BP701T · B.Pharm Semester VII · Theory Notes

Biostatistics and Research Methodology

Unit IV — Inferential Statistics (12 hours)

Unit IV (12 hours)

Pills & Python Notes

From-zero teaching standard: every technical term is defined before use.

Syllabus coverage: Unit IV (12 hours).

PILLS & PYTHON

Pharmacy • Data • AI • Healthcare

Instagram @pillsandpython · pillsandpython.com · YouTube @PillsAndPython

CHAPTER 9: Statistical Estimation and Confidence Intervals

Learning Objectives

- Define statistical estimation and distinguish point estimates from interval estimates
- Explain what a population parameter and a sample statistic are
- Define a confidence interval and interpret it correctly
- Construct confidence intervals for a population mean using t values
- Explain when and why the t-distribution (instead of the normal z) is used
- Solve numerical problems on generating confidence intervals for means
- State the effect of sample size and confidence level on interval width
- Apply confidence-interval thinking to pharmaceutical examples (assay values, tablet weights)

Definitions

- **Statistical estimation** (simple meaning): Using the data from a small sample to make an educated guess about a characteristic of the whole population. We can never measure every tablet in a million-tablet batch, so we measure a few and *estimate* the batch's true average.
- **Population**: The entire group we care about (e.g., all tablets in a batch, all patients who could take a drug).
- **Population parameter**: A true, fixed (but usually unknown) number describing the population — e.g., the true mean weight μ (mu) of all tablets in the batch, or the true population standard deviation σ (sigma).
- **Sample statistic**: A number computed from the sample data that we use to guess the parameter — e.g., the sample mean $\bar{x}$ guesses μ; the sample standard deviation s guesses σ.
- **Point estimate**: A single-number guess of the parameter (e.g., "the batch mean is 500.4 mg").
- **Interval estimate**: A range of numbers ("from ... to ...") that we believe contains the true parameter, stated with a confidence level (e.g., "we are 95% confident the true mean lies between 498.1 mg and 502.7 mg").
- **Confidence interval (CI)**: An interval estimate of the form: point estimate ± margin of error, computed so that a stated percentage (e.g., 95%) of such intervals, built from repeated samples, would contain the true parameter.
- **Confidence level**: The percentage (90%, 95%, 99%) expressing how sure the method is — 95% means the procedure captures the true parameter in 95 out of 100 repeated samples.
- **Margin of error (E)**: The "±" part of the interval — half the width of the confidence interval.

- **t-distribution / t value**: A bell-shaped distribution like the normal curve but with heavier tails, used when the sample is small and the population standard deviation is unknown. The t critical value (t*) comes from a standard t table using degrees of freedom (df = n − 1).

Terminology

Term	Simple meaning
Population	The whole group of interest
Sample	The small part actually measured
Parameter (μ, σ)	True population number (unknown, fixed) — Greek letters
Statistic ($\bar{x}$, s)	Sample number used to guess the parameter — Roman letters
Point estimate	One-number guess (e.g., $\bar{x}$ = 500.4 mg)
Interval estimate	Range guess ("498 to 503 mg")
Confidence interval	Interval estimate with a stated confidence level
Confidence level	% of repeated intervals that would capture the parameter (95% usual)
Margin of error	The ± amount added/subtracted to the point estimate
Standard error of the mean (SE)	s/√n — how much sample means wobble around μ
t critical value (t*)	Multiplier from the t table for chosen confidence level and df
Degrees of freedom (df)	n − 1 for a single mean; row selector in the t table
Lower / upper confidence limit	The two ends of the interval

Simple Explanations — Starting from Zero

Technical concept: Statistical estimation infers population parameters from sample statistics.

Simple explanation: A "population" is everything you care about — say, all 200,000 tablets in a production batch. You cannot weigh every tablet; that would destroy the batch and take forever. So you weigh a *sample* — maybe 20 tablets. From the sample you compute a *statistic* (the sample mean $\bar{x}$). That statistic is your best single guess of the *parameter* (the true batch mean μ, which you can never know exactly). This leap from sample to population is called *statistical estimation*. The key honesty of statistics: we always report how uncertain the guess is.

Example: You weigh 20 tablets; the sample mean is 500.4 mg. That 500.4 mg is a *point estimate* of the true batch mean μ. But another sample of 20 might give 499.8 mg. So instead of one number, we give a range: "we are 95% confident the true batch mean is between 498.1 mg and 502.7 mg." That range is the *interval estimate* — the confidence interval.

Diagram/flowchart:

```
POPULATION (all 200,000 tablets)          SAMPLE (20 tablets weighed)
True mean μ = ??? (unknown, fixed)   →    Sample mean x¯ = 500.4 mg (known)
True SD σ = ??? (unknown)            →    Sample SD s = 4.6 mg (known)
                                                  |
                                     POINT ESTIMATE: μ ≈ 500.4 mg
                                                  |
                                     INTERVAL ESTIMATE (95% CI):
                                     500.4 ± margin of error
                                     = (498.1 mg, 502.7 mg)
                                     "95% confident μ is inside"
```

Key takeaway: We measure a few, then estimate the whole — a point estimate gives one number, an interval estimate honestly shows the uncertainty.

Technical concept: Confidence intervals — the idea and the correct interpretation.

Simple explanation: A 95% confidence interval is built by a *procedure* that works 95% of the time: if you repeated the sampling-and-interval-building process 100 times, about 95 of the resulting intervals would contain the true parameter, and about 5 would miss it. The confidence is in the *method*, not in any single interval. Correct statement: "We are 95% confident that the true population mean lies between L and U." WRONG statement: "There is a 95% probability that μ is between L and U" — μ is fixed, not random; it is either inside or outside. Also wrong: "95% of the tablets weigh between L and U" — the interval is about the *mean*, not about individual tablets.

Example: A 95% CI for mean tablet hardness is (6.2 kp, 7.0 kp). Correct reading: the procedure that made this interval captures the true mean hardness in 95% of repeated uses; the true mean is plausibly anywhere in 6.2–7.0 kp. It does NOT mean 95% of individual tablets have hardness in that range (individual tablets vary much more widely).

Diagram/flowchart:

```
Imagine 20 different samples, 20 intervals built the same way:

Sample 1:   |--------●--------|   contains μ ✓
Sample 2:        |-----●-----|      contains μ ✓
Sample 3:   |-----●-----|           misses μ ✗ (rare!)
...
Sample 20:      |--------●--------|

About 95% contain the true μ — that is what
"95% confident" means. It describes the method's
long-run success rate, not one interval's probability.
```

Key takeaway: "95% confident" = the interval-building recipe succeeds about 95 times in 100; it is not a probability about one fixed interval.

Technical concept: Confidence interval for a population mean using t values.

Simple explanation: When the sample is small (roughly n < 30) and we don't know the population standard deviation σ (the usual real-life case), sample means vary a bit more than the normal curve predicts. The *t-distribution* accounts for this extra wobble — it looks like the normal bell but with thicker

tails, and it gets closer to normal as n grows. The formula for the confidence interval of the mean is:

$\bar{x} \pm t^* \times (s / \sqrt{n})$

where t *is the* critical value *read from a standard t table for your chosen confidence level and degrees of freedom df = n − 1. The quantity s/√n is the* standard error of the mean* (SE) — it measures how much sample means typically bounce around the true μ. Bigger n → smaller SE → narrower, more precise interval.

Example: n = 16 tablets, $\bar{x}$ = 250.0 mg, s = 8.0 mg, 95% confidence. df = 15; from the standard t table, t* = 2.131. SE = 8/√16 = 8/4 = 2.0 mg. Margin of error = 2.131 × 2.0 = 4.262 mg. CI = 250.0 ± 4.26 = (245.74 mg, 254.26 mg).

Diagram/flowchart:

```
GOAL: 95% CI for population mean μ (σ unknown, small n)
                  |
1. Compute x̄ (sample mean) and s (sample SD)
                  |
2. df = n − 1; look up t* in the standard t table
   (column = confidence level, row = df)
                  |
3. Standard error SE = s / √n
                  |
4. Margin of error E = t* × SE
                  |
5. CI = (x̄ − E, x̄ + E)
                  |
6. Interpret: "95% confident the true mean μ
   lies between ___ and ___."
```

Key takeaway: Mean CI = sample mean ± (t-table value × s/√n); the t value corrects for the extra uncertainty of small samples.

Technical concept: What makes a confidence interval wider or narrower.

Simple explanation: Three things control the width: (1) **Confidence level** — demanding 99% instead of 95% confidence needs a bigger t *and gives a wider interval (more confidence costs more width). (2)* ***Sample size n*** *— more data shrinks s/√n, narrowing the interval (precision is bought with sample size). (3)* ***Variability s*** *— noisier data (bigger s) widens the interval. Rule of thumb: to halve the width, you must roughly* quadruple* the sample size (because of the √n).

Example: With n = 16, the margin of error above was 4.26 mg. With n = 64 (four times more tablets) and the same s, SE halves to 1.0 mg and the margin roughly halves too — the interval gets about twice as precise.

Key takeaway: Higher confidence → wider; bigger sample → narrower; noisier data → wider. Width is the price of certainty.

Classifications

Types of estimates:

- **Point estimate:** single value ($\bar{x}$ for μ, s for σ, $\hat{p}$ for a proportion).
- **Interval estimate:** range with confidence level — the confidence interval.

Common confidence levels:

- **90% CI:** narrower; used when a little more risk is acceptable.
- **95% CI:** the standard workhorse in pharmacy and health sciences.
- **99% CI:** wider; used when high assurance is needed (e.g., safety-critical decisions).

When to use t vs z (normal):

- **t-distribution:** σ unknown and estimated by s (the usual case), especially with small samples — always safe for means in this course.
- **z (normal):** σ known, or very large n (t and z become nearly identical beyond about n = 100).

Processes

Process 1: Building a confidence interval for the mean using t values (step-by-step)

1 From the sample compute n, $\bar{x}$ and s.

2 Choose the confidence level (usually 95%).

3 Compute degrees of freedom: df = n − 1.

4 Read the critical value t* from a standard t table (column = confidence level, row = df).

5 Compute the standard error: SE = s / √n.

6 Compute the margin of error: E = t* × SE.

7 Form the interval: lower limit = $\bar{x}$ − E, upper limit = $\bar{x}$ + E.

8 State the conclusion: "We are ___% confident that the true population mean lies between ___ and ___."

Process 2: Reading a t table (step-by-step)

1 Find the column for your confidence level (95% → the "0.05 two-tailed" column in most tables).

2 Find the row for df = n − 1.

3 The number at their crossing is t. *Example: 95% confidence, df = 9 → t = 2.262 (as per standard t table).*

Process 3: Checking whether a target value is plausible (step-by-step)

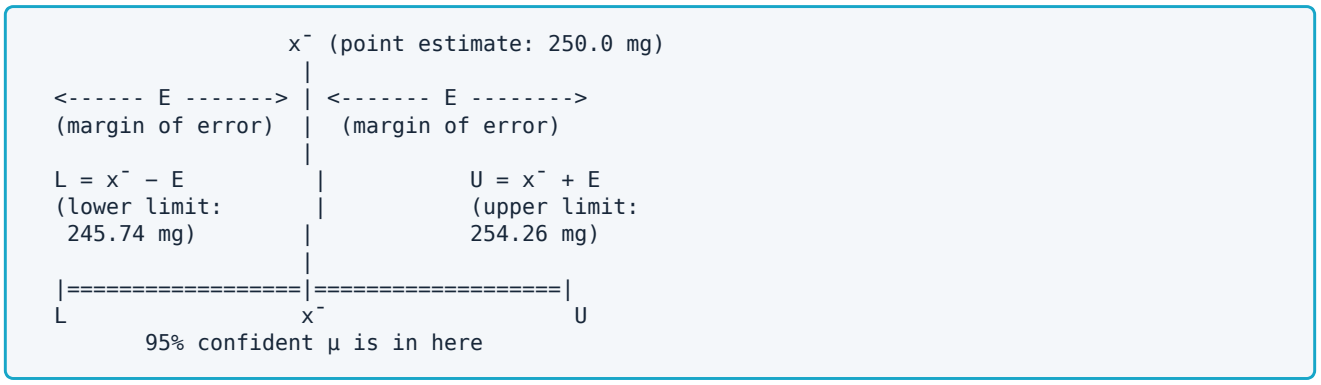
1 Build the confidence interval for the mean.

2 Ask: does the target/specification value fall inside the interval?

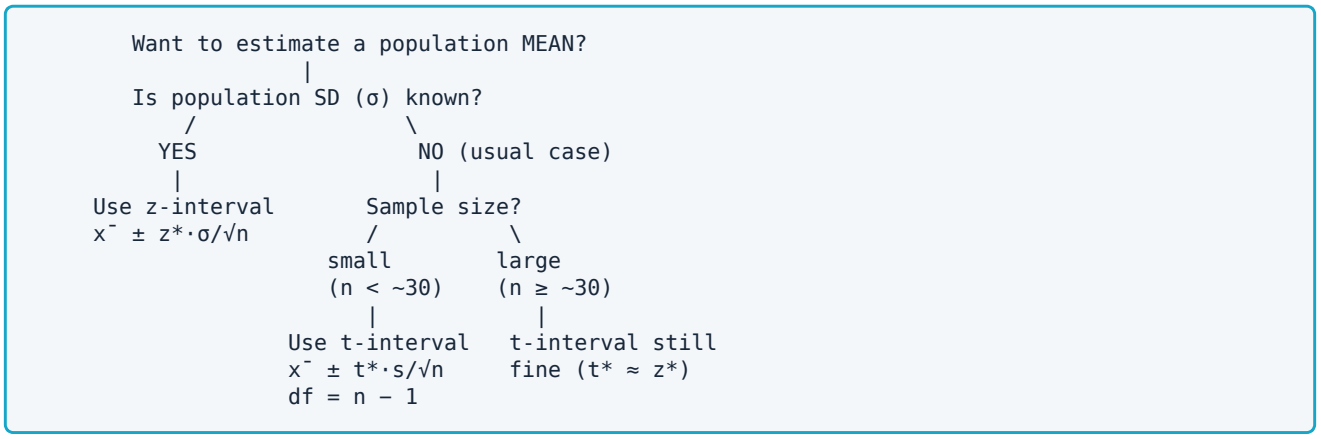
3 If yes — the data are *consistent* with the target (not proof of equality). If no — the data suggest the true mean differs from the target.

Textual Diagrams and Flowcharts

Anatomy of a confidence interval



Choosing the right tool — flowchart



> INFOGRAPHIC CONCEPT: The Fisherman's Net

The Fisherman's Net. Draw a fisherman casting a net into a lake. The lake's true average fish weight (μ, the parameter) is a big fish hiding somewhere underwater — fixed but unseen. Each cast of the net is one sample; the net's span is the confidence interval, centred on the sample mean $\bar{x}$ (a float marking the net's middle). Show 20 casts: 19 nets trap the big fish (green nets, "captured"), 1 net misses it (red net, "missed — the unlucky 5%"). Caption: "A 95% confidence net catches the true mean in about 95 of 100 casts." Three sliders beside the fisherman control the net: "Confidence 90/95/99%" (higher = wider net), "Sample size n" (bigger = tighter, more precise net), "Data variability s" (noisier water = wider net). Warning tag on the net: "The net estimates the AVERAGE fish — it does not describe every individual fish."

Tables

Table 1: Point estimate vs interval estimate

Feature	Point estimate	Interval estimate (CI)
Form	Single number	Range (lower, upper)
Example	$\bar{x}$ = 500.4 mg	(498.1 mg, 502.7 mg), 95% confident

Feature	Point estimate	Interval estimate (CI)
Shows uncertainty?	No — looks falsely exact	Yes — width shows precision
Use	Quick best guess	Honest reporting, decisions, specifications

Table 2: Parameter vs statistic

	Population parameter	Sample statistic
Symbol (mean)	μ (mu)	$\bar{x}$ (x-bar)
Symbol (SD)	σ (sigma)	s
Nature	Fixed but unknown	Known, varies sample to sample
Example	True mean weight of all 200,000 tablets	Mean weight of the 20 tablets weighed

Table 3: Common t* values (as per standard t table; two-tailed)

df (n − 1)	90% confidence	95% confidence	99% confidence
5	2.015	2.571	4.032
9	1.833	2.262	3.250
15	1.753	2.131	2.947
24	1.711	2.064	2.797
29	1.699	2.045	2.756

Worked Examples and Calculation Boxes

CALCULATION BOX 1 — 95% CI for a mean using t, full working (pharmaceutical example)

Problem: A QC inspector weighs 10 tablets from a batch (label claim 500 mg). Weights (mg): 498, 502, 501, 499, 503, 500, 497, 502, 501, 499. Construct a 95% confidence interval for the true mean tablet weight μ. (Use: s = 1.94 mg as computed below; t* for df = 9 at 95% = 2.262 as per standard t table.)

Solution — every step shown:

1 n = 10. Σx = 498+502+501+499+503+500+497+502+501+499 = 5002. $\bar{x}$ = 5002/10 = **500.2 mg**.

2 Deviations squared: $(498-500.2)^2 = 4.84$; $(502-500.2)^2 = 3.24$; $(501-500.2)^2 = 0.64$; $(499-500.2)^2 = 1.44$; $(503-500.2)^2 = 7.84$; $(500-500.2)^2 = 0.04$; $(497-500.2)^2 = 10.24$; $(502-500.2)^2 = 3.24$; $(501-500.2)^2 = 0.64$; $(499-500.2)^2 = 1.44$. Sum = 33.60.

3 Sample variance $s^2 = 33.60/(10-1) = 33.60/9 = 3.733 \rightarrow s = \sqrt{3.733} \approx$ **1.93 mg**.

4 df = 10 − 1 = 9; 95% confidence → t* = **2.262** (standard t table).

5 SE = $s/\sqrt{n} = 1.93/\sqrt{10} = 1.93/3.162 \approx$ **0.610 mg**.

6 Margin of error E = t* × SE = 2.262 × 0.610 ≈ **1.38 mg**.

7 CI = 500.2 ± 1.38 = (**498.82 mg**, **501.58 mg**).

8 **Conclusion:** We are 95% confident that the true mean weight of all tablets in the batch lies between 498.8 mg and 501.6 mg. The label claim of 500 mg falls inside the interval — the data are consistent with the claim.

Key takeaway: Follow the eight steps — mean, SD, df, t*, SE, margin, interval, plain-language conclusion.

CALCULATION BOX 2 — 99% CI, full working (shows effect of confidence level)

Problem: Using the same tablet data (n = 10, $\bar{x}$ = 500.2 mg, s = 1.93 mg, SE = 0.610 mg), construct a 99% confidence interval for μ. (t* for df = 9 at 99% = 3.250 as per standard t table.)

Solution — every step shown:

1 Everything is the same except t*: df = 9, 99% confidence → t* = **3.250**.

2 E = 3.250 × 0.610 ≈ **1.98 mg**.

3 CI = 500.2 ± 1.98 = (**498.22 mg**, **502.18 mg**).

4 **Compare:** the 99% interval (width ≈ 3.96 mg) is wider than the 95% interval (width ≈ 2.76 mg). **Interpretation:** demanding more confidence forces a wider net — certainty costs precision.

Key takeaway: Same data, higher confidence → larger t* → wider interval. Choose 95% unless there is a reason for more (or less).

Example 3 — Effect of sample size (illustrative)

Suppose a second inspector weighs 40 tablets and finds the same $\bar{x}$ = 500.2 mg and s = 1.93 mg. Then SE = 1.93/√40 = 1.93/6.325 ≈ 0.305 mg (half the earlier SE), df = 39, t* ≈ 2.02, E ≈ 0.62 mg, and the 95% CI is (499.58 mg, 500.82 mg) — about half as wide as with n = 10. Quadrupling the sample roughly halved the width, as the √n rule predicts.

Key takeaway: Precision is bought with sample size — and the price follows √n, so gains slow down as n grows.

Analogies

- **The fisherman's net (see infographic):** Each sample casts a net (interval); 95 of 100 casts trap the true mean fish. One cast missing doesn't mean the method is broken — it's the expected 5%.
- **Weighing rice:** You can't weigh every grain in a sack to find the average grain weight — you weigh a handful (sample) and estimate. The confidence interval is your honest statement: "the average is about this much, give or take this much."

Pharmacy Relevance

- **Content uniformity / assay:** A 95% CI for mean drug content per tablet tells whether the batch truly meets the label claim — regulators think in intervals, not single numbers.

- **Tablet weight variation:** CI for mean weight supports batch release decisions: if the entire interval sits within specification limits, release is well justified.
- **Stability studies:** CIs for mean assay at each time point show the degradation trend with honest uncertainty bands.
- **Bioequivalence thinking:** Comparing confidence intervals of test vs reference product means is the conceptual backbone of bioequivalence assessment (formal bioequivalence uses 90% CIs on ratios — a topic for advanced study).
- **Clinical measurements:** A 95% CI for mean blood-pressure reduction tells a clinician the plausible range of the drug's true average effect, not just one trial's lucky number.

India Context

- Indian pharmaceutical QC laboratories report assay results with an understanding of uncertainty; confidence intervals formalise the "give or take" that analysts already discuss when a result sits near a specification limit.
- In pharmacopoeial testing performed across India, batch-acceptance decisions implicitly weigh sampling uncertainty — the CI framework makes that reasoning explicit and defensible during audits.
- Public-health surveys in India (e.g., district-level estimates of anaemia prevalence or immunisation coverage) are reported with confidence intervals so programme managers know how much trust to place in each estimate before allocating funds.
- Clinical researchers in Indian hospitals use 95% CIs for mean treatment effects so physicians see the plausible range of benefit, not just a single point estimate from one small trial.

Key Points

- Estimation = guessing population parameters from sample statistics; parameters (μ, σ, Greek) are fixed and unknown; statistics ($\bar{x}$, s, Roman) are known and vary.
- Point estimate = one number; interval estimate = range; confidence interval = interval + confidence level.
- 95% CI for the mean (σ unknown): $\bar{x} \pm t \times (s/\sqrt{n})$, *with* $df = n - 1$ *and* t from the standard t table.
- "95% confident" describes the method's long-run success rate (95 of 100 intervals capture μ) — not the probability that one fixed interval contains μ, and not a statement about individual values.
- Wider intervals come from: higher confidence level, smaller n, larger s. To halve the width, roughly quadruple n.
- Use t (not z) when σ is unknown — the standard situation with small samples.
- Always end with a plain-language conclusion naming the parameter, the limits and the confidence level.

Common Confusions

- **"95% probability μ is in this interval"** — wrong. μ is fixed; the interval either contains it or not. The 95% belongs to the *procedure* over many repetitions.
- **"95% of tablets lie in the interval"** — wrong. The interval estimates the *mean*, not individuals. Individual tablets scatter far more widely (their spread is described by s, not by the CI).
- **Narrower is not automatically "better":** a narrow interval from a flawed sample (biased sampling, wrong method) is precisely wrong. Validity first, precision second.
- **t vs z mix-up:** z needs σ known (rare in practice). With small samples and s estimated from data, t is the correct choice — it gives slightly wider, more honest intervals.
- **CI vs prediction:** a confidence interval for the mean is much narrower than the range where a *single new observation* would fall — don't use a mean-CI to judge individual tablets.

Common Mistakes

- Using z *instead of t* with small samples and unknown σ (gives a falsely narrow interval).
- Using n instead of $n - 1$ for degrees of freedom when reading the t table.
- Forgetting the square root: writing s/n instead of $s/\sqrt{n}$ for the standard error.
- Dividing by n instead of $n - 1$ when computing the sample variance s^2.
- Interpreting the interval as being about individual values rather than the mean.
- Reporting the interval without stating the confidence level.
- Building a CI from a biased or non-random sample and trusting it — the method assumes a proper random sample.

Mnemonics

- **"Greek is Hidden, Roman is Known":** Parameters wear Greek letters (μ, σ) and stay hidden; statistics wear Roman letters ($\bar{x}$, s) and are known from data.
- **"Mean ± t-times-wobble":** CI = $\bar{x} \pm t^* \times (s/\sqrt{n})$ — the mean plus/minus the t-value times the wobble (standard error).
- **"df = n minus 1":** Degrees of freedom for one mean is always one less than the sample size.
- **"95-95 rule":** 95% confidence ≈ the method wins 95 times out of 100.
- **"Quadruple to halve":** To halve the interval width, quadruple the sample size ($\sqrt{n}$ rule).

Exam Summary

- Estimation: population vs sample; parameter (μ, σ) vs statistic ($\bar{x}$, s).
- Point estimate vs interval estimate; definition of confidence interval and confidence level.
- Correct interpretation of "95% confident" (method's long-run rate; not P(μ in interval); not about individuals).

- CI for population mean with σ unknown: $\bar{x} \pm t(s/\sqrt{n})$; *df* = $n - 1$; *t* from standard t table; when to use t vs z.
- Standard error SE = s/√n; margin of error E = t* × SE.
- Factors affecting width: confidence level, sample size (√n rule), variability.
- Numericals: full 8-step construction of 95% (and 99%) CIs for a mean from raw data, with interpretation and comparison.
- Pharmaceutical applications: assay/content uniformity, weight variation, stability, clinical mean effects.

One-Minute Revision

Estimation guesses population parameters (μ, σ — Greek, hidden) from sample statistics ($\bar{x}$, s — Roman, known). Point estimate = one number; confidence interval = range + confidence level. Mean CI: $\bar{x} \pm t \cdot (s/\sqrt{n})$, *df* = $n-1$, *t* from the t table. 95% confident = the recipe captures μ in ~95 of 100 repeats. Higher confidence → wider; bigger n → narrower (quadruple n to halve width); bigger s → wider. The interval is about the mean, not individuals.

Questions

Short-answer questions

1 Define statistical estimation.

2 Distinguish a population parameter from a sample statistic, with symbols.

3 What is the difference between a point estimate and an interval estimate?

4 Define a confidence interval.

5 Write the formula for a confidence interval for the population mean using t values, naming each term.

6 What are degrees of freedom for a single mean?

7 When is the t-distribution used instead of the normal (z) distribution?

8 What is the standard error of the mean?

9 State two factors that widen a confidence interval.

10 Correct this statement: "There is a 95% probability that μ lies in this interval."

Long-answer questions

1 Explain point estimation and interval estimation. Define the confidence interval and interpret the 95% confidence level correctly, with a diagram.

2 Derive (present) the formula for the confidence interval of a population mean when σ is unknown. Explain each component and the role of the t-distribution.

3 Describe the step-by-step procedure for constructing a confidence interval for a mean, and illustrate it with a worked pharmaceutical example.

4 Discuss the factors affecting the width of a confidence interval. Why must the sample size roughly quadruple to halve the width?

5 Explain the applications of confidence intervals in pharmaceutical quality control and clinical research.

MCQs (10) with answer key

1 The symbol μ denotes the: (a) sample mean (b) population mean (c) sample SD (d) margin of error. **Answer: (b)**

2 A point estimate of μ is: (a) s (b) σ (c) $\bar{x}$ (d) t*. **Answer: (c)**

3 The 95% CI for a mean with σ unknown is: (a) $\bar{x} \pm z \cdot \sigma/\sqrt{n}$ (b) $\bar{x} \pm t \cdot s/\sqrt{n}$ (c) $\bar{x} \pm s$ (d) $\mu \pm t^* \cdot s$. **Answer: (b)**

4 Degrees of freedom for a single-mean CI with n = 12 is: (a) 12 (b) 11 (c) 10 (d) 13. **Answer: (b)**

5 "95% confident" correctly means: (a) P(μ in this interval) = 0.95 (b) 95% of data lie in the interval (c) the procedure captures μ in ~95 of 100 repetitions (d) the sample mean equals μ with 95% probability. **Answer: (c)**

6 Increasing the confidence level from 95% to 99% makes the interval: (a) narrower (b) wider (c) unchanged (d) centred on s. **Answer: (b)**

7 The standard error of the mean is: (a) s/n (b) $s/\sqrt{n}$ (c) s^2/n (d) t*·s. **Answer: (b)**

8 To roughly halve the CI width, the sample size must be multiplied by about: (a) 2 (b) 4 (c) 10 (d) 1.5. **Answer: (b)**

9 The t-distribution is preferred over z when: (a) σ is known (b) σ is unknown and estimated by s (c) n > 1000 (d) data are categorical. **Answer: (b)**

10 A 95% CI for mean tablet weight is (498.8, 501.6) mg. This interval estimates: (a) the weight of each tablet (b) the population mean weight (c) the sample standard deviation (d) the label claim exactly. **Answer: (b)**

Viva questions

1 What is the difference between μ and $\bar{x}$?

2 Why do we use t values instead of z values for small samples?

3 What does the margin of error represent?

4 If the label claim falls outside the 95% CI for mean content, what would you conclude?

5 Why is "95% of tablets lie inside the CI" a wrong interpretation?

Application-based questions

1 A batch's 95% CI for mean assay is (98.2%, 101.4%) with specification 95.0–105.0%. Should the batch be released? Justify. (Answer: yes — the entire interval lies comfortably within specifications, so the true mean is confidently within limits.)

2 Two analysts report the same point estimate 500 mg but CIs of (499, 501) mg and (490, 510) mg. Which estimate is more useful and why? (Answer: the first — same best guess but much higher precision; the second admits the true mean could be far off.)

3 A researcher wants a margin of error half the current one without changing the confidence level. What must be done to the sample size, and what is the practical trade-off? (Answer: roughly quadruple n; trade-off is cost and time of extra sampling/testing.)

References

- Bolton (Pharmaceutical Statistics)
- Daniel (Biostatistics)
- Montgomery (Design and Analysis of Experiments)

CHAPTER 10: Hypothesis Testing and Parametric Tests — t-Tests and ANOVA

1. Learning Objectives

After studying this chapter, you will be able to:

- Define a statistical hypothesis, null hypothesis, and alternative hypothesis in simple words.
- List and apply the six steps of hypothesis testing in the correct order.
- Explain Type I and Type II errors using the court-trial analogy.
- Explain sample size, power of a test, and p-value.
- Perform single sample, independent two-sample, and paired t tests by hand on small datasets.
- Carry out a one-way ANOVA (and describe two-way ANOVA) and compute the F ratio.
- Test hypotheses about a regression slope and a correlation coefficient.
- Apply these tests to pharmacy case studies: comparing tablet formulations, checking batch consistency, analysing stability data.

2. Definitions

- **Hypothesis:** A claim about a population (a large group we cannot fully measure) that we check using data from a small sample.
- **Hypothesis testing:** A fixed step-by-step procedure in which sample data are used to decide, with a known risk of being wrong, whether a claim about the population should be rejected.
- **Null hypothesis (H0):** The "no effect / no difference" statement, e.g., "the new tablet is no different from the old tablet." We assume it is true until the data strongly say otherwise.
- **Alternative hypothesis (H1):** The opposite claim — "there IS a difference." It is accepted only when the data give strong evidence against H0.

- **Parametric test:** A test that works by estimating population parameters (mainly the mean and variance) and assumes the data follow a known shape, usually the normal (bell-shaped) distribution. t-tests and ANOVA are parametric tests.
- **p-value:** The probability of getting data as extreme as ours (or more extreme) if the null hypothesis were really true. A very small p-value means "this would be shocking under H0," so we reject H0.
- **Significance level (α):** The fixed risk of wrongly rejecting H0 that we accept before the experiment — almost always 0.05 (5%). We compare the p-value with α.

3. Terminology Table

Term	Simple meaning
Hypothesis	A testable claim about a big group, checked with small sample data
Null hypothesis (H0)	The "nothing is happening / no difference" starting assumption
Alternative hypothesis (H1)	The "something is happening / there is a difference" claim
Test statistic	One number (t, F, z) computed from the sample; its size shows how unusual the data are
Critical value	The cut-off line from standard tables; crossing it means "reject H0"
Type I error	Rejecting H0 when it is actually true (a false alarm); probability = α
Type II error	Failing to reject H0 when it is actually false (a missed signal); probability = β
Power of the test	The chance of correctly detecting a real difference = 1 − β
p-value	How surprising the data would be if H0 were true; reject H0 if p < α
Degrees of freedom (df)	The number of independent pieces of information in the data; roughly n − 1
One-tailed test	Looking for a change in only one direction (more than, or less than)
Two-tailed test	Looking for a change in either direction (different from)

4. Simple Explanations

4.1 What is hypothesis testing, from absolute zero?

The label on a paracetamol bottle says "each tablet contains 500 mg." You cannot test all 2,00,000 tablets in the batch, so you take a sample — say 10 tablets — and measure each. Their average is 496 mg. Question: is the batch really under-dosed, or is 496 just the bad luck of the sample? Any small sample's average wobbles a little around the true batch average.

Hypothesis testing is the formal rule-book for answering that question. Start by assuming the batch is fine (H0: true mean = 500 mg). Then ask: how likely is a sample average as low as 496 purely by chance? If that probability (the p-value) is extremely small — smaller than our fixed risk limit α = 0.05 — we stop assuming the batch is fine and reject H0. Otherwise we say the data are not strong enough to reject H0.

We never say "we proved the batch is perfect." We only say "reject H0" or "fail to reject H0." The verdict is always about the evidence, never absolute proof.

4.2 Null and alternative hypotheses

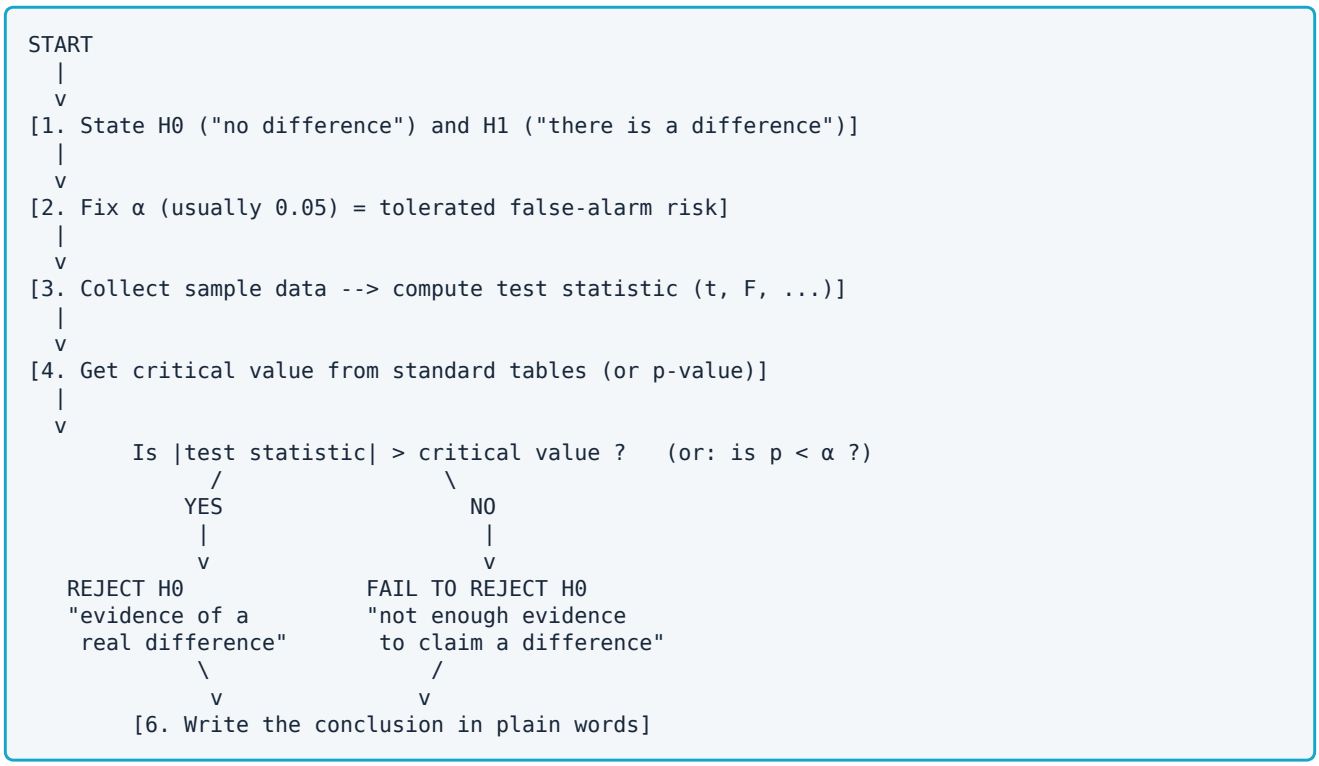
- **H0 (null):** written with an equals sign. Example: H0: μ = 500 mg (the true batch mean equals the label claim). "μ" (mu) is the symbol for the population mean — the true average of the whole batch.
- **H1 (alternative):** written with ≠, <, or >. Example: H1: μ ≠ 500 mg (two-tailed: a difference in either direction) or H1: μ < 500 mg (one-tailed: only under-dosing matters).

Example. A syrup maker claims 100 mL average fill. H0: μ = 100 mL (claim true); H1: μ ≠ 100 mL (claim false). **Diagram:** think of H0 as an accused person — innocent until proven guilty; conviction needs overwhelming evidence. **Key takeaway:** H0 is the "innocent / no-difference" statement; H1 wins only with strong evidence.

4.3 The six steps of hypothesis testing

Step 1 — State the hypotheses. Write H0 and H1 in symbols (H0: μ = 500; H1: μ ≠ 500). **Step 2 — Choose the significance level (α).** Almost always α = 0.05: "we accept a 5% chance of rejecting H0 even when it is true." **Step 3 — Compute the test statistic.** Pick the right test for the data (see the decision tree in Chapter 12), then crunch the sample into one number — the test statistic (t, F, or z). **Step 4 — Find the critical value (or p-value).** From standard tables, find the cut-off for our α and degrees of freedom, or compute the p-value. **Step 5 — Decide.** If |test statistic| > critical value (equivalently p < α), **reject H0**; otherwise **fail to reject H0**. **Step 6 — Conclude in plain words.** Example: "The data give significant evidence that the batch mean differs from 500 mg."

Flowchart of hypothesis testing (text):



Key takeaway: Every hypothesis test is the same six steps — only the formula for the test statistic changes.

4.4 Type I and Type II errors (the court-trial analogy)

H0 = "the accused is innocent." The mapping is exact:

- **Type I error (false alarm):** convicting an innocent person. In statistics: rejecting H0 when H0 is actually true. Probability = α (0.05). Like wrongly recalling a good batch — costly.
- **Type II error (missed signal):** letting a guilty person go free. In statistics: failing to reject H0 when H0 is actually false. Probability = β. Like releasing an under-dosed batch — dangerous for patients.

Example. Testing a batch for under-potency (H0: batch is fine). Type I error: recall a fine batch → money wasted. Type II error: release a weak batch → patients harmed.

Power of the test = 1 − β: the probability of correctly catching a real difference — the court's skill at convicting the truly guilty. Power increases when we (a) increase the sample size n, (b) use a larger α, or (c) the real difference is bigger.

> **Key takeaway:** α = risk of false alarm (Type I); β = risk of missed signal (Type II); power = 1 − β = chance of catching the truth.

4.5 Sample size and power

A bigger sample gives a clearer picture: the sample mean wobbles less around the true mean because the standard error (s/√n) shrinks. With more tablets tested, even a small real under-dosing stands out — power goes up.

Example. With n = 4 tablets, a true shortfall of 3 mg may hide inside the sampling wobble (low power). With n = 100, the same shortfall stands out clearly (high power). Studies fix n beforehand to reach the needed power (usually 80%) without wasting resources. **Key takeaway:** bigger n → smaller wobble → higher power → fewer missed signals.

4.6 p-values

The p-value answers: "If H0 were perfectly true, how surprising would my data be?" p = 0.03 means: "If the batch were really fine, data this extreme would occur only 3 times in 100 by pure chance." Since 0.03 < 0.05, we reject H0 — the data are too surprising to keep believing H0.

Example. Comparing two brands gives p = 0.40: even if the brands were identical, this much sample difference would happen 40 times in 100 by chance — not surprising, so we fail to reject H0. **Key takeaway:** small p (< 0.05) → surprising under H0 → reject H0; large p → nothing surprising → fail to reject.

4.7 Parametric tests — what makes them "parametric"

A parametric test estimates **parameters** — the population mean (μ) and variance (σ^2) — and assumes the data come from a known distribution shape, almost always the **normal (bell-shaped) distribution**: most values cluster near the mean and taper off symmetrically, like tablet weights from a well-controlled press.

The three parametric tools of this chapter: **t-tests** (compare means), **ANOVA** (compare 3+ means at once), and **t-tests in regression/correlation** (check whether a slope or correlation is really non-zero).

Assumptions of parametric tests:

1 **Normality:** data in each group are approximately bell-shaped.

2 **Independence:** one observation does not influence another.

3 **Equal variances:** groups have roughly equal spread (needed for two-sample t and ANOVA).

If assumption 1 fails badly (heavily skewed data), switch to the non-parametric tests of Chapter 11. **Example.** Tablet hardness from a good process is bell-shaped → t-test is fine. Pain scores (0–10, bunched at extremes) are not → use a non-parametric test. **Key takeaway:** parametric = uses means, assumes bell-shaped independent data.

4.8 The three t-tests

A. Single sample t test. One sample's mean vs a fixed target. $t = (\bar{x} - \mu 0)/(s/\sqrt{n})$; $df = n - 1$. **Example.** 10 tablet weights vs the 500 mg label claim (worked in Calculation Box 1). **Key takeaway:** one sample vs one target.

B. Two independent samples t test. Two *separate* groups — the 10 tablets of Formulation A are different tablets from the 10 of Formulation B. $t = (\bar{x}1 - \bar{x}2)/SE$, with pooled variance $s^2p = [(n1-1)s1^2 + (n2-1)s2^2]/(n1+n2-2)$; $df = n1 + n2 - 2$. **Example.** Dissolution of Brand A vs Brand B tablets. **Key takeaway:** two separate groups → pooled variance.

C. Paired t test. The *same* subjects measured twice — each "before" tied to one "after," like two shoes of one pair. Compute the difference d for each pair, then run a single sample t test on the differences: $t = \bar{d}/(sd/\sqrt{n})$; $df = n - 1$. **Example.** BP of 8 patients before and after a drug (worked in Calculation Box 2). **Key takeaway:** same subjects twice → analyse the differences.

4.9 ANOVA — Analysis of Variance (one-way and two-way)

The t-test compares two means. With **three or more** groups, many t-tests are wrong — each adds another 5% false-alarm risk, so the total risk balloons. **ANOVA compares all group means in one test.**

The big idea. ANOVA splits total variation into:

- **Between-group variation (MSB):** how far the group means are from each other. Big MSB = treatments really differ.
- **Within-group variation (MSW):** random wobble inside each group (noise).
- **F ratio = MSB/MSW.** If groups are identical, F hovers near 1 (both parts are noise). If groups truly differ, F becomes large. Reject H0 (all means equal) when F exceeds the critical value.

One-way ANOVA: one factor (formulations A, B, C) → "does formulation affect hardness?" **Two-way ANOVA:** two factors at once (formulation × storage condition: room/accelerated) → "does formulation matter? does storage matter? is there an interaction (storage hits some formulations harder)?" One dataset answers both questions — more efficient than two experiments. Same three assumptions as t-tests. **Example.** Three batches of one product: H0 = all batch means equal (process consistent); a large F means the process is drifting. **Key takeaway:** 3+ groups → one ANOVA, not many t-tests; F = between-group variation ÷ within-group variation.

4.10 Hypothesis testing in regression and correlation

After fitting a line $y = a + bx$, two t-tests check whether the relationship is real:

- **Is the slope non-zero?** H0: $\beta = 0$ ("x has no effect on y"). $t = b/SE(b)$, $df = n - 2$.
- **Is the correlation non-zero?** H0: $\rho = 0$ ("no linear relationship"). $t = r\sqrt{(n-2)}/\sqrt{(1-r^2)}$, $df = n - 2$.

Example. A calibration curve (absorbance vs concentration): testing H0: slope = 0 checks whether absorbance truly rises with concentration — failing to reject means the "calibration" is useless. **Diagram:** a scatter of points with a fitted line; the t-test asks "is this line really tilted, or is the tilt a trick of random scatter?" **Key takeaway:** a slope or correlation is believable only after its t-test rejects "no relationship."

5. Classifications

- **By what is compared:** one sample vs target → single sample t; two independent groups → independent samples t; two related measurements → paired t; 3+ groups, one factor → one-way ANOVA; 3+ groups, two factors → two-way ANOVA; relationship between two variables → t-test for slope/correlation.
- **By hypothesis direction:** two-tailed (H1: ≠) vs one-tailed (H1: > or <).
- **By error type:** Type I (false alarm, α) vs Type II (missed signal, β); power = 1 − β.
- **Parametric vs non-parametric:** parametric tests (this chapter) assume normality and use means; non-parametric tests (Chapter 11) assume no shape and use ranks.

6. Processes

6.1 Procedure of a one-way ANOVA (step-by-step)

1 H0: μ1 = μ2 = μ3 (all means equal); H1: at least one differs. Fix α = 0.05.

2 Compute each group mean and the grand mean.

3 SSB = $\Sigma\, n_i(\text{group mean}_i - \text{grand mean})^2$; df = k − 1 (k = groups).

4 SSW = Σ(each value − its group mean)^2; df = N − k (N = total observations).

5 MSB = SSB/(k−1); MSW = SSW/(N−k).

6 F = MSB/MSW. Reject H0 if F > critical F (df1 = k−1, df2 = N−k) or p < 0.05.

7 Conclude in words; a post-hoc test then shows *which* pairs differ.

6.2 Choosing among the tests of this chapter

```
How many groups / what is the question?
  |
  +-- ONE sample vs a fixed target? --> single sample t test
  |
  +-- TWO groups?
  |      +-- same subjects twice (before/after)? --> paired t test
  |      +-- two separate groups? --> independent samples t test
  |
  +-- THREE OR MORE groups?
  |      +-- one factor varied? --> one-way ANOVA
  |      +-- two factors varied? --> two-way ANOVA
  |
  +-- Relationship between two variables?
         --> regression line; t-test for slope; t-test for correlation
```

7. Textual Diagrams/Flowcharts

The two diagrams of this chapter are the hypothesis-testing flowchart (section 4.3) and the test-selection tree (section 6.2). Read the flowchart top to bottom: every test follows the same six-step spine (hypotheses → α → statistic → critical value/p → decide → conclude); the tree only decides *which* statistic to compute.

8. Infographic Concept

INFOGRAPHIC CONCEPT: "The Justice System of Statistics." A courtroom: H0 is the accused wearing a badge "INNOCENT until proven guilty." Six gavels on the judge's bench = the six steps. Two red doors: left "TYPE I ERROR — convicted an innocent batch (false alarm, α = 5%)", right "TYPE II ERROR — freed a guilty batch (missed signal, β)". Centre: a balance scale "F = between-group evidence ÷ within-group noise" tipping when evidence outweighs noise. Bottom: three t-test ID cards — "Single sample = me vs target", "Independent = my group vs your group", "Paired = me before vs me after" — plus an ANOVA badge: "3+ groups, one trial instead of many."

9. Tables

9.1 Paired vs independent samples t test

Feature	Paired t test	Independent samples t test
Data	Same subjects measured twice	Two separate, unrelated groups
We analyse	The differences d	The two means $\bar{x}1$, $\bar{x}2$
df	n − 1 (n = pairs)	n1 + n2 − 2
Extra assumption	Differences approx. normal	Equal variances in both groups
Pharmacy example	BP before/after drug, same patients	Dissolution: Brand A vs Brand B

9.2 Parametric tests at a glance

Test	H0	Test statistic	df
Single sample t	μ = μ0	$t = (\bar{x} - \mu 0)/(s/\sqrt{n})$	n − 1
Independent samples t	μ1 = μ2	$t = (\bar{x}1 - \bar{x}2)/SE$ (pooled)	n1 + n2 − 2
Paired t	mean difference = 0	$t = \bar{d}/(sd/\sqrt{n})$	n − 1
One-way ANOVA	μ1 = μ2 = ...	F = MSB/MSW	(k−1, N−k)
Slope in regression	β = 0	t = b/SE(b)	n − 2
Correlation	ρ = 0	$t = r\sqrt{(n-2)}/\sqrt{(1-r^2)}$	n − 2

10. Worked Examples and Calculation Boxes

CALCULATION BOX 1 — Single sample t test (tablet weight vs label claim). μ0 = 500 mg. Ten tablets (mg): 498, 501, 497, 500, 499, 502, 496, 500, 499, 498. H0: μ = 500; H1: μ ≠ 500; α = 0.05.

- $\bar{x}$ = 4990/10 = 499 mg.
- Squared deviations from 499: 1, 4, 4, 1, 0, 9, 9, 1, 0, 1 → sum = 30. s^2 = 30/9 = 3.333; s = 1.826.
- SE = 1.826/√10 = 1.826/3.162 = 0.577.

- t = (499 − 500)/0.577 = −1.732; df = 9. Critical t (two-tailed, 0.05, df 9) = 2.262 as per standard tables. |t| = 1.732 < 2.262.
- **Decision:** fail to reject H0. **Conclusion:** no significant evidence the batch mean differs from 500 mg.

CALCULATION BOX 2 — Paired t test (BP fall, before − after, 8 patients, mmHg). Differences: 12, 15, 10, 14, 11, 13, 16, 9. H0: mean difference = 0; H1: ≠ 0; α = 0.05.

- $\bar{d}$ = 100/8 = 12.5.
- Squared deviations: 0.25, 6.25, 6.25, 2.25, 2.25, 0.25, 12.25, 12.25 → sum = 42. sd^2 = 42/7 = 6; sd = 2.449.
- SE = 2.449/√8 = 0.866. t = 12.5/0.866 = 14.43; df = 7. Critical t = 2.365.
- 14.43 > 2.365 → **reject H0**. **Conclusion:** the drug significantly lowers systolic BP.

CALCULATION BOX 3 — Independent samples t test (hardness, kg/cm²). A (n1 = 5): 8.2, 8.5, 8.0, 8.4, 8.3 → $\bar{x}A$ = 8.28, sA^2 = 0.037. B (n2 = 5): 7.6, 7.9, 7.5, 7.8, 7.7 → $\bar{x}B$ = 7.70, sB^2 = 0.025. H0: μA = μB; α = 0.05.

- s^2p = [(4)(0.037) + (4)(0.025)]/8 = 0.248/8 = 0.031.
- SE = √[0.031(1/5 + 1/5)] = √0.0124 = 0.1114.
- t = (8.28 − 7.70)/0.1114 = 5.21; df = 8. Critical t = 2.306.
- 5.21 > 2.306 → **reject H0**. **Conclusion:** the formulations differ significantly in hardness.

CALCULATION BOX 4 — One-way ANOVA (dissolution %, 3 formulations × 5 tablets). A: 82, 85, 83, 84, 86 (mean 84). B: 78, 80, 79, 81, 77 (mean 79). C: 88, 90, 89, 87, 91 (mean 89). Grand mean = 84. H0: μA = μB = μC; α = 0.05.

- SSB = $5(0)^2 + 5(-5)^2 + 5(5)^2$ = 250; df = 2; MSB = 125.
- SSW = 10 + 10 + 10 = 30; df = 12; MSW = 2.5.
- F = 125/2.5 = 50. Critical F (2, 12, 0.05) = 3.89 as per standard tables.
- 50 > 3.89 → **reject H0**. **Conclusion:** the formulations differ significantly in dissolution.

CALCULATION BOX 5 — Test for a correlation coefficient. Stability study: r = 0.85 between storage time and degradant level, n = 10. H0: ρ = 0; α = 0.05.

- t = 0.85√8/√(1 − 0.7225) = 2.404/0.5268 = 4.56; df = 8. Critical t = 2.306.
- 4.56 > 2.306 → **reject H0**. **Conclusion:** the correlation is statistically significant.

11. Analogies

- **Court trial:** H0 = "innocent." Type I error convicts an innocent person (false alarm); Type II frees a guilty one (missed signal). Power = the court's skill at convicting the truly guilty.
- **Faulty bathroom scale:** the scale wobbles ±0.5 kg, so a 0.3 kg change proves nothing. Hypothesis testing is the rule for deciding when a change is bigger than the wobble.
- **p-value as coincidence meter:** p = 0.001 means "a one-in-a-thousand coincidence if nothing were happening." Tiny p = too strange to be coincidence.

12. Pharmacy Relevance

- **Batch release:** single sample t test of mean assay vs the label claim.
- **Generic vs innovator:** independent samples t test on dissolution or hardness.
- **Clinical efficacy:** paired t test on before/after BP, blood sugar, or pain scores.
- **Process consistency:** one-way ANOVA across three batches; two-way ANOVA adds a second factor (site, storage condition).
- **Stability studies:** t-test of the degradant-vs-time slope confirms real degradation and supports shelf-life claims.
- **Analyst comparison:** two analysts assaying the same samples (paired by sample) — a paired t test exposes analyst bias.

13. India Context

Indian companies file batch-release and stability data with the national drug regulator, and hypothesis tests — t-tests of assay vs specification, ANOVA for batch-to-batch consistency, slope tests for stability trends — are the standard statistical language of these dossiers. The Indian Pharmacopoeia's uniformity tests rest on the same logic: compare sample statistics with fixed limits while controlling error risks. In B.Pharm/M.Pharm projects, comparing a developed formulation with a marketed product by t test, or proving batch consistency by ANOVA, is the expected analysis in dissertations and journal papers.

14. Key Points

- H0 = "no difference"; reject H0 only with strong evidence ($p < \alpha$). Never "prove" H0.
- Six steps: hypotheses → α → statistic → critical value/p → decide → conclude.
- Type I (α) = false alarm; Type II (β) = missed signal; power = $1 - \beta$; larger n → more power.
- Single sample t: vs target. Independent t: two separate groups (pooled variance). Paired t: same subjects twice (differences).
- ANOVA: 3+ groups in one test; F = MSB/MSW; one-way = one factor, two-way = two factors + interaction.
- Slope and correlation each get a t-test (df = $n - 2$).

15. Common Confusions

- **"Fail to reject" ≠ "H0 is true":** it means "not enough evidence." Absence of evidence is not evidence of absence.
- **p-value is NOT the probability H0 is true** — it is the probability of such data *assuming* H0 is true.
- **One- vs two-tailed:** decide before seeing the data; switching afterwards is cheating.
- **Paired vs independent:** same patients before/after = paired; different tablets of A vs B = independent.
- **ANOVA says *that* groups differ, not *which*** — a post-hoc test gives the pairwise detail.

16. Common Mistakes

- Three t-tests instead of one ANOVA for three groups (false-alarm rate balloons).
- Paired t test on independent groups, or vice versa.
- Writing "the drug has no effect" when $p > 0.05$ (correct: "no significant evidence of an effect").
- Ignoring assumptions: t-test on badly skewed data instead of a non-parametric test (Chapter 11).
- Reporting "$p < 0.05$" without the statistic, df, and means — the reader cannot judge effect size.

17. Mnemonics

- **"TIP" for t-tests: T**arget = single sample; **I**ndependent = two separate groups; **P**aired = same people twice.
- **"F = Fight Between Means":** the F ratio pits between-group differences against within-group noise; big F = the groups really differ.
- **Type I = "I falsely accused"** (false alarm); **Type II = "Too blind to see"** (missed signal).

18. Exam Summary

Six-step hypothesis testing (hypotheses → α → statistic → critical value → decision → conclusion). H0 = no difference; Type I (α) = false alarm, Type II (β) = missed signal, power = $1 - \beta$ (rises with n). Reject H0 when $p < \alpha$. Parametric tests assume normality, independence, equal variances. t-tests: single sample (vs target), independent samples (two groups, pooled variance), paired (before–after differences). ANOVA: 3+ means in one test, F = MSB/MSW; one-way = one factor, two-way = two factors + interaction. Slope and correlation: t-tests with df = $n - 2$.

19. One-Minute Revision

Six steps; H0 = no difference; $p < 0.05$ → reject. Type I = false alarm (α), Type II = missed signal (β), power = $1 - \beta$, bigger n = more power. t-tests: single = vs target, independent = two groups, paired = before/after. ANOVA: 3+ groups, F = between ÷ within. Slope and r: t-test, df = $n - 2$. Never "prove" H0.

20. Questions

Short-answer:

1 Define null and alternative hypotheses with a pharmacy example.

2 Distinguish Type I and Type II errors.

3 What is power? Name two ways to increase it.

4 Define p-value in one sentence.

5 When is a paired t test preferred over an independent samples t test?

6 Write the F ratio and explain MSB and MSW.

7 State the three assumptions of parametric tests.

8 Why use ANOVA rather than three t-tests for three groups?

Long-answer:

1 Describe the six steps of hypothesis testing, illustrating with the court-trial analogy.

2 Explain the procedure of one-way ANOVA with a worked example.

3 Explain one-way vs two-way ANOVA with pharmacy examples.

4 Describe hypothesis testing for a regression slope and a correlation coefficient.

5 A QC chemist tests 12 tablets against a 250 mg claim: $\bar{x} = 247$ mg, $s = 4$ mg. Perform the full test at $\alpha = 0.05$.

MCQs (answer key below):

1 The null hypothesis states: (a) there is a difference (b) there is no difference (c) the sample is large (d) $\alpha = 0.01$

2 A Type I error is: (a) accepting a false H0 (b) rejecting a true H0 (c) using the wrong test (d) a calculation mistake

3 Power of a test equals: (a) α (b) β (c) $1 - \beta$ (d) $1 - \alpha$

4 Reject H0 when: (a) $p > \alpha$ (b) $p < \alpha$ (c) $p = \alpha$ (d) $t = 0$

5 Paired t test df for 10 pairs: (a) 10 (b) 9 (c) 18 (d) 8

6 Independent samples t, $n1 = 6$, $n2 = 8$: df = (a) 14 (b) 12 (c) 13 (d) 11

7 ANOVA, 4 groups, 20 observations: df for MSB = (a) 3 (b) 16 (c) 19 (d) 4

8 The F ratio is: (a) MSW/MSB (b) MSB/MSW (c) SSB/SSW (d) $s1^2/s2^2$

9 Which increases power? (a) smaller n (b) larger n (c) smaller α (d) larger β

10 t-test for correlation uses df: (a) $n - 1$ (b) $n - 2$ (c) $2n - 2$ (d) n

Answer key: 1-b, 2-b, 3-c, 4-b, 5-b, 6-b, 7-a, 8-b, 9-b, 10-b.

Viva questions:

1 What does "fail to reject H0" actually mean?

2 Why is α usually fixed at 0.05?

3 Your p-value is 0.06. Significant? Why?

4 When would you choose a one-tailed test?

5 What is pooled variance and when is it needed?

Application-based:

1 A stability study gives $r = 0.78$ ($n = 12$) between storage time and degradant. Test significance at $\alpha = 0.05$.

2 Two coating machines; 8 tablets weighed from each. Outline how to test equal mean weight — which t-test and why?

21. References

- Bolton (Pharmaceutical Statistics)
- Daniel (Biostatistics)
- Montgomery (Design and Analysis of Experiments)

CHAPTER 11: Non-Parametric Tests — Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman and Chi-Square

1. Learning Objectives

After studying this chapter, you will be able to:

- Define a non-parametric test and explain when it is used instead of a parametric test.
- Explain the idea of ranks in simple words.
- State the assumptions, procedure, and applications of the Mann-Whitney U test, Wilcoxon signed-rank test, Kruskal-Wallis test, Friedman test, and chi-square tests.
- Compute the Mann-Whitney U statistic, Wilcoxon signed-rank statistic, Kruskal-Wallis H, Friedman chi-square, and chi-square statistic by hand on small datasets.
- Match each non-parametric test with its parametric counterpart.
- Solve examination problems on non-parametric tests and apply them to pharmacy examples.

2. Definitions

- **Non-parametric test:** A statistical test that does not estimate population parameters (mean, variance) and does not assume the data follow a bell-shaped (normal) distribution. Also called a **distribution-free test**.
- **Rank:** The position number of a value when all values are arranged in order from smallest to largest. The smallest value gets rank 1, the next rank 2, and so on.
- **Median:** The middle value of ordered data — half the values are below it, half above. Non-parametric tests often compare medians rather than means.
- **Tied ranks:** When two or more values are equal, each gets the average of the rank positions they share (e.g., two values tied for positions 2 and 3 each get rank 2.5).
- **Chi-square (χ^2) test:** A non-parametric test for counted (frequency) data that compares observed counts with the counts expected under the null hypothesis.

3. Terminology Table

Term	Simple meaning
Non-parametric test	A test that makes no assumption about the shape of the data
Distribution-free	Another name for non-parametric — no bell-shape assumed
Rank	Position number after sorting smallest → largest
U statistic	The Mann-Whitney test value; small U means the groups are well separated
Signed rank	Rank of a difference, carrying its + or − sign (Wilcoxon)
Kruskal-Wallis H	The test statistic comparing 3+ independent groups
Friedman test	The non-parametric version of repeated-measures comparison (3+ related groups)
Observed frequency (O)	The count actually seen in the data
Expected frequency (E)	The count expected if H0 were true
Contingency table	A table of counts with rows and columns (e.g., treatment × outcome)
Goodness of fit	Chi-square test asking "do the observed counts fit a claimed pattern?"
Test of independence	Chi-square test asking "are the row and column categories related?"

4. Simple Explanations

4.1 Non-parametric tests, from absolute zero

Parametric tests (Chapter 10) need bell-shaped data. But much pharmacy data is not bell-shaped: patient pain scores (0–10) bunch at the ends; tablet defect counts are whole numbers with many zeros; taste scores are "1 = poor … 5 = excellent" (ordered labels, not real measurements).

Non-parametric tests solve this by forgetting the actual numbers and using only their order — their ranks. Take disintegration times 45, 12, 30 minutes. Sorted: 12 (rank 1), 30 (rank 2), 45 (rank 3). The test works with 1, 2, 3 instead of 12, 30, 45. Ranks have no shape assumption — they are just positions 1 to n — so the test works on any data, even skewed or ordinal data.

Example. Five volunteers rate bitterness 1–5: 2, 5, 3, 5, 1. Sorted: 1 (rank 1), 2 (rank 2), 3 (rank 3), 5 and 5 tied for positions 4, 5 → each gets rank (4+5)/2 = 4.5. **Key takeaway:** non-parametric = no bell-shape needed; rank everything, then test the ranks.

4.2 Mann-Whitney U test (two independent groups)

Technical concept: the non-parametric counterpart of the independent samples t test. H0: the two groups have the same distribution (same median).

Simple explanation: Mix both groups' values together, rank them all, then add up the ranks of each group. If the groups are truly alike, the ranks mix evenly and both rank-sums are middling. If one group is genuinely higher, its values hog the top ranks and its rank-sum swells — the U statistic becomes small, and we reject H0.

Procedure: (1) Combine both samples, rank all N values (average ties). (2) R1 = sum of ranks in group 1. (3) U1 = n1·n2 + n1(n1+1)/2 − R1; U2 = n1·n2 − U1. (4) Test statistic U = smaller of U1, U2. (5) Compare with the critical U from standard tables (or $p < 0.05$); reject H0 if U ≤ critical U. **Note the reversed rule:** here a *small* statistic rejects H0.

Example. Disintegration times (min) of Formulation A vs B are ranked and compared (full numbers in Calculation Box 1). **Diagram:**

```
Group A values:  ●       ●    ●  ● ●
Group B values: ● ●  ● ●    ● ●
Ranks mixed 1..11 --> add A's ranks = R1
If A truly slower, A's ranks pile at the top --> R1 large --> U small --> reject H0
```

Key takeaway: two separate groups, non-normal data → Mann-Whitney U; rank all, sum ranks, small U rejects.

4.3 Wilcoxon signed-rank test (two related measurements)

Technical concept: the non-parametric counterpart of the paired t test. H0: the median difference between the paired measurements is zero.

Simple explanation: For each pair compute the difference (after − before), throw away any zero differences, rank the *absolute* differences, then put the +/− sign back on each rank. Add the positive ranks (W+) and the negative ranks (W−). If the treatment does nothing, + and − ranks balance. If it works, one side dominates and the smaller sum W = min(W+, W−) becomes tiny → reject H0.

Procedure: (1) Differences di = after − before; drop zeros. (2) Rank |di| (average ties). (3) Attach each rank its sign; W+ = sum of positive ranks, W− = sum of negative ranks. (4) W = min(W+, W−). (5) Reject H0 if W ≤ critical W from standard tables.

Example. Pain scores before/after an analgesic in 8 patients (full numbers in Calculation Box 2). **Key takeaway:** before/after same subjects, non-normal differences → Wilcoxon signed-rank; small W rejects.

4.4 Kruskal-Wallis test (three or more independent groups)

Technical concept: the non-parametric counterpart of one-way ANOVA. H0: all groups have the same distribution.

Simple explanation: It is the Mann-Whitney idea stretched to 3+ groups: rank *all* observations together, compute each group's rank-sum Ri, and measure how uneven the rank-sums are with the H statistic: $H = [12/(N(N+1))] \times \Sigma(Ri^2/ni) - 3(N+1)$, where N = total observations, ni = size of group i. Evenly mixed ranks → H near 0; one group hogging high ranks → H large. Compare H with the critical chi-square value (df = k − 1 groups); reject H0 if H exceeds it.

Example. Dissolution ranks of three coating formulations compared (full numbers in Calculation Box 3). **Key takeaway:** 3+ separate groups, non-normal → Kruskal-Wallis H; large H rejects.

4.5 Friedman test (three or more related groups)

Technical concept: the non-parametric counterpart of repeated-measures ANOVA. H0: the treatments have identical effects.

Simple explanation: Each subject (block) tries *all* treatments — e.g., each volunteer tastes all three syrup flavours and ranks them 1, 2, 3 *within* the volunteer. Then add up each treatment's ranks across volunteers

(R1, R2, R3). If flavours are equally liked, the rank-sums are similar; if one flavour wins everywhere, its rank-sum stands out. Statistic: $\chi^2 r = [12/(b \cdot k(k+1))] \times \Sigma Rj^2 - 3b(k+1)$, where b = number of blocks (volunteers), k = number of treatments. Compare with critical chi-square (df = k − 1); large value rejects H0.

Example. Four volunteers rank three formulations (full numbers in Calculation Box 4). **Diagram:** think of b judges each giving 1st/2nd/3rd place to k contestants; the Friedman test asks whether the total scores differ more than judge-to-judge wobble allows. **Key takeaway:** same subjects × 3+ treatments → Friedman; rank within each subject, then compare rank-sums.

4.6 Chi-square tests (counted data)

Technical concept: for frequencies (counts), not measurements. $\chi^2 = \Sigma(O - E)^2/E$ summed over all cells; df = (rows − 1)(columns − 1) for a table. Large χ^2 (≥ critical value) rejects H0.

Simple explanation: We count how many observations fall in each category (the observed counts, O), then compute what the counts *would* be if nothing were going on (the expected counts, E). The χ^2 statistic totals the squared mismatches, scaled by E. Big total mismatch → the pattern is real.

Two uses:

- **Goodness of fit:** one categorical variable. H0: "the counts follow a claimed ratio." Example: a genetics cross expects a 3:1 ratio of round to wrinkled seeds — do 90 round and 30 wrinkled fit? (E = 90, 30 → χ^2 = 0, perfect fit.)
- **Test of independence:** two categorical variables in a contingency table. H0: "the row variable and column variable are unrelated." Expected count per cell = (row total × column total)/grand total.

Example. Adverse effects (yes/no) vs Formulation A/B in 80 patients (full numbers in Calculation Box 5). **Key takeaway:** counted data → chi-square; $\chi^2 = \Sigma(O-E)^2/E$; big value rejects "no pattern / no association."

5. Classifications

- **By parametric counterpart:** Mann-Whitney U ↔ independent samples t test; Wilcoxon signed-rank ↔ paired t test; Kruskal-Wallis ↔ one-way ANOVA; Friedman ↔ repeated-measures ANOVA; chi-square has no direct parametric twin (it handles counts).
- **By data structure:** independent groups (Mann-Whitney, Kruskal-Wallis, chi-square) vs related/matched data (Wilcoxon signed-rank, Friedman).
- **By number of groups:** two groups (Mann-Whitney, Wilcoxon, 2×2 chi-square) vs three or more (Kruskal-Wallis, Friedman, larger chi-square tables).
- **By variable type:** ranked/ordinal/continuous non-normal (rank tests) vs categorical counts (chi-square).

6. Processes

6.1 Mann-Whitney U procedure (step-by-step)

1 H0: the two populations have the same distribution; H1: they differ. Fix α = 0.05.

2 Pool both samples; rank all N values from smallest (1) upward; average the ranks of ties.

3 R1 = sum of ranks belonging to group 1 (n1 observations); n2 = size of group 2.

4 $U1 = n1 \cdot n2 + n1(n1+1)/2 - R1$; $U2 = n1 \cdot n2 - U1$. Test statistic $U = \min(U1, U2)$.

5 Find critical U for (n1, n2) at α from standard tables. Reject H0 if $U \leq$ critical U (small U rejects).

6 Conclude in plain words.

6.2 General non-parametric workflow

```
Is the data bell-shaped and are n's reasonable?
  |-- YES --> use parametric tests (Chapter 10: t, ANOVA)
  |-- NO (skewed, ordinal scores, small n, outliers) --> rank-based test:
        Two independent groups?  --> Mann-Whitney U
        Two related measurements? --> Wilcoxon signed-rank
        3+ independent groups?   --> Kruskal-Wallis
        3+ related measurements? --> Friedman
        Counted categories?      --> Chi-square
```

7. Textual Diagrams/Flowcharts

The ranking illustration (section 4.2) and the selection workflow (section 6.2) are the diagrams of this chapter. The core visual to remember: two rows of dots (two groups) merged into one ranked line — when one group's dots cluster at the high end of the line, the groups differ.

8. Infographic Concept

INFOGRAPHIC CONCEPT: "The Ranking Olympics." A podium where every data value is an athlete wearing a bib showing its rank (1, 2, 3 ...). Five event cards hang above: "Mann-Whitney — two teams, separate tracks"; "Wilcoxon — same athlete, before vs after"; "Kruskal-Wallis — three-plus teams, one race"; "Friedman — every judge scores every contestant"; "Chi-square — counting spectators in each stand, expected vs observed." A referee holds a flag: "Data not bell-shaped? Don't measure the athletes — rank them!"

9. Tables

9.1 Parametric vs non-parametric tests

Feature	Parametric (Ch. 10)	Non-parametric (this chapter)
Assumption about shape	Bell-shaped (normal) data	No shape assumed
What is compared	Means	Medians / distributions (ranks)
Works on	Measured numbers	Ranks, ordered scores, counts
Power when data are normal	Higher	Slightly lower
When data are skewed	Can mislead	The correct choice
Examples	t-tests, ANOVA	Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman, chi-square

9.2 Test mapping: parametric counterpart

Situation	Parametric test	Non-parametric test
Two independent groups	Independent samples t test	Mann-Whitney U test
Two related measurements	Paired t test	Wilcoxon signed-rank test
3+ independent groups	One-way ANOVA	Kruskal-Wallis test
3+ related measurements	Repeated-measures ANOVA	Friedman test
Categorical counts	(none)	Chi-square test

10. Worked Examples and Calculation Boxes

CALCULATION BOX 1 — Mann-Whitney U (disintegration time, min). Formulation A (n1 = 5): 12, 15, 18, 20, 24. Formulation B (n2 = 6): 10, 11, 13, 14, 16, 17. H0: same distribution; α = 0.05, two-tailed.

- Pooled and ranked: 10(1), 11(2), 12(3), 13(4), 14(5), 15(6), 16(7), 17(8), 18(9), 20(10), 24(11).
- Ranks of A: 3 + 6 + 9 + 10 + 11 = 39 = R1.
- U1 = (5)(6) + 5(6)/2 − 39 = 30 + 15 − 39 = 6. U2 = 30 − 6 = 24. U = min(6, 24) = 6.
- Critical U (n1 = 5, n2 = 6, α = 0.05 two-tailed) = 3 as per standard tables. Reject only if U ≤ 3. Here 6 > 3.
- **Decision:** fail to reject H0. **Conclusion:** no significant difference in disintegration time between the formulations.

CALCULATION BOX 2 — Wilcoxon signed-rank (pain-score fall after − before, 8 patients). Differences: +3, −1, +4, +2, −5, +6, −2, +7 (no zeros). H0: median difference = 0; α = 0.05.

- Absolute values: 3, 1, 4, 2, 5, 6, 2, 7. Ranks: 1→1; 2,2 tie → (2+3)/2 = 2.5 each; 3→4; 4→5; 5→6; 6→7; 7→8.
- Positive ranks (W+): 4 + 5 + 2.5 + 7 + 8 = 26.5. Negative ranks (W−): 1 + 6 + 2.5 = 9.5. Check: 26.5 + 9.5 = 36 = 8(9)/2. ✓
- W = min(26.5, 9.5) = 9.5. Critical W (n = 8, α = 0.05 two-tailed) = 3 as per standard tables. 9.5 > 3.
- **Decision:** fail to reject H0. **Conclusion:** no significant median change in pain score.

CALCULATION BOX 3 — Kruskal-Wallis (dissolution %, 3 formulations × 4 tablets). A: 5, 7, 9, 11. B: 2, 4, 6, 8. C: 12, 14, 16, 18. H0: same distribution; α = 0.05.

- Ranks of all 12: 2(1), 4(2), 5(3), 6(4), 7(5), 8(6), 9(7), 11(8), 12(9), 14(10), 16(11), 18(12).
- RA = 3+5+7+8 = 23; RB = 1+2+4+6 = 13; RC = 9+10+11+12 = 42. Check: 78 = 12(13)/2. ✓
- H = [12/(12×13)] × ($23^2/4 + 13^2/4 + 42^2/4$) − 3(13) = (12/156) × (529+169+1764)/4 − 39 = 0.07692 × 615.5 − 39 = 47.35 − 39 = 8.35.
- df = 3 − 1 = 2. Critical χ^2 (0.05, df 2) = 5.991 as per standard tables. 8.35 > 5.991.
- **Decision:** reject H0. **Conclusion:** the three formulations differ significantly in dissolution.

CALCULATION BOX 4 — Friedman (4 volunteers rank 3 syrup flavours 1 = best). Rank-sums: flavour 1 = 5, flavour 2 = 9, flavour 3 = 10. Check: 5+9+10 = 24 = 4×3×4/2. ✓ H0: no preference difference; α = 0.05.

- χ^2r = [12/(4×3×4)] × ($5^2 + 9^2 + 10^2$) − 3(4)(4) = (12/48) × (25+81+100) − 48 = 0.25 × 206 − 48 = 51.5 − 48 = 3.5.

- df = 3 − 1 = 2. Critical χ^2 = 5.991. 3.5 < 5.991.
- **Decision:** fail to reject H0. **Conclusion:** no significant preference difference among the flavours.

CALCULATION BOX 5 — Chi-square test of independence (adverse effect × formulation, 80 patients). Observed: A: yes 8, no 32 (row 40); B: yes 16, no 24 (row 40). Column totals: yes 24, no 56; grand total 80. H0: adverse effect independent of formulation; α = 0.05.

- Expected = row × column / grand: A-yes = 40×24/80 = 12; A-no = 40×56/80 = 28; B-yes = 12; B-no = 28.
- $\chi^2 = (8-12)^2/12 + (32-28)^2/28 + (16-12)^2/12 + (24-28)^2/28 = 16/12 + 16/28 + 16/12 + 16/28 = 1.333 + 0.571 + 1.333 + 0.571 = 3.808$.
- df = (2−1)(2−1) = 1. Critical χ^2 (0.05, df 1) = 3.841 as per standard tables. 3.808 < 3.841.
- **Decision:** fail to reject H0. **Conclusion:** no significant association between formulation and adverse effects.

11. Analogies

- **Race positions (ranks):** You do not need each runner's exact time to know who won — finishing positions (1st, 2nd, 3rd) are enough. Rank tests likewise ignore exact values and use positions.
- **Party guests (chi-square):** You expect 50 guests (E) but 70 arrive (O). The χ^2 statistic measures "how wrong was the expectation?" across every category; a huge mismatch means your assumption was wrong.
- **Judges' scorecards (Friedman):** Four judges each rank three dishes 1-2-3. If one dish keeps getting 1st place from every judge, the rank-sums expose the winner without ever using the actual scores.
- **Tug of war (Wilcoxon):** Positive ranks pull one way, negative ranks the other; W is the weaker team's total pull — if it is tiny, the stronger side clearly won.

12. Pharmacy Relevance

- **Dissolution comparison of two generics** with skewed data → Mann-Whitney U instead of a t test.
- **Before/after symptom scores** (pain, acidity) that are ordinal → Wilcoxon signed-rank.
- **Comparing 3+ suppliers' assay values** when normality fails → Kruskal-Wallis.
- **Taste/palatability panels** where each volunteer tries every formulation → Friedman test.
- **Adverse-event counts** across treatment arms, or **defect counts** (cracked vs intact tablets across batches) → chi-square tests.
- **Questionnaire data** (satisfaction: poor/average/good/excellent) are ordinal — rank tests are the natural analysis.

13. India Context

Indian pharmacy dissertations routinely collect ordinal data — patient feedback scores, taste-panel ratings, visual grades of formulations — for which supervisors expect Mann-Whitney, Wilcoxon, or

Friedman analyses rather than t-tests. Chi-square tests appear in Indian pharmacoepidemiology work, e.g., testing whether adverse-drug-reaction reporting is associated with dosage form or patient group. Knowing the non-parametric toolkit lets a B.Pharm student analyse real project data correctly instead of forcing skewed data into a t test.

14. Key Points

- Non-parametric = no bell-shape assumed; work with ranks (or counts), not raw values.
- Ties share the average of their rank positions.
- Mann-Whitney U: 2 independent groups; small U rejects H0.
- Wilcoxon signed-rank: 2 related measurements; rank |differences|, re-attach signs; small W rejects.
- Kruskal-Wallis H: 3+ independent groups; $H = [12/(N(N+1))]\Sigma(Ri^2/ni) - 3(N+1)$; large H rejects (df = k − 1).
- Friedman: same subjects × 3+ treatments; rank within each subject; large χ^2r rejects (df = k − 1).
- Chi-square: $\chi^2 = \Sigma(O-E)^2/E$; goodness of fit (one variable) vs test of independence (two variables, E = row×column/grand).
- Expected frequencies below 5 weaken the chi-square test — combine small categories or collect more data.

15. Common Confusions

- **Mann-Whitney vs Wilcoxon:** Mann-Whitney = two *separate* groups; Wilcoxon signed-rank = *paired* before/after. Names sound alike; the data structure decides.
- **Small statistic rejects:** in Mann-Whitney and Wilcoxon, a *small* U or W rejects H0 — the reverse of t and F. Do not mix up the direction.
- **Chi-square needs counts, not means:** you cannot run chi-square on average tablet weights — only on frequencies in categories.
- **"Non-parametric" ≠ "no assumptions":** independence of observations is still required, and chi-square needs adequate expected counts.
- **Kruskal-Wallis vs Friedman:** Kruskal-Wallis = different subjects per group; Friedman = same subjects in every group.

16. Common Mistakes

- Using a t test on clearly skewed data or ordinal scores instead of the rank test.
- Forgetting to drop zero differences in the Wilcoxon signed-rank test.
- Not averaging tied ranks (assigning 2 and 3 instead of 2.5 and 2.5).
- Using chi-square when expected frequencies are tiny (< 5) without combining categories.
- Concluding "the groups are equal" from a non-significant result — correct: "no significant evidence of a difference."

17. Mnemonics

- **"Rank the Tank":** when data are *skewed like a dented tank*, don't measure — rank.
- **"U-W-K-F-Chi": U**nrelated pairs → Mann-Whitney **U**; **W**ithin-subject pairs → **W**ilcoxon; **K**ount 3+ groups → **K**ruskal-Wallis; **F**lavours every judge tastes → **F**riedman; **C**ounts in categories → **Chi**-square.
- **"Small U, big news":** tiny U or W means reject H0 — the opposite of t and F.

18. Exam Summary

Non-parametric (distribution-free) tests assume no bell-shaped distribution; they analyse ranks (or counts) instead of raw values. The Mann-Whitney U test compares two independent groups (U = min(U1, U2); small U rejects). The Wilcoxon signed-rank test compares two related measurements (rank absolute differences, re-attach signs; small W rejects). The Kruskal-Wallis test extends the idea to three or more independent groups (H statistic, df = k − 1). The Friedman test handles three or more related measurements by ranking within each subject. Chi-square tests analyse counted data: $\chi^2 = \Sigma(O-E)^2/E$, used for goodness of fit and for tests of independence in contingency tables (E = row total × column total / grand total).

19. One-Minute Revision

No bell shape? Rank it. Mann-Whitney U = 2 separate groups (small U rejects). Wilcoxon = before/after pairs, drop zeros (small W rejects). Kruskal-Wallis = 3+ separate groups (big H rejects). Friedman = same subjects, 3+ treatments, rank within subject. Chi-square = counted data, $\Sigma(O-E)^2/E$. Ties get averaged ranks. Never "prove equality" — only fail to reject.

20. Questions

Short-answer:

1 Define a non-parametric test and give one situation where it beats a t test.

2 What is a rank? How are tied ranks handled?

3 State H0 for the Mann-Whitney U test.

4 Why are zero differences dropped in the Wilcoxon signed-rank test?

5 Write the Kruskal-Wallis H formula and define each symbol.

6 Distinguish goodness of fit from the test of independence.

7 How is the expected frequency of a contingency-table cell computed?

8 Name the non-parametric counterpart of (a) paired t test, (b) one-way ANOVA.

Long-answer:

1 Describe the procedure of the Mann-Whitney U test with a worked example.

2 Explain the Wilcoxon signed-rank test step by step with a worked example.

3 Derive the logic of the Kruskal-Wallis test and compare it with one-way ANOVA.

4 Describe the Friedman test with an example; when is it preferred over Kruskal-Wallis?

5 Explain the chi-square test of independence with a worked 2×2 example, including expected frequencies.

MCQs (answer key below):

1 Non-parametric tests assume: (a) normal data (b) equal variances (c) no specific distribution (d) large samples

2 Ranks are: (a) actual values (b) position numbers after sorting (c) squared values (d) percentages

3 Mann-Whitney U is the counterpart of: (a) paired t (b) independent samples t (c) ANOVA (d) chi-square

4 In Mann-Whitney, H0 is rejected when U is: (a) large (b) small (c) zero (d) negative

5 Wilcoxon signed-rank needs: (a) two independent groups (b) paired data (c) counts (d) 3+ groups

6 Kruskal-Wallis compares: (a) 2 means (b) 3+ independent groups (c) paired data (d) variances

7 Friedman test ranks: (a) all data together (b) within each subject (c) only the best (d) randomly

8 χ^2 = (a) $\Sigma(O-E)/E$ (b) $\Sigma(O-E)^2/E$ (c) $\Sigma(O+E)^2/E$ (d) $\Sigma O/E$

9 df for a 3×4 contingency table: (a) 12 (b) 6 (c) 7 (d) 11

10 Chi-square is unsuitable when: (a) data are counts (b) expected frequencies are very small (c) n is large (d) $\alpha = 0.05$

Answer key: 1-c, 2-b, 3-b, 4-b, 5-b, 6-b, 7-b, 8-b, 9-b, 10-b.

Viva questions:

1 Your data are badly skewed with outliers. t test or Mann-Whitney? Why?

2 Two values tie for ranks 4 and 5. What rank does each get?

3 Why does a small U reject H0 while a large F rejects H0?

4 When do you use Friedman instead of Kruskal-Wallis?

5 What goes wrong if an expected frequency is 2 in a chi-square test?

Application-based:

1 Twelve patients rate nausea (0–10) before and after an antiemetic; the differences are skewed. Which test do you choose? Outline the steps.

2 Defect counts (cracked/intact) are recorded for three tablet batches: 6/94, 12/88, 9/91. Set up the chi-square test of independence (H0, expected frequencies, df) and state the decision rule.

21. References

- Bolton (Pharmaceutical Statistics)
- Daniel (Biostatistics)
- Montgomery (Design and Analysis of Experiments)