INTERVAL ESTIMATION, THE BOOTSTRAP AND ERROR BARS
(ESTIMATION II DA_2022)

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

Lecture n. 13, Rome 13th April 2022

Dipartimento di Fisica

Sapienza
Università di Roma

Outline L 13

Point estimation/Interval estimation

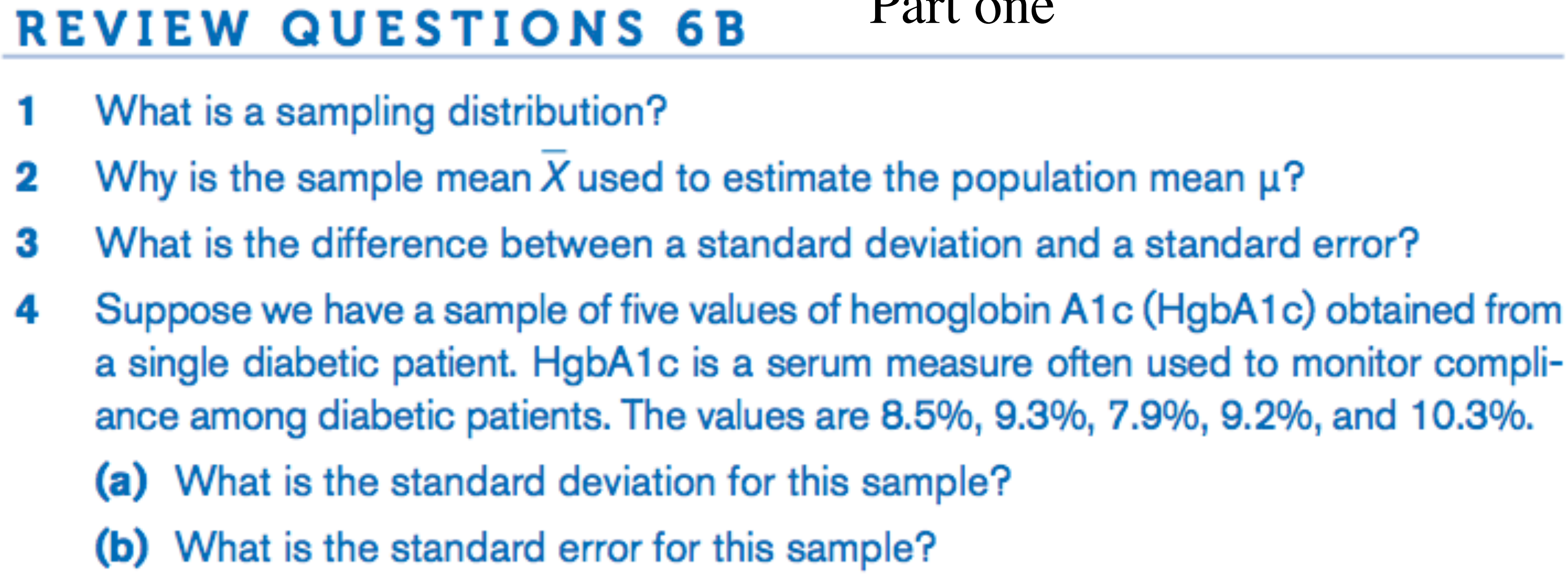
DA_2022 LECTURE 13 WED APR 13 11-13 auletta 2 CU026
-The bootstrap Montecarlo method to evaluate estimates and confidence interval
from just one sample (see R chap 6.7, and W&S chap 19)
-confidence intervals definition and evaluation using t-distributions and chi-square distributions formulas and examples.

HOMEWORK to be done on the LOG-BOOK
REVIEW QUESTION 6B AND & 6C in Rosners' textbook

The study material for this lecture can be found in chap. 6 of Rosner's textbook

WRITTEN HOMEWORK DUE MONDAY APRIL 25
DA_2022_HW_25_4
To be collected in your DA_2022 **logbook**

Part one



REVIEW QUESTIONS 6B

1 What is a sampling distribution?

2 Why is the sample mean $\overline{X}$ used to estimate the population mean μ?

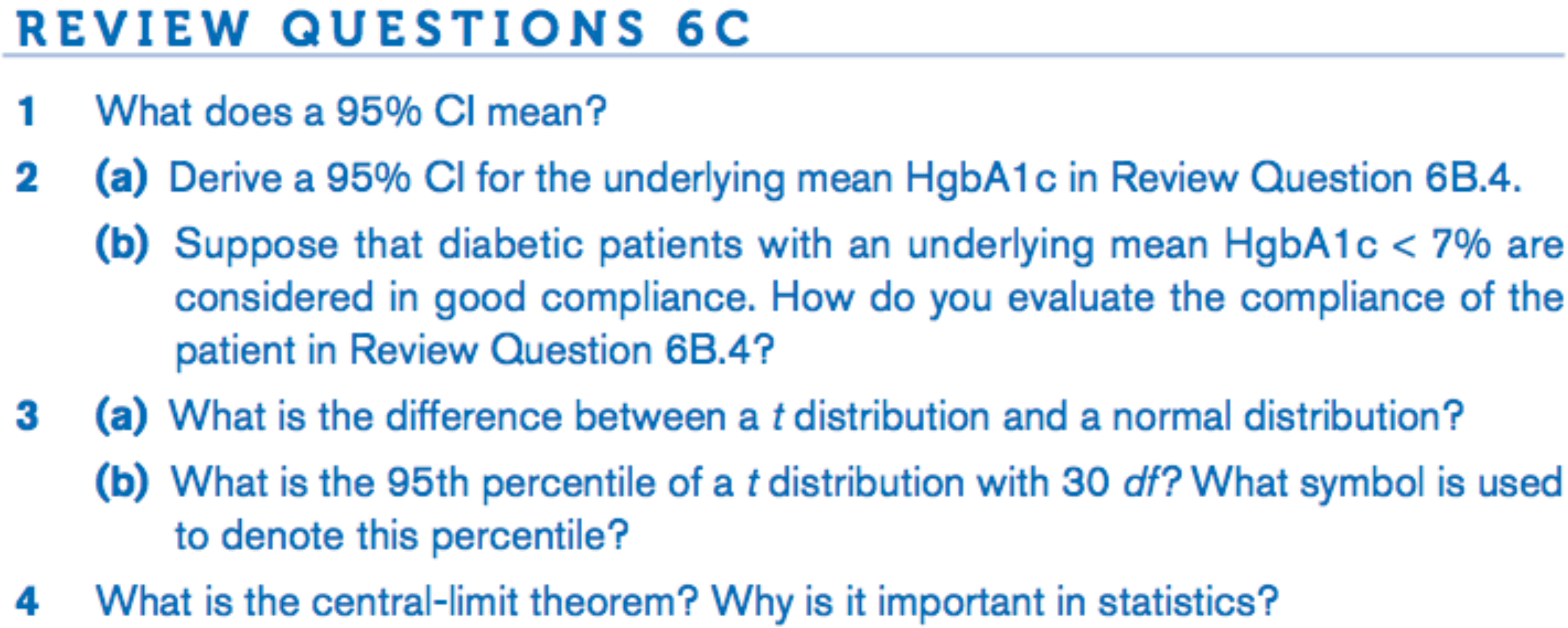
3 What is the difference between a standard deviation and a standard error?

4 Suppose we have a sample of five values of hemoglobin A1c (HgbA1c) obtained from a single diabetic patient. HgbA1c is a serum measure often used to monitor compliance among diabetic patients. The values are 8.5%, 9.3%, 7.9%, 9.2%, and 10.3%.

(a) What is the standard deviation for this sample?

(b) What is the standard error for this sample?

Part two



REVIEW QUESTIONS 6C

1 What does a 95% CI mean?

2 (a) Derive a 95% CI for the underlying mean HgbA1c in Review Question 6B.4.

(b) Suppose that diabetic patients with an underlying mean HgbA1c < 7% are considered in good compliance. How do you evaluate the compliance of the patient in Review Question 6B.4?

3 (a) What is the difference between a *t* distribution and a normal distribution?

(b) What is the 95th percentile of a *t* distribution with 30 *df*? What symbol is used to denote this percentile?

4 What is the central-limit theorem? Why is it important in statistics?

Exercise check the central limit theorem

6.12 Summary

This chapter introduced the concept of a sampling distribution. This concept is crucial to understanding the principles of statistical inference. The fundamental idea is to forget about our sample as a unique entity and instead regard it as a random sample from all possible samples of size n that could have been drawn from the population under study. Using this concept, $\bar{X}$ was shown to be an unbiased estimator of the population mean μ; that is, the average of all sample means over all possible random samples of size n that could have been drawn will equal the population mean. Furthermore, if our population follows a normal distribution, then $\bar{X}$ has minimum variance among all possible unbiased estimators and is thus called a *minimum-variance unbiased estimator* of μ. Finally, if our population follows a normal distribution, then $\bar{X}$ also follows a normal distribution. However, even if our population is not normal, the sample mean still approximately follows a normal distribution for a sufficiently large sample size. This very important idea, which justifies many of the hypothesis tests we study in the rest of this book, is called the *central-limit theorem*.

The idea of an interval estimate (or CI) was then introduced. Specifically, a 95% CI is defined as an interval that will contain the true parameter for 95% of all random samples that could have been obtained from the reference population. The preceding principles of point and interval estimation were applied to the following:

(1) Estimating the mean μ of a normal distribution

(2) Estimating the variance σ^2 of a normal distribution

(3) Estimating the parameter p of a binomial distribution

(4) Estimating the parameter λ of a Poisson distribution

(5) Estimating the expected value μ of a Poisson distribution

The t and chi-square distributions were introduced to obtain interval estimates for (1) and (2), respectively. Finally, the bootstrap CI was introduced to obtain confidence limits for the mean when the assumption of normality is questionable, and can also be applied to obtain confidence limits for other parameters from other distributions.

In Chapters 7 through 14, the discussion of statistical inference continues, focusing primarily on testing hypotheses rather than on parameter estimation. In this regard, some parallels between inference from the points of view of hypothesis testing and CIs are discussed.

6.7 ESTIMATION OF THE VARIANCE OF A DISTRIBUTION

Point Estimation

In Chapter 2, the sample variance was defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

This definition is somewhat counterintuitive because the denominator would be expected to be n rather than $n - 1$. A more formal justification for this definition is now given. If our sample $x_1, \ldots, x_n$ is considered as coming from some population with mean μ and variance σ^2, then how can the unknown population variance σ^2 be estimated from our sample? The following principle is useful in this regard:

6.10 Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2. The **sample variance S^2 is an unbiased estimator** of σ^2 over all possible random samples of size n that could have been drawn from this population; that is, $E(S^2) = \sigma^2$.

Therefore, if repeated random samples of size n are selected from the population, as was done in Table 6.3, and the sample variance s^2 is computed from each sample, then the average of these sample variances over a large number of such samples of size n is the population variance σ^2. This statement holds for any underlying distribution.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

tends to underestimate the underlying variance σ^2 by a factor of $(n - 1)/n$. This factor is considerable for small samples but tends to be negligible for large samples. A more complete discussion of the relative merits of different estimators for σ^2 is given in [3].

- Example 2: We are looking for an estimator of the variance of a random variable, we propose the following two estimators:

$$S_n^{\ 2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}{n} \quad \text{and} \quad R_n^{\ 2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}{n-1}$$

Which, in your opinion, is the best estimator of σ^2 ?

$S_n^{\ 2}=1/n\sum(X_i^{\ 2}-2X_i\bar{X}_n+\bar{X}_n^{\ 2})$ and

$S_n^{\ 2}=1/n\sum X_i^{\ 2}-2T_n\sum X_i/n+T_n^2=1/n\sum X_i^{\ 2}-T_n^2$

then $E(S_n^2)=E(X^2)-E(T_n^2)=\sigma^2+m^2-V(T_n)-E(T_n)^2$

and $E(S_n^{\ 2})=\sigma^2-\dfrac{\sigma^2}{n}$ so $E(S_n^{\ 2})=\dfrac{n-1}{n}\sigma^2$ and $b_{S_n^2}=-\dfrac{\sigma^2}{n}$

.

By a quite similar calculation we find $b_{R_n^2}=0$.

This quantities are two estimators of the variance because they are unbiased or asymptotically unbiased.

$R_n^{\ 2}$ is the best estimator because it has no bias.

To obtain an interval estimate for σ^2, a new family of distributions, called chi-square (χ^2) distributions, must be introduced to enable us to find the sampling distribution of S^2 from sample to sample.

DEFINITION 6.14

If $G = \sum_{i=1}^{n} X_i^2$

where $X_1, \ldots, X_n \sim N(0,1)$

and the X_i's are independent, then G is said to follow a **chi-square distribution with n degrees of freedom** (*df*). The distribution is often denoted by χ_n^2.

The chi-square distribution is actually a family of distributions indexed by the parameter n referred to, again, as the degrees of freedom, as was the case for the t distribution. Unlike the t distribution, which is always symmetric about 0 for any degrees of freedom, the chi-square distribution only takes on positive values and is always skewed to the right. The general shape of these distributions is indicated in Figure 6.8.

For $n = 1, 2$, the distribution has a mode at 0 [3]. For $n \geq 3$, the distribution has a mode greater than 0 and is skewed to the right. The skewness diminishes as n increases. It can be shown that the expected value of a χ_n^2 distribution is n and the variance is $2n$.

DEFINITION 6.15

The uth percentile of a χ_d^2 distribution (i.e., a chi-square distribution with d *df*) is denoted by $\chi_{d,u}^2$, where $Pr(\chi_d^2 < \chi_{d,u}^2) \equiv u$. These percentiles are shown in Figure 6.9 for a chi-square distribution with 5 *df* and appear in Table 6 in the Appendix.

FIGURE 6.8 General shape of various χ^2 distributions with *d df*

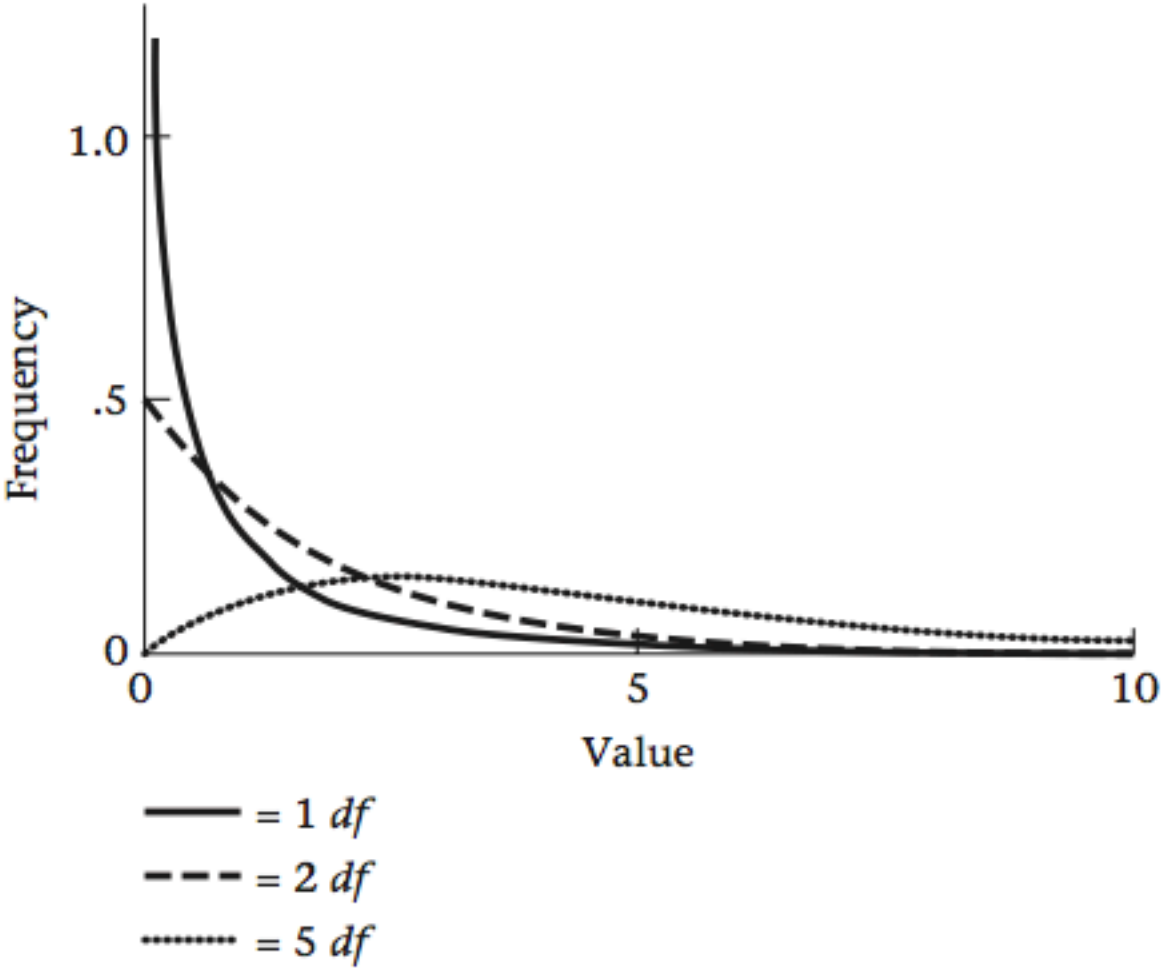
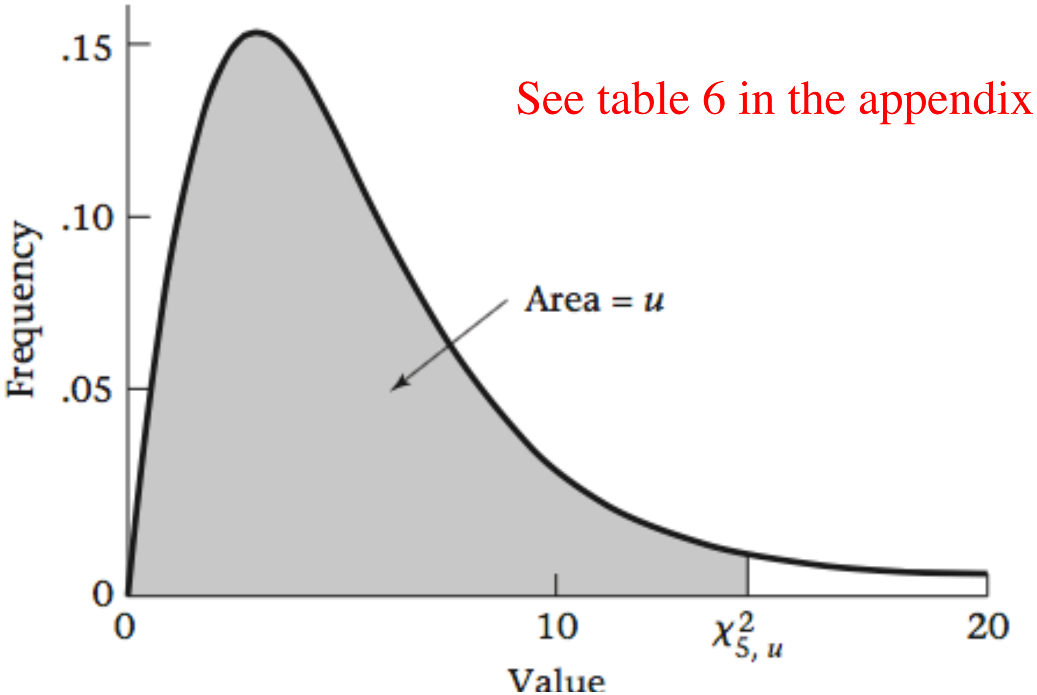


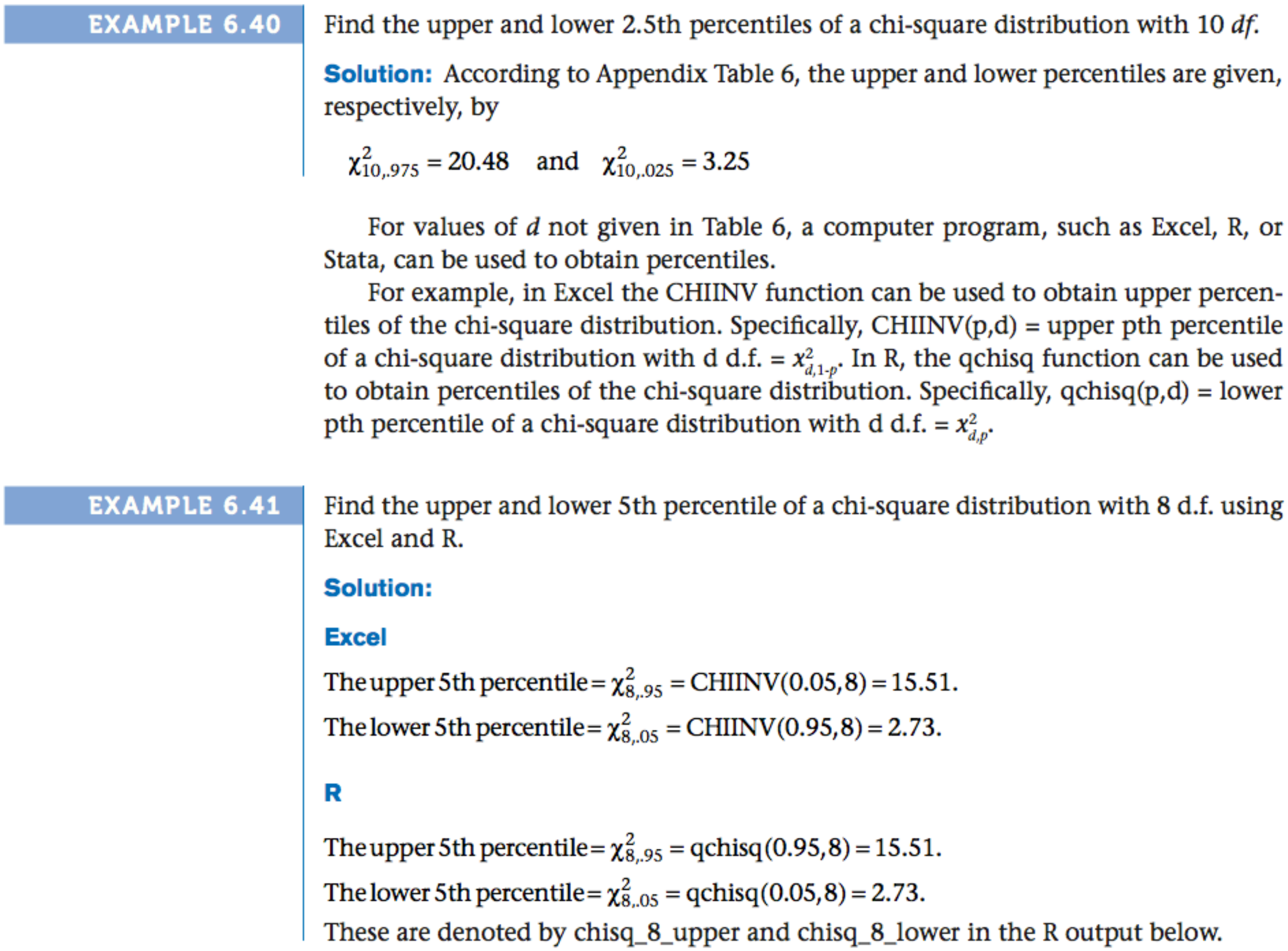
FIGURE 6.9 Graphic display of the percentiles of a χ_5^2 distribution



See table 6 in the appendix

Table 6 is constructed like the *t* table (Table 5), with the degrees of freedom (*d*) indexed in the first column and the percentile (*u*) indexed in the first row. The main difference between the two tables is that both *lower* ($u \le 0.5$) and *upper* ($u > 0.5$) percentiles are given for the chi-square distribution, whereas only upper percentiles are given for the *t* distribution. The *t* distribution is symmetric about 0, so any lower percentile can be obtained as the negative of the corresponding upper percentile. Because the chi-square distribution is, in general, a skewed distribution, there is no simple relationship between the upper and lower percentiles.

Practical numerical aspects



EXAMPLE 6.40 Find the upper and lower 2.5th percentiles of a chi-square distribution with 10 *df*.

Solution: According to Appendix Table 6, the upper and lower percentiles are given, respectively, by

$$\chi^2_{10,.975} = 20.48 \quad \text{and} \quad \chi^2_{10,.025} = 3.25$$

For values of *d* not given in Table 6, a computer program, such as Excel, R, or Stata, can be used to obtain percentiles.

For example, in Excel the CHIINV function can be used to obtain upper percentiles of the chi-square distribution. Specifically, CHIINV(p,d) = upper pth percentile of a chi-square distribution with d d.f. = $x^2_{d,1-p}$. In R, the qchisq function can be used to obtain percentiles of the chi-square distribution. Specifically, qchisq(p,d) = lower pth percentile of a chi-square distribution with d d.f. = $x^2_{d,p}$.

EXAMPLE 6.41 Find the upper and lower 5th percentile of a chi-square distribution with 8 d.f. using Excel and R.

Solution:

Excel

The upper 5th percentile = $\chi^2_{8,.95}$ = CHIINV(0.05,8) = 15.51.

The lower 5th percentile = $\chi^2_{8,.05}$ = CHIINV(0.95,8) = 2.73.

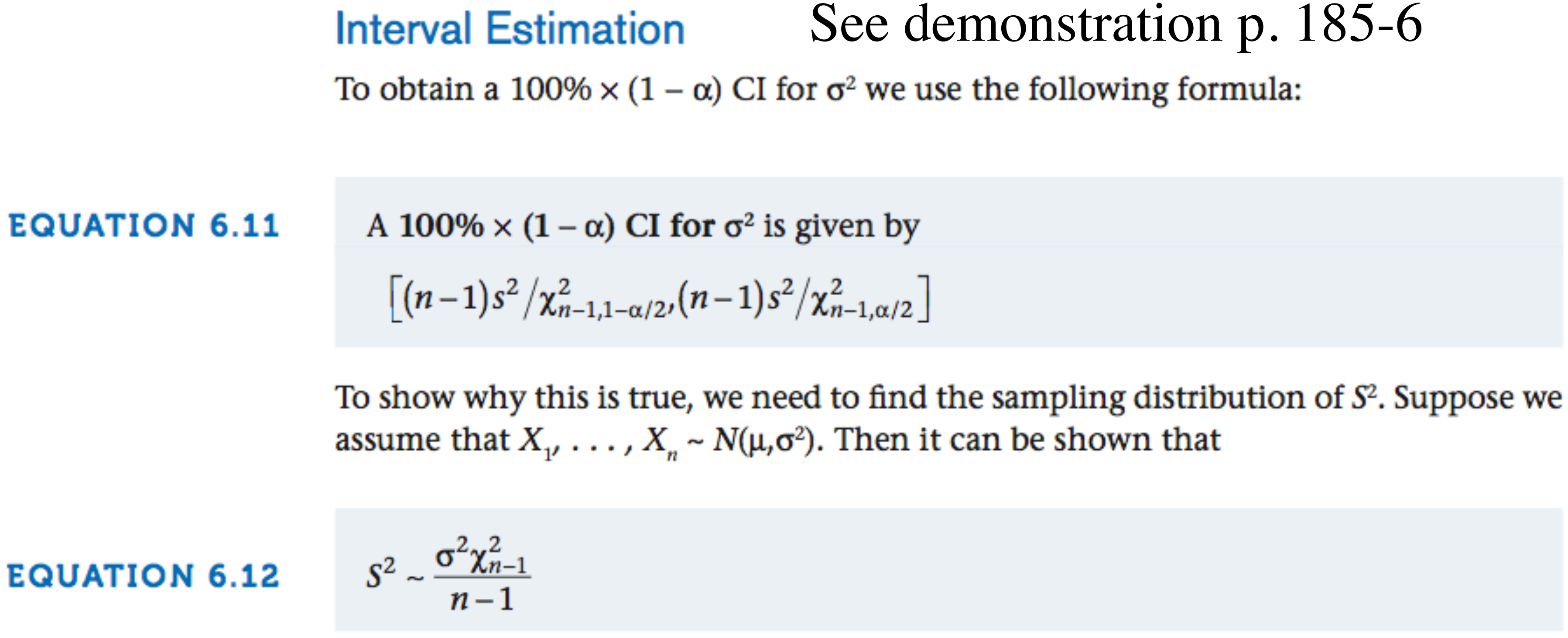
R

The upper 5th percentile = $\chi^2_{8,.95}$ = qchisq(0.95,8) = 15.51.

The lower 5th percentile = $\chi^2_{8,.05}$ = qchisq(0.05,8) = 2.73.

These are denoted by chisq_8_upper and chisq_8_lower in the R output below.

See demonstration p. 185-6



Interval Estimation

To obtain a 100% × $(1-\alpha)$ CI for σ^2 we use the following formula:

EQUATION 6.11

A 100% × $(1-\alpha)$ CI for σ^2 is given by

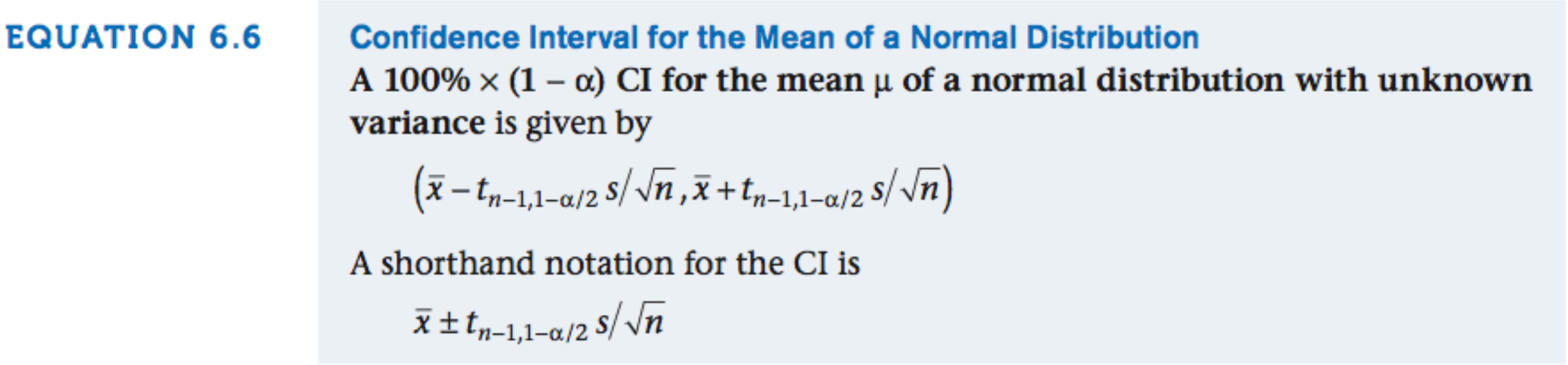
$$\left[(n-1)s^2/\chi^2_{n-1,1-\alpha/2},(n-1)s^2/\chi^2_{n-1,\alpha/2}\right]$$

To show why this is true, we need to find the sampling distribution of S^2. Suppose we assume that $X_1, \ldots, X_n \sim N(\mu,\sigma^2)$. Then it can be shown that

EQUATION 6.12

$$S^2 \sim \frac{\sigma^2\chi^2_{n-1}}{n-1}$$

REMINDER



EQUATION 6.6

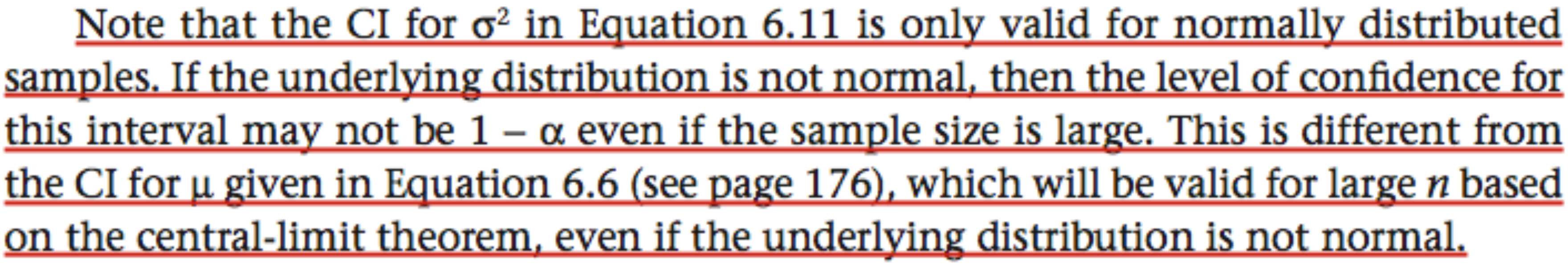
Confidence Interval for the Mean of a Normal Distribution

A 100% × (1 – α) CI for the mean μ of a normal distribution with unknown variance is given by

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

A shorthand notation for the CI is

$$\bar{x} \pm t_{n-1,1-\alpha/2}\, s/\sqrt{n}$$



Note that the CI for σ^2 in Equation 6.11 is only valid for normally distributed samples. If the underlying distribution is not normal, then the level of confidence for this interval may not be 1 – α even if the sample size is large. This is different from the CI for μ given in Equation 6.6 (see page 176), which will be valid for large *n* based on the central-limit theorem, even if the underlying distribution is not normal.

EXAMPLE 6.39 **Hypertension** An *Arteriosonde machine* "prints" blood-pressure readings on a tape so that the measurement can be read rather than heard. A major argument for using such a machine is that the variability of measurements obtained by different observers on the same person will be lower than with a standard blood-pressure cuff.

Suppose we have the data in Table 6.6, consisting of systolic blood pressure (SBP) measurements obtained on 10 people and read by two observers. We use the difference d_i between the first and second observers to assess interobserver variability. In particular, if we assume the underlying distribution of these differences is normal with mean μ and variance σ^2, then it is of primary interest to estimate σ^2. The higher σ^2 is, the higher the interobserver variability.

TABLE 6.6 **SBP measurements (mm Hg) from an Arteriosonde machine obtained from 10 people and read by two observers**

	Observer		
Person (i)	1	2	Difference (d_i)
1	194	200	−6
2	126	123	+3
3	130	128	+2
4	98	101	−3
5	136	135	+1
6	145	145	0
7	110	111	−1
8	108	107	+1
9	102	99	+3
10	126	128	−2

We have seen previously that an unbiased estimator of the variance σ^2 is given by the sample variance S^2. In this case,

$$\text{Mean difference} = (-6+3+\cdots-2)/10 = -0.2 = \bar{d}$$

$$\text{Sample variance} = s^2 = \sum_{i=1}^{n}(d_i-\bar{d})^2/9$$

$$= \left[(-6+0.2)^2+\cdots+(-2+0.2)^2\right]/9 = 8.178$$

How can an interval estimate for σ^2 be obtained?

EXAMPLE 6.42

Hypertension We now return to the specific data set in Example 6.39 (see page 182). Suppose we want to construct a 95% CI for the interobserver variability as defined by σ^2.

Solution: Because there are 10 people and $s^2 = 8.178$, the required interval is given by

$$\left(9s^2/\chi^2_{9,.975}, 9s^2/\chi^2_{9,.025}\right) = \left[9(8.178)/19.02, 9(8.178)/2.70\right] = (3.87, 27.26)$$

Similarly, a 95% CI for σ is given by $\left(\sqrt{3.87}, \sqrt{27.26}\right) = (1.97, 5.22)$. Notice that the CI for σ^2 is *not* symmetric about $s^2 = 8.178$, in contrast to the CI for μ, which *was* symmetric about $\bar{x}$. This characteristic is common in CIs for the variance.

We could use the CI for σ^2 to make decisions concerning the variability of the Arteriosonde machine if we had a good estimate of the interobserver variability of blood-pressure readings from a standard cuff. For example, suppose we know from previous work that if two people are listening to blood-pressure recordings from a standard cuff, then the interobserver variability as measured by the variance of the set of differences between the readings of two observers is 35. This value is outside the range of the 95% CI for σ^2(3.87, 27.26), and we thus conclude that the interobserver variability is reduced by using an Arteriosonde machine. Alternatively, if this prior variance were 15, then we could not say that the variances obtained from using the two methods are different.

REMARKS ON THE SAMPLE VARIANCE

- Let us assume $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ as an estimator for σ^2; we shall show that S_n^2 is biased

- $X_i \sim N(\mu, \sigma^2) \; \forall i$; $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$; $E[\bar{X}] = \mu$

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i^2 - 2\bar{X} X_i + X_i^2)$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + \bar{X}^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 \qquad \bar{X} \longrightarrow T_n$$

- REM if $E[S_n^2 - \sigma^2] = 0$ then S_n^2 is unbiased

$$E[S_n^2] = \frac{1}{n} \sum_{i} E[X_i^2] - E[T_n^2] = E[X^2] - E[T_n^2]$$

$$= var[X] + E[X]^2 - E[T_n^2]$$

$$- var[T_n] - E[T_n]^2$$

$$E[S_n^2] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2$$

$$= \sigma^2 \left(1 - \frac{1}{n}\right)$$

REMARKS ON THE SAMPLE VARIANCE

- Let us assume $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ as an estimator for σ^2; we shall show that S_n^2 is biased

- $X_i \sim N(\mu, \sigma^2)\ \forall i$; $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$; $E[\bar{X}] = \mu$

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + X_1^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{X}\sum_{i=1}^{n} x_i + \bar{X}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{X}^2 \qquad \bar{X} \longrightarrow T_n$$

- REM if $E[S_n^2 - \sigma^2] = 0$ then S_n^2 is unbiased

$$E[S_n^2] = \frac{1}{n}\sum_i E[X_i^2] - E[T_n^2] = E[X^2] - E[T_n^2]$$

$$= var[X] + E[X]^2 - E[T_n^2]$$

$$- var[T_n] - E[T_n]^2$$

$$E[S_n^2] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2$$

$$= \sigma^2\left(1 - \frac{1}{n}\right)$$

RESULT #0 IS, SYNTHETICALLY, THE MEAN SQUARE ERROR OF $\bar{X}$ is $\frac{\sigma^2}{n}$

LET US DEFINE THE MEAN SQUARE ERROR OF AN ESTIMATOR T_n

$$R_{T_n} \equiv E[(T_n-\theta)^2] = E[T_n^2 - 2\theta T_n + \theta^2]$$

$$= E[T_n^2] - 2\theta E[T_n] + \theta^2$$

$$= E[T_n^2] \underbrace{- 2\theta(B_{T_n}+\theta) + \theta^2}_{-2B_{T_n}\theta - \theta^2}$$

$$R_{T_n} = VAR[T_n] + B_{T_n}^2$$

(indeed, $VAR[T_n] \equiv E[T_n^2] - E[T_n]^2 =$

$$= E[T_n^2] - (B_{T_n}+\theta)^2$$

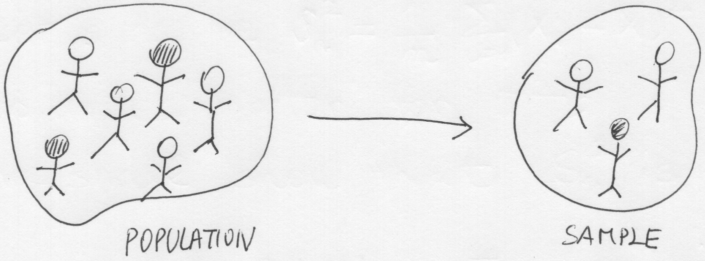
$$= E[T_n^2] - B_{T_n}^2 - 2B_{T_n}\theta - \theta^2$$

$$VAR[T_n] + B_{T_n}^2 = E[T_n^2] - 2B_{T_n}\theta - \theta^2$$

(note: $-2\theta B_{T_n} - 2\theta^2 + \theta^2 = -2B_{T_n}\theta - \theta^2$))

THEOREM: when the estimator is unbiased then its MEAN SQUARE ERROR IS equal to its VARIANCE

THE IDEA OF BOOTSTRAP



$\downarrow$

$$X = \{x_1, x_2, \dots, x_n\}$$

$\downarrow$ BOOTSTRAP SAMPLING (USE TABLES RND)

$$\left\{\begin{array}{l}\{Y_1^1, Y_2^1, \dots Y_n^1\} \equiv Y^1 \\ \quad\downarrow \text{BS} \\ \{Y_1^2, Y_2^2, \dots Y_n^2\} \equiv Y^2 \\ \quad\vdots \\ \{Y_1^B, Y_2^B, \dots, Y_n^B \quad B >> 1\end{array}\right.$$

BOOTSTRAP STATISTICS

sample mean

sample variance

tat cau be used to have point estimates of

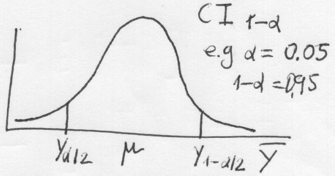
$E[f(\underline{Y})]$

$\sqrt{var[f(\underline{y})]} = SD$

$SE[f(\underline{y})]$

$CI[f(\underline{y})]$

e.g. $P[\bar{Y}]$

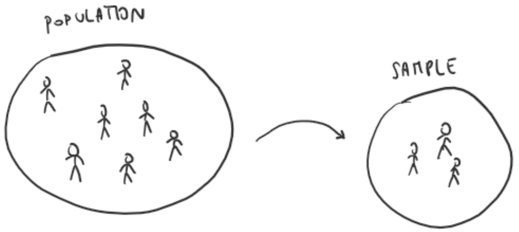


This is a MONTECARLO PROCEDURE !

INDEED IT IS BASED ON RANDOM NUMBER GENERATION (BOOTSTRAP SAMPLING)

Bootstrap sampling

https://datasciencechalktalk.com/2019/11/12/bootstrap-sampling-an-implementation-with-python/



$$\downarrow$$

$$\underline{X} = (x_1, x_2, \dots x_n)$$

Empirical CDF: $\hat{F}(x)$

$\downarrow$ drawing another sample of size n

$$\underline{X}^* = (x_1^*, x_2^*, \dots, x_n^*)$$

$\downarrow$ re-sampling with replacement B times

$$\underline{X}^{1*} = (x_1^{1*}, x_1^{1*}, x_5^{1*}, x_2^{1*} \dots) \longrightarrow T(\underline{X}^{1*})$$

$$\vdots$$

$$\underline{X}^{B*} = (x_n^{B*}, x_{n-1}^{B*}, x_6^{B*} \dots) \longrightarrow T(\underline{X}^{B*})$$

for each sample, compute the statistic

$$\longrightarrow \begin{bmatrix} T_1^* \\ T_2^* \\ \vdots \\ T_B^* \end{bmatrix}_{B\times 1} \longrightarrow$$

from here, we can work out Expectation, Variance, distribution of T^*

Bootstrap standard errors and confidence intervals

The **bootstrap** is a computer-intensive procedure used to approximate the sampling distribution of an estimate. Bootstrapping creates this sampling distribution by taking new samples randomly and repeatedly *from the data themselves*. Unlike simulation, the bootstrap is not directly intended for testing hypotheses. Instead, the bootstrap is used to find a standard error or confidence interval for a parameter estimate. The bootstrap is especially useful when no formula is available for the standard error or when the sampling distribution of the estimate of interest is unknown.

> ***Bootstrapping*** uses resampling from the data to approximate the sampling distribution of an estimate.

Recall from Section 4.1 that the sampling distribution is the probability distribution of sample estimates when a population is sampled repeatedly in the same way. The standard error is the standard deviation of this sampling distribution. In principle, therefore, we might obtain a standard error of an estimate by taking repeated samples from the population, calculating the sample estimate each time, and then taking the standard deviation of the many sample estimates. In reality, however, we can't do repeated sampling; collecting data is expensive, and it is best to put all individuals collected into one sample if we had more data. However, if the size of our sample from the population is large, then we do have easy access to a part of the popula-tion—namely, the part that was already sampled. Bootstrapping is a kind of repeated sampling, but instead of taking individuals from the population directly, we use a computer to draw the samples from the *data*, a procedure called "resampling." If the data set is large enough, then bootstrap samples drawn in this way will have statistical properties very similar to the distribution of possible sample estimates obtained from the population itself.

The bootstrap is therefore a bit strange: we resample from the data itself to generate many new data sets, and from these we infer the sampling distribution of the estimate. If you think about it, this is almost cheating, because we use the one and only data set to infer the distribution of estimates from all possible data sets. Hence the name "bootstrap," coming from the idea of picking yourself up by your own bootstraps.[2] The method was proposed by Bradley Efron in 1979, when desktop computers started to become available. The bootstrap is now commonly used in biology and other sciences.

Example 19.2 shows how to calculate a standard error and a confidence interval using the bootstrap. This particular example estimates a median, a simple quantity for which it is otherwise difficult to calculate a sampling distribution. Bootstrapping, however, can be applied to essentially any type of estimate.[3]

EXAMPLE 19.2 The language center in chimps' brains

One of the things that makes humans different from other organisms is our well-developed capacity for complex speech. Chimps and gorillas can learn some rudimentary language, but with a capacity far below that of humans. Speech production in humans is associated with a part of the brain called "Brodmann's area 44," which is part of Broca's area. In humans, this area is larger in the left hemisphere of the brain than in the right, and this asymmetry has been shown to be important for language development. With the advent of magnetic resonance imaging (MRI), it is possible to ask whether this area is asymmetric in other apes' brains as well. A sample of 20 chimpanzees were scanned with MRI, and the asymmetry of their Brodmann's area 44 was recorded (Cantalupo and Hopkins 2001). This asymmetry score is left measurement minus the right, divided by the average of the two sides. The raw data are listed in Table 19.2-1. The sample median asymmetry score was 0.14. We want to quantify the uncertainty of this estimate of the population median by calculating its standard error.

M. Watson / ardea.c/AGE Fotostock

TABLE 19.2-1 Asymmetry scores for Brodmann's area 44 in 20 chimpanzees.

Name of chimp	Asymmetry score
Austin	0.30
Carmichael	0.16
Chuck	−0.24
Dobbs	−0.25
Donald	0.36
Hoboh	0.17
Jimmy Carter	0.11
Lazarus	0.12
Merv	0.34
Storer	0.32
Ada	0.71
Anna	0.09
Atlanta	1.12
Cheri	−0.22
Jeannie	1.19
Kengee	0.01
Lana	−0.24
Lulu	0.24
Mary	−0.30
Panzee	−0.16

The frequency distribution of asymmetry scores shown in Figure 19.2-1 is skewed to the right and might even be bimodal. A transformation of these data would be difficult to find because the range of values includes negative numbers. What to do? Bootstrapping provides a suitable approach.

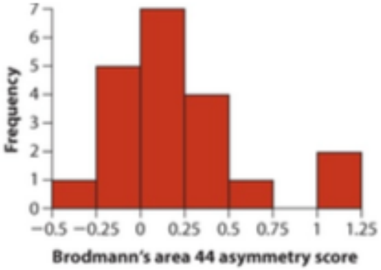


Figure 19.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 19.2-1 The frequency distribution of asymmetry scores for Brodmann's area 44 in 20 chimpanzees. A negative score indicates that the area is larger on the right side of the chimp's brain, while chimps with positive scores show a larger area in the left hemisphere.

because the range of values includes negative numbers. What to do? Bootstrapping provides a suitable approach.

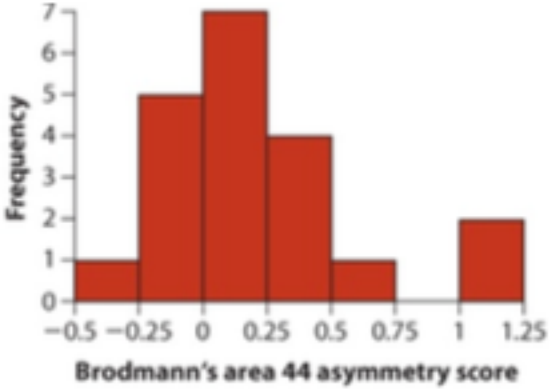
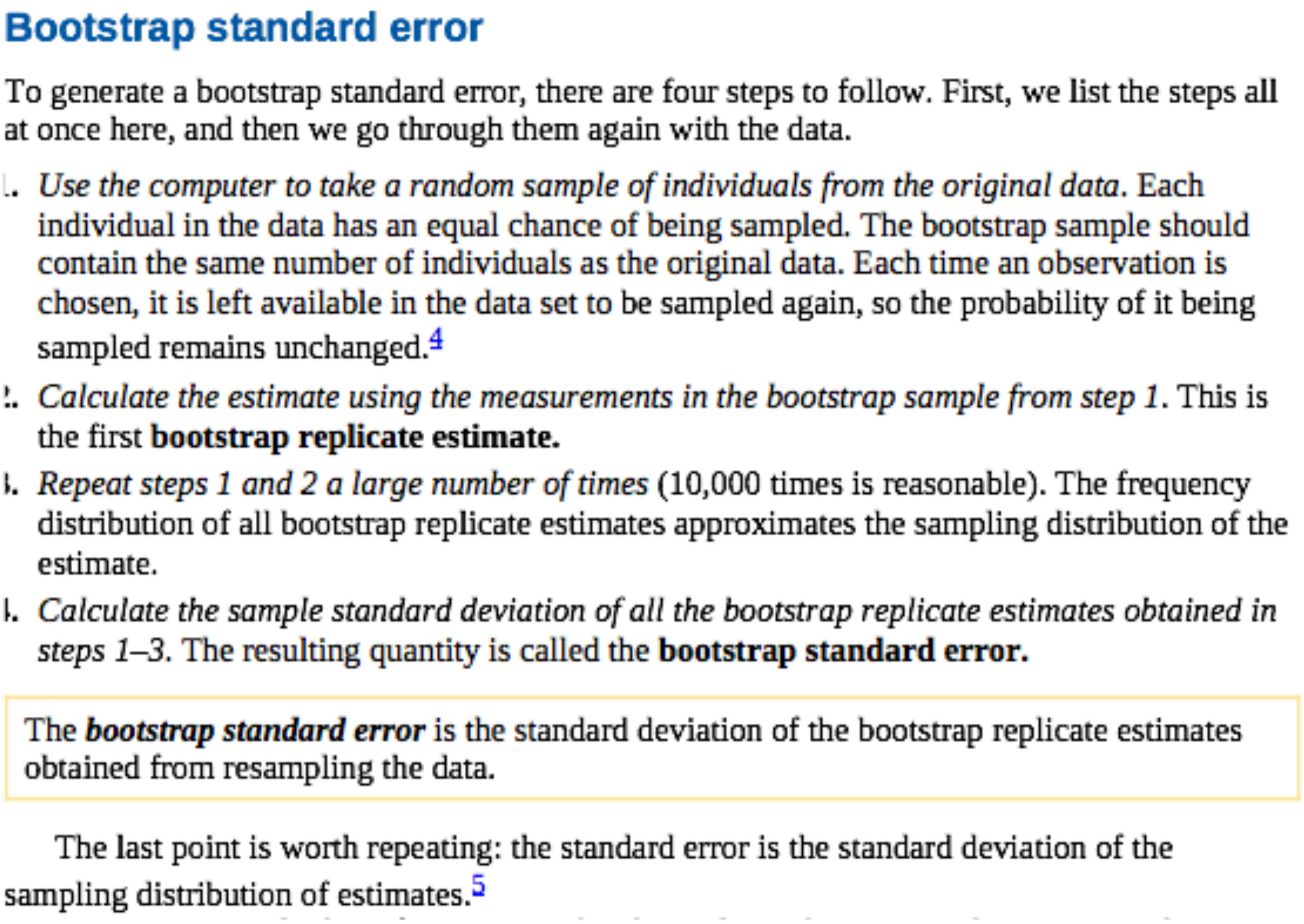


Figure 19.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e,
© 2015 W. H. Freeman and Company

FIGURE 19.2-1 The frequency distribution of asymmetry scores for Brodmann's area 44 in 20 chimpanzees. A negative score indicates that the area is larger on the right side of the chimp's brain, while chimps with positive scores show a larger area in the left hemisphere.



Bootstrap standard error

To generate a bootstrap standard error, there are four steps to follow. First, we list the steps all at once here, and then we go through them again with the data.

1. *Use the computer to take a random sample of individuals from the original data*. Each individual in the data has an equal chance of being sampled. The bootstrap sample should contain the same number of individuals as the original data. Each time an observation is chosen, it is left available in the data set to be sampled again, so the probability of it being sampled remains unchanged.[4]
2. *Calculate the estimate using the measurements in the bootstrap sample from step 1*. This is the first **bootstrap replicate estimate.**
3. *Repeat steps 1 and 2 a large number of times* (10,000 times is reasonable). The frequency distribution of all bootstrap replicate estimates approximates the sampling distribution of the estimate.
4. *Calculate the sample standard deviation of all the bootstrap replicate estimates obtained in steps 1–3*. The resulting quantity is called the **bootstrap standard error.**

> The ***bootstrap standard error*** is the standard deviation of the bootstrap replicate estimates obtained from resampling the data.

The last point is worth repeating: the standard error is the standard deviation of the sampling distribution of estimates.[5]

We can now apply these four steps to the chimp data. There are 20 data points in the sample, so each bootstrap sample must also have 20 measurements. Each of the 20 measurements in the bootstrap sample is chosen with equal probability from the values in the original data. Applying step 1, the following is the first bootstrap replicate that we obtained:

0.24	0.36	0.30	0.16	0.34	−0.24	0.30	1.19	0.32	0.32
0.36	0.01	0.01	0.11	0.11	−0.25	0.12	0.32	−0.24	0.17

Each of the measurements in this first bootstrap sample is present in the original data set. By chance, some of the original data points are present more than once in the bootstrap sample. For example, the score 0.32 (from the chimp named Storer) is present three times. Also by chance, some of the original data points are absent from this first bootstrap sample. For example, the score 0.71 (from the chimp named Ada) was not sampled. The sample median of this bootstrap sample is 0.205, so this is our first bootstrap replicate estimate of the median asymmetry score (step 2).

We repeated this process 10,000 times, calculating the sample median of the measurements each time (step 3). Figure 19.2-2 shows the frequency distribution of the bootstrap replicate estimates of the sample median.

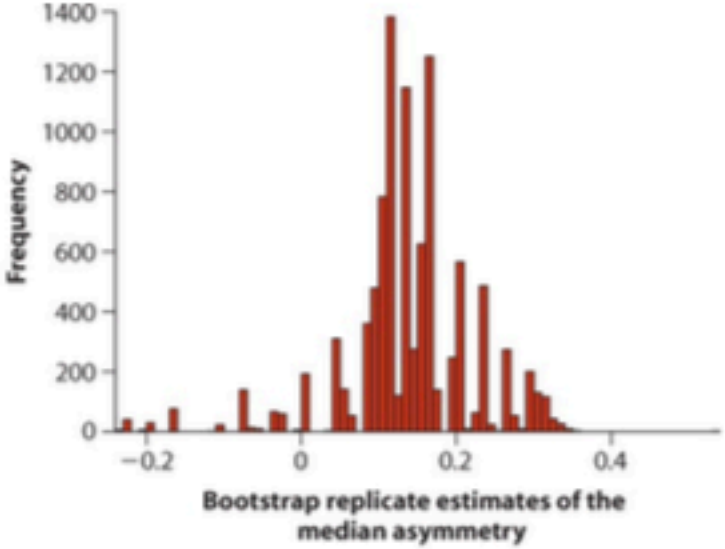


Figure 19.2-2

Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

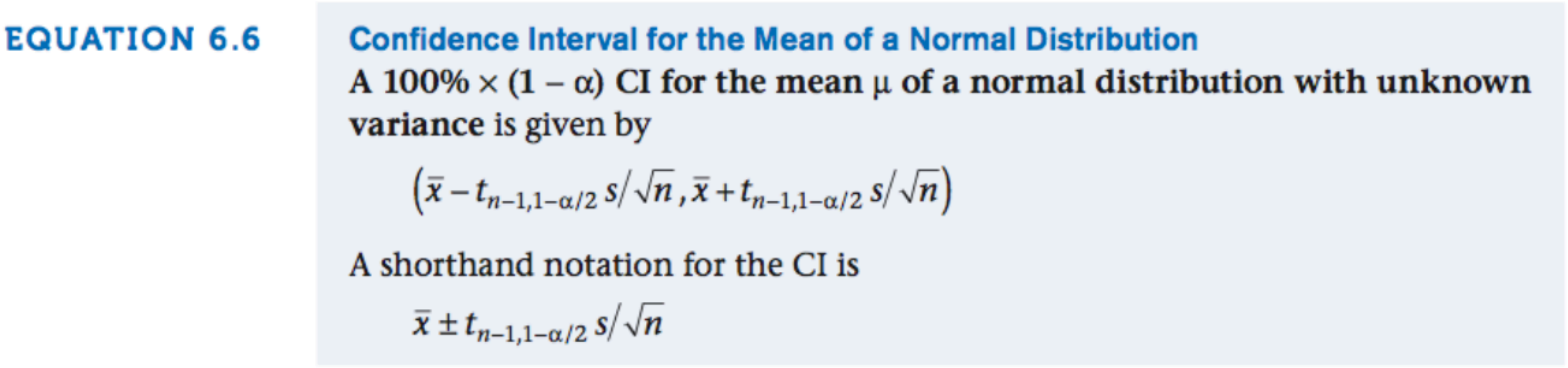
FIGURE 19.2-2 The distribution of 10,000 bootstrap replicate estimates for the median asymmetry of Brodmann's area 44 in chimpanzees.

The mean of the bootstrap replicate estimates is 0.142, which is very close to the estimated median from the original data (0.14). Remember that the bootstrap procedure is calculating a sampling distribution for an estimate, not a null distribution for a hypothesis test. As such, the overall mean of the bootstrap replicate estimates should be close to the estimate first calculated on the original data.[6]

The standard deviation of these bootstrap replicate estimates is 0.085 (step 4). This is the bootstrap standard error of our sample median: SE = 0.085.

Because the bootstrap samples come from the data, which generally do not represent the full population, the bootstrap standard error tends to be slightly smaller than the true standard error. This effect is negligible when the sample size is large.

Beyond normality



EQUATION 6.6

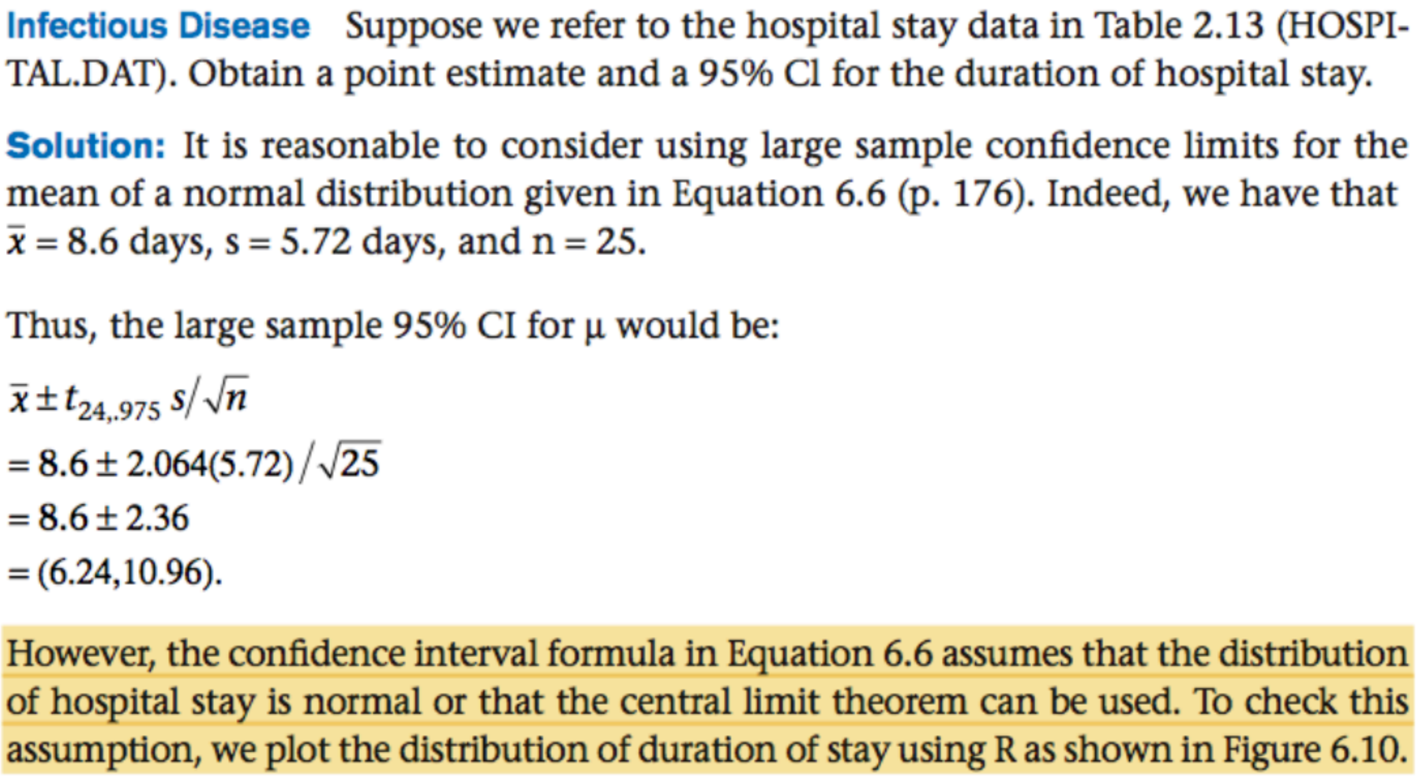
Confidence Interval for the Mean of a Normal Distribution

A 100% × (1 – α) CI for the mean μ of a normal distribution with unknown variance is given by

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

A shorthand notation for the CI is

$$\bar{x} \pm t_{n-1,1-\alpha/2}\, s/\sqrt{n}$$



Infectious Disease Suppose we refer to the hospital stay data in Table 2.13 (HOSPITAL.DAT). Obtain a point estimate and a 95% Cl for the duration of hospital stay.

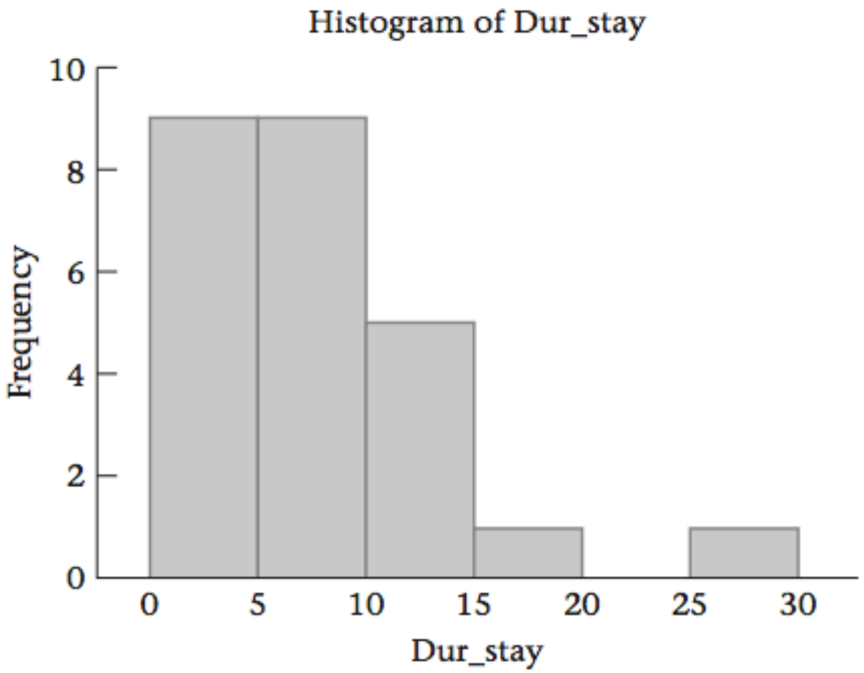
Solution: It is reasonable to consider using large sample confidence limits for the mean of a normal distribution given in Equation 6.6 (p. 176). Indeed, we have that $\bar{x} = 8.6$ days, s = 5.72 days, and n = 25.

Thus, the large sample 95% CI for μ would be:

$$\begin{aligned}&\bar{x} \pm t_{24,.975}\, s/\sqrt{n}\\&= 8.6 \pm 2.064(5.72)/\sqrt{25}\\&= 8.6 \pm 2.36\\&= (6.24,10.96).\end{aligned}$$

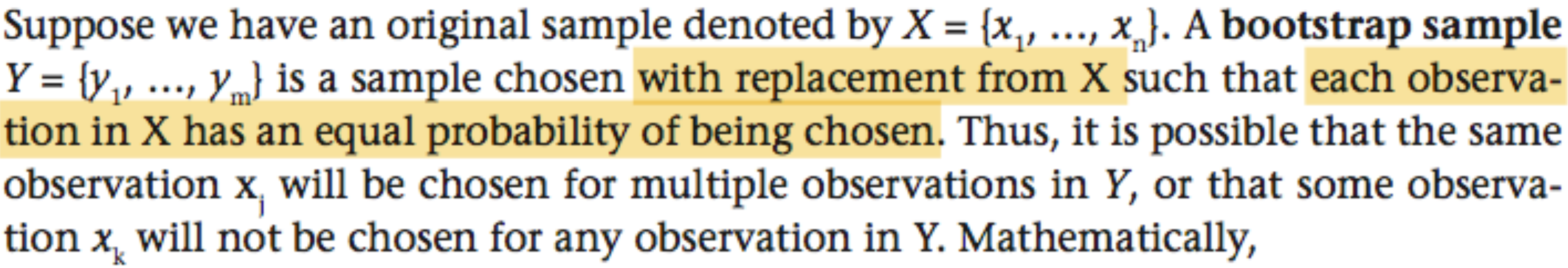
However, the confidence interval formula in Equation 6.6 assumes that the distribution of hospital stay is normal or that the central limit theorem can be used. To check this assumption, we plot the distribution of duration of stay using R as shown in Figure 6.10.

Plot of duration of stay in HOSPITAL.DAT

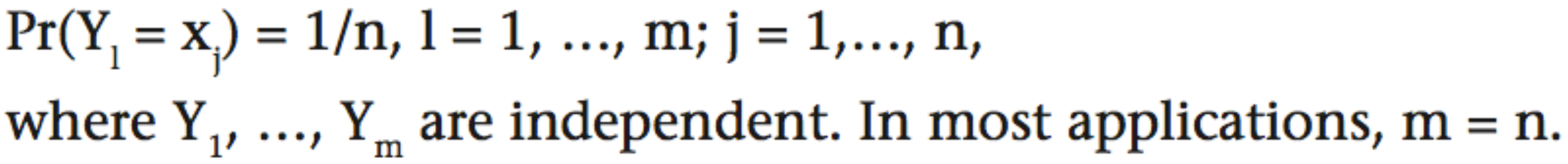


The distribution appears right-skewed and far from being normal. How can we check the validity of the 95% Cl computed in Example 6.63? A simulation-based approach, known as the Bootstrap approach, can be used for this purpose for estimating confidence intervals.

Bootstrap sample

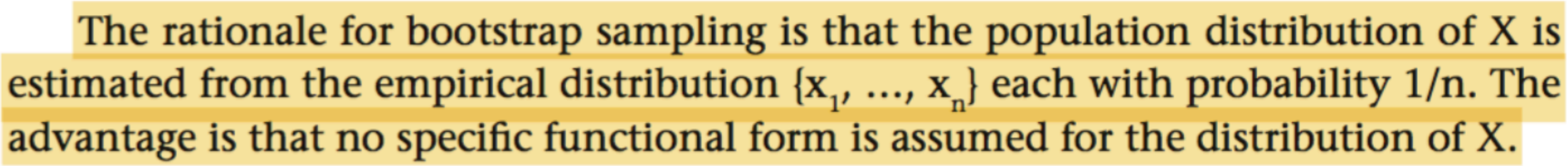


Suppose we have an original sample denoted by $X = \{x_1, \ldots, x_n\}$. A **bootstrap sample** $Y = \{y_1, \ldots, y_m\}$ is a sample chosen with replacement from X such that each observation in X has an equal probability of being chosen. Thus, it is possible that the same observation x_j will be chosen for multiple observations in Y, or that some observation x_k will not be chosen for any observation in Y. Mathematically,

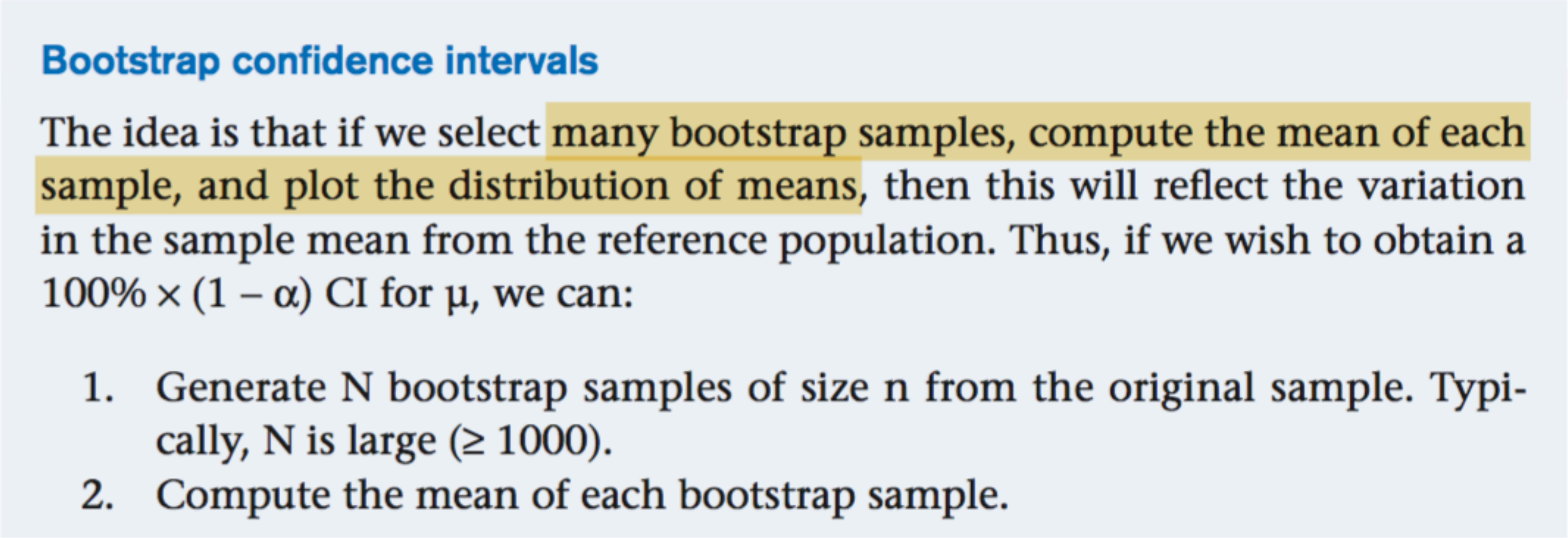


$$\Pr(Y_l = x_j) = 1/n,\ l = 1, \ldots, m;\ j = 1, \ldots, n,$$

where $Y_1, \ldots, Y_m$ are independent. In most applications, $m = n$.



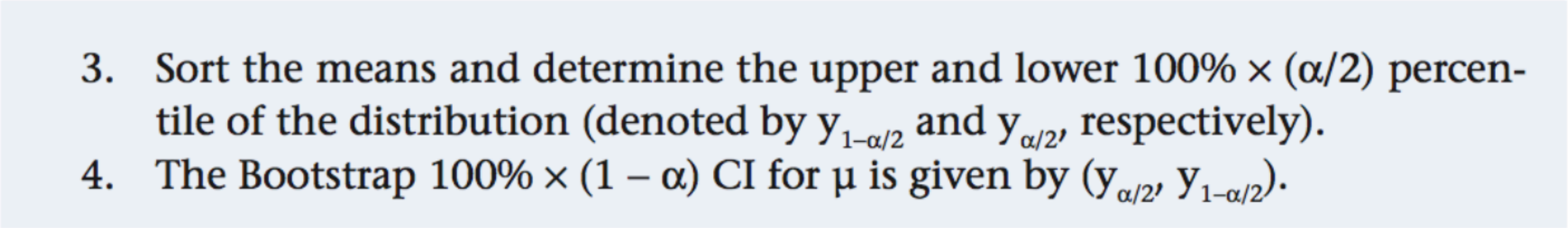
The rationale for bootstrap sampling is that the population distribution of X is estimated from the empirical distribution $\{x_1, \ldots, x_n\}$ each with probability 1/n. The advantage is that no specific functional form is assumed for the distribution of X.



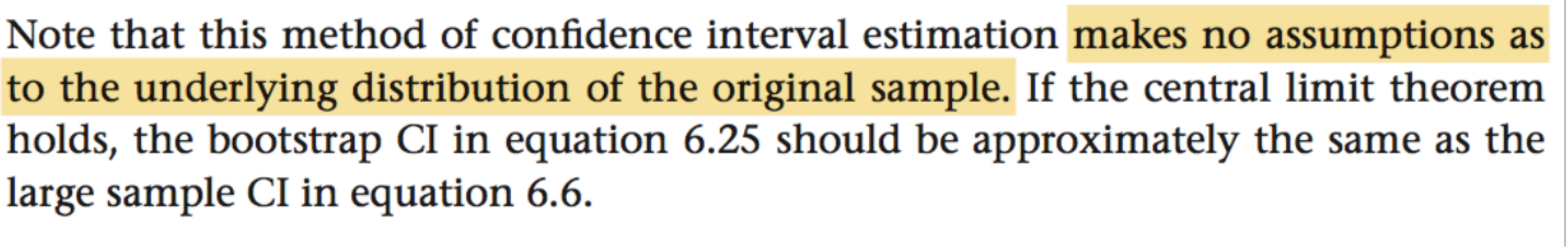
Bootstrap confidence intervals

The idea is that if we select many bootstrap samples, compute the mean of each sample, and plot the distribution of means, then this will reflect the variation in the sample mean from the reference population. Thus, if we wish to obtain a 100% × (1 – α) CI for μ, we can:

1. Generate N bootstrap samples of size n from the original sample. Typically, N is large (≥ 1000).
2. Compute the mean of each bootstrap sample.



3. Sort the means and determine the upper and lower 100% × (α/2) percentile of the distribution (denoted by $y_{1-\alpha/2}$ and $y_{\alpha/2}$, respectively).
4. The Bootstrap 100% × (1 – α) CI for μ is given by $(y_{\alpha/2}, y_{1-\alpha/2})$.



Note that this method of confidence interval estimation makes no assumptions as to the underlying distribution of the original sample. If the central limit theorem holds, the bootstrap CI in equation 6.25 should be approximately the same as the large sample CI in equation 6.6.

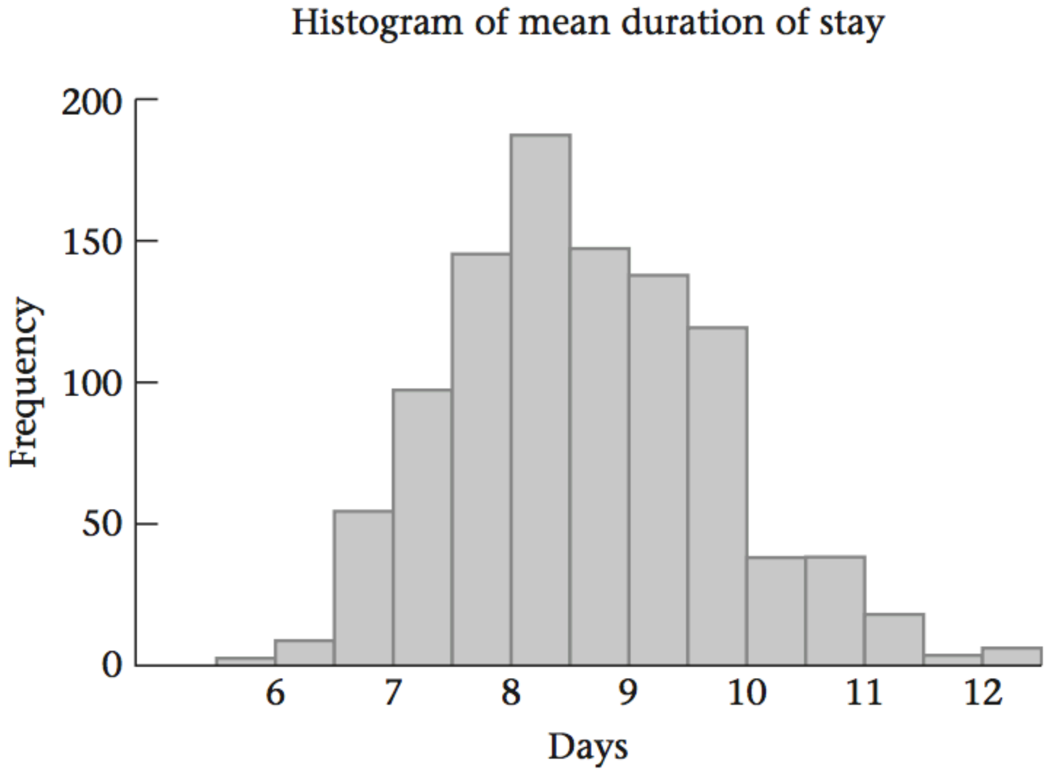
Infectious Disease Determine a 95% CI for the mean duration of stay in the Data Set HOSPITAL.DAT (Table 2.13) using bootstrap methods.

Solution: We use the sample command of R to select N = 1000 bootstrap samples and the mean command to calculate the mean of each of the samples. We then use the quantile command to determine the 2.5th and 97.5th percentiles of the 1000 sample means. The R code used for this purpose is given in Table 6.9.

The R code for obtaining 95% Bootstrap confidence limits for the mean duration of stay in HOSPITAL.DAT.

```
> a<– numeric(1000)
> for (i in 1:1000){
+ a[i]<–mean(sample(Dur_stay,25,replace=T))}
> quantile(a,c(.025,.975))
2.5% 97.5%
6.68 11.04
```

We see that the 95% CI for μ = (6.68, 11.04).
A histogram of the means of the 1000 bootstrap samples is given in Figure 6.11.



The distribution of sample means looks slightly positively skewed, which is consistent with the bootstrap 95% CI (6.68, 11.04) being asymmetric with respect to the mean in the original sample (8.6) and notably different from the large sample 95% CI (6.24, 10.96) given in the solution to Example 6.63.

Thus, the large sample 95% CI for μ based on n = 25 is probably not appropriate for this type of data and the bootstrap CI is preferable. The bootstrap method for obtaining CI can also be used to obtain confidence limits for other parameters. More details about bootstrap sampling is provided in Efron and Tibshirani [4].