Statistics

Chapter 9

Agenda

- Population and Sample Review (from Chapter 8- Barron)
- Parameter Estimation
- Central Limit Theorem (Recall)
- Confidence Interval
- Hypothesis Testing
  - Z Test (tables based on left side of the bell curve)
  - T-Test (tables created based on the right side of the bell curve)

Population and Sample (Ch 8-Barron's book)

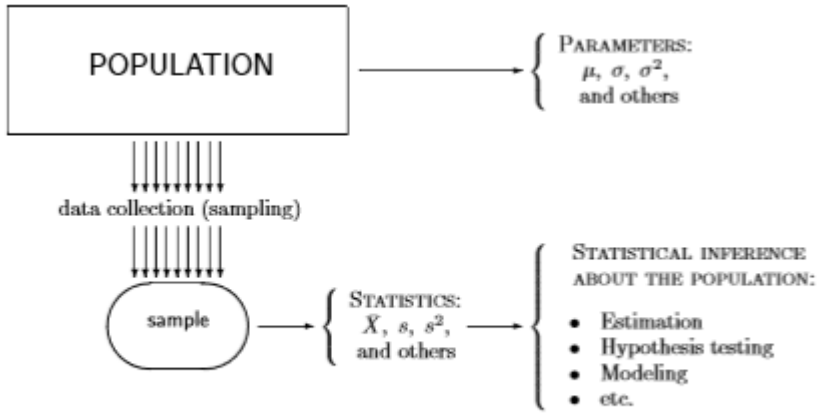

FIGURE 8.1: *Population parameters and sample statistics.*

8.1 Population and sample, parameters and statistics

Data collection is a crucially important step in Statistics. We use the collected and observed *sample* to make statements about a much larger set — the *population*.

DEFINITION 8.1

A population consists of all units of interest. Any numerical characteristic of a population is a **parameter**. **A sample** consists of observed units collected from the population. It is used to make statements about the population. Any function of a sample is called **statistic**.

Sampling and non-sampling errors

Sampling and non-sampling errors refer to any discrepancy between a collected sample and a whole population.

Sampling errors are caused by the mere fact that only a sample, a portion of a population, is observed. For most of reasonable statistical procedures, sampling errors decrease (and converge to zero) as the sample size increases.

Non-sampling errors are caused by inappropriate sampling schemes or wrong statistical techniques. Often no wise statistical techniques can rescue a poorly collected sample of data.

Source: Michael Baron's Probability and Statistics for Computer Scientists

Population and Sample Parameters Contd…

NOTATION			
	μ	=	population mean
	$\bar{X}$	=	sample mean, estimator of μ
	σ	=	population standard deviation
	s	=	sample standard deviation, estimator of σ
	σ^2	=	population variance
	s^2	=	sample variance, estimator of σ

Source: Michael Baron's Probability and Statistics for Computer Scientists

Standard Error of the Mean

- Std error of the mean SE = Std. Deviation of the sample distribution/Sqrt(n)
- SEM =s/sqrt(n)
- SEM decreases as n increases
- SEM measures how accurately the sample mean xbar estimates the actual population mean
- Std deviation (s) measures scatter of sample data around estimated sample mean x bar

Source: Michael Baron's Probability and Statistics for Computer Scientists

Confidence Intervals

DEFINITION 9.4

An interval $[a, b]$ is a $(1-\alpha)100\%$ **confidence interval** for the parameter θ if it contains the parameter with probability $(1-\alpha)$,

$$P\{a \leq \theta \leq b\} = 1 - \alpha.$$

The **coverage probability** $(1-\alpha)$ is also called a **confidence level**.

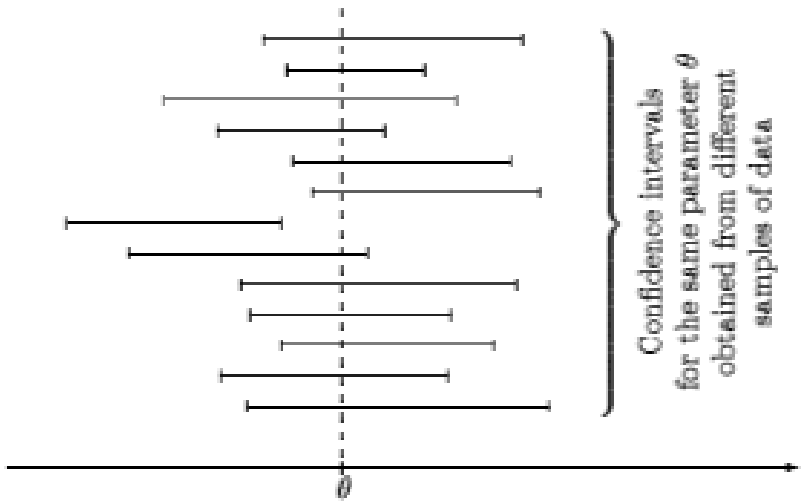

FIGURE 9.2: *Confidence intervals and coverage of parameter θ.*

Source: Michael Baron's Probability and Statistics for Computer Scientists

Confidence Interval Contd…

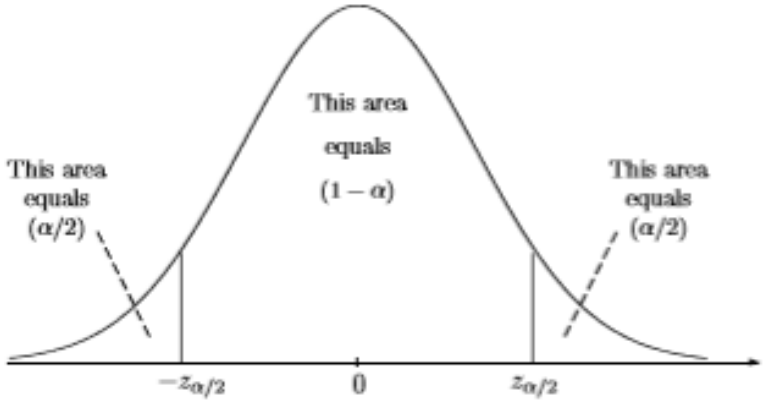

FIGURE 9.3: *Standard Normal quantiles* $\pm z_{\alpha/2}$ *and partition of the area under the density curve.*

Confidence interval, Normal distribution

If parameter θ has an unbiased, Normally distributed estimator $\hat{\theta}$, then

$$\hat{\theta} \pm z_{\alpha/2} \cdot \sigma(\hat{\theta}) = \left[\hat{\theta} - z_{\alpha/2} \cdot \sigma(\hat{\theta}),\ \hat{\theta} + z_{\alpha/2} \cdot \sigma(\hat{\theta})\right]$$

is a $(1-\alpha)100\%$ confidence interval for θ.

If the distribution of $\hat{\theta}$ is *approximately* Normal, we get an *approximately* $(1-\alpha)100\%$ confidence interval.

The problem is solved! We have obtained two numbers

$$a = \hat{\theta} - z_{\alpha/2} \cdot \sigma(\hat{\theta})$$
$$b = \hat{\theta} + z_{\alpha/2} \cdot \sigma(\hat{\theta})$$

such that

$$\boldsymbol{P}\{a \le \theta \le b\} = 1 - \alpha.$$

Source: Michael Baron's Probability and Statistics for Computer Scientists

Confidence Interval for the population Mean..

9.2.2 Confidence interval for the population mean

Let us construct a confidence interval for the population mean

$$\theta = \mu = \mathbf{E}(X).$$

Start with an estimator,

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

The rule (9.3) is applicable in two cases.

1. If a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ comes from Normal distribution, then $\bar{X}$ is also Normal, and rule (9.3) can be applied.

2. If a sample comes from *any* distribution, but the sample size n is large, then $\bar{X}$ has an approximately Normal distribution according to the Central Limit Theorem on p. 93. Then rule (9.3) gives an approximately $(1-\alpha)100\%$ confidence interval.

Source: Michael Baron's Probability and Statistics for Computer Scientists

Confidence Interval for the population Mean Contd..

In Section 8.2.1, we derived

$$\begin{aligned} \mathbf{E}(\bar{X}) &= \mu \qquad \text{(thus, it is an unbiased estimator);} \\ \sigma(\bar{X}) &= \sigma/\sqrt{n}. \end{aligned}$$

Then, (9.3) reduces to the following $(1-\alpha)100\%$ confidence interval for μ.

Confidence interval for the mean; σ is known

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (9.5)$$

Source: Michael Baron's Probability and Statistics for Computer Scientists

Problem

Example 9.13. Construct a 95% confidence interval for the population mean based on a sample of measurements

$$2.5,\ 7.4,\ 8.0,\ 4.5,\ 7.4,\ 9.2$$

if measurement errors have Normal distribution, and the measurement device guarantees a standard deviation of $\sigma = 2.2$.

Source: Michael Baron's Probability and Statistics for Computer Scientists

Confidence Interval for a population Mean

- A computer server processed 250 print jobs on average of 12.5 min/job. If the margin of the error is 1.7 min/job at 95% confidence, construct the confidence intervals

Source: Michael Baron’s Probability and Statistics for Computer Scientists

Another problem on CI

- An online airline reservation system reserves tickets 900 tickets average of 12 minutes/reservation. If the margin of the error is 4.1 min/reservation at 98% confidence, construct the confidence intervals

Source: Michael Baron's Probability and Statistics for Computer Scientists

Confidence Interval for the difference bet two means

Under the same conditions as in the previous section,

– Normal distribution of data or
– sufficiently large sample size,

we can construct a confidence interval for the *difference* between two means.

This problem arises when we compare two populations. It may be a comparison of two materials, two suppliers, two service providers, two communication channels, two labs, etc. From each population, a sample is collected (Figure 9.4),

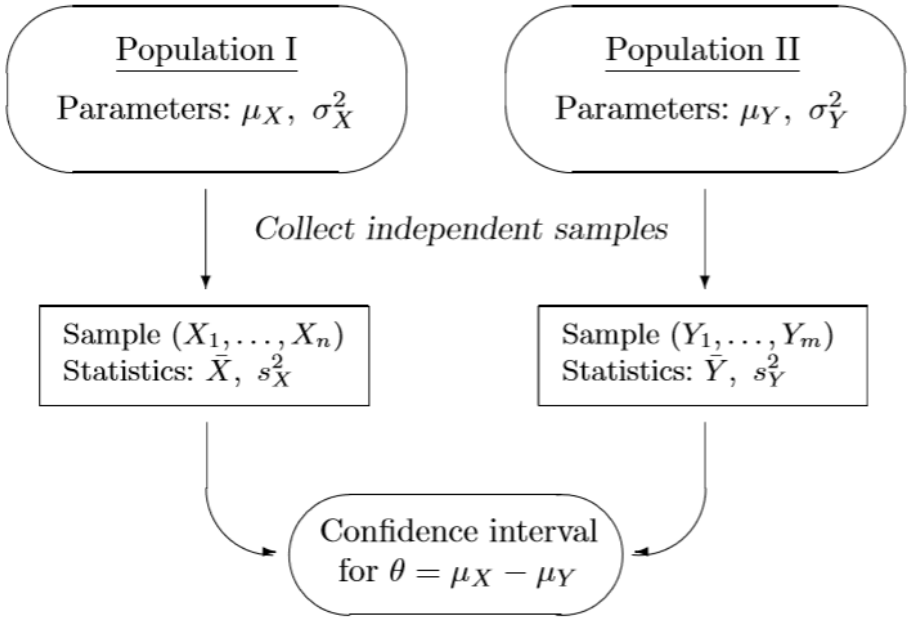

FIGURE 9.4: *Comparison of two populations.*

$\boldsymbol{X} = (X_1, \ldots, X_n)$ from one population,
$\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ from the other population.

Suppose that the two samples are collected **independently** of each other.

To construct a confidence interval for the difference between population means

$$\theta = \mu_X - \mu_Y,$$

we complete the usual steps (a)–(e) below.

Source: Michael Baron's Probability and Statistics for Computer Scientists

Confidence Interval for the difference bet two means Contd…

(a) Propose an estimator of θ,

$$\hat{\theta} = \bar{X} - \bar{Y}.$$

It is natural to come up with this estimator because $\bar{X}$ estimates μ_X and $\bar{Y}$ estimates μ_Y.

(b) Check that $\hat{\theta}$ is unbiased. Indeed,

$$\mathbf{E}(\hat{\theta}) = \mathbf{E}\left(\bar{X} - \bar{Y}\right) = \mathbf{E}\left(\bar{X}\right) - \mathbf{E}\left(\bar{Y}\right) = \mu_X - \mu_Y = \theta.$$

(c) Check that $\hat{\theta}$ has a Normal or approximately Normal distribution. This is true if the observations are Normal or *both* sample sizes m and n are large.

(d) Find the standard error of $\hat{\theta}$ (using independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$),

$$\sigma(\hat{\theta}) = \sqrt{\mathrm{Var}\left(\bar{X} - \bar{Y}\right)} = \sqrt{\mathrm{Var}\left(\bar{X}\right) + \mathrm{Var}\left(\bar{Y}\right)} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

(e) Find quantiles $\pm z_{\alpha/2}$ and compute the confidence interval according to (9.3). This results in the following formula.

Confidence interval for the difference of means; known standard deviations

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \qquad (9.6)$$

Source: Michael Baron's Probability and Statistics for Computer Scientists

Margin of Error -FYI

- The term $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is also called margin of error

Source: Michael Baron's Probability and Statistics for Computer Scientists

Example 9.14 Review

Example 9.14 (EFFECT OF AN UPGRADE). A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 7.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. Construct a 90% confidence interval showing how much the mean running time reduced due to the hardware upgrade.

Source: Michael Baron's Probability and Statistics for Computer Scientists

Calculating sample size

9.2.4 Selection of a sample size

Formula (9.3) describes a confidence interval as

$$\text{center} \pm \text{margin}$$

where

$$\begin{aligned} \text{center} &= \hat{\theta}, \\ \text{margin} &= z_{\alpha/2} \cdot \sigma(\hat{\theta}). \end{aligned}$$

9.2.5 Estimating means with a given precision

When we estimate a population mean, the margin of error is

$$\text{margin} = z_{\alpha/2} \cdot \sigma/\sqrt{n}.$$

Solving inequality (9.7) for n results in the following rule.

Sample size for a given precision

In order to attain a margin of error Δ for estimating a population mean with a confidence level $(1-\alpha)$,

a sample of size $n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{\Delta}\right)^2$ is required. (9.8)

Source: Michael Baron's Probability and Statistics for Computer Scientists

Example 9.13 Review

Example 9.15. In Example 9.13, we constructed a 95% confidence with the center 6.50 and margin 1.76 based on a sample of size 6. Now, that was too wide, right? How large a sample do we need to estimate the population mean with a margin of at most 0.4 units with 95% confidence?

Source: Michael Baron's Probability and Statistics for Computer Scientists

How about Confidence intervals for unknown Std Deviation

9.3 Unknown standard deviation

A rather heavy condition was assumed when we constructed all the confidence intervals. We assumed a *known standard deviation* σ and used it in all the derived formulas.

Sometimes this assumption is perfectly valid. We may know the variance from a large archive of historical data, or it may be given as precision of a measuring device.

Much more often, however, the population variance is unknown. We'll then estimate it from data and see if we can still apply methods of the previous section.

Two broad situations will be considered:

– large samples from any distribution,

– samples of any size from a Normal distribution.

In the only remaining case, a small non-Normal sample, a confidence interval will be constructed by special methods. A popular modern approach called *bootstrap* is discussed in Section 10.3.3.

Source: Michael Baron's Probability and Statistics for Computer Scientists

Unknown Std Deviation

9.3.1 Large samples

A large sample should produce a rather accurate estimator of a variance. We can then replace the true standard error $\sigma(\hat{\theta})$ in (9.3) by its estimator $s(\hat{\theta})$, and obtain an approximate confidence interval

$$\hat{\theta} \pm z_{\alpha/2} \cdot s(\hat{\theta}).$$

Source: Michael Baron's Probability and Statistics for Computer Scientists

Example 9.16 review

Example 9.16 (DELAYS AT NODES). Internet connections are often slowed by delays at nodes. Let us determine if the delay time increases during heavy-volume times.

Five hundred packets are sent through the same network between 5 pm and 6 pm (sample $\boldsymbol{X}$), and three hundred packets are sent between 10 pm and 11 pm (sample $\boldsymbol{Y}$). The early sample has a mean delay time of 0.8 sec with a standard deviation of 0.1 sec whereas the second sample has a mean delay time of 0.5 sec with a standard deviation of 0.08 sec. Construct a 99.5% confidence interval for the difference between the mean delay times.

Solution. We have $n = 500$, $\bar{X} = 0.8$, $s_X = 0.1$; $m = 300$, $\bar{Y} = 0.5$, $s_Y = 0.08$. Large sample sizes allow us to replace unknown population standard deviations by their estimates and use an approximately Normal distribution of sample means.

For a confidence level of $1 - \alpha = 0.995$, we need

$$z_{\alpha/2} = z_{0.0025} = q_{0.9975}.$$

Look for the *probability* 0.9975 in the body of Table A4 and find the corresponding value of z,

$$z_{0.0025} = 2.81.$$

Then, a 99.5% confidence interval for the difference of mean execution times is

$$\begin{aligned} \bar{X} - \bar{Y} &\pm z_{0.0025}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} = (0.8 - 0.5) \pm (2.81)\sqrt{\frac{(0.1)^2}{500} + \frac{(0.08)^2}{300}} \\ &= \underline{0.3 \pm 0.018} \text{ or } \underline{[0.282, 0.318]}. \end{aligned}$$

Source: Michael Baron's Probability and Statistics for Computer Scientists ◊

What about smaller samples and unknown standard Deviation

9.3.4 Small samples: Student's t distribution

Having a small sample, we can no longer pretend that a sample standard deviation s is an accurate estimator of the population standard deviation σ. Then, how should we adjust the confidence interval when we replace σ by s, or more generally, when we replace the standard error $\sigma(\hat{\theta})$ by its estimator $s(\hat{\theta})$?

A famous solution was proposed by *William Gosset* (1876–1937), known by his pseudonym *Student*. Working for the Irish brewery Guinness, he derived the T-distribution for the quality control problems in brewing.

Source: Michael Baron's Probability and Statistics for Computer Scientists

T-Distribution Contd…

So, using T-distribution instead of Standard Normal and estimated standard error instead of the unknown true one, we obtain the confidence interval for the population mean.

Confidence interval for the mean; σ is unknown

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \qquad (9.9)$$

where $t_{\alpha/2}$ is a critical value from T-distribution with $n-1$ degrees of freedom

Recollect Population Confidence Intetvals

In Section 8.2.1, we derived

$$\begin{aligned} \mathbf{E}(\bar{X}) &= \mu \qquad \text{(thus, it is an unbiased estimator);} \\ \sigma(\bar{X}) &= \sigma/\sqrt{n}. \end{aligned}$$

Then, (9.3) reduces to the following $(1-\alpha)100\%$ confidence interval for μ.

Confidence interval for the mean; σ is known

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad (9.5)$$

Source: Michael Baron's Probability and Statistics for Computer Scientists

Example 9.19 Review

Example 9.19 (UNAUTHORIZED USE OF A COMPUTER ACCOUNT). If an unauthorized person accesses a computer account with the correct username and password (stolen or cracked), can this intrusion be detected? Recently, a number of methods have been proposed to detect such unauthorized use. The time between keystrokes, the time a key is depressed, the frequency of various keywords are measured and compared with those of the account owner. If there are significant differences, an intruder is detected.

The following times between keystrokes were recorded when a user typed the username and password:

.24, .22, .26, .34, .35, .32, .33, .29, .19, .36, .30, .15, .17, .28, .38, .40, .37, .27 seconds

As the first step in detecting an intrusion, let's construct a 99% confidence interval for the mean time between keystrokes assuming Normal distribution of these times.

Source: Michael Baron's Probability and Statistics for Computer Scientists

Home work Reading

- Comparison of Two populations with unknown Variances
  - Case 1 : Equal Variances
  - Case 2 : Unequal Variances