

Week 4

Estimation of Parameters

Dr. Esam Mahdi

Department of Mathematics, Statistics and Physics
Qatar University

Learning Objectives

- Draw inferences about the population of interest based on output from a sample.
- Compute the point and interval estimation for the population mean for small and large samples.
- Compute the confidence interval for difference of means of two independent populations.
- Compute the confidence interval for difference of means of two paired populations (before and after intervention, or dependent case and control groups).
- Compute the point and interval estimation for the population proportion.
- Compute the confidence interval for difference of population proportions of two independent populations.
- Compute the confidence interval for the odds ratios.
- Compute the confidence interval for the population correlation coefficient.

Estimation of Population Parameters

- A numerical characteristic of a target population is called a **parameter**.
 - Generally, it would be a time consuming or too costly to obtain the entire population.
 - Sometimes the population does not even exist (e.g., investigative drug for leukemia).
- Decisions in health science are often use a small sample of a population.
- The estimator of the parameter, obtained from a sample, is called a **statistic**.

	Population	Sample
Size	N	n
Mean	μ (mu)	\bar{x} (x-bar)
Variance	σ^2 (sigma-squared)	S^2 (s-squared)
Standard Deviation	σ	S
Proportion	π (pi)	P
Covariance	σ_{xy}	S_{xy}
Coefficient of Correlation	ρ (rho)	r

Research's Objectives

- Depending on the research's objectives, Inferences about the population of interest are classified into two categories:
 1. Estimate the value of a parameter (e.g., estimate the response rate of a leukemia investigative drug).
 2. Compare the parameters for two subpopulations using statistical tests of significance (e.g., decide whether men have higher cholesterol levels, on average, than women?).
- In formal statistical estimation, we can determine the amount of **uncertainty** (and so the error) in the estimate. It gives us the best guess and then tells us how “wrong” the guess could be, in quite precise terms.

Basic Concepts

- **Variable (Random variable)** is a characteristic that changes or varies over time and/or for different individuals or objects under consideration.
- **Examples:** Hair color, white blood cell count, weight, height, blood pressure, or the presence or absence of a certain habit or practice, such as smoking or use of drugs.
- The **distribution** of a random variable is often assumed to belong to a certain family of distributions, such as binomial, Poisson, or normal.
- The family of distributions is specified or indexed by one or several parameters, such as a population mean μ or a population proportion π .

Sampling Distributions

- The distribution of values of a statistic obtained from repeated samples of the same size from a given population is called the *sampling distribution* of that statistic.
- **Unbiased estimator:** The estimator $\hat{\theta}$ (computed from a sample) is an unbiased estimator of the parameter θ (unknown value in a population) if the expected value of $\hat{\theta}$ equals θ .
 - In other words, if we use the sample mean (sample proportion) to estimate the population mean (population proportion), we are correct on the average that the sample mean (sample proportion) is the same as the mean of the original distribution.
- The **variance** of a sampling distribution of a statistic can be used as a measure of precision of that statistic.
 - A small variance for a sampling distribution indicates that most possible values for the statistic are close to each other, so that a particular value is more likely to be reproduced.
 - The smaller this quantity, the better the statistic as an estimate of the corresponding parameter.
- The square root of this variance is called the **standard error** of the statistic; for example, we will have the standard error of the sample mean, or $SE(\bar{x})$; the standard error of the sample proportion, $SE(p)$; and so on.

Example 4.1

- Consider a population consisting of six subjects.
- Table 4.1 gives the subject names and values of a variable under investigation (e.g., 1 for a smoker and 0 for a nonsmoker).
- In this case the population mean μ (also the population proportion π for this very special dichotomous variable) is 0.5 ($= 3/6$).
- We now consider all possible samples, without replacement, of size 3.
- Table 4.2 represents the sampling distribution of the sample mean.

Example 4.1 (Cont.)

TABLE 4.1

Subject	Value
A	1
B	1
C	1
D	0
E	0
F	0

TABLE 4.2

Samples	Number of Samples	Value of Sample Mean, \bar{x}
(D, E, F)	1	0
(A, D, E), (A, D, F), (A, E, F) (B, D, E), (B, D, F), (B, E, F) (C, D, E), (C, D, F), (C, E, F)	9	$\frac{1}{3}$
(A, B, D), (A, B, E), (A, B, F) (A, C, D), (A, C, E), (A, C, F) (B, C, D), (B, C, E), (B, C, F)	9	$\frac{2}{3}$
(A, B, C)	1	1
Total	20	

This sampling distribution gives the following properties:

- Its mean (i.e., the mean of all possible sample means) is unbiased estimator

$$\frac{1 \times 0 + 9 \times 1/3 + 9 \times 2/3 + 1 \times 1}{20} = 0.5$$

- If we form a **bar graph** for this sampling distribution (Figure 4.1). It shows a shape somewhat similar to that of a symmetric, bell-shaped normal curve. This is much clearer with real populations and larger sample sizes.

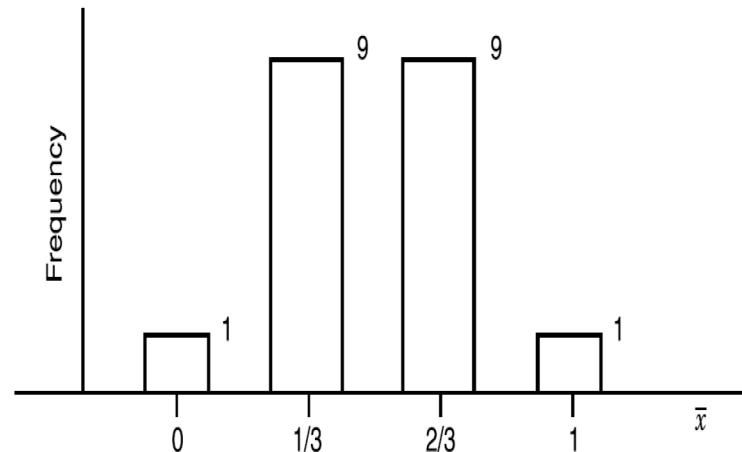


Figure 4.1 Bar graph for sampling distribution in Example 4.1.

Estimation of Means

- **Central limit theorem (CLT):** Given any population with mean μ and variance σ^2 , the sampling distribution of \bar{x} will be approximately normal with mean μ and variance σ^2/n when the sample size n is large.
 - In practice, $n \geq 25$ or more could be considered adequately large).
- This means that we have the two properties

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

- It follows that $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Example 4.2

- Birth weights obtained from deliveries over a long period of time at a certain hospital show a mean μ of 112 oz and a standard deviation σ of 20.6 oz. Let us suppose that we want to compute the probability that the mean birth weight from a sample of 25 infants will fall between 107 and 117 oz (i.e., the estimate is off the mark by no more than 5 oz).
- The central limit theorem is applied and it indicates that \bar{x} follows a normal distribution with mean $\mu_{\bar{x}} = 112$ and variance $\sigma_{\bar{x}}^2 = \frac{(20.6)^2}{25}$ or standard error $\sigma_{\bar{x}} = 4.12$.
- It follows that

$$\begin{aligned}\Pr(107 \leq \bar{x} \leq 117) &= \Pr\left(\frac{107 - 112}{4.12} \leq z \leq \frac{117 - 112}{4.12}\right) \\ &= \Pr(-1.21 \leq z \leq 1.21) = (2)(0.3869) \\ &= 0.7738\end{aligned}$$



- In other words, if we use the mean of a sample of size $n = 25$ to estimate the population mean, about 80% of the time we are correct within 5 oz.
- This figure would be 98.5% if the sample size were 100.

Confidence Estimation

- Statistical estimation can be classified into two categories:
 1. **Point estimator:** A single numerical value used to estimate the corresponding population parameter (e.g., $\bar{x} = 120\$$ estimates the truth unknown income mean μ , the sample proportion p estimates the population one, π , and so on).
 2. **Interval estimator** (e.g., the truth income mean is almost certainly between 100\$ and 145\$. i.e., $100 < \mu < 145$).
- The point estimate and its **standard error** are combined to form an interval estimate (or confidence interval).
- In general, If θ is an unknown parameter, its confidence interval

$$\hat{\theta} \mp d \times SE(\hat{\theta}),$$

where d is the degrees of confidence associated with the $(1 - \alpha)\%$ level of confidence around the θ ; usually $d = 1.96$ for $\alpha = 0.05$.

- Note that $d \times SE(\hat{\theta})$ is called as **the margin of error of $\hat{\theta}$** .

95% Confidence Interval (C.I) for a Mean

- In practice, the most conventional degree of confidence is the 95% level of confidence around the mean μ .
- A 95% **confidence interval** for a mean μ is the interval of the endpoints

$$\bar{x} \mp 1.96 \times \frac{\sigma}{\sqrt{n}}$$

- Equivalently, we have the interval $a \leq \mu \leq b$, where

$$a = \bar{x} - 1.96 \times \sigma/\sqrt{n} \quad \& \quad b = \bar{x} + 1.96 \times \sigma/\sqrt{n}$$

- **Interpretation:** In 95% of the samples we take, the true population mean will be in the interval

$$\bar{x} - 1.96 \times \sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1.96 \times \sigma/\sqrt{n}$$

- If σ is unknown (usually σ is unknown) and $n \geq 25$, we can replace σ by its estimator s . In this case, the 95% **confidence interval** for a mean μ is

$$\bar{x} \mp 1.96 \times SE(\bar{x}) = \bar{x} \mp 1.96 \times \frac{s}{\sqrt{n}}$$

Confidence Intervals for a Mean

- Table 4.4 provides some common degrees of confidence. For example $d = 2.576$ is associated with a 99% level of confidence.
- Coefficients in Table 4.4 are taken from the standard normal distribution table (2.576, 1.960, 1.645, and 1.282 corresponding for degrees of confidence 99%, 95%, 90%, and 80% respectively).

TABLE 4.4

Degree of Confidence	$d = \text{Coefficient}$
99%	2.576
→ 95%	1.960
90%	1.645
80%	1.282

Example 4.3

- For the data on percentage saturation of bile for 31 male patients of Example 2.4:

40, 86, 111, 86, 106, 66, 123, 90, 112, 52, 88, 137, 88, 88, 65, 79, 87, 56, 110,
106, 110, 78, 80, 47, 74, 58, 88, 73, 118, 67, 57

we have $n = 31$, $\bar{x} = 84.65$, $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = 24$

leading to a standard error $SE(\bar{x}) = \frac{24}{\sqrt{31}} = 4.31$

and a 95% confidence interval for the population mean:

$$84.65 \mp 1.96 \times 4.31 = (76.2, 93.1)$$

- The resulting interval is wide, due to a large standard deviation as observed from the sample, $s = 24$, reflecting heterogeneity of sample subjects.

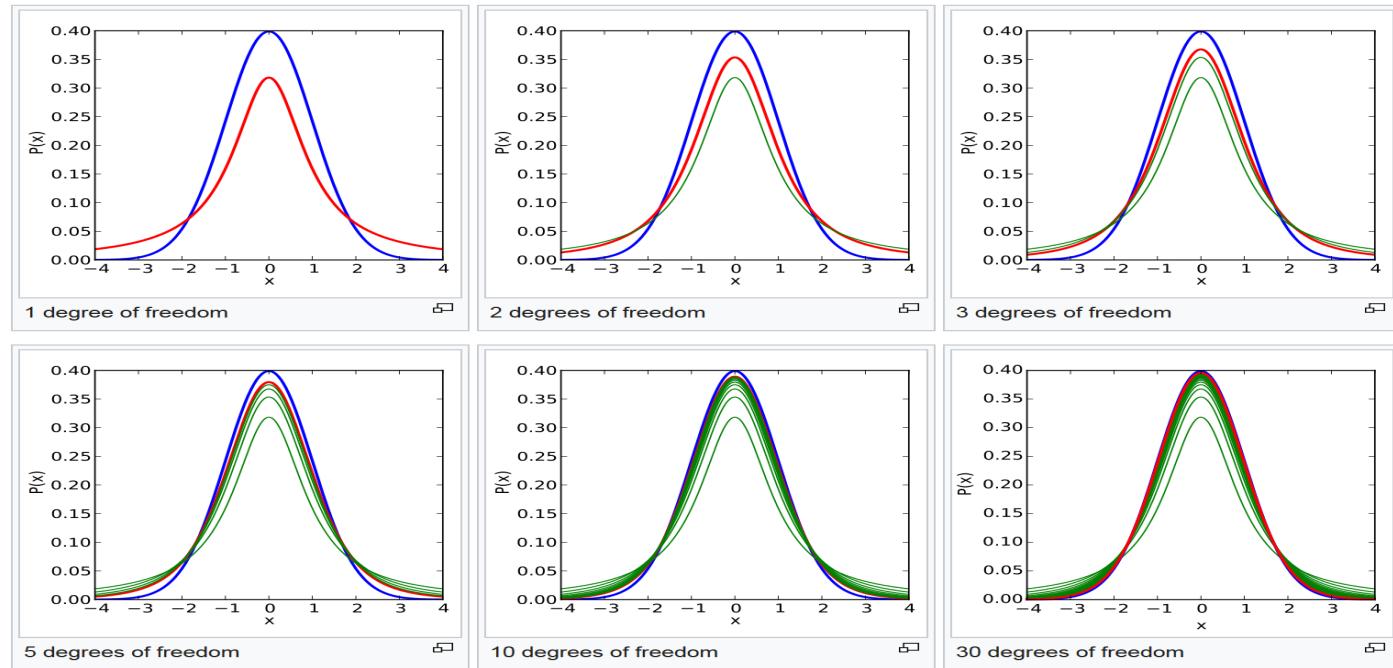


Uses of Small Samples

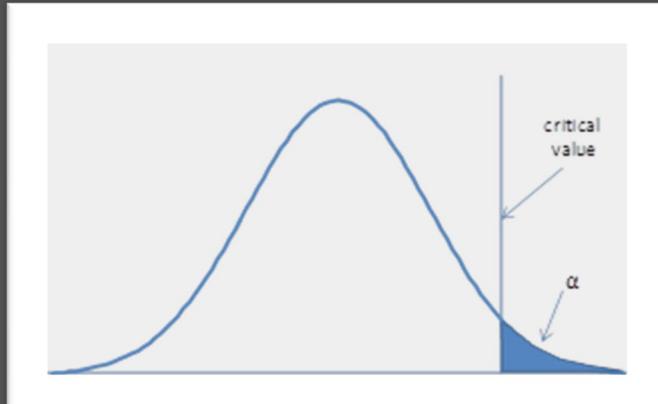
- If σ is unknown (usually σ is unknown) and $n < 25$, we can still replace σ by its estimator s . However, In this case, we will use corresponding numbers from the t curves (see Appendix C) where the quantity of information is indexed by the degree of freedom ($df = n - 1$).

Density of the t -distribution (red) for 1, 2, 3, 5, 10, and 30 degrees of freedom compared to the standard normal distribution (blue).

Previous plots shown in green.



Appendix C: Percentiles of the t- distribution



Entries in the table give t_α values, where α is the area or probability in the upper tail of the t distribution. For example, with 10 degrees of freedom and a 0.025 area in the upper tail, $t_{.025} = 2.228$.

Degrees of Freedom	Area in Upper Tail				
	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Confidence Intervals for a Mean

- In practice, the **confidence interval** for a mean μ is

$$\bar{x} \mp t_{\alpha,n-1} \times \frac{s}{\sqrt{n}},$$

where $t_{\alpha,n-1}$ is a value given in the table of Appendix C where α is the area or probability in the upper tail of the t distribution.

- The column to read is the one with the correct normal coefficient on the bottom row (marked with $df = \infty$). See, for example, Table 4.5 for the case where the degree of confidence is 0.95.
- **For better results, it is always a good practice to use the t table regardless of sample size because coefficients such as 1.96 are only for very large sample sizes.**

TABLE 4.5

df	t Coefficient (percentile)
5	2.571
10	2.228
15	2.131
20	2.086
24	2.064
$\rightarrow \infty$	1.960

Example 4.4

- In an attempt to assess the physical condition of joggers, a sample of $n = 25$ joggers was selected and maximum volume oxygen (VO_2) uptake was measured, with the following results:

$$\bar{x} = 47.5 \text{ mL/kg}$$

$$s = 4.8 \text{ mL/kg}$$

$$\begin{aligned}\text{SE}(\bar{x}) &= \frac{4.8}{\sqrt{25}} \\ &= 0.96\end{aligned}$$



From Appendix C we find that the t coefficient with 24 df for use with a 95% confidence interval is 2.064, leading to a 95% confidence interval for the population mean μ (this is the population of joggers' VO_2 uptake) of

$$47.5 \mp 2.064 \times 0.96 = (45.5, 49.5)$$

Example 4.5

- In addition to the data in Example 4.4, we have data from a second sample consisting of 26 non joggers which were summarized into these statistics:

$$n_2 = 26$$

$$\bar{x}_2 = 37.5 \text{ mL/kg}$$

$$s_2 = 5.1 \text{ mL/kg}$$

$$\begin{aligned} \text{SE}(\bar{x}) &= \frac{s_2}{\sqrt{n_2}} \\ &= \frac{5.1}{\sqrt{26}} \\ &= 1.0 \end{aligned}$$



- Solution:** From Appendix C we find that the t coefficient with 25 df for use with a 95% confidence interval is 2.060, leading to a 95% confidence interval for the population mean μ (this is the population of joggers' VO_2 uptake) of

$$37.5 \pm (2.060)(1.0) = (35.4, 39.6)$$

Evaluation of Interventions

To determine the effect of a risk factor or an intervention, we may want to estimate the difference of means: say, between the population of cases and the population of controls.

Confidence interval for the difference of the means of a paired samples ($\mu_d = \mu_x - \mu_y$)

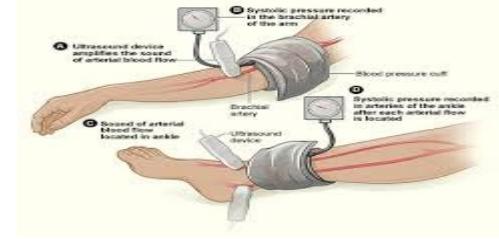
- The procedure is to reduce the data to a one sample problem by computing before-and-after (or control-and-case) differences for each subject (or pairs of matched subjects).
- By doing this with paired observations, we get a set of differences that can be handled as a single sample problem.
- The **confidence interval** for the difference mean $\mu_d = \mu_x - \mu_y$ is

$$\bar{d} \mp t_{\alpha,n-1} \times \frac{s_d}{\sqrt{n}}$$

Sample 1 (x)	Sample 2 (y)	Difference (d=x-y)
x_1	x_1	$d_1 = x_1 - y_1$
x_2	y_2	$d_2 = x_2 - y_2$
:	:	:
x_n	y_n	$d_n = x_n - y_n$

Note that sample 1 and sample 2 are dependent (or matched paired). Usually they represent sample before and after intervention.

Example 4.6



- The systolic blood pressures of 12 women between the ages of 20 and 35 were measured before and after administration of a newly developed oral contraceptive.

Given the data in Table 4.6, we have from the column of differences, the d_i 's, leading to

$$\bar{d} = \text{average difference} = \frac{31}{12} = 2.58 \text{ mmHg}$$

$$s^2 = \frac{185 - (31)^2/12}{11} = 9.54$$

$$s = 3.09$$

$$\text{SE}(\bar{d}) = \frac{3.09}{\sqrt{12}} = 0.89$$

TABLE 4.6

Subject	Systolic Blood Pressure (mmHg)		After-Before Difference, d_i	d_i^2
	Before	After		
1	122	127	5	25
2	126	128	2	4
3	132	140	8	64
4	120	119	-1	1
5	142	145	3	9
6	130	130	0	0
7	142	148	6	36
8	137	135	-2	4
9	128	129	1	1
10	132	137	5	25
11	128	128	0	0
12	129	133	4	16

With a degree of confidence of 0.95 the t coefficient from Appendix C is 2.201, for 11 degrees of freedom, so that a 95% confidence interval for the mean difference is

$$2.58 \mp (2.201)(0.89) = (0.62, 4.54)$$

This means that the “after” mean is larger than the “before” mean, an increase of between 0.62 and 4.54.

Confidence interval for the difference of the means of independent samples ($\mu_d = \mu_x - \mu_y$)

- In many other interventions, or in studies to determine possible effects of a risk factor, it may not be possible to employ matched design.
- The comparison of means is based on data from two independent samples.
- The process of estimating the difference of means is summarized briefly as follows:

1. Data are summarized separately to obtain

$$\text{sample 1: } n_1, \bar{x}_1, s_1^2$$

$$\text{sample 2: } n_2, \bar{x}_2, s_2^2$$

2. The standard error of the difference of means is given by

$$\text{SE}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

3. Finally, a 95% confidence interval for the difference of population means, $\mu_1 - \mu_2$, can be calculated from the formula

$$(\bar{x}_1 - \bar{x}_2) \pm (\text{coefficient})\text{SE}(\bar{x}_1 - \bar{x}_2)$$

where the coefficient is 1.96 if $n_1 + n_2$ is large; otherwise, a t coefficient is used with approximately

$$\text{df} = n_1 + n_2 - 2$$

Example



- Biologists took samples of the crawfish species *Orconectes sanborii* from two rivers in central Ohio, Y = the Upper Cuyahoga River (CUY) and X = East Fork of Pine Creek (EFP), and measured the length (mm) of each crawfish captured. Table shows the summary statistics

$$\bar{y}_1 - \bar{y}_2 = 22.91 - 21.97 = 0.94$$

$$SE(\bar{Y}_1 - \bar{Y}_2) = \sqrt{0.69^2 + 0.53^2} = 0.87$$

If we construct a 95% confidence interval for $(\mu_1 - \mu_2)$ we get

$$0.94 \pm 1.96 \times 0.87$$

or

$$(-0.7652, 2.6452)$$

Crawfish data:length (mm)		
	CUY	EFP
n	30	30
\bar{y}	22.91	21.97
s	3.78	2.90

Note: The confidence interval includes zero, which indicates that we have no significant evidence that the mean of CUY is differ from the mean EFP.

Estimation of Proportions

- Let π represents the population proportion and the sample proportion, $p = \frac{x}{n}$ (where x is the number of positive outcomes and n is the sample size).
- The [central limit theorem](#) implies that the sampling distribution of the sample proportion, p , will be approximately normal when the sample size n is large.
- The mean and variance of this sampling distribution are $\mu_p = \pi$ and $\sigma_p^2 = \frac{\pi(1-\pi)}{n}$ respectively.
- That is; when n is large then $p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$ so that $z = \frac{P-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$.
- The approximate 95% confidence interval for a population proportion π :

$$p \mp 1.96 \times SE(p),$$

where $SE(p)$ is the standard error of the sample proportion and calculated from

$$SE(P) = \sqrt{p(1-p)/n}$$

Example 4.7

- Suppose that the true proportion of smokers in a community is known to be in the vicinity of $\pi = 0.4$, and we want to estimate it using a sample of size $n = 100$. The central limit theorem indicates that p follows a normal distribution with mean $\mu_p = 0.4$ and variance $\sigma_p^2 = \frac{0.4(0.6)}{100}$, or standard error $\sigma_p = 0.049$.
- Suppose that we want our estimate to be correct within $\mp 3\%$; it follows that

$$\begin{aligned}\Pr(0.37 \leq p \leq 0.43) &= \Pr\left(\frac{0.37 - 0.40}{0.049} \leq z \leq \frac{0.43 - 0.40}{0.049}\right) \\ &= \Pr(-0.61 \leq z \leq 0.61) \\ &= (2)(0.2291) \\ &= 0.4582 \quad \text{or} \quad \text{approximately } 46\%\end{aligned}$$

That means if we use the proportion of smokers from a sample of $n = 100$ to estimate the true proportion of smokers, only about 46% of the time are we correct within $\mp 3\%$.

Example 4.8

Example 4.8 Consider the problem of estimating the prevalence of malignant melanoma in 45- to 54-year-old women in the United States. Suppose that a random sample of $n = 5000$ women is selected from this age group and $x = 28$ are found to have the disease. Our point estimate for the prevalence of this disease is

$$\begin{aligned} p &= \frac{28}{5000} \\ &= 0.0056 \end{aligned}$$

Its standard error is

$$\begin{aligned} \text{SE}(p) &= \sqrt{\frac{(0.0056)(1 - 0.0056)}{5000}} \\ &= 0.0011 \end{aligned}$$



Therefore, a 95% confidence interval for the prevalence π of malignant melanoma in 45- to 54-year-old women in the United States is given by

$$0.0056 \pm (1.96)(0.0011) = (0.0034, 0.0078)$$

Example 4.9

Example 4.9 A public health official wishes to know how effective health education efforts are regarding smoking. Of $n_1 = 100$ males sampled in 1965 at the time of the release of the Surgeon General's Report on the health consequences of smoking, $x_1 = 51$ were found to be smokers. In 1980 a second random sample of $n_2 = 100$ males, gathered similarly, indicated that $x_2 = 43$ were smokers. Application of the method above yields the following 95% confidence intervals for the smoking rates:

(a) In 1965, the estimated rate was

$$\begin{aligned} p_1 &= \frac{51}{100} \\ &= 0.51 \end{aligned}$$

with its standard error

$$\begin{aligned} \text{SE}(p_1) &= \sqrt{\frac{(0.51)(1 - 0.51)}{100}} \\ &= 0.05 \end{aligned}$$

leading to a 95% confidence interval of

$$0.51 \pm (1.96)(0.05) = (0.41, 0.61)$$

(b) In 1980, the estimated rate was

$$\begin{aligned} p_2 &= \frac{43}{100} \\ &= 0.43 \end{aligned}$$

with its standard error

$$\begin{aligned} \text{SE}(p_2) &= \sqrt{\frac{(0.43)(1 - 0.43)}{100}} \\ &= 0.05 \end{aligned}$$

leading to a 95% confidence interval of

$$0.43 \pm (1.96)(0.05) = (0.33, 0.53)$$

It can be seen that the two confidence intervals, one for 1965 and one for 1980, are both quite long and overlapped, even though the estimated rates show a decrease of 8% in smoking rate, because the sample sizes are rather small.

Example 4.10

- A study was conducted to look at the effects of oral contraceptives (OC) on heart disease in women 40–44 years of age.
- It is found that among $n_1 = 5,000$ current OC users, 13 develop a myocardial infarction (MI) over a three-year period,
- while among $n_1 = 10,000$ non-OC users, seven develop an MI over a three-year period.



Application of the method described above yields the following 95% confidence intervals for the MI rates:

(a) For OC users, the estimated rate was

$$p_1 = \frac{13}{5000} = 0.0026$$

with its standard error

$$\begin{aligned}\text{SE}(p_1) &= \sqrt{\frac{(0.0026)(1 - 0.0026)}{5000}} \\ &= 0.0007\end{aligned}$$

leading to a 95% confidence interval of

$$0.0026 \pm (1.96)(0.0007) = (0.0012, 0.0040)$$

(b) For non-OC users, the estimated rate was

$$p_2 = \frac{7}{10,000} = 0.0007$$

with its standard error

$$\begin{aligned}\text{SE}(p_2) &= \sqrt{\frac{(0.0007)(1 - 0.0007)}{10,000}} \\ &= 0.0003\end{aligned}$$

leading to a 95% confidence interval of

$$0.0007 \pm (1.96)(0.0003) = (0.0002, 0.0012)$$

It can be seen that the two confidence intervals, one for OC users and one for non-OC users, do not overlap, a strong indication that the two population MI rates are probably not the same.

Comparison of Proportions based on Data from Two Independent Samples

- The 95% confidence interval for the difference of proportions from two independent populations is

$$(p_1 - p_2) \pm (1.96) \text{SE}(p_1 - p_2)$$

where

$$\text{SE}(p_1 - p_2) = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

$p_1 = \frac{x}{n_1}$, and $p_2 = \frac{y}{n_2}$ (where x and y are the number of positive outcomes from population 1 and 2 respectively and n_1 and n_2 are the corresponding sample sizes.

Estimation of Odds Ratios

- The sampling distributions of the sample odds ratio is positively skewed.
- These sampling distributions can be almost normalized by taking the logarithm.
- Data from a case-control study, for example, may be summarized in a 2×2 table (Table 4.7).
- A 95% confidence interval for the odds ratio under investigation is obtained by exponentiating the two endpoints:

$$\ln \frac{ad}{bc} \pm 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

TABLE 4.7

Exposure	Cases	Controls
Exposed	a	c
Unexposed	b	d

(a) The odds that a case was exposed is

$$\text{odds for cases} = \frac{a}{b}$$

(b) The odds that a control was exposed is

$$\text{odds for controls} = \frac{c}{d}$$

Therefore, the (observed) odds ratio from the samples is

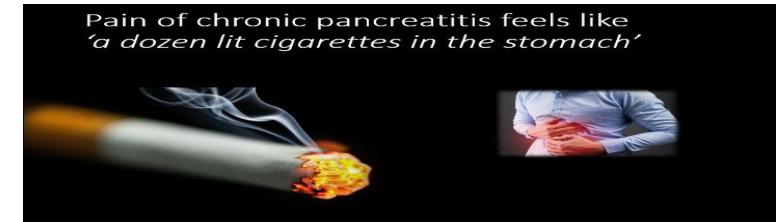
$$\text{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Example 4.11

- The role of smoking in pancreatitis has been recognized for many years; the data shown in Table 4.8 are from a case-control study carried out in eastern Massachusetts and Rhode Island in 1975-1979 (see Example 1.14). We have

TABLE 4.8

Use of Cigarettes	Cases	Controls
Current smokers	38	81
Ex-smokers	13	80
Never	2	56



a) For ex-smokers compared to those who never smoked, $OR = \frac{13 \times 56}{80 \times 2} = 4.55$

and a 95% confidence interval for the population odds ratio on the log scale is from

$$\ln(4.55) \pm 1.96 \sqrt{\frac{1}{13} + \frac{1}{56} + \frac{1}{80} + \frac{1}{2}} = (-0.01, 3.04)$$

and hence the corresponding 95% confidence interval for the population odds ratio obtained by exponentiating (the reverse log operation or antilog) is $(0.99, 2.96)$

b) For current smokers, compared to those who never smoked, $OR = \frac{38 \times 56}{81 \times 2} = 13.14$

and a 95% confidence interval for the population odds ratio on the log scale is from

$$\ln(13.14) \pm 1.96 \sqrt{\frac{1}{38} + \frac{1}{56} + \frac{1}{81} + \frac{1}{2}} = (1.11, 4.04)$$

and hence the corresponding 95% confidence interval for the population odds ratio obtained by exponentiating (the reverse log operation or antilog) is $(3.04, 56.70)$

Example 4.12

- Toxic shock syndrome (TSS) is a disease first recognized in the 1980s, characterized by sudden onset of high fever ($> 102^{\circ}\text{F}$), vomiting, diarrhea, rapid progression to hypotension, and in most cases, shock.
- Because of the striking association with menses, several studies have been undertaken to look at various practices associated with the menstrual cycle.
- In a study by the Centers for Disease Control, 30 of 40 TSS cases and 30 of 114 controls who used a single brand of tampons used the Rely brand.

TABLE 4.9

Brand	Cases	Controls
Rely	30	30
Others	10	84
Total	40	114



- Data are presented in a 2×2 table (Table 4.9). We have $\text{OR} = \frac{30 \times 84}{10 \times 30} = 8.4$ and a 95% confidence interval for the population odds ratio on the log scale is from

$$\ln(8.4) \mp 1.96 \sqrt{\frac{1}{30} + \frac{1}{10} + \frac{1}{30} + \frac{1}{84}} = (1.30, 2.96)$$

- The reverse antilog 95% confidence interval for the population odds ratio is $(3.67, 19.30)$, indicating a very high risk elevation for Rely users.

Estimation of Correlation Coefficients

- Pearson's coefficient of correlation (denoted as r) is a measure of the linear correlation between two variables X and Y .

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{[\sum(x - \bar{x})^2][\sum(y - \bar{y})^2]}} = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}}$$

- The sampling distributions of the sample coefficient of correlation is positively skewed.
- The sample coefficient of correlation r is a number between 0 and 1.
 - Values near 1 indicate a strong positive association.
 - Values near -1 indicate a strong negative association.
 - Values around 0 indicate a weak association.

Confidence Interval for Population Coefficient of Correlation

- The sampling distributions of the sample coefficient of correlation r is positively skewed (not normal).
- The transformed scale $z = \frac{1}{2} \ln \frac{1+r}{1-r}$ follows a normal distribution with mean r and variance $\frac{1}{n-3}$.
- Consequently, an approximate 95% confidence for the population correlation coefficients interval, on z , for Pearson's correlation coefficients is given by

$$z \mp 1.96 \times \sqrt{1/(n-3)}$$

- A 95% confidence interval for the coefficients of correlation under investigation is (r_l, r_u)

where the lower endpoint $r_l = \frac{\exp(2z_l)-1}{\exp(2z_l)+1}$ and the upper endpoint $r_u = \frac{\exp(2z_u)-1}{\exp(2z_u)+1}$,

$$z_l = z - 1.96 \times \sqrt{1/(n-3)} \text{ and } z_u = z + 1.96 \times \sqrt{1/(n-3)}$$



Example 4.13

- The data shown in Table 4.10 represent systolic blood pressure readings on 15 women. The descriptive analysis in Example 2.9 yields $r = 0.566$ and we have

$$z = \frac{1}{2} \ln \frac{1 + 0.566}{1 - 0.566} = 0.642$$

$$z_l = 0.642 - 1.96\sqrt{\frac{1}{12}} = 0.076$$

$$z_u = 0.642 + 1.96\sqrt{\frac{1}{12}} = 1.207$$

$$r_l = \frac{\exp(0.152) - 1}{\exp(0.152) + 1} = 0.076$$

$$r_u = \frac{\exp(2.414) - 1}{\exp(2.414) + 1} = 0.836$$

TABLE 4.10

Age (x)	SBP (y)	Age (x)	SBP (y)
42	130	80	156
46	115	74	162
42	148	70	151
71	100	80	156
80	156	41	125
74	162	61	150
70	151	75	165
80	156		

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{[\sum(x - \bar{x})^2][\sum(y - \bar{y})^2]}} = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}}$$

or a 95% confidence interval for the population coefficient of correlation of **(0.076, 0.836)**, indicating a positive association between a woman's age and her systolic blood pressure; that is, older women are likely to have higher systolic blood pressure.

Example 4.14

- Table 4.11 gives the values for the birth weight (x) and the increase in weight between days 70 and 100 of life, expressed as a percentage of the birth weight (y) for 12 infants. The descriptive analysis in Example 2.8 yields $r = 0.946$ and we have

$$z = \frac{1}{2} \ln \frac{1 - 0.946}{1 + 0.946} = -1.792$$

$$z_l = 0.014 - 1.96 \sqrt{\frac{1}{9}} = -2.446$$

$$z_u = 0.014 + 1.96 \sqrt{\frac{1}{9}} = -1.139$$

$$r_l = \frac{\exp(-4.892) - 1}{\exp(-4.892) + 1} = -0.985$$

$$r_u = \frac{\exp(-2.278) - 1}{\exp(-2.278) + 1} = -0.814$$

TABLE 4.11

x (oz)	y (%)	x (oz)	y (%)
112	63	81	120
111	66	84	114
107	72	118	42
119	52	106	72
92	75	103	90
80	118	94	91

or a 95% confidence interval for the population coefficient of correlation of $(-0.985, -0.814)$, indicating a very strong negative association between a baby's birth weight and his or her increase in weight between days 70 and 100 of life; that is, smaller babies are likely to grow faster during that period (that may be why, at three months, most babies look the same size!).