MSIT 431 Probability and Statistical Methods

Chapter 4. Probability: The Study of Randomness

Dongning Guo

Fall 2017

Chapter 4 Probability: The Study of Randomness

4.1 Randomness

4.2 Probability Models

4.3 Random Variables

4.4 Means and Variances of Random Variables

4.5 General Probability Rules* (we discuss this section's material along with materials of 4.1-4.4)

2

4.1 Randomness

- The language of probability
- Thinking about randomness
- The uses of probability

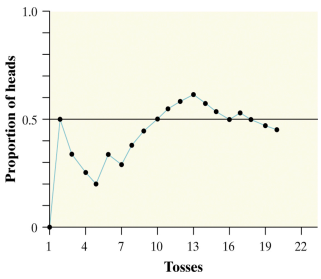
3

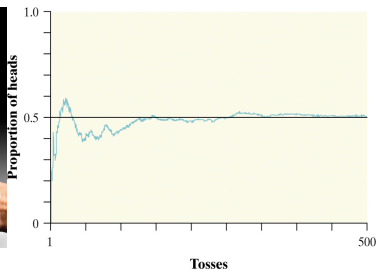
The Language of Probability

Chance behavior is unpredictable in the short run but has a regular and predictable pattern in the long run.

We call a phenomenon **random** if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

The **probability** of any outcome of a chance process is the proportion of times the outcome would occur in a very long series of repetitions.

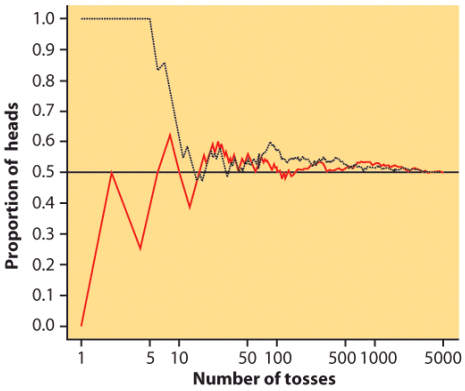




4

Thinking About Randomness

The result of any single coin toss is random. But the result over many tosses is predictable, as long as the trials are **independent** (i.e., the outcome of a new coin flip is not influenced by the result of the previous flip).



The probability of heads is 0.5 = the proportion of times you get heads in many repeated trials.

5

View 1: Classical probability

- If an experiment has *N* equally likely outcomes, and an event *E* consists of *M* of the outcomes, then *P(E) = M/N*.
- Examples:
  - A fair coin toss has two equally likely outcomes, Head and Tail. The event Head has probability 1/2.
  - 3 different letters are randomly put into 3 distinctly addressed envelops, one letter in each envelop. What is the probability that all letters are found in their corresponding envelop? The number of different permutations of 3 letters is 6, all equally likely, so the answer is 1/6.
- Classical probability has difficulty with the situation where the outcomes are not equally likely. For example, what is the probability that a given flight will be on time?

6

View 2: Frequencist's view

- The probability of an event is the long-term relative frequency of the event.
- Examples: Coin flips, free-throw percentage, hitting average... The frequency can be computed using empirical data.
- The limitation of this view is due to the fact that the frequency obtained from any given number of trials is only an estimate of the limit.

7

View 3: Subjective probability

- Many experiments cannot be repeated. Or at least repeating is irrelevant to a specific situation, e.g., should someone take a heart transplant with potential risks and benefits? It may be a life or death problem in a single operation.
- Probability is how likely you believe an event will happen (there may or may not be scientific basis for the belief).
- Examples:
  - Probability that I win a lottery tomorrow.
  - Probability that I will get an A for MSIT 431.
  - Probability that someone with cancer survives a treatment.
  - A game of betting $1 on the outcome of a particular dice for the return $4.
- Question: Would you bet? Perhaps it depends on how you believe the fairness of the dice, and how you value win and loss?

8

View 4: Axiomatic definition of probability

- A unifying probability theory has to wait until 1930s to be established by Kolmogorov. Probability is a relatively young branch of mathematics.
- Kolmogorov noticed that the common ground of the classical and the frequencist's perspectives is "the additively" of probability. Namely:

  P(E happens) + P(F happens) = P(E or F happens)

- If E and F do not happen at the same time. P can be frequency or belief.
- We shall adopt this view.
- More on this later.

9

4.2 Probability Models

- Sample spaces
- Probability rules
- Assigning probabilities
- Independence and the multiplication rule

10

Probability Models

Descriptions of chance behavior contain two parts: a list of possible outcomes and a probability for each outcome.

The **sample space *S*** of a chance process is the set of all possible outcomes.

An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.

A **probability model** is a description of some chance process that consists of two parts: a sample space *S* and a probability for each outcome.

11

Experiments and events

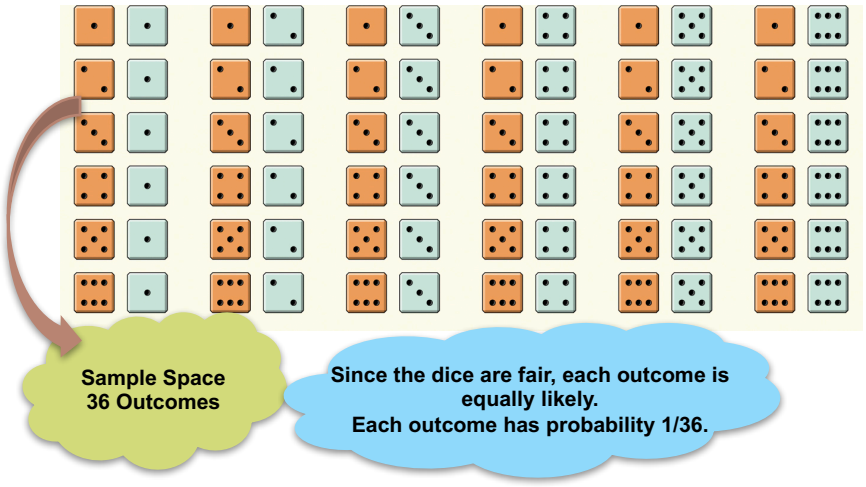
- Example 1. Experiment: To roll a dice.
  Outcomes: 1, 2, 3, 4, 5, 6.
  Events: get an even number {2,4,6}, get a prime number {2,3,5}, get a positive number {1,2,3,4,5,6}
  Total number of different events: $2^6 = 64$.
- We are often interested in the probability of an event, e.g., what is the probability of throwing an odd number?
- Example 2. A = { it's going to rain today } (event)

  The experiment (the weather system) and its two possible outcomes (rain, no rain).
- Before an experiment is carried out, we already know all possible outcomes (although we do not know which one will actually take place), all the events, and the probability of all the events.

12

Probability Models

Example: Give a probability model for the chance process of rolling two fair, six-sided dice—one that's red and one that's green.



13

Classical probability and its calculation

- Combinatorics
  - n distinct items, the number of permutations is *n!=nx(n-1)x(n-2)x…x1*.
  - Choose r out of n distinct items, the number of permutations is *n!/(n-r)!*.
  - Choose r out of n distinct items, the number of combinations is *n!/(n-r)!/r!*, also denoted as $\binom{n}{r}$
- Example: n different letters are randomly put into n distinctly addressed envelops, one letter in each envelop. What is the probability that all letters are found in their corresponding envelops?
- The answer is 1/(n!) because there are n! equally likely permutations of n letters, where there is only one correct permutation.

14

Illinois Powerball

- Choose 5 distinct numbers out of 1-69 (white balls) and an additional powerball number out of 1-26 (the red ball).
- Consider all possible combinations of 5 distinct white balls and a red ball. How many possible outcomes are there?

$$N = \binom{69}{5} \times 26 = 292{,}201{,}338$$

- Are they equally likely? Yes, they are.
- The chance of winning the first prize is *1/N*.
- The second prize ($1 million): all five white balls match but the red ball misses. What is the probability?
- The number of winning outcomes is 26-1=25. So the probability is 25/N.

15

Where does the lottery dollar go?

- 59% - Prizes paid to winners;
- 30% - Common school fund, capital projects fund, and specialty tickers;
- 11% - Retailer and vendor commissions and other expenses.
- If the average return is only 60%, why do people buy lottery?

16

- The 2017 National League Division Series consist of two best-of-five-game series played in the Major League Baseball postseason that will determine the participating teams of the 2017 National League Championship Series.
- The Cubs won the first game against Washington Nationals last night. Suppose the Cubs wins each game with probability 1/2 independent of other games.
- What is the probability that the Cubs wins in all the first 3 games?
- What is the probability that only 4 games are played to determine the winner?
- What is the probability that no more than 4 games are played?

17

Axioms of probability

Axiom 1. The probability $P(A)$ of any event A satisfies **$0 \le P(A) \le 1$.**

Any probability is a number between 0 and 1.

Axiom 2. If S is the sample space in a probability model, then **$P(S) = 1$.**

All possible outcomes together must have probability 1.

Axiom 3. If A and B are **disjoint,** **$P(A \text{ or } B) = P(A) + P(B)$.**

This is the **addition rule for disjoint events.**

If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities.

Corollary: The probability that an event does not occur is 1 minus the probability that the event does occur. The **complement** of any event A is the event that A does not occur, written A^C. **$P(A^C) = 1 - P(A)$.**

18

Probability Rules

Randomly select an undergraduate student who is taking distance-learning courses for credit and record the student's age. Here is the probability model:

Age group (yr):	18 to 23	24 to 29	30 to 39	40 or over
Probability:	0.57	0.17	0.14	0.12

(a) Show that this is a legitimate probability model.

Each probability is between 0 and 1 and
0.57 + 0.17 + 0.14 + 0.12 = 1

(b) Find the probability that the chosen student is not in the traditional college age group (18 to 23 years).

***P*(not 18 to 23 years) = 1 – *P*(18 to 23 years)**
= 1 – 0.57 = 0.43

19

Finite Probability Models

One way to assign probabilities to events is to assign a probability to every individual outcome, then add these probabilities to find the probability of any event. This idea works well when there are only a finite (fixed and limited) number of outcomes.

A probability model with a finite sample space is called **finite.**

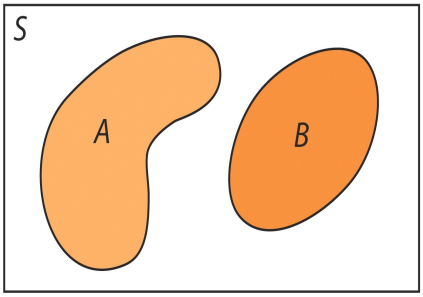
To assign probabilities in a finite model, list the probabilities of all the individual outcomes. These probabilities must be numbers between 0 and 1 that add up to exactly 1. The probability of any event is the sum of the probabilities of the outcomes making up the event.

20

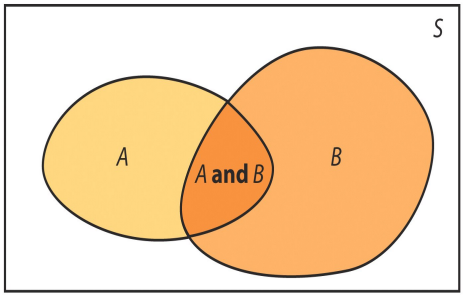
Venn Diagrams

Sometimes, it is helpful to draw a picture to display relations among several events. A picture that shows the sample space *S* as a rectangular area and events as areas within *S* is called a **Venn diagram.**

Two disjoint events:



Two events that are not disjoint, and the event "*A* and *B*" consisting of the outcomes they have in common:

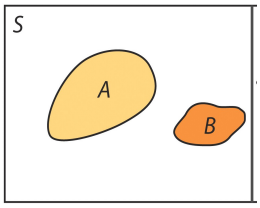


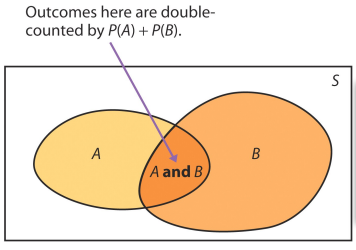
21

The General Addition Rule

Addition Rule for Disjoint Events
If *A*, *B,* and *C* are **disjoint** in the sense that no two have any outcomes in common, then:

***P*(one or more of *A*, *B, C*) = *P*(*A*) + *P*(*B*) + *P*(*C*)**





Addition Rule for Unions of Two Events
For any two events *A* and *B*:

***P*(*A* or *B*) = *P*(*A*) + *P*(*B*) – *P*(*A* and *B*)**

22

Multiplication Rule for Independent Events

If two events *A* and *B* do not influence each other, and if knowledge about one does not change the probability of the other, the events are said to be **independent** of each other.

Definition: Two events *A* and *B* are said to be **independent** if
***P*(*A* and *B*) = *P*(*A*) × *P*(*B*)**

Caution: Disjoint events are dependent in general! In fact if one event happens, the other cannot happen.

23

Conditional Probability

The probability we assign to an event can change if we know that some other event has occurred. This idea is the key to many applications of probability.

The probability that one event happens given that another event is already known to have happened is called a **conditional probability.**

When $P(A) > 0$, the probability that event *B* happens *given* that event *A* has happened is found by:

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

24

The General Multiplication Rule

The definition of conditional probability reminds us that in principle all probabilities, including conditional probabilities, can be found from the assignment of probabilities to events that describe a random phenomenon. The definition of conditional probability then turns into a rule for finding the probability that both of two events occur.

The probability that events A and B both occur can be found using the **general multiplication rule:**

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

where $P(B \mid A)$ is the conditional probability that event B occurs given that event A has already occurred.

Note: Two events A and B that both have positive probability are **independent** if:

$$P(B|A) = P(B)$$

25

Tree Diagrams

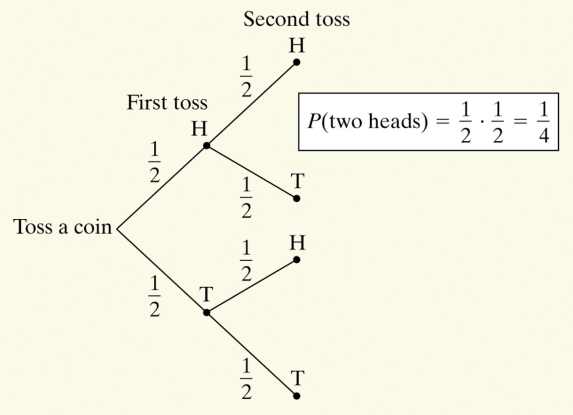
Probability problems often require us to combine several of the basic rules into a more elaborate calculation. One way to model chance behavior that involves a sequence of outcomes is to construct a **tree diagram.**

Consider flipping a coin twice.

What is the probability of getting two heads?

Sample Space:
HH HT TH TT

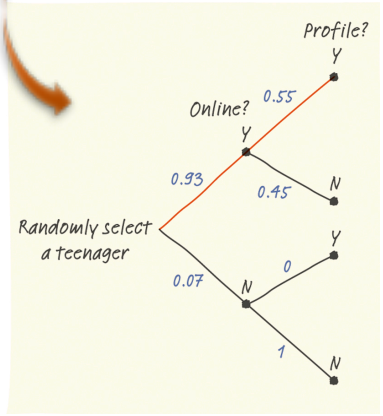
So, *P*(two heads) = *P*(HH) = 1/4



26

Example

The Pew Internet and American Life Project finds that 93% of teenagers (ages 12 to 17) use the Internet, and that 55% of online teens have posted a profile on a social-networking site.
What percent of teens are online *and* have posted a profile?



$$P(\text{online}) = 0.93$$
$$P(\text{profile} \mid \text{online}) = 0.55$$

$$P(\text{online and have profile}) = P(\text{online})\ P(\text{profile} \mid \text{online})$$
$$= (0.93)(0.55)$$
$$= 0.5115$$

51.15% of teens are online *and* have posted a profile.

27

Bayes's Rule

An important application of conditional probabilities is Bayes's rule. It is the foundation of many modern statistical applications.

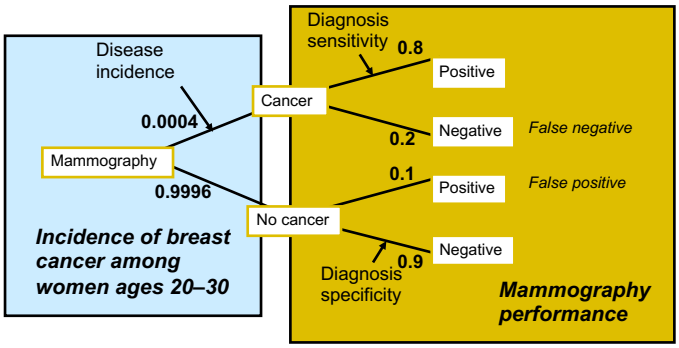
* Suppose that a sample space is decomposed into *k* disjoint events $A_1, A_2, \ldots, A_k$ —none of which has a 0 probability—such that $P(A_1) + P(A_2) + \ldots + P(A_k) = 1$,

- Let *C* be any other event such that $P(C)$ is not 0. Then

$$P(A_i \mid C) = \frac{P(C \mid A_i)P(A_i)}{P(C \mid A_1)P(A_1) + P(C \mid A_2)P(A_2) + \cdots + P(C \mid A_k)P(A_k)}$$

28

Example



If a woman in her 20s gets screened for breast cancer and receives a positive test result, what is the probability that she does have breast cancer?

$$P(\text{cancer}|\text{pos}) = \frac{P(\text{pos}|\text{cancer})P(\text{cancer})}{P(\text{pos} \mid \text{cancer})P(\text{cancer}) + P(\text{pos} \mid \text{no cancer})P(\text{no cancer})}$$
$$= \frac{0.8(0.0004)}{0.8(0.004) + 0.1(0.9996)} \approx 0.3\%$$

29

Bayesian spam filtering

- Particular words have particular probabilities of occurring in spam email and in legitimate email. For instance, most email users will frequently encounter the word ”Viagra” in spam email, but will seldom see it in other email. The filter doesn't know these probabilities in advance, and must first be trained so it can build them up. To train the filter, the user (or the designer) must manually indicate whether a new email is spam or not. For all words in each training email, the filter will adjust the probabilities that each word will appear in spam or legitimate email in its database. For instance, Bayesian spam filters will typically have learned a very high spam probability for the words ”Viagra” and ”refinance”, but a very low spam probability for words seen only in legitimate email, such as the names of friends and family members.
- Each suspicious word in the email contributes to the email's spam probability. The email's spam probability is computed over all words in the email, and if the total exceeds a certain threshold (say 95%), the filter will mark the email as a spam.
- Let's suppose the suspected message contains the word “replica.” Most people who are used to receiving e-mail know that this message is likely to be spam, more precisely a proposal to sell counterfeit copies of well-known brands of watches.

30

Spamicity

- Define events: S = {the email is spam} H = {the email is ham}
  W = {the email has word "replica" in it}
- The formula used by the software to determine that is derived from Bayes' theorem
  $P(S|W) = P(W|S)\,P(S) / (P(W|S)\,P(S) + P(W|H)\,P(H))$
- Statistics show that the current probability of any message being spam is 80%, at the very least. Most bayesian spam detection software makes the assumption that there is no a priori reason for any incoming message to be spam rather than ham, and considers both cases to have equal probabilities: P(S) = 0.5; P(H) = 0.5. That is, we assume the prior distribution to be uniform.
- Under the uniform prior assumption, the spamicity is:
  $P(S|W) = P(W|S) / (P(W|S) + P(W|H))$
  $= 1 / (1 + P(W|H)/P(W|S)).$

31

Multiple suspicious words

- Consider two words, W_1 and W_2. It is reasonable (and convenient) to assume W_1 and W2 are independent conditioned on either S or H. Thus $P(W_1\,W_2\,|S) = P(W_1\,|S)P(W_2\,|S)$, $P(W_1\,W_2\,|H) = P(W_1\,|H)P(W_2\,|H)$.
- Consider an email that contains both W_1 and W_2. What is the spamicity?
  $P(S|\,W_1\,W_2)$
  $= P(W_1\,W_2\,|S\,)P(S\,) / (P(W_1\,W_2\,|S)P(S) + P(W_1\,W_2\,|H)P(H))$
  $= P(W_1\,|S\,)P(W_2\,|S\,)P(S\,) / (P(W_1\,|S)P(W_2\,|S)P(S) + P(W_1\,|H)P(W_2\,|H)P(H))$
  $= 1 / (1 + (P(W_1\,|H)/P(W_1\,|S\,)) \times (P(W_2\,|H)/P(W_2\,|S\,)))$
- The preceding result can be generalized to n words.
- If the spamicity is lower than a certain threshold, the message is considered as likely ham, otherwise it is considered as likely spam.

32

4.3 Random Variables

- Random variable
- Discrete random variables
- Continuous random variables
- Normal distributions as probability distributions

33

Random Variables

A **probability model** describes the possible outcomes of a chance process and the likelihood that those outcomes will occur.

A numerical variable that describes the outcomes of a chance process is called a **random variable.** The probability model for a random variable is its probability distribution.

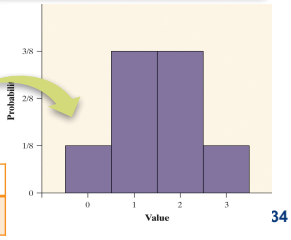
A **random variable** takes numerical values that describe the outcomes of some chance process.

The **probability distribution** of a random variable gives its possible values and their probabilities.

Example: Consider tossing a fair coin 3 times.
Define X = the number of heads obtained

X = 0: TTT
X = 1: HTT THT TTH
X = 2: HHT HTH THH
X = 3: HHH

Value	0	1	2	3
Probability	1/8	3/8	3/8	1/8



34

Discrete Random Variable

There are two main types of random variables: *discrete* and *continuous*. If we can find a way to list all possible outcomes for a random variable and assign probabilities to each one, we have a **discrete random variable.**

A **discrete random variable *X*** takes a fixed set of possible values with gaps between. The probability distribution of a discrete random variable *X* lists the values x_i and their probabilities p_i:

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

The probabilities p_i must satisfy two requirements:

1. Every probability p_i is a number between 0 and 1.

2. The sum of the probabilities is 1.

To find the probability of any event, add the probabilities p_i of the particular values x_i that make up the event.

35

Continuous Random Variable

Discrete random variables commonly arise from situations that involve counting something. Situations that involve measuring something often result in a **continuous random variable.**

A **continuous random variable *Y*** takes on all values in an interval of numbers. The probability distribution of *Y* is described by a **density curve.** The probability of any event is the area under the density curve and above the values of *Y* that make up the event.

The probability model of a discrete random variable *X* assigns a probability between 0 and 1 to each possible value of *X*.

A continuous random variable *Y* has *infinitely many* possible values. All continuous probability models assign probability 0 to every individual outcome. Only *intervals* of values have positive probability.

36

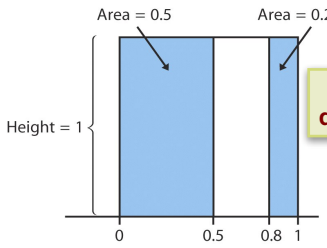
Continuous Probability Models

Suppose we want to choose a number at random between 0 and 1, allowing *any* number between 0 and 1 as the outcome. We cannot assign probabilities to each individual value because there is an infinite interval of possible values.

A **continuous probability model** assigns probabilities as areas under a density curve. The area under the curve and above any range of values is the probability of an outcome in that range.

Example: Find the probability of getting a random number that is less than or equal to 0.5 OR greater than 0.8.

$P(X \le 0.5 \text{ or } X > 0.8)$
$= P(X \le 0.5) + P(X > 0.8)$
$= 0.5 + 0.2$
$= 0.7$



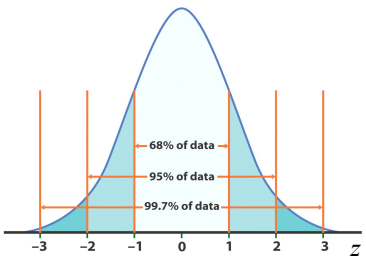
Uniform distribution

37

Normal Probability Models

- Often, the density curve used to assign probabilities to intervals of outcomes is the Normal curve.
- Probabilities can be assigned to intervals of outcomes using the Standard Normal probabilities in Table A.
- We **standardize** normal data by calculating *z*-scores so that any Normal curve $N(\mu,\sigma)$ can be transformed into the standard Normal curve $N(0,1)$.

$$z = \frac{x - \mu}{\sigma}$$



38

Normal Probability Models

Often the density curve used to assign probabilities to intervals of outcomes is the Normal curve.

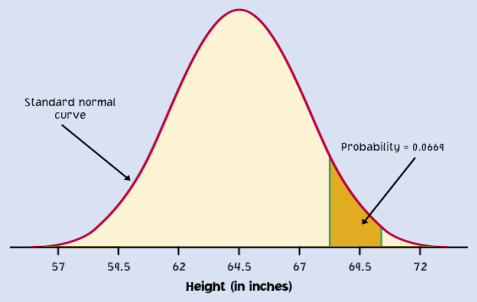
Women's heights are Normally distributed with mean 64.5 and standard deviation 2.5 in. If we pick one woman at random, what is the probability that her height will be between 68 and 70 inches $P(68 < X < 70)$? Because the woman is selected at random, X is a random variable.

$$z = \frac{x - \mu}{\sigma}$$

As before, we calculate the z-scores for 68 and 70.

For x = 68", $z = \frac{68 - 64.5}{2.5} = 1.4$

For x = 70", $z = \frac{70 - 64.5}{2.5} = 2.2$



39

4.4 Means and Variances of Random Variables

- The mean of a random variable
- The law of large numbers
- Rules for means
- The variance of a random variable
- Rules for variances and standard deviations

40

The Mean of a Random Variable

When analyzing discrete random variables, we'll follow the same strategy we used with quantitative data—describe the shape, center, and spread, and identify any outliers.

The mean of any discrete random variable is an average of the possible outcomes, with each outcome weighted by its probability.

Mean of a Discrete Random Variable

Suppose that X is a discrete random variable whose probability distribution is

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

To find the **mean (expected value)** of X, multiply each possible value by its probability, then add all the products:

$$E(X) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots$$
$$= \sum x_ip_i$$

41

Example: Babies' Health at Birth

The probability distribution for X = Apgar scores is shown below:

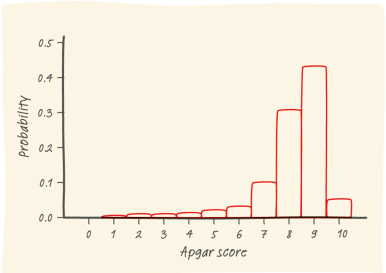
a. Show that the probability distribution for X is legitimate.

b. Make a histogram of the probability distribution. Describe what you see.

c. Apgar scores of 7 or higher indicate a healthy baby. What is $P(X \geq 7)$?

Value:	0	1	2	3	4	5	6	7	8	9	10
Probability:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

(a) All probabilities are between 0 and 1, and they add up to 1. This is a legitimate probability distribution.



(c) $P(X \geq 7) = .908$. We'd have a 91% chance of randomly choosing a healthy baby.

(b) The left-skewed shape of the distribution suggests a randomly selected newborn will have an Apgar score at the high end of the scale. There is a small chance of getting a baby with a score of 5 or lower.

42

Example: Apgar Scores—What's Typical?

Consider the random variable X = Apgar Score.

Compute the mean of the random variable *X* and interpret it in context.

Value:	0	1	2	3	4	5	6	7	8	9	10
Probability:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

$$E(X) = \sum x_i p_i$$
$$= (0)(0.001) + (1)(0.006) + (2)(0.007) + \ldots + (10)(0.053)$$
$$= 8.128$$

The mean Apgar score of a randomly selected newborn is 8.128. This is the long-term average Apgar score of many, many randomly chosen babies.

Note: The expected value does not need to be a possible value of X or an integer! It is a long-term average over many repetitions.

43

Rules for Means

Rule 1: If X is a random variable and a and b are fixed numbers, then:

$$\mu_{a+bX} = a + b\mu_X$$

Rule 2: If X and Y are random variables, then:

$$\mu_{X+Y} = \mu_X + \mu_Y$$

44

Variance of a Random Variable

Since we use the mean as the measure of center for a discrete random variable, we'll use the standard deviation as our measure of spread. The definition of the **variance of a random variable** is similar to the definition of the variance for a set of quantitative data.

Variance of a Discrete Random Variable

Suppose that X is a discrete random variable whose probability distribution is:

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

and that μ_X is the mean of X. The **variance** of X is:

$$Var(X) = \sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + (x_3 - \mu_X)^2 p_3 + \ldots$$
$$= \sum (x_i - \mu_X)^2 p_i$$

To get the **standard deviation of a random variable,** take the square root of the variance.

45

Example: Apgar Scores—How Variable Are They?

Consider the random variable X = Apgar Score

Compute the standard deviation of the random variable *X* and interpret it in context.

Value:	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Probability:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

$$\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$$
$$= (0 - 8.128)^2(0.001) + (1 - 8.128)^2(0.006) + \ldots + (10 - 8.128)^2(0.053)$$
$$= 2.066$$

Variance

$$\sigma_X = \sqrt{2.066} = 1.437$$

The standard deviation of X is 1.437. On average, a randomly selected baby's Apgar score will differ from the mean 8.128 by about 1.4 units.

46

Rules for Variances

Rule 1: If X is a random variable and a and b are fixed numbers, then:

$$\sigma^2_{a+bX} = b^2\sigma^2_X$$

Rule 2: If X and Y are *independent* random variables, then:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

Rule 3: If X and Y have correlation ρ, then:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\rho\sigma_X\sigma_Y$$

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$$

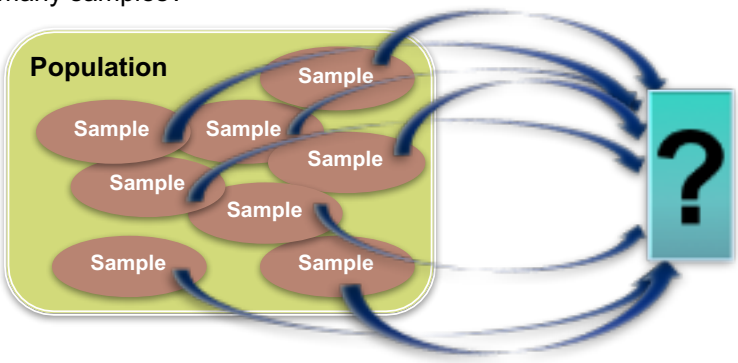
47

Statistical Estimation

Suppose we would like to estimate an unknown μ. We could select an SRS and base our estimate on the sample mean. However, a different SRS would probably yield a different sample mean.

This basic fact is called **sampling variability**: The value of a statistic varies in repeated random sampling.

To make sense of sampling variability, we ask, "What would happen if we took many samples?"



48

The Law of Large Numbers

How can $\bar{x}$ be an accurate estimate of μ? After all, different random samples would produce different values of $\bar{x}$.

If we keep on taking larger and larger samples, the statistic $\bar{x}$ is guaranteed to get closer and closer to the parameter .

Draw independent observations at random from any population with finite mean μ. The **law of large numbers** says that, as the number of observations drawn increases, the sample mean of the observed values gets closer and closer to the mean μ of the population.

49

The law of large numbers

- Theorem: Suppose $X_1, X_2, \ldots$ are independent and identically distributed. Suppose $E(X_1)$ is finite. Then as n increase to infinity, we have the following convergence result

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to E(X_1)$$

- Under the additional condition that σ^2_{X1}, the variance of X_1, is finite (this is usually satisfied), the theorem can be proved as follows.
- We know the mean of the sample average is $E(X_1)$. Consider the variance of the sample average:

  $\sigma^2_{Xbar} = (1/n^2)(\sigma^2_{X1} + \sigma^2_{X2} + \ldots + \sigma^2_{Xn}) = (1/n)\,\sigma^2_{X1}$

- As n increases, the sample variance vanishes. Therefore, the sample mean concentrates at the statistical average. The theorem is thus proved.

50

Example 1: Insurance

- Suppose an insurance company sells the same policies to members of a demographic group.
- Some customer files a (large) claim and receives coverage, some don't.
- The variance of a customer (individual risk) is typically quite large.
- The variance of the customer average (group risk) is much smaller. The larger the customer base, the smaller the group risk.
- The insurance company reduce a customer's risk significantly for a (relatively small) premium.

51

Example 2: Cloud-based storage

- Think of n employees in a company. Some need very little storage. Some work with a lot of data and need very large storage space. It is usually difficult to predict who needs how much, as well as the maximum an employee may need. The storage needed by employee j can be thought of as a random variable X_j.
- If data is stored in each employee's own hard disk, all hard disks have to be very large (and expensive).
- If data is stored in the cloud, the average storage per employee is much more predictable and much smaller than the maximum.
- Many companies now provide cloud-based storage with no quota limit on how much each employee uses!

52

Example 3: Histogram

- The histogram, when property normalized, converges to the probability mass function or the probability density function as the number of samples increases.
- Here is why: When we make the histogram, we divide the real number set into m intervals: $I_1, I_2, \ldots, I_m$. Out of n samples, the number of samples in interval I_j is n_j . Clearly, $n_1 + \cdots + n_m = n$.
- By the law of large numbers, as $n \to \infty$,

  $n_j/n \to P(X \in I_j)$

53

Applications of probability and statistics

- Electrical engineering: Noise canceling headphones mitigate unknown, random noise
- Computer science: Autocompletion is a feature provided by many web browsers, search engine interfaces, and word processors to predict the user input correctly with high probability.
- Computer engineering: To predict which data or instructions will be used by the CPU in the near future and pre-fetch to the local cache. The more successful the prediction the faster the CPU in real terms.
- Financial engineering: 1) Insurance companies make profit on random events. 2) Market goes up and down. How to manage an investment portfolio to maximize return subject to some risk constraint?
- Physics: Quantum mechanics tells us that at least at microscopic scale, the physical world has inherent uncertainties.
- Economics: How to model an economy subject to unpredictable events, e.g., natural disasters (flooding, storms, drought, etc.) and man made disasters (political turmoil, wars, etc.)?
- Medicine: Given some test results, the diagnosis is often said to be a probability that the patient has a certain disease. What does this mean?
- Public health: “The increase in the risk of death from eating red meat is about 1 percent a year.” How to interpret this?
- Other things we encounter in life: Should I buy a lottery ticket? Is it worth it to take vitamins? Should I move to another job?

54