



Chapter 5

1. The importance of learning statistics

Captured Tank Serial Numbers Reveal Population Size

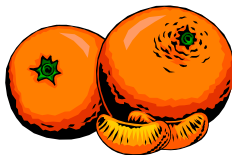


During World War II, Allied intelligence specialists wanted to determine the number of tanks Germany was producing. Traditional spy techniques provided unreliable results, but statisticians obtained accurate estimates by analyzing serial numbers on captured tanks. As one example, records show that Germany actually produced 271 tanks in June 1941.

The estimate based on serial numbers was 244, but traditional intelligence methods resulted in the extreme estimate of 1550. (See “An empirical Approach to Economic Intelligence in World War II,” by Ruggles and Brodie, *Journal of the American Statistical Association*, Vol 42.)

2. Estimating the population mean μ , \bar{x} , the point estimator

Estimating Sugar in Oranges



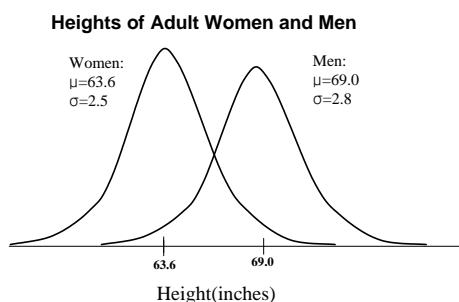
In Florida, members of the citrus industry make extensive use of statistical methods. One particular application involves the way in which growers are paid for oranges used to make orange juice. An arriving truckload of oranges is first weighed at the receiving plant, then a sample of about a dozen oranges is randomly selected. The sample is weighed

and then squeezed, and the amount of sugar in the juice is measured. Based on the sample results, an estimate is made of the total amount of sugar in the entire truckload. Payment for the load of oranges is based on the estimate of

the amount of sugar because sweeter oranges are more valuable than those less sweet, even though the amounts of juice may be the same.

3. The Bell-Curve. Normal Distributions

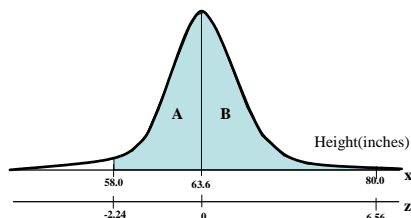
EXAMPLE 5.1 The U.S. Army requires that women's heights be between 58 in. and 80 in. Find the percentage of women satisfying that requirement. Again assume that women have heights that are normally distributed with a mean of 63.6 in. and a standard deviation of 2.5 in.



SOLUTION

$$z = \frac{x - \mu}{\sigma} = \frac{58.0 - 63.6}{2.5} = -2.24 \quad z = \frac{80.0 - 63.6}{2.5} = 6.56$$

Area of regions A and B combined = 0.4875 + 0.4999 = 0.9874.



4. Distribution of the sample mean– the Central Limit Theorem(CLT)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \text{or} \quad Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1)$$

The above approximation is good when the sample size is large. ($n \geq 30$)

Note 1 When the population is normal

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad (\text{For any sample size})$$

5. Constructing Confidence Interval for μ

Q: How close is \bar{x} to μ ?

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$
$$\implies P\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

We are $100(1-\alpha)\%$ sure that μ is between $\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$ and $\left(\bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$

Equivalently, we say that a $100(1 - \alpha)\%$ Confidence Interval for μ is

$$\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

1. When σ is unknown and $n \geq 30$, we can substitute σ with the sample standard deviation S ;
2. Some common values for α are 0.05, 0.01, 0.1.

EXAMPLE 5.2

In a random sample of $n = 36$ parochial schools throughout the south, the average number of pupils per school is 379.2, with a standard deviation of 124. Use the sample to construct a 95% confidence interval for μ , the mean number of pupils per school for all parochial schools in the south.

SOLUTION The sample data indicate that $\bar{y} = 379.2$ and $s = 124$. The appropriate 95% confidence interval is then computed by using the formula

$$\bar{y} \pm 1.96\sigma_{\bar{y}}$$

where $\sigma_{\bar{y}} = s/\sqrt{n}$. In section 5.8 we present a procedure for obtaining a confidence interval for μ when σ is unknown. However, for all practical purposes, if the sample size is 30 or more, we can estimate the population standard deviation σ with s in the confidence interval formula. With s replacing σ , our interval is

$$379.2 \pm 1.96 \frac{124}{\sqrt{36}} \quad \text{or} \quad 379.2 \pm 40.51$$

The interval from 338.69 to 419.71 forms a 95% confidence interval for μ . In other words, we are 95% sure that the average number of pupils per school for parochial schools throughout the south lies between 338.69 and 419.71.

EXAMPLE 5.3

A forester is interested in estimating the average number of “count trees” per acre (trees) larger than a specified size) on a 2,000-acre plantation. She can then use this information to determine the total timber volume for trees in the plantation. A random sample of $n = 50$ 1-acre plots is selected and examined. The average(mean) number of count trees per acre is found to be 27.3, with a standard deviation of 12.1. Use this information to construct a 99% confidence interval for μ , the mean number of count trees per acre for the entire plantation.

SOLUTION We use the general confidence interval with confidence coefficient equal to .99 and a $z_{\alpha/2}$ -value equal to 2.58 (see Table 5.1). Substituting into the formula $\bar{y} \pm 2.58\sigma_{\bar{y}}$ where $\sigma_{\bar{y}} = s/\sqrt{n}$, we have

$$27.3 \pm 2.58 \frac{12.1}{\sqrt{50}}.$$

This corresponds to the confidence interval 27.3 ± 4.42 , that is, the interval from 22.88 to 31.72. Thus, we are 99% sure that the average number of count trees per acre is between 22.88 and 31.72

6. Choosing the sample size for estimating μ

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

where E is the maximum error. (Note, $E = \frac{L}{2}$ where L is the length of the $100(1 - \alpha)\%$ Confidence interval for μ .)

Derivation

$$\begin{aligned} P(|\bar{X} - \mu| < E) &= 1 - \alpha \\ \implies P(-E < \bar{X} - \mu < E) &= 1 - \alpha \\ \implies P\left(\frac{-E}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{E}{\sigma/\sqrt{n}}\right) &= 1 - \alpha \\ \implies P\left(\frac{-E}{\sigma/\sqrt{n}} < Z < \frac{E}{\sigma/\sqrt{n}}\right) &= 1 - \alpha \\ \therefore \frac{E}{\sigma/\sqrt{n}} = Z_{\alpha/2} &\iff n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2. \end{aligned}$$

EXAMPLE 5.4

Union officials are concerned about reports of inferior wages paid to a company’s employees under their jurisdiction. It is decided to take a random sample of n wage sheets from the company to estimate the average hourly wage. If it

is known that wages in the company have a range of \$10 per hour, determine the sample size required to estimate the average hourly wage μ using a 95% confidence interval with width equal to \$1.20.

SOLUTION Since we want a 95% confidence interval with width \$1.20, $E = \$0.60$. The value that we use to substitute for σ is $\text{range}/4 = 2.50$. Substituting into the formula for n we have

$$n = \frac{(1.96)^2(2.5)^2}{(.60)^2} = 66.69.$$

To be on the safe side, we will round this number up to the next integer. A sample size of 67 should give a 95% confidence interval with the desired width of \$1.20.

EXAMPLE 5.5

A federal agency has decided to investigate the advertised weight printed on cartons of a certain brand of cereal. The company in question periodically samples cartons of cereal coming off the production line to check their weight. A summary of 1,500 of the weights made available to the agency indicates a mean weight of 11.80 ounces per carton and a standard deviation of .75 ounce. Use this information to determine the number of cereal cartons the federal agency must examine to estimate the average weight of cartons being produced now, using a 99% confidence interval of width .50.

SOLUTION The federal agency has specified that the width of the confidence interval is to be .50, so $E = .25$. Assuming that the weights made available to the agency by the company are accurate, we can take $\sigma = .75$. The required sample size with $z_{\alpha/2} = 2.58$ is

$$n = \frac{(2.58)^2(.75)^2}{(.25)^2} = 59.91.$$

That is, the federal agency must obtain a random sample of 60 cereal cartons to estimate the mean weight to within $\pm .25$.

7. Hypothesis testing — a brief overview

- A hypothesis test is rather similar to a criminal trial.
- The null hypothesis (H_0) is our original belief (the prejudice). In a trial, the defendant is assumed innocent until proven guilty. That is, H_0 : the defendant is innocent.
- The alternative hypothesis (H_1 or H_a) is the belief contradicting to H_0 . That is, H_a : the defendant is guilty.
- The aim of the hypothesis testing is to reject H_0 in favor of H_a . We make the judgement based on the amount of evidence we have (whether the test statistic falls into the rejection region or not).

- The Type I and Type II Errors.
- The probability of committing the Type I error (or the Type I error rate) is called the significance level of the test and is denoted by the Greek letter α .
- The probability of committing the Type II error (or the Type II error rate) is denoted by the Greek letter β .
- The power of the test is the probability of rejecting H_0 when H_a is true. It equals to $(1 - \beta)$.
- That is $\alpha = P(\text{reject } H_0 | H_0)$; $\beta = P(\text{fail to reject } H_0 | H_a)$ and $1 - \beta = P(\text{reject } H_0 | H_a)$.
- Since Type I error is often considered more severe, we usually fix the Type I error rate for a given test. The fixed Type I error rate will help to determine the rejection region of the test.
- The P-value: it is the probability of observing a value for the test statistic that is at least as extreme as the one observed assuming the null hypothesis is true (under the null hypothesis). The p -value is also called the observed significance level.



The experimental evidence against the null hypothesis that Lucy did not act deliberately reaches a P -value of one in ten billion. Nevertheless, Charlie Brown repeats the experiment each year.

p-value for one-tailed test To summarize, the **level of significance for a one-tailed test** can be computed as follows: For $H_a : \mu > \mu_0$,

$$p = P[z > \text{computed } z].$$

For $H_a : \mu < \mu_0$,

$$p = P[z < \text{computed } z].$$

p-value for two-tailed test For two-tailed tests (as determined by the form of H_a), we still compute the probability of obtaining a sample outcome more contradictory to H_0 than the observed result, but the level of significance is commonly taken to be twice this probability. For a two-tailed test, the level of significance can be written as

$$p = 2P[z > |\text{computed } z|].$$

		Truth	
		H_0 is true (O.J. is innocent)	H_a is true (O.J. is guilty)
Jury's Verdict	Reject H_0 (O.J. guilty)	Type I error	✓
	Fail to Reject H_0	✓	Type II error

8. Large sample test for a population mean

(* If the population is normal and σ is known, the sample size does not matter.)

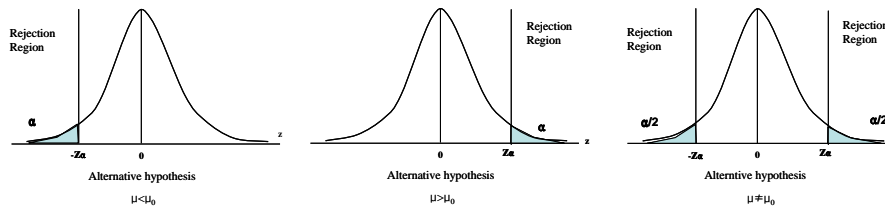
- The following test is appropriate for normal population with population standard deviation σ is known; or for any population when the sample size n is large ($n \geq 30$) (in the large sample scenario, we can substitute the sample standard deviation S for σ when σ is unknown.).
- **Null Hypothesis:** $H_0 : \mu = \mu_0$ (note: here μ represents the unknown population mean, and μ_0 is a given value believed to be the true value of μ (the original claim, belief, etc.).
- **Test Statistic:** The test statistic z is the z -score of the observed sample mean \bar{x} under the null hypothesis (when the null hypothesis is assumed to be true).

The formula is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \quad \text{when } \sigma \text{ is known}$$

or

$$z \approx \frac{\bar{x} - \mu_0}{S/\sqrt{n}}, \quad \text{when } \sigma \text{ is unknown}$$



- There are two ways to do the hypothesis testing: **the traditional method** and **the p-value method**. Please read the following flow-charts and plots for more details.

<i>Alternative hypothesis</i>	<i>Reject the null hypothesis if</i>
$\mu < \mu_0$	$z \leq -z_\alpha$
$\mu > \mu_0$	$z \geq z_\alpha$
$\mu \neq \mu_0$	$z \leq -z_{\alpha/2}$ OR $z \geq z_{\alpha/2}$

1. One-tailed test:

$$\beta = P\left(z < z_\alpha - \frac{|\mu_0 - \mu_a|}{\sigma_{\bar{y}}}\right); \quad \text{power} = 1 - \beta$$

2. Two-tailed test:

$$\beta \approx P\left(z < z_{\alpha/2} - \frac{|\mu_0 - \mu_a|}{\sigma_{\bar{y}}}\right); \quad \text{power} = 1 - \beta$$

EXAMPLE 5.6

Prospective salespeople for an encyclopedia company are now being offered a sales training program. Previous data indicate that the average number of sales per month for those who do not participate in the program is 33. To determine whether the training program is effective. A random sample of 35 new employees is given the sales training and then sent out into the field. One month later, the mean and standard deviation for the number of sets of encyclopedias sold are 35 and 8.4, respectively. Do the data present sufficient evidence to indicate that the training program enhances sales? Use $\alpha = .05$.

SOLUTION The five parts to our statistical test are as follows

$$H_0: \mu = 33$$

$$H_a: \mu > 33$$

T.S. :

$$z = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \approx \frac{35 - 33}{8.4/\sqrt{35}} = 1.41$$

R.R. : For $\alpha = 0.05$ we will reject the null hypothesis if $z > z_{.05} = 1.645$.

Conclusion: Since the observed value of z does not fall into the rejection region, we reserve judgment on accepting H_0 until we calculate β . That is, we conclude that there is insufficient evidence to reject the null hypothesis that persons on the sales program have the same mean number of sales per month as those not under the program.

EXAMPLE 5.7

Refer to Example 5.6. Suppose that the encyclopedia company thinks that the cost of financing the sales program will be offset by increased sales if those on the program average 38 sales per month. compute β for $\mu_a = 38$ and, based on the value of β , indicate whether you would accept the null hypothesis.

SOLUTION Using the computational formula for β with $\mu_0 = 33, \mu_a = 38$, and $\alpha = 0.05$, we have

$$\begin{aligned}\beta &= P\left[z < z_{.05} - \frac{|\mu_0 - \mu_a|}{s/\sqrt{n}}\right] = P\left[z < 1.645 - \frac{|33 - 38|}{8.4/\sqrt{35}}\right] \\ &= P[z < -1.88].\end{aligned}$$

The area corresponding to $z = 1.88$ in Table 2 of the Appendix is .0301. Hence,

$$\beta = .0301; \quad \text{power} = 1 - .0301 = 0.9699.$$

Because β is relatively small, we accept the null hypothesis and conclude that the training program has not increased the average sales per month above the point where increased sales would offset the cost of the training program.

9. Choosing the sample size for testing μ

Sample Size for a One-sided Test of μ

$$n = \frac{\sigma^2}{\Delta^2} (z_\alpha + z_\beta)^2$$

Note: If σ^2 is unknown, substitute an estimated value to get an approximate sample size.

Approximate Sample Size for a Two-sided test of $H_0 : \mu = \mu_0$.

$$n = \frac{\sigma^2}{\Delta^2} (z_{\alpha/2} + z_\beta)^2$$

Note: If σ^2 is unknown, substitute an estimated value to get an approximate sample size.

EXAMPLE 5.8

A cereal packager is concerned that one of its machines has a mean fill per package or more than 16 ounces, the labelled net weight. While this is not bad from a public relations standpoint, it could cost the packager a great deal of money. Previous experience suggests that the standard deviation of the package fill weights is approximately .225. For $H_0 : \mu = 16$ v.s. $H_a : \mu > 16$ with $\alpha = .05$, determine the sample size required to make $\beta = .01$ or less if the actual mean is 16.1 ounces or more. By putting this restriction on β , the packager is saying that it wants a very small probability of falsely accepting $H_0 : \mu = 16$, when in fact the actual mean is 16.1 ounces or more.

SOLUTION From previous data the fill weights have a standard deviation approximately equal to .225. The appropriate z -values, $z_{.05}$ and $z_{.01}$, for $\alpha = .05$ and $\beta = .01$ are 1.645 and 2.33, respectively. Using $\Delta = 16.1 - 16 = .1$, the require sample size is

$$n = \frac{(.225)^2(1.645 + 2.33)^2}{(.1)^2} = 79.99 \approx 80.$$

That is, the packager must obtain a random sample of $n = 80$ cartons to conduct this test under the specified conditions. Suppose that after obtaining the sample, the computed value of

$$z = \frac{\bar{y} - 16}{\sigma_{\bar{y}}}$$

does not fall in the rejection region. What is our conclusion? In similar situations in previous sections, our conclusion would have been that there was insufficient evidence to reject H_0 . Now, however, knowing that $\beta \leq .01$ when $\mu \geq 16.1$, we would feel safe in our conclusion to accept $H_0 : \mu = 16$. No further testing would be required.

10. Small sample test for a population mean when the population is normal and σ is unknown – the t-test

EXAMPLE 5.9

In a psychological depth-perception test, a random sample of $n = 14$ airline pilots were asked to judge the distance between two markers at the other end of a laboratory. The sample data (recorded in feet) are listed below.

2.7	2.4	1.9	2.6	2.4	1.9	2.3
2.2	2.5	2.3	1.8	2.5	2.0	2.2

use the sample data to place a 95% confidence interval on μ , the average recorded distance for this psychological test.

SOLUTION Before setting up a 95% confidence interval on μ , we must compute \bar{y} and s . You can verify that

$$\sum_i y_i = 31.70 \quad \text{and} \quad \sum_i y_i^2 = 72.79.$$

The sample mean, variance, and standard deviation are then

$$\bar{y} = \frac{\sum_i y_i}{14} = \frac{31.70}{14} = 2.26 \quad (1)$$

$$s^2 = \frac{1}{13} \left[72.79 - \frac{(31.7)^2}{14} \right] = .078 \quad (2)$$

$$s = \sqrt{.078} = .28. \quad (3)$$

Referring to Table 4 in the Appendix, the t -value corresponding to $\alpha = .025$ and $df = 13$ is 2.160. Hence, the 95% confidence interval is

$$\bar{y} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \quad \text{or} \quad 2.26 \pm \frac{2.160(.28)}{\sqrt{14}}.$$

which is the interval $2.26 \pm .16$, or 2.10 to 2.42. Thus, we are 95% confident that the interval from 2.10 to 2.42 will encompass the mean μ .

$$H_0: \mu = \mu_0$$

$$H_a:$$

1. $\mu > \mu_0$
2. $\mu < \mu_0$
3. $\mu \neq \mu_0$

T.S. :

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

R.R. : For a probability α of a Type I error and $df = n - 1$,

1. reject H_0 if $t > t_\alpha$
2. reject H_0 if $t < -t_\alpha$
3. reject H_0 if $|t| > t_{\alpha/2}$

EXAMPLE 5.10

A tire company guarantees that a particular brand of tire has a mean useful lifetime of 42,000 miles or more. A consumer test agency, wishing to verify this claim, observed $n = 10$ tires on a test wheel that simulated normal road conditions. The lifetimes (in thousands of miles) were as follows:

42 36 46 43 41 35 43 45 40 39

Use these data to determine whether there is sufficient evidence to contradict the manufactures claim. Set $\alpha = 0.05$.

SOLUTION The null and research hypotheses for this example are

$$H_0 : \mu = 42$$

and

$$H_a : \mu < 42.$$

Note that we are giving the manufacturer the benefit of the doubt by setting $\mu = 42$ for H_0 . Before setting up the test statistic, and rejection region, we must first compute the sample mean and standard deviation. You can verify that

$$\sum_i y_i = 410 \quad \text{and} \quad \sum_i y_i^2 = 16,926.$$

Then

$$\bar{y} = \frac{\sum_i y_i}{10} = \frac{410}{10} = 41.$$

Similarly, substituting into the shortcut formula for s^2 , we find

$$s^2 = \frac{1}{9} \left[\sum_i y_i^2 - \frac{(\sum_i y_i)^2}{10} \right] = 12.89 \quad (4)$$

$$s = \sqrt{12.89} = 3.59. \quad (5)$$

The test statistic, then, is

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{41 - 42}{3.59/\sqrt{10}} = -.88$$

and the rejection region is

$$\text{R.R.: Reject } H_0 \text{ if } t < -t_{.05}.$$

From Table 4 in the Appendix, the critical t -value with $\alpha = .05$ and $df = 9$ is 1.833, so $-t_{.05}$ is -1.833. Since the observed value of t is not less than -1.833, we have insufficient evidence to indicate that the mean lifetime of this brand of tires is less than 42,000 miles.