Business Statistics 41000

Lecture 3

Mladen Kolar

Probability and Decisions

Suppose you are deciding whether or not to target a customer with a promotion.

It will cost you \$0.80 (eighty cents) to do the promotion and on average a customer spends \$40 **if** they respond to the promotion.

Should you do it ???

Suppose you are deciding whether or not to target a customer with a promotion.

It will cost you $0.80 (eighty cents) to do the promotion and on average a customer spends $40 **if** they respond to the promotion.

Should you do it ???

Well, it depends on how likely it is that the customer will respond!!

If they respond, you get $40 - 0.80 = 39.20$.

If they do not respond, you get -0.80
(you lose eighty cents).

To decide you need **the probability that they will respond!!**

This is where you need your Predictive Analytics team.
They have developed a model which gives you

the probability the customer responds
given
information about the customer.

To decide you need **the probability that they will respond!!**

This is where you need your Predictive Analytics team.
They have developed a model which gives you

the probability the customer responds
given
information about the customer.

What information about a customer do they use?

- nTab: number of past orders.
- moCbook: months since last order.
- iRecMer1 : 1/months since last order in merchandise category 1.
- llDol: log of the dollar value of past purchases

The PA team has a model that gives us the probability a customer will respond given the values of these 4 variables.

You actually have 5,000 prospective customers you could target.

```
##   ProbRespond nTab moCbook iRecMer1   llDol
## 1    0.022521    2      50 0.315637 -2.3026
## 2    0.020908    1      50 0.019608  3.9524
## 3    0.029421    5      50 0.378419 -2.3026
## 4    0.020201    1      50 0.283321 -2.3026
## 5    0.016380    3      50 0.095591 -2.3026
## 6    0.012221    0      50 0.019608 -2.3026
```

How do they come up with something like this ?!

Suppose for a particular customer, the probability of a response is 0.05.

Should we do the promotion?

If we do the promotion what can happen?

How likely are these outcomes?

Let M denote the amount of money you get if you do the promotion.

M is a number you are not sure about,
  M **is a random variable**!

Specifying a random variable

To work with a random variable you specify its distribution.

- List all the numbers it could turn out to be.
- Assign a probability to each possible **outcome**.

The probabilities have to be between 0 and 1 and add up to 1.

The distribution of M:

m	$P(M = m)$
-0.80	0.95
39.20	0.05

We use m to denote a possible **outcome** for M.

Notation:
$P(M = m)$ is "the probability M turns out to be m".
We will also write $p(m) = P(M = m)$.

Okay, now what is your decision?

You can get \$0 if you don’t target, or the random outcome of M if you do.

We also say, you get a "draw" from the distribution of M.

m	$P(M=m)$
-0.80	0.95
39.20	0.05

One way to decide is to compute the **expected value of** M.

You compute the probability weighted average of the outcomes:

$$E(M) = 0.95 \cdot (-0.8) + 0.05 \cdot 39.20 = 1.2.$$

$E(M)$ is "the expected value of M".

The probability weighted average of M is 1.2 which is bigger than 0,
so you go for it!!

Does the Targeting Work ?

You have 5,000 customers on your file you could target.

According to your predictive model and decision analysis, it is optimal to target anyone with a probability of responding greater than 0.02.

Let's suppose you are a little skeptical about the predictive model and you go ahead and target all the customers (all 5,000 get the promotion).

`response` tells us whether or not they responded to the promotion
`dotarget` indicates whether or not the probability of a response from the model is greater than the cutoff of 0.02.

```
##   dotarget response ProbRespond
## 1      YES    NOBUY    0.022521
## 2      YES    NOBUY    0.020908
## 3      YES    NOBUY    0.029421
## 4      YES    NOBUY    0.020201
## 5       NO    NOBUY    0.016380
## 6       NO    NOBUY    0.012221
```

```
## response
## NOBUY   BUY
##  4888   112
```

```
## dotarget
##   NO  YES
## 3767 1233
```

This two-way table displays and relates response and dotarget.

```
##          response
## dotarget NOBUY  BUY
##      NO   3730   37
##      YES  1158   75
```

Fraction of $p > 0.02$ customers who bought: $75/(1158+75) = 0.06$
Fraction of $p \leq 0.02$ customers who bought: $37/(3730+37) = 0.01$

How much money did you get by targeting everyone?

$$(37+75) \cdot 40 - 5000 \cdot (0.8) = 480$$

How much money would you have made if you had used the probabilities and cutoff of 0.02?

$$75 \cdot 40 - (1158+75) \cdot (0.8) = 2013.6$$

Big Picture

Often, to make decisions quantitatively, we need probabilities.

More specifically, we need **conditional probabilities**.

Given a set of information, what is the probability something will happen.

Example:

- What is the probability you get cancer, given you smoke?
- What is the probability you get cancer, given you don’t smoke?

To the extent that these are different, smoking and cancer are related and this might affect your decision about whether or not to smoke.

How do we come up with probabilities?

In some simple cases we can use our judgment.

Example:
You are about to toss a coin. $X = 1$ if it comes up heads, 0, otherwise.
What is the probability X turns out to be 1?

Often we look at the past and collect data on variables of interest and use statistical methods to help is figure out the model.

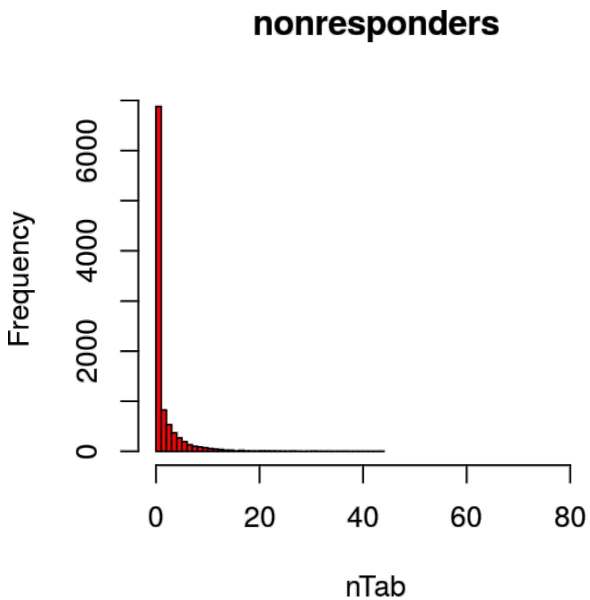
Ideally, we use **both** judgment and **stats** to come up with our models.

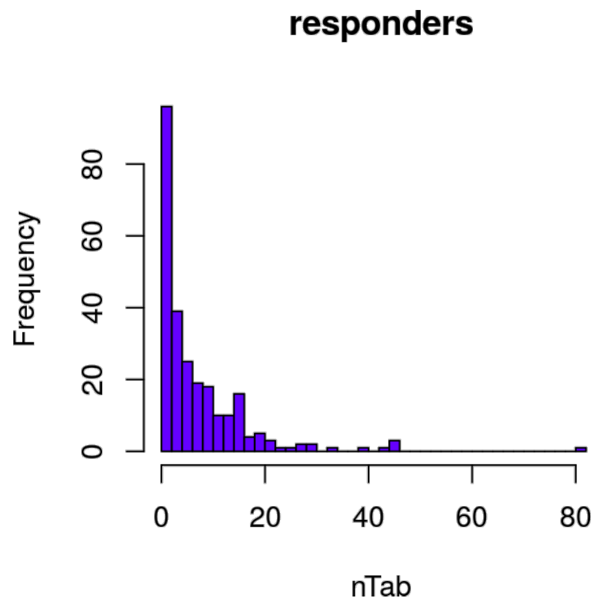
- What is the probability the Hawks win the Stanley Cup?
- What is the probability the Fed raises rates at the next meeting?

Burning question

How do we come up with a predictive model that gives the probability of a response given values for `nTab`, `moCbook`, `iRecMer1`, and `llDol`?

At left is the histogram of `nTab` values for non-responders and at right is the histogram of `nTab` values for responders using some data from the past.





We can see that `nTab` is related to response but how do you quantify the relationship between the probability for response and all 4 variables?

Discrete Random Variables

A random variable is *a number we're not sure about*.

Its **distribution** describes what we think it might turn out to be.

For a discrete random variable, we specify the distribution by:

- Listing all the possible numbers it can turn out to be.
- Assigning a probability to each possible outcome.
- Each probability is between 0 and 1.
- The probabilities add up to 1.

Note: "discrete" refers to the situation where can make the list.
Later we will look at continuous random variable where such a list is not practical.

Example: Tossing two coins

Suppose we are about to toss two coins.

Let X denote the number of heads.

Then the distribution of X might be given by

x	$P(X = x)$
0	0.25
1	0.5
2	0.25

Example: At-bat outcomes

We can associate a 0, 1, 2, 3, or 4 to the outcomes of a baseball at-bat.

Event	x	$P(X = x)$
Out	0	0.820
Base hit	1	0.115
Double	2	0.033
Triple	3	0.008
Home run	4	0.024

What is $P(X > 2)$?
What is $P(X \geq 2)$?

The Bernoulli Distribution

A very common situation is that we are wondering whether something will happen or not.

Examples:

- heads or tails
- respond or don't respond
- pearson has or does not have a disease
- a tennis player serves an ace or does not

It turns out to be very convenient to code up one possibility as a 0, and the other possibility as a 1.

The gives us the *Bernoulli* distribution.

$X \sim \text{Bernoulli}(p)$ means:

x	$P(X = x)$
0	$1 - p$
1	p

The value 1 is often called a "success."

Example:

You are about to toss a coin.

Let X be 1 if it comes up Heads and 0 if tails.

$$X \sim \text{Bernoulli}(0.5)$$

Example:

You are about to target a customer.

Let R be 1 if the respond (buy) and 0 otherwise.

For a particular customer we might have:

$$R \sim \text{Bernoulli}(0.05)$$

The discrete Uniform distribution

$X \sim \text{Discrete Uniform}$ means that X is a discrete random variable taking on a finite number of values with equal probabilities.

If there are N outcomes, the probabilities are all $1/N$.

Example: Suppose we roll a fair six-sided die. Let X denote the outcome.

x	$P(X = x)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$

What is the probability that $P(2 < X < 5)$?

Probabilities of Subsets of Outcomes

To compute the probability that any one of a group of outcomes occurs we sum up their probabilities.

$$P(a < X < b) = \sum_{a<x<b} P(X = x)$$

Example: Rolling a die

$$P(2 < X < 5) = P(X = 3) + P(x = 4) = \frac{1}{3}$$

$$P(2 \leq X < 5) = P(X = 2) + P(X = 3) + P(X = 4) = \frac{1}{2}$$

$$P(X \neq 6) = 1 - P(X = 6) = 5/6$$

Probabilities of Subsets of Outcomes

Here are two helpful reminders

OR means ADD

$$P(X = a \text{ OR } X = b) = P(X = a) + P(X = b)$$

- As long as two events cannot **both** happen, the probability of either is the sum of the probabilities.

NOT means ONE MINUS

$$P(X \neq a) = 1 - P(X = a)$$

- The probability that something does **NOT** happen is one minus the probability that it does.

Mean and Variance

The Mean or Expected Value is defined as (for a discrete X):

$$E(X) = \sum_{j=1}^{m} P(X = x_j) \cdot x_j$$

We weight each possible value by how likely they are.

This provides us with a measure of centrality of the distribution, a "good" prediction for X.

The Variance is defined as (for a discrete X):

$$\mathrm{Var}(X) = \sum_{j=1}^{m} P(X = x_j) \cdot [x_j - E(X)]^2$$

Weighted average of squared prediction errors.

This is a measure of **spread** of a distribution.

A more intuitive way to understand the spread of a distribution is to look at the **standard deviation**:

$$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$$

Expected Value and Variance of a Bernoulli

$X \sim \text{Bernoulli}(p)$ means:

x	$P(X = x)$
0	$1 - p$
1	p

$$E(X) = (1 - p) \cdot 0 + p \cdot 1 = p$$

$$\begin{aligned} \text{Var}(X) &= \sum_{j=1}^{m} P(X = x_j) \cdot [x_j - E(X)]^2 \\ &= (1 - p) \cdot (0 - p)^2 + p \cdot (1 - p)^2 \\ &= p \cdot (1 - p) \end{aligned}$$

The standard deviation is therefore $\sqrt{p \cdot (1 - p)}$

Question: For which value of p is the variance the largest?

Example: medical expenditures

We can *bin* household medical expenditures and think of the distribution over medical expenses.

Event	x	$P(X = x) \times 10,000$
Between 0 and \$100	50	$2,600$
Between \$100 and \$1000	550	$3,300$
Between \$1000 and \$5000	$3K$	$2,500$
Between \$5000 and \$10,000	$7.5K$	800
Between \$10,000 and \$20,000	$15K$	500
Between \$20,000 and \$30,000	$25K$	200
Between \$30,000 and \$40,000	$35K$	60
Between \$40,000 and \$50,000	$45K$	30
Between \$50,000 and \$100,000	$75K$	7
Between \$100,000 and \$600,000	$350K$	3

Example: medical expenditures (con't)

For our medical costs random variable we can calculate the mean and variance as

$$
\begin{aligned}
E(X) &= \sum_{j=1}^{m} x_j \cdot P(X = x_j) \\
&= 50(0.26) + 550(0.33) + 3000(0.25) + 7500(0.08) \\
&\quad + 15000(0.05) + 25000(0.02) + 35000(0.006) \\
&\quad + 45000(0.003) + 75000(0.0007) + 350000(0.0003) \\
&= 3297.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(X) &= \sum_{j=1}^{m} [x_j - E(x)]^2 \cdot P(X = x_j) \\
&= (50 - 3297)^2(0.26) + (550 - 3297)^2(0.33) + (3000 - 3297)^2(0.25) + (7500 - 3297)^2(0.08) \\
&\quad + (15000 - 3297)^2(0.05) + (25000 - 3297)^2(0.02) + (35000 - 3297)^2(0.006) \\
&\quad + (45000 - 3297)^2(0.003) + (75000 - 3297)^2(0.0007) + (350000 - 3297)^2(0.0003) \\
&= 73842766
\end{aligned}
$$

The standard deviation is 8593.18.

Sample Mean vs. the Expected Value

The Sample Mean

- **variable**: in a dataset, it is the observed set of values
- **sample mean**: of a variable in our data is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- It is the average of the observed values in the data set.

The Expected Value

- **random variable**: a mathematical model for an uncertain quantity
- **expected value**: of a random variable is $E(X) = \sum_{j=1}^{m} P(X = x_j) \cdot x_j$
- Average of the possible values taken by a r.v. weighted by their probabilities.

The sample mean, sample variance, and sample standard deviation of a set of numbers are sample statistics computed from observed data.

The mean, variance, and standard deviation of a random variable are properties of its probability distribution which is a mathematical model of uncertainty.

They do share a lot of the same properties.

The distinction between them is subtle but important for later on in the course!

Expected value of a function

The expected value of a function $g(X)$ is defined as:

$$E\left(g(X)\right) = \sum_{j=1}^{m} g(x_j) \cdot P(X = x_j).$$

So far we have been using the identity function $g(x) = x$.

Example: Deal or No Deal

George has cases worth \$5, \$400, \$10,000, and \$1,000,000 remaining. There are 4 outcomes and each is equally likely. The banker’s offer is \$189,000.

Let W be the prize a game show contestant ends up with.

$$E(W) = (5 + 400 + 10000 + 1000000)/4 = 252601.25$$

Is the banker's offer a good or bad deal?

Example: Deal or No Deal

George has cases worth \$5, \$400, \$10,000, and \$1,000,000 remaining. There are 4 outcomes and each is equally likely. The banker’s offer is \$189,000.

Let W be the prize a game show contestant ends up with.

$$E(W) = (5 + 400 + 10000 + 1000000)/4 = 252601.25$$

Is the banker's offer a good or bad deal?

Economists believe in *diminishing marginal utility of income*. The more wealth you have the less utility you get from each additional \$1.

This is often modeled with a utility function over wealth.
For example, $U(W) = \sqrt{W}$.

$$E\left[U(W)\right] = \left(\sqrt{5} + \sqrt{400} + \sqrt{10000} + \sqrt{1000000}\right)/4 = 280.56$$

Compare this to the utility of the banker’s offer: $\sqrt{189000} = 434.74$.

Aside: computational shortcut for variance

A convenient way to compute the variance is via the identity

$$\mathrm{Var}(\mathrm{X}) = E\left[(X - E(X))^2\right] = E\left(X^2\right) - (E(X))^2$$

The variance is the "expected value of the square minus the square of the expected value."

Properties of expectation and variance

Here are some rules that make calculating expectations and variances easier.

Here, X, Y are random variables and a, b, c are any numbers.

- $E\left(a + b \cdot X\right) = a + b \cdot E(X)$
- $\mathrm{Var}\left(a + b \cdot X\right) = b^2 \cdot \mathrm{Var}(X)$
- $E\left(a \cdot X + b \cdot Y + c\right) = a \cdot E(X) + b \cdot E(Y) + c$

The facts are not hard to show directly from the definition
(but we will not do so here).

Conditional, Marginal, and Joint Distributions

What happens when there are two (or more) variables that we are uncertain about?

How do we describe them probabilistically?

We want to use probability to understand how two (or more) variables are related.

In this section, we extend the results above to more than one variable.

How are my sales impacted by the overall economy?

Let E denote the performance of the economy next quarter.

For simplicity, say $E = 1$ if the economy is expanding and $E = 0$ if the economy is contracting.

Assume $P(E = 1) = 0.7$, that is, $E \sim \text{Bernoulli}(0.7)$

Let S denote my sales next quarter.

Let us suppose the following probability statements:

s	$P(S = s \mid E = 1)$	$P(S = s \mid E = 0)$
1	0.05	0.20
2	0.20	0.30
3	0.50	0.30
4	0.25	0.20

These are called **Conditional Distributions**

s	$P(S = s \mid E = 1)$	$P(S = s \mid E = 0)$
1	0.05	0.20
2	0.20	0.30
3	0.50	0.30
4	0.25	0.20

The left distirbution is the conditional distribution of S given $E = 1$.

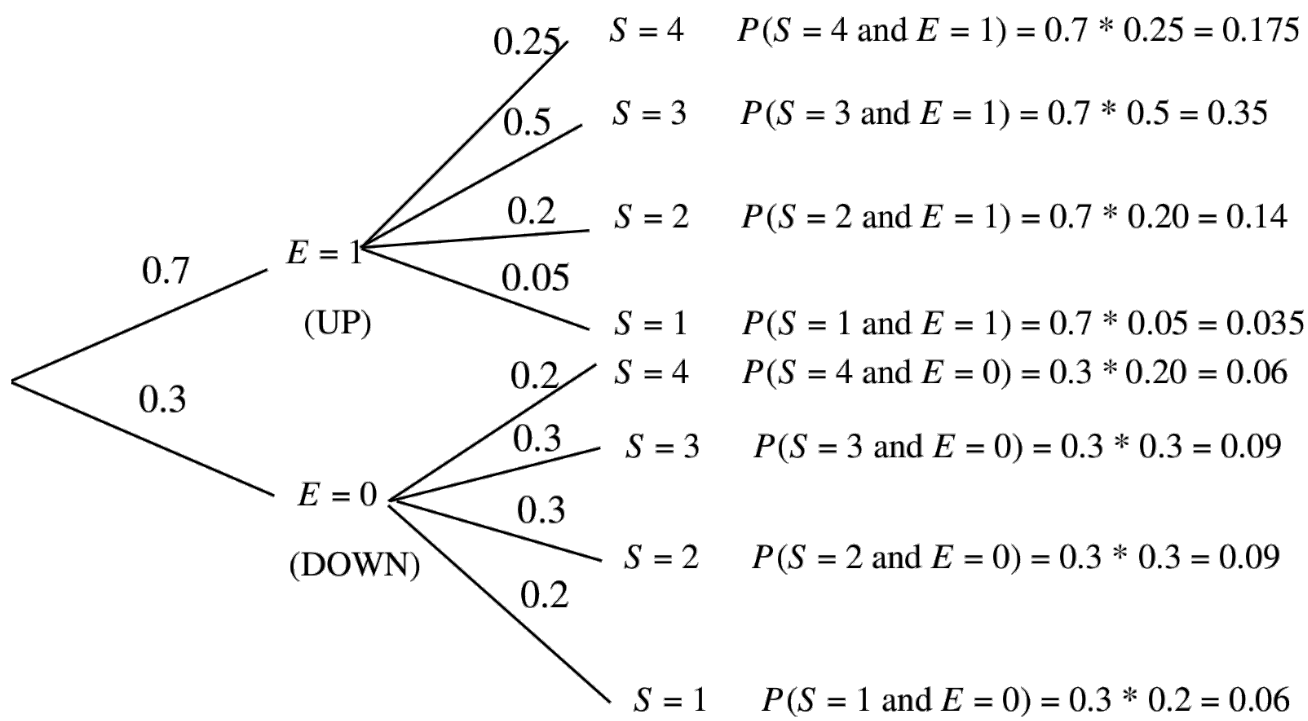
The right distirbution is the conditional distribution of S given $E = 0$.

The probability of Sales equals 4 $(S = 4)$ *given (or conditional on)* the economy is growing $(E = 1)$ is 0.25.

The conditional distributions tell us about about what can happen to S for a given value of E, but what about S and E **jointly**?

$$P(S=4 \text{ and } E=1) = P(E=1) \cdot P(S=4 \mid E=1) = 0.7 \cdot 0.25 = 0.175$$

70% of the time the economy goes up, and 1/4 of those times sales equals 4. 25% of 70% is 17.5%



We can specify the distribution of the pair of random variables (S, E) by listing all possible pairs and the corresponding probability.

(e, s)	$P(E = e \text{ and } S = s)$
$(1, 4)$	0.175
$(1, 3)$	0.350
$(1, 2)$	0.140
$(1, 1)$	0.035
$(0, 4)$	0.060
$(0, 3)$	0.090
$(0, 2)$	0.090
$(0, 1)$	0.060

When there are only two discrete random variables, we can also display the joint distribution of E and S in a different table.

		S			
		1	2	3	4
E	0	0.060	0.090	0.090	0.060
	1	0.035	0.140	0.350	0.175

What is the probability that $P(E = 1 \text{ and } S = 4)$?

If we don't know anything about E, what is $P(S = 4)$?

Marginal distribution

What is the probability of S if we know nothing about E?

What is the probability of E if we know nothing about S?

		S 1	2	3	4	$P(E=e)$
E	0	0.060	0.090	0.090	0.060	0.3
	1	0.035	0.140	0.350	0.175	0.7
	$P(S=s)$	0.095	0.230	0.440	0.235	1

Summary so far

The **joint probability** that Y turns out to be y **and** that X turns out to be x is denoted by

$$P(Y = y, X = x) = P(Y = y \text{ and } X = x)$$

The **conditional probability** that Y turns out to be y **given you know that** $X = x$ is denoted by

$$P(Y = y \mid X = x)$$

The **marginal probabilities** of $Y = y$ and $X = x$ are denoted as $P(Y = y)$ and $P(X = x)$, respectively.

Two Important Relationships

Relationship between Joint and Conditional

$$\begin{aligned} P(Y = y, X = x) &= P(X = x) \cdot P(Y = y \mid X = x) \\ &= P(Y = y) \cdot P(X = x \mid Y = y) \end{aligned}$$

Relationship between Joint and Marginal

$$P(X = x) = \sum_y P(X = x, Y = y)$$

$$P(Y = y) = \sum_x P(X = x, Y = y)$$

Conditionals from Joints

We derived the joint distribution of (E, S) by first considering the marginal of E and then thinking about the conditional distribution of $S \mid E$.

An alternative approach is to start with a joint distribution $P(Y = y, X = x)$ and the marginal $P(X = x)$ and then obtain the conditional distribution.

$$P(Y = y, X = x) = P(X = x) \cdot P(Y = y \mid X = x)$$

$$\Longrightarrow$$

$$P(Y = y \mid X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

Example: given that the economy is up $(E = 1)$, what is the probability that sales is 4?

		S				
		1	2	3	4	$P(E = e)$
E	0	0.060	0.090	0.090	0.060	0.3
	1	0.035	0.140	0.350	0.175	0.7
	$P(S = s)$	0.095	0.230	0.440	0.235	1

$$P(S = 4 \mid E = 1) = \frac{P(S = 4, E = 1)}{P(E = 1)} = \frac{0.175}{0.7} = 0.25$$

Example: given that sales is $(S = 4)$, what is the probability that the economy is up?

		S				
		1	2	3	4	$P(E = e)$
E	0	0.060	0.090	0.090	0.060	0.3
	1	0.035	0.140	0.350	0.175	0.7
	$P(S = s)$	0.095	0.230	0.440	0.235	1

$$P(E = 1 \mid S = 4) = \frac{P(S = 4, E = 1)}{P(S = 4)} = \frac{0.175}{0.235} = 0.745$$

In general, you can compute the joint from marginals and conditionals and the other way around.

How you think about stuff depends on what's easiest or what you know, or what you care about.

Example: suppose you toss two fair coins: X is the first, Y is the second.

What is $P(X = 1 \text{ and } Y = 1) = P(\text{two heads})$?

- There are 4 possible outcomes for the two coins and each is equally likely so it is $\frac{1}{4}$.
- $P(X = 1 \text{ and } Y = 1) = P(X = 1) \cdot P(Y = 1|X = 1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

Bayes Theorem

Bayes Theorem

In many situations, you will know

- one conditional distribution $P(Y = y \mid X = x)$ and
- the marginal distribution $P(X = x)$,

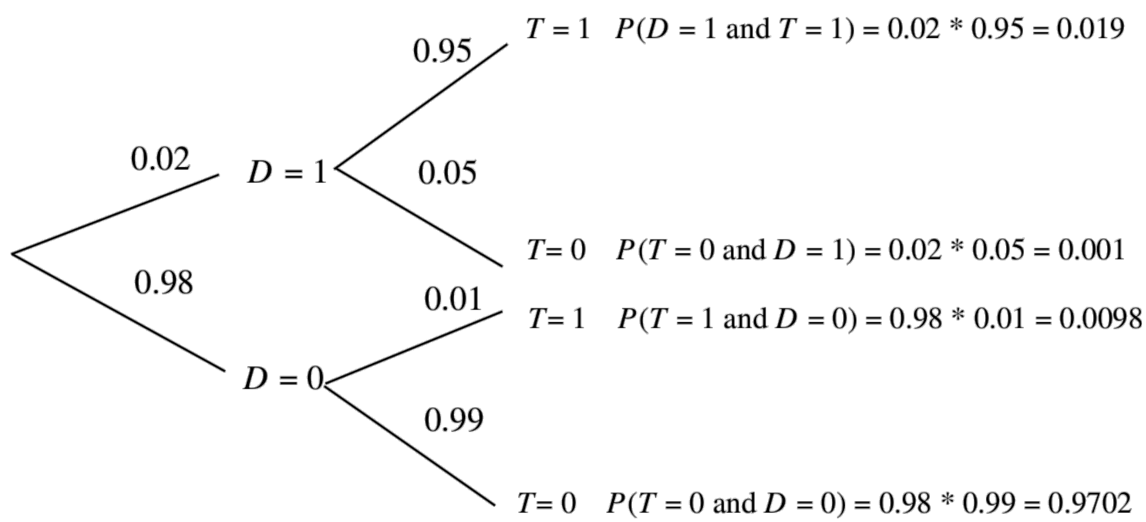
but you are really interested in the other conditional distribution $P(X = x \mid Y = y)$.

$$
\begin{aligned}
P(X = x \mid Y = y) = \frac{P(Y = y, X = x)}{P(Y = y)} &= \frac{P(Y = y, X = x)}{\sum_x P(Y = y, X = x)} \\
&= \frac{P(X = x)P(Y = y \mid X = x)}{\sum_x P(X = x)P(Y = y \mid X = x)}
\end{aligned}
$$

Example: Testing for a Disease

Let $D = 1$ indicate you have a certain (rare) disease.
Let $T = 1$ indicate that you tested positive for it.

Suppose we know the marginal of D and conditional of T given D.



We start with info about D and $T \mid D$. But if you are the patient who tests positive for a disease you care about $P(D = 1 \mid T = 1)$!

Given that you have tested positive, what is the probability that you have the disease?

		D = 0	D = 1
T	0	0.9702	0.001
	1	0.0098	0.019

$$P(D = 1 \mid T = 1) = \frac{P(D = 1, T = 1)}{P(T = 1)} = \frac{.019}{(.019 + .0098)} = 0.66$$

Using Bayes Theorem

$$P(D = 1 \mid T = 1) = \frac{P(T = 1 \mid D = 1)P(D = 1)}{P(T = 1 \mid D = 1)P(D = 1) + P(T = 1 \mid D = 0)P(D = 0)}$$

Try to think about this intuitively.
Imagine you are about to test 100,000 people.

- we assume that about 2,000 of those have the disease
- we also expect 1% of the disease-free people to test positive, that is, 980, and
- 95% of the sick people to test positive, that is 1,900.

So, we expect a total of 2,880 positive tests.

Choose one of the 2,880 people at random. What is the probability that he/she has the disease?

$$P(D = 1 \mid T = 1) = 1900/2880 = 0.66$$

The same !!

Several variables

As we have seen in looking at data, we often want to think about more than two variables at a time.

We can extend the approach we used with two variables.

Suppose we have three random variables (Y_1, Y_2, Y_3).

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = P(Y_3 = y_3 \mid Y_2 = y_2, Y_1 = y_1)P(Y_2 = y_2 \mid Y_1 = y_1)P(Y_1 = y_1)$$

Using more succinct notation:

$$p(y_1, y_2, y_3) = p(y_3 \mid y_2, y_1)p(y_2 \mid y_1)p(y_1)$$

The joint distribution of all three variables can be broken down into the marginal and conditionals distributions.

This is important because it allows us to extend our results to n random variables.

$$p\left(y_1, y_2, \ldots, y_n\right) = p\left(y_n \mid y_{n-1}, y_{n-2}, \ldots, y_2, y_1\right) \ldots p\left(y_3 \mid y_2, y_1\right) p\left(y_2 \mid y_1\right) p\left(y_1\right)$$

Example: Sampling without Replacement

Imagine grabbing Skittles out of a bag one at a time.
You hate the yellow ones and are upset to get three yellows in a row.
What is the probability of this event?

Assume each bag has 80 Skittles, 15 of which are yellow.

- Let A_j denote that the jth draw was yellow.
- Let $B =$ "draw three yellow Skittles in a row."

$P(B)$ can be expressed as

$$P(B) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_2, A_1) = \left(\frac{15}{80}\right)\left(\frac{14}{79}\right)\left(\frac{13}{78}\right).$$

Independence

Given a bunch of random variables, we say they are independent of each other if the conditional distribution of any one of them does not depend on anything you might observe for any of the others.

Example: Suppose I am about to toss 100 coins.

Let Y_i be 1 if the i-th coin is a head and 0 otherwise.

What is $P(Y_3 = 1)$?

What is $P(Y_3 = 1 \mid Y_1 = 1, Y_2 = 0)$?
What is $P(Y_3 = 1 \mid Y_1 = 0, Y_2 = 1)$?
What is $P(Y_3 = 1 \mid Y_1 = 0, Y_2 = 0)$?
What is $P(Y_3 = 1 \mid Y_1 = 1, Y_2 = 1)$?

Given a bunch of random variables, we say they are independent of each other if the conditional distribution of any one of them does not depend on anything you might observe for any of the others.

Example: Suppose I am about to toss 100 coins.

Let Y_i be 1 if the i-th coin is a head and 0 otherwise.

What is $P(Y_3 = 1)$?

What is $P(Y_3 = 1 \mid Y_1 = 1, Y_2 = 0)$?
What is $P(Y_3 = 1 \mid Y_1 = 0, Y_2 = 1)$?
What is $P(Y_3 = 1 \mid Y_1 = 0, Y_2 = 0)$?
What is $P(Y_3 = 1 \mid Y_1 = 1, Y_2 = 1)$?

What is $P(Y_{100} = 1 \mid Y_1 = 1, Y_2 = 1, \ldots, Y_{99} = 1)$? (first 99 are heads)

Given a bunch of random variables, we say they are independent of each other if the conditional distribution of any one of them does not depend on anything you might observe for any of the others.

Example: Suppose I am about to toss 100 coins.

Let Y_i be 1 if the i-th coin is a head and 0 otherwise.

What is $P(Y_3 = 1)$?

What is $P(Y_3 = 1 \mid Y_1 = 1, Y_2 = 0)$?
What is $P(Y_3 = 1 \mid Y_1 = 0, Y_2 = 1)$?
What is $P(Y_3 = 1 \mid Y_1 = 0, Y_2 = 0)$?
What is $P(Y_3 = 1 \mid Y_1 = 1, Y_2 = 1)$?

What is $P(Y_{100} = 1 \mid Y_1 = 1, Y_2 = 1, \ldots, Y_{99} = 1)$? (first 99 are heads)

What is $P(Y_1 = 1, Y_2 = 1, \ldots, Y_{99} = 1, Y_{100} = 1)$? (100 heads in a row)

Independence

The random variable Y is **independent** of the random variable X if the conditional distribution of Y

$$P(Y = y \mid X = x)$$

does not depend on x. In particular, we have

$$P(Y = y \mid X = x) = P(Y = y)$$

Intuitively, knowing that $X = x$ occurred imparts no information to change the probability that $Y = y$ occurred.

Independence

The random variable Y is **independent** of the random variable X if the conditional distribution of Y

$$P(Y = y \mid X = x)$$

does not depend on x. In particular, we have

$$P(Y = y \mid X = x) = P(Y = y)$$

Intuitively, knowing that $X = x$ occurred imparts no information to change the probability that $Y = y$ occurred.

If X and Y are independent then the joint is the product of the marginals

$$P(X = x, Y = y) = P(X = x)P(Y = y \mid X = x) = P(X = x)P(Y = y)$$

This also works "the other way," that is, if the joint is the product of the marginals then they are independent.

Example: Gambler's fallacy

You've been watching the roulette table for the past five spins and noticed it has come up red each and every time.
What is the probability that the next spin will be black?

Example: Gambler's fallacy

You've been watching the roulette table for the past five spins and noticed it has come up red each and every time.
What is the probability that the next spin will be black?

If the roulette wheel is fair and the spins are independent, then the probability should be the same irrespective of the previous rolls, so $P(black) = 18/37$.

Would this logic apply to sunshine after 5 consecutive days of rain?

Example: You are about to manufacture two parts

$X = 1$ if part one fails. $Y = 1$ if part two fails.
The table below gives the joint distribution of X and Y.

		X		
		0	1	$P(Y = y)$
Y	0	0.72	0.08	0.8
	1	0.18	0.02	0.2
	$P(X = x)$	0.90	0.10	1

Here are two ways to check that the variables X are Y independent:

1. Verify that the conditional distribution is equal to marginal.

$$P(Y = 1 \mid X = 0) = 0.18/0.9 = 0.2$$
$$P(Y = 1 \mid X = 1) = 0.02/0.1 = 0.2$$

2. Verify that $P(X = x, Y = y) = P(X = x)P(Y = y)$

Example: diagnostic testing (two positive results)

The prevalence of a rare disease is 1 in ten thousand. A clinical test detects the disease correctly 98% of the time and has a false positive rate of 5%. What is the probability that the patient is disease free, even though two independent tests comes back positive?

Let $D = 1$ if a patient has a rare disease.
Let $T = 1$ if two positive tests are returned.

Example: diagnostic testing (two positive results)

The prevalence of a rare disease is 1 in ten thousand. A clinical test detects the disease correctly 98% of the time and has a false positive rate of 5%. What is the probability that the patient is disease free, even though two independent tests comes back positive?

Let $D = 1$ if a patient has a rare disease.
Let $T = 1$ if two positive tests are returned.

$$P(D = 1) = .0001 \qquad P(T = 1 \mid D = 1) = 0.98^2$$
$$P(D = 0) = .9999 \qquad P(T = 1 \mid D = 0) = 0.05^2$$

$$P(D = 0 \mid T = 1) = \frac{P(D = 0)P(T = 1 \mid D = 0)}{P(D = 0)P(T = 1 \mid D = 0) + P(D = 1)P(T = 1 \mid D = 1)}$$

$$= \frac{0.9999(0.05^2)}{0.9999(0.05^2) + 0.0001(0.98^2)} = 0.963.$$

For random variables Y_i, if they are independent we have

$$\begin{aligned} p\left(y_1, y_2, \ldots, y_n\right) &= p\left(y_1\right) p\left(y_2 \mid y_1\right) p\left(y_3 \mid y_2, y_1\right) \ldots p\left(y_n \mid y_{n-1}, y_{n-2}, \ldots, y_2, y_1\right) \\ &= p\left(y_1\right) p\left(y_2\right) p\left(y_3\right) \ldots \left(y_n\right) \end{aligned}$$

Example: Let Y_i be 1 if the i-th coin is a head and 0 otherwise.

What is a probability of 10 heads in a row?

$$\begin{aligned} & P(Y_1 = 1, Y_2 = 1, Y_3 = 1, \ldots, Y_{10} = 1) \\ & \quad = P(Y_1 = 1)P(Y_2 = 1)P(Y_3 = 1) \ldots P(Y_{10} = 1) \\ & \quad = 0.5^{10} \end{aligned}$$

IID Random Variables

Random variables $Y_1, Y_2, \ldots, Y_n$ are **independent and identically distributed** (abbreviated **i.i.d.**) if

- each Y_i is independent of all the others (**i.**)
- and, each Y_i has the same marginal distribution. (**i.d.**)

IMPORTANT NOTE: *i.i.d.* random variables are **both** independent **and** identically distributed.

You will see in a homework assignement that variables can be identically distributed, but not independent.

Example: Suppose I am about to toss 100 coins.

Let Y_i be 1 if the i-th coin is a head and 0 otherwise.

We usually think that the coin tosses are independent.

In addition, we usually think they are **identically distributed**, that is, each one has the same **marginal distribution**.

Here, each $Y_i \sim \text{Bernoulli}(0.5)$

We can succinctly describe coin tossing by

$$Y_i \sim \text{Bernoulli}(0.5), \quad i.i.d.$$

Example:

Suppose we have 1,000,000 voters.

- 400,000 are republican;
- 600,000 are democratic.

We randomly choose 3, sampling without replacement.
Let Y_i be 1 if the i-th voter is a democrat and 0 otherwise, $i = 1, 2, 3$.

How can we describe the joint distribution of (Y_1, Y_2, Y_3)?
Are they i.i.d.?

More examples

Suppose I am about to toss a die 100 times.
Let D_i be the outcome for the i-th toss (a number in {1, 2, 3, 4, 5, 6}). Are D_i i.i.d.?

- See https://mlakolar.shinyapps.io/RV_distributions/

Suppose an experienced NBA player is about to take repeated free-throws.
Let Y_i be 1 if he makes the i-th attempt and 0 otherwise. Are these Y_i i.i.d. Bernoulli?

Suppose the first penalty in an NHL game is on team A. For subsequent penalties $P_i = 1$ if the penalty is on a different team than the previous one and 0 otherwise. Are P_i independent? Are they i.i.d.?

Suppose you are monitoring a stock and for every 10 minute interval, you record whether the price went up or down. Let U_i be 1 if it goes up in the i-th interval, 0 otherwise. Are U_i i.i.d.?

Suppose that P_t is the price of a stock today and that P_{t+1} is its price tomorrow. Are P_t and P_{t+1} independent?

Model Building

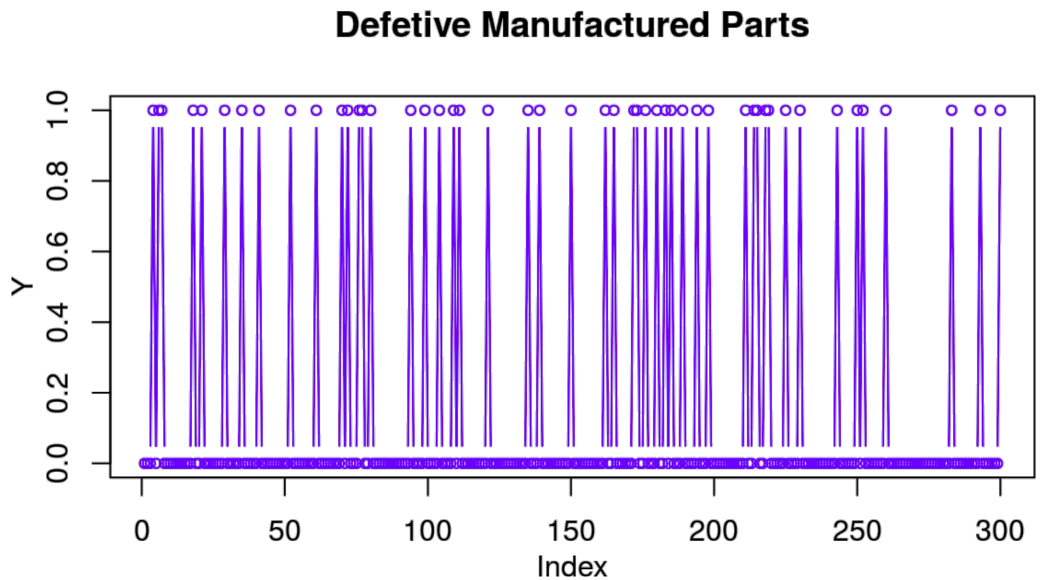
Up until now we have used the i.i.d. model by thinking about the process and trying to decide if the i.i.d. model is appropriate.

Usually, in statistics we look at the data to try to think up an appropriate model that captures the patterns (or lack of pattern) that we see.

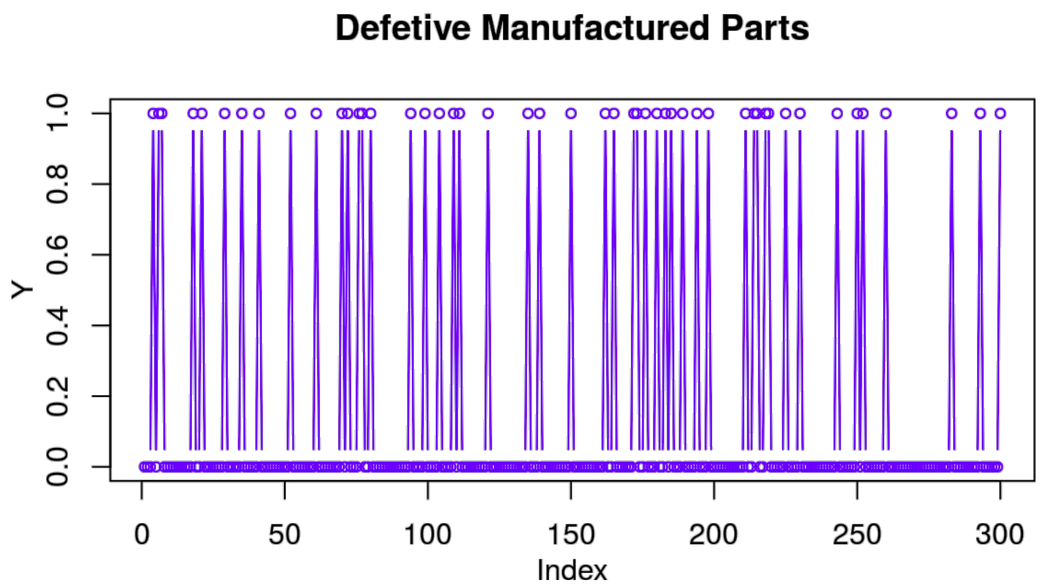
What kind of model could have generated the data that we see?

Example: The Defects Data

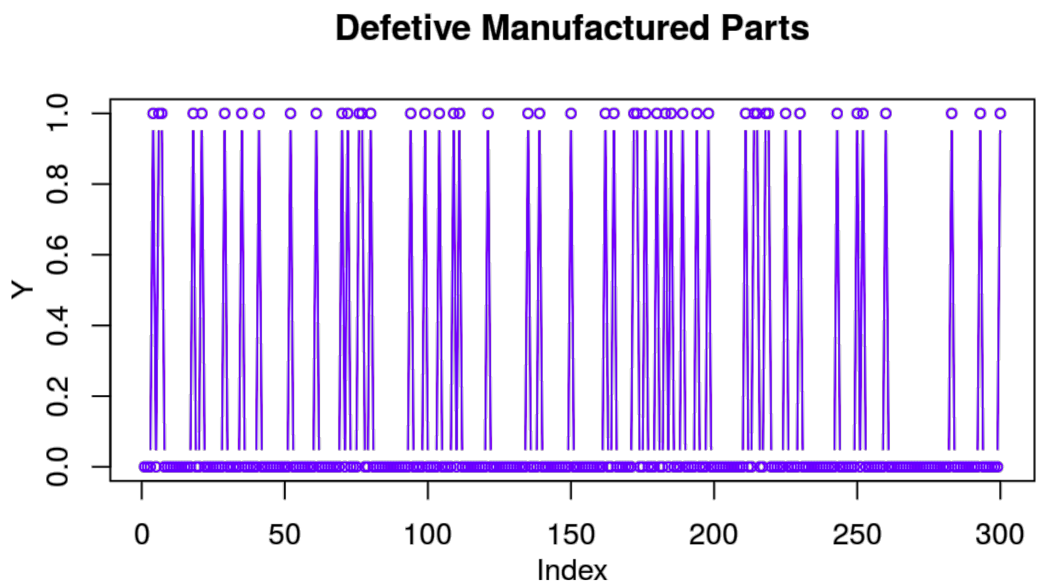
You are in charge of the process that manufactures a part.
You collect data on 300 parts that were made successively.
For each part, you write 1 if the part is defective, and 0 else.



49 of the 300 are defective or 49/300 = 16.3%

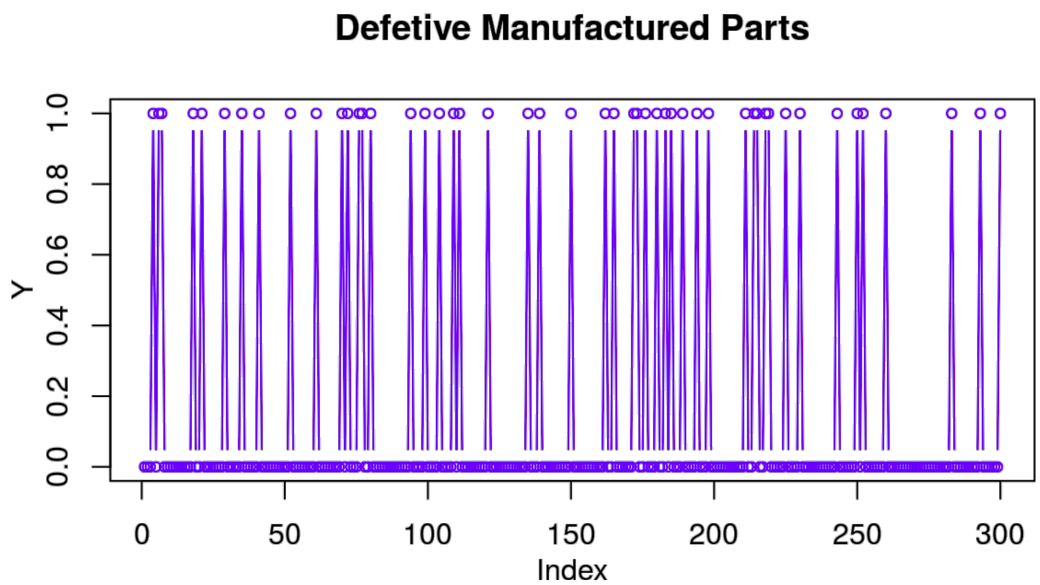


What is the probability that the next part is defective?
What is the probability that none of the next 10 parts are defective?



What is the probability that the next part is defective?
What is the probability that none of the next 10 parts are defective?

It does not look like $P(Y_i = 1 \mid Y_{i-1} = y_{i-1})$ depends on y_{i-1}.

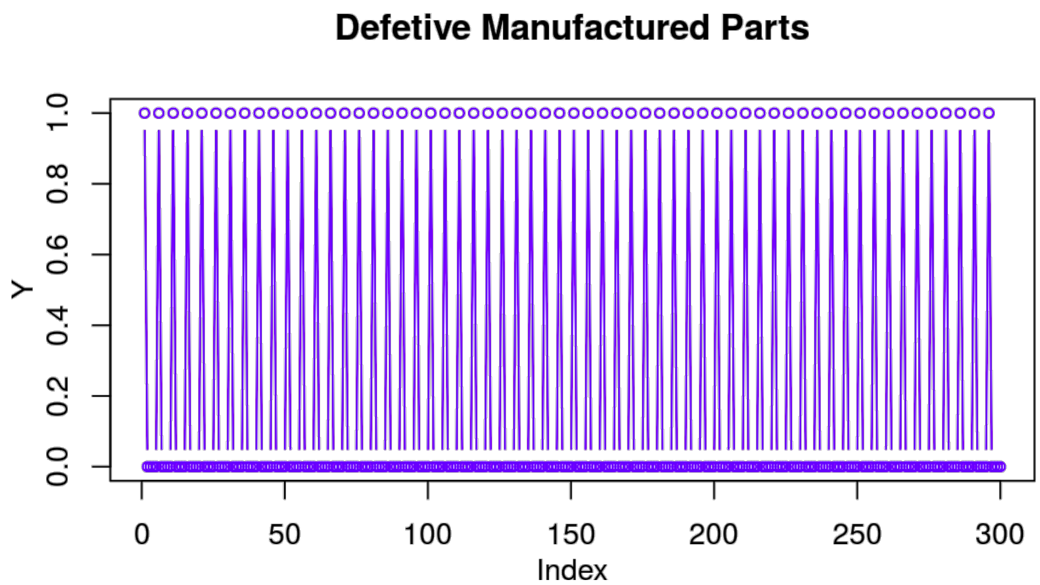


What is the probability that the next part is defective?
What is the probability that none of the next 10 parts are defective?

It does not look like $P(Y_i = 1 \mid Y_{i-1} = y_{i-1})$ depends on y_{i-1}.

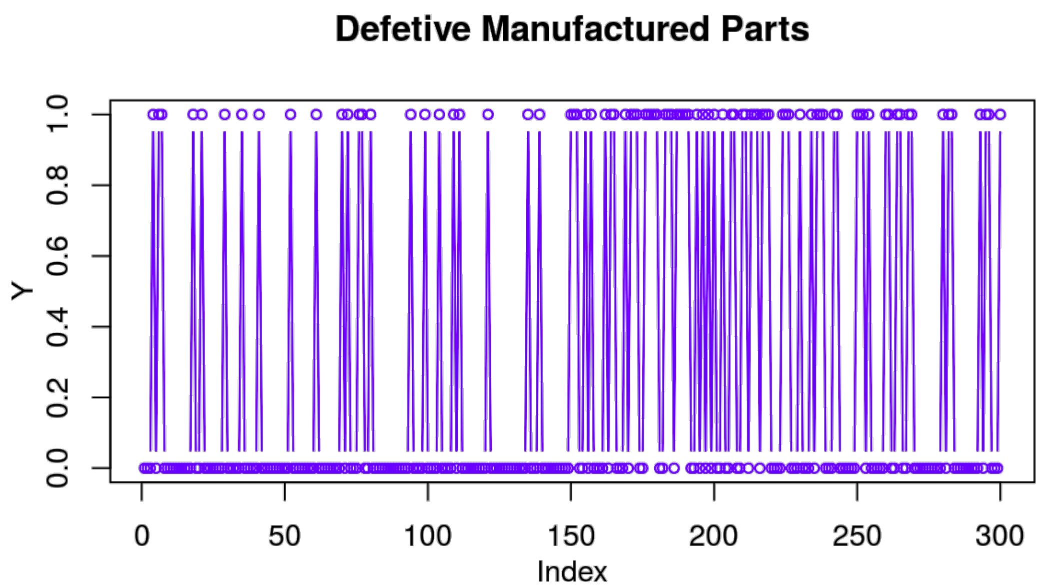
It does not look like the fraction of the defectives is changing over time.

Suppose you had this data. It always goes $1, 0, 0, 0, 0, \ldots, 1, 0, 0, 0, 0, \ldots$



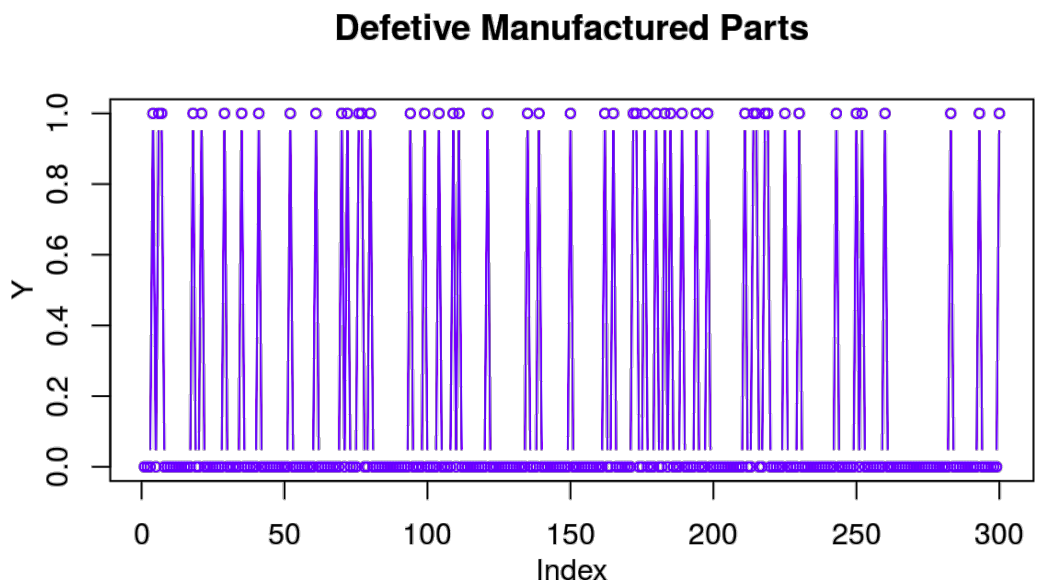
20% are 1's.
What is the probability that the next one is a 1?

Suppose you had this data:



It seems like you get more 1's (defects) in the second half of the data.

It seems like you were more likely to get a defect in the second half.



Back to our actual data.

Since there is no obvious pattern over successive values, we might think that whether we get a 1 or a 0 are independent draws from a Bernoulli.

Since the proportion of defective is not changed over time, we might assume the for each part, we have the same chance of getting a 1.

What is the probability of a 1? We might guess 16.3%.

The data looks like what you might get if $Y_i \sim \text{Bernoulli}(0.163), \quad i.i.d.$

We **model** the defects as i.i.d. $\mathrm{Bernoulli}(0.163)$

Assuming our model is correct (or close enough),

what is the probability that the next part is defective?

$$16.3$$

What is the probability that none of the next 10 parts are defective?

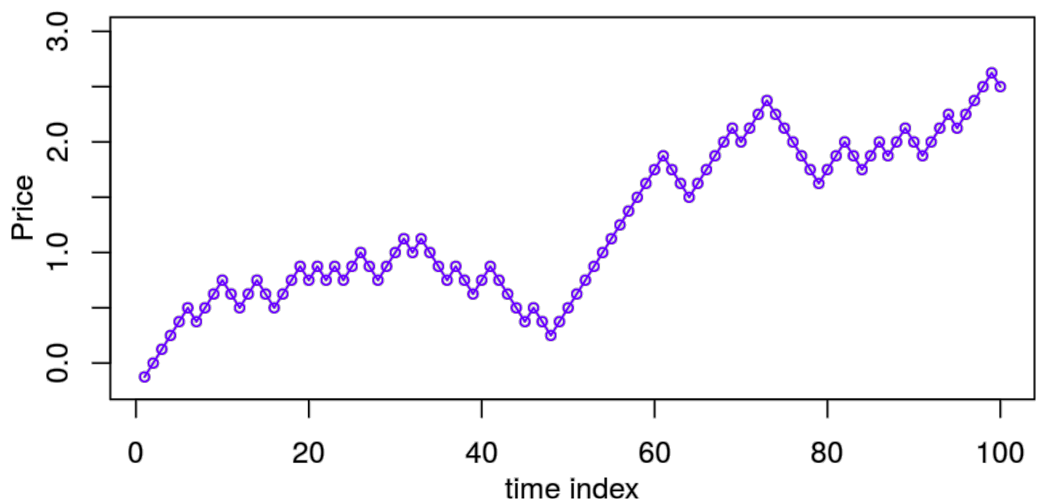
$$(1 - 0.163)^{10}$$

This is an example of what we often do in statistics.

- We build models that could have plausibly generated our observed data.
- The models have parameters which we estimate using the data.
- We can use the models to make predictions.

Example: The Random Walk Model

Below is the time series plot of the price of a stock.
The first price was subtracted from all the prices, so the first price is 0.
Each price change is one tick up or down where a tick is 0.125.

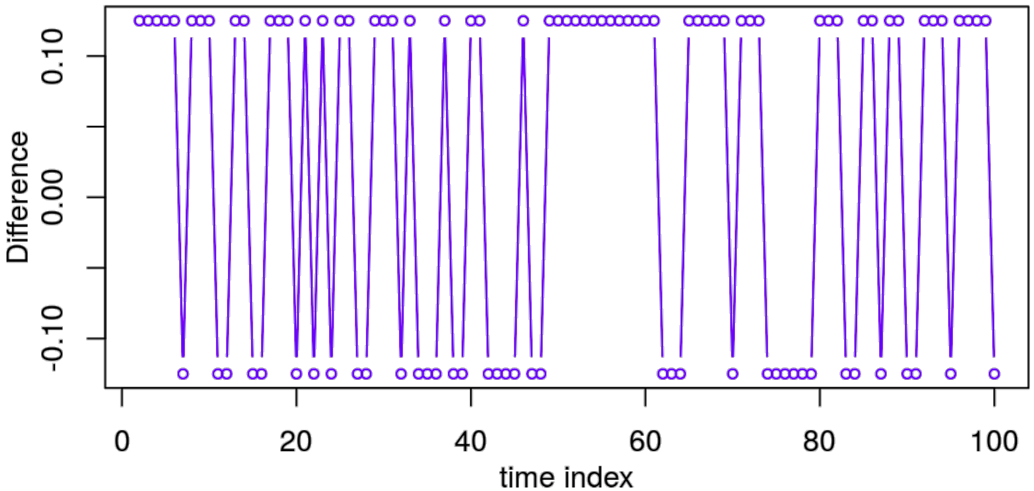


How would you model this data?
Could i.i.d. draws from some distribution have generated it?
What is your prediction for the next price?

The trick here is to look at the differences

$$D_t = P_t - P_{t-1}, \quad t = 2, 3, \ldots, 100$$

Here is the time series plot of the 99 differences.

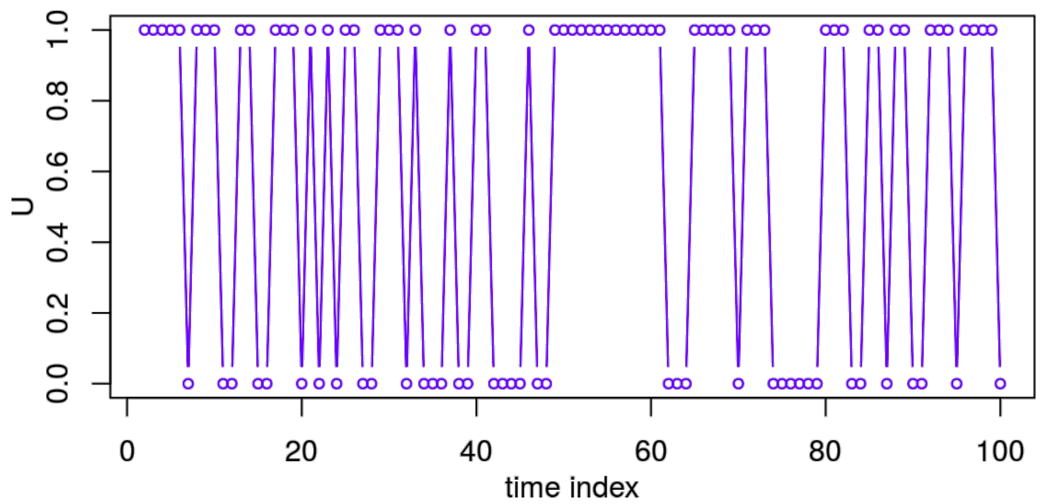


This looks like it might have come from an i.i.d. Bernoulli model,
but the outcomes are not 0 or 1.

Let U_t be 1 if the price went up and 0 otherwise.

$$D_t = 0.125 \cdot (2 \cdot U_t - 1)$$

Here is the time series plot of the 99 U_t.



Now, this does look like it might have come from an i.i.d. Bernoulli model.

The stock price went up 60 times out of 99.

We can model the U_t as i.i.d. Bernoulli, and estimate the parameter p with the sample proportion of ups $= 60/99 \approx 0.61$

$$U_t \sim \text{Bernoulli}(0.61), \quad \text{i.i.d.}$$

Both U_t and $D_t = 0.125 \cdot (2 \cdot U_t - 1)$ are i.i.d. with distributions

u	$P(U = u)$	d	$P(D = d)$
0	0.39	-0.125	0.39
1	0.61	0.125	0.61

Our model for the prices is

$$P_t = P_{t-1} + D_t, \qquad D_i \quad \text{i.i.d.}$$

Each price is the last price plus an increment, and each increment D_t is an i.i.d. draw, independent of the past.

The last price was 2.5.
What is your prediction for the next price ?

Note that

$$
\begin{aligned}
P_t &= P_{t-1} + D_t \\
D_t &= 0.125 \cdot (2 \cdot U_t - 1) \\
U_t &\sim \text{Bernoulli}(0.61) \quad \text{i.i.d.}
\end{aligned}
$$

is another way of specifying the conditional distribution $P(P_t = p_t \mid P_{t-1} = p_{t-1})$

p_t	$P(P_t = p_t \mid P_{t-1} = p_{t-1})$
$p_{t-1} - 0.125$	0.39
$p_{t-1} + 0.125$	0.61

We have a model for the conditional distribution of the next price given the current price.

The last price was 2.5.
What is your prediction for the next price ?

Note that

$$
\begin{aligned}
P_t &= P_{t-1} + D_t \\
D_t &= 0.125 \cdot (2 \cdot U_t - 1) \\
U_t &\sim \text{Bernoulli}(0.61) \quad \text{i.i.d.}
\end{aligned}
$$

is another way of specifying the conditional distribution $P(P_t = p_t \mid P_{t-1} = p_{t-1})$

p_t	$P(P_t = p_t \mid P_{t-1} = p_{t-1})$
$p_{t-1} - 0.125$	0.39
$p_{t-1} + 0.125$	0.61

We have a model for the conditional distribution of the next price given the current price.

Our model implies that

$$p(p_t \mid p_{t-1}, p_{t-2}, \ldots, p_1) = p(p_t \mid p_{t-1})$$

To predict the next price given all the previous prices, you only use the previous price.

Example: The binomial distribution

We have already seen that we can take a sum of two (or more) random variables to create a new random variable.

For example, the price of a stock was a sum of the price yesterday and the price change.

A binomial random variable can be constructed as the sum of independent Bernoulli random variables.

Examples:

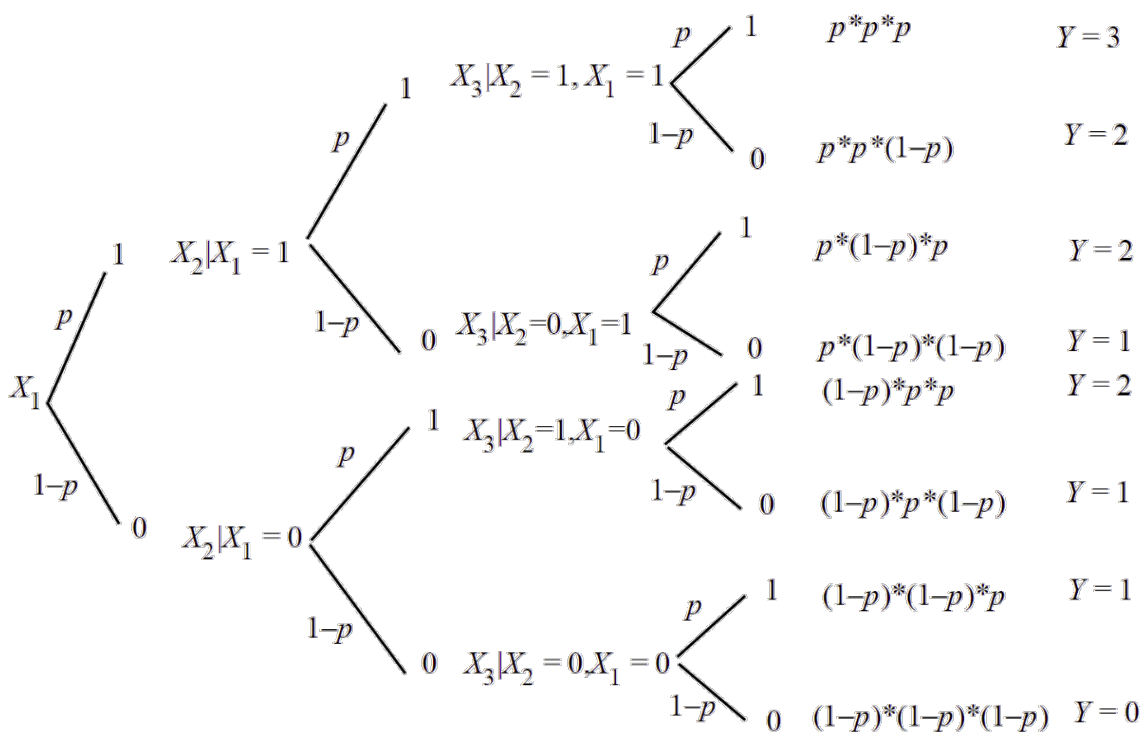
- Number of heads in 10 coin tosses
- Number of defective parts comming of assembly line
- Number of successful free throws in 20 trials

Familiarity with the binomial distribution eases many practical probability calculations.

Suppose you are about to make three parts.
Let $X_i = 1$ if the i-th part is good and 0 if it is defective for $i = 1, 2, 3$.

$$X_i \sim \text{Bernoulli}(p) \quad \text{i.i.d.}$$

How many parts will be good ?
To answer this, we consider the distribution of $Y = X_1 + X_2 + X_3$.



We can summarize the distribution of $Y = X_1 + X_2 + X_3$.

Event	y	$P(Y = y)$
000	0	$(1-p)^3$
001 or 100 or 010	1	$(1-p)(1-p)p + p(1-p)(1-p) + (1-p)p(1-p)$
011 or 110 or 101	2	$(1-p)p^2 + p^2(1-p) + p(1-p)p$
111	3	p^3

Let $X_1, X_2, \ldots, X_n$ denote n i.i.d. $\mathrm{Bernoulli}(p)$ random variables.
The binomial distribution is the probability distribution for the total number of successes:

$$Y = X_1 + X_2 + \ldots X_n = \sum_{i=1}^{n} X_i$$

What does this mean?

- You try something n times (X_i denotes i-th outcome).
- Each time you have the same chance p of success.
- Each time you try your probability of a success does not depend on any of the other outcomes.
- Y simply counts the number of success in n tries.

Let $X_1, X_2, \ldots, X_n$ denote n i.i.d. $\text{Bernoulli}(p)$ random variables.
The binomial distribution is the probability distribution for the total number of successes:

$$Y = X_1 + X_2 + \ldots X_n = \sum_{i=1}^{n} X_i$$

Y has a **binomial distribution** written $Y \sim \text{binomial}(n, p)$.

$$E[Y] = np \qquad \text{Var}[Y] = np(1-p)$$

$P(Y = y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \text{ for } y = 0, 1, 2, \ldots, n$

```
dbinom(y, n, p)
```

Remark: $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$is called "n factorial"

There is no need to memorize the function for $P(Y = y)$.
All statistical software packages can compute this for you.

However, remember the two special cases of the binomial distribution:

- If the probability of getting heads on each toss is 0.5, the probability of 10 heads in $n = 10$ tosses is

$$P(Y = 10) = 0.5^{10}$$

- Suppose the probability of a defect is 0.01 and you make 100 parts. What is the probability they are all good?

$$P(X = 100) = (1 - 0.01)^{100} = 0.99^{100}$$

Example: rural vs. urban hospitals

About as many boys as girls are born in hospitals. In a small Country Hospital only a few babies are born every week. In the urban center, many babies are born every week at City General. Say that a normal week is one where between 45% and 55% of the babies are female. An unusual week is one where more than 55% are girls or more than 55% are boys.

Which of the following is true?

- Unusual weeks occur equally often at Country Hospital and at City General.
- Unusual weeks are more common at Country Hospital than at City General.
- Unusual weeks are less common at Country Hospital than at City General.

Example: rural vs. urban hospitals

We can model the births in the two hospitals as two independent random variables.
Let $X =$ "number of baby girls born at Country Hospital."
Let $Y =$ "number of baby girls born at City General."

$$X \sim \text{Binomial}(N_1, p)$$

$$Y \sim \text{Binomial}(N_2, p)$$

Assume that $p = 0.5$. The key difference is that N_1 is much smaller than N_2. To illustrate, assume that $N_1 = 20$ and $N_2 = 500$.

Example: rural vs. urban hospitals

During a usual week at the rural hospital between $0.45N_1 = 0.45(20) = 9$ and $0.55N_1 = 0.55(20) = 11$ baby girls are born.

The probability of usual week is

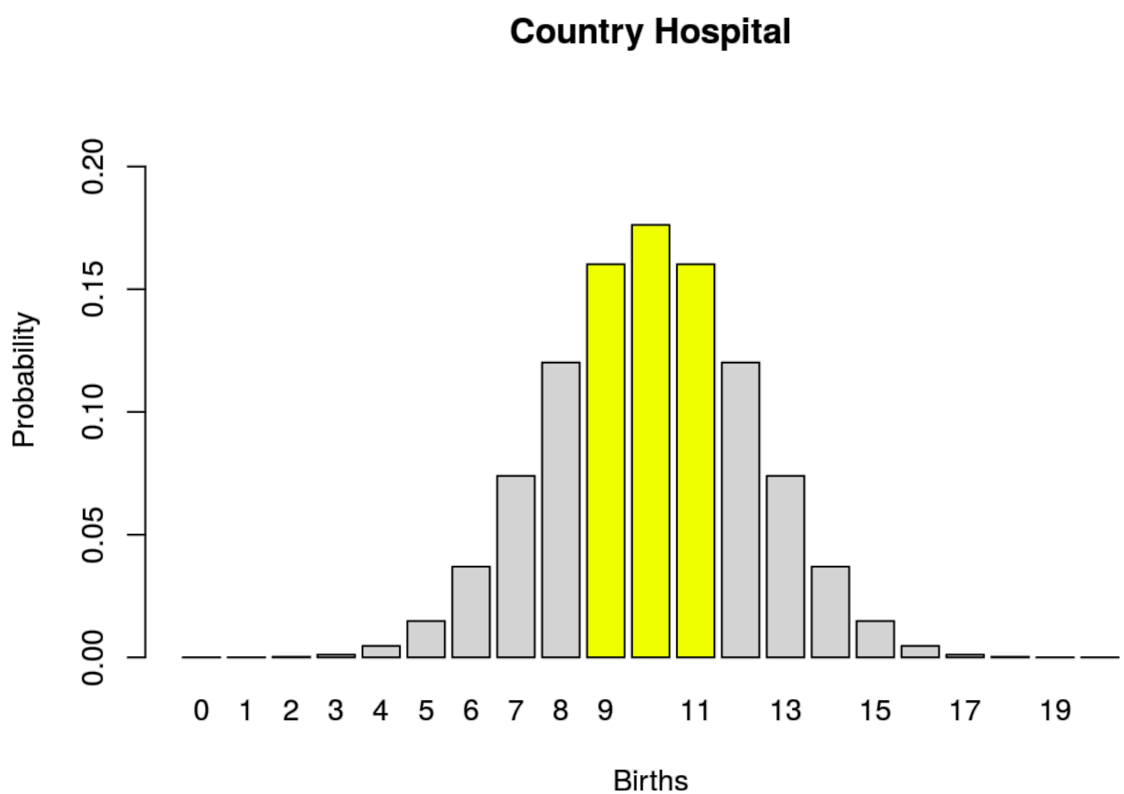
$$P(9 \leq X \leq 11) \approx 0.50,$$

so the probability of an unusual week is

$$1 - P(9 \leq X \leq 11) = P(X < 9) + P(X > 11) \approx 0.5.$$

Note: satisfying the condition $X < 9$ is the same as **not** satisfying the condition $X \geq 9$; strict versus non-strict inequalities make a difference.

Example: rural vs. urban hospitals



Example: rural vs. urban hospitals

In a usual week at the city hospital between $0.45N_2 = 0.45(500) = 225$and $0.55N_2 = 0.55(500) = 275$baby girls are born.

Then the probability of a usual week is

$$P(225 \leq X \leq 275) = 0.978,$$

so the probability of an unusual week is

$$1 - P(225 \leq X \leq 275) = P(X < 225) + P(X > 275) = 0.022.$$

Example: rural vs. urban hospitals