Probability

A quick, easy and painless review

Auburn University INSY 7970
By: Patrick Almas
Graduate Assistant

What is Probability ?

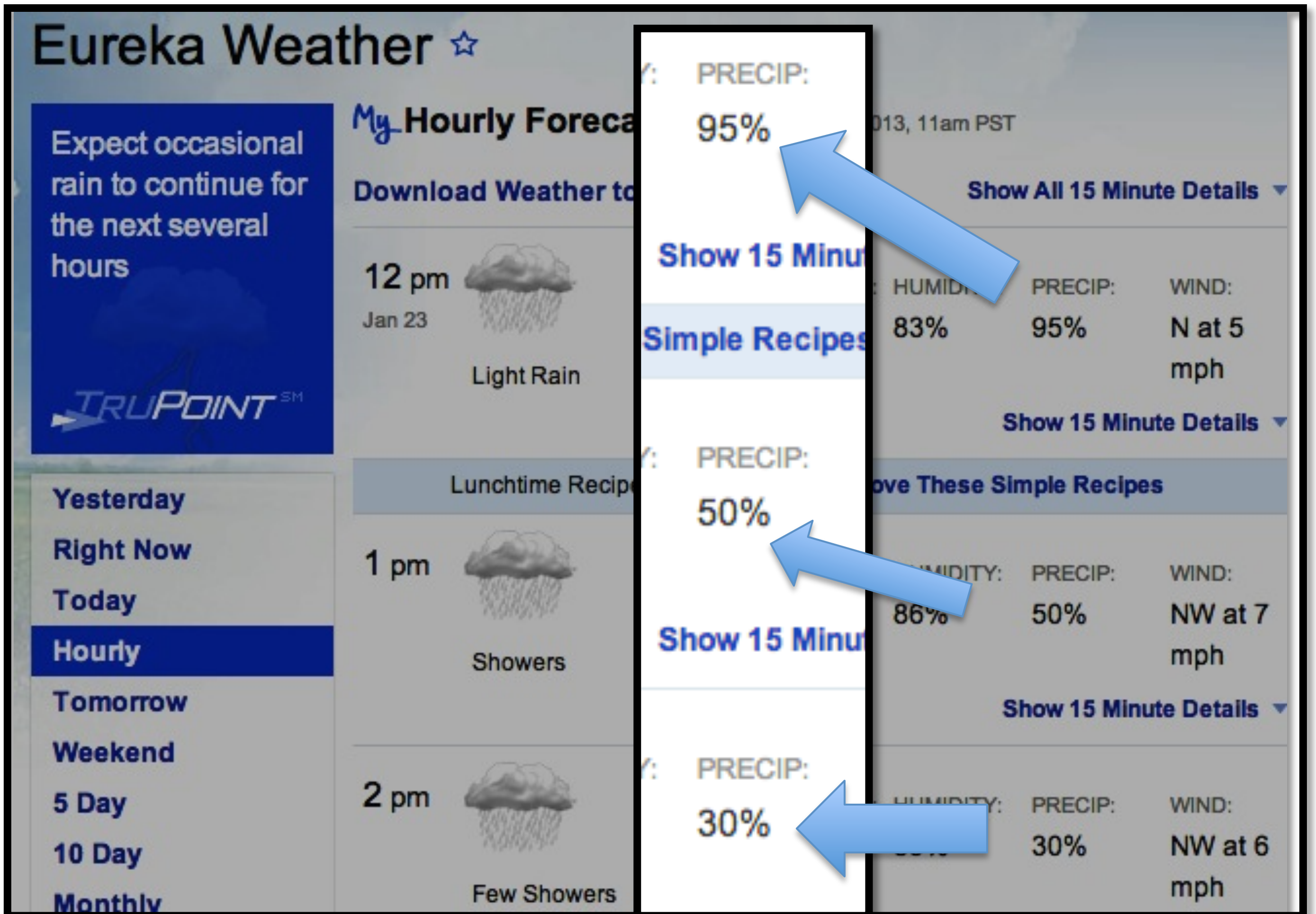
- Probability is a branch of mathematics that deals with calculating the likelihood of a given event's occurrence, which is expressed as a number between 1 and 0.

- An event with a probability of 1 can be considered a certainty.

- How is this used in our everyday lives?

Weather forecasts !

Weather forecasts !

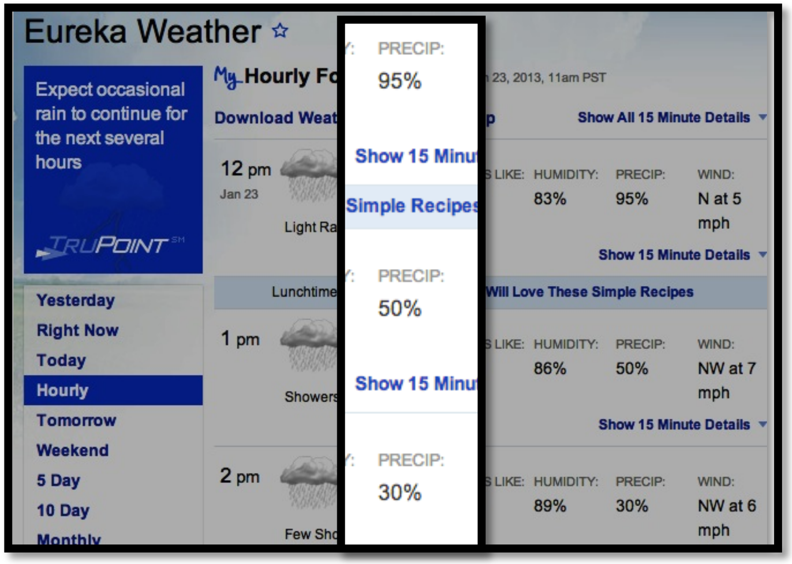


What does this mean?

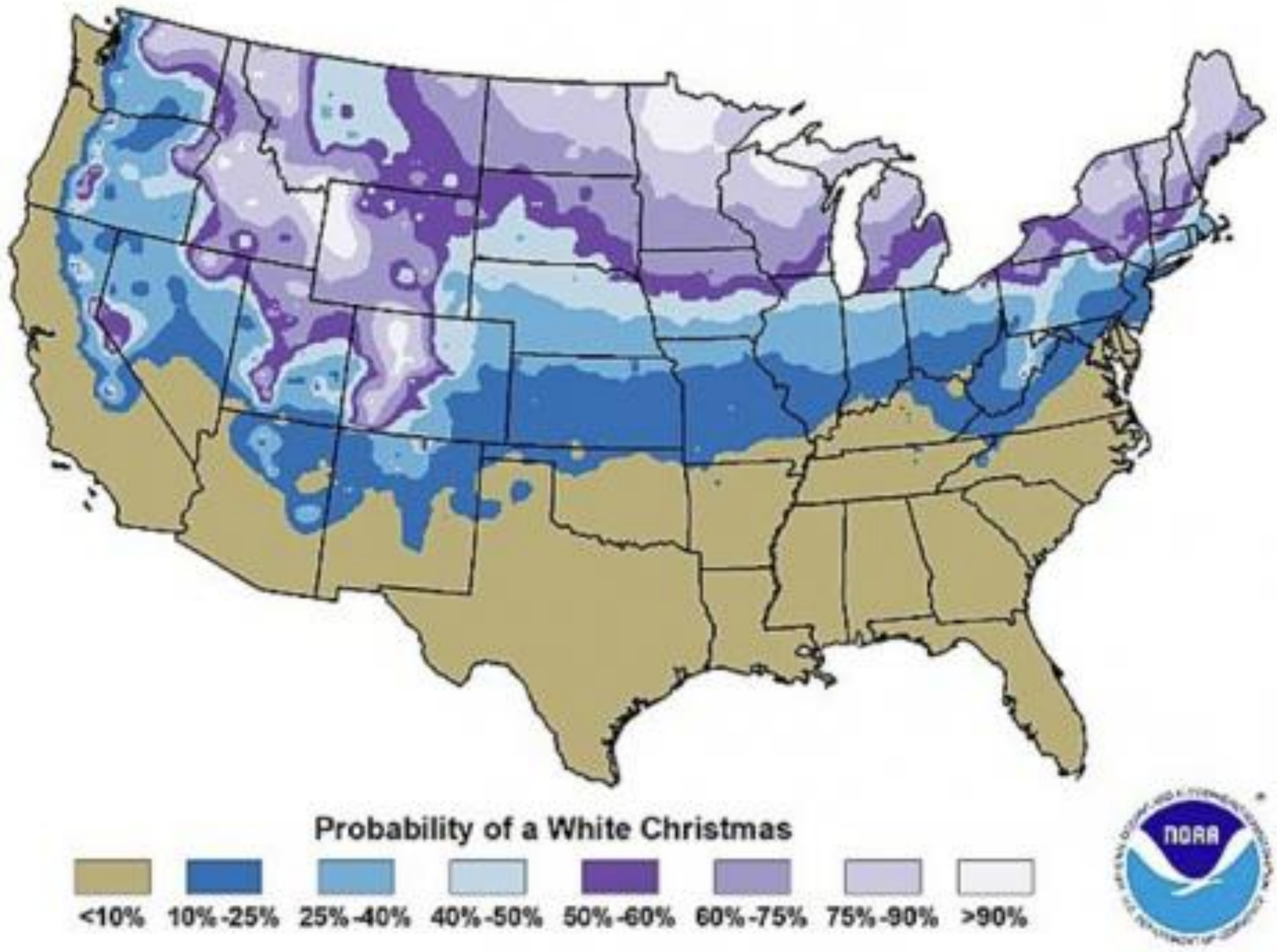
It can mean one of two things…

1. How strongly you believe in the data
   - I am 95% sure it will rain

2. If you had all the data and results
   - Out of 100 identical days, in 95 of those, it will rain



Probability of a White Christmas



Lets see how smart we are!

- If we flip a coin once, what is the probability of the coin landing on heads?

Lets see how smart we are!

- If we flip a coin once, what is the probability of the coin landing on heads?

50%!

So easy a cave man can do it!

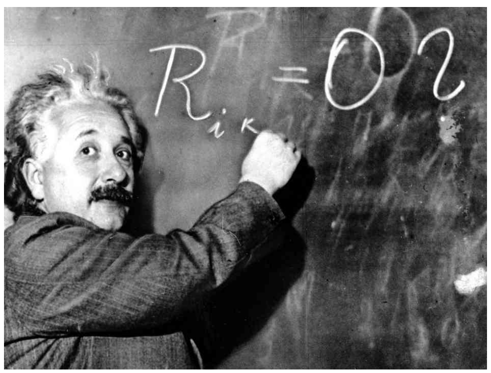
Lets see how smart we are!

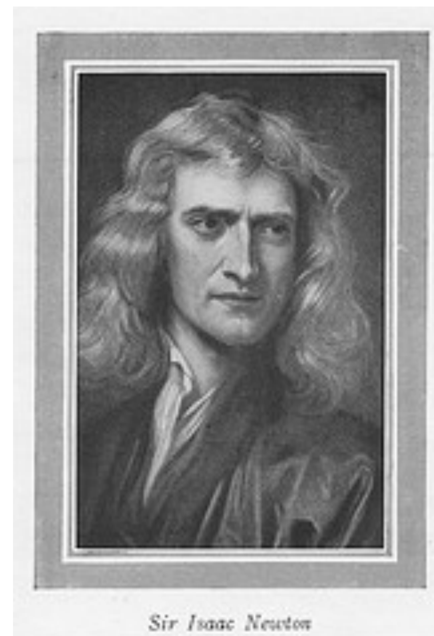
- If we flip a coin twice, what is the probability of the coin landing on heads twice?

× 2

Lets see how smart we are!

- If we flip a coin twice, what is the probability of the coin landing on heads twice?





25%!

Unbelievable!

Lets switch gears (just a bit)

- What are odds?
  - Think Sports

2010 Ironman Hawaii Men's Odds

Written by: Timothy Carlson
Date: Tue Oct 05 2010

Despite the polished excellence of Craig Alexander, a horde of fearless contenders believe they can swim bike and run to glory. The contenders include Rasmus Henning, Andreas Raelert, Chris Lieto, Chris McCormack, Eneko Llanos, Andy Potts, Marino Vanhoenacker, Dirk Bockel, Terenzo Bozzone, Timo Bracht...

You get the picture. It's a wide open shootout in the Kona Corral

Craig Alexander, 37 (AUS) 5-2

Once you establish yourself as one of the greatest clutch performers and best-prepared professionals in Ironman history, one of only eight men to win this race two or more times in its 33-year history, it's a natural inclination is to look for chinks in the armor. Especially when the man has dead aim on joining Dave Scott (6 IMH wins) Mark Allen (6), and Peter Reid (3) as the only men to win 3 or more Ironman Hawaii titles -- and leave behind certified 2-win greats Scott Tinley, Luc Van Lierde, Tim DeBoom and Normann Stadler.

Lets switch gears (just a bit)

- What are odds?

odds plural of **odds** (Noun)

Noun

1. The ratio between the amounts staked by the parties to a bet, based on the expected probability either way: "odds of 8-1".
2. The chances or likelihood of something happening or being the case.

Synonyms

difference - inequality

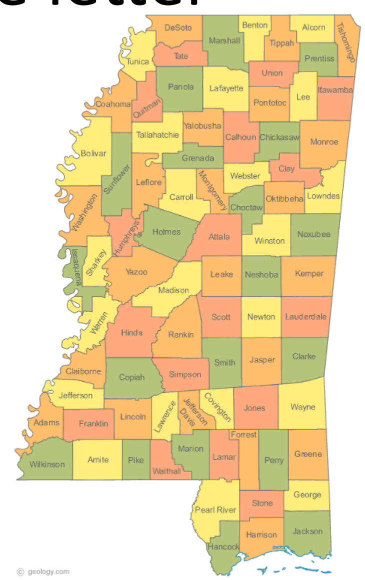
Equations

Definition of Probability	**probability of an event** or $P(\text{event}) = \dfrac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$
Definition of Odds	**odds** of an event = number of ways the event can occur : number of ways the event cannot occur = successes : failures

Quick Practice

Definition of Probability	**probability of an event** or P(event) $= \dfrac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$
Definition of Odds	**odds** of an event $=$ number of ways the event can occur : number of ways the event cannot occur $=$ successes : failures

1)Find the probability of randomly choosing the letter ‘s’ in the word “Mississippi.”



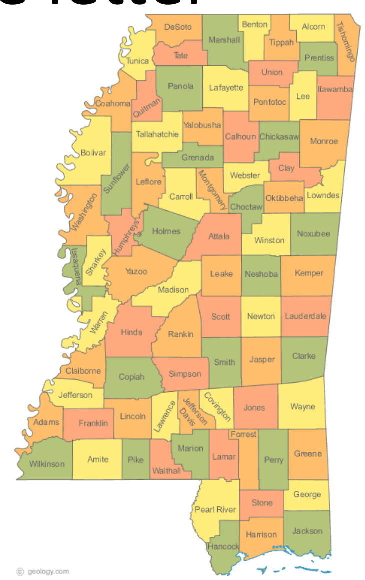
2) Find the odds of randomly selecting the letter ‘p’ in the word “Mississippi.”

3)How do you spell “Mississippi” with one ‘i’ ?

Quick Practice

Definition of Probability	**probability of an event** or $P(\text{event}) = \dfrac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$
Definition of Odds	**odds** of an event = number of ways the event can occur : number of ways the event cannot occur = successes : failures

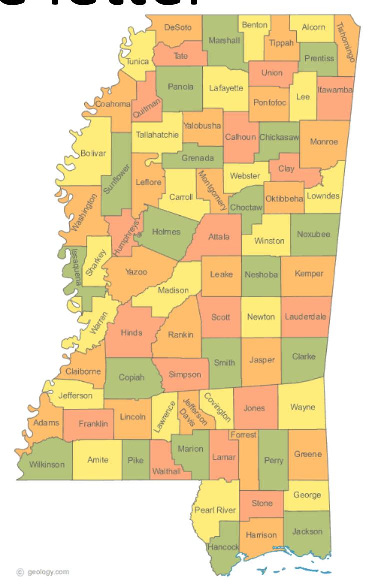
1)Find the probability of randomly choosing the letter 's' in the word "Mississippi."



Quick Practice

Definition of Probability	**probability of an event** or P(event) $= \dfrac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$
Definition of Odds	**odds** of an event = number of ways the event can occur : number of ways the event cannot occur = successes : failures

1)Find the probability of randomly choosing the letter ‘s’ in the word “Mississippi.”



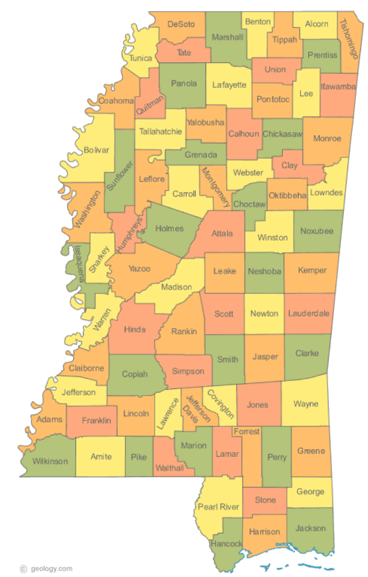
There are 4 s’s and 11 letter in all.

P(choosing s)= (4/11)=0.363636 or 36.36%

Quick Practice

Definition of Probability	**probability of an event** or $P(\text{event}) = \dfrac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$
Definition of Odds	**odds** of an event = number of ways the event can occur : number of ways the event cannot occur = successes : failures

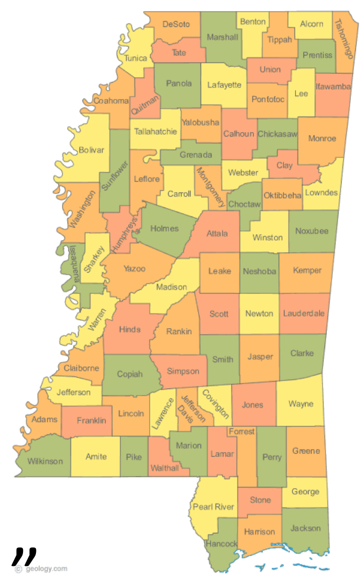
2) Find the odds of randomly selecting the letter ‘p’ in the word “Mississippi.”



Quick Practice

Definition of Probability	**probability of an event** or $P(\text{event}) = \frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$
Definition of Odds	**odds** of an event = number of ways the event can occur : number of ways the event cannot occur = successes : failures

2) Find the odds of randomly selecting the letter ‘p’ in the word “Mississippi.”



There are 11 letters in the word. Two letters are p’s and 11-2 or 9 letters are not p’s.

Odds of selecting a ‘p’=

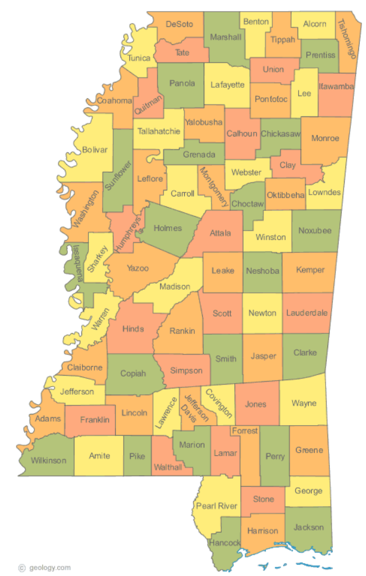
Number of p’s : number not p’s

So the odds are 2:9 and it is read as “2 to 9.”

<u>Quick Practice</u>

Definition of Probability	**probability of an event** or P(event) $= \dfrac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$
Definition of Odds	**odds** of an event = number of ways the event can occur : number of ways the event cannot occur = successes : failures

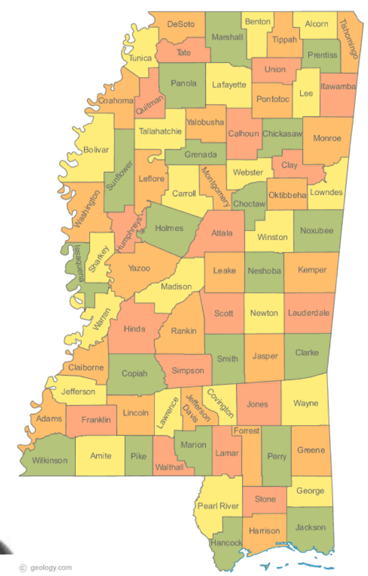
3)How do you spell “Mississippi” with one ‘i’ ?



Quick Practice

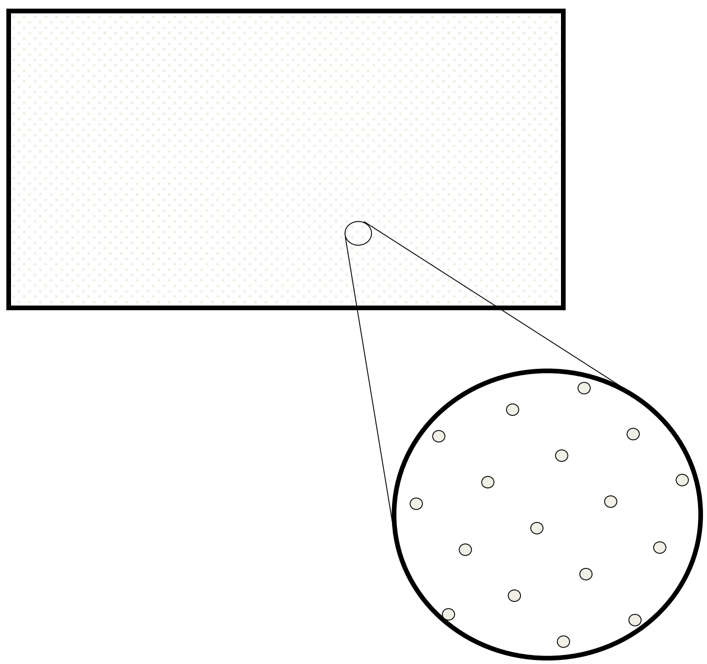
Definition of Probability	**probability of an event** or $P(\text{event}) = \dfrac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$
Definition of Odds	**odds** of an event = number of ways the event can occur : number of ways the event cannot occur = successes : failures

3)How do you spell “Mississippi” with one ‘i’ ?



Sample Space

Definition
The set, S, of all possible outcomes of a particular experiment is called the ***sample space*** for the experiment

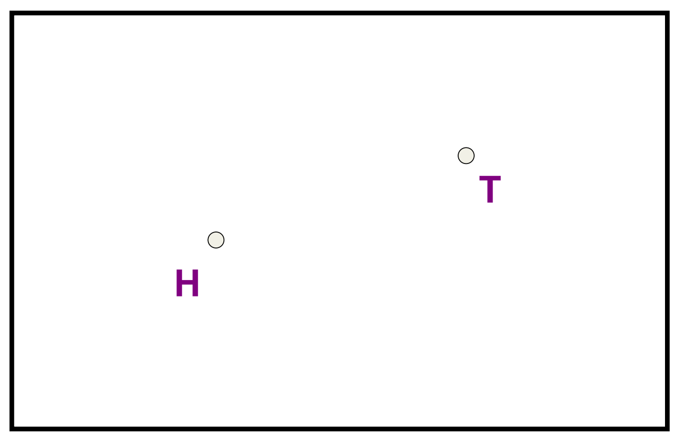


The elements of the sample space are called ***outcomes***.

Sample Spaces

Sample space of a coin flip:

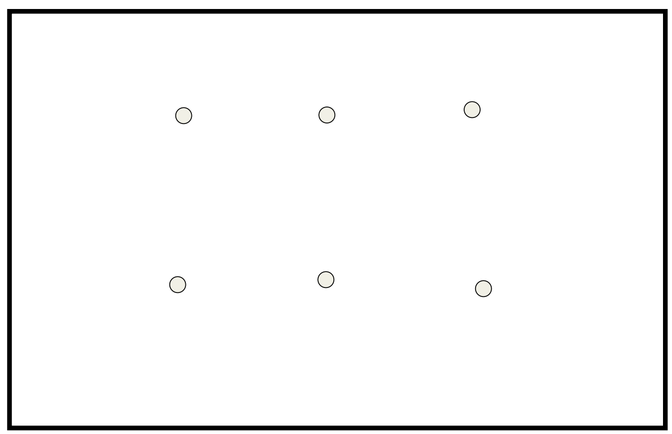
S = {H, T}



Sample Spaces

Sample space of a die roll:

$S = \{1, 2, 3, 4, 5, 6\}$

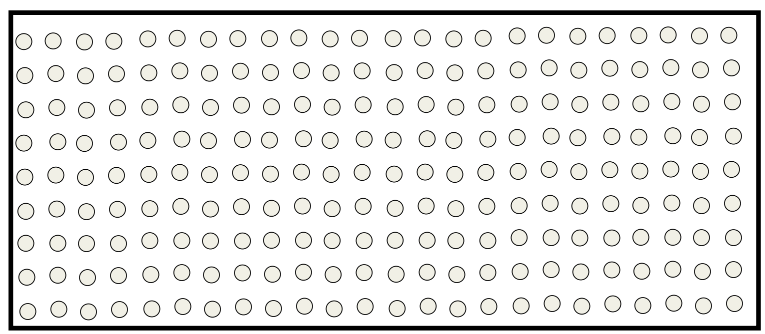


Sample Spaces

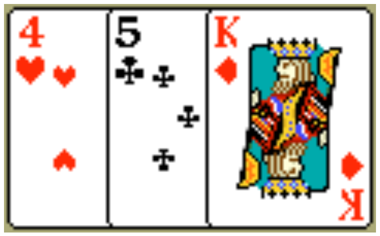
Sample space of *three* die rolls?

S = {111,112,113,…,

…,664,665,666}



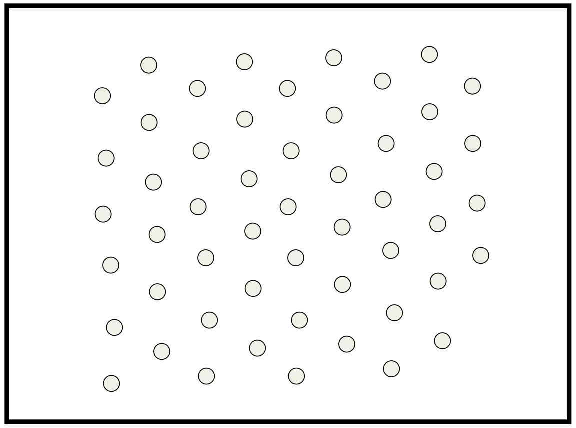
Sample Spaces



Sample space of a single draw from a deck of cards:

S={As,Ac,Ah,Ad,2s,2c,2h,…

…,Ks,Kc,Kd,Kh}

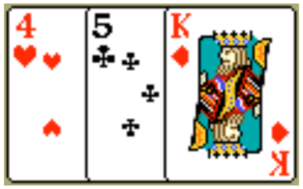


So Far…

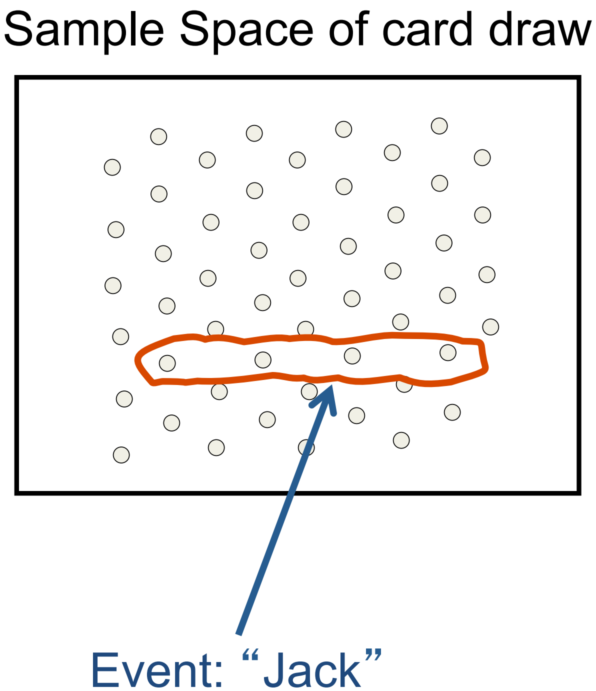
Definition	**Example**
The **Sample Space** is the set of all outcomes	{As,Ac,Ah,Ad,2s,2c,2h,… …,Ks,Kc,Kd,Kh}
An ***outcome*** is an element of the sample space.	2c

Events

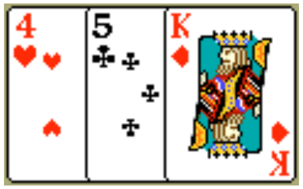
Definition
An ***event*** is any subset of S (including S itself)



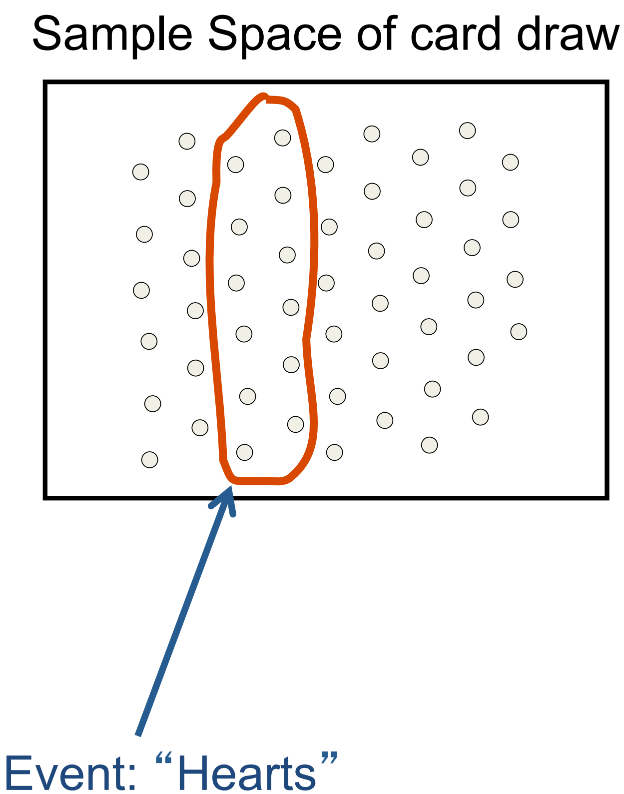
Events



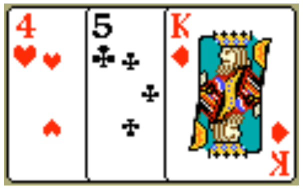
- The ***Sample Space*** is the set of all outcomes.
- An ***Outcome*** is a possible element.
- An ***Event*** is a set of outcomes



Events

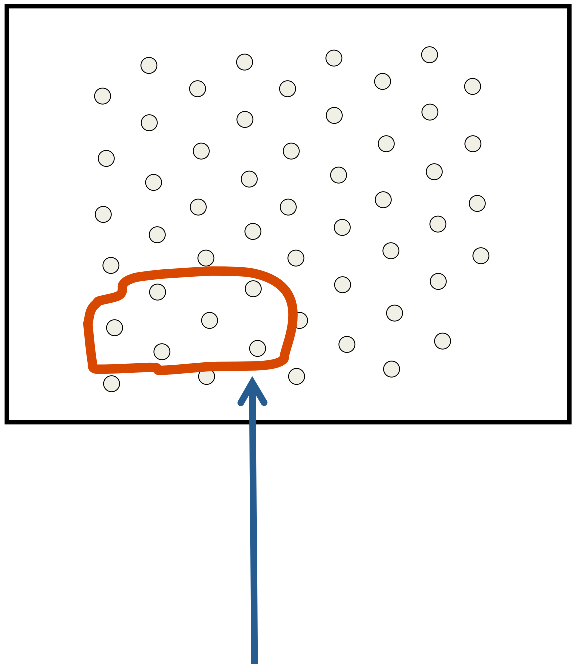


- The ***Sample Space*** is the set of all outcomes.
- An ***Outcome*** is a possible element.
- An ***Event*** is a set of outcomes



Events

Sample Space of card draw



Event: “Red and Face”

- The ***Sample Space*** is the set of all outcomes.
- An ***Outcome*** is a possible element.
- An ***Event*** is a set of outcomes

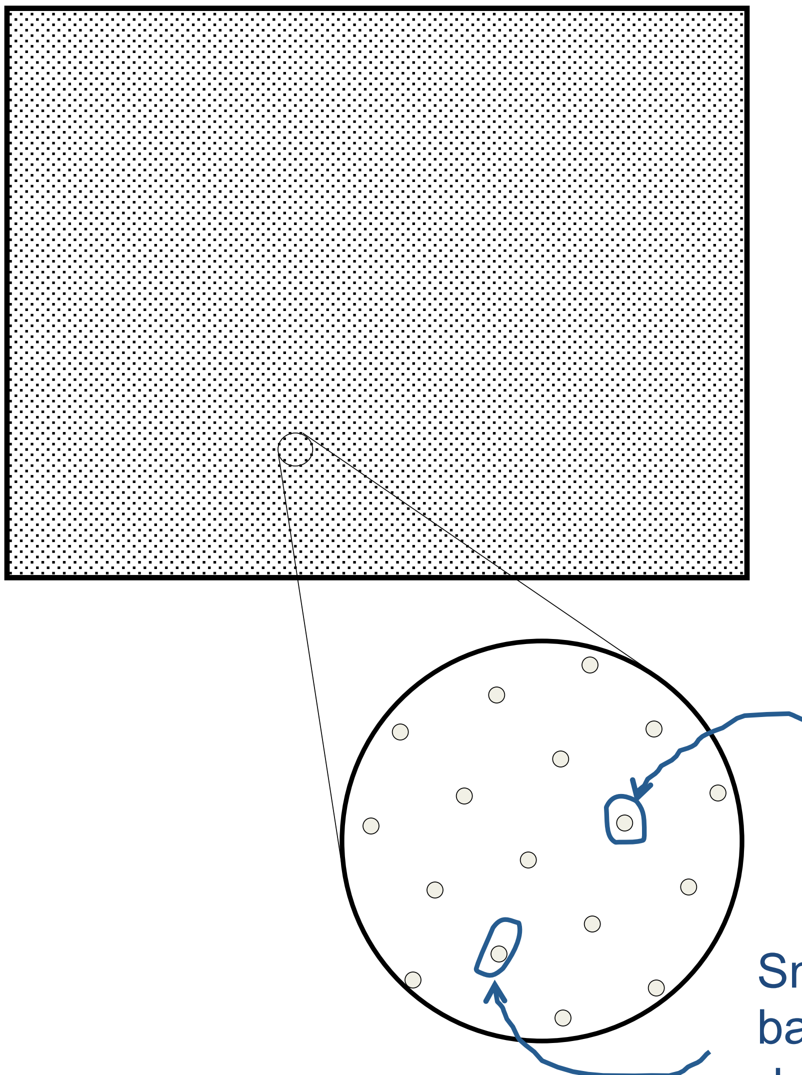
Discrete Random Variables

- A is a Boolean-value random variable if A denotes an event, and there is some degree of uncertainty as to weather A occurs.
- Examples
  - A = Miss Auburn will be a sorority girl
  - A= You will make an "A" in this class
- A ***discrete random variable*** is one which may take on only a countable number of distinct values such as 0,1,2,3,4,........
- What is the P(A)?

Definitions

Definition	**Example**
The ***sample space*** is the set of all possible outcomes.	{As,Ac,Ah,Ad,2s,2c,2h,… …,Ks,Kc,Kd,Kh}
An ***outcome*** is a single point in the sample space.	4c
An ***event*** is a set of one or more outcomes	Card is "Red" and "Face"
P(A) maps event A to a real number	P(Red) = 0.50 P(Black) = 0.50

Everyday Example



Assume you are a doctor.

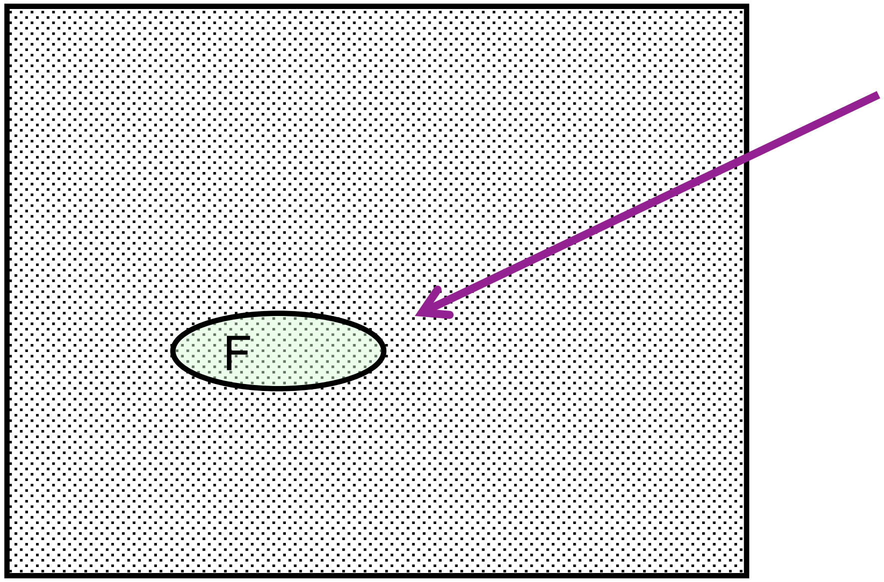
This is the sample space of "patients you might see on any given day".

Outcomes

Non-smoker, female, diabetic, headache, good insurance, etc…

Smoker, male, herniated disk, back pain, mildly schizophrenic, delinquent medical bills, etc…

Everyday Example



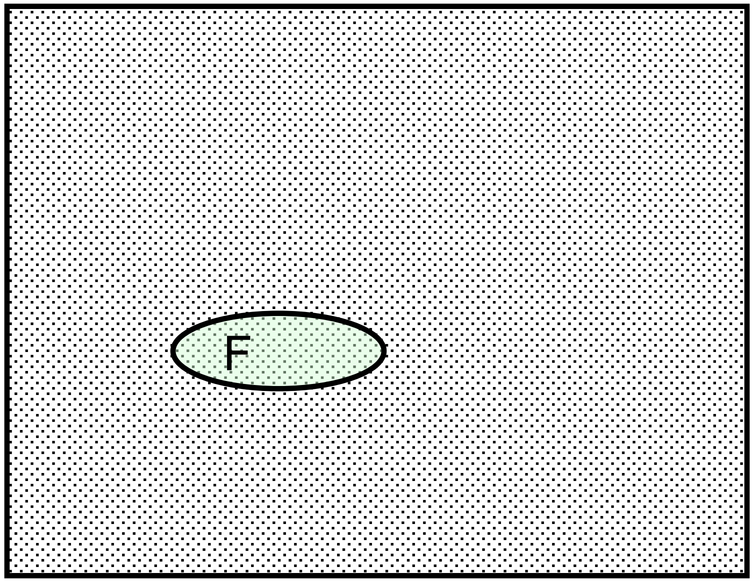
Event: Patient has Flu

Size of set "F":
2 jillion
(Exactly 2 jillion of the points in the sample space have flu.)

Size of "patient space":
100 jillion

$$P_{patientSpace}(F) = \frac{2\ \text{jillion}}{100\ \text{jillion}} = 0.02$$

Everyday Example

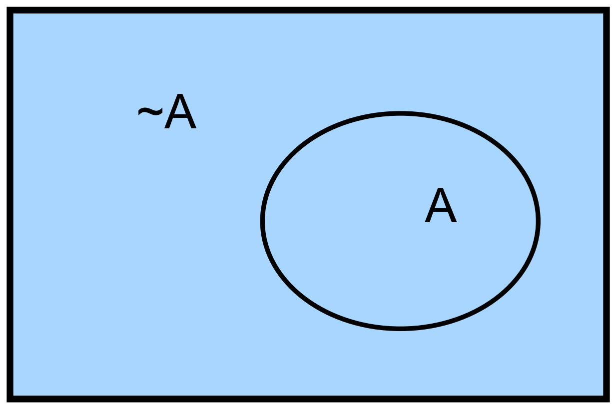


$$P_{patientSpace}(F) = \frac{2\text{ jillion}}{100\text{ jillion}} = 0.02$$

From now on, the subscript on P() will be omitted…

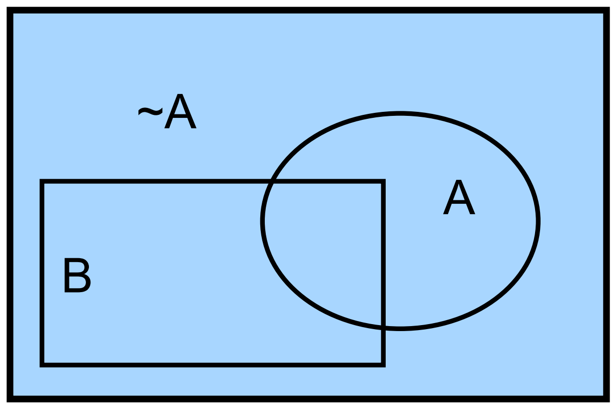
Elementary Probability in Pictures

$P(\sim A) + P(A) = 1$



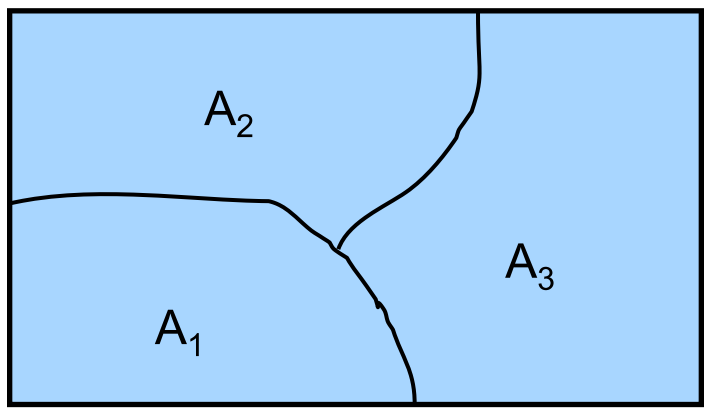
Elementary Probability in Pictures

$P(B) = P(B, A) + P(B, \sim A)$



Elementary Probability in Pictures

$$\sum_{j=1}^{k} P(A_j) = 1$$



Birthday Game

- What is the probability that two people in this classroom share the same birthday?
  - (given: 45 students in the class)

Birthday Game

- We really need to think… What is the probability that no two people will share a birthday?

  - ✓ P(event happens) + P(event doesn't happen) = 1
  - ✓ P(two people share birthday) + P(no two people share birthday) = 1
  - ✓ P(two people share birthday) = 1 - P(no two people share birthday).

Birthday Game

- The formula for k people not sharing the same birthday is:

$$\frac{365!}{(365-k)! \times 365^k}$$

  - With two people it would be: (364/365)
  - With three people it would be: (364/365) * (363/365)

- Therefore, 1 minus this probability tells you the chances of two people sharing the same birthday!

Birthday Game

- So, what is the probability that two people in this classroom share the same birthday?
    - (given: 45 students in the class)

Birthday Game

- So, what is the probability that two people in this classroom share the same birthday?
    - (given: 45 students in the class)

$$\frac{365!}{(365-k)! \times 365^k}$$

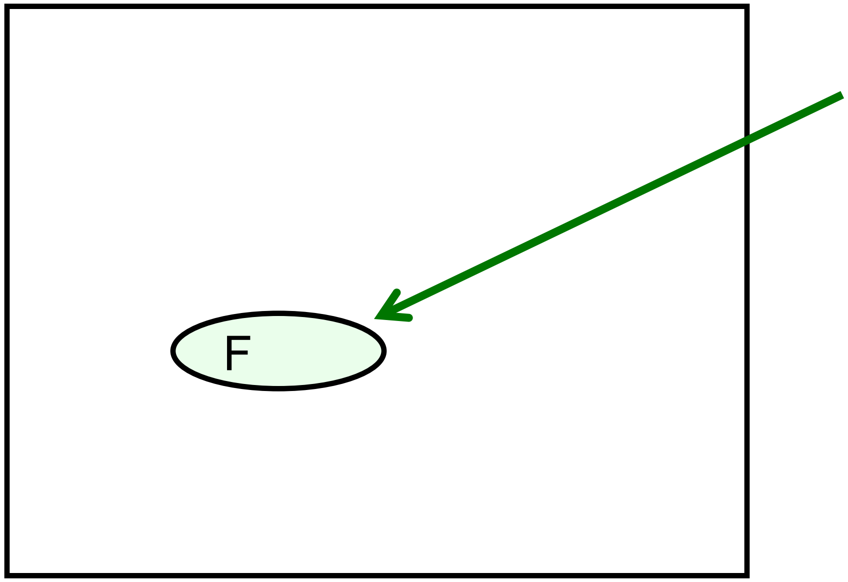
- (1-0.0590241) * 100% = 94.098%

Conditional Probability

Assume once more that you are a doctor.

Again, this is the sample space of “patients you might see on any given day”.

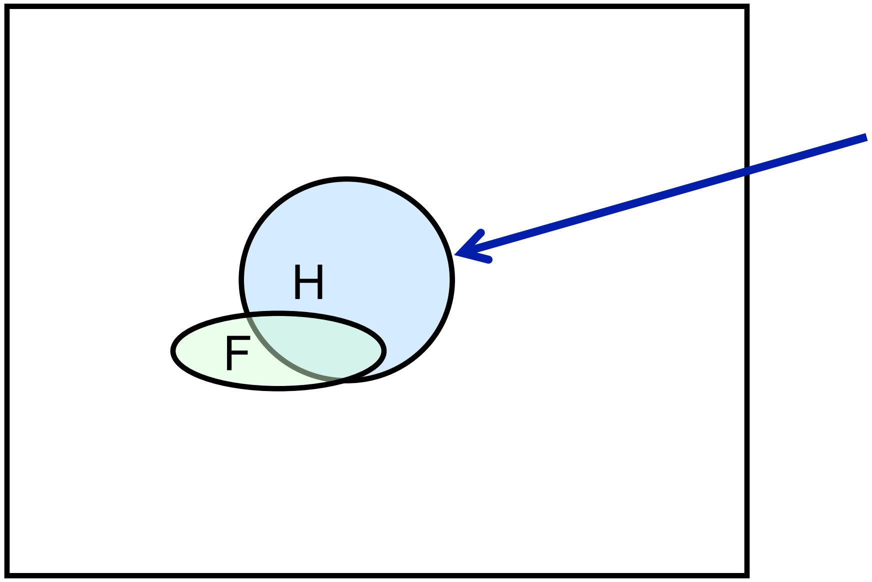
Conditional Probability



Event: Flu

P(F) = 0.02

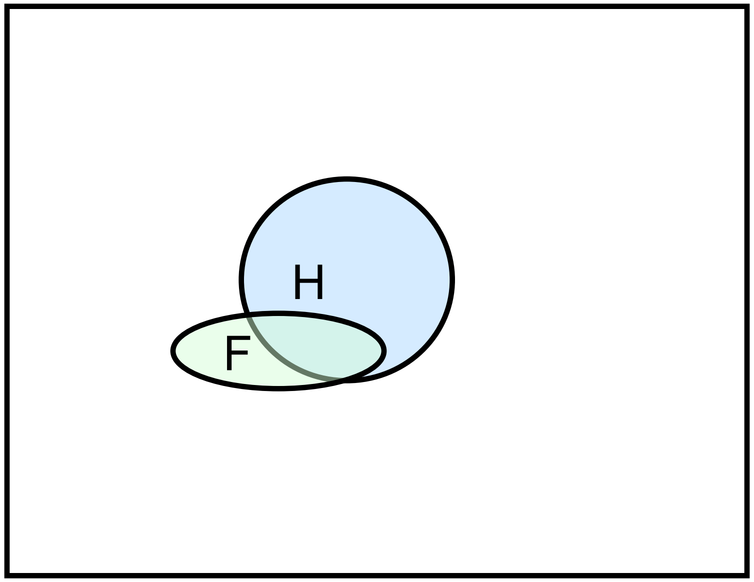
Conditional Probability



Event: Headache

P(H) = 0.10

Conditional Probability



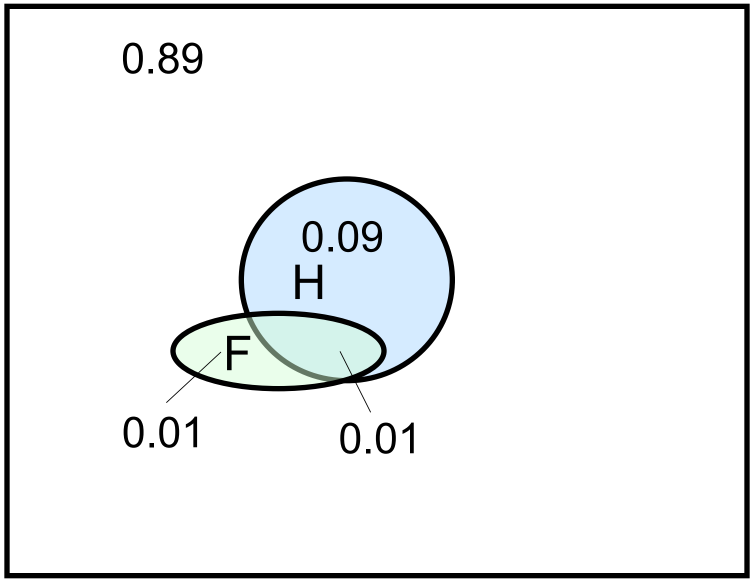
P(Flu) = 0.02
P(Headache) = 0.10

…we still need to specify the *interaction* between flu and headache…

Define

P(H|F) = Fraction of F's outcomes which are also in H

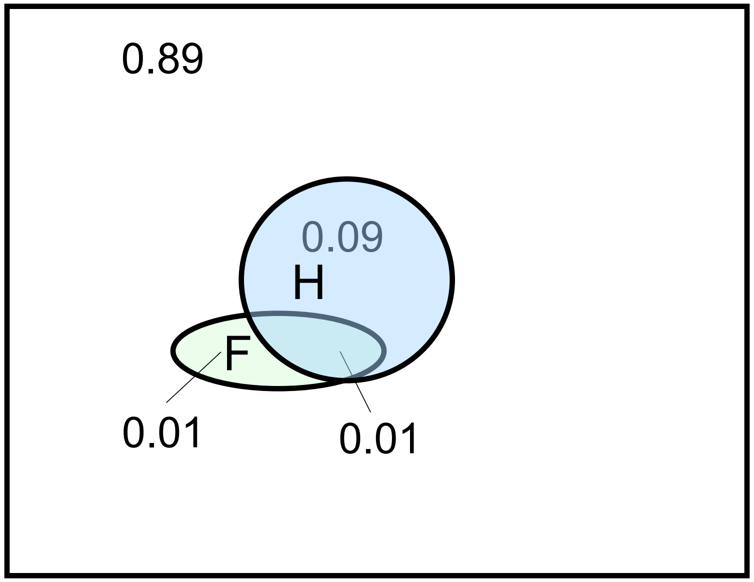
Conditional Probability



H = “headache”
F = “flu”

P(Flu) = 0.02
P(Headache) = 0.10
P(H|F) = 0.50

Conditional Probability



H = "headache"
F = "flu"

P(H|F) = Fraction of flu worlds in which patient has a headache

= #worlds with flu and headache

#worlds with flu

= Size of "H and F" region

Size of "F" region

= P(H, F)

P(F)

Conditional Probability

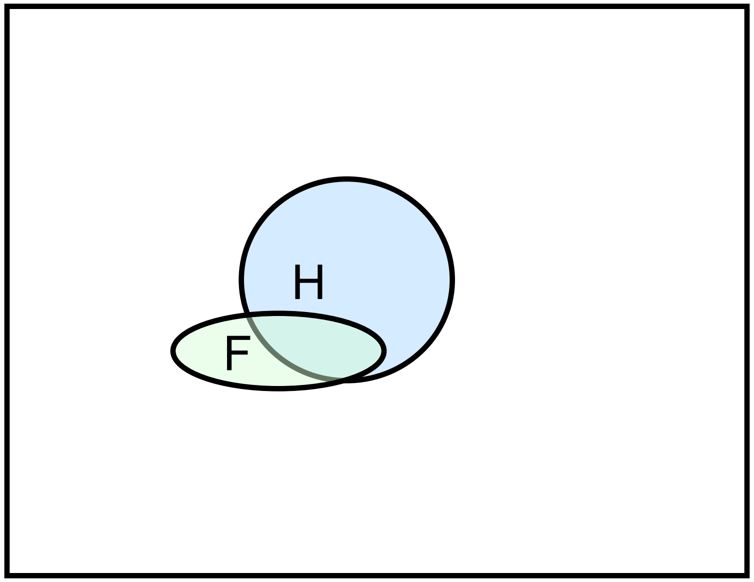
Definition.
If A and B are events in S, and P(B) > 0, then the ***conditional probability*** of A given B, written P(A|B), is

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

The Chain Rule
A simple rearrangement of the above equation yields

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

Probabilistic Inference



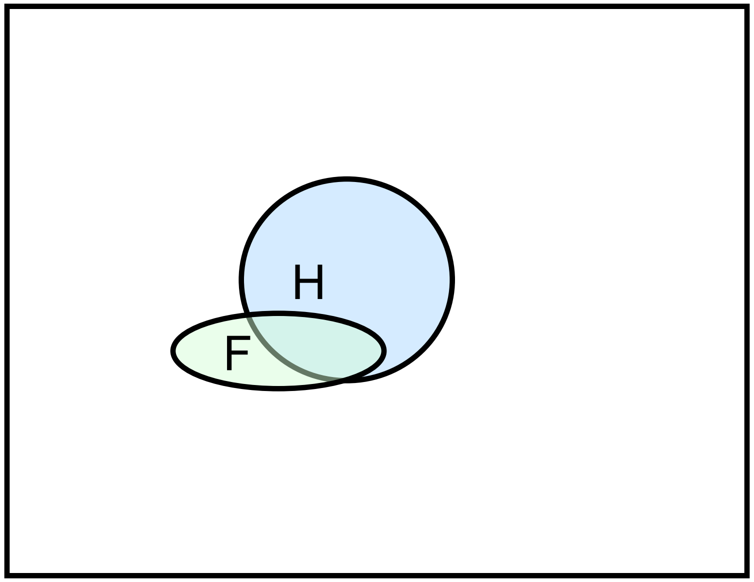
H = "Have a headache"
F = "Coming down with Flu"

P(H) = 0.10
P(F) = 0.02
P(H|F) = 0.50

One day you wake up with a headache. You think: "Oh,NO! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"

Is this reasoning good?   NO!!

Probabilistic Inference



H = "Have a headache"
F = "Coming down with Flu"

P(H) = 0.10
P(F) = 0.02
P(H|F) = 0.50

What is the probability of having the flu given that you woke up with a headache?

$$P(F \mid H) = \frac{P(F,H)}{P(H)} = \frac{P(H \mid F)P(F)}{P(H)} = \frac{(0.50)(0.02)}{0.1} = 0.10$$

Look what we just did…

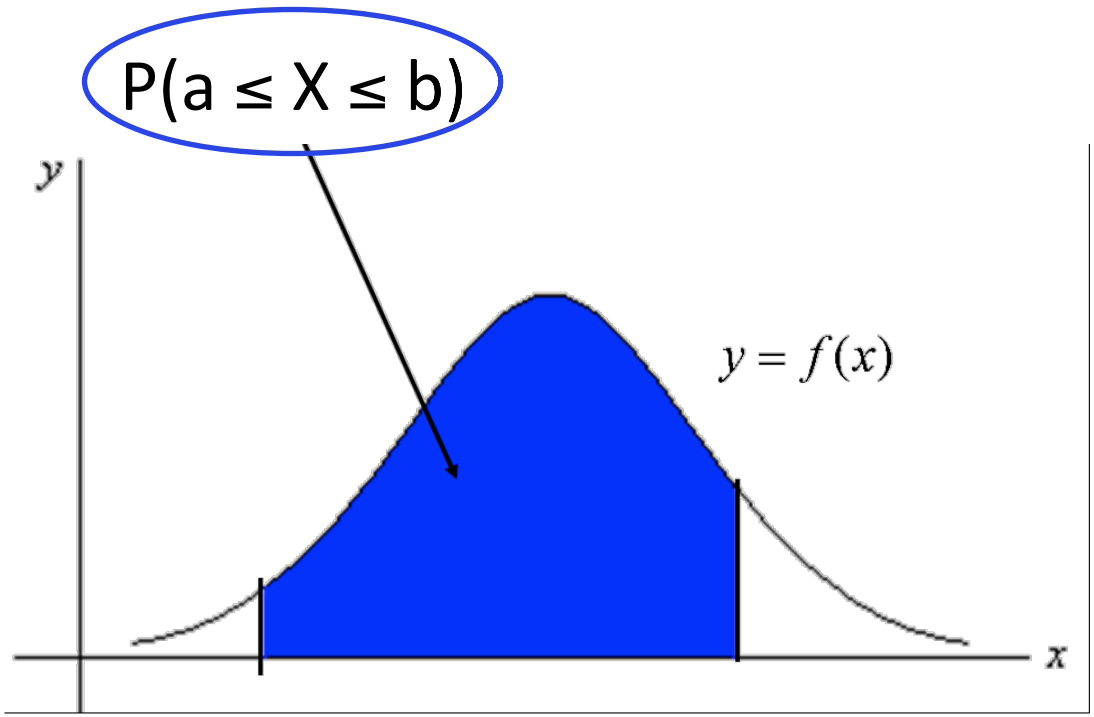
$$P(B|A) = \frac{P(A,B)}{P(A)} = \frac{P(A|B)\,P(B)}{P(A)}$$

This is Bayes Rule

Bayes, Thomas

Probability Density Functions

- The area under a pdf curve for an interval is the probability that an event mapped into that interval will occur.



$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Cumulative Distribution Functions

- $P(X \leq a)$

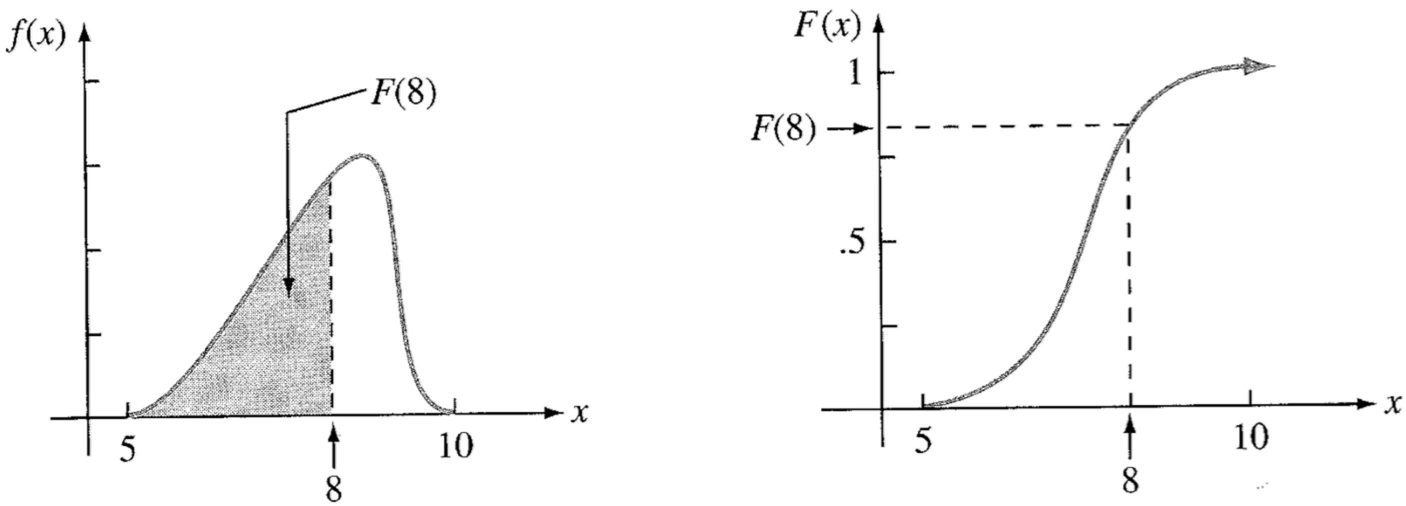


Figure 4.5 A pdf and associated cdf

Properties of pdfs

- Mean

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- E(h(x))

$$\mu = E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) dx$$

- Variance

$$\sigma^2 = Var(X) = E[(X - \mu)^2] = E(X^2) - [E(X)]^2$$

Standard Distribution Functions

There are some standard distributions that are commonly used. These are determined from either from the experiment or from analysis of the data

Binomial Distribution

Experimental Conditions

1. Know the number of trials
2. Each trial can have only two outcomes.
3. The trials are independent.
4. The probability of success is constant.

Formula:

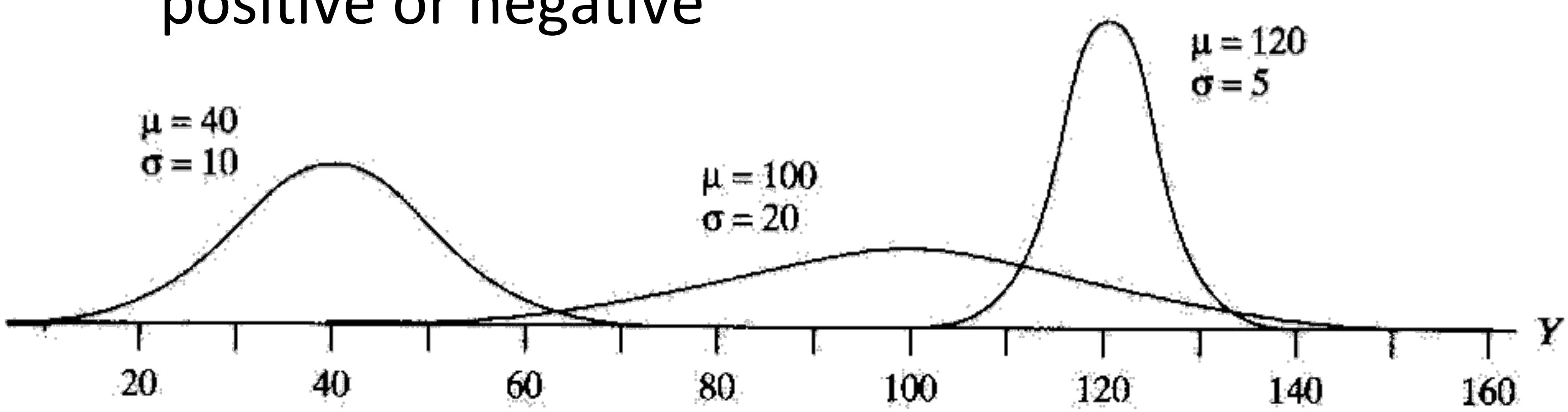
$$P(X = r) = \binom{n}{r} p^r (1 - p)^{n-r}$$

Other Discrete Distributions

- Hypergeometric
  - Like binomial but without replacement
- Poisson
  - Like a binomial but with very low probability of success
- Negative Binomial
  - Like binomial but want to know how my trials until a certain number of successes.

Normal Distribution Function

- Continuous
- This is the most commonly occurring distribution.
  - Systematic errors
  - A large number of small equally likely to be positive or negative



Now let's see probability in practice

orange

DATA MINING
FRUITFUL&FUN

If we have more time…

The Earthscope

- **The Earthscope is the world's largest science project.**

- Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data.

- It analyzes seismic slips in the San Andreas fault (think California and earth quakes), but also the plume of magma underneath Yellowstone.

- Find More info Here:

http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI

Type of Data

- Relational Data
  - (Tables/Transaction/Legacy Data)
- Text Data
  - (Web)
- Semi-structured Data
  - (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once

Additional Random Sample and Statistics

- *Population:* is used to refer to the set or universe of all entities under study.
- Looking at the entire population may not be feasible, or may be too expensive.
- Instead, we draw a random sample from the population, and compute appropriate *statistics* from the sample, that give estimates of the corresponding population parameters of interest.

5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	4.9	7.0	5.5	6.3	6.8	6.1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	5.0	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.7	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	6.9	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1

Statistic

- Let Si denote the random variable corresponding to data point xi , then a *statistic* ˆθ is a function ˆθ : (S1, S2, · · · , Sn) → R.

- If we use the value of a statistic to estimate a population parameter, this value is called a *point estimate of the parameter, and the statistic is called as an estimator of the* parameter.

Empirical Cumulative Distribution Function

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I(S_i \leq x)}{n}$$

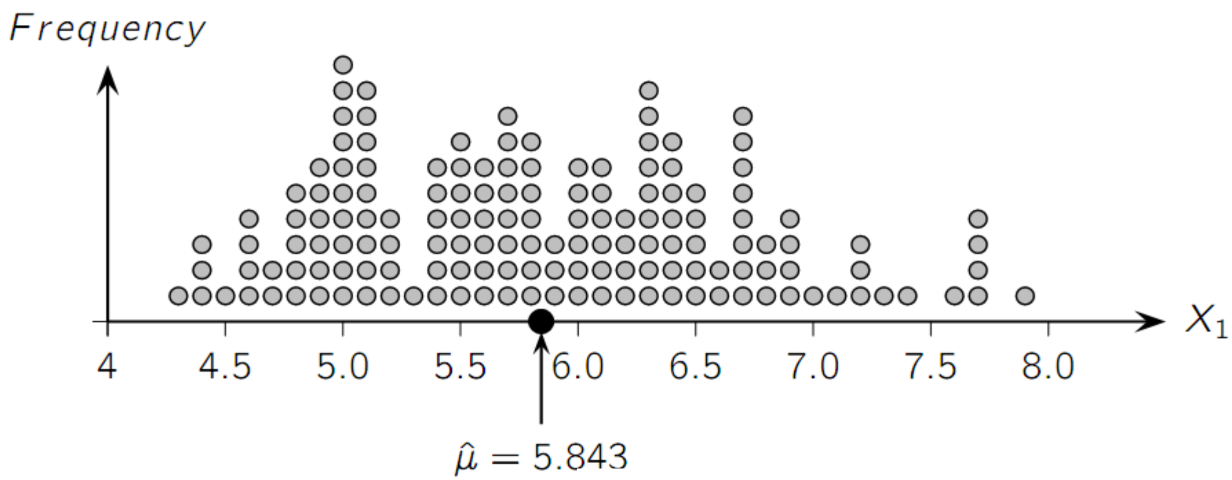
Where

$$I(S_i \leq x) = \begin{cases} 1 & \text{if } S_i \leq x \\ 0 & \text{if } S_i > x \end{cases}$$

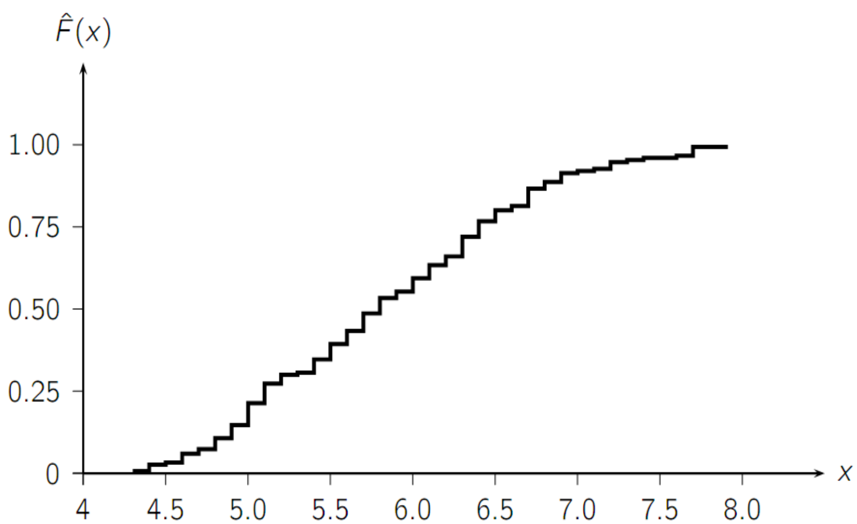
Inverse Cumulative Distribution Function

$$F^{-1}(q) = \min\{x : F(x) > q\} \qquad \text{for } q \in [0, 1]$$

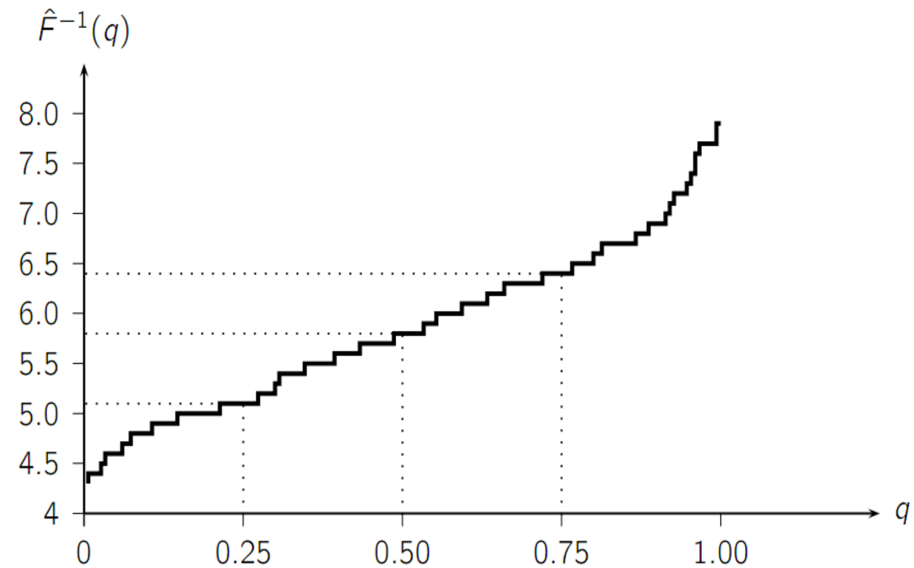
Example



Frequency
X_1
4 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
$\hat{\mu} = 5.843$



$\hat{F}(x)$
1.00
0.75
0.50
0.25
0
4 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
x



$\hat{F}^{-1}(q)$
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4
0 0.25 0.50 0.75 1.00
q

Measures of Central Tendency (Mean)

Population Mean:

$$\mu = E[X] = \sum_x x\, f(x)$$

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

Sample Mean (Unbiased, not robust):

$$\hat{\mu} = \sum_x x \hat{f}(x) = \sum_x x \left(\frac{\sum_{i=1}^n I(S_i = x)}{n} \right) = \frac{\sum_{i=1}^n S_i}{n}$$

$$E[\hat{\mu}] = E\left[\frac{\sum_{i=1}^n S_i}{n}\right] = \frac{1}{n}\sum_{i=1}^n E[S_i] = \frac{1}{n}\sum_{i=1}^n \mu = \mu$$

Measures of Central Tendency (Median)

Population Median:

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

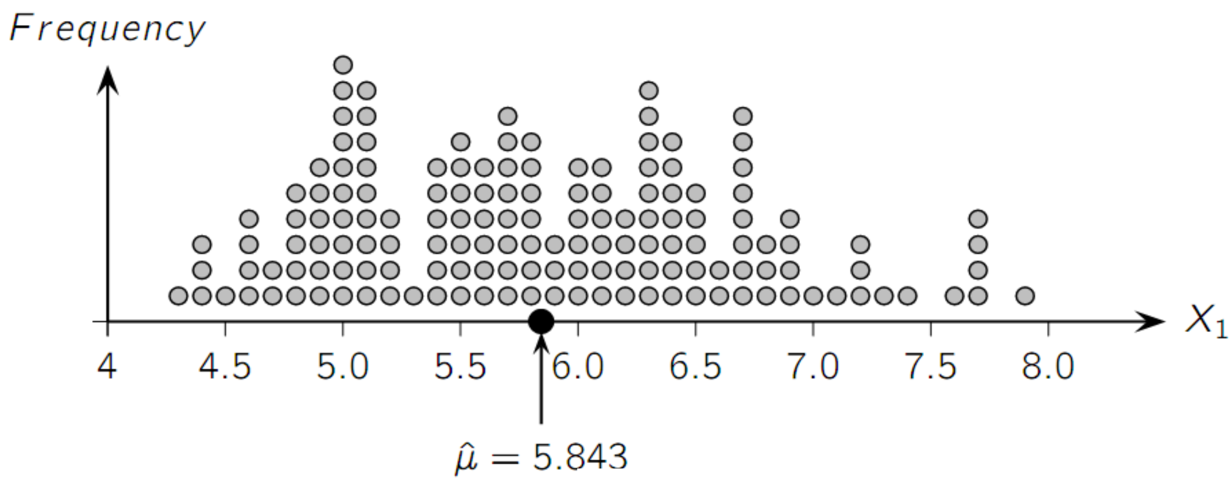
or

$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5)$$

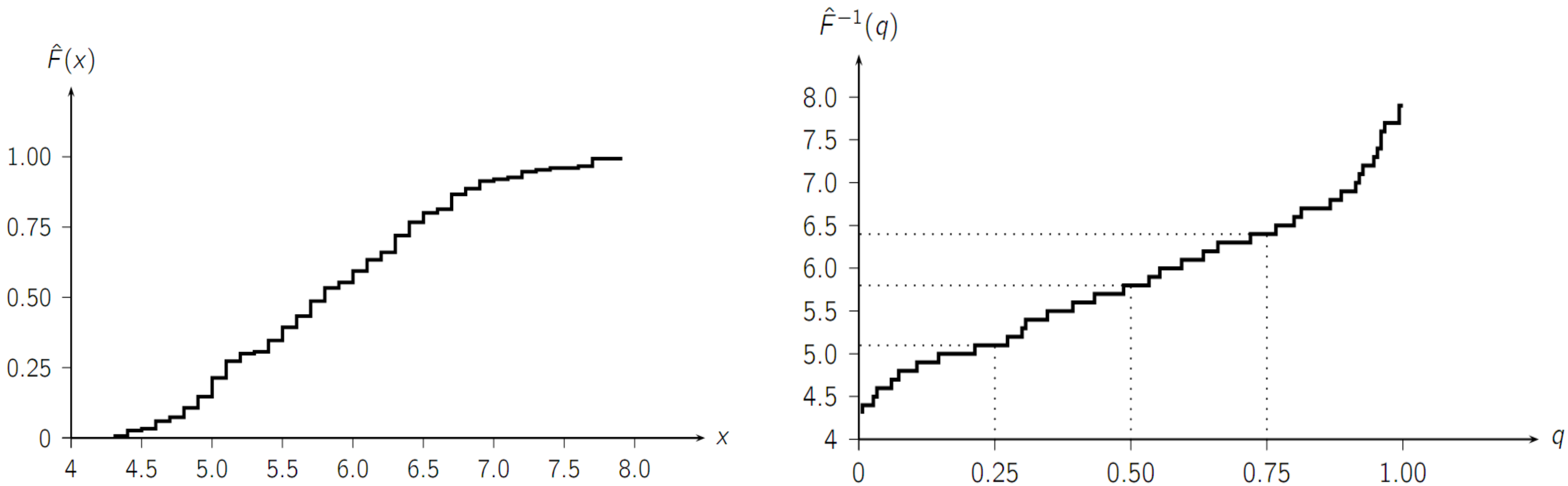
Sample Median:

$$\hat{F}(m) = 0.5 \text{ or } m = \hat{F}^{-1}(0.5)$$

Example



Frequency
X_1
4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
$\hat{\mu} = 5.843$



$\hat{F}(x)$
1.00
0.75
0.50
0.25
0
4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
x
$\hat{F}^{-1}(q)$
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4
0
0.25
0.50
0.75
1.00
q

Measures of Dispersion (Range)

Range:

$$r = \max_x\{x\} - \min_x\{x\}$$

Sample Range:

$$\hat{r} = \max_i\{S_i\} - \min_i\{S_i\} = \max_i\{x_i\} - \min_i\{x_i\}$$

- Not robust, sensitive to extreme values

Measures of Dispersion (Inter-Quartile Range)

Inter-Quartile Range (IQR):

$$IQR = F^{-1}(0.75) - F^{-1}(0.25)$$

Sample IQR:

$$\widehat{IQR} = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

Measures of Dispersion
(Variance and Standard Deviation)

Variance:

$$var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 \ f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 \ f(x) \ dx & \text{if } X \text{ is continuous} \end{cases}$$

Standard Deviation:

$$\begin{aligned} \sigma^2 = var(X) = E[(X-\mu)^2] &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Measures of Dispersion (Variance and Standard Deviation)

Variance:

$$var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 \, f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

Standard Deviation:

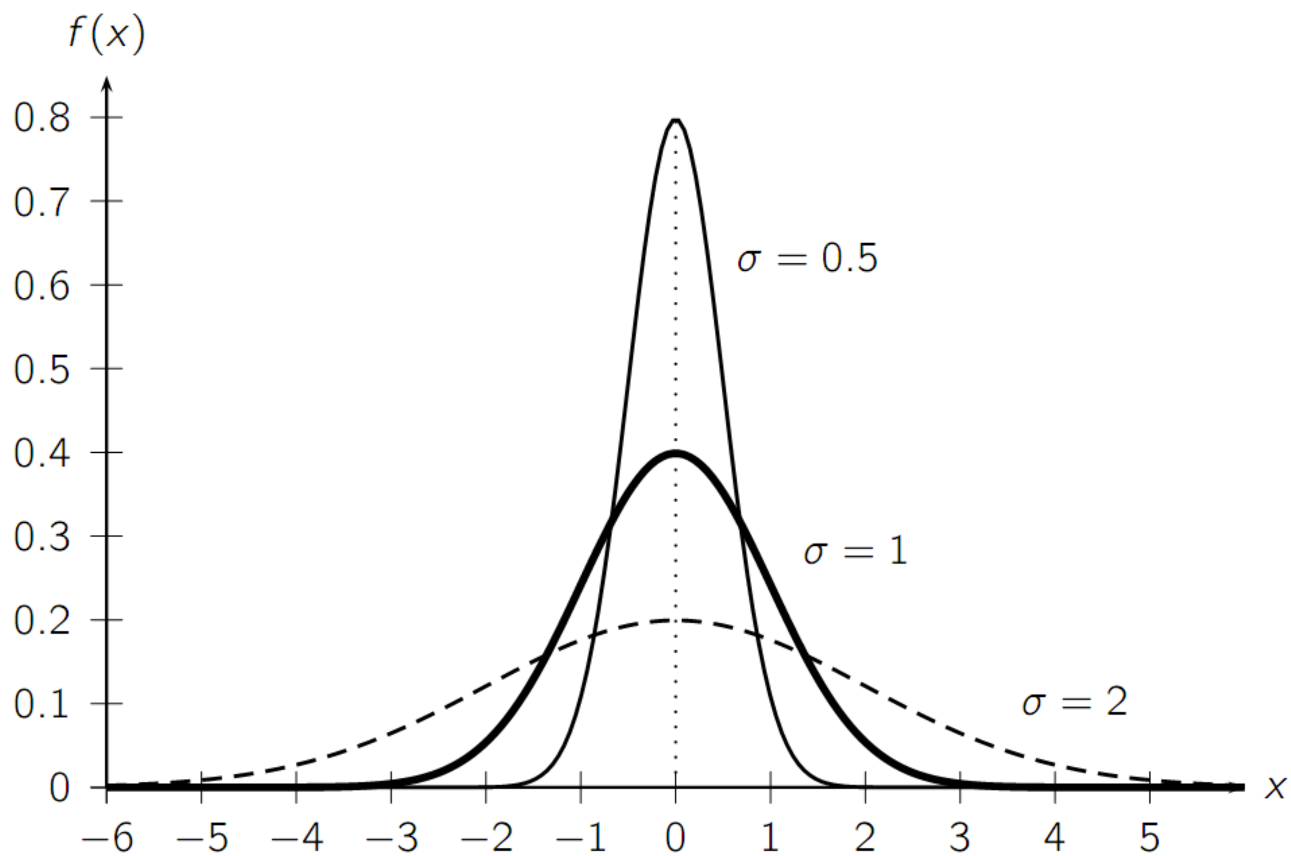
$$\begin{aligned} \sigma^2 = var(X) = E[(X-\mu)^2] &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Sample Variance & Standard Deviation:

$$\hat{\sigma}^2 = \sum_x (x-\hat{\mu})^2 \hat{f}(x) = \sum_x (x-\hat{\mu})^2 \left(\frac{\sum_{i=1}^n I(S_i = x)}{n} \right) = \frac{\sum_{i=1}^n (S_i - \hat{\mu})^2}{n}$$

Univariate Normal Distribution

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



Multivariate Normal Distribution

$$f(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right\}$$