Probability and Statistics course

Introduction

Gabor Vigh

TTK Department of Probability Theory and Statistics
ELTE

November 15, 2025

Week 10

Introduction

When we report an estimator $\hat{\theta}$ of a population parameter θ, we know that most likely

$$\hat{\theta} \neq \theta$$

due to a sampling error. We realize that we have estimated θ up to some error. Likewise, nobody understands the internet connection of 11 megabytes per second as exactly 11 megabytes going through the network every second, and nobody takes a meteorological forecast as the promise of exactly the predicted temperature.

Then how much can we trust the reported estimator? How far can it be from the actual parameter of interest? What is the probability that it will be reasonably close? And if we observed an estimator $\hat{\theta}$, then what can the actual parameter θ be?

To answer these questions, statisticians use **confidence intervals**, which contain parameter values that deserve some confidence, given the observed data.

Confidence Interval

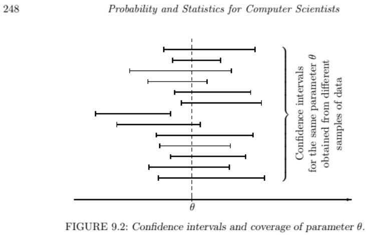
Definition

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**.

Confidence Intervals



248 *Probability and Statistics for Computer Scientists*

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

Figure: Confidence intervals and coverage of parameter θ

Scenario: A sample of students was taken to estimate the **true mean amount of sleep** (μ) all students at a school get per night.

- **Confidence Level:** 95%
- **Calculated Confidence Interval (CI):** $[7.5, 8.5]$ hours

The Common Mistake (Incorrect):"There is a 95% chance that the true mean μ falls between 7.5 and 8.5 hours." (The true mean is a fixed value, not a random variable, so the probability it is in the fixed interval is either 0 or 1.)
The Correct Interpretation (Shorthand): The most practical and commonly used correct interpretation is:
"We are 95% confident that the true mean amount of sleep (μ) is between 7.5 and 8.5 hours."
The Most Correct Interpretation (Technical): The 95% confidence level represents the method's reliability:
"If we repeatedly took samples and calculated confidence intervals, 95% of those intervals would contain the true population mean μ."

Construction of confidence intervals: a general method

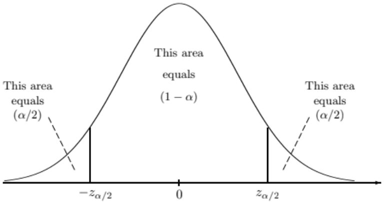


Figure: he standard normal distribution, denoted $Z \sim N(0,1)$, is centered at $\mu = 0$ with a standard deviation $\sigma = 1$.

We start by estimating a population parameter θ. Assume there is an **unbiased estimator** $\hat{\theta}$ that has a **Normal distribution**. When we standardize this estimator, we obtain a **Standard Normal variable** $Z = \frac{\hat{\theta}-E[\hat{\theta}]}{\mathrm{SD}(\hat{\theta})} = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ This variable $Z = \frac{\hat{\theta}-\theta}{\sigma(\hat{\theta})}$ falls between the Standard Normal quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$, denoted by $-z_{\alpha/2} = q_{\alpha/2}$ and $z_{\alpha/2} = q_{1-\alpha/2}$ with probability $(1-\alpha)$, as you can see in figure above. Then, we can write the probability statement for the standardized variable Z:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sigma(\hat{\theta})} \leq z_{\alpha/2}\right) = 1-\alpha.$$

Solving the inequality inside the parentheses for the parameter θ, we get the following probability statement:

$$P\left\{\hat{\theta} - z_{\alpha/2}\cdot\sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{\alpha/2}\cdot\sigma(\hat{\theta})\right\} = 1-\alpha.$$

Confidence interval for the population mean

Let us construct a confidence interval for the population mean $\theta = \mu = E(X)$. Start with an estimator,$\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We can assume normality for the estimator

1. **Normal Population.** If a sample $X = (X_1, \dots, X_n)$ comes from a Normal distribution, then $\overline{X}$ is also Normal, and rule (9.3) can be applied.
2. **Large Sample (CLT).** If a sample comes from any distribution, but the sample size n is large, then $\overline{X}$ has an **approximately Normal distribution** according to the **Central Limit Theorem**.

$E(\overline{X}) = \mu$ (thus, it is an unbiased estimator); $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$.
This reduces to the following $(1-\alpha)100\%$ confidence interval for μ.

$$\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Example

Example

Let X_1, X_2, X_3, X_4 be a sample from the $N(\mu, 2^2)$ distribution, with the following values: $14.8; 12.2; 16.8; 11.1$.
a) Give a 95% confidence interval for μ!
b) How many sample elements are needed if we intend the interval to be shorter than 1.6?
c) How does the problem change if we do not know the standard deviation? (later)

Confidence interval for the difference between two means

Under the same conditions: 1) Normal distribution of data or 2) sufficiently large sample size. To construct a confidence interval for the difference between population means$\theta = \mu_X - \mu_Y$,, we complete the usual steps (a)–(e) below.

(a) Propose an estimator of θ: $\hat{\theta} = \overline{X} - \overline{Y}$. It is natural to come up with this estimator because $\overline{X}$ estimates μ_X and $\overline{Y}$ estimates μ_Y.

(b) Check that $\hat{\theta}$ is unbiased. Indeed, $E(\hat{\theta}) = E[\overline{X} - \overline{Y}] = E[\overline{X}] - E[\overline{Y}] = \mu_X - \mu_Y = \theta$.

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

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

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

(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

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

Example

Example

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.

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. Two broad situations will be considered:

1. large samples from any distribution,
2. samples of any size from a Normal distribution,
3. small non-Normal sample (won't be covered here, you can take a look at [**?**])

Large samples

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

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

Example

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 X), and three hundred packets are sent between 10 pm and 11 pm (sample 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.

Confidence intervals for proportions

In particular, we surely don't know the variance when we estimate a population proportion.

Definition

We assume a subpopulation A of items that have a certain attribute. By the population proportion we mean the probability

$$p = P\{i \in A\}$$

for a randomly selected item i to have this attribute. A sample proportion

$$\hat{p} = \frac{\text{number of sampled items from } A}{n}$$

is used to estimate p.

Confidence intervals for proportions

Let us use the indicator variables $X_i = \begin{cases} 1 & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases}$. Each X_i has Bernoulli distribution with parameter p. In particular, $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$. Also, the sample proportion $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is nothing but a sample mean of X_i.

Therefore, $E(\hat{p}) = p$ and $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$, as we know from properties of sample means. We conclude that

1. a sample proportion $\hat{p}$ is unbiased for the population proportion p;
2. it has approximately Normal distribution for large samples, because it has a form of a sample mean (by the Central Limit Theorem);
3. when we construct a confidence interval for p, we do not know the standard deviation $\text{Std}(\hat{p})$.

Thus, we estimate the unknown standard error $\sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$ by the estimated standard error $s(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and use it in the general formula $\hat{p} \pm z_{\alpha/2} \cdot s(\hat{p})$ to construct an approximate $(1-\alpha)100\%$ confidence interval.

Confidence interval for a population proportion p:

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

Confidence interval for the difference between two proportions

Similarly, we can construct a confidence interval for the difference between two proportions. Summarizing the two-sample case, we have:

$$\begin{aligned}
\text{Parameter of interest:}\quad & \theta = p_1 - p_2 \\
\text{Estimated by:}\quad & \hat{\theta} = \hat{p}_1 - \hat{p}_2 \\
\text{True standard error:}\quad & \sigma(\hat{\theta}) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\
\text{Estimated standard error:}\quad & s(\hat{\theta}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}
\end{aligned}$$

Confidence interval for the difference of proportions $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example

candidate prepares for the local elections. During his campaign, 42 out of 70 randomly selected people in town A and 59 out of 100 randomly selected people in town B showed they would vote for this candidate. Estimate the difference in support that this candidate is getting in towns A and B with 95% confidence. Can we state affirmatively that the candidate gets a stronger support in town A?

Small Samples Student t distribution

If m and σ are both unknown parameters, then the following interval is an $1-\alpha$ confidence interval for m (that is, it contains m with probability at least $1-\alpha$):

$$\left(\overline{X} - t_{n-1,1-\alpha/2} \cdot \frac{s_n^*}{\sqrt{n}}, \overline{X} + t_{n-1,1-\alpha/2} \cdot \frac{s_n^*}{\sqrt{n}}\right).$$

Here $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$-quantile of the t-distribution with $n-1$ degrees of freedom .

Why t distribution?

Why We Use the t-Distribution When σ is Unknown Then constructing a confidence interval for the population mean (m), we start with the standardized variable Z:

$$Z = \frac{\overline{X} - m}{\sigma/\sqrt{n}}, \quad \text{where } Z \sim N(0,1) \text{ when } \sigma \text{ is known.}$$

The Problem: Estimating σ When the population standard deviation (σ) is unknown, we must use the sample standard deviation (s_n^*) as an estimator. Substituting s_n^* for σ introduces additional uncertainty and variability. The resulting variable is denoted T:

$$T = \frac{\overline{X} - m}{s_n^*/\sqrt{n}}$$

The Solution: Student's t-Distribution Because s_n^* is itself a random variable, T no longer follows the Standard Normal (Z) distribution. Instead, T follows the **Student's t-distribution** with $\text{df} = n - 1$ degrees of freedom.

$$T \sim t_{n-1}$$

- The t-distribution has **heavier tails** than the Standard Normal distribution, correctly reflecting the greater uncertainty due to the estimation of σ.
- The shape of the t-distribution depends on the degrees of freedom ($\text{df} = n - 1$).
- As the sample size n increases (i.e., $\text{df} \to \infty$), the t-distribution converges to the Standard Normal distribution, so $t_{\alpha/2} \to z_{\alpha/2}$.

In summary, the t-distribution is used because it correctly accounts for the extra error or uncertainty created when we use the sample standard deviation (s_n^*) to estimate the true population standard deviation (σ).

Student t distribution

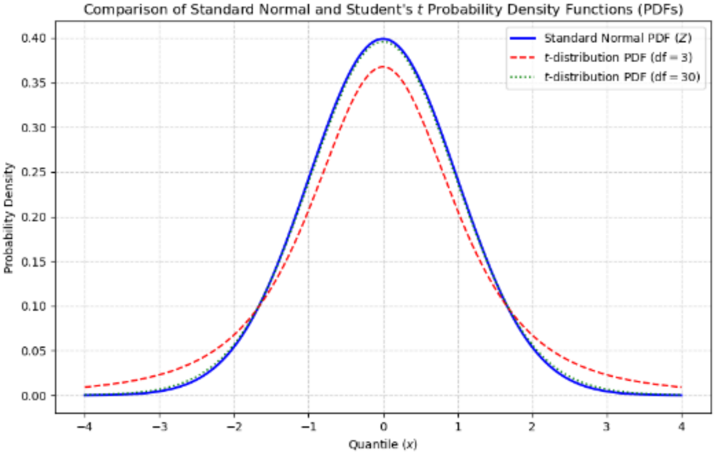


Figure: comparison student T and Z

Small samples: Student's t distribution

Example

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, and 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 (in seconds) were recorded when a user typed the username and password:

$$0.24, 0.22, 0.26, 0.34, 0.35, 0.32, 0.33, 0.29, 0.19, 0.36, 0.30, 0.15, 0.17, 0.28, 0.38, 0.40, 0.37, 0.27$$

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

Comparison of two populations with unknown variances - Case I

Two important cases need to be considered here. In one case, there exists an exact and simple solution based on T-distribution. The other case suddenly appears to be a famous Behrens-Fisher problem, where no exact solution exists, and only approximations are available.

Case I: Equal variances Suppose there are reasons to assume that the two populations have equal variances, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.For example, two sets of data are collected with the same measurement device, thus, measurements have different means but the same precision. In this case, there is only one variance σ^2 to estimate instead of two. We should use both samples X and Y to estimate their common variance. This estimator of σ^2 is called a **pooled sample variance**, and it is computed as

$$s_p^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2} = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}.$$

Substituting this variance estimator to previous formula for σ_X^2 and σ_Y^2, we get the following confidence interval.

$$(\bar{X} - \bar{Y}) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

where s_p is the pooled standard deviation, and $t_{\alpha/2}$ is a critical value from T-distribution with $(n+m-2)$ degrees of freedom.

Example - Case I

Example

CD writing is energy consuming; therefore, it affects the battery lifetime on laptops. To estimate the effect of CD writing, 30 users are asked to work on their laptops until the “low battery” sign comes on. Eighteen users without a CD writer worked an average of 5.3 hours with a standard deviation of 1.4 hours. The other twelve, who used their CD writer, worked an average of 4.8 hours with a standard deviation of 1.6 hours. Assuming Normal distributions with equal population variances ($\sigma_X^2 = \sigma_Y^2$), construct a 95% confidence interval for the battery life reduction caused by CD writing.

Case II - Unequal variances

The most difficult case is when both variances are unknown and unequal. Confidence estimation of $\mu_X - \mu_Y$ in this case is known as the **Behrens-Fisher problem**. Certainly, we can replace unknown variances σ_X^2, σ_Y^2 by their estimates s_X^2, s_Y^2 and form a T-ratio

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}.$$

However, it won't have a T-distribution.
An approximate solution was proposed: the method of moments to estimate degrees of freedom ν of a T-distribution that is "closest" to this T-ratio. This number depends on unknown variances. Estimating them by sample variances, he obtained the formula that is now known as the **Satterthwaite approximation**,

$$\nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}}. \quad (9.12)$$

This number of degrees of freedom often appears non-integer and we need to find the closest integer to get degree of freedom

Confidence interval for the difference of means; unequal, unknown standard deviations

$$\bar{X} - \bar{Y} \pm t_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

Example for Case II

Example

An account on server A is more expensive than an account on server B. However, server A is faster. To see if it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 30 times on server A and 20 times on server B with the following results:

	Server A (Population 1)	**Server B (Population 2)**
Sample Size (n_1, n_2)	$n_1 = 30$	$n_2 = 20$
Sample Mean $(\bar{X}_1, \bar{X}_2)$	$\bar{X}_1 = 6.7$ min	$\bar{X}_2 = 7.5$ min
Sample Standard Deviation (s_1, s_2)	$s_1 = 0.6$ min	$s_2 = 1.2$ min

Construct a 95% confidence interval for the difference $\mu_1 - \mu_2$ between the mean execution times on server A and server B, assuming that the observed times are approximately Normal.