NOTESKARTS

noteskarts.com

B. Pharmacy — 8th Semester

BIOSTATISTICS AND RESEARCH METHODOLOGY

UNIT 2 — Regression | Probability | Hypothesis Testing | Parametric Tests

Subject Code	Semester	Unit	Platform
BP801T	8th Semester	Unit 2 of 5	Noteskarts.com

UNIT 2 SYLLABUS AT A GLANCE

S.N.	Topic	Sub-topics
1	**Regression Analysis**	*Least squares, y=a+bx, x=a+by, Multiple regression, SE of regression*
2	**Probability**	*Definition, Binomial, Normal, Poisson distributions, properties, problems*
3	**Sampling & Hypothesis**	*Sample/Population, Null/Alternative hypothesis, Types, Type I & II Error, SEM*
4	**Parametric Tests**	*t-test (One-sample, Unpaired, Paired), ANOVA (One-way, Two-way), LSD*

Regression Analysis

Regression Analysis is a statistical method used to examine the relationship between a dependent variable (Y) and one or more independent variables (X). While correlation measures the strength of relationship, regression predicts the value of one variable from another.

◆ Method of Least Squares — Curve Fitting

The method of least squares fits the best line through a set of data points such that the sum of squares of vertical deviations (residuals) is minimized.

Principle: Minimize $S = \Sigma(y_i - \hat{y}_i)^2$ where y_i = observed value, $\hat{y}_i$ = predicted value. The line that minimizes this sum is called the 'Line of Best Fit' or 'Regression Line'.

▸ Normal Equations for fitting y = a + bx (Y on X):

$\Sigma y = na + b\Sigma x$	$\Sigma y = na + b\Sigma x$	Normal equation 1
$\Sigma xy = a\Sigma x + b\Sigma x^2$	$\Sigma xy = a\Sigma x + b\Sigma x^2$	Normal equation 2

Solving for regression coefficients directly:

Slope b	$b = [n\Sigma xy - \Sigma x\Sigma y] / [n\Sigma x^2 - (\Sigma x)^2]$	Regression coeff of y on x
Intercept a	$a = (\Sigma y - b\cdot\Sigma x) / n = \bar{y} - b\cdot\bar{x}$	y-intercept

◆ Regression Line of Y on X: y = a + bx

Used to predict Y (dependent) from X (independent). The regression coefficient b (= byx) represents the change in Y for unit change in X.

Line y on x	$\hat{y} = a + byx \cdot x$ OR $(y - \bar{y}) = byx(x - \bar{x})$	Predict Y from X
byx	$byx = r \cdot (Sy / Sx)$	r=correlation, Sy,Sx=SD

◆ Regression Line of X on Y: x = a + by

Used to predict X (independent) from Y (dependent). This is a separate regression line — not the same as y on x.

Line x on y	$x = a + bxy \cdot y$ OR $(x - \bar{x}) = bxy(y - \bar{y})$	Predict X from Y
bxy	$bxy = r \cdot (Sx / Sy)$	Reverse regression coeff

Note: byx × bxy = r^2 | The two regression lines intersect at the point $(\bar{x}, \bar{y})$ | If r = ±1, both lines coincide. If r = 0, lines are perpendicular to each other.

◆ Multiple Regression

Multiple Regression extends simple regression to include two or more independent variables. It predicts a dependent variable (Y) from multiple predictors (X1, X2, X3...).

Multiple Regression	$Y = b0 + b1X1 + b2X2 + ... + bkXk$	b0=intercept, b1...bk=partial regression coefficients

Partial Regression Coefficients:

- b1 = change in Y per unit change in X1, keeping X2 constant
- b2 = change in Y per unit change in X2, keeping X1 constant
- R^2 (Coefficient of multiple determination) = proportion of variance in Y explained by all X variables

◆ Standard Error of Regression (SE of Estimate)

The Standard Error of Regression (Syx) measures how accurately the regression equation predicts Y. It is the standard deviation of residuals.

SE of Regression	$Syx = \sqrt{[\Sigma(y - \hat{y})^2 / (n-2)]}$	= $\sqrt{[(\Sigma y^2 - a\Sigma y - b\Sigma xy)/(n-2)]}$

Interpretation:

- Small Syx → regression equation fits data well (good prediction)
- Large Syx → high residual error → poor fit
- 95% Prediction interval: $\hat{y} \pm 2 \cdot Syx$ (approx.)
- Pharmaceutical use: Validates calibration curves (HPLC, UV spectroscopy)

Aspect	Correlation (r)	Regression (b)
Purpose	Measures strength of relationship	Predicts value of Y from X
Result	r (dimensionless, −1 to +1)	Equation: $\hat{y} = a + bx$
Symmetry	r(x,y) = r(y,x)	y on x ≠ x on y
Causality	Shows association only	Implies functional dependence
Pharma use	Dose-response strength	Calibration curve equations

Probability Theory & Distributions

Probability is the numerical measure of the likelihood of occurrence of an event. It ranges from 0 (impossible) to 1 (certain). Probability forms the mathematical foundation for all inferential statistics.

Probability	**P(A) = Number of favourable outcomes / Total possible outcomes**	$0 \leq P(A) \leq 1$

◆ Basic Probability Terms

Term	Definition with Example
Experiment	Any process whose outcome is uncertain. Ex: Testing a tablet for dissolution
Sample Space (S)	Set of all possible outcomes. Ex: {Pass, Fail} for a tablet QC test
Event (A)	Subset of sample space. Ex: Tablet passes dissolution test
Complementary Event	$P(A') = 1 - P(A)$. Ex: If P(pass)=0.95, P(fail)=0.05
Mutually Exclusive	Events cannot occur together. Ex: Tablet cannot both pass and fail
Independent Events	Occurrence of A does not affect B. Ex: Two separate tablet tests
Addition Rule	$P(A \cup B) = P(A)+P(B)-P(A \cap B)$
Multiplication Rule	$P(A \cap B) = P(A) \times P(B\|A)$

◆ Binomial Distribution

Binomial Distribution gives the probability of exactly 'x' successes in 'n' independent trials, where each trial has only two outcomes (success/failure) with constant probability p.

Binomial P(X=x)	$P(X=x) = nCx \cdot p^x \cdot q^{n-x}$	n=trials, p=P(success), q=1−p, x=0,1,2...n

Properties of Binomial Distribution:

- Mean (μ) = np
- Variance (σ^2) = npq
- Standard Deviation (σ) = √(npq)
- Shape: Symmetric when p=0.5; Right-skewed when p<0.5; Left-skewed when p>0.5
- Each trial is independent with constant p

Pharmaceutical Problem — Binomial Distribution

Problem: A tablet batch has a defect rate of 10% (p=0.10). A sample of n=5 tablets is selected. Find P(exactly 2 defective tablets). Given: n=5, x=2, p=0.10, q=0.90 P(X=2) = 5C2 × $(0.10)^2$ × $(0.90)^3$ = 10 × 0.01 × 0.729 = 0.0729 Result: P(exactly 2 defectives) = 0.0729 = 7.29% Mean = np = 5×0.10 = 0.5 tablets expected defective SD = √(npq) = √(5×0.10×0.90) = √0.45 = 0.67

◆ Normal Distribution

The Normal (Gaussian) Distribution is a continuous, symmetrical, bell-shaped distribution. It is the most important distribution in statistics and is used extensively in pharmaceutical sciences.

Normal PDF	$f(x) = (1/\sigma\sqrt{2\pi}) \cdot e^{\wedge}[-(x-\mu)^2/2\sigma^2]$	μ=mean, σ=SD
Standard Normal (Z)	$Z = (x - \mu) / \sigma$	Converts to standard normal N(0,1)

Properties of Normal Distribution:

- Mean = Median = Mode (perfectly symmetric)
- Bell-shaped, asymptotic to x-axis
- Total area under curve = 1 (100%)
- 68.27% of data lies within μ ± 1σ
- 95.45% of data lies within μ ± 2σ
- 99.73% of data lies within μ ± 3σ (3-sigma rule / empirical rule)

- Defined by only two parameters: μ (location) and σ (shape/spread)

Range	Area Under Curve	Pharmaceutical Use
μ ± 1σ	68.27%	*Normal biological variation*
μ ± 2σ	95.45%	*95% confidence interval*
μ ± 3σ	99.73%	*Process control limits (±3σ rule)*
μ ± 1.96σ	95.00%	*Exact 95% CI in clinical trials*
μ ± 2.576σ	99.00%	*99% confidence interval*

◆ Poisson's Distribution

Poisson Distribution gives the probability of a given number of events occurring in a fixed interval (time, area, volume), when events occur independently at a constant average rate (λ).

Poisson P(X=x)	$P(X=x) = (e^{-\lambda} \cdot \lambda^{x}) / x!$	λ=mean rate, e=2.71828, x=0,1,2...

Properties of Poisson Distribution:

- Mean = Variance = λ (unique property — equal mean and variance)
- Standard Deviation = $\sqrt{\lambda}$
- Applicable when n is large, p is very small, and np = λ (constant)
- Discrete distribution (counts of events)
- Shape: Right-skewed for small λ; approaches normal as λ increases

◆ Comparison of Probability Distributions

Feature	Binomial	Normal	Poisson
Data type	Discrete	Continuous	Discrete (counts)
Outcomes	2 (success/fail)	Any value	Count of events
Parameters	n, p	μ, σ	λ (= mean = variance)
Mean	np	μ	λ
Variance	npq	σ^2	λ
Shape	Symmetric (p=0.5)	Bell-shaped	Right-skewed
Pharma use	Defect rate	Tablet weight, PK	Particle counts, ADR rate

Sampling, Population & Hypothesis Testing

◆ Population vs Sample

Feature	Population	Sample
Definition	Complete set of all items under study	Subset selected from the population
Size	N (usually large or infinite)	n (smaller, manageable)
Parameters	μ (mean), σ (SD), π (proportion)	$\bar{x}$ (mean), s (SD), p (proportion)
Study	Census / Complete enumeration	Sampling study
Cost & Time	Expensive, time-consuming	Economical, faster
Pharma example	All tablets in a batch	20 tablets selected for QC

◆ Large Sample vs Small Sample

Feature	Large Sample ($n \geq 30$)	Small Sample ($n < 30$)
Distribution	Normal (Z-distribution)	t-distribution (Gosset)
Test used	Z-test	t-test
Parameters	Known or estimated	Unknown, estimated from sample
Accuracy	High	Lower, needs t-correction
Pharma use	Large clinical trials	Pilot studies, early phase trials

◆ Null & Alternative Hypothesis

Hypothesis Testing is a formal procedure to decide whether evidence from a sample is sufficient to reject a stated assumption (null hypothesis) about a population parameter.

Hypothesis	Definition & Example
Null Hypothesis (H_0)	States that there is NO significant difference or effect. It is the hypothesis being tested and assumed true until evidence says otherwise. Ex: H_0: New drug has same efficacy as standard drug ($\mu 1 = \mu 2$)

Alternative Hypothesis (H_1 or Ha)	States that there IS a significant difference or effect. Accepted when H_0 is rejected. Ex: H_1: New drug is more efficacious ($\mu 1 \neq \mu 2$, or $\mu 1 > \mu 2$, or $\mu 1 < \mu 2$)

Types of Alternative Hypothesis:

- Two-tailed test: H_1: $\mu 1 \neq \mu 2$ (difference in either direction — used when direction is unknown)
- Right-tailed test: H_1: $\mu 1 > \mu 2$ (testing if new drug is better)
- Left-tailed test: H_1: $\mu 1 < \mu 2$ (testing if new drug is worse)

◆ Types of Sampling

Type of Sampling	Description & Pharmaceutical Example
Simple Random Sampling	Every item has equal chance of selection. Ex: Every tablet in batch assigned a number; 20 drawn randomly using random number table
Stratified Sampling	Population divided into strata (layers), samples drawn from each stratum. Ex: Clinical trial patients stratified by age group (<40, 40–60, >60 years)
Systematic Sampling	Every kth item selected. Ex: Every 10th tablet from production line tested (k=batch size/sample size)
Cluster Sampling	Population divided into clusters; entire cluster selected. Ex: All patients from 5 randomly selected hospitals in a multicentric study
Convenience Sampling	Easiest available subjects selected. Ex: Patients attending OPD on a particular day
Purposive/Judgement Sampling	Researcher selects based on knowledge. Ex: Expert selecting reference standards for analytical validation

◆ Type I and Type II Errors

In hypothesis testing, two types of errors can be made when deciding to reject or not reject H_0.

	H_0 is TRUE (reality)	H_0 is FALSE (reality)
Reject H_0 (decision)	**TYPE I ERROR (α) — False Positive Rejecting a true H_0**	**CORRECT DECISION True Positive (Power = 1−β)**
Accept H_0 (decision)	**CORRECT DECISION True Negative (Confidence = 1−α)**	**TYPE II ERROR (β) — False Negative Accepting a false H_0**

Pharmaceutical Significance:

- Type I Error ($\alpha = 0.05$): Concluding a drug works when it actually does NOT → False efficacy claim → Dangerous to patients
- Type II Error ($\beta = 0.20$): Concluding a drug does NOT work when it actually DOES → Missed effective treatment
- Power of test = $1 - \beta$ = probability of correctly detecting a real effect
- In drug trials: Regulatory agencies set $\alpha = 0.05$ (5% significance level)

Memory trick: Type I = FALSE ALARM (fire drill when no fire); Type II = MISSED ALARM (no drill when there IS a fire). Both errors have serious consequences in clinical research.

◆ Standard Error of Mean (SEM)

Standard Error of Mean (SEM) measures the precision of the sample mean as an estimate of the population mean. It indicates how much the sample mean is expected to vary from the true population mean.

SEM	**SEM = s / √n OR σ / √n**	s=sample SD, n=sample size

Points about SEM:

- SEM decreases as sample size (n) increases → larger samples give more precise estimates
- SEM < SD always (SEM = SD/√n)
- 95% Confidence Interval for μ: $\bar{x} \pm 1.96 \times$ SEM (large sample)
- 95% CI (small sample): $\bar{x} \pm t(\alpha/2, n-1) \times$ SEM
- Reported as: Mean ± SEM in pharmaceutical research papers

Pharmaceutical Example — SEM

A pharmacist measured serum drug concentration in 16 patients: Mean ($\bar{x}$) = 12.4 μg/mL, SD (s) = 2.8 μg/mL, n = 16 SEM = s / √n = 2.8 / √16 = 2.8 / 4 = 0.70 μg/mL Reported as: Serum concentration = 12.4 ± 0.70 μg/mL (Mean ± SEM) 95% CI = $\bar{x} \pm 1.96 \times$ SEM = 12.4 ± 1.96 × 0.70 = 12.4 ± 1.37 = (11.03, 13.77) μg/mL Interpretation: We are 95% confident that the true population mean serum concentration lies between 11.03 and 13.77 μg/mL.

Parametric Tests

Parametric tests are statistical tests that assume the data follows a specific distribution (usually normal). They use population parameters (μ, σ). They require: normally distributed data, continuous variables, and homogeneity of variance.

◆ Student's t-test

The t-test compares means when sample size is small (n<30) or when population SD is unknown. It uses the t-distribution with (n−1) degrees of freedom.

▸ A) One-Sample t-test (Single Sample t-test)

Tests whether a sample mean differs significantly from a known or hypothesized population mean.

One-sample t	$t = (\bar{x} - \mu_0) / (s / \sqrt{n})$	df = n−1, μ_0 = hypothesized mean

▸ B) Unpaired (Pooled/Independent) t-test

Compares means of two independent groups. Used when two separate groups of subjects receive different treatments.

Unpaired t	$t = (\bar{x}_1 - \bar{x}_2) / Sp\sqrt{(1/n_1 + 1/n_2)}$	df = n_1+n_2-2
Pooled SD (Sp)	$Sp = \sqrt{[(n_1-1)s_1^2 + (n_2-1)s_2^2] / (n_1+n_2-2)}$	Weighted average of both SDs

▸ C) Paired t-test

Compares means of the same group measured twice (before-after study). Uses differences (d = $x_1 - x_2$) between paired observations.

Paired t	$t = \bar{d} / (Sd / \sqrt{n})$	df = n−1, $\bar{d}$=mean difference, Sd=SD of differences

Feature	One-sample t	Unpaired t	Paired t
Compares	Sample vs known μ	Two independent groups	Same group: before vs after
Groups	1	2 (independent)	1 (measured twice)
df	n−1	n_1+n_2-2	n−1 (pairs)

Pharma use	QC vs pharmacopoeial standard	Drug A vs Drug B	Pre-post drug effect

◆ Analysis of Variance (ANOVA)

ANOVA tests whether the means of three or more groups are significantly different. It compares the variance BETWEEN groups to variance WITHIN groups using the F-ratio. It avoids the inflated Type I error of multiple t-tests.

F-ratio (ANOVA)	**F = MSB / MSW = Variance Between Groups / Variance Within Groups**	F ≥ 1; larger F = more significant

▸ A) One-Way ANOVA

Compares means of 3 or more independent groups based on ONE factor (independent variable).

Steps in One-Way ANOVA:

- Step 1: State H_0: $\mu_1=\mu_2=\mu_3=...=\mu k$ (all group means are equal)
- Step 2: Calculate Grand Mean ($\bar{\bar{x}}$), Group Means ($\bar{x}_i$)
- Step 3: Calculate SS Between (SSB) = $\Sigma n_i(\bar{x}_i - \bar{\bar{x}})^2$
- Step 4: Calculate SS Within (SSW) = $\Sigma\Sigma(x_{ij} - \bar{x}_i)^2$
- Step 5: Total SS (SST) = SSB + SSW
- Step 6: Degrees of freedom: dfB = k−1, dfW = N−k
- Step 7: MSB = SSB/dfB; MSW = SSW/dfW
- Step 8: F_calc = MSB/MSW; Compare with F_table (dfB, dfW, α=0.05)
- Step 9: If F_calc > F_table → Reject H_0 → Significant difference exists

Source	SS	df	MS	F	p-value
Between (Treatment)	SSB	k−1	MSB=SSB/(k−1)	**MSB/MSW**	*Compare with F_table*
Within (Error)	SSW	N−k	MSW=SSW/(N−k)	—	—
Total	SST	N−1	—	—	—

▸ B) Two-Way ANOVA

Tests the effect of TWO factors simultaneously and their interaction on the dependent variable.

Two-Way ANOVA tests: H_0A: No significant effect of Factor A H_0B: No significant effect of Factor B H_0AB: No significant interaction between A and B

Source	SS	df	MS	F	p-value
Factor A (Row)	SSA	r−1	MSA=SSA/(r−1)	**MSA/MSE**	*p<0.05?*
Factor B (Column)	SSB	c−1	MSB=SSB/(c−1)	**MSB/MSE**	*p<0.05?*
Interaction A×B	SSAB	(r−1)(c−1)	MSAB	**MSAB/MSE**	*p<0.05?*
Error (Within)	SSE	rc(n−1)	MSE	—	—
Total	SST	N−1	—	—	—

Two-Way ANOVA — Pharmaceutical Example

- **Study design:** Effect of (Factor A) Drug Dose [Low/High] and (Factor B) Route [Oral/IV] on drug absorption (%).
- **Factor A (Dose):** 2 levels (r=2) Factor B (Route): 2 levels (c=2) n=3 replicates per cell → Total N=12
- **Hypotheses tested:** H_0A: Low and High dose give same absorption
- **H_0B:** Oral and IV give same absorption
- **H_0AB:** There is no interaction between dose and route If Interaction F is significant → Effect of dose depends on route (and vice versa). Both main effects must be interpreted with caution. If Interaction NOT significant → Interpret each main effect independently.

◆ Least Significant Difference (LSD) Test

LSD (Fisher's LSD) is a post-hoc test performed AFTER ANOVA, when H_0 is rejected, to identify WHICH specific pairs of group means differ significantly.

LSD	**$LSD = t(\alpha/2, dfW) \times \sqrt{[MSW \times (1/n_1 + 1/n_2)]}$**	Compare each pair: if $\lvert\bar{x}_i - \bar{x}_j\rvert > LSD$ → significant

Steps for LSD:

- Step 1: Perform ANOVA and confirm F is significant (reject H_0)
- Step 2: Get MSW and dfW from ANOVA table
- Step 3: Calculate LSD using formula above with t_table value
- Step 4: Compare all pairwise differences $|\bar{x}_i - \bar{x}_j|$ with LSD
- Step 5: If $|\bar{x}_i - \bar{x}_j| > LSD \rightarrow$ That pair is significantly different

◆ Expected Exam Questions — Unit 2

Q	Question	Marks
1	What is regression? Explain the method of least squares. Derive the normal equations for y=a+bx.	5–10
2	Differentiate correlation and regression. State their pharmaceutical applications.	5
3	Explain Binomial distribution with properties and pharmaceutical problem.	5–10
4	What is Normal distribution? Explain its properties and significance in pharmacy (68-95-99.7 rule).	5–10
5	Explain Poisson distribution. A batch has λ=2 defects. Find P(0), P(1), P(2) defects.	5
6	What is a null hypothesis and alternative hypothesis? Explain Type I and Type II errors with examples.	5
7	Explain different types of sampling with pharmaceutical examples.	5
8	What is SEM? How is it different from SD? Calculate SEM for given data.	3–5
9	Explain paired t-test with a pharmaceutical example (before-after design).	5–10
10	What is ANOVA? Explain one-way ANOVA with the ANOVA table. Solve a problem with 3 batches.	10
11	What is LSD test? When is it used? Perform LSD test on ANOVA results.	5
12	Write short notes on: (a) Multiple regression (b) Standard error of regression.	5