

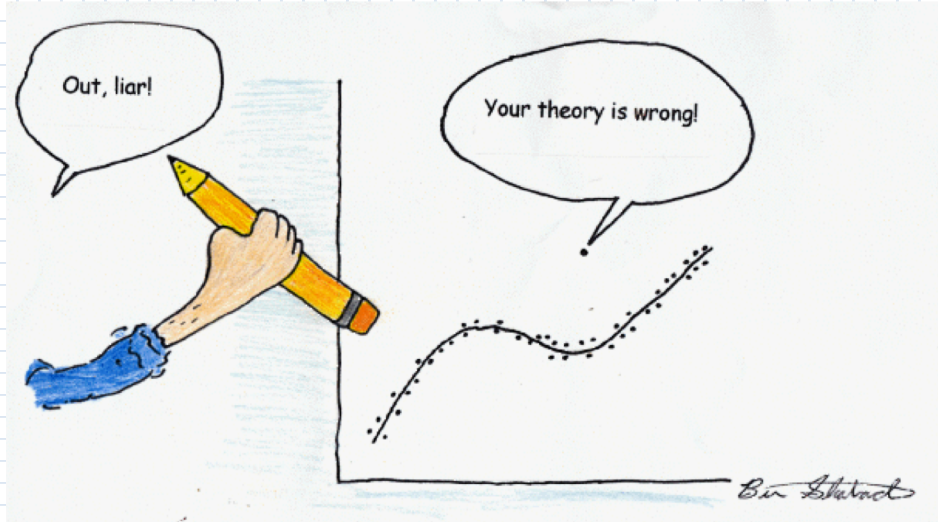
Biostatistics 200

Basically, cleaning your data &
then Basic Probability



Data Cleaning

- Always necessary with a new dataset
- Assume your data has errors
- Tables & summary statistics & graphs can help
- Outliers – extreme values
- Always keep a copy of the original data



Data Cleaning

- Data cleaning tasks
 - Examine your data
 - Fill in missing values (many ways)
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data
 - Noisy
 - Redundant
 - Biased

***2005 Leff, Allen et al, Inside Employment Interventions: Effects of Job Development and Job Support on Competitive Employment of Persons with Severe Mental Illness, *Psychiatric Services*, 56(10): 1237-1245.**

Outliers – what do we do?

- Is the value physically possible?
- Yes, keep them but...
- Do sensitivity analyses (what's that?)
- Which summary measures are they likely to affect?
- Trimmed means, robust statistics, transformations

Prevent Errant Data

- Check that you are coding numbers as numbers and not text
- Range checks, skip patterns
- Can you do double entry?
- Even if you are using a very large dataset, check the data early, often, and look at samples
- If you have an error in the data, then what?

- And now for something almost completely different...quantifying your data very simply
- You want to start looking at your data...
- But first, remember that being a statistician means you never have to be certain...

Probability!

Probability:

It's all about uncertainty

- Probability is the numerical measure of the likelihood that an event will occur
- Value is between 0 and 1
- Sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1

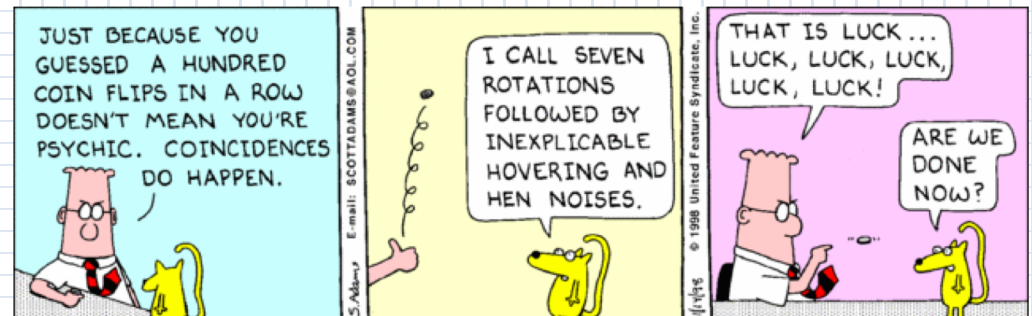
Events & Sample Spaces

- Simple event
 - Outcome from a sample space with one characteristic
 - e.g.: Probability of a baby's delivery @ 36 weeks
- Joint event
 - Involves two outcomes simultaneously
 - e.g.: Probability of delivery @ 36 weeks AND the baby is female

The collection of all possible outcomes is the Sample space

Defining an Event

- The result of an experiment or observation
- The event occurs or does not occur
- We all apply probability to events and we want to estimate the relative frequency that an event occurs over a large number of identical trials
- An event can be a disease occurrence or an abnormal lab value



Definition of Probability

- If an experiment is repeated n times under essentially identical conditions, and if the event A occurs m times, then as n increases,
 - The ratio m/n approaches a fixed number which is the probability of A occurring

$$P(A) = m/n$$

Probability of an event

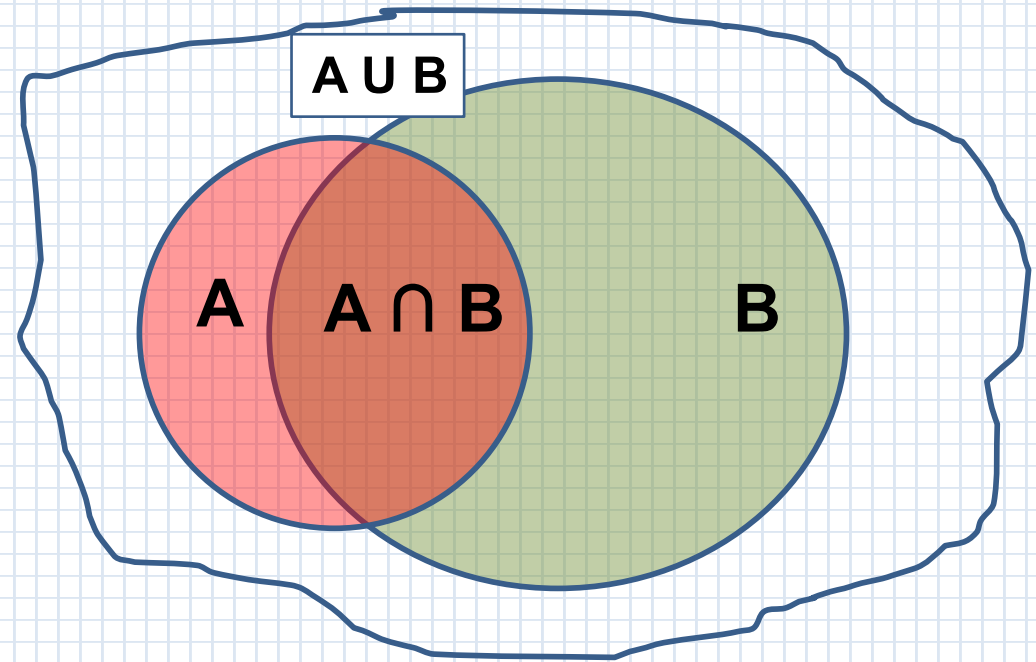
- Probability of an event is the relative frequency of its occurrence in a large number of trials repeated under the same conditions
 - Always lies between 0 and 1 (inclusive)
 - Denoted $P(A)$ or $P(X)$ etc...
 - Examples?
 - Efficacy trials?
 - Vaccine trials?

Complement & Universe

- The Complement of an event, A^c (A complement or 'Not A')
 - e.g. if A is the event that a person has a disease, A^c is the event that a person does not have a disease
 - $P(A) = 1 - P(A^c)$
 - The universe of all possible events is $P(A) + P(A^c)$ and this = 1

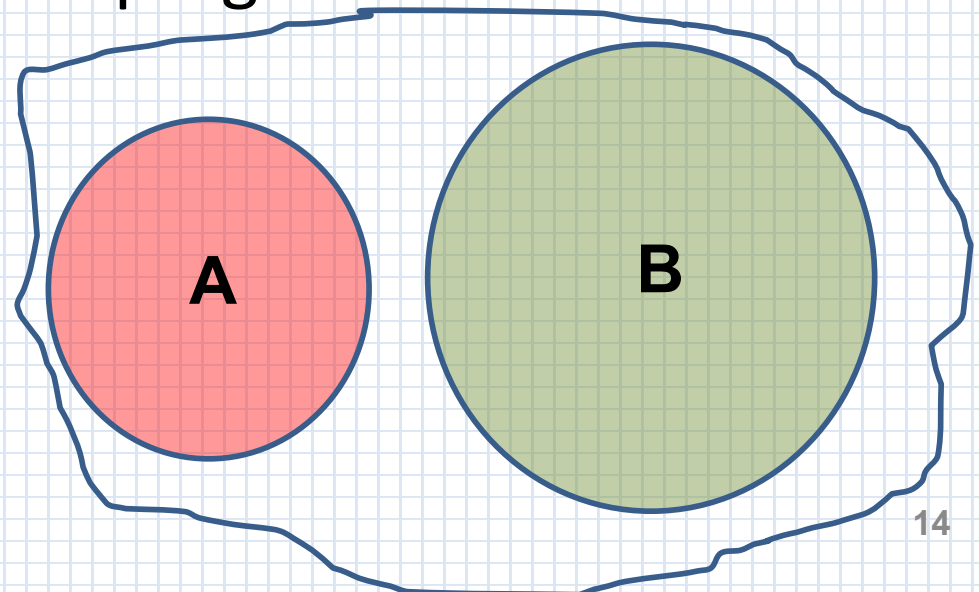
Union & Intersection

- The Union of 2 events is written $A \cup B$
- The Union is if either A or B or both occur
- The Intersection is when both occur at the same time
- $P(A \cap B)$ is the Intersection
- $P(A \cup B)$ is the Union
- $= P(A) + P(B) - P(A \cap B)$



Mutually Exclusive

- Two events are mutually exclusive if they cannot occur together
 - There is no overlap area because both can't happen together
 - The intersection is empty = 0
 - e.g. being pregnant & not pregnant



Additive rule

- The additive rule of probability can be applied to mutually exclusive events
- For 3 events: If none of the events can occur together, then

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$

- For n events:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots P(A_n)$$

Probability summary

- Complement: $P(A) = 1 - P(A^c)$
- Intersection: $P(A \text{ and } B) = P(A \cap B)$
- Union: $P(A \text{ or } B \text{ or both}) = P(A \cup B)$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Mutually exclusive events:

$$P(A \cap B) = 0$$

If A and B are mutually exclusive: then

$$P(A \cup B) = P(A) + P(B) \text{ additive rule}$$

Simple Probability Example

- A is the event that you are exposed to high levels of carbon monoxide (CO)
- B is the event that you are exposed to high levels of nitrogen dioxide (NO₂)
- In words:
 - What is the intersection of A and B
 - What is the union of A and B
 - Are these events mutually exclusive?

Conditional Probability & Rev. Bayes

- The probability that an event B will occur given that event A has occurred
 - The probability of B given A = $P(B | A)$
 - The probability that it rains given you brought your umbrella
 - $P(\text{Rain} | \text{Have umbrella})$
- $P(B | A) = P(A \cap B) / P(A)$
 - The relative size of the probability of the intersection compared to the relative size of $P(A)$

Frequency tables

- Choice of cutpoints for categories
 - Even intervals
 - E.g. 10-year age categories
 - Meaningful cutpoints related to a health outcome or decision
 - E.g. $CD4 < 50$ cells/mm³
 - Equal percentage of the data falling into each category
 - Tertiles – 33%
 - Quartiles – 25%
 - Quantiles – 20%

Conditional Probability Example

- Very often used in diagnostic tests
- $P(\text{HIV} | \text{given a positive test})$ OR $P(\text{Positive test} | \text{HIV})$
 - In diagnostic testing these are called positive predictive value and sensitivity
- What are each of these telling us about a test?
- $P(\text{not HIV} | \text{negative test})$ OR $P(\text{Negative test} | \text{not HIV})$
 - Negative predictive value and specificity

Conditional Probability Example:

The dangers of screening everyone

- $P(\text{Disease}) = 0.001$ or 100 of 100,000
- $P(\text{No Disease}) = 1 - P(\text{Disease}) = 0.999$
- $P(\text{Positive+} | \text{Disease}) = 0.99$ or 99 of 100
- $P(\text{Negative-} | \text{No Disease}) = 0.99$ or 98,901 of 99,900

	Disease	No Disease	
Test Positive	99	999	1098
Test Negative	1	98901	98902
	100	99900	100000

$P(\text{Disease}) =$	$100/100000 =$	0.001	
$P(\text{No Disease}) =$	$99900/100000 =$	0.999	
$P(\text{Test+} \text{Disease}) =$	$99/100 =$	0.99	
$P(\text{Test-} \text{No Disease}) =$	$98901/99900 =$	0.99	
$P(\text{No Disease} \text{Test+}) =$	$999/1098 =$	0.91	

Probability example using a table

2014 State of Michigan Live Births

- Probability that a mother's age is 20-24 = .237
- Probability that a mother's age was ≤ 24 =
 - $.001 + .061 + .237 = .299$
 - By what probability rule?
- *Conditional probability*: Given that a mother's age is at least 30, what is the probability that she is at least 40?
- $P(\text{Mother's age} \geq 30 \text{ and Mother's age} \geq 40) / P(\text{Mother's age} \geq 30)$
 $= P(\text{Mother's age} \geq 40) / P(\text{Mother's age} \geq 30)$
 (why?)
 $= (.024) / ((30165 + 12475 + 2757) / 114460)$
 $= \mathbf{0.061}$

Maternal age group	Totals	% of total
<15	69	0.1%
15 - 19	6,968	6.1%
20 - 24	27,134	23.7%
25 - 29	34,884	30.4%
30 - 34	30,165	26.4%
35 - 39	12,475	10.9%
40+	2,757	2.4%
Total	114,460	100%

Examples of conditional probabilities

- Sensitivity & Specificity
- Relative risk

$$P(\text{disease} \mid \text{exposure}) / P(\text{disease} \mid \text{no exposure})$$

- Odds ratios

$$\frac{P(\text{exposure} \mid \text{disease}) / (1 - P(\text{exposure} \mid \text{disease}))}{P(\text{exposure} \mid \text{no disease}) / (1 - P(\text{exposure} \mid \text{no disease}))}$$

...we'll see these later in the course and use them in models as well as separate calculations

Independence

- If the occurrence of event B does not depend on the occurrence of event A, B and A are independent.
- Then $P(B|A)=P(B)$
- Examples:
 - Probability of becoming infected with malaria given that you have 2 sisters = Probability of becoming infected with malaria
 - Probability of a (fair) coin toss landing on heads given that the previous coin toss was a heads
 - Probability you are a banker and having green eyes

Multiplicative rule of probability

- The multiplicative rule if you have independence:

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

- Why?
- Rearrangement of the definition of conditional probability

$$P(B | A) = P(B \cap A) / P(A) = P(A \cap B) / P(A)$$

$$\text{so } P(A \cap B) = P(A) P(B | A)$$

- If you have independence: $P(B | A) = P(B)$ so you can substitute into above

Independence

Note that independence $\neq$ mutual exclusivity!

- Mutual exclusivity

- 2 events cannot both occur
- $P(A \cap B) = 0$

- Independence

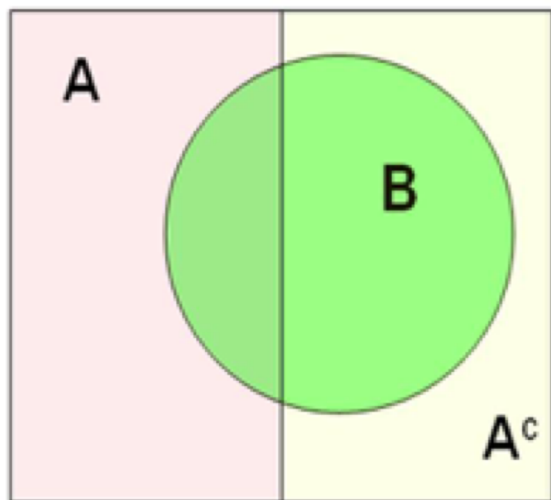
- 2 events do not depend on each other
- $P(B|A) = P(B)$

$$P(A \cap B) = P(A) P(B|A) = P(A) P(B)$$

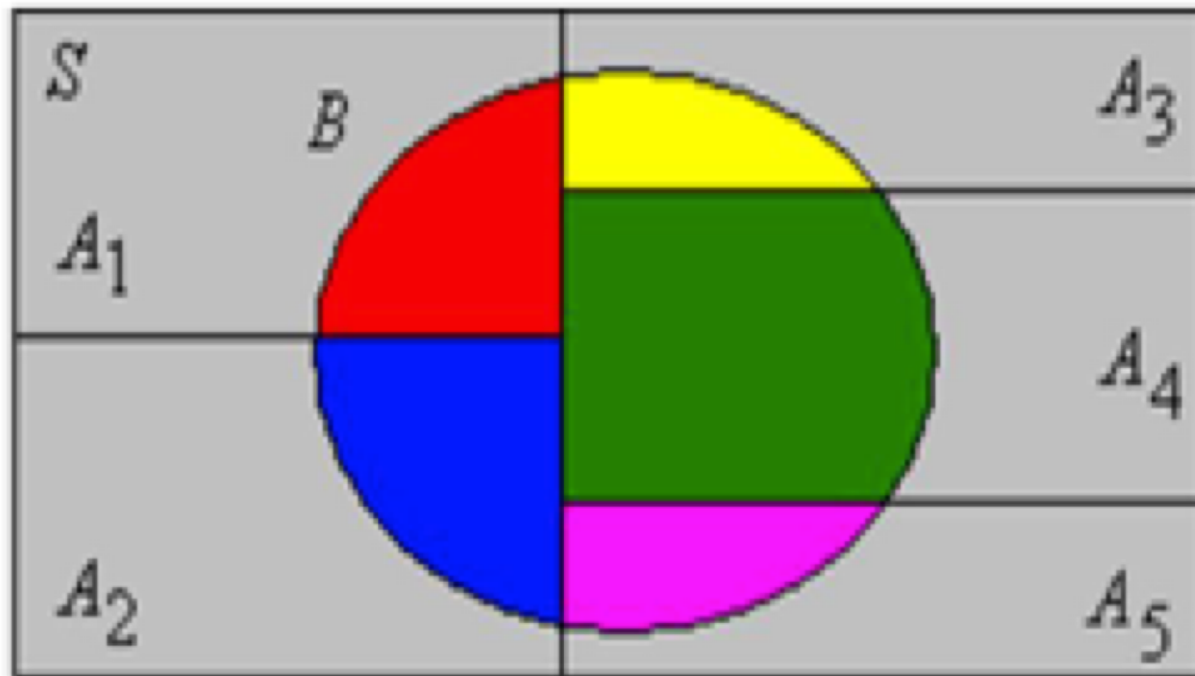
Law of Total Probability

- Helpful when you cannot directly calculate a probability and also if dealing with prevalence
 - Example
 - Suppose you know the prevalence of TB in the counties of an area, but you want to know the overall prevalence of TB in the region
- $$P(\text{TB+ in the region}) = P(\text{TB+} \cap \text{live in county1}) + P(\text{TB+} \cap \text{live in county2}) + P(\text{TB+} \cap \text{live in county3}) \dots \text{etc}$$
- $$P(\text{TB+ in country 1} \mid \text{live in county 1}) * P(\text{live in county 1}) + P(\text{TB+ in country 2} \mid \text{live in county 2}) * P(\text{live in county 2}) + \dots \text{etc.}$$
- This is a weighted average
 - This is very messy to calculate

Law of Total Probability Diagram



$$P(B) = P(A \cap B) + P(A^c \cap B)$$



$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n)$$

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_n)P(A_n)$$

Using Probability: Diagnostic Tests

- True positives – the test is positive given the person has the disease
 - The probability of this is $P(T^+ | D^+) = \text{Sensitivity}$
- False positives – the test is positive although the person does not have the disease
- True negatives – the test is negative given the person does not have the disease
 - The probability of this is $P(T^- | D^-) = \text{Specificity}$
- False negatives – the test is negative even though the person has the disease

Diagnostic Tests

- Divide the world into 4 quadrants based on disease and test result

		TRUTH	
		Disease D ⁺	No Disease D ⁻
Test	T ⁺	TP	FP
	T ⁻	FN	TN

- Sensitivity = $P(T^+ | D^+)$
= $P(T^+ \cap D^+) / P(D^+)$
= $TP / (TP + FN)$
- Specificity = $P(T^- | D^-)$
= $P(T^- \cap D^-) / P(D^-)$
= $TN / (FP + TN)$

Other examples similar to this?

Applying probability to diagnostic test results

- Suppose you have a panel of diagnostic tests that give false positive results 2% of the time.
- If you test a disease-free patient, what is the probability of the correct negative result?
 - This is the _____

Applying probability to diagnostic test results

- Suppose you test a disease-free patient twice with this test. What are the probabilities:
- Negative Negative
- Negative Positive
- Positive Negative
- Positive Positive
- What is the $P(1 \text{ or more tests are positive})$?

Applying probability to diagnostic test results

- Suppose you test a disease-free patient twice with this test. What are the probabilities:
- Negative Negative = $0.98 * 0.98 = 0.9604$
- Negative Positive = $0.98 * 0.02 = 0.0196$
- Positive Negative = $0.02 * 0.98 = 0.0196$
- Positive Positive = $0.02 * 0.02 = 0.0004$
- What is the $P(1 \text{ or more tests are positive})$?
 - Which of the above tests have at least 1 positive?
 - $= 0.0196 + 0.0196 + 0.0004 = 0.0396$

Applying probability to diagnostic test results

- What is the $P(1 \text{ or more tests are positive})$?
- This can also be written as $1 - (\text{Neg Neg})$
 $0.0396 = 1 - 0.9604$
- Then the probability of at least 1 false positive if 5 tests are run $= 1 - (0.98)^5 = 0.096$
- Suppose the false positive probability was $= 0.05$
then $= 1 - (0.95)^5 = 0.226$

Back to Bayes' theorem

- Suppose there is a super rapid HIV antibody test
 - Sensitivity = $P(T+ | HIV+) = 0.96$
 - Specificity = $P(T- | HIV-) = 0.99$
- You want to know the probability that someone with a positive test is truly HIV+
 - $P(HIV+ | T+)$
 - Positive Predictive Value (PPV)
 - You also know the $P(HIV+) = 0.02$ in this region

Recall Bayes' Theorem

$$P(A|B) = P(B|A)P(A) / P(B)$$

Proof:

- By definition of conditional probability
- $P(A|B) = P(A \cap B) / P(B)$
 $\rightarrow P(A \cap B) = P(A|B)P(B)$
- $P(B|A) = P(A \cap B) / P(A)$
 $\rightarrow P(A \cap B) = P(B|A)P(A)$
- $\rightarrow$ so $P(A|B) P(B) = P(B|A)P(A)$
- rearrange to get $\rightarrow P(A|B) = P(B|A)P(A) / P(B)$

Bayes' Theorem

By Bayes' theorem:

$$P(\text{HIV+} | T^+) = P(T^+ | \text{HIV+})P(\text{HIV+}) / P(T^+) \text{ using}$$

$$P(A | B) = P(B | A)P(A) / P(B)$$

Want to know $P(\text{HIV+} | T^+)$

Instead we know:

Sensitivity $P(T^+ | \text{HIV+})$ and

Specificity $P(T^- | \text{HIV-})$

and $P(T^- | \text{HIV+}) = 1 - \text{sensitivity}$ (false negatives)

and $P(T^+ | \text{HIV-}) = 1 - \text{specificity}$ (false positives)

Bayes' theorem

$$P(\text{HIV+} | T^+) = P(T^+ | \text{HIV+}) * P(\text{HIV+}) / P(T^+)$$

$$P(T^+ | \text{HIV+}) = 0.96 \text{ (sensitivity)}$$

$$P(\text{HIV+}) \text{ in region of study is } = 0.02$$

$P(T^+)$ = the overall chances of having a positive test

$$P(T^+) = P(T^+ | \text{HIV+}) P(\text{HIV+}) + P(T^+ | \text{HIV-}) P(\text{HIV-}) \text{ by the law of total probability (remember previous slides?)}$$

$$= 0.96 * 0.02 + 0.01 * 0.98 = 0.029$$

So this gives us:

$$\mathbf{P(\text{HIV+} | T^+) = 0.96 * 0.02 / 0.029 = 0.662}$$

Moving from the prior to posterior probability

- The prevalence of HIV+ was assumed to be = 2%
- Prior to testing the probability that a randomly selected person is infected with HIV = 0.02
 - This is the Prior probability
- The probability that someone who tests positive is actually infected with HIV = 0.662
 - This is the Posterior probability incorporating information gained from doing the test.

Random variables

- A random variable is some numerical outcomes of a random process
- Toss a coin 10 times
 $X = \#$ of heads
- Toss a coin until a head
 $X = \#$ of tosses needed

More random variables

- Toss a die
X=points showing
- Plant 100 seeds of pumpkins
X=% germinating
- Test a light bulb
X=lifetime of bulb
- Test 20 light bulbs
X=average lifetime of bulbs

Types of random variables

- Discrete or Countable valued

Counts, finite-possible values

- Continuous

Lifetimes, time

Probability distributions

- For a discrete random variable, the probability of for each outcome x to occur, denoted by $f(x)$, with properties

$$0 \leq f(x) \leq 1, \quad \sum f(x) = 1$$

Binomial distribution

- In many applied problems, we are interested in the probability that an event will occur x times out of n .

Binomial distribution

- Roll a die 3 times. $X = \#$ of sixes.

S=a six, N=not a six

No six: ($x=0$)

$$NNN \rightarrow (5/6)(5/6)(5/6)$$

One six: ($x=1$)

$$NNS \rightarrow (5/6)(5/6)(1/6)$$

$$NSN \rightarrow \text{same}$$

$$SNN \rightarrow \text{same}$$

Two sixes: ($x=2$)

$$NSS \rightarrow (5/6)(1/6)(1/6)$$

$$SNS \rightarrow \text{same}$$

$$SSN \rightarrow \text{same}$$

Three sixes: ($x=3$)

$$SSS \rightarrow (1/6)(1/6)(1/6)$$

Binomial distribution

- | x | $f(x)$ |
|-----|-----------------|
| 0 | $(5/6)^3$ |
| 1 | $3(1/6)(5/6)^2$ |
| 2 | $3(1/6)^2(5/6)$ |
| 3 | $(1/6)^3$ |

$$f(x) = \binom{3}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{3-x}$$

Binomial distribution

- Roll a die 20 times. $X = \#$ of 6's,
 $n=20, p=1/6$

$$f(x) = \binom{20}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{20-x}$$

$$p(x=4) = \binom{20}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^{16}$$

- Flip a fair coin 10 times. $X = \#$ of heads

$$f(x) = \binom{10}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x} = \binom{10}{x} \left(\frac{1}{2}\right)^{10}$$

Hypergeometric distribution

- Sampling with replacement

If we sample with replacement and the trials are all independent, the binomial distribution applies.

- Sampling without replacement

If we sample without replacement, a different probability distribution applies.

Hypergeometric distribution

Example

- 52 cards. Pick $n=5$.

X = # of aces,

then $a=4$, $b=48$

$$P(X = 2) = \frac{\binom{4}{2} \binom{48}{3}}{\binom{52}{5}}$$

Poisson distribution

- Events happen independently in time or space with, on average, λ events per unit time or space.
- Waiting for BART
- $\lambda=1$ train per 10 minutes
- Radioactive decay
 $\lambda=2$ particles per minute
- Lightning strikes
 $\lambda=0.01$ strikes per acre

For next time

- Review today's slides for Lab Thursday
- Lab this Thursday – Probability, identifying outliers
- Lab 2 & Homework 1 are posted on the CLE
- Next week's material will cover important probability distributions: Binomial, Normal etc