Basic Statistics and Probability

Chapter 9:
Inferences Based on Two Samples:
Confidence Intervals and Tests of Hypotheses

- Identifying the Target Parameter
- Comparing Two Population Means: Independent Sampling
- Comparing Two Population Means: Paired Difference Experiments
- Comparing Two Population Proportions: Independent Sampling
- Determining the Sample Size

Identifying the Target Parameter

Many experiments involve a comparison of two populations.

Example 9.1: Brain Size

Do children diagnosed with attention deficit/hyperactivity disorder (ADHD) have smaller brains than children without this condition?

This was the topic of a research study described in the paper "Developmental Trajectories of Brain Volume Abnormalities in Children and Adolescents with Attention Deficit/Hyperactivity Disorder" (*Journal of the American Medical Association* 2002).

Brain scans were completed for 152 children with ADHD and 139 children of similar age without ADHD. Summary values for total cerebral volume (in milliliters) are:

	n	mean	s
Children with ADHD	152	1059.4	117.5
Children without ADHD	139	1104.5	111.3

Example 1.4: Fish Diet. Source: DASL

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer. The original study actually used pairs of twins, which enabled the researchers to discern that the risk of cancer for those who never ate fish actually was substantially greater.

Data and description at:
`https://dasl.datadescription.com/datafile/fish-diet/`

Do these data provide sample evidence to conclude that prostate cancer incidence depends on fish consumption?

We observe the same type of random variable (e.g., cerebral volume) in two different populations (children with and without ADHD). We denote the variable by X in the Population 1 and by Y in Population 2.

In each of the two populations, the probability distribution of the random variable depends on an unknown parameter (for example, the population mean μ or variance σ^2 or a proportion p).

We want to compare the same parameter of interest in the two populations:

$$\begin{array}{c|c|c} \text{Population 1} & & \text{Population 2} \\ \hline \mu_1 = E(X) & \leftrightarrow & \mu_2 = E(Y) \\ \sigma_1^2 = V(X) & \leftrightarrow & \sigma_2^2 = V(Y) \\ p_1 & \leftrightarrow & p_2 \end{array}$$

To this end, we use the information provided by a sample $x_1, \ldots, x_{n_1}$ from X and a sample $y_1, \ldots, y_{n_2}$ from Y.

Determining the Target Parameter

Parameter	Key Words or Phrases	Type of Data
$\mu_1 - \mu_2$	Mean difference; difference in averages	Quantitative
$p_1 - p_2$	Difference between proportions, percentages, fractions, or rates; compare proportions	Qualitative
$(\sigma_1)^2/(\sigma_2)^2$	Ratio of variances; difference in variability or spread; compare variation	Quantitative

Throughout this chapter, we will use the following notation:

- For general r.v.'s X and Y,

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \qquad \text{and} \qquad \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$$

are the sample means of X and Y respectively, and

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2 \qquad \text{and} \qquad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$$

are the sample variances of X and Y respectively.

- For $X \sim \text{Bernoulli}(p_1)$ and $Y \sim \text{Bernoulli}(p_2)$,

$$\hat{p}_1 = \bar{x} \qquad \text{and} \qquad \hat{p}_2 = \bar{y}$$

are the sample proportions of "successes" in Populations 1 and 2 respectively.

Comparing Two Population Means: Independent Sampling

We assume that the two populations, X and Y, are independent (intuitively, whether one of the variables takes some value or other does not have any influence on the values taken by the other variable).

We want to compare the population means of X and Y:

$$\mu_1 = E(X) \qquad \text{and} \qquad \mu_2 = E(Y),$$

via a confidence interval for their difference:

$$\text{CI}_{1-\alpha}(\mu_1 - \mu_2)$$

or via hypothesis testing

$$\begin{array}{lll} H_0 : \mu_1 - \mu_2 = D_0 & H_0 : \mu_1 - \mu_2 \geq D_0 & H_0 : \mu_1 - \mu_2 \leq D_0 \\ H_1 : \mu_1 - \mu_2 \neq D_0 & H_1 : \mu_1 - \mu_2 < D_0 & H_1 : \mu_1 - \mu_2 > D_0 \end{array}$$

LARGE SAMPLES

We assume that both sample sizes are large ($n_1 \geq 20$ and $n_2 \geq 20$).

There are no assumptions on the probability distributions of X and Y.

By the Central Limit Theorem (CLT), the sampling distribution of $\bar{x} - \bar{y}$ is approximately normal.

Then, a confidence interval for $\mu_1 - \mu_2$ with approximate confidence level $1 - \alpha$ is

$$\text{CI}_{1-\alpha}(\mu_1 - \mu_2) = \left(\bar{x} - \bar{y} \mp z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

Example 9.1: Brain Size

Let D_0 be a fixed numerical value (typically, $D_0 = 0$).

Hypothesis test	Rejection region for significance level α
$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 \neq D_0$	$R = \{\|z\| > z_{\alpha/2}\}$
$H_0 : \mu_1 - \mu_2 \geq D_0$ $H_1 : \mu_1 - \mu_2 < D_0$	$R = \{z < -z_\alpha\}$
$H_0 : \mu_1 - \mu_2 \leq D_0$ $H_1 : \mu_1 - \mu_2 > D_0$	$R = \{z > z_\alpha\}$

where the test statistic is

$$z = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

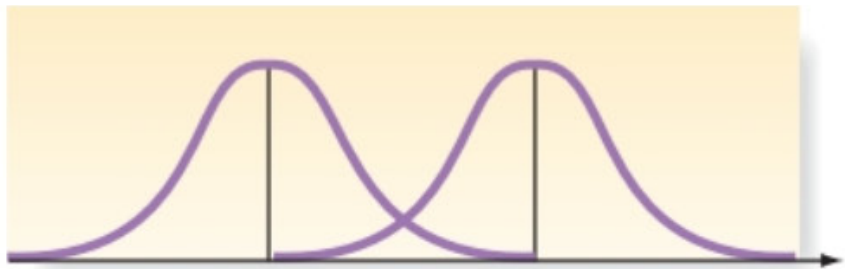
Example 9.1: Brain Size

Do the data provide evidence that the mean brain volume of children with ADHD is smaller than the mean brain volume for children without ADHD? Use a .05 level of significance.

SMALL SAMPLES

Here at least one of the sample sizes is small ($n_1 < 20$ or $n_2 < 20$).

Consequently, the CLT cannot be applied. We have to assume that X and Y are normally distributed and have the same variance ($\sigma_1^2 = \sigma_2^2 = \sigma_2$).



We use the information in both samples to estimate the common variance σ^2 of X and Y via the pooled sample variance s_p^2, a weighted average of the sample variances s_1^2 and s_2^2,

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{i=1}^{n_2}(y_i-\bar{y})^2}{n_1+n_2-2}.$$

Example 9.2: White Coat Effect on Blood Pressure

The tendency for blood pressure to be higher when measured in a doctor's office than when measured in a less stressful environment is called "white coat effect".

In a study, patients with high blood pressure were randomly assigned to one of two groups. Those in the first group (talking group) were asked questions about their medical history and about the sources of stress in their lives in the minutes before their blood pressure was measured. Those in the second group (counting group) were asked to count aloud from 1 to 100 four times before their blood pressure was measured. The following are the data values for diastolic blood pressure (in mm Hg):

Talking	104	110	107	112	108	103	108	118
	$n_1 = 8$	$\bar{x} = 108.75$	$s_1^2 = 4.74$					
Counting	110	96	103	98	100	109	97	105
	$n_2 = 8$	$\bar{y} = 102.25$	$s_2^2 = 5.39$					

A confidence interval for $\mu_1 - \mu_2$ with confidence level $1 - \alpha$ is

$$\text{CI}_{1-\alpha}(\mu_1 - \mu_2) = \left(\bar{x} - \bar{y} \mp t_{n_1+n_2-2;\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right)$$

Example 9.2: White Coat Effect on Blood Pressure

Let D_0 be a fixed numerical value (typically, $D_0 = 0$).

Hypothesis test	Rejection region for significance level α
$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 \neq D_0$	$R = \{\lvert t\rvert > t_{n_1+n_2-2;\alpha/2}\}$
$H_0 : \mu_1 - \mu_2 \geq D_0$ $H_1 : \mu_1 - \mu_2 < D_0$	$R = \{t < -t_{n_1+n_2-2;\alpha}\}$
$H_0 : \mu_1 - \mu_2 \leq D_0$ $H_1 : \mu_1 - \mu_2 > D_0$	$R = \{t > t_{n_1+n_2-2;\alpha}\}$

where the test statistic is

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Example 9.2: White Coat Effect on Blood Pressure

Is there enough sample evidence to conclude that counting reduces the white coat effect?

Comparing Two Population Means: Paired Difference Experiments

We observe the **same random variable** on the same individual before (X) and after an intervening "treatment" (Y), or on two individuals which are somehow related (e.g., twins). Then X and Y are paired, a specific type of dependency between variables.

Example 9.3: Benefits of Ultrasound

Ultrasound is often used in the treatment of soft tissue injuries. An experiment investigated the effect of an ultrasound and stretch therapy on knee extension, by measuring range of motion both before and after treatment in physical therapy patients.

	Range of Motion						
Subject	1	2	3	4	5	6	7
Pre-treatment	31	53	45	57	50	43	32
Post-treatment	32	59	46	64	49	45	40

Data are from "Location of Ultrasound Does Not Enhance Range of Motion Benefits of Ultrasound and Stretch Treatment" (Univ. of Virginia Thesis, T. Tashiro, 2003)

We want to compare $E(X) = \mu_1$ and $E(Y) = \mu_2$, using the information from a sample $(x_1, y_1), \ldots, (x_n, y_n)$ of (X, Y).

With paired data the procedure is to define $D = X - Y$ and assume that $D \sim N(\mu_d = \mu_1 - \mu_2, \sigma_d)$. Then $d_1 = x_1 - y_1, \ldots, d_n = x_n - y_n$ is a sample from D.

We express the confidence intervals and hypothesis tests on $\mu_1 - \mu_2$ in terms of μ and then apply the methods of Ch. 7 and 8 for the mean of a Gaussian population:

$$\mathrm{CI}_{1-\alpha}(\mu_1 - \mu_2) = \mathrm{CI}_{1-\alpha}(\mu_d)$$

$$\begin{array}{ll} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{array} \Leftrightarrow \begin{array}{ll} H_0: & \mu_d = 0 \\ H_1: & \mu_d \neq 0 \end{array}$$

$$\begin{array}{ll} H_0: & \mu_1 \leq \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{array} \Leftrightarrow \begin{array}{ll} H_0: & \mu_d \leq 0 \\ H_1: & \mu_d > 0 \end{array}$$

$$\begin{array}{ll} H_0: & \mu_1 \geq \mu_2 \\ H_1: & \mu_1 < \mu_2 \end{array} \Leftrightarrow \begin{array}{ll} H_0: & \mu_d \geq 0 \\ H_1: & \mu_d < 0 \end{array}$$

Example 9.3: Benefits of Ultrasound

Is there evidence that the ultrasound and stretch treatment increases range of motion?

	Range of Motion						
Subject	1	2	3	4	5	6	7
Pre-treatment (X)	$x_1 = 31$	$x_2 = 53$	$x_3 = 45$	$x_4 = 57$	$x_5 = 50$	$x_6 = 43$	$x_7 = 32$
Post-treatment (Y)	$y_1 = 32$	$y_2 = 59$	$y_3 = 46$	$y_4 = 64$	$y_5 = 49$	$y_6 = 45$	$y_7 = 40$
$D = X - Y$	$d_1 = -1$	$d_2 = -6$	$d_3 = -1$	$d_4 = -7$	$d_5 = 1$	$d_6 = -2$	$d_7 = -8$

$$\bar{d} = \text{Sample mean of differences} = \frac{1}{7}\sum_{i=1}^{7} d_i = -3.4286$$

$$s_d = \text{Sample s.d. of differences} = \sqrt{\frac{1}{6}\sum_{i=1}^{7}(d_i - \bar{d})^2} = 3.5051$$

Comparing Two Population Proportions: Independent Sampling

Suppose $X \sim \text{Bernoulli}(p_1)$ and $Y \sim \text{Bernoulli}(p_2)$ are two independent r.v. We want to compare p_1 and p_2 based on a sample $x_1, \ldots, x_{n_1}$ from X and a sample $y_1, \ldots, y_{n_2}$ from Y.

Example 1.4: Fish Diet. Source: DASL

The data are summarized as follows:

		Fish Consumption			
		Large	Moderate	Small	Never
Cancer	Yes	42	209	201	14
	No	507	2769	2420	110
Total		549	2978	2621	124

To check that those who never eat fish have a higher incidence of prostrate cancer, we might consider comparing

p_1 = proportion of prostate cancer in group "Never"

p_2 = proportion of prostate cancer in group "Large"

If sample sizes are large ($n_1 \geq 20$ and $n_2 \geq 20$), then by the CLT, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

Then, a large-sample $(1 - \alpha)$ confidence interval for the diffrence of the two proportions is

$$\text{CI}_{1-\alpha}(p_1 - p_2) = \left(\hat{p}_1 - \hat{p}_2 \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right).$$

Example 1.4: Fish Diet

Hypothesis test	Rejection region for significance level α
$H_0 : p_1 - p_2 = 0$ $H_1 : p_1 - p_2 \neq 0$	$R = \{\|z\| > z_{\alpha/2}\}$
$H_0 : p_1 - p_2 \geq 0$ $H_1 : p_1 - p_2 < 0$	$R = \{z < -z_{\alpha}\}$
$H_0 : p_1 - p_2 \leq 0$ $H_1 : p_1 - p_2 > 0$	$R = \{z > z_{\alpha}\}$

where the test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

and

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i}{n_1 + n_2}.$$

Example 1.4: Fish Diet

At a 5% significance level, is there sample evidence that the proportion of men who develop prostrate cancer is higher in the group who never eats fish than in the group with large fish consumption?

Determining the Sample Size

You can find the appropriate sample size to estimate the difference between a pair of parameters with a specified sampling error (SE) and degree of reliability $1-\alpha$, by equating the half-width of the confidence interval for the parameters difference to the SE.

We assume that $n_1 = n_2 = n$.

To estimate $\mu_1 - \mu_2$ to within a sampling error SE and with confidence level $1-\alpha$, take

$$n = \frac{(z_{\alpha/2})^2(s_1^2 + s_2^2)}{(\text{SE})^2},$$

where s_1^2 and s_2^2 are sample variances from prior pilot samples of X and Y.

To estimate $p_1 - p_2$ to within a sampling error SE and with confidence level $1 - \alpha$, take

$$n = \frac{(z_{\alpha/2})^2(\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2))}{(\text{SE})^2},$$

where $\hat{p}_1$ and $\hat{p}_2$ are estimates of p_1 and p_2 respectively from prior pilot samples of X and Y.