B.PHARM SEMESTER II | BP201T

APPLIED BIOSTATISTICS AND DATA ANALYTICS

FOR PHARMACEUTICAL SCIENCES (THEORY)

UNIT II

Probability and Statistical Distributions in Healthcare

Short Notes as per PCI B.Pharm Syllabus

What is covered in this unit

UNIT II TOPICS

Part A – Probability: basic concepts, addition and multiplication rules, conditional probability, Bayes' theorem and clinical decision-making.

Part B – Random variables and normal distribution: discrete and continuous variables, the normal curve and its importance in biological and pharmaceutical measurements.

Part C – Binomial and Poisson distributions: clinical trial outcomes and rare adverse drug reactions.

Part D – Python: probabilities, simulations and graphs of the distributions using SciPy and Matplotlib.

Part E – Practice: solved numerical problems and formulae.

noteskarts.com | Noteskarts Prime App | YouTube: Noteskarts

Contents of Unit II

Part	Section	Topic
A	1	Basic concepts of probability
A	2	Laws of probability: addition and multiplication
A	3	Conditional probability and its interpretation
A	4	Bayes' theorem and clinical decision-making
B	5	Random variables (discrete and continuous)
B	6	Normal distribution
C	7	Binomial distribution
C	8	Poisson distribution
C	9	Comparison of distributions and how to choose
D	10	Calculating probabilities with SciPy
D	11	Bayes, simulation and graphs in Python
E	12	Solved numerical problems
E	13	Formulae used in this unit

Noteskarts
Smart Notes, Bright Future

PART A: PROBABILITY AND BAYES' THEOREM

1. Basic Concepts of Probability

Probability is a number between 0 and 1 that measures how likely an event is to occur. In healthcare almost nothing is certain: a drug may work in 70% of patients, a test may give a false result, an adverse reaction may occur in 1 of 1000 patients. Probability gives a language and a set of rules to handle this uncertainty, and it is the foundation of inferential statistics (Unit III).

Basic terms

Term	Meaning	Example
Experiment (trial)	A process whose result is uncertain	Giving a drug to a patient and observing the response
Outcome	One possible result of the experiment	Patient responds / does not respond
Sample space (S)	The set of all possible outcomes	S = {responds, does not respond}; for a die S = {1, 2, 3, 4, 5, 6}
Event (A, B…)	A set of one or more outcomes	'The patient has nausea'; 'the tablet is defective'
Simple and compound event	Simple: one outcome. Compound: two or more outcomes combined	Getting a 6 (simple); getting an even number (compound)
Mutually exclusive events	Cannot occur together	A patient is either male or female; a tablet is either defective or not
Exhaustive events	Together cover the whole sample space	A, B, AB and O blood groups
Independent events	Occurrence of one does not change the probability of the other	Responses of two unrelated patients to a drug
Complement (A′ or A^c)	'Not A': everything in S that is outside A	A = ADR occurs; A′ = no ADR

Definitions of probability

Approach	Definition	Example
Classical (a priori)	P(A) = number of favourable outcomes / total number of equally likely outcomes	A box holds 20 ampoules, 4 defective. P(defective) = 4/20 = **0.20**
Empirical (relative frequency)	P(A) = number of times A occurred / total number of trials, for a large number of trials	In 1000 patients, 30 developed rash. P(rash) ≈ 30/1000 = **0.03**
Subjective	A degree of belief based on experience or expert judgement	A physician estimates a 60% chance that a patient will respond

The relative-frequency approach is the one used in medical statistics. As the number of trials increases, the observed proportion settles close to the true probability (**law of large numbers**, demonstrated with Python in Part D).

Expressing risk: probability, percentage, odds and natural frequencies

Form	Example (same information)	Remarks
Probability	0.15	Between 0 and 1; used in calculations
Percentage	15%	Probability × 100; common in reports
Odds	0.15 / 0.85 = 0.176, about 3 : 17	Used in logistic regression and case-control studies. **Odds = p / (1 − p); p = odds / (1 + odds)**
Natural frequency	15 out of every 100 patients	The easiest form for patient counselling; people misunderstand percentages and conditional probabilities more often than counts

- **Absolute risk versus relative risk:** a drug that raises the risk of a reaction from 1 in 10,000 to 3 in 10,000 has a relative risk of 3 (a 'tripling') but an absolute increase of only 2 in 10,000. A pharmacist should give **both** figures when counselling.
- For rare events probability and odds are almost equal; for common events they differ greatly (p = 0.5 means odds = 1).

Counting rules (needed for the binomial distribution)

- **Factorial:** $n! = n \times (n-1) \times \ldots \times 2 \times 1$; by definition $0! = 1$. Example: $5! = 120$.

- **Permutation** (order matters): nPr = n! / (n − r)!
- **Combination** (order does not matter): **nCr = n! / [r! (n − r)!]**. Example: the number of ways to choose 3 patients out of 10 for a trial arm = 10C3 = 120. Also nC0 = nCn = 1 and nCr = nC(n − r).

2. Laws of Probability: Addition and Multiplication Rules

A. Addition rule (the 'OR' rule)

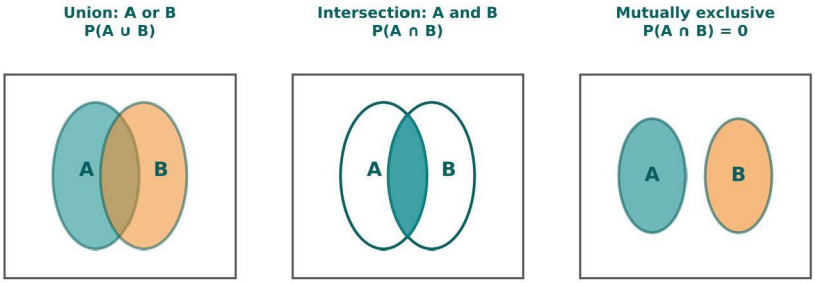

Figure 1: Venn diagrams for union, intersection and mutually exclusive events

- **Mutually exclusive events:** P(A or B) = P(A) + P(B)
- **Any two events (general rule): P(A or B) = P(A) + P(B) − P(A and B)**. The overlap is subtracted because it was counted twice.
- **Three or more mutually exclusive events:** add all the probabilities.

WORKED EXAMPLE 1: ADDITION RULE

In a drug study of 200 patients, 30 had nausea (N), 20 had headache (H) and 8 had both.

P(N) = 30/200 = 0.15; P(H) = 20/200 = 0.10; P(N and H) = 8/200 = 0.04.

P(nausea or headache) = 0.15 + 0.10 − 0.04 = 0.21. P(neither) = 1 − 0.21 = **0.79**.

Mutually exclusive case: in a population, blood group probabilities are O = 0.37, B = 0.32, A = 0.22, AB = 0.09. A donor is either A or B: P(A or B) = 0.22 + 0.32 = **0.54** (no overlap, since a person has only one group).

B. Multiplication rule (the 'AND' rule)

- **Independent events:** P(A and B) = P(A) × P(B)
- **Dependent events (general rule): P(A and B) = P(A) × P(B | A)**, where P(B | A) is the probability of B after A has occurred (conditional probability, next section).
- For several independent events, multiply all the probabilities.

WORKED EXAMPLE 2: MULTIPLICATION RULE

Independent: a drug works in 70% of patients. Two unrelated patients are treated. P(both respond) = 0.7 × 0.7 = **0.49**. P(at least one responds) = 1 − P(neither) = 1 − 0.3 × 0.3 = **0.91**.

Dependent (sampling without replacement): a box has 20 ampoules, 4 defective. Two are drawn one after another without replacement. P(both defective) = (4/20) × (3/19) = 12/380 = **0.0316**.

P(at least one defective in 2) = 1 − P(none) = 1 − (16/20)(15/19) = 1 − 0.6316 = **0.368**.

WORKED EXAMPLE: COUNTING AND PROBABILITY (HYPERGEOMETRIC IDEA)

A ward has 10 patients, of whom 4 are responders to a new drug. A pharmacist selects 3 patients at random for a drug-utilisation review.

Number of ways to choose any 3 patients = 10C3 = **120.** Number of ways to choose 3 responders = 4C3 = **4. P(all three are responders) = 4/120 = 0.033.**

P(exactly 2 responders) = (4C2 × 6C1) / 120 = (6 × 6) / 120 = **0.30. P(no responder)** = 6C3 / 120 = 20/120 = **0.167.**

This is sampling **without replacement**, so the binomial formula cannot be used; combinations are used instead.

Situation	Rule	Key word
Either A or B (one or the other)	Addition: P(A) + P(B) − P(A and B)	OR → add
Both A and B	Multiplication: P(A) × P(B | A)	AND → multiply
At least one…	Use complement: 1 − P(none)	Complement
Not A	1 − P(A)	Complement

COMMON MISTAKES

Mutually exclusive ≠ independent. Mutually exclusive events cannot happen together (P(A and B) = 0), so knowing A has occurred tells us B has not: they are strongly **dependent**.

Do not add probabilities of events that overlap without subtracting the intersection.

3. Conditional Probability and its Interpretation

Conditional probability is the probability of an event **given that another event has already occurred**. It narrows the sample space to the group in which the given event is true.

FORMULA

P(A | B) = P(A and B) / P(B), provided P(B) > 0. Read as 'probability of A given B'.

Rearranged, this gives the multiplication rule: P(A and B) = P(B) × P(A | B).

WORKED EXAMPLE 3: ADR IN DRUG AND PLACEBO GROUPS

A trial of 400 patients: 200 received Drug X and 200 received placebo. Adverse reactions (ADR):

Drug X: ADR 30, no ADR 170 (total 200). **Placebo:** ADR 10, no ADR 190 (total 200). **All patients:** ADR 40, no ADR 360.

P(ADR) = 40/400 = 0.10. P(ADR | Drug X) = 30/200 = 0.15. P(ADR | Placebo) = 10/200 = 0.05.

P(Drug X | ADR) = 30/40 = 0.75. Among patients who had an ADR, 75% were on the drug.

Interpretation: the chance of an ADR is three times higher with the drug (0.15 versus 0.05, a relative risk of 3). Because P(ADR | Drug X) ≠ P(ADR), the ADR is **not independent** of the treatment.

Important points of interpretation

- **P(A | B) is generally not equal to P(B | A).** P(ADR | Drug X) = 0.15 but P(Drug X | ADR) = 0.75. Confusing the two is a common error in reading medical results (the 'prosecutor's fallacy').
- **Test for independence:** A and B are independent if P(A | B) = P(A), or equivalently P(A and B) = P(A) × P(B). If knowing B changes the probability of A, the events are dependent.
- **Risk and odds:** the conditional probability of disease given exposure is called **risk**. The ratio of two risks is the **relative risk**, a measure used in clinical trials and cohort studies. Odds = P / (1 − P).
- In a **contingency (2 × 2) table**, a conditional probability is simply a **row or column percentage**: choose the row (or column) of the given event as the denominator.

Risk, relative risk and odds ratio from conditional probabilities

Conditional probabilities from a 2 × 2 table lead directly to the measures used in clinical research. Using Worked example 3 (Drug X versus placebo):

Measure	Formula	Value in the example	Meaning
Risk (incidence) in drug group	P(ADR | Drug)	30/200 = 0.15	15% of treated patients had an ADR
Risk in placebo group	P(ADR | Placebo)	10/200 = 0.05	5% had an ADR
Relative risk (RR)	Risk(drug) / Risk(placebo)	0.15 / 0.05 = **3.0**	ADR is 3 times as likely with the drug; RR = 1 means no association
Risk difference	Risk(drug) − Risk(placebo)	0.10	10 extra ADRs per 100 treated patients
Number needed to harm (NNH)	1 / risk difference	1 / 0.10 = **10**	On average 1 extra ADR for every 10 patients treated
Odds	p / (1 − p)	Drug 0.15/0.85 = 0.176; Placebo 0.05/0.95 = 0.053	Chance of event relative to non-event
Odds ratio (OR)	Odds(drug) / Odds(placebo)	0.176 / 0.053 = **3.35**	Used in case-control studies; close to RR when the event is rare

For a **beneficial** outcome the same calculation gives the **absolute risk reduction (ARR)** and the **number needed to treat (NNT = 1/ARR)**. These measures are explained further in Unit III.

Total probability

If an event B can occur through several mutually exclusive and exhaustive routes A_1, A_2, …, A_n, then **$P(B) = P(A_1)\cdot P(B \mid A_1) + P(A_2)\cdot P(B \mid A_2) + \ldots + P(A_n)\cdot P(B \mid A_n)$**. This 'law of total probability' is the denominator of Bayes' theorem.

4. Bayes' Theorem and Clinical Decision-Making

Bayes' theorem reverses a conditional probability. It lets us start from what is easy to know, such as the chance of a positive test **if** the patient has the disease, and calculate what is clinically needed: the chance of disease **if** the test is positive. It updates an earlier ('prior') probability using new evidence.

BAYES' THEOREM

$P(A \mid B) = P(B \mid A) \times P(A) / P(B)$

For events $A_1 \ldots A_n$: **$P(A_i \mid B) = P(B \mid A_i)\, P(A_i) / \Sigma\, P(B \mid A_j)\, P(A_j)$**

P(A) = prior probability (before the evidence); **P(A | B)** = posterior probability (after the evidence); **P(B | A)** = likelihood.

WORKED EXAMPLE 4: BAYES' THEOREM IN QUALITY CONTROL

Machine A makes 60% of the tablets and 2% of its tablets are defective. Machine B makes 40% and 5% of its tablets are defective. A tablet is picked at random and found defective. Which machine made it?

Total probability of defective: P(D) = 0.60 × 0.02 + 0.40 × 0.05 = 0.012 + 0.020 = **0.032**.

P(A | D) = 0.012 / 0.032 = 0.375; P(B | D) = 0.020 / 0.032 = 0.625.

Although Machine B makes fewer tablets, it is the more likely source of a defective tablet (62.5%) because of its higher defect rate.

Diagnostic test terms

Term	Definition (conditional probability)	Formula from 2 × 2 table
Sensitivity	P(test positive | disease present): ability to detect the sick	TP / (TP + FN)
Specificity	P(test negative | disease absent): ability to clear the healthy	TN / (TN + FP)
Positive predictive value (PPV)	P(disease | test positive): how much a positive result can be trusted	TP / (TP + FP)
Negative predictive value (NPV)	P(no disease | test negative)	TN / (TN + FN)
Prevalence	P(disease) in the population tested (the prior)	(TP + FN) / total
Likelihood ratio, LR+	Sensitivity / (1 − specificity): how much a positive result raises the odds of disease	–
Likelihood ratio, LR−	(1 − sensitivity) / specificity: how much a negative result lowers the odds	–

TP = true positive, FP = false positive, TN = true negative, FN = false negative. **Sensitivity and specificity** are properties of the test; **PPV and NPV** also depend on the **prevalence**.

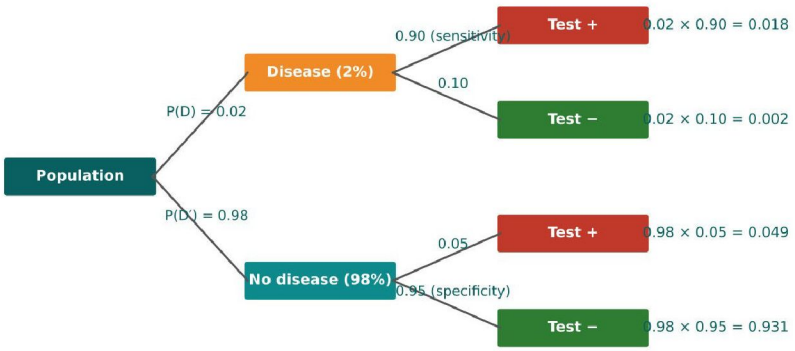

Figure 2: Probability tree for a screening test: prevalence 2%, sensitivity 90%, specificity 95%

WORKED EXAMPLE 5: SCREENING TEST

A screening test has sensitivity 90% and specificity 95%. The disease prevalence is 2%. Take 10,000 people.

Diseased (2%) = 200: TP = 200 × 0.90 = 180; FN = 20. **Healthy = 9,800:** FP = 9,800 × 0.05 = 490; TN = 9,310.

PPV = 180 / (180 + 490) = 180/670 = 0.269 (27%). NPV = 9,310 / (9,310 + 20) = 0.998 (99.8%).

By Bayes' formula: P(D | +) = (0.90 × 0.02) / (0.90 × 0.02 + 0.05 × 0.98) = 0.018 / 0.067 = **0.269**.

Interpretation: only about **1 in 4** positive people truly has the disease, although the test looks accurate. A negative result is very reliable (99.8%). **LR+ = 0.90/0.05 = 18; LR− = 0.10/0.95 = 0.105.**

If prevalence is 20% (a high-risk group): PPV = 1,800 / (1,800 + 400) = **81.8%**. The same test is far more useful in a high-risk population.

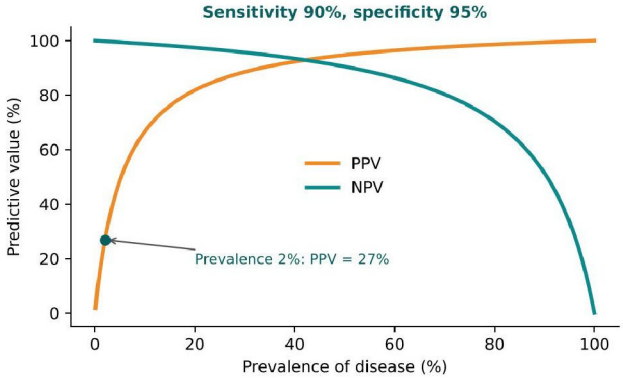

Figure 3: Effect of disease prevalence on PPV and NPV (sensitivity 90%, specificity 95%)

Choosing a cut-off: the ROC curve

For tests with a continuous result (blood glucose, a biomarker), changing the **cut-off** changes sensitivity and specificity in opposite directions. The **receiver operating characteristic (ROC) curve** plots sensitivity (y-axis) against 1 − specificity (x-axis) for every cut-off. A curve close to the top-left corner shows a good test; the diagonal line shows a useless test. The **area under the curve (AUC)** summarises accuracy: 0.5 = chance, 1.0 = perfect, 0.8–0.9 = good.

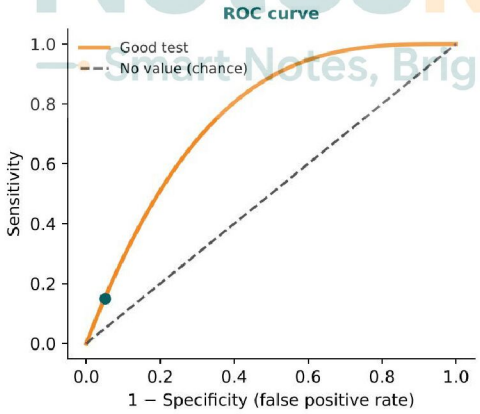

Figure 4: ROC curve of a good diagnostic test compared with chance

Applications in clinical decision-making

- **Screening versus diagnosis:** screening healthy people (low prevalence) produces many false positives, so positive results need a confirmatory test. In patients with symptoms (high prevalence) the same test gives a much higher PPV.
- **Pre-test and post-test probability:** the doctor's initial estimate (prior) is revised after the test result (posterior). **Post-test odds = pre-test odds × likelihood ratio.**
- **Pharmacogenomic testing:** a genetic test before giving a drug (to avoid a serious hypersensitivity reaction) is judged by its predictive values in the population treated.
- **Drug safety signals:** Bayesian methods are used in pharmacovigilance (for example the Bayesian Confidence Propagation Neural Network used by WHO-UMC) to decide whether a drug-reaction pair is reported more often than expected.
- **Choosing the test:** a **highly sensitive** test (few false negatives, high NPV) is best to **rule out** disease; a **highly specific** test (few false positives, high PPV) is best to **rule in** (confirm) disease.
- **Probabilistic thinking:** results should never be read as 'positive = diseased'; they change the probability, and decisions must also weigh the harm of treating or missing the disease.

MEMORY AID

SnNout: a Sensitive test, when Negative, rules out. **SpPin:** a Specific test, when Positive, rules in.

PART B: RANDOM VARIABLES AND THE NORMAL DISTRIBUTION

5. Random Variables

A **random variable (X)** is a variable whose value is a numerical outcome of a chance experiment. It converts outcomes such as 'responder' or 'non-responder' into numbers so that they can be analysed. We use a capital letter (X) for the variable and a small letter (x) for a particular value.

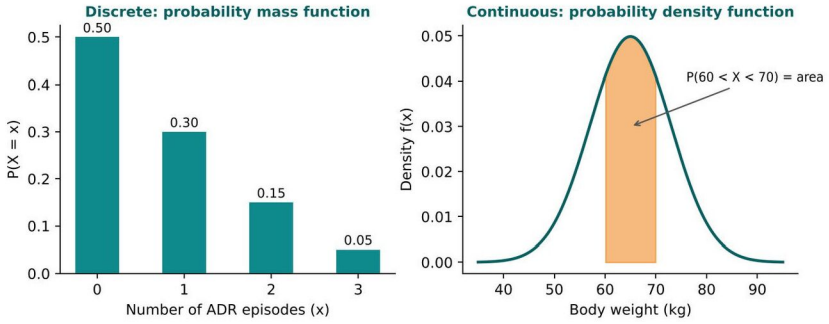

Figure 5: A discrete random variable (probability mass function) and a continuous random variable (probability density function)

Feature	Discrete random variable	Continuous random variable
Values	Countable (whole numbers): 0, 1, 2, 3…	Any value in an interval, including decimals
Examples	Number of ADRs in a patient, number of responders among 10 patients, number of defective tablets in a sample, number of seizures per month	Body weight, blood pressure, plasma concentration, tablet weight, time to relief
Probability function	**Probability mass function (PMF):** P(X = x) gives the probability of each value	**Probability density function (PDF):** f(x); probability is the **area under the curve** between two values
Probability at one exact value	Can be greater than 0	**Exactly 0**; P(X = 60) = 0. Only intervals have probability
Total probability	ΣP(X = x) = 1	Total area under the curve = 1
Cumulative function (CDF)	F(x) = P(X ≤ x): add the probabilities up to x	F(x) = P(X ≤ x): area to the left of x
Distributions covered	Binomial, Poisson	Normal

Probability distribution of a discrete random variable

A **probability distribution** lists every possible value of X with its probability. The rules are: each P(x) lies between 0 and 1, and the probabilities add up to 1.

WORKED EXAMPLE 8: MEAN AND VARIANCE OF A DISCRETE VARIABLE

X = number of ADR episodes per patient in a 6-month study. The probability distribution is:

x: 0, 1, 2, 3 **P(x):** 0.50, 0.30, 0.15, 0.05 (sum = 1.00, so it is a valid distribution).

Expected value (mean): E(X) = μ = Σ x·P(x) = 0(0.50) + 1(0.30) + 2(0.15) + 3(0.05) = 0.75.

$E(X^2) = 0 + 1(0.30) + 4(0.15) + 9(0.05) = 1.35$. Variance $\sigma^2 = E(X^2) - \mu^2 = 1.35 - 0.5625 = 0.7875$; SD $\sigma = 0.887$.

P(at least one episode) = 1 − P(0) = 0.50. Interpretation: on average a patient has 0.75 episodes; half the patients have none.

- **Mean (expected value):** $\mu = E(X) = \Sigma\, x \cdot P(x)$: the long-run average of X over many repetitions.
- **Variance:** $\sigma^2 = \Sigma\,(x - \mu)^2 \cdot P(x) = E(X^2) - \mu^2$; **SD $\sigma = \sqrt{\sigma^2}$.**
- For a continuous variable, the sums are replaced by integrals of the density function; software calculates these.
- **Properties:** $E(aX + b) = a \cdot E(X) + b$; $Var(aX + b) = a^2 \cdot Var(X)$.

6. Normal Distribution

The **normal (Gaussian) distribution** is the most important continuous distribution. Its graph is a **symmetrical bell-shaped curve** described completely by two parameters: the **mean μ** (centre) and the **standard deviation σ** (spread). It is written **$X \sim N(\mu, \sigma^2)$**.

Properties of the normal curve

- Bell-shaped and **symmetrical** about the mean; **Mean = Median = Mode.** Skewness = 0, excess kurtosis = 0.
- The curve is highest at μ, falls on both sides and **never touches the x-axis** (asymptotic).
- The **total area under the curve = 1.** Half the area (0.5) lies on each side of the mean.
- **Empirical rule:** about **68.3%** of values lie within μ ± 1σ, **95.4%** within μ ± 2σ and **99.7%** within μ ± 3σ. Precisely, **95% lie within μ ± 1.96σ**.
- **Points of inflection** (where the curve changes from bending down to bending up) lie at μ ± σ.
- A change in **μ shifts** the curve left or right; a change in **σ** makes it taller and narrower (small σ) or lower and wider (large σ).
- The PDF is f(x) = [1 / (σ√(2π))] · e^{ −(x − μ)² / (2σ²) }; it is not needed for calculations because tables and software are used.

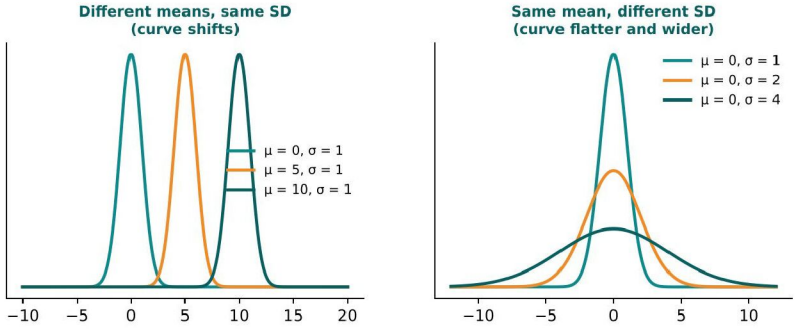

Figure 7: Effect of the mean (shifts the curve) and the SD (changes the width) on the normal curve

The standard normal distribution and z-scores

Normal distributions with different μ and σ are all converted to one common curve, the **standard normal distribution** with **mean 0 and SD 1**, written **$Z \sim N(0, 1)$**, by the **z-score** (standardisation):

STANDARDISATION

$z = (x - \mu) / \sigma$ and the reverse **$x = \mu + z \cdot \sigma$**

z tells how many standard deviations the value lies from the mean. Areas under the standard normal curve are found from **z-tables** or software.

z	0	0.5	1.0	1.5	1.645	1.96	2.0	2.326	2.576	3.0
P(Z ≤ z)	0.5000	0.6915	0.8413	0.9332	0.9500	0.9750	0.9772	0.9900	0.9950	0.9987

- For **negative z** use symmetry: P(Z ≤ −z) = 1 − P(Z ≤ z). Example: P(Z ≤ −1.5) = 1 − 0.9332 = 0.0668.
- **P(Z > z) = 1 − P(Z ≤ z). P(a < Z < b) = P(Z ≤ b) − P(Z ≤ a).**
- **Useful critical values:** z = 1.645 (5% in one tail, 90% central), **z = 1.96 (2.5% in each tail, 95% central)**, z = 2.576 (99% central). These are used for confidence intervals and tests in Unit III.

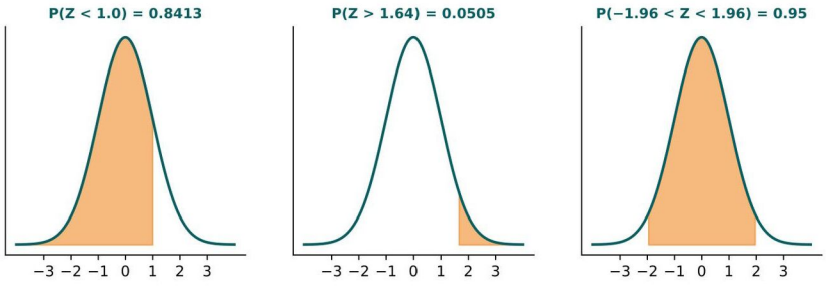

Figure 8: Areas under the standard normal curve (shaded): left tail, right tail and central 95%

How to solve a normal probability problem

Step	Action
1	Identify μ and σ and write the event (for example P(490 < X < 510)).

Step	Action
2	Sketch the bell curve and shade the required area.
3	Convert x to z using $z = (x - \mu) / \sigma$.
4	Read the area from the z-table (or use Python) and use complement or subtraction as needed.
5	State the answer as a probability or percentage and interpret it in the real context.

WORKED EXAMPLE 9: TABLET WEIGHT

Tablet weight in a batch is normal with $\mu = 500$ mg and $\sigma = 5$ mg.

(a) P(490 < X < 510): $z = (490 - 500)/5 = -2$ and $z = (510 - 500)/5 = +2$. Area = 0.9772 − 0.0228 = **0.9545 (95.45%)**.

(b) P(X < 492): $z = -1.6$; P = 1 − 0.9452 = **0.0548 (5.5%)** of tablets weigh less than 492 mg.

(c) 90th percentile: $z = 1.2816 \rightarrow x = 500 + 1.2816 \times 5 =$ **506.4 mg**. Nine out of ten tablets weigh less than 506.4 mg.

Interpretation: a 5.5% share of tablets below 492 mg may be too high for the specification; the process must be tightened (smaller σ).

WORKED EXAMPLE 10: BLOOD PRESSURE AND CHOLESTEROL

Systolic BP of adults is $N(120, 15^2)$ mmHg. **P(110 < X < 140):** $z = -0.667$ and $z = +1.333$; area = 0.9088 − 0.2525 = **0.656**. About 66% of adults have systolic BP between 110 and 140 mmHg.

Cholesterol is $N(190, 20^2)$ mg/dL. **P(X > 240):** $z = 2.5$, P = 1 − 0.9938 = **0.0062** (0.6%).

Haemoglobin in women is $N(13.5, 1.2^2)$ g/dL. **P(X < 12):** $z = -1.25$; **P = 0.106**: about 10.6% of healthy women fall below 12 g/dL, so a fixed 'cut-off' at 12 would label many healthy women as anaemic.

Reference range: the normal range of a laboratory test is taken as **$\mu \pm 1.96\sigma$**, covering the central 95% of healthy people. By definition 5% of healthy people lie outside it.

Importance of the normal distribution in biological and pharmaceutical data

- **Many measurements are approximately normal:** height, weight, blood pressure, haemoglobin, tablet weight, assay content in a well-controlled process, and random errors of measurement.
- **Central limit theorem (CLT):** the distribution of the **sample mean** becomes approximately normal as the sample size grows (usually $n \geq 30$), **whatever the shape of the original data**. This allows confidence intervals and tests for means even for skewed data.
- **Basis of statistical tests:** the t-test, z-test, ANOVA and regression assume (approximately) normal data or normal errors (Units III and IV).
- **Reference ranges and cut-offs** in clinical laboratories are built on $\mu \pm 1.96\sigma$.
- **Quality control and process capability:** the percentage of tablets within limits, control chart limits ($\mu \pm 3\sigma$) and acceptance criteria use normal probabilities.
- **Approximation:** binomial and Poisson distributions with large n or large λ can be approximated by the normal distribution (continuity correction ±0.5).
- **Dose-response and bioassay:** log-dose or log-concentration data are often normal, so that log-normal data (Cmax, AUC) are made normal by log transformation (Unit I).

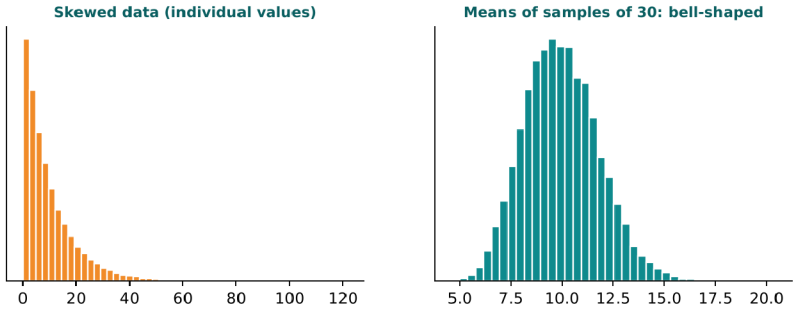

Figure 9: Central limit theorem: skewed individual values give bell-shaped sample means (simulation of 10,000 samples of size 30)

Log-normal distribution for pharmacokinetic data

A variable is **log-normal** if its natural logarithm is normal. It is always positive and **right-skewed**, and it describes many biological quantities: Cmax, AUC, drug clearance, viral load and antibody titres.

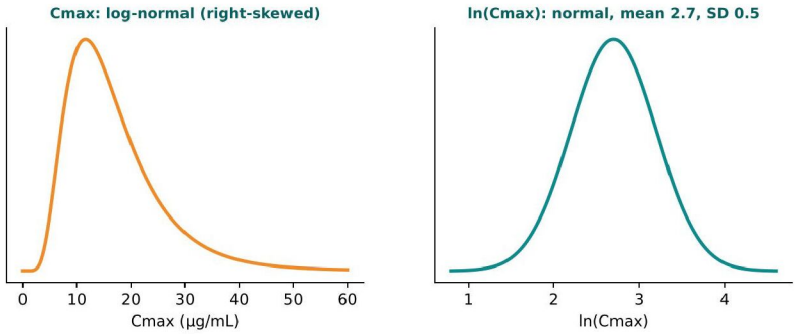

Figure 10: A log-normal variable (Cmax) and its natural logarithm (normal)

WORKED EXAMPLE 12: LOG-NORMAL CMAX

ln(Cmax) is normal with mean 2.7 and SD 0.5. **Geometric mean of Cmax = $e^{2.7}$ = 14.88 µg/mL** (this is also the median).

P(Cmax > 25): ln 25 = 3.219; z = (3.219 − 2.7)/0.5 = 1.04; **P = 0.150.** About 15% of volunteers exceed 25 µg/mL.

Central 95% of volunteers: exp(2.7 ± 1.96 × 0.5) = **5.6 to 39.6 µg/mL**, a very wide, asymmetric range.

In **bioequivalence studies**, ln Cmax and ln AUC are analysed, and the test and reference products are accepted as equivalent when the **90% confidence interval of the ratio of geometric means lies within 80–125%**.

Checking normality

- **Histogram and box plot:** symmetrical and bell-shaped; **mean ≈ median**; skewness near 0.
- **Q-Q plot:** points fall close to a straight line.
- **Formal tests:** Shapiro–Wilk or Kolmogorov–Smirnov (a p-value above 0.05 gives no evidence against normality).
- If data are **not** normal: transform (log), use the median, or use non-parametric tests.

PART C: BINOMIAL AND POISSON DISTRIBUTIONS

7. Binomial Distribution

The **binomial distribution** gives the probability of obtaining a given number of **successes** in a fixed number of independent trials, each with only two possible outcomes. It is the main distribution for **clinical trial outcomes** such as responder or non-responder, cured or not cured, alive or dead, ADR present or absent.

Conditions (Bernoulli trials)

Condition	Meaning	Clinical example
B – Binary	Each trial has only two outcomes: success or failure	Patient responds or does not respond
I – Independent	The result of one trial does not affect another	One patient's response does not alter another's
N – Number fixed	The number of trials n is fixed in advance	Exactly 10 patients are treated
S – Same probability	The probability of success p is the same in every trial	Each patient has a 30% chance of response

Here '**success**' only means the outcome we are counting (it may even be an adverse event). Probability of failure q = 1 − p. The variable X = number of successes in n trials follows a binomial distribution, written **X ~ B(n, p)**.

BINOMIAL PROBABILITY FORMULA

$P(X = r) = nCr \cdot p^r \cdot q^{n-r}$, r = 0, 1, 2, …, n, where nCr = n! / [r!(n − r)!]

Mean: µ = n·p **Variance: $\sigma^2 = n \cdot p \cdot q$** **SD: $\sigma = \sqrt{(n \cdot p \cdot q)}$**

Cumulative probability: P(X ≤ r) = P(0) + P(1) + … + P(r); **P(X ≥ r) = 1 − P(X ≤ r − 1).**

WORKED EXAMPLE 15: RESPONSE TO A DRUG

A drug produces a response in 30% of patients (p = 0.3). Ten patients (n = 10) are treated.

(a) Exactly 4 respond: $P(X = 4) = 10C4 \times (0.3)^4 \times (0.7)^6 = 210 \times 0.0081 \times 0.117649 =$ **0.2001**.

(b) At least 3 respond: $P(X \geq 3) = 1 - [P(0) + P(1) + P(2)] = 1 - (0.0282 + 0.1211 + 0.2335) = 1 - 0.3828 =$ **0.617**.

(c) Mean and SD: $\mu = 10 \times 0.3 =$ **3** responders; $\sigma = \sqrt{(10 \times 0.3 \times 0.7)} = \sqrt{2.1} =$ **1.45**.

Interpretation: on average 3 of 10 patients respond, but anything from 1 to 5 is common. The chance of seeing 4 responders is 20%.

r	0	1	2	3	4	5	6	7	8
P(X = r) for n = 10, p = 0.3	0.0282	0.1211	0.2335	0.2668	0.2001	0.1029	0.0368	0.0090	0.0014

Shape of the binomial distribution

- **p = 0.5:** symmetrical. **p < 0.5:** right-skewed (tail to the right). **p > 0.5:** left-skewed.
- The distribution becomes **more symmetrical and bell-shaped as n increases**, so for large n it can be approximated by a normal distribution with $\mu = np$ and $\sigma = \sqrt{(npq)}$ (use when **np ≥ 5 and nq ≥ 5**, with continuity correction).
- The mean np is the expected number of successes; the **variance npq is always less than the mean**.

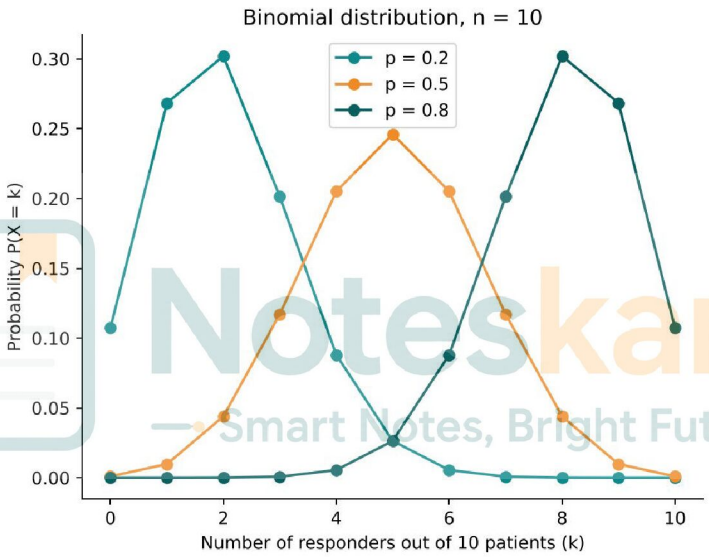

Figure 11: Binomial distributions for n = 10 with p = 0.2, 0.5 and 0.8 (output of the Python code in Part D)

Cumulative binomial probabilities

k	0	1	2	3	4	5	6	7
P(X ≤ k), n = 10, p = 0.3	0.0282	0.1493	0.3828	0.6496	0.8497	0.9527	0.9894	0.9984

- From the table: $P(X \leq 2) = 0.3828$, so **$P(X \geq 3) = 1 - 0.3828 = 0.617$**; **$P(2 \leq X \leq 5) = 0.9527 - 0.1493 = 0.8034$**.
- **Mode** of a binomial distribution = the integer part of (n + 1)p. Here (11 × 0.3) = 3.3, so the most likely number of responders is **3** (P = 0.2668).

Applications of the binomial distribution in clinical trials

WORKED EXAMPLE 19: MORE CLINICAL EXAMPLES

Adverse events: the probability of an ADR in a patient is 0.10. In 12 patients (n = 12, p = 0.1): $P(\text{no ADR}) = 0.9^{12} =$ **0.282**; P(3 or more ADR) = **0.111**; expected number of ADRs = 12 × 0.1 = **1.2**.

Cure rate: a drug cures 80% of patients. Among 20 patients, P(exactly 18 cured) = $20C18\ (0.8)^{18}\ (0.2)^2 =$ **0.137**; P(at least 15 cured) = **0.804**.

At least one event: with p = 0.05 per patient and 20 patients, $P(\text{at least one ADR}) = 1 - 0.95^{20} =$ **0.642**: even a 5% risk will almost certainly be seen at least once in 20 patients.

Sign test (is a drug better than chance?): among 10 patients, 8 prefer the new formulation. If there were no real preference (p = 0.5), P(8 or more) = **0.0547**. This is slightly above 0.05, so 8 of 10 is **not quite enough** to claim a real preference at the 5% level.

- **Response rate** in phase II trials, **success or failure** of treatment, **survival at a fixed time**, **proportion of defective** products in a lot sampled for inspection (acceptance sampling), **vaccine efficacy**, and **screening test results** are all binomial.
- Statistical tests for **proportions** (z-test for a proportion, chi-square, exact binomial test) are built on this distribution (Unit III).
- **Sample size planning:** choosing how many patients are needed to observe a target number of responders uses binomial probabilities.

CAUTION

Binomial calculations need **independent** patients and a **constant p**. If p differs between patient groups (for example by age) or patients influence each other (infectious disease), the binomial model is not suitable.

8. Poisson Distribution

The **Poisson distribution** (Siméon Poisson, 1837) models the **number of times a rare event occurs in a fixed interval of time, area or volume**, when events occur independently and at a constant average rate. It is the natural model for **rare adverse drug reactions** and counts such as colonies on a plate.

Conditions

- Events occur **independently** of each other.
- The **average rate (λ)** of occurrence is **constant** over the interval.
- Two events cannot occur at exactly the same instant; events are **rare** compared with the number of opportunities.
- There is **no fixed upper limit** to the number of events (unlike the binomial, where X cannot exceed n).

POISSON PROBABILITY FORMULA

$P(X = r) = e^{-\lambda} \cdot \lambda^r / r!$, r = 0, 1, 2, 3, … where e = 2.7183 and λ = average number of events in the interval.

Mean = λ **Variance = λ** **SD = √λ** (mean and variance are **equal**: a key feature).

X ~ Poisson(λ). If the rate is known per unit, **λ = rate × size of the interval** (for example 2 per 1000 patients × 500 patients = 1).

WORKED EXAMPLE 20: RARE ADVERSE REACTION

A serious ADR occurs in 2 per 1000 patients treated with a drug. In 500 patients the expected number is **λ = 500 × 2/1000 = 1**.

$P(0 \text{ cases}) = e^{-1} = 0.3679$. **$P(1 \text{ case}) = e^{-1} \times 1 = 0.3679$.** **$P(2 \text{ cases}) = e^{-1}/2 = 0.1839$.** P(3) = 0.0613; P(4) = 0.0153.

P(2 or more cases) = 1 − P(0) − P(1) = 1 − 0.3679 − 0.3679 = 0.2642.

Interpretation: even if the drug causes the reaction, there is a 37% chance that **no** case is seen in 500 patients, so a premarketing trial of a few hundred patients can easily miss rare reactions. This is why post-marketing surveillance (pharmacovigilance) is essential.

WORKED EXAMPLE 21: COUNTS ON A PLATE

A water sample gives on average 2 colonies per plate (λ = 2). **$P(\text{no colony}) = e^{-2} = 0.1353$.** **$P(3 \text{ or more colonies}) = 1 - [P(0) + P(1) + P(2)] = 1 - 0.6767 = 0.3233$.**

Rule of three: if **no** adverse event is seen in n patients, the upper 95% confidence limit of the true rate is about **3 / n**. If 0 events occur in 300 patients, the true rate could still be as high as 3/300 = **1%**.

r	0	1	2	3	4	5
P(X = r), λ = 1	0.3679	0.3679	0.1839	0.0613	0.0153	0.0031
P(X = r), λ = 2	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361

Cumulative Poisson probabilities and use of the rate

- **$P(X \le k) = P(0) + P(1) + \ldots + P(k)$; $P(X \ge k) = 1 - P(X \le k - 1)$.**
- The rate scales with the interval: if a reaction occurs 0.5 times a month on average, the mean over 6 months is **λ = 3**; then P(no case in 6 months) = e^{-3} = 0.0498.
- **Mean and variance equal λ:** for λ = 4, the SD is 2. About 95% of the values lie within $\lambda \pm 2\sqrt{\lambda}$ for larger λ.

WORKED EXAMPLE 22: A PHARMACOVIGILANCE SIGNAL

A drug is expected to produce 2 reports of a particular reaction per month (λ = 2). This month **7 reports** are received.

P(X ≥ 7) = 1 − P(X ≤ 6) = 0.0045. Such a high count would happen by chance in only 0.45% of months.

Interpretation: the excess is unlikely to be chance alone, so it is a **possible safety signal** that must be investigated (check reporting practice, publicity, batch problems and a true increase). A signal is not proof of cause.

Shape and properties

- For **small λ** the distribution is strongly **right-skewed** with its highest bar at 0 or 1; as **λ increases** it becomes more symmetrical, and for **λ above about 10** it is close to normal N(λ, λ).
- The sum of independent Poisson variables is Poisson with the sum of the λ values (rates add up).

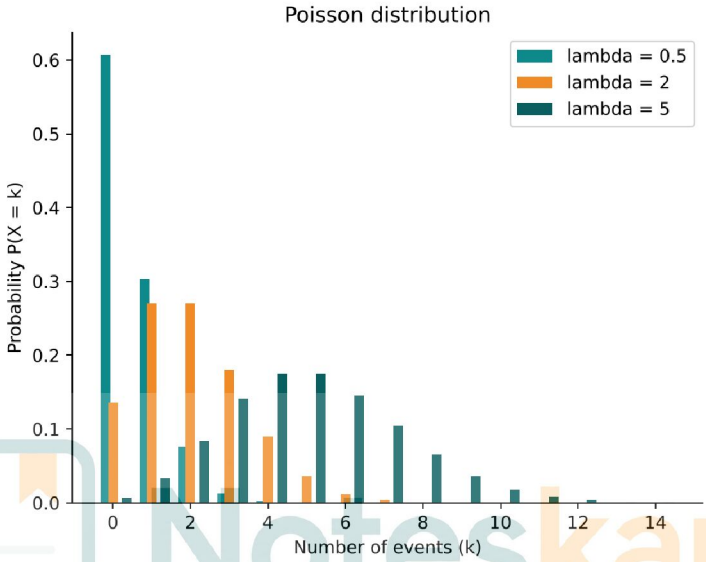

Figure 12: Poisson distributions for λ = 0.5, 2 and 5 (output of the Python code in Part D)

Poisson approximation to the binomial

- When **n is large (n ≥ 50) and p is very small (np < 5 or so)**, the binomial B(n, p) is well approximated by Poisson with **λ = n·p**. This avoids long binomial calculations.
- **Example:** n = 1000, p = 0.002. Binomial P(X = 2) = 0.2709; Poisson (λ = 2) gives **0.2707**: almost the same (see the figure below).

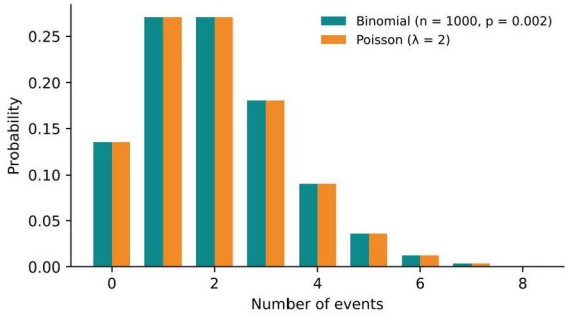

Figure 13: Poisson distribution approximates the binomial when n is large and p is small

Applications in pharmacy and medicine

- **Rare ADRs:** number of cases of a reaction per 1,000 or 10,000 patients, or per month in a pharmacovigilance database (are there more reports than expected?).
- **Microbiology and sterility testing:** colonies per plate, particles per mL in an injectable (particulate matter limits), contamination counts in cleanrooms.
- **Hospital and community pharmacy:** number of prescription errors per day, number of patients arriving per hour, drug-induced emergency admissions per month.
- **Epidemiology:** new cases of a rare disease per year; **radioactivity counts** in radiopharmaceuticals (decays per minute).
- **Overdispersion:** if the observed variance is much larger than the mean, a simple Poisson model is not adequate and other models are needed.

9. Comparison of the Three Distributions and How to Choose

Feature	Binomial	Poisson	Normal
Type of variable	Discrete	Discrete	Continuous
Parameters	n and p	λ	μ and σ
Possible values	0, 1, …, n (upper limit)	0, 1, 2, … (no upper limit)	−∞ to +∞
Mean	np	λ	μ
Variance	npq (less than mean)	λ (equal to mean)	σ^2
Shape	Skewed unless p = 0.5; symmetrical for large n	Right-skewed; symmetrical for large λ	Symmetrical bell
Typical use in pharmacy	Responders in a fixed number of patients; proportion defective in a sample	Rare ADRs; counts per unit of time or space	Tablet weight, BP, Hb, assay, test errors
Example	8 of 10 patients cured	3 ADR reports this month	Weight 502 mg

Which distribution should I use?

Question	Use
Counting successes in a fixed number of independent trials with two outcomes?	**Binomial**
Counting rare events in a time, area or volume with no fixed number of trials?	**Poisson**
Measured (continuous) quantity with a symmetrical bell-shaped histogram?	**Normal**
Binomial with large n and p near 0.5 (np ≥ 5 and nq ≥ 5)?	**Normal approximation** with continuity correction
Binomial with large n and very small p?	**Poisson approximation** (λ = np)
Positive, right-skewed measurement (Cmax, AUC)?	**Log-normal**: take logarithms, then treat as normal

WORKED EXAMPLE 23: NORMAL APPROXIMATION TO THE BINOMIAL

A treatment has a response rate of 30%. Among 100 patients (n = 100, p = 0.3), what is P(35 or fewer respond)?

np = 30 and nq = 70 (both ≥ 5), so use the normal approximation: μ = 30, σ = √(100 × 0.3 × 0.7) = √21 = 4.58.

With **continuity correction** use 35.5: z = (35.5 − 30) / 4.58 = 1.20; **P ≈ 0.885**. The exact binomial value is **0.884**, so the approximation is excellent.

PART D: PROBABILITY DISTRIBUTIONS USING PYTHON

Probability tables are slow and limited. In Python, the **SciPy** library (module **scipy.stats**) contains every common distribution, and **NumPy** can generate random data for simulations. **Matplotlib** draws the graphs. As in Unit I, the aim is to **understand and interpret** the output.

10. Calculating Probabilities with SciPy

The scipy.stats toolkit

Distribution	Create the object	Main parameters
Normal	stats.norm(loc=μ, scale=σ)	loc = mean, **scale = SD** (not variance)
Binomial	stats.binom(n=n, p=p)	n = trials, p = probability of success
Poisson	stats.poisson(mu=λ)	mu = average number of events (λ)

Method	Gives	Used for
.pmf(k)	P(X = k) for discrete distributions	Binomial, Poisson: probability of exactly k
.pdf(x)	Height of the density curve at x	Normal: drawing the curve (not a probability)

Method	Gives	Used for
.cdf(x)	P(X ≤ x), cumulative probability	P(X ≤ k); P(a < X < b) = cdf(b) − cdf(a)
.sf(x)	P(X > x) = 1 − cdf(x) (survival function)	Upper-tail probabilities
.ppf(q)	The x with P(X ≤ x) = q (inverse of cdf, percent-point function)	Percentiles, critical values such as norm.ppf(0.975) = 1.96
.rvs(size=n)	n random values from the distribution	Simulation
.mean(), .std(), .var()	Theoretical mean, SD and variance	Check μ = np, λ

Example: normal, binomial and Poisson probabilities

The program below solves Worked examples 9, 15 and 20 in a few lines.

PYTHON CODE

```
from scipy import stats

# ---- Normal: tablet weight ~ Normal(mean = 500 mg, SD = 5 mg)
w = stats.norm(loc=500, scale=5)
print("P(490 < X < 510)      :", round(w.cdf(510) - w.cdf(490), 4))
print("P(X < 492)            :", round(w.cdf(492), 4))
print("90th percentile (mg)  :", round(w.ppf(0.90), 2))

# ---- Binomial: 10 patients, probability of response p = 0.3
b = stats.binom(n=10, p=0.3)
print("P(X = 4)              :", round(b.pmf(4), 4))
print("P(X >= 3)             :", round(1 - b.cdf(2), 4))
print("Mean, SD              :", b.mean(), round(b.std(), 3))

# ---- Poisson: average of 1 adverse reaction per 500 patients
p = stats.poisson(mu=1)
print("P(X = 0)              :", round(p.pmf(0), 4))
print("P(X >= 2)             :", round(1 - p.cdf(1), 4))
```

OUTPUT

```
P(490 < X < 510)      : 0.9545
P(X < 492)            : 0.0548
90th percentile (mg)  : 506.41
P(X = 4)              : 0.2001
P(X >= 3)             : 0.6172
Mean, SD              : 3.0 1.449
P(X = 0)              : 0.3679
P(X >= 2)             : 0.2642
```

Output	Interpretation
P(490 < X < 510) = 0.9545	95.45% of tablets lie between 490 and 510 mg, matching the ±2 SD rule.
P(X < 492) = 0.0548	About 5.5% of tablets are below 492 mg.
90th percentile = 506.41 mg	90% of tablets weigh less than 506.4 mg.
Binomial P(X = 4) = 0.2001; P(X ≥ 3) = 0.6172	There is a 20% chance that exactly 4 of 10 patients respond and a 62% chance that at least 3 respond.
Mean 3.0, SD 1.449	On average 3 responders, typically within about 1.4 of that.
Poisson P(X = 0) = 0.3679; P(X ≥ 2) = 0.2642	Even when the reaction is real, 37% of 500-patient trials would show no case.

IMPORTANT POINTS

P(X ≥ 3) for a discrete variable = 1 − cdf(2) (not cdf(3)); P(X > 3) = 1 − cdf(3) or b.sf(3).

In **norm()**, **scale** is the **standard deviation**, not the variance. Writing scale=25 instead of scale=5 is a very common error.

pdf gives the height of a curve, not a probability. Probabilities for a continuous variable need an **interval** (cdf differences).

11. Bayes' Theorem, Simulation and Graphs in Python

Bayes' theorem and predictive values

The function below calculates PPV and NPV for any prevalence and shows how prevalence changes the usefulness of a test (compare Worked example 5 and the PPV graph).

PYTHON CODE

```
def predictive_values(prevalence, sensitivity, specificity):
    tp = prevalence * sensitivity                 # true positives
    fn = prevalence * (1 - sensitivity)           # false negatives
    fp = (1 - prevalence) * (1 - specificity)     # false positives
    tn = (1 - prevalence) * specificity           # true negatives
    ppv = tp / (tp + fp)                          # Bayes' theorem
    npv = tn / (tn + fn)
    return ppv, npv

print("Prevalence   PPV      NPV")
for prev in [0.01, 0.02, 0.05, 0.10, 0.20, 0.50]:
    ppv, npv = predictive_values(prev, sensitivity=0.90, specificity=0.95)
    print(f"{prev:>8.0%}   {ppv:6.1%}   {npv:6.1%}")

print("LR+ =", round(0.90 / (1 - 0.95), 1), "  LR- =", round((1 - 0.90) / 0.95, 3))
```

OUTPUT

```
Prevalence   PPV      NPV
      1%    15.4%    99.9%
      2%    26.9%    99.8%
      5%    48.6%    99.4%
     10%    66.7%    98.8%
     20%    81.8%    97.4%
     50%    94.7%    90.5%
LR+ = 18.0   LR- = 0.105
```

- At 1% prevalence only **15.4%** of positives are true; at 50% prevalence **94.7%**. **PPV rises with prevalence** while NPV falls.
- Even with a good test (sensitivity 90%, specificity 95%), a positive result in a low-prevalence screening programme must be confirmed.

Simulation: law of large numbers and central limit theorem

Simulation means generating many random results with the computer to see what probability theory predicts. The seed in default_rng(42) makes the random numbers repeatable.

PYTHON CODE

```
import numpy as np

rng = np.random.default_rng(42)          # fixed seed so the result can be repeated

# Law of large numbers: probability of response is really 0.30
for n in [10, 100, 1000, 100000]:
    responders = rng.binomial(1, 0.30, size=n)
    print(f"n = {n:>6}: observed proportion = {responders.mean():.4f}")

# Central limit theorem: skewed data (exponential), means of samples of size 30
skewed = rng.exponential(scale=10, size=100000)
means = rng.exponential(scale=10, size=(10000, 30)).mean(axis=1)
```

```
print("Skewed data : mean =", round(skewed.mean(), 2), " SD =", round(skewed.std(), 2))
print("Sample means: mean =", round(means.mean(), 2), " SD =", round(means.std(), 2),
      " (theory: 10/sqrt(30) =", round(10 / 30 ** 0.5, 2), ")")
```

OUTPUT

```
n =     10: observed proportion = 0.5000
n =    100: observed proportion = 0.2700
n =   1000: observed proportion = 0.3050
n = 100000: observed proportion = 0.3008
Skewed data : mean = 9.95  SD = 9.98
Sample means: mean = 9.97  SD = 1.8  (theory: 10/sqrt(30) = 1.83 )
```

- **Law of large numbers:** with 10 patients the observed response rate is 0.50, far from the true 0.30; with 100,000 patients it is **0.3008**. Small trials are unreliable; observed proportions approach the true probability as n grows.
- **Central limit theorem:** the individual values are strongly skewed (mean 9.95, SD 9.98), yet the means of samples of 30 have SD **1.80**, very close to the theoretical σ/√n = 1.83, and their histogram is bell-shaped (see the CLT figure in Part B).

Graphical visualisation of the distributions

Graphs of the distributions are drawn with **Matplotlib**. For discrete distributions we plot the **pmf** (bars or points) and for continuous distributions the **pdf** (a smooth curve). Colours and labels can be changed freely.

1. Normal distribution (continuous curve with shaded probability)

PYTHON CODE

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

x = np.linspace(470, 530, 400)
mean, sd = 500, 5
y = stats.norm.pdf(x, mean, sd)

plt.plot(x, y, color="teal", linewidth=2)
inside = (x >= 490) & (x <= 510)
plt.fill_between(x[inside], y[inside], color="orange", alpha=0.5, label="490 to 510 mg
(95.45%)")
plt.xlabel("Tablet weight (mg)")
plt.ylabel("Probability density")
plt.title("Normal distribution: mean 500 mg, SD 5 mg")
plt.legend()
plt.show()
```

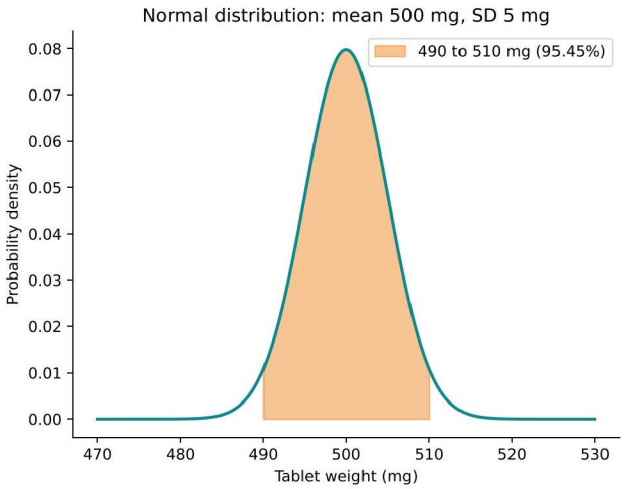

Figure 14: Output of the normal distribution code: the shaded area is P(490 < X < 510)

- np.linspace creates 400 evenly spaced x-values; stats.norm.pdf gives the curve height; **fill_between** shades the area that represents the probability.
- **Reading the graph:** the orange area is 95.45% of the total area under the curve, i.e. the probability that a tablet lies within 10 mg of the mean.

2. Binomial distribution (discrete points)

PYTHON CODE

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

k = np.arange(0, 11)                        # possible number of responders, 0 to 10
for p, colour in [(0.2, "teal"), (0.5, "orange"), (0.8, "purple")]:
    plt.plot(k, stats.binom.pmf(k, n=10, p=p), "o-", color=colour, label=f"p = {p}")

plt.xlabel("Number of responders out of 10 patients (k)")
plt.ylabel("Probability P(X = k)")
plt.title("Binomial distribution, n = 10")
plt.legend()
plt.show()
```

- The output is the binomial figure in Part C. **p = 0.2** is right-skewed with its peak at 2, **p = 0.5** is symmetrical with its peak at 5, and **p = 0.8** is left-skewed with its peak at 8. The sum of every curve's probabilities is 1.
- Plot a **cumulative** curve with stats.binom.cdf(k, 10, 0.3) to read P(X ≤ k) directly; use plt.step or plt.bar.

3. Poisson distribution (discrete bars)

PYTHON CODE

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

k = np.arange(0, 15)
for lam, colour in [(0.5, "teal"), (2, "orange"), (5, "purple")]:
    plt.bar(k + (lam - 2) * 0.12, stats.poisson.pmf(k, lam), width=0.25,
            color=colour, label=f"lambda = {lam}")

plt.xlabel("Number of events (k)")
plt.ylabel("Probability P(X = k)")
plt.title("Poisson distribution")
plt.legend()
plt.show()
```

- The output is the Poisson figure in Part C. For **λ = 0.5** most of the probability sits at 0; for **λ = 2** the peak moves to 1 and 2; for **λ = 5** the shape becomes broader and nearly symmetrical.
- The small shift added to k (k + (lam − 2) × 0.12) only places the three bars side by side; it does not change the probabilities.

5. Checking normality: histogram with fitted curve, Q-Q plot and Shapiro-Wilk test

PYTHON CODE

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

rng = np.random.default_rng(7)
weights = rng.normal(500, 5, 200)               # 200 simulated tablet weights

print("Shapiro-Wilk p-value:", round(stats.shapiro(weights).pvalue, 3))

fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(9, 3.2))
ax1.hist(weights, bins=15, density=True, color="teal", edgecolor="white")
x = np.linspace(480, 520, 200)
ax1.plot(x, stats.norm.pdf(x, weights.mean(), weights.std(ddof=1)), color="orange",
linewidth=2)
```

```
ax1.set_xlabel("Tablet weight (mg)")
ax1.set_title("Histogram with fitted normal curve")
stats.probplot(weights, dist="norm", plot=ax2)  # Q-Q plot
ax2.set_title("Normal Q-Q plot")
plt.show()
```

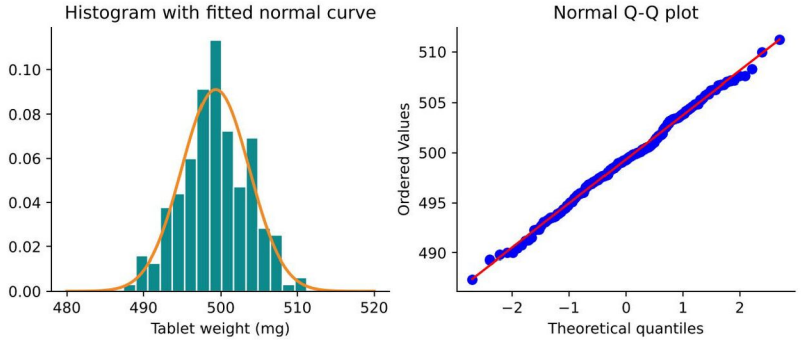

Figure 16: Output of the normality-checking code for 200 simulated tablet weights

- **Shapiro-Wilk p-value = 0.883**, much larger than 0.05, so there is **no evidence against normality**. (A small p-value below 0.05 would suggest the data are not normal.)
- **Histogram:** follows the fitted bell curve. **Q-Q plot:** points lie on the straight line, which is what is expected for normal data. Skewed data would bend away from the line at the ends.
- Tests of normality are **sensitive to sample size**: with thousands of values trivial departures become 'significant', while with very few values a test has little power. Always inspect the graphs too.

Quick reference: distribution calls in Python

Question	Code
P(X = 4), binomial n = 10, p = 0.3	stats.binom.pmf(4, 10, 0.3)
P(X ≤ 4), binomial	stats.binom.cdf(4, 10, 0.3)
P(X ≥ 5), binomial	1 - stats.binom.cdf(4, 10, 0.3) or stats.binom.sf(4, 10, 0.3)
P(X = 3), Poisson λ = 2	stats.poisson.pmf(3, 2)
P(X ≥ 3), Poisson λ = 2	stats.poisson.sf(2, 2)
P(X < 492), normal mean 500, SD 5	stats.norm.cdf(492, 500, 5)
P(490 < X < 510)	stats.norm.cdf(510, 500, 5) - stats.norm.cdf(490, 500, 5)
95th percentile, normal	stats.norm.ppf(0.95, 500, 5)
z-score critical value (two-sided 95%)	stats.norm.ppf(0.975) gives 1.96
1000 random values from a distribution	stats.norm.rvs(500, 5, size=1000) or rng.normal(500, 5, 1000)

PART E: SOLVED NUMERICAL PROBLEMS AND FORMULAE

12. Solved Numerical Problems

PROBLEM 1: CONDITIONAL PROBABILITY FROM A TABLE

In a hospital 500 patients were studied: 200 smokers (of whom 60 had an ulcer) and 300 non-smokers (of whom 30 had an ulcer).

P(ulcer) = 90/500 = 0.18. P(ulcer | smoker) = 60/200 = 0.30. P(ulcer | non-smoker) = 30/300 = 0.10. P(smoker | ulcer) = 60/90 = 0.667.

Ulcer and smoking are **dependent** (0.30 ≠ 0.18). The risk is 3 times higher in smokers (0.30/0.10).

PROBLEM 2: ADDITION AND MULTIPLICATION RULES

Two independent tests are done on a diseased patient; each has a sensitivity of 90%. **P(both positive) = 0.9 × 0.9 = 0.81. P(at least one positive) = 1 − (0.1 × 0.1) = 0.99. P(exactly one positive) = 0.9 × 0.1 × 2 = 0.18.**

Check: 0.81 + 0.18 + 0.01 = 1.00.

PROBLEM 3: BAYES' THEOREM IN A SCREENING PROGRAMME

A test has sensitivity 95% and specificity 98%. Prevalence of disease is 0.5%. Find PPV.

PPV = (0.95 × 0.005) / (0.95 × 0.005 + 0.02 × 0.995) = 0.00475 / (0.00475 + 0.0199) = 0.1927 ≈ 19%.

Only 1 in 5 positive results is a true case, because the disease is rare. A second, different test is advised for every positive screen.

PROBLEM 4: NORMAL DISTRIBUTION (BLOOD PRESSURE)

Systolic BP is N(120, 15^2). Find the proportion of adults with SBP above 140 mmHg.

z = (140 − 120)/15 = 1.33; **P(Z > 1.33) ≈ 1 − 0.9082 = 0.092 (about 9%).** (Software, using z = 1.3333, gives 0.0912.)

The SBP **below which 95% of adults** lie: x = 120 + 1.645 × 15 = **144.7 mmHg**.

PROBLEM 5: LIMITS THAT CONTAIN THE CENTRAL 95%

Tablet weight is N(500, 5^2). The central 95% of tablets lie within 500 ± 1.96 × 5 = **490.2 mg to 509.8 mg**.

A pharmacopoeial limit of 475–525 mg is z = ±5 from the mean, which is practically never exceeded (probability about 0.0000006).

PROBLEM 6: BINOMIAL

A vaccine causes mild fever in 10% of recipients. Five people are vaccinated (n = 5, p = 0.1).

P(exactly 2 have fever) = $5C2\ (0.1)^2\ (0.9)^3$ = 10 × 0.01 × 0.729 = **0.0729. P(none)** = 0.9^5 = **0.5905. P(at most 1)** = 0.5905 + 5 × 0.1 × 0.9^4 = 0.5905 + 0.3281 = **0.9185. P(at least one)** = 1 − 0.5905 = **0.4095.**

Mean = 0.5 persons with fever; SD = √(5 × 0.1 × 0.9) = 0.67.

Noteskarts — Smart Notes, Bright Future

PROBLEM 7: POISSON

A community pharmacy makes on average 3 dispensing errors in a day (λ = 3), counted as a Poisson variable.

P(no error) = e^{-3} = 0.0498. P(at least one error) = 1 − 0.0498 = 0.9502. P(4 or fewer errors) = 0.8153.

On 95% of days there is at least one error, so the pharmacy should look at the system rather than individual events.

PROBLEM 8: CHOOSING THE DISTRIBUTION

(a) Number of the next 25 patients who develop nausea when p = 0.15 → **binomial** (n = 25, p = 0.15).

(b) Number of reports of anaphylaxis to a regional centre per week, average 2 → **Poisson** (λ = 2).

(c) Plasma glucose of healthy adults, symmetrical histogram → **normal**.

(d) Number of defective tablets in a sample of 1000 with p = 0.001 → **Poisson approximation** (λ = 1).

PROBLEM 9: SUM OF NORMAL VARIABLES

Fill volume of a vial is N(10.0, 0.1^2) mL. Four vials are pooled. The pooled volume has mean 40.0 mL and SD = 0.1 × √4 = **0.2 mL**.

P(pooled volume < 39.6 mL) = P(Z < −2) = **0.0228**.

PROBLEM 10: POISSON COUNT WITH A CHANGED INTERVAL

A hospital pharmacy records 0.8 medication errors per 100 prescriptions. For 250 prescriptions in a day, **λ = 0.8 × 2.5 = 2.0**.

P(5 or more errors) = 1 − P(X ≤ 4) = 1 − 0.9473 = 0.0527. An alarm for 5 or more errors in a day would be raised by chance on about 5% of days.

PROBLEM 11: BAYES' THEOREM IN PHARMACOGENOMICS

6% of a population carries a gene variant. A genetic test has sensitivity 98% and specificity 99%. A patient tests positive.

PPV = (0.98 × 0.06) / (0.98 × 0.06 + 0.01 × 0.94) = 0.0588 / (0.0588 + 0.0094) = 0.862.

About 86% of positives truly carry the variant; the drug is withheld or the dose modified for those patients, and borderline results are confirmed by repeat testing.

PROBLEM 12: READING PROBABILITIES IN WORDS

A student states: 'P(ADR) = 0.02 means that out of every 100 patients exactly 2 will have an ADR.' This is **wrong**. It means that **on average 2 of every 100** patients are expected to have an ADR; in a given group of 100 the number may be 0, 1, 2, 3, 4 or more. Using the Poisson/binomial distribution, with n = 100 and p = 0.02, P(exactly 2) is only **0.273**.

PROBLEM 13: GENERAL ADDITION RULE WITH THREE OUTCOMES

Among 300 patients on a drug, 45 had nausea, 30 had dizziness and 12 had both. **P(nausea or dizziness) = 45/300 + 30/300 − 12/300 = 63/300 = 0.21.**

P(neither) = 1 − 0.21 = 0.79. P(nausea but not dizziness) = (45 − 12)/300 = 0.11. P(exactly one of the two) = 0.11 + (30 − 12)/300 = 0.11 + 0.06 = 0.17.

PROBLEM 14: COMPARING STANDARDISED VALUES

A patient's systolic BP is 150 mmHg (population $N(120, 15^2)$) and the serum cholesterol is 235 mg/dL (population $N(190, 20^2)$). Which value is more unusual?

z(BP) = (150 − 120)/15 = 2.0; z(cholesterol) = (235 − 190)/20 = 2.25.

Cholesterol is further from its mean in SD units, so it is the **more unusual** value (about 1.2% of people are higher, against 2.3% for BP).

13. Formulae Used in This Unit

Topic	Formula
Classical probability	$P(A)$ = favourable outcomes / total outcomes; $P(A') = 1 - P(A)$
Addition rule	$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$; mutually exclusive: $P(A) + P(B)$
Multiplication rule	$P(A \text{ and } B) = P(A) \times P(B \mid A)$; independent: $P(A) \times P(B)$
Conditional probability	$P(A \mid B) = P(A \text{ and } B) / P(B)$
Bayes' theorem	$P(A \mid B) = P(B \mid A)\, P(A) / P(B)$; $P(B) = \Sigma\, P(B \mid A_i)\, P(A_i)$
Sensitivity, specificity	$Sn = TP / (TP + FN)$; $Sp = TN / (TN + FP)$
Predictive values	$PPV = TP / (TP + FP)$; $NPV = TN / (TN + FN)$; $LR+ = Sn / (1 - Sp)$; $LR- = (1 - Sn) / Sp$
Discrete random variable	$\mu = \Sigma\, x \cdot P(x)$; $\sigma^2 = \Sigma\, x^2 \cdot P(x) - \mu^2$
Normal distribution	$z = (x - \mu) / \sigma$; $x = \mu + z\sigma$; 95% central: $\mu \pm 1.96\sigma$
Binomial	$P(X = r) = nCr\, p^r q^{n-r}$; $\mu = np$; $\sigma^2 = npq$
Poisson	$P(X = r) = e^{-\lambda} \lambda^r / r!$; $\mu = \sigma^2 = \lambda$
Approximations	Binomial → normal if $np \geq 5$ and $nq \geq 5$ ($\mu = np$, $\sigma = \sqrt{npq}$, continuity correction ±0.5); binomial → Poisson if n large, p small ($\lambda = np$)
Rule of three	If 0 events in n subjects, upper 95% limit of the rate ≈ 3 / n

Noteskarts
Smart Notes, Bright Future

STUDY SMARTER WITH NOTESKARTS

Free and premium resources for D.Pharm and B.Pharm students

WHERE TO FIND US

Noteskarts Prime App – complete notes, test series, PYQs and video courses in one place.

Noteskarts.com – free study material, blog articles and exam updates.

YouTube – Noteskarts – lecture videos and revision sessions.

WhatsApp and Telegram communities – daily updates and doubt discussion.

NOTE FOR STUDENTS

These notes are prepared for study and revision. Reference values and pharmacopoeial limits vary between editions and laboratories; always follow your prescribed textbook and the current pharmacopoeia.

Noteskarts
Smart Notes, Bright Future