6.1 Statistics and Their Distributions

6.2 The Distribution of Sample Totals, Means, and Proportions

DEFINITION A **statistic** is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, *a statistic is a random variable* and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

PROPOSITION Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

1. $E(T_o) = n\mu$
2. $V(T_o) = n\sigma^2$ and $\sigma_{T_o} = \sqrt{n}\sigma$
3. If the X_i's are normally distributed, then T_o is also normally distributed.

1. $E(\overline{X}) = \mu$
2. $V(\overline{X}) = \dfrac{\sigma^2}{n}$ and $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$
3. If the X_i's are normally distributed, then $\overline{X}$ is also normally distributed.

11. The inside diameter of a randomly selected piston ring is a random variable with mean value 12 cm and standard deviation .04 cm.

   a. If $\overline{X}$ is the sample mean diameter for a random sample of $n = 16$ rings, where is the sampling distribution of $\overline{X}$ centered, and what is the standard deviation of the $\overline{X}$ distribution?
   b. Answer the questions posed in part (a) for a sample size of $n = 64$ rings.
   c. For which of the two random samples, the one of part (a) or the one of part (b), is $\overline{X}$ more likely to be within .01 cm of 12 cm? Explain your reasoning.

CENTRAL LIMIT THEOREM (CLT) Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Then, in the limit as $n \to \infty$, the standardized versions of $\overline{X}$ and T_o have the standard normal distribution. That is,

$$\lim_{n\to\infty} P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z\right) = P(Z \le z) = \Phi(z)$$

and

$$\lim_{n\to\infty} P\left(\frac{T_o - n\mu}{\sqrt{n}\sigma} \le z\right) = P(Z \le z) = \Phi(z)$$

where Z is a standard normal rv. It is customary to say that $\overline{X}$ and T_o are **asymptotically normal**, and that their standardized versions **converge in distribution** to Z. Thus when n is sufficiently large, $\overline{X}$ has approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Equivalently, for large n the sum T_o has approximately a normal distribution with $\mu_{T_o} = n\mu$ and $\sigma_{T_o} = \sqrt{n}\sigma$.

13. The National Health Statistics Reports dated Oct. 22, 2008 stated that for a sample size of 277 18-year-old American males, the sample mean waist circumference was 86.3 cm. A somewhat complicated method was used to *estimate* various population percentiles, resulting in the following values:

5th	10th	25th	50th	75th	90th	95th
69.6	70.9	75.2	81.3	95.4	107.1	116.4

a. Is it plausible that the waist size distribution is at least approximately normal? Explain your reasoning. If your answer is no, conjecture the shape of the population distribution.

b. Suppose that the population mean waist size is 85 cm and that the population standard deviation is 15 cm. How likely is it that a random sample of 277 individuals will result in a sample mean waist size of at least 86.3 cm?

8.1 Basic Properties of Confidence Intervals

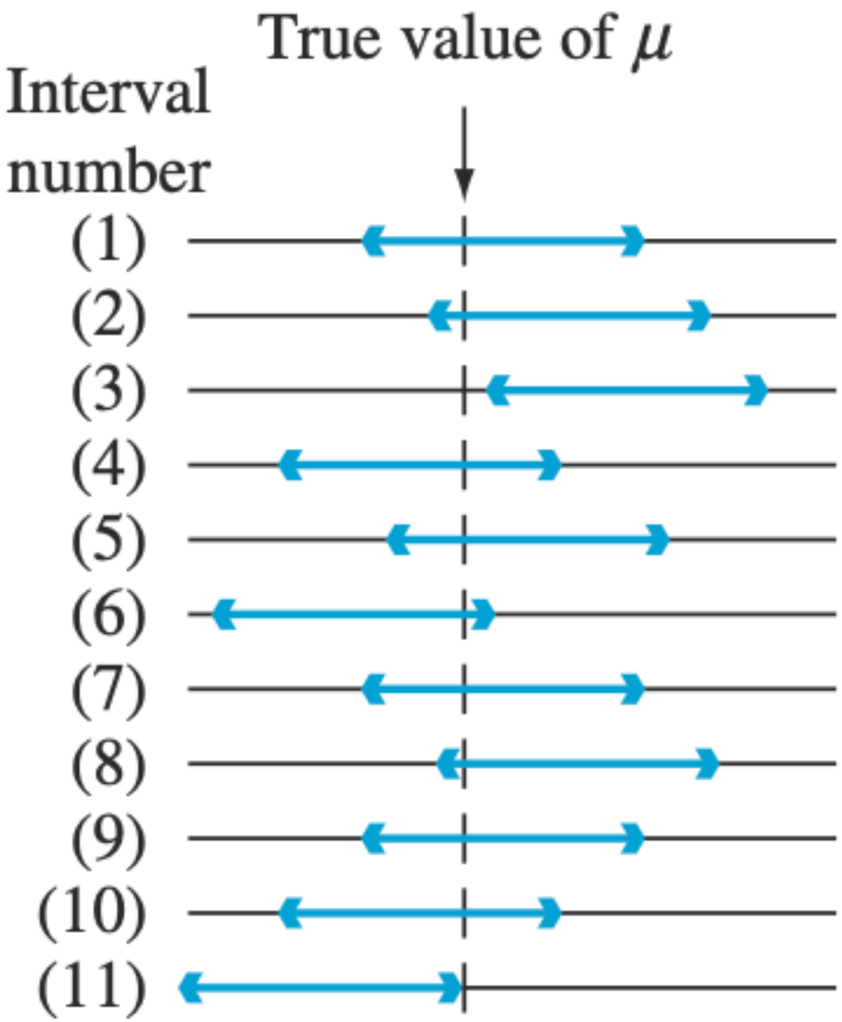
DEFINITION If after observing $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n,$ we compute the observed sample mean $\bar{x}$ and then substitute $\bar{x}$ into (8.4) in place of $\overline{X}$, the resulting fixed interval is called a **95% confidence interval for μ**. This CI can be expressed either as

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \text{ is a 95\% confidence interval for } \mu$$

or as

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ with 95\% confidence}$$

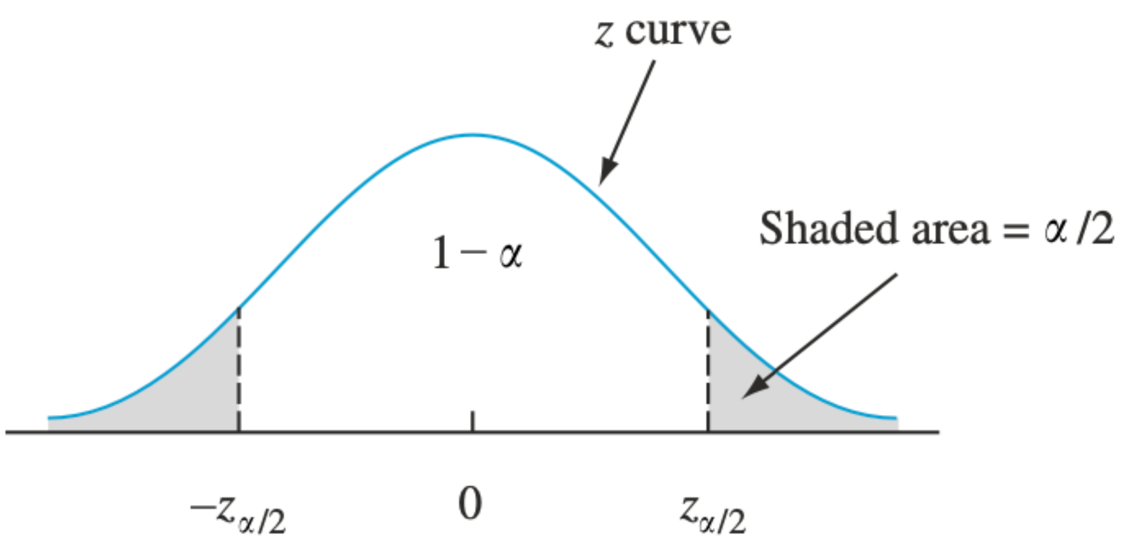
A concise expression for the interval is $\bar{x} \pm 1.96 \cdot \sigma/\sqrt{n}$, where $-$ gives the left endpoint (lower limit) and + gives the right endpoint (upper limit).


True value of μ
Interval number
(1)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
(9)
(10)
(11)

DEFINITION A **100(1 – α)% confidence interval** for the mean μ of a normal population when the value of σ is known is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{8.5}$$

or, equivalently, by $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.



1. Consider a normal population distribution with the value of σ known.

   a. What is the confidence level for the interval $\bar{x} \pm 2.81\sigma/\sqrt{n}$?

   b. What is the confidence level for the interval $\bar{x} \pm 1.44\sigma/\sqrt{n}$?

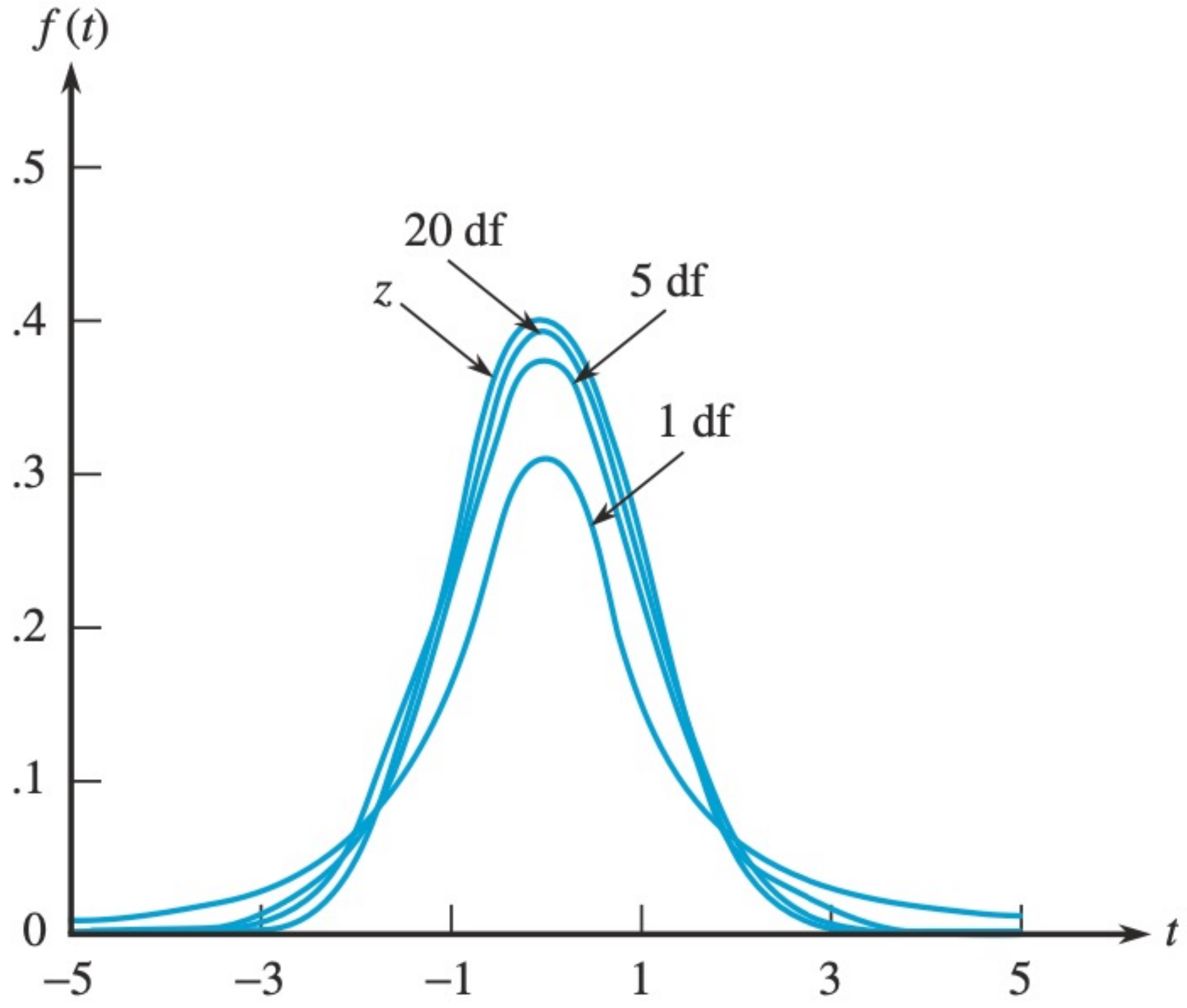
3. Suppose that a random sample of 50 bottles of a particular brand of cough syrup is selected and the alcohol content of each bottle is determined. Let μ denote the average alcohol content for the population of all bottles of the brand under study. Suppose that the resulting 95% confidence interval is (7.8, 9.4).

   a. Would a 90% confidence interval calculated from this same sample have been narrower or wider than the given interval? Explain your reasoning.
   b. Consider the following statement: There is a 95% chance that μ is between 7.8 and 9.4. Is this statement correct? Why or why not?
   c. Consider the following statement: We can be highly confident that 95% of all bottles of this type of cough syrup have an alcohol content that is between 7.8 and 9.4. Is this statement correct? Why or why not?
   d. Consider the following statement: If the process of selecting a sample of size 50 and then computing the corresponding 95% interval is repeated 100 times, 95 of the resulting intervals will include μ. Is this statement correct? Why or why not?

8.2 The One-Sample t Interval and Its Relatives

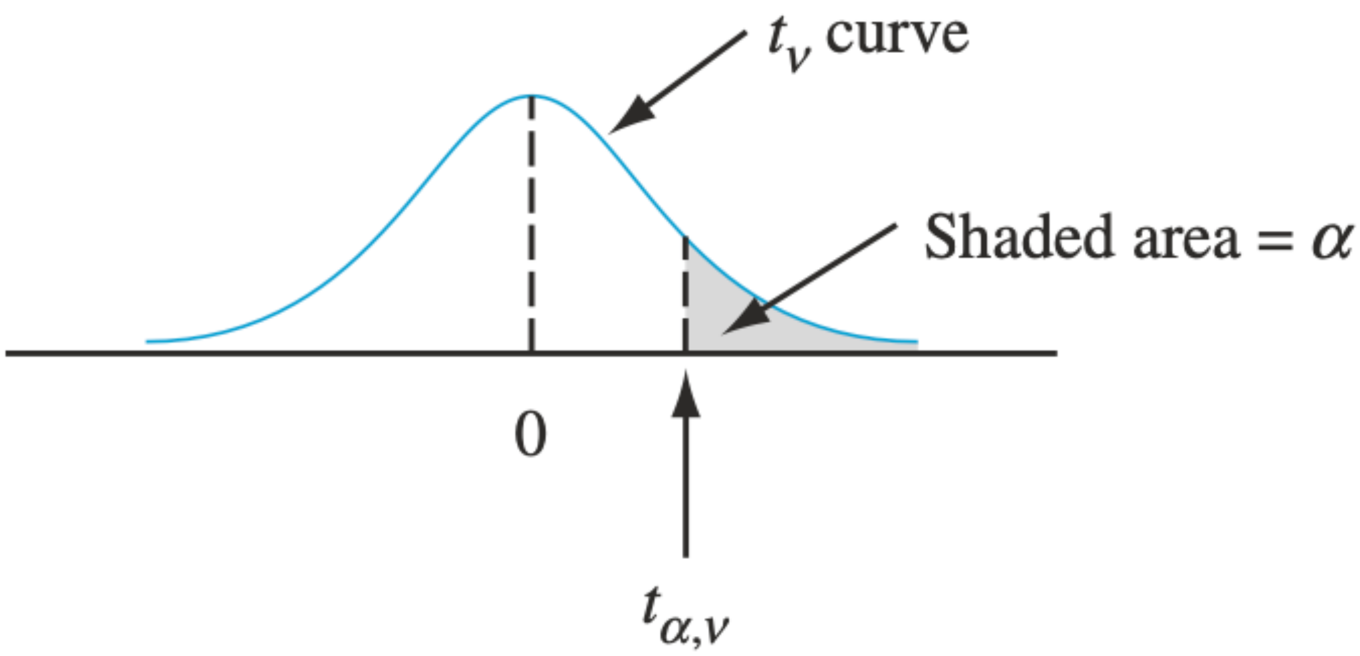
PROPOSITION The pdf of a random variable T having a t distribution with ν degrees of freedom is

$$f(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{(1+t^2/\nu)^{(\nu+1)/2}} \qquad -\infty < t < \infty$$



f (t)
.5
.4
.3
.2
.1
0
20 df
z
5 df
1 df
–5
–3
–1
1
3
5
t

NOTATION Let $t_{\alpha,v}$ = the number on the measurement axis for which the area under the t curve with v df to the right of $t_{\alpha,v}$ is α; $t_{\alpha,v}$ is called a ***t* critical value**.



PROPOSITION Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from the results of a random sample from a *normal* population with mean μ. Then a **100(1 − $\boldsymbol{\alpha}$)% confidence interval for $\boldsymbol{\mu}$**, also called the **one-sample $\boldsymbol{t}$ CI**, is

$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right) \tag{8.11}$$

or, more compactly, $\bar{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}$.

13. Determine the values of the following quantities:

a. $t_{.1,15}$ b. $t_{.05,15}$ c. $t_{.05,25}$ d. $t_{.05,40}$ e. $t_{.005,40}$

15. Determine the t critical value for a two-sided confidence interval in each of the following situations:

a. Confidence level = 95%, df = 10
b. Confidence level = 95%, df = 15
c. Confidence level = 99%, df = 15
d. Confidence level = 99%, $n = 5$
e. Confidence level = 98%, df = 24
f. Confidence level = 99%, $n = 38$

17. Here are the alcohol percentages for a random sample of 16 beers (light beers excluded):

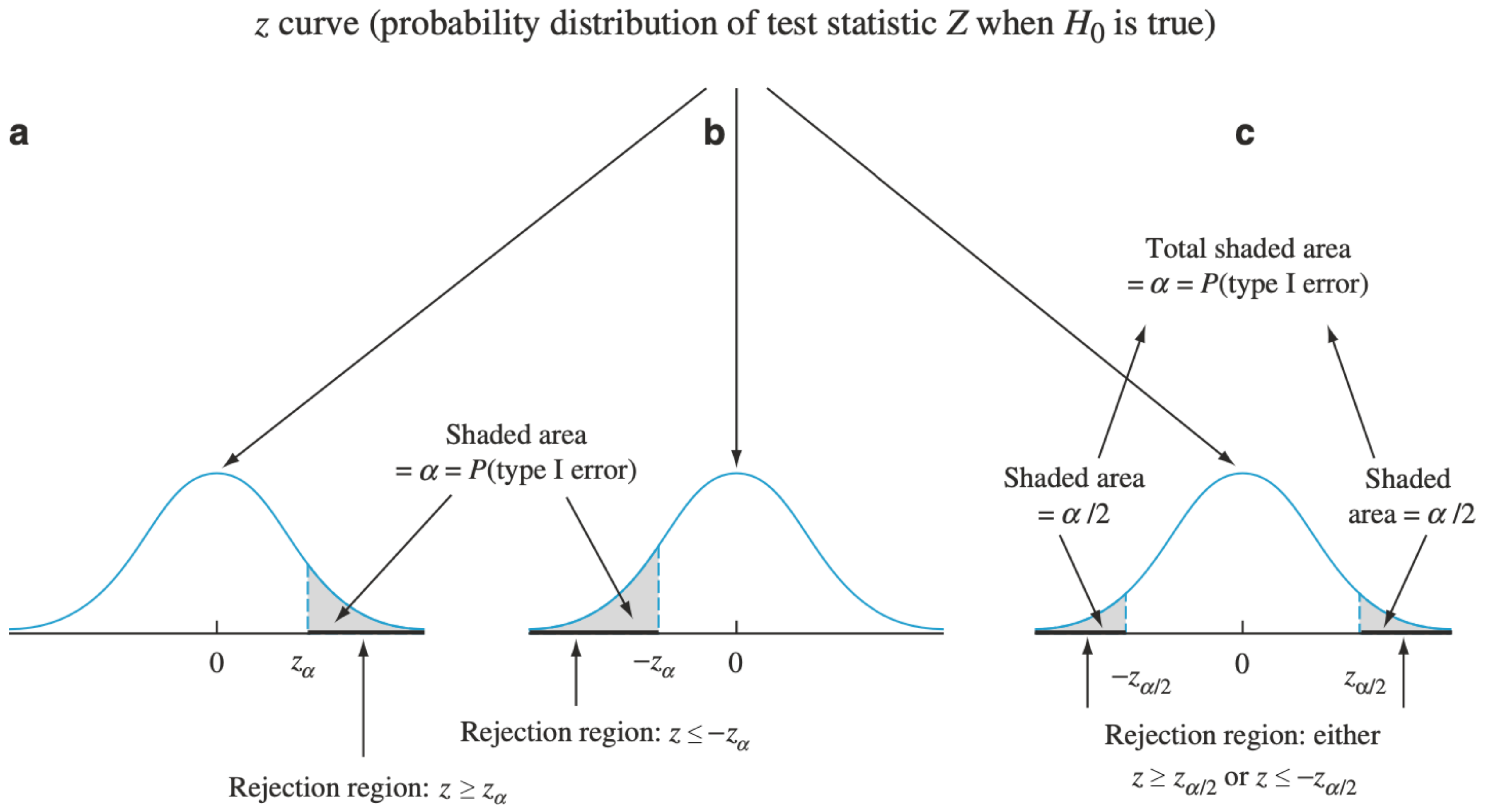
4.68	4.13	4.80	4.63	5.08	5.79	6.29	6.79
4.93	4.25	5.70	4.74	5.88	6.77	6.04	4.95

a. Construct a normal probability plot of the data. Is it plausible that these values represent a sample from a normal distribution?
b. Calculate a 95% CI for the mean alcohol percentage of all nonlight beers.
c. Calculate a 95% CI for the mean amount of alcohol, in ounces, in a 12-oz. serving of (again, nonlight) beer.

9.1 Hypotheses and Test Procedures

DEFINITION The **null hypothesis**, denoted by H_0, is the claim that is initially assumed to be true (the "prior belief" claim). The **alternative hypothesis**, denoted by H_a, is the assertion that is contradictory to H_0.

9.2 Tests About a Population Mean


z curve (probability distribution of test statistic Z when H_0 is true)
a
b
c
Shaded area
$= \alpha = P$(type I error)
0
z_α
Rejection region: $z \geq z_\alpha$
$-z_\alpha$
0
Rejection region: $z \leq -z_\alpha$
Total shaded area
$= \alpha = P$(type I error)
Shaded area
$= \alpha/2$
Shaded
area $= \alpha/2$
$-z_{\alpha/2}$
0
$z_{\alpha/2}$
Rejection region: either
$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

THE ONE-SAMPLE z TEST

Null hypothesis: H_0: $\mu = \mu_0$

Test statistic value: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
H_a: $\mu > \mu_0$	$z \geq z_\alpha$ (upper-tailed test)
H_a: $\mu < \mu_0$	$z \leq -z_\alpha$ (lower-tailed test)
H_a: $\mu \neq \mu_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

15. Let the test statistic Z have a standard normal distribution when H_0 is true. Give the significance level for each of the following situations:

a. H_a: $\mu > \mu_0$, rejection region $z \geq 1.88$
b. H_a: $\mu < \mu_0$, rejection region $z \leq -2.75$
c. H_a: $\mu \neq \mu_0$, rejection region $z \geq 2.88$ or $z \leq -2.88$

Example 9.10 If the activation temperature of an automated sprinkler system used for fire protection in an office building is too high, a fire could do substantial damage before water is dispersed. On the other hand, activation at too low a temperature could cause water damage when there is little fire threat. A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°. A sample of $n = 9$ systems, when tested, yields a sample average activation temperature of 131.08 °F. If the distribution of activation times is normal with standard deviation 1.5 °F, does the data contradict the manufacturer's claim at significance level $\alpha = .01$?

THE ONE-SAMPLE t TEST

Null hypothesis: H_0: $\mu = \mu_0$

Test statistic value: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	Rejection Region for a Level α Test
H_a: $\mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed)
H_a: $\mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed)
H_a: $\mu \neq \mu_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed)

Example 9.12 Particulate matter from roads contributes to pollution when those particles are washed into nearby waterways by rain. The size of the particles can impact the effectiveness of various stormwater control measures. The authors of the article “Characterizing Runoff from Roads: Particle Size Distributions, Nutrients, and Gross Solids” (*J. Environ. Engr.* 2016) took roadside measurements at several sites in North Carolina. For each assay they recorded d_{50}, the median size of particles in the assay (a standard measure of particle size in such studies). Here are the d_{50} values (microns) for $n = 9$ assays performed off I-40 near Black Mountain:

$$82.9 \quad 56.8 \quad 66.5 \quad 49.4 \quad 105.4 \quad 79.5 \quad 82.5 \quad 50.7 \quad 43.0$$

Previous studies indicated that the typical d_{50} value alongside roads of this type is 44 microns. Does the sample data provide convincing statistical evidence that the true mean d_{50} value differs from 44 microns? Let’s carry out a test using a significance level of $\alpha = .01$.

27. On the label, Pepperidge Farm bagels are said to weigh four ounces each (113 g). A random sample of six bagels resulted in the following weights (in grams):

117.6 109.5 111.6 109.2 119.1 110.8

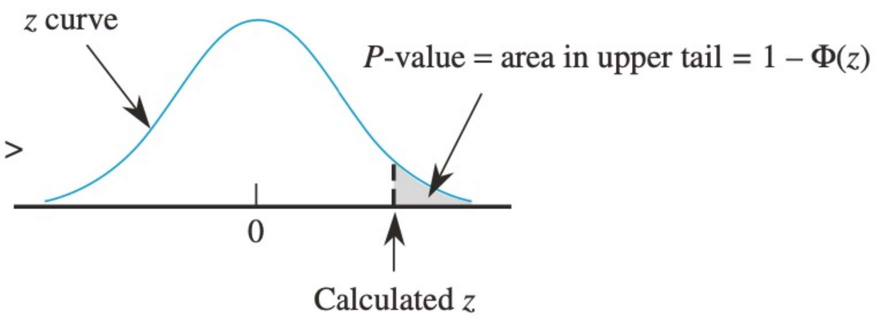
a. Based on this sample, is there any reason to doubt that the population mean is at least 113 g?

9.4 *P*-Values

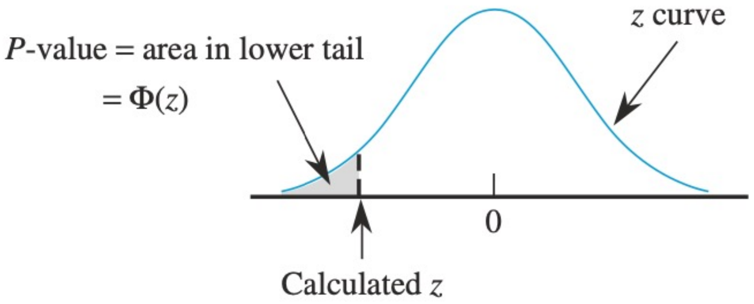
DEFINITION The ***P*-value** is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample.

- The P-value is a probability.
- This probability is calculated assuming that H_0 is true.
- The P-value is a function of the sample data.
- To determine the P-value, we must decide which values of the test statistic are "at least as contradictory to H_0" as the value obtained from our sample.

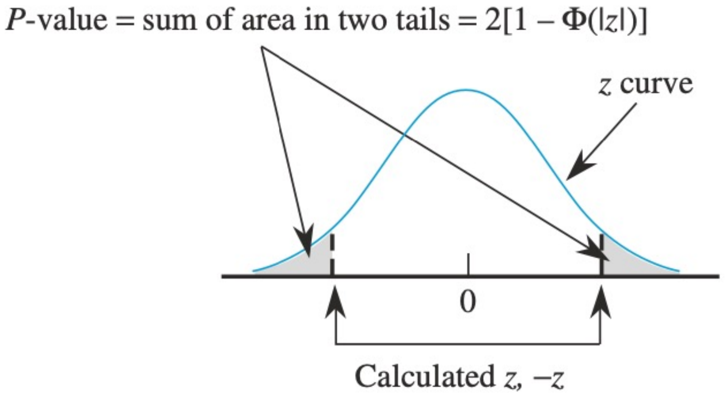
1. **Upper-tailed test**
H_a contains the inequality $>$



2. **Lower-tailed test**
H_a contains the inequality $<$



3. **Two-tailed test**
H_a contains the inequality $\neq$



49. For which of the given P-values would the null hypothesis be rejected when performing a level .05 test?

a. 001 b. .021 c. .078 d. .047 e. .148

51. Let μ denote the mean reaction time to a certain stimulus. For a one-sample z test of H_0: $\mu = 5$ versus H_a: $\mu > 5$ (i.e., assuming σ is known), find the P-value associated with each of the given values of the z test statistic.

a. 1.42 b. .90 c. 1.96 d. 2.48 e. –.11

Example 9.21 The recommended daily intake of calcium for adults ages 18–30 is 1000 mg/day. The article “Dietary and Total Calcium Intakes Are Associated with Lower Percentage Total Body and Truncal Fat in Young, Healthy Adults” (*J. Amer. College of Nutr.* 2011: 484–490) reported the following summary data for a sample of 76 healthy Caucasian males from southwestern Ontario, Canada: $n = 76$, $\bar{x} = 1093$, $s = 477$. Let’s carry out a test at significance level .01 to see whether the population mean daily intake exceeds the recommended value.

10.1 The Two-Sample z Confidence Interval and Test

ASSUMPTIONS

1. $X_1, X_2, \ldots, X_m$ is a random sample from a population with mean μ_1 and standard deviation σ_1.
2. $Y_1, Y_2, \ldots, Y_n$ is a random sample from a population with mean μ_2 and standard deviation σ_2.
3. The X and Y samples are independent of each other.

PROPOSITION The expected value of $\overline{X} - \overline{Y}$ is $\mu_1 - \mu_2$, so $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. The standard deviation of $\overline{X} - \overline{Y}$ is

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

TWO-SAMPLE z INTERVAL

Assuming independent random samples from normal population distributions, a CI for $\mu_1 - \mu_2$ with a confidence level of $100(1 - \alpha)\%$ has endpoints

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

TWO-SAMPLE *z* TEST

Null hypothesis: H_0: $\mu_1 - \mu_2 = \Delta_0$

Test statistic value: $z = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
H_a: $\mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (upper-tailed test)
H_a: $\mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (lower-tailed test)
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

3. Let μ_1 and μ_2 denote true average tread lives (miles) for two competing brands of size P205/65R15 tires.

   a. Test H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 \neq 0$ at level .05 using the following information: $m = 45$, $\bar{x} = 42,500$, $\sigma_1 = 2200$, $n = 45$, $\bar{y} = 40{,}400$, and $\sigma_2 = 1900$.

   b. Use the information in part (a) to compute a 95% CI for $\mu_1 - \mu_2$. Does the resulting interval suggest that $\mu_1 - \mu_2$ has been precisely estimated?

10.2 The Two-Sample t Confidence Interval and Test

WELCH'S THEOREM When the population distributions are both normal, the standardized variable

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{m} + \dfrac{S_2^2}{n}}} \tag{10.2}$$

has approximately a t distribution with df v estimated from the data by

$$v = \frac{\left(\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}\right)^2}{\dfrac{(s_1^2/m)^2}{m-1} + \dfrac{(s_2^2/n)^2}{n-1}} = \frac{\left[(se_1)^2 + (se_2)^2\right]^2}{\dfrac{(se_1)^4}{m-1} + \dfrac{(se_2)^4}{n-1}} \tag{10.3}$$

where $se_1 = s_1/\sqrt{m}$ and $se_2 = s_2/\sqrt{n}$ (round v down to the nearest integer).

TWO-SAMPLE *t* PROCEDURES The **two-sample *t* confidence interval for $\mu_1 - \mu_2$** with approximate confidence level $100(1 - \alpha)\%$ is

$$\bar{x} - \bar{y} \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where ν = Welch's df formula (10.3). One-sided confidence bounds can be calculated by retaining the appropriate sign (+ or −) and replacing $t_{\alpha/2,\nu}$ by $t_{\alpha,\nu}$.

The **two-sample *t* test** for testing H_0: $\mu_1 - \mu_2 = \Delta_0$ is as follows:

Alternative Hypothesis	**Rejection Region for Approximate Level α Test**
H_a: $\mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha,\nu}$ (upper-tailed test)
H_a: $\mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha,\nu}$ (lower-tailed test)
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	either $t \geq t_{\alpha/2,\nu}$ or $t \leq -t_{\alpha/2,\nu}$ (two-tailed test)

Test statistic value: $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}}$

A *P*-value can be computed as described in Section 9.4 for the one-sample *t* test.

2. Suppose that a certain drug A was administered to eight patients selected at random, and after a fixed time period, the concentration of the drug in certain body cells of each patient was measured in appropriate units. Suppose that these concentrations for the eight patients were found to be as follows:

$$1.23,\ 1.42,\ 1.41,\ 1.62,\ 1.55,\ 1.51,\ 1.60,\ \text{ and } 1.76.$$

Suppose also that a second drug B was administered to six different patients selected at random, and when the concentration of drug B was measured in a similar way for these six patients, the results were as follows:

$$1.76,\ 1.41,\ 1.87,\ 1.49,\ 1.67,\ \text{ and } 1.81.$$

Assuming that all the observations have a normal distribution with a common unknown variance, test the following hypotheses at the level of significance 0.10: The null hypothesis is that the mean concentration of drug A among all patients is at least as large as the mean concentration of drug B. The alternative hypothesis is that the mean concentration of drug B is larger than that of drug A.