

The Relevance of t -Statistics for Small Sample Sizes

An Introductory Class to Higher Statistics

Xinyu Chen

December 6, 2024



Outline

Content:

- ➊ How was t -statistic developed?
- ➋ Normal distribution vs. student t -distribution?
- ➌ What is t -statistic?
- ➍ How to calculate a t -test?
- ➎ What are the hypotheses and the assumptions?
- ➏ How to interpret results?

Development

The problems of **small sample sizes**

1876

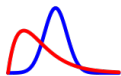
t-distribution as
a posterior distribution



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)

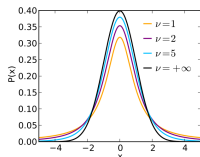
Diagram of the Pearson system

1908

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



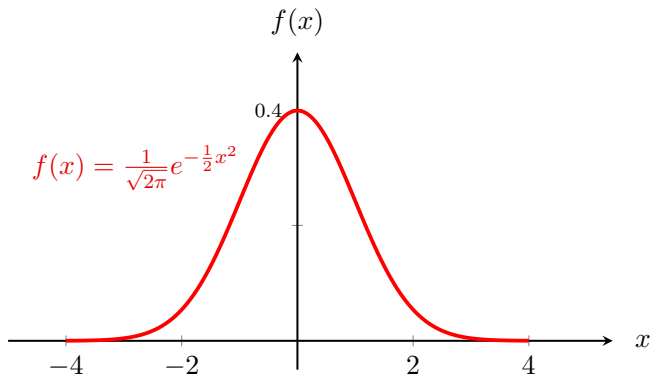
W. S. Gosset



(Source: Wiki)

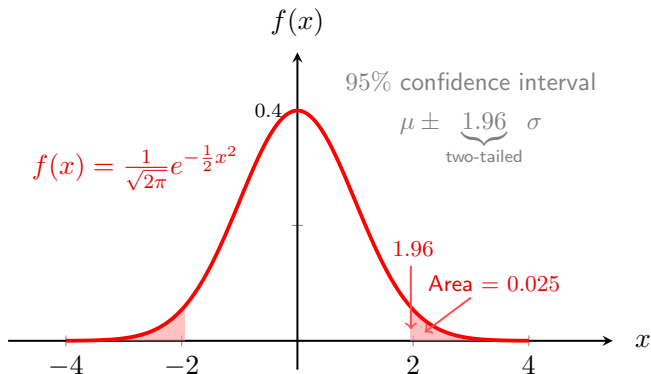
Probability density function

Revisiting Normal Distribution



Probability density function of the standard normal distribution

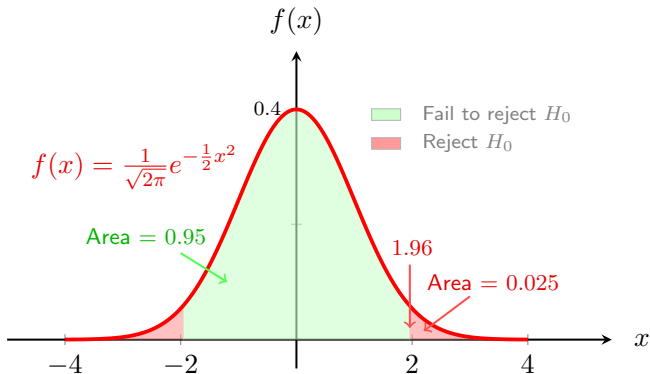
Revisiting Normal Distribution



Probability density function of the standard normal distribution

Connecting with Hypothesis Test

- Hypothesis test
 - Population: mean μ , standard deviation σ
 - Sample: mean \bar{x} , sample size n
 - Null hypothesis (H_0): The population mean is μ
 - z-statistic: $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ ($z \uparrow$ implies statistically significant difference)
- 95% confidence interval



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.

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

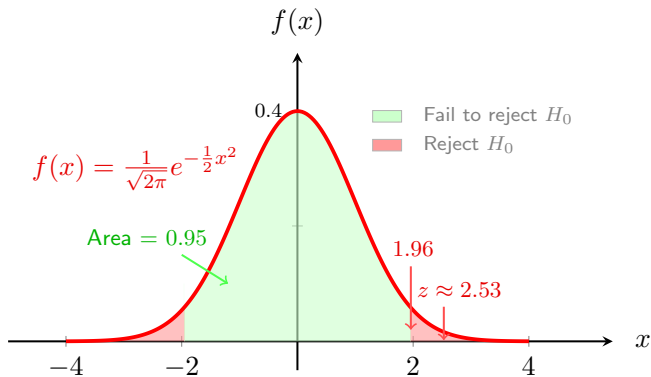
Steps:

① Formulate Hypotheses

- Null Hypothesis (H_0): The population mean is $\mu = 30$ kWh.
- Alternative Hypothesis (H_a): The population mean is $\mu \neq 30$ kWh.

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

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



Test statistic $|z| > 1.96 \Rightarrow$ Reject the null hypothesis

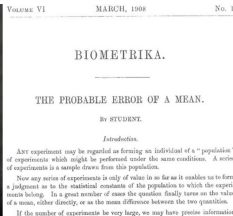
Student t -Distribution

In the case of **small sample sizes**?

- Switch to student t -distribution and t -test
- Probability density function:

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

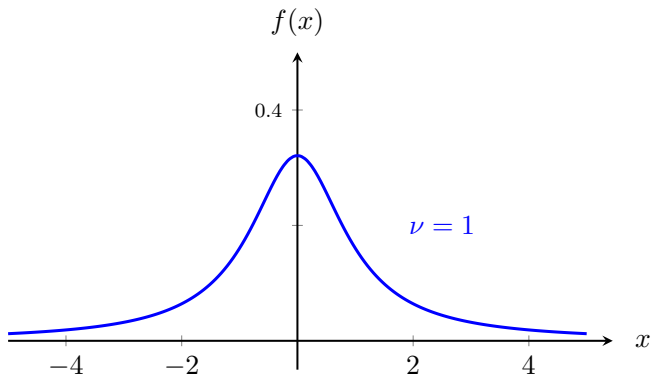
- $x \in \mathbb{R}$: random variable
- $\nu \in \mathbb{Z}^+$: degrees of freedom
- $\Gamma(\cdot)$: Gamma function



Gosset'1908 (known as “student”)

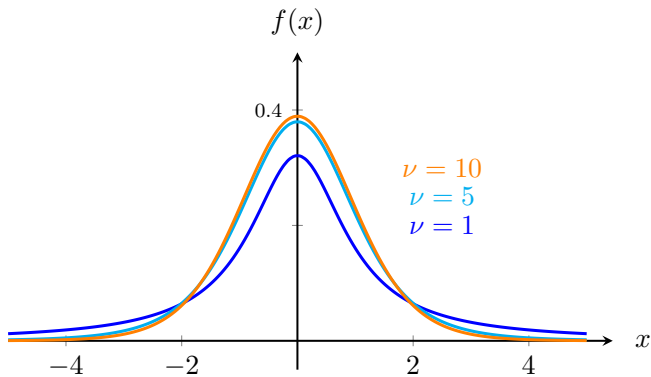
(Source: [link](#))

Student t -Distribution



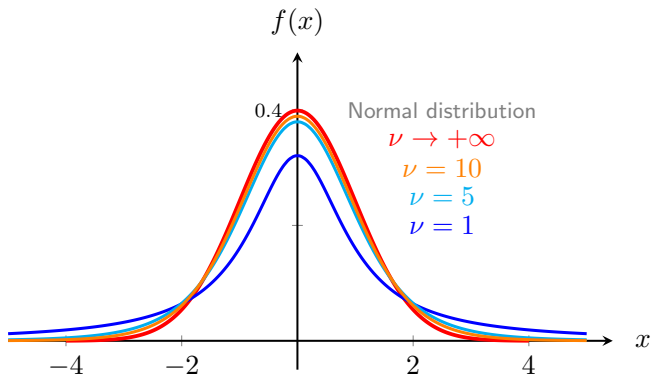
Student t -distribution of ν degrees of freedom

Student t -Distribution



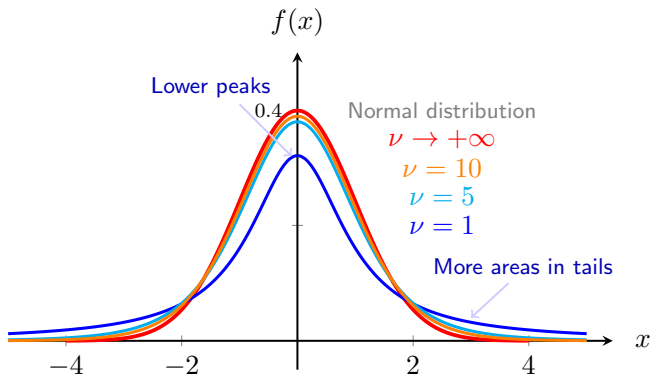
Student t -distribution of ν degrees of freedom

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 (✓) and standard deviation σ (✓/X)

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

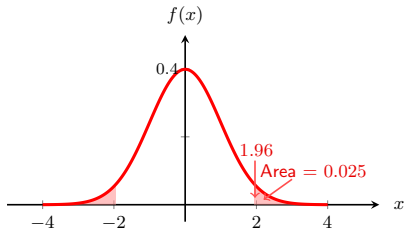
$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

- Use **z-test**

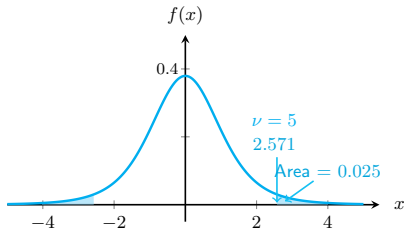
- If σ is unknown, using **sample standard deviation s** instead

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

- Use **t-test**



Standard normal distribution



Student t -distribution

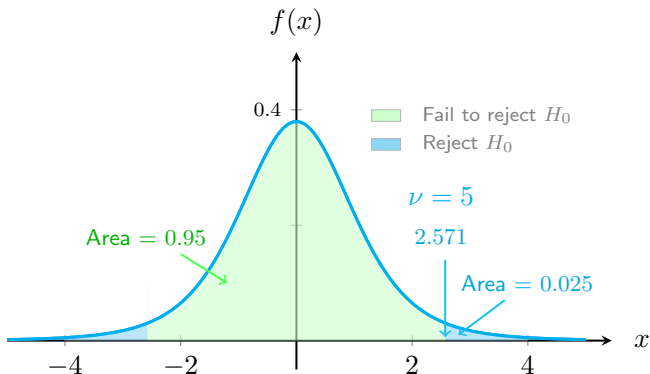
- **Heavy tail** in student t -distribution ($\nu = n - 1$ degrees of freedom) is important for small sample size n

Definition of t -Statistic

- Formula of t -statistic for small sample sizes

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

- Population: mean μ
- Sample: mean \bar{x} , standard deviation s , sample size n (small value)
- A high absolute value of t suggests a statistically significant difference.



Implementing t -Test for Small Sample Size

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.

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

Steps:

① Formulate Hypotheses

- Null Hypothesis (H_0): The population mean is $\mu = 30$ kWh.
- Alternative Hypothesis (H_a): The population mean is $\mu \neq 30$ kWh.

② Use the t -test formula since the population standard deviation is not known:

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

Critical Values in t -Table

Small sample sizes

- Degrees of freedom for a t -test:

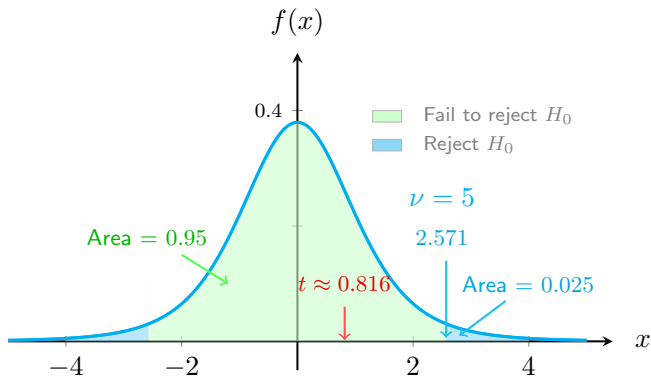
$$\nu = \underbrace{n}_{\text{sample size}} - 1 = 6 - 1 = 5$$

- 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

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



Test statistic $|t| < 2.571 \Rightarrow$ Fail to reject the null hypothesis

Implementing t -Test for Small Sample Size

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

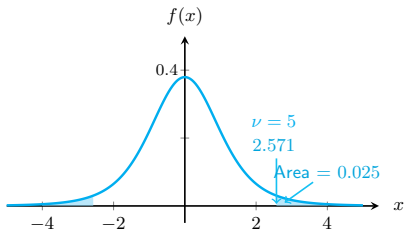
- Reject H_0 if $|t| > 2.571$.
- Otherwise, fail to reject H_0 .

- 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 for small sample size: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
- t -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

(Source: [link](#))

Thank you!

Any Questions?

Slides: https://xinychen.github.io/slides/t_stat.pdf

About me:

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

 How to reach me: chenxy346@gmail.com