Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 6: Probability

Jinko Graham

2018-09-13

Basics (Section 6.1)

Definitions

- A phenomenon is *random* if individual outcomes are uncertain but there is a predictable behaviour in a large number of repetitions (random $\neq$ haphazard).
- The *probability* of any outcome is the proportion of times the outcome would occur in a very long (infinitely long) series of repetitions.
- Repetitions, or *trials*, are said to be *independent* if the outcome of one trial does not affect the outcome of another (e.g., coin tosses).

Probability Models

Three parts:

1. the sample space $\mathcal{S}$,
2. a list $\mathcal{E}$ of all possible events E, and
3. a way of assigning probabilities to events.

Sample Space, $\mathcal{S}$

The set of all possible outcomes.

- *Example:* Two coin tosses.

$$\mathcal{S} = \{HH, HT, TH, TT\}$$

(the braces { and } denote a set in between)

- *Example:* Roll of two dice.

$$\mathcal{S} = \{11, 12, 13, \ldots, 56, 66\}$$

- *Example:* Estrogen + progestin (EP) or not ($\overline{\text{EP}}$) and breast cancer (BC) or not ($\overline{\text{BC}}$).

$$\mathcal{S} = \{\text{EP\&BC},\ \ \text{EP\&}\overline{\text{BC}},\ \ \overline{\text{EP}}\text{\&BC},\ \ \overline{\text{EP}}\text{\&}\overline{\text{BC}}\}$$

List of All Possible Events, $\mathcal{E}$

- *Example:* Two coin tosses.

$$
\begin{aligned}
E_1 &= \{HH\} = \{\text{2 heads}\} \\
E_2 &= \{HT, TH\} = \{\text{1 head, 1 tail}\} \\
E_3 &= \{HT, TH, TT\} = \{\text{at least 1 tail}\} \\
E_4 &= \{HT, TH, HH\} = \{\text{at least 1 head}\} \\
E_5 &= \{HH, HT\} = \{\text{head on 1st toss}\} \\
E_6 &= \{HH, TH\} = \{\text{head on 2nd toss}\} \\
E_7 &= \{TH\} \\
&\quad \text{etc.}
\end{aligned}
$$

$$
\mathcal{E} = \{E_1, E_2, E_3, \ldots\}
$$

- *Example:* Roll of two dice.

$$
\begin{aligned}
E_1 &= \{11, 12, 21\} = \{\text{sum} \leq 3\} \\
E_2 &= \{11, 13, \ldots, 55, 66\} = \{\text{sum is even}\} \\
&\quad \text{etc.} \\
\\
\mathcal{E} &= \{E_1, E_2, \ldots\}
\end{aligned}
$$

Events, cont.

- *Example:* Estrogen + progestin (EP) or not ($\overline{\text{EP}}$) and breast cancer (BC) or not ($\overline{\text{BC}}$).

$$
\begin{aligned}
E_1 &= \{\text{EP\&BC},\ \text{EP\&}\overline{\text{BC}}\} = \{\text{EP}\} \\
E_2 &= \{\text{EP\&BC},\ \overline{\text{EP}}\text{\&BC}\} = \{\text{BC}\} \\
E_3 &= \{\overline{\text{EP}}\text{\&BC},\ \overline{\text{EP}}\text{\&}\overline{\text{BC}}\} = \{\overline{\text{EP}}\} \\
E_4 &= \text{EP\&BC} \\
E_5 &= \text{EP\&}\overline{\text{BC}} \\
&\quad \text{etc.}
\end{aligned}
$$

$$
\mathcal{E} = \{E_1, E_2, E_3, \ldots\}
$$

Operations on Events

- The *intersection* of two events A and B is the collection of outcomes that are in both A and B:
  - Denoted $A \cap B$ and read as "A *and* B".
- The *union* of two events A and B is the collection of outcomes that are in either A or B (or both):
  - Denoted $A \cup B$ and read as "A *or* B".
- The *complement* of an event A is the collection of outcomes not in A:
  - Denoted $\overline{A}$ and read as "not A".

Probabilities of Events: Four Rules

A probability model tells us the probability of each possible event.
Notation: $P(E)$ for "probability of event E".

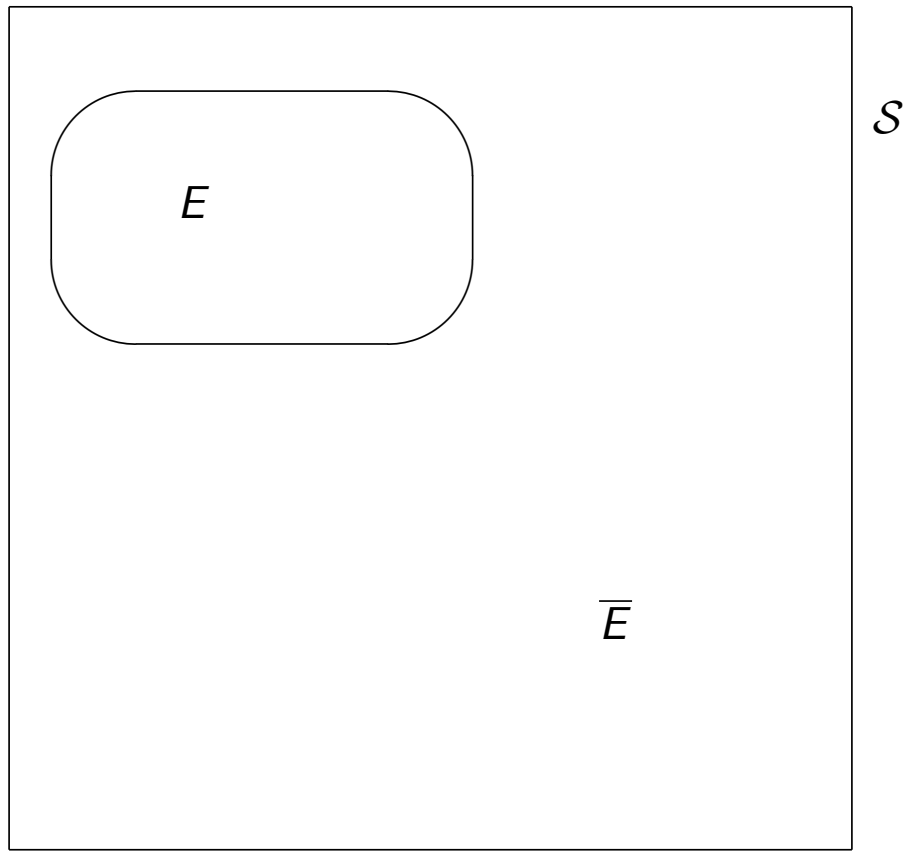
Basic rules of probability:

1. For any event E, $0 \leq P(E) \leq 1$.
2. $P(\mathcal{S}) = 1$ (something must happen)
3. For an event E, $P(\overline{E}) = 1 - P(E)$.
4. Two events are *disjoint* if they have no outcomes in common.
   - If A and B are disjoint, then the *addition rule* holds:
     $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$

Venn Diagrams

A way to picture events and probabilities.

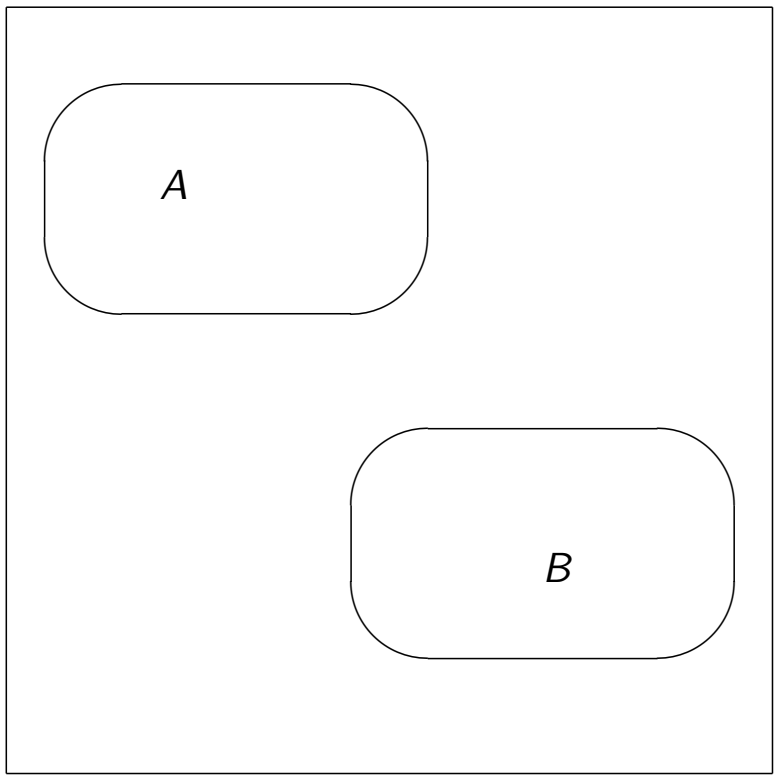
The event E and its complement $\overline{E}$ within the sample space $\mathcal{S}$.



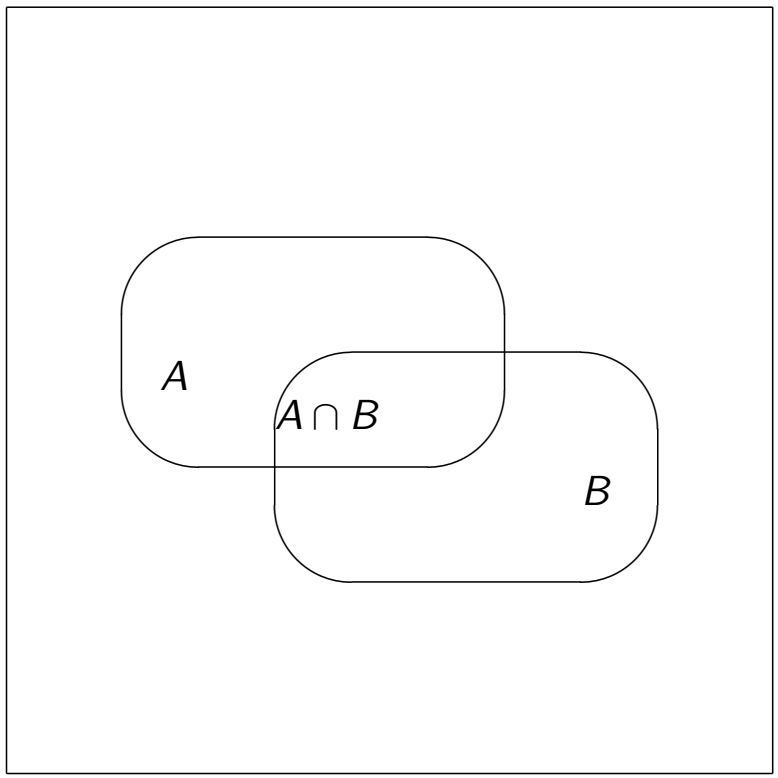
The area of an event represents its probability; e.g., we can see $P(E) < P(\overline{E})$.

Venn Diagrams, cont.

Disjoint sets A and B

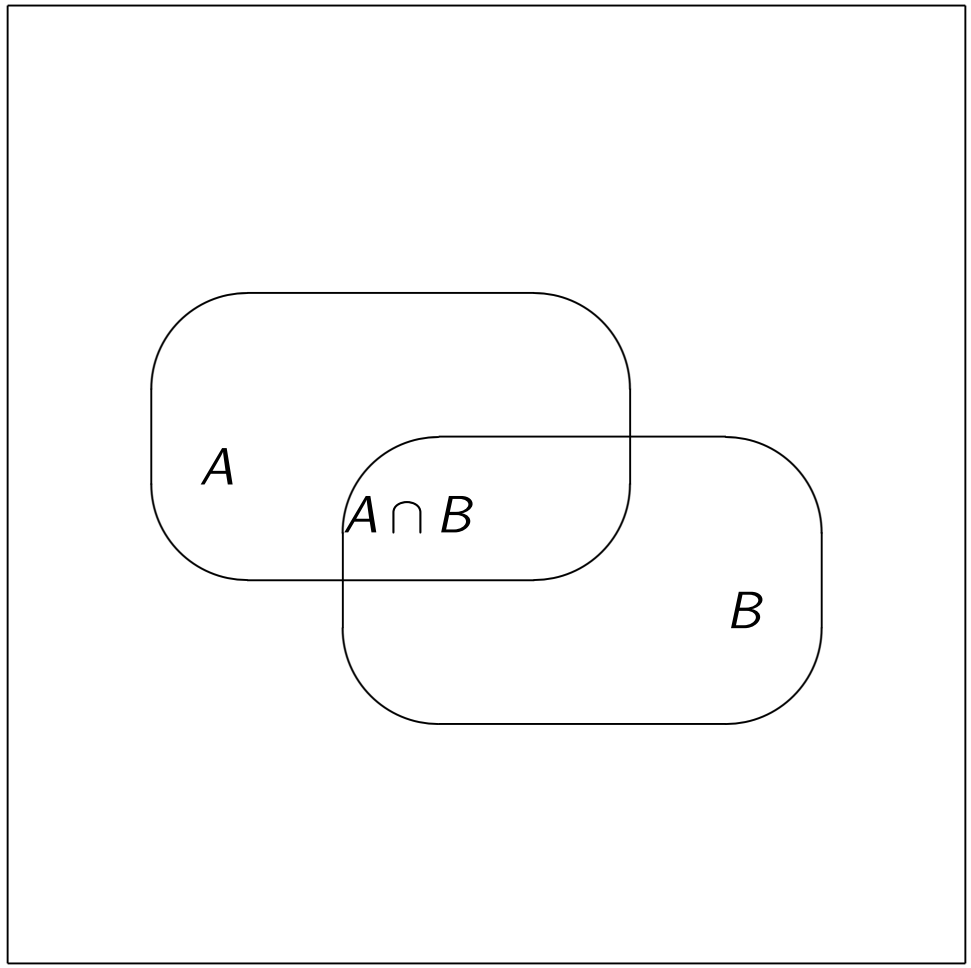


Non-disjoint sets with intersection $A \cap B$.



The General Addition Rule

General addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$:



- Adjusts for double-counting the intersection.

Multiplication rule for independent events

- Events A and B are *independent* if knowing that one has occurred provides no information about whether the other will.
- E.G. Random experiment of 2 tosses of a fair coin has sample space $\mathcal{S} = \{HH, HT, TH, TT\}$, with all 4 outcomes having chance $1/4$.
- Knowing H occurs on the 1st toss provides no info about whether H will occur in the 2nd toss.
  - Let $A = \{HH, HT\}$ be the event of an H on the 1st toss and $B = \{HH, TH\}$ be the event of an H on the 2nd toss.
  - The events A and B are independent.
- Multiplication rule for independent events:
  - If A and B are independent, then $P(A \cap B) = P(A) \times P(B)$.
- In the coin-toss experiment, $A \cap B = \{HH\}$.
  - $P(A) = P(\{HH, HT\}) = P(HH) + P(HT) = 1/4 + 1/4 = 1/2$, $P(B) = P(\{HH, TH\}) = P(HH) + P(TH) = 1/4 + 1/4 = 1/2$.
  - So, $P(HH) = P(A \cap B) = P(A) \times P(B) = \frac{1}{2} \times \frac{1}{2} = 1/4$.

Conditional Probability (Section 6.2)

General Multiplication Rule

- For 2 (not necessarily independent) events, the rule is:

$$\begin{aligned} P(A \cap B) &= P(A \mid B) \times P(B), \text{ or} \\ P(A \cap B) &= P(B \mid A) \times P(A) \end{aligned}$$

- Read $A \mid B$ as "A given B" and $P(A \mid B)$ as "the conditional probability of A given B."
- When A and B are independent,
  - $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$, so that

$$P(A \cap B) = P(A) \times P(B).$$

- In general, though, knowing B has happened may modify the probability of A;
  - i.e., the *conditional probability* of A given B may not be the probability of A; or $P(A \mid B) \neq P(A)$.
- The general multiplication holds regardless.

Conditional Probability

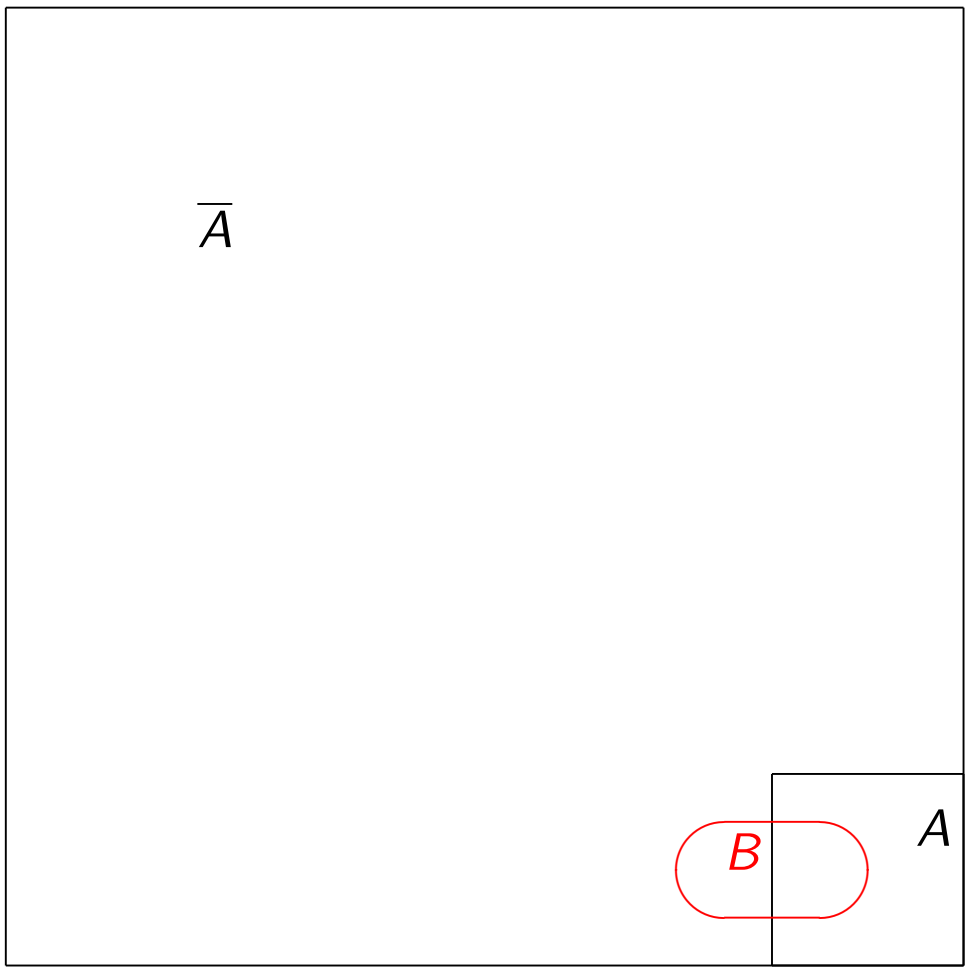
- The general multiplication rule follows from the definition of conditional probabilities:

$$\begin{aligned} P(A \mid B) &= P(A \cap B)/P(B) \quad \text{normalizing by size of } B \\ P(B \mid A) &= P(A \cap B)/P(A) \end{aligned}$$

- E.G., for $P(B \mid A)$, we are restricting our sample space to the outcomes in A and considering those that are also in B.
  - the probability is therefore the proportion of outcomes in A that are also in B.

Conditional Probability Picture

- On a Venn diagram, $P(B|A)$ is the proportion of the area of A occupied by $\{A \cap B\}$:



$P(B \cap A) = P(B \cap \overline{A})$.

Now $P(B \mid A) = P(B \cap A)/P(A)$

and $P(B \mid \overline{A}) = P(B \cap \overline{A})/P(\overline{A})$.

Numerators same but denominators different;

Since $P(A) < P(\overline{A})$, dividing by smaller number for $P(B \mid A)$ than for $P(B \cap \overline{A})$.

So $P(B \mid A) > P(B \mid \overline{A})$ as visualized on diagram.

Bayes' Theorem (Section 6.3)

Bayes' Theorem and Partitioning the Sample Space

- Connects conditional probabilities:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

- Writing B as $\{A \cap B\} \cup \{\bar{A} \cap B\}$ and noticing that these two sets are disjoint (see previous diagram) means that we can write

$$P(B) = P(A \cap B) + P(\bar{A} \cap B).$$

- Applying the general multiplication rule to both terms in the sum for $P(B)$, we obtain

$$P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A}).$$

- Which means an alternate form of Bayes' Theorem is

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})}.$$

- See text, pg 134, for an extension to partitioning of the sample space into more than two (A and $\bar{A}$) events.

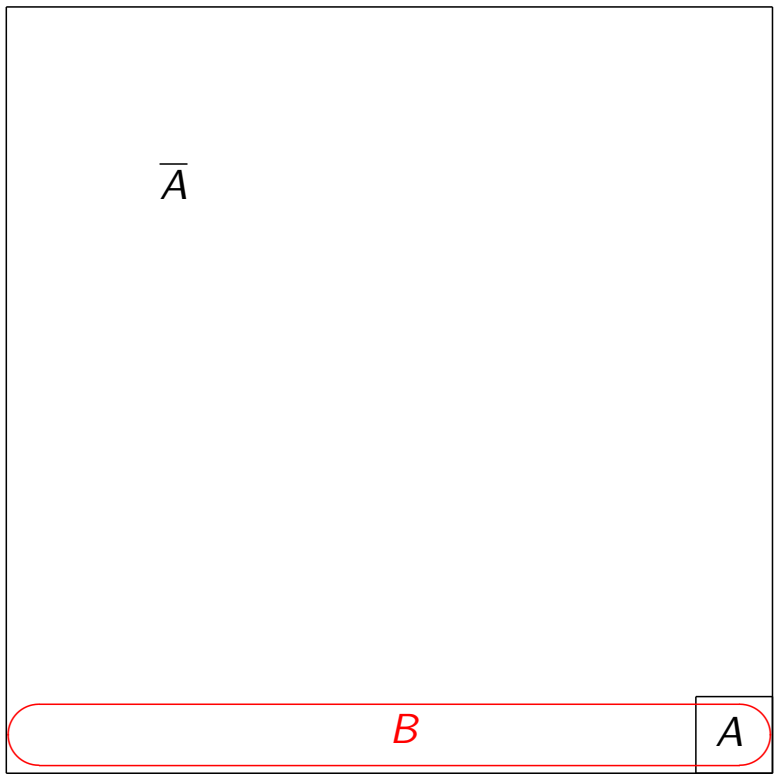
Application of Bayes' Theorem: Diagnostic Testing

- Let A={disease}, B={test positive},
- Suppose $P(B \mid A) = 0.99$, $P(B \mid \bar{A}) = 1/1000$; a good diagnostic test. Suppose rare disease with $P(A) = 1/1000000$.
- What is $P(A \mid B)$, the probability you have the disease given that you tested positive?

$$
\begin{aligned}
P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})} \\
&= \frac{0.99(1/1000000)}{0.99(1/1000000) + (1/1000)(0.999999)} \\
&\approx 1/1000! \quad \text{very low despite a good test}
\end{aligned}
$$

Diagnostic Testing Picture

The following picture is not to scale, but gives the idea:



E.G. A=disease, B=test +

- $P(B \mid A)$ may be nearly one, but $P(A \mid B)$ can still be quite small if A is rare.
- More on diagnostic testing in the next section.

Diagnostic Testing (Section 6.4)

Diagnostic Testing

- Diagnostic tests are used to screen for diseases.
  - E.G., pap smear to screen for cervical cancer
- The test can be positive (T^+) or negative (T^-), and an individual may have the disease (D^+) or not (D^-);
  - e.g., for n individuals:

		Disease Status D^+	Disease Status D^-	
Test Result	T^+	True Positive (TP)	False Positive (FP)	TP+FP
	T^-	False Negative (FN)	True Negative (TN)	FN+TN
		TP+FN	FP+TN	n

Diagnostic Test Performance

- The utility of a diagnostic test can be described in terms of
  - the probability of *test outcomes* given true disease status (sensitivity and specificity), or
  - the probability of *disease status* given test outcomes (positive and negative predictive value).

Sensitivity of a Test

- Let's focus on the diseased ($D+$): they can either test positive or negative.
  - Those who are diseased and test positive are truly positive (TP)
  - Those who are diseased and test negative are falsely positive (FN)
- The observed true-positive rate is the proportion of diseased who are truly positive, or TP/(TP+FN).
  - TP/(TP+FN) is an estimate of $P(T^+ \mid D^+)$, known as the *sensitivity* of the test.
  - A test that has sensitivity near one is good.
- The complement, FN/(TP+FN) is an estimate of the *false*-negative rate, $P(T^- \mid D^+)$.

Specifity of a Test

- Now let's focus on the non-diseased: they can either test positive or negative.
  - Those who are non-diseased and test positive are falsely positive (FP)
  - Those who are non-diseased and test negative are truly negative (TN)
- The observed true-negative rate is the proportion of non-diseased (D^-) who are truly negative (TN), or TN/(TN+FP).
  - TN/(TN+FP) is an estimate of the true-negative rate $P(T^- \mid D^-)$, known as the *specificity* of the test.
  - Specificity near one is good.
- The complement, FP/(TN+FP), is an estimate of the false-positive rate $P(T^+ \mid D^-)$.
- We can estimate the sensitivity and specificity of a diagnostic test from samples of diseased and non-diseased subjects who receive the test.

Positive Predictive Value of a Test

- Let's focus on those who test positive ($T+$): they can either be diseased or non-diseased.
  - Those who test positive and are diseased are truly positive (TP)
  - Those who test positive and are non-diseased are falsely positive (FP)
- The observed proportion of those who test positive that are truly diseased is TP/(TP+FP).
  - TP/(TP+FP) is an estimate of $P(D^+ \mid T^+)$, the true-discovery rate or *positive predictive value* of the test (aka *precision*).
  - The complement, FP/(TP+FP), is an estimate of the *false-discovery rate* of the test, $P(D^- \mid T^+)$.

Negative Predictive Value of a Test

- Finally, let's focus on those who test negative (T^-): they can either be diseased or non-diseased.
  - Those who test negative and are diseased are falsely negative (FN)
  - Those who test negative and are non-diseased are truly negative (TN)
- The observed proportion of those who test negative that are non-diseased is TN/(TN+FN).
  - TN/(TN+FN) is an estimate of $P(D^- \mid T^-)$, the *negative predictive value* of the test.
  - The complement, FN/(TN+FN), that estimates $P(D^+ \mid T^-)$ is unnamed.
- For rare diseases, we may estimate the positive and negative predictive values using Bayes' Theorem and the population prevalence of disease, $P(D^+)$.

Predictive Values by Bayes' Theorem: Motivation

- When the disease we're testing for is rare, we expect to have very few diseased individuals in our random sample.
- This is a problem for the positive and negative predictive values.
  - Samples with few diseased persons yield unreliable estimates of
    - $P(D^+ \mid T^+)$, the positive predictive value, or
    - $P(D^- \mid T^-)$, the negative predictive value $(= 1 - P(D^+ \mid T^-))$.
- Fortunately, we can use Bayes' Theorem to express the positive- and negative-predictive values in terms of:
  - the sensitivity, $P(T^+ \mid D^+)$;
  - the specificity, $P(T^- \mid D^-)$; and
  - the disease prevalence, $P(D^+)$.
- The sensitivity, specificity and disease prevalence can be reliably estimated if a population registry of cases (diseased persons) and census information is available.
  - E.G. BC Cancer Agency's Registry and the Canadian Census

Public Health Professionals Need Population Registries and the Census to Make Informed Policy Decisions

- A decent-sized sample of diseased individuals can be obtained from a population disease registry, which allows us to estimate the sensitivity of a diagnostic test, $P(T^+ \mid D^+)$.
- A large sample of non-diseased individuals allows reliable estimation of the specificity of a test, $P(T^- \mid D^-)$.
- The population registry and census data also provide the numbers with the disease and in the population, respectively, which lead to an estimate of the disease prevalence, $P(D^+)$.
- Note: Other approaches to estimating the prevalence will not be covered, but see Section 6.4.4 of the text if you are interested.
- Raises an interesting point: Governments that care about making informed decisions should think twice before eliminating public resources such as disease registries and the census.

Predictive Values by Bayes' Theorem: Details

- The positive predictive value is

$$\begin{aligned} P(D^+ \mid T^+) &= \frac{P(T^+ \mid D^+)P(D^+)}{P(T^+ \mid D^+)P(D^+) + P(T^+ \mid D^-)P(D^-)} \\ &= \frac{\text{sens} \times P(D^+)}{\text{sens} \times P(D^+) + (1 - \text{spec}) \times (1 - P(D^+))} \end{aligned}$$

- The negative predictive value is

$$\begin{aligned} P(D^- \mid T^-) &= \frac{P(T^- \mid D^-)P(D^-)}{P(T^- \mid D^-)P(D^-) + P(T^- \mid D^+)P(D^+)} \\ &= \frac{\text{spec} \times (1 - P(D^+))}{\text{spec} \times (1 - P(D^+)) + (1 - \text{sens}) \times P(D^+)} \end{aligned}$$

Positive Predictive Value Example

- Pap smear performance as a diagnostic test for cervical cancer (see the text, pgs 136-137).
- Estimated sensitivity of the test is 0.8375
- Estimated specificity of the test is 0.8136
- Estimated prevalence of cervical cancer is 8.3 per 100,000 or 0.000083.
- Hence the positive predictive value is

$$
\begin{aligned}
P(D^+ \mid T^+) &= \frac{\text{sens} \times P(D^+)}{\text{sens} \times P(D^+) + (1 - \text{spec}) \times (1 - P(D^+))} \\
&= \frac{0.8375 \times 0.000083}{0.8375 \times 0.000083 + (1 - 0.8136) \times (1 - 0.000083)} \\
&= 0.000373
\end{aligned}
$$

or only 37.3 expected to actually have cervical cancer per 100,000 positive tests.

The Relative Risk and Odds Ratio (Section 6.5)

The Relative Risk

- Example: Organochlorines (OGC) are used in farm pesticides. Does exposure to OGC increase the risk of developing non-Hodgkin lymphoma (NHL)?

Int. J. Cancer: **121**, 2767–2775 (2007)

Organochlorines and risk of non-Hodgkin lymphoma

John J. Spinelli[1]*, Carmen H. Ng[1], Jean-Philippe Weber[2], Joseph M. Connors[1], Randy D. Gascoyne[1], Agnes S. Lai[1], Angela R. Brooks-Wilson[1], Nhu D. Le[1], Brian R. Berry[1] and Richard P. Gallagher[1]

[1]*BC Cancer Agency, Vancouver, British Columbia, Canada*
[2]*Centre de Toxicologie du Québec, Sainte-Foy, Québec*

- Compare the risk of NHL in those exposed to OGC to the risk in those unexposed to OGC.
- The multiplicative factor by which OGC exposure increases the risk of NHL is the relative risk, or RR:

$$RR = \frac{P(\text{disease} \mid \text{exposed})}{P(\text{disease} \mid \text{unexposed})}$$

Estimating the Relative Risk

- If we obtain a sample of exposed individuals, we can estimate $P(\text{disease} \mid \text{exposed})$ by the proportion of diseased individuals in the sample.
- Similarly, if we obtain a sample of unexposed individuals, we can estimate $P(\text{disease} \mid \text{unexposed})$ by the proportion of diseased individuals in the sample.
- The ratio of our estimates is then an estimate of the RR.

Rare Diseases

- For rare diseases it may be difficult to collect enough data to estimate the RR
  - E.G., about 20 cases of NHL per 100,000 people per year (about 1 per 5,000 per year).
  - With few disease cases in either exposed or unexposed groups, estimates of $P(\text{disease} \mid \text{exposed})$ and $P(\text{disease} \mid \text{unexposed})$ will be unreliable.
  - This is where population-based disease registries come in. Disease registries are an ample source of cases.
- The BC Cancer Agency study of NHL used a cost-effective study design called the case-control design and the BC Cancer Agency's provincial registry of NHL cases.
- The classic example of a case-control study: Doll and Hill's smoking and lung cancer study.

Doll and Hill's Study

- Research Question: Is smoking associated with lung cancer?
  - (We know now, yes, but in the 1950's this was a controversial hypothesis.)
- Data on hospitalized patients from about 20 hospitals around London between Apr. 1948 - Feb. 1952.
- Hospital staff contact investigators whenever a new patient is admitted for lung cancer (case) and also interview patient about smoking habits (exposure).
- Also interview a patient admitted for something other than lung cancer (controls) at the same hospital.

Doll and Hill's Data

		case	control
Smoke	Yes	1350	1296
(E)	No	7	61
		1357	1357

- Problem: Proportions of cases in the smokers or non-smokers do not reflect population proportions because of the sampling design, which over-samples cases relative to their frequency in the population.
  - In the table above, half our sample is cases and half controls. But in the population, a small proportion is cases and a large proportion is controls (because lung cancer is rare).
- **Can't estimate the *RR* from case-control data**, but we can estimate a related quantity called the odds-ratio ...

The Odds Ratio

- An alternative to the relative risk is the *odds ratio* OR
- Let $p_1 = P(\text{disease} \mid \text{exposed})$ and $p_0 = P(\text{disease} \mid \text{unexposed})$.
- The odds of disease in the exposed group is $p_1/(1-p_1)$
- The odds of disease in the unexposed group is $p_0/(1-p_0)$
- The odds ratio is

$$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

- Compare to $RR = p_1/p_0$.

Connections between RR and OR

- If exposure does not affect disease risk, then $p_0 = p_1$, which implies that both $RR = 1$ and $OR = 1$.
- When the disease is rare, both p_0 and p_1 will be close to zero, so that $1 - p_0$ and $1 - p_1$ will be close to one. Then the OR is approximately equal to the RR.
- When can the OR be used to approximate the RR? How rare does the disease have to be? No consensus. In this class we will use a disease prevalence of less than 5%.

More on the Odds Ratio

- After some algebra and applications of Bayes' theorem, can show that the OR is also the ratio of the odds of exposure in the disease group to the odds of exposure in the no-disease group:

$$
\begin{aligned}
OR &= \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \\
&= \frac{P(\text{disease} \mid \text{exposed})/P(\text{no disease} \mid \text{exposed})}{P(\text{disease} \mid \text{unexposed})/P(\text{no disease} \mid \text{unexposed})} \\
&= \ldots \text{algebra} \ldots \\
&= \frac{P(\text{exposed} \mid \text{disease})/P(\text{unexposed} \mid \text{disease})}{P(\text{exposed} \mid \text{no disease})/P(\text{unexposed} \mid \text{no disease})}
\end{aligned}
$$

More on the Odds Ratio, cont.

- In other words,

$$\begin{aligned} OR &= \frac{\text{odds of disease in exposed}}{\text{odds of disease in unexposed}} \\ &= \frac{\text{odds of exposure in diseased}}{\text{odds of exposure in not diseased}} \end{aligned}$$

- Importantly, the odds of exposure in either cases or controls *can* be estimated from case-control data.
- For example, with Doll and Hill's data,
  - $P(\text{exposed} \mid \text{not diseased})$ is estimated by the proportion of smokers in the control group, or $1296/1357$, and
  - $P(\text{exposed} \mid \text{diseased})$ is estimated by the proportion of smokers in the case group, or $1350/1357$
  - Together, these lead to an estimated odds ratio of

$$\frac{\frac{1350}{1357} / \frac{7}{1357}}{\frac{1296}{1357} / \frac{61}{1357}} = \frac{1350/7}{1296/61}.$$

- Hence, we **can estimate the OR from case-control data**.

Example OR Estimates

- Doll and Hill's lung cancer study gave an estimated *OR* for smoking of 9.1.
- The BC Cancer Agency NHL study gave an estimated *OR* for OGC of 2.7.
- Both lung cancer and NHL are relatively rare, so we may interpret these estimates as relative risks:
  - smoking increases your risk of lung cancer nine-fold
  - exposure to certain pesticides increases your risk of NHL almost three-fold
- We will attach measures of uncertainty to such estimates (i.e., confidence intervals) in Chapter 15.

Chapter Summary

- Discussed the basic definitions and rules of probability, including the definition of conditional probability.
- Use Bayes' Theorem to relate $P(A \mid B)$ to $P(B \mid A)$,$P(A)$ and $P(B)$.
- Public-health and medical practitioners work with many conditional probabilities every day; e.g.,
  - diagnostic test sensitivity and specificity
  - relative risks and odds ratios
- Case-control studies oversample cases relative to their frequency in the population.
  - As a result, we cannot estimate relative risks of disease from case-control data.
  - We can however estimate odds-ratios for disease because the relative odds of exposure is estimable and equal to the relative odds of disease.