Basic Probability Theory (I)

Intro to Bayesian Data Analysis & Cognitive Modeling

Adrian Brasoveanu

[partly based on slides by Sharon Goldwater & Frank Keller and John K. Kruschke]

Fall 2012 · UCSC Linguistics

Terminology

Terminology for probability theory:

- *experiment:* process of observation or measurement; e.g., coin flip;
- *outcome:* result obtained through an experiment; e.g., coin shows tails;
- *sample space:* set of all possible outcomes of an experiment; e.g., sample space for coin flip: $S = \{\mathrm{H}, \mathrm{T}\}$.

Sample spaces can be finite or infinite.

Example: Finite Sample Space

Roll two dice, each with numbers 1–6. Sample space:

$$S_1 = \{\langle x, y\rangle : x \in \{1, 2, \ldots, 6\} \wedge y \in \{1, 2, \ldots, 6\}\}$$

Alternative sample space for this experiment – sum of the dice:

$$S_2 = \{x + y : x \in \{1, 2, \ldots, 6\} \wedge y \in \{1, 2, \ldots, 6\}\}$$

$$S_2 = \{z : z \in \{2, 3, \ldots, 12\}\} = \{2, 3, \ldots, 12\}$$

Example: Infinite Sample Space

Flip a coin until heads appears for the first time:

$$S_3 = \{\mathrm{H}, \mathrm{TH}, \mathrm{TTH}, \mathrm{TTTH}, \mathrm{TTTTH}, \ldots\}$$

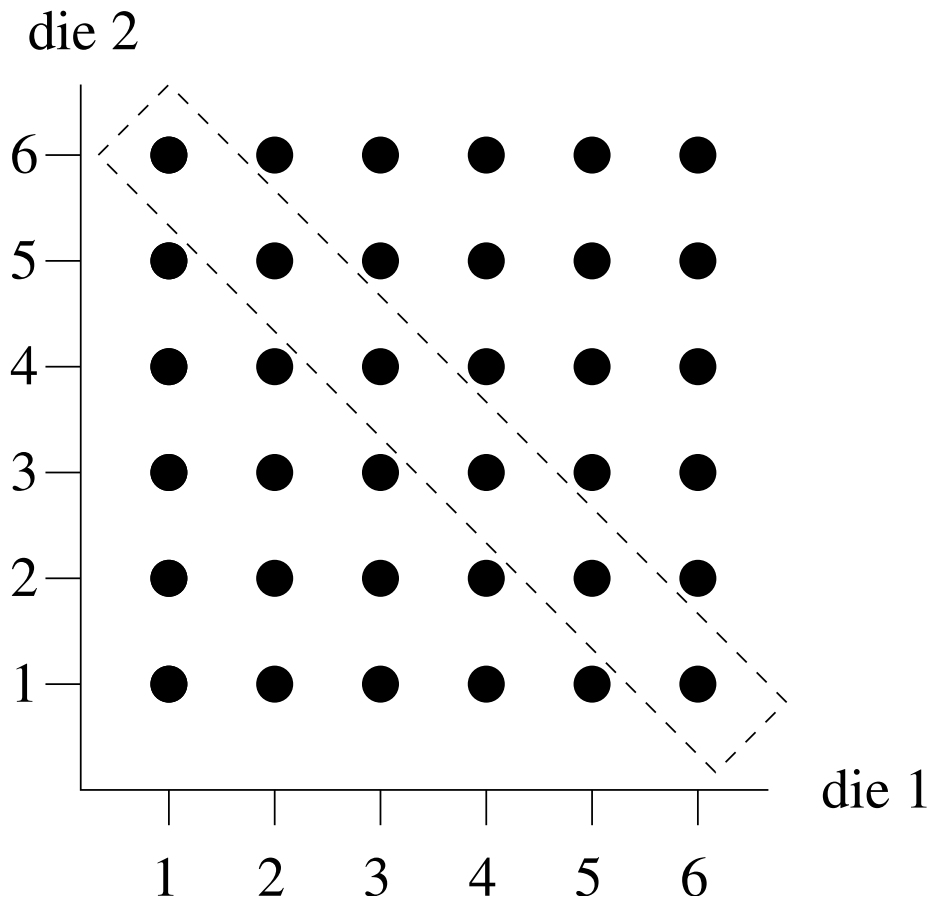
Often we are not interested in individual outcomes, but in events. An *event* is a subset of a sample space.

Example

With respect to S_1, describe the event B of rolling a total of 7 with the two dice.

$$B = \{\langle 1,6\rangle, \langle 2,5\rangle, \langle 3,4\rangle, \langle 4,3\rangle, \langle 5,2\rangle, \langle 6,1\rangle\}$$

The event B can be represented graphically:

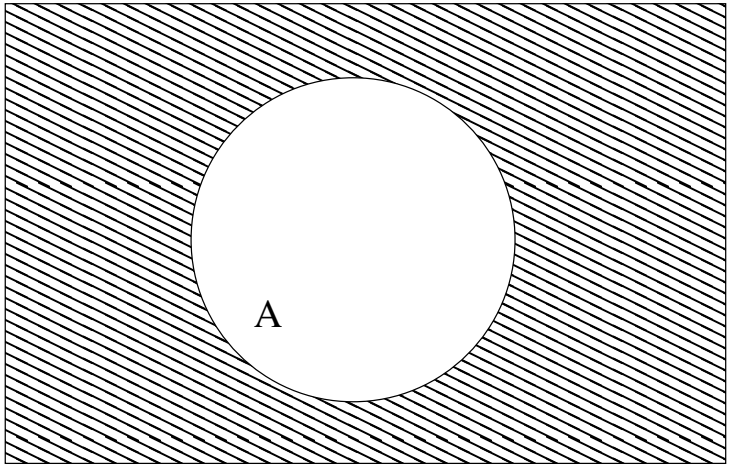


Often we are interested in combinations of two or more events.
This can be represented using set theoretic operations.
Assume a sample space S and two events A and B:

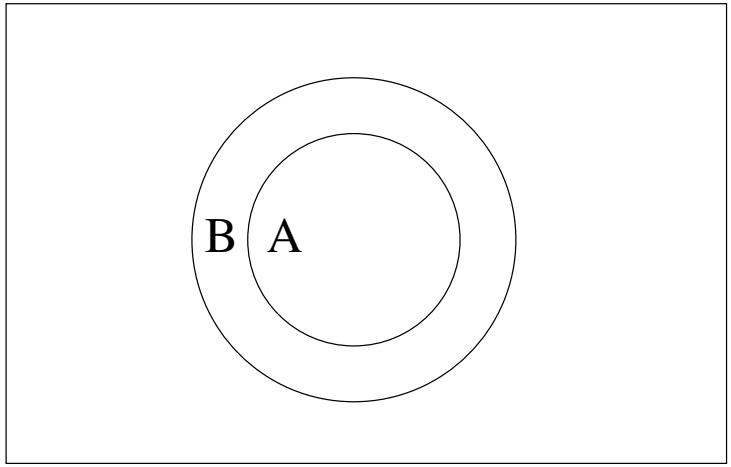
- *complement* $\overline{A}$ *(also* A'*):* all elements of S that are not in A;
- *subset* $A \subseteq B$*:* all elements of A are also elements of B;
- *union* $A \cup B$*:* all elements of S that are in A or B;
- *intersection* $A \cap B$*:* all elements of S that are in A and B.

These operations can be represented graphically using *Venn diagrams.*

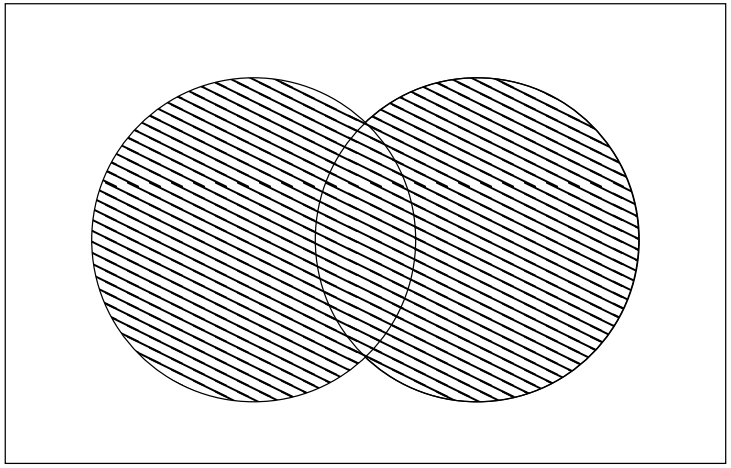
Venn Diagrams



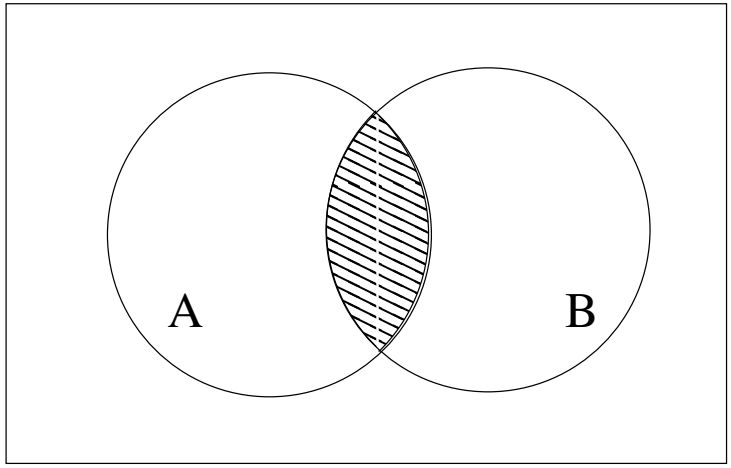
$\bar{A}$



$A \subseteq B$



$A \cup B$



$A \cap B$

Events are denoted by capital letters A, B, C, etc. The *probability* of an event A is denoted by $p(A)$.

Axioms of Probability

1. The probability of an event is a nonnegative real number: $p(A) \geq 0$ for any $A \subseteq S$.
2. $p(S) = 1$.
3. If $A_1, A_2, A_3, \ldots,$ is a set of mutually exclusive events of S, then:

$$p(A_1 \cup A_2 \cup A_3 \cup \ldots) = p(A_1) + p(A_2) + p(A_3) + \ldots$$

Theorem: Probability of an Event

If A is an event in a sample space S and $O_1, O_2, \ldots, O_n$, are the individual outcomes comprising A, then $p(A) = \sum_{i=1}^{n} p(O_i)$

Example

Assume all strings of three lowercase letters are equally probable. Then what's the probability of a string of three vowels?

There are 26 letters, of which 5 are vowels. So there are $N = 26^3$ three letter strings, and $n = 5^3$ consisting only of vowels. Each outcome (string) is equally likely, with probability $\frac{1}{N}$, so event A (a string of three vowels) has probability $p(A) = \frac{n}{N} = \frac{5^3}{26^3} \approx 0.00711$.

Theorems: Rules of Probability

1. If A and $\overline{A}$ are complementary events in the sample space S, then $p(\overline{A}) = 1 - p(A)$.
2. $p(\emptyset) = 0$ for any sample space S.
3. If A and B are events in a sample space S and $A \subseteq B$, then $p(A) \leq p(B)$.
4. $0 \leq p(A) \leq 1$ for any event A.

Addition Rule

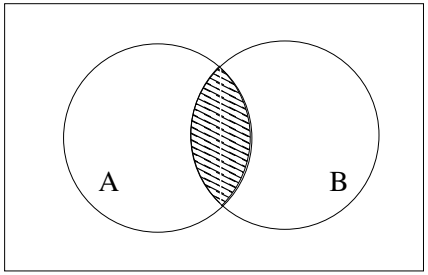
Axiom 3 allows us to add the probabilities of mutually exclusive events. What about events that are not mutually exclusive?

Theorem: General Addition Rule

If A and B are two events in a sample space S, then:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Ex: A = "has glasses", B = "is blond".
$p(A) + p(B)$ counts blondes with glasses twice, need to subtract once.



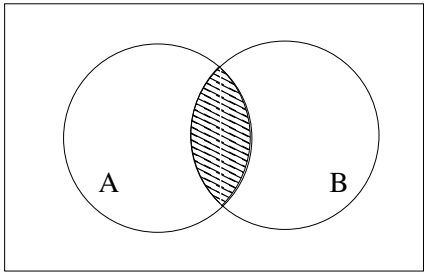
Definition: Conditional Probability, Joint Probability

If A and B are two events in a sample space S, and $p(A) \neq 0$ then the *conditional probability* of B given A is:

$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

$p(A \cap B)$ is the *joint probability* of A and B, also written $p(A, B)$.

Intuitively, $p(B|A)$ is the probability that B will occur given that A has occurred.
Ex: The probability of being blond given that one wears glasses: $p(\text{blond}|\text{glasses})$.



Example

A manufacturer knows that the probability of an order being ready on time is 0.80, and the probability of an order being ready on time and being delivered on time is 0.72.

What is the probability of an order being delivered on time, given that it is ready on time?

R: order is ready on time; D: order is delivered on time. $p(R) = 0.80$, $p(R, D) = 0.72$. Therefore:

$$p(D|R) = \frac{p(R, D)}{p(R)} = \frac{0.72}{0.80} = 0.90$$

Example

Consider sampling an adjacent pair of words (bigram) from a large text T. Let $\mathcal{BI}$ = the set of bigrams in T (this is our sample space), A = "first word is **run**" = $\{\langle \textbf{run}, w_2\rangle : w_2 \in T\} \subseteq \mathcal{BI}$ and B = "second word is **amok**" = $\{\langle w_1, \textbf{amok}\rangle : w_1 \in T\} \subseteq \mathcal{BI}$.

If $p(A) = 10^{-3.5}$, $p(B) = 10^{-5.6}$, and $p(A, B) = 10^{-6.5}$, what is the probability of seeing **amok** following **run**, i.e., $p(B|A)$? How about **run** preceding **amok**, i.e., $p(A|B)$?

$$p(\text{"}\textbf{run} \text{ before } \textbf{amok}\text{"}) = p(A|B) = \frac{p(A,B)}{p(B)} = \frac{10^{-6.5}}{10^{-5.6}} = .126$$

$$p(\text{"}\textbf{amok} \text{ after } \textbf{run}\text{"}) = p(B|A) = \frac{p(A,B)}{p(A)} = \frac{10^{-6.5}}{10^{-3.5}} = .001$$

[How do we determine $p(A)$, $p(B)$, $p(A, B)$ in the first place?]

(Con)Joint Probability and the Multiplication Rule

From the definition of conditional probability, we obtain:

Theorem: Multiplication Rule

If A and B are two events in a sample space S and $p(A) \neq 0$, then:

$$p(A, B) = p(A)p(B|A)$$

Since $A \cap B = B \cap A$, we also have that:

$$p(A, B) = p(B)p(A|B)$$

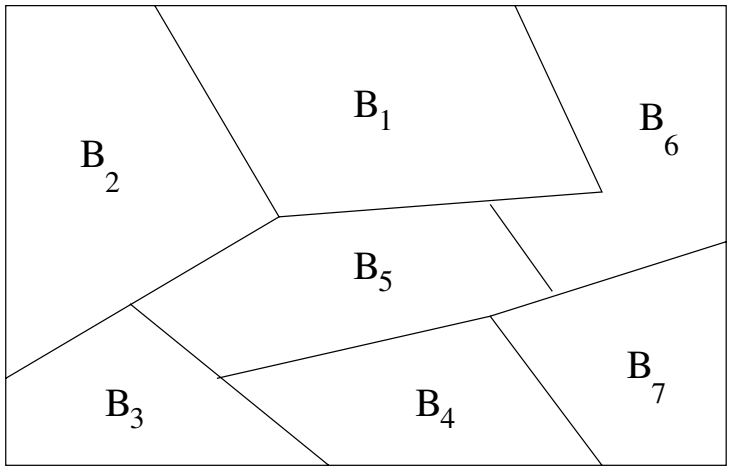
Marginal Probability and the Rule of Total Probability

Theorem: Marginalization (a.k.a. Rule of Total Probability)

If events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S and $p(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S:

$$p(A) = \sum_{i=1}^{k} p(A, B_i) = \sum_{i=1}^{k} p(A|B_i)p(B_i)$$

$B_1, B_2, \ldots, B_k$ form a *partition* of S if they are pairwise mutually exclusive and if $B_1 \cup B_2 \cup \ldots \cup B_k = S$.



Example

In an experiment on human memory, participants have to memorize a set of words (B_1), numbers (B_2), and pictures (B_3). These occur in the experiment with the probabilities $p(B_1) = 0.5$, $p(B_2) = 0.4$, $p(B_3) = 0.1$.

Then participants have to recall the items (where A is the recall event). The results show that $p(A|B_1) = 0.4$, $p(A|B_2) = 0.2$, $p(A|B_3) = 0.1$. Compute $p(A)$, the probability of recalling an item.

By the theorem of total probability:

$$\begin{aligned} p(A) &= \sum_{i=1}^{k} p(B_i)p(A|B_i) \\ &= p(B_1)p(A|B_1) + p(B_2)p(A|B_2) + p(B_3)p(A|B_3) \\ &= 0.5 \cdot 0.4 + 0.4 \cdot 0.2 + 0.1 \cdot 0.1 = 0.29 \end{aligned}$$

Example

Proportions for a sample of University of Delaware students 1974, $N = 592$. Data adapted from Snee (1974).

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Example

These are the joint probabilities $p(\textbf{eyeColor}, \textbf{hairColor})$.

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Joint, Marginal & Conditional Probability

Example

E.g., $p(\textbf{eyeColor} = \texttt{brown}, \textbf{hairColor} = \texttt{brunette}) = .20.$

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Example

These are the marginal probabilities $p($**eyeColor**$)$.

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Joint, Marginal & Conditional Probability

Example

E.g.,

$$p(\textbf{eyeColor} = \texttt{brown}) =$$

$$\sum_{\textbf{hairColor}} p(\textbf{eyeColor} = \texttt{brown}, \textbf{hairColor}) =$$

$$.12 + .20 + .01 + .04 = .37$$

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Example

These are the marginal probabilities $p($**hairColor**$)$.

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Joint, Marginal & Conditional Probability

Example

E.g., $p(\textbf{hairColor} = \texttt{brunette}) =$

$$\sum_{\textbf{eyeColor}} p(\textbf{eyeColor}, \textbf{hairColor} = \texttt{brunette}) =$$

$$.14 + .20 + .14 = .48$$

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Joint, Marginal & Conditional Probability

Example

To obtain the cond. prob. $p(\textbf{eyeColor}|\textbf{hairColor} = \texttt{brunette})$, we do two things:

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Example

To obtain the cond. prob. $p(\textbf{eyeColor}|\textbf{hairColor} = \texttt{brunette})$, we do two things:

i. reduction: we consider only the probabilities in the `brunette` column;

	hairColor				
eyeColor	black	brunette	blond	red	
blue		.14			
brown		.20			
hazel/green		.14			
		.48			

Example

To obtain the cond. prob. $p(\textbf{eyeColor}|\textbf{hairColor} = \texttt{brunette})$, we do two things:

ii. normalization: we divide by the marginal $p(\texttt{brunette})$, since all the probability mass is now concentrated here.

	hairColor				
eyeColor	black	brunette	blond	red	
blue		.14/.48			
brown		.20/.48			
hazel/green		.14/.48			
		.48			

Joint, Marginal & Conditional Probability

Example

E.g., $p(\textbf{eyeColor} = \texttt{brown} | \textbf{hairColor} = \texttt{brunette}) = .20/.48$.

	hairColor				
eyeColor	black	brunette	blond	red	
blue		.14/.48			
brown		.20/.48			
hazel/green		.14/.48			
		.48			

Example

Moreover:

$p(\textbf{eyeColor} = \texttt{brown} | \textbf{hairColor} = \texttt{brunette}) \neq$
$p(\textbf{hairColor} = \texttt{brunette} | \textbf{eyeColor} = \texttt{brown})$

Consider $p(\textbf{hairColor} | \textbf{eyeColor} = \texttt{brown})$:

	hairColor				
eyeColor	black	brunette	blond	red	
blue	.03	.14	.16	.03	.36
brown	.12	.20	.01	.04	.37
hazel/green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Joint, Marginal & Conditional Probability

Example

To obtain $p(\textbf{hairColor}|\textbf{eyeColor} = \texttt{brown})$, we reduce,

	hairColor				
eyeColor	black	brunette	blond	red	
blue					
brown	.12	.20	.01	.04	.37
hazel/green					

and we normalize.

	hairColor				
eyeColor	black	brunette	blond	red	
blue					
brown	.12/.37	.20/.37	.01/.37	.04/.37	.37
hazel/green					

Example

So $p(\textbf{hairColor} = \texttt{brunette} | \textbf{eyeColor} = \texttt{brown}) = .20/.37$,

	hairColor				
eyeColor	black	brunette	blond	red	
blue					
brown	.12/.37	.20/.37	.01/.37	.04/.37	.37
hazel/green					

but $p(\textbf{eyeColor} = \texttt{brown} | \textbf{hairColor} = \texttt{brunette}) = .20/.48$.

	hairColor				
eyeColor	black	brunette	blond	red	
blue		.14/.48			
brown		.20/.48			
hazel/green		.14/.48			
		.48			

Conditional Probability: $p(A|B)$ vs $p(B|A)$

Example 1: Disease Symptoms (from Lindley 2006)

- Doctors studying a disease D noticed that 90% of patients with the disease exhibited a symptom S.
- Later, another doctor sees a patient and notices that she exhibits symptom S.
- As a result, the doctor concludes that there is a 90% chance that the new patient has the disease D.

But: while $p(S|D) = .9$, $p(D|S)$ might be very different.

Conditional Probability: $p(A|B)$ vs $p(B|A)$

Example 2: Forensic Evidence (from Lindley 2006)

- A crime has been committed and a forensic scientist reports that the perpetrator must have attribute `P`. E.g., the DNA of the guilty party is of type `P`.
- The police find someone with `P`, who is charged with the crime. In court, the forensic scientist reports that attribute `P` only occurs in a proportion α of the population.
- Since α is very small, the court infers that the defendant is highly likely to be guilty, going on to assess the chance of guilt as $1 - \alpha$ since an innocent person would only have a chance α of having `P`.

But: while $p(\texttt{P}|\texttt{innocent}) = \alpha$, $p(\texttt{innocent}|\texttt{P})$ might be much bigger.

Conditional Probability: $p(A|B)$ vs $p(B|A)$

Example 3: Significance Tests (from Lindley 2006)

- As scientistis, we often set up a straw-man/null hypothesis. E.g., we may suppose that a chemical has no effect on a reaction and then perform an experiment which, if the effect does not exist, gives numbers that are very small.
- If we obtain large numbers compared to expectation, we say the null is rejected and the effect exists.
- "Large" means numbers that would only arise a small proportion α of times if the null hypothesis is true.
- So we say that we have confidence $1 - \alpha$ that the effect exists, and α (often .05) is the significance level of the test.

But: while $p(\texttt{effect}|\texttt{null}) = \alpha$, $p(\texttt{null}|\texttt{effect})$ might be bigger.

Relating $p(A|B)$ and $p(B|A)$

We can infer something about a disease from a symptom, but we need to do it with some care – the proper inversion is accomplished by the Bayes' rule

Bayes' Theorem

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

- Derived using mult. rule: $p(A, B) = p(A|B)p(B) = p(B|A)p(A)$.
- Denominator $p(A)$ can be computed using theorem of total probability: $p(A) = \sum_{i=1}^{k} p(A|B_i)p(B_i)$.
- Denominator is a normalizing constant: ensures $p(B|A)$ sums to 1. If we only care about relative sizes of probabilities, we can ignore it: $p(B|A) \propto p(A|B)p(B)$.

Example

Consider the memory example again. What is the probability that an item that is correctly recalled (A) is a picture (B_3)?

By Bayes' theorem:

$$\begin{aligned} p(B_3|A) &= \frac{p(B_3)p(A|B_3)}{\sum_{i=1}^{k} p(B_i)p(A|B_i)} \\ &= \frac{0.1 \cdot 0.1}{0.29} = 0.0345 \end{aligned}$$

The process of computing $p(B|A)$ from $p(A|B)$ is sometimes called *Bayesian inversion.*

Example

A fair coin is flipped three times. There are 8 possible outcomes, and each of them is equally likely.

For each outcome, we can count the number of heads and the number of switches (i.e., *HT* or *TH* subsequences):

outcome	**probability**	**#heads**	**#switches**
HHH	1/8	3	0
THH	1/8	2	1
HTH	1/8	2	2
HHT	1/8	2	1
TTH	1/8	1	1
THT	1/8	1	2
HTT	1/8	1	1
TTT	1/8	0	0

Example

The joint probability $p(\#\textbf{heads}, \#\textbf{switches})$ is therefore:

	#heads				
#switches	0	1	2	3	
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
	1/8	3/8	3/8	1/8	

Let us use Bayes' theorem to relate the two conditional probabilities:

$$p(\#\textbf{switches} = 1 | \#\textbf{heads} = 1)$$

$$p(\#\textbf{heads} = 1 | \#\textbf{switches} = 1)$$

Example

	#heads				
#switches	0	1	2	3	
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
	1/8	3/8	3/8	1/8	

Note that:

$$p(\#\textbf{switches} = 1 | \#\textbf{heads} = 1) = 2/3$$

$$p(\#\textbf{heads} = 1 | \#\textbf{switches} = 1) = 1/2$$

Example

#switches \ #heads	0	1	2	3	
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
	1/8	3/8	3/8	1/8	

The joint probability $p(\#\textbf{switches} = 1, \#\textbf{heads} = 1) = \frac{2}{8}$ can be expressed in two ways:

$$p(\#\textbf{switches} = 1 | \#\textbf{heads} = 1) \cdot p(\#\textbf{heads} = 1) = \frac{2}{3} \cdot \frac{3}{8} = \frac{2}{8}$$

Example

	#heads				
#switches	0	1	2	3	
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
	1/8	3/8	3/8	1/8	

The joint probability $p(\#\textbf{switches} = 1, \#\textbf{heads} = 1) = \frac{2}{8}$ can be expressed in two ways:

$$p(\#\textbf{heads} = 1 | \#\textbf{switches} = 1) \cdot p(\#\textbf{switches} = 1) = \frac{1}{2} \cdot \frac{4}{8} = \frac{2}{8}$$

Example

	#heads				
#switches	0	1	2	3	
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
	1/8	3/8	3/8	1/8	

Bayes' theorem is a consequence of the fact that we can reach the joint $p(\#\textbf{switches} = 1, \#\textbf{heads} = 1)$ in these two ways:

- by restricting attention to the row $\#\textbf{switches} = 1$
- by restricting attention to the column $\#\textbf{heads} = 1$

Example: Selenium and cancer (from Lindley 2006)

- A clinical trial tests the effect of a selenium-based treatment on cancer.

Example: Selenium and cancer (from Lindley 2006)

- A clinical trial tests the effect of a selenium-based treatment on cancer.
- We assume the existence of a parameter ϕ such that: if $\phi = 0$, selenium has no effect on cancer; if $\phi > 0$, selenium has a beneficial effect; finally, if $\phi < 0$, selenium has a harmful effect.
- The trial would not have been set up if the negative value was reasonably probable, i.e., $p(\phi < 0 | \texttt{cancer})$ is small.

Example: Selenium and cancer (from Lindley 2006)

- A clinical trial tests the effect of a selenium-based treatment on cancer.
- We assume the existence of a parameter ϕ such that: if $\phi = 0$, selenium has no effect on cancer; if $\phi > 0$, selenium has a beneficial effect; finally, if $\phi < 0$, selenium has a harmful effect.
- The trial would not have been set up if the negative value was reasonably probable, i.e., $p(\phi < 0|\texttt{cancer})$ is small.
- The value $\phi = 0$ is of special interest: it is the null value. The hypothesis that $\phi = 0$ is the null hypothesis.
- The non-null values of ϕ are the alternative hypothese(s), and the procedure to be developed is a test of the null hypothesis.
- The null hypothesis is a straw man that the trial attempts to reject: we hope the trial will show selenium to be of value.

Example: Selenium and cancer (from Lindley 2006)

- Assume the trial data is a single number d: the difference in recovery rates between the patients receiving selenium and those on the placebo.

Example: Selenium and cancer (from Lindley 2006)

- Assume the trial data is a single number *d*: the difference in recovery rates between the patients receiving selenium and those on the placebo.
- Before seeing the data *d* provided by the trial, the procedure selects values of *d* that in total have small probability if $\phi = 0$.
- We declare the result "significant" if the actual value of *d* obtained in the trial is one of them.
- The small probability is the significance level α. The trial is significant at the α level if the actually observed *d* is in this set.

Example: Selenium and cancer (from Lindley 2006)

- Assume the trial data is a single number *d*: the difference in recovery rates between the patients receiving selenium and those on the placebo.
- Before seeing the data *d* provided by the trial, the procedure selects values of *d* that in total have small probability if $\phi = 0$.
- We declare the result "significant" if the actual value of *d* obtained in the trial is one of them.
- The small probability is the significance level α. The trial is significant at the α level if the actually observed *d* is in this set.
- Assume the actual *d* is one of these improbable values. Since improbable events happen (very) rarely, doubt is cast on the assumption that $\phi = 0$, i.e., that the null hypothesis is true.
- That is: either an improbable event has occurred or the null hypothesis is false.

Example: Selenium and cancer (from Lindley 2006)

- The test uses only one probability α of the form $p(d|\phi = 0)$, i.e., the probability of data when the null is true.
- Importantly: α is not the probability of the actual difference d observed in the trial, but the (small) probability of the set of extreme values.
- Thus, a significance test does not use only the observed value d, but also those values that might have occurred but did not.
- Determining what might have occurred is the major source of problems with null hypothesis significance testing (NHST). See Kruschke (2011), ch. 11, for more details.

Example: Selenium and cancer (from Lindley 2006)

The test uses only $p(d|\phi = 0)$, but its goal is to make **inferences about the inverse probability** $p(\phi = 0|d)$, i.e., the probability of the null given the data. Two Bayesian ways (Kruschke 2011, ch. 12):

- Bayesian model comparison: we want the posterior odds, i.e., odds after the trial, of the null relative to the alternative(s):

$$o(\phi{=}0|d) = \frac{p(\phi{=}0|d)}{p(\phi{\neq}0|d)} = \frac{\frac{p(d|\phi{=}0)p(\phi{=}0)}{p(d)}}{\frac{p(d|\phi{\neq}0)p(\phi{\neq}0)}{p(d)}} = \frac{p(d|\phi{=}0)p(\phi{=}0)}{p(d|\phi{\neq}0)p(\phi{\neq}0)} = \frac{p(d|\phi{=}0)}{p(d|\phi{\neq}0)} o(\phi{=}0)$$

- Bayesian parameter estimation: we compute the posterior probability of all the (relevant) values of the parameter ϕ and examine it to see if the null value is credible:

compute $p(\phi|d) = \frac{p(d|\phi)p(\phi)}{p(d)}$, then check whether the null value is in the interval of ϕ values with the highest posterior probability.

Definition: Independent Events

Two events A and B are independent iff:

$$p(A, B) = p(A)p(B)$$

Intuition: two events are independent if knowing whether one event occurred does not change the probability of the other.

Note that the following are equivalent:

$$p(A, B) = p(A)p(B) \quad (1)$$
$$p(A|B) = p(A) \quad (2)$$
$$p(B|A) = p(B) \quad (3)$$

Example

A coin is flipped three times. Each of the eight outcomes is equally likely. A: heads occurs on each of the first two flips, B: tails occurs on the third flip, C: exactly two tails occur in the three flips. Show that A and B are independent, B and C dependent.

$$
\begin{array}{ll}
A = \{\mathrm{HHH}, \mathrm{HHT}\} & p(A) = \frac{1}{4} \\
B = \{\mathrm{HHT}, \mathrm{HTT}, \mathrm{THT}, \mathrm{TTT}\} & p(A) = \frac{1}{2} \\
C = \{\mathrm{HTT}, \mathrm{THT}, \mathrm{TTH}\} & p(C) = \frac{3}{8} \\
A \cap B = \{\mathrm{HHT}\} & p(A \cap B) = \frac{1}{8} \\
B \cap C = \{\mathrm{HTT}, \mathrm{THT}\} & p(B \cap C) = \frac{1}{4}
\end{array}
$$

$p(A)p(B) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} = p(A \cap B)$, hence A and B are independent.
$p(B)p(C) = \frac{1}{2} \cdot \frac{3}{8} = \frac{3}{16} \neq p(B \cap C)$, hence B and C are dependent.

Example

A simple example of two attributes that are independent: the **suit** and **value** of cards in a standard deck: there are 4 suits $\{\diamondsuit, \spadesuit, \clubsuit, \heartsuit\}$ and 13 values of each suit $\{\mathtt{2}, \cdots, \mathtt{10}, \mathtt{J}, \mathtt{Q}, \mathtt{K}, \mathtt{A}\}$, for a total of 52 cards.

Consider a randomly dealt card:

- marginal probability it's a heart:
  $p(\textbf{suit} = \heartsuit) = 13/52 = 1/4$
- conditional probability it's a heart given that it's a queen:
  $p(\textbf{suit} = \heartsuit | \textbf{value} = \mathtt{Q}) = 1/4$
- in general, $p(\textbf{suit}|\textbf{value}) = p(\textbf{suit})$, hence **suit** and **value** are independent

Example

We can verify independence by cross-multiplying marginal probabilities too. For every suit $\mathtt{s} \in \{\diamondsuit, \spadesuit, \clubsuit, \heartsuit\}$ and value $\mathtt{v} \in \{\mathtt{2}, \cdots, \mathtt{10}, \mathtt{J}, \mathtt{Q}, \mathtt{K}, \mathtt{A}\}$:

- $p(\textbf{suit} = \mathtt{s}, \textbf{value} = \mathtt{v}) = \frac{1}{52}$ (in a well-shuffled deck)
- $p(\textbf{suit} = s) = \frac{13}{52} = \frac{1}{4}$
- $p(\textbf{value} = v) = \frac{4}{52} = \frac{1}{13}$
- $p(\textbf{suit} = s) \cdot p(\textbf{value} = v) = \frac{1}{4} \cdot \frac{1}{13} = \frac{1}{52}$

Independence comes up when we construct mathematical descriptions of our beliefs about more than one attribute: to describe what we believe about combinations of attributes, we often assume independence and simply multiply the separate beliefs about individual attributes to specify the joint beliefs.

Definition: Conditionally Independent Events

Two events A and B are conditionally independent given event C iff:

$$p(A, B|C) = p(A|C)p(B|C)$$

Intuition: Once we know whether C occurred, knowing about A or B doesn't change the probability of the other.

Show that the following are equivalent:

$$p(A, B|C) = p(A|C)p(B|C) \quad (4)$$
$$p(A|B, C) = p(A|C) \quad (5)$$
$$p(B|A, C) = p(B|C) \quad (6)$$

Example

In a noisy room, I whisper the same number $n \in \{1, \ldots, 10\}$ to two people A and B on two separate occasions. A and B imperfectly (and independently) draw a conclusion about what number I whispered. Let the numbers A and B think they heard be n_a and n_b, respectively.

Are n_a and n_b independent (a.k.a. marginally independent)?
No. E.g., we'd expect $p(n_a = 1 | n_b = 1) > p(n_a = 1)$.

Are n_a and n_b conditionally independent given n? Yes: if you know the number that I actually whispered, the two variables are no longer correlated.
E.g., $p(n_a = 1 | n_b = 1, n = 2) = p(n_a = 1 | n = 2)$

Conditional Independence Example & the Chain Rule

The Anderson (1990) memory model: A is the event that an item is needed from memory; A depends on contextual cues Q and usage history H_A, but Q is independent of H_A given A.

Show that $p(A|H_A, Q) \propto p(A|H_A)p(Q|A)$.

Solution:

$$
\begin{aligned}
p(A|H_A, Q) &= \frac{p(A, H_A, Q)}{p(H_A, Q)} \\
&= \frac{p(Q|A, H_A)p(A|H_A)p(H_A)}{p(Q|H_A)p(H_A)} && \text{[chain rule]} \\
&= \frac{p(Q|A, H_A)p(A|H_A)}{p(Q|H_A)} \\
&= \frac{p(Q|A)p(A|H_A)}{p(Q|H_A)} \\
&\propto p(Q|A)p(A|H_A)
\end{aligned}
$$

Definition: Random Variable

If S is a sample space with a probability measure and X is a real-valued function defined over the elements of S, then X is called a random variable.

We symbolize random variables (r.v.s) by capital letters (e.g., X), and their values by lower-case letters (e.g., x).

Example

Given an experiment in which we roll a pair of 4-sided dice, let the random variable X be the total number of points rolled with the two dice.

E.g. $X = 5$ 'picks out' the set $\{\langle 1, 4\rangle, \langle 2, 3\rangle, \langle 3, 2\rangle, \langle 4, 1\rangle\}$.

Specify the full function denoted by X and determine the probabilities associated with each value of X.

Example

Assume a balanced coin is flipped three times. Let X be the random variable denoting the total number of heads obtained.

Outcome	Probability	x
HHH	$\frac{1}{8}$	3
HHT	$\frac{1}{8}$	2
HTH	$\frac{1}{8}$	2
THH	$\frac{1}{8}$	2

Outcome	Probability	x
TTH	$\frac{1}{8}$	1
THT	$\frac{1}{8}$	1
HTT	$\frac{1}{8}$	1
TTT	$\frac{1}{8}$	0

Hence, $p(X = 0) = \frac{1}{8}$, $p(X = 1) = p(X = 2) = \frac{3}{8}$, $p(X = 3) = \frac{1}{8}$.

Definition: Probability Distribution

If X is a random variable, the function $f(x)$ whose value is $p(X = x)$ for each value x in the range of X is called the probability distribution of X.

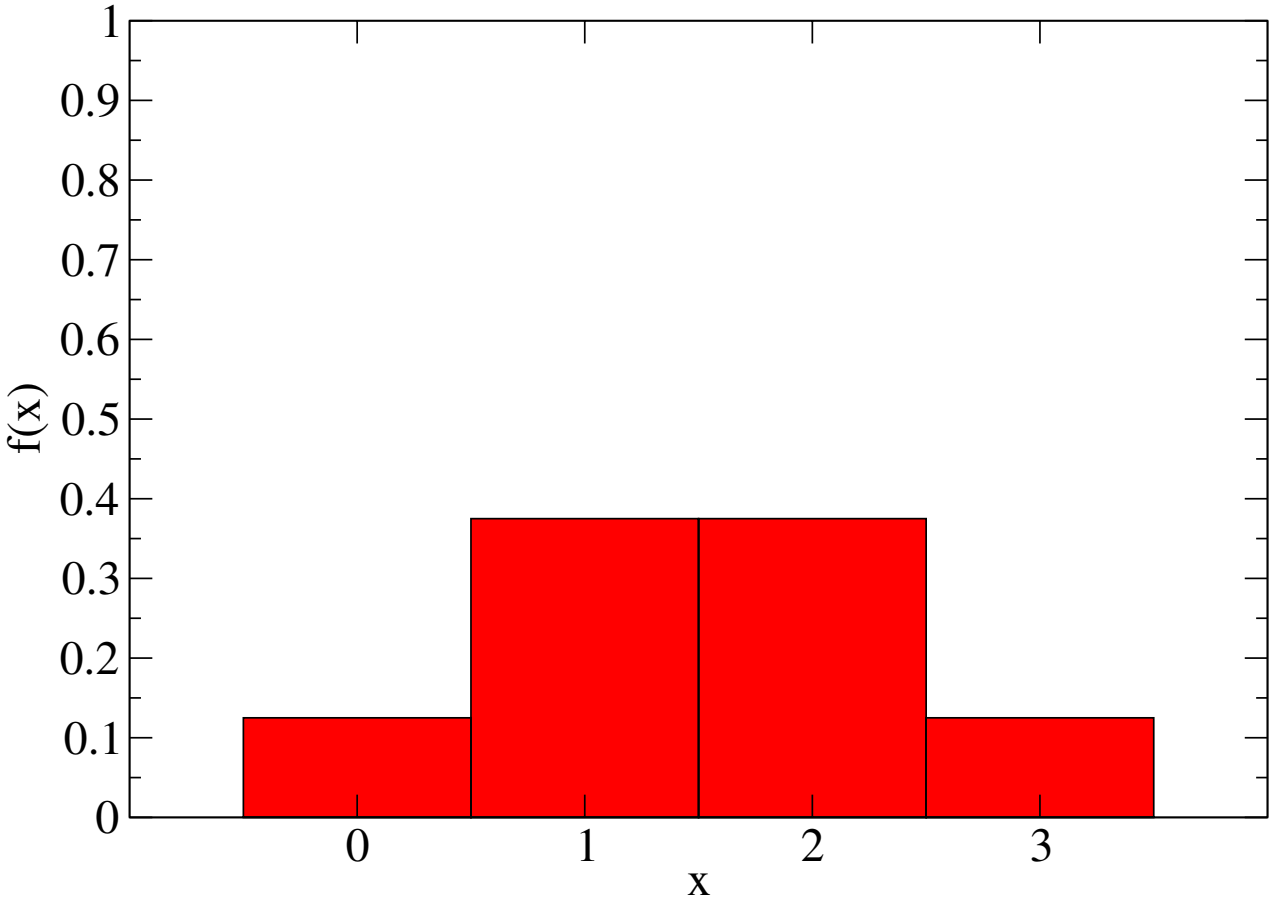
Note: the set of values x ('the support') = the domain of f = the range of X.

Example

For the probability function defined in the previous example:

x	$f(x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

A probability distribution is often represented as a *probability histogram.* For the previous example:



Any probability distribution function (or simply: probability distribution) f of a random variable X is such that:

1. $f(x) \geq 0, \forall x \in \textbf{Domain}(f)$
2. $\sum_{x \in \textbf{Domain}(f)} f(x) = 1.$

Example: geometric distribution

Let X be the number of coin flips needed before getting heads, where p_h is the probability of heads on a single flip. What is the distribution of X?

Assume flips are independent, so:

$$p(\mathrm{T}^{n-1}\mathrm{H}) = p(\mathrm{T})^{n-1}p(\mathrm{H})$$

Therefore:

$$p(X = n) = (1 - p_h)^{n-1}p_h$$

The notion of mathematical expectation derives from games of chance. It's the product of the amount a player can win and the probability of wining.

Example

In a raffle, there are 10,000 tickets. The probability of winning is therefore $\frac{1}{10{,}000}$ for each ticket. The prize is worth \$4,800.

Hence the expectation per ticket is $\frac{\$4{,}800}{10{,}000} = \$0.48$.

In this example, the expectation can be thought of as the average win per ticket.

This intuition can be formalized as the *expected value* (or *mean*) of a random variable:

Definition: Expected Value

If X is a random variable and $f(x)$ is the value of its probability distribution at x, then the expected value of X is:

$$E(X) = \sum_x x \cdot f(x)$$

Example

A balanced coin is flipped three times. Let X be the number of heads. Then the probability distribution of X is:

$$f(x) = \begin{cases} \frac{1}{8} & \text{for } x = 0 \\ \frac{3}{8} & \text{for } x = 1 \\ \frac{3}{8} & \text{for } x = 2 \\ \frac{1}{8} & \text{for } x = 3 \end{cases}$$

The expected value of X is:

$$E(X) = \sum_x x \cdot f(x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

The notion of expectation can be generalized to cases in which a function $g(X)$ is applied to a random variable X.

Theorem: Expected Value of a Function

If X is a random variable and $f(x)$ is the value of its probability distribution at x, then the expected value of $g(X)$ is:

$$E[g(X)] = \sum_x g(x)f(x)$$

Example

Let X be the number of points rolled with a balanced (6-sided) die. Find the expected value of X and of $g(X) = 2X^2 + 1$.

The probability distribution for X is $f(x) = \frac{1}{6}$. Therefore:

$$E(X) = \sum_x x \cdot f(x) = \sum_{x=1}^{6} x \cdot \frac{1}{6} = \frac{21}{6}$$

$$E[g(X)] = \sum_x g(x) f(x) = \sum_{x=1}^{6} (2x^2 + 1)\frac{1}{6} = \frac{94}{6}$$

Summary

- Sample space S contains all possible outcomes of an experiment; events A and B are subsets of S.
- rules of probability: $p(\bar{A}) = 1 - p(A)$.
  if $A \subseteq B$, then $p(A) \leq p(B)$.
  $0 \leq p(B) \leq 1$.
- addition rule: $p(A \cup B) = p(A) + p(B) - p(A, B)$.
- conditional probability: $p(B|A) = \frac{p(A,B)}{p(A)}$.
- independence: $p(A, B) = p(A)p(B)$.
- marginalization: $p(A) = \sum_{B_i} p(B_i)p(A|B_i)$.
- Bayes' theorem: $p(B|A) = \frac{p(B)p(A|B)}{p(A)}$.
- any value of an r.v. 'picks out' a subset of the sample space.
- for any value of an r.v., a distribution returns a probability.
- the expectation of an r.v. is its average value over a distribution.

References

Anderson, John R.: 1990, *The adaptive character of thought*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Kruschke, John K.: 2011, *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press/Elsevier.

Lindley, Dennis V.: 2006, *Understanding Uncertainty*. Wiley, Hoboken, NJ.

Snee, R. D.: 1974, 'Graphical display of two-way contingency tables', *The American Statistician* **38**, 9–12.