Class 21

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-23

Independence

What does it mean for 2 random processes to be ***independent***?

Independence

What does it mean for 2 random processes to be ***independent***?

- The outcome for event A (e.g. A or A') provides no information about the outcome for event B (e.g. B or B'), and vice versa.

Independence

What does it mean for 2 random processes to be ***independent***?

- The outcome for event A (e.g. A or A') provides no information about the outcome for event B (e.g. B or B'), and vice versa.
- The probability of event B is the same, regardless of whether A is true or A' is true.

Independence

What does it mean for 2 random processes to be ***independent***?

- The outcome for event A (e.g. A or A') provides no information about the outcome for event B (e.g. B or B'), and vice versa.
- The probability of event B is the same, regardless of whether A is true or A' is true.
- $P(B|A) = P(B) = P(B|A')$

Dependence

What does it mean for 2 random processes to be ***dependent***?

Dependence

What does it mean for 2 random processes to be ***dependent***?

- You know more about how likely event B is to occur when you know event A or A' has occurred (i.e. $P(B|A)$ is more informative than $P(B)$)

Dependence

What does it mean for 2 random processes to be ***dependent***?

- You know more about how likely event B is to occur when you know event A or A' has occurred (i.e. $P(B|A)$ is more informative than $P(B)$)
- The probability of event B is different when A is true than when A' is true.

Dependence

What does it mean for 2 random processes to be ***dependent***?

- You know more about how likely event B is to occur when you know event A or A' has occurred (i.e. $P(B|A)$ is more informative than $P(B)$)
- The probability of event B is different when A is true than when A' is true.
- $P(B|A) \neq P(B)$, $P(B|A) \neq P(B|A')$

The General Multiplication Rule

The probability of event A ***and*** event B occurring is the product of the probability that A occurs and the *conditional probability* that B occurs given that A has already occurred.

$$\begin{aligned} P(A \text{ and } B) &= P(A) \times P(B \text{ given } A) \\ &= P(A) \times P(B|A) \\ &= P(A \cup B) \end{aligned}$$

Conditional Probability of Independent Events

- When event B is independent of event A, it's probability is not *conditional* on the outcome of A

Conditional Probability of Independent Events

- When event B is independent of event A, it's probability is not *conditional* on the outcome of A
- When event B is independent of event A, the probability $P(B|A)$ is just the probability $P(B)$

Conditional Probability of Independent Events

- When event B is independent of event A, it's probability is not *conditional* on the outcome of A
- When event B is independent of event A, the probability $P(B|A)$ is just the probability $P(B)$
- The probability of event B does not depend on event A

Multiplication Rule for Independent Processes

The probability of event A ***and*** event B occurring is the product of the probability that A occurs and the probability that B occurs, because the probability of B does not change based on the outcome of A.

$$\begin{aligned} P(A \text{ and } B) &= P(A) \times P(B \text{ given } A) \\ &= P(A) \times P(B|A) \\ &= P(A) \times P(B) \\ &= P(A \cup B) \end{aligned}$$

Determining Independence for Categorical Variables

- Compare conditional probabilities.
  $P(B|A_1) \approx P(B|A_2) \approx ... \approx P(B|A_k)$ for all k possible outcomes of event A when event B is independent of event A.

Determining Independence for Categorical Variables

- Compare conditional probabilities.
  $P(B|A_1) \approx P(B|A_2) \approx ... \approx P(B|A_k)$ for all k possible outcomes of event A when event B is independent of event A.
- Check if the *observed* outcomes are consistent with the *expected* outcomes, assuming event B is independent of event A. $P(A \text{ and } B) = P(A) \times P(B)$ when event B is independent of event A.

Homework Problem: Marketing

Your company is pilot-testing a new email campaign in a random sample of customers. They sent out 1000 emails. Of those 1000 emails, 176 were marked as read but not opened, 106 were opened but no links clicked, 159 were opened and a link clicked, 100 were opened then deleted, 196 were deleted without being opened, 195 were left unread, and 68 were undeliverable.

Homework Problem: Marketing

Your company is pilot-testing a new email campaign in a random sample of customers. They sent out 1000 emails. Of those 1000 emails, 176 were marked as read but not opened, 106 were opened but no links clicked, 159 were opened and a link clicked, 100 were opened then deleted, 196 were deleted without being opened, 195 were left unread, and 68 were undeliverable.

385 of the 1000 sampled customers purchased the advertised product, 142 of which also opened the email.

Homework Problem: Marketing

Your company is pilot-testing a new email campaign in a random sample of customers. They sent out 1000 emails. Of those 1000 emails, 176 were marked as read but not opened, 106 were opened but no links clicked, 159 were opened and a link clicked, 100 were opened then deleted, 196 were deleted without being opened, 195 were left unread, and 68 were undeliverable.

385 of the 1000 sampled customers purchased the advertised product, 142 of which also opened the email.

Are making a purchase and opening the email dependent or independent processes?

What are the 2 random processes?

What are the 2 random processes?

- Made a purchase

$$P(\text{bought})$$

- Opened the email

$$P(\text{opened})$$

Contingency Table

Table 1: Purchases by Email Opens

opened	TRUE	FALSE	Total
TRUE	142	223	365
FALSE	243	392	635
Total	385	615	1000

Approach 1: Conditional Probabilities

$P(\text{bought} \,|\, \text{opened}) \approx P(\text{bought} \,|\, \text{notopened})$ or
$P(\text{bought} \,|\, \text{opened}) \approx P(\text{bought})$ suggests independence.

Approach 1: Conditional Probabilities

$P(\text{bought} \,|\, \text{opened}) \approx P(\text{bought} \,|\, \text{notopened})$ or
$P(\text{bought} \,|\, \text{opened}) \approx P(\text{bought})$ suggests independence.

- Of the 1000 customers, 385 made a purchase.
- Of the 365 people who opened the email, 142 of them made a purchase.
- Of the 635 people who did not open the email, 243 of them made a purchase.

Approach 1: Conditional Probabilities

$P(\text{bought} \,|\, \text{opened}) \approx P(\text{bought} \,|\, \text{notopened})$ or
$P(\text{bought} \,|\, \text{opened}) \approx P(\text{bought})$ suggests independence.

- Of the 1000 customers, 385 made a purchase.
- Of the 365 people who opened the email, 142 of them made a purchase.
- Of the 635 people who did not open the email, 243 of them made a purchase.

Are these probabilities the same??

What is the probability a customer made a purchase?

"385 of the 1000 sampled customers purchased the advertised product"

What is the probability a customer made a purchase?

"385 of the 1000 sampled customers purchased the advertised product"

$$
\begin{aligned}
P(\text{bought}) &= \frac{\text{count}(\text{count})}{\text{count}(\text{customers})} \\
&= \frac{385}{1000} \\
&= 0.385
\end{aligned}
$$

Calculating Conditional Probabilities

$$
\begin{aligned}
P(\text{bought} \mid \text{opened}) &= \frac{\text{count}(\text{bought})}{\text{count}(\text{opened})} \\
&= \frac{142}{365} \\
&= 0.389
\end{aligned}
$$

Calculating Conditional Probabilities

$$
\begin{aligned}
P(\text{bought} \,|\, \text{opened}) &= \frac{\text{count}(\text{bought})}{\text{count}(\text{opened})} \\
&= \frac{142}{365} \\
&= 0.389
\end{aligned}
$$

$$
\begin{aligned}
P(\text{bought} \,|\, \text{notopened}) &= \frac{\text{count}(\text{bought})}{\text{count}(\text{notopened})} \\
&= \frac{243}{635} \\
&= 0.383
\end{aligned}
$$

Calculating Conditional Probabilities

Variable 2

Variable 1	Category 1	Category 2	Total
Category 1	A / (A + C)	B / (B + D)	1
Category 2	C / (A + C)	D / (B + D)	1
Total	(A + C) / (A + B + C + D)	(B + D) / (A + B + C + D)	1

Proportions by Row

Table 2: Proportion of Purchases by Email Opens

opened	TRUE	FALSE	Total
TRUE	0.3890411	0.6109589	1
FALSE	0.3826772	0.6173228	1
Total	0.3850000	0.6150000	1

Approach 2: Assume Independence

If making a purchase is independent of opening the email, then $P(\text{boughtandopened}) \approx P(\text{opened}) \times P(\text{bought})$.

Approach 2: Assume Independence

If making a purchase is independent of opening the email, then $P(\text{boughtandopened}) \approx P(\text{opened}) \times P(\text{bought})$.

- Of the 1000 customers, 385 made a purchase.
- Of the 1000 customers, 365 opened the email.
- Of the 1000 customers, 142 opened the email and made a purchase.

Approach 2: Assume Independence

If making a purchase is independent of opening the email, then $P(\text{boughtandopened}) \approx P(\text{opened}) \times P(\text{bought})$.

- Of the 1000 customers, 385 made a purchase.
- Of the 1000 customers, 365 opened the email.
- Of the 1000 customers, 142 opened the email and made a purchase.

Are these results consistent with the expected results, assuming these events are independent?

What is the probability a customer opened the email?

"***159 were opened and a link clicked, 100 were opened then deleted***"

What is the probability a customer opened the email?

"***159 were opened and a link clicked, 100 were opened then deleted***"

$$
\begin{aligned}
P(\text{open}) &= \frac{\text{count}(\text{opened})}{\text{count}(\text{sent})} \\
&= \frac{106 + 159 + 100}{1000} \\
&= \frac{365}{1000} \\
&= 0.365
\end{aligned}
$$

Calculating Expected Results

If we assume making a purchase is independent of opening the email, we can apply ***The Multiplication Rule for Independent Events*** to estimate the proportion of customers who we would expect to do both and compare it to the observed data.

Calculating Expected Results

If we assume making a purchase is independent of opening the email, we can apply ***The Multiplication Rule for Independent Events*** to estimate the proportion of customers who we would expect to do both and compare it to the observed data.

- H_0: $P(\text{boughtandopened}) = P(\text{opened}) \times P(\text{bought})$
- H_A: $P(\text{boughtandopened}) \neq P(\text{opened}) \times P(\text{bought})$

Calculating Expected Results

$$\begin{aligned} P(\text{boughtandopened}) &= P(\text{opened}) \times P(\text{bought}) \\ &= 0.365 \times 0.385 \\ &= 0.141 \end{aligned}$$

Calculating Expected Results

$$
\begin{aligned}
P(\text{boughtandopened}) &= P(\text{opened}) \times P(\text{bought}) \\
&= 0.365 \times 0.385 \\
&= 0.141
\end{aligned}
$$

$$
\begin{aligned}
P(\text{boughtandopened}) &= \frac{\text{count}(\text{boughtandopened})}{n} \\
&= \frac{142}{1000} \\
&= 0.142
\end{aligned}
$$

Homework Problem: Marketing

Your company is pilot-testing a new email campaign in a random sample of customers. They sent out 1000 emails. Of those 1000 emails, 176 were marked as read but not opened, 106 were opened but no links clicked, 159 were opened and a link clicked, 100 were opened then deleted, 196 were deleted without being opened, 195 were left unread, and 68 were undeliverable.

Homework Problem: Marketing

Your company is pilot-testing a new email campaign in a random sample of customers. They sent out 1000 emails. Of those 1000 emails, 176 were marked as read but not opened, 106 were opened but no links clicked, 159 were opened and a link clicked, 100 were opened then deleted, 196 were deleted without being opened, 195 were left unread, and 68 were undeliverable.

385 of the 1000 sampled customers purchased the advertised product, 107 of which also opened the email and clicked a link.

Homework Problem: Marketing

Your company is pilot-testing a new email campaign in a random sample of customers. They sent out 1000 emails. Of those 1000 emails, 176 were marked as read but not opened, 106 were opened but no links clicked, 159 were opened and a link clicked, 100 were opened then deleted, 196 were deleted without being opened, 195 were left unread, and 68 were undeliverable.

385 of the 1000 sampled customers purchased the advertised product, 107 of which also opened the email and clicked a link.

Are making a purchase and clicking a link after opening the email dependent or independent processes?

What are the 2 random processes?

What are the 2 random processes?

- Made a purchase

$$P(\text{bought})$$

- Clicked a link

$$P(\text{linked})$$

Contingency Table

Table 3: Purchases by Link Clicks

linked	TRUE	FALSE	Total
TRUE	107	52	159
FALSE	278	563	841
Total	385	615	1000

Approach 1: Conditional Probabilities

$P(\text{bought} \,|\, \text{linked}) \approx P(\text{bought} \,|\, \text{notlinked})$ or
$P(\text{bought} \,|\, \text{linked}) \approx P(\text{bought})$ suggests independence.

Approach 1: Conditional Probabilities

$P(\text{bought} \,|\, \text{linked}) \approx P(\text{bought} \,|\, \text{notlinked})$ or
$P(\text{bought} \,|\, \text{linked}) \approx P(\text{bought})$ suggests independence.

- Of the 1000 customers, 385 made a purchase.
- Of the 159 people who clicked a link, 107 of them made a purchase.
- Of the 841 people who did not click a link, 278 of them made a purchase.

Approach 1: Conditional Probabilities

$P(\text{bought} \,|\, \text{linked}) \approx P(\text{bought} \,|\, \text{notlinked})$ or
$P(\text{bought} \,|\, \text{linked}) \approx P(\text{bought})$ suggests independence.

- Of the 1000 customers, 385 made a purchase.
- Of the 159 people who clicked a link, 107 of them made a purchase.
- Of the 841 people who did not click a link, 278 of them made a purchase.

Are these probabilities the same??

What is the probability a customer made a purchase?

"385 of the 1000 sampled customers purchased the advertised product"

What is the probability a customer made a purchase?

"385 of the 1000 sampled customers purchased the advertised product"

$$
\begin{aligned}
P(\text{buy}) &= \frac{\text{count}(\text{purchases})}{\text{count}(\text{customers})} \\
&= \frac{385}{1000} \\
&= 0.385
\end{aligned}
$$

Calculating Conditional Probabilities

$$
\begin{aligned}
P(\text{bought} \mid \text{linked}) &= \frac{\text{count}(\text{bought})}{\text{count}(\text{linked})} \\
&= \frac{107}{159} \\
&= 0.673
\end{aligned}
$$

Calculating Conditional Probabilities

$$
\begin{aligned}
P(\text{bought} \mid \text{linked}) &= \frac{\text{count}(\text{bought})}{\text{count}(\text{linked})} \\
&= \frac{107}{159} \\
&= 0.673
\end{aligned}
$$

$$
\begin{aligned}
P(\text{bought} \mid \text{notlinked}) &= \frac{\text{count}(\text{bought})}{\text{count}(\text{notlinked})} \\
&= \frac{278}{841} \\
&= 0.331
\end{aligned}
$$

Calculating Conditional Probabilities

Variable 2

Variable 1	Category 1	Category 2	Total
Category 1	A / (A + C)	B / (B + D)	1
Category 2	C / (A + C)	D / (B + D)	1
Total	(A + C) / (A + B + C + D)	(B + D) / (A + B + C + D)	1

Proportions by Row

Table 4: Proportion of Purchases by Link Clicks

linked	TRUE	FALSE	Total
TRUE	0.6729560	0.3270440	1
FALSE	0.3305589	0.6694411	1
Total	0.3850000	0.6150000	1

Approach 2: Assume Independence

If making a purchase is independent of clicking a link, then $P(\text{boughtandlinked}) \approx P(\text{linked}) \times P(\text{bought})$.

Approach 2: Assume Independence

If making a purchase is independent of clicking a link, then $P(\text{boughtandlinked}) \approx P(\text{linked}) \times P(\text{bought})$.

- Of the 1000 customers, 385 made a purchase.
- Of the 1000 customers, 159 clicked a link.
- Of the 1000 customers, 107 clicked a link and made a purchase.

Approach 2: Assume Independence

If making a purchase is independent of clicking a link, then $P(\text{boughtandlinked}) \approx P(\text{linked}) \times P(\text{bought})$.

- Of the 1000 customers, 385 made a purchase.
- Of the 1000 customers, 159 clicked a link.
- Of the 1000 customers, 107 clicked a link and made a purchase.

Are these results consistent with the expected results, assuming these events are independent?

What is the probability a customer clicked a link?

"***159 were opened and a link clicked***"

What is the probability a customer clicked a link?

*"**159 were opened and a link clicked**"*

$$
\begin{aligned}
P(\text{linked}) &= \frac{\text{count}(\text{linked})}{\text{count}(\text{sent})} \\
&= \frac{159}{1000} \\
&= 0.159
\end{aligned}
$$

Calculating Expected Results

If we assume making a purchase is independent of clicking a link, we can apply ***The Multiplication Rule for Independent Events*** to estimate the proportion of customers who we would expect to do both and compare it to the observed data.

Calculating Expected Results

If we assume making a purchase is independent of clicking a link, we can apply ***The Multiplication Rule for Independent Events*** to estimate the proportion of customers who we would expect to do both and compare it to the observed data.

- H_0: $P(\text{boughtandlinked}) = P(\text{linked}) \times P(\text{bought})$
- H_A: $P(\text{boughtandlinked}) \neq P(\text{linked}) \times P(\text{bought})$

Calculating Expected Results

$$\begin{aligned} P(\text{boughtandlinked}) &= P(\text{linked}) \times P(\text{bought}) \\ &= 0.159 \times 0.385 \\ &= 0.061 \end{aligned}$$

Calculating Expected Results

$$
\begin{aligned}
P(\text{boughtandlinked}) &= P(\text{linked}) \times P(\text{bought}) \\
&= 0.159 \times 0.385 \\
&= 0.061
\end{aligned}
$$

$$
\begin{aligned}
P(\text{boughtandlinked}) &= \frac{\text{count}(\text{boughtandlinked})}{n} \\
&= \frac{107}{1000} \\
&= 0.107
\end{aligned}
$$

Homework: Car Crash Injuries

You're trying to investigate if the risk of injury following a car accident differs between men and women. You took a random sample of 1000 car accident reports involving 672 women and 428 men. An injury occurred in 697 of the reports.

Homework: Car Crash Injuries

You're trying to investigate if the risk of injury following a car accident differs between men and women. You took a random sample of 1000 car accident reports involving 672 women and 428 men. An injury occurred in 697 of the reports.

If these two processes are independent, what is the probability that a woman was injured?

Homework: Car Crash Injuries

If getting an injury in a car accident is independent of the sex of the occupant, then we would expect
$P(\text{womaninjured}) = P(\text{injury}) \times P(\text{woman})$.

Homework: Car Crash Injuries

If getting an injury in a car accident is independent of the sex of the occupant, then we would expect $P(\text{womaninjured}) = P(\text{injury}) \times P(\text{woman})$.

- Out of 1000 accident reports, 697 involved injuries. $P(\text{injury}) = \frac{697}{1000} = 0.697$

Homework: Car Crash Injuries

If getting an injury in a car accident is independent of the sex of the occupant, then we would expect $P(\text{womaninjured}) = P(\text{injury}) \times P(\text{woman})$.

- Out of 1000 accident reports, 697 involved injuries. $P(\text{injury}) = \frac{697}{1000} = 0.697$
- Out of 1000 accident reports, 672 involved women. $P(\text{woman}) = \frac{672}{1000} = 0.672$

Homework: Car Crash Injuries

If getting an injury in a car accident is independent of the sex of the occupant, then we would expect $P(\text{womaninjured}) = P(\text{injury}) \times P(\text{woman})$.

- Out of 1000 accident reports, 697 involved injuries. $P(\text{injury}) = \frac{697}{1000} = 0.697$
- Out of 1000 accident reports, 672 involved women. $P(\text{woman}) = \frac{672}{1000} = 0.672$

If these events were independent, the probability that a woman is injured in an accident would be $0.697 \times 0.672 = 0.468$.

Homework: Hurricanes

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

Assuming that events are independent, what is the probability that 2 hurricanes in a row make landfall in 2024?

Homework: Hurricanes

If one hurricane making landfall is independent of another hurricane making landfall, we would expect
$P(\text{landfall}_1 \text{ and} \text{landfall}_2) = P(\text{landfall}_1) \times P(\text{landfall}_2)$

Homework: Hurricanes

If one hurricane making landfall is independent of another hurricane making landfall, we would expect
$P(\text{landfall}_1 \text{ and} \text{landfall}_2) = P(\text{landfall}_1) \times P(\text{landfall}_2)$

Out of 298 hurricanes from 1980-2023, 72 made landfall.
$P(\text{landfall}) = \frac{72}{298} = 0.242$

Homework: Hurricanes

If one hurricane making landfall is independent of another hurricane making landfall, we would expect
$P(\text{landfall}_1 \text{ and} \text{landfall}_2) = P(\text{landfall}_1) \times P(\text{landfall}_2)$

Out of 298 hurricanes from 1980-2023, 72 made landfall.
$P(\text{landfall}) = \frac{72}{298} = 0.242$

If these events were independent, the probability two hurricanes make landfall would be $0.242 \times 0.242 = 0.058$.