Hypothesis testing

Pritha Banerjee

University of Calcutta

banerjee.pritha74@gmail.com

April 7, 2025

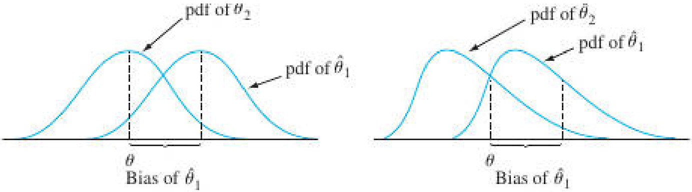
Overview

Point Estimation (1)

- From **sample parameters** mean $\bar{x}$ and variance s^2, **population parameters** μ and σ^2 can be **inferred** or estimated
- θ is the parameter to be estimated/ inferred.
- **Definition:** A **point estimate** of a parameter θ is a single number that gives a sensible value for θ that is obtained by selecting a suitable statistic and computation of its vale from sample data. The selected statistic is called the **point estimator** of θ
- **Example:** Find point estimate of population variance σ^2, given $*$ samples data $x = \{44.2, 43, 9, 44.7, 44.2, 44.0, 43.8, 44.6, 43.1\}$ from a population distribution.
- **Solution:** $\hat{\sigma}^2 = \frac{\sum_{i=1}^{8}(x_i - \bar{x})^2}{n-1} = 0.25125$; Thus, $\hat{\sigma} = 0.501$
- It is possible to find estimator $\hat{\theta}$ for which $\hat{\theta} = \theta$, ie. $\hat{\theta} = \theta +$ error of estimation, minimize $(\hat{\theta} - \theta)^2$, ie. Mean squared error $MSE = E((\hat{\theta} - \theta)^2)$, difficult to find smallest MSE

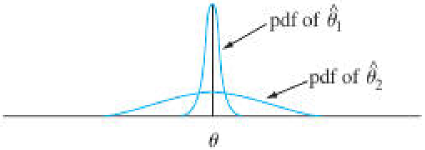
Point Estimation (2)

- **Definition:** A point estimator $\hat{\theta}$ is said to be **unbiased edtimator** of θ if $E(\hat{\theta}) = \theta$ for every possibe value of θ. If $\hat{\theta}$ is **not unbiased**, $E(\hat{\theta}) - \theta$ is the **bias** of $\hat{\theta}$
- pdf of **biased estimator** $\hat{\theta}_1$ and **unbiased estimator** $\hat{\theta}_2$ for a parameter θ
- **Example: sample proportion** $\hat{p} = \frac{X}{n}$ is an unbiased estimator, where p is probability of successes in binomial distribution with parameters n and p. $E(\hat{\theta}) = E(\hat{p}) = E(\frac{X}{n}) = \frac{1}{n}E(X) = \frac{1}{n}np = p = \theta$, since binomial distribution is normal distribution with $E(X) = np$



Some examples of point estimator

- **Proposition:** Let $X_1, X_2, \cdots, X_n$ be the random sample from a distribution with μ and variance σ^2, then the estimator $\hat{\sigma}^2 = s^2 = \frac{\sum(X_i - \bar{(X)})^2}{n-1}$ is an **unbiased** estimator of σ^2,**but** $\hat{s}$ is **biased estimator** of σ
- **Proposition:** If $X_1, X_2, \cdots, X_n$ are the random sample from a distribution with μ and variance σ^2, then $\bar{X}$ is an **unbiased** estimator of μ
- **Estimator with minimum variance:** Among all estimators of θ that are unbiased, choose one that has minimum variance. Such $\hat{\theta}$ is called **Minimum Variance Unbiased Estimator**(MVUE). For $X_1, X_2, \cdots, X_n$ random sample from Normal distribution with μ and σ^2, $\hat{\mu} = \bar{X}$ is MVUE for μ, $\hat{\theta}_1$ is MVUE



Standard error of estimator

- **Standard error(SE) of an estimator** $\hat{\theta}$ is standard deviation $\sigma_\theta = \sqrt{Var(\hat{\theta})}$
- If SE has unknow parameter, SE needs to be estimated , then $\hat{\sigma}_\theta$ is the **estimated SE**
- **Example:** SE of $\hat{p} = \frac{X}{n}$ is

$$\sigma_{\hat{p}} = \sqrt{Var(\frac{X}{n})} = \sqrt{\frac{Var(X)}{n^2}} = \sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}}$$

- Since p is unknown, substitute $\hat{p} = X/n$ and $\hat{q} = 1 - X/n$ in $\sigma_{\hat{p}}$ to obtain $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- If $\hat{\theta}$ has **approximately normal distribution** for large n, typically, **true value** of θ lies within 2 SE or Standard deviation of $\hat{\theta}$
- Point estimates are obtained by bootstrap method or maximum likelihood estimator method

Normal Distribution and Standard Normal Distribution

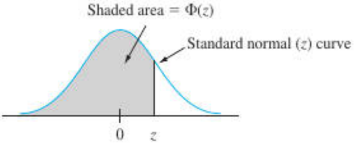
- A continuous rv X is said to have a **normal distribution** with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the PDF of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1}$$

- The **normal distribution** with parameters $\mu = 0$ and $\sigma = 1$ is called **standard normal distribution**. A **random variable** having standard normal distribution is called a **standard normal variable**, denoted by Z. $f(z; 0, 1)$ is standard normal or z curve; the pdf of Z is

$$f(z; 0, 1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{-z^2}{2}}; -\infty < z < \infty \tag{2}$$

- The cdf of Z is $P(Z \leq z) = \int\limits_{-infty}^{z} f(y; 0, 1)dy = \phi(z)$, where $z = \frac{x-\mu}{\sigma}$



Percentile of Standard Normal Distribution

- 99^{th} **percentile** of Std. Normal is that value on x axis such that area under z curve to the left of the value is 0.99, ie, for which value of z $P(Z \leq z) = 0.99$? the $100p^{th}$ percentile is row and column of normal distribution cdf table in which the entry p is found.
- z_α notation: For statistical inference we need value on z axis for which α of the area under z curve lies to the right of z_α, z_α is called z **critical value**

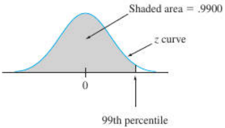
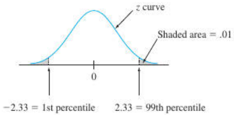


Figure 4.17 Finding the 99th percentile



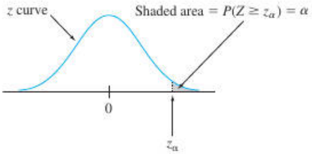
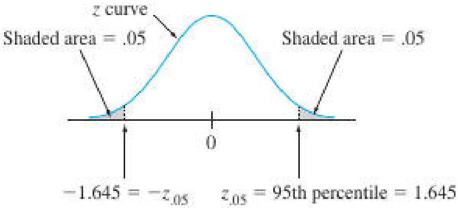


Table 4.1 Standard Normal Percentiles and Critical Values

Percentile	90	95	97.5	99	99.5	99.9	99.95
α (tail area)	.1	.05	.025	.01	.005	.001	.0005
$z_\alpha = 100(1-\alpha)$th percentile	1.28	1.645	1.96	2.33	2.58	3.08	3.27

Percentile of Standard Normal Distribution: Example

$z_{.05}$ is the $100(1 - .05)$th $= 95$th percentile of the standard normal distribution, so $z_{.05} = 1.645$. The area under the standard normal curve to the left of $-z_{.05}$ is also .05. (See Figure 4.20.)



Percentile of Non-Standard Normal Distribution

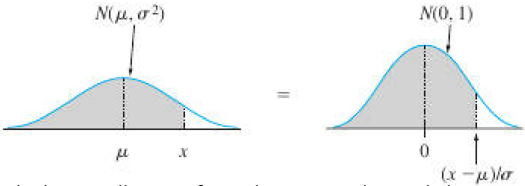
If X has a normal distribution with mean μ and standard deviation σ, then

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution. Thus

$$P(a \le X \le b) = P\left(\frac{a - \mu}{\sigma} \le Z \le \frac{b - \mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

$$P(X \le a) = \Phi\left(\frac{a - \mu}{\sigma}\right) \qquad P(X \ge b) = 1 - \Phi\left(\frac{b - \mu}{\sigma}\right)$$



- Standardizing calculates a distance from the mean value and then reexpresses the distance as some number of standard deviations. Thus, if $\mu = 100$ and $s = 15$, then $x = 130$ corresponds to $z = (130 - 100)/15 = 30/15 = 2$. That is, 130 is 2 standard deviations above (to the right of) the mean value.

Empirical rule

- If the population distribution of a variable is (approximately) normal, then
  - Roughly 68% of the values are within 1 SD of the mean.
  - Roughly 95% of the values are within 2 SDs of the mean.
  - Roughly 99.7% of the values are within 3 SDs of the mean.
- $(100p)^{th}$ percentile for normal $(\mu, \sigma) = \mu + (100p)^{th}$ percentile for standard normal . σ

Statistical Interval based on Single Sample

- Point estimate states nothing about how close an estimator $\hat{\theta}$ will be to the true θ
- **Confidence Interval** (CI) gives an interval around the true θ with a **confidence level** (measures degree of reliability of interval)
- A **Confidence level** of 95% means 95% of all samples would give an interval that includes true θ and only 5% of would yield erroneous interval; frequently used **Confidence levels** are 95%, 99%, 90%
- if **confidence level is high** and **interval is narrow**, it implies knowledge of parameter is **reasonably precise**, **wide CI** giver uncertainity

Basic properties of CI (1)

- Let the parameter of interest be population mean μ, population distribution is normal and value of the population standard deviation σ is known
- The actual sample observations $x_1, x_2, \cdots, x_n$ are assumed to be the result of a random sample $X_1, \cdots, X_n$ from a normal distribution with mean value μ and standard deviation σ
- The sample mean $\bar{X}$ is normally distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$. Standardizing X gives standard normal variable $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
- Because the area under the standard normal curve between -1.96 and 1.96 is .95,

$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95$$

- Write above as $l < \mu < u$, where the endpoints l and u involve X and $\sigma/\sqrt{n}$.

Basic properties of CI (2)

1. Multiply through by $\sigma/\sqrt{n}$:

$$-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

2. Subtract $\overline{X}$ from each term:

$$-\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

3. Multiply through by -1 to eliminate the minus sign in front of μ (which reverses the direction of each inequality):

$$\overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

that is,

$$\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

The equivalence of each set of inequalities to the original set implies that

$$P\left(\overline{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = .95$$

- random quantity appears on two ends, unknown constant μ appears in the middle. In interval notation, $\left(\bar{X} - 1.96.\sigma/\sqrt{n}, \bar{X} + 1.96.\sigma/\sqrt{n}\right)$
- The probability is .95 that the random interval includes or covers the true value of μ

Confidence Interval: Definition

If, after observing $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, we compute the observed sample mean $\bar{x}$ and then substitute $\bar{x}$ into (7.4) in place of $\bar{X}$, the resulting fixed interval is called a **95% confidence interval for $\boldsymbol{\mu}$**. This CI can be expressed either as

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \text{ is a 95\% CI for } \mu$$

or as

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ with 95\% confidence}$$

A concise expression for the interval is $\bar{x} \pm 1.96 \cdot \sigma/\sqrt{n}$, where $-$ gives the left endpoint (lower limit) and $+$ gives the right endpoint (upper limit).

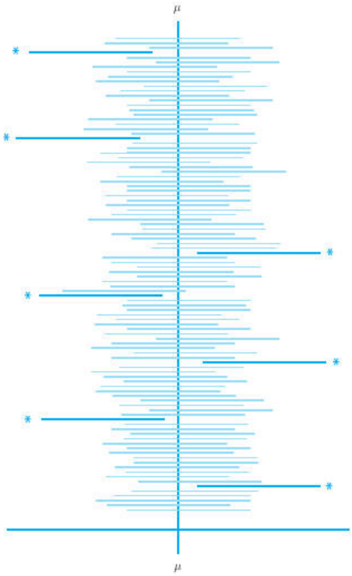
The quantities needed for computation of the 95% CI for true average preferred height are $\sigma = 2.0$, $n = 31$, and $\bar{x} = 80.0$. The resulting interval is

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 80.0 \pm (1.96)\frac{2.0}{\sqrt{31}} = 80.0 \pm .7 = (79.3, 80.7)$$

That is, we can be highly confident, at the 95% confidence level, that $79.3 < \mu < 80.7$. This interval is relatively narrow, indicating that μ has been rather precisely estimated.

Confidence Interval: Interpretation

- An event A has probability .95 implies if the experiment on which A is defined is performed over and over again, in the long run A will occur 95% of the time
- CI is **not** $P(\mu \text{ lies in}(79.3, 80.7)) = .95$
- Let A be the event that $\bar{X} - 1.96.\sigma/\sqrt{n} < \mu < \bar{X} + 1.96.\sigma/\sqrt{n}$. Since $P(A) = .95$, in the long run (in repeated experiments on different samples) 95% of computed CIs will contain μ.



Other Confidence Interval: Example

A **100(1 − α)% confidence interval** for the mean μ of a normal population when the value of σ is known is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{7.5}$$

or, equivalently, by $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

The production process for engine control housing units of a particular type has recently been modified. Prior to this modification, historical data had suggested that the distribution of hole diameters for bushings on the housings was normal with a standard deviation of .100 mm. It is believed that the modification has not affected the shape of the distribution or the standard deviation, but that the value of the mean diameter may have changed. A sample of 40 housing units is selected and hole diameter is determined for each one, resulting in a sample mean diameter of 5.426 mm. Let's calculate a confidence interval for true average hole diameter using a confidence level of 90%. This requires that $100(1 - \alpha) = 90$, from which $\alpha = .10$ and $z_{\alpha/2} = z_{.05} = 1.645$ (corresponding to a cumulative z-curve area of .9500). The desired interval is then

$$5.426 \pm (1.645)\frac{.100}{\sqrt{40}} = 5.426 \pm .026 = (5.400, 5.452)$$

With a reasonably high degree of confidence, we can say that $5.400 < \mu < 5.452$. This interval is rather narrow because of the small amount of variability in hole diameter ($\sigma = .100$). ■

One sided CI

A **large-sample upper confidence bound for μ** is

$$\mu < \bar{x} + z_\alpha \cdot \frac{s}{\sqrt{n}}$$

and a **large-sample lower confidence bound for μ** is

$$\mu > \bar{x} - z_\alpha \cdot \frac{s}{\sqrt{n}}$$

A **one-sided confidence bound for p** results from replacing $z_{\alpha/2}$ by z_α and $\pm$ by either $+$ or $-$ in the CI formula (7.10) for p. In all cases the confidence level is approximately $100(1 - \alpha)\%$.

The slant shear test is the most widely accepted procedure for assessing the quality of a bond between a repair material and its concrete substrate. The article "Testing the Bond Between Repair Materials and Concrete Substrate" (*ACI Materials J.*, 1996: 553–558) reported that in one particular investigation, a sample of 48 shear strength observations gave a sample mean strength of 17.17 N/mm^2 and a sample standard deviation of 3.28 N/mm^2. A lower confidence bound for true average shear strength μ with confidence level 95% is

$$17.17 - (1.645)\frac{(3.28)}{\sqrt{48}} = 17.17 - .78 = 16.39$$

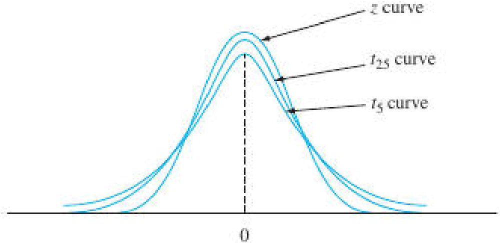
That is, with a confidence level of 95%, we can say that $\mu > 16.39$.

Normal population distribution: t distribution

- **Assumption:** The population of interest is normal, so that $X_1, \cdots, X_n$ constitutes a random sample from a normal distribution with both μ and σ unknown.
- For large n, the random variable $Z = (\bar{X} - \mu)/(S/\sqrt{n})$ has approximately a **standard normal distribution**. When n is small, S is no longer likely to be close to σ, so the variability in the distribution of Z arises from randomness in both the numerator and the denominator.
- This implies that the probability distribution of $Z = (\bar{X} - \mu)/(S/\sqrt{n})$ will be more spread out than the standard normal distribution. This introduces a new family of probability distributions called t-distributions.
- **Definition:** When $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, the rv $T = (\bar{X} - \mu)/(S/\sqrt{n})$ has a probability distribution called a t-distribution with $n-1$ degrees of freedom (df).

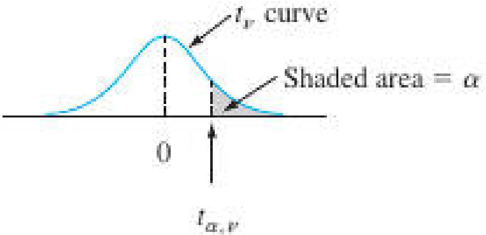
Properties of t distribution (1)

- T **does not have standard normal distribution**, when n is **small**.
- Any particular t distribution results from specifying the value of a single parameter, called the **number of degrees of freedom (df)**, ν; different curves for different ν values. Let t_ν denote the t-distribution with ν df
- Each t_ν curve is bell-shaped and centered at 0.
- Each t_ν curve is more spread out than the standard normal (z) curve.
- As ν increases, the spread of the corresponding t_ν curve decreases.
- As $n \to \infty$, the sequence of t_ν curves approaches the standard normal curve (so the z curve is often called the t curve with df $= \infty$).



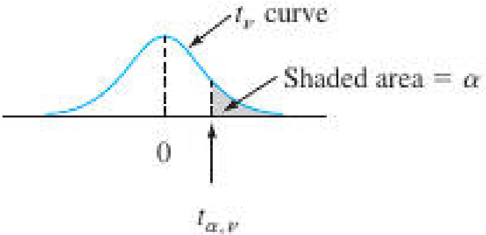
Properties of t distribution (2)

- The number of df ν for T $n-1$ because, although S is based on the n deviations $X_1 - \bar{X}, \cdots, X_n - \bar{X}$, $\sum(X_i - \bar{X}) = 0$ implies that only $n-1$ of these are **freely determined.** The number of df for a t variable is the number of **freely determined deviations** on which the estimated standard deviation in the denominator of T is based.
- t distribution is used for making inferences along with ν by capturing **t-curve tail areas** like z_α for z curve.
- **t critical value:** Let $t_{\alpha,n}$ is the number on the measurement axis for which the area under the t curve with n df to the right of $t_{\alpha,n}$ is α; $t_{\alpha,n}$ is called a t critical value.



Properties of t distribution (2)

- The number of df ν for T $n-1$ because, although S is based on the n deviations $X_1 - \bar{X}, \cdots, X_n - \bar{X}$, $\sum(X_i - \bar{X}) = 0$ implies that only $n-1$ of these are **freely determined.** The number of df for a t variable is the number of **freely determined deviations** on which the estimated standard deviation in the denominator of T is based.
- t distribution is used for making inferences along with ν by capturing **t-curve tail areas** like z_α for z curve.
- **t critical value:** Let $t_{\alpha,n}$ is the number on the measurement axis for which the area under the t curve with n df to the right of $t_{\alpha,n}$ is α; $t_{\alpha,n}$ is called a t critical value.



One sample t CI

- The standardized variable T has a t distribution with $n-1$ df, and the area under the t density curve between $-t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$ is $1-\alpha$ (area $\alpha/2$ lies in each tail), so $P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1-\alpha$
- **Proposition:**

Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean μ. Then a **$100(1-\alpha)\%$ confidence interval for μ** is

$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right) \tag{7.15}$$

or, more compactly, $\bar{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}$.

An **upper confidence bound for μ** is

$$\bar{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$$

and replacing $+$ by $-$ in this latter expression gives a **lower confidence bound for μ**, both with confidence level $100(1-\alpha)\%$.

One sample t CI

- The standardized variable T has a t distribution with $n-1$ df, and the area under the t density curve between $-t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$ is $1-\alpha$ (area $\alpha/2$ lies in each tail), so $P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1-\alpha$
- **Proposition:**

Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean μ. Then a **$100(1-\alpha)\%$ confidence interval for μ** is

$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right) \qquad (7.15)$$

or, more compactly, $\bar{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}$.

An **upper confidence bound for μ** is

$$\bar{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$$

and replacing $+$ by $-$ in this latter expression gives a **lower confidence bound for μ**, both with confidence level $100(1-\alpha)\%$.

One sample t CI: Example

- Let there is a sample data X with $n = 30$ observations. Determine the mean μ of the population with 95% confidence level.
- Let $\bar{x}$ of sample data is 1203.191 and sample $s = 543.54$.
- The CI is based on $n - 1 = 29$ degrees of freedom, so the necessary t **critical value** is $t_{.025,29} = 2.045$.
- The interval estimate is $\pm t_{.025,29}.s/\sqrt{n} = (7000.253, 7406.129)$
- $7000.253 < \mu < 7406.129$ with 95% confidence.
- A **lower** 95% confidence bound would result from retaining only the lower confidence limit (the one with $-$ sign) and replacing 2.045 with $t_{.05,29} = 1.699$.
- When sample size $n < 40$, apply t distribution rather than z or standard normal.