SUBJECT: FOUNDATIONS FOR BUSINESS ANALYTICS

Module 2: Theoretical Distributions

Simplified Intuitive Edition • Advanced Degree Level Notes (Units 5 – 8)

5 Random Variables, Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of a Continuous Random Variable

Intuitive Overview of Random Variables

A **Random Variable** is simply a mathematical rule that assigns a clean numerical value to the outcomes of a real-world experiment. Instead of tracking qualitative stories (like "Customer Accepted the Offer" or "Customer Rejected the Offer"), we map them to numbers (Success = 1, Failure = 0) so we can run calculations. They are structured into two distinct domains based on what values they can take:

- **Discrete Random Variables:** Things you can count clearly with isolated numbers (e.g., 0, 1, 2, 3 default alerts, or the count of clients waiting in a checkout line).
- **Continuous Random Variables:** Things you measure across an unbroken range or timeline where values can have infinite decimal places (e.g., the exact time a user spends on a webpage, a client's annual salary, or stock market return percentages). Because there are infinite points on a continuous line, the chance of hitting an exact decimal point value like exactly 4.123456... seconds is technically zero. Therefore, we always measure the probability of a continuous variable falling within a *range of values* by calculating the area under a curve.

Simplified: The Probability Density Function (PDF)

For a continuous variable, the **Probability Density Function (PDF)**, written as $f(x)$, draws the shape of the data curve over a graph. The height of the curve shows where values cluster, but the height itself is not a probability. To find an actual probability, you must look at the **area under the curve** between two

points. In calculus, finding this area is done using an integration tool, written with the long S symbol (∫), which simply means "sum up the area."

Every valid PDF curve must satisfy two simple baseline rules:

> ***RULE 1: NO NEGATIVE SPACE***
>
> $f(x) \geq 0$ *for all values of* x.
>
> *Plain English Translation:* The curve can never dip below the ground line (zero). You cannot have a negative likelihood of an event occurring.

> ***RULE 2: TOTAL CERTAINTY EQUAL TO 1***
>
> $$\int_{-\infty}^{\infty} f(x)\,dx = 1$$
>
> *Plain English Translation:* If you calculate the total area under the entire curve from the far left ($-\infty$) to the far right (∞), it must add up to exactly 1 (which represents 100% total probability).

To calculate the specific probability that a continuous variable X lands between target values a and b, you calculate the area trapped between those two vertical walls:

> $$P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$$
>
> *Plain English Translation:* The probability of landing between a and b equals the area under the curve sliced from point a to point b.

Simplified: The Cumulative Distribution Function (CDF)

The **Cumulative Distribution Function (CDF)**, written as $F(x)$, is a running total. It answers the question: "What is the probability that our variable is **less than or equal to** a certain value x?" It continuously accumulates the probability area starting from the extreme left tail up to your target point:

> $$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$
>
> *Plain English Translation:* The cumulative total at point x is the running sum of the area under the curve from the beginning up to that point.

Because the CDF is a running total, it has three intuitive rules: it never decreases as you move right; it starts at 0 on the far left (where nothing has accumulated yet); and it finishes at 1 on the far right (where the entire 100% probability space has been gathered).

Simplified: Expected Value (Mean) and Variance

For a continuous variable, the summary statistics are found by continuous integration rather than simple addition:

- **Expected Value (The Long-Run Average, μ):** Think of this as the balancing center of gravity of the distribution curve:

$$E[X] = \mu = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Concept: Multiply each possible numeric value x by its density weight $f(x)$, and integrate across the entire spectrum.

- **Variance (The Spread Metric, σ^2):** Measures how far individual data points are scattered away from the average center of gravity:

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx - \mu^2$$

Concept: The average squared value minus the squared mean.

6 Binomial Distribution

The Core Scenario Requirements

The Binomial distribution models a very specific workplace scenario: when you repeat an identical task multiple times, and each single attempt has only two possible outcomes, which you classify as a **Success** (the target event you are tracking, with probability p) or a **Failure** (the opposite event, with probability $q = 1 - p$).

To use this model accurately, you must satisfy four clear real-world criteria: the number of attempts (n) is fixed in advance; every attempt is completely independent of the others; the success rate (p) stays constant throughout; and outcomes are strictly binary.

The Simplified Binomial Formula Breakdown

The probability of getting exactly x successes out of n attempts is calculated using this structured equation:

$$P(X = x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x}$$

Component *Component*	*What it Actually Represents (Plain English)*
$P(X = x)$	*The final probability of scoring exactly x successes.*
${}^{n}C_{x} = n! / (x!(n-x)!)$	***The Combinations Counter:*** *The number of different ways or orders you can arrange x successes across n trials. (The exclamation mark $!$ means factorial: multiply the number down to 1, e.g., $3! = 3 \times 2 \times 1 = 6$).*
p^{x}	*The individual probability of success, multiplied by itself for each success achieved.*
q^{n-x}	*The individual probability of failure, multiplied by itself for all the remaining attempts ($n - x$).*

Simplified Summary Parameters

Instead of running long calculations, you can find the characteristics of a Binomial distribution instantly using these simplified parameters:

- **Expected Average Successes (μ):** $n \times p$ (Attempts multiplied by success rate).
- **Variance (σ^2):** $n \times p \times q$ (Always smaller than the mean because q is a fraction).

Step-by-Step Practical Case Review

An analytics platform reviews credit applications. The systemic default rate is 10% ($p = 0.10$, which means failure rate $q = 0.90$). A manager reviews a random batch of $n = 5$ independent applications. What is the chance that exactly 2 default ($x = 2$)?

Step-by-Step Math Breakdown:

Step 1: Calculate the Combinations Counter ($^5C_2$):

$^5C_2 = 5! / (2! \times 3!) = (5 \times 4 \times 3 \times 2 \times 1) / [\ (2 \times 1) \times (3 \times 2 \times 1)\ ] = 20 / 2 = 10$.

Meaning: There are exactly 10 different sequences where 2 defaults can occur among 5 people.

Step 2: Calculate the Success Component (p^x):

$(0.10)^2 = 0.01$ (The probability of two people defaulting).

Step 3: Calculate the Remaining Failure Component (q^{n-x}):

$(0.90)^{5-2} = (0.90)^3 = 0.729$ (The probability of the other three people safely non-defaulting).

Step 4: Multiply the Components Together:

$P(X = 2) = 10 \times 0.01 \times 0.729 = 0.0729$ or **7.29%**.

7 Poisson Distribution

The Core Scenario Requirements

The Poisson distribution switches focus from fixed trials to **continuous intervals**. It models the count of occurrences of an event over a defined segment of time, workspace, volume, or distance (e.g., the number of website crashes per hour, or the number of blemishes per square meter of fabric). It is used when events occur randomly and independently, at a known stable average rate, and cannot occur at the exact same fraction of a second.

The Simplified Poisson Formula Breakdown

Given an average occurrence rate of lambda (λ) over an interval, the probability of seeing exactly x events is calculated as follows:

$$P(X = x) = [\ e^{-\lambda} \cdot \lambda^x\ ] / x!$$

Component Component	What it Actually Represents (Plain English)
$P(X = x)$	The final probability of tracking exactly **x occurrences.**
λ (Lambda)	**The Expected Average Rate:** The long-run historical baseline average count of events within that specific interval size.
$e^{-\lambda}$	The mathematical constant base **e** ($ pprox 2.71828$) raised to the negative power of your average rate. This acts as a dampening factor that decreases as the rate increases.
λ^x	The historical average rate raised to the power of your targeted occurrence count.
$x!$	**The Success Factorial:** Scales down the probability for high extreme counts ($x imes (x-1) imes ... imes 1$).

Unique Identity: Mean Equals Variance

The Poisson model has a unique mathematical shortcut: the expected mean and the variance are completely identical and equal to lambda ($E[X] = Var(X) = \lambda$). If your operational logs show that the average number of server crashes is 3, and the variance is also approximately 3, this model is an optimal fit.

Step-by-Step Practical Case Review

A network server faces an average of $\lambda = 3$ dropouts per hour. What is the exact probability that the center experiences exactly 1 dropout during the next hour ($x = 1$)?

Step-by-Step Math Breakdown:

Step 1: Calculate the Dampening Base Component ($e^{-\lambda}$):

$e^{-3} = 1 / (2.71828)^3 \approx 0.049787$.

Step 2: Calculate the Target Rate Power Component (λ^x):

$3^1 = 3$.

Step 3: Calculate the Success Factorial (x!):

$1! = 1$.

Step 4: Combine the Values:

$P(X = 1) = (0.049787 \times 3) / 1 = 0.14936$ or **14.94%**.

8 Normal Distribution, Chi-Square Distribution, Student's t-Distribution, F-Distribution

Continuous sampling distributions are the foundations for advanced business hypothesis testing and analytics audits. We break down their equations and components into intuitive logic:

I. The Normal Distribution (The Symmetrical Bell Curve)

The Normal distribution maps data that clusters symmetrically around a central average. Its shape is driven by the **Central Limit Theorem**, which states that if you take large random samples ($n \geq 30$) from *any* population, the sample averages will always form a smooth normal bell curve, making it the bedrock of statistical inference.

The raw density equation looks intimidating, but it breaks down into simple structural parts:

$$f(x) = [\ 1 / (\sigma \sqrt{2\pi})\ ] \cdot e^{-1/2\ [\ (x - \mu) / \sigma\ ]^2}$$

Simplified Component Legend:

- $[\ 1 / (\sigma \sqrt{2\pi})\ ]$ ***(The Scaling Wall):*** *This front block uses the population standard deviation (σ) and pi ($\pi \approx 3.14159$). Its sole function is to scale the height of the curve so that the total area underneath stays exactly equal to 1.*
- $[\ (x - \mu) / \sigma\ ]$ ***(The Core Distance):*** *This is simply the **Z-score** formula. It measures how many standard deviations away your target data value x is from the population average (μ).*
- ***The Exponent Base e Block:*** *Symmetrically slopes the curve down to zero on both sides as the distance from the average grows.*

The area under this bell curve follows the **Empirical Rule**: 68.27% of data falls within 1 standard deviation from the center, 95.45% falls within 2 standard deviations, and 99.73% falls within 3 standard deviations (forming the statistical boundaries for Six Sigma corporate quality controls).

II. The Chi-Square (χ^2) Distribution (The Variance Tracker)

If you take multiple independent standard normal data points (Z), square each one to remove negative signs, and add them together, you get a **Chi-Square Distribution**:

> $$\chi^2 = Z_1^2 + Z_2^2 + \dots + Z_k^2$$
>
> *Intuitive Rule:* Because every component is squared, a Chi-Square value can never be negative ($[0, \infty)$). It is skewed to the right, but shapes into a symmetrical curve as you add more variables (degrees of freedom k).

Business Use: Powers **Goodness-of-Fit Tests** (checking if observed retail customer counts match corporate forecasts) and **Independence Tests** across data sheets.

III. Student's t-Distribution (The Small-Sample Guardrail)

When you want to run an analysis on a population mean but your sample size is small ($n < 30$) and you don't know the true population standard deviation, the normal curve becomes unreliable. You must use the **Student's t-Distribution** instead. Mathematically, it is a standard normal variable divided by a chi-square variance tracker:

> $$t = Z / \sqrt{(\chi^2 / v)}$$
>
> *Intuitive Rule:* The t-distribution looks like a normal curve but has **heavier, thicker tails**. This thickness adds an extra margin of safety (guardrail) to protect your analysis against the uncertainty of small sample sizes. As your sample size grows towards infinity, the tails thin out and it becomes identical to the normal Z-distribution.

IV. Snedecor's F-Distribution (The Multi-Group Variance Ratio)

The F-distribution is built to compare variances. It is simply the mathematical ratio of two separate, independent Chi-Square variance trackers, each divided by its own degrees of freedom:

> $$F = [\chi_1^2 / v_1] / [\chi_2^2 / v_2]$$
>
> *Intuitive Verbal Summary:* Think of the F-statistic as:
> F = (Variance measured BETWEEN different groups) / (Variance measured WITHIN the same groups)

If the variance between groups is significantly larger than the internal variance within groups, the F-value spikes high. This serves as the core calculation engine for **ANOVA (Analysis of Variance)** models, which check for performance differences across multiple market testing segments or validate multi-variable linear regression models.

End of Module 2 • Subject: Foundations for Business Analytics