Introduction to Probability Distributions

Utkarsh Kulshrestha
Data Scientist
kuls.utkarsh1205@gmail.com

Random Variable

- A random variable *x* takes on a defined set of values with different probabilities.
  - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
  - For example, if you poll people about their voting preferences, the percentage of the sample that responds “Yes on Proposition 100” is a also a random variable (the percentage will be slightly differently every time you poll).

- Roughly, <u>probability</u> is how frequently we expect different outcomes to occur if we repeat the experiment over and over (“frequentist” view)

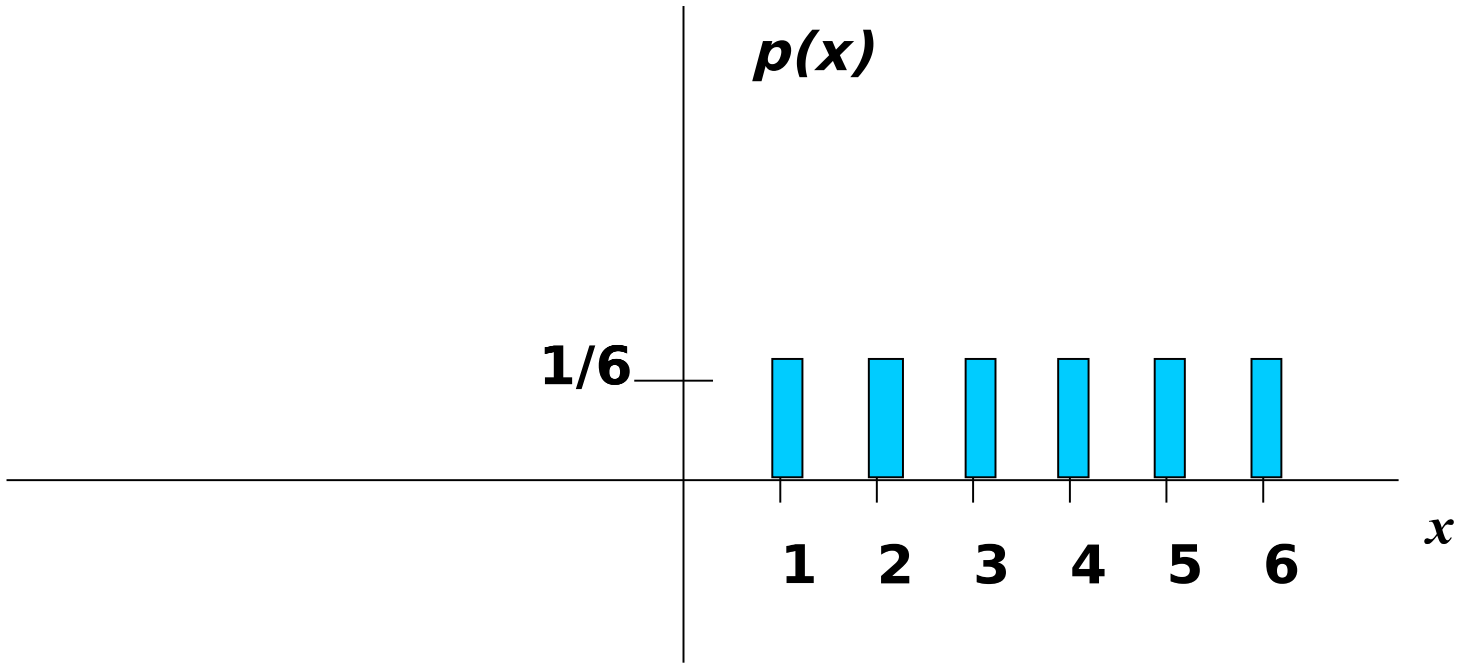
Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

Probability functions

- A probability function maps the possible values of *x* against their respective probabilities of occurrence, *p(x)*
- *p(x)* is a number from 0 to 1.0.
- The area under a probability function is always 1.

Discrete example: roll of a die

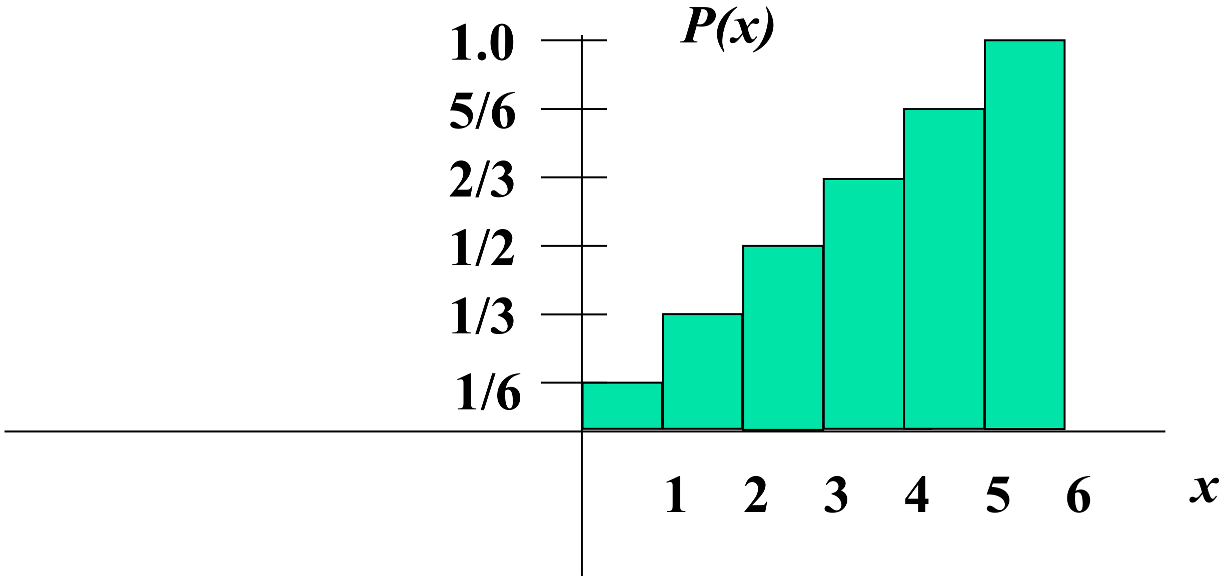


$$\sum_{\text{all } x} P(x) = 1$$

Probability mass function (pmf)

x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$p(x=6)=1/6$
	1.0

Cumulative distribution function (CDF)



Practice Problem:

- The number of patients seen in the ER in any given hour is a random variable represented by *x*. The probability distribution for *x* is:

x	**10**	**11**	**12**	**13**	**14**
P(x)	**.4**	**.2**	**.2**	**.1**	**.1**

Find the probability that in a given hour:

a. exactly 14 patients arrive $p(x=14) = .1$

b. At least 12 patients arrive $p(x \geq 12) = (.2 + .1 + .1) = .4$

c. At most 11 patients arrive $p(x \leq 11) = (.4 + .2) = .6$

Definitions

- **Probability**: the chance that an uncertain event will occur (always between 0 and 1)
- **Event**: Each possible type of occurrence or outcome
- **Simple Event**: an event that can be described by a single characteristic
- **Sample Space**: the collection of all possible events

Types of Probability

There are three approaches to assessing the probability of an uncertain event:

1. ***a priori* classical probability**: the probability of an event is based on prior knowledge of the process involved.

2. **empirical classical probability**: the probability of an event is based on observed data.

3. **subjective probability**: the probability of an event is determined by an individual, based on that person's past experience, personal opinion, and/or analysis of a particular situation.

Calculating Probability

1. ***a priori* classical probability**

$$\text{Probability of Occurrence} = \frac{X}{T} = \frac{\text{number of ways the event can occur}}{\text{total number of possible outcomes}}$$

2. **empirical classical probability**

$$\text{Probability of Occurrence} = \frac{\text{number of favorable outcomes observed}}{\text{total number of outcomes observed}}$$

These equations assume all outcomes are equally likely.

Example of *a priori* classical probability

Find the probability of selecting a face card (Jack, Queen, or King) from a standard deck of 52 cards.

$$\text{Probability of Face Card} = \frac{X}{T} = \frac{\text{number of face cards}}{\text{total number of cards}}$$

$$\frac{X}{T} = \frac{12 \text{ face cards}}{52 \text{ total cards}} = \frac{3}{13}$$

Example of empirical classical probability

Find the probability of selecting a male taking statistics from the population described in the following table:

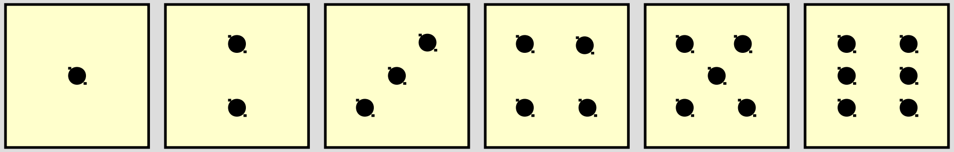
	Taking Stats	Not Taking Stats	Total
Male	84	145	229
Female	76	134	210
Total	160	279	439

$$\text{Probability of Male Taking Stats} = \frac{\text{number of males taking stats}}{\text{total number of people}} = \frac{84}{439} = 0.191$$

Examples of Sample Space

The Sample Space is the collection of all possible events

ex. All 6 faces of a die:



ex. All 52 cards in a deck of cards

ex. All possible outcomes when having a child: Boy or Girl

Events in Sample Space

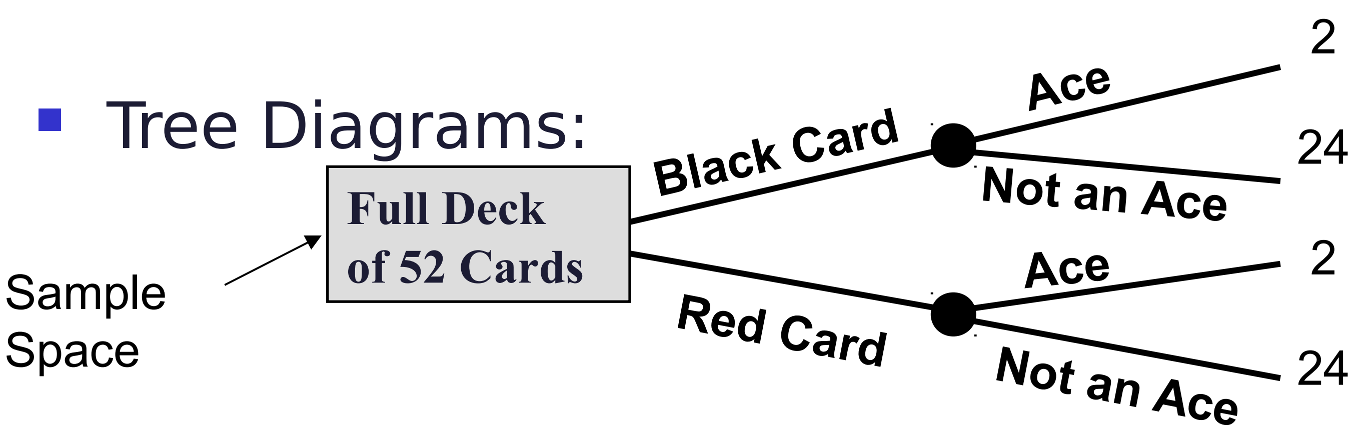
- Simple event
  - An outcome from a sample space with one characteristic
  - ex. A red card from a deck of cards
- Complement of an event A (denoted $A^{/}$)
  - All outcomes that are not part of event A
  - ex. All cards that are not diamonds
- Joint event
  - Involves two or more characteristics simultaneously
  - ex. An ace that is also red from a deck of cards

Visualizing Events in Sample Space

- Contingency Tables:

	Ace	Not Ace	Total
Black	2	24	26
Red	2	24	26
Total	4	48	52

- Tree Diagrams:



Definitions
Simple vs. Joint Probability

- Simple (Marginal) Probability refers to the probability of a simple event.
  - ex. P(King)
- Joint Probability refers to the probability of an occurrence of two or more events.
  - ex. P(King and Spade)

Definitions Mutually Exclusive Events

- **Mutually exclusive events** are events that cannot occur together (simultaneously).

- example:
  - A = queen of diamonds; B = queen of clubs
  - Events A and B are mutually exclusive if only one card is selected
- example:
  - B = having a boy; G = having a girl
  - Events B and G are mutually exclusive if only one child is born

Definitions Collectively Exhaustive Events

- **Collectively exhaustive events**
  - One of the events must occur
  - The set of events covers the entire sample space
- example:
  - A = aces; B = black cards; C = diamonds; D = hearts
  - Events A, B, C and D are **collectively exhaustive** (but not mutually exclusive – a selected ace may also be a heart)
  - Events B, C and D are **collectively exhaustive** and also mutually exclusive

Computing Joint and Marginal Probabilities

- The probability of a joint event, A and B:

$$P(\text{A and B}) = \frac{\text{number of outcomes satisfying A and B}}{\text{total number of elementary outcomes}}$$

- Computing a marginal (or simple) probability:

$$P(A) = P(\text{A and } B_1) + P(\text{A and } B_2) + \cdots + P(\text{A and } B_k)$$

  - Where B_1, B_2, …, B_k are k mutually exclusive and collectively exhaustive events

Example: Joint Probability

P(Red and Ace)

$$= \frac{\text{number of cards that are red and ace}}{\text{total number of cards}} = \frac{2}{52}$$

	Ace	Not Ace	Total
Black	2	24	26
Red	2	24	26
Total	4	48	52

Example: Marginal (Simple) Probability

P(Ace)

$$= P(\text{Ace and Red}) + P(\text{Ace and Black}) = \frac{2}{52} + \frac{2}{52} = \frac{4}{52}$$

	Ace	Not Ace	Total
Black	2	24	26
Red	2	24	26
Total	4	48	52

Joint Probability Using a Contingency Table

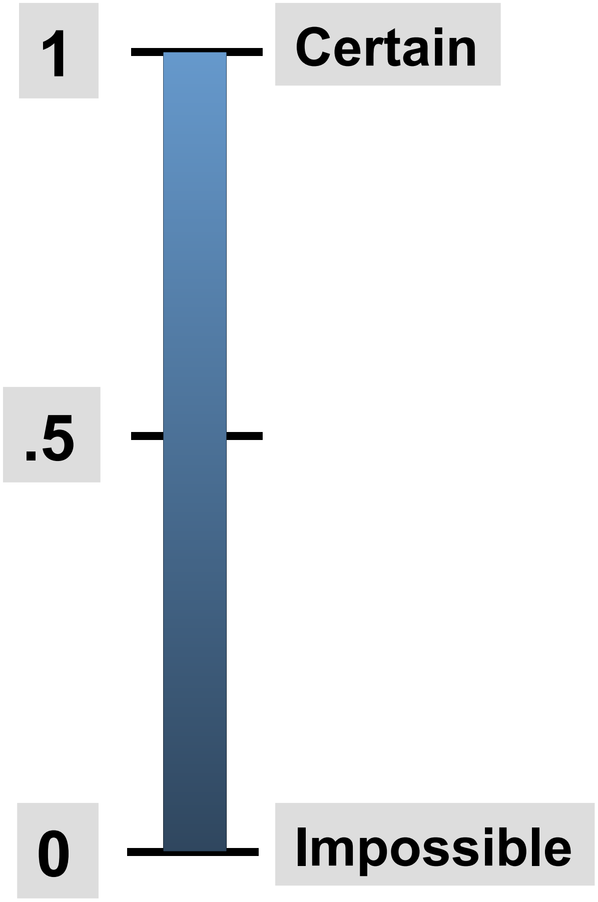
Event	Event: B_1	B_2	Total
A_1	$P(A_1 \text{ and } B_1)$	$P(A_1 \text{ and } B_2)$	$P(A_1)$
A_2	$P(A_2 \text{ and } B_1)$	$P(A_2 \text{ and } B_2)$	$P(A_2)$
Total	$P(B_1)$	$P(B_2)$	1

Joint Probabilities

Marginal (Simple) Probabilities

Probability Summary So Far

- Probability is the numerical measure of the likelihood that an event will occur.
- The probability of any event must be between 0 and 1, inclusively
  - $0 \le P(A) \le 1$ for any event A.
- The sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1.
  - $P(A) + P(B) + P(C) = 1$
  - A, B, and C are mutually exclusive and collectively exhaustive



General Addition Rule

General Addition Rule:

P(A or B) = P(A) + P(B) - P(A and B)

If A and B are mutually exclusive, then

P(A and B) = 0, so the rule can be simplified:

P(A or B) = P(A) + P(B)

for mutually exclusive events A and B

General Addition Rule Example

Find the probability of selecting a male or a statistics student from the population described in the following table:

	Taking Stats	Not Taking Stats	Total
Male	84	145	229
Female	76	134	210
Total	160	279	439

P(Male or Stat) = P(M) + P(S) – P(M AND S)

= 229/439 + 160/439 – 84/439 = 305/439

Conditional Probability

- A conditional probability is the probability of one event, given that another event has occurred:

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

The conditional probability of A given that B has occurred

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

The conditional probability of B given that A has occurred

Where $P(A \text{ and } B)$ = joint probability of A and B

$P(A)$ = marginal probability of A

$P(B)$ = marginal probability of B

Computing Conditional Probability

- Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both.
- What is the probability that a car has a CD player, given that it has AC ?
- We want to find P(CD | AC).

Computing Conditional Probability

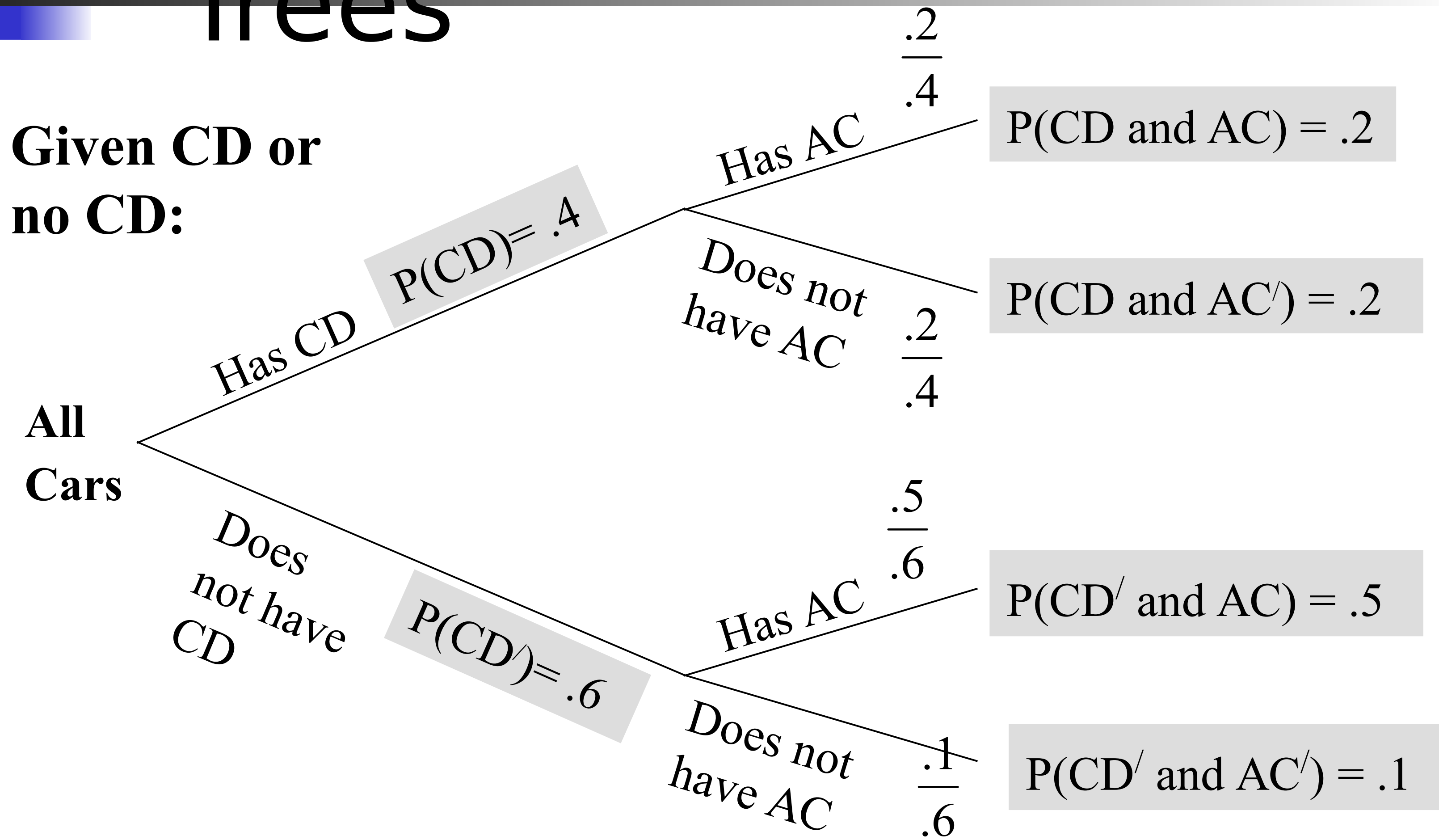
	CD	No CD	Total
AC	0.2	0.5	0.7
No AC	0.2	0.1	0.3
Total	0.4	0.6	1.0

Given

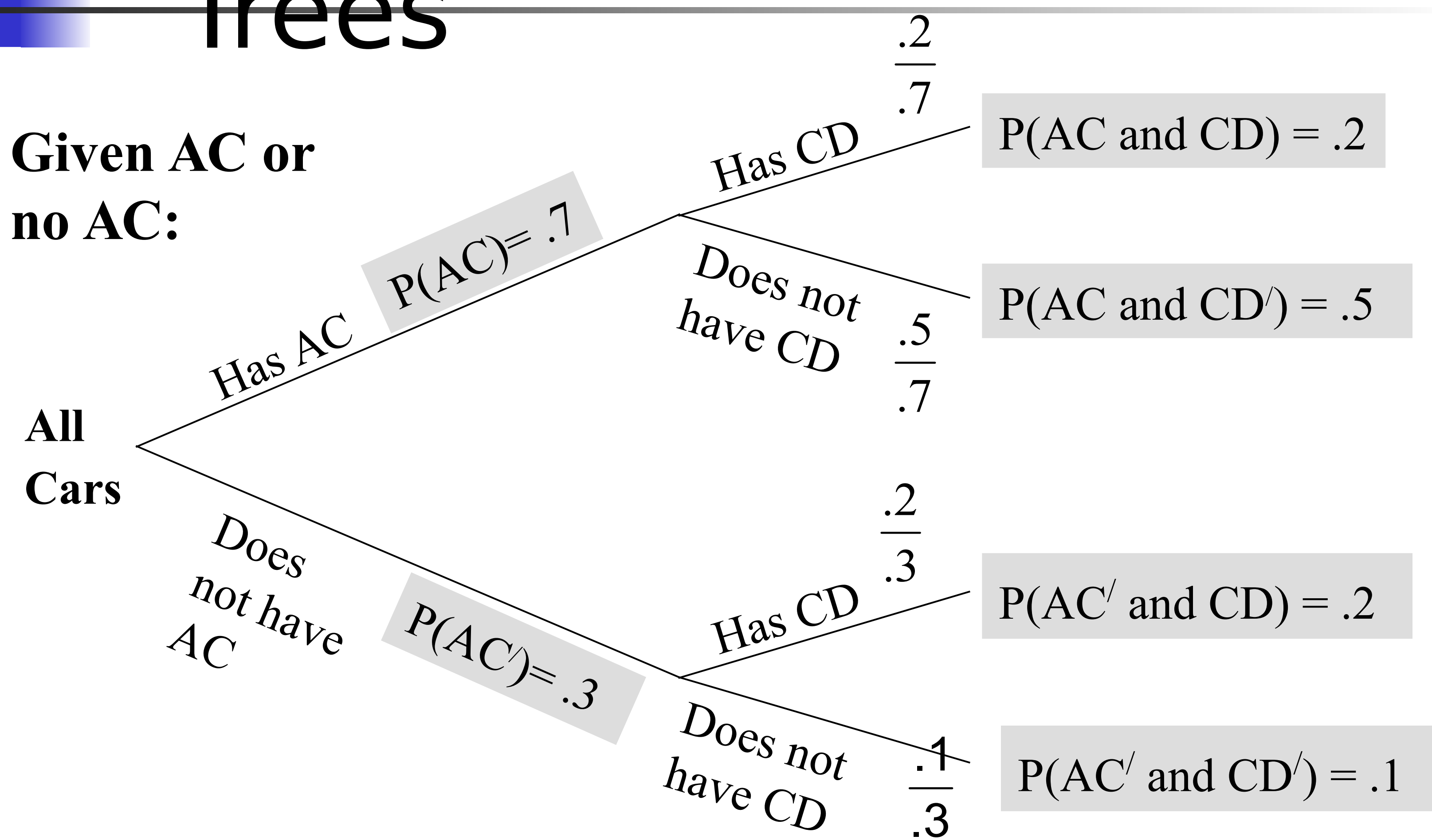
$$P(CD \mid AC) = \frac{P(CD \text{ and } AC)}{P(AC)} = \frac{.2}{.7} = .2857$$

Given AC, we only consider the top row (70% of the cars). Of these, 20% have a CD player. 20% of 70% is about 28.57%.

Computing Conditional Probability: Decision Trees



Computing Conditional Probability: Decision Trees



Statistical Independence

- Two events are **independent** if and only if:

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

- Events A and B are independent when the probability of one event is not affected by the other event

Multiplication Rules

- Multiplication rule for two events A and B:

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

- If A and B are independent, then

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

and the multiplication rule simplifies to:

$$P(A \text{ and } B) = P(A)\,P(B)$$

Multiplication Rules

- Suppose a city council is composed of 5 democrats, 4 republicans, and 3 independents. Find the probability of randomly selecting a democrat followed by an independent.

$$P(\text{I and D}) = P(\text{I} \mid \text{D})\,P(\text{D}) = (3/11)(5/12) = 5/44 = .114$$

- Note that after the democrat is selected (out of 12 people), there are only 11 people left in the sample space.

Marginal Probability Using Multiplication Rules

- Marginal probability for event A:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)$$

  - Where B_1, B_2, …, B_k are k mutually exclusive and collectively exhaustive events

Bayes' Theorem

- Bayes' Theorem is used to revise previously calculated probabilities based on new information.
- Developed by Thomas Bayes in the 18^{th} Century.
- It is an extension of conditional probability.

Bayes' Theorem

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \ldots + P(A \mid B_k)P(B_k)}$$

where:

B_i = i^{th} event of k mutually exclusive and collectively exhaustive events

A = new event that might impact $P(B_i)$

Bayes' Theorem Example

- A drilling company has estimated a 40% chance of striking oil for their new well.
- A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests.
- Given that this well has been scheduled for a detailed test, what is the probability that the well will be successful?

Bayes' Theorem Example

- Let S = successful well

  U = unsuccessful well
- $P(S) = .4$, $P(U) = .6$ (prior probabilities)
- Define the detailed test event as D
- Conditional probabilities:
  - $P(D|S) = .6$ $P(D|U) = .2$
- **Goal: To find $P(S|D)$**

Bayes' Theorem Example

Apply Bayes' Theorem:

$$
\begin{aligned}
P(S \mid D) &= \frac{P(D \mid S)P(S)}{P(D \mid S)P(S) + P(D \mid U)P(U)} \\
&= \frac{(.6)(.4)}{(.6)(.4) + (.2)(.6)} \\
&= \frac{.24}{.24 + .12} = .667
\end{aligned}
$$

So, the revised probability of success, given that this well has been scheduled for a detailed test, is .667

Bayes' Theorem Example

- Given the detailed test, the revised probability of a successful well has risen to .667 from the original estimate of 0.4.

Event	Prior Prob.	Conditional Prob.	Joint Prob.	Revised Prob.
S (successful)	.4	.6	.4*.6 = .24	.24/.36 = .667
U (unsuccessful)	.6	.2	.6*.2 = .12 Σ= .36	.12/.36 = .333

Review Question 1

If you toss a die, what’s the probability that you roll a 3 or less?

a. 1/6
b. 1/3
c. 1/2
d. 5/6
e. 1.0

Review Question 1

If you toss a die, what's the probability that you roll a 3 or less?

a. 1/6
b. 1/3
c. 1/2
d. 5/6
e. 1.0

Review Question 2

Two dice are rolled and the sum of the face values is six? What is the probability that at least one of the dice came up a 3?

a. 1/5
b. 2/3
c. 1/2
d. 5/6
e. 1.0

Review Question 2

Two dice are rolled and the sum of the face values is six. What is the probability that at least one of the dice came up a 3?

a. **1/5**
b. 2/3
c. 1/2
d. 5/6
e. 1.0

How can you get a 6 on two dice? 1-5, 5-1, 2-4, 4-2, 3-3

One of these five has a 3.

∴1/5

Important discrete probability distribution: The binomial

Binomial Probability Distribution

- A fixed number of observations (trials), n
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
  - e.g., head or tail in each toss of a coin; disease or no disease
  - Generally called “success” and “failure”
  - Probability of success is p, probability of failure is 1 – p
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same

Binomial distribution

Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

Binomial distribution

Solution:

One way to get exactly 3 heads: HHHTT

What's the probability of this <u>exact</u> arrangement?

P(heads)xP(heads) xP(heads)xP(tails)xP(tails) $=(1/2)^3 \times (1/2)^2$

Another way to get exactly 3 heads: THHHT

Probability of this exact outcome = $(1/2)^1 \times (1/2)^3 \times (1/2)^1 = (1/2)^3 \times (1/2)^2$

Binomial distribution

In fact, $(1/2)^3 \times (1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:

$(1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2$ $+$ ….. for as many unique arrangements as there are—but how many are there??

Outcome	Probability
THHHT	$(1/2)^3 \; x \; (1/2)^2$
HHHTT	$(1/2)^3 \; x \; (1/2)^2$
TTHHH	$(1/2)^3 \; x \; (1/2)^2$
HTTHH	$(1/2)^3 \; x \; (1/2)^2$
HHTTH	$(1/2)^3 \; x \; (1/2)^2$
HTHHT	$(1/2)^3 \; x \; (1/2)^2$
THTHH	$(1/2)^3 \; x \; (1/2)^2$
HTHTH	$(1/2)^3 \; x \; (1/2)^2$
HHTHT	$(1/2)^3 \; x \; (1/2)^2$
THHTH	$(1/2)^3 \; x \; (1/2)^2$

10 arrangements $x \; (1/2)^3 \; x \; (1/2)^2$

$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

The probability of each unique outcome (note: they are all equal)

${}_5C_3 = 5!/3!2! \; = 10$

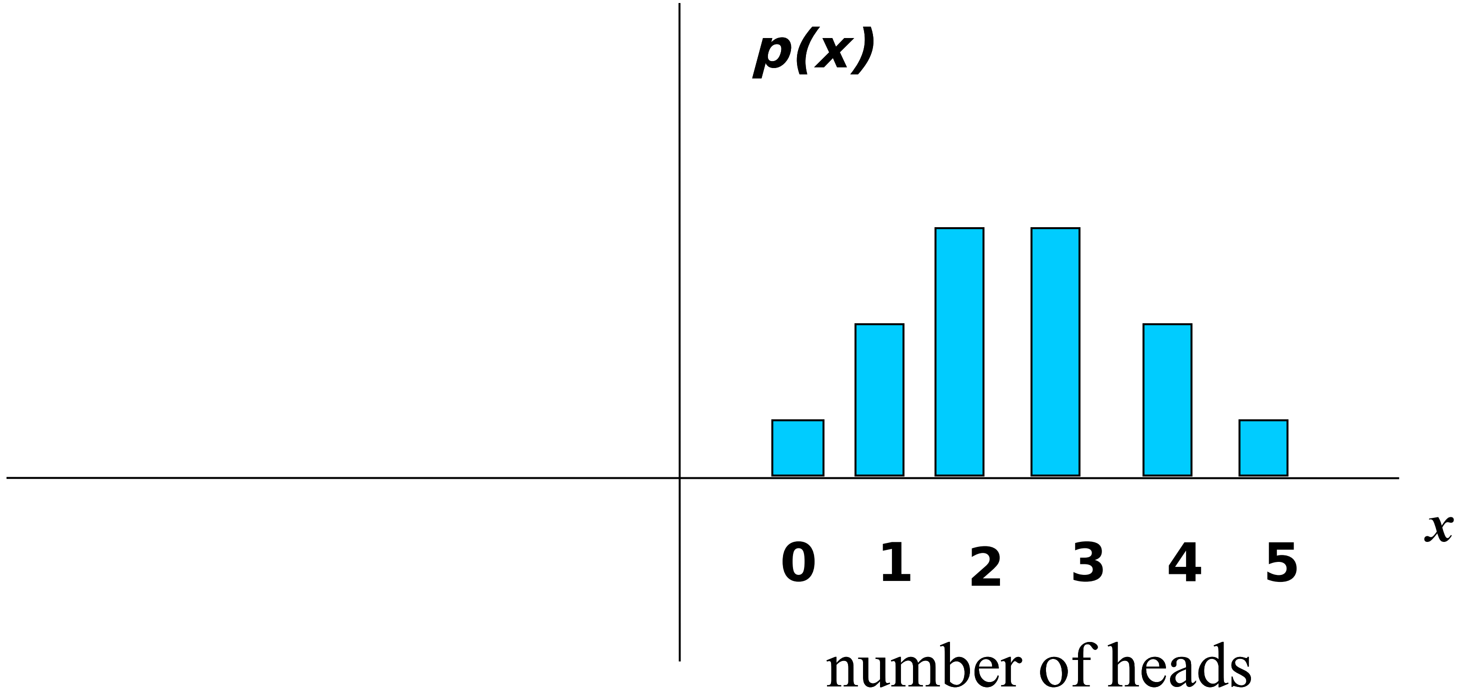
Factorial review: n! = n(n-1)(n-2)…

- **P(3 heads and 2 tails) =** $\binom{5}{3}$ ***x P(heads)³ x P(tails)² =***

10 x (½)⁵=31.25%

Binomial distribution function:

X= the number of heads tossed in 5 coin tosses



Binomial distribution, generally

Note the general pattern emerging → if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in n independent trials, then the probability of exactly X "successes"=

n = number of trials

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

X = # successes out of n trials

p = probability of success

$1-p$ = probability of failure

Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting exactly 10 heads?

$$\binom{20}{10}(.5)^{10}(.5)^{10} = .176$$

Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting of getting 2 or fewer heads?

$$\binom{20}{0}(.5)^0(.5)^{20}=\frac{20!}{20!0!}(.5)^{20}=9.5x10^{-7}+$$

$$\binom{20}{1}(.5)^1(.5)^{19}=\frac{20!}{19!1!}(.5)^{20}=20x9.5x10^{-7}=1.9x10^{-5}+$$

$$\binom{20}{2}(.5)^2(.5)^{18}=\frac{20!}{18!2!}(.5)^{20}=190x9.5x10^{-7}=1.8x10^{-4}$$

$$¿1.8x10^{-4}$$

**All probability distributions are characterized by an expected value and a variance:

<u>If</u> *X* follows a binomial distribution with parameters *n* and *p*: ***X ~ Bin (n, p)***

<u>Then:</u>

$E(X) = np$

$Var\ (X) = np(1-p)$

$\sqrt{np(1-p)}$

$SD\ (X) =$

Note: the variance will always lie between

0*N-.25 *N

p(1-p) reaches maximum at p=.5

P(1-p)=.25

Practice Problem

- 1. You are performing a cohort study. If the probability of developing disease in the exposed group is .05 for the study duration, then if you (randomly) sample 500 exposed people, how many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.

- 2. What's the probability that **at most** 10 exposed people develop the disease?

Answer

1. How many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.

$X \sim \text{binomial}(500, .05)$

$E(X) = 500\,(.05) = 25$

$Var(X) = 500\,(.05)\,(.95) = 23.75$

$StdDev(X) = \text{square root}\,(23.75) = 4.87$

$\therefore 25 \pm 4.87$

Answer

2. What's the probability that **at most** 10 exposed subjects develop the disease?

This is asking for a CUMULATIVE PROBABILITY: the probability of 0 getting the disease or 1 or 2 or 3 or 4 or up to 10.

$P(X \leq 10) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + \ldots + P(X=10) =$

$$\binom{500}{0}(.05)^{0}(.95)^{500} + \binom{500}{1}(.05)^{1}(.95)^{499} + \binom{500}{2}(.05)^{2}(.95)^{498} + \ldots + \binom{500}{10}(.05)^{10}(.95)^{490} < .01$$

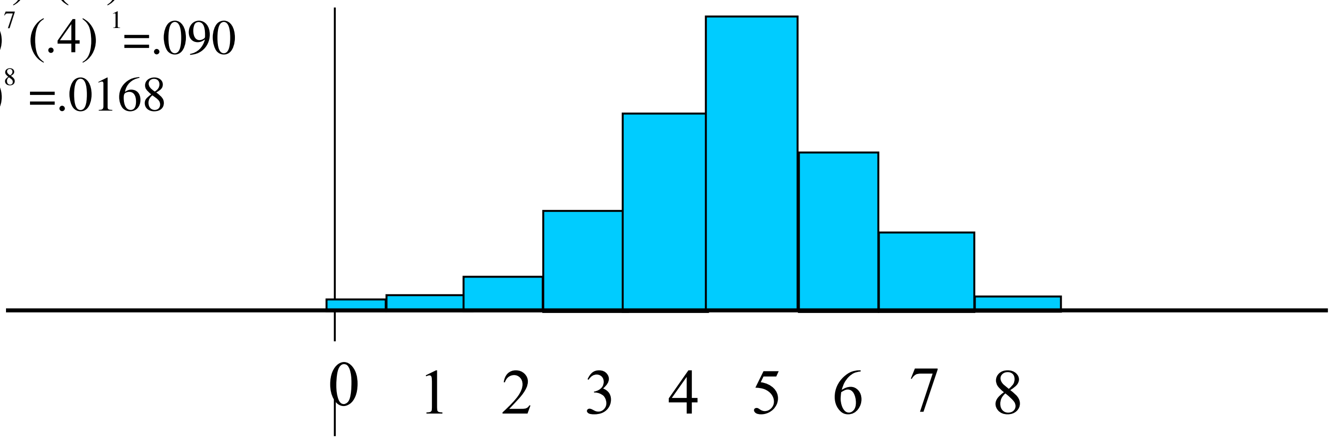
Practice Problem:

You are conducting a case-control study of smoking and lung cancer. If the probability of being a smoker among lung cancer cases is .6, what's the probability that in a group of 8 cases you have:

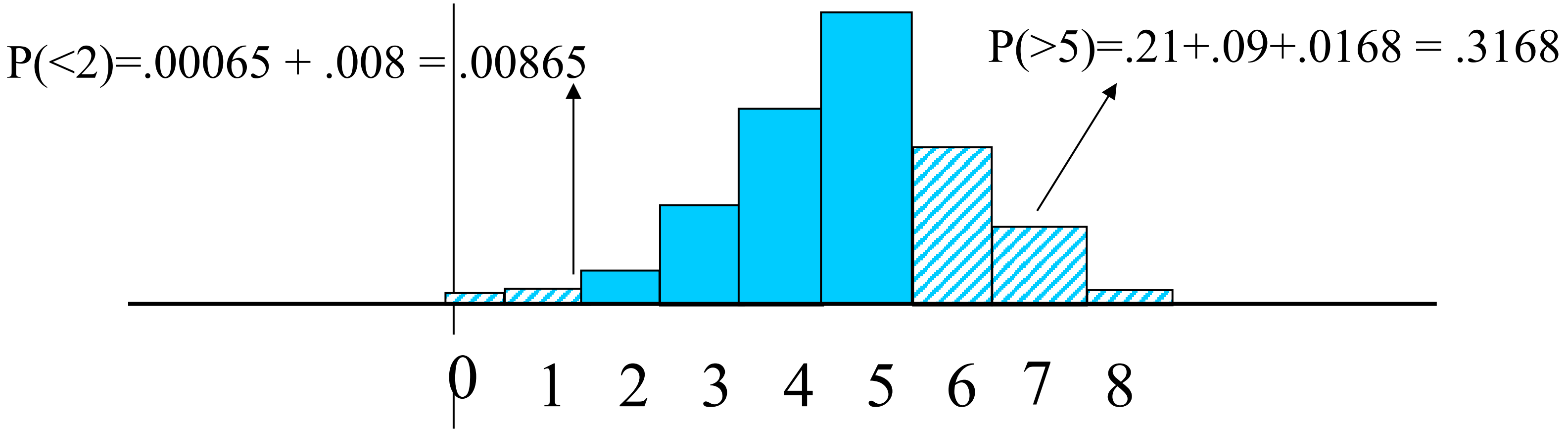
a. Less than 2 smokers?
b. More than 5?
c. What are the expected value and variance of the number of smokers?

Answer

X	P(X)
0	$1(.4)^8 = .00065$
1	$8(.6)^1\ (.4)^7 = .008$
2	$28(.6)^2\ (.4)^6 = .04$
3	$56(.6)^3\ (.4)^5 = .12$
4	$70(.6)^4\ (.4)^4 = .23$
5	$56(.6)^5\ (.4)^3 = .28$
6	$28(.6)^6\ (.4)^2 = .21$
7	$8(.6)^7\ (.4)^1 = .090$
8	$1(.6)^8 = .0168$



Answer, continued



$E(X) = 8\ (.6) = 4.8$

$Var(X) = 8\ (.6)\ (.4) = 1.92$

$StdDev(X) = 1.38$

Review Question 4

In your case-control study of smoking and lung-cancer, 60% of cases are smokers versus only 10% of controls. What is the odds ratio between smoking and lung cancer?

a. 2.5
b. 13.5
c. 15.0
d. 6.0
e. .05

Review Question 4

In your case-control study of smoking and lung-cancer, 60% of cases are smokers versus only 10% of controls. What is the odds ratio between smoking and lung cancer?

a. 2.5
b. **13.5**
c. 15.0
d. 6.0
e. .05

$$\frac{\frac{.6}{.4}}{\frac{.1}{.9}} = \frac{3}{2} x \frac{9}{1} = \frac{27}{2} = 13.5$$

Review Question 5

What's the probability of getting exactly 5 heads in 10 coin tosses?

a. $\binom{10}{0}(.50)^5(.50)^5$

b. $\binom{10}{5}(.50)^5(.50)^5$

c. $\binom{10}{5}(.50)^{10}(.50)^5$

d. $\binom{10}{10}(.50)^{10}(.50)^0$

Review Question 5

What's the probability of getting exactly 5 heads in 10 coin tosses?

a. $\binom{10}{0}(.50)^5(.50)^5$

b. $\binom{10}{5}(.50)^5(.50)^5$

c. $\binom{10}{5}(.50)^{10}(.50)^5$

d. $\binom{10}{10}(.50)^{10}(.50)^0$

Review Question 6

A coin toss can be thought of as an example of a binomial distribution with N=1 and p=.5. What are the expected value and variance of a coin toss?

a. .5, .25
b. 1.0, 1.0
c. 1.5, .5
d. .25, .5
e. .5, .5

Review Question 6

A coin toss can be thought of as an example of a binomial distribution with N=1 and p=.5. What are the expected value and variance of a coin toss?

a. **.5, .25**
b. 1.0, 1.0
c. 1.5, .5
d. .25, .5
e. .5, .5

Review Question 7

If I toss a coin 10 times, what is the expected value and variance of the number of heads?

a. 5, 5
b. 10, 5
c. 2.5, 5
d. 5, 2.5
e. 2.5, 10

Review Question 7

If I toss a coin 10 times, what is the expected value and variance of the number of heads?

a. 5, 5
b. 10, 5
c. 2.5, 5
d. 5, 2.5
e. 2.5, 10

Review Question 8

In a randomized trial with n=150, the goal is to randomize half to treatment and half to control. The number of people randomized to treatment is a random variable X. What is the probability distribution of X?

a. X~Normal(μ=75,σ=10)
b. X~Exponential(μ=75)
c. X~Uniform
d. X~Binomial(N=150, p=.5)
e. X~Binomial(N=75, p=.5)

Review Question 8

In a randomized trial with n=150, every subject has a 50% chance of being randomized to treatment. The number of people randomized to treatment is a random variable X. What is the probability distribution of X?

a. X~Normal(μ=75,σ=10)
b. X~Exponential(μ=75)
c. X~Uniform
d. X~Binomial(N=150, p=.5)
e. X~Binomial(N=75, p=.5)

Review Question 9

In the same RCT with n=150, if 69 end up in the treatment group and 81 in the control group, how far off is that from expected?

a. Less than 1 standard deviation
b. 1 standard deviation
c. Between 1 and 2 standard deviations
d. More than 2 standard deviations

Review Question 9

In the same RCT with n=150, if 69 end up in the treatment group and 81 in the control group, how far off is that from expected?

a. Less than 1 standard deviation
b. 1 standard deviation
c. Between 1 and 2 standard deviat
d. More than 2 standard deviations

Expected = 75

81 and 69 are both 6 away from the expected.

Variance = 150(.25) = 37.5

Std Dev ≅ 6

Therefore, about 1 SD away from expected.

Proportions…

- The binomial distribution forms the basis of statistics for proportions.
- A proportion is just a binomial count divided by n.
  - For example, if we sample 200 cases and find 60 smokers, X=60 but the observed proportion=.30.
- Statistics for proportions are similar to binomial counts, but differ by a factor of n.

Stats for proportions

For binomial:

$$\mu_x = np$$

$$\sigma_x^2 = np(1-p)$$

$$\sigma_x = \sqrt{np(1-p)}$$

Differs by a factor of n.

Differs by a factor of n.

For proportion:

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}}^2 = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

P-hat stands for “sample proportion.”

It all comes back to normal…

- Statistics for proportions are based on a normal distribution, because the binomial can be approximated as normal if np>5

Multinomial distribution
(beyond the scope of this course)

The multinomial is a generalization of the binomial. It is used when there are more than 2 possible outcomes (for ordinal or nominal, rather than binary, random variables).

- Instead of partitioning *n* trials into 2 outcomes (yes with probability *p* / no with probability *1-p*), you are partitioning *n* trials into 3 or more outcomes (with probabilities: $p_1, p_2, p_3, \ldots$)
  - General formula for 3 outcomes:

$$P(D=x, R=y, G=z) = \frac{n!}{x!\,y!\,z!} p_D^x p_R^y (1 - p_D - p_R)^z$$

Multinomial example

Specific Example: if you are randomly choosing 8 people from an audience that contains 50% democrats, 30% republicans, and 20% green party, what's the probability of choosing exactly 4 democrats, 3 republicans, and 1 green party member?

$$P(D=4, R=3, G=1) = \frac{8!}{4!\,3!\,1!}(.5)^4(.3)^3(.2)^1$$

You can see that it gets hard to calculate very fast! The multinomial has many uses in genetics where a person may have 1 of many possible alleles (that occur with certain probabilities in a given population) at a gene locus.

Introduction to the **Poisson Distribution**

- Poisson distribution is for counts—if events happen at a constant rate over time, the Poisson distribution gives the probability of X number of events occurring in time T.

Poisson Mean and Variance

- Mean $$\mu = \lambda$$

- Variance and Standard Deviation

$$\sigma^2 = \lambda$$

$$\sigma = \sqrt{\lambda}$$

For a Poisson random variable, the variance and mean are the same!

where λ = expected number of hits in a given time period

Poisson Distribution, example

The Poisson distribution models counts, such as the number of new cases of SARS that occur in women in New England next month.

The distribution tells you the probability of all possible numbers of new cases, from 0 to infinity.

If X= # of new cases next month and $X \sim$ Poisson (λ), then the probability that $X=k$ (a particular count) is:

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

Example

- For example, if new cases of West Nile Virus in New England are occurring at a rate of about 2 per month, then these are the probabilities that: 0,1, 2, 3, 4, 5, 6, to 1000 to 1 million to… cases will occur in New England in the next month:

Poisson Probability table

X	P(X)
0	$\frac{2^0 e^{-2}}{0!}$ =.135
1	$\frac{2^1 e^{-2}}{1!}$ =.27
2	$\frac{2^2 e^{-2}}{2!}$ =.27
3	$\frac{2^3 e^{-2}}{3!}$ =.18
4	=.09
5	
…	…

Example: Poisson distribution

Suppose that a rare disease has an incidence of 1 in 1000 person-years. Assuming that members of the population are affected independently, find the probability of k cases in a population of 10,000 (followed over 1 year) for k=0,1,2.

The expected value (mean) =λ = .001*10,000 = 10
10 new cases expected in this population per year→

$$P(X=0)=\frac{(10)^0 e^{-(10)}}{0!}=.0000454$$

$$P(X=1)=\frac{(10)^1 e^{-(10)}}{1!}=.000454$$

$$P(X=2)=\frac{(10)^2 e^{-(10)}}{2!}=.00227$$

more on Poisson…

"Poisson Process" (rates)

Note that the Poisson parameter λ can be given as the mean number of events that occur in a defined time period OR, equivalently, λ can be given as a rate, such as λ=2/month (2 events per 1 month) that must be multiplied by *t*=time (called a "Poisson Process") →

X ~ Poisson (λ)

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

E(X) = λt

Var(X) = λt

Example

For example, if new cases of West Nile in New England are occurring at a rate of about 2 per month, then what's the probability that exactly 4 cases will occur in the next 3 months?

X ~ Poisson (λ=2/month)

$$P(X = 4 \text{ in } 3 \text{ months}) = \frac{(2*3)^4 e^{-(2*3)}}{4!} = \frac{6^4 e^{-(6)}}{4!} = 13.4\%$$

Exactly 6 cases?

$$P(X = 6 \text{ in } 3 \text{ months}) = \frac{(2*3)^6 e^{-(2*3)}}{6!} = \frac{6^6 e^{-(6)}}{6!} = 16\%$$

Practice problems

1a. If calls to your cell phone are a Poisson process with a constant rate λ=2 calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour movie, your phone rings during that time?

1b. How many phone calls do you expect to get during the movie?

Answer

1a. If calls to your cell phone are a Poisson process with a constant rate λ=2 calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour movie, your phone rings during that time?

X ~ Poisson (λ=2 calls/hour)

$P(X \geq 1) = 1 - P(X=0)$

$$P(X=0) = \frac{(2*1.5)^0 e^{-2(1.5)}}{0!} \frac{(3)^0 e^{-3}}{0!} = e^{-3} = .05$$

$\therefore P(X \geq 1) = 1 - .05 = 95\%$ chance

1b. How many phone calls do you expect to get during the movie?

$E(X) = \lambda t = 2(1.5) = 3$