PubH 6414
Lesson 5

Probability

Course Overview so far

- Describing the Data: Categorical & Numerical
- Measuring the strength of relationship between two variables
  - Correlation coefficient for numerical data
  - RR and OR for nominal variables
- Probability and Probability Models – this is where we are now
- Estimating Parameters – Confidence Intervals
- Statistical Tests of Significance – Hypothesis Tests and p-values
- Statistical Models: ANOVA and Regression

Lesson 5 Outline

- Definition and properties of Probability
- Mutually exclusive events and the addition rule
- Non-mutually exclusive events
  - Marginal, joint and conditional probabilities
- Independent events and the multiplication rule
- Non-independent events

Lesson 5 Outline

- Sensitivity, Specificity, NPV and PPV as Conditional Probabilities
- Calculating PPV and NPV using Bayes' Theorem

Randomness

- Definition:
  - The phenomenon is *random* if the outcome is uncertain.
    - The outcome of a *single* flip of a coin.
    - The next participant's blood type.
    - Your blood pressure.
    - Tomorrow's weather.

Sources of Uncertainty

- Why are measurements of variables uncertain (i.e. variable)?
  - **Sampling Variability**: different samples give different results.
  - **Measurement Variability**: instrument calibration, observer skill.
  - **Intrinsic Variability**: circadian rhythm, hormonal cycle.
  - **Modeling Variability**: different models, applied to the same data, can give different results.

Handling such uncertainty is the foundation of statistics!!

Terminology and Definitions

- Terminology:
  - Trial: One repetition of an experiment that can have one or more possible outcomes
  - Event: one of the possible outcomes of a trial
- Definition of Probability
  - The probability that a given event occurs is the long-term relative frequency of the the event over many trials

Classic Probability Example

- A classic example of probability is tossing a coin
  - A **trial** is a single toss of the coin
  - The **event** is getting ‘heads’
- The outcome for a single toss is random
  - It could be ‘heads’ or ‘tails’
- After many repetitions of the trial, the long-term relative frequency of ‘heads’ can be calculated and the probability of getting heads is known: P(heads) = 0.5

Probability Rules:

1. The probability P(A) of any event A satisfies $0 \leq P(A) \leq 1$.
2. If S is the sample space in a probability model, then P(S) = 1.
3. The complement of any event A is the event that A does not occur, written as A^c. The complement rule states that $P(A^c) = 1-P(A)$.
4. Two events A and B are disjoint if they have no outcomes in common and so can never occur simultaneously. If A and B are disjoint, P(A or B) = P(A) + P(B).
5. If events A and B are not disjoint P(A or B) = P(A) + P(B) – P(A and B).

Examples used to describe Probability Rules 1-5

The following Blood Type distributions will be used to illustrate some probability rules

- Distribution of blood types in the US
- Distribution of blood types by gender
- Distribution of blood types in Australia and Finland

Probability: Rule 1

- The probability P(A) of any event A satisfies $0<P(A)<1$.
- Probability = 1 if the event is certain.
- Probability = 0 if the event is impossible.
- Probability values between 0 and 1 express some degree of uncertainty about whether or not the event will occur in a single trial

Distribution of US Blood Types

Blood Type	**Probability**
O	0.42
A	0.43
B	0.11
AB	0.04

Event A: Blood Type A

What is P(A)?

Addition Rule of Probability for Mutually Exclusive events: Rule 2

If S is the sample space in a probability model, then P(S) = 1.

- The sum of the probabilities for all 4 blood types = 1.0

Complementary events: Rule 3

- The complement of any event A is the event that A does not occur, written as A^c. The complement rule states that $P(A^c) = 1 - P(A)$.

- What is the probability of not having blood type O.

Mutually Exclusive (Disjoint) Events

- Two or more events are *mutually exclusive* if they cannot occur simultaneously (that is, they do not have any elements in common).

- Blood Types are *disjoint* events because each individual can have one (and only one) blood type

Addition Rule of Probability for Mutually Exclusive events: Rule 4

Two events A and B are disjoint if they have no outcomes in common and so can never occur simultaneously. If A and B are disjoint, P(A or B) = P(A) + P(B).

- What is the probability of a randomly selected person having blood type A, *or* B *or* AB?

Addition Rule of Probability for Mutually Exclusive events: Rule 4

Two events A and B are disjoint if they have no outcomes in common and so can never occur simultaneously. If A and B are disjoint, P(A or B) = P(A) + P(B).

- What is the probability of a randomly selected person having *either* blood type O or blood type A?

Gender and US Blood type distributions

	Probabilities		
Blood Type	Male	Female	Total
O	0.21	0.21	0.42
A	0.215	0.215	0.43
B	0.055	0.055	0.11
AB	0.02	0.02	0.04

Non-mutually Exclusive Events

- Events are *non-mutually exclusive* if they can occur simultaneously (*not disjoint*)
- Gender and blood type are non-mutually exclusive events because each person has both events

Non-mutually Exclusive Events

- The following probabilities can be identified for non-mutually exclusive events
  - Marginal probability: the probability that one of the events occurs
  - Joint probability: the probability that two events occur simultaneously
  - Conditional probability: the probability that one event occurs given that the other event has occurred

Marginal Probability

The marginal probability is the probability of a single event

These probabilities are in the margins of the table of gender / blood type distributions

For example, the probability of having blood type O, regardless of gender, is 0.42

Joint Probability

- A joint probability is the probability that two events occur simultaneously'

- Using the blood type distribution by gender, the P(Male and Type A) =

Conditional Probability

A conditional probability is the probability of one event given that the other event has occurred.

The notation for the conditional probability of having blood type O given that you are female is P(Type O | Female).

Conditional Probability: Rule 6

- When P(A) >0, the conditional probability of event B, given A has occurred:

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

Conditional Probability Example

Using the Distribution of blood types by gender, what is the P(Type O | Female)?

Independent Events: Rule 7

- Two events are independent if the probability of one does not affect the probability of the other.

- If two events A and B are independent,

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

Multiplication rule for independent events

Independent Events

For example: Is male gender independent of type O blood?

Independent Events and Conditional Probabilities

- If two events are ***independent,*** the conditional probability of the event is equal to the marginal probability of the event.

$$\boxed{P(A \mid B) = P(A)}$$

Independent Events and Conditional Probabilities

- For example, given that a participant is male, does this change the probability of type A blood?

Non-independent Events

- When two events are not independent
  - The joint probabilities are **not** equal to the product of the two marginal probabilities

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

  - The conditional probability is **not** equal to the marginal probability

$$P(A|B) \neq P(A)$$

Non-independent Events

Non-independence between two events suggests that a statistical relationship may exist between them.

Non-independence: the probability of one event is affected by the outcome of the other event

Example of Non-independent Events

- Blood type distributions are not the same across different countries and ethnic groups.
- For simplicity, the probability table on the next slide assumes an equal proportion of individuals from each country

Distribution of blood types by country

	Probabilities		
Blood Type	Australia	Finland	Total
O	0.245	0.155	0.40
A	0.19	0.22	0.41
B	0.05	0.085	0.135
AB	0.015	0.04	0.055
Total	0.50	0.50	1.00

Joint probabilities and Non-independent events

- Is blood type O independent of country?

Conditional Probabilities and Non-independent events

- Is blood type O independent of country?

Toss two fair coins once. What is the sample space (outcome space)?

---

1. {H T}
2. {(HH), (HT), (TT)}
3. {(HH), (HT), (TH), (TT)}
4. {½}
5. {¼, ½}
6. {¼, ½,¼}

How many trials did we do?

1. One
2. Two
3. Three
4. Four

If we were to toss two coins again, in the same way, would that be an identical trial?

1. No
2. Yes

In this experiment, what is the probability of getting two heads?

1. 1/2
2. 1/3
3. 1/4
4. 1/6

In this experiment, what is the probability of getting one head and one tail?

1. 1/2
2. 1/3
3. 1/4
4. 1/6

Consider tossing two dice. Below is the sample space for this experiment.

$$\therefore S = \left\{\begin{matrix} (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \end{matrix}\right\}$$

What is the probability you roll (3,3)?

1. 1/2
2. 1/6
3. 1/12
4. 1/36

What is the probability the sum of the numbers is even (i.e. (1,1)=>1+1=2)?

1. 1/2
2. 1/6
3. 1/12
4. 1/36

What is the probability the sum of the numbers is even ***AND*** you roll (3,3)?

1. 1/2
2. 17/36
3. 1/6
4. 1/12
5. 1/36

What is the probability the sum of the numbers is even ***OR*** you roll (3,3)?

1. 1/2
2. 17/36
3. 1/6
4. 1/12
5. 1/36

Screening Test Measures

- Screening tests are used to classify people as healthy or as falling into one or more disease categories.
- Screening tests are not 100% accurate and therefore misclassification is unavoidable.

- Examples: HIV test, Colonoscopy, Skin tests for TB, mammograms

Screening test measures are conditional probabilities

Screening tests involve two events: disease (D+ or D-) and Test result (T+ or T-).

- Sensitivity is the probability of a positive test given that the disease is present

  $P(T+ \mid D+)$

- Specificity is the probability of a negative test given that the disease is absent

  $P(T- \mid D-)$

Cancer Screening Data

	Test Result (T)		
Disease (D)	T+	T-	Total
D+	154	225	379
D-	362	23,362	23,724
Total	516	23,587	24,103

Sensitivity and Specificity

Sensitivity: $P\ (T+ \mid D+) =$

Specificity: $P\ (T- \mid D-) =$

Screening Test errors

Screening tests do not always accurately identify disease state of an individual. The two possible errors are called 'False Negative' and 'False Positive'

False Negative:

Probability of False Negative = P(T-|D+)

P(T-|D+)=

False Positive:

Probability of False Positive = P(T+|D-)

P(T+|D-) =

Positive and Negative Predictive Values

In addition to Sensitivity and Specificity, two other conditional probabilities are used to evaluate screening tests

Positive Predictive Value (PPV or PV+):

- the probability that the disease is present given that the test is positive.
- A conditional probability: P (D+ | T+)

Negative Predictive Value (NPV or PV-):

- The probability that the disease is not present given that the test is negative
- A conditional probability: P(D- | T-)

PPV and NPV

PPV = P (D+ | T+) =

NPV = P (D- | T -) =

Interpretation of Positive and Negative Predictive Values

If the test is positive, what is the probability that the individual actually has the disease?

- For the cancer screening data, PPV = 0.285 so only 28.5% of those with positive screening tests actually have cancer.
- In this example, NPV = 0.99 so 99% of those with a negative test result don't have cancer.

Effect of Prevalence on Screening test results

Example 1: AIDS screening test for 100,000 people in the general population

Screening	AIDS		
Result	Yes	No	Total
T+	98	1998	2,096
T-	2	97,902	97,904
Total	100	99,900	100,000

Example 1 Screening test measures

- Sensitivity = 98 / 100 = 98%
- Specificity = 97,902 / 99,900 = 98%

- The prevalence of AIDS for those screened can be calculated from the table. Prevalence = the proportion of the total with disease =

- PPV for this example = 98 / 2096 = 0.047

Effect of Prevalence on Screening test results

- The PPV of a screening test depends not only on the sensitivity and specificity of the test but also on disease prevalence in the population screened

- The higher the prevalence, the higher the test's positive predictive value.

Using Bayes Theorem to Calculate PPV and NPV

- In the previous calculations of PPV and NPV we had data about the true disease status of the individuals (D+ and D-).
- Often we have the screening test results but are lacking information on the true disease state of the individual being screened
  - Solution: Use Baye's Theorem.

Thomas Bayes

- Bayes, Thomas (b. 1702, London - d. 1761, Tunbridge Wells, Kent), mathematician who first used probability inductively and established a mathematical basis for probability inference.
- Source:http://www.bayesian.org/resources/bayes.html

Application of Bayes' Theorem to obtain PPV

Bayes' Theorem:

$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}}$$

Replace A with D+ and replace B with T+:

$$\mathrm{P(D+ \mid T+)} = \frac{\mathrm{P(T+ \mid D+)P(D+)}}{\mathrm{P(T+ \mid D+)P(D+)} + P\mathrm{(T+ \mid D-)(D-)}}$$

Using Bayes' Theorem to calculate PPV and NPV

- To begin:

$P(D+) =$

$P(D-) =$

$P(T-|D-) =$

$P(T+|D+) =$

$P(T+|D-) =$

$P(T-|D+) =$

Bayes's rule: An Example

- Suppose the University decides to screen the faculty for illegal drug use.
- Suppose the true prevalence of regular illegal drug use among the faculty is 0.1%
- The sensitivity of the screening test for illegal drug use has a sensitivity of 99.5%
- The specificity of the screening test is 98%

Bayes's rule: An Example

What is the probability a faculty member is a regular illegal drug user given he/she has a positive test result?

That is,

$$P(D=+|T=+)$$

Bayes's Rule: An Example

- To begin:
  - $P(D=+) =$
  - $P(D=-) =$
  - $P(T=+|D=-) =$
  - $P(T=-|D=-) =$
  - $P(T=+|D=+) =$
  - $P(T=-|D=+) =$

Bayes's Rule: An Example

- We want the probability of illegal drug use given a positive test result P(D=+|T=+).
- Use Bayes's Rule (Positive Predictivity) :

$$\boxed{P(D=+\mid T=+)=\frac{P(T=+\mid D=+)P(D=+)}{P(T=+\mid D=+)P(D=+)+P(T=+\mid D=-)P(D=-)}}$$

Bayes's Rule: An Example

- The probability of illegal drug use given a positive test result $P(D=+|T=+) = 0.0474$. This is the positive predictivity of the test.

- So, even with a highly sensitive and highly specific test the probability that someone is a regular illegal drug user given they have a positive test result is only 4.74% (for a population with a prevalence of 1 in 1000).

BAYES' RULE =
Inversion of Probabilities

Getting from $P(T+|D+)$ to $P(D+|T+)$

and

SNIFFY

SNIFFY tested positive for Lyme's Disease.

Do you think SNIFFY has the disease?

1. No
2. Maybe
3. Yes
4. What is Lyme's Disease?

What is $P(T+|D+)$?

1. Sensitivity
2. The probability of the disease given a positive test result.
3. 1-Sensitivity
4. Prevalence

What is P(D+|T+)?

1. Sensitivity
2. The probability of the disease given a positive test result.
3. 1-Sensitivty
4. Prevalence

What is the probability SNIFFY really has Lyme's Disease?

$$P(D+|T+)$$

Suppose

Sensitivity = 98%

Specificity = 97%

Prevalence = 5%

What is P(D+)?

1. 98%
2. 97%
3. 95%
4. 5%
5. 3%

What is $P(T+|D+)$?

1. 98%
2. 97%
3. 95%
4. 5%
5. 3%

What is $P(T+|D-)$?

1. 98%
2. 97%
3. 95%
4. 5%
5. 3%

What is P(D-)?

1. 98%
2. 97%
3. 95%
4. 5%
5. 3%

What is $P(D+|T+)$?

Positive Predictive Value

1. 98%
2. 97%
3. 95%
4. 85%
5. 63%

What is the probability SNIFFY really has Lyme's Disease?

$$\frac{\text{Sensitivity} * \text{Prevalence}}{\text{Sensitivity} * \text{Prevalence} + \text{False Positives} * (1 - \text{Prevalence})}$$

Online Clinical Calculator

- http://www.intmed.mcw.edu/clincalc/bayes.html
- Online Clinical Calculator from Medical College of Wisconsin
  - Enter Prevalence, Sensitivity, Specificity as decimals
  - The Online Clinical Calculator Returns PPV and NPV

A link to this website has been added to the course weblinks

Readings and Assignments

- Reading
  - Chapter 4 pgs. 63-68
  - Chapter 12 pgs. 306 - 309
  - Note on Pg. 63 of text: “Our experience indicates that the concepts underlying statistical inference are not easily absorbed in a first reading”
- Work through probability calculations on the Lesson 5 Practice Exercises
- Work through examples in Excel Module 5
- Complete Homework 3 by the due date