

The Relevance of t -Statistics for Small Sample Sizes

An Introductory Class to Higher Statistics

Xinyu Chen

December 6, 2024



Outline

Answering a lot questions, e.g.,

- ❶ How was t -statistic developed?
- ❷ Normal distribution vs. student t -distribution?
- ❸ What is t -statistic?
- ❹ How to calculate a t -test?
- ❺ What are the hypotheses and the assumptions?
- ❻ How to interpret results?

Development

The problems of **small sample sizes**

1876

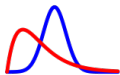
t-distribution as
a posterior distribution



F. R. Helmert



J. Lüroth



(Source: Wiki)

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

1895

t-distribution as Pearson
type IV distribution



K. Pearson



(Source: Wiki)

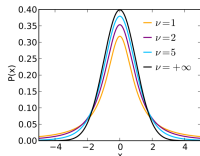
Diagram of the Pearson system

1908

Student *t*-distribution
"The Probable Error of a Mean" (Biometrika)



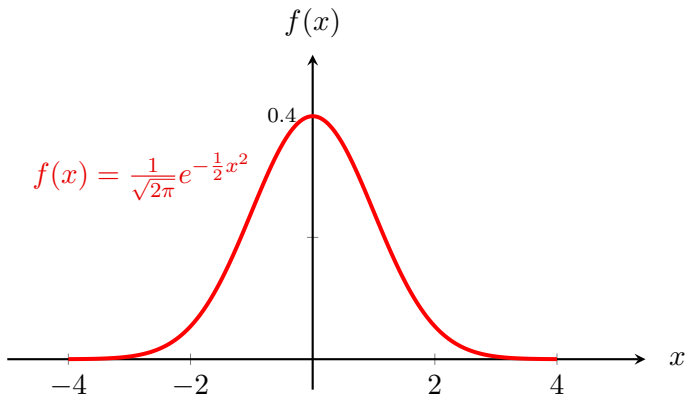
W. S. Gosset



(Source: Wiki)

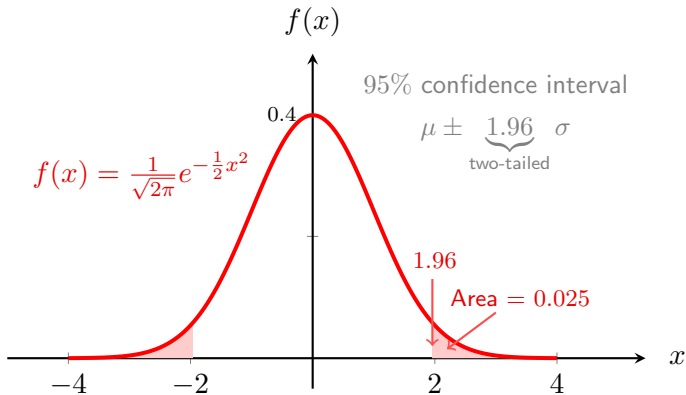
Probability density function

Revisiting Normal Distribution



Probability density function of the standard normal distribution

Revisiting Normal Distribution



Probability density function of the standard normal distribution

Connecting with Hypothesis Test

-

Implementing z -Test

Problem Statement

A company claims that the average daily energy consumption of households is 30 kWh with a population standard deviation of 5 kWh. A random sample of 40 households has an average daily energy consumption of 32 kWh. Conduct a two-tailed hypothesis test at a 95% confidence interval to determine if the sample provides sufficient evidence to reject the company's claim.

Implementing z -Test

Problem Statement

A company claims that the average daily energy consumption of households is 30 kWh with a population standard deviation of 5 kWh. A random sample of 40 households has an average daily energy consumption of 32 kWh. Conduct a two-tailed hypothesis test at a 95% confidence interval to determine if the sample provides sufficient evidence to reject the company's claim.

Steps:

❶ Formulate Hypotheses

- Null Hypothesis (H_0): The population mean is $\mu = 30$ kWh.
- Alternative Hypothesis (H_a): The population mean is not $\mu = 30$ kWh ($\mu \neq 30$).

❷ Use the z -test formula since the population standard deviation is known:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{32 - 30}{5/\sqrt{40}} = \frac{2}{5/6.32} = \frac{2}{0.79} \approx 2.53$$

- $\bar{x} = 32$ (sample mean)
- $\mu = 30$ (population mean)
- $n = 40$ (sample size)
- $\sigma = 5$ (population standard deviation)

Implementing z -Test

Problem Statement

A company claims that the average daily energy consumption of households is 30 kWh with a population standard deviation of 5 kWh. A random sample of 40 households has an average daily energy consumption of 32 kWh. Conduct a two-tailed hypothesis test at a 95% confidence interval to determine if the sample provides sufficient evidence to reject the company's claim.

Steps:

- ② Use the z -test formula since the population standard deviation is known:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{32 - 30}{5/\sqrt{40}} = \frac{2}{5/6.32} = \frac{2}{0.79} \approx 2.53$$

- ③ Decision rule at a 95% confidence interval

- Reject H_0 if $|z| > 1.96$.
- Otherwise, fail to reject H_0 .

- ④ Interpretation

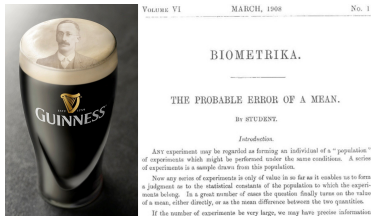
- The test statistic $|z| = 2.53 > 1.96$ (exceeding the critical value).
- Thus, we reject the null hypothesis.
- The sample provides sufficient evidence to conclude that the average daily energy consumption is not 30 kWh.

Student t -Distribution

- Probability density function:

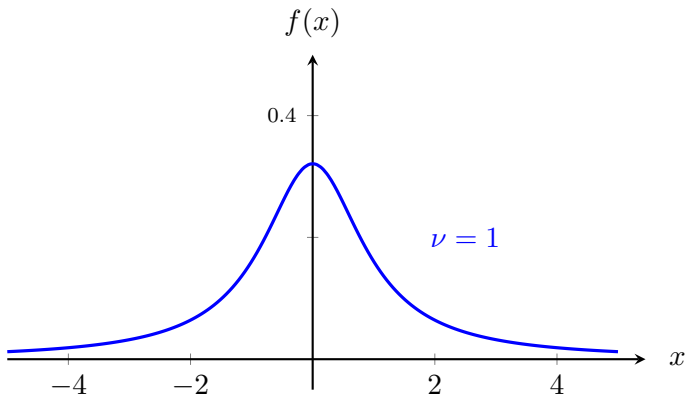
$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- $x \in \mathbb{R}$: The random variable
- $\nu \in \mathbb{Z}^+$: Degrees of freedom
- $\Gamma(\cdot)$: The Gamma function



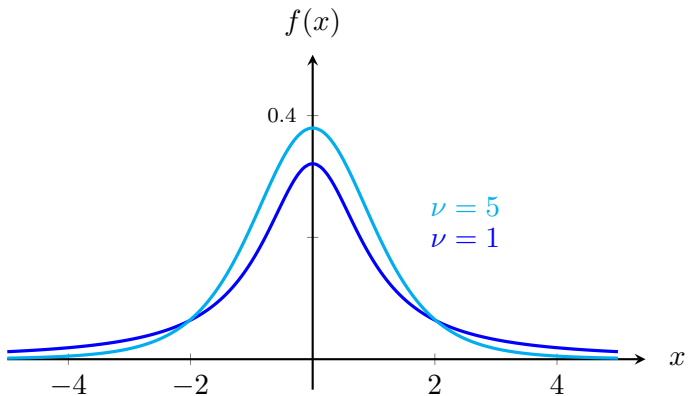
Gosset'1908 (known as "student")

Student t -Distribution



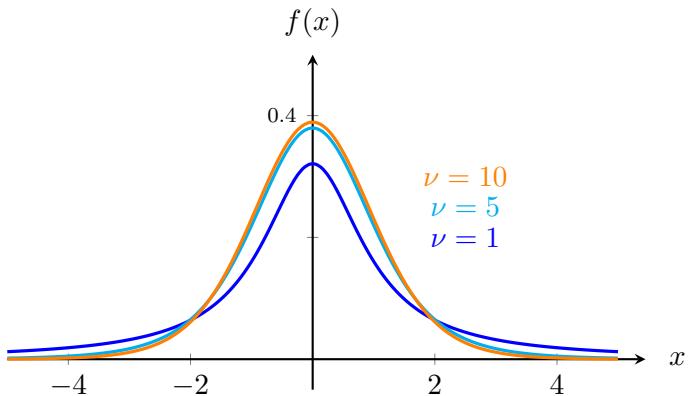
Student t -distribution of ν degrees of freedom

Student t -Distribution



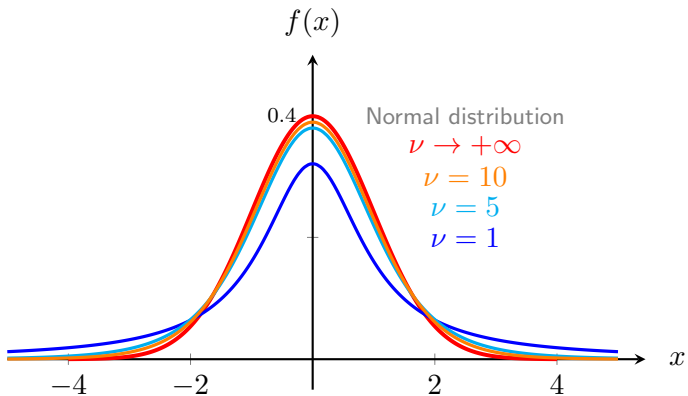
Student t -distribution of ν degrees of freedom

Student t -Distribution



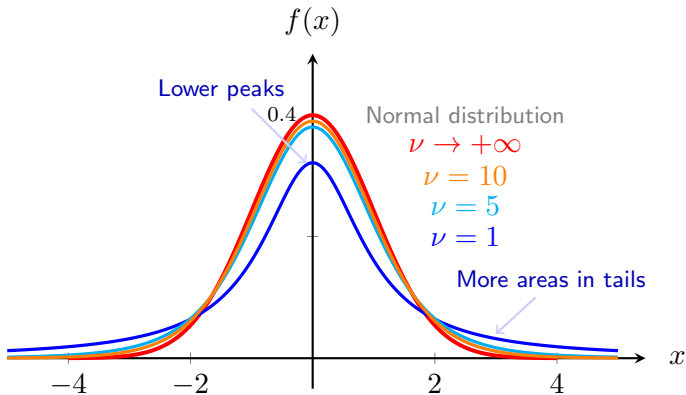
Student t -distribution of ν degrees of freedom

Student t -Distribution



Student t -distribution of ν degrees of freedom

Student t -Distribution



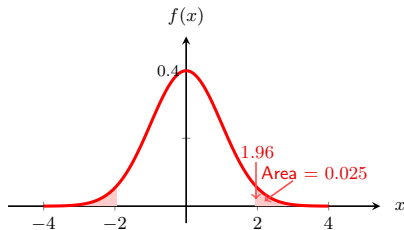
Student t -distribution of ν degrees of freedom

95% Confidence Interval

For the population mean μ (given) and standard deviation σ (given or not?)

- If **population standard deviation σ** is known

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$



Standard normal distribution

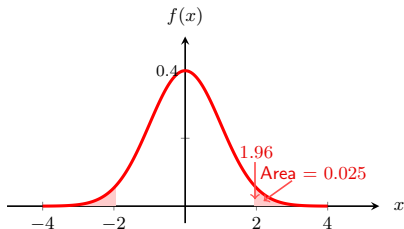
95% Confidence Interval

For the population mean μ (given) and standard deviation σ (given or not?)

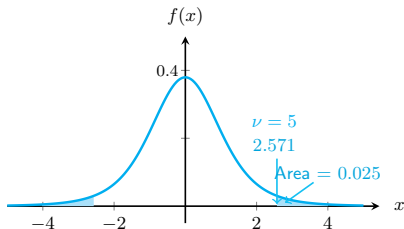
- If **population standard deviation σ** is known
- If σ is unknown, using **sample standard deviation s** instead

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} \pm ? \times \frac{s}{\sqrt{n}}$$



Standard normal distribution



Student t -distribution

- **Heavier tail** in student t -distribution ($\nu = n - 1$ degrees of freedom) is important for small sample size n

- Formula of t -statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- μ population mean
 - \bar{x} sample mean
 - s sample standard deviation
 - n sample size (usually small value)
- The t -statistic quantifies the difference relative to variability in the data.
- (Interpretation) A high absolute value of t (larger than the critical value from the t -table) suggests a statistically significant difference.
- The problem of small sample size!

Implementing t -Test

Problem Statement

A company claims that the average daily energy consumption of households is 30 kWh. A random sample of 6 households has an average daily energy consumption of 32 kWh, with a sample standard deviation of 6 kWh. Conduct a two-tailed hypothesis test at a 95% confidence interval to determine if the sample provides sufficient evidence to reject the company's claim.

Implementing t -Test

Problem Statement

A company claims that the average daily energy consumption of households is 30 kWh. A random sample of 6 households has an average daily energy consumption of 32 kWh, with a sample standard deviation of 6 kWh. Conduct a two-tailed hypothesis test at a 95% confidence interval to determine if the sample provides sufficient evidence to reject the company's claim.

Steps:

❶ Formulate Hypotheses

- Null Hypothesis (H_0): The population mean is $\mu = 30$ kWh.
- Alternative Hypothesis (H_a): The population mean is not $\mu = 30$ kWh ($\mu \neq 30$).

❷ Use the t -test formula since the population standard deviation is not known:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{32 - 30}{6/\sqrt{6}} = \frac{2}{6/2.449} \approx 0.816$$

- $\bar{x} = 32$ (sample mean)
- $s = 6$ (sample standard deviation)
- $n = 6$ (sample size)
- $\sigma = 30$ (population mean)

t-Table

Small sample sizes

- Degrees of freedom for a *t*-test:

$$\nu = \underbrace{n}_{\text{sample size}} - 1 = 6 - 1 = 5$$

- t*-distributions with ν degrees of freedom at a 95% confidence interval (two-tailed)

$\nu = 1$	$\nu = 5$	$\nu = 10$	$\nu \rightarrow +\infty$
12.706	2.571	2.228	1.960

- The critical *t*-value

$$t_{\nu, (1-0.95)/2} = t_{5, 0.025} = 2.571$$

Implementing t -Test

Problem Statement

A company claims that the average daily energy consumption of households is 30 kWh. A random sample of 6 households has an average daily energy consumption of 32 kWh, with a sample standard deviation of 6 kWh. Conduct a two-tailed hypothesis test at a 95% confidence interval to determine if the sample provides sufficient evidence to reject the company's claim.

Steps:

- ② Use the t -test formula:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{32 - 30}{6/\sqrt{6}} = \frac{2}{6/2.449} \approx 0.816$$

- ③ Decision rule at a 95% confidence interval

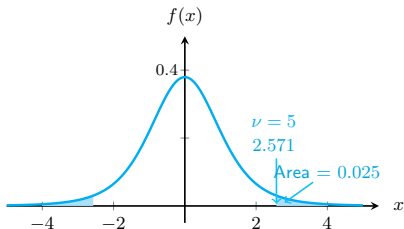
- Reject H_0 if $|t| > 2.571$.
- Otherwise, fail to reject H_0 .

- ④ Interpretation

- The test statistic $|t| = 0.816 < 2.571$.
- Thus, we fail to reject the null hypothesis.
- There is not enough evidence to conclude that the average daily energy consumption differs from the company's claim of 30 kWh.

Summary

- Student t -distribution of ν degrees of freedom



Student t -distribution

- Population: mean μ (✓), standard deviation σ (X)
- Sample: mean \bar{x} , standard deviation s , and small sample size n

- t -statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \Rightarrow t\text{-test}$

- 95% confidence interval: $\bar{x} \pm \underbrace{t_{\nu, 0.025}}_{\nu = n - 1} \times \frac{s}{\sqrt{n}}$



W. S. Gosset in Guinness

Teaching Concept

Method

- use math
- use figures
- use examples
- use data
- use codes
- use latex to create all examples

Thanks for your attention!

Any Questions?

About me:

 Homepage: <https://xinychen.github.io>

 How to reach me: chenxy346@gmail.com