Quantifying Distributions

Several metrics characterize shape of distributions

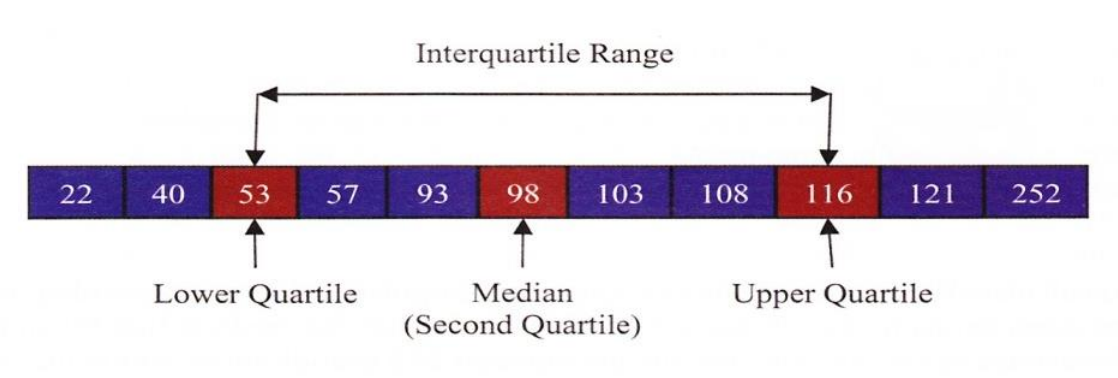
Interquartile Range:
Range of values between the 75% and the 25% percentiles of the data

Dist1: 4 – 2 = 2 (from 2 to 4)
Dist2: 4 – 2 = 2 (from 2 to 4)

What are 25% and 75% quartiles of data below ?

dist1: 1, 2, 3, 4, 5

dist2: 1, 2, 3, 4, 50



75%: 116
25%: 53

IQ Range: 63
(from 53 to 116)

Quantifying Distributions

Several metrics characterize shape of distributions

Mean: The average of the scores in the distribution

$$\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n}$$

What are the means of dist1 and dist2, below?

dist1: 1, 2, 3, 4, 5

mean1: 15 / 5 = 3

dist2: 1, 2, 3, 4, 50

mean2: 60 / 5 = 12

Quantifying Distributions

Distribution shapes categorized by symmetry (skew)

Skew: Measure of the symmetry of a distribution.

Symmetric distributions have a skew = 0.

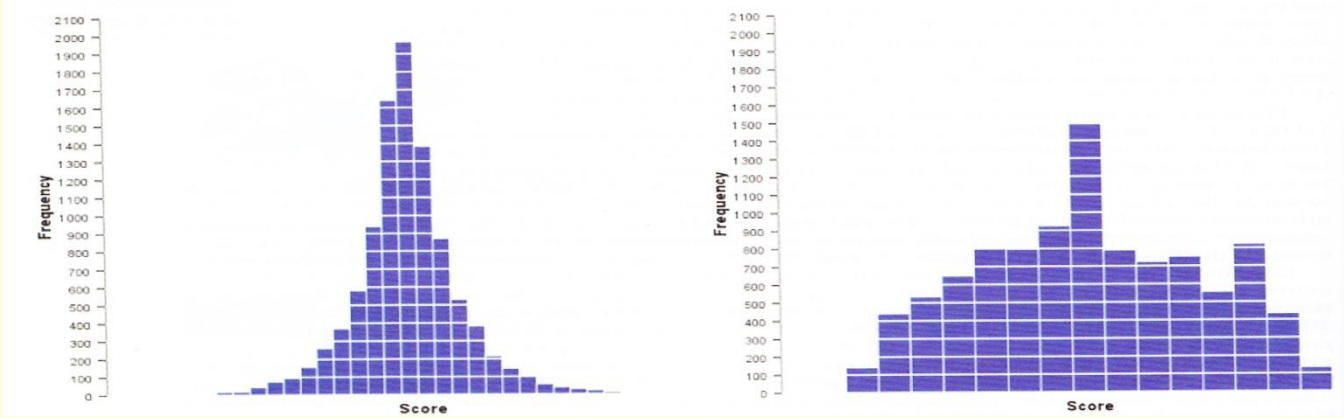
Positive skew:
the mean is larger
than the median,
skewness > 0

Negative skew:
the mean is smaller
than the median,
skewness < 0

Quantifying Distributions

Distribution shapes categorized by kurtosis

Kurtosis: Measure of the degree to which observations cluster in the tails or the center of the distribution.



Positive kurtosis:
Less values in tails and more values close to mean.
Leptokurtic.

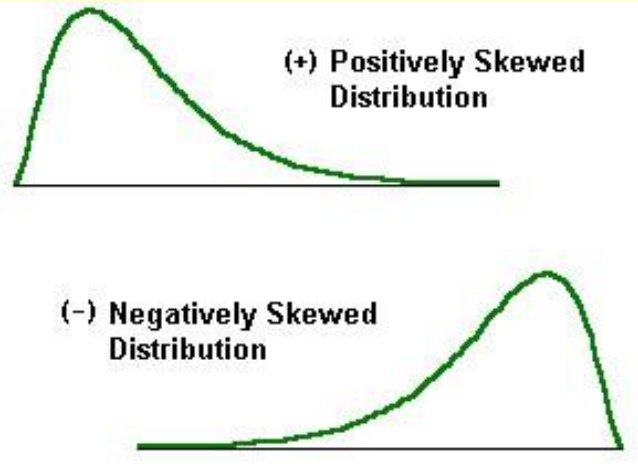
Negative kurtosis:
More values in tails and less values close to mean.
Platykurtic.

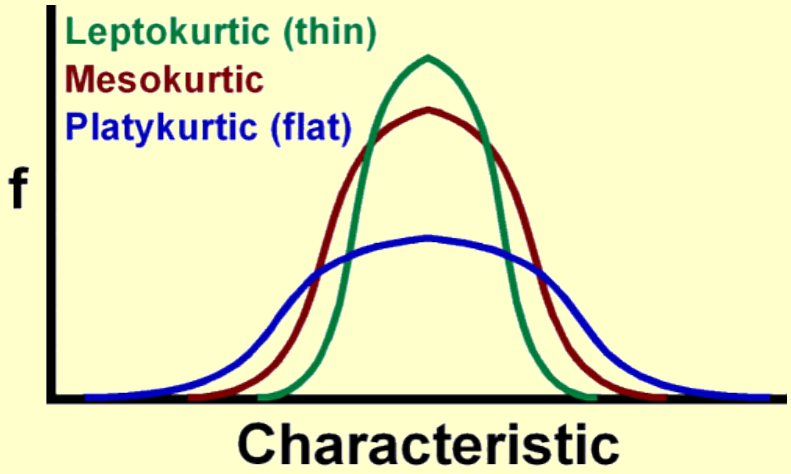
Quantifying Distributions

DO NOT NEED TO MEMORIZE: skewness and kurtosis:

$$\text{Skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

$$\text{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$





Summary

Quantifying Probability

Consider sample space and conditional probability

Be ready to perform calculations like in examples

Defining Variables

Influence how we measure / quantify observations

Memorize important definitions

Characterizing Variables

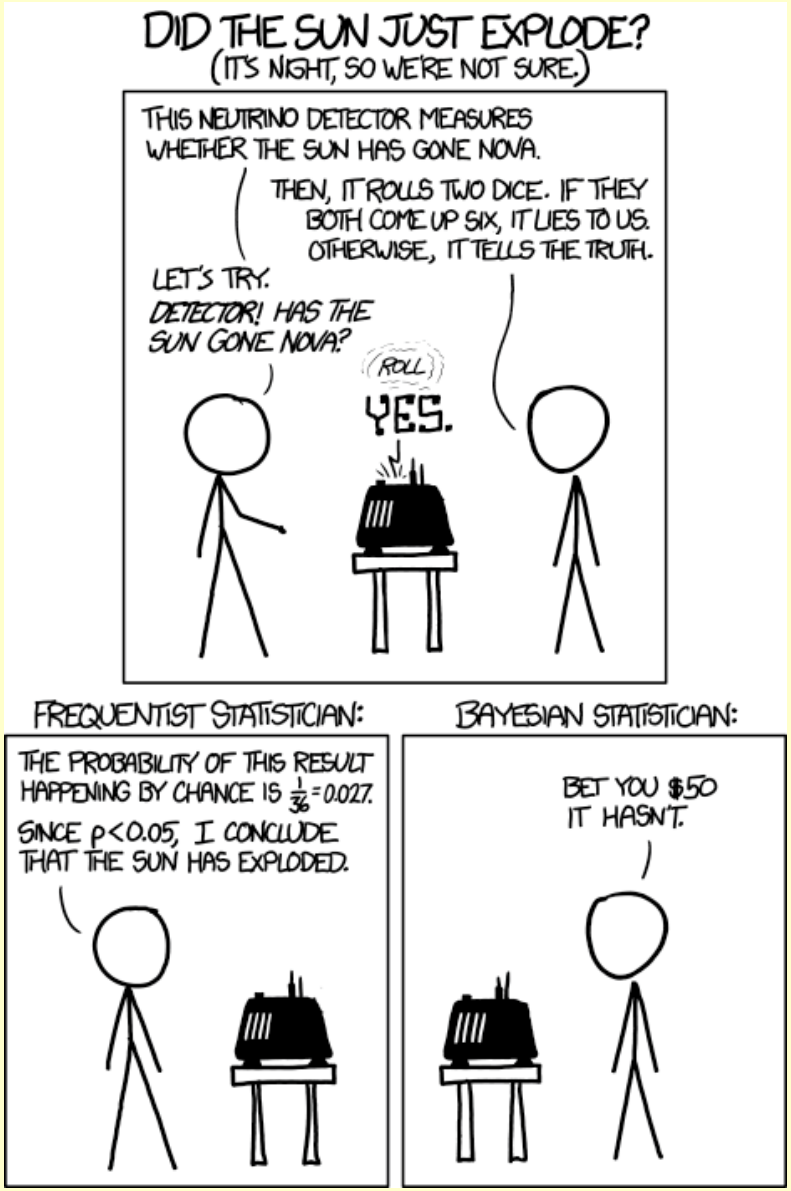
Use frequency tables and distributions

Distributions characterized with certain metrics:

range, median, mean, mode, skew, kurtosis

Memorize important definitions

Conditional Probability & Bayesian Statistics



http://www.pelagicos.net/classes_biometry_fa16.htm

Fisherian (Frequentist) Statistics

Frequentist statistics provides a generally applicable scheme for making statistical inference.

Based on estimating probability of the data, if the null hypothesis is, in fact, true.

It implies drawing conclusions from sample data by the emphasis on the frequency or proportion of the data.

- Define Population of Interest
- Sample Randomly from that population
- Estimate Parameters of Population, using the Sample

Probabilistic Independence

Definition: Two events, A and B, are independent if the fact that A occurs does not affect the probability of B occurring.

Some examples of independent events:

Landing on heads after tossing a coin AND rolling a 5 on a single 6-sided die.

Choosing a marble from a jar AND landing on heads after tossing a coin.

Choosing a 3 from a deck of cards, replacing it, AND then choosing an ace as the second card.

Conceptually, how can you show whether two events are independent ?

P (sun)
= 2 / 5 = 0.4

P (rain)
= 3 / 5 = 0.6

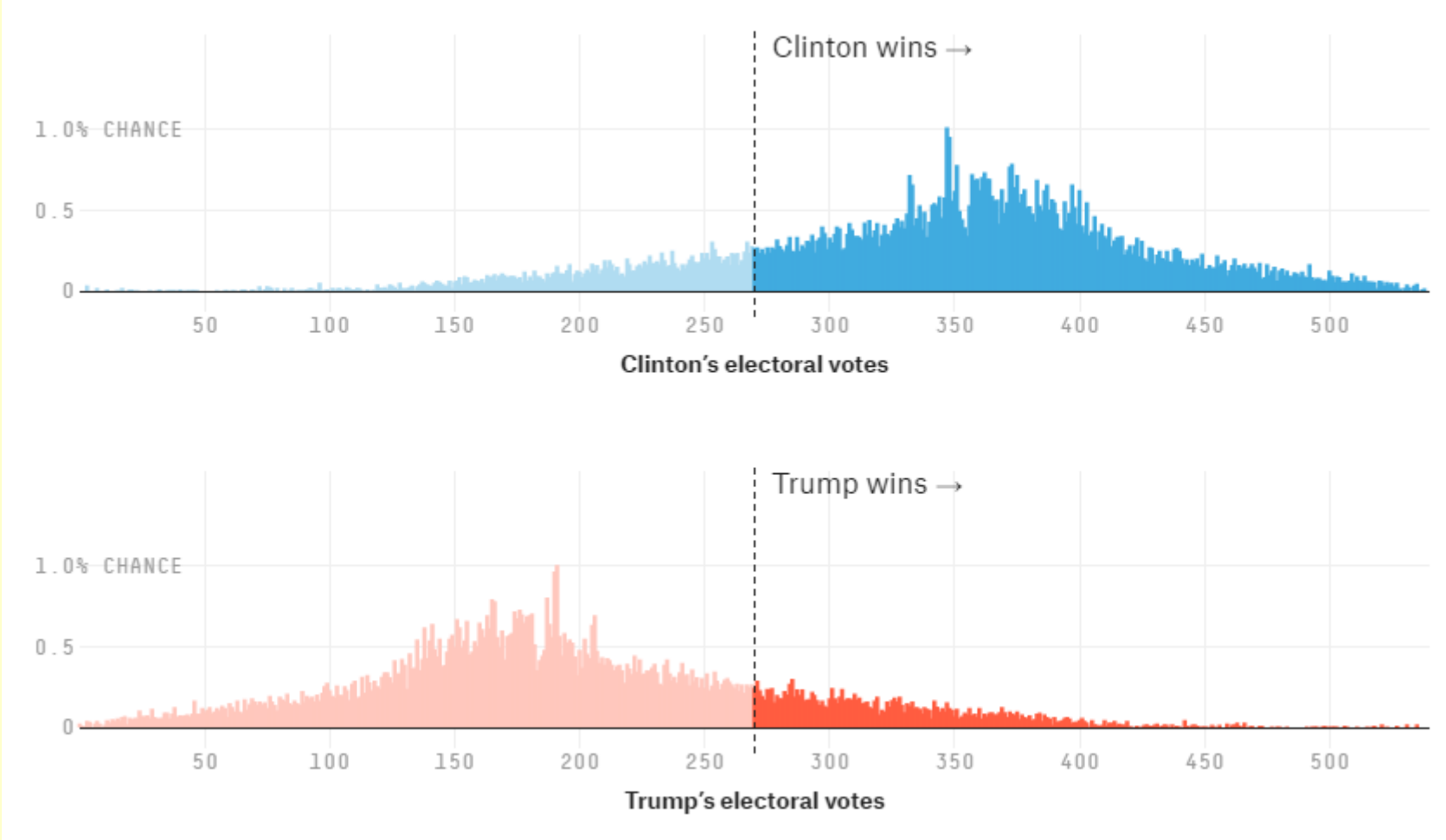
	EXPECTED	OBSERVED (longer dataset)
P (sun - sun)	= 0.4 * 0.4 = 0.16	16 / 100 = 0.16
P (sun - rain)	= 0.6 * 0.4 = 0.24	24 / 100 = 0.24
P (rain - sun)	= 0.6 * 0.4 = 0.24	24 / 100 = 0.24
P (rain - rain)	= 0.6 * 0.6 = 0.36	36 / 100 = 0.36

RESULT: Independent

Probability Trees

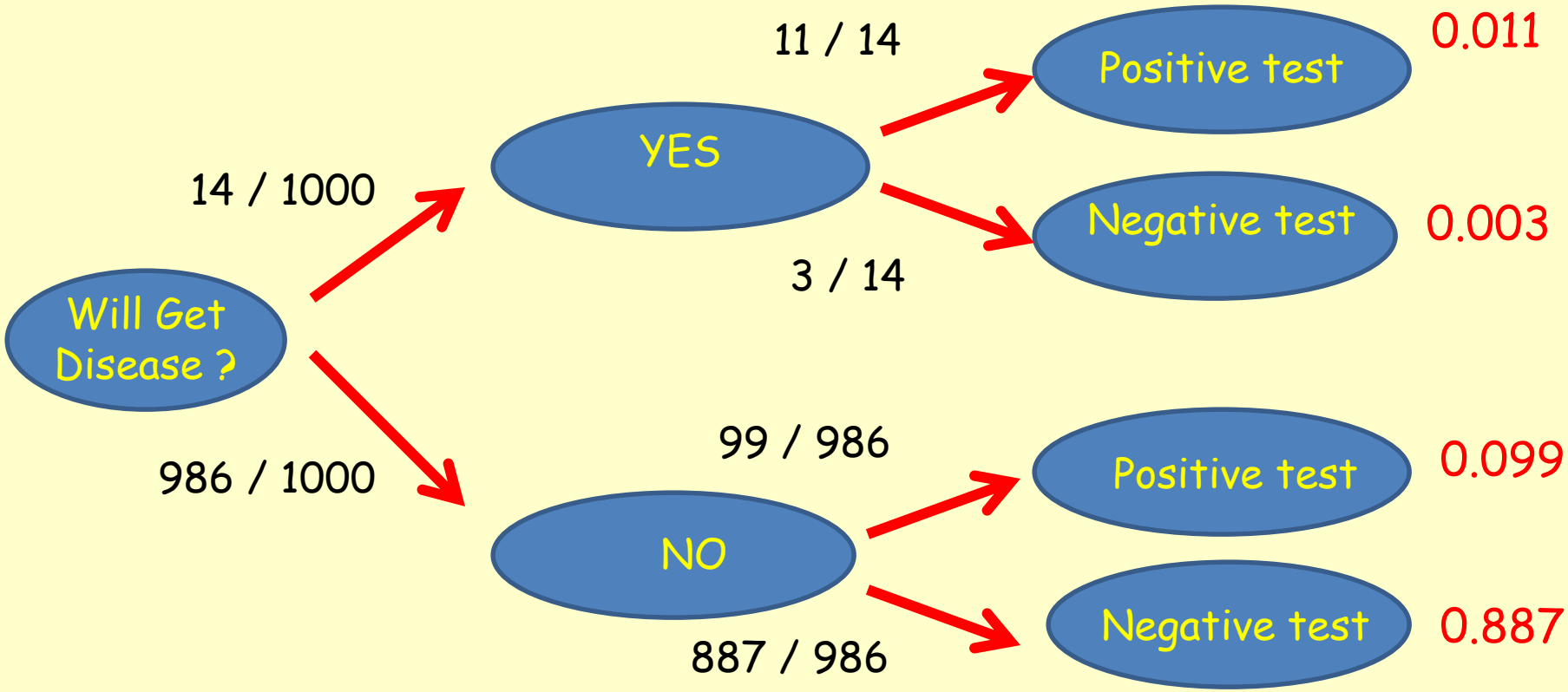
FiveThirtyEight

Question: Who will win 2016 Presidential Election?



Example – Probability Trees

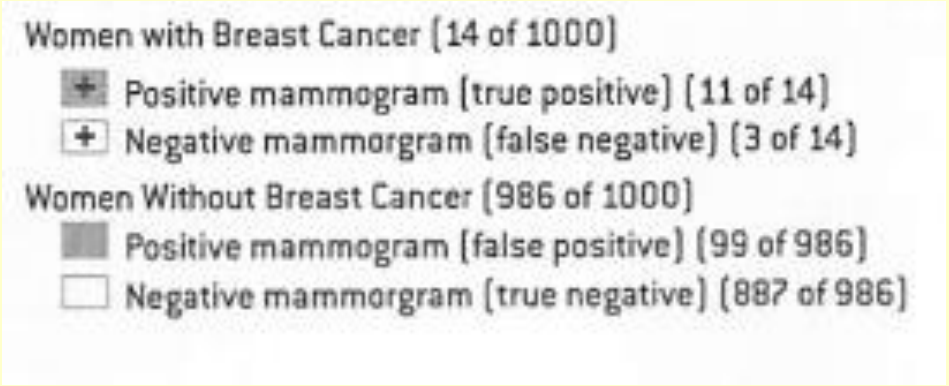
Question: What is the probability of getting a disease if you tested positive ?



Hypothesis Testing with Frequentist Stats

- Approach:

Outcome is binary:
accept / reject



Compare hypothesis and decision about hypothesis

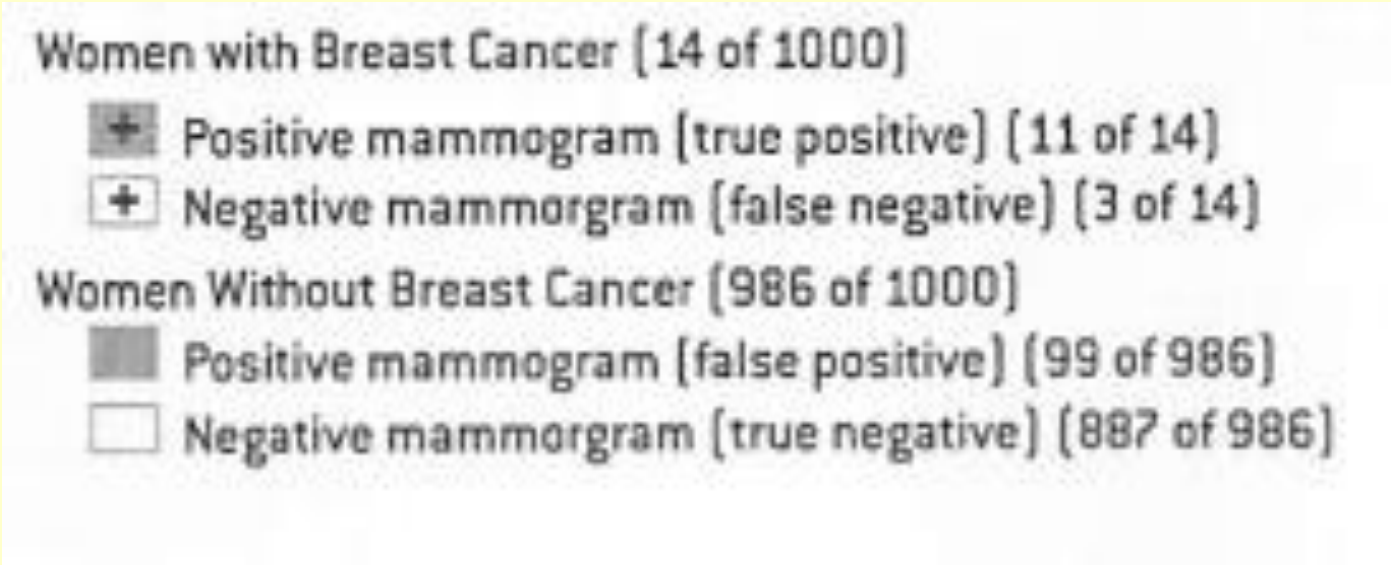
Decision about Null Hypothesis	Null Hypothesis: True	Null Hypothesis: False
Reject	99 / 986	11 / 14
Accept	887 / 986	3 / 14

Hypothesis Testing with Bayesian Stats

- Approach:

Based on estimating probability of the various hypotheses, given the data (multiple observations)

Outcome is continuous: actual probabilities



Conditional Probability

When calculating the probability of a complex event, information on known outcomes can be used to revise the probability calculations.

These updated estimates are termed "conditional probabilities" $P(A \mid B)$

Probability of event A, given than event B occurred

e.g., How would knowing that the first coin was a head, modify the probability of getting two heads OR one head & 1 tail ?

Conditional Probability

When calculating the probability of a complex event, information on known outcomes can be used to revise the probability calculations.

What is the probability of 2 heads, 2 tails, 1 head & 1 tail in 2 tosses?

(1/4)

(1/2)

(1/4)

How would knowing that first coin was a head, modify the probabilities?

Event 1

Event 2

(1/2)

(1/2)

Bayes Theorem – getting closer and closer to the truth as we gather more evidence

Concerned with conditional probability.

It quantifies the probability that a hypothesis is true if some event has happened.

Goal: to calculate the "Posterior Probability", incorporating the new observation (event)

Bayes Theorem – getting closer and closer to the truth as we gather more evidence

- Approach: Need to consider three quantities:

- Probability of observation, if hypothesis is true

- Probability of observation, if hypothesis is not true

- The Prior: Probability of the hypothesis, before the (new) observation was made

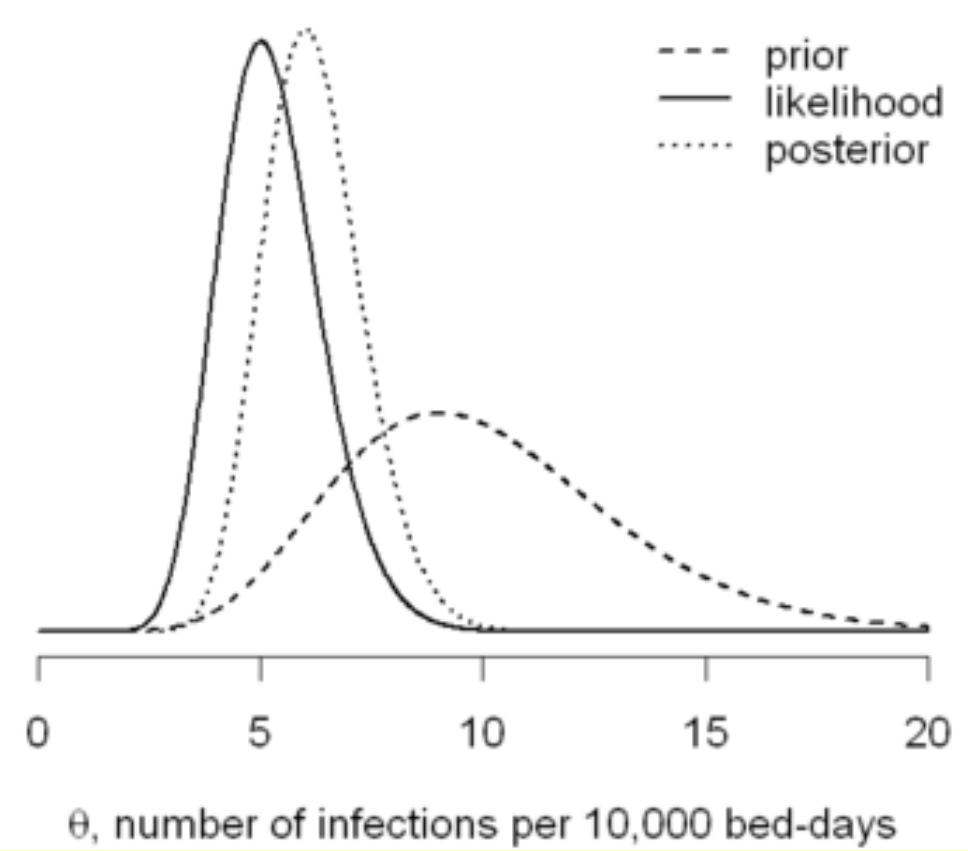
- Examples:

What is probability of a growing (bull) stock market ?
What is the expected number of disease infections ?

Bayes Theorem – Modifying the Prior

Prior, likelihood and posterior distributions for θ, the rate of disease infections per 10,000 bed-days.

The posterior distribution is a formal compromise between the likelihood, summarizing the evidence in the data alone, and the prior distribution, which summarizes the external evidence of higher rates.

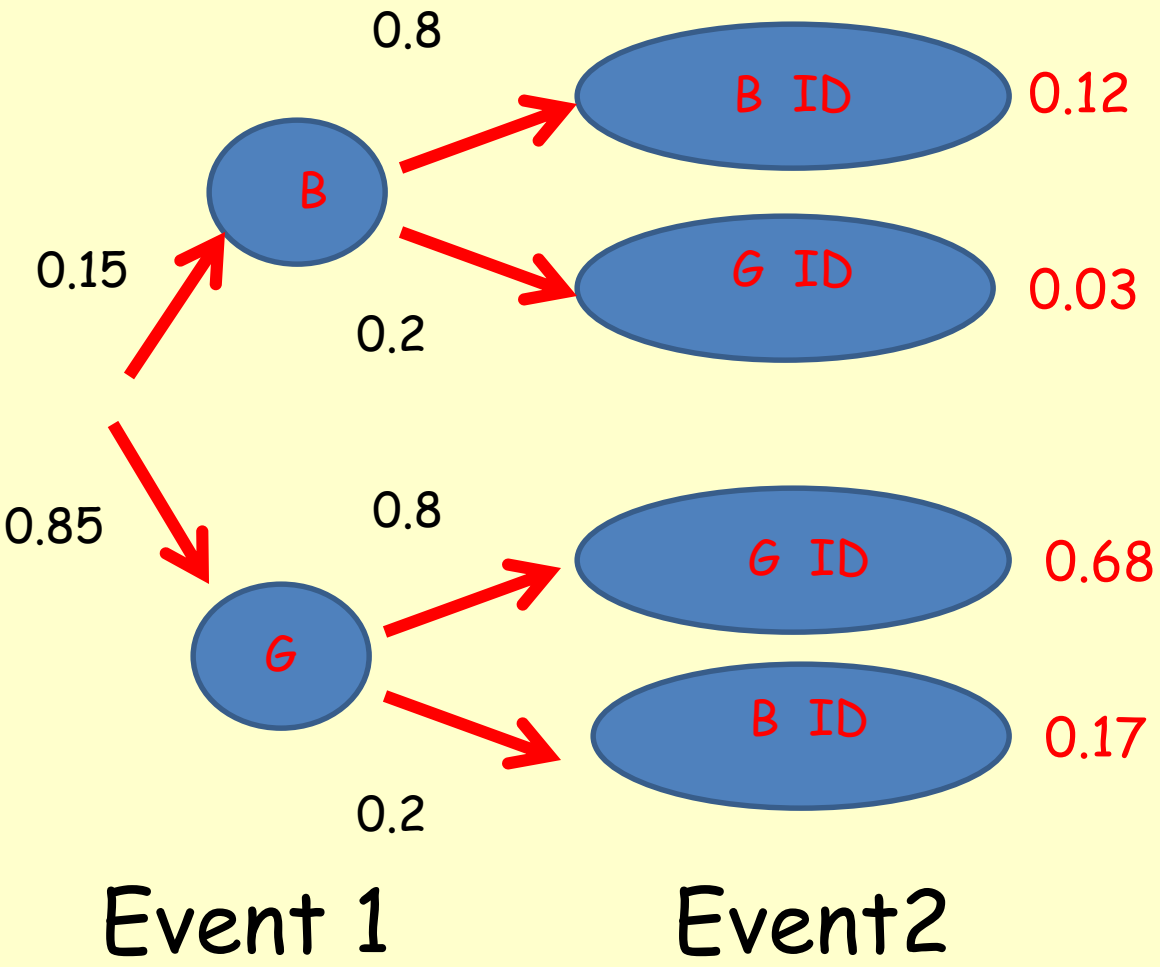


Conditional Probability

Question: What is probability that car was blue if identified as a blue car?

Car colors:
85% green
15% blue

ID:
80% correct
20% incorrect



Solution:

$$\frac{\text{Blue ID \& Blue Car}}{\text{Blue ID}}$$

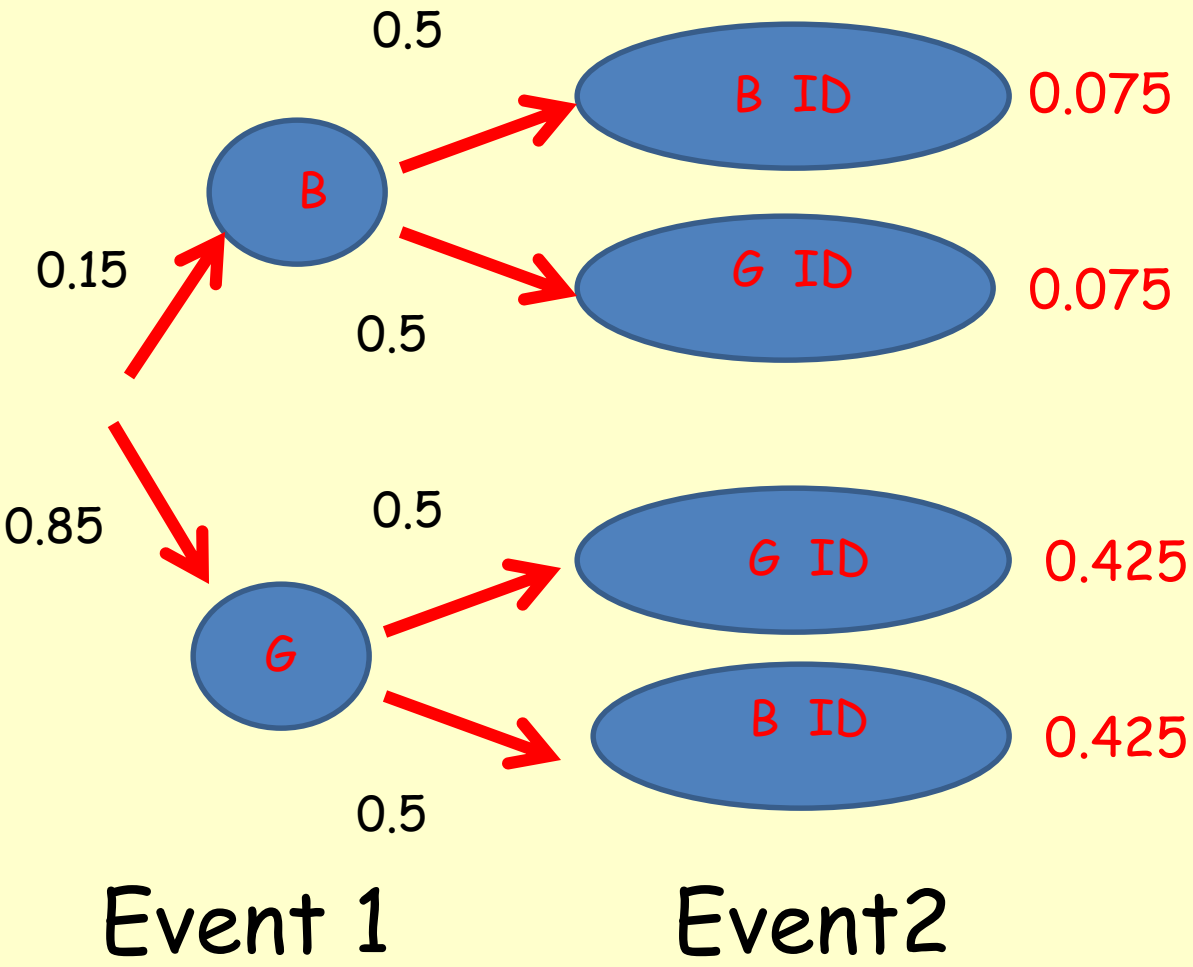
0.12 / (0.12 + 0.17) =
0.12 / 0.29 = 41.3%

Probability Trees

Question: What is probability that car was blue if identified as a blue car?

Car colors:
- 85% green
- 15% blue

ID:
- 50% correct
- 50% incorrect



Solution:

$$\frac{\text{Blue ID \& Blue Car}}{\text{Blue ID}}$$

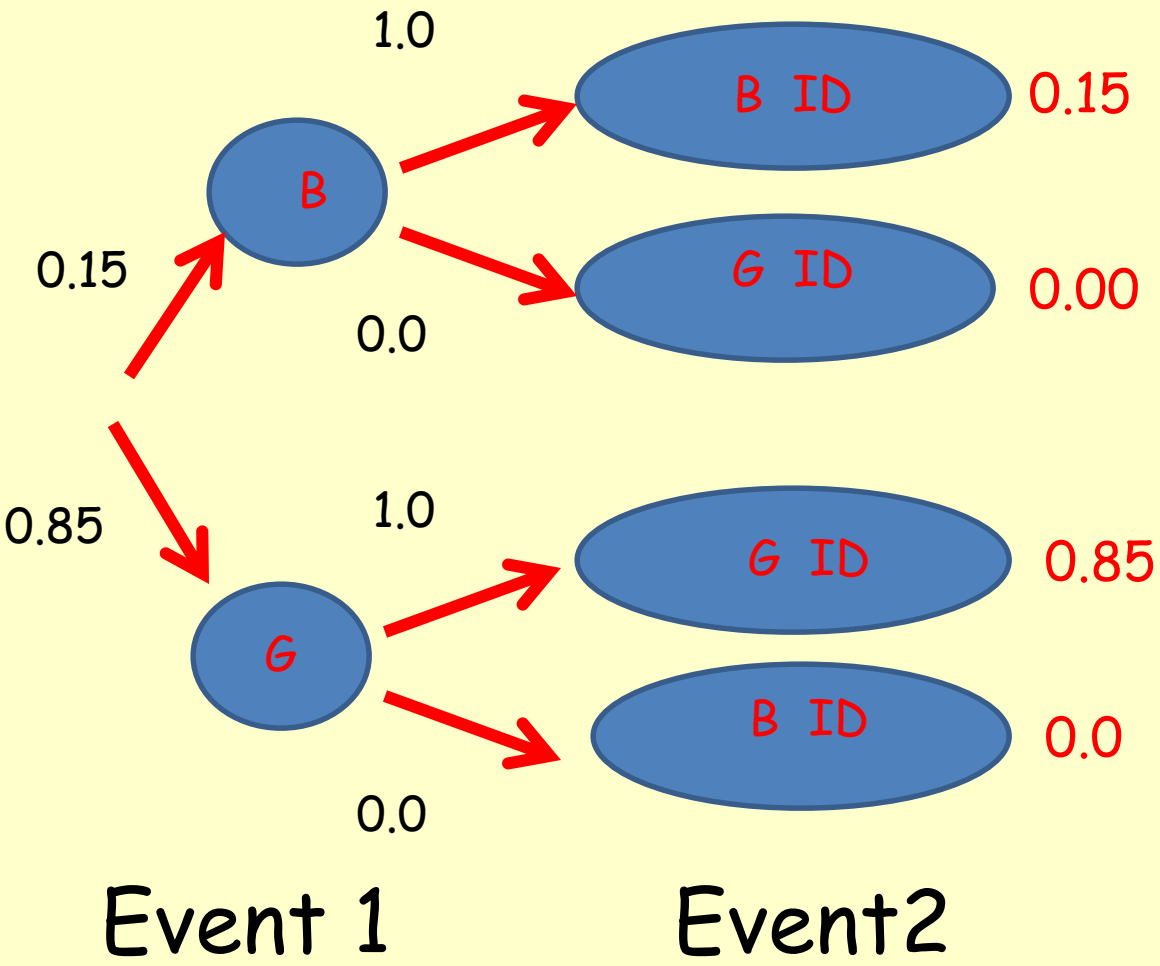
= 0.075 / (0.075 + 0.425)
= 0.075 / 0.5 = 15.0 %

Probability Trees

Question: What is probability that car was blue if identified as a blue car?

Car colors:
85% green
15% blue

ID:
100% correct
0% incorrect

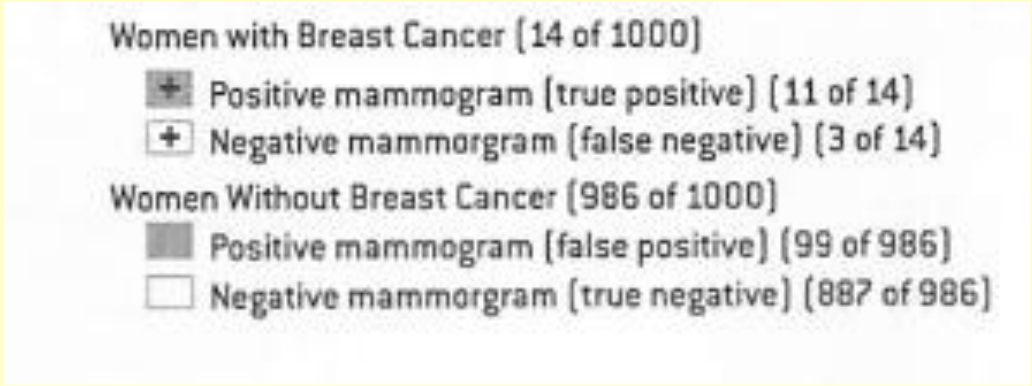


Solution:

$$\frac{\text{Blue ID \& Blue Car}}{\text{Blue ID}}$$

= 0.15 / 0.15
= 0.15 / 0.15 = 100 %

Example - Applying Bayes Theorem



X = Prior Probability
P (developing cancer)
= 14 / 1000 = 0.0140

Y = P (developing cancer, if test was positive) = 11 / 14 = 0.7857
Z = P (not developing cancer, if test was positive) = 99 / 986 = 0.1004

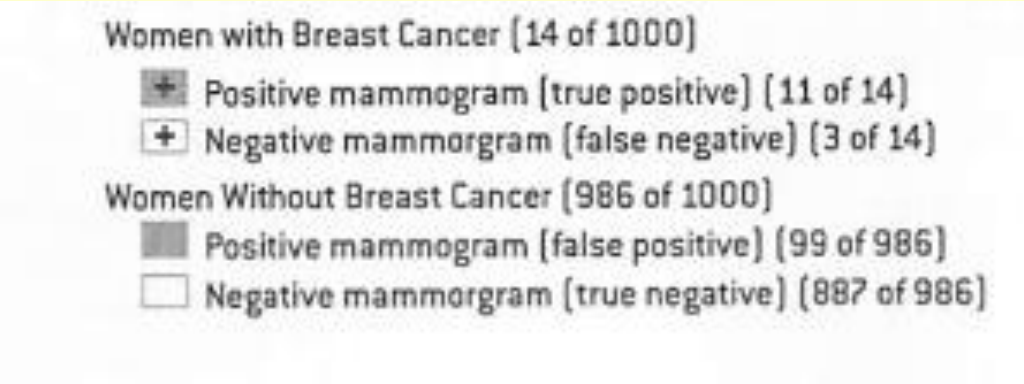
Posterior = (X * Y) / [(X * Y) + (Z * (1 – X)]

- Numerator: 0.0140 * 0.7857 = 0.011

- Denominator: (0.0140 * 0.7857) + (0.1004 * (1 - 0.9860))
  (0.0110) + (0.0990)

- Posterior = 0.0110 / 0.1100 = 0.1000

Example - Applying Bayes Theorem



X = Prior Probability
P (not developing cancer)
= 986 / 1000 = 0.9860

Y = P (not developing cancer, if test was positive) = 99 / 986 = 0.1004
Z = P (developing cancer, if test was positive) = 11 / 14 = 0.7857

Posterior = (X * Y) / [(X * Y) + (Z * (1 – X)]

- Numerator: 0.9860 * 0.1004 = 0.0989

- Denominator: (0.9860 * 0.1004) + (0.7857 * (1 - 0.986))
  (0.0990) + (0.0110)

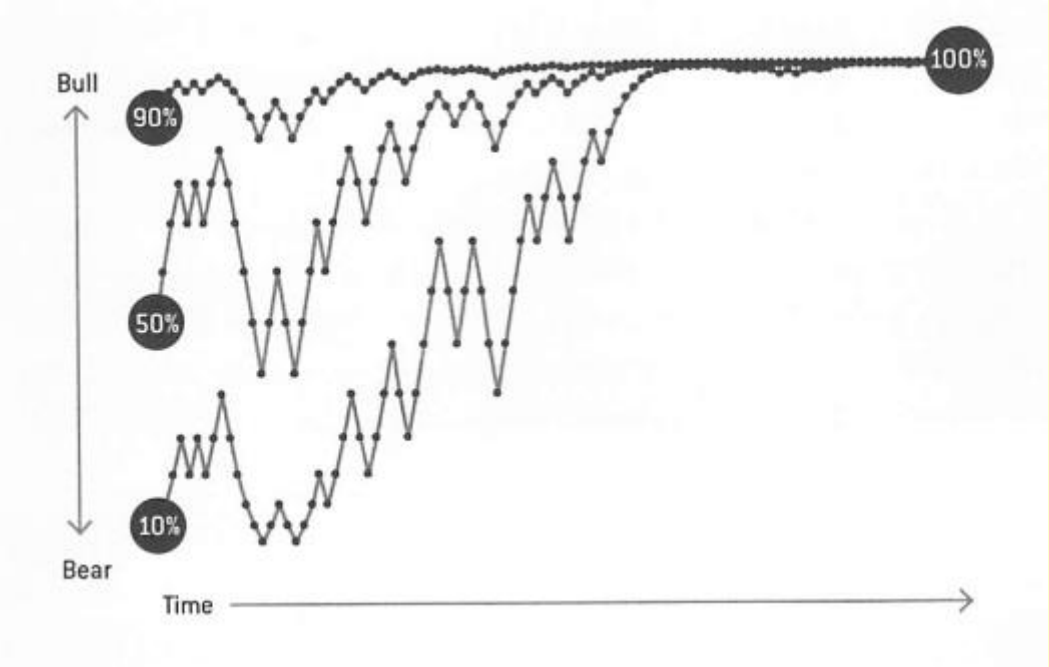
- Posterior = 0.0990 / 0.1099 = 0.9000

Bayes Theorem – Modifying the Prior

Because the Prior Probability influences the Posterior Probability, the selection of a Prior should be objective.

Otherwise, it will have an undue influence in the analysis.

However, given enough observations… varying priors should converge towards the actual probabilities of the phenomenon being tested (the truth).



Bayes VS. Frequentist Statistics

To oversimplify:

- "Bayesian probability" is the interpretation of probability as the degree of belief in a hypothesis
- Yields probabilities to hypotheses, which vary as additional observations are collected

- "Frequentist probability" is the interpretation of probability as the frequency of a particular outcome in a large number of experimental trials
- Yields the acceptance / rejection of the null hypothesis

“Is this needed for a Bayesian analysis?”

What Do I Expect You to Know

Concepts of the sample space and conditional probability

Be ready to perform calculations like this in quiz #2

Homework #1:

Given two 8-sided dice, with the faces numbered 1 to 8, what is probability of:

1) a total count of 7 by summing the values from the two top surfaces in a single toss?
2) a total count of 7 or 10 by summing the values from the two top surfaces – in a single toss?
3) a total count of 7 or 10 – at least once – by summing the values from the two top surfaces – in 4 tosses of both dice?
4) a total count of 7 or 10 – at least twice – by summing the values from the two top surfaces – in 5 tosses of both dice?