DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Lecture 19.4 - Statistical Analysis

Sample Means

The sample mean $\bar{x}$ is an estimate of the population mean, μ.

Given *n* samples $\{X_1, X_2, \ldots, X_n\}$, drawn from the population with mean μ and standard deviation σ,

$$\bar{x} = \frac{1}{n}\sum_i X_i$$

The mean and variance of the sample mean are:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem (CLT)

For a large enough sample size n, the distribution of the sample mean $\bar{x}$ will approach a normal distribution.

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- This is true for a sample of independent random variables from any population distribution, as long as the population has a finite standard deviation σ.

Variance

Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Population variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Confidence Interval

When the population standard deviation σ is assumed to be known, according to the CLT, we can estimate how accurately the sample mean estimates the population mean. This is often expressed as a **confidence interval**.

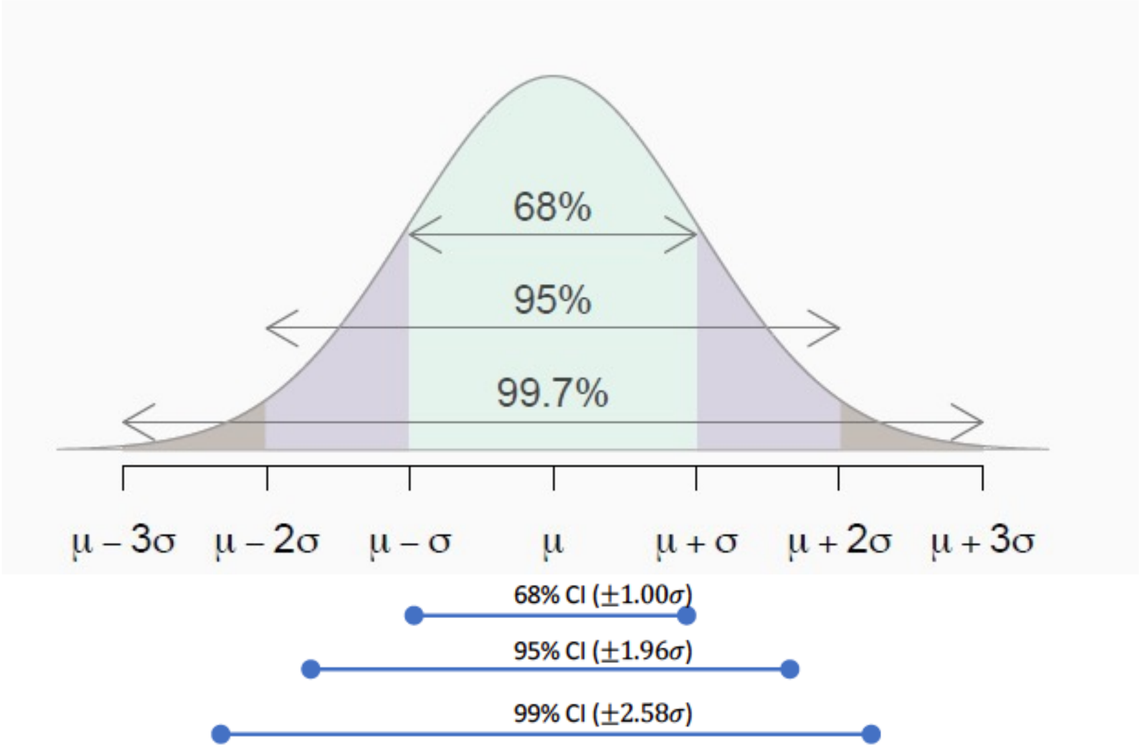
In other words, for random sample of sufficient size (usually >20), the confidence interval at $1 - \alpha$ confidence level is given as:

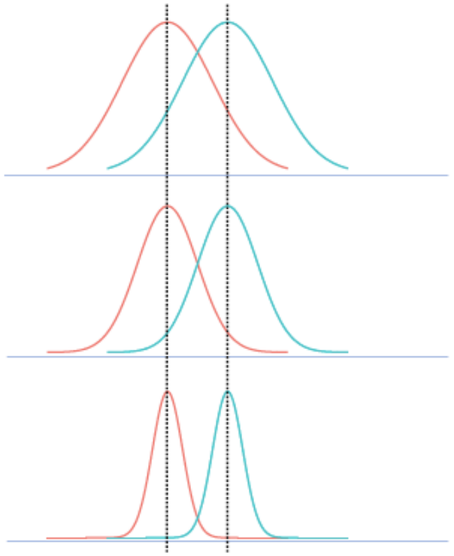
$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\alpha}{\sqrt{n}}$$

where $z_{\frac{\alpha}{2}}$ is $100(1 - \frac{\alpha}{2})$ percentile of the standard normal distribution z.

Confidence Interval

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\alpha}{\sqrt{n}}$$





High variability

Medium variability

Low variability

Internal Measures: Cohesion and Separation

Cluster Cohesion measures how closely related objects are in a cluster, and the within cluster sum of square (WSS) can be used to quantify it.

$$WSS = \sum_{i} \sum_{x \in Ci} (x - m_i)^2$$

Cluster Separation measures how distinct or well-separated a cluster is from other clusters, and the between cluster sum of squares (BSS) can be used to quantify it.

$$BSS = \sum_{i} |C_i| \, (m - mi)^2$$

where $|C_i|$ is the size of a cluster C_i and m is the centroid of all the samples.

Hypothesis Testing

Statistical tests of population mean with known variance

Lower Tailed Test with Population Mean with Known Variance

Null Hypothesis

$$H_0: \mu_0 \leq \mu$$

where μ_0 is the hypothesized population mean.

Test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

H_0 is rejected if $z \leq -z_\alpha$, where z_α is the 100 $(1 - \alpha)$ percentile of the standard normal distribution.

Hypothesis Testing

Statistical tests of population mean with known variance

Upper Tailed Test with Population Mean with Known Variance

Null Hypothesis

$$H_0: \mu_0 \geq \mu$$

where μ_0 is the hypothesized population mean.

Test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

H_0 is rejected if $z \geq -z_\alpha$, where z_α is the 100 (1 − α) percentile of the standard normal distribution.

Hypothesis Testing

Statistical tests of population mean with known variance

Two-Tailed Test with Population Mean with Known Variance

Null Hypothesis

$$H_0: \mu_0 = \mu$$

where μ_0 is the hypothesized population mean.

Test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

H_0 is rejected if $z \geq z_{\frac{\alpha}{2}}$ or $z \leq -z_{\frac{\alpha}{2}}$, where $z_{\frac{\alpha}{2}}$ is the 100 $(1 - \frac{\alpha}{2})$ percentile of the standard normal distribution.

Hypothesis Testing

Statistical tests of population mean with unknown variance

- Lower tailed test
- Upper tailed test
- Two-tailed test

Basically, the same step with only one difference.

- Test statistic:

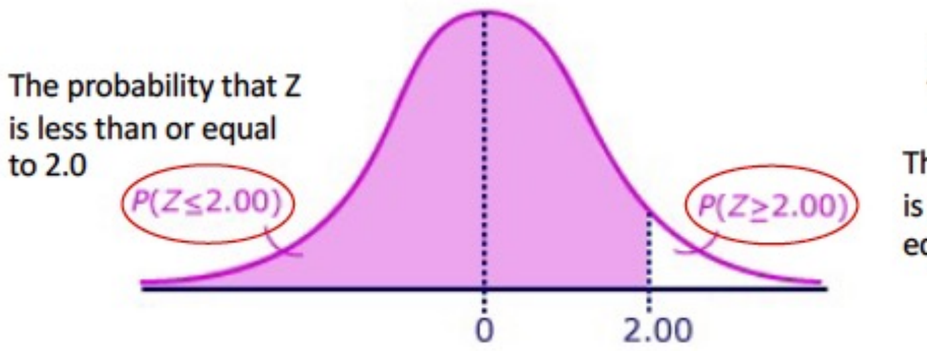
$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

where s is the sample variance, and t follows Student t distribution with $n - 1$ degrees of freedom.

What is p-value?

a.k.a., the probability value or asymptotic significance

For a given statistical model and observations, the probability that the null hypothesis (H0) is not rejected, and hence, true.



Statistical Analysis for Genomics

Differentially Expressed Genes is taking the normalized read count data and performing **statistical analysis** to discover quantitative changes in expression levels between experimental groups.

- For example, we use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

Gene Set Enrichment Analysis (GSEA) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, which may have an association with disease phenotypes.

- The method uses **statistical approaches** to identify significantly enriched or depleted groups of genes.
- Transcriptomics technologies and proteomics results often identify thousands of genes which are used for the analysis.

DLI Accelerated Data Science Teaching Kit

Thank You