Sample Spaces and Uniform Probability

Gregory M. Shinault

Goal for This Lesson

We want to describe the basic setup of a probability model for a random experiment.

This material roughly corresponds to Section 1.1 of the textbook. We will cover Appendices B and C (set theory and counting), before fully tackling Section 1.1.

Introduction

What is Probability Theory?

- The mathematical formalism used to quantify uncertainty or randomness, usually in a scientific or social experiment
- Probability is NOT statistics, nor is it a subset of statistics
- Probability is a mathematical theory

process to establish a probabilistic model
(1) identify the Ω set of all outcomes
(2) Events underlying set subset of Ω

Relationship Between Probability and Statistics

So why is this course required to understand statistics?

- The mean, mode, etc. of a data set are *descriptive statistics*
- Want to use descriptive statistics to make predictions for the future (*inferential statistics*)
- Probability theory is the mathematical tool we use to bridge this gap

Warning: We will never analyze a data set in this course, that is for statistics courses. This is not a statistics course.

The Sample Space

Starting Point for this Course: Sample Space

Question: You have an experiment with some form of randomness in it. What is the first thing you should do to describe this experiment?

- **Answer:** Identify all possible outcomes. Sample Space

Ω is a set NOT A Number

- **Definition:** For a random experiment the set of all basic outcomes is called the *sample space*. It is denoted by Ω.
- *Comment:* The sample space Ω is the fundamental "number system" for probability.

Examples

Identify the sample space in the following random experiments.

1. We roll a six-sided die. $\Omega = \{1,2,3,4,5,6\}$
2. We conduct a medical drug trial on 73 patients and count the number of patients for whom the drug is effective. $\Omega = \{0,1,2,\ldots,73\}$
3. We count the number of car crashes in Madison today. $\Omega = \{0,1,2,\ldots\}$
4. We identify the location of the first lightning strike in Dane county today. $\Omega = \{\text{All land in Dane county}\} \cup \{N\}$ N: Never occured
5. We roll a six-sided die twice. $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$

Nuances

The sample spaces can be classified by the number of outcomes in them.

1. Discrete sample space, finite
2. Discrete sample space, finite
3. Discrete sample space, (countably) infinite
4. Continuous sample space
5. Discrete sample space, finite

We will later see that the theory requires different tools for each classification.

Events

Definition

Motivation: Are we always interested in an exact outcome? No, sometimes we care about a range of outcomes.

Definition: An *event E* is a subset of the *sample space*.

Comment: We want to assign probabilities to events, not necessarily outcomes.

Examples

Identify the sample space in the following random experiments and the event described.

1. We roll a six-sided die twice and would like to know the probability the sum is 7.
2. In a family with two children we would like to know if the older child is a boy.
3. We measure the temperature of this room and would like to know the probability that this temperature is between 67F and 75F.

Probability Measure

Main Idea

Starting point: Assume all outcomes are equally likely. Probability was studied primarily with this assumption starting with Gerolamo Cardano in ~1564 continuing through Pierre-Simon, marquis de Laplace in 1812

Warning: This assumption (equal likelihood) is not always justified. We will address the more general situation soon.

Formal Definition

The *uniform probability* on a finite sample space gives the probability of an event E by

$$\mathbb{P}(E) = \frac{\#E}{\#\Omega} = \frac{\text{Number of outcomes in } E}{\text{Number of outcomes in } \Omega}.$$

Examples

Compute the probability of the stated event under the equal likelihood assumption where appropriate.

1. We roll a six-sided die twice and would like to know the probability the sum is 7.
2. In a family with two children we would like to know if the older child is a boy.
3. We measure the temperature of this room and would like to know the probability that this temperature is between 67F and 75F.

The Wrap Up

Summary

We have learned three ideas:

1. The *sample space* Ω is the set of all outcomes.
2. An *event* is a subset of the the sample space. These are what we compute probabilities of.
3. If all outcomes are equally likely we assign the *uniform probability measure* to the events.

finite discrete

Next Steps

1. If we wish to combine events which are represented as sets, we are going to need a little bit of set theory.
2. The only probabilities we can now compute require counting. Let's learn more about counting techniques.

Event Algebra

Gregory M. Shinault

Sample Spaces and Events

Recall the following facts for a random experiment:

1. The *sample space* Ω is the set of all outcomes.
2. An *event* is a subset of the the sample space. These are what we compute probabilities of.
3. A *probability measure* is needed to assign probabilities to events.

Goal for this Lesson

Learn how to combine events mathematically.

The corresponding textbook section is Appendix B.

Operations on Sets

Working Example

$$\Omega = \text{Die roll outcomes} = \{1,2,3,4,5,6\}$$

$$A = \text{Roll is even} = \{2,4,6\}$$

$$B = \text{Roll is prime} = \{2,3,5\}$$

The "AND" Operation / Set Intersection

$3 \in B$

$3 \notin A$

In English: Event A AND Event B occur

Mathematical Notation: $A \cap B = AB$

Computation: $\{2,4,6\} \cap \{2,3,5\} = \{2\}$

The "OR" Operation / Set Union

In English: Event A OR Event B occurs (not XOR)

Mathematical Notation: $A \cup B$

Computation: $\{2,4,6\} \cup \{2,3,5\} = \{2,3,4,5,6\}$

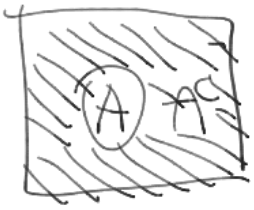
The "NOT" Operation / Set Complement

In English: Event A does NOT occur

Mathematical Notation: $A^c = \Omega \backslash A$

Computation: $\{2,4,6\}^c = \{1,2,3,4,5,6\} \backslash \{2,4,6\} = \{1,3,5\}$

$A^C \quad \bar{A} \quad (\neg A)$



Algebraic Properties of Set Operations

Distributive Laws

$$A(B \cup C) = AB \cup AC$$
$$A \cup (BC) = (A \cup B)(A \cup C)$$

De Morgan's Law

$$(AB)^c = A^c \cup B^c$$
$$(A \cup B)^c = A^c B^c$$

Examples

A simple lottery is being held for a large sum of money. The lottery tickets are labeled 1 through 500. I buy tickets labeled 37, 311, 104, 425, 117. My arch nemesis Matthew McConaughey buys tickets labeled 37, 117, 87, 75. Let G be the event that I win the lottery and M be the event that McConaughey wins the lottery.

Describe the following events in terms of G and M, then explicitly write out the set corresponding to the event (if it is reasonably small).

$\Omega = \{1, 2, \ldots 500\}$

$G = \{37, 311, 104, 425, 117\}$

$M = \{37, 117, 87, 75\}$

$GM = \{37, 117\}$

$G\bar{M} = \{311, 104, 425\}$

$\bar{G}\bar{M} = \Omega$

(a) I win the lottery.
(b) McConaughey wins the lottery.
(c) We both win the lottery.
(d) I win and McConaughey does not.
(e) We both lose.

Partitions

Big Idea

A common problem-solving technique is to decompose your main problem into smaller, easier problems.

nonoverlapping subsets

Within probability theory, this often means breaking your event into smaller events.

With appropriate restrictions, we call this decomposition a **partition**.

Formalism

In English: Event A can be broken down into the distinct cases $A_1, A_2, \ldots$

Mathematical Notation: $A = A_1 \cup A_2 \cup \cdots$ where $A_i A_j = \varnothing$ for $i \neq j$. Some people write $A = A_1 \sqcup A_2 \sqcup \cdots$

Definition: We say that A and B are *disjoint* if $AB = \varnothing$.

Definition: We say that $A_1, A_2, \ldots$ form a *partition* of A if 1. $A_1, A_2, \ldots$ are pairwise disjoint, and 2. $A = A_1 \cup A_2 \cup \cdots$.

Partitioning Countably Infinite Sample Spaces

Give three different ways to partition the sample space $\Omega = \mathbb{N} = \{1, 2, 3, \ldots\}$ into a finite or countable number of events.

Partitioning Uncountably Infinite Sample Spaces

Give three different ways to partition the sample space

- $\Omega = [0, 10]$
- $\Omega = \mathbb{R}^+ = [0, \infty)$
- $\Omega = [0, 3] \times [2, 4]$

into a finite or countable number of events.

Comment: The difference between bounded and unbounded sample spaces is significant. We will address this mysterious comment in detail a bit later.

Wrap Up

Summary

The event algebra is a mathematically formal way to describe events.

1. The fundamental operations we will use are set intersection, union, and complement.

2. The fundamental properties we will use in conjunction with these operations are the distributive and De Morgan's laws.

Next Step

Okay, we have enough set theory to do some serious probability. Now we need to learn how to count.

Counting

Gregory M. Shinault

Goal for This Lesson

We have seen the uniform probability measure for finite sets. It is defined as

$$P(E) = \frac{\#E}{\#\Omega} = \frac{\text{Number of outcomes in } E}{\text{Number of outcomes in } \Omega}.$$

To use this for events and samples spaces that are not tiny, we need a systematic method to count the number of elements in a set. This branch of mathematics is called combinatorics.

The material in this section will correspond to Appendix C in the textbook, and some parts of Section 1.2.

The Multiplication Rule

Introductory Example

Suppose you are buying a new robot. There are three ways to customize your robot.

1. It can be a dog robot or kitty robot.
2. The robot can be black, brown, or magenta.
3. The robot can either speak English or make animal sounds.

How many different robots can be made?

$2 \times 3 \times 2 = 12$

Statement

If we have k successive decisions to make with $n_1, n_2, \ldots, n_k$ choices respectively, then there are $n_1 \cdot n_2 \cdots n_k$ ways to make the k decisions.

Examples

Apply the multiplication rule to determine the number of elements in the following sets.

include or exclude

1. Glass Nickel Pizza offers 37 different toppings. How many different pizzas can Glass Nickel make?

$\underset{\text{1st}}{2} \quad \underset{\text{2nd}}{2} \quad \underset{\text{3rd}}{2} \quad \underset{\text{4th}}{2} \quad \cdots \quad \underset{\text{37th}}{2}$

2^{37}

205 204 203 ... 1 = 205!
1st C 2nd C 3rd C ... 205th

≈ 2.7 × 10^386

2. 205 countries participated in the 2012 Summer Olympics. We must select the order for the countries to march in for the opening ceremonies. How many ways can we do that?
3. The bottom bookshelf in my office has 37 books. I have decided to group them by subject: Analysis, probability, algebra, and physics. There are 9 analysis books, 11 probability books, 6 algebra books, and 8 physics books. How many different ways can my books be arranged, given I group them by subject?

4! 11! 6! 9! 8!
order the subjects | order the prob | order alg | order analysis | order phy

Sample with Replacement, Ordered

Suppose you have an urn filled with 12 marbles, numbered 1 through 12. You take one marble from the urn, record its number, then replace it in the urn. You repeat this process 4 times, keeping track of the order the marbles are sampled from the urn.

(1) $S = \{1, \ldots, 12\}$ Outcome of 1 draw
4 times $\Omega = S^4$

$\{(w_1, w_2, w_3, w_4) \mid w_i \in S\}$
↓ draw 1 ↓ draw 4 for $i = 1, 2, 3, 4$

1. What is an appropriate sample space for this experiment?
2. How many outcomes are in this sample space?
3. Suppose we generalize this experiment to n marbles and k samples. Answer questions 1 and 2 with these more general assumptions.

(3) $S = \{1, \ldots, n\}$ $\Omega = S^k$ assuming ordered
$|\Omega| = n^k$

Permutations

Introductory Example

(2) $\#\Omega = 12^4$

There are 43 instructors in the the math department. Of those instructors, there must be a Department Chair, Vice Chair, and Advising Director. How many ways are there for the department to fill these roles?

$\frac{43 \cdot 42 \cdot 41}{DC \quad UC \quad AD} = (43)_3$

OR

$\frac{43 C 3}{\text{choose 3 people}} \quad \frac{3!}{\text{assign them to the pos}}$

Definition

Suppose there is a set of n elements. Any arrangement of k elements of the set is call a *permutation*.

The number of size k permutations from a set with n elements is denoted by $(n)_k$, and spoken as "n permute k"

$(n)_k = (n-1)(n-2) \cdots (n-k+1)$

$= \frac{n!}{(n-k)!}$

Fact

The number of size k permutations from a set with n elements is

$$(n)_k = n(n-1)(n-2)\cdots(n-k+2)(n-k+1) = \frac{n!}{(n-k)!}.$$

Note this is just a special case of the multiplication rule.

Examples

1. I have 200 books, and 41 of them will fit on the top shelf of my bookcase. How many different arrangements of books are possible on the top shelf?
2. A man forgets his 5-digit PIN, but remembers there is a 37 and 85 in there. What is the maximum number of guesses will he need to get into his account?
3. There are five couples at a fancy dinner party. How many different seating arrangements are there for each couple to be seated together at a round table with 10 chairs? Now the couples are going to the a midnight screening of the cult classic *Billy Jack*. If the party secures 10 seats in a row, how many seating arrangements are there for each couple to be seated together?

Sample Without Replacement, Ordered

Suppose you have an urn filled with 12 marbles, numbered 1 through 12. You take one marble from the urn, record its number, then throw the marble away. You repeat this process 4 times, keeping track of the order the marbles are sampled from the urn.

1. What is an appropriate sample space for this experiment?
2. How many outcomes are in this sample space?
3. Suppose we generalize this experiment to n marbles and k samples. Answer questions 1 and 2 with these more general assumptions.

Combinations

Introductory Example

There are 43 instructors in the the math department. Of those instructors, a 3 person committee must be formed to determine graduate school admission. How many ways are there for the department to form the committee?

Idea

A *combination* is like a permutation, but the ordering does not matter.

Definition: Let Ω be a set of size n. A subset of size k is called a *combination* of k elements of Ω.

The number of combinations of size k is denoted by $\binom{n}{k}$, which is pronounced "n choose k."

Fact

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{(n)_k}{k!}$$

Examples

1. A senate subcommittee is being formed to investigate the actions of **INSERT TIMELY CULTURAL REFERENCE**. There will be 6 members on this committee. 3 will be Democrats, 2 will be Republicans, 1 will be independent. There are currently 53 Democrats, 45 Republicans, and 2 independents in the senate. How many ways are there to form the subcommittee?
2. We have a box of 50 screws. 47 of them have a Phillips head, but 3 of them have a slotted head. We needed 10 screws for our current job, so we pull 10 out of the box at random. What is the probability we pulled out all 3 slotted head screws?
3. Which of the following is larger/largest:

$$\binom{20}{9}, \binom{20}{10}, \text{ or } \binom{20}{11}?$$

$$\binom{20}{1} \text{ or } \binom{20}{19}?$$

Use this (and a few more examples if necessary) to make a conjecture for binomial coefficients in general.

Bonus: If possible, prove your conjectures.

Sample Without Replacement, Unordered

Suppose you have an urn filled with 12 marbles, numbered 1 through 12. You take one marble from the urn, record its number, then throw the marble away. You repeat this process 4 times, but you do not keep track of the order the marbles are sampled from the urn.

1. What is an appropriate sample space for this experiment?

2. How many outcomes are in this sample space?
3. Suppose we generalize this experiment to n marbles and k samples. Answer questions 1 and 2 with these more general assumptions.

Additional Problems

Statements

1. There are about 40 people in this room. What is the probability that 2 people share a birthday?
2. Poker dice is played by rolling 5 dice. What is the probability of rolling a pair? Two pairs?

The Wrap Up

Summary

We learned a few things about counting.

1. The *multiplication rule* is fundamental to solving most counting problems.
2. *Permutations* and *combinations* are just simple counting problems that are usually part of a larger problem.

Next Step

We know enough to handle equally likely outcomes. Most outcomes are not equally likely, so we have to develop a more general mathematical framework.

Axioms of Probability and Infinite Sample Spaces

Gregory M. Shinault

Goal for this Lecture

- Learn the axioms of probability theory in general.
- Learn how to use the additivity axiom when the sample space contains infinitely many outcomes.

The material of this lecture roughly corresponds to 1.1 and 1.3 of the textbook.

The Axioms of Probability

Definition: A *probability measure* and a sample space Ω is a function $\mathbb{P}$: Events $\to \mathbb{R}$ that satisfies

1. $\mathbb{P}(E) \geq 0$,
2. $\mathbb{P}(\Omega) = 1$, and
3. if $E_1, E_2, E_3, \ldots$ are pairwise disjoint then

$$\mathbb{P}\left(\cup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mathbb{P}\left(E_k\right).$$

Comment: These are the only assumptions needed for probability theory to work.

The First of Many Coin Toss Problems

Prove that a fair coin tossed repeatedly will eventually come up heads.

Follow Up Question

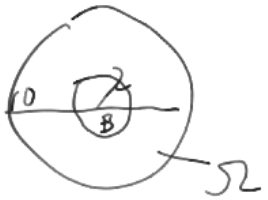
What is the probability that the first heads occurs on an odd flip?

Uncountable Sample Spaces

Equally Likely Outcomes (Uncountable Setting)

Introductory Example: I throw a dart at a dartboard that is 10 inches across. Assuming my aim is terrible and there is no consistency in

my throws, what is the probability that the dart is within 2 inches of the bullseye?



$P(B) = \frac{\text{Area } B}{\text{Area } \Omega}$

uniform measure on continuous Ω

$P(B) = \frac{4}{25}$

Equally Likely Outcomes (Uncountable Setting)

Definition: Suppose Ω is a bounded, uncountable sample space in 1 dimension. The *uniform probability measure* on Ω is given by

$$\mathbb{P}(A) = \frac{\text{Length}(A)}{\text{Length}(\Omega)}.$$

Comment: This generalizes in 2 dimensions to Area, 3 dimensions to Volume, and so on.

$\forall E \quad P(E) = \frac{\text{Area } E}{\text{Area } \Omega} \geq 0 \quad \forall E \in \Omega \quad P(\Omega) = 1$

$P(E) = \frac{\text{Area } E}{\text{Area } \Omega}$ can be splitted w partitions

Examples

1. You drop a piece of chalk that is 3 inches long, and it breaks into exactly two pieces. The break is equally likely to occur at any position on the chalk. What is the probability that the larger of the two broken pieces is more than 2 inches long?

Ω = all pos where break can occur

Ω = [0, 3]

E = larger of 2 pieces > 2 inch

2. We choose a point uniformly from the triangle with corners at $(0,2)$, $(2,0)$, $(4,2)$. What is the probability that the point is above the line $y = 1$?

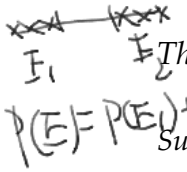
$P(E) = \frac{\text{Area}(E)}{\text{Area}(\Omega)} = \frac{3}{4}$

$P(E) = \frac{\text{Area } B_1 + \text{Area } B_2}{\text{Area } \Omega}$

P(pick P) a point

$\frac{\text{Area of P}}{\text{Area}(\Omega)} = \frac{0}{4} = 0$

The Wrap Up



$P(E) = P(E_1) + P(E_2) = \frac{2}{3}$

Summary

1. The additivity axiom is pretty much the best tool we have for dealing with countably many outcomes.
2. The uniform probability measure is the only tool we have (for now) for dealing with uncountably many outcomes.

Next Step

After seeing how the axioms can be useful for dealing with infinitely many outcomes, we are going to look at what consequences they have for sample spaces in general.

$P(A) = 0$ does not imply impossible

Consequences of the Probability Axioms

Gregory M. Shinault

Goal for this Lesson

Learn how to use the basic properties of probability measures for computation. This requires three pieces of knowledge.

1. The statements of the three axioms of probability
2. The formulas that are consequences of the axioms
3. How to use the formulas to solve problems

The material of this lecture roughly corresponds to Section 1.4 of the textbook.

The Axioms of Probability

Recall the axioms of probability:

1. $\mathbb{P}(E) \geq 0$,
2. $\mathbb{P}(\Omega) = 1$, and
3. if $E_1, E_2, E_3, \ldots$ are pairwise disjoint then

$$\mathbb{P}\left(\cup_k E_k\right) = \sum_k \mathbb{P}\left(E_k\right).$$

Today we are going to look at what consequences these simple rules must have on all sample spaces and events.

Complement Rule

Statement

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^c)$$

Exercise: Prove this.

Examples

1. Compute the probability that two people in this room share a birthday.
2. Approximately 20% of cats are calico or have 6 toes. A cat walks by the window. Find the probability that the kitty is neither calico nor has 6 toes.

Difference Rule

Statement

Proof

$B = A \cup A^c B$ is a partition

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

≥ 0

If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$ and $\mathbb{P}(BA^c) = \mathbb{P}(B) - \mathbb{P}(A)$.

Exercise: Prove this.

Personal Comment: I feel like I rarely actually use this one, but it is important to keep in mind for probability intuition.

Conceptual Example / Conjunction Fallacy

Ex. 1.2

Survey given to study participants

Rate the following statements according to their likelihood:
Linda is active in the feminist movement
Linda is a psychiatric social worker
Linda works in a bookstore and takes yoga classes
Linda is a bank teller and is active in the feminist movement
Linda is a teacher
Linda is a member of the League of Women Voters
Linda is a bank teller
Linda sells insurance

Tversky, A. and Kahneman, D. (1982) "Judgments of and by representativeness". In D. Kahneman, P. Slovic & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. Cambridge, UK: Cambridge University Press.

Inclusion-Exclusion Principle

Simple Version

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$$

Exercise: Prove this.

More General Version

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) = \quad & \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ & - \mathbb{P}(AB) - \mathbb{P}(AC) - \mathbb{P}(BC) \\ & + \mathbb{P}(ABC)\end{aligned}$$

Fully General Version

$$\mathbb{P}\left(\cup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} \left((-1)^{j+1} \sum_{\substack{I \subset \{1,\ldots,n\} \\ |I|=j}} \mathbb{P}\left(\cap_{\ell \in I} A_\ell\right) \right)$$

Comment: This formula is not that helpful to memorize. It is better to understand how to use it.

Example

There are currently courses being offered in analysis, botany, and continuum mechanics. Of 100 students chosen, we have the following statistics:

Courses	Enrollment	Courses	Enrollment
Analysis	35	Analysis and CM	25
Botany	50	Botany and CM	30
Cotinuum Mechanics	40	All three	15
Analysis and Botany	20		

If we select one of the 100 students at random, what is the probability that they are taking at least one of the three courses?

Classic Envelope Problem

You are an important executive, sending n letters to n different people. Unfortunately you hired a fool to stuff the letters into the appropriate envelope. He randomly grabs a letter and stuffs it in an envelope that is already addressed, repeating the procedure until all the envelopes are stuffed.

What is the probability that at least one person gets the correct letter? What is this probability as we consider $n \to \infty$?

The Wrap Up

Summary

1. All probability facts follow from the axioms.
2. Computing a probability is usually some mixture of the additive axiom, complement rule, inclusion-exclusion principle, and differ-

ence rule.

Next Step

We have all the basics of the sample space, which is the fundamental object of probability theory. Now we learn how functions behave in probability.

Introduction to Random Variables

Gregory M. Shinault

Goal for this Lecture

Learn some of the basic concepts surrounding random variables. This will include the following topics.

1. The definition of a random variable, and its distribution
2. Discrete random variables and their probability mass functions

The material of this lecture roughly corresponds to Section 1.5 of the textbook.

Random Variable: It is Just a Function

Definition: A *random variable* (RV) is a function whose domain is a sample space and codomain is the set of real numbers, $X : \Omega \to \mathbb{R}$.

Why RVs?

1. We cannot do conventional algebra on sample spaces.
2. We cannot always identify the sample space.

Random Variables, Explicit Sample Space

We roll 2 dice. Let X be their sum.

0. Give the sample space for the random experiment, and describe the random variable in terms of elements from the sample space.
1. $X = 7$ is an event. Write out this event as a set and find its probability.
2. Identify all the values X can take and compute the probability that X takes each value.
3. Find $\mathbb{P}(X \leq 4)$.

The Distribution of a Random Variable

The **(probability) distribution** of a random variable X is the collection of probabilities $\mathbb{P}(X \in B)$ for any set B of real numbers.

There are many possible ways to define the distribution of a RV. We will use the word "distribution" to refer to all of them.

Discrete Random Variables

range of X finite Countably infinite

Vocabulary

For now we only look at random variables with a countable range.

Defiinition: A random variable X is called *discrete* if the range of X is countably infinite or finite.

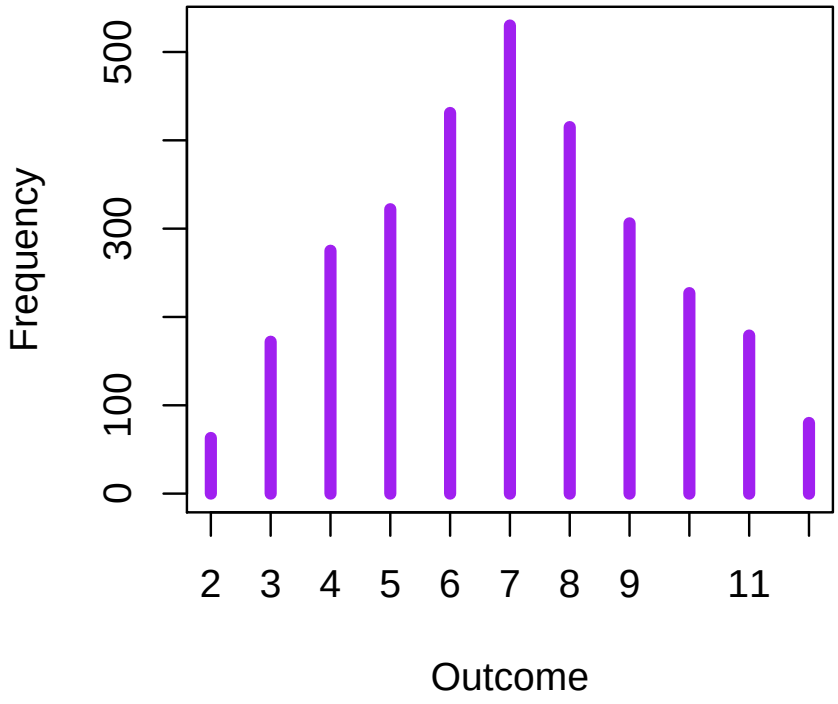
Definition: For each value $k \in \text{Range}(X)$, we define the *probability mass function* (pmf) of X as

$$p_X(k) = \mathbb{P}(X = k).$$

Comment: The PMF is analogous to the relative frequency bar chart for datasets.

PMF vs. Relative Frequency Bar Chart

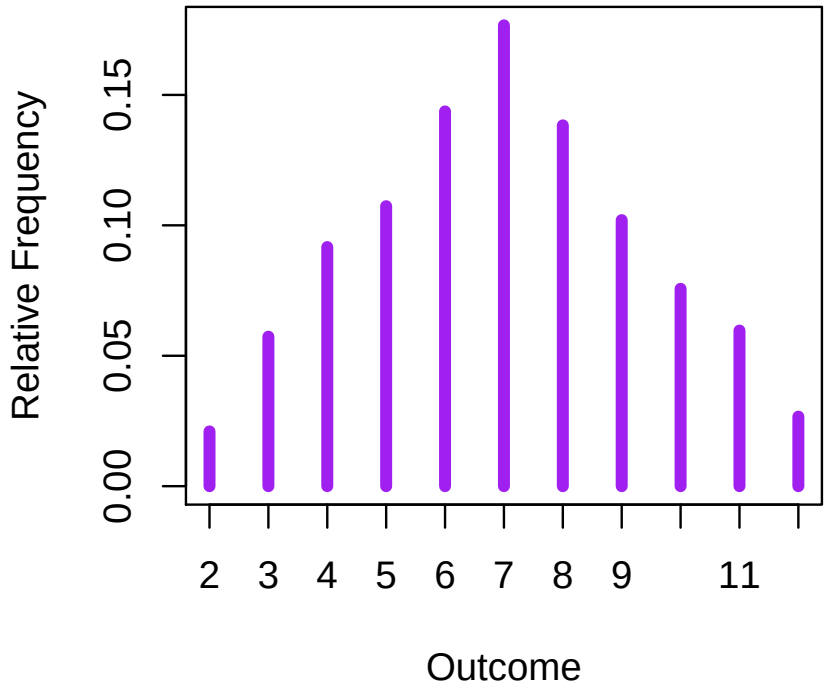
```
## diceSum
##   2   3   4   5   6   7   8   9  10  11  12
##  63 172 275 322 431 530 415 306 227 179  80
```



PMF vs. Relative Frequency Bar Chart

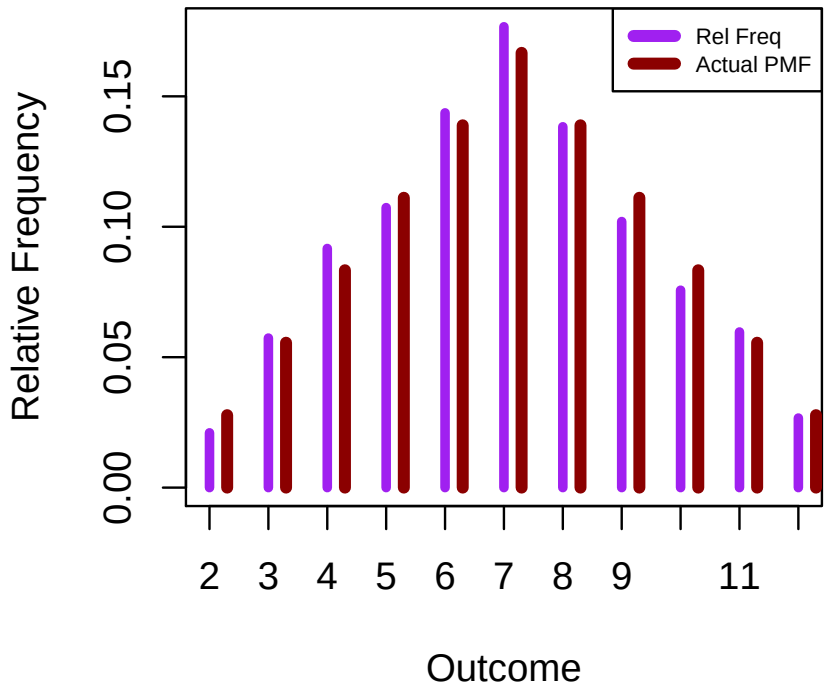
```
## diceSum
```

```
##   2   3   4   5   6   7   8   9  10  11  12
##  63 172 275 322 431 530 415 306 227 179  80
```



PMF vs. Relative Frequency Bar Chart

```
## diceSum
##   2   3   4   5   6   7   8   9  10  11  12
##  63 172 275 322 431 530 415 306 227 179  80
```



PMF vs. Relative Frequency Bar Chart

The Takeaway Lesson: The PMF tells us how frequently an outcome should occur in repeated simulations of the random variable.

Random Variables, Implicit Sample Space

Later:
May not clearly specify Ω bc $X: \Omega \to \mathbb{R}$

Suppose the average number of customers that come to the antique shop I own during the lunch hour is 4. We can model this with X as a random variable with probability mass function

$$\mathbb{P}(X = k) = p_X(k) = e^{-4}\frac{4^k}{k!} \text{ for } k = 0, 1, 2, \ldots$$

1. First verify that this is a valid pmf. That is, check that for all k $p_X(k) \geq 0$ and the probability for all outcomes of X sum to 1.
2. Compute the probability that I have more than 2 customers during the lunch hour.

want to compute $P(X>2) = \sum_{x=3}^{\infty} p(x=k) = 1 - p(X \leq 2)$
$= 1 - P(x=0 \cup x=1 \cup x=2)$
$= 1 - \left(e^{-4}\frac{4^0}{0!} + e^{-4}\frac{4}{1!} + e^{-4}\frac{4^2}{2!}\right) = 1 - 13e^{-4} \approx 0.76$

X = # lunch customers we see on 09/25/2018

1. (1) $e^{-4}4^k \geq 0$, $k! \geq 0$, so ≥ 0
   (2) check $\sum_{k \in Range(X)} P_X(k) = 1$

$$\sum_{k=0}^{\infty} e^{-4}\frac{4^k}{k!}$$

$$e^{-4}\sum_{k=0}^{\infty}\frac{4^k}{k!} = e^{-4}e^{4} = 1 \quad \text{(exponential series)}$$

$$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$$

A Mixed Discrete/Continuous RV

We throw a dart at a circular board that has radius 12 inches. There is a ring that is 0.5 inches thick, with the center halfway from the center of the board to the edge in which the point value of your throw is

worth twice as much. Let X be the distance from where your dart lands to this ring. Compute $\mathbb{P}(X \leq 2)$.

$P = \frac{\pi \cdot 8.25^2 - \pi \cdot 3.75^2}{\pi \cdot 12^2}$

Comment: X is an example of a random variable that is a mixture of discrete and continuous. $\mathbb{P}(X = 0) > 0$, but for all other values $\mathbb{P}(X = d) = 0$.

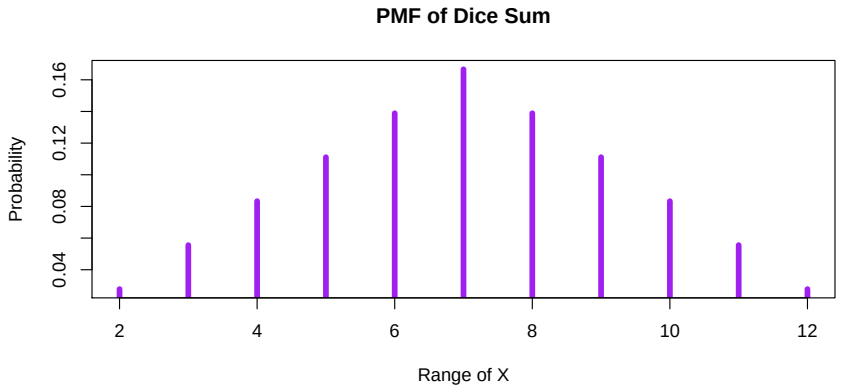
for $d > 0$

Visualization for RVs

- To understand conventional functions $f : \mathbb{R} \to \mathbb{R}$ we usually graph them
- The domain of RVs (sample spaces) do not have a consistent structure
- Thus, no graphs
- Next best thing: graph the pmf

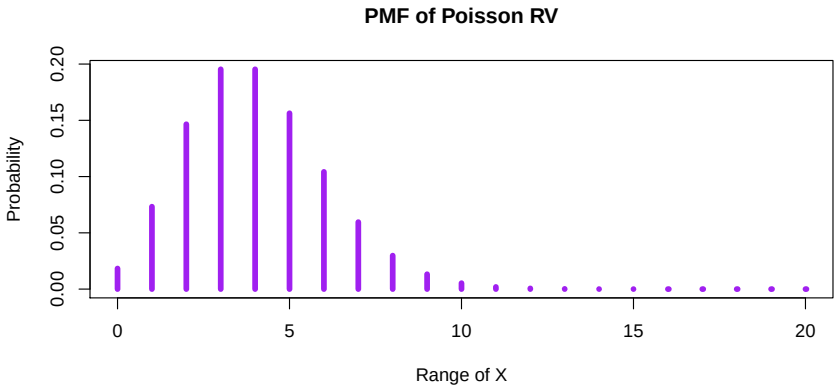
PMF graph

```
ranX <- 2:12; pmfX <- (1/36)*c(1,2,3,4,5,6,5,4,3,2,1)
plot(ranX , pmfX, type = 'h', col="purple", lwd=5,
     xlab="Range of X", ylab="Probability", main="PMF of Dice Sum")
```



PMF graph

```
ranX <- 0:20; pmfX <- dpois(ranX, lambda = 4)
plot(ranX , pmfX, type = 'h', col="purple", lwd=5,
     xlab="Range of X", ylab="Probability", main="PMF of Poisson RV")
```



Applied Example

Tweet Lengths

- Let X be the length of a randomly selected tweet. What is it's PMF?
- The issue: There is no clear theoretical/mathematical approach
- Idea: Collect a lot of tweets, count the length of each tweet, create a frequency table of tweet length
- I did this with some python scripts. One to collect raw tweets, one to process the tweets.
- I wrote the results to a file called "tweet_length.csv"

Frequency Table

We can use R to look at this directly if we like.

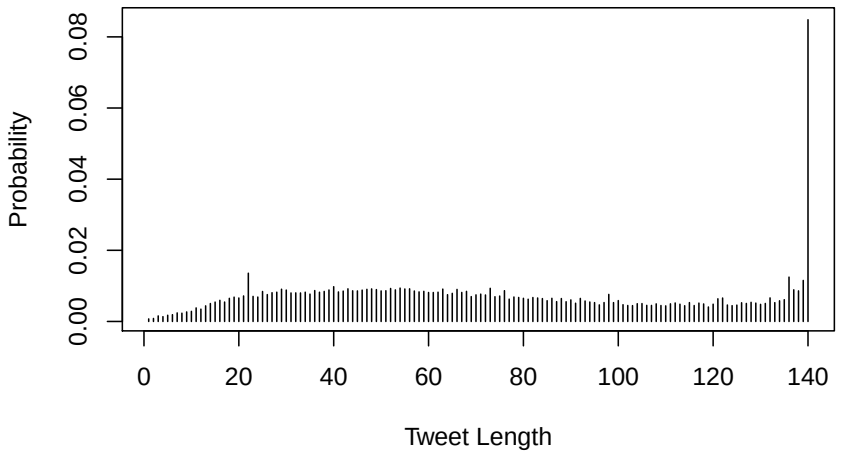
```
tweet_data <- read.csv("tweet_length.csv", header=FALSE,
                       col.names=c("TweetLength", "Freq"),
                       colClasses=c("integer", "numeric"))
head(tweet_data)
```

```
##   TweetLength         Freq
## 1           1 0.0007181124
## 2           2 0.0008434971
## 3           3 0.0015616095
## 4           4 0.0013564345
## 5           5 0.0017439872
## 6           6 0.0019035678
```

Frequency Chart

Just looking at the PMF as a table is not helpful. We should graph it.

```
plot(1:140, tweet_data$Freq, type = 'h', xlab="Tweet Length", ylab="Probability")
```

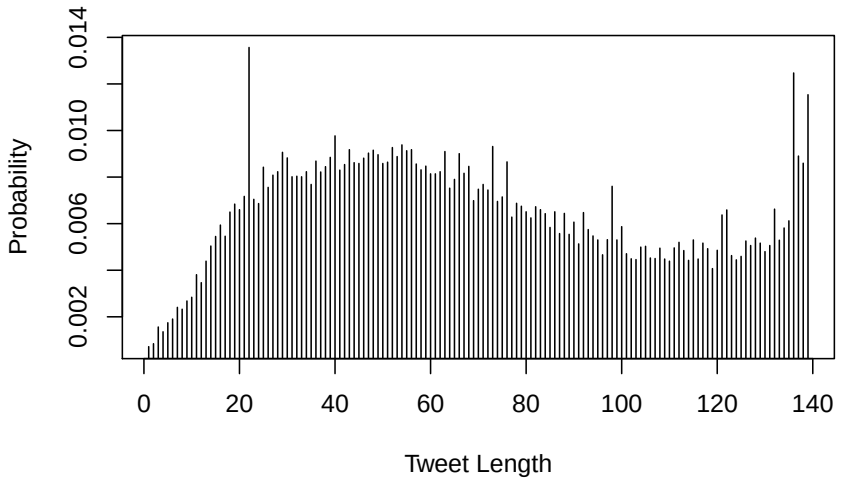


Definitive peak at $k = 140$, probably due to people running out of space.

Trimming Outliers

We can ignore $k = 140$ for additional perspective.

```
plot(1:139, tweet_data$Freq[1:139], type = 'h', xlab="Tweet Length", ylab="Probability")
```



Why is $k = 22$ so big? That seems strange.

Investigation

- Write a script to print tweets of length 22 from the collection of raw tweets.
- Many are of the form "`http://t.co/ABCDEFGHIJ`".

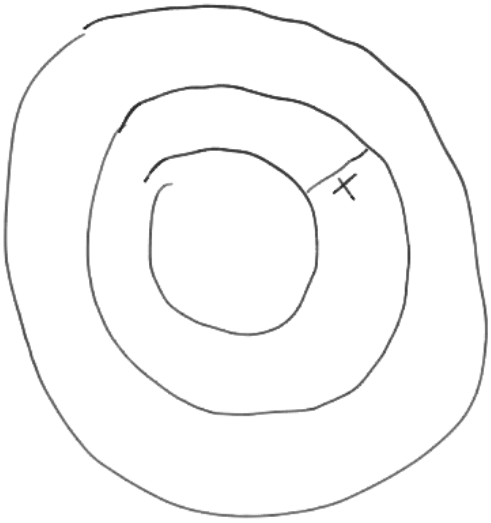
- If someone just posts a link, it is length 22 by default.
- Conclusion: People post a lot of links on twitter.
- Larger punchline: The PMF is not any sort of deep theory, but inspecting the closely related bar chart can reveal some interesting insights in practice.

The Wrap Up

Summary

1. A random variable is just a function with a sample space for a domain
2. The PMF is the most important tool for computing probabilities for discrete RVs
3. To visualize outcomes for a discrete RV, graph the PMF

Mixed Random variables



$$P(X \leq 2) = \frac{\text{Area (shaded region)}}{\text{Area (dart board)}}$$

An Introduction to Conditional Probability

Gregory M. Shinault

How do we update probability with knowledge that certain events occur

Goal for this Lecture

Learn the basic concepts surrounding conditional probability. This will include the following topics.

1. The definition and computation of conditional probability
2. The use of the multiplication rule to compute the probabilities of intersections
3. The use of the law of total probability to compute the probability of a single event, using conditional probabilities

The material of this lecture roughly corresponds to Section 2.1 of the textbook.

Conditional Probability

Restrict sample space to A Shrink attentions to certain outcomes

$\Omega \to \Omega_* = A = \{BB, BG, GB\}$ $P_*(C) = \frac{1}{3}$

Motivating Example

A family has two children. One of them is a boy. What is the probability the other one is as well?

$\Omega = \{BB, BG, GB, GG\}$

Definition

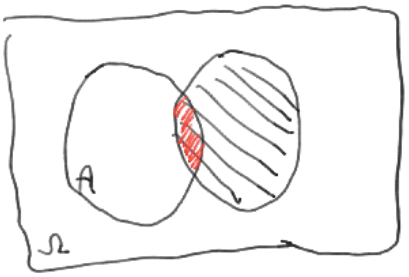
A = at least 1 child is a boy C = both are boys

Definition: The *conditional probability* of A given B is

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

provided that $\mathbb{P}(B) \neq 0$.

Visual Interpretation



Examples

Two cards are dealt from a 52-card deck. The first is a face card. What is the probability the second is a king? (Do not do this with counting. Use the definition of conditional probability for practice.)

① A = first card is a face Want $P(B|A) = \frac{P(BA)}{P(A)} = \frac{4 \cdot 4}{51 \cdot 12}$

B = second card is a king

$P(A) = \frac{12}{52}$

Examples

A family has two children. One of them is a boy that was born on a Tuesday. What is the probability the other one is a boy?

$\Omega = \left(\{B, G\} \times \{1, 2, 3, 4, 5, 6, 7\}\right)^2$ $P(AB) = \frac{4 \cdot 11}{52 \cdot 51}$

$|\Omega| = 14^2$ B_k = kth child is a boy T_k = kth child was born on tuesday

$P(B_1 B_2 | B_1 T_1 \cup B_2 T_2)$

Simulation

```
gender1 <- rbinom(1000000, p=0.5, size=1)
day1 <- sample(1:7, 1000000, replace = TRUE)
gender2 <- rbinom(1000000, p=0.5, size=1)
day2 <- sample(1:7, 1000000, replace = TRUE)
bdayDF <- data.frame(cbind(gender1, day1, gender2, day2))
cond.DF <- subset(bdayDF, (gender1==1 & day1==2) | (gender2==1 & day2==2))
twoboy.DF <- subset(cond.DF, gender1==1 & gender2==1 )
nrow(twoboy.DF)/nrow(cond.DF)
```

```
## [1] 0.4789174
```

13/27

```
## [1] 0.4814815
```

Conditional Probability is a Probability Measure

Fact: The conditional probability $\mathbb{P}(\,\cdot\,|B)$ satisfies the 3 axioms on the smaller sample space $\Omega_* = B$.

Exercise: Prove this statement.

The Multiplication Rule

Key Idea

Suppose we deal three cards from the deck. What is the probability the first card is a jack, the second is a queen, and the third is another jack?

Multiplication Rule for Two Events

Fact:

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B|A)$$

Multiplication Rule for Three Events

More General Fact:

$$\mathbb{P}(ABC) = \mathbb{P}(A)\mathbb{P}(B|A)\mathbb{P}(C|AB)$$

Multiplication Rule for Many Events

Most General Fact:

$$\mathbb{P}(A_1A_2\cdots A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1A_2)\cdots\mathbb{P}(A_n|A_1A_2\cdots A_{n-1})$$

repeated application of 1st version

Examples

$C_k :=$ kth question is one you can correctly answer Want $P(C_1C_2C_3)$

Suppose I am going to give a three question quiz (I am not). Questions are randomly drawn from my problem bank, which has 50 problems in it. If you know the answer to 45 of my questions, what is the probability that you answer all of the questions correctly?

$P(C_1C_2C_3) = P(C_1)P(C_2|C_1)P(C_3|C_2C_1) = \frac{45}{50}\frac{44}{49}\frac{43}{48}$

The Law of Total Probability

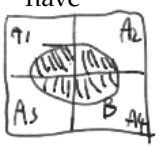
Main Idea

We can use the axiom of countable additivity to decompose a probability for an event into a sum of probabilities for a partition of that event.

If the probabilities are still difficult, we can use the multiplication rule to assist in those computations.

Statement of The Law of Total Probability

Suppose $A_1, A_2, \ldots, A_n$ is a partition of Ω. Then for any event B we have $n=4$



$$\mathbb{P}(B) = \sum_{j=1}^{n} \mathbb{P}(B|A_j)\mathbb{P}(A_j).$$

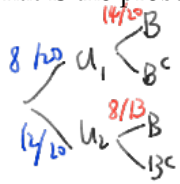
$P(B) = P(BA_1)+P(BA_2)+P(BA_3)+P(BA_4)$

$= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + P(A_4)P(B|A_4)$

Example

We have two urns and a 20-sided die.
Urn 1 has 14 blue marbles and 6 yellow marbles. $B :=$ choose a blue marble
Urn 2 has 8 blue marbles and 5 yellow marbles. $U_k :=$ kth urn is chosen
We roll the die. $D_L :=$ outcome of die roll is L
If the outcome is a prime number we choose Urn 1.
Otherwise we choose Urn 2.
Next we randomly select a marble from the chosen urn.

What is the probability the marble is blue?



$P(B) = P(U_1)P(B|U_1) + P(U_2)P(B|U_2)$

$= \frac{8}{20}\cdot\frac{14}{20} + \frac{12}{20}\frac{8}{13}$

Comment: All of the "marbles in urn" problems seem contrived, but mathematically they are identical to sampling from a population.

Example

There is a diagnostic test for cancer that has a true positive rate of 0.95 and a false positive rate of 0.02. 1 in 10000 people in the population have this type of cancer. If we select a person from the population at random what is the probability they test positive?

C = Person has cancer

T = Person tests positive

$P(T|C) = 0.95 \quad P(T|C^c) = 0.02$

want $P(T)$

$P(T) = P(C)P(T|C) + P(C^c)P(T|C^c) = 10^{-4} \cdot 0.95 + (1 - 10^{-4}) \cdot 0.02$

Example

You are conducting a sociology experiment that requires you to determine the percentage of the population that regularly engages in various types of illegal drug use. The plan requires interviewing the experiment participants, but obviously many of the participants will lie. Design an experimental procedure to remedy this problem.

roll 1 or 2 — say Yes

19 or 20 — No

other — answer truthfully

The Wrap Up

S_y = says "Yes" in survey

T_y = Actual "Yes" person engaged in behavior

At end of experiment we can estimate

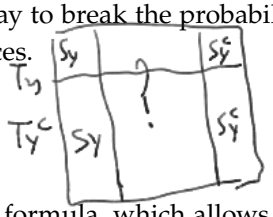
$P(S_y) \approx \frac{\#\text{ say Yes}}{\#\text{ participants}}$ what we really want is $P(T_y)$

Summary

1. The idea behind conditional probability given an event B is to restrict the sample space to B.
2. The multiplication rule is useful if there are sequential events for which conditional probabilities are easy to determine.
3. The law of total probability is a useful way to break the probability of a complicated event into easier pieces.

How do we recover $P(T_y)$ from $P(S_y)$?

$P(S_y) = P(1 \text{ or } 2)P(S_y|1 \text{ or } 2) + P(19 \text{ or } 20)P(S_y|19 \text{ or } 20) + P(\text{other})P(S_y|\text{other})$



$\Rightarrow P(S_y) = \frac{2}{20} \cdot 1 + \frac{2}{20} \cdot 0 + \frac{16}{20} \cdot P(T_y)$

$\Rightarrow P(T_y) = \frac{20}{16}\left(P(S_y) - \frac{2}{20}\right)$

Next Step

We put these pieces together to form Bayes' formula, which allows us to update beliefs in the face of new information.

Bayes' Theorem

Gregory M. Shinault

Goal for this Lecture

Learn about the famous rule of Thomas Bayes. We will see

1. the statement and proof of the Bayes formula,
2. learn how to use the Bayes formula, and
3. see how significant proper use of the Bayes formula is in decision making.

The material of this lecture roughly corresponds to Section 2.2 of the textbook.

Importance of Bayes' Formula

Bayes' Theorem gives us a way to update probabilistic beliefs in the face of new information.

Statement

Let A be an event, and $B_1, \ldots, B_n$ a partition of Ω. Then

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_{\ell=1}^{n} \mathbb{P}(A|B_\ell)\mathbb{P}(B_\ell)}.$$

Comment: This can be memorized, but understanding the derivation is more important.

Example

We have three numbered urns.
Urn 1 has 2 red marbles and 3 blue marbles.
Urn 2 has 2 red marbles and 1 yellow marble.
Urn 3 has 3 red marbles and 2 blue marbles.

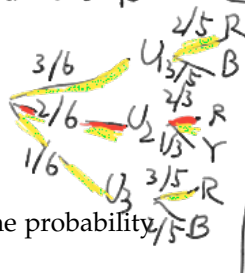
Behind a curtain I conduct an experiment.
I roll a die.
If the die is 1, 2, 3 then I draw a marble from Urn 1.
If the die is 4, 5 then I draw a marble from Urn 2.
If the die is 6 then I draw a marble from Urn 3.

Now I reveal to you that I drew a red marble. What is the probability that I drew from Urn 2?

$P(B_k|A) = \frac{P(AB_k)}{P(A)} = \frac{P(A|B_k)P(B_k)}{P(A)} = \frac{P(A|B_k)P(B_k)}{\sum_{\ell=1}^{n} P(A|B_\ell)P(B_\ell)}$ law of total probability

U_k = choose k^{th} urn
R = marble is Red
B = marble is Blue
Y = marble is Yellow

Want $P(U_2|R)$



$P(U_2|R) = \frac{P(U_2 R)}{P(R)} = \frac{P(R|U_2)P(U_2)}{P(R)} = \frac{\frac{2}{3}\cdot\frac{2}{6}}{P(R)} = \frac{\frac{2}{3}\cdot\frac{2}{6}}{P(U_1)P(R|U_1) + P(U_2)P(R|U_2) + P(U_3)P(R|U_3)}$

≈ 0.4255

Example

Suppose we randomly select a person from the population and apply a diagnostic test for leukemia. Leukemia affects roughly 1/10000 people. This diagnostic test is very accurate:

$$\mathbb{P}(+|L) = 0.995, \quad \mathbb{P}(-|L^c) = 0.999.$$

If the test comes back positive, what is the probability the person tested actually has leukemia?

Example

Let's repeat the previous diagnostic test with a variation. This time we will perform the test on someone with a symptom that is indicative of leukemia in 5% of known cases. This diagnostic test still has accuracy given by:

$$\mathbb{P}(+|L) = 0.995, \quad \mathbb{P}(-|L^c) = 0.999.$$

If the test comes back positive, what is the probability the person tested actually has leukemia?

Visualization of Diagnostic Testing (Sample Size 6000)

Parameters

Let's consider the diagnostic problem with

$$\mathbb{P}(+|\text{Sick}) = 0.9995$$
$$\mathbb{P}(-|\text{Healthy}) = 0.999$$
$$\mathbb{P}(\text{Sick}) = 0.001$$

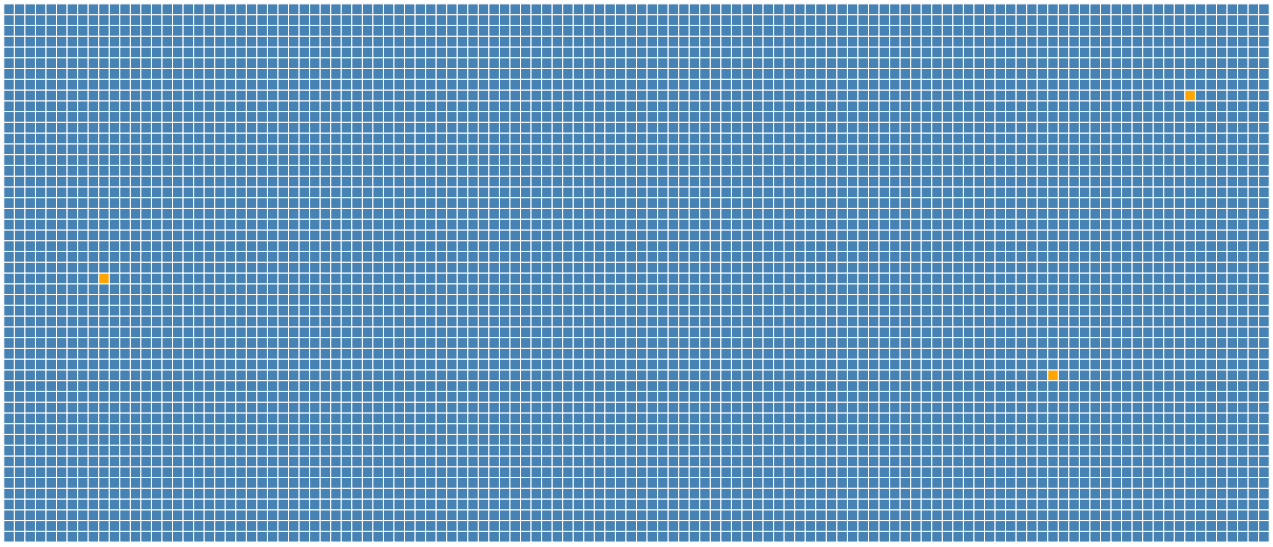
By Bayes'

$$\mathbb{P}(\text{Healthy}|+) = \frac{\mathbb{P}(+|\text{Healthy})\mathbb{P}(\text{Healthy})}{\mathbb{P}(+|\text{Healthy})\mathbb{P}(\text{Healthy}) + \mathbb{P}(+|\text{Sick})\mathbb{P}(\text{Sick})} = 0.4998749$$

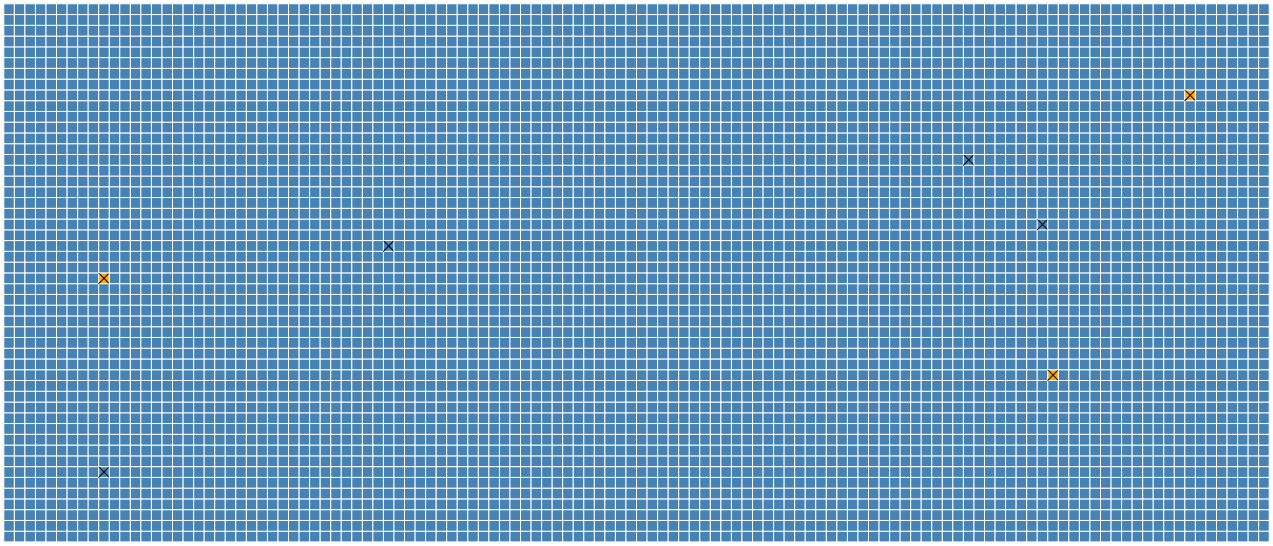
Numerical Analysis of the Simulation

```
## [1] "Number of True Positives:  3"
```

Population



Positive Tests



4 false positive

3 true pos

```
## [1] "Number of False Negatives:  0"

## [1] "Number of False Positives:  4"

## [1] "Number of True Negatives:  5993"

## [1] "Empirical Probability of No Disease Given Positive Test:  0.571428571428571"

## [1] "Actual Probability of No Disease Given Positive Test:  0.499874906179635"
```

The Wrap Up

for P(L) = 0.05 ⇒ 0.98 is accurate rate

Summary

1. Bayes' formula is just a combination of the multiplication rule and the law of total probability.
2. Bayes' formula gives us a method to update probabilities for beliefs when faced with new evidence.

Next step

We have all the basic results for conditional probability now. Next we consider events for which conditioning tells us nothing.

Independence

Gregory M. Shinault

Goal for this Lecture

Learn about the concept of independence. Topics covered will include

1. the definition of independence for two events,
2. the definition of independence for many events, and
3. the definition of independence for random variables.

The material of this lecture roughly corresponds to Section 2.3 of the textbook.

Independence of 2 Events

Big Idea

Independent is the term we use for when conditioning tells us nothing:

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

The Flaw: $\mathbb{P}(A|B)$ is not defined if $\mathbb{P}(B) = 0$. only defined when $P(B) > 0$

Definition

This is an a priori non-symmetric relation

$A \perp B$

We say the events A and B are *independent* if This equation fixes both issues

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B).$$

Example

Prove the outcomes in a sample with replacement are always independent. R_1 := 1st draw is red B_2 := 2nd draw is blue NTS $P(R_1B_2) = P(R_1)P(B_2)$

More rigorously, we have a bin with r red marbles and b blue marbles. We randomly select 1 marble and record its color, and place it back in the bin. Then we select another marble and record its color. Let R_1 be the event the first marble is red and B_2 be the event the second marble is blue. Determine if R_1 and B_2 are dependent for any choice of parameters r and b.

$$P(R_1) = \frac{r}{r+b} \quad P(B_2) = \frac{rb+bb}{(r+b)^2} = \frac{(r+b)b}{(r+b)^2} = \frac{b}{r+b}$$

$$P(R_1B_2) = \frac{rb}{(r+b)^2} = P(R_1)P(B_2)$$

Examples

Is it ever possible for two distinct outcomes in a sample without replacement to be independent?

More rigorously, we have a bin with r red marbles and b blue marbles. We randomly select 1 marble and record its color. Then we select another marble and record its color. Let R_1 be the event the first marble is red and B_2 be the event the second marble is blue. Determine if R_1 and B_2 are independent for any choice of parameters r and b. $P(R_1)=\frac{r}{r+b}$ $P(B_2)=\frac{rb+b(b-1)}{(r+b)(r+b-1)}=\frac{b(r+b-1)}{(r+b)(r+b-1)}=\frac{b}{r+b}$

Useful Facts $P(R_1B_2)=\frac{rb}{(r+b)(r+b-1)}\neq\frac{rb}{(r+b)(r+b)}$ They are not independent they are dependent

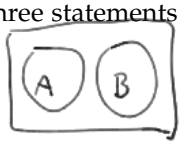
The following are equivalent.

1. A and B are independent.
2. A and B^c are independent.
3. A^c and B^c are independent.

$P(AB^c)=A\setminus AB=P(A)-P(AB)=P(A)-P(A)P(B)=P(A)(1-P(B))$
A and B^c are independent

Exercise: Prove the three statements are equivalent.

Common Mistake



$AB=\emptyset$ A B are independent
B occur
Complete information A will not occur
WRONG
Disjoint AB can only be indt if $P(A)=0$ or $P(B)=0$
Why? $P(AB)=P(A)P(B)$
$P(A)P(B)=P(\emptyset)=0$

WARNING: Do not confuse *independent* and *disjoint*.

Independence of Many Events

Definition

Idea: If ABCD independent you would hope $P(AC)=P(A)P(C)$ or $P(BCD)=P(B)P(C)P(D)$ or something similar for any subcollection of ABCD

We call the events $A_1, A_2, \ldots, A_n$ *(mutually) independent* if

$$\mathbb{P}(A_1^* A_2^* \cdots A_n^*) = \mathbb{P}(A_1^*)\mathbb{P}(A_2^*)\cdots\mathbb{P}(A_n^*).$$

where A_j^* is either A_j or A_j^c.

We need strong conditions to ensure this works

(So that is 2^n equations to check!)

WARNING:

$A_1, A_2, \ldots, A_n$ are pairwise independent
$\not\Rightarrow$ $A_1, A_2, \ldots, A_n$ are independent

$P(A_jA_k)=P(A_j)P(A_k)$ for $j\neq k$

Pairwise independent is not sufficient to prove independence

Equivalent Definition

Fact: The events $A_1, A_2, \ldots, A_n$ are independent if and only if for any subcollection $A_{j_1}, \ldots, A_{j_k}$ we have

$$\mathbb{P}(A_{j_1} \cdots A_{j_k}) = \mathbb{P}(A_{j_1})\mathbb{P}(A_{j_2}) \cdots \mathbb{P}(A_{j_k}).$$

Example

The perfect-use failure rate of the combined oral contraceptive pill is 0.3% That means for a woman that uses the pill exactly as prescribed there is a 0.003 probability that she will become pregnant in one year of use. Let A_j denote the event that the woman becomes pregnant in the j-th year of use, and assume $A_1, \ldots, A_{10}$ are independent.

What is the probability (under these possibly flawed assumptions) a woman using the combined oral contraceptive pill experiences an unplanned pregnancy in the next 10 years?

Now update this probability for the typical-use failure rate of 9%.

Comments

I suspect the assumptions of the previous problem are flawed, but even if they are not these types of figures often lead our mind into the *gambler's fallacy*.

There is a good visualization of these numbers under the same assumptions in an old NYT article. `https://www.nytimes.com/interactive/2014/09/14/sunday-review/unplanned-pregnancies.html?_r=0`

An astute blogger quickly pointed out the flaws in the probability calculations. `https://andrewwhitby.com/2014/09/15/averages-deceive-birth-control-is-better-than-the-nyt-cred`

Independence of Random Variables

Definition

Let $X_1, X_2, \ldots, X_n$ be random variables on the same sample space. Then $X_1, X_2, \ldots, X_n$ are *independent* if

$$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{k=1}^{n} \mathbb{P}(X_k \in B_k)$$

for all subsets $B_1, B_2, \ldots, B_n$ of $\mathbb{R}$.

Discrete Case

Fact: The discrete RVs $X_1, X_2, \ldots, X_n$ are independent if

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=1}^{n} \mathbb{P}(X_k = x_k)$$

for all choices $x_1, \ldots, x_n$.

Example

Suppose we take a size k sample from the set $\{1, 2, \ldots, n\}$. Let $X_1, \ldots, X_k$ denote the outcomes. Determine if the random variables are independent for

- a sample with replacement and
- a sample without replacement.

The Wrap Up

Summary

1. Independent $\Leftrightarrow \mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$.
2. Do not confuse independence and mutual disjointness.
3. Increasing independence to multiple events is complicated, but you should still know the definitions.
4. Independence of discrete RVs requires only looking at the individual outcomes.
5. The outcomes of a sample
   a. with replacement are independent,
   b. without replacement are dependent.

Next step

We take what we know from independence and look at a few classes of random variables that occur frequently.

Exam wed. 5:30-7:00pm

Independent Trials and Special Distributions

Gregory M. Shinault

Lec 1-9

Goal for this Lecture

Cover the idea of independent trials, and the associated special discrete distributions. Topics covered will include

1. the definition of independent trials and the Bernoulli distribution,
2. the definition and application of the Binomial distribution, and
3. the definition and application of the Geometric distribution.

The material of this lecture roughly corresponds to Section 2.4 of the textbook.

Independent Trials

Big Idea

Repeat a random experiment multiple times, in such a way that the experiments do not affect one another.

Suppose the independent trials we conduct only have two possible outcomes, which we call success and failure. The probability of success is p (thus the failure probability is $q = 1 - p$). We call this a Bernoulli(p) trial.

Bernoulli Distribution

Definition: We say a random variable X has the Bernoulli(p) distribution if

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

We denote this as $X \sim \text{Ber}(p)$. The success probability p is the only parameter of this distribution.

Equivalently, $X \sim \text{Ber}(p)$ if and only if the probability mass function is

$$p_X(k) = p^k(1-p)^{1-k}$$

for $k = 0, 1$.

Binomial Distribution

Introduction

Suppose we are conducting an early clinical drug trial. The medicine we are testing was effective in treating insomnia in 80% of lab rats. The trial is only on 8 human patients. What is the probability at least 6 patients have a positive response to the drug?

Definition

A random variable X is said to have a Binomial(n, p) distribution if it has the probability mass function

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

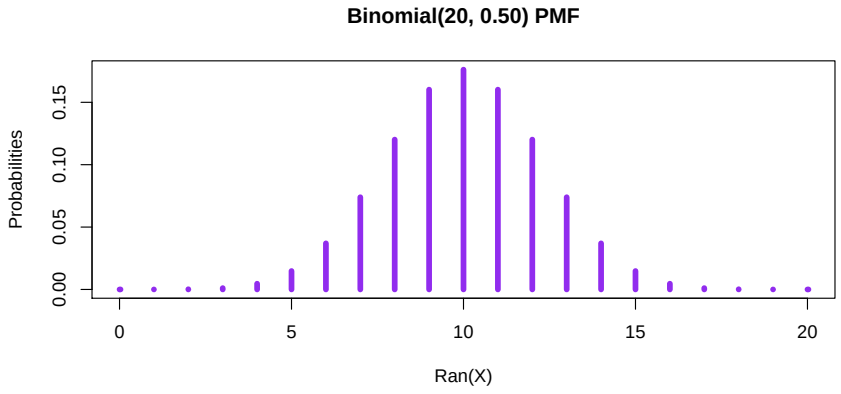
```
n = number of independent trials
p = success probability
X = number of trials resulting
X1,X2 Xn are individual outcomes for each trial
X=X1+X2+...+Xn
```

for $k = 0, 1, 2, \ldots, n$. We denote this by $X \sim \text{Bin}(n, p)$.

The success probability p and number of trials n are the parameters of the Binomial distribution.

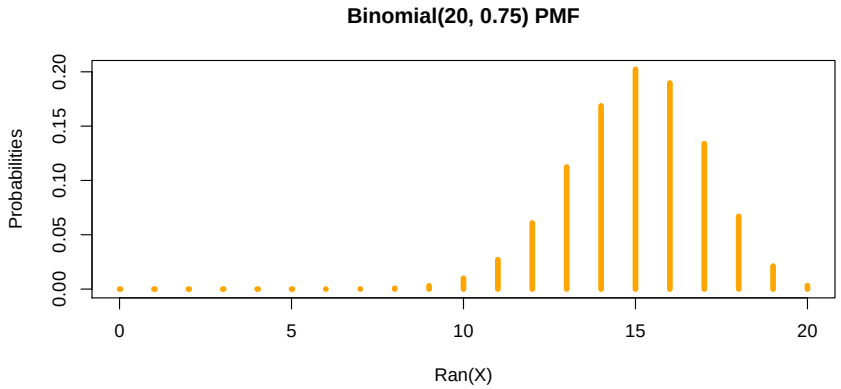
Probability Mass Function

```
PMF1 <- dbinom(0:20, size=20, prob=0.50); kVals <- 0:20
plot(kVals, PMF1, type='h', col="purple2", lwd=5,
     main="Binomial(20, 0.50) PMF", xlab="Ran(X)", ylab="Probabilities")
```



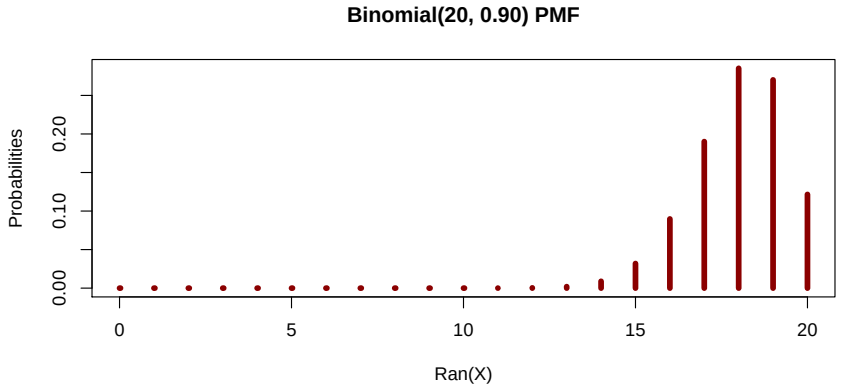
Probability Mass Function

```
PMF2 <- dbinom(0:20, size=20, prob=0.75); kVals <- 0:20
plot(kVals, PMF2, type='h', col="orange", lwd=5,
     main="Binomial(20, 0.75) PMF", xlab="Ran(X)", ylab="Probabilities")
```

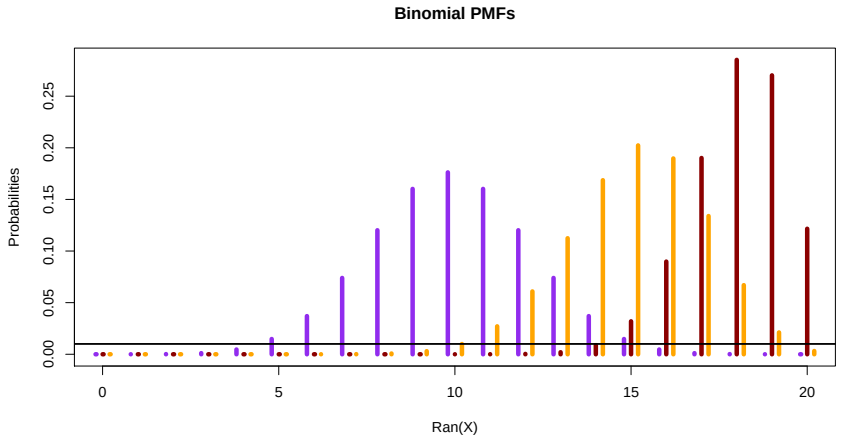


Probability Mass Function

```
PMF3 <- dbinom(0:20, size=20, prob=0.90); kVals <- 0:20
plot(kVals, PMF3, type='h', col="darkred", lwd=5,
     main="Binomial(20, 0.90) PMF", xlab="Ran(X)", ylab="Probabilities")
```



Binomial Distribution



Example

You are conducting a medical trial with 80 patients of a drug that was 75% effective in treating lung cancer in mice. Let X be the number of patients that respond positively to treatment. Assuming this level of effectiveness holds for humans, what is $\mathbb{P}(X \geq 60)$?

Computation in R

The following all yield the same answer.

```
sum(dbinom(60:80, size=80, prob=0.75))
```

```
## [1] 0.5597063
```

```
pbinom(60-1, size=80, prob=0.75, lower.tail = FALSE)
```

```
## [1] 0.5597063
```

```
1-pbinom(60-1, size=80, prob=0.75)
```

```
## [1] 0.5597063
```

Example

You ask 10 people if the prefer raspberry or blueberry cobbler. Raspberry is obviously way better so 70% of the entire population prefers raspberry cobbler.

```
X=#people surveyed out of 10 who prefer raspberry
X~Bin(10, 0.7)
P(X=6)=10C6*0.7^6*0.3^4
p(X=6|X>=3)=p(X=6)/1-P(X=0)-P(X=1)-P(X=2)
```

- What is the probability 6 people prefer raspberry?
- Given at least 3 people prefer raspberry, what is the probability exactly 6 people prefer raspberry?
- Given the first three people you ask prefer raspberry, what is the probability that exactly 6 people prefer raspberry?

```
p(X=6|X1=1,X2=1,X3=1)
=p(X=6,X1=1,X2=1,X3=1)/P(X1=1,X2=1,X3=1)
=P(X1+X2+X3+...+X10=6,X1=1,X2=1,X3=1)/
P(X1=1)P(X2=1)P(X3=1)
=P(X4+X5+...+X10=3, X1=1,X2=1,X3=1)/0.7^3
=P(X4+X5+...+X10=3)P(X1=1)P(X2=1)P(X3=1)/0.7^3
=P(X4+X5+...+X10=3)=7C3*0.7^3*0.3^4
```

Computation in R

```
dbinom(6, size=10, prob=0.7)
```

```
## [1] 0.2001209
```

```
dbinom(6, size=10, prob=0.7)/pbinom(3-1, size=10, prob=0.7, lower.tail = FALSE)
```

```
## [1] 0.2004397
```

```
dbinom(3, size=7, prob=0.7)
```

```
## [1] 0.0972405
```

Applied Example

The Survey of Professional Forecasters is conducted by the Federal Reserve. Predictions are collected from economists and data is compiled to form a 90% confidence interval for GDP growth rate. Collect the data for the past 20 years for the confidence interval (predicted growth rate) and the actual growth rate. How well did the confidence intervals match the actual growth rate? What is the probability a set of true 90% confidence intervals would perform this way?

This analysis can be read in *The Signal and the Noise* by Nate Silver. It is an engaging book that tells stories involving the use of Bayesian statistics. This book is the most enjoyable read to accompany this course, or the website `http://fivethirtyeight.com`.

```
X=#TIMES IN LAST 20 YEARS ACTUAL RATE
WAS IN 90%CONFIDENCE INTERVAL
X~Binomial(20,0.9)
Actual result:X=12
How bad is this?
An unfair measure:P(X=12)=20C12*0.9^12*0.1^8
=0.005
more fair: what is the probability of doing this bad or worse?
P(x<=12)is about 0.006
```

Computation in R

```
dbinom(12, size = 18, prob = 0.90)
```

```
## [1] 0.005243022
```

```
pbinom(12, size = 18, prob = 0.90)
```

```
## [1] 0.00641515
```

Hypothesis Test in R

```
binom.test(12, n = 18, p = 0.90)
```

```
##
##  Exact binomial test
##
## data:  12 and 18
## number of successes = 12, number of
## trials = 18, p-value = 0.006415
## alternative hypothesis: true probability of success is not equal to 0.9
## 95 percent confidence interval:
```

```
##  0.4099252 0.8665726
## sample estimates:
## probability of success
##              0.6666667
```

Geometric Distribution

Introduction

Conduct independent Bernoulli trials. Stop when you have the first success. What is the probability you conduct k trials?

Repeat Bernoulli trials until the first success
X=#trials needed
What is the pmf of X

Definition

A random variable X is said to have a Geometric(p) distribution if it has the probability mass function

$$p_X(k) = (1-p)^{k-1}p$$

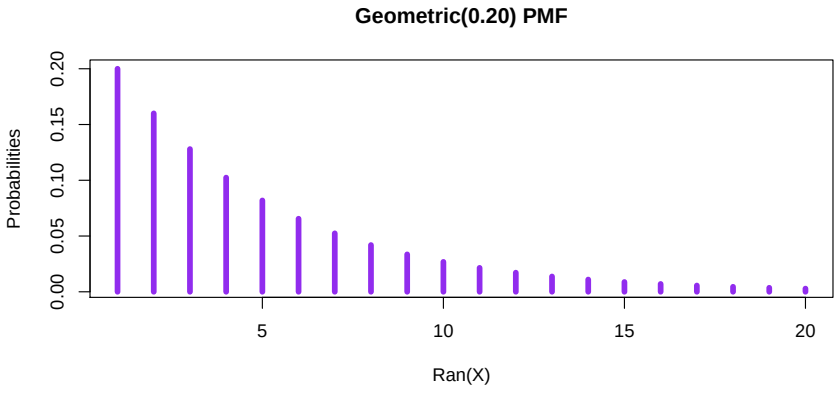
for $k = 1, 2, 3, \ldots$. We denote this by $X \sim \text{Geo}(p)$.

The success probability p is the only parameter in the Geometric distribution.

WARNING: Some sources use a slightly different convention, counting the number of failed trials.

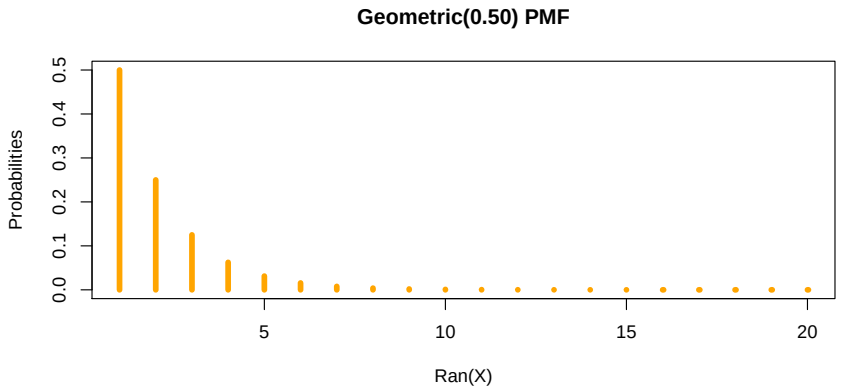
Probability Mass Function

```
GPMF1 <- dgeom(0:19, prob=0.20); kVals <- 1:20
plot(kVals, GPMF1, type='h', col="purple2", lwd=5,
     main="Geometric(0.20) PMF", xlab="Ran(X)", ylab="Probabilities")
```



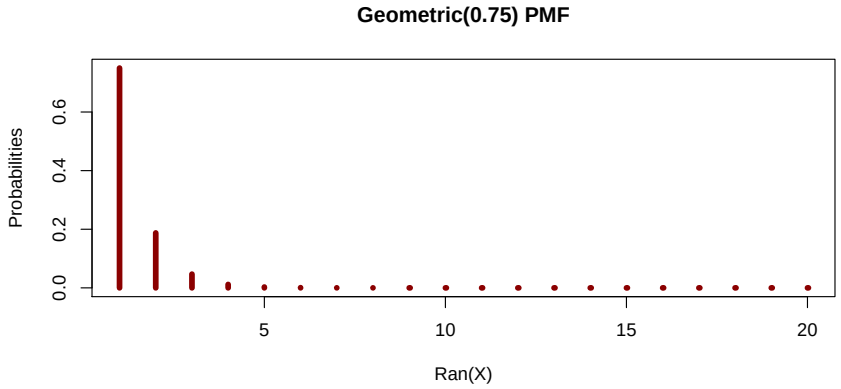
Probability Mass Function

```
GPMF2 <- dgeom(0:19, prob=0.50); kVals <- 1:20
plot(kVals, GPMF2, type='h', col="orange", lwd=5,
     main="Geometric(0.50) PMF", xlab="Ran(X)", ylab="Probabilities")
```

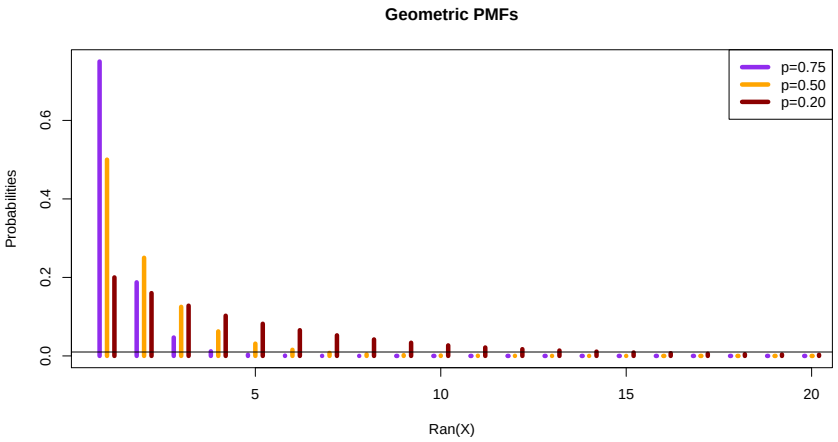


Probability Mass Function

```
GPMF3 <- dgeom(0:19, prob=0.75); kVals <- 1:20
plot(kVals, GPMF3, type='h', col="darkred", lwd=5,
     main="Geometric(0.75) PMF", xlab="Ran(X)", ylab="Probabilities")
```



Geometric Distribution



P(Y>k)=(1-p)^k

Example

We have a bin with 6 purple marbles and 2 orange marbles in it. We draw 4 marbles without replacement from the bin. If there are exactly 3 purple marbles in our sample we stop. If not we put the marbles back in the bin and repeat this procedure. Let X be the number of times we do not draw 3 purples. What is the probability mass function of X? What is $\mathbb{P}(X \leq 5)$?

independent trials
x+1=Y #trials required to draw 3 purples

Y~geometric(p)
X={0,1,2,3…} P(x=k)=P(Y=k+1)=(1-p)^k for k = 0,1,2 P(x<=5)
=1-P(X>5) = 1-P(Y>6) = 1-(1-p)^6

Example

X~Geom(p)

Suppose X is a Geometric(p) RV. Find an expression for $\mathbb{P}(X > k)$ in two ways:

l=k+1 to INF SUM:p(1-p)^l-1
= P
l=k to INF SUM: (1-p)^l

- Use $\mathbb{P}(X > k) = \sum_{\ell=k+1}^{\infty} \mathbb{P}(X = \ell)$ and manipulate the series.
- Make an argument based upon the meaning of $\{X > k\}$ in terms of independent trials.

p(1-p)^k(1+1-p+(1-p)^2…)
=p(1-p)^k(1/(1-(1-p)))=(1-p)^k
x>k first k trials fails

The Wrap Up

Summary

1. Independent trials give us ways to construct interesting random variables with special distributions (we will see at least one more).
2. The binomial disribution gives probabilities for the number of successes in a fixed number of trials.
3. The geometric distribution gives probabilities for the number of trials until the first success.

Next Step

We are done with the basics of conditional probability. Now we just have to cover a few odds and ends to finish the topic.

Miscellaneous Topics in Conditional Probability

Gregory M. Shinault

Goal for this Lecture

Cover a few odds and ends related to conditional probability. Topics covered will include

1. the concept of conditional independence,
2. independence for combinations of independent events, and
3. the Hypergeometric distribution.

The material of this lecture corresponds to Section 2.5 of the textbook.

Conditional Independence

Key Idea

Suppose there is a fair coin, and a coin that has heads on both sides. I randomly choose one of the coins, and flip it twenty times. The result of all twenty tosses is heads.

Do you think the next toss will be heads? Is your answer based on the previous flips? If so, coin flips are supposed to be independent. How do we explain this apparent contradiction? The answer is with conditional independence.

Formal Definition

Let $A_1, A_2, \ldots, A_n$ and B be the events with $\mathbb{P}(B) > 0$. We say that $A_1, A_2, \ldots, A_n$ are **conditionally independent given** B if and only if for

$$\mathbb{P}(A_{i_1} A_{i_2} \cdots A_{i_k} | B) = \mathbb{P}(A_1^* | B)\mathbb{P}(A_2^* | B) \cdots \mathbb{P}(A_n^*) | B).$$

is true for all cases wehre A_k^* is either A_k or A_k^c.

For two events you only check

$$P(A_1 A_2 | B) = P(A_1 | B) P(A_2 | B).$$

Typically use as assumption (No need to check)

Equivalent Definition

Let $A_1, A_2, \ldots, A_n$ and B be the events with $\mathbb{P}(B) > 0$. We say that $A_1, A_2, \ldots, A_n$ are **conditionally independent given** B if and only if

for any $k \in \{2, \ldots, n\}$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, we have

$$\mathbb{P}(A_{i_1} A_{i_2} \cdots A_{i_k} | B) = \mathbb{P}(A_{i_1}|B)\mathbb{P}(A_{i_2}|B) \cdots \mathbb{P}(A_{i_k}|B).$$

Example (2.42 from textbook)

Suppose 90% of coins in circulation are fair, and the remaining 10% of coins are biased.

The biased coins give heads with probability 3/5. We have a random coin and flip it 3 times. What is the proability of getting heads exactly twice?

Example (2.43 from textbook)

We roll two fair dice. Let

A = Event that first die is 1 or 2

B = Event that 3 appears at least once

C = Event that the sum of dice is 5

Show that A and B are dependent, but also A and B are conditionally independent give C.

Independence from Constructed Events

Example (2.41 from textbook)

Let A, B, and C be independent events. Prove that A and $B^c \cup C$ are independent.

The Point of the Previous Example

It might have seemed obvious that A and $B^c \cup C$ are independent, but the actual proof is more complicated than it would seem.

You may use the general fact that constructions like the one in the previous example are independent. However, you should be aware that an elegant proof of this general fact requires mathematical machinery reserved for an advanced course.

The Case for Random Variables

There is a similar fact for random variables: If X, Y, and Z are independent random variables then X and $g(Y,Z)$ are independent for a real-valued function g.

$f: \mathbb{R} \to \mathbb{R}$ $g: \mathbb{R}^2 \to \mathbb{R}$

$f(X)$ and $g(Y,Z)$ are independent.

clean operation $\not\Rightarrow$ independent

ex: X^2 and $\frac{1}{2}(Y+Z)$

Y^2 $\frac{1}{2}(X+Z)$

Not ex: X^2 and $\frac{1}{2}(X+Y)$ might be dependent

$\frac{1}{2}(X+Y)$ $\frac{1}{2}(Y+Z)$ might be dep

Hypergeometric Distribution

Key Idea

We have introduced the idea of a sample without replacement. This is a common enough tool in statistics that we should identify the formula for its probabilities as one of our special distributions.

Motivation in Terms of Samples

Suppose a population is size N. In this population, there are N_A items of type A and $N_B = N - N_A$ items of type B. We take a sample of size n. Let X denote the number of items in our sample that have type A. Then we say that X has the hypergeometric distribution with parameters N, N_A, n.

- What is the PMF of X is the sample is taken with replacement? All n samples are independent, probability of selecting a member of the population with trait A is N_A/N (always)
- What is the PMF of X is the sample is taken without replacement?

$\Rightarrow X \sim \text{Binomial}\left(n, \frac{N_A}{N}\right)$ $P_X(k) = \binom{n}{k}\left(\frac{N_A}{N}\right)^k \cdot \left(1-\frac{N_A}{N}\right)^{n-k}$

$P(X=k)$ = #ways sample can have exactly k members with trait A / #ways sample of size n can occur

$= \frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}}$

Definition in Terms of PMF

We say X follows the hypergeometric distribution for with parameters (N, N_A, n) if its probability mass function is

$$p_X(k) = \frac{\binom{N_a}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}}$$

for $k = 0, 1, \ldots, N$.

$X \sim \text{hypergeom}(N, N_A, n)$ if it has the pmf

This is denoted by $X \sim \text{Hypergeom}(N, N_A, n)$.

Example

$X, Y \sim \text{Binomial}(n, p)$ are independent

Suppose X and Y are independent random variables with the distribution Binomial(n, p). Prove that

original p gets all canceled becomes comb prob

$P(X = k | X + Y = N)$ is equal to pmf of Hypergeom $(2n, N, n)$

$= P(X=k, X+Y=N) / P(X+Y=N)$

$= P(X=k, Y=N-k) / P(X+Y=N) = \frac{P(X=k)P(Y=N-k)}{P(X+Y=N)} = \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{n}{n-k}p^{N-k}(1-p)^{n-(N-k)}}{\binom{2n}{N}p^N(1-p)^{2n-N}}$

$= \frac{\binom{n}{k}\binom{n-n}{N-k}}{\binom{2n}{N}}$

$X+Y \sim \text{Bin}(2n, p)$

is equal to the PMF for the Hypergeom$(2n, N, n)$ distribution.

Hint: What should the distribution of $X + Y$ be? It is fine to make an appeal to intuition for now. We will formally find the PMF of $X + Y$ later using a technique based on moment generating functions.

The Wrap Up

Summary

- There is a subtle difference between independence and conditional independence. Neither necessarily implies the other.
- If you have mutual independence for several events, it is typically possible to create news events using unions, complements, and intersections that are also independent.
- The **hypergeometric distribution** is the special name we give to the PMF of random variables that count the number of outcomes in a particular category for a sample without replacement.

Next Steps

With the development of conditional probability, we have all of the fundamental probability tools in place. Now we dive into the topic of random variables, and build the remainder of our theory around them.

Random Variables, Densities, and CDFs

Gregory M. Shinault

Intro

Reminder

For a discrete RV X the most important object to understand the probabilities related to X is the probability mass function,

$$p_X(k) = \mathbb{P}(X = k)$$

for all $k \in \text{Ran}(X)$.

Goals for this Lecture

1. Define, use, and interpret the probability density function. This is the like the PMF, but used for continuous random variables.
2. Define and use the cumulative distribution function for any random variable.

well defined for ∀ RVs

This material corresponds to section 3.1 and 3.2 of the textbook.

Continuous Random Variables

Definition

The *probability density function* (PDF) of a continuous random variable is the function such that

$$\mathbb{P}(X \leq b) = \int_{-\infty}^{b} f_X(x)\,dx.$$

Facts

0. For $a \leq b$ and for integrable subsets $B \subset \mathbb{R}$

$$\mathbb{P}(a < X \leq b) = \int_{a}^{b} f_X(x)\,dx, \quad = P(X \leq b) - P(X \leq a) = \int_{-\infty}^{b} f_X(x)dx - \int_{-\infty}^{a} f_X(x)dx$$

$$= \int_{a}^{b} f_X(x)\,dx$$

$$\mathbb{P}(X \in B) = \int_{B} f_X(x)\,dx.$$

1. A function f is a PDF if and only if $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
2. For a continuous RV X, $\mathbb{P}(X = c) = 0$ for all c.

$P(-\infty < X \leq \infty)$

$X \in \{c\}$

Example

We say X is *uniformly distributed on* $[a, b]$ if it has the PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

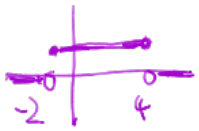
all points from a and b are equally likely

We denote this by $X \sim \text{Unif}([a, b])$.

Let $X \sim \text{Unif}([-2, 4])$. Find $\mathbb{P}(X^2 \le 2)$. $= P(-\sqrt{2} \le X \le \sqrt{2})$

$= P(X = -\sqrt{2}) + P(-\sqrt{2} < X \le \sqrt{2})$

$= 0 + \int_{-\sqrt{2}}^{\sqrt{2}} \frac{1}{6}\,dx = \frac{\sqrt{2}}{3}$



Example

Let X be a continuous random variable with PDF

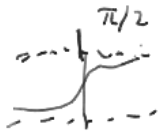
$$f_X(x) = \frac{c}{1+x^2}$$

for all real x. Find the value of c.

use fact $\int_{-\infty}^{\infty} f_X(x) = 1$

$\int_{-\infty}^{\infty} \frac{c}{1+x^2}\,dx = 1$

$c\,[\arctan x]_{-\infty}^{\infty} = 1$



$c[\pi/2 + \pi/2] = c\pi = 1$

$c = \frac{1}{\pi}$

$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$

Interpretation

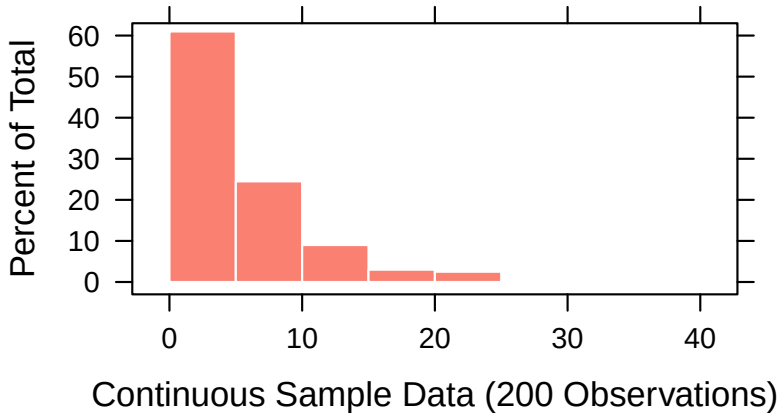
Naive Idea

We want to see how the PDF connects to actual data. To this end we look at a large sample of continuous data.

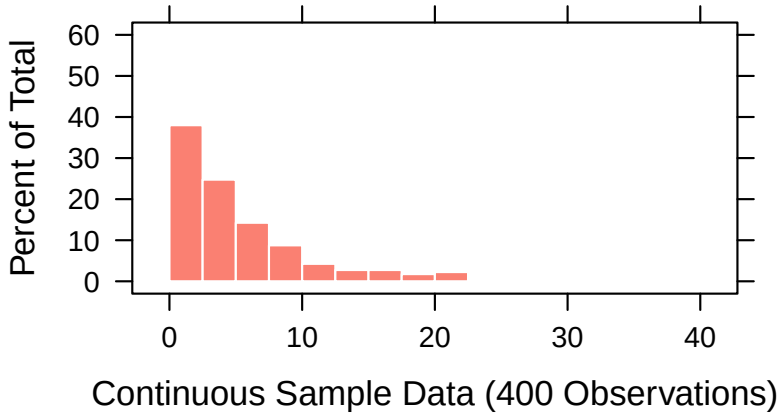
Reminder: The PMF of a distribution is analogous to the frequency bar chart for data generated from that distribution.

Expectation: The PDF of a distribution is analogous to the frequency histogram for data generated from that distribution. This turns out to be incorrect.

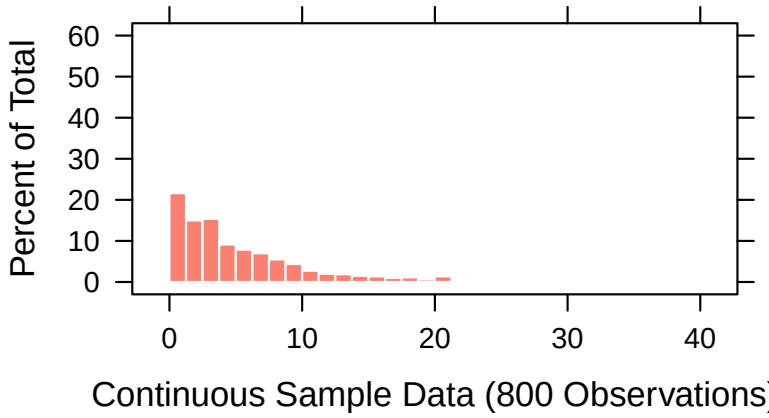
Probability Bins for 200 Samples



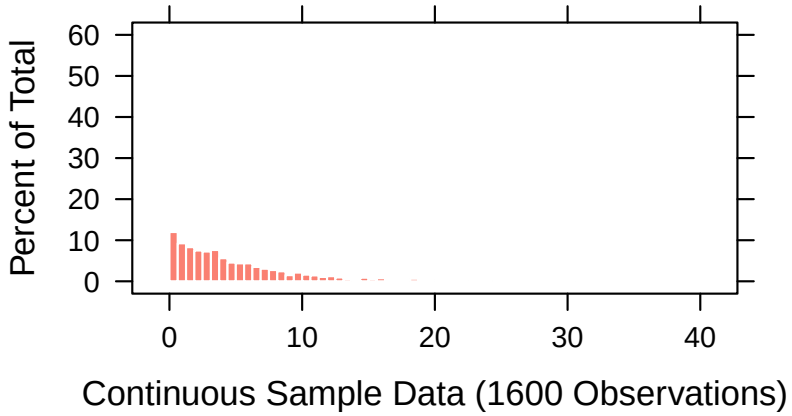
Probability Bins for 400 Samples



Probability Bins for 800 Samples



Probability goes to 0 as number of bins increases

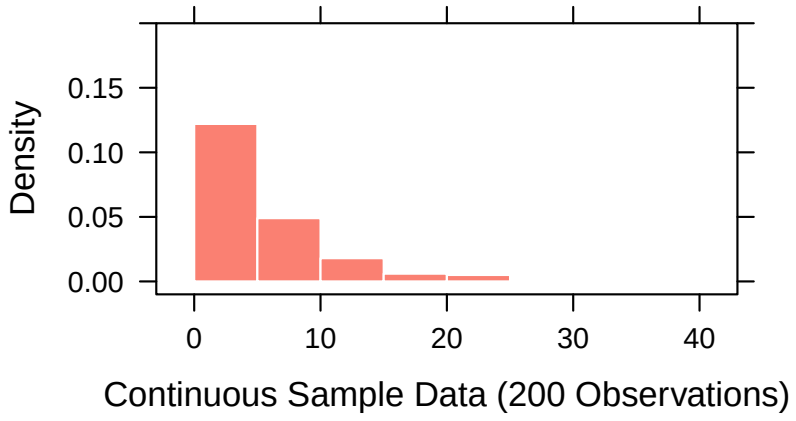


Why does this fail?

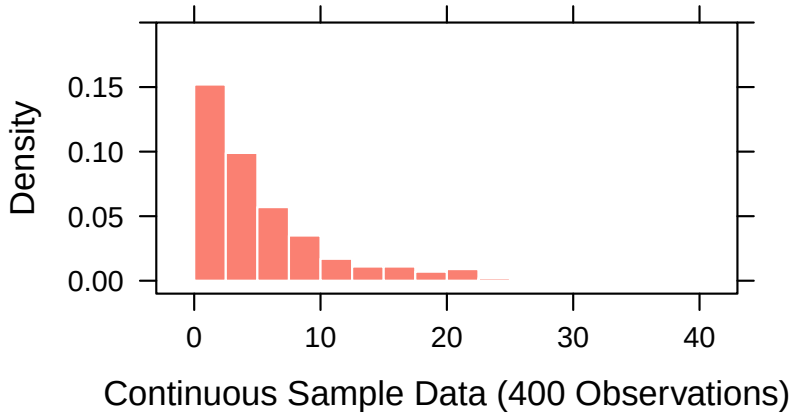
As the bins we are using to construct our frequency histogram get more narrow, the probability of a data point falling into that bin decreases to 0. So the frequency histogram cannot correspond to the PDF.

Fix: Plot the Frequency/Bin Width, rather than the frequency alone.

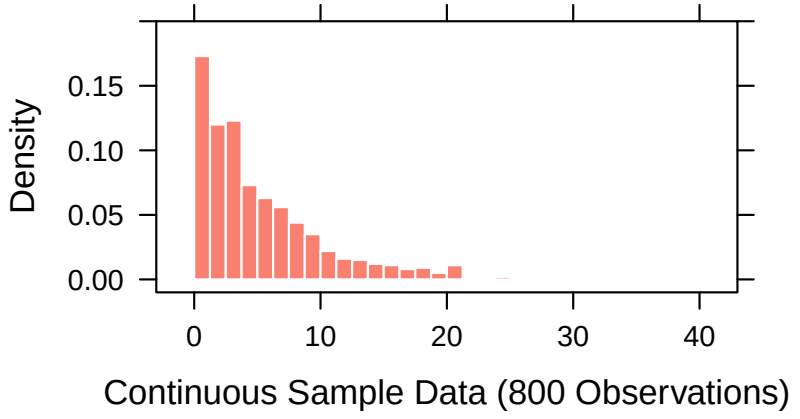
Fix: Plot (Frequency/Bin Width), Probability Density



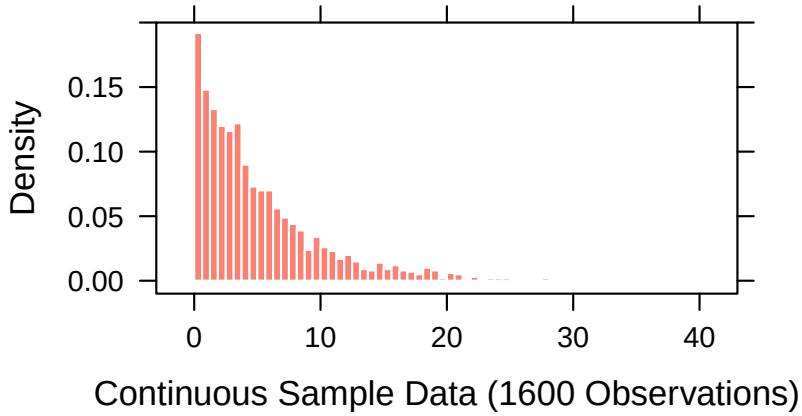
Fix: Plot (Frequency/Bin Width), Probability Density



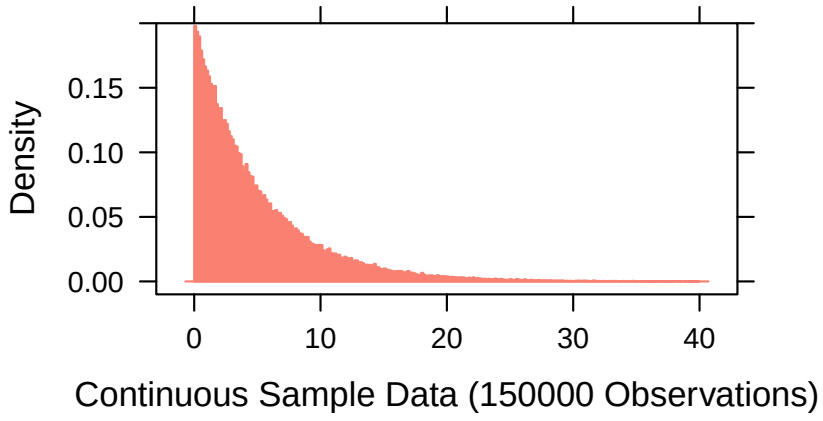
Fix: Plot (Frequency/Bin Width), Probability Density



Fix: Plot (Frequency/Bin Width), Probability Density



Fix: Plot (Frequency/Bin Width), Probability Density



Conclusion

The PDF corresponds to the frequency histogram, divided by bin width.

This is why it is a probability *density* function. The units are probability per unit of X.

Cumulative Distribution Function

Definition

The *cumulative distribution function* (CDF) of a random variable X is defined by

$$F_X(s) = \mathbb{P}(X \leq s).$$

Note the CDF is defined for all random variables: discrete, continuous, and mixed.

Relationship to PDF

If X is a continuous RV then

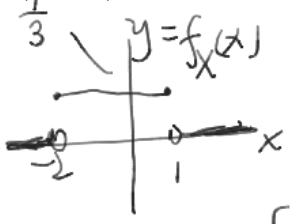
$$f_X(t) = \frac{d}{dt}\left[F_X(t)\right].$$

$$\frac{d}{dt}[F_X(t)] = \frac{d}{dt}[P(X \leq t)]$$
$$= \frac{d}{dt} \int_{-\infty}^{t} f_X(x)\,dx$$
$$= f_X(t)$$

Exercise: Prove this fact.

What is the significance of this fact? If we want to find the PDF of a random variable, it is usually easiest to start by finding $P(X \leq x) = F_X(x)$. We then just differentiate to get the ~~CDF.~~ PDF

find CDF take derivative

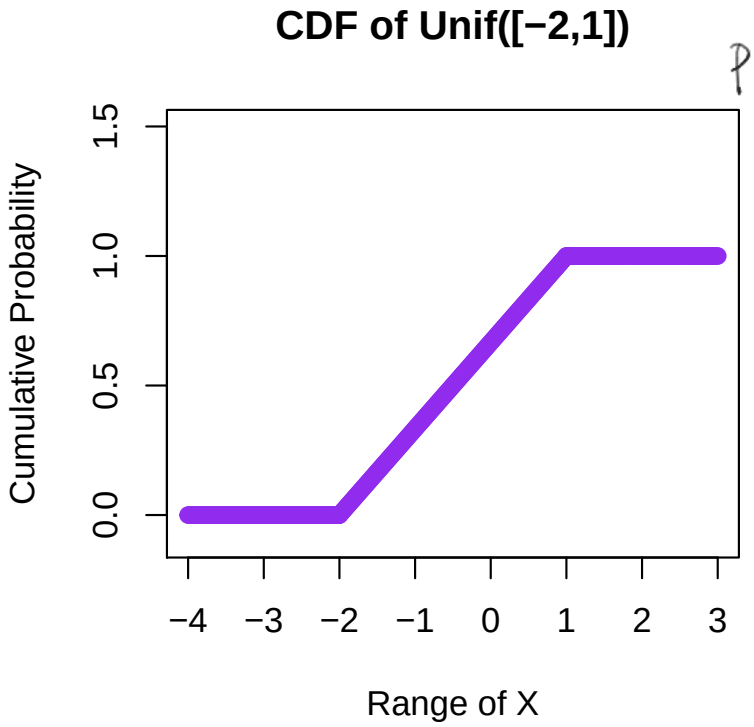


Example

Find and sketch the CDF of $X \sim \text{Unif}(-2, 1)$.

CDF Plot

$$F_X(x) = \begin{cases} 0 & t \leq -2 \\ \frac{t}{3} + \frac{2}{3} & -2 \leq t < 1 \\ 1 & t \geq 1 \end{cases}$$



$$P(X \leq t) = \int_{-\infty}^{t} f_X(x)\,dx$$

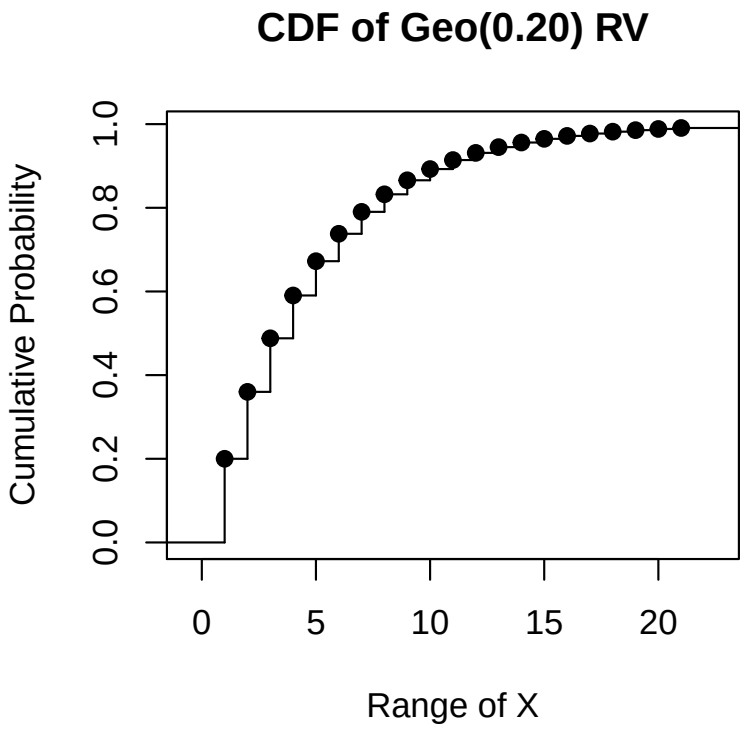
$$= \int_{-\infty}^{-2} 0\,dx + \int_{-2}^{t} \frac{1}{3}\,dx$$

$$= \frac{t}{3} + \frac{2}{3}$$

Example

1. Find the CDF of $X \sim \text{Geo}(0.20)$.
2. Use the CDF to find $\mathbb{P}(4 \leq X \leq 7)$.

CDF Plot

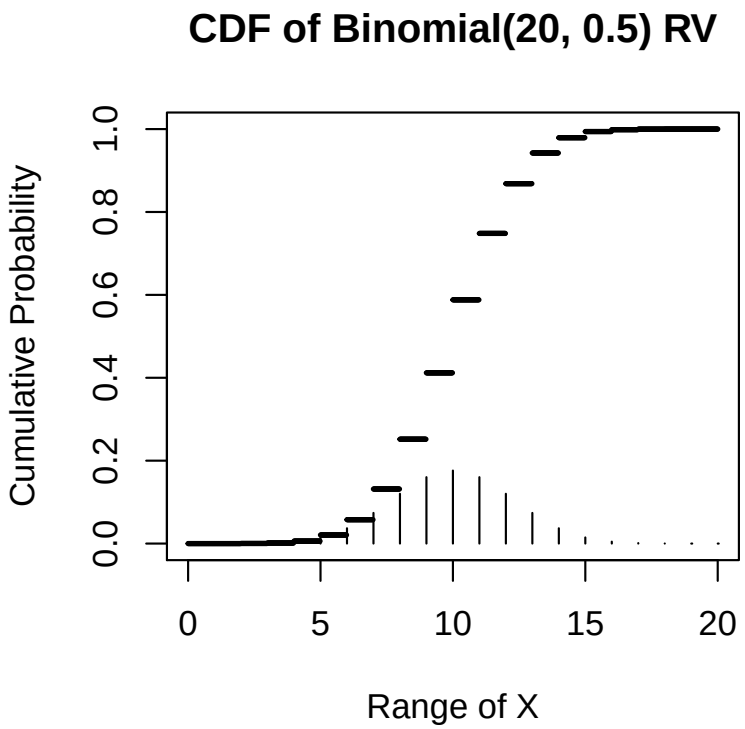


Example

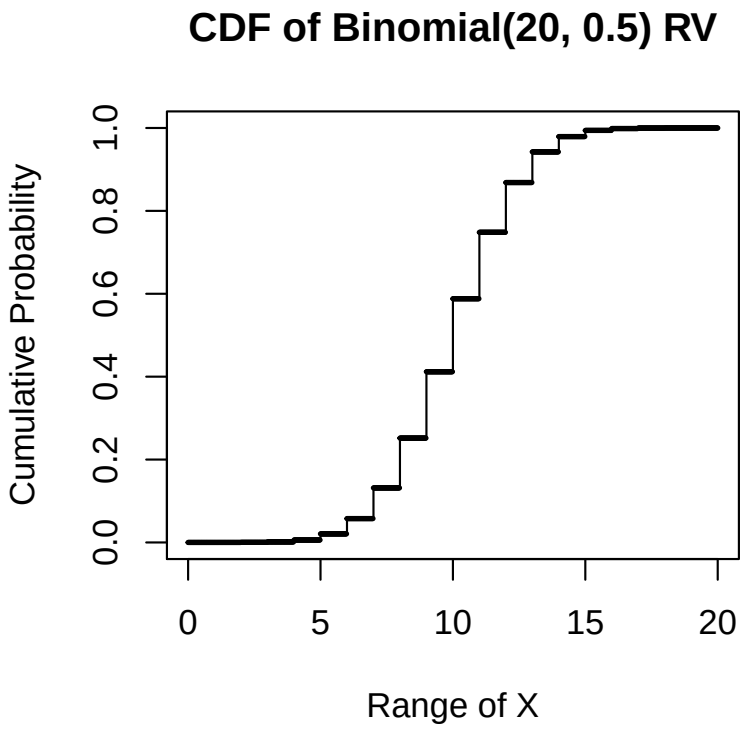
Throw a dart at a triangular board, whose corners we can place at $(0,0)$, $(0,3)$, and $(2,0)$. Let X be the x-coordinate of the point where the dart hits. Find the PDF of X.

More Graphs

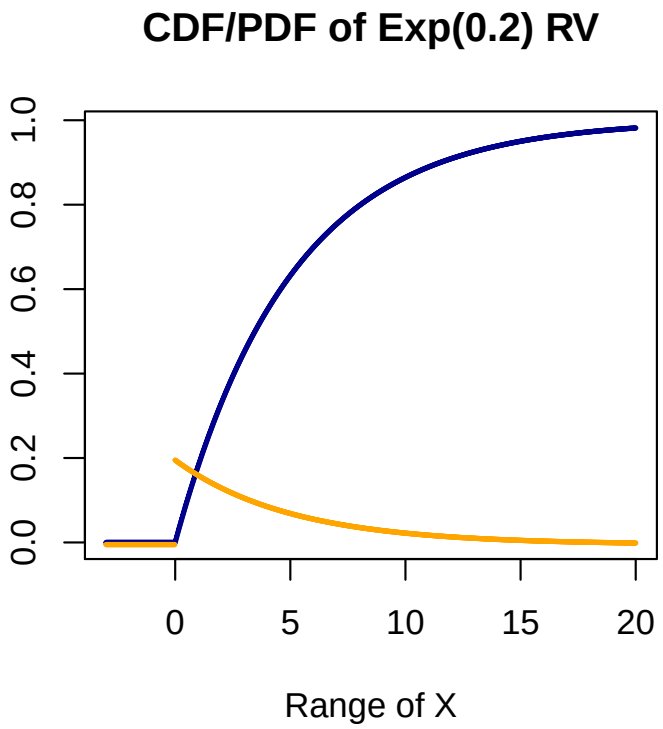
Binomial CDF and PMF



Binomial CDF connected by PMF



The CDF and PDF Plotted Together



A bit harder to interpret!

The Wrap Up

Summary

1. For continuous RVs we can use integrals of the PDF to compute probabilities.
2. The CDF is defined for all types of RVs and fully classifies its distribution.
3. For continuous RVs the CDF is often easier to use for probability computations.
4. To find the PDF of X, usually you should first find the CDF of X.

Next Step

The PMF and PDF give complete information about a random variable. The expectation and variance are quick summaries of the ran-

dom variable.

Expectation

Gregory M. Shinault

Goals for this Lecture

1. Define, use, and interpret the expectation of a random variable.
2. Learn the key properties that mathematical expectation possesses.

This material corresponds to section 3.3 of the textbook.

Basics of Expectation

Motivation

The definition of the expected value is motivated by the definition of the average.

Definition

For a discrete RV X the *expected value* is given by

$$\mathbb{E}X = \sum_{k \in \text{Ran}(X)} k \cdot \mathbb{P}(X = k).$$

For a continuous RV X the *expected value* is given by

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Example

Find the expected value of a Binomial(n, p) random variable.

Example

Let X be a continuous random variable with pdf

$$f_X(x) = \begin{cases} \frac{x^2}{3} & \text{for } -1 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

Find the expected value of X.

Example

Your dumb friend forgot his 8-digit PIN and decides to start guessing at random. What is the expected number of guesses he will make until he is correct?

This is really a computer security problem. If someone were to attack your system, you should be more interested in how long it would typically take to gain access than maximum possible.

History of the Problem of Points

1494: Proposed by Pacioli (an accountant)
~1550: Tartaglia realized his solution was flawed
1654: Resolved by Pascal and Fermat in a series of letters
1657: Huygens ran with their solution to write the first treatise on probability, *On Reasoning in Games of Chance*

Subsequently: Casinos use expectation to determine the fair price of games of chance, then set the price unfairly in their favor. Insurance companies pay actuaries to do essentially the same thing.

Example of the Problem of Points

You are playing a game with a friend in which the winner takes home the prize pot of $20 dollars. The game consists of independent rounds in which each player is equally likely to win. The winner of the game is the first to win 5 rounds. I have won 2 rounds, my opponent has won 3 rounds. The game is interrupted and we must divide the pot. What is a fair quantity to give to me?

C = cost to enter a game at a disadvantage

X = Net prize money for player with disadvantage

$$X = \begin{cases} 20 - C & \text{if player wins} \\ -C & \text{if player loses} \end{cases}$$

Fair game
$EX = 0$

Let p = prob player wins the game

$EX = (20-C)p + (-C)(1-p)$
$= 0$

$20p - C = 0 \quad C = 20p$

p = P(player wins 3 or 4 of remaining rounds)

$= \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) + \binom{4}{4}\left(\frac{1}{2}\right)^4$

$= \frac{5}{16} \qquad C = 6.25$

Properties of Expectation

Law of the Unconscious Statistician

Fact: For a real function $g(x)$ and discrete RV X we have

$Y = g(X)$

$$\mathbb{E}[g(X)] = \sum_{k \in \mathrm{Ran}(X)} g(k) \cdot \mathbb{P}(X = k).$$

For a real function $g(x)$ and continuous RV X we have

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

$EY = \sum_{j \in \mathrm{Ran}(Y)} j\, P(Y=j)$

$= \sum_{j \in \mathrm{Ran}(g(X))} j\, P(g(X)=j)$

$= \sum_{k \in \mathrm{Ran}(X)} g(k) P(X=k)$

Scaling of Expectation

The following fact is a corollary to the Law of the Unconscious Statistician.

Fact: For real constants a, b we have

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Example

Let X be a continuous RV with PDF

$$f_X(x) = \begin{cases} \frac{x^2}{3} & \text{for } -1 \le x \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

Define $Y = X^3$. Find $\mathbb{E}Y$.

St. Petersburg Paradox

Problem Statement

There is a boring casino game based on coin tossing. You flip a fair coin until you get heads. If you do this in k flips you win $\$2^k$. What is a fair price to play this game?

A Few Ideas

There are many proposed solutions to this paradox. The most practical (in my opinion) requires you to consider the number of games to be played and the bankrolls of the player and casino (which are finite).

Other good solutions discard the notion that expectation is the best determination of a fair price. See Feller's classic book on introductory probability for a good example.

Simulation of 100 Games

```
NumFlips <- rgeom(100, 0.5)+1
Winnings <- 2^NumFlips
mean(Winnings)
```

```
## [1] 5.08
```

Simulation of 1000 Games

```
    NumFlips <- rgeom(1000, 0.5)+1
    Winnings <- 2^NumFlips
    mean(Winnings)
```

```
## [1] 19.326
```

Simulation of 10000 Games

```
    NumFlips <- rgeom(10000, 0.5)+1
    Winnings <- 2^NumFlips
    mean(Winnings)
```

```
## [1] 18.3316
```

Simulation of 100000 Games

```
    NumFlips <- rgeom(100000, 0.5)+1
    Winnings <- 2^NumFlips
    mean(Winnings)
```

```
## [1] 26.20968
```

Simulation of 1000000 Games

```
    NumFlips <- rgeom(1000000, 0.5)+1
    Winnings <- 2^NumFlips
    mean(Winnings)
```

```
## [1] 27.048
```

The Wrap Up

Summary

1. Mathematical expectation is the theoretical foundation for the average of a randomly generated dataset.
2. The Law of the Unconscious Statistician significantly simplifies expected value calculations.
3. Expectation is used to determine fairness in situations involving randomness (gambling, insurance, etc.).

4. This is only helpful for a large number of repetitions of games of chance. If the game will only be played once, the most likely outcome is best to determine your appropriate course of action.

Variance and Standard Deviation

Gregory M. Shinault

Goals for this Lecture

1. Define and interpret the variance and standard deviation of a random variable.
2. Learn some techniques for computation of these quantities.
3. Learn a couple numerical properties.

This material corresponds to section 3.4 of the textbook.

Basic Definitions and Properties

Introduction

Consider the RVs X_1 outcomes spreadout wrt EX_1

$$X_1 = \begin{cases} 1000000 & \text{with probability } 0.50 \\ 0 & \text{with probability } 0.50, \end{cases}$$

$EX_1 = 1000000 \times 0.5 + 0 = 500000$

$$X_2 = \begin{cases} 499999 & \text{with probability } 1/3, \\ 500000 & \text{with probability } 1/3, \\ 500001 & \text{with probability } 1/3. \end{cases}$$

$EX_2 = \frac{499999 + 500000 + 500001}{3} = 500000$

different measure $E[(X-\mu)^4]$

Same expected value, wildly different behavior.

Exercise: Verify X_1 and X_2 have the same expected value.

$E[|X-\mu|]$

Variance

Definition How spread out (far away) from mean the RV X is distributed

The *variance* of a random variable X with mean $\mu = \mathbb{E}X$ is given by

$$\text{Var } X = \mathbb{E}[(X-\mu)^2].$$

(outcome − mean)$^{\text{square}}$ ⇒ positive distance; penalize those that are far

This is often denoted by σ_X^2.

Computation

Fact: The variance of X can be computed by the formula

$$\text{Var } X = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Exercise: Prove this fact.

prove continuous version

$$\text{Var}(X) = E[(X-\mu)^2]$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx$$

$$= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f_X(x)\,dx$$

$$= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx - 2\mu \int_{-\infty}^{\infty} x f_X(x)\,dx + \mu^2 \int_{-\infty}^{\infty} f_X(x)\,dx$$

$$\text{Var}(X) = EX^2 - 2\mu EX + \mu^2 \cdot 1$$

$$= EX^2 - 2\mu^2 + \mu^2$$

$$= EX^2 - \mu^2$$

Standard Deviation

Definition

The *standard deviation* of a random variable X is given by

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}\ X}.$$

This is often denoted by σ_X.

The standard deviation is a more meaningful measure of a random variables fluctuation, but the square root makes it more difficult to work with.

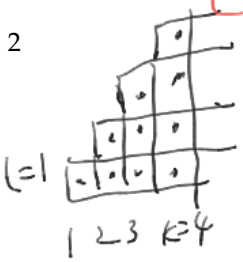
Example

Find the variance and standard deviation of a Geo(p) RV.

Example

Find the variance and standard deviation of a continuous RV with PDF

$$f_X(x) = \begin{cases} \frac{x^2}{3} & \text{for } -1 \le x \le 2 \\ 0 & \text{otherwise.} \end{cases}$$



Basic Properties

Scaling and Translation

Fact: For real numbers a, b we have

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X),$$

$$\mathrm{SD}(aX + b) = |a|\mathrm{SD}(X).$$

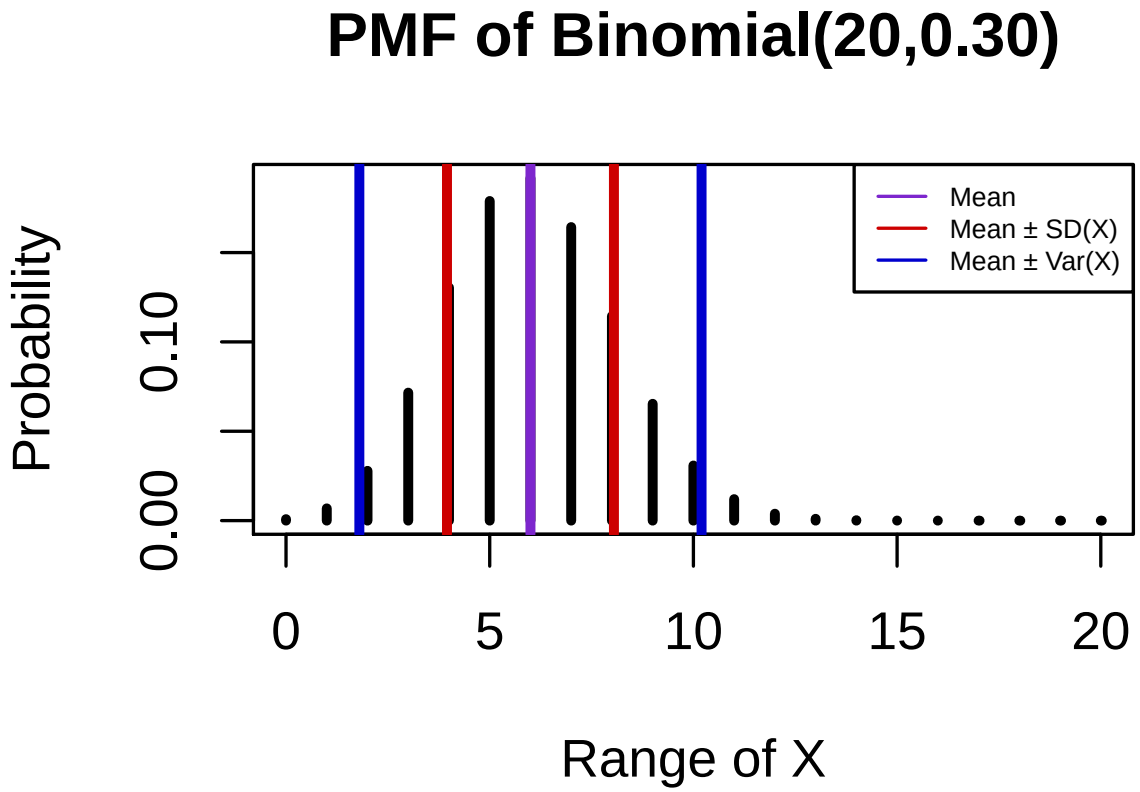
Exercise: Prove this fact.

Comparison of the Two

Binomial(20, 0.30)

Binomial(20, 0.30)

Let's compare $\mathbb{P}(\mu - \sigma_X < X \le \mu + \sigma_X)$ to $\mathbb{P}(\mu - \sigma_X^2 < X \le \mu + \sigma_X^2)$.


PMF of Binomial(20,0.30)
Probability
0.00
0.10
0
5
10
15
20
Range of X
Mean
Mean ± SD(X)
Mean ± Var(X)

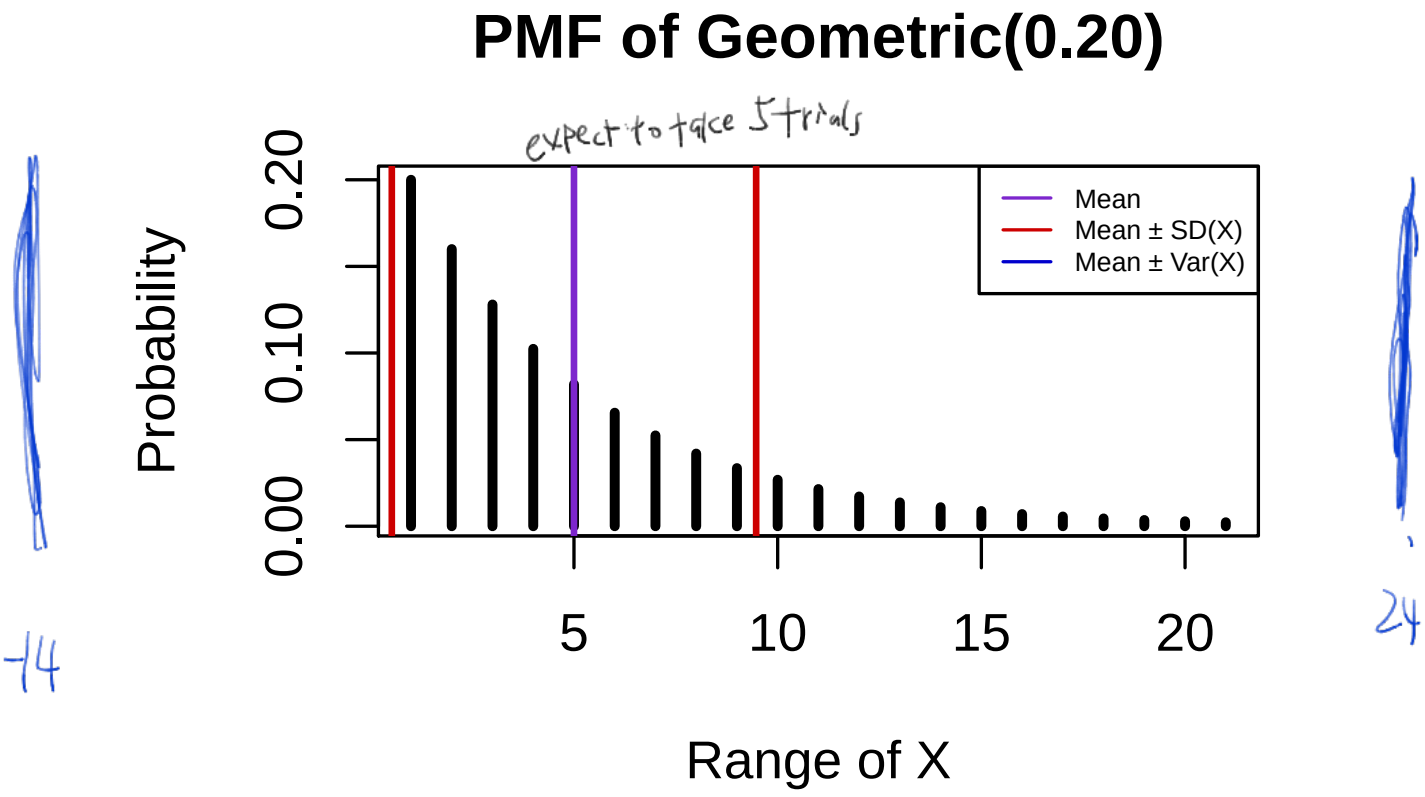
```
    pbinom(MeanX+sdX, size = n, prob = p)- pbinom(MeanX-sdX, size = n, prob = p)
```

```
## [1] 0.7795817
```

```
    pbinom(MeanX+VarX, size = n, prob = p)- pbinom(MeanX-VarX, size = n, prob = p)
```

```
## [1] 0.9752179
```

Geometric(0.20)



Geometric(0.20)

Let's compare $\mathbb{P}(\mu - \sigma_X < X \leq \mu + \sigma_X)$ to $\mathbb{P}(\mu - \sigma_X^2 < X \leq \mu + \sigma_X^2)$.

```
    p=0.20 ; MeanX = 1/p ; VarX <- (1-p)/(p^2) ; sdX <- sqrt(VarX)
    print(c(VarX, sdX))
```

```
## [1] 20.000000  4.472136
```

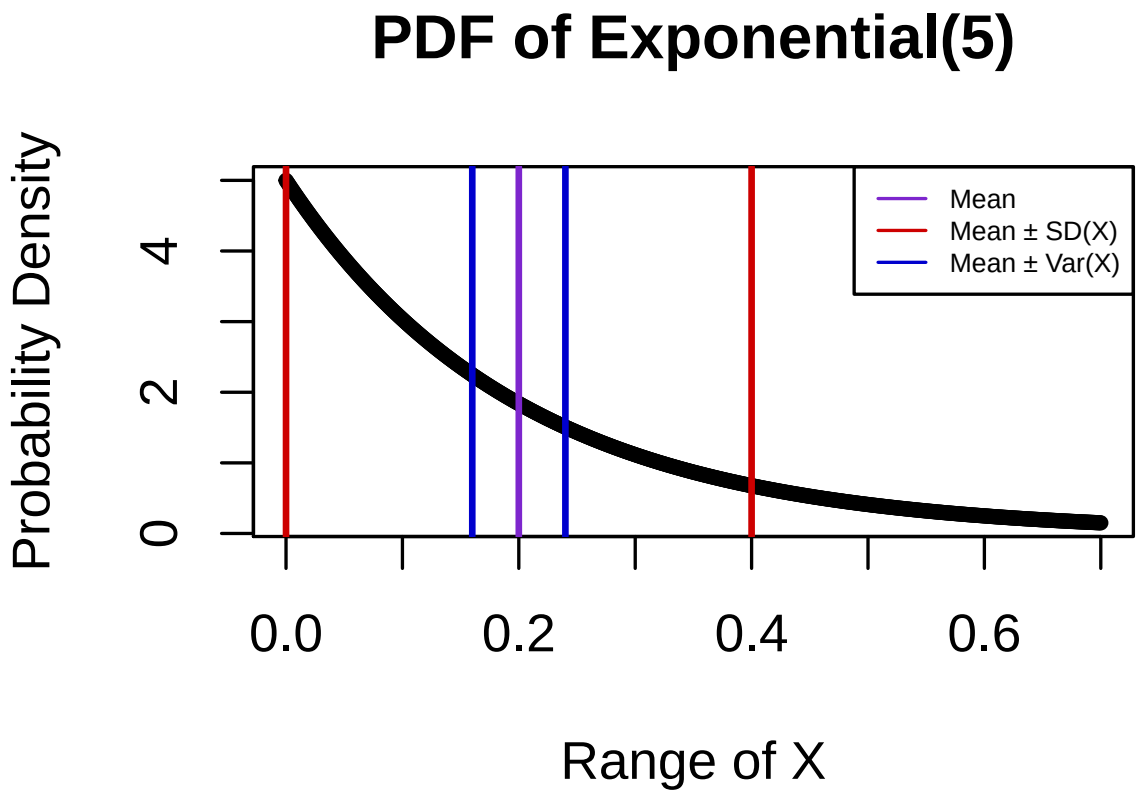
```
    pgeom(MeanX+sdX, prob = p) - pgeom(MeanX-sdX, prob = p)
```

```
## [1] 0.6926258
```

```
    pgeom(MeanX+VarX, prob = p) - pgeom(MeanX-VarX, prob = p)
```

```
## [1] 0.9969777
```

Exp(5)



Exp(5)

```
    Rate=5 ;MeanX = 1/Rate ; VarX <- 1/(Rate^2) ; sdX <- sqrt(VarX)
    print(c(VarX, sdX))
```

```
## [1] 0.04 0.20
```

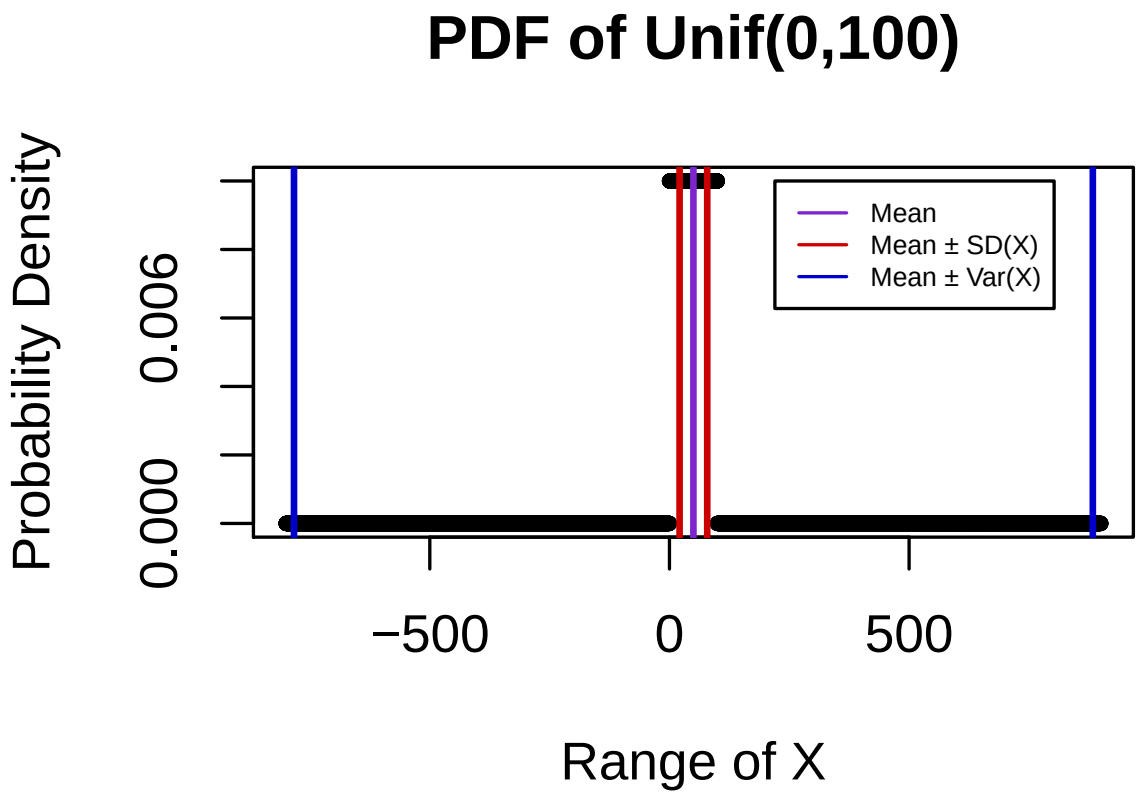
```
    pexp(MeanX+sdX, rate = Rate) - pexp(MeanX-sdX, rate = Rate)
```

```
## [1] 0.8646647
```

```
    pexp(MeanX+VarX, rate = Rate) - pexp(MeanX-VarX, rate = Rate)
```

```
## [1] 0.1481348
```

Unif(0,100)



The Wrap Up

Summary

1. Variance and standard deviation are used to measure the spread of a RV from its mean.
2. Compute variance using $\mathbb{E}X^2 - \mu^2$.
3. Remember the scaling laws, which tell you standard deviation is the more natural, but frustrating, measure of spread.

Next Step

Now we have all the tools necessary to define the most important continuous distribution in classical probability theory, the Normal distribution.

The Normal Distribution

Gregory M. Shinault

Goals for this Lecture

1. Define the standard normal/Gaussian distribution, and the general normal/Gaussian distribution. This is the most important distribution in classical probability.
2. Learn the nice properties that the standard normal distribution possesses.
3. Learn how to compute probabilities for a general normal distribution from a table of standard normal values.

This material corresponds to section 3.5 of the textbook.

Introduction

The bell curve is known more formerly as the normal distribution or the Gaussian distribution. Today we present some facts about it, with no information about where it comes from.

Standard Normal Distribution

Definition

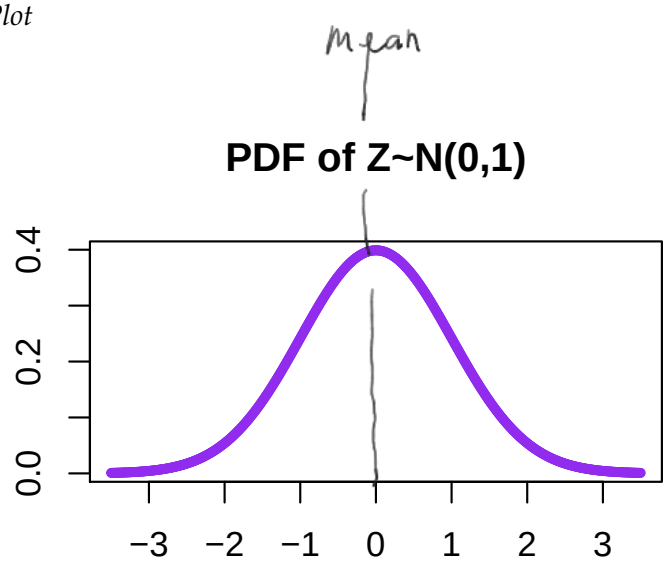
A continuous RV Z is called *standard normal* if it has the PDF

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for all real x. We denote this by $Z \sim N(0,1)$.

The CDF a standard normal RV is denoted by $\Phi(x) = \mathbb{P}(Z \leq x)$.

PDF Plot



decay rapidly

$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ ① $\frac{1}{\sqrt{2\pi}}>0$ $e^{-x^2/2}>0 \Rightarrow \varphi(x)>0 \ \forall x$

② $\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx=I$ $\quad I^2=\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx\right)^2$

$=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy$

$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}dx\,dy$

$=\int_0^{2\pi}\int_0^{\infty}\frac{1}{2\pi}e^{-\frac{r^2}{2}}\cdot r\,dr\,d\theta$

$=\frac{1}{2\pi}\int_0^{2\pi}\left[e^{-r^2/2}\right]_0^{\infty}d\theta$

$=\frac{1}{2\pi}\int_0^{2\pi}[0-(-1)]\,d\theta$

$=1$

$I^2=1 \Rightarrow I=1$ or -1

$\Rightarrow I=1$ bcz $\varphi(x)>0$.

Example

Verify the the standard normal density is a PDF.

Find $EZ=0$ $\quad Var=1$

Example

Find the mean and variance of a standard normal RV.

$EZ=\int_{-\infty}^{\infty}x\varphi(x)dx$ $\quad x^2+y^2=r^2$ $\quad dxdy=r\,dr\,d\theta$ Fubini's Thm

$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}xe^{-x^2/2}dx$

$=\frac{1}{\sqrt{2\pi}}\left[e^{-x^2/2}\right]_{-\infty}^{\infty}=\frac{1}{\sqrt{2\pi}}\left[\lim_{a\to\infty}\left(e^{-a^2/2}\right)-\lim_{b\to-\infty}\left(e^{-b^2/2}\right)\right]$ Switch to polar coordinate

$=\frac{1}{\sqrt{2\pi}}[0-0]=0$

Algebra of Standard Normal CDF

Fact:

1. $\Phi(x)=1-\Phi(-x)$.
2. $\Phi(x)-\Phi(-x)=2\Phi(x)-1$.

Exercise: Prove these facts.

$\varphi(-x)=1-\varphi(x)$ symmetric $\quad \varphi(x)-\varphi(-x)=\varphi(x)-(1-\varphi(x))=2\varphi(x)-1$

General Normal Distribution

Definition

The standard normal distribution has very nice properties and its shape is useful, but most observables in application will not have

mean 0 and variance 1.

A random variable X is Normal(μ, σ^2) if it is defined by

$$X = \mu + \sigma Z.$$

Alternate Definition: A random variable X is Normal(μ, σ^2) if it has the PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

for all x.

Example

Prove that $X \sim$ Normal(μ, σ^2) has the PDF given in the alternate definition.

Example

The height of *The Average American Male* is 70 inches.
The standard deviation of American males' height is 3 inches.
We choose an adult American male at random.
Assuming heights are normally distributed (this is approximately true), what is the probability a randomly chosen man is between 68 and 73 inches tall?

Solution using R

```
pnorm((73-70)/3) - pnorm((68-70)/3)
```

```
## [1] 0.5888522
```

```
pnorm(73, mean=70, sd=3) - pnorm(68, mean=70, sd=3)
```

```
## [1] 0.5888522
```

Normally Distributed Data

The normal distribution is important because of there are countless types of data that are normally distributed, or can at least be accurately modelled by the normal distribution.

Student exam grades, heights, lumosity of stars, the modulus squared of the wave function for a quantum harmonic oscillator at its lowest energy level, some growth rates in finance, error analysis, etc.

The Wrap Up

Summary

1. The standard normal distribution is denoted by $N(0,1)$, usually as a RV Z. Its PDF is $f_Z(z) = e^{-z^2/2}/\sqrt{2\pi}$,

$$\mathbb{E}Z = 0, \quad \text{Var}(Z) = 1.$$

2. The more general normal distribution with mean μ and variance σ^2 can be obtained by $X = \mu + \sigma Z$.
3. This is the most important continuous distribution in classical statistics.

Next Step

Now that we have introduced the Gaussian distribution, we are going to see where it actually comes from.

The Normal Approximation

Gregory M. Shinault

Overview

Goals for this Lecture

1. Learn how to use the Normal distribution to approximate the Binomial distribution, and when it is useful.
2. Learn how to derive the standard formula for confidence intervals of proportions.
3. Learn how to use the *continuity correction* to make a more accurate approximation.

This material corresponds to section 4.1-4.3 of the textbook.

Introduction

The binomial distribution is incredibly useful in applications. Today we learn to approximate it with the normal distribution for several reasons.

1. Binomial coefficients can be computationally expensive (and unstable).
2. Limit theorems are where we see universal behavior arise from randomness.

Big Idea

binomial dist converges to normal dist

Suppose $X \sim \text{Bin}(n, p)$.
Set $\mu = \mathbb{E}X = np$ and $\sigma^2 = \text{Var}(X) = np(1-p)$.

If n is a large number then

$$X \overset{d}{\approx} Y \sim N(\mu, \sigma^2).$$

Alternately,

$$\frac{X - \mathbb{E}X}{SD(X)} \overset{d}{\approx} Z \sim N(0,1).$$

approximately standard normal

$X \overset{d}{\approx} Y = \mu + \sigma Z \iff \frac{X-\mu}{\sigma} \approx Z$

$\Updownarrow$

$P(a \le X \le b) \approx P(a \le Y \le b)$

$\Downarrow$

$P(a \le X \le b) \simeq P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$

$\simeq \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

Formal Statement

de Moivre-Laplace Theorem (1738): Suppose $S_n \sim \text{Bin}(n, p)$ is a sequence of random variables for a fixed value of p.

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \Phi(b) - \Phi(a).$$

Typical Use

Generally we will have $X \sim \text{Bin}(n, p)$ and want to compute $\mathbb{P}(k \leq X \leq \ell)$.

$$\begin{aligned} \mathbb{P}(k \leq X \leq \ell) &= \mathbb{P}\left(\frac{k-np}{\sqrt{np(1-p)}} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq \frac{\ell-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{\ell-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Proof

Interactive Explanation

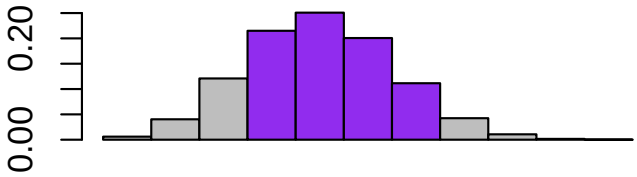
I wrote a simple app to illustrate how closely the Normal distribution fits the Binomial distribution: `http://shinault.shinyapps.io/BinomialApprox`

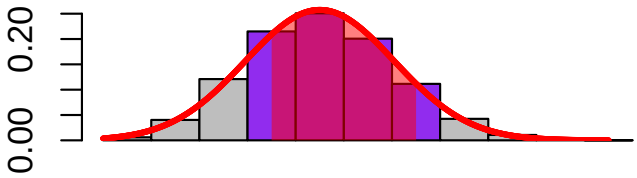
You can experiment with a web app to see how well the Gaussian PDF fits the Binomial PMF for various choices of the parameters n and p.

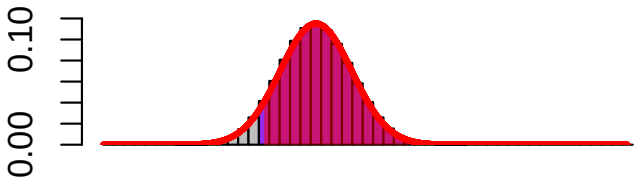
Proof Idea

Use Stirling's formula $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ and Riemann sums.

This is a limit theorem, so the primary mathematical ideas used are from analysis (Math 421 or Math 521). We will not cover the formal proof in lecture. Instead, we will look at some pictures to illustrate the ideas of the proof.


0.00
0.20


0.00
0.20



Practical Examples

When is n big enough?

Rule of Thumb: If $Var(X) = np(1-p) > 10$, the normal approximation to the binomial distribution should be fairly accurate.

Example

This is Wisconsin, so we ask 100 people if the prefer mozzarella sticks or cheese curds. The probability a random person will prefer ~~cheese curds~~ is 0.6. Approximate the probability that at least 70 people prefer ~~mozzarella sticks~~. cheese curds

Let X be # people surveyed who prefer Cheese curds is the normal approximation valid 69 more accurate than 70

$\Rightarrow X \sim Bin(100, 0.6)$

moz Want $P(X \geq 70) = P(70 \leq X \leq 100)$

$= P(69 < X \leq 100)$

$Var X = 100 \cdot 0.6 \cdot 0.4 = 24 > 10$

$P(69 < X \leq 100) \approx P\left(\frac{69-60}{\sqrt{24}} < Z \leq \frac{100-60}{\sqrt{24}}\right)$

$= \Phi\left(\frac{40}{\sqrt{24}}\right) - \Phi\left(\frac{9}{\sqrt{24}}\right)$

≈ 0.248

Solution in R

```
SampleSize <- 100; Prob <- 0.6;
MeanX <- SampleSize*Prob; SDX <- sqrt(SampleSize*Prob*(1-Prob))
Actual <- pbinom(100, size = SampleSize, prob = Prob) -
            pbinom(69, size = SampleSize, prob = Prob)
Approx <- pnorm((100-MeanX)/SDX)-pnorm((70-MeanX)/SDX)
c(Actual, Approx)
```

```
## [1] 0.02478282 0.02061342
```

Example (Confidence intervals)

Suppose we want to predict the next election. To do so we take a public opinion poll and let $\hat{p}$ denote the fraction of people who will vote for the Democratic candidate.
How many people must we survey in order to be 95% certain that $\hat{p}$ will be within 0.005 of the fraction of people who will vote for the Democratic candidate for the whole population, p?

X : number of pp surveyed who who say they will vote democrat

$\Rightarrow X \sim Bin(n, p)$

Want to estimate p

prob they will vote demo

p is unknown

n : # pp survey

estimator $\hat{p} = \frac{X}{n}$

want n big enough

Want $P(|\hat{p} - p| < 0.005) \geq 0.95$

we must choose n big enough so that we are confident about it

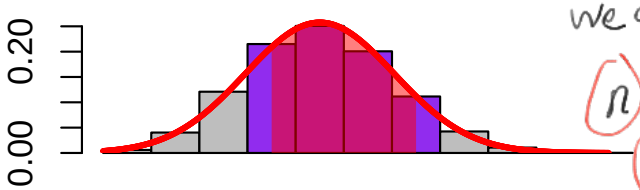
n: sample size

$\varepsilon = 0.005$ acceptable margin of error

$C = 0.95$ confidence level

Continuity Correction

Missing The Right and Left Edges of Rectangles!



by inequality 不知道是什麼

$P\left(-\varepsilon < \frac{X}{n} - p < \varepsilon\right) \geq C$

$P(np - n\varepsilon < X < np + n\varepsilon) \geq C$

$P\left(\frac{-n\varepsilon}{\sqrt{np(1-p)}} < Z < \frac{n\varepsilon}{\sqrt{np(1-p)}}\right) \geq C$

use $P(|\hat{p} - p| < \varepsilon) \geq C$

with $X \sim Bin(n, p)$ to get how large sample size should be

$\Phi\left(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{-\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) \geq C$

$n \geq \max_{0 \leq p \leq 1}\{p(1-p)\} \frac{\Phi^{-1}(0.975)^2}{0.005^2}$

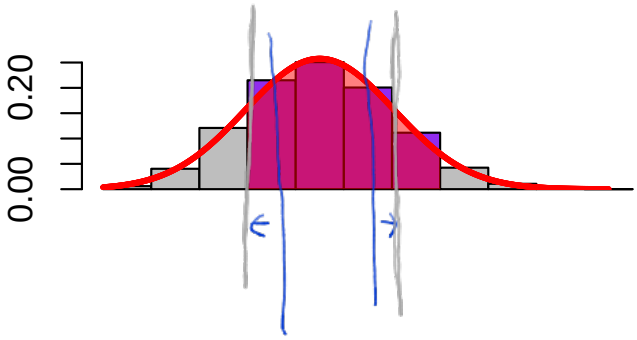
$\geq 2\Phi\left(\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}\right) - 1 \geq C$, $\sqrt{n} \geq \sqrt{p(1-p)}\,\Phi^{-1}\left(\frac{1+C}{2}\right)$

$n \geq p(1-p)\frac{\Phi^{-1}\left(\frac{1+C}{2}\right)^2}{\varepsilon^2}$

$p = \arg\max_p (p(1-p))$

$p = \frac{1}{2}$

The Fix: Extend the Integral



Formula for Continuity Correction

For integers k, ℓ and $X \sim \text{Bin}(n, p)$, we can use the improved approximation

More accurate approximation

$$\begin{aligned} \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(a - 1/2 \leq X \leq b + 1/2) \\ &\approx \Phi\left(\frac{b+1/2-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-1/2-np}{\sqrt{np(1-p)}}\right). \end{aligned}$$

Example

This is Wisconsin, so we ask 100 people if the prefer mozzarella sticks or cheese curds. The probability a random person will prefer cheese curds is 0.6. Approximate the probability that at least 70 people prefer mozzarella sticks, using the continuity correction this time.

X: # people prefer cheese curds

$\Rightarrow X \sim \text{Bin}(100, 0.6)$ $\mu = 100 \cdot 0.6 = 60$ $\sigma^2 = 60 \times 0.4 = 24$

cheese

Solution in R

want to compute $P(X \geq 70) = P(70 \leq X \leq 100) = P(69.5 \leq X \leq 100.5)$

$\approx \Phi\left(\frac{100.5-60}{\sqrt{24}}\right) - \Phi\left(\frac{69.5-60}{\sqrt{24}}\right)$

≈ 0.026

```
SampleSize <- 100; Prob <- 0.6;
MeanX <- SampleSize*Prob; SDX <- sqrt(SampleSize*Prob*(1-Prob))
Actual <- pbinom(100, size = SampleSize, prob = Prob) -
            pbinom(69, size = SampleSize, prob = Prob)
Approx <- pnorm((100-MeanX)/SDX)-pnorm((70-MeanX)/SDX)
ApproxImp <- pnorm((100+0.5-MeanX)/SDX)-pnorm((70-0.5-MeanX)/SDX)
c(Actual, Approx, ApproxImp)
```

```
## [1] 0.02478282 0.02061342 0.02623975
```

Example

We buy 2 gross of eggs for the Lion's Club pancake breakfast. In most packages the probability each egg is broken is 0.05. What is the approximate probability that less than 10 of our eggs are broken? Be certain to use the continuity correction.

Solution in R

```
SampleSize <- 2*144; Prob <- 0.05;
MeanX <- SampleSize*Prob; SDX <- sqrt(SampleSize*Prob*(1-Prob))
Actual <- pbinom(9, size = SampleSize, prob = Prob)
Approx <- pnorm((9-MeanX)/SDX)-pnorm((0-MeanX)/SDX)
ApproxCor <- pnorm((9+0.5-MeanX)/SDX)-pnorm((0-0.5-MeanX)/SDX)
c(Actual, Approx, ApproxCor)
```

```
## [1] 0.08619932 0.07209659 0.09258931
```

The Wrap Up

Summary

1. If n is big, then $\text{Bin}(n, p)$ is approximately the same distribution as $N(np, np(1-p))$.
2. Our rule of thumb for using this approximation is that $\text{Var}(X) > 10$.
3. The continuity correction can increase the accuracy significantly. However, you are only required to use it if explicitly requested in this course.

Next Step

What do we do if the rule of thumb says not to use the normal approximation?

Finer Points

The proof of the de Moivre-Laplace Theorem is provided in your textbook. I strongly recommend reading it.

Just how accurate is the normal approximation? If you are interested in precise statements about the error in this approximation, look up the Berry-Esseen inequality.

The Poisson Approximation

Gregory M. Shinault

Introduction

Reminder

If $Var(X) = np(1-p) > 10$, the normal approximation to the binomial distribution should be fairly accurate.

What can we do if n is large, but p makes this condition fail?

Goals for this Lecture

1. Derive the correct probability mass function for approximating the binomial distribution for small success probabilities.
2. Define the Poisson(λ) distribution.
3. Learn how to use the Poisson approximation, and when it is appropriate to use.

This material corresponds to section 4.4 of the textbook.

Poisson Distribution

Definition: We say that X is a Poisson(λ) RV if it has PMF

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

for $k = 0, 1, 2, \ldots$. We denote this by $X \sim \text{Pois}(\lambda)$.

Fact

If $X \sim \text{Pois}(\lambda)$ then $\mathbb{E}X = \lambda$ and $\text{Var}(X) = \lambda$.

Exercise: Prove this fact.

Poisson Approximation

Big Idea

Suppose $X \sim \text{Bin}(n, p)$. If $p \ll 1$ then $X \overset{d}{\approx} Y \sim \text{Pois}(np)$.

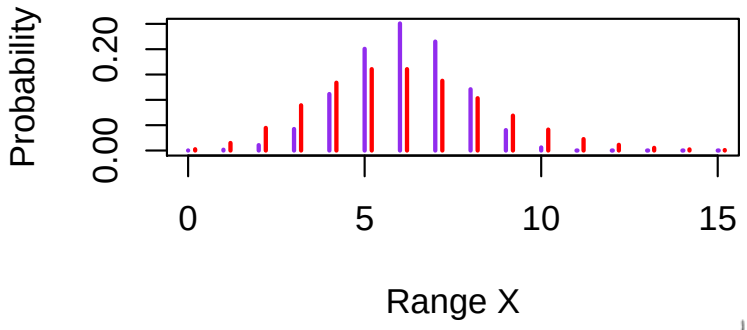
Formal Statement

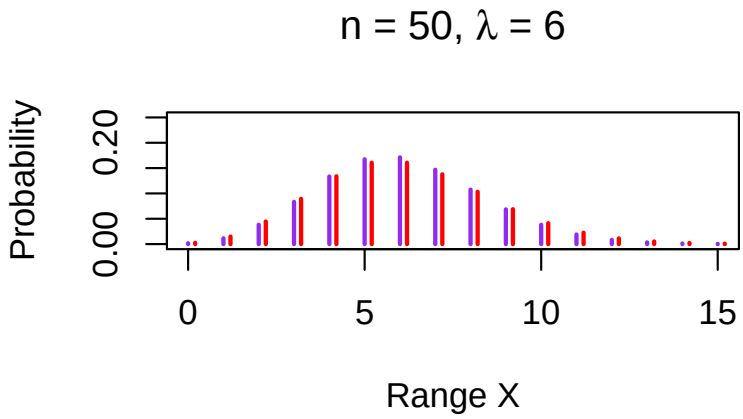
Fact: Let $\lambda > 0$, set $p_n = \lambda/n$. Suppose $S_n \sim \text{Bin}(n, p_n)$ (so that $\mathbb{E}S_n = \lambda$). Then

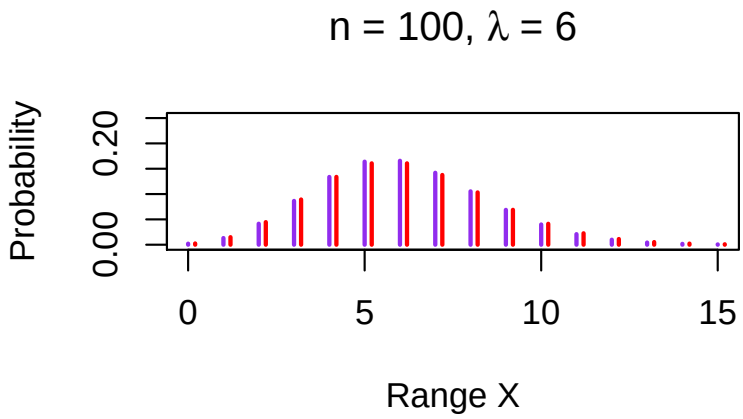
$$\lim_{n\to\infty} \mathbb{P}(S_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

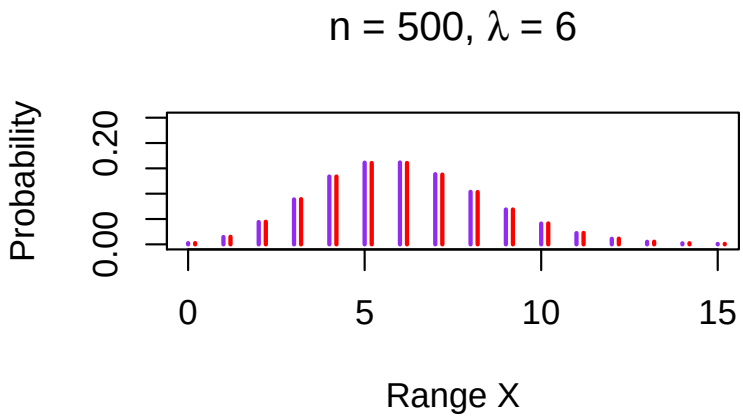
Proof Picture

We can look at some graphs comparing the Binomial and Poisson PMFs to get a feeling for when the approximation is useful.









n↑ p↓ accuracy↑

Practical Considerations and Examples

Example

The probability of dying by horse kick in the Prussian cavalry in the late 19th century is $\frac{1}{10000}$. There are 150,000 soldiers in the cavalry.

What is the probability that more than 20 people die by horse kick in 1893? (Is normal approximation appropriate? Why not?)

X = #l people die by horse kick in 1893

$X \sim$ Binomial $(150000, \frac{1}{10000})$

$\text{Var}\, X = 150000 \times \frac{1}{10000} \times \frac{9999}{10000} \approx 15 > 10$ Normal should be alright

$p = \frac{1}{10000} \ll 1$ possion alright $Y \sim \text{possion}(15)$

$P(X>20) = 1 - P(X \leq 20) = 1 - P(Y \leq 20) = 1 - \sum_{i=0}^{20} e^{-15} \cdot \frac{15^i}{i!}$

Solution in R

```
SampleSize <- 150000; Prob <- 0.0001; MeanX <- SampleSize*Prob
Actual <- 1-pbinom(20, size = SampleSize, prob = Prob)
Approx <- 1-ppois(20, lambda = MeanX)
ApproxNorm <- 1-pnorm(20, mean=MeanX, sd= sqrt(MeanX*(1-Prob)))
c(Actual, Approx, ApproxNorm)
```

```
## [1] 0.08296046 0.08297091 0.09834161
```

Binomial is great but poisson needs only 1 parameter (variance)

It's good when success probability population information is not sufficient

Example

We have a weekly card game with a friend, in which we play 50 hands of poker. Our friend is a cheater and deals the cards so that the probability of getting a 2 of a kind more than 5 times in a week is 0.01. We have this card game every week for 5 years. What is the approximate probability that in at least 2 weeks we get 2 of a kind more than 5 times?

X = # weeks in which A occurs where A = get 2 of a kind more than 5 ≥ 2

$X \sim \text{Binomial}(5 \cdot 52, P(A))$

$X \sim \text{Binomial}(260, 0.01)$

$Y \sim \text{Poisson}(2.6)$

Want $P(X \geq 2) = 1 - P(X < 2)$

$= 1 - P(Y < 2)$

$= 1 - P(Y=0) - P(Y=1)$

$= 1 \cdot e^{-2.6} \frac{2.6^0}{0!} - e^{-2.6} \frac{2.6^1}{1!}$

≈ 0.7526

Solution in R

```
SampleSize <- 5*52; Prob <- 0.01; MeanX <- SampleSize*Prob
Actual <- 1-pbinom(1, size = SampleSize, prob = Prob)
Approx <- 1-ppois(1, lambda = MeanX)
ApproxNorm <- 1-pnorm(1, mean=MeanX, sd= sqrt(MeanX*(1-Prob)))
c(Actual, Approx, ApproxNorm)
```

```
## [1] 0.7341664 0.7326151 0.8406849
```

Error Bound

Fact: Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Pois}(np)$. Then for any a, b

$$|\mathbb{P}(a \leq X \leq b) - \mathbb{P}(a \leq Y \leq b)| \leq np^2.$$

difference in estimate $\leq np^2$

Note no dependence on a, b!

Proof is a bit advanced for this course.

Applying the Error Bound

Note this explains why the previous example is well approximated by the Poisson distribution:

$$\text{Var}(X) = 150000 \cdot 0.0001 \cdot 0.9999 = 14.9985 > 10,$$

so normal approximation is probably pretty good. However,

$$|\mathbb{P}(X \geq 21) - \mathbb{P}(Y \geq 21)| \leq 150000 \cdot (0.0001)^2 = 0.0015.$$

So Poisson approximation will be great.

Lesson: If p is really tiny, Poisson approximation will be excellent.

Note on Applications

The Poisson distribution is used for many applications largely due to the approximation theorem.

However, it has the fairly strong condition that

$$\mathbb{E}X = \mathrm{Var}X.$$

The Wrap Up

Summary

1. The Poisson(λ) distribution has PMF

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

   for $k = 0, 1, 2, \ldots$.
2. If p is tiny, then Bin(n, p) is approximately the same distribution as Pois(np).
3. Sometimes Poisson approximation is better than normal approximation even when $\mathrm{Var}(X) > 10$.

Next Step

We have seen how the binomial distribution can be approximated by the normal distribution, which is continuous. Is there something similar we can do for the geometric distribution?

The Exponential Distribution

Gregory M. Shinault

Motivation

Remember that a Geo(p) RV is representative of the waiting time for the first success in a sequence of Bernoulli trials. This is a discrete waiting time.

There must be a distribution that is representative of a continuous waiting time. We arrive at this distribution by taking the limit of the geometric distribution.

Goals for this Lecture

1. See how we can use a limit to pass from a discrete distribution to a continuous distribution.
2. Define the Exponential(λ) distribution.
3. Learn how to perform some computations for the exponential distribution.

This material corresponds to section 4.5 of the textbook.

Derivation

Setup

Suppose we need to model a continuous waiting time and we know the average waiting time is $1/\lambda$.

We can approximate this in discrete time.

$$X_n \sim \text{Geo}(\lambda/n)$$

$$T_n = \frac{X_n}{n} \quad \Rightarrow \quad \mathbb{E}T_n = \frac{\mathbb{E}X_n}{n} = \frac{1}{\lambda}$$

So T_n subdivides the intervals into units of $1/n$, while the expected waiting time is always $1/\lambda$.

Limit

Fact: Let $t > 0$. Then

$$\lim_{n\to\infty} \mathbb{P}(T_n > t) = e^{-\lambda t}.$$

Definition

A continuous RV X is said to have the *exponential distribution* with *rate* λ if it has the CDF

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

We denote this by $X \sim \text{Exp}(\lambda)$.

Alternate Definition

A continuous RV X is said to have the *exponential distribution* with *rate* λ if it has the PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that the exponential distribution has only one parameter, λ.

Properties of the Exponential Distribution

Example

Find the mean and variance of the exponential distribution.

Example

Let $X \sim \text{Exp}(\lambda)$. Find a general expression for $\mathbb{P}(a \leq X \leq b)$ where $a, b > 0$.

Memoryless Property

Fact: Let $X \sim \text{Exp}(\lambda)$. Then for $s,t > 0$ we have

$$\mathbb{P}(X > t + s \mid X > t) = \mathbb{P}(X > s).$$

Exercise: Prove this fact.

Radioactive Decay

Setup

Unstable isotopes eventually emit an α-particle (two protons and two neutrons) or a β-particle (one electron) thus making it a different isotope.

This is of interest in a probability course because the waiting time for the emission of the particle is a random variable. Specifically, it is an exponential random variable.

If we are going to model this as an exponential random variable how can we find the rate λ?

Getting the Rate from Half-life

$T \sim Exp(\lambda)$ (λ: rate) $P(T < 5730) = \frac{1}{2}$

$1 - e^{-\lambda \cdot 5730} = \frac{1}{2}$

$\frac{1}{2} = e^{-\lambda 5730}$

$\log \frac{1}{2} = -\lambda 5730$

$\lambda = -\frac{\log \frac{1}{2}}{5730} = \frac{\log 2}{5730}$

Most decays are described in terms of their half-life. This is defined in terms of a large number of atoms. For example, radiocarbon (carbon-14) has a half-life of 5730 years. It emits a β-particle and becomes nitrogen-14 which is a stable isotope. This means that if we start out with 8 grams of carbon-14, 5730 years later we can expect to be left with 4 grams of carbon-14. For individual atoms this is $\mathbb{P}(T < 5730) = .5$. From this fact, how do we recover the rate λ?

Longer Chains

Many of the more important isotopes do not decay directly into a stable isotope. There is a long chain of decays, all having different decay rates. Often times it is important to have information about the intermediate states of decay. The first four steps in the decay chain for Uranium-235 is as follows:

multiple exponential decay processes

$$U^{235} \xrightarrow[\alpha]{7E8a} Th^{231} \xrightarrow[\beta]{1.06d} Pa^{231} \xrightarrow[\alpha]{3.3E4a} Ac^{227} \xrightarrow[\beta]{21.8a} Th^{227}$$

Let $T_{U^{235} \to Th^{227}}$ be the time at which an isotope originating as U^{235} decays into Th^{227}.

$T_{U^{235} \to Th^{227}} \sim$ Hyperexponential distribution

How do we model this radioactive decay using exponential RVs?

$T_{A \to B}$: Decay from $A \to B$ $\sim Exp(\lambda)_{A \to B}$

$T_{U^{235} \to Th^{227}} = T_{U^{235} \to Th^{231}} + \ldots + T_{Ac^{227} \to Th^{227}}$ Linear independent

Summary

The Key Facts

1. The $\text{Exp}(\lambda)$ distribution is continuous with PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

   We call the parameter $\lambda > 0$ the *rate* of the distribution.
2. The $\text{Exp}(\lambda)$ distribution is generally used to model a waiting time, or survival time.
3. The $\text{Exp}(\lambda)$ distribution is the continuous version of the geometric distribution.

Exp distribution is a limit of geom

Moment Generating Functions

Gregory M. Shinault

Motivation

Reminder

$\mathbb{E}X$ tells us the center of X

If we know $\mathbb{E}X^2$ we can find $\text{Var}(X)$ which tells us the spread of X.

A Couple Steps Further

We can go up to EX^3 to find

$$\text{Skew}(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

and EX^4 to find

$$\text{Kurt}(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$

how sharp the peak is

These measure the skewness and peakedness of the distribution of X.

If we know all the moments $\mathbb{E}X^n$ of X, does this fully describe the distribution of X? In many cases, yes.

Goals for this Lecture

1. Define the *moment generating function* for any random variable.
2. See exactly how the moment generating function generates moments.
3. Use the MGF theory to easily solve some otherwise difficult problems.

This material corresponds to section 5.1 of the textbook.

Theory

Definition

The *moment generating function* (MGF) of a RV X is defined as

$$M_X(t) = \mathbb{E}e^{tX}.$$

$E[e^{tX}]$ t : any real #

Ex. $n=1$: $\frac{d}{dt}[M_X(t)] = \frac{d}{dt}[Ee^{tX}] = E\left[\frac{d}{dt}[e^{tX}]\right] = E[Xe^{tX}] = M'(t)$

Key Fact I

$M'(0) = E[Xe^0] = EX$ | In general LHS $= E\left[\frac{d^n}{t^n}e^{tX}\right] = E[X^n e^{tX}] = M^{(n)}(t)$

The moment generating function generates moments. That is,

take nth derivative get nth moment $\frac{d^n}{dt^n}[M_X(t)]_{t=0} = \mathbb{E}X^n$.

$M^n(0) = E[X^n \cdot e^{0X}] = EX^n$

Key Fact II

Suppose there exists some $\delta > 0$ such that $M_X(t) = M_Y(t)$ for all $-\delta < t < \delta$. Then $X \stackrel{d}{=} Y$. have the same distribution (same CDF) For us this means they have the same density or mass function

Examples flip a coin 5 times. X = #heads Y = #tails $X \stackrel{d}{=} Y \sim$ Binomial$(5, \frac{1}{2})$ they are

Example but $P(X=Y) = 0$ Same distribution doesn't nessarily means same r.v

Find the MGF of $X \sim \text{Poisson}(\lambda)$. Use it to find $\text{Var}(X)$

Example $M_X(t) = E[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} \cdot P(X=k) = \sum_{k=0}^{\infty} e^{tk} \cdot e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}$

Suppose X has the MGF $M'_X(t) = e^{\lambda(e^t-1)}\cdot\lambda e^t$ $M''_X(t) = e^{\lambda(e^t-1)}\lambda e^t\lambda e^t + \lambda e^t \cdot e^{\lambda(e^t-1)}$

$$M_X(t) = 0.2 + 0.1e^t + 0.2e^{2t} + 0.3e^{4t} + 0.2e^{7t}.$$

$\text{Var}(X) = EX^2 - (EX)^2 = \lambda$

Find the PMF of X.

$M_X(t) = P_X(0)e^{t\cdot 0} + P_X(1)e^{t\cdot 1} + \dots P_X(7)e^{t\cdot 7}$

$EX = M'_X(0) = \lambda$

$EX^2 = M_X''(0) = \lambda^2 + \lambda$

Example

Find the MGF of $X \sim \text{Exp}(\lambda)$. Use it to find $\text{Var}(X)$

$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx = \int_0^{\infty} e^{tx}\cdot\lambda e^{-\lambda x}\,dx = \lambda\int_0^{\infty} e^{(t-\lambda)x}\,dx$

The Wrap Up

we requires $t-\lambda < 0$ $t < \lambda$ so the integral is finite

Summary

1. $M_X(t) = \mathbb{E}e^{tX}$
2. $M_X^{(n)}(0) = \mathbb{E}X^n$
3. If $M_X(t) = M_Y(t)$ then $X \stackrel{d}{=} Y$ (with some restrictions).

$\lambda\left[\frac{1}{t-\lambda}e^{(t-\lambda)x}\right]_{x=0}^{\infty} = \lambda\left[\frac{1}{t-\lambda}\cdot 0 - \frac{1}{t-\lambda}e^0\right] = \frac{\lambda}{\lambda-t} = M_X(t)$ for $t<\lambda$

$M_X(t) = \lambda(\lambda-t)^{-1}$

$M'_X(t) = \lambda[-(\lambda-t)^{-2}](-1) = \lambda(\lambda-t)^{-2}$

$M''_X(t) = 2\lambda(\lambda-t)^{-3}$

$\text{Var}(X) = M''_X(0) - (M'_X(0))^2$

The Distribution of a Function of a RV

Gregory M. Shinault

The Problem

X is a RV, with known PDF $f_X(x)$ or PMF $p_X(k)$. The function $g : \mathbb{R} \to \mathbb{R}$ is given. Find the PDF/PMF of $Y = g(X)$.

This is the only type of problem we solve in this section. The case where X is a discrete RV is usually easy, so we will primarily focus on the case where X is a continuous RV.

Goals for this Lecture

1. Learn how to find the PMF of $g(X)$ is X is discrete.
2. Learn how to find the PDF of $g(X)$ is X is continuous.

This material corresponds to section 5.2 of the textbook.

Discrete Case

Example

Example 5.14. Suppose the range of X is $\{-1, 0, 1, 2\}$ with $P\{X = k\} = \frac{1}{4}$ for each $k \in \{-1, 0, 1, 2\}$. Let $Y = X^2$. Find the PMF of Y.

Example

Example 5.15. Suppose that X takes values in the set $\{-1, 0, 1, 2\}$ with

$$P\{X = -1\} = \frac{1}{10} \quad P\{X = 0\} = \frac{2}{10}$$
$$P\{X = 1\} = \frac{3}{10} \quad P\{X = 2\} = \frac{4}{10}.$$

Let $Y = 2X^3$. Find the PMF of Y.

Continuous Case

General Strategy

1. Find the CDF: $F_Y(y) = \mathbb{P}(g(X) \leq y)$.
2. Differentiate the CDF: $f_Y(y) = F_Y'(y)$.

Easiest Example

Suppose $X \sim U([0,4])$. Set $Y = X^2$. Find $f_Y(y)$.

Easy Example

Suppose $X \sim U([-4,4])$. Set $Y = X^2$. Find $f_Y(y)$.

Moderate Example

Suppose $X \sim U([-3,4])$. Set $Y = X^2$. Find $f_Y(y)$.

Simplest General Case

Fact: If $g(u)$ has a differentiable inverse then pdf for transformed x

$$f_{g(X)}(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}.$$

Most General Case

Fact: If $g(u)$ is differentiable and $g'(u) = 0$ at only finitely many points then

$$f_{g(X)}(y) = \sum_{x:g(x)=y} \frac{f_X(x)}{|g'(x)|}.$$

Special Example

Suppose X is $N(\mu, \sigma^2)$. Prove that the distribution of $Y = aX + b$ is $N(a\mu + b, a^2\sigma^2)$.

This is a general fact worth remembering, especially for the duration of this course.

Log-Normal Distribution

$y \le 0$: $P(Y \le y) = P(e^X \le y) = 0$

Suppose X is $N(\mu, \sigma^2)$. Find the PDF of $Y = e^X > 0$

Definition: The distribution of Y is called the *log-normal* distribution.

$y > 0$ $F_Y(y) = P(Y \le y)$

$= P(e^X \le y) = P(X \le \log y)$

Application of Log-Normal

1. Modeling in mathematical finance

$F_Y(y) = F_X(\log y)$

$\Rightarrow f_Y(y) = \frac{d}{dy}[F_X(\log Y)]$

$= F_X'(\log y) \cdot \frac{1}{y} = f_X(\log y) \cdot \frac{1}{y}$

2. Distribution of income in the USA
3. Number of alcoholic drinks consumed by an individual per week in every culture

$= \frac{1}{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\log y - \mu)^2/2\sigma^2}$

$= f_Y(y)$ for $y > 0$

The Wrap Up

Summary

1. Use the CDF method or the formula to find the PDF of $g(X)$.
2. The CDF method is probably better practice because it requires more thought about probabilities, rather than analysis. (Personal preference)

$r \sim N(\mu, \sigma^2)$ $\quad rt \sim N(\mu t, t^2\sigma^2)$

e^{rt} ← log-normally distributed

general case of 多变量 RV

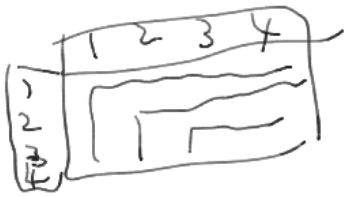
X, Y discrete RV joint PMF

ex. 1. $P(Y=4) = \sum_{j=1}^{3} P_{X,Y}(j,4)$ (sum over column) $= 0.2+0.15+0 = 0.35$

$P(X=2) = \sum_{k=1}^{4} P_{X,Y}(2,k) = 0.1+0.05+0+0.15 = 0.3$

$P(X+Y<4) = P_{X,Y}(1,1)+P_{X,Y}(1,2)+P_{X,Y}(2,1) = 0.05+0.1+0.1 = 0.25$

$\sum_{j,k:\, j+k<4} P_{X,Y}(j,k)$



$P(X=Y) = P_{X,Y}(1,1)+P_{X,Y}(2,2)+P_{X,Y}(3,3) = 0.15$

Sol $P_X(j) = P(X=j)$ $j=1,2,3$

$= \sum_{k=1}^{4} P_{X,Y}(j,k)$

$P_X(1) = 0.05+0.1+0.05+0.2 = 0.4$

$P_X(3) = 0.2+0.05+0.05+0 = 0.3$

j	1	2	3
$P_X(j)$	0.4	0	0.3

Range $(W) = \{1, 2, 3\}$

$P_W(1) = 0.2+0.1+0.05+0.1+0.05+2 = 0.7$

$P_W(2) = 0.05+0.05+0+0.15$

$P_W(3) = 0.05$

Jointly Discrete RVs

Gregory M. Shinault

Introduction

We have almost exclusively computed probabilities for a single RV so far (also called *univariate random variables*).

Now we look at the more general case of multiple RVs (*multivariate random variables* or *random vectors*).

Goals for this Lecture

We will learn how to

1. compute probabilities for multiple discrete RVs using their *joint probability mass function*,
2. determine the *marginal probability mass function* for a random variable from the joint PMF,
3. use the law of the unconscious statistician in a multivariate setting, and
4. use the joint PMF to determine if a collection of random variables is independent.

This material corresponds to section 6.1 and part of 6.3 of the textbook.

Joint PMF

Definition

Definition (for 2 RVs): Let X and Y be discrete RVs. Their *joint probability mass function* is defined by

$$p_{X,Y}(j,k) = \mathbb{P}(X = j, Y = k).$$

This is often represented as a table,

		Y			
		k_1	k_2	$\cdots$	k_m
X	j_1	$p_{X,Y}(j_1,k_1)$	$p_{X,Y}(j_1,k_2)$	$\cdots$	$p_{X,Y}(j_1,k_m)$
	j_2	$p_{X,Y}(j_2,k_1)$	$p_{X,Y}(j_2,k_2)$	$\cdots$	$p_{X,Y}(j_2,k_m)$
	$\vdots$	$\vdots$		$\ddots$	
	j_n	$p_{X,Y}(j_n,k_1)$	$p_{X,Y}(j_n,k_2)$	$\cdots$	$p_{X,Y}(j_n,k_m)$

Example

Suppose X and Y have the joint PMF

		Y			
		1	2	3	4
X	1	0.05	0.10	0.05	0.20
	2	0.10	0.05	0	0.15
	3	0.20	0.05	0.05	0

1. Compute $\mathbb{P}(Y = 4)$, $\mathbb{P}(X = 2)$, $\mathbb{P}(X + Y < 4)$ and $\mathbb{P}(X = Y)$.
2. Find the PMF of X.
3. Set $W = \min\{X, Y\}$. Find the PMF of W.

Properties

1. Probabilities are positive:
$$p_{X,Y}(j,k) \geq 0$$
2. Probabilities sum to 1:
$$\sum_{j \in \text{Ran}(X)} \sum_{k \in \text{Ran}(Y)} p_{X,Y}(j,k) = 1$$
3. To find the probabilities for only one of the RVs, you sum over all the values of the other RV:
$$p_X(j) = \sum_{k \in \text{Ran}(Y)} p_{X,Y}(j,k)$$

Joint PMF

Definition (for many RVs): Let $X_1, X_2, \ldots, X_n$ be discrete RVs. Their *joint probability mass function* is defined by

$$p_{X_1,X_2,\ldots,X_n}(k_1, k_2, \ldots, k_n) = \mathbb{P}(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n).$$

This is cannot be represented as a table.

Special Example (Multinomial Distribution)

Suppose we conduct a medical trial with 3 outcomes: improvement (1), deterioriation (2), no change (3). There are 60 patients. Let X_1, X_2, X_3 be the number of patients with outcome type 1, 2, 3 respectively.

The probabilities of outcomes 1, 2, and 3 are p_1, p_2, and p_3, respectively.

What is the joint PMF of X_1, X_2, X_3?

Multinomial Distribution

Definition: Let $X_1, X_2, \ldots, X_r$ be discrete RVs. They are said to have the *multinomial distribution* with parameters $n, p_1, \ldots, p_r$ if they have the joint PMF

$$p_{X_1,X_2,\ldots,X_r}(k_1, k_2, \ldots, k_r) = \binom{n}{k_1, k_2, \cdots, k_r} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}$$

k_1个类1 k_r个类r

for $k_1 + k_2 + \cdots + k_r = n$ and $p_1 + p_2 + \cdots + p_r = 1$.

We denote this by $(X_1, \ldots, X_r) \sim \text{Mult}(n, p_1, \ldots, p_r)$.

Comment: This is probably the only special discrete multivariate distribution we will discuss, compared to the many special discrete univariate distributions we can analyze.

Marginal PMF

Definition

Definition: Let $X_1, X_2, \ldots, X_n$ be discrete RVs with joint PMF $p(k_1, \ldots, k_n)$. The *marginal PMF* of $X_1, X_2, \ldots, X_m$ for $m < n$ is defined by

$$p_{X_1,\ldots,X_m}(k_1, \ldots, k_m) = \mathbb{P}(X_1 = k_1, \ldots, X_m = k_m).$$

This is computed by

$$p_{X_1,\ldots,X_m}(k_1, \ldots, k_m) = \sum_{j_{m+1},\ldots,j_n} p(k_1, \ldots, k_m, j_{m+1}, \ldots, j_n).$$

Comment: You have already found a marginal PMF in the first example.

Example

Let $(X_1, X_2, X_3) \sim \text{Mult}(n, p_1, p_2, p_3)$. Find the marginal pmf of X_2.

Expectation

Law of the Unconscious Statistician

Let $g : \mathbb{R}^n \to \mathbb{R}$. If $X_1, \ldots, X_n$ are discrete random variables with joint PMF p then

$$E[g(X_1, \ldots, X_n)] = \sum_{k_1} \cdots \sum_{k_n} g(k_1, \ldots, k_n) p(k_1, \ldots, k_n)$$

Example (6.6 in textbook)

A fair 4-sided die is rolled twice. Let X_k denote the k-th outcome. Set $Y_1 = \min\{X_1, X_2\}$ and $Y_2 = |X_1 - X_2|$.

- Find the joint PMF of X_1 and X_2.
- Find the joint PMF of Y_1 and Y_2.
- Find the expected value $E[Y_1 Y_2]$.

Independence

Fact

Recall: Discrete RVs are independent if and only if

$$\mathbb{P}(X = j, Y = k) = \mathbb{P}(X = j)\mathbb{P}(Y = k)$$

for all j, k.

Fact: $X_1, \ldots, X_n$ are independent if and only if

$$p_{X_1,\ldots,X_n}(k_1, \ldots, k_n) = p_{X_1}(k_1)p_{X_2}(k_2)\cdots p_{X_n}(k_n).$$

for $k_1, k_2, \ldots, k_n$.

Example

Suppose X and Y have the joint PMF

		Y			
		1	2	3	4
X	1	0.05	0.10	0.05	0.20
	2	0.10	0.05	0	0.15
	3	0.20	0.05	0.05	0

Are X and Y independent?

The Wrap Up

Summary

1. A joint PMF is based on the same idea and has the same properties as an ordinary PMF. The only difference is that it is multivariable.

2. A marginal PMF is obtained by summing over the ranges of the RVs that we are not interested in.
3. Independence is equivalent to splitting the joint PMF into a product of marginal PMFs.
4. The law of the unconscious statistician extends to multiple discrete random variables as you might hope.

Next Step

After completing the discrete version, of course we must cover the case of continuous random variables.

Jointly Continuous RVs

Gregory M. Shinault

Introduction

After working with jointly discrete RVs, the natural step to take is into jointly continuous RVs.

Goals for This Lecture

We will learn how to

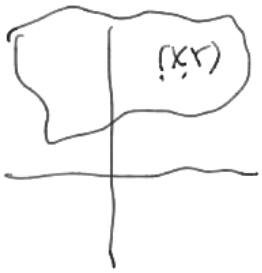
1. compute probabilities for multiple continuous RVs using their *joint probability density function*,
2. determine the *marginal probability density function* for a random variable from the joint PDF,
3. use the law of the unconscious statistician in a multivariate setting, and
4. use the joint PDF to determine if a collection of random variables is independent.

This material corresponds to section 6.2 and part of 6.3 of the textbook.

Joint PDF

Bivariate Case

Definition (for 2 RVs): Let X and Y be continuous RVs. Their *joint probability density function* is defined as the function which determines joint probabilities for X and Y:



$$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)\, dy\, dx$$

or for any region $R \subseteq \mathbb{R}^2$,

$$\mathbb{P}((X,Y) \in R) = \iint_R f_{X,Y}(x,y)\, dy\, dx$$

Example

Suppose X and Y have the joint PDF

$$f(x,y) = \begin{cases} \frac{1}{2} & \text{for } 0 \le x \le 1,\ 0 \le y \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

Compute $\mathbb{P}(Y = 1)$, $\mathbb{P}(X + Y < 1)$, $\mathbb{P}(X < Y)$, and $\mathbb{P}(X = Y)$.

Properties

1. Probabilities are positive:
$$f_{X,Y}(x,y) \ge 0$$
2. Total probability is 1:
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx = 1$$
3. To find the PDF for only one of the RVs, you integrate over all the values of the other RV:
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$$

$P(a \le X \le b) = \int_a^b f_X(x)\,dx$

$P(a \le X \le b) = P(a \le X \le b,\ -\infty < Y < \infty)$

$= \int_a^b \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx$

$f_X(x)$

Many Variables

Definition (for many RVs): Let $X_1, X_2, \ldots, X_n$ be continuous RVs. Their *joint probability density function* is defined as the function which determines joint probabilities for $X_1, X_2, \ldots, X_n$:

$$\begin{aligned}\mathbb{P}(a_1 \le\ & X_1 \le b_1, a_2 \le X_2 \le b_2, \ldots, a_1 \le X_n \le b_n) \\ &= \int_{a_n}^{b_n} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n,\end{aligned}$$

or more generally for region $R \subseteq \mathbb{R}^n$.

$$\begin{aligned}\mathbb{P}((X_1,\ & X_2, \ldots, X_n) \in R) \\ &= \int \cdots \int_R f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n.\end{aligned}$$

Example

Suppose X, Y, and Z have the joint PDF

$$f_{X,Y,Z}(x,y,z) = \begin{cases} 8xyz & \text{for } 0 \le x \le 1, 0 \le y \le 2 \\ 0 & \text{otherwise.} \end{cases} \quad 0 \le x, y, z \le 1$$

Compute $\mathbb{P}(Z \ge XY)$, and $\mathbb{P}(X < Y)$.

$\{Z \ge XY\} = \{0 \le X \le 1, 0 \le Y \le 1, XY \le Z \le 1\}$

$= \int_0^1 \int_0^1 [4xyz^2]_{z=xy}^1\, dy\, dx$

$= \int_0^1 \int_0^1 4xy - 4x^3y^3\, dy\, dx$

$P(Z \ge XY) = \int_0^1 \int_0^1 \int_{xy}^1 8xyz\, dz\, dy\, dx$

$= \int_0^1 \left[2xy^2 - x^3y^4\right]_{y=0}^1 dx = \int_0^1 2x - x^3\, dx$

Marginal PDF

Definition: Let $X_1, X_2, \ldots, X_n$ be continuous RVs with joint PDF $f(x_1, \ldots, x_n)$. The *marginal PDF* of $X_1, X_2, \ldots, X_m$ for $m < n$ is defined by

$$f_{X_1,\ldots,X_m}(x_1, \ldots, x_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)\, dx_{m+1}\, dx_{m+2} \cdots dx_n.$$

Uniform Distribution

Definition: Let $X_1, X_2, \ldots, X_n$ be continuous RVs. They are said to have the *uniform distribution* on the region A if they have the joint PDF

$$f(x_1, x_2, \ldots, x_n) = \begin{cases} \frac{1}{\text{Volume}(A)} & \text{for } (x_1, x_2, \ldots, x_n) \in A \\ 0 & \text{otherwise.} \end{cases}$$

Problem: Suppose X and Y are uniformly distributed on the unit circle, $x^2 + y^2 \leq 1$. Find the marginal PDF of Y.

Expectation

Law of the Unconscious Statistician

Let $g : \mathbb{R}^n \to \mathbb{R}$. If $X_1, \ldots, X_n$ are continuous random variables with joint PDF f then

$$E[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Example (6.14 in textbook)

Suppose X, Y have joint PDF

$$f(x, y) = \begin{cases} \frac{3}{2}(xy^2 + y) & \text{if } 0 \leq x, y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of $P(X < Y)$ and $E[X^2 Y]$.

Independence

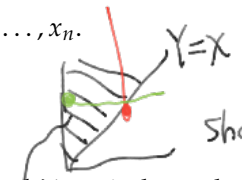
Fact

Fact: Continuous RVs $X_1, \ldots, X_n$ are independent if and only if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

for all $x_1, x_2, \ldots, x_n$.

Example



Suppose X and Y are independent RVs with distributions $\text{Exp}(\lambda)$ and $\text{Exp}(\gamma)$, respectively. Compute $\mathbb{P}(X < Y)$.

The Wrap Up

Summary

1. A joint PDF is based on the same idea and has the same properties as an ordinary PDF. The only difference is that it is multivariable.
2. A marginal PDF is obtained by integrating over the ranges of the RVs that we are not interested in.
3. Independence is equivalent to splitting the joint PDF into a product of marginal PDFs.
4. ~~Independence~~ is equivalent to splitting the joint PDF into a product of marginal PDFs.

Expectation in multivariate setting

Additional Topics in Jointly Distributed RVs

Gregory M. Shinault

Introduction

This is a lecture reserved for additionaly topics and examples on jointly distributed random variables. It would contain ideas from sections 6.3 and 6.4 of the textbook.

We will not be including this lecture this semester.

Convolution Formulas

Gregory M. Shinault

The Problem

Suppose X and Y are independent discrete/continuous RVs. What is the the PMF/PDF of $W = X + Y$?

This lecture corresponds to section 7.1 of the textbook.

Discrete Case

The Convolution Formula

Fact: Suppose X and Y are independent discrete RVs, and $\text{Ran}(X) = \{0, 1, \ldots, n\}$, $\text{Ran}(Y) = \{0, 1, \ldots, \ell\}$. Set $W = X + Y$. Then

$$p_W(k) = \sum_{j=0}^{k} p_X(j) p_Y(k-j) = p_X \star p_Y(k).$$

for $k = 0, 1, 2, \ldots, n + \ell$.

Comment: You can generalize to other ranges for the RVs. It is just necessary to be careful with the indices.

Example

Let X and Y be independent RVs with the distributions $\text{Pois}(\lambda_X)$ and $\text{Pois}(\lambda_Y)$. Find the PMF of $X + Y$.

Special Example

Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$ be independent RVs. Find the PMF of $X + Y$.

Negative Binomial Distribution

Definition

A RV is said to have the *negative binomial* distribution with parameters k and p if it has PMF

$$p_X(n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

P=0 until

for $n = k, k+1, \ldots$.

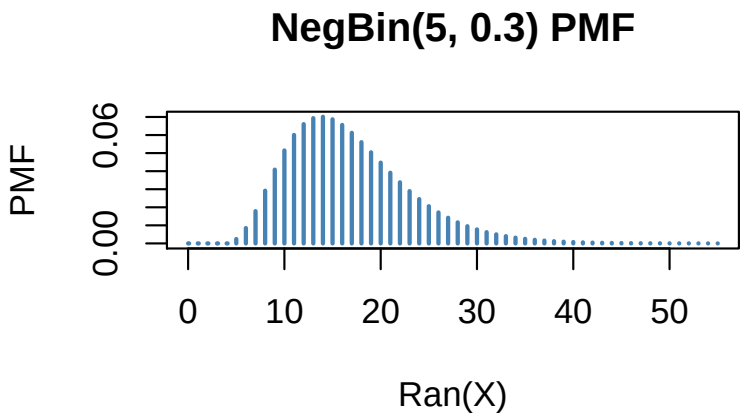
NegBin results as sum of geometric RVs

Interpretation: $X \sim \text{NegBin}(k, p)$ can be defined as

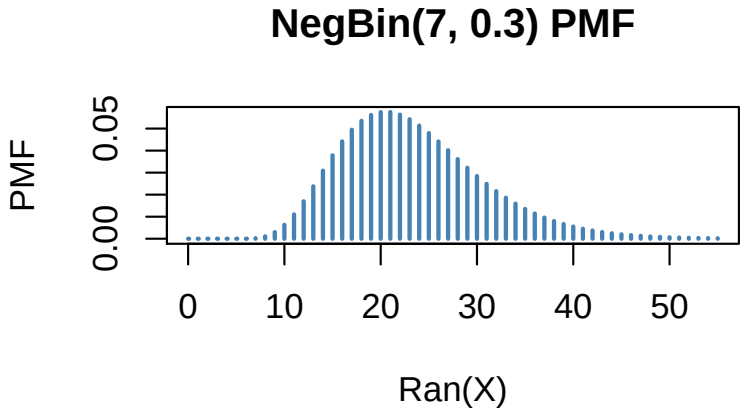
$$X = T_1 + T_2 + \cdots + T_k$$

for independent RVs $T_\ell \sim \text{Geo}(p)$.

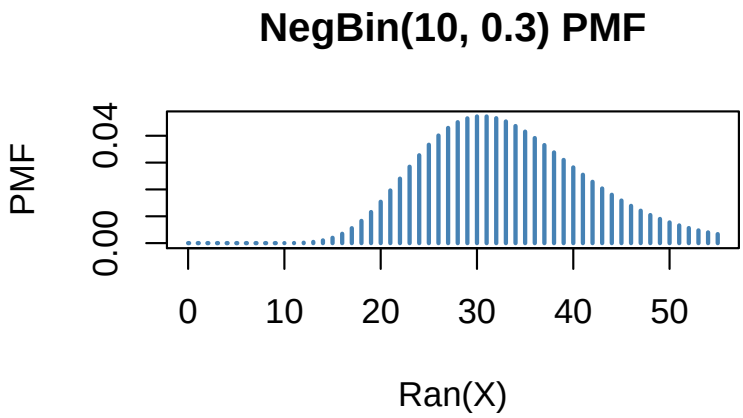
PMF, $k = 5$ and $p = 0.3$



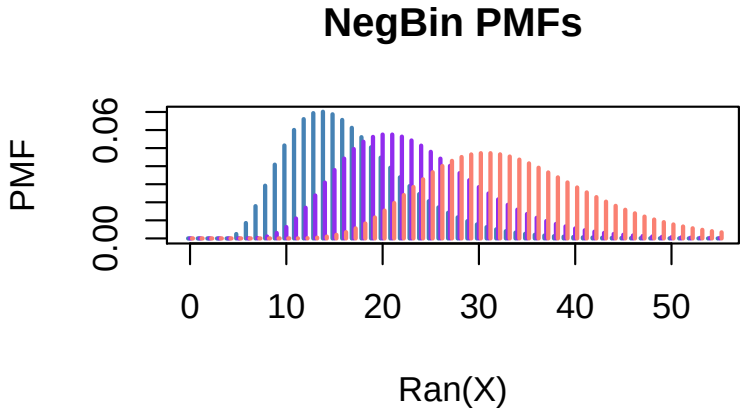
PMF, $k = 7$ and $p = 0.3$



PMF, $k = 10$ and $p = 0.3$



PMFs, Compared



Continuous Setting

Convolution Formula

卷积

Fact: Suppose X and Y are independent continuous RVs Set $W = X + Y$. Then

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w - x)\, dx = f_X \star f_Y(w).$$

for all w.

Example

Let X and Y be independent RVs with the distribution $N(0,1)$. Find the PDF of $X+Y$.

Important Fact

The following fact is a generalization of the previous example.

Fact: Suppose $X_1, \ldots, X_n$ are independent and $X_j \sim N(\mu_j, \sigma_j^2)$. Then

$$X_1 + X_2 + \cdots + X_n \sim N(\mu_1 + \mu_2 + \cdots + \mu_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2).$$

Example

Let $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\lambda)$ be independent RVs. Find the PDF of $X+Y$.

Gamma Distribution

Definition

A RV is said to have the *gamma* distribution with parameters k and λ if it has PDF

$$f_X(x) = \frac{\lambda^k x^{k-1}}{(k-1)!} e^{-\lambda x}$$

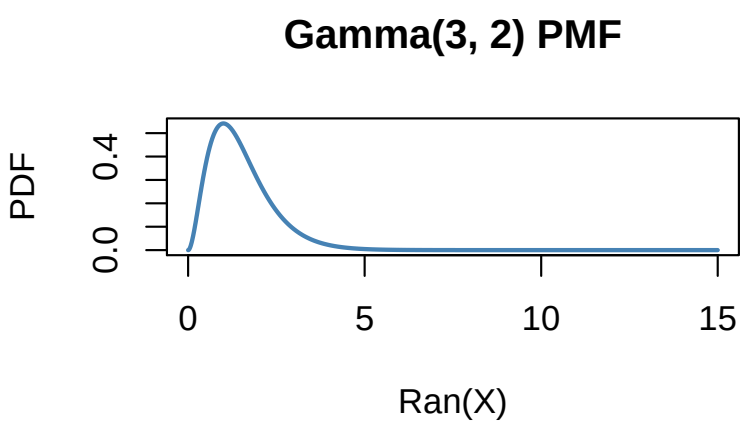
for $x \geq 0$.

Interpretation: $X \sim \text{Gamma}(k, \lambda)$ can be defined as

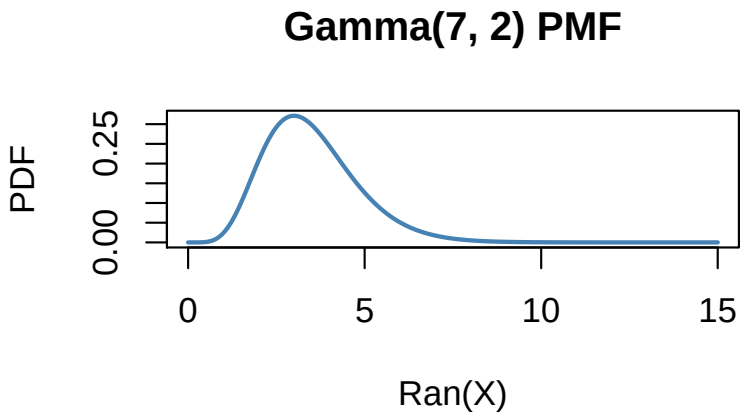
$$X = T_1 + T_2 + \cdots + T_k$$

for independent RVs $T_\ell \sim \text{Exp}(\lambda)$.

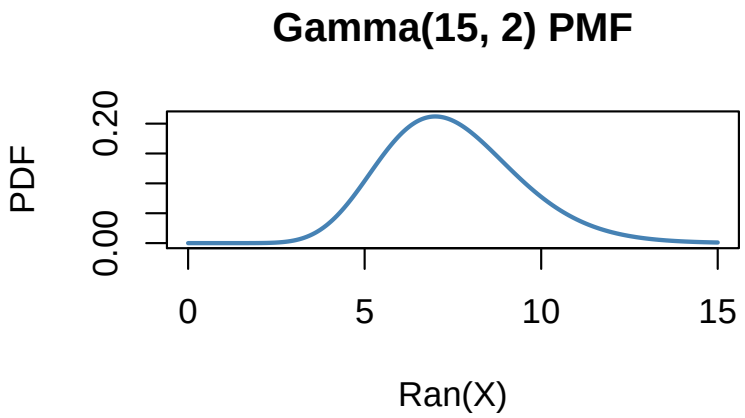
PDF, $n = 3$ *and* $\lambda = 2$



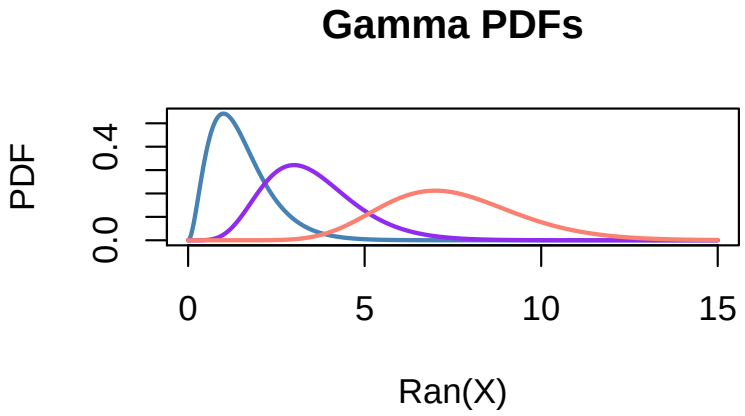
PDF, $n = 7$ *and* $\lambda = 2$



PDF, $n = 17$ and $\lambda = 2$



PDFs, Compared



A Few More Special Uses

1. The sum of independent Poisson RVs is Poisson.
2. The sum of independent Binomial RVs is Binomial.
3. The sum of independent Geometric RVs with the same parameter is NegBin.
4. The sum of independent Normal RVs is Normal.
5. The sum of independent Exponential RVs with the same parameter is Gamma.

6. The $\chi^2(n)$ distribution can be derived as a sum of independent RVs.
7. ...and so on.

The Wrap Up

Summary

1. The convolution formula can be used to find the PMF/PDF of $X + Y$ when X and Y are independent.
2. The sum of independent normal RVs has a normal distribution.
3. We have learned two new special distributions: Negative Binomial and the Gamma.

Covariance and Correlation

Gregory M. Shinault

Introduction

Goal

The joint PMF is complete information about the relationship between X and Y. Unfortunately it can be difficult to interpret the relationship between X and Y from the PMF.

Covariance and correlation provide a simple way to interpret this relationship.

We also have to cover a few prerequisites for this material.

The core of this topic is covered in Section 8.4. We also need a few ideas from 8.1 and 8.2, but not the entirety of those sections.

Linearity of Expectation

Theorem

For jointly distributed RVs $X_1, \ldots, X_n$ we have

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n].$$

Exercise: Prove this for the continuous case. You will want to use the law of the unconscious statistician.

Method of Indicators

Example: As I stroll down Williamson Street on a lovely Sunday morning, there are 4 bakeries I can duck into for a nice baked good: Lazy Jane's, Batch Bakehouse, the Co-Op, and Madison Sourdough. As a lover of baking, I might stop into any number of the 4. The probabilities I stop into each bakery are 0.7, 0.6, 0.05, and 0.3. What is the expected number of bakeries I stop into on Sunday?

We do not have time to address this topic this semester, but I highly recommend reading about it in section 8.1. The method of indicators is a beautiful example of good math making extremely difficult problems reasonable.

Independence and Expectation

Fact

If X and Y are independent then

$$\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y.$$

More generally, if $X_1, \ldots, X_n$ are independent and we have single variable functions $g_1, \ldots, g_n$ then

$$\mathbb{E}[g_1(X_1)g_2(X_2)\cdots g_n(X_n)] = \mathbb{E}[g_1(X_1)]\mathbb{E}[g_2(X_2)]\cdots\mathbb{E}[g_n(X_n)].$$

Independence and Variance

Fact

If $X_1, \ldots, X_n$ are independent then

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n).$$

Example

Find the variance of the negative binomial distribution.

Covariance

Definition

The *covariance* of random variables X and Y is given by

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

If $\mathrm{Cov}(X, Y) > 0$, we say X and Y are *positively correlated*.
If $\mathrm{Cov}(X, Y) < 0$, we say X and Y are *negatively correlated*.
If $\mathrm{Cov}(X, Y) = 0$, we say X and Y are *uncorrelated*.

Computation

Fact: The *covariance* of random variables X and Y is computed by

$$\mathrm{Cov}(X, Y) = \mathbb{E}XY - \mu_X\mu_Y.$$

Example

Suppose X and Y have the joint PMF

		Y		
		1	2	3
X	1	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{4}$
	2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$

Find $\text{Cov}(X, Y)$.

Example

Suppose (X, Y) are uniformly distributed on the circle of radius 2 at the origin. Find $\text{Cov}(X, Y)$.

Properties

1. Symmetry: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
2. Bilinearity:

$$\begin{aligned}&\text{Cov}(aX + bY, cW + dZ)\\ &= ac\text{Cov}(X, W) + ad\text{Cov}(X, Z) + bc\text{Cov}(Y, W) + bd\text{Cov}(Y, Z).\end{aligned}$$

Relationship to Variance

Fact: The variance of a sum of random variables can be found with the formula

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{j=1}^{n} \text{Var}(X_j) + 2\sum_{i<k} \text{Cov}(X_i, X_k).$$

Special Example | Hypergeometric Distribution

Suppose $X \sim \text{HyperGeo}(N, N_A, n)$. Find the variance of X.

Shortcomings

1. The magnitude of the covariance is not indicative of the strength of the relationship between X and Y. (Changing units changes the covariance, but the underlying relationship should not change)
2. Covariance only measures the linear relationship between X and Y. (We will not address this issue)

Correlation

Definition

Let

$$X_* = \frac{X - \mu_X}{\sigma_X}, \quad Y_* = \frac{Y - \mu_Y}{\sigma_Y}.$$

The *correlation* of X and Y is defined as

$$\text{Corr}(X, Y) = \text{Cov}(X_*, Y_*) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Key Properties

1. $-1 \le \text{Cor}(X, Y) \le 1$.
2. $\text{Corr}(X, Y) = 1$ if and only if $Y = aX + b$ for some positive a.
3. $\text{Corr}(X, Y) = -1$ if and only if $Y = -aX + b$ for some positive a.

Special Example | Multinomial Distribution

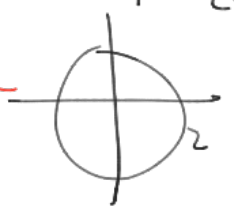
Suppose $(X_1, \ldots, X_r) \sim \text{Mult}(n, p_1, \ldots, p_r)$. Find $\text{Corr}(X_i, X_j)$.

Summary

Key Ideas

1. $\text{Cov}(X, Y) = \mathbb{E}XY - \mu_X \mu_Y$.
2. $\text{Cor}(X, Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$.
3. Both are used to measure the linear relationship between X and Y.
4. They possess many properties. You must know them all.

continuous example



Two Important Limit Theorems

Gregory M. Shinault

Introduction

Goals for this Lecture

1. Learn about the statement of the Weak Law of Large Numbers.
2. Learn the statement and use of the Central Limit Theorem.

The material in this lecture corresponds to sections 9.2 and 9.3 of the textbook.

IID Sequences

Definition: We say that a sequence of random variables X_1, X_2, ... are *indepedent and identically distributed* (IID) if two conditions are satisfied.

1. The random variables $X_1, X_2, \ldots$ are independent.
2. The random variables $X_1, X_2, \ldots$ all follow the same distribution.

Importance: IID sequences are an idealized version of a statistical sample. Each element of the sample follows the same distribution, and we can take a sample size to be arbitrarily large. This might not match realistic circumstances, but it is close enough to be useful.

Weak Law of Large Numbers

The Statement

Independent & same distribution

Theorem: Suppose $X_1, X_2, \ldots$ is an IID sequence of RVs with mean $\mathbb{E}X_1 = \mu$ and variance $\mathrm{Var}(X_1) = \sigma^2$. Let $\bar{X}_n$ be the sample mean,

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

sample mean

Then for any $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| < \varepsilon) = 1.$$

Sample mean - mean

Not data in general

Central Limit Theorem

iid sums ~ Normal distribution

Big Idea

Recall the normal approximation to the binomial distribution. A binomial random variable can be expressed as a sum of IID Bernoulli RVs.

It turns out this can be restated as a sum of (almost) any IID RVs and the normal approximation is still valid.

$X \sim Bin(n,p)$

n is big

$X \overset{d}{\approx} Y \sim N(EX, SD(X))$

Formal Statement

Alternate view $X_1 X_2 \cdots X_n$ iid, Ber(p)

$X \overset{d}{\approx} X_1 + X_2 + \cdots + X_n$.

Suppose $X_1, X_2, \ldots$ is an IID sequence of RVs with $\mathbb{E}X_1 = \mu < \infty$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Then for all real t we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le t\right) = \Phi(t)$$

where $\Phi(t)$ is the standard normal CDF and $S_n = X_1 + X_2 + \cdots + X_n$.

Sums of iid bernoulli(p) RVs is normally distributed $N(np, np(1-p))$

Intuitive Interpretation

1. Sums of RVs are approximately normal:

$$X_1 + \cdots + X_n \approx Y \sim N(n\mu, n\sigma^2).$$

2. Sample means are approximately normal:

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \approx Y \sim N(\mu, \sigma^2/n).$$

$E S_n = \sum_{i=1}^{n} E X_i = n\mu$

$\mathrm{Var}(S_n) = \mathrm{Var}(X_1 + \ldots + X_n)$

$= \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\sigma^2$

Key Idea in Proof of CLT | Continuity Theorem for MGFs

Assume the MGFs of $S_1, S_2, \ldots$ and Z satisfy

$$\lim_{n\to\infty} M_{S_n}(t) = M_Z(t)$$

for all $-\varepsilon < t < \varepsilon$ for some $\varepsilon > 0$. Then

$$\lim_{n\to\infty} \mathbb{P}(S_n \le s) = \mathbb{P}(Z \le s)$$

for all real s.

$M_{\frac{S_n - n\mu}{\sigma\sqrt{n}}}(t) = E\left[\exp\left(t \cdot \frac{S - n\mu}{\sigma\sqrt{n}}\right)\right]$

$= E\left[\exp\left(\frac{t}{\sigma\sqrt{n}} \cdot \sum_{k=1}^{n} (X_k - \mu)\right)\right]$

$= E\left[\prod_{k=1}^{n} \exp\left(\frac{t}{\sigma\sqrt{n}}(X_k - \mu)\right)\right]$

$= \prod_{k=1}^{n} E\left[\exp\left(\frac{t}{\sigma\sqrt{n}}(X_k - \mu)\right)\right]$

$= \left[E\left[\exp\left(\frac{t}{\sigma\sqrt{n}}(X_1 - \mu)\right)\right]\right]^n$

$= E\left[\sum_{j=0}^{\infty} \frac{1}{j!}\left(\frac{t}{\sigma\sqrt{n}}(X_1 - \mu)\right)^j\right]$

$= E[\ldots + \text{Error}]$ *

Example

We roll 1000 dice and add up the values. Estimate the probability that the sum is at least 3600.

0.032

Let X_k = outcome of kth die

$S_{1000} = \sum_{k=1}^{1000} X_k$

$E S_{1000} = E[X_1 + \ldots + X_{1000}]$

$= \sum_{k=1}^{1000} E X_k = 3.5 \cdot 1000 = 3500$

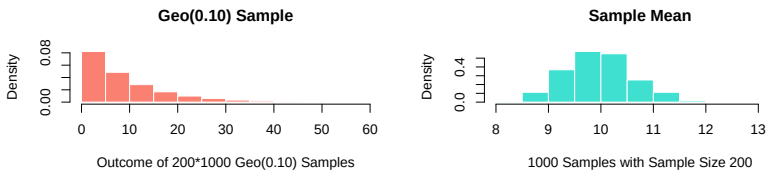
$\mathrm{Var}\, S_{1000} = \mathrm{Var}(X_1 + \ldots + X_{1000})$

$= \sum_{k=1}^{1000} \mathrm{Var}(X_k) = \frac{3500}{12}$

$P(S_{1000} > 3600) = 1 - P(S_{1000} \le 3600)$

$\approx 1 - \Phi\left(\frac{100}{\sqrt{\frac{3500}{12}}}\right)$

Simulation of Sample Mean

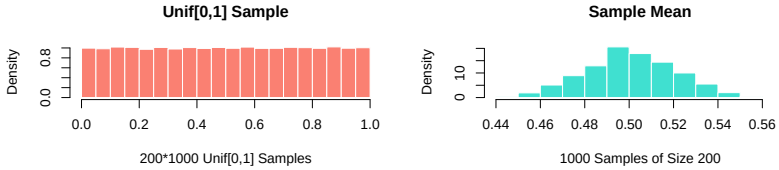


$$\chi = \left(1 + \frac{t^2/2}{n} + \text{ERROR}\right)^n$$

$$n \to \infty \quad \chi = e^{t^2/2}$$

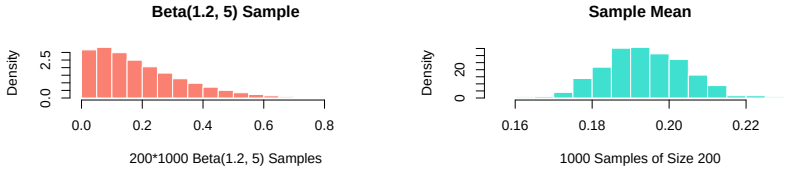
$$= M_Z(t)$$

Simulation of Sample Mean



$$\text{Tidsum} \sim \text{Normal}$$

Simulation of Sample Mean



Simulation of Sample Mean

Search for "Galton box" or "quincunx."

Nonexample of a Simulation

Search for "popcorn central limit theorem." The video is neat, but not related to the CLT.

Summary

1. The weak law of large numbers tells us that the empirical mean of an IID sequence will converge to the expected value of th RVs' shared distribution.
2. The MGF is a critical mathematical tool in proving the CLT.

3. The CLT roughly says that large sums are approximately normally distributed, and thus large sample means are approximately normally distributed.
4. The CLT does NOT say that a lot of data will be normally distributed, only that data based on sum of random variables will be approximately normally distributed.

Discrete Conditioning

Gregory M. Shinault

Big Idea

We just want to take what we already know about conditional probability and apply it to PMFs, to get conditional PMFs. Nothing new here.

We then take the conditional PMFs and compute conditional expectation. This is a new part of the theory. The point is to give a mathematical formulation to these types of questions:

1. What is the average life expectancy for a person randomly chosen in the whole world? (71.0 years)
2. What is the average life expectancy for a person randomly chosen in the United States? (79.8 years)

This material corresponds to section 10.1, and some portions of 10.3 of the textbook.

Conditioning on an Event

Conditional PMF

Definition: The *conditional PMF* of a RV X given an event A is defined as

$$p_{X|A}(k) = \mathbb{P}(X = k|A) = \frac{\mathbb{P}(X = k,\, A)}{\mathbb{P}(A)}.$$

Conditional Expectation

using conditional mass function

Definition: The *conditional expectation* of a RV X given an event A is defined as

$$\mathbb{E}[X|A] = \sum_{k \in \text{Ran}(X)} k\, p_{X|A}(k).$$

Example

The number of customers that go into The Soap Opera at lunch follows the Poisson(10) distribution on sunny days, but the Poisson(4) distribution on a snowy day. What is the expected number of customers at lunch on a randomly chosen day, given that it is sunny? What if it is snowy?

$X :=$ number of customers on given day

$A :=$ Event that it is sunny

$A^c :=$ Event that it is snowy

Want $E[X|A]$ $E[X|A^c]$

$E[X|A] = \sum_{k=0}^{\infty} k\, P_{X|A}(k) = \sum_{k=0}^{\infty} k \cdot \frac{e^{-10} 10^k}{k!} = 10$ poisson(10)

$E[X|A^c] = \sum_{k=0}^{\infty} k P(X|A^c) = 4$

$EX = E[X|A]P(A) + E[X|A^c]P(A^c) = 4 + 6P(A)$

How would you guess we find the expected number of customers at lunch on a randomly chosen day (no conditioning)?

Properties

Fact: The following are true.

1. $\mathbb{E}[cX + Y|A] = c\mathbb{E}[X|A] + \mathbb{E}[Y|A]$.
2. For a partition $A_1, \ldots, A_n$ we have

$$\mathbb{E}X = \sum_{j=1}^{n} \mathbb{E}[X|A_j]\mathbb{P}(A_j).$$

Conditioning on a RV

Conditional PMF

Definition: For two RVs X and Y the *conditional PMF* of X given $Y = y$ is defined as

$$p_{X|Y}(x|y) = \mathbb{P}(X = x \,|\, Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

for all $x \in \text{Ran}(X)$ and $y \in \text{Ran}(Y)$.

Conditional Expectation given $Y = y$

Definition: For two RVs X and Y the *conditional expectation* of X given $Y = y$ is defined as

$$\mathbb{E}[X|Y = y] = \sum_{x \in \text{Ran}(X)} x\, p_{X|Y}(x|y)$$

for all $y \in \text{Ran}(Y)$.

Conditional Expectation given Y

Definition: For two RVs X and Y the *conditional expectation* of X given Y is defined as

where

$$\mathbb{E}[X|Y] = g(Y)$$

$$g(y) = \mathbb{E}[X|Y = y].$$

Properties

1. Conditional expectation is linear:

$$\mathbb{E}[cX + Y|Z] = c\mathbb{E}[X|Z] + \mathbb{E}[Y|Z].$$

2. You can make substitutions based on the given information:

$$\mathbb{E}[h(X,Y)|Y = y] = \mathbb{E}[h(X,y)|Y = y].$$

3. Self-conditioning gives the original RV:

$$\mathbb{E}[X|X] = X.$$

4. If X and Y are independent then

$$\mathbb{E}[X|Y] = \mathbb{E}X.$$

Exercise: Prove these properties.

Example

Suppose X is the outcome of a fair die roll and Y is the outcome of a biased 4-sided die with PMF

k	1	2	3	4
$\mathbb{P}(Y = k)$	1/6	1/3	1/3	1/6

Set $Z = X + Y$. Find $\mathbb{E}[Z|Y = 2]$ and $\mathbb{E}[Z|Y]$.

Property

For any RVs X and Y we have

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}X.$$

Comment. This property is the most important property of conditional expectation. It simplifies many expectation computations. It goes by many names: law of total expectation, law of iterated expectation, the tower property, etc.

Special Example (Random Sums)

Suppose N is a discrete nonnegative RV, and $X_1, X_2, \ldots$ are IID RVs. Set

$$S_N = X_1 + X_2 + \cdots + X_N = \sum_{k=1}^{N} X_k.$$

What is $\mathbb{E}S_N$? (The result is also called *Wald's identity*)

To make this more concrete, you can set N to be the number of customers a store has over the lunch hour. X_k is the amount the k-th customer spends. Then S_N is the total amount the store makes in the lunch hour.

Example

California experiences X earthquakes over magnitude 4.0 each month, where $X \sim \text{Poisson}(65)$. For each earthquake above 4.0, there is a probability 0.10 that the earthquake is above magnitude 5.0. Assume the magnitude of each earthquake is independent.

Let X be the number of over magnitude 4.0 earthquakes this year and let Y be the number of over magnitude 5.0 earthquakes this year. Find $\mathbb{E}(Y|X)$, $\mathbb{E}Y$, and $\mathbb{E}(X|Y)$.

Summary

Key Ideas

1. Conditional PMFs are defined as
$$p_{X|Y}(j|k) = P(X = j|Y = k) = p_{X,Y}(j,k)/p_Y(k).$$
2. The conditional expectation given an event is
$$E[X|Y = k] = \sum_j j p_{X|Y}(j|k).$$
3. If $g(k) = E[X|Y = k]$ then $E[X|Y] = g(Y)$. So $E[X|Y]$ is a random variable.
4. $E[E(X|Y)] = EX$. Towerproperty