Probability and Statistics course

Introduction

Gabor Vigh

TTK Department of Probability Theory and Statistics
ELTE

November 9, 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

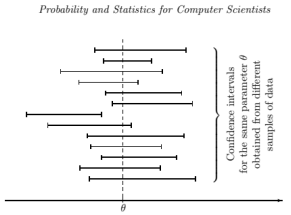

FIGURE 9.2: *Confidence intervals and coverage of parameter θ.*

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

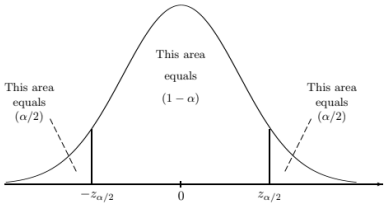

Figure: he standard normal distribution, denoted $Z \sim N(0,1)$, is centered at $\mu = 0$ with a standard deviation $\sigma = 1$.

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** $Z = \frac{\hat{\theta} - E[\hat{\theta}]}{\text{SD}(\hat{\theta})} = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$ 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

Proposition. Let $X_1, X_2, \ldots, X_n$ be independent random variables with normal distribution $N(m, \sigma^2)$. If we assume that σ is known, then the following interval is an $1-\alpha$ confidence interval for m (that is, it contains m with probability at least $1-\alpha$):

$$\left(\overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

Here $z_{1-\alpha/2}$ is the $1-\alpha/2$-quantile of the standard normal 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

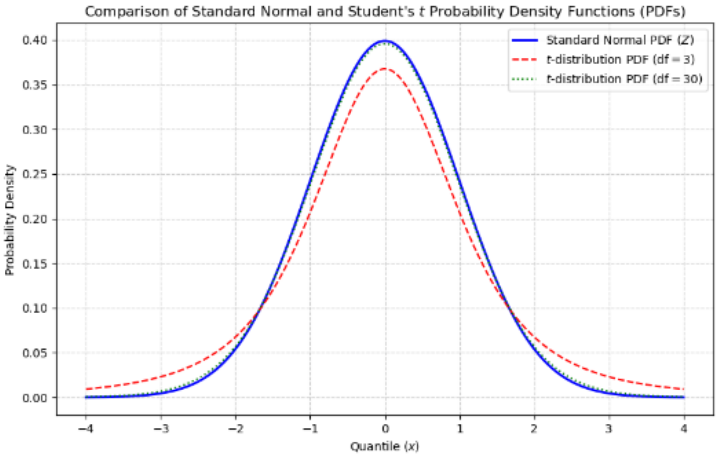

Figure: comparison student T and Z

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?