MM 225 – AI and Data Science

Day 17: Interval Estimation

Instructors: Hina Gokhale, MP Gururajan, N. Vishwanathan

5 SEPTEMBER 2024

Why Interval Estimation?

Point estimator does not indicate any accuracy of measurement

Point estimator $\bar{X}$ for population mean μ, only indicates that the value of $\bar{X}$ is "close to" population mean μ.

In order to answer "how close", one needs to make interval estimation

Interval Estimation with Normal population

Let $X_1, X_2, \dots, X_n$ be n random measurements of an experiment.

Assume that these measurement come from a Normal population with mean μ and standard deviation σ.

$E(\bar{X}) = \mu \; and \; Var(\bar{X}) = \frac{\sigma^2}{n}, hence \; Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

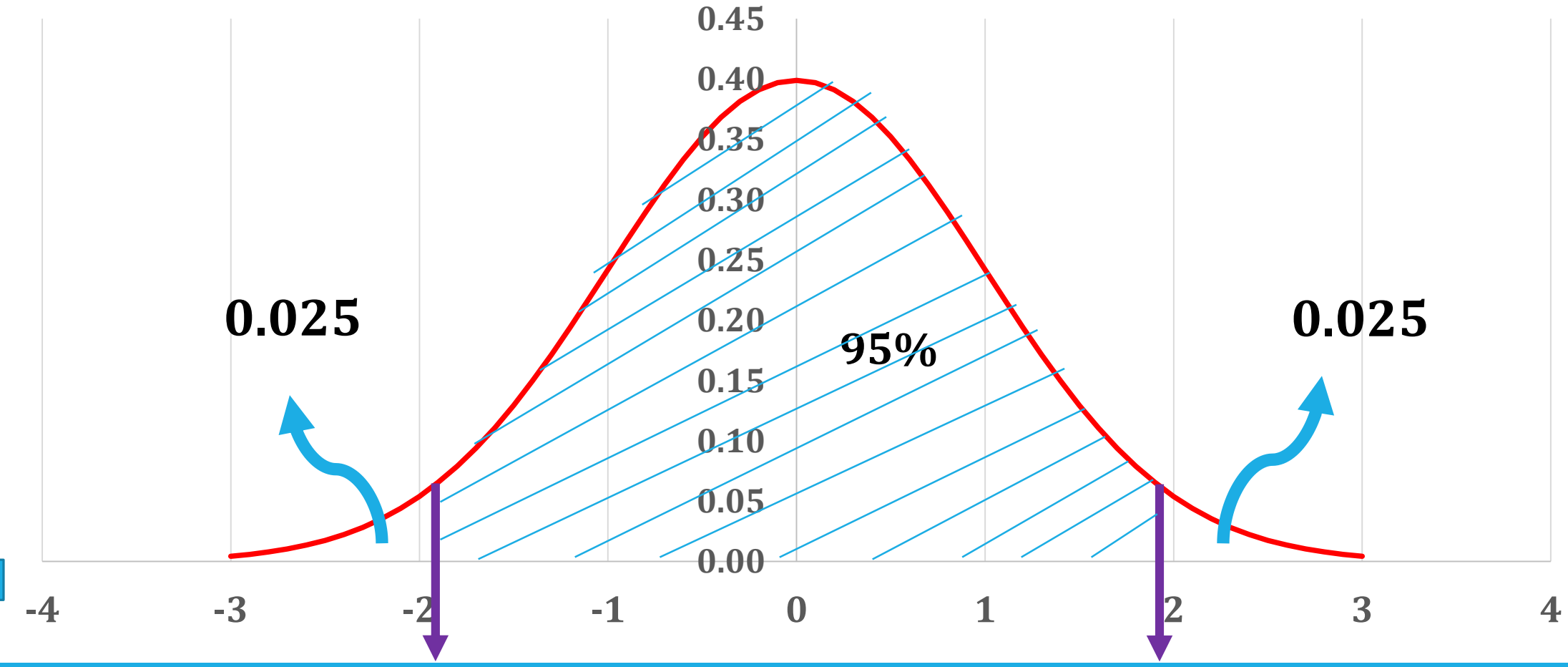
Therefore, from standard Normal table we find, $P[-1.96 < Z < +1.96] = 0.95$

$$\Rightarrow P\left[-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right] = 0.95$$

$$\Rightarrow P\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is called 95% interval estimator of μ

Plot for N(0, 1) and probabilities



Data Representation

In general, data is represented with error of $\pm 1\sigma$ limit. Here σ represents standard deviation of the distribution.

Statistically it can be expressed as

$$P[-1 < Z < 1] = 0.68$$

$$\Rightarrow P\left[\bar{X} - \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}\right] = 0.68$$

Note that it is also referred as “Data Accuracy”

The case of σ^2 unknown

95% interval estimator of μ:

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Note that this interval assumes that σ^2 : population variance is known.

What happens when it is unknown?

Replace the σ^2 by its estimate: s^2

When σ^2 is unknown…

Then note that

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/\sigma}{\sqrt{\frac{S^2}{\sigma^2 n}}} \sim t(n-1)$$

Hence would like to find a and b such that P[a < t < b] = 1- α, where 1- α indicates the confidence level.

Since t distributions are symmetric about 0, this would simplify to find a such that

$$P[-a < t < a] = 1 - \alpha$$

When σ is unknown

$$P[\ -a < t < a] = 1-\alpha$$

$$P[t < -a\ ] + P[a < t] = \alpha$$

Due to symmetry, $2*P[t > a] = \alpha$

$$\text{Or } P[t > a] = \frac{\alpha}{2}$$

$$P\left[-t_{(n-1),\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{(n-1),\alpha/2}\right] = 1-\alpha$$

Note that given α, percentage points $t_{(n-1),\alpha/2}$ can be found using the t-distribution tables, as shown graphically on the next slide.

When σ is unknown

$$P\left[\bar{X} - \frac{S}{\sqrt{n}} t_{(n-1),\alpha/2} < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{(n-1),\alpha/2}\right] = 1 - \alpha$$

Example:

If 1- α = 0.95 and n = 5, then n-1= 4 and $t_{4,0.025} = 2.776$

The 95% accuracy of the data can be given by

$$\bar{X} \pm \frac{S}{\sqrt{5}} 2.776 = \bar{X} \pm 1.241\, S$$

Interval estimator of population variance

Let $X_1, X_2, \dots, X_n$ be a random sample from N(μ, σ^2)

Want find interval estimator for σ^2

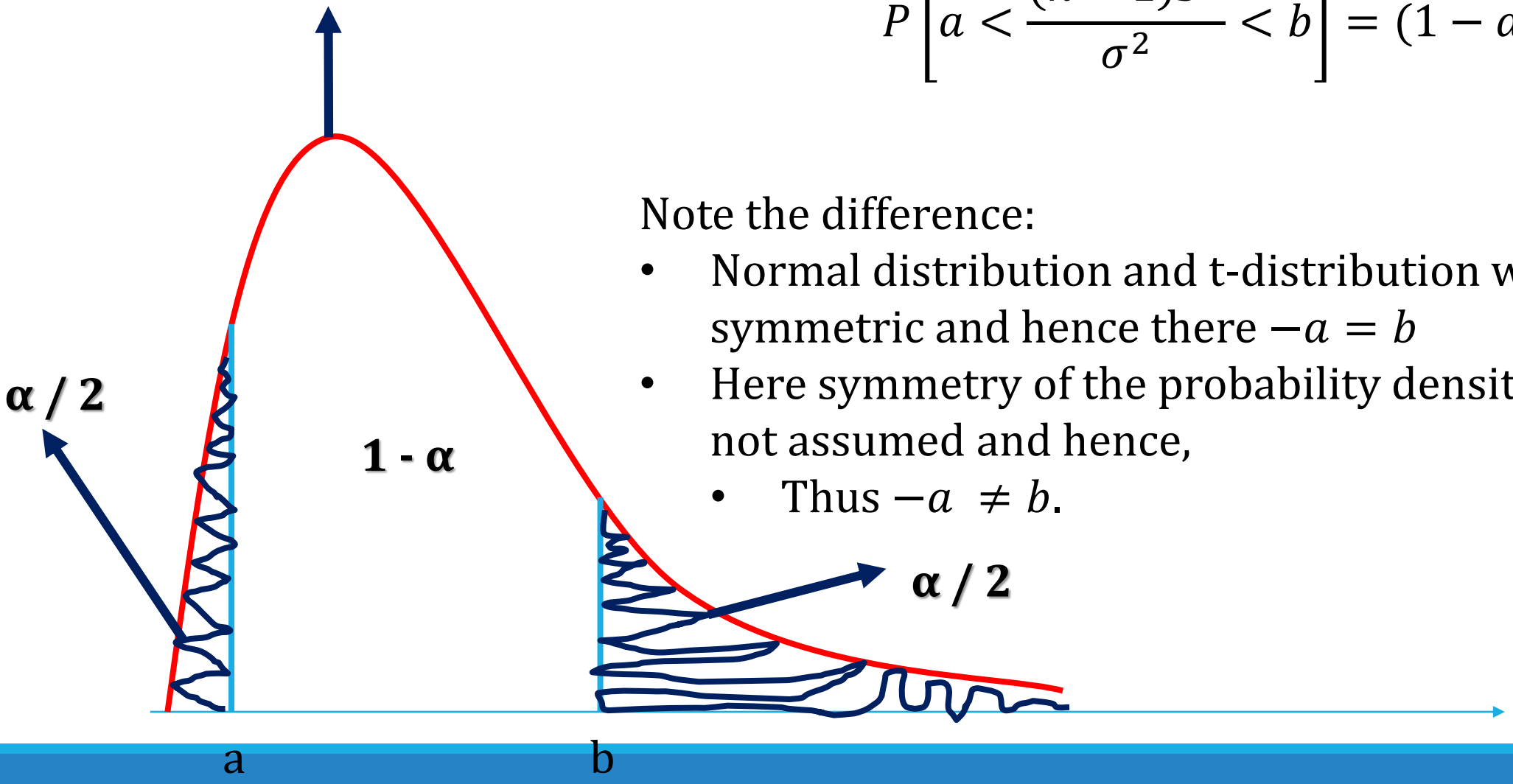
Note that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, therefore need to find a and b such that

$$P\left[a < \frac{(n-1)S^2}{\sigma^2} < b\right] = (1-\alpha)$$

Chi-Squared pdf

Need to find a and b such that

$$P\left[a < \frac{(n-1)S^2}{\sigma^2} < b\right] = (1-\alpha)$$



Note the difference:

- Normal distribution and t-distribution were symmetric and hence there $-a = b$
- Here symmetry of the probability density function is not assumed and hence,
  - Thus $-a \neq b$.

Interval estimator of population variance

Refer to the Chi-square table:

Accordingly, in $P\left[a < \frac{(n-1)S^2}{\sigma^2} < b\right] = (1-\alpha)$,

$a = \chi^2_{1-\alpha/2,(n-1)}$ and $b = \chi^2_{\alpha/2,(n-1)}$

Consider the case of n=10, and confidence level of 95%, then α = 0.05

$$a = \chi^2_{1-\alpha/2,(n-1)} = \chi^2_{0.975,9} = 2.70$$

$$b = \chi^2_{\alpha/2,(n-1)} = \chi^2_{0.025,9} = 19.02$$

$$P\left[2.70 < \frac{(n-1)S^2}{\sigma^2} < 19.02\right] = 0.95\text{ , therefore,}$$

$$P\left[\frac{(n-1)S^2}{19.02} < \sigma^2 < \frac{(n-1)S^2}{2.70}\right] = P\left[\frac{9S^2}{19.02} < \sigma^2 < \frac{9S^2}{2.70}\right] = 0.95$$

Important:

Do not forget to see how the table values are calculated.

If you are using Excel function to calculate percentage points, check the help to see how Excel calculates the values

If you are using Python or R use the help to check how the values are calculated

Example:

Each of 20 science students independently measured the melting point of lead. The sample mean and sample standard deviation of these measurements were (in degrees centigrade) 330.2 and 15.4, respectively. Construct (a) a 95 percent and (b) a 99 percent confidence interval estimate of the true melting point of lead.

Solution:

X = melting point of lead measured in degree centigrade

$\bar{x} = 330.2 \; and \; s = 15.4$

Let μ = true melting point

Sample size = 20,

Assume that $X \sim N(\mu, \sigma^2)$

mean and variance are unknown:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

$$\therefore P\left(-t(19, 0.025) < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < t(19, 0.025)\right) = 0.95$$

$$P\left(\bar{x} - 2.093\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{x} + 2.093\left(\frac{s}{\sqrt{n}}\right)\right) = 0.95$$

$$P(322.99 < \mu < 337.41) = 0.95$$

Summary

Interval Estimation for accuracy

The case of Normal population – variance known

The case of Normal population – variance unknown

The case of Normal population and interval estimate of Variance

Thank you…