PharmacoLearn

MRM301T Research Methodology and Biostatistics, M.Pharm Semester 3

Unit 2: Biostatistics

Free notes at pharmacolearn.fyi

2.1 Definition and applications

Biostatistics is the branch of statistics that deals with data arising from biological, medical and pharmaceutical work: measurements on living systems, which vary from subject to subject in ways that inanimate measurements do not. Its two arms are **descriptive statistics**, which summarise and present the data actually collected (means, standard deviations, percentages, tables and graphs), and **inferential statistics**, which use a sample to draw conclusions about the population it came from, with a stated margin of uncertainty. Nearly everything in this unit belongs to the second arm, because the purpose of a pharmacy experiment is never merely to describe the twelve rats studied: it is to claim something about the drug.

The applications in pharmaceutical research run through the whole pipeline of this degree. Formulation scientists compare dissolution profiles and assay values between batches and products. Pharmacologists compare responses between dose groups in animals and compute potency and dose response quantities. Clinical researchers size trials, compare treatments and test safety signals. Quality units set specification thinking on process capability and sampling. In every case the form of the question is the same: an observed difference, an estimate of the noise around it, and a disciplined decision about whether the difference is real. Biostatistics is that discipline.

2.2 Data and variables: types and measurement scales

The named tests of this unit cannot be chosen, let alone computed, without knowing what kind of data feed them, so the foundations are set out here briefly. A **variable** is any characteristic measured on each study unit: tablet weight, blood pressure, pain score, sex, cured or not cured. Variables are **quantitative** (numerical: weight, concentration, count) or **qualitative** (categorical: sex, blood group, grade of reaction). Quantitative variables are **discrete** (counts: number of tablets rejected) or **continuous** (any value in a range: weight, assay percent).

Measurement scales refine the classification, and the scale decides which statistics are legitimate. **Nominal** scales name categories with no order (blood group). **Ordinal** scales order categories without equal steps (pain: none, mild, moderate, severe). **Interval** scales have equal steps but no true zero (temperature in Celsius). **Ratio** scales have equal steps and a true zero (weight, concentration), and support every arithmetic operation. The practical rule used throughout this unit: parametric tests of sections 2.12 to 2.16 are built for interval and ratio data that are reasonably normally distributed; ordinal data, and numerical data that are badly skewed, go to the non parametric tests of sections 2.17 and 2.18.

2.3 Summarising data: central tendency and dispersion

A set of observations is first summarised by where it centres and how widely it spreads. The **mean** is the arithmetic average, the sum of observations divided by their number; it uses every value and is the centre the inferential tests are built on, but a single extreme value drags it. The **median** is the middle value of the ordered set, untouched by extremes, and the right centre for skewed data. The **mode** is the most frequent value, the only centre available for nominal data.

Spread is measured by the **variance** and its square root, the **standard deviation (SD)**. The variance averages the squared deviations from the mean, dividing by (n - 1) for a sample, because a sample's deviations are taken around its own mean and slightly understate the population spread; that (n - 1) is the first appearance of the degrees of freedom of section 2.9. The SD, back in the original units, describes how far a typical observation lies from the mean. The **standard error of the mean (SE)** is a different quantity: SE = SD / sqrt(n), and it measures how precisely the sample mean estimates the population mean. Confusing SD and SE in a thesis table is the commonest descriptive error in pharmacy research: SD describes the data, SE describes the precision of the mean.

Worked example 1 (mean, SD, SE). Ten tablets from a batch weigh (mg): 248, 252, 250, 246, 255, 249, 251, 253, 247, 250.

- Sum = 2501, so mean = 2501 / 10 = 250.1 mg. Ordered, the middle pair is 250 and 250, so the median is 250.0 mg.
- The squared deviations from 250.1 sum to 68.9, so the sample variance s squared = 68.9 / 9 = 7.66 and s = 2.77 mg.
- SE = 2.77 / sqrt(10) = 0.88 mg.
- Reported properly: mean tablet weight 250.1 mg, SD 2.77 mg (the spread of individual tablets); the mean itself is estimated to within about plus or minus 1.96 x 0.88, giving a 95 percent confidence interval of 248.4 to 251.8 mg. The coefficient of variation, 100 x 2.77 / 250.1 = 1.1 percent, expresses the spread relative to the size of the measurement.

2.4 The normal distribution

Most parametric machinery rests on one model of how measurements distribute: the **normal (Gaussian) distribution**, the symmetrical bell shaped curve fully described by its mean mu and standard deviation sigma. Biological measurements that result from many small independent influences (weights, volumes, many assay values) approximate it, and, crucially, **sample means** approximate it even when the raw data do not, which is why tests on means work as widely as they do.

The curve's working properties are few and exact. About **68.27 percent** of values lie within mean plus or minus 1 SD, about **95.45 percent** within plus or minus 2 SD, and about **99.73 percent** within plus or minus 3 SD. The conventional 95 percent limits are mean plus or minus 1.96 SD, leaving 2.5 percent in each tail; the 99 percent limits use 2.58 SD. Deviations measured in SD units are called standard normal deviates or z values, and every tail area used in this unit (0.025 beyond z = 1.96, for instance) comes from this curve. When data depart strongly from the bell (heavy skew, many ties, a hard boundary at zero), the parametric tests lose their footing and the rank based tests of section 2.17 take over.

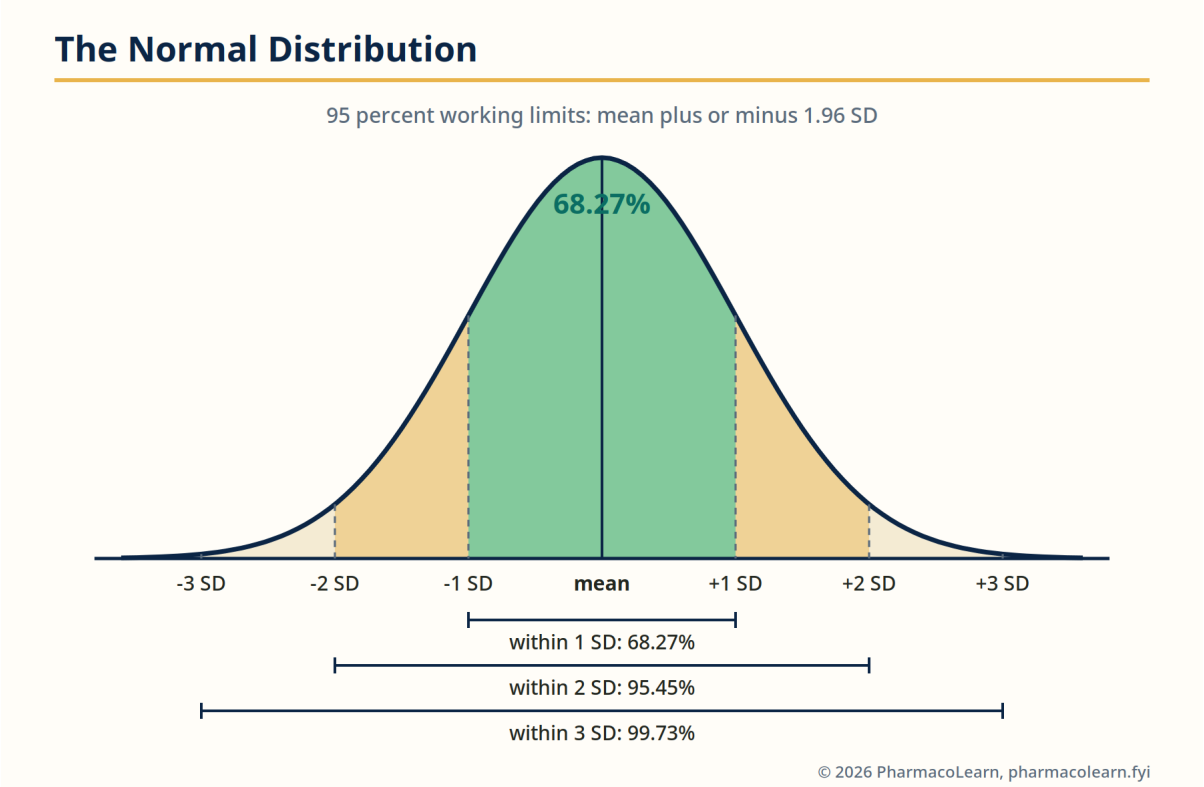

Figure: The normal curve. The areas are exact properties of the model: 68.27 percent within one standard deviation of the mean, 95.45 percent within two, 99.73 percent within three; the 95 percent working limits sit at 1.96 standard deviations.

2.5 Sample size: why it matters

The **sample size** is the number of study units per group (or in total), and fixing it is a design decision, not an administrative detail. A study that is **too small** wastes its subjects and its ethics approval: even a real, worthwhile effect will fail to reach significance because the noise swamps it (a Type II error, section 2.10), and the result "no significant difference" then means only "the study could not tell". A study that is **too large** wastes resources and, in animal and human work, exposes more subjects than the question needs, which is its own ethical fault; it can also make trivial, clinically meaningless differences "significant". The right size is the smallest number that gives the study a predeclared high chance (the power, section 2.10) of detecting the smallest effect worth detecting. Sample size is therefore fixed in the protocol, from the formula of the next section, before the first subject is enrolled; choosing it after seeing the data, or stopping when significance appears, destroys the meaning of the P value.

2.6 Factors influencing sample size

Four quantities decide the size, and the student should be able to state the direction of each:

- **The size of the difference to be detected (delta).** Smaller differences need larger samples: halving the target difference roughly quadruples the required size, other things equal.
- **The variability of the measurement (sigma, the SD).** Noisier outcomes need larger samples; the formula carries sigma squared, so a method twice as variable needs four times the subjects. Pilot data and the literature review exist largely to supply an honest sigma.
- **The significance level (alpha).** The stricter the false positive rate demanded (0.01 instead of 0.05), the larger the sample.
- **The power (1 minus beta).** The higher the chance demanded of detecting the effect (90 percent instead of 80 percent), the larger the sample.

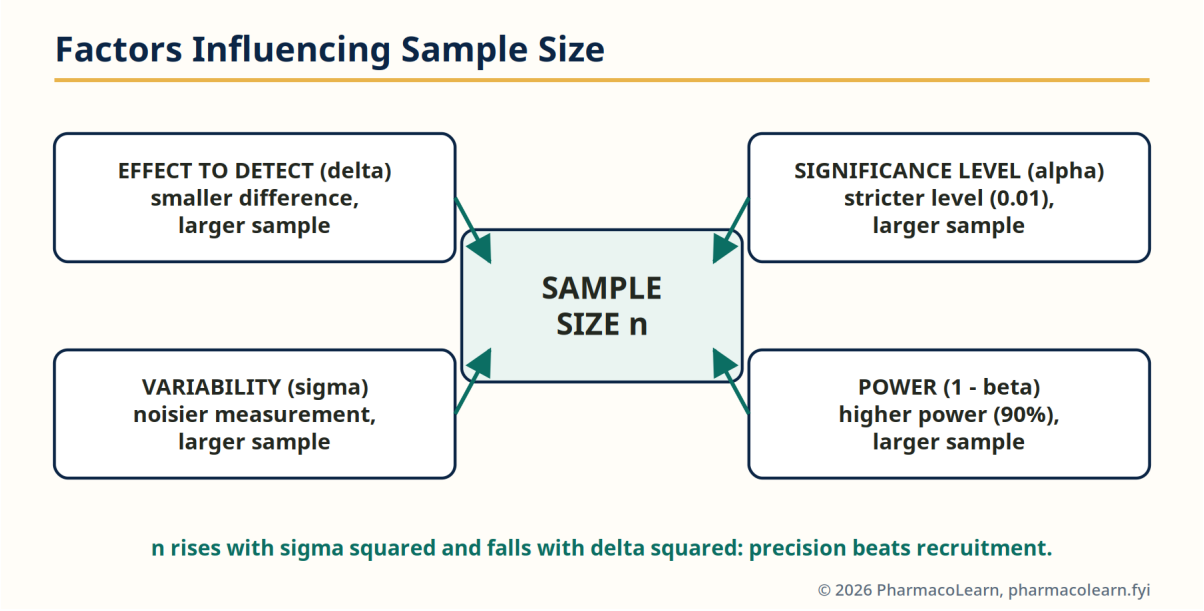

Figure: The four factors that fix a sample size. The difference worth detecting and the variability of the measurement dominate; alpha and power are chosen by convention within narrow ranges.

For comparing two means with equal group sizes, the standard formula is:

```
n per group = 2 x (z for alpha + z for power) squared x sigma squared / delta squared
```

with z = 1.96 for a two sided alpha of 0.05 and z = 0.84 for 80 percent power.

Worked example 2 (sample size). A trial is planned to detect a difference of 8 units between two treatments, where the outcome has SD 10, at alpha 0.05 (two sided) with 80 percent power. Then n per group = 2 x (1.96 + 0.84) squared x 100 / 64 = 2 x 7.84 x 100 / 64 = 24.5, rounded up to **25 subjects per group**. Note the leverage of variability: with an SD of 5 instead of 10, the same calculation gives 6.1, that is 7 per group. Precision of measurement is worth more than recruitment.

2.7 Dropouts

A **dropout** is a subject who enters the study but does not complete it or provide the final measurement: the patient who withdraws, moves away, stops the drug for an adverse event, or is lost to follow up; in animal work, an animal that dies or is excluded mid study. Dropouts shrink the achieved sample below the planned size and so erode power, and they are rarely random: patients who drop out because the drug made them ill differ systematically from those who stay, so a completers only analysis can bias the result toward the drug.

Two defences are standard. At the design stage, the computed sample size is **inflated for the expected dropout rate**: if 25 completers per group are required and 10 percent dropout is expected, enrol 25 / 0.90 = 27.8, that is **28 per group**. At the analysis stage, the **intention to treat** principle analyses every randomised subject in the group to which the dice assigned him, with predeclared rules for the missing values, because that is the comparison randomization bought. The dropout rate itself is always reported, group by group, in the results; a large imbalance in dropouts between groups is a finding, not a footnote.

2.8 Statistical tests of significance and the null hypothesis

A **statistical test of significance** answers one disciplined question: if there were really no effect, how surprising would data like ours be? The machinery starts with the **null hypothesis (H0)**, the hypothesis of no effect: the two population means are equal, the correlation is zero, the proportions do not differ. The

alternative hypothesis (H1) states that an effect exists (two sided: in either direction; one sided: in a named direction). The test computes a **test statistic** from the data, a standardised distance between what was observed and what H0 predicts, measured in units of the noise (the standard error). Large statistics are evidence against H0; the **P value** of section 2.9 converts "large" into a probability.

Two disciplines protect the logic. The hypotheses, the test and the significance level are fixed **before** the data are examined; testing twenty outcomes and reporting the one that crosses 0.05 manufactures false positives at a known rate (with twenty independent tests of true nulls, one crossing is expected by chance alone). And the test must match the design and the data: paired data to a paired test, counts to a count test, skewed data to a rank test. Section 2.20 gathers the matching rules.

2.9 P values, degrees of freedom and interpretation

The **P value** is the probability, computed assuming the null hypothesis is true, of obtaining a result at least as extreme as the one observed. It is not the probability that the null hypothesis is true, and it is not the probability that the result occurred by chance in any loose sense; it is a conditional probability about the data, given H0. By convention a result with **P at or below 0.05** is called statistically significant and H0 is rejected; at or below 0.01, highly significant. The 0.05 line is a convention, not a law of nature: P = 0.049 and P = 0.051 carry nearly identical evidence, and honest reporting gives the exact P where possible.

Degrees of freedom (df) count the number of values free to vary once the statistics already computed are fixed, and they set the exact shape of the t, chi square and F distributions used to judge the test statistic. A sample of n observations used to estimate one mean has n - 1 degrees of freedom for its variance (section 2.3); an unpaired t test on two groups has n1 + n2 - 2; a paired t test on n pairs has n - 1; a chi square test on a table has (rows - 1) x (columns - 1). Every test report in a thesis should carry its degrees of freedom, because the same test statistic means different things at different df.

Interpretation needs three cautions. First, significance is not importance: with a huge sample a trivial difference is significant, and with a small one a large difference may not be; report the size of the effect (the difference and its confidence interval) beside the P value. Second, "not significant" never proves equality; it states that the data do not exclude H0 at the chosen level, which an underpowered study will say about almost everything. Third, a P value says nothing about bias: it prices random error only, and the biased study of Unit 1 gets precise, significant, wrong answers.

The working sequence of a significance test, used by every example in this unit:

```
State H0 and H1, choose the test and alpha (usually 0.05)
        |
        v
Compute the test statistic from the data
        |
        v
Find its degrees of freedom
        |
        v
Compare with the table value at alpha (or read the P value)
        |
        v
Statistic beyond the table value: reject H0, result significant
Statistic within it: do not reject H0, result not significant
```

2.10 Type I and Type II errors, and power

Every significance test can err in exactly two ways, and the two errors trade against each other:

	H0 is actually true	H0 is actually false
Test says significant (reject H0)	Type I error, probability alpha	Correct: a real effect detected, probability = power
Test says not significant	Correct	Type II error, probability beta

A **Type I error** is a false alarm: declaring an effect that is not there. Its probability is the significance level alpha, set by the investigator, conventionally 0.05. A **Type II error** is a miss: failing to detect an effect that is there. Its probability beta falls as the sample size, the true effect size or the measurement precision rises. The **power** of a test, 1 minus beta, is the probability of detecting an effect of a stated size when it exists; 80 percent power is the conventional planning standard of section 2.6. The trade off is structural: demanding fewer false alarms (smaller alpha) at fixed sample size buys more misses, and the only way to reduce both errors together is more information, a larger sample or a less variable measurement.

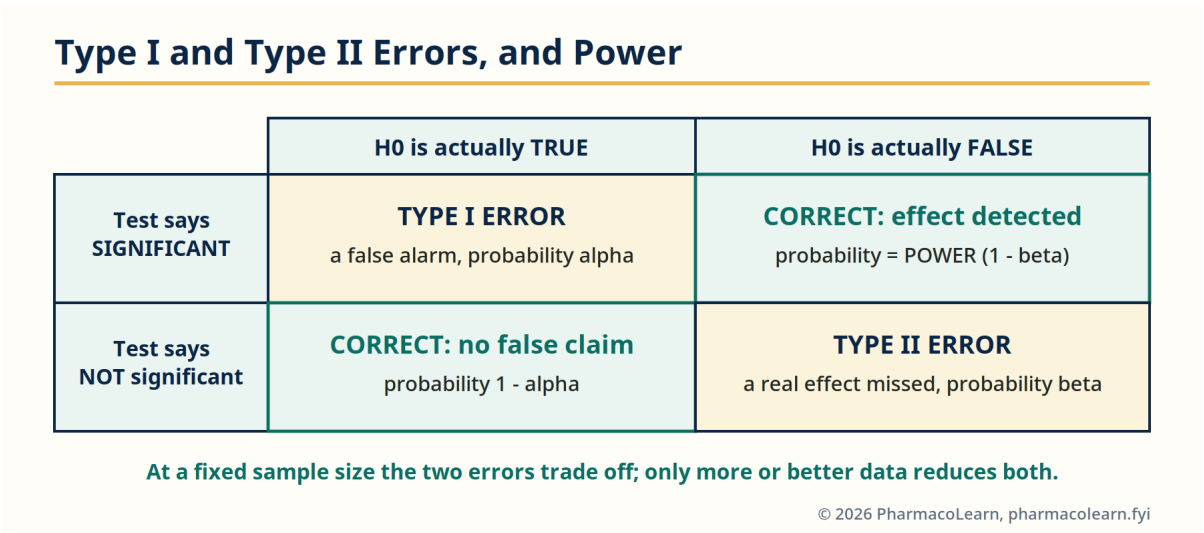

Figure: The decision table of a significance test. Alpha is chosen by the investigator; beta, and with it the power, is set by the sample size, the effect size and the variability.

2.11 Types of significance tests: parametric and non parametric

Significance tests divide into two families. **Parametric tests** assume the data follow a named distribution, in practice the normal distribution, and work with the actual values through their means and standard deviations. They are the more powerful family when their assumptions hold: the data are interval or ratio scale, roughly normally distributed (or the samples are large enough for means to behave), and the group variances are similar. The syllabus names four: Student's t test, ANOVA, the correlation coefficient and regression. **Non parametric (distribution free) tests** make no such assumption; most work on the **ranks** of the observations rather than the values, which makes them robust to skew and outliers and legitimate for ordinal data, at the price of some power when the data really are normal. The syllabus names the Wilcoxon rank tests, non parametric analysis of variance, non parametric correlation and the chi square test. The chi square test sits slightly apart: it works on counts and frequencies rather than measurements at all, and is treated in section 2.19.

2.12 Student's t test: comparing two means

Student's t test (published by W. S. Gosset under the pen name Student) compares means when the population standard deviation is unknown and estimated from the sample, which is always the practical case. Its statistic has the universal shape of this unit: an observed difference divided by its standard error. For **two independent groups** (the unpaired t test):

```
t = (mean1 - mean2) / (sp x sqrt(1/n1 + 1/n2))
```

where sp is the pooled standard deviation, combining both groups' variances, sp squared = ((n1 - 1) s1 squared + (n2 - 1) s2 squared) / (n1 + n2 - 2), and the degrees of freedom are n1 + n2 - 2. Assumptions: independent observations, roughly normal data in each group, similar variances between groups. A one sample variant tests a single mean against a known standard value.

Worked example 3 (unpaired t test). Percentage of drug released at 45 minutes from two tablet formulations, 8 tablets each. Formulation A: 82, 85, 79, 88, 84, 81, 86, 83 (mean 83.5, SD 2.88). Formulation B: 74, 78, 72, 77, 75, 73, 79, 76 (mean 75.5, SD 2.45).

- Difference of means = 83.5 - 75.5 = 8.0.
- Pooled SD: sp squared = (7 x 2.88 squared + 7 x 2.45 squared) / 14 = 7.14, so sp = 2.67.
- Standard error of the difference = 2.67 x sqrt(1/8 + 1/8) = 2.67 x 0.5 = 1.34.
- t = 8.0 / 1.34 = 5.99, with df = 8 + 8 - 2 = 14.
- The table value of t at df 14 and alpha 0.05 (two sided) is 2.145. Since 5.99 exceeds 2.145, the difference is statistically significant (P below 0.05): formulation A releases more drug at 45 minutes, and a gap of 8 percentage points on this variability is far beyond chance.

2.13 The paired t test

When each subject is measured twice (before and after treatment) or subjects come in matched pairs, the **paired t test** applies: it tests whether the mean of the within pair differences differs from zero, with t = mean difference / (SD of differences / sqrt(n)) and df = n - 1. Pairing is powerful for the same reason the crossover design is: it removes between subject variation from the error. Its assumptions concern the differences, which should be roughly normally distributed. Using an unpaired test on paired data, or the reverse, is a design analysis mismatch that examiners and reviewers both catch.

Worked example 4 (paired t test). Systolic blood pressure (mmHg) in 9 patients before and after 4 weeks of a drug. Before: 152, 148, 160, 146, 155, 150, 158, 149, 154. After: 138, 142, 148, 140, 146, 141, 150, 143, 145. The falls (before minus after) are: 14, 6, 12, 6, 9, 9, 8, 6, 9.

- Mean fall = 8.78 mmHg; SD of the falls = 2.77; standard error = 2.77 / sqrt(9) = 0.92.
- t = 8.78 / 0.92 = 9.49, with df = 9 - 1 = 8.
- Table value at df 8, alpha 0.05: 2.306. Since 9.49 far exceeds it, the fall is statistically significant (P below 0.05). Every patient fell, and the average fall of about 9 mmHg is estimated far too precisely to be chance fluctuation.

2.14 Analysis of variance (ANOVA)

Comparing three or more group means with repeated t tests inflates the Type I error: each test carries its 5 percent false alarm risk, and the risks accumulate. **Analysis of variance** answers the multi group question with a single test. It partitions the total variation of all observations into two components: variation **between groups** (the group means scattered around the grand mean, which a real treatment effect inflates) and variation **within groups** (subjects scattered around their own group mean, pure noise). The ratio of the two, each divided by its degrees of freedom to give mean squares, is the **F statistic**: F = MS between / MS within. If treatments do nothing, F hovers near 1; a real effect drives it up. The one way ANOVA of the example compares groups on one factor; two way ANOVA extends the partition to two factors and their interaction. A significant F says only that the group means are not all equal; **which** groups differ is a further question, answered by planned comparisons or post hoc tests, and the overall test must come first to license them.

Worked example 5 (one way ANOVA). Plasma concentrations (microgram per mL) 2 hours after dosing, 5 animals per formulation. Formulation 1: 12.1, 11.8, 12.5, 11.9, 12.2 (mean 12.10). Formulation 2: 10.4, 10.8, 10.2, 10.6, 10.5 (mean 10.50). Formulation 3: 13.2, 12.8, 13.5, 13.0, 13.3 (mean 13.16). Grand mean = 11.92.

Source of variation	Sum of squares	df	Mean square	F
Between formulations	17.93	2	8.97	135.9
Within formulations (error)	0.79	12	0.066	
Total	18.72	14		

F = 8.97 / 0.066 = 135.9, on (2, 12) degrees of freedom. The table value of F at alpha 0.05 for (2, 12) is 3.89. The computed F dwarfs it: the formulation means differ far beyond what within group noise can explain (P well below 0.05). Note how small the within group sum of squares is: the animals within a formulation agree closely, so even moderate mean differences would have shown. Post hoc comparisons would then identify formulation 2 as the low outlier against 1 and 3.

2.15 Correlation coefficient

Correlation measures the strength and direction of the linear association between two quantitative variables measured on the same units. The Pearson **correlation coefficient r** runs from -1 (perfect negative linear association) through 0 (no linear association) to +1 (perfect positive). It is computed as r = Sxy / sqrt(Sxx x Syy), where Sxy is the sum of products of the deviations of x and y from their means, and Sxx and Syy are the sums of squared deviations. Three cautions govern its use. Correlation is not causation: a third factor may drive both variables. r measures **linear** association only; a strong curved relationship can give r near zero. And r is hostage to range and outliers: one extreme point can manufacture or destroy it, which is why the scatter diagram is always drawn first. The significance of r is tested with t = r x sqrt((n - 2) / (1 - r squared)) on n - 2 df; r squared, the coefficient of determination, states the proportion of the variation in y that the linear relation with x accounts for.

Worked example 6, part 1 (correlation). In a pharmacology practical, six preparations were tested at six dose levels (x = 10, 20, 30, 40, 50, 60) and gave percentage inhibitions (y): 18, 22, 38, 40, 52, 63. Means: x bar = 35, y bar = 38.83. The deviation sums are Sxy = 1585, Sxx = 1750, Syy = 1476.8. Then r = 1585 / sqrt(1750 x 1476.8) = 1585 / 1607.5 = 0.986. A very strong positive linear association: as dose rises, inhibition rises almost in step. Its significance: t = 0.986 x sqrt(4 / (1 - 0.972)) = 11.79 on 4 df, against a table value of 2.776, so the correlation is statistically significant.

2.16 Regression

Regression goes one step beyond correlation: it fits the line that best predicts y from x, in the least squares sense of minimising the squared vertical deviations. The simple linear regression line is y = a + b x, with slope b = Sxy / Sxx and intercept a = y bar - b x bar. The slope is the working quantity of pharmacy: units of response per unit of dose, per unit of concentration, per hour. Every calibration curve in analytical work is a regression line, and its inverse use (reading x from a measured y) is the daily arithmetic of assay. Regression assumes the relation is linear, the scatter around the line is roughly even along its length, and (for the significance tests) the residuals are approximately normal. Extrapolation beyond the measured range of x is the standing abuse: the line is evidence only where data stand beneath it.

Worked example 6, part 2 (regression on the same data). Slope b = Sxy / Sxx = 1585 / 1750 = 0.906. Intercept a = 38.83 - 0.906 x 35 = 7.13. The fitted line is: predicted inhibition = 7.13 + 0.906 x dose. Readings: each additional 10 units of dose predicts about 9.1 more percentage points of inhibition; the

predicted inhibition at a dose of 45 is 7.13 + 0.906 x 45 = 47.9 percent. r squared = 0.972 says 97.2 percent of the variation in inhibition across these six preparations is accounted for by dose through this line. Correlation and regression are the same data wearing two hats: r summarises the tightness of the scatter, the line turns it into predictions, and neither, alone, proves that dose caused the inhibition; that claim belongs to the experimental design of Unit 1.

	Correlation	Regression
Question answered	How strongly are x and y linearly associated?	What y is predicted for a given x?
Output	One number, r, between -1 and +1, no units	An equation with slope (units of y per unit of x) and intercept
Symmetry	r of x on y equals r of y on x	The line of y on x differs from the line of x on y
Typical pharmacy use	Dose response association, method comparison	Calibration curves, prediction, potency estimation

2.17 Non parametric tests: the Wilcoxon rank tests

When data are ordinal, badly skewed, or the samples are too small for normality to be trusted, the values are replaced by their **ranks** and the test proceeds on ranks, which no outlier can dominate. The **Wilcoxon signed rank test** is the non parametric partner of the paired t test: take the differences within pairs, rank their absolute values, restore the signs to the ranks, and sum the positive ranks (W+) and the negative ranks (W-). If the treatment does nothing, the two sums should be about equal; the smaller sum, T, is compared with table values for the number of pairs. The **Wilcoxon rank sum test** (equivalently the Mann Whitney U test) is the partner of the unpaired t test for two independent groups: all observations are ranked together and the rank sums of the two groups are compared. Assumptions are mild: independent pairs or groups, differences at least ordinal, and (for the signed rank test) a roughly symmetrical distribution of differences. Zero differences are dropped and the sample size reduced accordingly.

Worked example 7 (Wilcoxon signed rank test). Eight subjects show the following changes in a response score after treatment (positive is improvement): +3.2, -1.1, +2.4, +4.0, +1.8, -0.6, +2.9, +3.6. Ranking the absolute values 0.6, 1.1, 1.8, 2.4, 2.9, 3.2, 3.6, 4.0 gives ranks 1 to 8; the two negative differences hold ranks 1 and 2. So W- = 1 + 2 = 3 and W+ = 3 + 4 + 5 + 6 + 7 + 8 = 33 (check: 33 + 3 = 36 = 8 x 9 / 2, the total of all ranks). T = the smaller sum = 3. The table value for n = 8 pairs at alpha 0.05 (two sided) is 4; since T = 3 is at or below it, the improvement is statistically significant. Six of eight subjects improved, and by the rank evidence that pattern is unlikely under no effect.

2.18 Non parametric analysis of variance and correlation

The rank idea extends to the other parametric procedures. For **three or more independent groups** the counterpart of one way ANOVA is the **Kruskal Wallis test**: all observations are ranked together, the rank sum of each group is compared with what equal distributions would give, and the statistic H is judged against the chi square distribution. For **paired or repeated measurements across three or more conditions** the counterpart is the Friedman test. The counterpart of the unpaired comparison remains the Mann Whitney U test of the previous section. For **association**, the counterpart of Pearson's r is **Spearman's rank correlation coefficient (rho)**, computed exactly like r but on the ranks of x and y, or directly as rho = 1 - 6 x (sum of squared rank differences) / (n x (n squared - 1)). It detects any consistent (monotonic) rise or fall, curved as well as straight, and is the right coefficient for ranked or ordinal data.

Worked example 8 (Spearman rho). Five formulations are ranked 1 to 5 by in vitro dissolution and, independently, by in vivo absorption: dissolution ranks 1, 2, 3, 4, 5; absorption ranks 2, 1, 3, 5, 4. The rank

differences are -1, +1, 0, -1, +1, whose squares sum to 4. Then rho = 1 - (6 x 4) / (5 x 24) = 1 - 24 / 120 = 0.80: a strong tendency for the dissolution order to match the absorption order, the rank form of the in vitro in vivo relationship.

2.19 The chi square test

The **chi square test** handles data that arrive as **counts in categories**: numbers of patients cured or not, tablets passing or failing, animals responding or not. Its most used form tests **association in a contingency table**. For each cell an expected frequency is computed from the row and column totals under the null hypothesis of no association, expected = (row total x column total) / grand total, and the statistic accumulates the standardised disagreements: chi square = sum of (observed - expected) squared / expected, over all cells, with df = (rows - 1) x (columns - 1). Conditions: observations independent, categories mutually exclusive, and expected frequencies not too small (the working rule: every expected count at least 5, or nearly so; for a 2 x 2 table with small numbers the Yates continuity correction, subtracting 0.5 from each absolute disagreement, or Fisher's exact test is used). The same statistic family tests goodness of fit of observed counts to a theoretical distribution.

Worked example 9 (chi square test of association). In a trial, nausea was reported by 12 of 60 patients on a drug and 24 of 60 on placebo. The 2 x 2 table, with expected counts under no association in brackets: drug 12 (18) nausea, 48 (42) none; placebo 24 (18) nausea, 36 (42) none.

- Chi square = (12 - 18) squared / 18 + (48 - 42) squared / 42 + (24 - 18) squared / 18 + (36 - 42) squared / 42 = 2.00 + 0.86 + 2.00 + 0.86 = 5.71.
- df = (2 - 1) x (2 - 1) = 1. Table value at alpha 0.05: 3.84. Since 5.71 exceeds 3.84, the association is statistically significant (P below 0.05): nausea is reported less often on the drug than on placebo in this trial. With the Yates correction the statistic is 4.80, still beyond 3.84, so the conclusion stands.
- Note the discipline of the example: the test compares proportions (20 percent versus 40 percent) through their counts. Running a chi square test on percentages alone, without the underlying numbers, is meaningless, because the same percentages on 12 patients would carry almost no evidence.

2.20 Choosing the right test

The examples of this unit reduce to one skill: matching the test to the design and the data. The questions to ask, in order, are: what type of data (measurements or counts, and if measurements, which scale); how many groups; are the observations paired or independent; and are the parametric assumptions defensible?

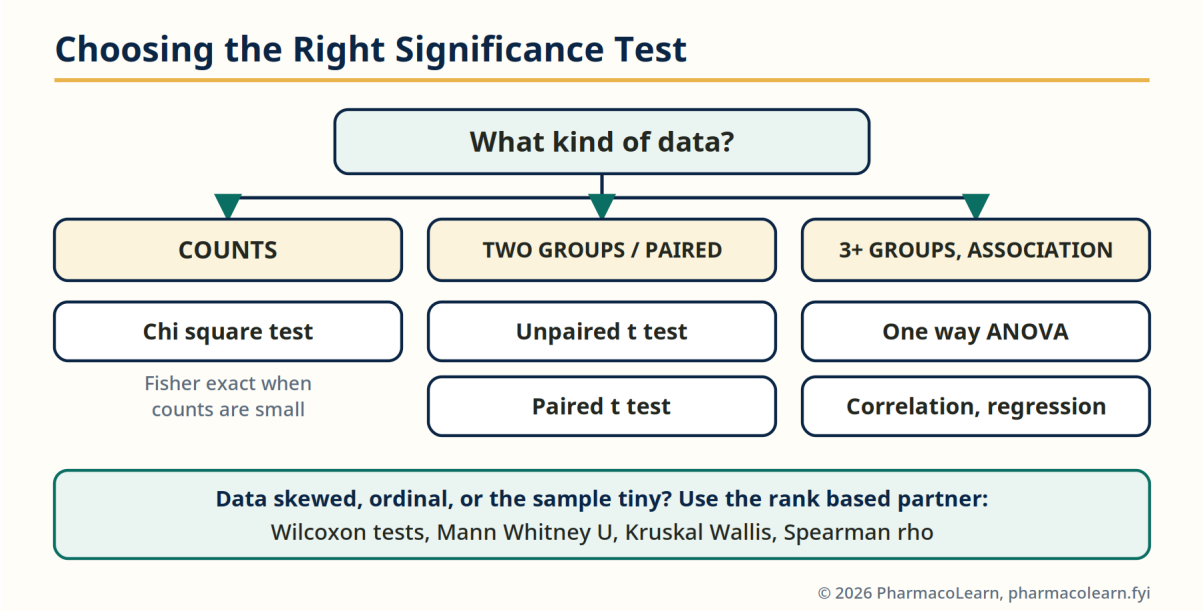

Figure: Choosing the test. Data type first, then structure (groups, pairing), then the parametric question; every road in this unit ends at one of the named tests.

Situation	Parametric test	Non parametric counterpart
One mean against a standard value	One sample t test	Wilcoxon signed rank (one sample form)
Two independent groups, means	Unpaired (Student's) t test	Mann Whitney U / Wilcoxon rank sum
Paired measurements, two conditions	Paired t test	Wilcoxon signed rank test
Three or more independent groups	One way ANOVA	Kruskal Wallis test
Repeated measures, three or more conditions	Repeated measures ANOVA	Friedman test
Association of two quantitative variables	Pearson correlation, regression	Spearman rank correlation
Counts in a contingency table	(none)	Chi square test (Fisher exact for small counts)

The table is the analysis plan of section 1.3 in one view: the row is chosen when the protocol is written, not after the data are seen.

More free M.Pharm notes and unit wise PDFs at pharmacolearn.fyi
Free video lectures on the PharmacoLearn YouTube channel