NOTESKARTS

noteskarts.com

B. Pharmacy — 8th Semester

BIOSTATISTICS AND RESEARCH METHODOLOGY

UNIT 1 — Comprehensive Notes

Subject Code	Semester	Unit	Source
BP801T	8th Semester	Unit 1 of 5	Noteskarts.com

UNIT 1 SYLLABUS COVERAGE

1.1	Introduction to Statistics & Biostatistics	*Definition, Scope, Importance*
1.2	Frequency Distribution	*Types, Tabulation, Histogram, Ogive*
1.3	Measures of Central Tendency	*Mean, Median, Mode with Pharma Examples*
1.4	Measures of Dispersion	*Range, Variance, Standard Deviation*
1.5	Correlation	*Karl Pearson's, Multiple Correlation – Pharma Problems*

Introduction to Statistics & Biostatistics

◆ What is Statistics?

Statistics is the science of collecting, organizing, summarizing, analyzing, and interpreting numerical data to draw meaningful conclusions and make informed decisions.

Statistics can be divided into two main branches:

Descriptive Statistics	Inferential Statistics
Summarizes and describes the collected data	Draws conclusions about a population from a sample
Mean, Median, Mode, SD, Range	*t-test, ANOVA, Chi-square, Regression*

◆ What is Biostatistics?

Biostatistics (Biological Statistics) is the application of statistical methods to biological, health, pharmaceutical, and medical sciences. It helps in designing experiments, analyzing drug data, and validating research findings.

Applications of Biostatistics in Pharmacy:

- Designing and analyzing clinical trials for new drugs
- Evaluating drug efficacy and safety profiles
- Quality control in pharmaceutical manufacturing
- Epidemiological studies and disease surveillance
- Stability testing of pharmaceutical formulations
- Bioavailability and bioequivalence studies
- Pharmacokinetic and pharmacodynamic modeling

◆ Importance of Biostatistics in Pharmacy

S.N.	Area	Role of Biostatistics
1	**Drug Development**	Determines sample size, analyzes Phase I-IV trial data
2	**Quality Control**	Monitors batch variability, acceptance sampling
3	**Pharmacokinetics**	Analyzes drug concentration-time data
4	**Epidemiology**	Studies disease prevalence, drug side-effects
5	**Regulatory Affairs**	CDSCO/FDA submission requires statistical analysis

◆ Basic Statistical Terms

Term	Definition
Population	Complete set of individuals/items from which data is collected (e.g., all diabetic patients in India)
Sample	A subset of population selected for study (e.g., 200 diabetic patients from Delhi)
Variable	Any characteristic that can vary between individuals (e.g., blood glucose level, drug dose)
Data	Observed values of variables (e.g., 120 mg/dL, 250 mg tablet)
Parameter	Numerical summary of a population (μ = population mean, σ = population SD)
Statistic	Numerical summary of a sample ($\bar{x}$ = sample mean, s = sample SD)

Frequency Distribution

A Frequency Distribution is an organized tabular arrangement showing how frequently each value or group of values occurs in a dataset. It converts raw data into a meaningful form.

◆ Terms in Frequency Distribution

Term	Meaning
Class Interval	A range of values grouped together (e.g., 10–20, 20–30)
Class Width	Difference between upper and lower limits of a class (e.g., 20–10 = 10)
Frequency (f)	Number of times a value falls within a class interval
Relative Frequency	Frequency of a class divided by total frequency (f/N)
Cumulative Frequency	Running total of frequencies from the beginning
Class Mark / Midpoint	Middle value of a class interval = (Lower + Upper) / 2

◆ Pharmaceutical Example — Frequency Distribution Table

Example: Dissolution (%) values of 30 tablets of Paracetamol 500 mg:

Raw data: 76, 82, 89, 74, 91, 85, 78, 83, 87, 90, 79, 84, 88, 76, 93, 81, 86, 77, 92, 84, 80, 85, 89, 75, 91, 88, 83, 78, 87, 86

Class Interval (%)	Frequency (f)	Class Mark (xi)	Rel. Freq.	Cumul. Freq.
70–75	2	72.5	0.067	2
75–80	6	77.5	0.200	8
80–85	8	82.5	0.267	16
85–90	9	87.5	0.300	25
90–95	5	92.5	0.167	30
Total	**30**	–	**1.000**	–

Interpretation: Maximum tablets (9 out of 30) showed dissolution between 85–90%, indicating most tablets meet pharmacopoeial dissolution standards (NLT 80%).

◆ Types of Frequency Distribution

Type	Description & Example
Simple / Ungrouped	Frequency of each individual value. Used for discrete data. Ex: Number of tablets passing/failing
Grouped Frequency	Values grouped into class intervals. Used for continuous data. Ex: Dissolution %, Particle size
Relative Frequency	Each class frequency expressed as proportion of total. Used for comparison
Cumulative Frequency	Running sum of frequencies. Basis for Ogive curve
Normal Distribution	Bell-shaped, symmetric frequency distribution around mean. Common in pharmacokinetics

Measures of Central Tendency

Measures of Central Tendency are statistical values that represent the 'centre' or 'typical value' of a dataset. The three main measures are Mean, Median, and Mode.

◆ Mean (Arithmetic Mean)

The Mean is the sum of all observations divided by the total number of observations. It is the most commonly used measure of central tendency.

MEAN	$\bar{x} = \Sigma x / n$	x=values, n=count

MEAN (Grouped)	$\bar{x} = \Sigma(f \cdot xi) / \Sigma f$	f=freq, xi=midpoint

Properties of Mean:

- Uses all data values — most sensitive to extreme values (outliers)
- Sum of deviations from mean is always zero: $\Sigma(x - \bar{x}) = 0$
- Algebraically tractable — can be used in further calculations
- Affected by skewed distribution and extreme values

Pharmaceutical Example — Mean

Problem: A pharmacist measured the weight of 8 tablets (mg): 498, 502, 497, 503, 499, 501, 500, 500 Mean = (498+502+497+503+499+501+500+500) / 8 = 4000 / 8 = 500 mg Interpretation: Average tablet weight = 500 mg, which meets USP ±5% limit (475–525 mg). All tablets within spec.

◆ Median

The Median is the middle value when data is arranged in ascending or descending order. It divides the data into two equal halves.

Odd Number (n) of Values	Even Number (n) of Values
Median = value at position (n+1)/2	**Median = Average of n/2 and (n/2+1) values**
n=7: Median = 4th value	*n=8: Median = avg of 4th & 5th values*

Formula for Grouped Data:

MEDIAN (Grouped)	$M = L + [(N/2 - cf) / f] \times h$	L=lower limit, cf=cumul freq before, f=class freq, h=width

Pharmaceutical Example — Median

Problem: Disintegration times (seconds) of 7 tablets: 45, 52, 38, 60, 47, 55, 42 Step 1: Arrange in order: 38, 42, 45, 47, 52, 55, 60 Step 2: n = 7 (odd), Median = (7+1)/2 = 4th value = 47 seconds Interpretation: 50% tablets disintegrate in ≤47 seconds. Not affected by extreme values.

◆ Mode

Mode is the value that appears most frequently in a dataset. A dataset can have No mode, One mode (Unimodal), Two modes (Bimodal), or Multiple modes (Multimodal).

MODE (Grouped)	$Mo = L + [(f1 - f0) / (2f1 - f0 - f2)] \times h$	f1=modal class freq, f0=preceding freq, f2=following freq

Pharmaceutical Example — Mode

Problem: Side effects reported by patients: Nausea(15), Headache(22), Dizziness(8), Headache(22), Vomiting(10) Mode = Headache (appears most frequently = 22 cases) Pharmaceutical Use: Identifies the most common adverse drug reaction (ADR) — useful in pharmacovigilance reporting.

◆ Comparison of Mean, Median & Mode

Property	Mean	Median	Mode
Definition	Arithmetic average	Middle value	Most frequent
Uses all values	Yes	No	No
Affected by outliers	Highly	Less	Not affected
Can be calculated for	Continuous data	Continuous data	All data types
Pharma application	Avg tablet weight	Median shelf life	Most common ADR
Mathematical use	Further calculations	Non-parametric tests	Categorical data

📐 Empirical Relationship: $Mode = 3 \times Median - 2 \times Mean$ (holds for moderately skewed distributions)

Measures of Dispersion

Measures of Dispersion quantify the spread or variability in a dataset. They show how much the individual values deviate from the central value (mean). High dispersion = less uniformity; Low dispersion = more uniformity.

◆ What is Dispersion?

Two datasets can have the same mean but very different variability. Example:

Dataset	Values	Mean
Batch A Tablet Weights (mg)	498, 500, 499, 501, 502	**500 mg**
Batch B Tablet Weights (mg)	480, 510, 490, 520, 500	**500 mg**

Both batches have the same mean (500 mg), but Batch B has much greater variability — this is pharmacologically unacceptable. Measures of dispersion help detect such differences.

◆ Range

Range is the simplest measure of dispersion. It is the difference between the largest and smallest values.

RANGE	$R = X_{max} - X_{min}$	Max minus Min value

Pharmaceutical Example — Range

Problem: Hardness (kg/cm²) of tablets: 5.2, 6.8, 5.9, 7.2, 6.1, 5.5, 7.0 Range = 7.2 − 5.2 = 2.0 kg/cm² Interpretation: The hardness varies by 2 kg/cm². USP limit: 4–8 kg/cm². All in range but variability should be minimized in production.

Limitations of Range:

- Considers only two extreme values — ignores all other data
- Highly affected by outliers
- Not suitable for open-ended frequency distributions
- Not useful when comparing datasets with different sizes

◆ Variance

Variance is the average of squared deviations from the mean. It gives a measure of how spread the data is.

Population Variance (σ^2)	Sample Variance (s^2)
$\sigma^2 = \Sigma(x - \mu)^2 / N$	$s^2 = \Sigma(x - \bar{x})^2 / (n-1)$

Used when all data of population is known	*Used for sample data (Bessel's correction: n−1)*

◆ 1.4.4 Standard Deviation (SD)

Standard Deviation (SD) is the square root of variance. It is expressed in the same units as the original data, making it the most widely used measure of dispersion in pharmaceutical sciences.

Population SD (σ)	Sample SD (s)
$\sigma = \sqrt{[\Sigma(x - \mu)^2 / N]}$	$s = \sqrt{[\Sigma(x - \bar{x})^2 / (n-1)]}$

Step-by-Step Formula for SD Calculation:

- Step 1: Calculate the mean ($\bar{x}$)
- Step 2: Find deviation of each value from mean: $(x - \bar{x})$
- Step 3: Square each deviation: $(x - \bar{x})^2$
- Step 4: Sum all squared deviations: $\Sigma(x - \bar{x})^2$
- Step 5: Divide by $(n-1)$ for sample: $s^2 = \Sigma(x - \bar{x})^2 / (n-1)$
- Step 6: Take square root: $s = \sqrt{s^2}$

Pharmaceutical Problem — Standard Deviation

Problem: Blood glucose levels (mg/dL) of 6 patients after Metformin therapy: 120, 135, 118, 142, 128, 133 Step 1: Mean = (120+135+118+142+128+133)/6 = 776/6 = 129.33 mg/dL Step 2: Deviations: −9.33, 5.67, −11.33, 12.67, −1.33, 3.67 Step 3: Squared deviations: 87.05, 32.15, 128.37, 160.53, 1.77, 13.47 Step 4: Sum = 423.34 Step 5: s^2 = 423.34 / (6−1) = 84.67 Step 6: s = √84.67 = 9.20 mg/dL Result: SD = 9.20 mg/dL | Interpretation: Average blood glucose = 129.33 ± 9.20 mg/dL

◆ Coefficient of Variation (CV)

CV expresses SD as a percentage of the mean. It is used to compare variability between datasets with different units or means.

CV	$CV = (s / \bar{x}) \times 100\ \%$	Lower CV = more uniform batch

📌 Pharmaceutical Standard: CV < 2% for tablet weight uniformity (USP). CV < 6% for content uniformity. If CV > 5%, batch-to-batch variability is high and must be investigated.

◆ Measures of Dispersion

Measure	Formula	Advantage	Disadvantage
Range	Xmax − Xmin	Simple, easy	Ignores other values
Variance	$\Sigma(x-\bar{x})^2/(n-1)$	Uses all values	Unit = square of data
Std. Dev.	√Variance	Same unit as data	Affected by outliers
CV	(SD/Mean)×100	Unitless comparison	Meaningless if mean≈0

Correlation

Correlation is a statistical measure that describes the degree and direction of the linear relationship between two or more variables. It helps understand how one variable changes with respect to another.

◆ Types of Correlation

Type	Direction	Example in Pharmacy
Positive Correlation	Both variables move in same direction	*Drug dose ↑ → Drug effect ↑*
Negative Correlation	Variables move in opposite direction	*Drug concentration ↑ → Survival rate ↓ (in toxicity)*
Zero / No Correlation	No relationship between variables	*Patient height vs drug efficacy*
Perfect Positive	r = +1.0, exact linear relationship	*Ideal standard calibration curve*
Perfect Negative	r = −1.0, exact inverse relationship	*Drug clearance vs drug half-life*

◆ Karl Pearson's Coefficient of Correlation (r)

Karl Pearson's coefficient of correlation (r) is the most widely used method to measure the linear correlation between two continuous variables.

Pearson's r	$r = \Sigma(x-\bar{x})(y-\bar{y}) / \sqrt{[\Sigma(x-\bar{x})^2 \cdot \Sigma(y-\bar{y})^2]}$	Population formula
r (simplified)	$r = [n\Sigma xy - \Sigma x\Sigma y] / \sqrt{\{[n\Sigma x^2-(\Sigma x)^2][n\Sigma y^2-(\Sigma y)^2]\}}$	Working formula

▸ Interpretation of r Value:

Value of r	Strength	Pharmaceutical Interpretation
+1.0	Perfect Positive	*Ideal calibration curve linearity*
+0.7 to +0.99	Strong Positive	*High dose-response correlation*
+0.3 to +0.69	Moderate Positive	*Drug conc. vs partial effect*
0 to +0.29 / 0 to −0.29	Weak / No correlation	*Unrelated variables*
−0.3 to −0.69	Moderate Negative	*Clearance vs blood level*
−0.7 to −0.99	Strong Negative	*Toxicity vs dose (therapeutic window)*
−1.0	Perfect Negative	*Inverse relationship*

▸ Properties of r:

- r is dimensionless (no units)
- $-1 \leq r \leq +1$ always
- r is independent of change of origin and scale
- If r = 0, variables are uncorrelated (but may have non-linear relationship)
- Pearson's r is sensitive to outliers — use Spearman's rank correlation for non-normal data

Pharmaceutical Problem — Karl Pearson's r

A study investigated correlation between drug concentration (x, µg/mL) and response (%inhibition, y): x: 2, 4, 6, 8, 10 y: 18, 32, 48, 65, 75 Calculations: n=5, $\Sigma x=30$, $\Sigma y=238$, $\Sigma x^2=220$, $\Sigma y^2=13282$, $\Sigma xy=1660$ $r = [5\times1660 - 30\times238] / \sqrt{\{[5\times220-30^2][5\times13282-238^2]\}}$ $r = [8300 - 7140] / \sqrt{\{[1100-900][66410-56644]\}}$ $r = 1160 / \sqrt{\{200 \times 9766\}}$ $r = 1160 / \sqrt{1953200}$ $r = 1160 / 1397.57$ $r = +0.830$ Result: r = +0.83 → Strong positive correlation between drug concentration and inhibitory response.

◆ Multiple Correlation

Multiple Correlation measures the relationship between one dependent variable (Y) and two or more independent variables (X1, X2, X3...). The Multiple Correlation Coefficient is denoted as R (capital R).

Multiple R (R^2)	$R^2y.12 = (r^2y1 + r^2y2 - 2ry1 \cdot ry2 \cdot r12) / (1 - r^212)$	R = multiple correlation coefficient

Where:

- $R^2y.12$ = Coefficient of multiple determination
- ry1 = correlation of Y with X1
- ry2 = correlation of Y with X2
- r12 = correlation between X1 and X2

▸ Pharmaceutical Example — Multiple Correlation:

Variable	Represents	Role
Y	Drug Dissolution (%)	*Dependent variable*
X1	Binder concentration (%)	*Independent variable 1*
X2	Compression Force (kN)	*Independent variable 2*

Given: ry1 = 0.85, ry2 = 0.78, r12 = 0.60

$R^2y.12 = (0.85^2 + 0.78^2 - 2 \times 0.85 \times 0.78 \times 0.60) / (1 - 0.60^2)$

$R^2y.12 = (0.7225 + 0.6084 - 0.7956) / (1 - 0.36)$

$R^2y.12 = 0.5353 / 0.64 = 0.8364$

$\boldsymbol{R\ y.12 = \sqrt{0.8364} = 0.915}$

Interpretation: R = 0.915 — Strong multiple correlation. Binder concentration and compression force together explain 83.6% of the variation in drug dissolution (R^2 = 0.836).

◆ Correlation vs Causation

⚠ IMPORTANT: Correlation does NOT imply causation. A high correlation between two variables does not mean one causes the other. Always interpret correlation in the context of pharmacological logic and experimental design.

Correlation	Causation

Statistical relationship between variables	One variable directly influences another
Measured by r or R coefficient	*Established by controlled experiments*
Example: High r between drug dose and response	Example: Increasing dose CAUSES increased response (pharmacodynamics)

◆ Expected Exam Questions

Q	Question	Marks
1	Define Statistics and Biostatistics. What is its importance in pharmacy?	2–5 Marks
2	What is a frequency distribution? Draw and explain frequency distribution table for given pharmaceutical data.	5 Marks
3	Explain Mean, Median, Mode with pharmaceutical examples.	5–10 Marks
4	What are measures of dispersion? Explain range, variance, and SD with examples.	5–10 Marks
5	Define standard deviation. Calculate SD for given tablet weight data.	5 Marks
6	Explain Karl Pearson's coefficient of correlation. Calculate r for a dose-response dataset.	5–10 Marks
7	Write a note on: (a) Multiple Correlation (b) Correlation vs Causation.	5 Marks
8	Differentiate between Descriptive and Inferential Statistics.	2 Marks
9	The dissolution % of 5 tablets are: 78, 82, 79, 85, 91. Find Mean, Median, SD.	5 Marks
10	What is coefficient of variation? How is it used in pharmaceutical QC?	3 Marks