Introduction to Applied Statistics
STAT 5005
Lecture 3: Probability and Probability Distributions (Chapter 4)

Jingyu Sun

Fall 2021

Introduction

Inferential Statistics

- Statistics is the science of **Learning from Data**, which consists of four steps:

1. Defining the problem
2. Collecting the data
3. **Summarizing the data** (this chapter)
4. Analyzing the data

- Graphical and numerical descriptive techniques are a means to summarize and describe a sample
- However, a sample is not identical to the population from which it was selected.
- Thus the need to assess the degree of accuracy to which the sample mean, sample standard deviation, etc. represent the corresponding population values

Probability is the language of uncertainty

- Most management decisions must be made in the presence of uncertainty
- Example 1: Prices and designs for new automobiles must be selected on the basis of shaky forecasts of consumer preference, national economic trends, and competitive actions
- Example 2: Medical screening tests for home are now widely spread for a variety of purposes (pregnancy, diabetes glucose monitoring, ...). How reliable are the testing kits?
- Suppose a company states its pregnancy test provides correct results in 75% of its applications by pregnant women
- At what point do we decide that the result of an observed sample is so improbable, assuming the company's claim is correct, that we disagree with its claim?
- Probability is the tool that enables us to make an inference

Basic Probability Rules

Probabilities: Definitions and Properties

- A **random experiment** is a mechanism that produces a definite **outcome** that cannot be predicted with certainty
  - Example: Rolling two dice
- The **sample space** associated with a random experiment is the set of all possible outcomes
- An **event** is a subset of the sample space
- An event E is said to **occur** on a particular trial of the experiment if the outcome observed is an element of the set E
- A graphical representation of a sample space and events is a **Venn diagram**

- The probability of an outcome e in a sample space S is a number that measures the *likelihood* that e will occur on a single trial of the corresponding random experiment
- The probability of an event E is the sum of the probabilities of the individual outcomes of which it is composed. It is denoted $P(E)$. If an event $E = \{e_1, \ldots, e_k\}$, then

$$P(E) = P(e_1) + \ldots + P(e_k)$$

- In the classical interpretation of probability, assuming that all outcomes are equally likely, the probability of an event E is computed by taking the ratio of the number of outcomes, $|E|$, favorable to event E to the total number of possible outcomes, $|S|$:

$$P(E) = \frac{|E|}{|S|}$$

- For any event E, $0 \leq P(E) \leq 1$
- Two events A and B are said to be **mutually exclusive** if (when the experiment is performed a single time) the occurrence of one of the events excludes the possibility of the occurrence of the other event
- If two events, A and B, are mutually exclusive, the probability that either event occurs is $P(A) + P(B)$
  - Example: Find the probability of a sum less than or equal to 4 on a single toss of two dice

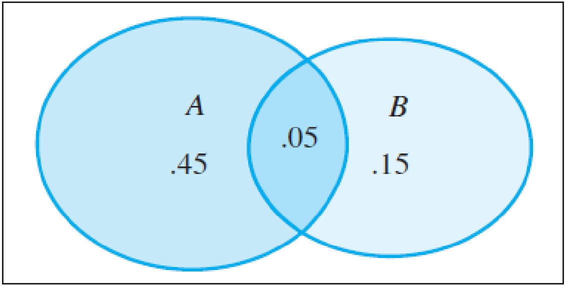
- The **complement** of an event E is the event that E does not occur. The complement of E is denoted by the symbol $\overline{E}$. It follows that $P(E) + P(\overline{E}) = 1$
- The **union** of two events A and B is the set of all outcomes that are included in either A or B (or both). The union is denoted as $A \cup B$
- The **intersection** of two events A and B is the set of all outcomes that are included in both A and B. The intersection is denoted as $A \cap B$
- The probability of the union of A and B is

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

Example

Events and event probabilities are shown in the Venn diagram below. Use this diagram to determine the following probabilities

a. $P(A), P(\overline{A})$
b. $P(B), P(\overline{B})$
c. $P(A \cap B)$
d. $P(A \cup B)$



Conditional Probability and Independence

- Example: The examination of a large number of insurance claims, categorized according to type of insurance and whether the claim was fraudulent, produced the results shown below

	Type of Policy (%)			
Category	Fire	Auto	Other	Total (%)
Fraudulent	6	1	3	10
Nonfraudulent	14	29	47	90
Total	20	30	50	100

- What is the probability of the event F, "the claim is fraudulent"?
- Suppose now you have the additional information that the claim was associated with a fire policy. How does it change the above probability?

- Consider two events A and B with $P(B) \neq 0$. The **conditional probability** of event A given event B is

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

- In the insurance example,

$$P(F|\text{fire policy}) = \frac{P(F \cap \text{fire policy})}{P(\text{fire policy})} \qquad = \frac{.06}{.20} = .30$$

- The probability $P(F) = .10$ is called the **unconditional** or **marginal probability** of the event F
- Clearly, the conditional probabilities of F, given the types of policies, will be of much greater assistance in measuring the risk of fraud than the unconditional probability of F

Multiplication Law

- The probability of the intersection of two events A and B, with nonzero probabilities $P(A)$ and $P(B)$, is

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

- Example: A corporation is proposing to select two of its current regional managers as vice presidents
- The corporation has 6 male regional managers and 4 female regional managers
- Make the assumption that the 10 regional managers are equally qualified and hence all possible groups of two managers should have the same chance of being selected as the vice presidents
- Find the probability that both vice presidents are male

Independent Events

- Two events A and B are **independent events** if $P(A|B) = P(A)$ or $P(B|A) = P(B)$, or equivalently, $P(A \cap B) = P(A)P(B)$
- A card is selected at random from an ordinary deck of 52 playing cards. If A is the event that the selected card is an ace and B is the event that it is a spade, prove that A and B are independent events

Example

- A meat inspector must decide whether a randomly selected meat sample contains *E. coli* bacteria.
- The inspector conducts a diagnostic test. However, the diagnostic test is occasionally in error
- The results of the test may be a false positive (indication of *E. coli presence* is incorrect), or a false negative (conclusion of *E. coli* absence is incorrect)

- The results of large-scale screening tests are summarized in the table below

Diagnostic Test Result	Meat Sample Status **E**	**NE**
Positive	9,500	100
Negative	500	9,900
Total	10,000	10,000

- **E**: *E. coli* present in the meat sample / **NE**: no traces of *E. coli*
- The **sensitivity** of the diagnostic test is the true positive rate
- The **specificity** of the diagnostic test is the true negative rate
- Compute the sensitivity and specificity of the diagnostic test

- Sensitivity is

$$P(\text{test is positive}|\textbf{E}) = \frac{9,500}{10,000} = 0.95$$

- Specificity is

$$P(\text{test is negative}|\textbf{NE}) = \frac{9,900}{10,000} = 0.99$$

Bayes' Formula 1

If A and B are any events whose probabilities are not 0 or 1, then

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

- The primary question facing the inspector is to evaluate the probability of *E. coli* being present in the meat sample when the test yields a positive result
- Suppose that *E. coli* is present in 4.5% of all meat samples

$$
\begin{aligned}
P(\textbf{E}|\text{test is positive}) &= \frac{P(\text{test is positive}|\textbf{E})P(\textbf{E})}{P(\text{test is positive}|\textbf{E})P(\textbf{E}) + P(\text{test is positive}|\textbf{NE})P(\textbf{NE})} \\
&= \frac{0.95(0.045)}{0.95(0.045) + (100/10,000)(1-0.045)} \\
&= 0.817
\end{aligned}
$$

Bayes' Formula 2

If $A_1, \ldots, A_k$ form a partition of the sample space S, that is $A_1, \ldots, A_k$ are mutually exclusive events and $A_1 \cup \ldots \cup A_k = S$, then, for any $i = 1, \ldots, k$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}$$

Variables: Discrete and Continuous

Variables: Discrete and Continuous

- Many times the events of interest in an experiment are quantitative outcomes associated with a **quantitative random variable**, since the possible responses vary in numerical magnitude, e.g. blood cell count in a sample, sleep duration among adults, ...
- When observations on a quantitative random variable can assume only a countable number of values, the variable is called a **discrete random variable**. Examples:
  - the number of persons voting for the incumbent in an upcoming election
  - the number of times you flip a coin until you get a heads

Variables: Discrete and Continuous

- Many times the events of interest in an experiment are quantitative outcomes associated with a **quantitative random variable**, since the possible responses vary in numerical magnitude, e.g. blood cell count in a sample, sleep duration among adults, ...

- When observations on a quantitative random variable can assume only a countable number of values, the variable is called a **discrete random variable**. Examples:
  - the number of persons voting for the incumbent in an upcoming election
  - the number of times you flip a coin until you get a heads

- When observations on a quantitative random variable can assume any one of the uncountable number of values in a line interval, the variable is called a **continuous random variable**. Examples:
  - the time per week you spend on working for my course
  - Federer's service speed

Probability Distributions for Discrete Random Variables

- The **probability distribution** of a discrete random variable Y is given by the probability of specific individual values y's occurring
- This distribution can be presented as a table, a graph, or a formula
- Example: Compute the probability distribution of the number of heads observed for the tossing of two fair coins

Properties of Discrete Random Variables

1. The probability associated with every value of Y lies between 0 and 1.
2. The sum of the probabilities for all values of Y is equal to 1.
3. The probabilities for a discrete random variable are additive.

Two Discrete Random Variables: The Binomial and the Poisson

The Binomial Distribution

A binomial experiment is one that has the following properties:

1. The experiment consists of n identical trials.
2. Each trial results in one of two outcomes. We will label one outcome a success and the other a failure.
3. The probability of success on a single trial is equal to p and p remains the same from trial to trial.
4. The trials are independent; that is, the outcome of one trial does not influence the outcome of any other trial.
5. The random variable Y is the number of successes observed during the n trials.

Did this study satisfy the properties of a binomial experiment?

- An article in the March 5, 1998, issue of The New England Journal of Medicine discussed a large outbreak of tuberculosis.
- One person, called the index patient, was diagnosed with tuberculosis in 1995.
- The 232 co-workers of the index patient were given a tuberculin screening test.
- The number of co-workers recording a positive reading on the test was the random variable of interest.

Formulas for a Binomial Experiment

- The probability of observing y successes in n trials of a binomial experiment is

$$p(y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}$$

- Mean: $\mu = np$
- Standard Deviation: $\sigma = \sqrt{np(1-p)}$

Binomial Distribution in R

- For $n = 20$ and $p = 0.3$, $p(4)$ can be computed using the R command

```r
dbinom(4, 20, 0.3)
```

```
## [1] 0.130421
```

and $P(Y \leq 5)$ is given by

```r
pbinom(5, 20, 0.3)
```

```
## [1] 0.4163708
```

Example

- Suppose that a sample of households is randomly selected from all the households in the city in order to estimate the percentage in which the head of the household is unemployed.
- To illustrate the computation of a binomial probability, suppose that the unknown percentage is actually 10% and that a sample of $n = 5$ is selected from the population.
- What is the probability that all five heads of the households are employed?

- Thus far, the binomial probabilities were calculated for small values of n
- In practice, n is usually large (in national surveys, sample sizes as large as 1,500 are common), and the computation of the binomial probabilities is tedious
- Later in this chapter, we will present a simple procedure for obtaining approximate values to the probabilities we need in making inferences
- In order to obtain very accurate calculations when n is large, use a computer software program

The Poisson Distribution

- The number of times an event occurs in an interval of time or space may be suitably modeled using a **Poisson distribution**
- Examples:
  - the number of typos on a printed page
  - the number of cars passing through the intersection of Storrs Rd and N Eagleville Rd in five minutes
  - the number of students arriving during office hours

- Let Y be the number of events occurring during a fixed time or space interval. Then the probability distribution of Y is Poisson, provided certain conditions are satisfied:

1. Events occur one at a time; two or more events do not occur precisely at the same time or same space
2. The occurrence (or nonoccurrence) of an event in a given period of time or region of space is independent of the occurrence of the event in a nonoverlapping time period or region of space
3. The expected number of events during one period or region, μ, is the same as the expected number of events in any other period or region

Formulas for a Poisson Distribution

- Assuming that the above conditions hold, the Poisson probability of observing y events in a unit of time or space is given by the formula

$$p(y) = \frac{\mu^y \, e^{-\mu}}{y!}$$

- Mean: μ
- Standard Deviation: $\sigma = \sqrt{\mu}$

Poisson Distribution in R

- For $\mu = 2.2$, $p(3)$ can be computed using the R command

```r
dpois(3, 2.2)
```

```
## [1] 0.1966387
```

and $P(Y \leq 4)$ is given by

```r
ppois(4, 2.2)
```

```
## [1] 0.9275037
```

Example

- A team of wildlife scientists is surveying the number and types of small mammals in the region
- The average number of field mice captured per trap over a 24-hour period is 2.3.
- What is the probability of finding exactly four field mice in a randomly selected trap?
- What is the probability of finding at most four field mice in a randomly selected trap?
- What is the probability of finding more than four field mice in a randomly selected trap?

Approximation of Binomial Probabilities with Poisson Probabilities

- When n is large and p is small in a binomial experiment, the Poisson distribution provides a good approximation to the binomial distribution
- As a general rule, use the Poisson distribution approximation when $n \geq 100, p \leq 0.01$, and $np \leq 20$
- Take $\mu = np$ for the parameter of the Poisson distribution

Approximation of Binomial Probabilities with Poisson Probabilities

- When n is large and p is small in a binomial experiment, the Poisson distribution provides a good approximation to the binomial distribution
- As a general rule, use the Poisson distribution approximation when $n \geq 100, p \leq 0.01$, and $np \leq 20$
- Take $\mu = np$ for the parameter of the Poisson distribution

Example

- The prevalence of side effects for a certain drug is reported to be 0.1%
- Compute the probability that none of a random sample of 1,000 patients administered the drug experiences a particular side effect
- Compare the above probability with a Poisson approximation

Probability Distributions for Continuous Random Variables

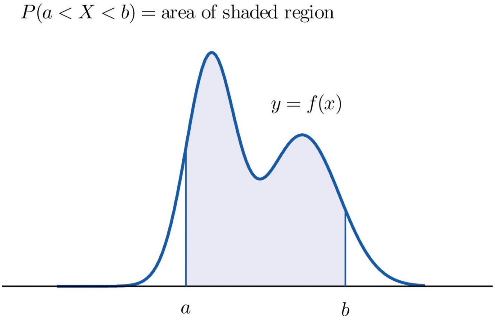
Probability Distributions for Continuous Random Variables

- Let X be the distance between Federer's position when he serves and the location where the ball lands. What is the probability that X is exactly 15 meters next time Federer serves?

- See Hawk eye system: https://www.youtube.com/watch?v=XhQyVnwBXBs

- For a discrete random variable X the probability that X assumes one of its possible values on a single trial of the experiment makes good sense
- We state, without elaboration, that it is impossible to assign a small amount of probability to each value of x
- More meaningful questions are those of the form: What is the probability that the distance is less than 16 meters, or is between 14.9 and 15.1 meters?

The probability distribution of a continuous random variable X is an assignment of probabilities to intervals of decimal numbers using a function $f(x)$, called a density function, in the following way: the probability that X assumes a value in the interval $[a, b]$ is equal to the area of the region that is bounded above by the graph of the equation $y = f(x)$, bounded below by the x-axis, and bounded on the left and right by the vertical lines through a and b.



- Every density function $f(x)$ must satisfy the following two conditions:
  1. For all numbers x, $f(x) \geq 0$
  2. The area of the region under the graph of $y = f(x)$ and above the x-axis is 1.

- Every density function $f(x)$ must satisfy the following two conditions:
  1. For all numbers x, $f(x) \geq 0$
  2. The area of the region under the graph of $y = f(x)$ and above the x-axis is 1.
- We have the following property: For any continuous random variable X:

$$\begin{aligned} P(a \leq X \leq b) = P(a < X \leq b) &= P(a \leq X < b) \\ &= P(a < X < b) \end{aligned}$$

A Continuous Probability Distribution: The Normal Distribution

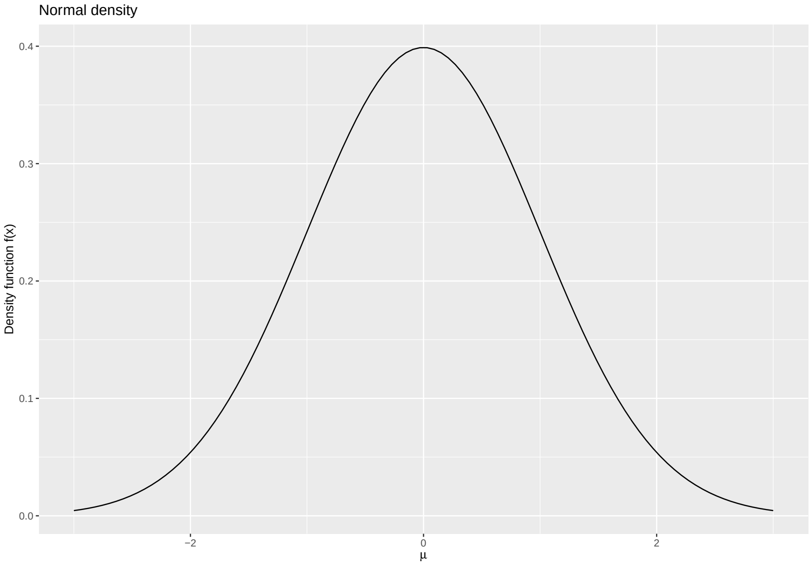
Example

- A restaurant wants to advertise a new burger they call **The Quarter-kilogram**. What probability distribution do you expect for the weights of those burgers?

The Normal Distribution

- The normal distribution is one of the most (even perhaps the most) important distributions in Probability and Statistics. It allows to model many natural, physical and social phenomenons.
- A normal (or Gaussian) random variable is determined by two parameters, its mean μ and its standard deviation σ

- The graph of a normal density is a bell curve that is symmetric about the mean



Normal Distribution in R

- Assume X is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 2.5$
- The density function at $x = 11.3$, $f(11.3)$ is

```r
dnorm(11.3, 15, 2.5)
```

```
## [1] 0.05337412
```

- $P(X \leq 16.7)$ is

```r
pnorm(16.7, 15, 2.5)
```

```
## [1] 0.7517478
```

The Standard Normal Distribution

- A standard normal random variable is a normally distributed random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$. It is usually denoted by the letter Z

- Using R, find the following probabilities where Z denotes a standard normal random variable

    1. $P(Z < 1.48)$
    2. $P(Z < -0.25)$
    3. $P(Z > 1.60)$
    4. $P(Z > -1.02)$
    5. $P(1.13 < Z < 4.16)$
    6. $P(-5.22 < Z < 2.15)$

Probability Computations for General Normal Random Variables

- If X is a normally distributed random variable with mean μ and standard deviation σ, then

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

  where Z denotes a standard normal random variable, a can be any real number or $-\infty$, b can be any real number or ∞

- The new end points $(a-\mu)/\sigma$ and $(b-\mu)/\sigma$ are called the z-**scores** of a and b

Example

- The United States Environmental Protection Agency (EPA) has developed procedures for measuring vehicle emission levels of nitrogen oxide
- The amount of this pollutant in a randomly selected automobile in Houston, Texas is normally distributed with a mean level of 70 ppb (parts per billion) and standard deviation of 13 ppb
- Answer the following questions using **only** cumulative probabilities of a standard normal distribution
  1. What is the probability that a randomly selected vehicle will have emission levels less than 60 ppb?
  2. What is the probability that a randomly selected vehicle will have emission levels greater than 90 ppb?
  3. What is the probability that a randomly selected vehicle will have emission levels between 60 and 90 ppb?

Quantile of a Normal Distribution

- The value of the standard normal random variable Z that cuts off a left tail of area c is denoted z_c
- By symmetry, the value of Z that cuts off a right tail of area c is $-z_c$
- You can find $z_{0.80}$ with the R command:

```
qnorm(0.8, 0, 1)
```

```
## [1] 0.8416212
```

- Suppose X is a normally distributed random variable with mean μ and standard deviation σ
- To find the value x_c of X that cuts off a left or right tail of area c in the distribution of X:
  1. find the value z_c of Z that cuts off a left tail of area c in the standard normal distribution;
  2. compute x_c using the destandardization formula

$$x_c = \mu + z_c \cdot \sigma$$

Example

- A State of Texas environmental agency, using the vehicle inspection process described in the previous example, is going to offer a reduced vehicle license fee to those vehicles having very low emission levels
- They will offer this incentive to the group of vehicle owners having the best 10% of emission levels
- What emission level should the agency use in order to identify the best 10% of all emission levels?

Sampling Distributions

Random Sampling

- A sample of n measurements selected from a population is said to be a **random sample** if every different sample of size n from the population has an equal probability of being selected
- Example: A study of crimes related to handguns is being planned for the ten largest cities in the United States
- The study will randomly select two of the ten largest cities for an in-depth study following the preliminary findings
- The population of interest is the ten largest cities C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}
- How many possible different samples consisting of two cities that could be selected from the population of ten cities are there?
- Give the probability associated with each sample in a random sample of 2 cities selected from the population

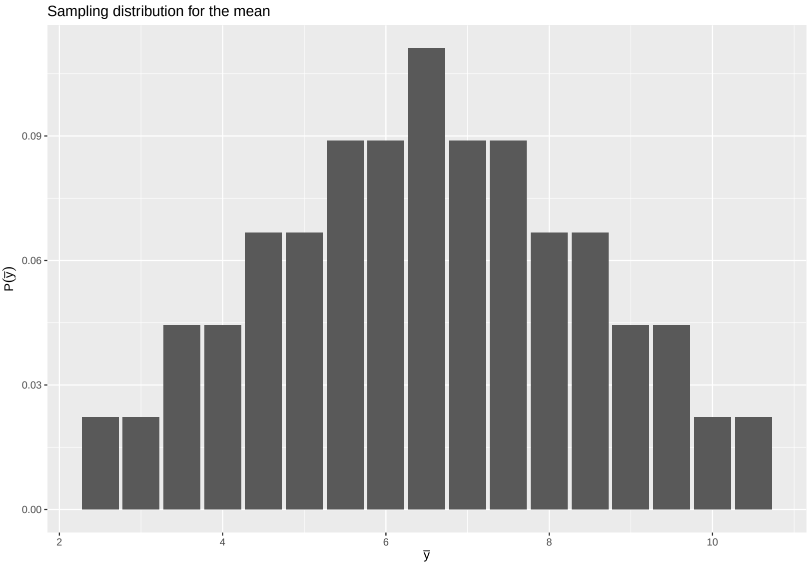
Constructing a Random Sample

- Let N be the population size
- In practice, we construct a list of elements in the population by assigning a number from 1 to N to each element in the population
- We then randomly select n integers from the integers $(1, 2, \ldots, N)$
- Most statistical software programs contain routines for randomly selecting n integers from the integers $(1, 2, \ldots, N)$
- In R, use the function `sample()`

Sampling Distributions

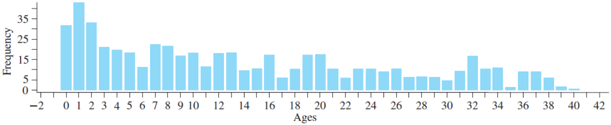
- The numerical value of a sample statistic cannot be predicted exactly in advance
- Even if we knew that a population mean is $\mu = 216.37$, we could not say that the sample mean $\bar{y}$ would be exactly equal to 216.37
- A sample statistic is a random variable
- Like any other random variable, a sample statistic has a probability distribution, called the **sampling distribution** of that statistic
- Example: a population consists of 10 values $(2, 3, 4, 5, 6, 7, 8, 9, 10, 11)$. Find the sampling distribution of the sample mean $\overline{Y}$, based on a random sample of size 2.

Sample	Value of $\bar{y}$	Sample	Value of $\bar{y}$	Sample	Value of $\bar{y}$
2,3	2.5	3,10	6.5	6,7	6.5
2,4	3	3,11	7	6,8	7
2,5	3.5	4,5	4.5	6,9	7.5
2,6	4	4,6	5	6,10	8
2,7	4.5	4,7	5.5	6,11	8.5
2,8	5	4,8	6	7,8	7.5
2,9	5.5	4,9	6.5	7,9	8
2,10	6	4,10	7	7,10	8.5
2,11	6.5	4,11	7.5	7,11	9
3,4	3.5	5,6	5.5	8,9	8.5
3,5	4	5,7	6	8,10	9
3,6	4.5	5,8	6.5	8,11	9.5
3,7	5	5,9	7	9,10	9.5
3,8	5.5	5,10	7.5	9,11	10
3,9	6	5,11	8	10,11	10.5

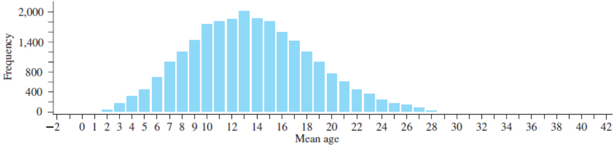

Sampling distribution for the mean
P(ȳ)
0.09
0.06
0.03
0.00
2
4
6
8
10
ȳ

Example

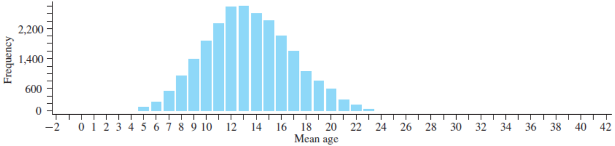
- A population consists of 500 pennies from which we compute the age of each penny
- The histogram of the 500 ages is displayed below



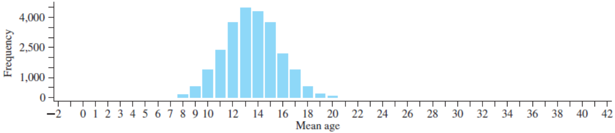
- Generate the sampling distribution of $\overline{Y}$ for samples of sizes $n = 5, n = 10$ and $n = 25$
- We will use a computer program to select 25,000 samples in order to obtain an approximation of the sampling distribution of $\overline{Y}$



(b) Sampling distribution of $\overline{y}$ for $n = 5$



(c) Sampling distribution of $\overline{y}$ for $n = 10$



(d) Sampling distribution of $\overline{y}$ for $n = 25$

- What do you notice about the shape of these three histograms?

Central Limit Theorem

- Let $\bar{y}$ denote the sample mean computed from a random sample of n measurements from a population having a mean, μ, and finite standard deviation σ
- Let $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ denote the mean and standard deviation of the sampling distribution of $\overline{Y}$, respectively
- Based on repeated random samples of size n from the population, we can conclude the following:
  1. $\mu_{\bar{y}} = \mu$
  2. $\sigma_{\bar{y}} = \sigma/\sqrt{n}$
  3. When n is large, the sampling distribution of $\overline{Y}$ will be approximately normal (with the approximation becoming more precise as n increases)
  4. When the population distribution is normal, the sampling distribution of $\overline{Y}$ is exactly normal for any sample size n

- How large should the sample size be for the Central Limit Theorem to hold?
- Numerous simulation studies have been conducted over the years and suggest that, in general, the Central Limit Theorem holds for $n > 30$
- However, one should not apply this rule blindly. It depends on the shape (skewness) of the population distribution

Example

- Suppose a patient's systolic blood pressure readings during a given day have a normal distribution with a mean 160 mm mercury and a standard deviation 20 mm
- If the systolic blood pressure exceeds 150, the patient is considered to have high blood pressure
    a. What is the probability that a single blood pressure measurement will fail to detect that the patient has high blood pressure?
    b. If five blood pressure measurements are taken at various times during the day, what is the probability that the average of the five measurements will be less than 150?
    c. How many measurements would be required in a given day so that there is at most 1% probability of failing to detect that the patient has high blood pressure?

Central Limit Theorem for Sums

- Suppose we have a random sample of n measurements $(Y_1, \ldots, Y_n)$ from a population and let $\sum Y = Y_1 + \ldots + Y_n$
- Let $\mu_{\sum y}$ and $\sigma_{\sum y}$ denote the mean and standard deviation of the sampling distribution of $\sum Y$, respectively
- Based on repeated random samples of size n from the population, we can conclude the following:
  1. $\mu_{\sum y} = n\mu$
  2. $\sigma_{\sum y} = \sigma\sqrt{n}$
  3. When n is large, the sampling distribution of $\sum Y$ will be approximately normal (with the approximation becoming more precise as n increases)
  4. When the population distribution is normal, the sampling distribution of $\sum Y$ is exactly normal for any sample size n

Normal Approximation to the Binomial

Example

- Suppose a sample of 1,000 voters is polled to determine sentiment toward the consolidation of city and county government
- What would be the probability of observing 460 or fewer favoring consolidation if we assume that 50% of the entire population favor the change?

Example (ctn'd)

Use the normal approximation to the binomial to compute the probability of observing 460 or fewer in a sample of 1,000 favoring consolidation if we assume that 50% of the entire population favor the change