Math 140
Introductory Statistics

Conditional probability

Conditional Probability refers to the probability of a particular event where additional information is known.

An example

A laboratory technician is testing contaminated blood samples. Out of 100 samples, 20 are contaminated. Suppose she makes the correct decision 90% of the time (regardless of contamination or not). Make a table. What is the false positive rate? The false negative rate?

Detection of contamination

Contaminated?

	Pos	Neg	Total
Yes			
No			
Total			100

An example

	Pos	Neg	Total
Yes			20
No			80
Total			100

90% of the time she makes no mistakes

An example

	Pos	Neg	Total
Yes	18	2	20
No	8	72	80
Total			100

An example

	Pos	Neg	Total
Yes	18	2	20
No	8	72	80
Total	26	74	100

	Pos	Neg	Total
Yes	18	2	20
No	8	72	80
Total	26	74	100

False Pos. Rate= P(no disease | test positive)
False Neg. Rate= P(disease present | test negative)
Sensitivity= P(test positive | disease present)
Specificity= P(test negative | no disease)

	Pos	Neg	Total
Yes	18	2	20
No	8	72	80
Total	26	74	100

False Positive rate = 8/26 = 0.31

False Negative rate = 2/74 = 0.027

Sensitivity = 18/20 = 0.9

Specificity = 72/80 = 0.9

Conditional probability and statistical interference

We start with models.

For example, what is the probability of drawing an even number while rolling a die?

$P = 3/6 = 0.5$

True if the model is that DIE IS FAIR

Conditional probability and statistical interference

Let's now suppose that we suspect the die is not fair and that positive values are preferred.
How do we know for sure?

If I launch the die 20 times, what is the probability of getting all fair numbers?

For a fair die that is

$\underbrace{0.5 * 0.5 * 0.5 \ldots\ldots \text{etc}}_{\text{20 times}} = 0.5^{20} \sim 1$ in a million!

Conditional probability and statistical interference

1 in a million is such a small number that we can assume that if we find positive values after 20 rolls the "fair die" model can be abandoned.

There is something wrong!

What we calculated is

P(get even on all 20 rolls | the die is fair)

Conditional probability and statistical interference

But what we CANNOT calculate is

P(the die is fair | we get even numbers on 20 rolls)

The wise statistician says

If you start by assuming the model is true, you can compute the chances of various results.

But if you're trying to start from the results and compute the chance that the model is right or wrong:

You can't do that!

Independent events

		Male (%)	Female (%)	Row Total (%)
Weight Category	Not Obese	38	38	76
	Obese ($BMI \geq 30$)	12	12	24
	Column Total	50	50	100

It makes no difference whether we are discussing males or females. The percentage of obese people is the same!

Calculate

P(obese | male)

Independent events

		Male (%)	Female (%)	Row Total (%)
Weight Category	Not Obese	38	38	76
	Obese (*BMI* $\geq$ 30)	12	12	24
	Column Total	50	50	100

Recall

P(A and B) = P(A)* P (B | A)

A = male
B = obese

Independent events

		Male (%)	Female (%)	Row Total (%)
Weight Category	Not Obese	38	38	76
	Obese (*BMI* ≥ 30)	12	12	24
	Column Total	50	50	100

Recall

P(male and obese) = P(male)* P (obese | male)
0.12 = 0.5 * P(obese | male)

P(obese|male) = 0.12/0.5 = 0.24

exact same thing for women and general population

Independent events

		Male (%)	Female (%)	Row Total (%)
Weight Category	Not Obese	38	38	76
	Obese ($BMI \geq 30$)	12	12	24
	Column Total	50	50	100

They are all the same

P(obese | male) = 0.12/0.5 = 0.24

P(obese | female) = 0.12/0.5 = 0.24

P(obese) = 24/100 = 0.24

Independent events

Weight Category		Male (%)	Female (%)	Row Total (%)
	Not Obese	38	38	76
	Obese ($BMI \geq 30$)	12	12	24
	Column Total	50	50	100

The event OBESE is independent of the event MALE

A person at random is obese with probability 24%

Is obesity correlated with education?

		College or Technical School Graduate (%)	Not a College or Technical School Graduate (%)
Weight Category	Not Obese	81	73
	Obese ($BMI \geq 30$)	19	27
	Column Total	100	100

P(obese | college or technical school grad) = 19%

P(obese | not a college or technical school grad) = 27%

Is obesity correlated with education?

		College or Technical School Graduate (%)	Not a College or Technical School Graduate (%)
Weight Category	Not Obese	81	73
	Obese ($BMI \geq 30$)	19	27
	Column Total	100	100

We knew that is we randomly selected a person by chance, we would get 24%

These values are different, so the event OBESE is dependent on the event EDUCATION LEVEL

Is obesity correlated with education?

This is important knowledge in publc policy.

For example is we had to design a campaign to reduce obesity, we would know NOT to be concerned with gender differences but to focus more on non-college graduates

Independent events

Events A and B are independent if and only if

$P(A \mid B) = P(A)$ P(obese | male) = P(obese)

Equivalently, A and Bare independent if and only if

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

Knowing that B happened does not affect the probability of A happening. Knowing that A happened does not affect the probability of B happening.

We knew that

$$P(A \text{ and } B) = P(A)^* \, P(B \mid A)$$

For independent events $P(B|A) = P(B)$
this becomes

$$P(A \text{ and } B) = P(A)^* \, P(B)$$

Multiplication Rule for Independent Events

Two events A and B where $P(A) > 0$ and $P(B) > 0$ are independent if and only if

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

More generally, events $A_1, A_2, \ldots, A_n$ are independent if and only if

$$P(A_1 \text{ and } A_2 \text{ and } \ldots \text{ and } A_n) = P(A_1) \cdot P(A_2) \cdot \cdots \cdot P(A_n)$$

So the test for independence is

Check if $P(B|A) = P(B)$
or $P(A|B) = P(A)$

from here we can say

$P(A \text{ and } B) = P(A) * P(B)$

For coin flipping

If we flip a coin four times what is the probability that all flips give heads?

These are independent events so that

P(heads 1 and heads 2 and heads 3 and heads 4) =
P(heads 1) * P(heads2) * P(heads3) * P(heads4)=

1/2* 1/2* 1/2*1/2 =

1/16

You try to check for independence

Are the events being male and identifying tap water independent?

The test is: P(A | B) = P(A)?
If yes, they are independent

Identified Tap Water ?

Gender		Yes	No	Total
	Male	21	14	35
	Female	39	26	65
	Total	60	40	100

You try to check for independence

P(male|identify) = P(male)?

		Identified Tap Water ?		
		Yes	**No**	**Total**
Gender	**Male**	21	14	35
	Female	39	26	65
	Total	60	40	100

P(male|identify) = 21/60
P(male) = 35/100

are these the same?

You try to check for independence

P(male|identify) = P(male)= 7/20 YES

		Identified Tap Water ?		
		Yes	No	Total
Gender	Male	21	14	35
	Female	39	26	65
	Total	60	40	100

THE EVENTS ARE INDEPENDENT

Another case

		Identified Tap Water ?		
		Yes	**No**	**Total**
Drinks Bottled Water ?	**Yes**	24	6	30
	No	36	34	70
	Total	60	40	100

Are the events identifying tap water and drinking bottled water independent?

Another case

		Identified Tap Water ?		
		Yes	**No**	**Total**
Drinks Bottled Water ?	**Yes**	24	6	30
	No	36	34	70
	Total	60	40	100

P(drinking bottled water | identifying tap water) = 24/60

P(drinking bottled water) = 30/100

They are different

P(drinking bottled water | identifying tap water) = 4/10

P(drinking bottled water) = 3/10

The two events are DEPENDENT

Identified Tap Water ?

Drinks Bottled Water ?

	Yes	No	Total
Yes	24	6	30
No	36	34	70
Total	60	40	100

Does knowing that event A happened increase, decrease or leave unchanged the probability of event B?

A: The student is a football player.
B: The student weighs less than 120 pounds.

A: The student has long fingernails.
B: The student is female.

A: The student is a freshman.
B: The student is male.

A: The student is a freshman.
B: The student is a senior.

The Dodger games

Won the Game ?

Time of Game

	Yes	No	Total
Day	11	10	21
Night	30	27	57
Total	41	37	78

The Los Angeles Dodgers won a 41 games and lost 37.

Are the events win and day game independent?

Calculate P(win) vs. P(win | day)

A grain of salt

		Won the Game ?		
		Yes	**No**	**Total**
Time of Game	**Day**	11	10	21
	Night	30	27	57
	Total	41	37	78

$$P(\text{win}) = 41/78 = 0.526$$

$$P(\text{win} \mid \text{day}) = 11/21 = 0.524$$

They seem dependent, since the numbers are different, but they are so close! In practice, because the datais small, we can conclude they are independent.

We will work with other tests, later

Health care in America

About 30% of Americans between 18 and 24
don't have health insurance.

What is the chance that if I select 2 people at random,
One will have health insurance and the other will not?

Health care in America

About 30% of Americans between 18 and 24
don't have health insurance.

What is the chance that if I select 2 people at random,
One will have health insurance and the other wll not?

These are independent events, so that
P(1st yes health insurance and 2nd no heath insurance)
= P(yes h.i) * P(no h.i | yes h. I.)
P(yes h.i) * P(no h.i)
=0.3 * 0.7 =0.21

Health care in America

About 30% of young American adults ages 19 to 29 don't have health insurance.

Suppose you take a random sample of ten American adults in this age group. What is the probability that at least one of them
doesn't have health insurance?

Let’s think

P(at least one DOES NOT have health insurance) =

1 - P(all have it)

Let's think

P(at least one DOES NOT have health insurance) =

1 - P(all have it) =

1 - P(1st has it AND 2nd has it AND .. 10th has it)

Let’s think

P(at least one DOES NOT have health insurance) =

1 - P(all have it) =

1 - P(1st has it AND 2nd has it AND .. 10th has it) =

1 - P(1st has it) * P(2nd has it)… *P(10th has it)

Since they are independent

Let's think

P(at least one DOES NOT have health insurance) =

1 - P(all have it) =

1 - P(1st has it AND 2nd has it AND .. 10th has it) =

1 - P(1st has it) * P(2nd has it)… *P(10th has it) =

1 - 0.7 * 0.7 * 0.7 … *0.7

ten times =

$1 - (0.7)^{10}$

= 0.972

A sad story - Sally Clark

2 of her kids died of sudden infant death syndrome
Assume these are independent events and calculate

P(baby 1 died and baby 2 dies)

Assuming P(baby dies) = 1/8500

A sad story - Sally Clark

If the events were independent

P(baby 1 died and baby 2 dies) =
P(baby 1 died) * P(baby2 died | baby 1 died)
= P(baby 1 died) * P(baby 2 died) =
1/8500 * 1/8500

1 in 70 million

In the UK there are only about
200,000 second births per year

She was sentenced to life in prison

A sad story - Sally Clark

The Royal Statistical Society of the UK
argued that two babies dying in the same family
ARE NOT independent

and concluded that the previous analysis does not apply.

P(baby 1 died and baby 2 dies) =
P(baby 1 died) * P(baby2 died | baby 1 died)
= 1/8500 * 1/100

This translates to one or two per year for the UK data

A sad story - Sally Clark

Sally Clark was released from prison

She died after 4 years.

Her family says she never recovered from the miscarriage of justice.

Hk