Biostat 200 Lecture 2

Today

- Discussion of data cleaning
- Probability
  - Definitions
  - Examples of use in diagnostic tests

Data cleaning

- Data cleaning is <u>always necessary</u> with a new data set
- Assume your data set has errors and <u>your job is to find them</u>
- Use tables and summary statistics and graphs to identify problems, outliers and anomalies
- Outliers
  - Extreme values, numerically distant from the bulk of the data
- We do NOT automatically remove outliers !!!
- Carefully document all decisions to drop or change values
- Always keep a copy of the original data

Outliers – what do we do?

- First consider if the value is physically possible
- Look at the other variables for clues
- Document all changes to data – and reasons
- **Always retain a copy of the original data set**

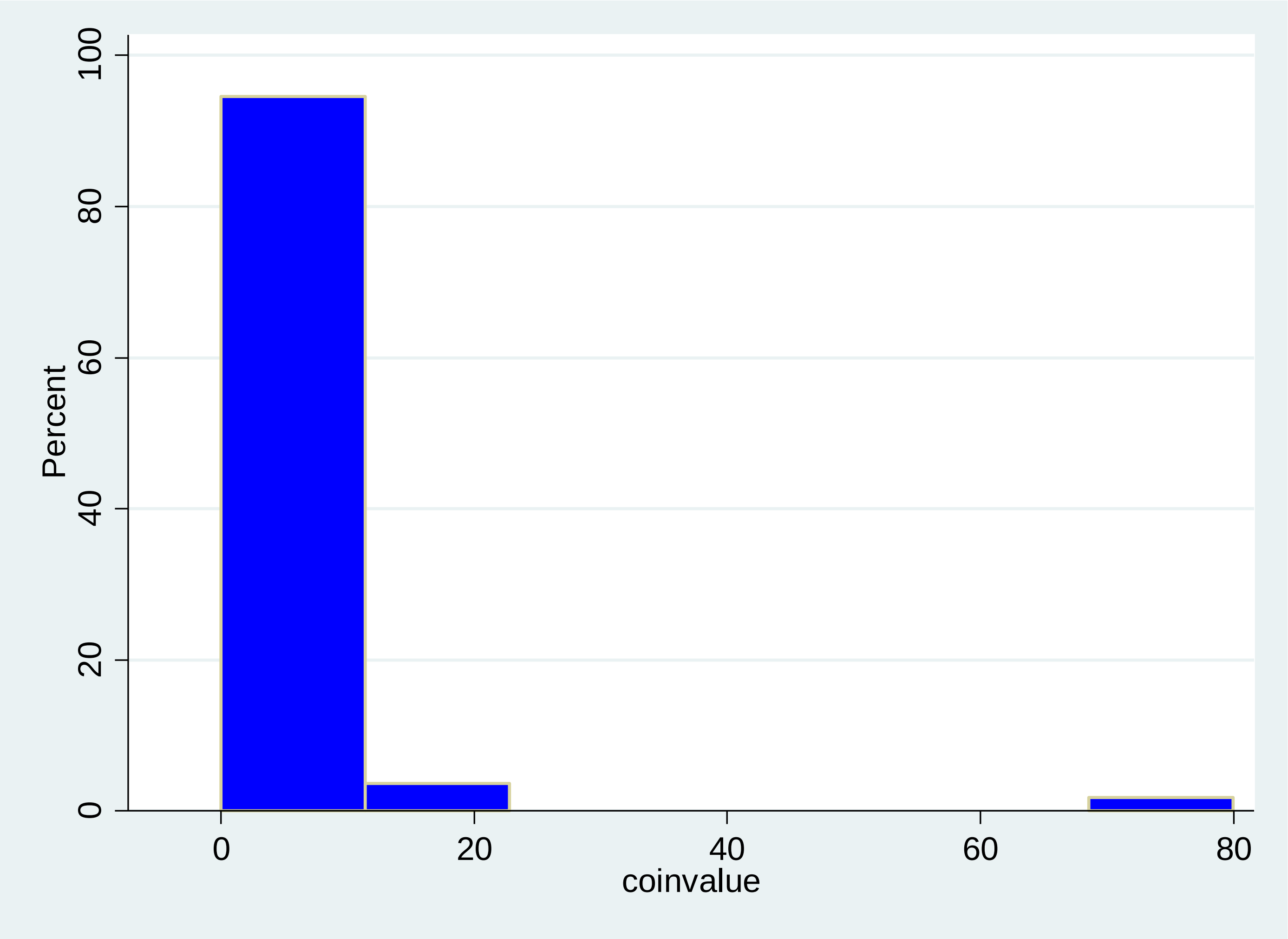
Outliers – what do we do?

- What about outliers that seem reasonable?
- Remember outliers are distant from the rest of the data, that doesn't mean they are wrong
- Keep them, but
  - Be aware that they may have large influence on some analyses
  - Think about more robust analyses (analyses not so sensitive to extreme values)
    - E.g. which measures of central tendency might you use?

Prevention

- Make sure numbers are coded as such – do not allow text
- Range checks, skip patterns
- Most programs have such capabilities
- Do double entry if data were not collected electronically
- Examine the data as early and as frequently as possible to check for problems

Coin value?




```
hist coinvalue, fcolor(blue) percent
(bin=7, start=0, width=11.428571)
```

Coin value?

```
. summm coinvalue, detail

                          coinvalue
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                  55
25%            0              0       Sum of Wgt.          55

50%          .67                      Mean           2.932727
                        Largest       Std. Dev.       10.9957
75%         1.25             10
90%            5             12       Variance       120.9054
95%           12             16       Skewness       6.471767
99%           80             80       Kurtosis       45.56809
```

Coin value?

```
. list if coinvalue>5

     +------------------------------------------------------------------------------+
     | sex    height   coinva~e   cashva~e   epi203   iphone6   child~yn   child~_n |
     |------------------------------------------------------------------------------|
 17. |   2       167         12          5        2         2          2          1 |
 33. |   2        63        5.3         80        2         2          1          0 |
 35. |   1        72         16      41.16        2         1          1          0 |
 45. |   1         6         10          .        2         2          1          0 |
 50. |   2   64.5669         80         85        2         1          1          0 |
     +------------------------------------------------------------------------------+
```

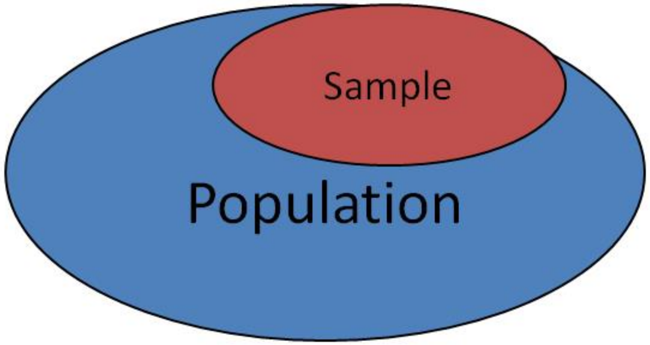
Coinvalue?

- Histogram looked odd
- Use summary command, thought about the range of values
- Use list command – see other relevant variables
- Back to the source documents (or study subjects) – did this person really have $80 coins?
- Make a decision and <u>document</u> it – use a .do file

Basic probability

Basic probability

- Probability is the foundation of statistical inference
  - Statistical inference is what is needed to make statements about the characteristics of the population from which a sample was drawn
  - p-values and confidence intervals tell us how our sample might relate to the population



- Many of the entities we use daily are probabilities – e.g. the probability of developing breast or ovarian cancer given a patient carries a BRCA1 or BRCA2 mutation

Event

- <u>Event</u>
  - Result of an experiment or observation
  - Occurs or does not occur
  - Denoted by uppercase letters e.g. A,B, X
  - We will apply probability to events
    - We want to estimate the relative frequency that an event occurs over a large number of identical trials
  - E.g.
    - An event may be a disease occurrence or the occurrence of a large laboratory value

Definition of probability

Frequentist definition of <u>probability</u>

- If an experiment is repeated n times under essentially identical conditions, and if the event A occurs m times, then <u>as n increases</u>,
  - the ratio m/n approaches a fixed number which is the probability of A occurring
- $P(A) = m/n$

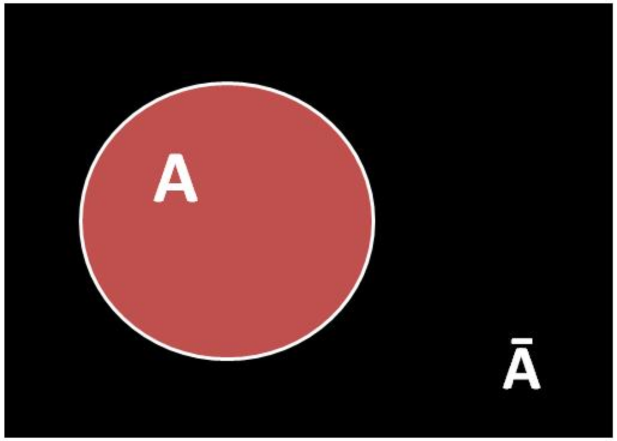
Probability of an event

- Probability of an event – relative frequency of its occurrence in a large number of trials repeated under the same conditions
  - E.g. Probability of picking a red ball out of a bag of red and black balls
  - Always lies between 0 and 1 (inclusive)
  - Denoted P(A) or P(X)

Complement

<u>Complement</u> of an event, $\bar{A}$ or A^C (read Not A or A complement)

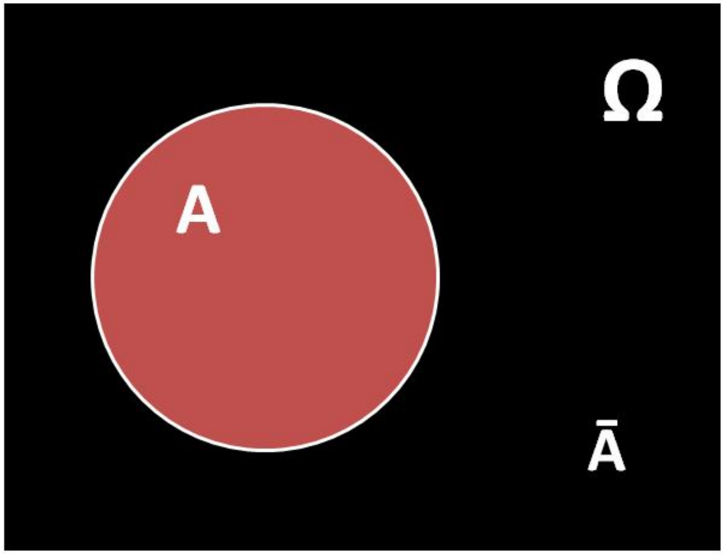
- E.g. A is the event that a person has Plasmodium falciparum (P.f.) malaria, $\bar{A}$ is the event of no P.f. malaria
- $P(A) = 1 - P(\bar{A})$
  - Often E for exposed and $\bar{E}$ for not exposed



Universe

<u>Universe</u> Ω is all the possible outcomes of an event

- $P(\Omega) = P(A) + P(\bar{A}) = 1$



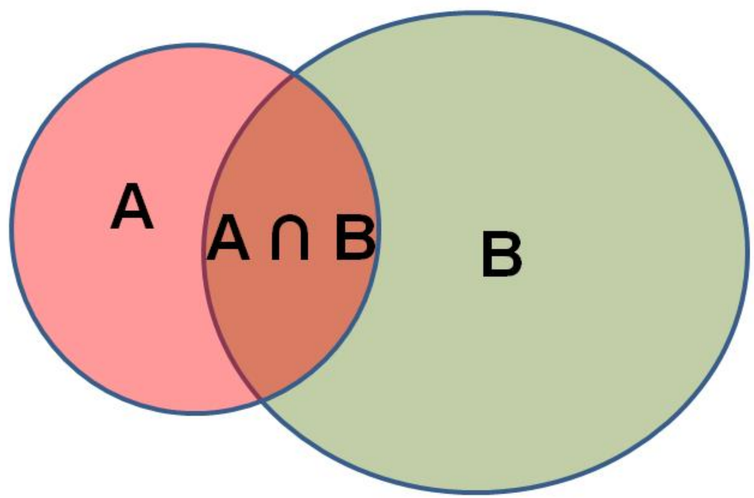
Complement example

- Probability that someone has Plasmodium falciparum (P.f.) malaria plus the probability that they do not
- $P(P.f.+) + P(P.f.-) = 1$

Intersection

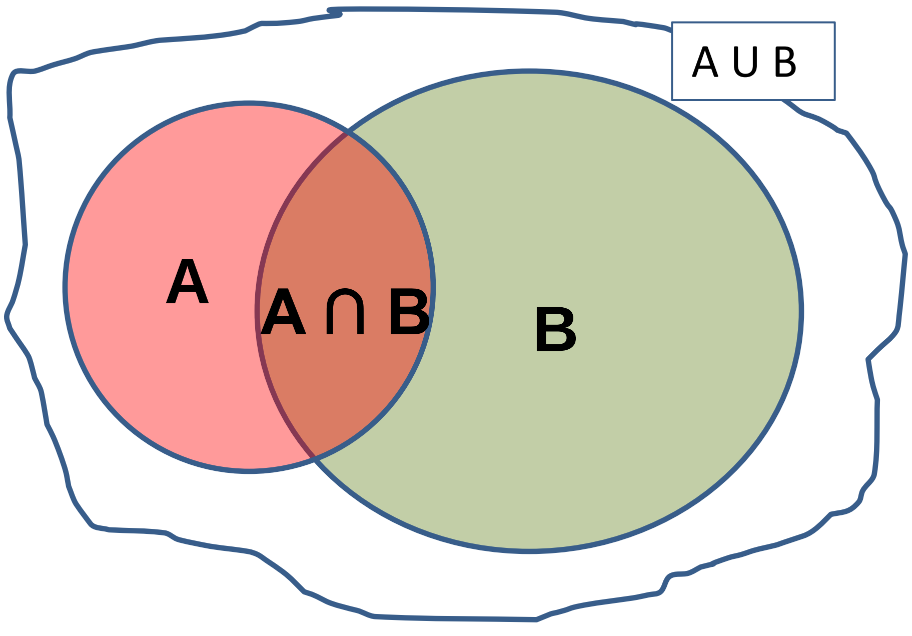
- The intersection of 2 events is written A ∩ B
- The intersection is when both A and B occur
  - E.g. The event that a person has both malaria and pulmonary tuberculosis
  - The probability that both occur is written

    P(A ∩ B)



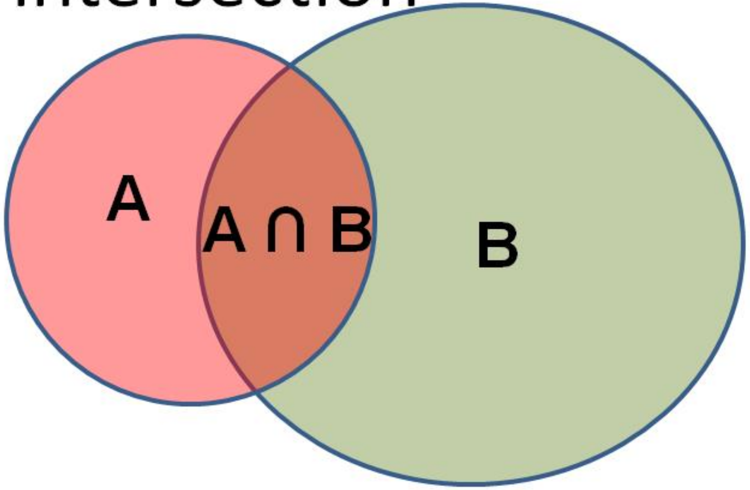
Union

- The union of 2 events is written A U B
- The union is if either A or B or both occur



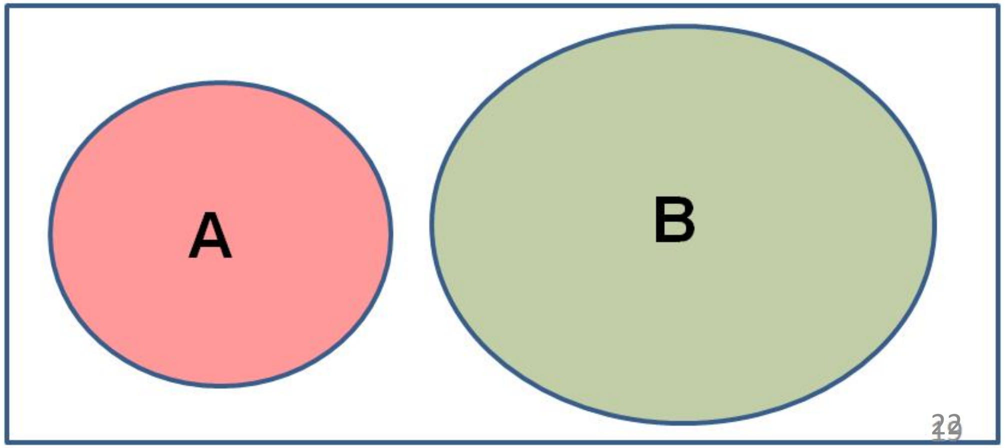
Union

- E.g. The event that a person has either malaria or tuberculosis or both
- **$P(A \cup B) = P(A) + P(B) - P(A \cap B)$**
- The probability of A or B or both is the sum of their individual probabilities minus the probability of their intersection



Mutual exclusivity

- Two events are <u>mutually exclusive</u> if they cannot occur together
  - There is no overlap area because both can't happen together
  - The intersection is empty
  - E.g.
    - Being pregnant and not pregnant
    - You cannot be both



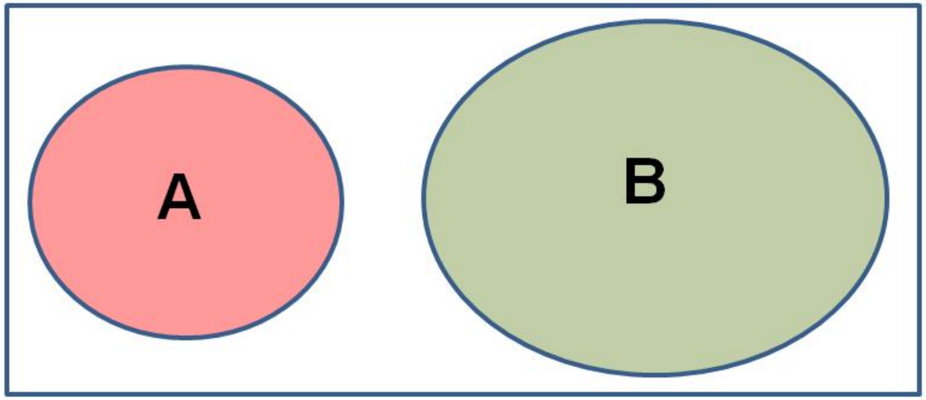
Mutual exclusivity

- For mutually exclusive events, the probability of A or B occurring is the sum of their individual probabilities
- Both A and B cannot occur together so

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

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

$$= P(A) + P(B)$$



Additive rule

- If A and B are **mutually exclusive**, the probability of either or both occurring is sum of the probability of either occurring

  $P(A \cup B) = P(A) + P(B)$

- This is the additive rule of probability

- E.g. Among persons with HCV infection (assuming only one infection at a time is possible)

  P(HCV genotype 1) in the US = .7

  P(HCV genotype 2) in the US = .15

  P(All other HCV genotypes) = .15

  → P(HCV genotype 1 or 2) = .7 + .15 = .85

Additive rule

- The additive rule of probability can be applied to several 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:

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

Probability summary

- Complement: $P(A)= 1-P(\bar{A})$
- Intersection: Prob (A *and* B)= $P(A \cap B)$
- Union: Prob (A *or* B *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

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

Probability summary

- Are A and $\bar{A}$ (i.e. an event and its complement) mutually exclusive?

- What is $P(A \cap \bar{A})$?

- What is $P(A \cup \bar{A})$?

Example

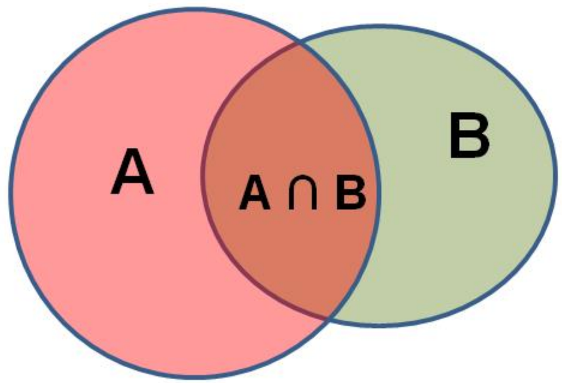
- A = the event that an individual is exposed to high levels of carbon monoxide (CO)
- B = the event that an individual is exposed to high levels of nitrogen dioxide (NO_2)
  - What is the event $A \cap B$?
  - What is the event A U B?
  - Are A and B mutually exclusive?

Conditional probability

- The probability that an event B will occur given that event A has occurred
  - **Notation: P(B|A)**
  - Read: the probability of B given A
- Example: Probability of a person becoming infected with malaria given that he/she uses a bed net at night
  - Event A is using a bed net
  - Event B is becoming infected with P.f. malaria

Conditional probability

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



The probability of B given A is the probability of the intersection (i.e. both events occurring) divided by the probability of A occurring

It is the relative size of the probability of the intersection $A \cap B$ compared to the relative size of the probability of A occurring

Conditional probability example

P(becoming infected with malaria | use a bed net)

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

= P(Becoming infected <u>and</u> using a bed net) / P(using a bed net)

Estimated by:

proportion of people who become infected with malaria and who use a bed net / proportion of people who use a bed net

or

number of people who become infected with malaria and who use a bed net / number of people who use a bed net

Probability example

1992 U.S. birth statistics

- Because the sample is so large the proportion approaches the probability
- Probability that mother's age was 20-24 = 0.263
- Probability that mother's age was ≤24

  = 0.003 + 0.124 + 0.263 = 0.390

- By what probability rule?

Age of mother	Probability
<15	0.003
15-19	0.124
20-24	0.263
25-29	0.290
30-34	0.220
35-39	0.085
40-44	0.014
45-49	0.001
Total	1.000

Probability example

1992 U.S. birth statistics

- Given that a mother is under age 30, what is the probability that she is under age 15?

P(Mother's age<15 | Mother's age<30)

= P (Mother's age<15 **and** <30) / P(Mother's age <30)

= (0.003) / (0.003 + 0.124 + 0.263 + 0.290)

= 0.003 / 0.68 = 0.004

Age of mother	Probability
<15	0.003
15-19	0.124
20-24	0.263
25-29	0.290
30-34	0.220
35-39	0.085
40-44	0.014
45-49	0.001
Total	1.000

Examples of conditional probabilities

- Relative risk is the ratio of 2 conditional probabilities

  RR = P(disease | exposed) / P(disease | not exposed)

- Odds ratios also include conditional probabilities

  OR = P(disease | exposed) / (1- P(disease | exposed)) /
  P(disease | not exposed) / (1- P(disease | not exposed))

also OR = P(exposed | disease) / (1- P(exposed | disease)) /
P(exposed | no disease) / (1- P(exposed | no disease))

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 coin toss landing on heads given that the previous coin toss was a heads

Multiplicative rule of probability

- $P(A \cap B) = P(A)\,P(B|A)$
  - This is a rearrangement of the definition of conditional probability $P(B|A) = P(A \cap B) / P(A)$
- Also: $P(A \cap B) = P(B \cap A)$

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

- If you have independence: $P(B|A) = P(B)$
  - Then the multiplicative rule (remember, just a rearrangement of conditional probability)

$P(A \cap B) = P(A)\,P(B|A)$ reduces to

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

Independence

Note that independence ≠ mutual exclusivity!

- Mutual exclusivity
  - 2 events cannot both occur
  - $P(A \cap B) = 0$
- Independence
  - 2 events do not depend on each other
  - $P(B|A) = P(B)$

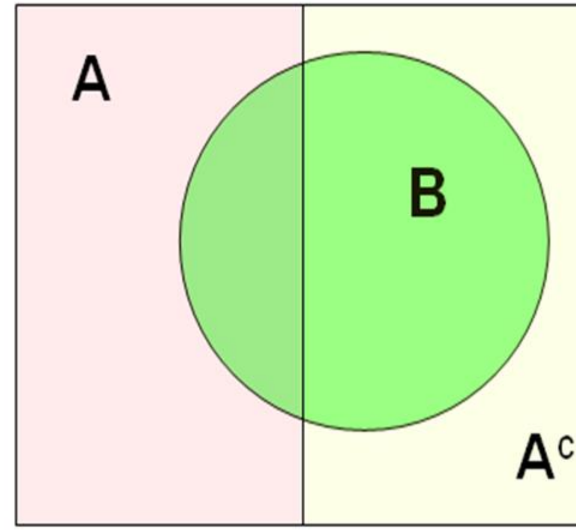
  $P(A \cap B) = P(A)\,P(B|A) = P(A)\,P(B)$

Law of Total Probability

- The law of total probability:

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

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

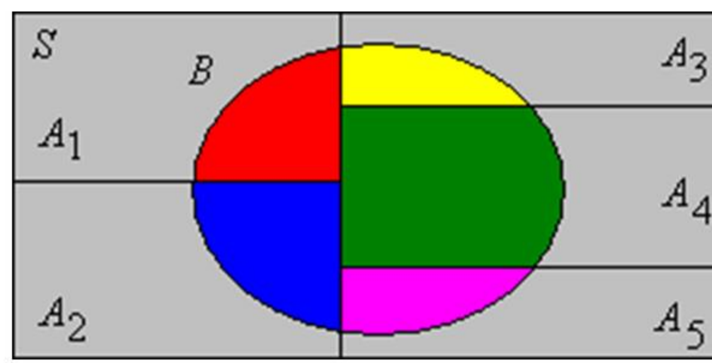


Law of Total Probability

More generally, if you can divide A into multiple mutually exclusive (non-overlapping) sections (here n=5):

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

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



Law of Total Probability

- Helpful when you cannot directly calculate a probability
- Example:
  - Suppose you know the TB prevalence in different areas and the population size in those areas, and you want to know the worldwide TB prevalence
  - P(TB+) = P(TB+| live in low/middle income country)*P(live in low/middle income country) + P(TB+| live in upper income country)*P(live in upper income country)
  - Weighted average of the 2 TB prevalences

Diagnostic tests

- Diagnostic tests of disease are not perfect
  - True positives – the test is positive given the person has the disease
    - The probability of this is $P(T^+|D^+)$ = 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 $P(T^-|D^-)$ = 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 $= P(T^- | D^-)$

  $= P(T^- \cap D^-)/P(D^-)$

  $= TN/(FP+TN)$

Diagnostic tests

- Diagnostic test characteristics (sensitivity and specificity) are based on experiments in which a test is compared to a "gold standard"="truth"

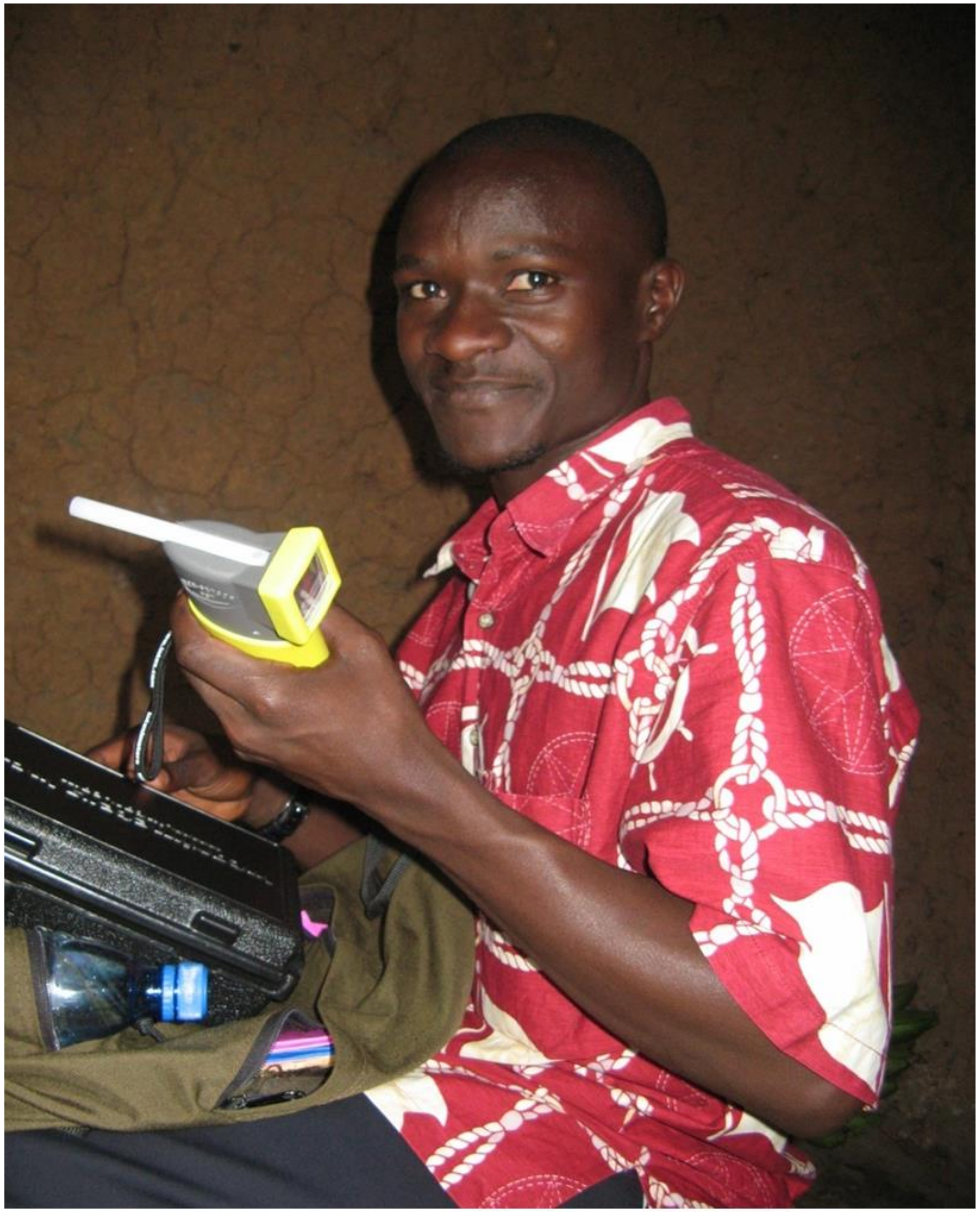
Diagnostic test validation example

- New biological markers of alcohol consumption are being developed. Phosphatidylethanol (PEth) is a metabolite of alcohol that is formed only in the presence of alcohol and lasts 2-3 weeks after moderate drinking.

- We examined 77 persons with HIV in Mbarara, Uganda. We followed them for 21 days and conducted daily breathalyzer tests and administered drinking surveys. If the breathalyzer result was ever >0 and/or the participant reported drinking, we considered this any alcohol consumption ("true drink+").

- We drew blood at the end of the 21-days to test for PEth.

Breathalyzer in Uganda



Diagnostic test example

- Sensitivity

  = P(PEth Test+ | true drink+)

  = TP/(TP+FN)

  = ???

- Specificity

  = P(PEth Test - | true drink-)

  = TN /(TN + FP)

  = ???

		“TRUTH”	
		drink+	drink-
Test	PEth+	45	3
	PEth-	6	23

Diagnostic tests

- Test results are often a continuous value (i.e. an optical density in an ELISA test)
- The level of the cutoff for a diagnostic test can be set to
  - Maximize sensitivity
    - This might be ideal if a follow up confirmatory test is easy and you want to be sure not to miss any positives
  - Maximize specificity
    - This might be necessary if there are grave ramifications of a false positive test
- When you change the cut point to maximize sensitivity you decrease specificity and vice-versa

Example: Test of T4 hormone levels to diagnose hypothyroidism

Total T4 (μg/dL)	True thyroid state	
	n Normal thyroid	n Hypothyroid
3 – 5	1	18
5.1 – 7	17	7
7.1 – 9	36	4
9.1 – 15	39	3
Total = 125	93	32

- If we decided that T4 values of 5 or less would be used to diagnose hypothyrodism, what would we get?

Total T4 (µg/dL)		True thyroid state	
		n Normal thyroid	n Hypothyroid
<= 5	Test +	1	18
> 5	Test -	92	14
		93	32

- Sensitivity = 18/32 = 0.56
- Specificity = 92/93 = 0.99

- If we decided that T4 values of 7 or less would be used to diagnose hypothyrodism, what would we get?

Total T4 (µg/dL)		True thyroid state	
		n Normal thyroid	n Hypothyroid
<=7	Test +	18	25
>7	Test -	75	7
		93	32

- Sensitivity = 25/32 = 0.78
- Specificity = 75/93 = 0.81

- If we decided that T4 values of 9 or less would be used to diagnose hypothyrodism, what would we get?

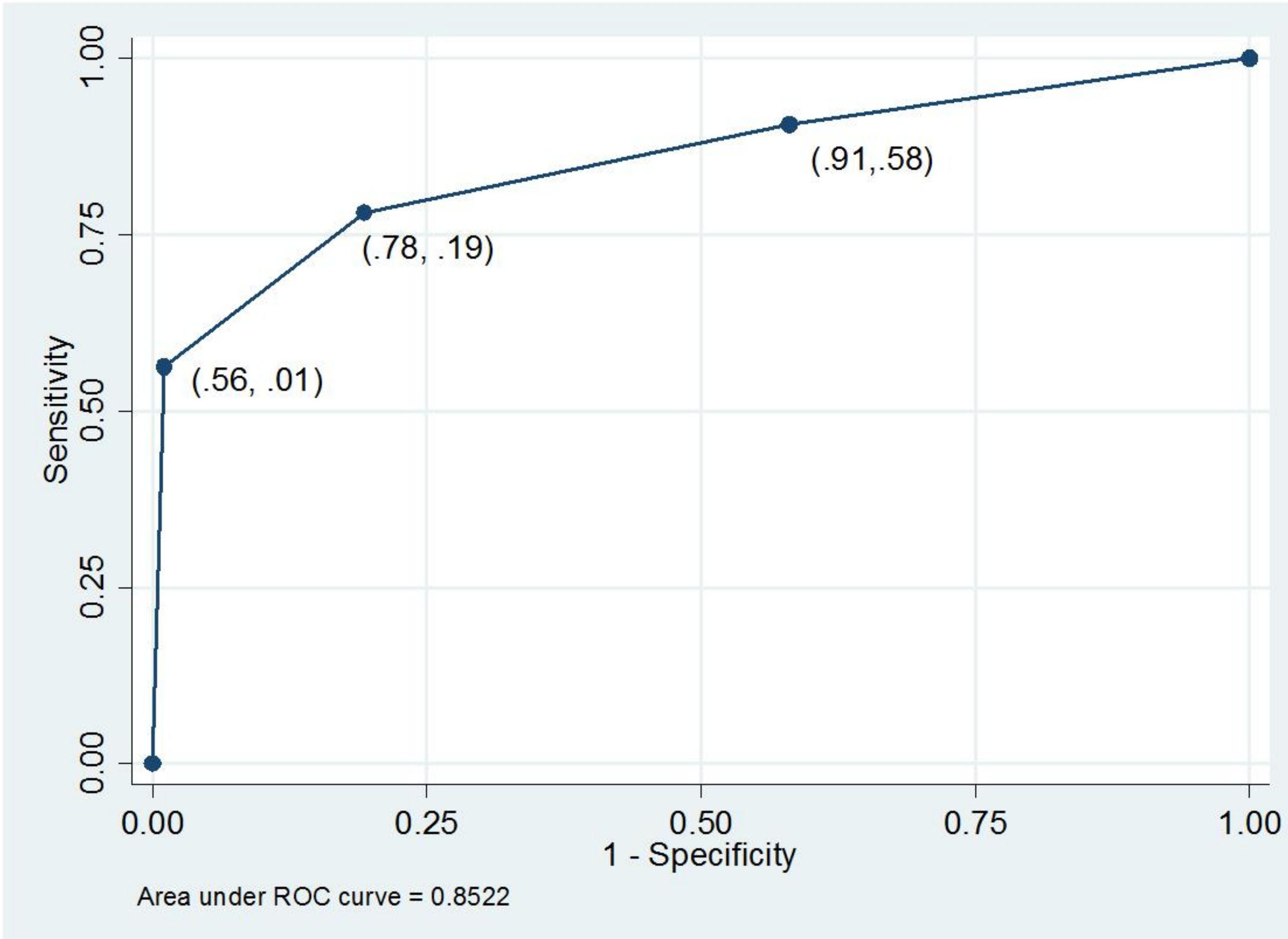
Total T4 (µg/dL)		**True thyroid state**	
		n Normal thyroid	n Hypothyroid
<=9	Test +	54	29
>9	Test -	39	3
		93	32

- Sensitivity = 29/32 = 0.91
- Specificity = 39/93 = 0.42

- Receiver-operator curves plots the sensitivity versus the 1-specificity for a test at every possible test cutoff

Cut point (value or less is pos)	Sensitivity	Specificity	1-Specificity
5	.56	.99	.01
7	.78	.81	.19
9	.91	.42	.58

ROC curve


1.00
0.75
0.50
0.25
0.00
Sensitivity
(.91,.58)
(.78, .19)
(.56, .01)
0.00
0.25
0.50
0.75
1.00
1 - Specificity
Area under ROC curve = 0.8522

Application of laws of probability to diagnostic tests

- Suppose you have a panel of diagnostic tests and each give false positive results 2% of the time (98% specificity)
- If you test your disease free patient with one of the tests, there is a 2% chance you'll get a false positive result
- There is a 98% chance you will get the correct negative result.

Application of laws of probability to diagnostic tests

- If you test a patient (without disease) with two tests, what are the possible results that you would get and what is the probability of those results?

- Neg Neg

  $P(\text{Neg test 1} \cap \text{Neg test 2}) = 0.98*0.98=.9604$ (why?)
- Neg Pos

  $P(\text{Neg test 1} \cap \text{Pos test 2}) = 0.98*0.02=.0196$
- Pos Neg

  $P(\text{Pos test 1} \cap \text{Neg test 2}) = 0.02*0.98=.0196$
- Pos Pos

  $P(\text{Pos test 1} \cap \text{Pos test 2}) = 0.02*0.02=.0004$

- All 4 of these possibilities add to 1 (why?)

  $.9604 + .0196 + .0196 + .0004 = 1$

Application of laws of probability to diagnostic tests

- What is the probability one or more tests will be positive?
- P(1 or more test is pos) =

  P(Neg test 1 ∩ Pos test 2) +

  P(Pos test 1 ∩ Neg test 2) +

  P(Pos test 1 ∩ Pos test 2)

  = .0196 + .0196 + .0004

  =.0396

Application of laws of probability to diagnostic tests

- An easier way:

  P(1 or more test is pos) = **1-P(both tests are neg)**

- P(both tests are neg) = P(Neg test 1 $\cap$ Neg test 2)

  $=0.98 * 0.98 = 0.98^2 = 0.9604$
- So P(1 or more test is pos) = $1 - 0.98^2 = 0.0396$
- In general, P(At least one false positive)

  = 1-P(no false positives occur over all tests)

  = 1-Test specificity$^{\text{\# of tests}}$

Application of laws of probability to diagnostic tests

- What is the probability of at least one false positive if 5 tests were run?

  $1-0.98^5 = 0.096$

- What if the false positive probability was .05?

  $1-0.95^5 = 0.226$

- What is the probability of at least one false positive if 10 tests were run (where P(FP)=0.02)?

  $1-0.98^{10} = 0.183$

- What if the false positive proportion was .05?

  $1-0.95^{10} = 0.401$

Bayes' theorem for diagnostic tests

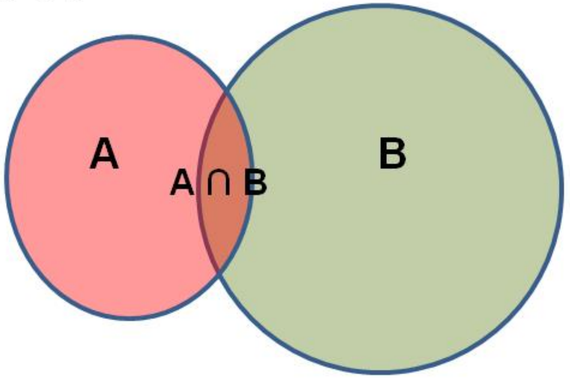
- Suppose you know from diagnostic testing that
  - The sensitivity of a new super rapid HIV antibody test ($P(T^+|HIV+)$) is 0.96
  - The specificity $P(T^-|HIV-)$ of the test is 0.99
- You want to know the probability that someone with a positive test using this test is truly infected with HIV
  - What is $P(HIV+|T^+)$?
  - This is called the Positive Predictive Value (PPV) of the test

Bayes' theorem

- Suppose we know the sensitivity of a test P(T+ | D+), but we want to know the probability of disease given a positive test P(D+ | T+) ?

- Bayes theorem allows us to switch between conditional probabilities
- $P(A|B)=P(B|A)P(A) / P(B)$

Bayes' theorem



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

- Proof:
  - By definition of conditional probability
  - $P(A|B)=P(A\cap B)/P(B)$

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

    → $P(A\cap B) = P(B|A)P(A)$

  → so $P(A|B)\ P(B) = P(B|A)P(A)$

  rearrange to get → $P(A|B)=P(B|A)P(A) / P(B)$

Bayes' theorem for diagnostic tests

By Bayes' theorem:

$P(HIV+|T^+) = P(T^+|HIV+)P(HIV+) \,/\, P(T^+)$ using

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

Probability of being truly infected with HIV (HIV+) if you have a positive test result

Bayes' theorem for diagnostic tests

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

Instead we know:

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

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

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

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

Bayes' theorem for diagnostic tests

$P(HIV+|T^+) = P(T^+|HIV+)*P(HIV+) / P(T^+)$

$P(T^+|HIV+) = 0.96$ (sensitivity)

$P(HIV+)$ in region of study is $= 0.02$

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

$P(T^+) = P(T+|HIV+)\ P(HIV+) + P(T+|HIV-)\ P(HIV-)$ by the law of total probability (remember slides 38-39?)

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

$P(HIV+|T^+) = 0.96*0.02/0.029 = 0.662$

Prior and posterior probability

The prevalence of HIV was assumed to be 2%

So before testing, the probability that a randomly selected person is infected with HIV is .02

This is the prior probability.

The probability that someone who tests positive in this setting actually is infected with HIV is 0.662

This is the posterior probability

It incorporates the information gained by doing the test

In reality, HIV tests have much higher sensitivity than 96% … So the PPV is higher

Bayes' theorem for diagnostic tests

What is $P(HIV+|T^{+})$ in a population in which the HIV prevalence is 0.004?

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

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

$P(HIV+)$ is $=0.004$

$P(T^{+}) = P(T+|HIV+)\ P(HIV+) + P(T+|HIV-)\ P(HIV-)$

$= 0.96*0.004 + 0.01*0.996 = 0.0138$

$P(HIV+|T^{+}) = 0.96*0.004/0.0138 = 0.278$

Bayes’ theorem

Bayes’ theorem allows you to use what you know about the conditional probability of one event on another to help you understand the inverse

For next time

- Read Pagano and Gauvreau
  - Chapter 6 (Review of today's material)
  - Chapter 7