>%
  select(class,hwy)
with(df_testing,t.test(hwy~class,var.equal=FALSE,alternative="two.sided",mu=0))
```

```
##
##  Welch Two Sample t-test
##
## data:  hwy by class
## t = 1.5593, df = 74.36, p-value = 0.1232
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2791517  2.2895305
## sample estimates:
## mean in group compact mean in group midsize
##              28.29787              27.29268
```

Based on the results of the test, the data does not provide enough justification for rejecting the null hypothesis that there is no difference in the population mean for highway gas mileage between compact and midsize vehicles.

What we need to understand first in order to grasp the details of the previous test are answers to the following questions:

1) What are the relevant **parameters** for stating the null hypothesis "numerically"?
2) What is an appropriate **random variable(s)** and **probability distribution** to use in order to compute the probability of getting a result that is at least as extreme as the data under the null hypothesis? That is, what is the relevant probability model and how do we use it to compute the *p*-value?

In order to answer these questions and to go even further in order to carry out hypothesis tests for the broadest possible range of applications, it is worth some effort to understand probability models in a general abstract mathematical way. This will save us time in the long run since then statistics will not be “cookbook” but very natural. Thus, theory becomes very practical. Motivated as such, we now begin the course in earnest.