Statistical dependence

Business Statistics 41000

Fall 2015
Topics

1. Conditional probability and conditional distributions

2. Independence

3. Multivariate random variables

4. Continuous distributions

5. Correlation and linear prediction
See OpenIntro section 2.2, exercises 2.6.2.

“Pruning possibilities, renormalizing to make probabilities.”
Example: single die

What is the probability that you roll a six (event $A$), given that you roll an even number (event $B$)?

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>two</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>three</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>four</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>five</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>six</td>
<td>$\frac{1}{6}$</td>
</tr>
</tbody>
</table>

\[
P(A \mid B) = \frac{P(A \& B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{6} + \frac{1}{6}} = \frac{1}{3}
\]
Example: single die

What is the probability that you roll an outcome ending in a vowel, given that you roll a number less than six?

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>( \frac{1}{6} )</td>
</tr>
<tr>
<td>two</td>
<td>( \frac{1}{6} )</td>
</tr>
<tr>
<td>three</td>
<td>( \frac{1}{6} )</td>
</tr>
<tr>
<td>four</td>
<td>( \frac{1}{6} )</td>
</tr>
<tr>
<td>five</td>
<td>( \frac{1}{6} )</td>
</tr>
<tr>
<td>six</td>
<td>( \frac{1}{6} )</td>
</tr>
</tbody>
</table>

\[
P(A \mid B) = \frac{P(A \& B)}{P(B)} = \frac{4}{5}
\]
Definition: conditioning

This process of *restricting* the sample space and *renormalizing* to get valid probabilities is called **conditioning**. It is one of the most useful ideas in probability.

**Conditional Probability**

\[
P(A \mid B) = \frac{P(A \& B)}{P(B)}.\]

We read the bar symbol as “given” or “conditional on”.

This formula underpins the use of probability to learn from new information. The new information is precisely the restriction of the sample space; by ruling out some possibilities we know more than we did before.
Example: is marijuana a gateway drug?

Let $H =$“uses heroin” and $M =$“smokes marijuana”. Suppose the probabilities of the four outcomes are as follows, per thousand people.

<table>
<thead>
<tr>
<th></th>
<th>$H$</th>
<th>not-$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>9</td>
<td>191</td>
</tr>
<tr>
<td>not-$M$</td>
<td>1</td>
<td>799</td>
</tr>
</tbody>
</table>

(divide by 1,000 to get the probabilities on the correct scale.)
Example: is marijuana a gateway drug? (cont’d)

According to these (completely fake) numbers, a person who smokes marijuana is quite a lot more likely to use heroin compared to someone who does not, but the rise is from 0.13% probability to 4.5% probability. Among heroin users, on the other hand, pot smoking is the norm at 90%.

<table>
<thead>
<tr>
<th></th>
<th>$H$</th>
<th>not-$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>9</td>
<td>191</td>
</tr>
<tr>
<td>not-$M$</td>
<td>1</td>
<td>799</td>
</tr>
</tbody>
</table>

$$P(M \mid H) = 0.90,$$
$$P(M \mid \text{not-}H) = 0.19,$$
$$P(H \mid M) = 0.045,$$
$$P(H \mid \text{not-}M) = 0.00125.$$
Example: diagnostic false positive

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 the test comes back positive?

We can answer this problem using a probability formula called Bayes’ rule.
We can rearrange the conditional probability formula to get a new way to think about joint probabilities.

**Compositional representation of joint probabilities**

\[ P(A \& B) = P(A \mid B)P(B). \]

This is useful for thinking about joint probabilities mechanistically: \( P(A \mid B) \) and \( P(B) \) are sometimes easier to deal with directly.
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 $j$th draw was yellow, and $B$ = “draw three yellow Skittles in a row”. $P(B)$ can be expressed as

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

When he’s sober, he hits 0.350; when he’s drunk he hits a mere 0.180. He’s drunk 20% of the time.

So on any given at-bat the probability he both \( A = \) “gets a hit” and \( B = \) “is drunk” is given by

\[
P(A \& B) = P(A \mid B)P(B),
\]
\[
= 0.18(0.2) = .036.
\]

The overall probability that he gets a hit on a random at bat are given by the law of total probability

\[
P(A) = P(A \mid B)P(B) + P(A \mid \text{not-}B)P(\text{not-}B)
\]
\[
= 0.18(0.2) + 0.35(0.8) = 0.316.
\]
Example: the alcoholic baseball hitter (cont’d)

The next year, his batting average goes up to 0.324 despite the fact that his sober hitting drops to 0.340 and his drunk hitting drops to 0.175! How is that possible? By cutting back the frequency of his drunk at bats to 10%.

\[
P(A) = P(A \mid B)P(B) + P(A \mid \text{not-}B)P(\text{not-}B)
\]
\[
= 0.175(0.1) + 0.34(0.9) = 0.3235.
\]

This phenomenon is called **Simpson’s paradox**. Does it seem paradoxical to you?
Example: the Berkeley gender bias case

Let $A =$ “admitted to Berkeley”. In 1973 it was noted that $P(A \mid \text{male}) = 0.44$ while $P(A \mid \text{female}) = 0.35$. Meanwhile, individual departments showed no signs of discrimination. Consider the chemistry department and the psychology department.

<table>
<thead>
<tr>
<th></th>
<th>Chemistry</th>
<th>Psychology</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(A \mid \text{female})$</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>$P(A \mid \text{male})$</td>
<td>0.5</td>
<td>0.25</td>
</tr>
</tbody>
</table>

What is going on? (These figures are made up.)
Example: the Berkeley gender bias case (cont’d)

<table>
<thead>
<tr>
<th></th>
<th>Chemistry</th>
<th>Psychology</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(A \mid F)$</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>$P(A \mid M)$</td>
<td>0.5</td>
<td>0.25</td>
</tr>
</tbody>
</table>

\[
P(A \mid F) = P(A \mid F, \text{chem}) P(\text{chem} \mid F) + P(A \mid F, \text{psych}) P(\text{psych} \mid F) \\
= 0.6q_f + 0.3(1 - q_f) = 0.35.
\]

\[
P(A \mid M) = P(A \mid M, \text{chem}) P(\text{chem} \mid M) + P(A \mid M, \text{psych}) P(\text{psych} \mid M) \\
= 0.5q_m + 0.25(1 - q_m) = 0.44.
\]
Definition: Bayes’ rule

Bayes rule comes from substituting in the compositional representation of the joint probabilities in the numerator of the definition of conditional probability.

Bayes’ rule

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \]
Example: diagnostic false positive (cont’d)

Let $D = \text{“has rare disease”}$. Let $Y = \text{“positive test”}$. Then we have

\[
P(D) = .0001 \quad P(Y \mid D) = .98
\]
\[
P(\text{not-}D) = .9999 \quad P(Y \mid \text{not-}D) = 0.05
\]

\[
P(\text{not-}D \mid Y) = \frac{P(\text{not-}D)P(Y \mid \text{not-}D)}{P(\text{not-}D)P(Y \mid \text{not-}D) + P(D)P(Y \mid D)}.
\]

\[
= \frac{0.9999(0.05)}{0.9999(0.05) + 0.0001(0.98)} = 0.998.
\]

Notice the use of the Law of Total Probability in the denominator of Bayes’ rule here.
Example: the drunk batter

If our boozing ball player gets a hit in his first at bat, what is the probability that he is drunk? Recall that $A =$ “gets a hit” and $B =$ “is drunk”.

\[
P(B) = 0.2 \quad \quad \quad P(A \mid B) = 0.180
\]

\[
P(\text{not-}B) = 0.8 \quad \quad \quad P(A \mid \text{not-}B) = 0.350
\]

\[
P(B \mid A) = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(\text{not-}B)P(A \mid \text{not-}B)}.
\]

\[
= \frac{0.2(0.180)}{0.2(0.180) + 0.8(0.350)} = 0.114.
\]

The number goes down from the prior probability of being drunk (20%).
You are on a jury for a hit-and-run case involving a taxi cab. In your town there are 15% blue taxis and 85% green taxis. A witness reports that the cab in the incident was blue. However, this witness is only correct 80% of the time in foggy conditions like the night of the accident. What is the probability the culprit was a blue cab?
Let $W$ = “witness reports blue cab” and $B$ = “is a blue cab”. We want to know $P(B \mid W)$, the probability that the offending cab was blue, given that the witness says it was.

\[
P(B) = 0.15 \quad P(W \mid B) = 0.80
\]

\[
P(\text{not-}B) = 0.85 \quad P(W \mid \text{not-}B) = 0.20
\]

\[
P(B \mid W) = \frac{P(B)P(W \mid B)}{P(B)P(W \mid B) + P(\text{not-}B)P(W \mid \text{not-}B)}.
\]

\[
= \frac{0.15(0.80)}{0.15(0.80) + 0.85(0.20)} = 0.4137.
\]
Conditional probabilities can be used to describe the random processes which generate events (data) that we observe.

In light of these probability models, these observations (data) become evidence for or against different explanations (hypotheses) about how the world might be.

\[
\text{Data + Model} \rightarrow \text{Inference}
\]

We will illustrated this process via an example.
Example: two girls (version one)

You are house-sitting for a colleague. You know she has two teenage children, but you don’t know how many boys/girls. The bathroom of the main hallway is shared by the two kids and you observe several bottles of glittery lip gloss, a staple of any teenage girl’s bathroom.

What is the probability that your colleague has two girls in light of this evidence?

A tree diagram can help tabulate our joint probabilities so we can apply Bayes’ rule.
Example: two girls (version one, cont’d)

\[
\begin{align*}
\frac{1}{4} & \quad BB \\
\frac{1}{2} & \quad BG,GB \\
\frac{1}{4} & \quad GG
\end{align*}
\]

- \(\frac{1}{4} BB\) with:
  - 0: glitter gloss
  - 1: no glitter gloss

- \(\frac{1}{2} BG,GB\) with:
  - 1: glitter gloss
  - 0: no glitter gloss

- \(\frac{1}{4} GG\) with:
  - 1: glitter gloss
  - 0: no glitter gloss
You are house-sitting for a colleague. You know she has two teenage children, but you don’t know how many boys/girls. The main hallway has two bathrooms off it, one for each child. In the first bathroom you observe several bottles of glittery lip gloss, a staple of any teenage girl’s bathroom.

What is the probability that your colleague has two girls in light of this evidence?
Example: two girls (version 2, cont’d)

\[\frac{1}{4} \quad BB \quad \frac{1}{2} \quad BG, GB \quad \frac{1}{4} \quad GG\]

- \(0\) → glitter gloss
- \(1\) → no glitter gloss
- \(\frac{1}{2}\) → glitter gloss
- \(\frac{1}{2}\) → no glitter gloss
- \(\frac{1}{4}\) → glitter gloss
- \(\frac{1}{4}\) → no glitter gloss
Remark: probability models

The two girls example illustrates that the key to probabilistic thinking is being mindful of the process that produced the data we are evaluating as a way to contextualize evidence.

The answers provided by Bayes’ rule depend on the joint probabilities we start with! This is what is meant by a probability model. We speak of having “models for data” that allow us to evaluate that data to answer questions.

Different models will yield different conclusions. For example, in our first model for the glitter gloss data, not seeing glitter gloss is taken as proof-positive of a two-boy household! If we think this is not what we would have concluded in that scenario, we would not want to use that model.
A better model for glitter gloss

\[
\begin{align*}
\frac{1}{4} & \quad \text{BB} \\
& \quad \begin{cases}
0.02 & \text{glitter gloss} \\
0.98 & \text{no glitter gloss}
\end{cases} \\
\frac{1}{2} & \quad \text{BG,GB} \\
& \quad \begin{cases}
0.30 & \text{glitter gloss} \\
0.70 & \text{no glitter gloss}
\end{cases} \\
\frac{1}{4} & \quad \text{GG} \\
& \quad \begin{cases}
0.85 & \text{glitter gloss} \\
0.15 & \text{no glitter gloss}
\end{cases}
\end{align*}
\]
We can apply the Law of Total Probability to several outcomes at once using some special notation.

Recall that we write \( p(x) \) as shorthand for \( P(X = x) \). Then, if we apply the conditional probability formula to each possible value of \( x \) individually, we get the conditional distribution.

If we were conditioning on an event \( A \), for example, we could use the LTP to write

\[
p(x) = p(x \mid A)P(A) + p(x \mid \text{not-}A)P(\text{not-}A).
\]
Example: at-bat outcomes

What is the probability of getting a single, double, triple or home run GIVEN that you got a hit?

<table>
<thead>
<tr>
<th>Event</th>
<th>x</th>
<th>( P(X = x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out</td>
<td>0</td>
<td>( \frac{820}{1000} )</td>
</tr>
<tr>
<td>Base hit</td>
<td>1</td>
<td>( \frac{115}{1000} )</td>
</tr>
<tr>
<td>Double</td>
<td>2</td>
<td>( \frac{33}{1000} )</td>
</tr>
<tr>
<td>Triple</td>
<td>3</td>
<td>( \frac{8}{1000} )</td>
</tr>
<tr>
<td>Home run</td>
<td>4</td>
<td>( \frac{24}{1000} )</td>
</tr>
</tbody>
</table>

To apply the conditional probability formula to each individual outcome we divide each probability by \( P(\text{hit}) = (115 + 33 + 8 + 24)/1000 \).
Example: at-bat outcomes (cont’d)

The resulting table gives a **conditional distribution**.

<table>
<thead>
<tr>
<th>Event</th>
<th>x</th>
<th>$P(X = x \mid X \neq 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base hit</td>
<td>1</td>
<td>$\frac{115}{180}$</td>
</tr>
<tr>
<td>Double</td>
<td>2</td>
<td>$\frac{33}{180}$</td>
</tr>
<tr>
<td>Triple</td>
<td>3</td>
<td>$\frac{8}{180}$</td>
</tr>
<tr>
<td>Home run</td>
<td>4</td>
<td>$\frac{24}{180}$</td>
</tr>
</tbody>
</table>

The relative likelihoods do not change, we have just rescaled the numbers so they sum to 1.
An important special case of the conditional probability formula is when the conditional probability of one event, given the other, is the same as the unconditional probability. In this case we say the events are independent.

See OpenIntro 2.1.6.
Definition: Independence

Events $A$ and $B$ are independent if

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

Intuitively, knowing that $B$ occurred imparts no information to change the probability that $A$ occurred.
Definition: Independence

An important consequence of the probability definition of independence is that.

Probability of independent events multiply

If events $A$ and $B$ are independent, then

$$P(A \& B) = P(A)P(B).$$

From this we can see that independence is symmetric, because the order we multiply events does not change the answer.
### Example: subsequent rolls of a die

We roll a die once, record the result, and then roll it again.

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roll a 6 followed by a 3</td>
<td>$\frac{1}{6} \times \frac{1}{6}$</td>
</tr>
<tr>
<td>Roll a 6 followed by any even number.</td>
<td>$\frac{1}{6} \times \frac{1}{2}$</td>
</tr>
<tr>
<td>Roll a number divisible by 3 followed by a number not divisible by three.</td>
<td>$\frac{1}{3} \times \frac{2}{3}$</td>
</tr>
<tr>
<td>Roll a number greater than 1 followed by a number less than 3.</td>
<td>$\frac{5}{6} \times \frac{1}{3}$</td>
</tr>
</tbody>
</table>
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 (cont’d)

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

What is the probability of that sequence of 5 reds?

Would this logic apply to sunshine after 5 consecutive days of rain?
Example: the birthday problem

In a room of 25 people, is it more or less likely that no one has the same birthday or that at least two people share a common birthday?

Assume that the probability of being born on a given day are $\frac{1}{365}$ for any day.

$$P(\text{no match}) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365-24}{365} = 0.43.$$  

When events are believed to be independent, the probability model becomes simpler.
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?
Example: diagnostic testing (two positive results, cont’d)

Let $D$ = “has rare disease”. Let $T$ = “two positive tests”. Then we have

\[
P(D) = 0.0001 \quad \quad \quad P(T \mid D) = 0.98^2
\]

\[
P(\text{not}-D) = 0.9999 \quad \quad \quad P(T \mid \text{not}-D) = 0.05^2
\]

\[
P(\text{not}-D \mid T) = \frac{P(\text{not}-D)P(T \mid \text{not}-D)}{P(\text{not}-D)P(T \mid \text{not}-D) + P(D)P(T \mid D)}.
\]

\[
= \frac{0.9999(0.05^2)}{0.9999(0.05^2) + 0.0001(0.98^2)} = 0.963.
\]
Example: diagnostic testing (two positives, one negative)

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, if two independent tests come back positive but a third comes back negative?
Example: diagnostic testing (two positives, one negative, cont’d)

Let $D$ = “has rare disease”. Let $M$ = “two positive tests, one negative test”. Then we have

$$P(D) = 0.0001$$
$$P(\text{not-}D) = 0.9999$$

$$P(M \mid D) = 0.98^2(1 - 0.98) = 0.98^2(0.02)$$
$$P(M \mid \text{not-}D) = 0.05^2(1 - 0.05) = 0.05^2(0.95)$$

$$P(\text{not-}D \mid M) = \frac{P(\text{not-}D)P(M \mid \text{not-}D)}{P(\text{not-}D)P(M \mid \text{not-}D) + P(D)P(M \mid D)}.$$

$$= \frac{0.9999(0.05^2)(0.95)}{0.9999(0.05^2)(0.95) + 0.0001(0.98^2)(0.02)} = 0.9992.$$
Example: diagnostic testing \((m\) positive tests, \(n\) negative tests\)

Let \(E_{m,n}\) denote the event of \(m\) positive tests and \(n\) negative tests. What is the probability of no-disease given this evidence?

\[
P(\text{not-}D \mid E_{m,n}) = \frac{P(\text{not-}D)P(E_{m,n} \mid \text{not-}D)}{P(\text{not-}D)P(E_{m,n} \mid \text{not-}D) + P(D)P(E_{m,n} \mid D)}.
\]

\[
= \frac{0.9999(0.05^m)(0.95^n)}{0.9999(0.05^m)(0.95^n) + 0.0001(0.98^m)(0.02^n)}.
\]
We can extract more than one number from events in our sample space to make a multivariate random variable.

\[ X = (X_1, X_2) \]

<table>
<thead>
<tr>
<th>Event</th>
<th>((x_1, x_2))</th>
<th>(P(X_1 = x_1, X_2 = x_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>(2,1)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(1,2)</td>
<td>(3,2)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(1,3)</td>
<td>(4,3)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(2,1)</td>
<td>(3,2)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(2,2)</td>
<td>(4,4)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(2,3)</td>
<td>(5,6)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(3,1)</td>
<td>(4,3)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(3,2)</td>
<td>(5,6)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(3,3)</td>
<td>(6,9)</td>
<td>(\frac{1}{9})</td>
</tr>
</tbody>
</table>

For example, consider rolling two three-sided dice. Let \(X_1\) be the sum of the two rolls and let \(X_2\) be the product. For simplicity, these notes focus on bivariate random variables.
Joint distribution

The probabilities are now assigned to *n-tuples* or *vectors*.

\[
\mathbf{X} = (X_1, X_2)
\]

<table>
<thead>
<tr>
<th>Event</th>
<th>((x_1, x_2))</th>
<th>(P(X_1 = x_1, X_2 = x_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>(2,1)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(1,2) or (2,1)</td>
<td>(3,2)</td>
<td>(\frac{2}{9})</td>
</tr>
<tr>
<td>(1,3) or (3,1)</td>
<td>(4,3)</td>
<td>(\frac{2}{9})</td>
</tr>
<tr>
<td>(2,2)</td>
<td>(4,4)</td>
<td>(\frac{1}{9})</td>
</tr>
<tr>
<td>(2,3) or (3,2)</td>
<td>(5,6)</td>
<td>(\frac{2}{9})</td>
</tr>
<tr>
<td>(3,3)</td>
<td>(6,9)</td>
<td>(\frac{1}{9})</td>
</tr>
</tbody>
</table>
Joint distribution

Joint distributions can be visualized in a variety of ways. Here is a bubble chart.

Joint distribution of $X$
A **marginal distribution** is a list of the possible values of one variable and the corresponding probabilities.

### Joint distribution

\[ X = (X_1, X_2) \]

<table>
<thead>
<tr>
<th>( (x_1, x_2) )</th>
<th>( P(X_1 = x_1, X_2 = x_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2,1)</td>
<td>( \frac{1}{9} )</td>
</tr>
<tr>
<td>(3,2)</td>
<td>( \frac{2}{9} )</td>
</tr>
<tr>
<td>(4,3)</td>
<td>( \frac{2}{9} )</td>
</tr>
<tr>
<td>(4,4)</td>
<td>( \frac{1}{9} )</td>
</tr>
<tr>
<td>(5,6)</td>
<td>( \frac{2}{9} )</td>
</tr>
<tr>
<td>(6,9)</td>
<td>( \frac{1}{9} )</td>
</tr>
</tbody>
</table>

### Marginal distribution

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( P(X_1 = x_1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( \frac{1}{9} )</td>
</tr>
<tr>
<td>3</td>
<td>( \frac{2}{9} )</td>
</tr>
<tr>
<td>4</td>
<td>( \frac{3}{9} )</td>
</tr>
<tr>
<td>5</td>
<td>( \frac{2}{9} )</td>
</tr>
<tr>
<td>6</td>
<td>( \frac{1}{9} )</td>
</tr>
</tbody>
</table>

These probabilities are obtained by applying the Law of Total Probability — summing over the possible values of the other variable.
Conditional distribution

For a fixed value of one variable, we can ask about the **conditional distribution** of the other variable.

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>$X_1 = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_2$</td>
<td>$P(X_2 = x_2</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{2}{3}$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{1}{3}$</td>
</tr>
</tbody>
</table>

In this case, for all other values of $X_1$, the conditional distribution of $X_2 | X_1$ puts all its mass on a single value.
“We choose to [do these] things, not because they are easy, but because they are hard…”

Consider drawing words from this sentence uniformly at random.

Let $X_1$ be the length of the selected word. Let $X_2$ be the proportion of consonants in the selected word ($\#$ consonants / word length).

What is the distribution of $\mathbf{X} = (X_1, X_2)$?
Kennedy quote

\[ \mathbf{X} = (X_1, X_2) \]

<table>
<thead>
<tr>
<th>( (x_1, x_2) )</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>((2, \frac{1}{2}))</td>
<td>3</td>
</tr>
<tr>
<td>((3, \frac{1}{3}))</td>
<td>2</td>
</tr>
<tr>
<td>((3, \frac{2}{3}))</td>
<td>2</td>
</tr>
<tr>
<td>((4, \frac{1}{4}))</td>
<td>1</td>
</tr>
<tr>
<td>((4, \frac{1}{2}))</td>
<td>2</td>
</tr>
<tr>
<td>((4, \frac{3}{4}))</td>
<td>1</td>
</tr>
<tr>
<td>((5, \frac{3}{5}))</td>
<td>1</td>
</tr>
<tr>
<td>((6, \frac{1}{2}))</td>
<td>1</td>
</tr>
<tr>
<td>((6, \frac{5}{6}))</td>
<td>1</td>
</tr>
<tr>
<td>((7, \frac{3}{7}))</td>
<td>2</td>
</tr>
</tbody>
</table>
Kennedy quote
The marginal distribution of $X_1$ is

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

This table arises from the previous one by adding up the counts which share a given $X_1$ value. To get probabilities, we divide the counts by their total sum so that the total probability adds to one.
The conditional distribution of $X_2$ given that $X_1 = 4$ is

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{4}$</td>
<td>1</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>2</td>
</tr>
<tr>
<td>$\frac{3}{4}$</td>
<td>1</td>
</tr>
</tbody>
</table>

This table arises from the previous one by restricting our attention to cells where $X_1 = 4$. To get probabilities, we divide the counts by their total sum so that the total probability adds to one.
It is often convenient to represent a joint distribution in terms of a set of conditional distributions and marginal distributions directly.

Recall our drunk baseball player. Let $X_1$ take values $\{0, 1, 2\}$ corresponding to whether he is sober, hung-over, or drunk.

Assume $X_1$ has the following marginal distribution

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$P(X_1 = x_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Drunk batter

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$P(X_1 = x_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

In addition to this marginal distribution, let $X_2$ given $X_1 = x_1$ have a Binomial($5, p_{x_1}$) distribution, for $(p_0, p_1, p_2) = (0.310, 0.200, 0.160)$.

From this information we can construct a joint distribution using the expression $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2 | X_1 = x_1)$. 
Drunk batter
Independent random variables

Two random variables $X_1$ and $X_2$ are independent if $P(X_1 \mid X_2 = x_2) = P(X_1)$ for every value of $x_2$.

If every conditional distribution of $X_1$ given $X_2$ is equal to the marginal distribution of $X_1$, then $X_1$ and $X_2$ are independent. In symbols: $X_1 \perp \! \! \! \! \perp X_2$.

Independence is symmetric: we can switch the roles of $X_1$ and $X_2$ in this definition.

None of our examples so far have been independent.
Drunk batter

If $p_1 = p_2 = p_3 = 0.310$, then $X_1$ and $X_2$ would be independent. Why?

All the columns (rows) are rescaled versions of one another. The scale is set by $P(X_1)$ (resp. $P(X_2)$).
Dice example

Consider rolling two six-sided dice. Let $X_1$ be either the outcome of the first roll, or 3, whichever is smaller. Let $X_2$ be the outcome of the second roll, or 3, which is larger.

Here are the marginal distributions.

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$P(X_1 = x_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{2}{3}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x_2$</th>
<th>$P(X_2 = x_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>5</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{6}$</td>
</tr>
</tbody>
</table>
Dice example

Here is the corresponding joint distribution.

From this we can verify that $X_1 \perp \perp X_2$. 
Features of conditional and marginal distributions

Conditional and marginal distributions possess all of the same summary features that any distribution does and which we have already discussed: mean, mode, median, variance, etc.

The definitions only change in that we use the relevant conditional or marginal distribution anywhere probabilities are called for.
Conditional mean

For example, in the Kennedy quote example we can ask “what is the average proportion of constants among four letter words?”

\[ X_2 \mid X_1 = 4 \]

<table>
<thead>
<tr>
<th>( x_2 )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{4} )</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>( \frac{1}{2} )</td>
<td>( \frac{1}{2} )</td>
</tr>
<tr>
<td>( \frac{3}{4} )</td>
<td>( \frac{1}{4} )</td>
</tr>
</tbody>
</table>

\[
E(X_2 \mid X_1 = 4) = \sum_{x_2} \left( x_2 \cdot P(X_2 = x_2 \mid X_1 = 4) \right)
\]

\[
= \frac{1}{4} \left( \frac{1}{4} \right) + \frac{1}{2} \left( \frac{1}{2} \right) + \frac{3}{4} \left( \frac{1}{4} \right)
\]

\[
= \frac{1}{2}.
\]
See OpenIntro section 2.5 and 2.6.5.

When measurements/observations can take on continuous (rather than discrete) quantities, we have to assign probabilities to ranges of outcomes instead of individual outcomes.

Example: spinning a wheel of fortune.
Discrete quantities

So far we have focused on random variables which can take on a **discrete** set of values:

- number of hits in a game,
- number of wins in a series,
- number of babies born,
- temperature (in units of one degree).
Continuous quantities

It is useful to consider continuous quantities which can, in principle, be measured with arbitrary precision: between any two values we can always find yet another value. Examples are

- temperature,
- height,
- annual rainfall,
- speed of a fastball,
- blood alcohol level.
Example: NBA player heights

Consider the heights of NBA players measured with perfect precision. Assume the heights can be any value between 60 and 100 inches.

Trick question: what is the probability that a randomly selected hypothetical player is 6 foot 8 inches tall?
In fact, for a continuous random variable $X$, the probability of taking on any specific value is zero!

Instead, we associate probabilities to intervals. The probability of $X$ taking some value in a given region is equal to the area under the curve over that region.
Probability is assigned to intervals

For example, the area of the highlighted region corresponds to $P(72 < X < 84)$ where $X =$ “height of NBA player”.

The curve describing the distribution of probability mass is called the **probability density function**.
Mental picture: throwing darts

Regions of taller density are more likely to be observed. The proportion of darts landing within a region will approach the area under the curve as we throw more and more darts.

Darts landing above the curve are discarded entirely.
Definition: density functions

A density function must satisfy two properties in order to yield valid probabilities for any collection of intervals.

Probability density function

A function $f(x)$ is a probability density function if it satisfies the following two properties:

- $f(x) \geq 0$ for any $x$,
- the total area under the curve defined by $f(x)$ is equal to one.
Example: uniform distribution

Consider randomly selected points between 0 and 1/2 where any point is as likely as any other. What must be the height of the density?

Hint: we know the total area must sum to one.
Example: uniform distribution (cont’d)

In simple cases, determining probabilities amounts to calculating areas of geometric shapes with known formulas. The highlighted region depicts the set \( A = \{ x \mid \frac{2}{8} \leq x \leq \frac{3}{8} \} \).

What is \( P(A) \)?
Example: uniform distribution (cont’d)

We can consider disjoint intervals as well, in which case the probabilities add.

The set not-$A$ is shaded. It is straightforward to confirm that $P(\text{not}-A) = 1 - P(A)$ as it should.
Example: triangular distribution

The shaded region here is the set \( B = \{x \mid 0.2 \leq x \leq 0.4\} \).

What is \( P(B) \)? Recall that the area of a triangle is \( \frac{1}{2} \) (base \( \times \) height).
Definition: cumulative distribution function

Another way to visualize the distribution of a continuous random variable is to plot $P(X \leq x)$ across all possible $x$ values.

The CDF of the triangular density from the previous slide is shown above.
Definition: cumulative distribution functions

Cumulative distribution function

\[ F(x) = P(X \leq x) \] describes the distribution of a random variable \( X \) if

- \( 0 \leq F(x) \leq 1 \) for all \( x \),
- \( F(x) \) is increasing in \( x \).
Computing probabilities with CDFs

We can compute the probability of any interval via the expression:

\[ P(a \leq X \leq b) = F(b) - F(a). \]
PDFs and CDFs

Probability density functions and cumulative distribution functions are two different ways to represent the distribution of a continuous random variable.

PDF of triangular distribution

\[ f(x) = \begin{cases} 
4x & \text{if } x \leq \frac{1}{2} \\
4 - 4x & \text{if } x > \frac{1}{2}.
\end{cases} \]

CDF of triangular distribution

\[ F(x) = \begin{cases} 
2x^2 & \text{if } x \leq \frac{1}{2} \\
1 - 2(1 - x)^2 & \text{if } x > \frac{1}{2}.
\end{cases} \]

PDFs are better for visualization while CDFs are better for determining specific probabilities.
Mental image: throwing darts and CDFs

We can think of CDFs as transforming uniform random variables – dart throws against the vertical axis – to non-uniform random variables.

Steeper portions of the CDF have greater probability and correspond to taller regions of the density.
Unbounded values

The transformation viewpoint is especially helpful for thinking about random variables with **unbounded** values.
Unbounded values (cont’d)

The CDF gets closer and closer to 1 as $x$ gets larger and larger.

For any $x$, no matter how large, $P(X > x) > 0$. 
Unbounded values (cont’d)

This is what the corresponding PDF looks like. We shall revisit this particular bell-shaped curve in depth.

For any $x$, $f(x) > 0$: that is, strictly greater than 0.
Remarks: continuity and unboundedness

Do we really believe in infinitely precise and arbitrarily large measurements?

Probably not. However, working with continuous and/or unbounded random variables allows us to avoid two potentially difficult questions:

- What amount of precision do we have? What amount of discretization do we use?
- What is the absolutely largest (respectively, smallest) value we could possibly see?

The set-up is subtler, but by working with density functions instead of a finite list of probabilities we avoid having to answer these questions.
Example: NBA heights

Consider the distribution over heights of NBA players given by the following density function.

What is the probability of a randomly selected player being within 3 inches of 6.5 feet tall?
Example: NBA heights (cont’d)

Using the CDF to compute we find that

\[ P(75 \leq X \leq 81) = 0.59. \]
Example: NBA heights (cont’d)

This distribution is unbounded above. For example,

- a 6.7% chance of a randomly selected player being over 7 foot tall.
- a 1% chance of being over 7 foot 3 inches.
- a 0.00029% chance of being over 8 foot tall.
- \( P(X > 15 \text{ feet}) > 0 \), etc.
Summarizing dependence

Joint distributions provide us with an additional ability to characterize relationships between two measurements.

Comprehensive information is captured by the conditional distributions themselves. But, as with the mean, median, mode and variance, we may want a more parsimonious description of the relationship between two quantities.
Example: NBA heights and weights

![Graph showing the relationship between height and weight in NBA players.](image)

- Height in inches
- Weight in Kilograms
Standardizing

In order to summarize the relationship between two random variables $X$ and $Y$ we begin by **standardizing** them so they share a common scale. We do this by defining two new random variables

$$Z_X = \frac{X - \text{E}(X)}{\sqrt{\text{V}(X)}}$$

and

$$Z_Y = \frac{Y - \text{E}(Y)}{\sqrt{\text{V}(Y)}}.$$

It is a straightforward calculation to verify that $\text{E}(Z_X) = \text{E}(Z_Y) = 0$ and $\text{V}(Z_X) = \text{V}(Z_Y) = 1.$
The shape is the same, but the axis labels have changed so the probability mass is centered at the origin.
Correlation

The **correlation** between two random variables $X$ and $Y$ is

$$
\rho = \text{cor}(X, Y) = E(Z_X Z_Y),
$$

where $Z_X$ and $Z_Y$ are standardized versions of $X$ and $Y$.

Recall that

$$
E(Z_X Z_Y) = \sum_{z_X, z_Y} z_X z_Y P(Z_X = z_X, Z_Y = z_Y).
$$

Correlation is the expectation of the product of two standardized random variables.
Correlation facts

- Correlation is always between -1 and 1.

- Correlation is closer to 1 when big values of one variable frequently co-occur with big values of the other variable and small values of one variables frequently co-occur with small values of the other variable.

- Correlation is closer to -1 when big values of one variable frequently co-occur with small values of the other variable and vice-versa.
The correlation is the slope of the line which best* predicts $Y$ given $X$ or $X$ given $Y$ (in standardized units). Here $\rho = 0.85$. 

*Note: Predict weight given height
*Note: Predict height given weight
The line which best predicts $Y$, given $X$ (in the original units), in the sense of minimizing

$$E((a + bX - Y)^2),$$

has slope

$$b = \rho \frac{\sigma_Y}{\sigma_X}$$

and intercept

$$a = \mu_Y - b \mu_X.$$
First dice example

Joint distribution of $X$

$\rho = 0.96$
Kennedy quote
Drunk batter

\[ \rho = -0.3 \]
Drunk batter (independent)

If batter performance is unimpacted by alcohol

\[ \rho = 0 \]
Second dice example

\[ \rho = 0 \]
Independence implies zero correlation

These last two examples demonstrate that independent random variables have zero correlation.

This can be deduced from the definition directly, by making the substitution $P(X_1 = x_1, X_2 = x_2) \rightarrow P(X_1 = x_1)P(X_2 = x_2)$ and simplifying.

However, zero correlation does not imply independence…
\( \rho = 0 \) does not imply independence

Consider the following random variable:

\[
X = (X_1, X_2)
\]

<table>
<thead>
<tr>
<th>((x_1, x_2))</th>
<th>(P(X_1 = x_1, X_2 = x_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>(\frac{1}{12})</td>
</tr>
<tr>
<td>(1,2)</td>
<td>(\frac{2}{12})</td>
</tr>
<tr>
<td>(1,3)</td>
<td>(\frac{1}{12})</td>
</tr>
<tr>
<td>(2,4)</td>
<td>(\frac{1}{12})</td>
</tr>
<tr>
<td>(2,5)</td>
<td>(\frac{2}{12})</td>
</tr>
<tr>
<td>(2,6)</td>
<td>(\frac{1}{12})</td>
</tr>
<tr>
<td>(3,1)</td>
<td>(\frac{1}{12})</td>
</tr>
<tr>
<td>(3,2)</td>
<td>(\frac{2}{12})</td>
</tr>
<tr>
<td>(3,3)</td>
<td>(\frac{1}{12})</td>
</tr>
</tbody>
</table>
\( \rho = 0 \) does not imply independence

Because the conditional distributions differ from the marginal distributions, we can immediately observe that \( X_1 \) and \( X_2 \) are not independent.