

Chapter 4. Probability

- Probability is a branch of mathematics which addresses uncertainty. As such, it is essential for understanding and using statistics

Definitions

Experiment - a process that produces an outcome that cannot be predicted with certainty. If a process is an experiment, we *can* say with certainty that the outcome will be one of some set of possible outcomes

The sample space of an experiment is the set of all possible outcomes. We will use the symbol S to denote the sample space

Experiment 1: toss a coin once. Let H represent an outcome of heads, and T represent a tail. Then, $S = \{H, T\}$

Experiment 2: toss a coin two consecutive times and record the results (H or T) of each toss. $S = \{(H, H), (H, T), (T, H), (T, T)\}$.

Notation

- Braces, $\{$ and $\}$, indicate a set
- Parentheses indicate a pair, or an r -tuple, e.g., $(1, 3)$ is a pair, and $(1, 3, -2)$ is a 3-tuple.
- The order of elements in a r -tuple is important. Note that $(1, 3) \neq (3, 1)$. However, the order of elements in a set is meaningless; for example, $\{1, 3\} = \{3, 1\}$.
- Equality means something different for sets and pairs

The Algebra of Sets

- Let x denote an element in a set S . We write $x \in S$ if x is an element of S . If x is not in S , then we write $x \notin S$.
- Let A denote a *subset* of S . We write $A \subset S$. A is a subset of S if every element in A is in S . To show that $A \subset S$, we must show that every $x \in A$ is in S .

Note that

$$S \subset S \text{ and } A \subset A$$

- If A is not a subset of B , then we write $A \not\subset B$. If $A \not\subset B$, then there exists some element $x \in A$ which is not in B . To show that $A \not\subset B$, we must show that there is at least one $x \in A$ that is not in B
- Two sets are *equal* if and only if they contain exactly the same elements. To show $A = B$, we must show that

$$A \subset B \text{ and } B \subset A.$$

- If it is true that $A = B$, then it is true that $A \subset B$ and $B \subset A$
- For example: let $A = \{3, 2, 1\}$ and $B = \{1, 2, 3\}$. It is *true* that $A \subset B$ because every element in A is in B , and $B \subset A$ because every element in B is in A . Therefore, $A = B$.
- This illustrates an important property of sets: order is not important. In contrast, two pairs (a, b) and (c, d) are equal only if $a = c$ and $b = d$
- Which of the following are true?

a. $1 = \{1\}$

b. $\{1\} = \{1, 1\}$

c. $\{\{x\}, \{x, y\}\} = \{x, \{x, y\}\}$

d. $\{(1, 2), (3, 4)\} = \{(3, 4), (1, 2)\}$

- The *empty* set is a set with no elements. We denote the empty set by \emptyset . You cannot disprove the statement $\emptyset \subset A$ because it is impossible to identify an element in \emptyset that is not in A . Hence, $\emptyset \subset A$ is true for all sets A .

Suppose that $A \subset S$ and $B \subset S$. Then,

- The *complement* of A with respect to S are all elements in S that are not in A . The complement of A is denoted by \overline{A} . We cannot talk about complements without identifying the sample space S

- The *intersection* of A and B are all elements in both A and B . The intersection of A and B is denoted by $A \cap B$
- The union of A and B are all elements in either A or B . The union of A and B is denoted by $A \cup B$

Sample Spaces and Events

- The first probability topic is how to analyze an experiment and determine the sample space. There are no rules or formulas, so we must proceed using logical deduction

Experiment Roll a green and red die, observe the up-faces as pairs (red,green). It is useful to lay out the results in a table

Table 1. Outcomes of the first dice experiment. The first element in the pair is the green die and the second element in the pair is the red die.

Outcome of green die	Outcome of red die					
	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 6)
3	⋮	⋮				⋮
4	⋮	⋮				⋮
5	⋮	⋮				⋮
6	(6, 1)	(6, 2)	(6, 6)

- From the table, we can see that $S = \{(1, 1), (1, 2), \dots, (6, 6)\}$, and that there are 36 outcomes in S
- Note that there is only one way to get the outcome (2, 1): a green 2 and a red 1

Experiment: Roll a green and red die, observe the sum of up-faces. We can replace the pairs by sums in Table 1 and thereby construct Table 2

Table 2. Outcomes of the second dice experiment. The outcomes are the sum of the up-faces of the red and green dice.

Outcome of green die	Outcome of red die					
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	8
3	4	5				...
4
5			10	11
6	7	8	11	12

- The sample space is $S = \{2, 3, \dots, 12\}$
- Note that there are 3 ways to get a sum of 4
- Note the patterns and symmetries in the table

Definition: An *event* is a collection of outcomes, or a subset of the sample space. Generically, an event is denoted by E and the sample space by S . Therefore, $E \subset S$

• **An event *occurs* when the outcome of the experiment is an element in the event**

• Examples:

1. $E = \text{"a sum of 7"} = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$. We get a seven if the outcome is $(1, 6)$ or $(2, 5)$ or ... or $(6, 1)$

2. What event is certain to occur?

3. What event is certain not to occur?

• A and B are *mutually exclusive* events if it is impossible for them to occur simultaneously. Mathematically, A and B are *mutually exclusive* if $A \cap B = \emptyset$.

• The events $A = \text{"an even sum"}$ and $B = \text{"sum of seven"}$ are mutually exclusive.

• *Elementary events* contain only one outcome. If $S = \{2, 3, \dots, 12\}$, then the elementary events are $\{2\}, \{3\}, \dots, \{12\}$

More Examples

Roll a green and red die, observe the sum of up-faces. Consider the following events:

$$E_1 = \text{"sum is at most 3"}$$

$$E_2 = \text{"sum is at most 10"}$$

$$E_3 = \text{"sum is at least 7"}$$

$$E_4 = \text{"sum is odd"}$$

$$E_5 = \text{"sum is prime"}$$

- Suppose the outcome of the experiment is 4. Which events have occurred?
- Suppose that the outcome of the experiment is 7. Which events have occurred?
- Suppose that the outcome of the experiment is an even sum. With certainty, which events have occurred?
- Which two events are mutually exclusive?

Probability

- Probabilities are numbers assigned to events that express the likelihood that the event will occur.
- Probabilities are numbers between 0 and 1. A probability of 0 implies that it is impossible for the event to occur, and a probability of 1 implies that it is certain that the event will occur. If S is the sample space of an experiment, then
 1. $P(S) = 1$
 2. $P(\emptyset) = 0$
- There are several contexts in which we encounter probability. They are
 1. *Subjective probability* is an intuitive, personal assessment of the likelihood of an event
 - For example, I estimate the probability that Conrad Burns will be re-elected as a senator of Montana to be 0.5
 2. *Empirical probability* is the relative frequency of occurrence of an event. The experiment must be repeated to get an empirical probability. Therefore,

empirical probabilities are mainly applicable to simple, repeatable experiments.

- For example, to assess whether the subjective estimate of 0.5 for the probability of a head when tossing a fair coin is accurate, someone tossed a coin 10000 times and got 5050 heads. The empirical probability of a head, based on these data, is

$$P(\{H\}) = \frac{5050}{10000} = 0.505$$

- The empirical probability of an event should be recognized as an *estimate* of the true, but probably unknown, probability of the event

3. *Classical probability* is an approach to deducing probabilities for simple experiments in which all outcomes are equally likely (e.g., coin tossing, dice throwing)

- To deduce the probability of an event E , count the number of outcomes in the event, and divide by the number of outcomes in the sample space
- For example, suppose that the probability of a head is the same as a tail when tossing a (fair) coin. Then, $S = \{H, T\}$, and

$$P(\{H\}) = 1/2$$

is the (classical) probability of a head

- For example, suppose that 3 fair coins are tossed. Then, an analysis of the experiment suggests that each of the 8 outcomes in the sample space

$$S = \{(H, H, H), (H, H, T), \dots, (T, T, T)\}$$

are equally likely, and the probability getting exactly 2 heads is

$$P(2 \text{ heads}) = \frac{\#\{(H, H, T), (H, T, H), (T, H, H)\}}{\#S} = \frac{3}{8},$$

where $\#A$ is the number of elements in the set A

Probability Rules

Let S denote a sample space, and $A, B \subset S$ denote events. Then,

1. $0 \leq P(A) \leq 1$.

2. $P(S) = 1$ and $P(\emptyset) = 0$

3. If $A \cap B = \emptyset$, (A and B mutually exclusive), then

$$P(A \cup B) = P(A) + P(B).$$

From these rules, we can deduce the following:

4. $P(\bar{A}) = 1 - P(A)$

5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- Everyone should commit these five rules to memory
- *An example of classical probability:* Toss a red and a green die, and record the outcomes as pairs (red,green). Assume that all of the 36 pairs are equally likely.
- Let S be the sample space of the 36 pairs. Fill in the table below,

Outcome of green die	Outcome of red die					
	1	2	3	4	5	6
1	2	3	4			7
2	3	4				8
3	4					
4						
5						11
6	7	8			11	12

and answer the following questions:

1. Let $E_2 =$ "sum is a 2", $E_3 =$ "sum is a 3", ..., $E_{12} =$ "sum is a 12". What are the (classical) probabilities of the events $E_2, E_3,$ and E_{12} ?

2. Let $A =$ "sum is 7 or 11". Find $P(A)$.

3. Let $B =$ "sum is at least 10". Find $P(B), P(A \cap B), P(A \cup B), P(A \cap \bar{B}), P(A \cup \bar{B})$

Conditional Probability

- It is often the case in practical applications of probability that one event is known to have happened, and we are interested in assessing the probability of another event.
- For example, if the results of my tuberculosis tine (skin) test are positive, I will be very interested in the probability that I actually have TB. This probability is only about 0.15, assuming certain assumptions to be true
- This is because among those individuals that have positive tine tests, only about 15% of them have TB. To be more precise, let E be the event that an adult selected at random has TB, and let POS denote the event that an individual selected at random has a positive tine test. Then,

$$0.15 = P(E | POS)$$

is the probability that a randomly tested adult with a positive tine test actually has TB.

- We say that the probability of having TB conditional on a positive tine test is 0.15
- More generally, $P(E | F)$ is the probability of the event E given F
- When I stated that the probability that I actually have TB given a positive tine test is 0.15, I assumed that I was selected at random (or more likely, arbitrarily) for the test. If a doctor asks for the test because I complain of TB symptoms, then my probability is different (much greater than 0.15)

Contingency Tables and Conditional Probabilities

Example Doll, R. and Hill, A.B. 1952. A study of the aetiology of carcinoma of the lung. *British Medical Journal*, **2**, 1271-1286.

- Doll and Hill conducted a retrospective study of lung cancer and tobacco smoking in male patients in hospitals in several English cities. Retrospective refers to having selected a sample of male patients that had already contracted lung cancer. They also selected a similar control sample of hospitalized males that did not have lung cancer, and compared the extent of smoking between the two groups

Table. Numbers of male lung cancer and control patients classified by the reported average daily number of cigarettes smoked over a 10-year period preceding the onset of disease.

Daily average number of cigarettes	Disease Group		Total
	Lung Cancer patients	Control patients	
None	7	61	68
0-5	55	129	184
5-14	489	570	1059
15-24	475	431	908
25-49	293	154	447
50 +	38	12	50
Total	1357	1357	2714

- An application of conditional probability:

The empirical probability that a male patients has lung cancer, given that he was a nonsmoker is

$$P(\text{cancer} \mid \text{did not smoke}) = \frac{\#\{\text{have cancer and did not smoke}\}}{\#\{\text{did not smoke}\}} = \frac{7}{68} = .10$$

- To be precise, $P(\text{cancer} \mid \text{did not smoke})$ is empirical probability that a male patient selected at random from the collection of nonsmoking male patients has cancer.
- Specifically, there are 68 possible outcomes, all equally likely because of random sampling, and 7 outcomes (individuals) that have cancer. Hence, the probability is $7/68$.
- Similarly,

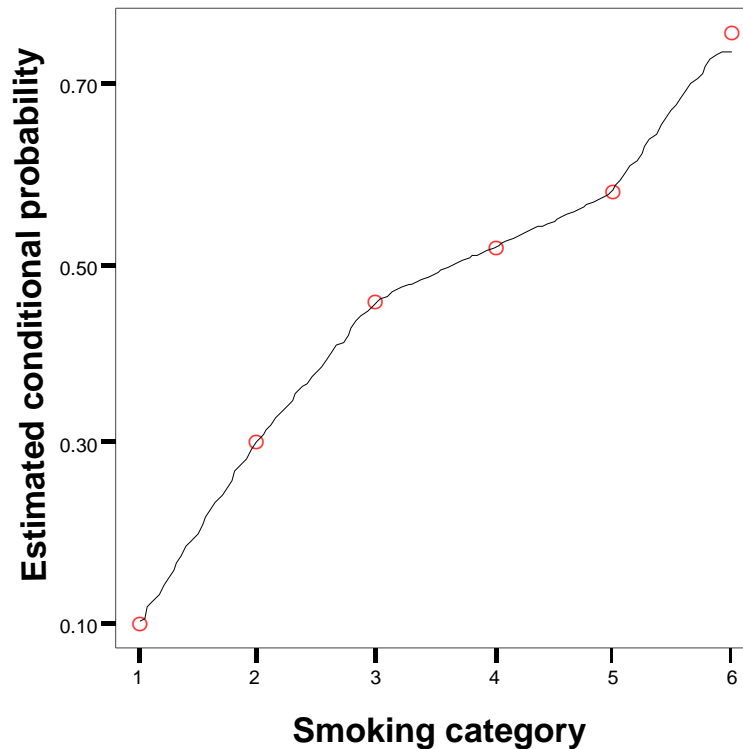
$$P(\text{cancer} \mid \text{smoked 0-5}) = \frac{\#\{\text{have cancer and smoked 0-5}\}}{\#\{\text{smoked 0-5}\}} = \frac{55}{184} = .30,$$

and

$$\begin{aligned} P(\text{cancer} \mid 50 +) &= \frac{\#\{\text{have cancer and smoked more than 50}\}}{\#\{\text{smoked more than 50}\}} \\ &= \frac{38}{50} = .76 \end{aligned}$$

- To summarize these empirical probabilities, I have plotted the empirical probabilities against the amount of smoking in the Figure below

Figure. Estimated conditional probability of lung cancer, given smoking categories, and a smooth. Smoking categories are ordered from least to greatest amount of smoking.



- Note that we are analyzing the set of $n = 2714$ patients, and no attempt is made at drawing inferences about a larger population. Inference must be done very carefully in this situation

The Mathematics of Conditional Probability

- Let S be a sample space, and suppose that A and B are events in S . Provided that $P(B) \neq 0$, the conditional probability of A given that B has happened is

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

- Dividing $P(A \cap B)$ by $P(B)$ amounts to revising the sample space to acknowledge that we are certain that the outcome belongs to the event B
- Now we have a conditional sample space comprised exclusively of the outcomes in B

- Whatever the outcome of the experiment, it is an outcome contained in the event B (since we know that B has happened). But, we know nothing further about the likelihood of outcomes in B
- The probability of \bar{B} is 0 (given that B has occurred)
- The computation

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

adjusts the probability of A to reflect this conditional sample space

- For example,

$$P(B | B) = \frac{P(B \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

and

$$P(\bar{B} | B) = \frac{P(\bar{B} \cap B)}{P(B)} = \frac{P(\emptyset)}{P(B)} = 0$$

- Recall the dice tossing experiment, where $S = \{2, 3, \dots, 12\}$ are the sums of the up-faces. Let $B = \text{"outcome is an odd sum"} = \{3, 5, 7, 9, 11\}$ and $A = \text{"sum is at least 11"} = \{11, 12\}$.
- Then, $P(B) = 18/36$, $P(A) = 3/36$ and $P(A \cap B) = 1/36$
- Given that B has occurred implies that whatever the outcome was, it must have been odd. We know nothing else, therefore, the 18 possible outcomes are $(1, 2), (1, 4), \dots, (5, 6)$, and each is equally likely
- There are 2 of 18 ways to get an odd outcome, and 2 of them give a sum of at least 11. (A 12 is impossible given that B has occurred). Therefore,

$$P(A | B) = \frac{2}{18} = \frac{1}{9}$$

- If we want to use the conditional probability formula, we get

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{18/36} = \frac{1}{9}$$

- Compare $P(A | B) = 1/9$ to $P(A) = 1/12$ (when we knew nothing about B)
- It is important to realize that the probability of $A \cap B$ can be computed from conditional probabilities using the formulas:

$$P(A \cap B) = P(A | B)P(B),$$

and

$$P(A \cap B) = P(B | A)P(A).$$

- For example, the probability of an ace when drawing a card randomly from a deck of 52 is

$$P(\text{ace on first draw}) = \frac{\#\text{Aces}}{\#\text{Cards}} = \frac{4}{52}$$

The probability of getting another ace if a second card is drawn is

$$P(\text{another ace} | \text{ace on first draw}) = \frac{\#\text{Aces left}}{\#\text{Cards left}} = \frac{3}{51}$$

- The probability of getting 2 aces when drawing 2 cards randomly and without replacement from a deck of 52 is

$$\begin{aligned} &P(\text{ace on first draw and ace on second draw}) \\ &= P(\text{another ace} | \text{ace on first draw}) \times P(\text{ace on first draw}) \\ &= \frac{3}{51} \times \frac{4}{52} = \frac{12}{2652} = 0.004, \end{aligned}$$

or about 1 in 250

Tree Diagrams (with probabilities)

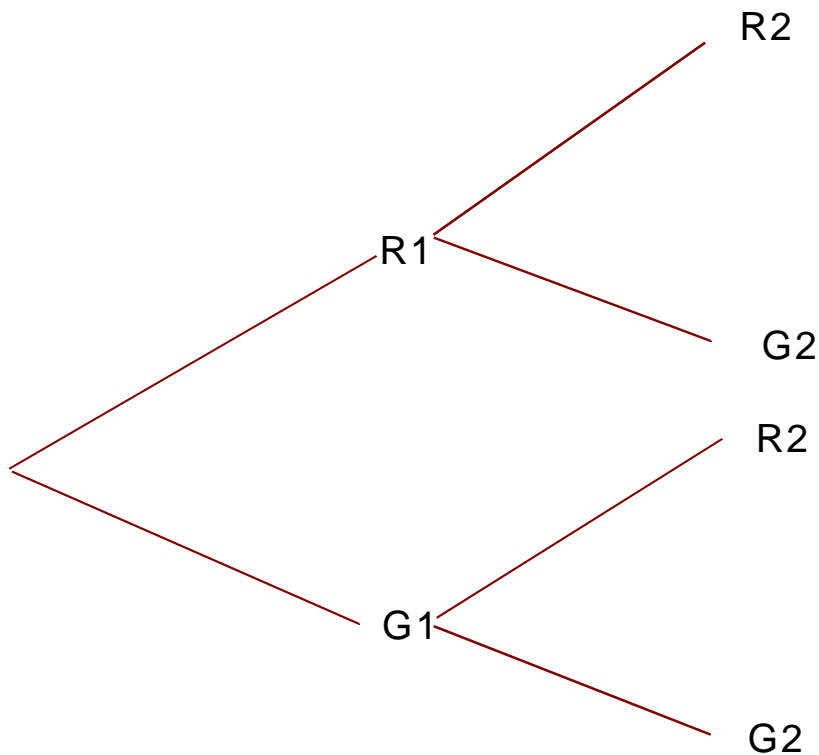
- These are very useful for analyzing certain experiments, such as experiments that consist of a short series of trials
- For example, an urn experiment. Suppose that 2 balls are randomly and without replacement from an urn containing 3 red and 7 green balls
- A sample space for the experiment is

$$S = \{(r, r), (r, g), (g, r), (g, g)\}$$

- Let R_1 denote the event "first ball is red" and G_1 denote the event "first ball is green"
- Then, $R_1 = \{(r, r), (r, g)\}$. What outcomes comprise G_1 ?
- Let R_2 denote the event "second ball is red" and G_2 denote the event "second ball is green"

Let A denote the event "draw two red balls". Hence, $A = R_1 \cap R_2$.

To find $P(A)$, we can use a tree diagram



- The first events R_1 and G_1 are *unconditional*, but the second set of events are *conditional*, because the path to the event depends on what has happened previously
- Applying classical probability ideas leads to $P(R_1) = 3/10$ and $P(R_2 | R_1) = 2/9$
- What are $P(G_1)$, $P(G_2 | R_1)$, $P(R_2 | G_1)$, and $P(G_2 | G_1)$?

- The probability of 2 reds is

$$P(R_1 \cap R_2) = P(R_1)P(R_2 | R_1) = \frac{6}{90}$$

- What are $P(R_1 \cap G_2)$, $P(G_1 \cap R_2)$ and $P(G_1 \cap G_2)$?

- What are $P(\text{"1 red"})$ and $P(\text{"no red"})$?

- What is the probability of getting a red on the second draw? A red on the second draw will happen if $R_1 \cap R_2$ or $G_1 \cap R_2$ occurs, so

$$\begin{aligned} P(R_2) &= P(R_1 \cap R_2) + P(G_1 \cap R_2) \\ &= \frac{6}{90} + \frac{21}{90} \\ &= \frac{27}{90} = \frac{3}{10} = 0.3 \end{aligned}$$

- We can also calculate the probability of having got a red on first draw if it is known that the second ball was red:

$$\begin{aligned}
 P(R_1 | R_2) &= \frac{P(R_1 \cap R_2)}{P(R_2)} \\
 &= \frac{6/90}{27/90} \\
 &= \frac{6}{27} = \frac{2}{9} = 0.2222\dots
 \end{aligned}$$

- Recall that the probability of a red on the first (knowing nothing about the second draw) was $3/10 = 0.3$. So, knowing that the second ball was red gave some information indicating that a red on the first was not quite so likely. Why is the probability reduced?
- Often, the unconditional probabilities are called *prior* probabilities (e.g., $P(R_1)$) and the conditional branch probabilities [e.g. $P(R_2 | R_1)$] are called *posterior* probabilities

Independent Events

- Let $A, B \subset S$, where S is a sample space.
- A and B are independent if knowing that A has occurred provides no information regarding whether B has occurred
- Likewise, A and B are independent if knowing that B has occurred provides no information regarding whether A has occurred
- Mathematically,

$$A \text{ and } B \text{ are independent if } P(A | B) = P(A)$$

- Also,

$$A \text{ and } B \text{ are independent if } P(B | A) = P(B)$$

- Suppose that A and B are independent. Then,

$$\begin{aligned}
 P(A \cap B) &= P(A | B)P(B) \\
 &= P(A)P(B)
 \end{aligned}$$

- This is a very useful result because it is often easy to determine $P(A)$ and $P(B)$

Homework problems for Friday, October 6:

p. 152: 4.35, 4.39, 4.40

p. 164: 4.52, 4.55, 4.57-4.60, 4.63, 4.64, 4.74, 4.76

(If Section 4.12 has been covered in Wednesday's lecture) p. 178: 4.87, 4.94, 4.97

Examples

- Suppose that you draw 2 cards randomly and *with* replacement from a set of 52 cards. What is the probability that both are aces?
- Let $A =$ "first card is ace" and $B =$ "second card is ace". Then $P(A) = P(B) = 4/52$, and

$$P(\text{"both aces"}) = P(A)P(B) = \frac{4}{52} \times \frac{4}{52} = 0.006,$$

or 1 in 170

- What is the probability of having three children, all of which are girls?
- How about four out of four?
- Recall the dice tossing experiment, where $S = \{2, 3, \dots, 12\}$ are the sums of the up-faces. Are the events $A =$ "an odd sum", and the event $B =$ "sum is at least 10" independent?
- We must determine if the following statement is true or false:

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

If the statement is true, then A and B are independent; if the statement is false, then A and B are dependent

Example: The Donner Party. Survival by age group, and gender within age group. The age groups are young (31 years or less), and old (at least 31)

Let

- A denote the event that an individual selected at random from the 45 is a female,
- B denote the event that an individual selected at random is young,
- C denote the event that an individual selected at random survived

- Tabled values below are the probability that an individual selected at random belongs to a particular cell.
- For example, what is the probability that an individual selected at random from the 45 is a young, female survivor? Since this event is $A \cap B \cap C$,

$$P(A \cap B \cap C) = \frac{\#A \cap B \cap C}{\#S} = \frac{7}{45} = 0.156$$

- Outcome probabilities for the Donner experiment

Age Group				Gender		Total
				Females	Males	
≤ 31	Survival	No	0.022	0.311	0.333	
		Yes	0.156	0.156	0.311	
	Total		0.178	0.467	0.644	

Age Group				Gender		Total
				Females	Males	
> 31	Survival	No	0.089	0.133	0.222	
		Yes	0.067	0.067	0.133	
	Total		0.156	0.200	0.356	

- From the table, we can compute

$$P(A) = 0.178 + 0.156 = 0.334$$

$$P(C) = 0.311 + 0.133 = 0.444$$

$$P(A \cap C) = 0.156 + 0.067 = 0.223$$

- Is there an association between survival and gender? In other words, does knowing gender provide information regarding the probability of survival? Yes, because $P(A \cap C) = 0.223 \neq 0.148 = P(A)P(C)$
- What is the probability that a female survived? This means: given the individual was female, what is the probability that she survived?

$$P(C | A) = \frac{P(A \cap C)}{P(A)} = \frac{0.223}{0.334} = 0.667$$

- Note that $P(C | A) \neq P(C)$, which confirms that survivorship and gender are not independent

- What is the probability that a young female survived? This means: given the individual was a young female, what is the probability that she survived?

- We need

$$P(\text{young and female}) = P(A \cap B) = 0.178,$$

and

$$P(\text{young and female and survivor}) = P(A \cap B \cap C) = 0.156$$

- Then

$$P(C | A \cap B) = \frac{P(A \cap B \cap C)}{P(A \cap B)} = \frac{0.156}{0.178} = 0.875$$

Bayes Rule

- Ott and Longnecker (p. 138) give the extension of Bayes rules to more general situations. We will only consider the following case

- Suppose that A are events C , and $P(C) \neq 0$. Then,

$$P(A | C) = \frac{P(A)P(C | A)}{P(A)P(C | A) + P(\bar{A})P(C | \bar{A})}$$

- The formula can be understood by noting that C is a union of two mutually exclusive events, $A \cap C$ and $\bar{A} \cap C$. Hence,

$$C = (A \cap C) \cup (\bar{A} \cap C)$$

- Because they are mutually exclusive

$$P(C) = P(A \cap C) + P(\bar{A} \cap C)$$

- The probability of $A \cap C$ can be expressed as

$$P(A \cap C) = P(A)P(C | A)$$

- Similarly,

$$P(\bar{A} \cap C) = P(\bar{A})P(C | \bar{A})$$

- In combination, we get

$$\begin{aligned} P(C) &= P(A \cap C) + P(\bar{A} \cap C) \\ &= P(A)P(C | A) + P(\bar{A})P(C | \bar{A}) \end{aligned}$$

- Collecting all these equations as a single formula for $P(A | C)$, we get Bayes rule:

$$P(A | C) = \frac{P(A)P(C | A)}{P(A)P(C | A) + P(\bar{A})P(C | \bar{A})}$$

- Bayes rule is best understood not as a formula, but as a method. Specifically, we can reverse a tree diagram in the sense that if we start knowing the conditional probability of C given A , and the probability of A , then we can find the probability of A given C .

- For example, the TB example can be analyzed using Bayes Rule:

- Researchers have established the *sensitivity of the test*. Specifically, the rate of correct positives is

$$P(POS | TB) = 0.92$$

where TB means that an individual has TB, and POS denotes a positive test

- They also have established that *specificity of the test*. Specifically, the rate of correct negatives is

$$P(\overline{POS} | \text{no } TB) = 0.96$$

- A false negative means there is no skin reaction when an individual has TB. The probability of a false negative is

$$P(\overline{POS} | TB) = 1 - P(POS | TB) = 0.08$$

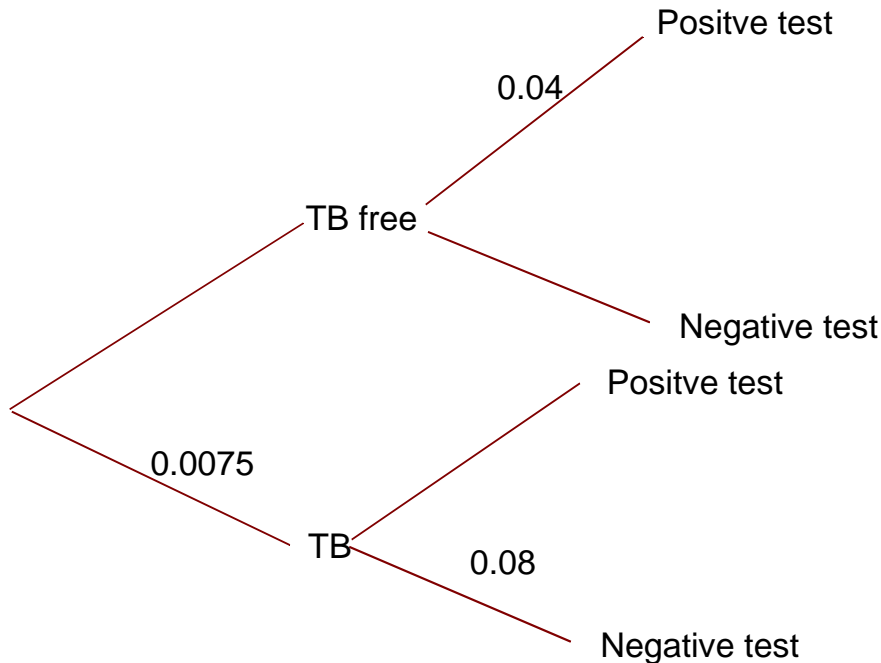
- The probability that standard TB tine test produces an incorrect positive is

$$P(POS | \text{no } TB) = 1 - P(\overline{POS} | \text{no } TB) = 0.04$$

- About 5 years ago, the probability that any individual selected at random from American adults has TB was

$$P(TB) = 0.0075$$

- Find the probability that an individual selected at random actually has TB, given a positive test.
- The best way to keep things straight is to use a tree diagram:



- Mathematically, we computed

$$\begin{aligned}
 P(TB | POS) &= \frac{P(TB \cap POS)}{P(TB \cap POS) + P(\overline{TB} \cap POS)} \\
 &= \frac{P(TB)P(POS | TB)}{P(TB)P(POS | TB) + P(\overline{TB})P(POS | \overline{TB})} \\
 &= \frac{.0075 \times .92}{.0075 \times .92 + .9925 \times .04} \approx .15.
 \end{aligned}$$

- One interpretation: if you test randomly, only 15% of the positive test results will be correct.
- How about drug testing in the Olympics? All athletes are tested. How confident can we be in a report of a positive test result?

4.6 Random Variables

- A quantitative *random variable* is rule that assigns numerical values to the outcomes of an experiment.
- A qualitative *random variable* is rule that assigns non-numerical values to the outcomes of an experiment.

- **Example** Suppose that 2 fair coins are tossed. Let

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

denote the sample space

- Counting the number of heads creates a quantitative random variable defined by

$$X = \# \text{ heads}$$

- X is a rule that assigns numbers to every one of the four outcomes in S . The assigned values are

$$X(H, H) = 2$$

$$X(H, T) = 1$$

$$X(T, H) = 1$$

$$X(T, T) = 0.$$

- For X to be a legitimate random variable, every outcome produces exactly one value for X
- If success is defined by getting at least one head, then I can create a random variable Y with two qualitative values, success and failure. Every outcome leads exactly to one value for Y
- For example, select a random sample of 2 people from a population of 4, and measure their heights. Suppose that the population of heights (in inches) are (70, 68, 72, 72). Then, the sample space is

$$S = \{(70, 68), (70, 72), (70, 72), (68, 72), (68, 72), (72, 72)\}$$

- The sample mean is a random variable. Specifically, it is defined according to

$$\bar{X} = \frac{X_1 + X_2}{2},$$

where X_1 and X_2 are the heights of the first and second sampled individuals

- \bar{X} is a random variable because that assigns numbers to the pairs. The assigned values are

$$\begin{aligned}\bar{X}(70, 68) &= \frac{70 + 68}{2} = 69 \\ \bar{X}(70, 72) &= \frac{70 + 72}{2} = 71 \\ &\vdots \\ \bar{X}(72, 72) &= 72.\end{aligned}$$

Some Definitions

- A set is *countable* if the elements can be identified, or listed. For example, $\{1, 2\}$ and $\{1, 2, 3, \dots\}$ are countable sets
- A set that is not countable is *uncountable*. For example, the interval of the real numbers $(0, 1) = \{x \in \mathbb{R} \mid 0 < x < 1\}$ is uncountable because between any two numbers, there are infinitely many numbers. In fact, we cannot even identify the smallest value in this set
- If a random variable can assume only countably many values, then it is called *discrete*.
- Examples:
 - 1) the number of heads when tossing a coin 3 times,
 - 2) number of times that a coin must be tossed before a head is observed,
 - 3) the number of females minus the number of males among a family of 4
- A *continuous* random variable takes uncountably many values.
- Example: a number selected at random from the set

$$\{x \in \mathbb{R} \mid 0 \leq x \leq 1\} = [0, 1]$$

is continuous random variable.

4.7 Probability Distributions for Discrete Random Variables

- Recall: an experiment produces an outcome, say (H, T) , and a random variable assigns a numerical value to the outcome, say $X(H, T) = 1$
- A probability distribution (for a discrete random variable) gives
 - 1) all possible values of the random variable
 - 2) the probabilities of observing the values

Notation

- Capital letters are reserved for the *random variable* (it's a rule)
- The *values* are lower-case. (Values are numbers)
- For example, the probability that the random variable X yields a value of 2 is denoted by $P(X = 2)$. Ott and Longnecker shorten this expression to $P(2)$.
- Often, we write $P(x)$ instead of $P(X = x)$

Examples

- Y is the number of heads in two tosses. The probability distribution of Y is

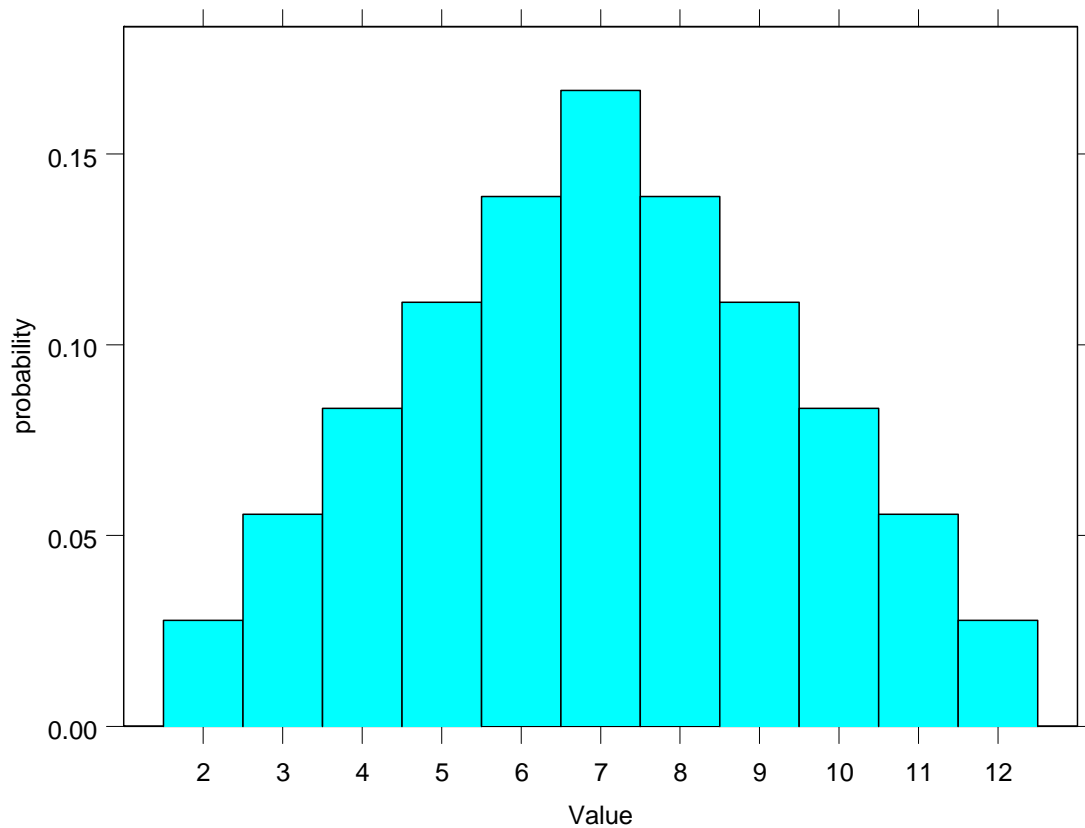
y		0	1	2
$P(y)$		1/4	1/2	1/4

- \bar{X} is the sample mean height. The probability distribution of \bar{X} is

\bar{x}		69	70	71	72
$P(\bar{x})$		2/12	4/12	4/12	2/12

- Do not list a particular value more than once
- Often a **probability histogram** is used to represent the distribution
- Figure. The probability distribution of $X =$ "sum of the up-faces of two dice".

Probability distribution for the sum of two fair dice



Properties of discrete random variables

Let X denote a discrete random variable, and x denote a real number. Then,

1. $0 \leq P(x) \leq 1$
2. The sum of $P(x)$ over all possible values of X is 1. We write $\sum P(x) = 1$.
3. The probabilities are additive, for example

$$P(X = 2 \text{ or } X = 3) = P(2) + P(3),$$

because the event $X = 2$ is mutually exclusive with respect to the event $X = 3$. X cannot be both 2 and 3. The probability of the event $\{X = 2 \text{ and } X = 3\}$ is 0

Binomial Experiments

An experiment is a *binomial experiment* if it satisfies all of the following conditions:

1. It consists of n identical trials.
2. Each trial results in one of two outcomes, generically denoted by S (success) and F (failure).
3. The probability of success is the same for all n trials.
4. The trials are independent.

A *binomial* random variable is the number of S 's over the n trials of a binomial experiment. The probability of success is denoted by π .

• Examples

1. The number of female children (out of $n = 5$) conceived by a couple; $\pi = 0.51$.
2. The number of heads when a fair coin is tossed $n = 100$ times; $\pi = 1/2$.
3. The number of times that a trained dog can distinguish between grizzly and black bear scat in $n = 5$ trials; $\pi = ?$

A Formula for Computing Binomial Probabilities

• Preliminaries:

$$n! = \begin{cases} n \times (n-1) \times (n-2) \cdots 2 \times 1, & \text{if } n = 1, 2, 3, \dots \\ 1, & \text{if } n = 0. \end{cases}$$

Also

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

- Example: verify that $\binom{4}{2} = 6$

- Suppose that Y has a binomial distribution with *parameters* n and π . I will write $Y \sim \text{Bin}(n, \pi)$. Then,

$$P(y) = \begin{cases} \binom{n}{y} \pi^y (1 - \pi)^{n-y}, & \text{if } y = 0, 1, 2, \dots, n, \\ 0, & \text{otherwise} \end{cases}$$

- Example: toss a coin 5 times. Let $X = \text{"# heads"}$. Then, $n = 5$ and $\pi = \frac{1}{2}$. Hence,

$$\begin{aligned} P(3) &= \binom{5}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{5-3} \\ &= \frac{5!}{3!2!} \left(\frac{1}{2}\right)^5 = \frac{5 \times 4}{2 \times 1} \times \frac{1}{32} = \frac{10}{32} \end{aligned}$$

- Also,

$$\begin{aligned} P(4) &= \binom{5}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{5-4} \\ &= \frac{5!}{4!1!} \left(\frac{1}{2}\right)^5 = 5 \times \frac{1}{32} = \frac{5}{32} \end{aligned}$$

- The probability of 3 or more heads is

$$\begin{aligned} P(y \geq 3) &= P(3) + P(4) + P(5) \\ &= \binom{5}{3} \frac{1}{32} + \binom{5}{4} \frac{1}{32} + \binom{5}{5} \frac{1}{32} \\ &= \frac{10}{32} + \frac{5}{32} + \frac{1}{32} \\ &= \frac{1}{2} \end{aligned}$$

- It turns out that the distribution of X is symmetric:

$$P(0) = P(5) = \frac{1}{32}$$

$$P(1) = P(4) = \frac{5}{32}$$

$$P(2) = P(3) = \frac{10}{32}$$

Example Suppose that the probability that an airplane engine will fail in flight is $\pi = 0.05$, and that engines on a particular plane are independent in this regard. A flight will be completed if at least 50% of the engines do not fail. Is a 2- or 4-engine airplane preferable in this respect?

- For the 2-engine plane, let $X = \#$ failed engines. Then, $X \sim \text{Bin}(2, 0.05)$

$$\begin{aligned} P(\text{flight is completed}) &= P(X \leq 1) \\ &= P(X = 0) + P(X = 1) \\ &= \binom{2}{0} 0.05^0 \times 0.95^2 + \binom{2}{1} 0.05^1 \times 0.95^1 \\ &= 0.95^2 + 2 \times 0.05 \times 0.95 \\ &= 0.9025 + 0.095 = 0.9975 \end{aligned}$$

- For the 4-engine plane, let $Y = \#$ failed engines. Then, $Y \sim \text{Bin}(4, 0.05)$

$$\begin{aligned} P(\text{flight is completed}) &= P(Y \leq 2) \\ &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= \binom{4}{0} 0.05^0 \times 0.95^4 + \binom{4}{1} 0.05^1 \times 0.95^3 \\ &\quad + \binom{4}{2} 0.05^2 \times 0.95^2 \\ &= 0.95^4 + 4 \times 0.05 \times 0.95^3 + 6 \times 0.05^2 \times 0.95^2 \\ &= 0.8145 + 0.1715 + 0.0135 = 0.9995 \end{aligned}$$

- Suppose instead that $\pi = 0.5$. For the 4-engine plane, $Y \sim \text{Bin}(4, 0.5)$

$$\begin{aligned}
 P(\text{flight is completed}) &= P(Y \leq 2) \\
 &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\
 &= \binom{4}{0} 0.5^0 \times 0.5^4 + \binom{4}{1} 0.5^1 \times 0.5^3 \\
 &\quad + \binom{4}{2} 0.5^2 \times 0.5^2 \\
 &= 0.5^4 + 4 \times 0.5^4 + 6 \times 0.5^4 \\
 &= 0.6875
 \end{aligned}$$

and for the 2-engine plane, $X \sim \text{Bin}(2, 0.5)$

$$\begin{aligned}
 P(\text{flight is completed}) &= P(X \leq 1) \\
 &= P(X = 0) + P(X = 1) \\
 &= \binom{2}{0} 0.5^0 \times 0.5^2 + \binom{2}{1} 0.5^1 \times 0.5^1 \\
 &= 0.5^2 + 2 \times 0.5^2 \\
 &= 0.75
 \end{aligned}$$

- Further analysis shows that the 2-engine plane is preferable if $\pi > 1/3$, whereas the 4-engine plane is preferable if $\pi \leq 1/3$.
- All finite probability distributions have a mean μ (center of mass) and standard deviation σ (average distance of observations to the mean)
- We also refer to the mean and standard deviation of the probability distribution as the *expected value* and *standard deviation* of the random variable
- The term *expected value* is used in reference to 1) A best guess of a future value of the random variable, 2) the average of many observations on the random variable
- The mean μ of the binomial probability distribution is easily calculated. It is

$$\mu = n\pi$$

- For example, if you guess randomly on a multiple choice exam with 4 possible answers on each of 20 questions, then $n = 20$, $\pi = 0.25$ and $\mu = 5$ is the expected number of correct answers

- The standard deviation is almost as simple. It is

$$\sigma = \sqrt{n\pi(1 - \pi)}.$$

- For the 4-engine airplane with $\pi = 0.05$, $\mu = 4 \times 0.05 = 0.2$ failures, and $\sigma = \sqrt{4 \times 0.05 \times 0.95} = 0.436$ failures.

- Later, we will use μ and σ when analyzing binomially distributed data.

Example If K. got 9 correct out of 20 on the multiple choice exam, is it likely that K. guessed randomly?

- To answer this question, let $X = \# \text{correct}$. If K. guesses randomly, then $X \sim \text{Bin}(20, 0.25)$

- The probability that K. did this well, or better, by guessing is

$$P(X \geq 9) = \sum_{k=9}^{20} \binom{20}{k} 0.25^k 0.75^{20-k} = 0.0409$$

- K.'s score of 9 out of 20 is quite improbable. Either

- 1) K. is very lucky, or

- 2) K. was not guessing randomly, and had some (limited) knowledge of the subject

- A common testing approach is to conclude that K. did not guess randomly if the probability of scoring as well, or better, than 9 is less than 0.05 (1 in 20)

- Should we use this test? For this one application of the test, it is impossible to judge whether we are right or wrong

- On the other hand, if we use this method (decide that they are not guessing if they get at least 9) on many individuals, then we will wrongly conclude that they are not guessing at most 4.09% of the time. That is,

$$P(\text{conclude not guessing} \mid \text{are guessing}) = 0.049$$

- This is true because only 4.09% of the time will they get 9 or better purely by guessing

- There is another error that can be made: we conclude that they are guessing randomly when they are not. The rate for this error is harder (but not impossible) to address

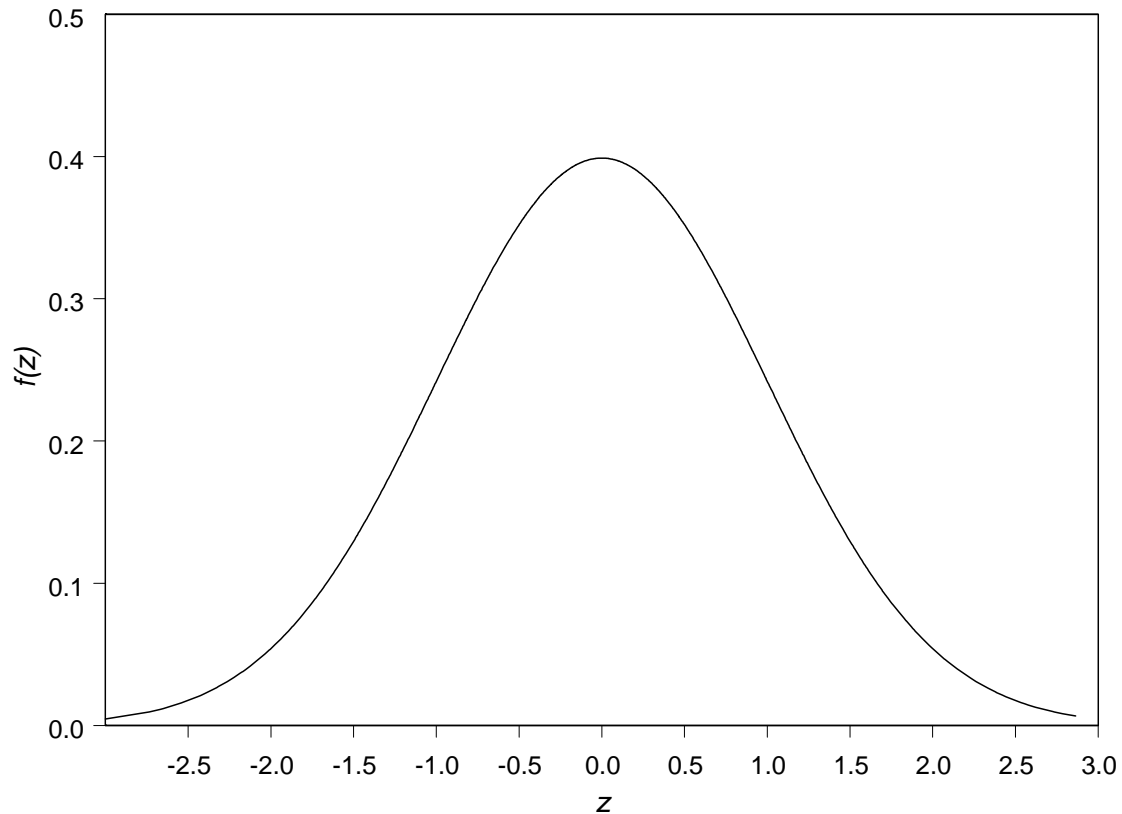
Continuous Random Variables

- Recall that a continuous random variable takes on uncountably many values
- The properties of discrete random variables do not hold for continuous r.v.'s. Specifically,
 1. A sum of uncountably many values cannot be computed. For this reason, the sum over all values is not defined, and it is not true that $\sum P(x) = 1$
 2. In addition, for any value x of X (a continuous random variable), $P(x) = 0$
- Instead of working with the probabilities of individual outcomes, we work with probabilities of events such as "the outcome of X is between 68 and 69". As a set, the event is the interval $\{x \in \mathbb{R} \mid 68 < x < 69\}$, and the probability of interest is denoted as $P(68 < X < 69)$
- A probability density function (p.d.f.) is used to determine the probability that the outcome of X is in some interval of \mathbb{R}
- For the standard normal random variable Z , the density function is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

- Like a probability histogram, the area under a p.d.f. is always 1

Figure. Standard normal density function $f(z)$ plotted against z



Properties of Continuous Random Variables

- Suppose that X is a continuous random variable with p.d.f. f . Then,
 1. $0 \leq f(x)$, for all possible values x .
 2. Let a and b be real numbers such that $a < b$. Then $P(a < X < b)$ is the probability that X takes on a value in the interval between a and b and

$$0 \leq P(a < X < b) \leq 1$$

3. $P(a < X < b)$ is the area under the curve described by $f(x)$ and the x -axis. This area is the integral of $f(x)$ evaluated from a to b

- For example, if X has a standard normal distribution, the empirical rule states that

$$0.68 = P(-1 < X < 1)$$

Normal random variables

- Because computing integrals for normal random variables is quite difficult, we tabulate these probabilities in normal tables
- The distributions of all normal random variables have the same shape, but they differ by mean μ and standard deviation σ
- The standard normal distribution is used to compute probabilities for other normal distributions.
- The *standard normal* random variable has $\mu = 0$ and $\sigma = 1$; usually it is denoted by Z .
- Ott and Longnecker use the lower case z instead of Z , and denote numerical values of Z generically by z_0 . In other words, my expression $P(Z \leq z)$ is denoted by Ott and Longnecker as $P(z \leq z_0)$
- Table 1 in the Appendix (page 1091) gives probabilities to the left of z . For example, $P(Z \leq -1.50) = 0.0668$
- Examples. Determine, or calculate:
 1. $P(Z \leq 1)$
 2. $P(Z < 1)$
 3. $P(Z > -1)$
 4. $P(0 \leq Z \leq 1)$
 5. $P(-1 \leq Z \leq 0)$
 6. $P(-1 \leq Z \leq 1)$
- Let X be a normal random variable with mean μ and standard deviation σ . We write $X \sim N(\mu, \sigma)$. E.g., if $\mu = 68$ and $\sigma = 2$, then $X \sim N(68, 2)$.
- The normal transformation converts X to a standard normal random variable according to the formula

$$Z = \frac{X - \mu}{\sigma}.$$

- Subtracting μ from X shifts the values to be centered about 0, and division by σ corrects the scale so that Z has standard deviation 1
- This transformation is used to compute probabilities for normal random variables besides the standard normal

Example Suppose that $X \sim N(68, 2)$. What is the probability that X is less than 66? The answer is

$$\begin{aligned}P(X < 66) &= P\left(\frac{X - \mu}{\sigma} < \frac{66 - \mu}{\sigma}\right) \\&= P\left(Z < \frac{66 - 68}{2}\right) \\&= P(Z < -1) \\&= 0.1587\end{aligned}$$

Example What is the probability that X takes on a value between 70 and 66? That is, what is $P(66 \leq X \leq 70)$?

- We need to break up the problem into two sub-problems because of the form of Table 1. It only gives areas to the left of a value, so we need to write $P(66 \leq X \leq 70)$ as a difference of two areas, one to the left of 70 and the other to the left of 66. The breakdown is

$$P(66 \leq X \leq 70) = P(X \leq 70) - P(X < 66)$$

- Now we can use Table 1:

$$\begin{aligned}P(66 \leq X \leq 70) &= P(X \leq 70) - P(X < 66) \\&= P\left(\frac{X - \mu}{\sigma} \leq \frac{70 - 68}{2}\right) - P\left(\frac{X - \mu}{\sigma} < \frac{66 - 68}{2}\right) \\&= P(Z \leq 1) - P(Z < -1) \\&= 0.8413 - 0.1587 = 0.6826.\end{aligned}$$

Example Suppose that $X \sim N(68, 2)$. Find the 95th percentile of the distribution of X . We want to determine the value x satisfying

$$0.95 = P(X \leq x).$$

This value of x is the 95th percentile.

Step 1: Find the 95th percentile $z_{0.95}$ of the standard normal distribution, i.e., determine the value $z_{0.95}$ satisfying

$$0.95 = P(Z \leq z_{0.95}).$$

From Table 1,

$$0.95 \approx P(Z \leq 1.645)$$

Therefore, $z_{0.95} = 1.645$ is the (approximate) 95th percentile of the standard normal distribution.

Step 2. To get the 95th percentile of the normal distribution with $\mu = 68$ and $\sigma = 2$, we must solve for x within the standard normal formula

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

- Plug in the known quantities $z = 1.645$, $\mu = 68$ and $\sigma = 2$. This yields

$$1.645 = \frac{x - 68}{2}.$$

- Solve for x :

$$x = 68 + 2 \times 1.645 = 71.29.$$

- This means that 71.29 is the 95th percentile. Therefore, $P(X < 71.29) = 0.95$.

The Sampling Distribution of \bar{Y}

- The most common task in applied statistics is that of estimating the population mean μ of some population
- The usual estimator of μ is the sample mean \bar{Y} .
- \bar{Y} is a random variable, hence, it has a probability distribution. The properties of this distribution are hugely important in applied statistics

- Consequently, it is essential to learn the basics about its distribution. We need to specifically know the mean and standard deviation of the distribution
- We can derive the exact distribution for artificial examples. Suppose that a population of heights is $\mathcal{P} = \{68, 68, 69, 71\}$, and \bar{Y} is computed from a random sample of $n = 2$.
- The sample space is

$$S = \{(68, 68), (68, 69), (68, 71), (68, 69), (68, 71), (69, 71)\},$$

and the distribution of \bar{Y} is determined by computing

$$\begin{aligned}\bar{Y}(68, 68) &= 68 \\ \bar{Y}(68, 69) &= 68.5 \\ \bar{Y}(68, 71) &= 69.5 \\ \bar{Y}(68, 69) &= 68.5 \\ \bar{Y}(68, 71) &= 69.5 \\ \bar{Y}(69, 71) &= 70.\end{aligned}$$

- All 6 pairs are equally likely (because the sample is collected by random sampling)
- The sampling distribution of \bar{Y} is

\bar{y}	68	68.5	69.5	70
$P(\bar{Y} = \bar{y})$	1/6	2/6	2/6	1/6

- What is the population mean?

$$\mu = \frac{68 + 68 + 69 + 71}{4} = 69$$

- What is the mean of the distribution of \bar{Y} ? By symmetry the mean is $\mu_{\bar{y}} = 69$
- The mean of the distribution of \bar{Y} (or, equivalently, the expected value of \bar{Y}) is *always* the same as the original distribution. In other words, $\mu_{\bar{y}} = \mu$
- We say that \bar{Y} is unbiased for μ . This means that the estimator \bar{Y} is neither consistently larger than μ , nor consistently smaller than μ

The Standard Deviation of the Distribution of \bar{Y}

- The sample mean is more precise when computed from large samples compared to small
- Thus, the distribution of \bar{Y} depends on the sample size n
- An application of integral calculus will show that σ/\sqrt{n} is the standard deviation of the distribution of \bar{Y} if σ is the standard deviation of the *sampled population*
- For example, I used as an example, a normal population with mean $\mu = 68$ and $\sigma = 2$ to represent the distribution of male heights (in inches)
- If a sample of size $n_1 = 4$ individuals is used to estimate μ , then the distribution of \bar{Y} will have standard deviation

$$\frac{\sigma}{\sqrt{n_1}} = \frac{2}{2} = 1$$

- If a sample of size $n_2 = 16$ individuals is used to estimate μ , then the distribution of \bar{Y} will have standard deviation

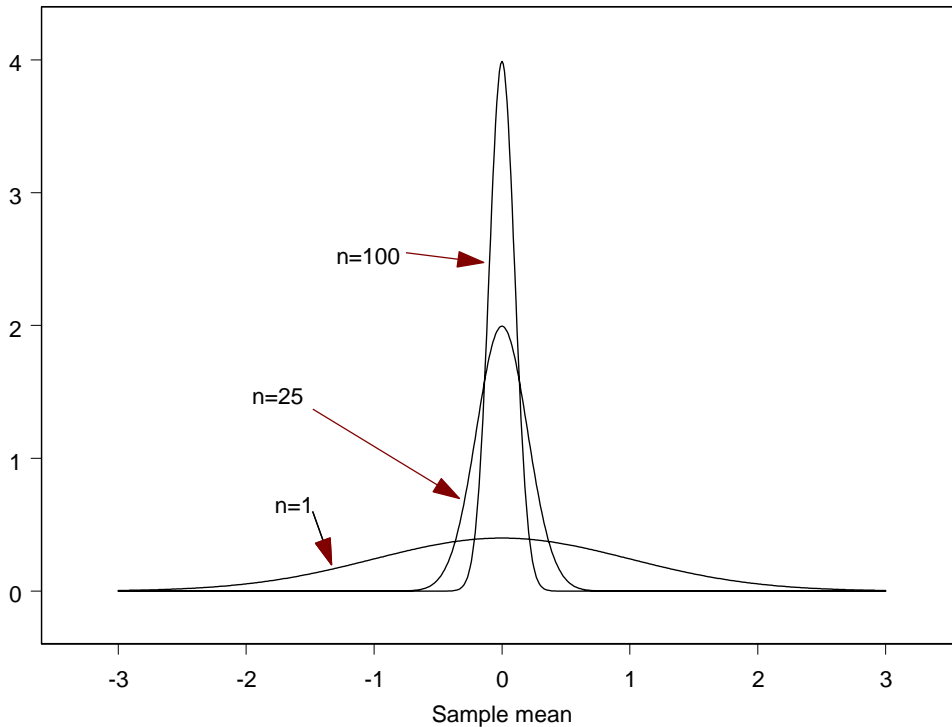
$$\frac{\sigma}{\sqrt{n_2}} = \frac{2}{4} = 0.5$$

- If a sample of size $n_3 = 64$ individuals is used to estimate μ , then the distribution of \bar{Y} will have standard deviation

$$\frac{\sigma}{\sqrt{n_3}} = \frac{2}{8} = 0.25$$

- The effect of sample size on the spread of the distribution of the sample mean is illustrated below:

Approximate sampling distribution of the sample mean when $\mu=0$ and $\sigma=1$.



Notation The standard deviation of the distribution of \bar{Y} is often denoted by $\sigma_{\bar{y}}$.

- We can calculate $\sigma_{\bar{y}}$ if we know the standard deviation σ of the population that is being sampled. Then,

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}.$$

- The mean of the distribution of \bar{Y} is occasionally denoted by $\mu_{\bar{y}}$. Because

$$\mu_{\bar{y}} = \mu,$$

where μ is the mean of the population that is being sampled, it is simpler to use μ to denote the mean of the distribution of \bar{Y}

- There are two important results that tell us more about the distribution of \bar{Y} :

1. If the sampled population is an exactly normal population, then \bar{Y} is *exactly* normal in distribution. That is,

$$\bar{Y} \sim N(\mu, \sigma_{\bar{y}}).$$

- This result is not very important because most populations are not exactly normal

2. \bar{Y} is *approximately* normal in distribution, provided that n is large. That is,

$$\bar{Y} \sim N(\mu, \sigma_{\bar{y}}).$$

- This statement is one version of the **Central Limit Theorem**. Though it is simple, it is extremely useful

- How large is enough? That depends on how skewed the original population is. A high degree of skew needs more observations to get approximate normality.

- How do we use this result? Given a set of data, we make a sample histogram. If the skew is not too bad, and the sample size is relatively large, then we use the Z transformation, as illustrated below, to compute probabilities

The **accuracy of the normal distribution** as an approximation of the true sampling distribution of \bar{Y} depends on two attributes

1. The *degree of symmetry* in the sample data relative to the sample size n

- If the sample data are approximately symmetric in distribution, then the normal approximation will be good with as few as 15 observations

- If the sample data is highly skewed (e.g., the mode is the same as the minimum), then at least 70 observations are needed to get a good approximation

- $n = 30$ is often thrown out as a safe number if the sample data are not highly skewed

2. The *data are a random sample*. If this is not true, then the approximation is not valid and should not be used

Example

- A machine produces computer hard disks. If the average asymmetry of hard disks produced by the machine is greater than 0.29, then the machine is considered to be malfunctioning
- It is reasonable to assume that asymmetry varies to some extent, even when the machine is operating properly. Consequently, it may be difficult to tell when the machine is malfunctioning.
- It is not unreasonable to suppose that the distribution of asymmetries in the population is symmetrically distributed. Consequently, the distribution of a sample of asymmetries is likely to be approximately symmetric. (This ought to be checked with a histogram or box plot)
- One method of testing for hard disk asymmetry is to take a sample of disks, measure asymmetry of each disk, and determine the sample mean
- Let X represent the asymmetry of a randomly selected hard disk, and \bar{X} represent the mean of a sample of hard disks
- Suppose that the population of hard disk asymmetries is normal in distribution with mean $\mu = 0.28$ and standard deviation $\sigma = 0.005$ (this implies that the machine is functioning properly)
- If \bar{X} is greater than 0.29, then we will say that the machine is malfunctioning

1. What is the probability that a sample of $n = 1$ disk will detect a malfunction?

- The solution is

$$\begin{aligned} P(X > 0.29) &= P\left(\frac{X - \mu}{\sigma} > \frac{0.29 - \mu}{\sigma}\right) \\ &= P\left(Z > \frac{0.29 - 0.28}{0.005}\right) = 0.0228 \end{aligned}$$

The percentage of disks in the population that are greater than 0.29 in asymmetry is 2.28%

2. If a sample of $n = 9$ disks is used, then what is the probability that the sample mean indicates a malfunction?

- The solution is

$$\begin{aligned}P(\bar{X} > 0.29) &= P\left(\sqrt{n}\frac{\bar{X} - \mu}{\sigma} > \sqrt{n}\frac{0.29 - \mu}{\sigma}\right) \\&= P\left(Z > 3 \times \frac{0.29 - 0.28}{0.005}\right) \\&= P(Z > 6) \\&< 0.0001\end{aligned}$$

3. Suppose that $\mu = 0.295$, so the machine is malfunctioning. What percentage of disks are greater than 0.29 in asymmetry?

- The solution is

$$\begin{aligned}P(X > 0.29) &= P\left(\frac{X - \mu}{\sigma} > \frac{0.29 - \mu}{\sigma}\right) \\&= P\left(Z > \frac{0.29 - 0.295}{0.005}\right) = 0.8413,\end{aligned}$$

that is, 84.1%.

- What is the probability that a sample of $n = 9$ disks will detect this state of malfunction?

- The solution is

$$\begin{aligned}P(\bar{X} > 0.29) &= P\left(\sqrt{n}\frac{\bar{X} - \mu}{\sigma} > \sqrt{n}\frac{0.29 - \mu}{\sigma}\right) \\&= P\left(Z > 3 \times \frac{0.29 - 0.295}{0.005}\right) \\&= P(Z > -3) \\&= 0.9986\end{aligned}$$

4. Suppose that the disks are tested destructively, so a small sample size is desirable. If we are willing to accept an error rate of 5%, how small can we make the sample?

• Step 1: find z satisfying $P(Z > z) = 0.05$. Inspection of Table 1 finds the value $z = 1.645$

• Step 2: equate Z and \bar{X} via the standard normal transformation:

$$\begin{aligned} Z &= \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \\ \Rightarrow 1.645 &= \sqrt{n} \frac{0.29 - 0.295}{0.005} \\ \Rightarrow \left(\frac{1.645 \times 0.005}{0.29 - 0.295} \right)^2 &= n. \end{aligned}$$

The solution is $n = 2.70$, so we must use a sample of at least $n = 3$ hard disks. The steps behind this calculation are

1. $0.05 = P(Z > 1.645) = P\left(\sqrt{n} \frac{\bar{X} - \mu}{\sigma} > 1.645\right)$

and

2. $\sqrt{n} \frac{\bar{X} - \mu}{\sigma} = \sqrt{n} \frac{0.29 - 0.295}{0.005},$

given that we say the machine is malfunctioning whenever \bar{X} is larger than 0.29

Homework: for Friday October 13:

p. 140: 4.28

p. 164: 4.52, 4.55, 4.57-4.60, 4.63, 4.64, 4.74, 4.76

p. 180: 4.87, 4.94, 4.97

Review Problems: p189: 4.110, 4.112, 4.121, 4.122. 4.123, 4.124

Other Applications of the Central Limit Theorem

• The Central Limit Theorem can also be used for the sum of n observations

$$\sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n.$$

- Specifically, if n is large, then

$$\sum_{i=1}^n X_i \sim N(n\mu, \sqrt{n}\sigma)$$

where μ and σ are the mean and standard deviation of the population that is being sampled

- The rationale for the CLT in this case is based on two facts:

$$1) \sum_{i=1}^n X_i = n\bar{X}$$

$$2) \bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

- 1) and 2) imply that $\sum_{i=1}^n X_i = n\bar{X} \sim N(n\mu, \sqrt{n}\sigma)$

- With the exception of the next topic, applications are fairly limited

4.13 Normal Approximation to the Binomial

- Suppose that $n = 100$ trials of an binomial experiment are to be carried out, and $X = \text{\#successes}$ in the n trials is the random variable of interest. Then $X \sim \text{Bin}(n, \pi)$, where $\pi = \text{probability of success on any given trial}$

- Recall that

$$P(X = x) = \begin{cases} \binom{n}{x} \pi^x (1 - \pi)^{n-x}, & \text{if } x = 0, 1, 2, \dots, n, \\ 0, & \text{otherwise} \end{cases}$$

where $\pi = \text{probability of success on any given trial}$

- We may want to compute $P(X \leq 40) = P(0) + P(1) + \dots + P(40)$. This sum cannot be computed easily without a computer

- The Central Limit Theorem can be used to calculate binomial probabilities and percentiles when the number of trials is sufficiently large

- The normal approximation says that if

$$1) X \sim \text{Bin}(n, \pi), \text{ and}$$

2) $n\pi \geq 5$ and $n(1 - \pi) \geq 5$,

then

$$X \sim N(n\pi, \sqrt{n\pi(1 - \pi)})$$

Example Suppose that 60% of a population favor a ballot issue (increased taxes for schools), but a super-majority of 67% must vote in favor of the tax for the issue to pass. If $n = 100$ people vote, what is the probability that a super-majority is in favor the the tax?

- We must compute $P(X \geq 67)$, where $X \sim \text{Bin}(100, 0.6)$. The normal approximation says that $X \sim N(n\pi, \sqrt{n\pi(1 - \pi)})$, or

$$X \sim N(60, 4.90)$$

- Hence,

$$\begin{aligned} P(X \geq 67) &= P\left(\frac{X - \mu}{\sigma} \geq \frac{67 - 60}{4.9}\right) \\ &\doteq P(Z \geq 1.43) \\ &= 0.076 \end{aligned}$$

- The exact value of this probability is $P(X \geq 67) = P(x = 67) + \dots + P(X = 100) = 0.0912$

The Continuity Correction

- Accuracy can be improved by correcting for (lack of) continuity
- If the calculation is an area to the *right* of a value, then subtract 0.5 from the numerator
- If the calculation is an area to the *left* of a value, then add 0.5 to the numerator

Example Applying the continuity correction to the previous example gives

$$\begin{aligned} P(X \geq 67) &= P\left(\frac{X - \mu}{\sigma} \geq \frac{67 - 60 - 0.5}{4.9}\right) \\ &\doteq P(Z \geq 1.33) \\ &= 0.092 \end{aligned}$$

- The normal approximation to the binomial is justified because

1) $X = \sum_{i=1}^n I_i$, where

$$I_i = \begin{cases} 1, & \text{if the } i\text{th trial is a success} \\ 0, & \text{if the } i\text{th trial is a failure.} \end{cases}$$

- The use of the symbol I_i is used because I_i *indicates* the outcome of the i th trial

2) All of the I_i 's are binomial random variables consisting of a single trial ($n = 1$) with mean $\mu = \pi$ and standard deviation $\sigma = \pi(1 - \pi)$

3) The CLT for the sum implies that $\sum_{i=1}^n I_i \sim N(n\mu, \sqrt{n}\sigma)$ provided that all of the I_i 's have mean μ and standard deviation σ . In this case,

$$\sqrt{n}\sigma = \sqrt{n}\sqrt{\pi(1 - \pi)} = \sqrt{n\pi(1 - \pi)}$$

Therefore,

$$X = \sum_{i=1}^n X_i \sim N(n\pi, \sqrt{n\pi(1 - \pi)})$$