ESTIMATION I (Find the best inferential estimates from finite samples)
(DA_2022)

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

Lecture n. 12, Rome 10th April 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

Outline L 12

- Estimators vs parameters
- POINT ESTIMATES VS INTERVAL ESTIMATES
- Random numbers and random samples
- Standard error of the mean
- Central limit theorem
- Interval estimation
- t-distribution (Student)
- Chi-square distribution
- Percentiles and confidence intervals

The study material for this lecture can be found in chap. 6 of Rosner's textbook and chap.10 and 11 of W&S

overwiew of Statistical Inference

The problem addressed in the rest of this text is that we have a data set and we want to **infer** the properties of the underlying distribution from this data set. This inference usually involves **inductive reasoning** rather than **deductive reasoning**; that is, in principle, a variety of different probability models must at least be explored to see which model best "fits" the data.

Statistical inference can be further subdivided into the two main areas of estimation and hypothesis testing. **Estimation** is concerned with estimating the values of specific population parameters; **hypothesis testing** is concerned with testing whether the value of a population parameter is equal to some specific value. Problems of estimation are covered in this chapter, and problems of hypothesis testing are discussed in Chapters 7 through 10.

From:Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015), chap. 6

6.2 THE RELATIONSHIP BETWEEN POPULATION AND SAMPLE

EXAMPLE 6.8

Obstetrics Suppose we want to characterize the distribution of birthweights of all liveborn infants born in the United States in 2013. Assume the underlying distribution of birthweight has an expected value (or mean) μ and variance σ^2. Ideally, we wish to estimate μ and σ^2 exactly, based on the entire population of U.S. liveborn infants in 2013. But this task is difficult with such a large group. Instead, we decide to select a random sample of n infants who are *representative* of this large group and use the birthweights $x_1, \ldots, x_n$ from this sample to help us estimate μ and σ^2. What is a random sample?

DEFINITION 6.1

A random sample is a selection of some members of the population such that each member is independently chosen and has a known nonzero probability of being selected.

DEFINITION 6.2

A simple random sample is a random sample in which each group member has the same probability of being selected.

6.3 RANDOM-NUMBER TABLES

In this section, practical methods for selecting random samples are discussed.

EXAMPLE 6.12 **Hypertension** Suppose we want to study how effective a hypertension treatment program is in controlling the blood pressure of its participants. We have a roster of all 1000 participants in the program, but because of limited resources only 20 can be surveyed. We would like the 20 people chosen to be a random sample from the population of all participants in the program. How should we select this random sample?

PROBLEM->

A computer-generated list of random numbers would probably be used to select this sample.

DEFINITION 6.4 A **random number** (or **random digit**) is a random variable X that takes on the values 0, 1, 2, . . . , 9 with equal probability. Thus,

$$Pr(X = 0) = Pr(X = 1) = \cdots = Pr(X = 9) = \frac{1}{10}$$

DEFINITION 6.5 Computer-generated **random numbers** are collections of digits that satisfy the following two properties:

(1) Each digit 0, 1, 2, . . . , 9 is equally likely to occur.

(2) The value of any particular digit is independent of the value of any other digit selected.

Table 4 in the Appendix lists 1000 random digits generated by a computer algorithm.

TABELLA 5.1. Numeri casuali (*random numbers*)

03474	37386	36964	73661	46986	37162	33261	68045
97742	46762	42811	45720	42533	23732	27073	60751
16766	22766	56502	67107	32907	97853	13553	85859
12568	59926	96966	82731	05037	29315	57121	01421
55595	63564	38548	24622	31624	30990	06184	43253
16227	79439	49544	35482	17379	32378	87352	09643
84421	75331	57245	50688	77047	44767	21763	35025
63016	37859	16955	56719	98105	07175	12867	35807
33211	23429	78645	60782	52420	74438	15510	01342
57608	63244	09472	79654	49174	60962	90528	47727
18180	79246	44171	65809	79838	61962	06765	00310
26623	89775	84160	74499	83114	63224	20148	58845
23424	06474	82977	77781	07453	21408	32989	40772
52362	81995	50922	61197	00567	63138	80220	25353
37859	43512	83395	00830	42340	79688	54420	68798
70291	71213	40332	03826	13895	10374	17763	71304
56621	83735	96835	08775	97122	59347	70332	40354
99495	72277	88429	54572	16643	61600	04431	86679
16081	50472	33271	43409	45593	46849	12720	73445
31169	33243	50278	98719	20153	70049	52856	66044
68343	01370	55743	07740	44227	88426	04334	60952
74572	56576	59299	76860	71913	86754	13581	82476
27423	78653	48559	06572	96576	93610	96469	24245
00396	82961	66373	22030	77845	70329	10456	50426
29949	89424	68496	91082	53759	19330	34252	05727
16908	26659	83626	41112	67190	07174	60472	12968
11279	47506	06091	97466	02943	73402	76709	03086
35241	01620	33325	12638	79784	50491	16925	35616
38231	68638	42389	70150	75876	68141	40017	49162
31962	59147	96443	34913	34868	25391	00524	34885
66674	06714	64057	19586	11056	50968	76832	03790
14908	44511	75738	80590	52274	11486	22981	22208
68055	11800	33960	27519	07606	29355	59338	24390
20467	87390	97514	01402	04023	33108	39541	64936
64195	89779	15061	59320	01901	07506	40787	88962

6.5 ESTIMATION OF THE MEAN OF A DISTRIBUTION

Now that we have discussed the meaning of a random sample from a population and have explored some practical methods for selecting such samples using computer-generated random numbers, let's move on to estimation. The question remains: How is a specific random sample $x_1, \ldots, x_n$ used to estimate μ and σ^2, the mean and variance of the underlying distribution? Estimating the mean is the focus of this section, while estimating the variance is covered in Section 6.7.

Point Estimation

A natural estimator to use for estimating the population mean μ is the sample mean

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

…making a sample of samples

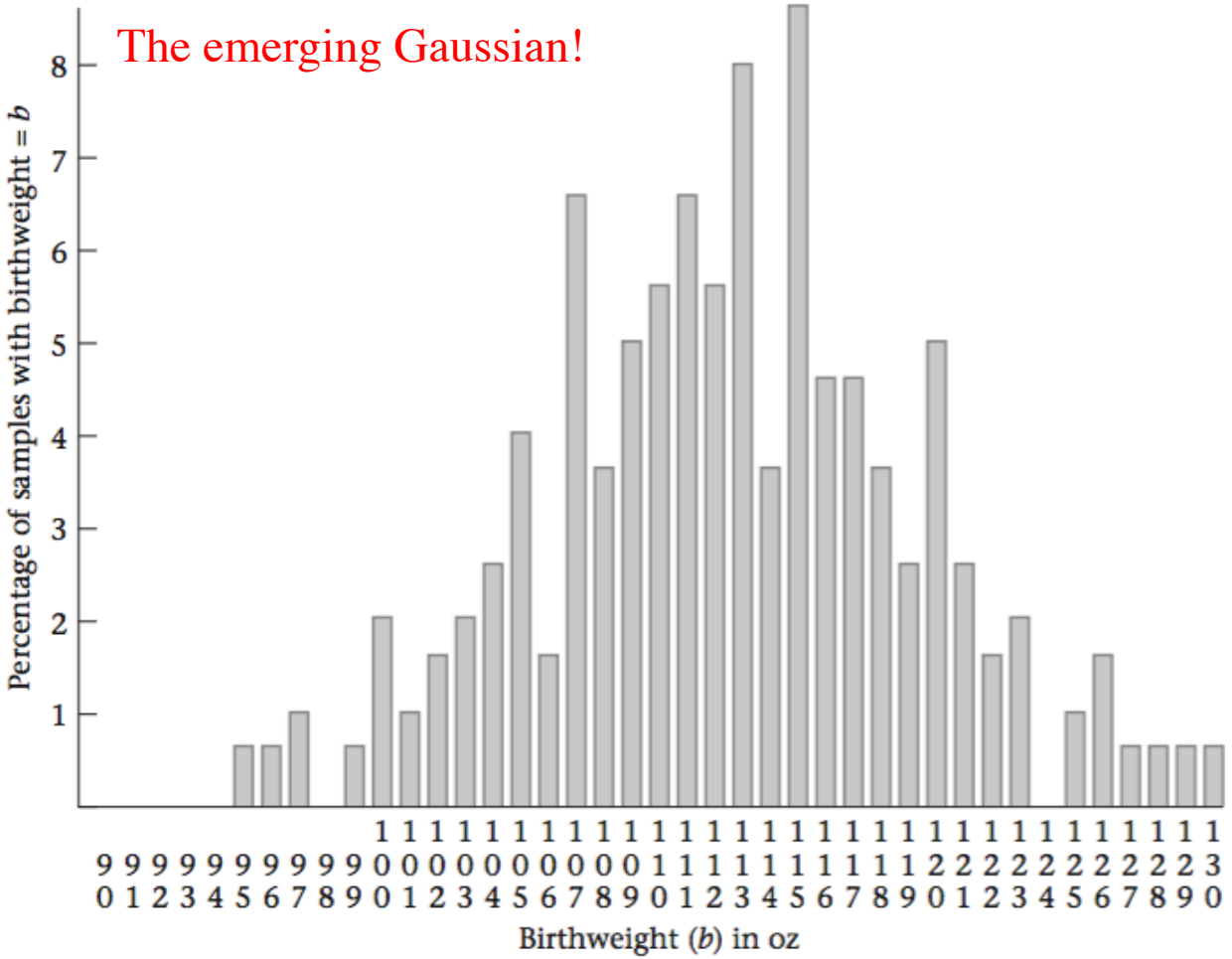
What properties of $\overline{X}$ make it a desirable estimator of μ? We must forget about our particular sample for the moment and consider the set of all possible samples of size n that could have been selected from the population. The values of $\overline{X}$ in each of these samples will, in general, be different. These values will be denoted by $\overline{x}_1$, $\overline{x}_2$, and so forth. In other words, we forget about our sample as a unique entity and consider it instead as representative of all possible samples of size n that could have been drawn from the population. Stated another way, $\overline{x}$ is a single realization of a random variable $\overline{X}$ over all possible samples of size n that could have been selected from the population. In the rest of this text, the symbol X denotes a random variable, and x denotes a specific realization of the random variable X in a sample.

6.10 The **sampling distribution** of $\overline{X}$ is the distribution of values of $\overline{x}$ over all possible samples of size n that could have been selected from the reference population.

Figure 6.1 gives an example of such a sampling distribution. This is a frequency distribution of the sample mean from 200 randomly selected samples of size 10 drawn from the distribution of 1000 birthweights given in Table 6.2, as displayed by the Statistical Analysis System (SAS) procedure PROC CHART.

We can show that the average of these sample means ($\overline{x}$'s), when taken over a large number of random samples of size n, approximates μ as the number of samples selected becomes large. In other words, the expected value of $\overline{X}$ over its sampling distribution is equal to μ. This result is summarized as follows:

FIGURE 6.1 Sampling distribution of $\overline{X}$ over 200 samples of size 10 selected from the population of 1000 birthweights given in Table 6.2 (100 = 100.0-100.9, etc.)



EQUATION 6.1

Let $X_1, \ldots, X_n$ be a random sample drawn from some population with mean μ. Then, for the sample mean $\overline{X}$, $E(\overline{X}) = \mu$.

Note that Equation 6.1 holds for any population regardless of its underlying distribution. In words, we refer to $\overline{X}$ as an unbiased estimator of μ.

Estimators

DEFINITION 6.11

We refer to an estimator of a parameter θ as $\hat{\theta}$. An **estimator** $\hat{\theta}$ of a parameter θ is **unbiased** if $E(\hat{\theta}) = \theta$. This means that the average value of $\hat{\theta}$ over a large number of random samples of size n is θ.

The unbiasedness of $\overline{X}$ is not sufficient reason to use it as an estimator of μ. For symmetric distributions, many unbiased estimators of μ exist, including the sample median and the average value of the largest and smallest data points in a sample. Why is $\overline{X}$ chosen rather than any of the other unbiased estimators? The reason is that if the underlying distribution of the population is normal, then it can be shown that the unbiased estimator with the smallest variance is given by $\overline{X}$. Thus, $\overline{X}$ is called the **minimum variance unbiased estimator** of μ.

Indeed, we would expect the sample means from repeated samples of size 100 to be less variable than those from samples of size 10. We can show this is true. Using the properties of linear combinations of independent random variables given in Equation 5.9,

See page 133 for **EQUATION 5.9**

$$Var(\overline{X}) = \left(\frac{1}{n^2}\right) Var\left(\sum_{i=1}^{n} X_i\right)$$
$$= \left(\frac{1}{n^2}\right) \sum_{i=1}^{n} Var(X_i)$$

However, by definition $Var(X_i) = \sigma^2$. Therefore,

$$Var(\overline{X}) = (1/n^2)(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = (1/n^2)(n\sigma^2) = \sigma^2/n$$

The standard deviation $(sd) = \sqrt{\text{variance}}$; thus, $sd(\overline{X}) = \sigma / \sqrt{n}$. We have the following summary:

EQUATION 6.2

Let $X_1, \ldots, X_n$ be a random sample from a population with underlying mean μ and variance σ^2. The set of sample means in repeated random samples of size n from this population has variance σ^2/n. The standard deviation of this set of sample means is thus $\sigma/\sqrt{n}$ and is referred to as the *standard error of the mean* or the *standard error.*

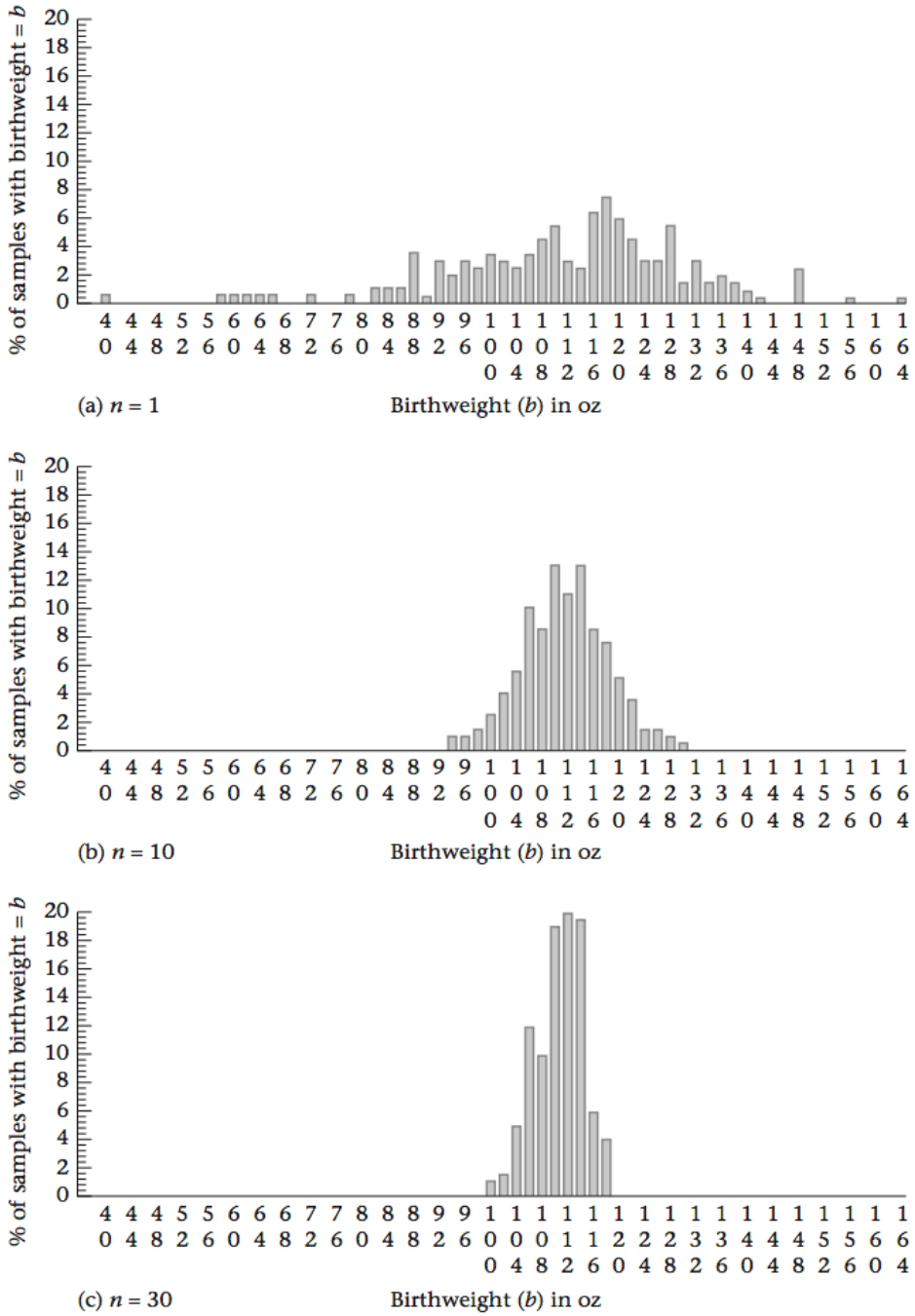
In practice, the population variance σ^2 is rarely known. We will see in Section 6.7 that a reasonable estimator for the population variance σ^2 is the sample variance s^2, which leads to the following definition:

DEFINITION 6.12

The **standard error of the mean (sem)**, or the **standard error (se)**, is given by $\sigma/\sqrt{n}$ and is estimated by $s/\sqrt{n}$. The standard error represents the estimated standard deviation obtained from a set of sample means from repeated samples of size n from a population with underlying variance σ^2.

Note that the standard error is *not* the standard deviation of an individual observation X_i but rather of the sample mean $\overline{X}$. The standard error of the mean is illustrated in Figure 6.3. In Figure 6.3a, the frequency distribution of the sample mean is plotted

FIGURE 6.3 Illustration of the standard error of the mean (100 = 100.0–103.9, etc.)



REVIEW QUESTIONS 6B

1. What is a sampling distribution?
2. Why is the sample mean $\bar{X}$ 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?

Central-Limit Theorem

If the underlying distribution is normal, then it can be shown that the sample mean is itself normally distributed with mean μ and variance σ^2/n (see Section 5.6). In other words, $\overline{X} \sim N\left(\mu, \sigma^2/n\right)$. If the underlying distribution is *not* normal, we would still like to make some statement about the sampling distribution of the sample mean. This statement is given by the following theorem:

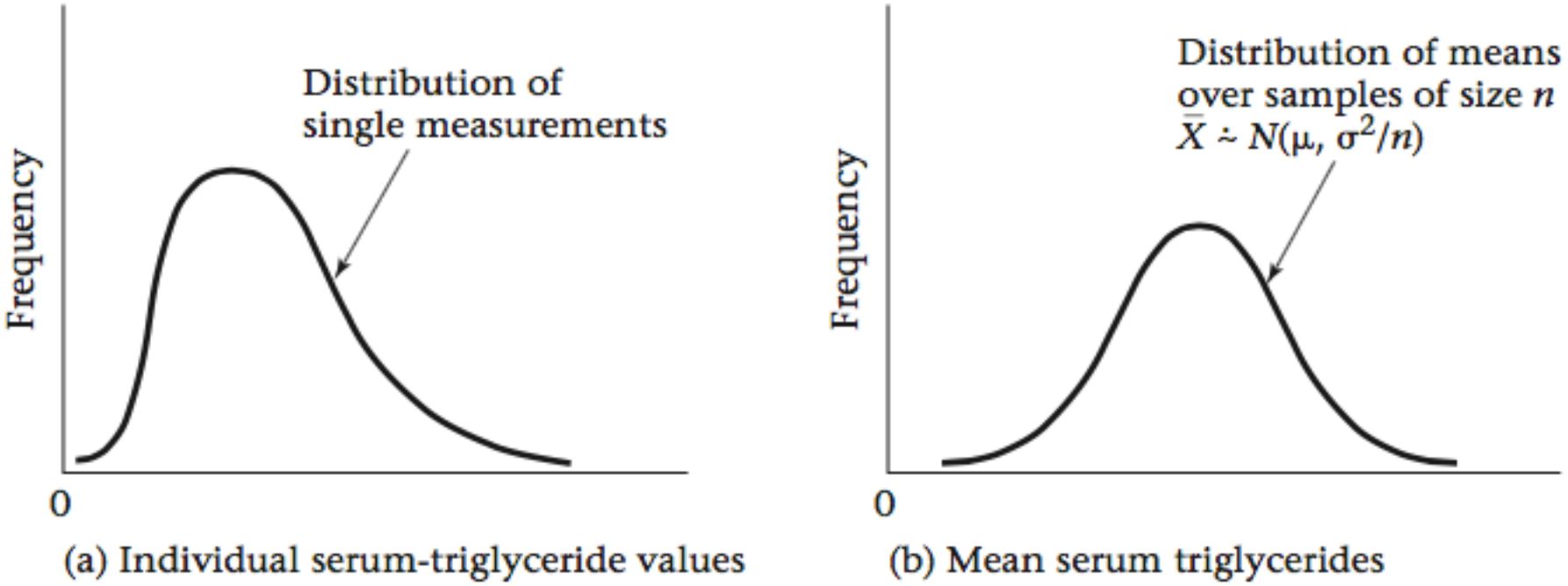
Central-Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2. Then, for large n, $\overline{X} \dot{\sim} N\left(\mu, \sigma^2/n\right)$ even if the underlying distribution of individual observations in the population is not normal. (The symbol $\dot{\sim}$ is used to represent "approximately distributed.")

Exercise check the central limit theorem

Why we do average measurements?

Distribution of single serum-triglyceride measurements and of means of such measurements over samples of size *n*



…Getting an universal distribution to be assumed as a standard

INTERVAL ESTIMATION

We have assumed previously that the distribution of birthweights in Table 6.2 was normal with mean μ and variance σ^2. It follows from our previous discussion of the properties of the sample mean that $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$. Thus, if μ and σ^2 were known,

then the behavior of the set of sample means over a large number of samples of size n would be precisely known. In particular, 95% of all such sample means will fall within the interval $\left(\mu - 1.96\,\sigma/\sqrt{n},\ \mu + 1.96\,\sigma/\sqrt{n}\right)$.

Alternatively, if we re-express $\bar{X}$ in standardized form by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

then Z should follow a standard normal distribution. Hence, 95% of the Z values from repeated samples of size n will fall between –1.96 and +1.96 because these values correspond to the 2.5th and 97.5th percentiles from a standard normal distribution. However, the assumption that σ is known is somewhat artificial, because σ is rarely known in practice.

t Distribution

Because σ is unknown, it is reasonable to estimate σ by the sample standard deviation *s* and to try to construct CIs using the quantity $(\bar{X}-\mu)/(S/\sqrt{n})$. The problem is that this quantity is no longer normally distributed.

This problem was first solved in 1908 by a statistician named William Gossett. For his entire professional life, Gossett worked for the Guinness Brewery in Ireland. He chose to identify himself by the pseudonym "Student," and thus the distribution of $(\bar{X}-\mu)/(S/\sqrt{n})$ is usually referred to as **Student's *t* distribution**. Gossett found that the shape of the distribution depends on the sample size *n*. Thus, the *t* distribution is not a unique distribution but is instead a family of distributions indexed by a parameter referred to as the **degrees of freedom** (*df*) of the distribution.

If $X_1, \ldots, X_n \sim N(\mu,\sigma^2)$ and are independent, then $(\bar{X}-\mu)/(S/\sqrt{n})$ is distributed as a *t* distribution with $(n-1)$ *df*.

Once again, Student's *t* distribution is not a unique distribution but is a family of distributions indexed by the degrees of freedom *d*. The *t* distribution with *d* degrees of freedom is sometimes referred to as the t_d distribution.

The **100 × *u*th percentile of a *t* distribution with *d* degrees of freedom** is denoted by $t_{d,u}$, that is,

$$Pr(t_d < t_{d,u}) \equiv u$$

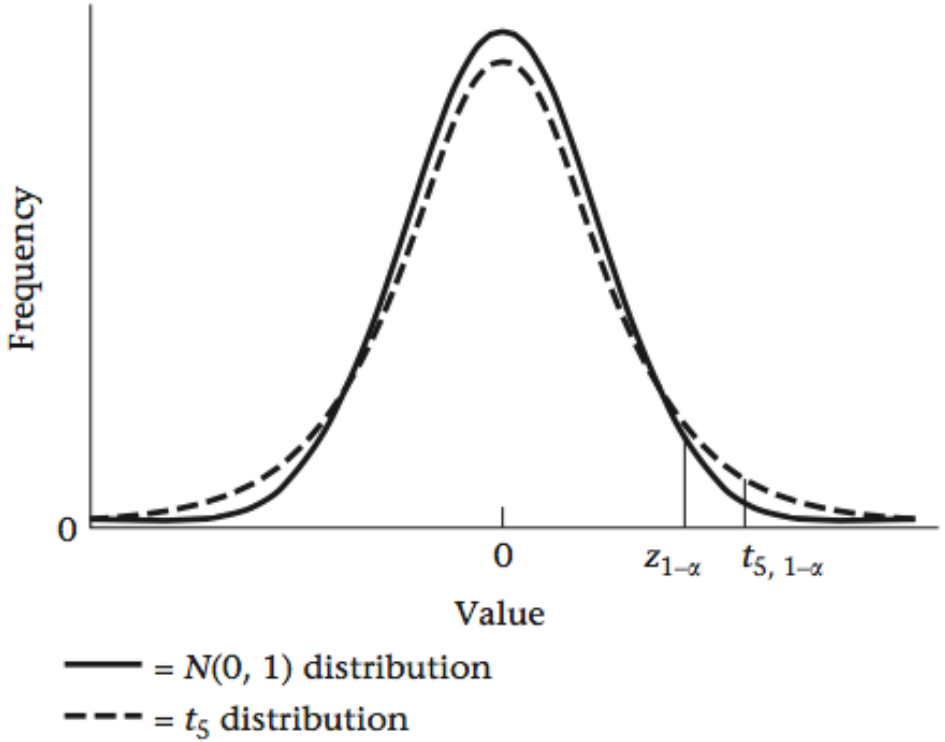
What does $t_{20,\,.95}$ mean?

Solution: $t_{20,\,.95}$ is the 95th percentile or the upper 5th percentile of a *t* distribution with 20 degrees of freedom.

It is interesting to compare a *t* distribution with *d* degrees of freedom with an $N(0, 1)$ distribution. The density functions corresponding to these distributions are depicted in Figure 6.6 for the special case where $d = 5$.

Notice that the *t* distribution is symmetric about 0 but is more spread out than the $N(0, 1)$ distribution. It can be shown that for any α, where $\alpha > .5$, $t_{d,1-\alpha}$ is always

FIGURE 6.6 Comparison of Student's t distribution with 5 degrees of freedom with an $N(0, 1)$ distribution



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].

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*

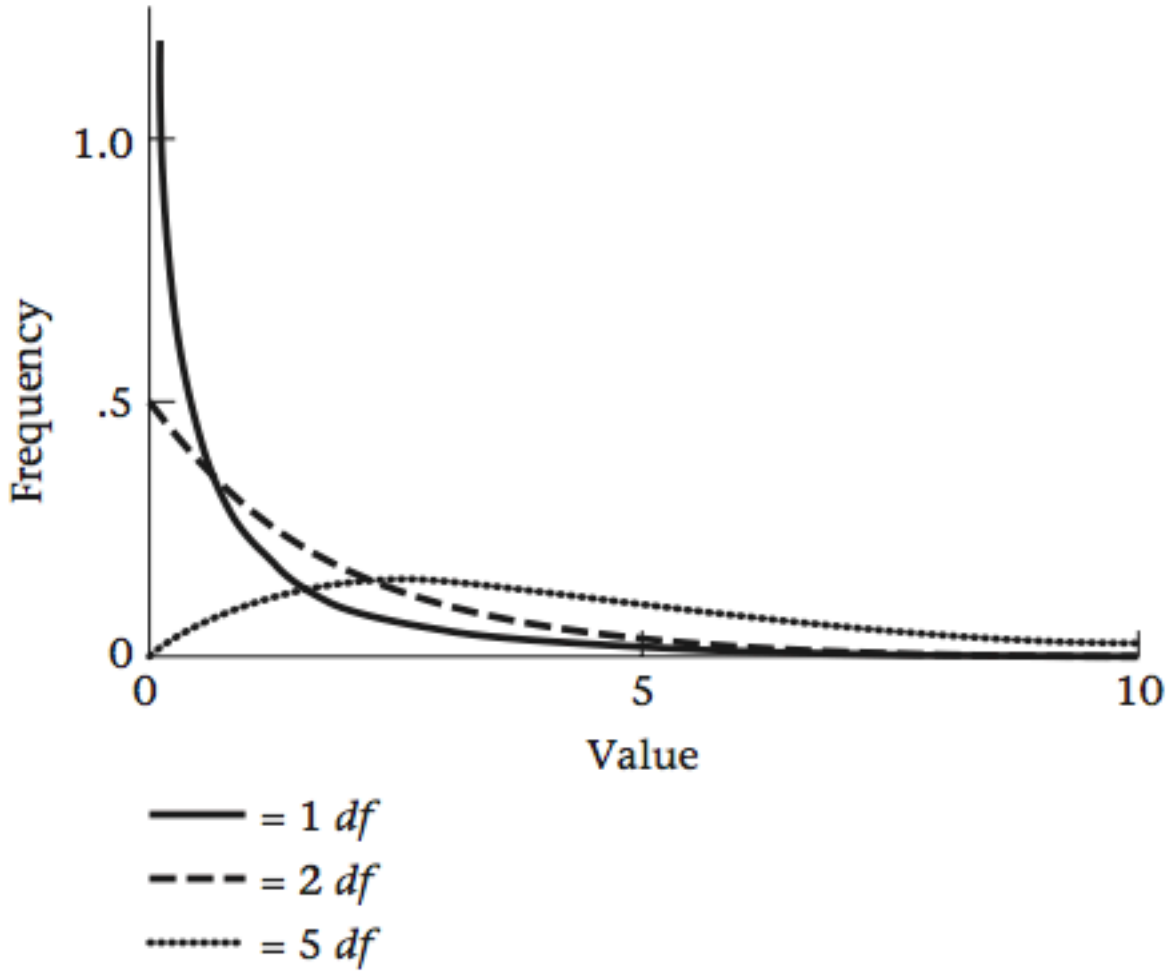
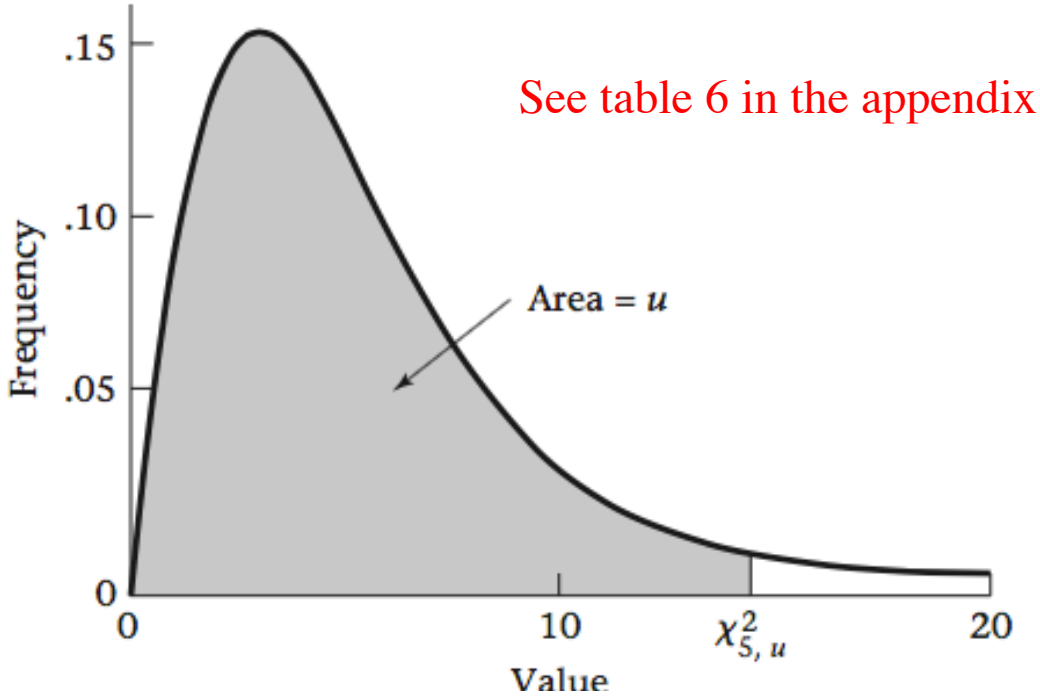


FIGURE 6.9 Graphic display of the percentiles of a χ^2_5 distribution



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* ($u \leq 0.5$) 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.