

Medical Statistics and Data Analysis

Lecture 7: Estimation

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)

Assistant Professor of Biomedical Engineering

ShanghaiTech University

October 19, 2026

Learning Goals of This Lecture

- Randomized experimental design and sampling
- Point estimations
- t Distribution
- Central-limit theorem
- Confidence intervals
- Chi-Square Distribution

Introduction

In this chapter, we will discuss

- How to **infer** the properties of the underlying distribution in a data set. This inference is usually based on **inductive reasoning** rather than **deductive reasoning**, that is, determining a best “fits” model among different probability models.

Two types of statistical inferences:

- **Estimation**: concerned with estimating the values of specific population parameters. These specific values are referred to as **point estimates**. Sometimes, **interval estimation** is carried out to specify a range within which the parameter values are likely to fall.
- **Hypothesis testing**: concerned with testing whether the value of a population parameter is equal to some specific value.

Relationship Between Population and Sample

- **Random sample:** a selection of some members of the population such that each member is independently chosen and has a known nonzero probability of being selected.
- **Simple random sample:** each group member has the same probability of being selected.
- **Reference, target or study population:** the group we want to study. The random sample is selected from the study population.
- Random sampling is not the only one used in practice. A popular alternative is **cluster sampling**.

Randomized Clinical Trials

Randomized clinical trial (RCT) is now accepted as the optimal study design in clinical research.

- A randomized clinical trial is a type of research design used for comparing different treatments, in which patients are assigned to a particular treatment by some random mechanism.
- The process of assigning treatments to patients is called **randomization**.
- Randomization means the types of patients assigned to different treatment modalities will be similar if the sample sizes are large. However, if sample sizes are small, then patient characteristics of treatment groups may not be comparable.
- It is customary to present a table of characteristics of different treatment groups in RCTs to check that the randomization process is working well.

TABLE 6.4 Baseline characteristics of randomized SHEP^a participants by treatment group

Characteristic	Active-treatment group	Placebo group	Total
Number randomized	2365	2371	4736
Age, y			
Average ^b	71.6 (6.7)	71.5 (6.7)	71.6 (6.7)
Percentage			
60–69	41.1	41.8	41.5
70–79	44.9	44.7	44.8
≥80	14.0	13.4	13.7
Race–sex, % ^c			
Black men	4.9	4.3	4.6
Black women	8.9	9.7	9.3
White men	38.8	38.4	38.6
White women	47.4	47.7	47.5
Education, y ^b	11.7 (3.5)	11.7 (3.4)	11.7 (3.5)
Blood pressure, mm Hg ^b			
Systolic	170.5 (9.5)	170.1 (9.2)	170.3 (9.4)
Diastolic	76.7 (9.7)	76.4 (9.8)	76.6 (9.7)
Antihypertensive medication at initial contact, %	33.0	33.5	33.3
Smoking, %			
Current smokers	12.6	12.9	12.7
Past smokers	36.6	37.6	37.1
Never smokers	50.8	49.6	50.2
Alcohol use, %			
Never	21.5	21.7	21.6
Formerly	9.6	10.4	10.0
Occasionally	55.2	53.9	54.5
Daily or nearly daily	13.7	14.0	13.8

Design Features of Randomized Clinical Trials

- Methods of randomization: random selection, random assignment, etc. In clinical trials, random assignment is sometimes called **block randomization**.
- A **block** size of $2n$ is predetermined, where for every $2n$ patients entering the study, n patients are randomly assigned to treatment A and the remaining n patients are assigned to treatment B .
- For more than two treatment groups: If there are k treatment groups, then the block size might be kn , where for every kn patients, n patients are randomly assigned to the first treatment, second treatment, and so on up to the k th treatment.

Stratification

- Another technique used in the randomization process.
- In some clinical studies, patients are subdivided into **subgroups**, or **strata**, according to characteristics thought important for patient outcome.
- Separate randomization lists are maintained for each stratum to ensure comparable patient populations within each stratum. This is called **stratification**.
- Each random selection (ordinary randomization) or random assignment (block randomization) might be used for each stratum.
- Typical characteristics used to define strata are **age**, **sex**, or **overall clinical condition** of the patient.

Blinding

- The use of blinding is an important advance in modern clinical research.
- A clinical trial is called **double blind** if neither the physician nor the patient knows what treatment he or she is getting. A clinical trial is called **single blind** if the patient is blinded as to treatment assignment but the physician is not.
- A clinical trial is **unblinded** if both the physician and patient are aware of the treatment assignment.
- The current gold standard of clinical research is the **randomized double-blind study**, in which patients are assigned to treatments at random and neither the patient nor the physician is aware of the treatment assignment.
- This prevents biased reporting of outcome. However, it may not always be feasible in research settings. In some cases, as treatment progresses the side effects may strongly indicate actual treatment received.

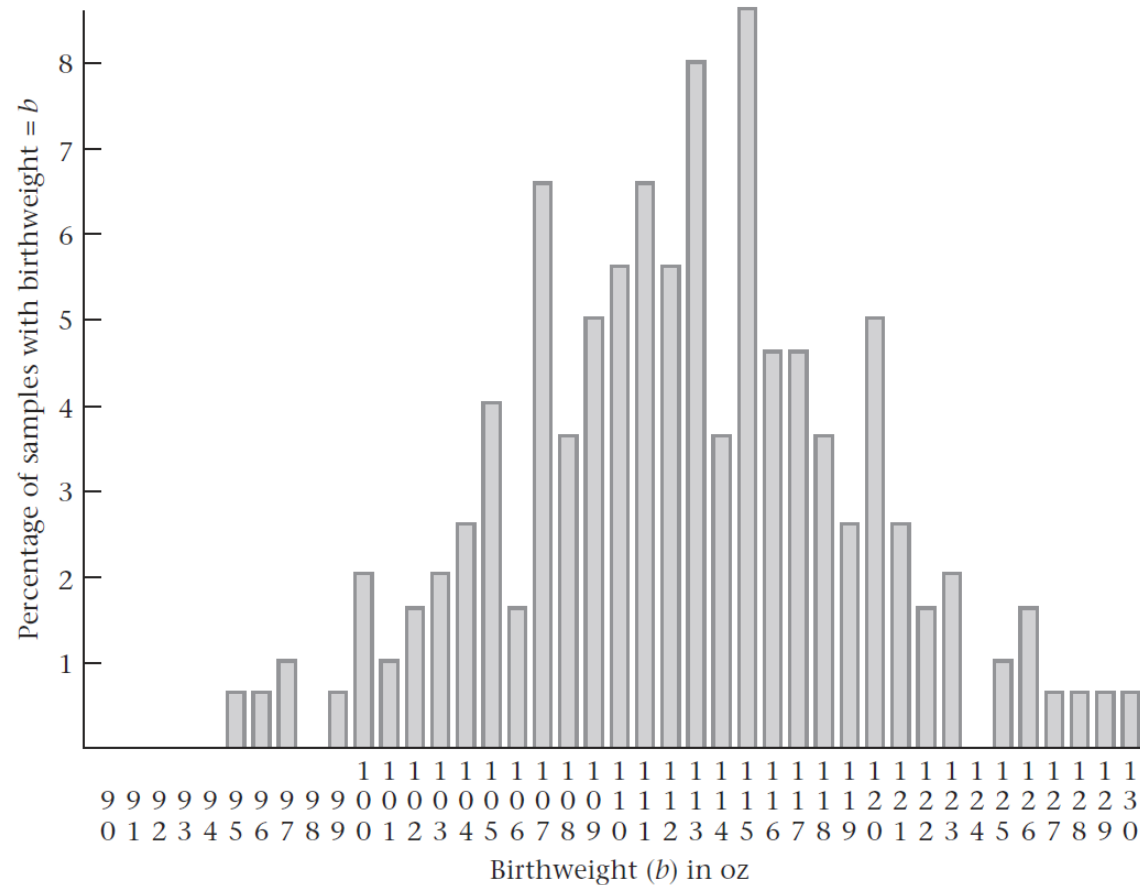
Estimation of the Mean of a Distribution

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

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

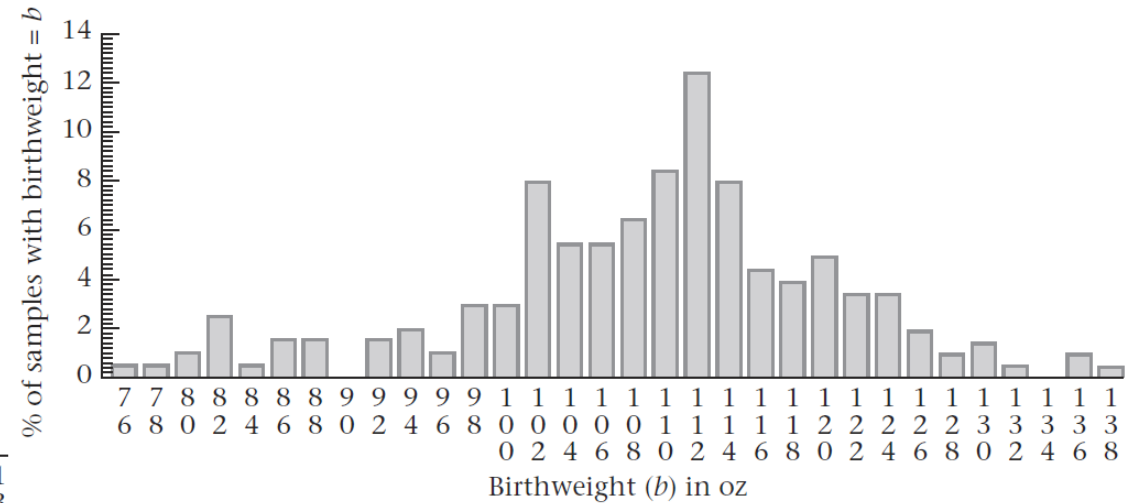
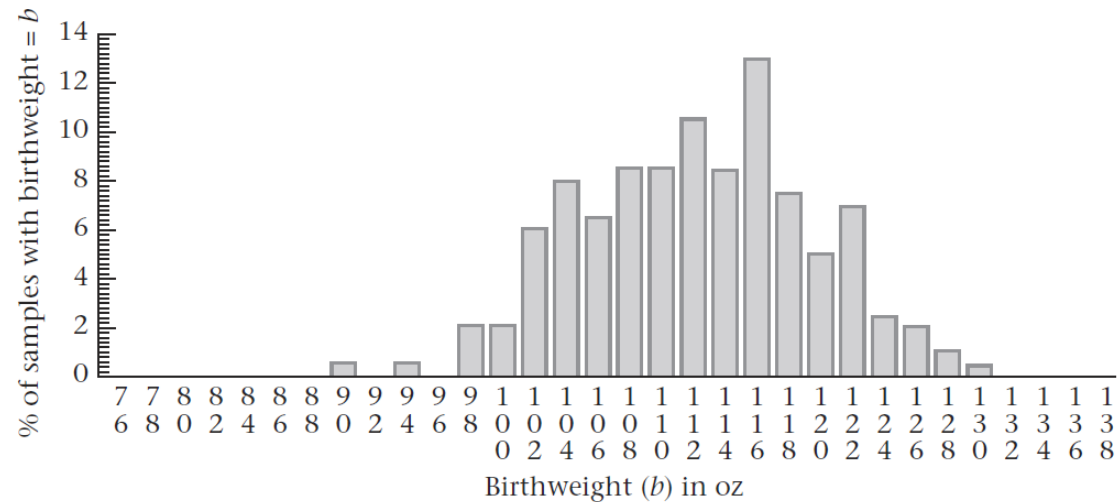
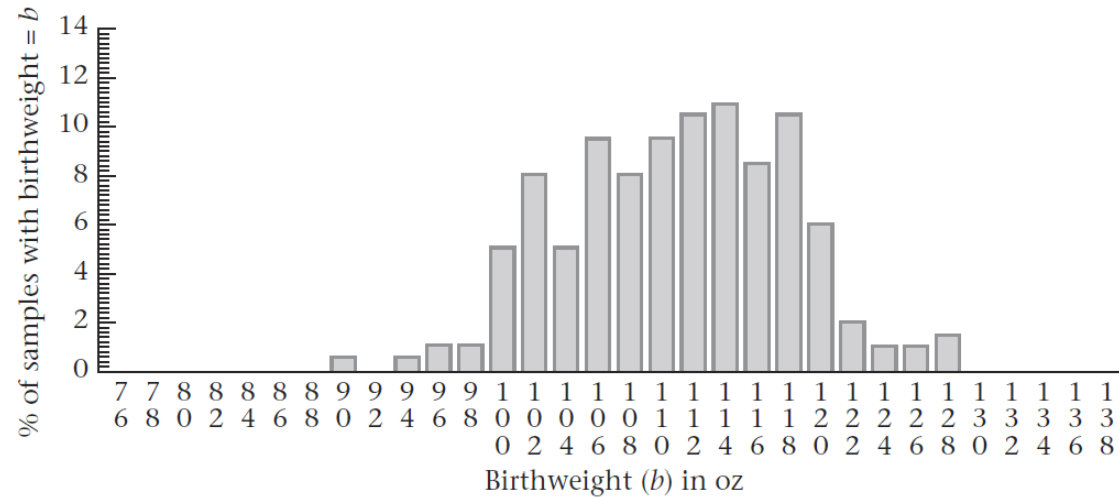
- $\bar{x}$ is a single realization of a random variable $\bar{X}$ over all possible samples of size n that could have been selected from the population.
- $\bar{X}$ denotes a random variable, and $\bar{x}$ denotes a specific realization of the random variable $\bar{X}$ in a sample.
- The sampling distribution of $\bar{X}$ is the distribution of values of $\bar{x}$ over all possible samples of size n that could have been selected from the reference population.

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



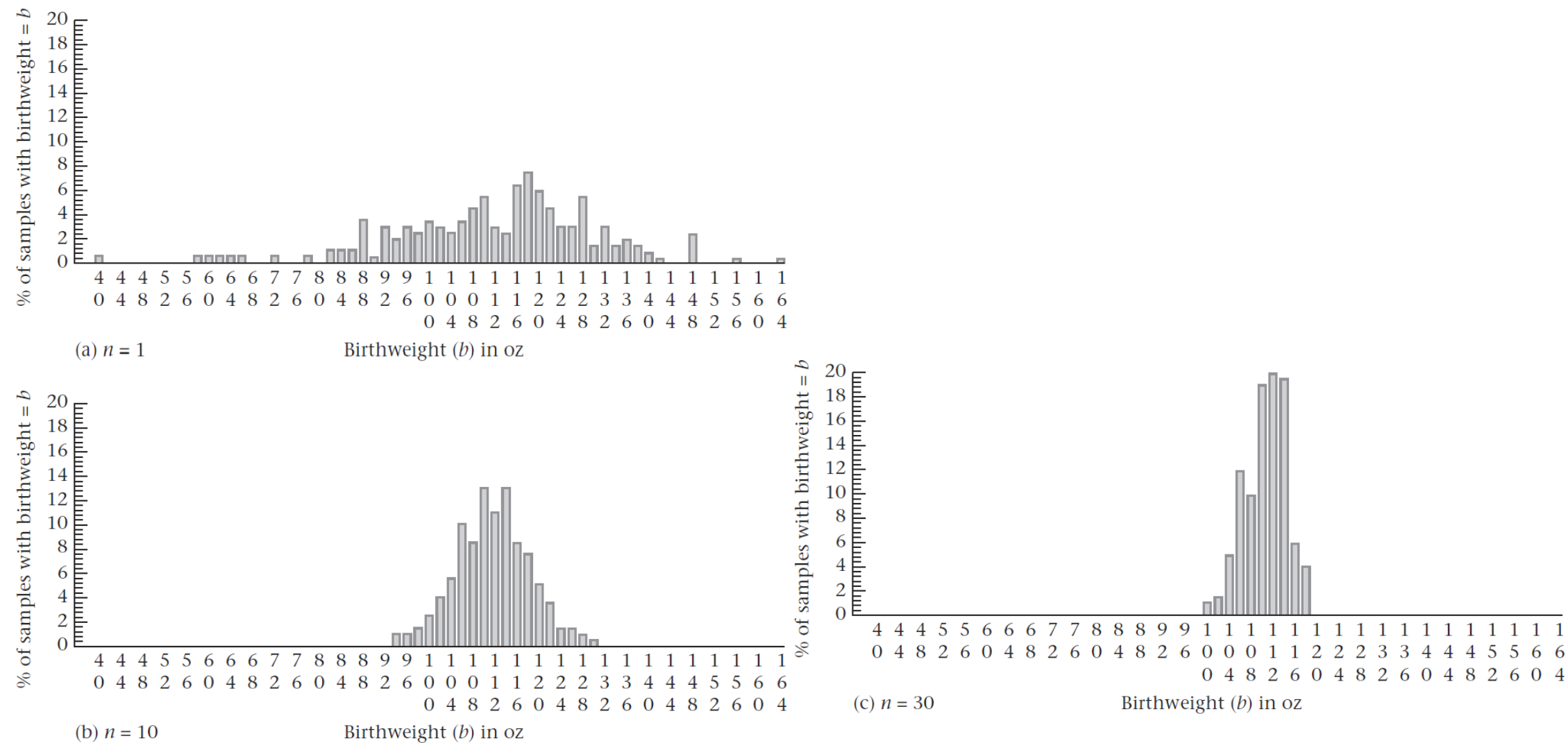
- Let $X_1, \dots, X_n$ be a random sample drawn from some population with mean μ .
- Then for the sample mean $\bar{X}$, $E(\bar{X}) = \mu$, regardless of its underlying distribution.
- $\bar{X}$ is regarded as an **unbiased estimator** of μ .
- An estimator of a parameter θ is referred to as $\hat{\theta}$.
- 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 repeated samples of size n is θ .
- 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.
- $\bar{X}$ is usually chosen over other unbiased estimators because if the underlying distribution is normal, then it can be shown that the unbiased estimator with smallest variance is given by $\bar{X}$. Thus, $\bar{X}$ is called the **minimum variance unbiased estimator** of μ .

FIGURE 6.2 Sampling distributions of several estimators of μ for 200 random samples of size 10 selected from the population of 1000 birthweights given in Table 6.2 (100 = 100.0–101.9, etc.)



- Let $X_1, \dots, X_n$ 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**.
- A reasonable estimator for the population variance σ^2 is the sample variance s^2 .
- 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 .

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



- The standard error is not the standard deviation of an individual observation X_i , rather of the sample mean $\bar{X}$.
- It is a quantitative measure of the variability of sample means obtained from repeated random samples of size n drawn from the sample population.
- It is directly proportional to both $1/\sqrt{n}$ and to the population standard deviation σ of individual observations.
- Precision of the estimate of μ is affected by the underlying variance σ^2 from the population of individual observations and the sample size.
- σ^2 can sometimes be affected by the experimental technique.

Central Limit Theorem

- Let $X_1, \dots, X_n$ be a random sample from some population with mean μ and variance σ^2 . Then for large n ,
$$\bar{X} \overset{\text{approx.}}{\sim} N(\mu, \sigma^2/n)$$
- even if the underlying distribution of individual observations in the population is not normal.
- The symbol $\overset{\text{approx.}}{\sim}$ is used to represent “approximately distributed.”
- This theorem allows us to perform statistical inference based on the approximate normality of the sample mean despite the nonnormality of the distribution of individual observations.
- The skewness of the distribution can be reduced by transformation of data using log scale. The central-limit theorem can then be applicable for smaller sizes than if the data are retained in the original scale.

FIGURE 6.4 Illustration of the central-limit theorem: $100 = 100-103.9$ in Figure 6.4(a): $= 100-101.9$ in Figures 6.4(b) and (c).

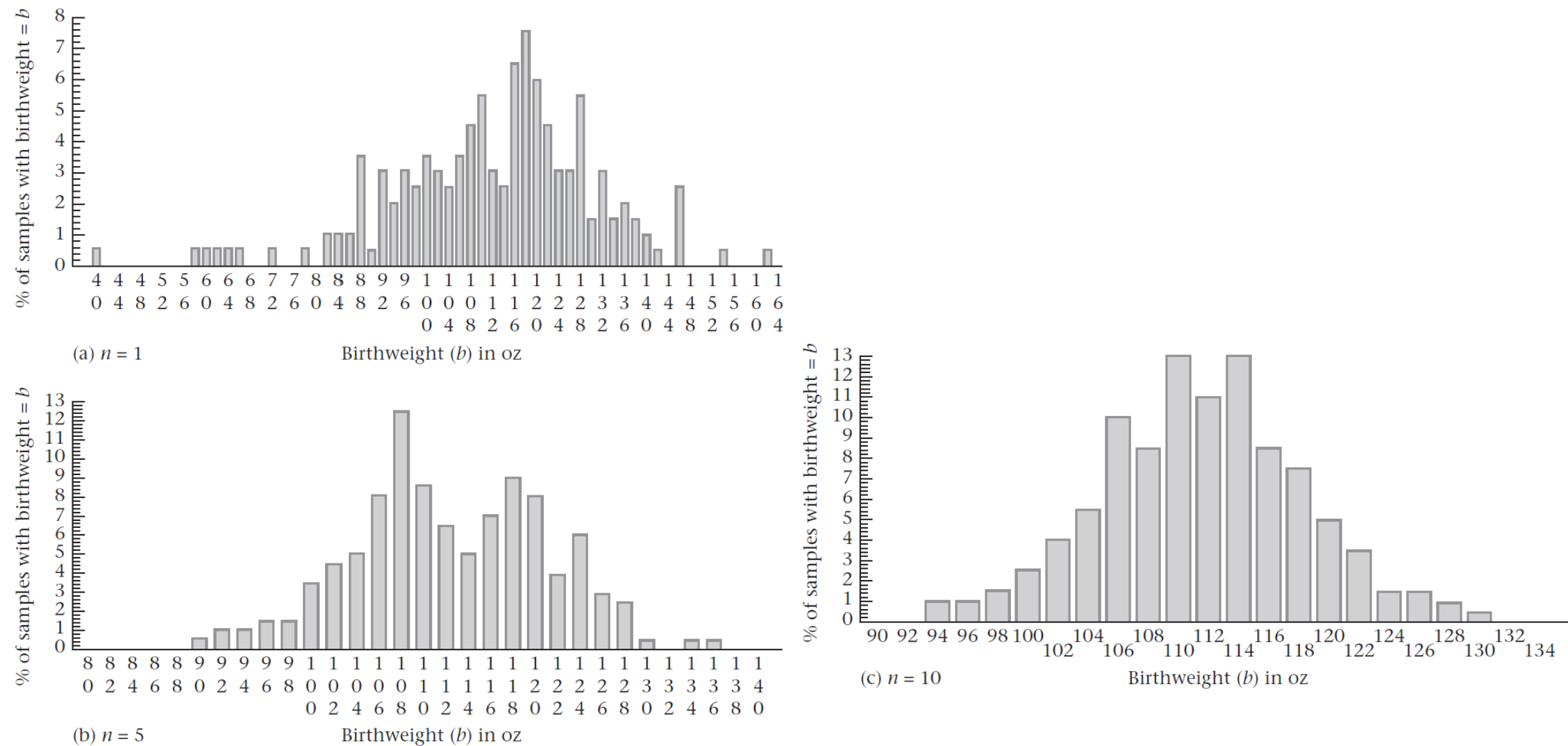
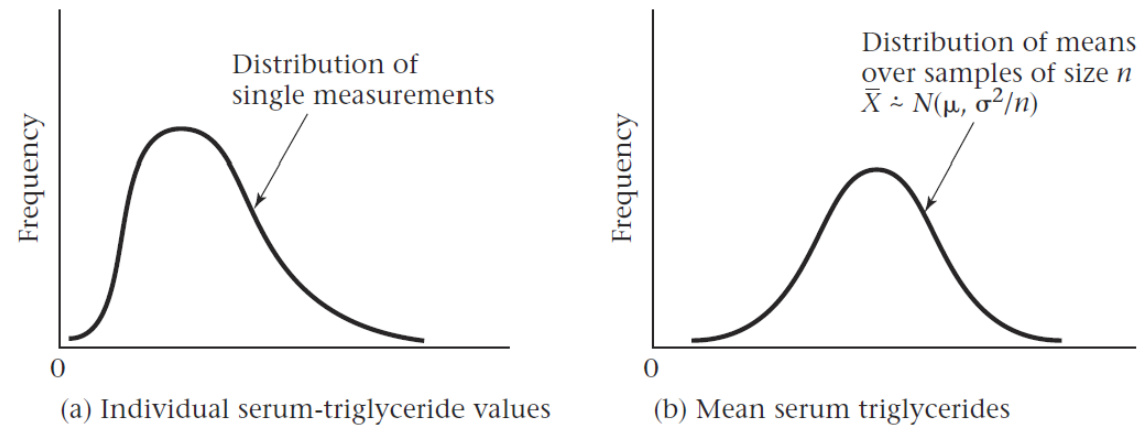


FIGURE 6.5 Distribution of single serum-triglyceride measurements and of means of such measurements over samples of size n



- **Interval Estimation:** Specify a range within which parameter values are likely to fall.
- 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 the 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

- σ can be estimated by the sample standard deviation s and to try to construct CIs using the quantity

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- however, this is not normally distributed.
- This problem was solved by a statistician named **William Gossett (“Student”)**. The distribution $(\bar{X} - \mu)/(s/\sqrt{n})$ is referred to as **Student’s t distribution**, the shape of which depends on the sample size n .
- The t distribution is not unique but is a family of distributions indexed by a parameter referred to as **degrees of freedom (df)** of the distribution.
- If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ and are independent, then

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

- that is, it follows a t distribution with $n - 1$ degrees of freedom.

- Student's t distribution is a family of distributions indexed by the degrees of freedom d .
- The t distribution with d degrees of freedom is referred to as the t_d distribution.
- The **$100 \times u$ th percentile of a t distribution with d degrees of freedom** is denoted by $t_{d,u}$, that is,
$$\Pr(T_d < t_{d,u}) = u$$

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

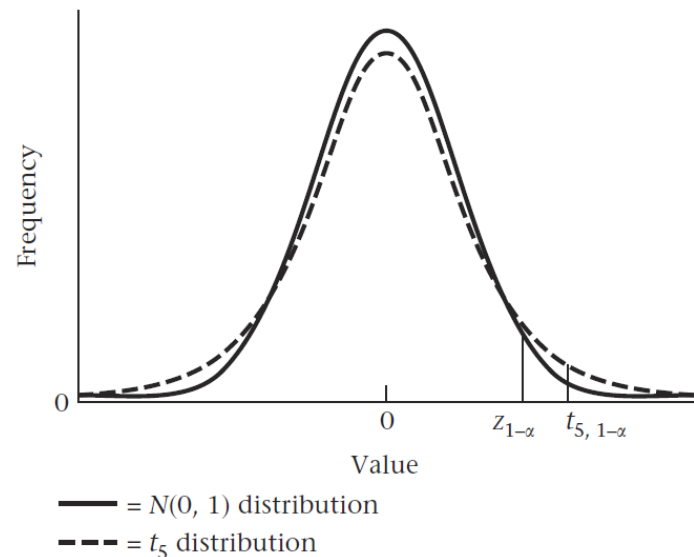


TABLE 6.5 Comparison of the 97.5th percentile of the t distribution and the normal distribution

d	$t_{d,975}$	$z_{.975}$	d	$t_{d,975}$	$z_{.975}$
4	2.776	1.960	60	2.000	1.960
9	2.262	1.960	∞	1.960	1.960
29	2.045	1.960			

- The percentage points of the t distribution for various degrees of freedom.
- The u th percentile of a t distribution with d degrees of freedom is found by reading across the row marked d and reading down the column marked u .
- A $100\% \times (1 - \alpha)$ CI for the mean μ of a normal distribution with unknown variance is given by

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

- Short hand notation:

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

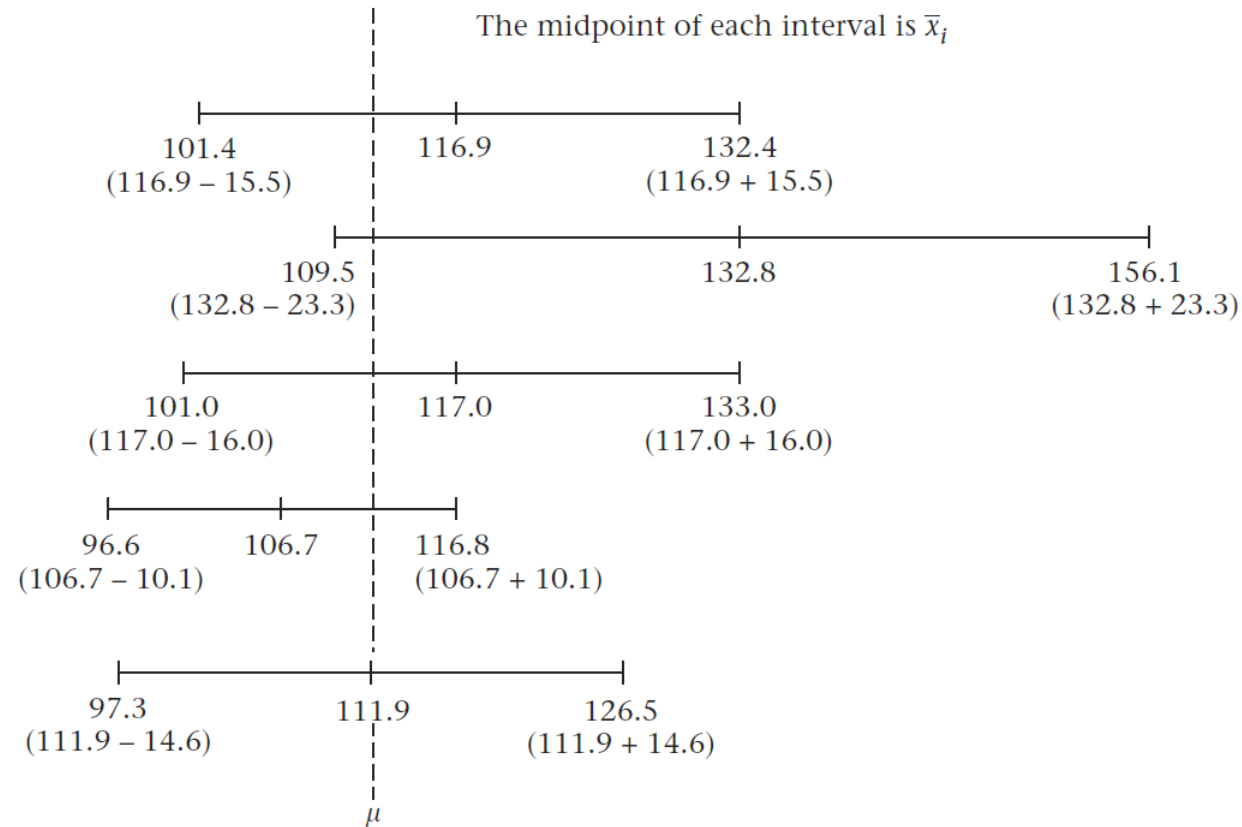
Confidence Interval for the Mean of a Normal Distribution (Large-Sample Case)

- An approximate $100\% \times (1 - \alpha)$ CI for the mean μ of a normal distribution with unknown variance is given by

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

- This interval should only be used if $n > 200$.
- However, it can also be used for $n \leq 200$ if the standard deviation (σ) is known, by replacing s with σ .
- The boundaries of the interval depend on the sample mean and sample variance and vary from sample to sample.

FIGURE 6.7 A collection of 95% CIs for the mean μ as computed from repeated samples of size 10 (see Table 6.3) from the population of birthweights given in Table 6.2



- Over the collection of all 95% CIs that could be constructed from repeated random samples of size n , 95% will contain the parameter μ .
- The length of the CI indicates the precision of the point estimate $\bar{x}$.

Factors Affecting the Length of a CI

- The length of a $100\% \times (1 - \alpha)$ CI for μ equals

$$2t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}$$

- and is determined by n , s , and α .
- n : As the sample size (n) increases, the length of the CI decreases.
- s : As the standard deviation (s), which reflects the variability of the distribution of individual observations, increases, the length of the CI increases.
- α : As the confidence desired increases (α decreases), the length of the CI increases.
- Usually only n and α can be controlled. s is a function of the type of variable being studied, although s itself can sometimes be decreased if changes in technique can reduce the amount of measurement error, day-to-day variability, and so forth.

Estimation of the Variance of a Distribution

Point Estimation

- Let $X_1, \dots, X_n$ 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

$$E(S^2) = \sigma^2$$

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

Person (i)	Observer		Difference (d_i)
	1	2	
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

The higher the σ^2 , the greater is the interobserver variability.

- 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.
- If

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

- where $X_1, \dots, X_n \sim N(0,1)$
- 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 only takes on positive values and is always skewed to the right.

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

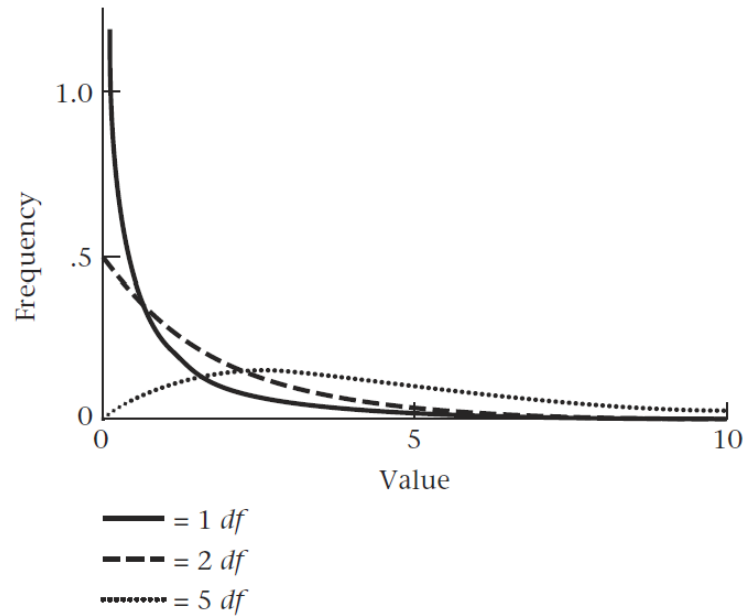
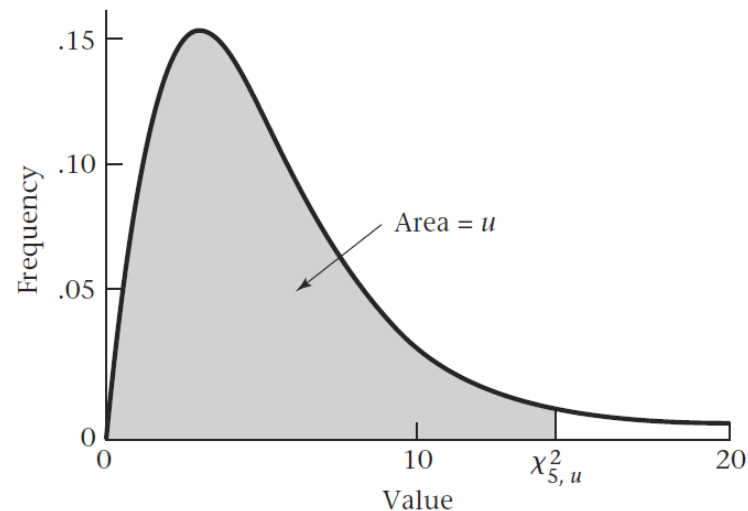


FIGURE 6.9 Graphic display of the percentiles of a χ^2_5 distribution



- For $n \geq 3$, the distribution has a mode greater than 0 and is skewed to the right. The skewness diminishes as n increases.

- The u th percentile of a χ^2_d distribution (that is, a chi-square distribution with d df) is denoted by $\chi^2_{d,u}$ where

$$\Pr(\chi^2_d < \chi^2_{d,u}) = u.$$

Interval Estimation

- To obtain an interval estimate of σ^2 , we need to find the sampling distribution of S^2 . We assume that $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Then it can be shown that

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

- If $X \sim N(\mu, \sigma^2)$ then standardization of X gives a new random variable
- $Z = (X - \mu)/\sigma$, then Z will be normally distributed with mean 0 and variance 1.

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

- chi-square distribution with n df.
- We estimate μ by $\bar{X}$. However, if we substitute $\bar{X}$ for μ , then we lose 1 df, resulting in

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

- The sample variance is given by

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

- Then, multiplying both sides by $(n - 1)$,

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

- which is

$$\frac{(n - 1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

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

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

Point Estimation for a Binomial Parameter p

- Let X be a binomial random variable with parameters n and p . An unbiased estimator of p is given by the sample proportion of events $\hat{p}$. Its standard error is given exactly by $\sqrt{pq/n}$ and is estimated by $\sqrt{\hat{p}\hat{q}/n}$.
- A better justification for the use of $\hat{p}$ to estimate the binomial parameter p is that $\hat{p}$ is a **maximum-likelihood estimator** of p . To understand this concept, we need to define a **likelihood**.
- The pmf of a discrete random variable X is a function of k parameters $p_1, \dots, p_k$, denoted by $\mathbf{p}$. If $x_1, \dots, x_n$ is a sample of n independent observations from X , then the likelihood of the sample $x_1, \dots, x_n$ given $\mathbf{p}$ is denoted by $L(\mathbf{x} \mid \mathbf{p})$ and represents the probability of obtaining our sample given specified values for the parameters $p_1, \dots, p_k$.

$$L(\mathbf{x} \mid \mathbf{p}) = \Pr(x_1 \mid \mathbf{p}) \times \cdots \times \Pr(x_n \mid \mathbf{p}) = \prod_{i=1}^n \Pr(x_i \mid \mathbf{p})$$

- Suppose we have a **pmf function** or a **probability-density function** that is a function of k parameters $p_1, \dots, p_k$.
- The **maximum-likelihood estimator (MLE)** of the parameters $p_1, \dots, p_k$ are the values $p_{1,ML}, \dots, p_{k,ML}$, which maximize the likelihood. Heuristically, the MLE can be thought of as the values of the parameters that maximize the probability of the observed data given the parameters.
- For large n , $\hat{p} = \bar{X}$ is approximately normally distributed with mean $\mu = p$ and variance $\sigma^2/n = pq/n$, or

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right)$$

- The number of successes in Bernoulli's trials is $X = n\hat{p}$.
- Multiplying by n on both sides, we have

$$X \sim N(np, npq)$$

- An approximate $100\% \times (1 - \alpha)$ CI for the binomial parameter p based on the normal approximation to the binomial distribution is given by

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- This method of interval estimation should only be used if $n\hat{p}\hat{q} \geq 5$.

- An exact $100\% \times (1 - \alpha)$ CI for the binomial parameter p that is always valid is given by (p_1, p_2) , where p_1, p_2 satisfy the equations.

$$\Pr(X \geq x \mid p = p_1) = \frac{\alpha}{2} = \sum_{k=\bar{x}x}^n \binom{n}{k} p_1^k (1 - p_1)^{n-k}$$

$$\Pr(X \leq x \mid p = p_2) = \frac{\alpha}{2} = \sum_{k=0}^x \binom{n}{k} p_2^k (1 - p_2)^{n-k}$$

Exact Confidence Limits for Binomial Proportions

(1) The sample size (n) is given along each curve. There are two curves corresponding to each given sample size. One curve is used to obtain the lower confidence limit and the other to obtain the upper confidence limit.

(2) If $0 \leq \hat{p} \leq .5$, then

- Refer to the lower horizontal axis, and find the point corresponding to $\hat{p}$.
- Draw a line perpendicular to the horizontal axis, and find the two points where this line intersects the two curves identified in (1).
- Read across to the left vertical axis; the smaller value corresponds to the lower confidence limit and the larger value to the upper confidence limit.

(3) If $.5 < \hat{p} \leq 1.0$, then

- Refer to the upper horizontal axis, and find the point corresponding to $\hat{p}$.
- Draw a line perpendicular to the horizontal axis, and find the two points where this line intersects the two curves identified in (1).
- Read across to the right vertical axis; the smaller value corresponds to the lower confidence limit and the larger value to the upper confidence limit.

Point Estimation for Poisson Distribution

- A person-year is a unit of time defined as 1 person being followed for 1 year.
- Let's assume the number of events X over T person-years is Poisson-distributed with parameter $\mu = \lambda T$.
- An unbiased estimator of λ is given by

$$\hat{\lambda} = \frac{X}{T}$$

- where X is the observed number of events over T person-years.
- If λ is the incidence rate per person-year, T = number of person-years of follow-up, and we assume a Poisson distribution for the number of events X over T person-years, then the expected value of X is given by

$$\begin{aligned} E(X) &= \lambda T \\ E(\hat{\lambda}) &= E(X)/T = \lambda T/T = \lambda \end{aligned}$$

- Thus, $\hat{\lambda}$ is an unbiased estimator of λ .

Exact Method for Obtaining a CI for the Poisson Parameter λ

- An exact $100\% \times (1 - \alpha)$ CI for the Poisson parameter λ is given by $(\mu_1/T, \mu_2/T)$
- where μ_1, μ_2 satisfy the equations

$$\Pr(X \geq x \mid \mu = \mu_1) = \frac{\alpha}{2} = \sum_{k=x}^{\infty} \frac{e^{-\mu_1} \mu_1^k}{k!} = 1 - \sum_{k=0}^{x-1} \frac{e^{-\mu_1} \mu_1^k}{k!}$$

$$\Pr(X \leq x \mid \mu = \mu_2) = \frac{\alpha}{2} = \sum_{k=0}^x \frac{e^{-\mu_2} \mu_2^k}{k!}$$

- x = observed number of events, T = number of person-years of follow-up.
- An upper one-sided $100\% \times (1 - \alpha)$ CI is of the form $p > p_1$ such that $\Pr(p > p_1) = 1 - \alpha$.
- If we assume that the normal approximation to the binomial holds true, then we can show that this CI is given approximately by

$$p > \hat{p} - z_{1-\alpha} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- This interval estimator is only used if $n\hat{p}\hat{q} \geq 5$.

Lower One-Sided CI for the Binomial Parameter p — Normal-Theory Method

- The interval $p < p_2$ such that
$$\Pr(p < p_2) = 1 - \alpha$$
- is referred to as a lower one-sided $100\% \times (1 - \alpha)$ CI and is given approximately by

$$p < \hat{p} + z_{1-\alpha} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Summary

- In this chapter, we discussed
- Sampling distribution, crucial to understanding the principles of statistical inference
- Minimum-variance unbiased estimator of μ
- Maximum-likelihood estimator (MLE) of μ
- Central-limit theorem
- Interval estimate or confidence interval
- t and chi-square distributions to obtain interval estimates

Key Concepts and Formulas of This Chapter

Concept	Formula
Sample Mean	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
Unbiased Estimation	$E(\hat{\theta}) = \theta$
Standard Error of the Sample Mean	$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
Central Limit Theorem	$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
Confidence Interval for the Population Mean	$\bar{x} \pm t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}$
Confidence Interval for the Population Variance	$\left(\frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2} \right)$
Estimation of a Binomial Proportion	$\hat{p} = \frac{X}{n}$
Estimation of a Poisson Incidence Rate	$\hat{\lambda} = \frac{X}{T}$