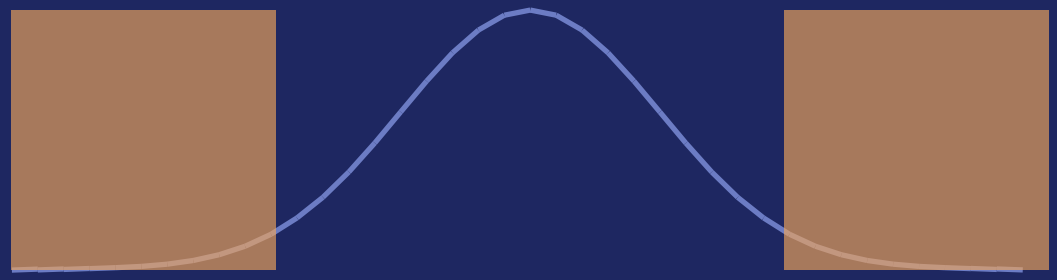
MODULE 5

Testing of Hypothesis

Track: Internet of Things / Cyber Security / Blockchain Technology



...deciding whether an observed difference is real, or just chance.

What We'll Cover

From the logic of a hypothesis test to eight concrete tests you can run

1. **The Logic of Hypothesis Testing**
   H0, H1, errors, and the general procedure
2. **Large-Sample (Z) Tests**
   Means, proportions, and their differences
3. **Small-Sample (t) Tests**
   Means and differences, when n is small
4. **Correlation & Variance Ratio**
   Testing a correlation coefficient, and the F-test
5. **Chi-Square Tests**
   Goodness of fit, and independence of attributes
6. **Putting It Together**
   Choosing the right test, pitfalls, and applications

What Is Hypothesis Testing?

A formal procedure for deciding between two competing claims about a population

The core question: we have a claim about a population parameter (a mean, a proportion, a variance). Sample data gives us some evidence. Is that evidence strong enough to reject the claim — or could the observed difference just be due to random sampling variation?

IoT	Does a new firmware actually reduce average power draw, or did this batch just get lucky sensors?
Cyber Security	Is the new detection rule really catching more threats, or is the improvement within normal noise?
Blockchain	Did a protocol change actually reduce block confirmation time, or does it just look that way in this sample?

Null and Alternative Hypotheses

Every test starts with two competing, precisely stated claims

Null Hypothesis H_0

The default, “no effect” or “no difference” claim — assumed true until evidence says otherwise. Always stated with an equality.

e.g. $H_0: \mu = 500$

Alternative Hypothesis H_1

What we conclude if the evidence is strong enough to reject H_0 — the claim we're actually trying to find support for.

e.g. $H_1: \mu \neq 500$

Key rule: we never “prove H_0 true.” We either reject H_0 in favor of H_1, or fail to reject H_0 — which is not the same as accepting it. Absence of evidence is not evidence of absence.

Analogy: like a criminal trial — H_0 is “innocent until proven guilty.” We reject innocence only if the evidence is strong enough, beyond reasonable doubt.

Type I and Type II Errors

Every decision rule carries two distinct risks of being wrong

	H_0 is actually True	H_0 is actually False
Reject H_0	Type I Error (α)	Correct decision
Fail to Reject H_0	Correct decision	Type II Error (β)

α (significance level)

P(Type I Error) — the probability we reject a true H_0; we choose this in advance (often 0.05)

β

P(Type II Error) — the probability we fail to reject a false H_0; harder to control directly

Power = 1 − β

the probability of correctly rejecting a false H_0 — higher power is better

The trade-off

lowering α (fewer false alarms) generally raises β (more missed real effects), for a fixed sample size

The General Hypothesis-Testing Procedure

Five steps that every test in this module follows

1	**State H_0 and H_1**	Write both hypotheses precisely, in terms of a population parameter
2	**Choose α and the test statistic**	Pick a significance level, and the formula appropriate to the situation (z, t, F, or χ^2)
3	**Compute the test statistic**	Plug the sample data into the formula
4	**Find the critical value (or p-value)**	Look up the boundary of the rejection region for the chosen α
5	**Decide and interpret**	Compare the statistic to the critical value; reject or fail to reject H_o, and state the conclusion in context

One-Tailed vs. Two-Tailed Tests

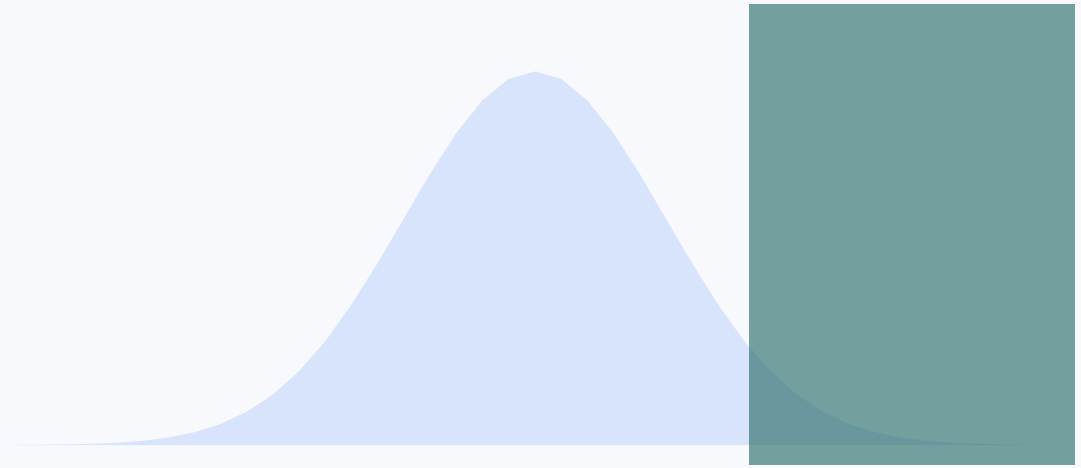
The direction of H_1 determines where the rejection region sits

Two-Tailed

$H_1: \mu \neq \mu_o$

One-Tailed (right)

$H_1: \mu > \mu_o$



How to choose: use a two-tailed test when H_1 simply claims "different from" (≠) — the rejection region splits between both tails. Use a one-tailed test only when H_1 specifically claims a direction ("greater than" or "less than") — decided before looking at the data, never after.

Critical Region, Critical Value & p-value

Two equivalent ways to reach the same decision

Critical Value Approach

1. Find the boundary value(s) that mark off α in the tail(s) of the test statistic's distribution.

2. If the computed statistic falls beyond that boundary (in the critical region), reject H_o.

p-value Approach

1. Compute the probability of getting a result at least as extreme as ours, if H_o were true.

2. If p-value < α, reject H_o. Smaller p-values are stronger evidence against H_o.

The two always agree: “test statistic is beyond the critical value” and “p-value < α” are two descriptions of exactly the same event — this module uses the critical-value approach throughout, since it's the more common approach when computing everything by hand.

Common mistake: a large p-value never proves H_o true — it only means the data didn't give enough evidence to reject it.

Large-Sample Z-Test for a Single Mean

EXAMPLE

Testing whether a population mean equals a claimed value, using a known σ

Question: A manufacturer claims average IoT sensor lifespan is 500 days (σ = 70, known). A sample of n = 49 sensors gives $\bar{x}$ = 485 days. Test H_0: μ = 500 vs H_1: μ ≠ 500 at α = 0.05.

Step-by-Step

1. **Standard error:** SE = 70/√49 = 10
2. **Test statistic:** z = (485−500)/10 = −1.50
3. **Critical value:** z(0.025) = ±1.96

|z| = 1.50 < 1.96 → Fail to Reject H_0

Interpretation: Not enough evidence at the 5% level to say the true mean lifespan differs from the manufacturer's claimed 500 days.

Large-Sample Z-Test for Difference of Two Means

EXAMPLE

Comparing two population means using large, independent samples

Question: IDS A ($n_1 = 100$) has mean response time 45 ms, $\sigma_1 = 12$. IDS B ($n_2 = 120$) has mean 42 ms, $\sigma_2 = 10$. Test H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 \neq \mu_2$ at $\alpha = 0.05$.

Step-by-Step

1. Standard error: $SE = \sqrt{(12^2/100 + 10^2/120)} = 1.51$
2. Test statistic: $z = (45-42)/1.51 = 1.99$
3. Critical value: $z(0.025) = \pm 1.96$

$|z| = 1.99 > 1.96 \rightarrow$ Reject H_0

Interpretation: At the 5% level, the two IDS systems' mean response times are significantly different — though the result sits right at the boundary of significance.

Large-Sample Z-Test for a Single Proportion

EXAMPLE

Testing whether a population proportion equals a claimed value

Question: A security team suspects more than the claimed 10% of traffic is improperly encrypted. A sample of n = 400 packets finds 52 improperly encrypted ($\hat{p}$ = 0.13). Test H_0: P = 0.10 vs H_1: P ≠ 0.10 at α = 0.05.

Step-by-Step

1. **Standard error:** SE = √(0.10×0.90/400) = 0.015
2. **Test statistic:** z = (0.13−0.10)/0.015 = 2.00
3. **Critical value:** z(0.025) = ±1.96

|z| = 2.00 > 1.96 → Reject H_0

Interpretation: Significant evidence the true improperly-encrypted rate differs from the claimed 10%.

Large-Sample Z-Test for Difference of Two Proportions

EXAMPLE

Comparing malicious-traffic rates between two network segments

Question: Segment A: $n_1 = 500$, 40 malicious packets ($\hat{p}_1 = 0.08$). Segment B: $n_2 = 600$, 54 malicious packets ($\hat{p}_2 = 0.09$). Test H_0: $P_1 = P_2$ at $\alpha = 0.05$.

Step-by-Step

1. **Pooled proportion:** $\hat{p} = (40+54)/1100 = 0.0855$
2. **Standard error:** $SE = \sqrt{(0.0855 \times 0.9145 \times (1/500+1/600))} = 0.0169$
3. **Test statistic:** $z = (0.08-0.09)/0.0169 = -0.59$

$|z| = 0.59 < 1.96 \rightarrow$ **Fail to Reject H_0**

Interpretation: No significant evidence the malicious-traffic rate differs between the two network segments.

Test for Difference of Standard Deviations

The large-sample counterpart to comparing two means, applied to spread instead

$$z = (s_1 - s_2) / \sqrt{(\sigma_1^2/2n_1 + \sigma_2^2/2n_2)}$$

H_0

$\sigma_1 = \sigma_2$ (the two populations are equally variable)

When σ_1, σ_2 unknown

substitute the sample standard deviations s_1, s_2 for large n

Used for

comparing consistency/variability, not central tendency — e.g. which of two sensors is more erratic

Small-sample version

use the F-test (covered later in this module) instead of this z-test

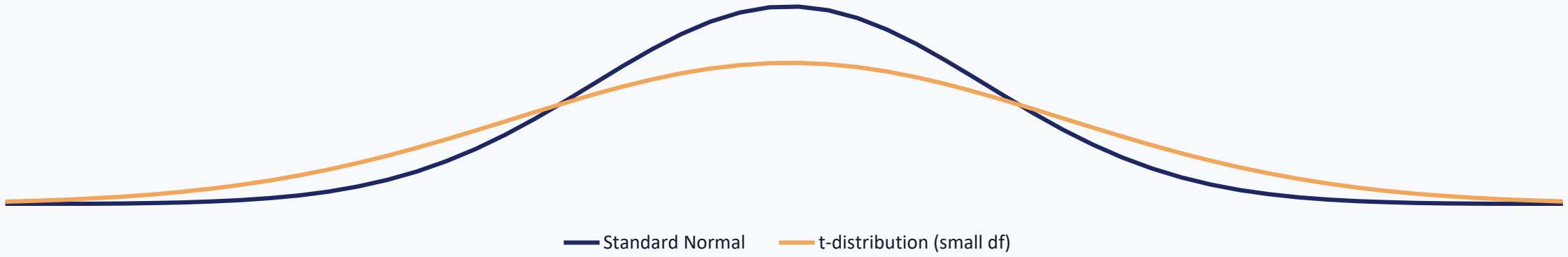
Example scenario: comparing the standard deviation of confirmation times for two blockchain networks — not just whether they average the same, but whether one is more erratic than the other. The mechanics mirror the difference-of-means test exactly.

Small Samples and the t-Distribution

When σ is unknown and n is small, the Normal approximation breaks down

The problem: for large n, replacing σ with s in a z-test works well. But for small n, s itself is a noisy estimate of σ — that extra uncertainty needs to be built into the distribution of the test statistic.

The t-distribution (df small) has heavier tails than the standard Normal



As n grows: the t-distribution's df = n−1 grows too, its tails shrink, and it converges to the standard Normal — by around n = 30 the two are nearly indistinguishable.

Small-Sample t-Test for a Single Mean

EXAMPLE

When σ is unknown and n is small, the t-distribution replaces the Normal

Question: A calibration check on n = 12 sensors gives $\bar{x} = 2.3$ against a target $\mu = 2.0$, with sample $s = 0.6$. Test $H_0: \mu = 2.0$ vs $H_1: \mu \neq 2.0$ at $\alpha = 0.05$.

Step-by-Step

1. **Standard error:** $SE = s/\sqrt{n} = 0.6/\sqrt{12} = 0.173$
2. **Test statistic:** $t = (2.3-2.0)/0.173 = 1.73$
3. **Critical value:** $df = 11$, $t(0.025,11) = 2.201$

$$|t| = 1.73 < 2.201 \rightarrow \text{Fail to Reject } H_0$$

Interpretation: Not enough evidence the sensors' true calibration mean differs from the 2.0 target.

Small-Sample t-Test for Difference of Two Means

EXAMPLE

Pooled-variance t-test, assuming the two population variances are equal

Question: Blockchain config A: $n_1 = 10$, $\bar{x}_1 = 12.4$, $s_1 = 2.1$. Config B: $n_2 = 12$, $\bar{x}_2 = 10.8$, $s_2 = 2.4$. Test H_0: $\mu_1 = \mu_2$ at $\alpha = 0.05$.

Step-by-Step

1. Pooled variance: $s_p^2 = [9\times4.41 + 11\times5.76]/20 = 5.15$
2. Standard error: $SE = \sqrt{5.15} \times \sqrt{(1/10+1/12)} = 0.972$
3. Test statistic: $t = (12.4-10.8)/0.972 = 1.65$
4. Critical value: $df = 20$, $t(0.025,20) = 2.086$

$|t| = 1.65 < 2.086 \rightarrow$ Fail to Reject H_0

Interpretation: No significant evidence the two configurations' mean confirmation times differ.

Test for a Correlation Coefficient

EXAMPLE

Is a sample correlation strong enough to conclude a real relationship exists?

Question: A sample of n = 15 paired observations gives sample correlation r = 0.65. Test H_0: ρ = 0 vs H_1: ρ ≠ 0 at α = 0.05.

Step-by-Step

1. **Test statistic:** $t = r\sqrt{(n-2)} / \sqrt{(1-r^2)}$
2. **Compute:** $= 0.65 \times \sqrt{13} / \sqrt{0.5775} = 3.08$
3. **Critical value:** df = 13, $t(0.025,13) = 2.160$

$|t| = 3.08 > 2.160 \rightarrow$ Reject H_0

Interpretation: The correlation is statistically significant — the two variables are genuinely related, not just by chance in this sample.

F-Test for the Ratio of Two Variances

EXAMPLE

Testing whether two populations have equal variability

Question: Batch A: $n_1 = 10$, $s_1^2 = 25$. Batch B: $n_2 = 13$, $s_2^2 = 12$. Test $H_0: \sigma_1^2 = \sigma_2^2$ vs $H_1: \sigma_1^2 > \sigma_2^2$ at $\alpha = 0.05$.

Step-by-Step

1. **Test statistic:** $F = s_1^2/s_2^2 = 25/12 = 2.08$
2. **Degrees of freedom:** $df_1 = 9$, $df_2 = 12$
3. **Critical value:** $F(0.05, 9, 12) = 2.80$

$F = 2.08 < 2.80 \rightarrow$ **Fail to Reject H_0**

Interpretation: No significant evidence that Batch A's variance is larger than Batch B's.

Chi-Square Test for Goodness of Fit

EXAMPLE

Does an observed distribution match an expected or historical pattern?

Question: Observed traffic (n = 400): HTTP = 120, HTTPS = 150, DNS = 80, Other = 50. Historically each type is expected at 25% (E = 100 each). Test H_o: the distribution matches history.

Step-by-Step

1. Test statistic: $\chi^2 = \Sigma(O-E)^2/E = 4 + 25 + 4 + 25$
2. Compute: $= 58.00$
3. Critical value: $df = 3,\ \chi^2(0.05,3) = 7.815$

$58.00 > 7.815 \rightarrow$ Reject H_o

Interpretation: Today's traffic mix is significantly different from the historical 25%-each pattern.

Chi-Square Test of Independence

EXAMPLE

Are two categorical attributes associated, or independent?

400 devices, cross-tabulated by type and infection status:

	Infected	Not Infected	Total
IoT	40	160	200
Non-IoT	15	185	200
Total	**55**	**345**	**400**

H_o: device type and infection status are independent — expected counts assume the same infection rate across both device types.

Step-by-Step

1. **Expected counts:** $E = 27.5, 172.5, 27.5, 172.5$
2. $\chi^2 = \Sigma(O-E)^2/E$: $5.68+0.91+5.68+0.91 = 13.18$

Conclusion

Degrees of freedom:

$df = (2-1)(2-1) = 1$

Critical value:

$\chi^2(0.05,1) = 3.841$

$13.18 > 3.841 \rightarrow$ Reject H_o

Device type and infection status are significantly associated — IoT devices show a different infection rate than non-IoT devices.

Statistical Significance vs. Practical Significance

A significant result is not automatically an important one

The trap: with a large enough sample, even a tiny, practically meaningless difference can become “statistically significant” — because standard error shrinks as n grows, making smaller and smaller true differences detectable.

Example: with n = 1,000,000 sensors, a mean battery life difference of just 0.2 days between two firmware versions could easily be statistically significant — but is 0.2 days out of ~500 days worth a firmware rollout?

Always ask

“How big is the effect?” (effect size), not just “is p < 0.05?”

Report both

the test result AND the actual estimated difference, with its confidence interval

Choosing the Right Test

One table to navigate every test in this module

Question	Sample Size	Test
One mean vs. a claimed value	Large (σ known)	**Z-test (single mean)**
One mean vs. a claimed value	Small (σ unknown)	**t-test (single mean)**
Two means, compared	Large	**Z-test (difference of means)**
Two means, compared	Small	**t-test (pooled variance)**
One proportion vs. a claim	Large	**Z-test (single proportion)**
Two proportions, compared	Large	**Z-test (difference of proportions)**
Two variances, compared	Small	**F-test**
A correlation, significant?	Any	**t-test for correlation**
Category counts vs. expected	Any	**Chi-square goodness of fit**
Two categorical attributes, related?	Any	**Chi-square independence**

Common Pitfalls to Avoid

Mistakes that show up again and again in exams and in practice

Choosing the tail after seeing data

the direction of H_1 (one-tailed vs two) must be decided before looking at the sample — never after

"Accepting" H_0

failing to reject H_0 is not the same as proving it true — it just means insufficient evidence to reject

Using z when n is small

with an unknown σ and small n, use the t-distribution, not a z-test — the tails are meaningfully different

Ignoring the equal-variance assumption

the pooled-variance t-test assumes $\sigma_1 = \sigma_2$ — check this (e.g. with an F-test) before applying it

Chasing significance, not size

a significant result with a tiny effect size may not be practically meaningful — see the previous slide

Multiple testing

running many tests at once inflates the overall chance of at least one false positive, even at $\alpha = 0.05$ each

Hypothesis Testing in Practice

The same tools, applied across your specialization tracks

Internet of Things

Z / t-test

Confirming a firmware update actually changed average power draw

Proportion test

Testing whether a fleet's failure rate exceeds a warranty threshold

F-test / χ^2

Comparing measurement consistency between two sensor batches

Cyber Security

Z / t-test

Testing whether a new rule changed average detection response time

Proportion test

Comparing malicious-traffic rates before and after a policy change

χ^2 independence

Testing whether attack type is associated with time of day

Blockchain

Z / t-test

Testing whether a protocol change reduced average confirmation time

Proportion test

Comparing transaction failure rates across two network versions

χ^2 goodness of fit

Testing whether mined-block timestamps follow the expected pattern

Key Takeaways

1. Hypothesis testing formally weighs sample evidence against a default claim H_o, using a test statistic compared to a critical value.

2. Type I error (α) and Type II error (β) are two distinct risks; lowering one generally raises the other for a fixed sample size.

3. Large samples use z-tests (mean, proportion, and their differences); small samples with unknown σ use t-tests instead.

4. The F-test compares two variances; chi-square tests compare observed counts to expected counts, or check independence of two attributes.

5. Statistical significance is not the same as practical importance — always look at the effect size alongside the test result.

6. This completes BSM301's core toolkit — from distributions, through curve fitting, sampling, estimation, to formal hypothesis testing.