Probability and Statistics

Module 4: Statistical Inference and Hypothesis Testing

Complete classroom and examination notes

B.Sc. Computer Science • Course Code DASH250011

Based on the NIELIT Deemed to be University 2025 curriculum

Contents

1 Module map 1

2 1. Statistical inference 1

3 2. Sampling techniques 1

3.1 Probability sampling 1

3.2 Non-probability sampling 1

3.3 Sampling error and bias 2

4 3. Sampling distributions and the standard error 2

5 4. Point estimation 2

5.1 Qualities of a good estimator 2

6 5. Interval estimation and confidence intervals 3

6.1 Confidence interval for a mean, known σ 3

6.2 Confidence interval for a mean, unknown σ 3

6.3 Confidence interval for a population proportion 3

6.4 Confidence interval for difference of two means 3

6.5 Width and sample size 4

7 6. Hypothesis testing framework 4

7.1 One-sided and two-sided alternatives 4

7.2 Test statistic and p-value 4

8 7. Z-test 4

9 8. t-test 5

9.1 One-sample t-test 5

9.2 Paired t-test 5

9.3 Two-sample t-test 5

10 9. Chi-square tests 6

10.1 Goodness-of-fit test 6

10.2 Test of independence 6

10.3 Test for a population variance 6

11 10. Type I and Type II errors 6

12 11. Statistical significance and practical significance 7

13 12. Multiple testing and responsible interpretation 7

14 13. Computing applications and case studies 7

15 14. Problem-solving checklist 8

16 15. Quick formula sheet 8

17 16. Practice questions with answers 8

1 Module map

The syllabus module covers sampling techniques; point and interval estimation; hypothesis testing; null and alternative hypotheses; Z-test, t-test, and chi-square test; Type I and Type II errors; confidence intervals; and computing case studies.

> **Learning outcomes.** You should be able to distinguish a population from a sample, select a sampling method, construct point and interval estimates, formulate hypotheses, choose and perform Z, t, and chi-square tests, interpret p-values and confidence intervals, explain Type I and Type II errors, and write statistically responsible conclusions.

2 1. Statistical inference

Statistical inference uses information from a sample to make statements about a population while accounting for sampling uncertainty.

- A **population** is the complete collection of units of interest.
- A **sample** is the observed subset selected from the population.
- A **parameter** is a fixed but usually unknown population quantity, such as μ, σ^2, or a population proportion p.
- A **statistic** is calculated from sample data, such as x |, s^2, or $\hat{p}$.
- An **estimator** is a rule or statistic used to estimate a parameter.
- An **estimate** is the numerical value produced by an estimator for one sample.

Inference is reliable only when the sample and analysis reasonably represent the population and the assumptions of the method are appropriate.

3 2. Sampling techniques

3.1 Probability sampling

In probability sampling, every population unit has a known, non-zero chance of selection.

Method	Description	Use / caution
Simple random	Every unit has an equal or known chance; select using random numbers.	Simple conceptually; needs a sampling frame.
Systematic	Choose a random start, then every k-th unit.	Efficient; beware periodic patterns.
Stratified	Divide into homogeneous strata, then sample within each.	Ensures subgroup representation.
Cluster	Divide into clusters and sample whole clusters or units within them.	Cost-efficient; clusters may be internally similar.

3.2 Non-probability sampling

Convenience, voluntary-response, quota, and purposive sampling may be useful in exploratory work, but selection probabilities are unknown and generalization is more difficult. A large convenience sample is not automatically representative.

3.3 Sampling error and bias

Sampling error is natural variation between a sample statistic and the population parameter. It tends to decrease as sample size increases under suitable random sampling.

Sampling bias is systematic error caused by the sampling process, such as undercoverage, nonresponse, voluntary response, or selection based on the outcome. Increasing the sample size does not necessarily remove bias.

> **Important distinction.** Random sampling supports generalization from sample to population. Random assignment in an experiment supports causal comparisons. They are different ideas.

4 3. Sampling distributions and the standard error

A **sampling distribution** is the probability distribution of a statistic over repeated samples of the same size from the same population.

For a sample mean from a population with mean μ and variance σ^2:

$$E[x\ |] = \mu, \quad V(x\ |) = \frac{\sigma^2}{n}, \quad \mathrm{SE}(x\ |) = \frac{\sigma}{\sqrt{n}}.$$

When σ is unknown, estimate the standard error by

$$\mathrm{SE}(x\ |) = \frac{s}{\sqrt{n}}.$$

For a sample proportion $\hat{p} = \frac{X}{n}$ under suitable independent Bernoulli sampling:

$$E[\hat{p}] = p, \quad \mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

The standard error measures sampling variability; it is not the standard deviation of individual observations.

5 4. Point estimation

A **point estimate** is one number used to estimate a population parameter.

Parameter	Common point estimator	Interpretation
Population mean μ	Sample mean $x\ \|$	Average outcome in the population.
Population proportion p	$\hat{p} = \frac{x}{n}$	Fraction with a characteristic.
Population variance σ^2	Sample variance s^2	Population spread.
Difference $\mu_1 - \mu_2$	$x\ \|_1 - x\ \|_2$	Difference between two population means.

5.1 Qualities of a good estimator

- **Unbiasedness**: $E\left[\hat{\theta}\right] = \theta$.
- **Consistency**: $\hat{\theta}$ approaches θ as sample size grows.
- **Efficiency**: among comparable unbiased estimators, it has smaller variance.
- **Sufficiency**: it captures the information in the sample relevant to the parameter under the model.

No estimator is automatically best in every situation; properties depend on the model and loss function.

6 5. Interval estimation and confidence intervals

A **confidence interval** gives a range of plausible values for a population parameter, together with a confidence level describing the long-run performance of the procedure.

General form:

$$\text{estimate} \pm \text{critical value} \times \text{standard error}.$$

A 95% confidence interval procedure captures the true parameter in about 95% of repeated samples when its assumptions and model are correct. It does **not** mean that there is a 95% probability that a fixed parameter lies inside one already-computed frequentist interval.

6.1 Confidence interval for a mean, known σ

For a random sample with known population standard deviation:

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

Common two-sided critical values:

Confidence level	α	Critical value $z_{\frac{\alpha}{2}}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.576

6.2 Confidence interval for a mean, unknown σ

When σ is unknown and the population is approximately Normal, use the Student t distribution with $n - 1$ degrees of freedom:

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}.$$

The t critical value is larger than the corresponding Normal critical value for finite samples because estimating σ adds uncertainty.

6.3 Confidence interval for a population proportion

A commonly used large-sample interval is

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

This approximation should not be used blindly when $n\hat{p}$ or $n(1-\hat{p})$ is small. Exact or improved intervals may be preferred.

6.4 Confidence interval for difference of two means

For independent samples, a common Welch interval is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where ν is the Welch–Satterthwaite approximate degrees of freedom. For a paired design, first compute within-pair differences and construct an interval for their mean.

Worked example: mean confidence interval

A sample of $n = 36$ requests has mean latency $x \mid = 120$ ms and known $\sigma = 18$ ms. A 95% interval is $120 + -1.96\left(\frac{18}{\sqrt{36}}\right) = 120 + -5.88$, giving $(114.12, 125.88)$ ms.

6.5 Width and sample size

For a mean interval with known σ, margin of error is $E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. To achieve a desired margin E:

$$n \geq \left(z_{\frac{\alpha}{2}} \frac{\sigma}{E}\right)^2.$$

Larger samples make intervals narrower; higher confidence makes them wider; greater population variability makes them wider.

7 6. Hypothesis testing framework

A hypothesis test evaluates evidence about a population claim.

1. State the **null hypothesis** H_0, usually representing no change, no difference, or a reference value.
2. State the **alternative hypothesis** H_1 or H_a.
3. Select significance level α before looking at the result.
4. Choose a test statistic and verify assumptions.
5. Calculate the test statistic and p-value, or compare with a critical value.
6. Reject H_0 if the evidence is sufficiently inconsistent with it; otherwise **fail to reject** H_0.
7. Report the result in context, including effect size and uncertainty where possible.

7.1 One-sided and two-sided alternatives

For a parameter θ and reference value θ_0:

- Two-sided: $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$.
- Right-sided: $H_0 : \theta \leq \theta_0$ versus $H_a : \theta > \theta_0$.
- Left-sided: $H_0 : \theta \geq \theta_0$ versus $H_a : \theta < \theta_0$.

The alternative determines where extreme evidence lies. Do not choose a one-sided test after seeing the data merely because it produces a smaller p-value.

7.2 Test statistic and p-value

A **test statistic** standardizes the distance between the observed estimate and the null value. A **p-value** is the probability, assuming H_0 is true, of obtaining a result at least as extreme as the observed result in the direction specified by H_a.

A small p-value means the data are unusual under H_0; it does not measure the probability that H_0 is true.

> **Decision language.** Say "reject H_0" or "fail to reject H_0." Avoid saying "accept H_0," because failure to find sufficient evidence against a hypothesis does not prove it true.

8 7. Z-test

A one-sample Z-test for a mean is appropriate when the population standard deviation σ is known or when a sufficiently large-sample Normal approximation is justified.

For $H_0 : \mu = \mu_0$:

$$z = \frac{x\ | - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

For a two-sided test, reject at level α when $|z| > z_{\frac{\alpha}{2}}$. For a right-sided test, reject when $z > z_\alpha$; for a left-sided test, reject when $z \leftarrow z_\alpha$.

A large-sample one-proportion Z-test uses

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 \frac{1-p_0}{n}}},$$

where the null proportion p_0 is used in the standard error.

Worked example: Z-test

A service claims mean latency is 100 ms. A random sample of 64 requests has $x\ | = 104$ ms and known $\sigma = 16$ ms. For $H_0 : \mu = 100$ versus $H_a : \mu \neq 100$, $z = \frac{104-100}{\frac{16}{8}} = 2$. Since $2 < 1.96$ is false, the two-sided 5% decision is to reject H_0; the evidence suggests the mean differs from 100 ms.

9 8. t-test

The Student t distribution is used for means when population σ is unknown, especially with small or moderate samples from an approximately Normal population.

9.1 One-sample t-test

For $H_0 : \mu = \mu_0$:

$$t = \frac{x\ | - \mu_0}{\frac{s}{\sqrt{n}}}, \quad \text{df} = n - 1.$$

Use the appropriate t critical value or calculate the p-value from the t distribution with $n - 1$ degrees of freedom.

9.2 Paired t-test

For paired observations, calculate differences $d_i = x_i - y_i$. Test the mean difference μ_d using

$$t = \frac{d\ | - d_0}{\frac{s_d}{\sqrt{n}}}, \quad \text{df} = n - 1.$$

Pairing controls for variation between matched units, such as the same users measured before and after a system change.

9.3 Two-sample t-test

For two independent means, Welch's test is generally safer when variances may differ:

$$t = \frac{(x\ |_1 - x\ |_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The degrees of freedom are approximated by the Welch–Satterthwaite formula. A pooled-variance t-test assumes equal population variances and should not be used without justification.

Worked example: one-sample t-test

A sample of $n = 16$ jobs has mean runtime $x \mid = 52$ seconds and sample standard deviation $s = 8$ seconds. To test $H_0 : \mu = 50$, $t = \frac{52-50}{\frac{8}{4}} = 1$. The degrees of freedom are 15. The result is not extreme for a two-sided 5% test, so fail to reject H_0.

10 9. Chi-square tests

The chi-square distribution is nonnegative and depends on degrees of freedom. It is used for tests involving counts and categorical data, and for inference about a population variance under Normality.

10.1 Goodness-of-fit test

This test compares observed counts O_i with expected counts E_i under a proposed categorical distribution:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

If k categories are specified without estimating parameters from the data, degrees of freedom are typically $k - 1$. If parameters are estimated, subtract the number of estimated independent parameters.
Expected counts should generally be sufficiently large; combine sparse categories when scientifically reasonable.

10.2 Test of independence

For a contingency table, the expected count in cell (i, j) under independence is

$$E_{ij} = \frac{(\text{row total})_{i(\text{column total})_j}}{n}.$$

Then calculate $\chi^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ with degrees of freedom

$$\text{df} = (r - 1)(c - 1).$$

A significant result indicates association between the categorical variables, not necessarily causation.

Worked example: independence

Suppose a table records browser type by failure status. Compute expected counts from row and column totals, add the cell contributions $\frac{(O-E)^2}{E}$, and compare the result with a chi-square distribution having $(r - 1)(c - 1)$ degrees of freedom. A small p-value indicates browser type and failure status are not independent in the population model.

10.3 Test for a population variance

For a Normal population, testing $H_0 : \sigma^2 = \sigma_0^2$ uses

$$\chi^2 = \frac{(n - 1)s^2}{\sigma_0^2}, \quad \text{df} = n - 1.$$

This test is sensitive to departures from Normality, so it is less robust than many mean-based methods.

11 10. Type I and Type II errors

Reality / decision	**Reject H_0**	**Fail to reject H_0**
H_0 true	Type I error; probability α	Correct decision

H_0 false	Correct decision; power	Type II error; probability β

- **Type I error**: reject a true null hypothesis; its maximum planned probability is α.
- **Type II error**: fail to reject a false null hypothesis; probability is β.
- **Power**: $1 - \beta$, the probability of detecting a specified real effect.

Power generally increases with larger sample size, larger effect size, lower variability, and larger α. Lowering α makes Type I errors less likely but can reduce power unless sample size increases.

12 11. Statistical significance and practical significance

A statistically significant result means the data provide evidence against the null model at the selected threshold. It does not tell us whether the effect is important, useful, or causal.

Always report:

- estimated effect and its direction;
- confidence interval or uncertainty measure;
- sample size and design;
- p-value and pre-specified significance level;
- practical or domain-specific importance;
- relevant assumptions and limitations.

A very large sample can make a tiny effect statistically significant. A practically important effect can fail to reach significance in a small or noisy sample.

13 12. Multiple testing and responsible interpretation

When many hypotheses are tested, the chance of at least one false positive increases. Possible responses include pre-specifying a primary hypothesis, controlling the familywise error rate, applying procedures such as Bonferroni correction, or controlling the false discovery rate.

Do not report only the most favorable analysis. Distinguish exploratory findings, which generate hypotheses, from confirmatory analyses with pre-specified questions and decisions.

14 13. Computing applications and case studies

- **Performance regression**: test whether mean latency changed after an optimization, using paired measurements or independent samples as appropriate.
- **A/B testing**: compare conversion proportions, while considering randomization, exposure, multiple metrics, and practical effect size.
- **System reliability**: test whether a failure rate differs from a service-level target.
- **Categorical monitoring**: use a chi-square test for association between error type and platform.
- **Model evaluation**: compare accuracy or error rates using paired observations when the same test cases are evaluated by two systems.
- **Data quality**: construct confidence intervals for missingness, defect, or incident proportions.

The test must match the data-generating process. Repeated observations from the same user, machine, or time series may violate independence and require paired, clustered, or time-series methods.

15 14. Problem-solving checklist

1. Identify the population parameter and the sample statistic.
2. State H_0, H_a, and whether the test is one- or two-sided.
3. Choose α before looking at the test result.
4. Check sampling design, independence, scale, and distributional assumptions.
5. Select Z, t, or chi-square based on the parameter and assumptions.
6. Calculate the test statistic, degrees of freedom, and p-value.
7. Compare p-value with α or statistic with the critical value.
8. State the decision in context without claiming proof or causality.
9. Include effect size and confidence interval when possible.

16 15. Quick formula sheet

Concept	Formula
Mean standard error	$\mathrm{SE}(x\ \vert) = \frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$
Known-σ mean CI	$x\ \vert + -z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
Unknown-σ mean CI	$x\ \vert + -t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$
Proportion estimate	$\hat{p} = \frac{x}{n}$
One-sample Z statistic	$z = \frac{x\ \vert -\mu_0}{\frac{\sigma}{\sqrt{n}}}$
One-sample t statistic	$t = \frac{x\ \vert -\mu_0}{\frac{s}{\sqrt{n}}}$, $\mathrm{df} = n-1$
Chi-square goodness-of-fit	$\chi^2 = \frac{\sum_{i(O_i-E_i)}}{E_i}$
Chi-square independence	$\mathrm{df} = (r-1)(c-1)$
Power	$1-\beta$

17 16. Practice questions with answers

1. A random sample of 100 users has mean session time 42 minutes and known $\sigma = 10$. Find a 95% CI. **Answer:** $42 + -1.96\left(\frac{10}{10}\right) = (40.04, 43.96)$ minutes.
2. A sample of 25 jobs has $x\ \vert = 8.4$ s and $s = 2$ s. Which test is suitable for $H_0 : \mu = 8$ s? **Answer:** One-sample t-test with $\mathrm{df} = 24$, assuming independent approximately Normal observations.
3. Interpret a p-value of 0.03 for a two-sided test at $\alpha = 0.05$. **Answer:** Reject H_0 at the 5% level; the data would be relatively unusual under H_0. It is not the probability that H_0 is true.
4. What is a Type I error in an A/B test? **Answer:** Concluding that the variants differ when the true population conversion rates are equal.
5. In a chi-square independence test with 4 rows and 3 columns, find df. **Answer:** $(4-1)(3-1) = 6$.
6. Why can a statistically significant result be practically unimportant? **Answer:** With a large sample, even a very small effect can produce a small p-value; inspect the effect size and confidence interval.
7. A confidence interval becomes wider when confidence level increases. Why? **Answer:** Greater coverage requires a larger critical value and therefore a larger margin of error.

Exam reminder. A complete hypothesis-test answer includes hypotheses, significance level, test statistic, degrees of freedom when relevant, p-value or critical comparison, decision, and a sentence interpreting the result in context.