CS 5602 Lecture 11 Probability

George Markowsky
Department of Computer Science
Missouri University of Science & Technology

1

The True Nature of Probability

- The traditional approach to probability, and its cousin statistics, uses a language that is often confusing
- For our purposes, we will revisit these concepts from a set theoretical point of view and avoid the ambiguous language so beloved by probabilists and statisticians

2

Probability

- Here we go revealing the secrets of probability
- It is a matter of translating into ordinary mathematical notation
- What is a **Sample Space**?
- It is just a set
- Generally, the set of objects that interests us

3

Events

- What is an **event**?
- It is just a subset of a sample space, S
- What is an **elementary event**?
- It is just a point in the sample space, S
- What is the **certain event**?
- S
- What is the **null event**?
- $\varnothing$

4

Events

- Two events are **mutually exclusive** iff their intersection is $\varnothing$
- In this course all our sample spaces will either be finite or at most subsets of the integers
- What is random about sample spaces?
- NOTHING!
- What is random about events?
- NOTHING!

5

Probability Distribution

- A **(discrete) probability distribution** on a sample space S is a **function** $f : S \to \mathbb{R}$ s.t.

1. $f(x) \geq 0$ for all $x \in S$
2. $\sum_{x \in S} f(x) = 1$

- What is random about this?
- NOTHING!

6

Probability Spaces

- A **probability space** is a pair (S,f) where S is sample space and f is a probability distribution on S

7

Probability of an Event

- Given a probability space (S,f) and $E \subseteq S$ an event
- The **probability of E**, denoted by $Pr(E) = \sum_{x \in E} f(x)$

8

Fundamental Theorem

- Let (S,f) be a probability space, and A, B are events in S

1. $Pr(A) \geq 0$
2. $Pr(S) = 1$
3. If $A \cap B = \varnothing$, then $Pr(A \cup B) = Pr(A)+Pr(B)$
4. $Pr(A') = 1 - Pr(A)$

9

Uniform Distribution

- The **uniform distribution** is defined for finite sample spaces as follows:
- $f : S \rightarrow \mathbf{R}$ is given by $f(x) = 1/|S|$ for all $x \in S$ (of course we assume $|S| \geq 1$)
- Usually assume uniform distribution unless have other evidence

10

Bernoulli Trial

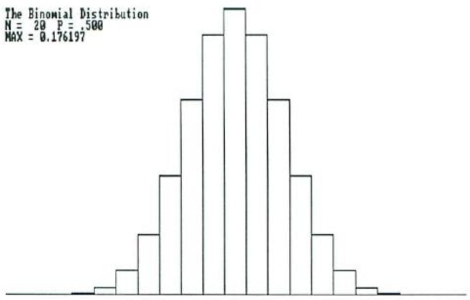
- Sample space = { H, T }
- Could be any other 2 point sample space
- $Pr(\{H\}) = p$ (often written Pr(H))
- $Pr(\{T\}) = q$ (often written Pr(T))
- $p + q = 1$
- $p, q \geq 0$
- Used to model coins

11

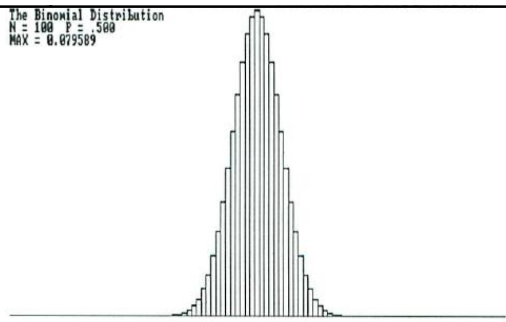
Binomial Distribution

- Have n "independent" coins
- Want to model the number of heads that come up on tosses
- S = { 0, 1, …, n } – will modify later
- $Pr(\{k\}) = C(n,k)p^k q^{(n-k)}$ where
- p is the probability of heads and q the probability of tails, $p+q = 1$, $p,q \geq 0$
- $\sum_{k=0\ldots n} Pr(\{k\}) = (p + q)^n = 1^n = 1$

12

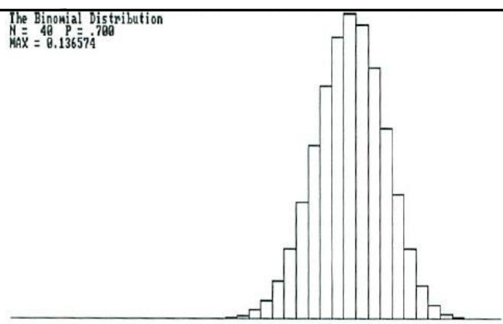


13



Another Binomial Distribution

14



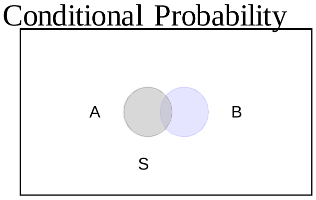
15

Normal Distribution

- This is a distribution over the reals, **R**, so we will not use it much
- Can model the normal distribution by the binomial distribution
- The "law of large numbers" roughly says that the binomial distribution approaches the normal distribution for large n

16

Conditional Probability



The **<u>conditional probability of A given B</u>**, denoted Pr(A|B) is <u>defined</u> as $Pr(A \cap B)/Pr(B)$

17

Independent Events

- Events A and B are said to be **<u>independent</u>** iff $Pr(A) \times Pr(B) = Pr(A \cap B)$
- Note that if $Pr(A) \neq 0$ this means that $P(B|A) = P(B)$
- If $Pr(B) \neq 0$ this means that $Pr(A|B) = Pr(A)$
- A set of events, $\{ A_i \}$ is **<u>pairwise independent</u>** iff A_i and A_j are independent for all $i \neq j$

18

Mutually Independent

- A set of events { A_i } is **mutually independent** iff for every subset of events

$$\Pr(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_j}) = \Pr(A_{i_1}) \times \Pr(A_{i_2}) \times \ldots \times \Pr(A_{i_j})$$

19

Counting

- Rule of Sums
- |A∪B| = |A| + |B| if A and B are disjoint
- |A∪B| = |A| + |B| - |A∩B| in general
- Rule of Products
- |A × B| = |A| × |B|

20

Bayes's Theorem

- It is nice to have an easy theorem named after you, especially when it has been around and is widely used
- Called Bayesian Analysis used on websites to predict what to show people
- Pr(A|B) = (Pr(A)Pr(B|A))/Pr(B)
- What is the proof?
- (Pr(A)P(B|A))/Pr(B)
- = (Pr(A)Pr(A∩B)/Pr(A))/P(B)
- = Pr(A∩B)/Pr(B) = Pr(A|B)

21

Bayesian Analysis

- Also used widely for spam filters
- The sample space is the set of all e-mail
- Event A is that a piece of e-mail is spam
- Event B is that a piece of mail has some properties
- Pr(Spam|Prop) = Pr(Spam)Pr(Prop|Spam)/Pr(Prop)
- Used for preference rankings

22

Simple Example

- Suppose you are given a fair coin and a biased coin (always comes up heads).
- You pick one of the coins at random and toss it twice.
- It comes up heads both times.
- What is the probability that you selected the biased coin?

23

Bayes's Theorem

- The "usual" way to solve this problem is as follows
- Let A be the event that you picked the biased coin
- Let B be the event that a coin would produce two heads in two throws

- Pr(A) = 1/2
- Pr(B|A) = 1
- Pr(B) = (1/2)*1 + (1/2)*(1/4) = 5/8
- Pr(A|B) = Pr(A)Pr(B|A)/Pr(B)
- = (1/2)*1/(5/8) = 4/5
- = 80%
- What does this mean?

24

A Better Way

- I believe that working directly with sample spaces is a better way to understand what you are doing
- What is the sample space here?
- S = { (f,HH), (f,HT), (f,TH), (f,TT), (b,HH), (b,HT), (b,TH), (b,TT) }
- Why show (b,TH) and other points?
- What is the probability space?

25

Using Sample Spaces

S = { (f,HH), (f,HT), (f,TH), (f,TT),
(b,HH),(b,HT), (b,TH), (b,TT)}

- How do we assign probabilities to the points?
- Here are two events. What should their probabilities be?
- Event J = you picked the fair coin
- Pr(J) = 1/2
- Event K = you picked the biased coin
- Pr(K) = 1/2

26

Using Sample Spaces

1/8 1/8 1/8 1/8
S = { (f,HH), (f,HT), (f,TH), (f,TT),
(b,HH),(b,HT), (b,TH), (b,TT)}
1/2 0 0 0

Event A – picked biased coin

Event B – got two heads in a row

Pr(A|B) = Pr(A∩B)/Pr(B) = (1/2)/(5/8) = 4/5 = .8

27

Random Variables

- What is the most important thing to remember about random variables?
- They are neither random, nor variables
- A **random variable** is a function from a sample space into the real numbers
- What is random about this?
- Is it a variable?
- Height, weight, age, etc.

28

Probability Density Function

- Let P = (S,Pr) be a probability space and $X : S \rightarrow \mathbf{R}$ be a random variable
- The **probability density function** of X, written $f_X(r) = Pr(X^{-1}(r))$ for all $r \in \mathbf{R}$
- Can define for "joint distributions" but will skip these definitions

29

Expected Value

- The **expected value of a random variable** V, E(V), is defined by
- $E(V) = \sum_{x \in S} V(x)Pr(\{x\})$
- This is just the "average value" of V
- Notice that E(V) depends very much on the probability distribution
- Different distributions can produce different average values
- Sometimes referred to as the "mean"
- Sometimes written $\mu(X)$ or μ

30

The Average Value Theorem

- For all probability spaces and random variables
- min $\{V(x) \mid x \in S\} \leq E(V) \leq$ max $\{V(x) \mid x \in S\}$
- Proof?

31

Sums of Random Variables

- Given a random variable V, you can multiply it by a constant c to get cV
- Given two random variables V and W, you can add them to get a random variable V + W
- $E(cV) = cE(V)$
- $E(V + W) = E(V) + E(W)$
- Often solve problems by breaking certain random variables into sums of simpler random variables

32

Variance

- If X is random variable, the **variance of X**, Var(X) is defined as
- $Var(X) = E((X-E(X))^2) = E((X- \mu)^2)$
- Note that $Var(X) = E(X^2) – E(X)^2 = E(X^2) – \mu^2$
- How about a proof?
- $Var(X) = E((X- \mu)^2) = E(X^2 – 2\mu X + \mu^2)$ (note that $\mu = E(X)$ is a constant)
- $= E(X^2) – 2\mu*\mu + \mu^2 = E(X^2) – \mu^2$
- $= E(X^2) –E(X)^2$
- This is called the "short form" for variance

33

Standard Deviation

- The **standard deviation of a random variable X**, SD(X), is defined by
- $SD(X) = Var(X)^{.5}$
- SD(X) is often denoted by $\sigma(X)$ or σ

34

Chebychev's Theorem

- Tschebyshev, Tchebyshev, Tchebychev, Chebyshev, etc.
- Very powerful result, and relatively simple to prove
- $Pr(|X - \mu| \geq t\sigma) \leq 1/t^2$
- Thus, for any distribution, the probability that a value is more than 3 standard deviations away from the mean is less that 1/9

35

Proof

- Let $A = \{ s \in S \mid |X(s) - \mu| < t\sigma \}$
- Let $B = \{ s \in S \mid |X(s) - \mu| \geq t\sigma \}$
- By definition,
- $\sigma^2 = E((X-\mu)^2)$
- $= \sum_{s\in A}(X(s)- \mu)^2 Pr(s) + \sum_{s\in B}(X(s)- \mu)^2 Pr(s)$
- $\geq \sum_{s\in B}(X(s)- \mu)^2 Pr(s)$
- $\geq (t\sigma)^2 Pr(B)$
- Thus, $\sigma^2 \geq t^2\sigma^2 Pr(B)$ or
- $1/t^2 \geq Pr(B)$

36

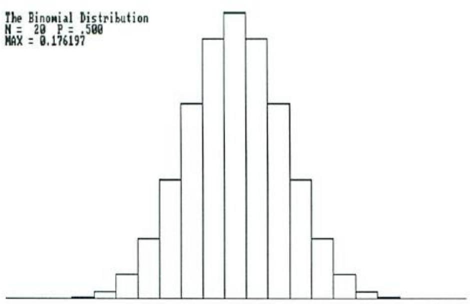
Binomial Distribution

- Will not calculate σ(Heads) – the algebra is a bit more serious than calculating μ(Heads)
- The result is:

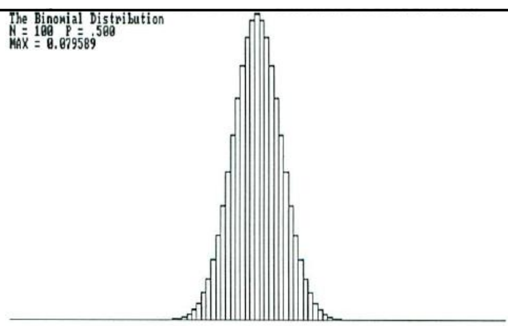
$$\sqrt{npq}$$

Why is this important?

37

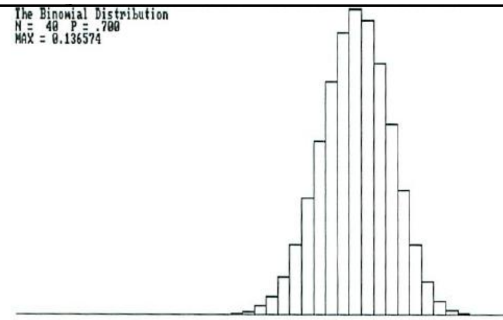


38

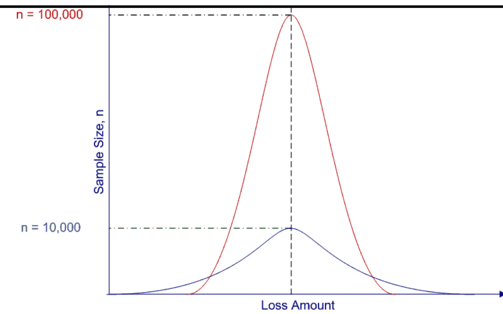


Another Binomial Distribution

39



40



41

The Weak Law of Large Numbers

- One way to see how having more coins makes the result more predictable is to notice that by Chebyshev's Theorem for any distribution at most 1/9 of the time you get a value more than 3 standard deviations away from the mean
- For the binomial distribution, the results are much sharper (about 1%)

42

The Weak Law of Large Numbers

- The SD deviation is $(npq)^{.5}$ and the graph runs from 0 to n, so 90% or more of the graph occupies a percentage of

$$\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

- Which goes to 0 as $n \to \infty$

43

The Birthday Problem

- How many people need to be in a room before you have a 50% chance that at least two of them have the same birthday?
- The surprising answer is that once you have 23 people in a room, you have at least a 50% chance that two of them have the same birthday
- How to see this?
- What is the sample space? the probability space?

44