Distributions and Sampling

Block 3: Making Decisions Under Uncertainty

Data Mining and Big Data Course

January 17, 2026

Learning Objectives

After this block, you will be able to:

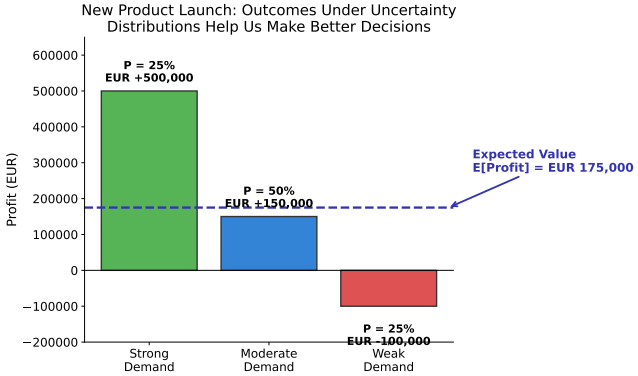
- Distinguish between discrete and continuous random variables
- Calculate and interpret expected value and variance
- Apply the normal distribution to business problems
- Understand sampling distributions and the Central Limit Theorem
- Construct and interpret confidence intervals
- Determine appropriate sample sizes for surveys

Focus: Practical business applications with marketing and pricing examples

Why Distributions Matter in Business

Every business decision involves uncertainty

- How many units will we sell tomorrow?
- What price will customers accept?
- Will this marketing campaign succeed?



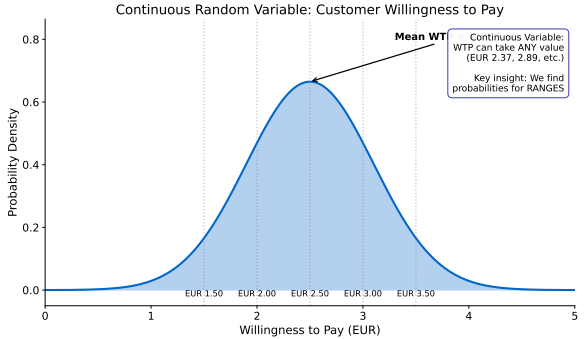
Key Insight: Distributions help us quantify and manage uncertainty

Random Variables

Discrete: Countable, distinct values (purchases: 0, 1, 2, ...; ratings: 1-5)

$$P(X = 0) = 0.30, \quad P(X = 1) = 0.40, \quad P(X = 2) = 0.20, \quad P(X = 3) = 0.10$$

Continuous: Any value in an interval (price, revenue, time)



Key Insight: Discrete: exact probabilities; Continuous: probabilities for ranges

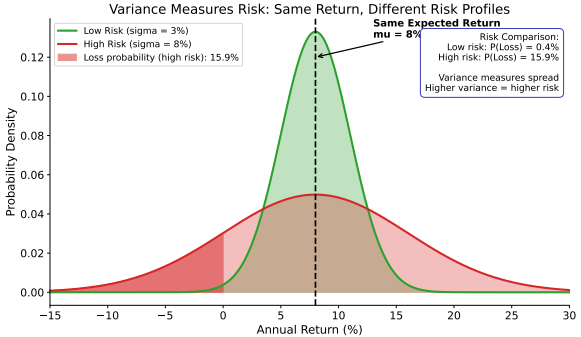
Expected Value and Variance

Expected Value: Weighted average of outcomes

$$E[X] = \sum_i x_i \cdot P(X = x_i) \qquad \text{(what we expect "on average")}$$

Variance: How spread out outcomes are from the mean

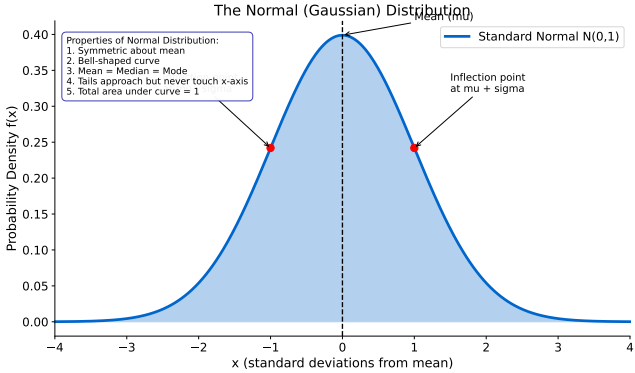
$$\text{Var}(X) = E[(X - \mu)^2] = \sigma^2 \qquad \text{(higher variance = higher risk)}$$



Key Insight: Same mean can have very different risk profiles

The Normal Distribution

The bell curve: Most important distribution in statistics

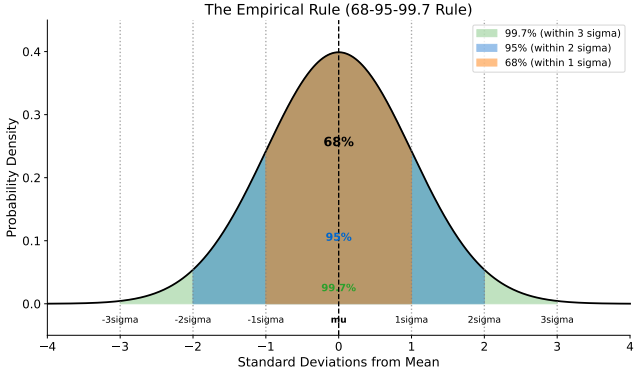


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Key Insight: Many business metrics approximately follow a normal distribution

Parameters and the Empirical Rule

Two parameters define a normal: μ (location) and σ (spread)



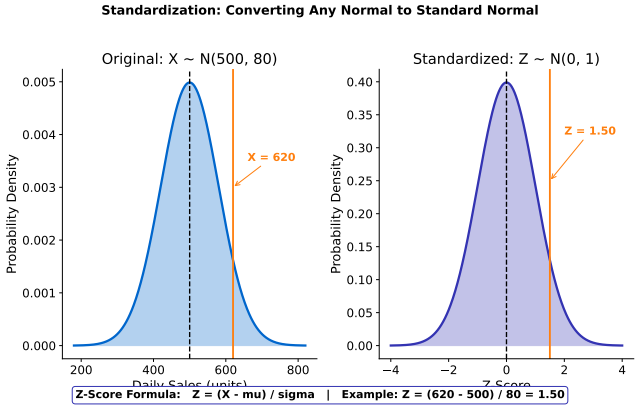
68-95-99.7 Rule: 68% within 1σ, 95% within 2σ, 99.7% within 3σ

Key Insight: Memorize this rule for quick probability assessments

Standardization: The Z-Score

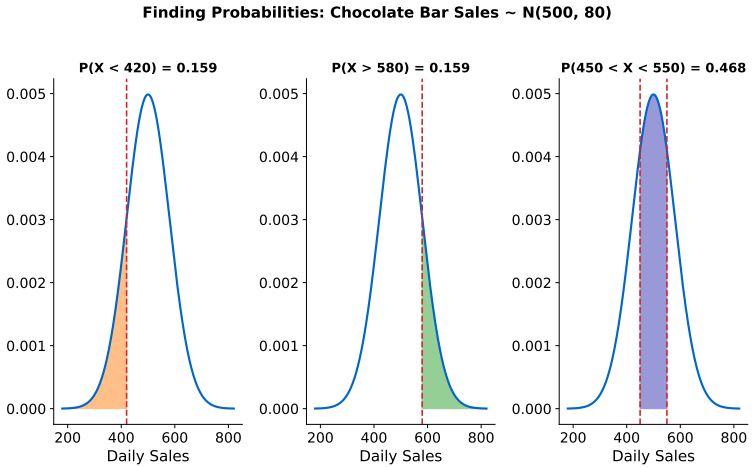
Z-Score: Converts any normal to the standard normal $N(0, 1)$

$$Z = \frac{X - \mu}{\sigma}$$



Key Insight: Z-score tells us how many standard deviations a value is from the mean

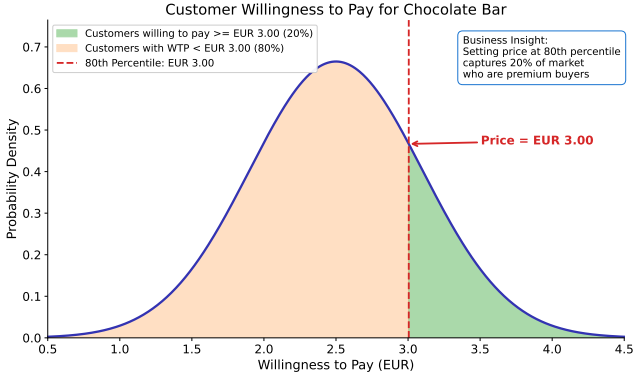
Finding Probabilities



Key Insight: Convert to Z-scores, then use tables or software

Business Example: Pricing Strategy

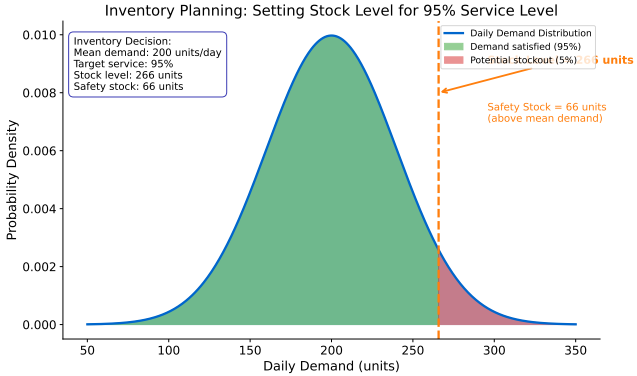
Question: What price captures 80% of customers?



Key Insight: Quantiles help set optimal price points for market segmentation

Business Example: Inventory Planning

Goal: Stock enough to satisfy 95% of demand days

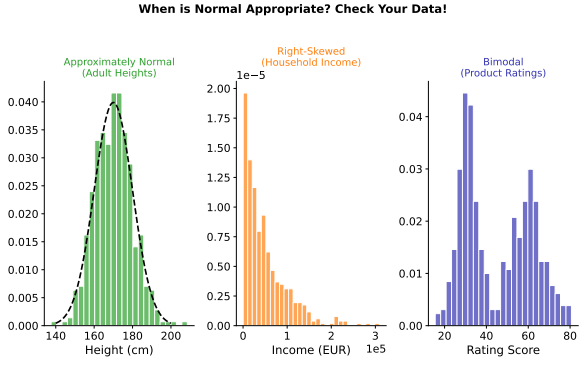


Key Insight: Safety stock = extra units above mean to handle demand variability

Checking Normality

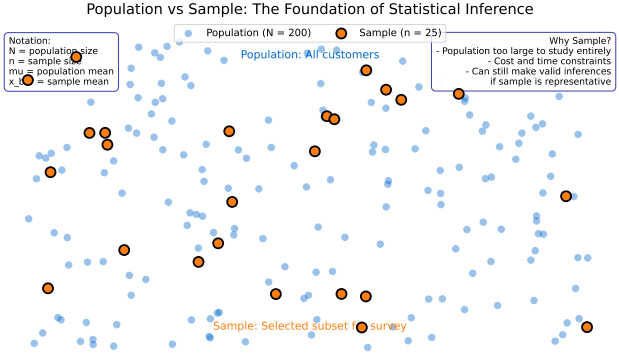
Before applying normal methods, verify your data:

- Visual check: histogram, density plot
- QQ-plot: points should follow diagonal line
- Sample size: CLT helps with $n \geq 30$



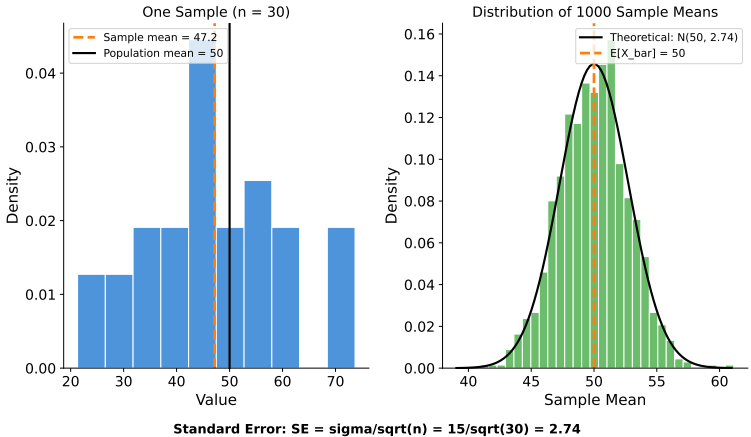
Key Insight: See appendix for detailed QQ-plot interpretation

Population vs Sample



Key Insight: We study samples to make inferences about populations

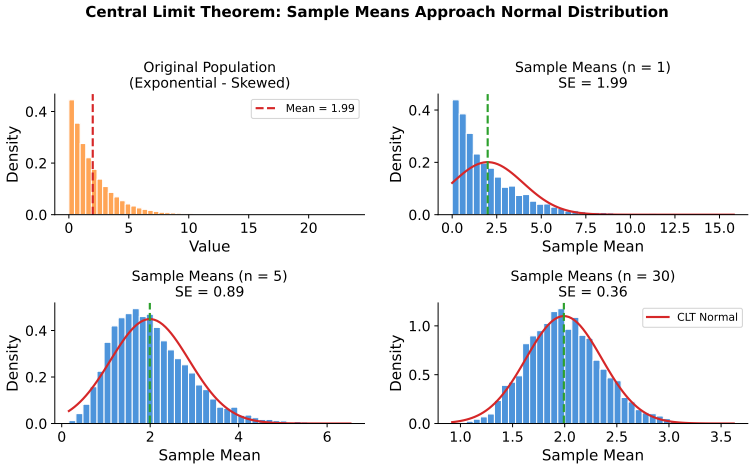
Sampling Distribution of the Mean



Key Insight: The sampling distribution shows how sample means vary from sample to sample

The Central Limit Theorem

CLT: The most important theorem in statistics

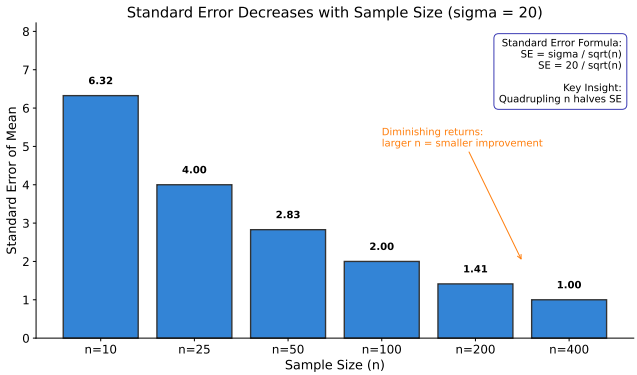


Key Insight: Regardless of population shape, sample means approach normal for large n

Standard Error

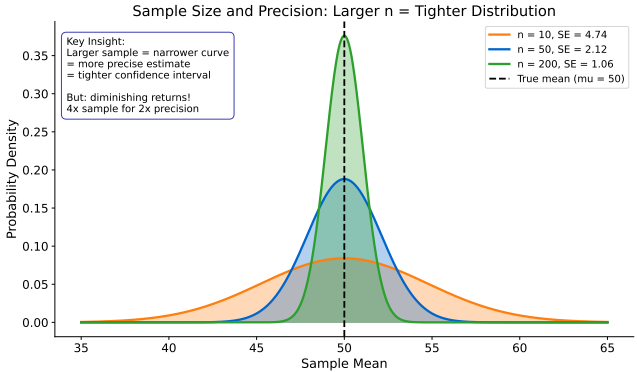
Standard Error: The standard deviation of the sampling distribution

$$SE = \frac{\sigma}{\sqrt{n}}$$



Key Insight: Larger samples = smaller SE = more precise estimates

Sample Size and Precision



Diminishing returns: Quadrupling n only halves the standard error

Key Insight: Balance cost vs accuracy in survey planning

Practical Sample Size Guidelines

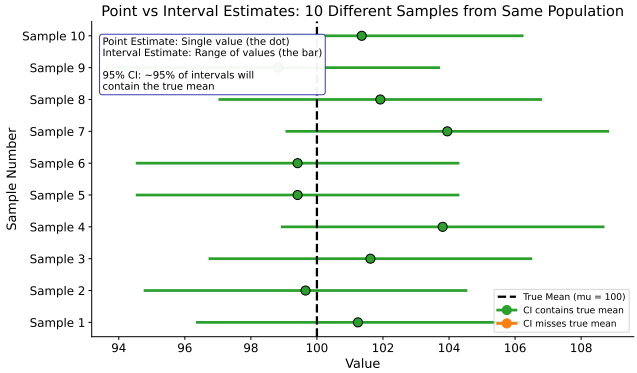
Rules of thumb for business surveys:

- Minimum $n = 30$ for CLT to apply
- $n = 100$ for basic confidence
- $n = 400$ for 5% margin of error (proportions)
- $n = 1000+$ for precise subgroup analysis

Factors: Expected variability, desired margin of error, budget, need for subgroups

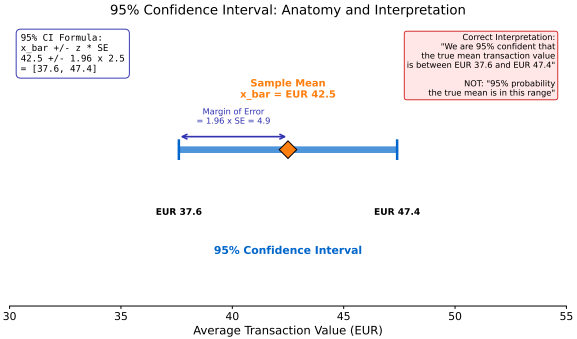
Key Insight: Precision has diminishing returns: balance cost vs accuracy

From Point to Interval Estimates



Key Insight: Intervals acknowledge uncertainty; points give false precision

Confidence Interval: Concept and Formula



Formula: $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ (e.g., $z = 1.96$ for 95% CI)

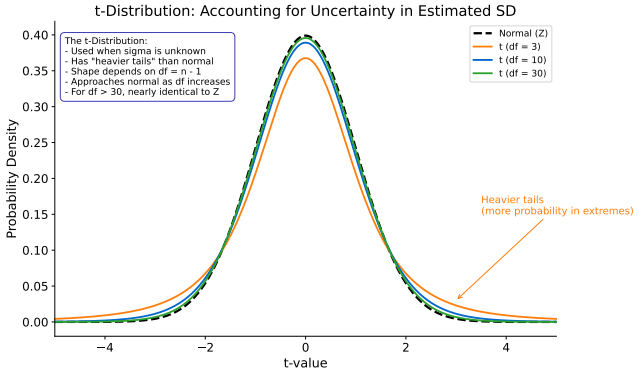
Interpretation: 95% of such intervals contain the true parameter

Key Insight: CI = Point Estimate $\pm$ Margin of Error

The t-Distribution

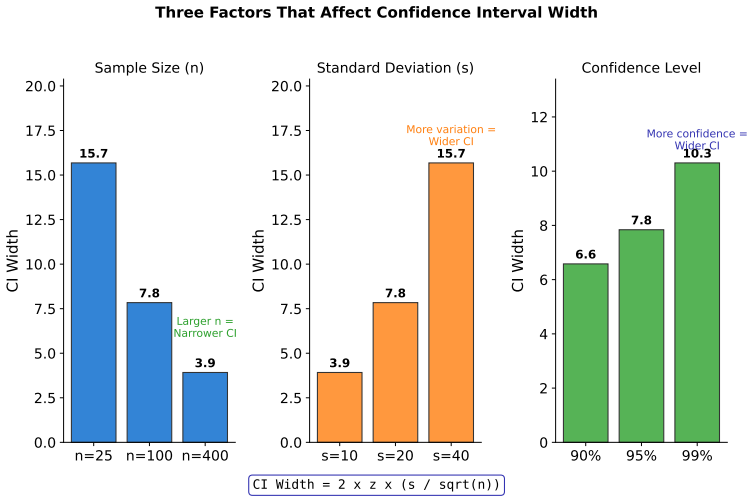
When σ is unknown (most common): Use t instead of z

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$



Key Insight: Heavier tails account for extra uncertainty; for $n > 30$, $t \approx z$

Factors Affecting CI Width

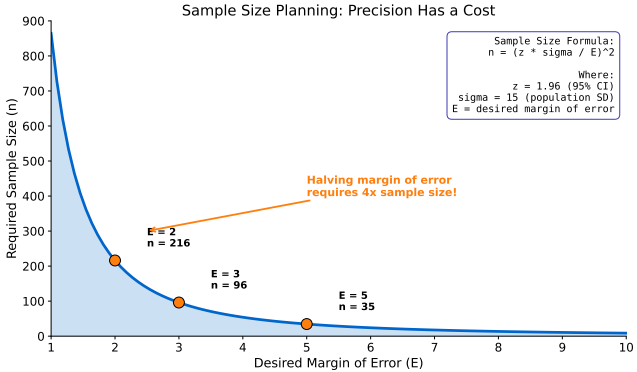


Key Insight: To narrow CI: increase n, reduce variability, or accept lower confidence

Sample Size for Desired Precision

Planning surveys: How many to sample?

$$n = \left(\frac{z \cdot \sigma}{E}\right)^2$$

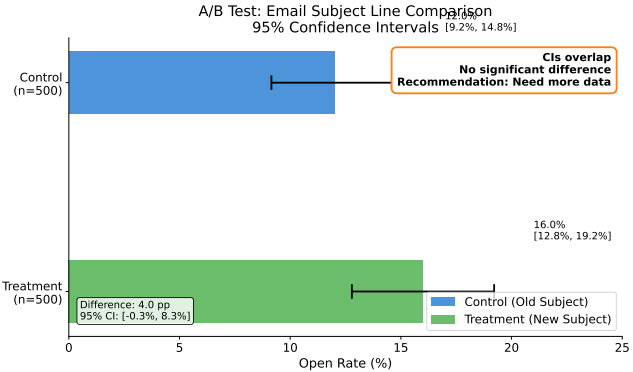


Key Insight: Halving margin of error requires quadrupling sample size

Interpreting Confidence Levels

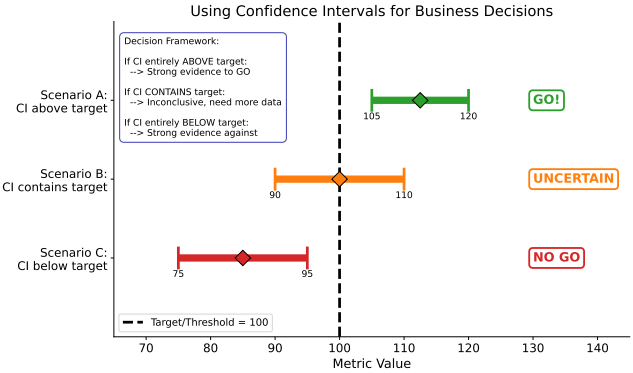
Key Insight: Choose confidence level based on decision consequences

A/B Testing Framework



Key Insight: Non-overlapping CIs indicate statistically significant difference

Decision Making with Uncertainty



Key Insight: Use CIs to make go/no-go decisions with appropriate uncertainty

Common Pitfalls to Avoid

Misinterpretations:

- "There is a 95% probability the true mean is in this CI" – WRONG
- Correct: "95% of such intervals contain the true mean"

Common mistakes:

- Using normal methods for highly skewed data
- Ignoring non-response bias in surveys
- Confusing confidence level with probability

Best practices: Report n and confidence level; check assumptions; consider practical significance

Summary and Key Takeaways

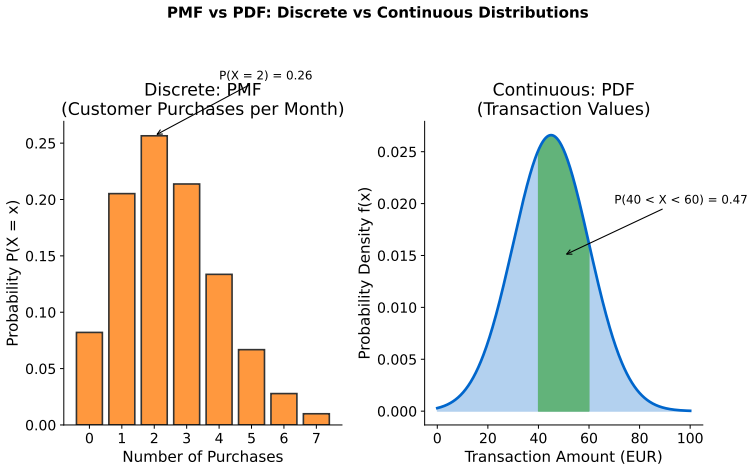
Core Concepts:

1. **Distributions** describe uncertainty in business outcomes
2. **Normal distribution** is central; characterized by μ and σ
3. **Sampling** allows inference from part to whole
4. **CLT** ensures sample means are approximately normal
5. **Confidence intervals** quantify estimation uncertainty

Key Formulas:

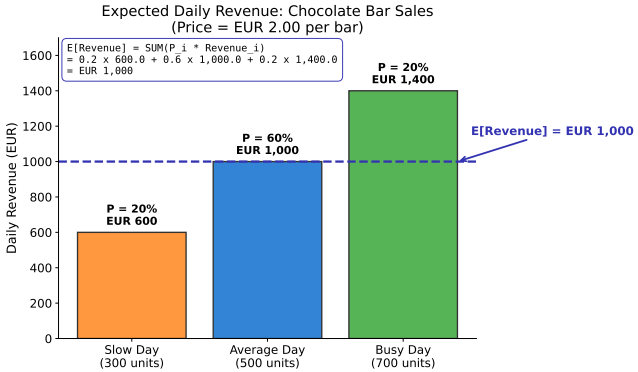
- Z-score: $Z = \frac{X-\mu}{\sigma}$
- Standard Error: $SE = \frac{\sigma}{\sqrt{n}}$
- 95% CI: $\bar{x} \pm 1.96 \cdot SE$
- Sample Size: $n = \left(\frac{z \cdot \sigma}{E}\right)^2$

Appendix: PMF vs PDF Details



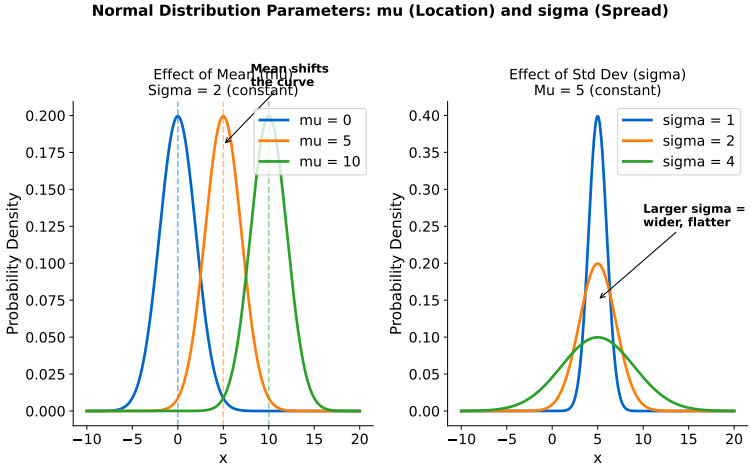
Key Insight: PMF gives exact probabilities; PDF requires integration over ranges

Appendix: Expected Revenue Calculation



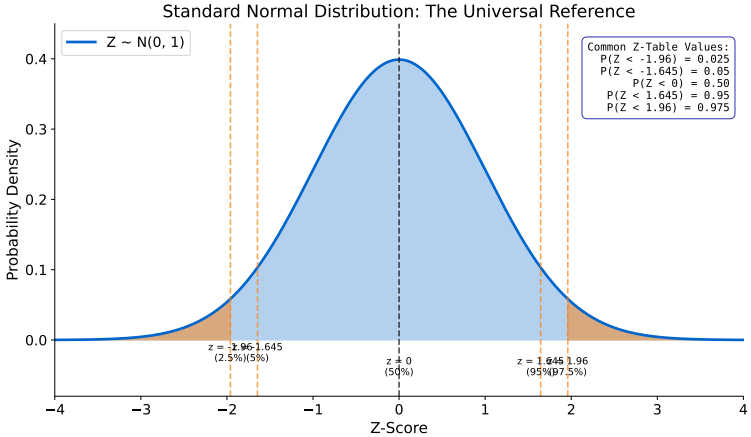
Key Insight: Expected value tells us what to expect "on average" over many trials

Appendix: Normal Parameters Visualization



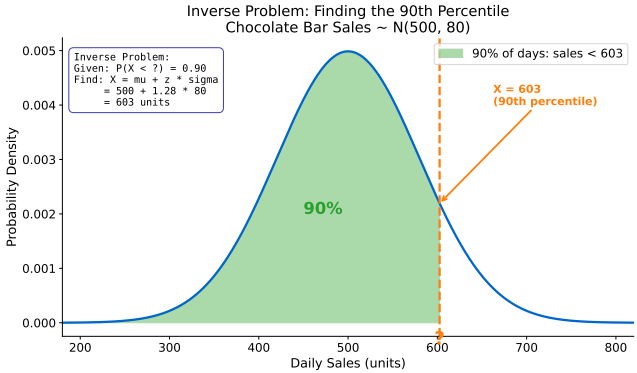
Key Insight: Two parameters completely define a normal distribution

Appendix: Standard Normal Distribution



Key Insight: The standard normal $Z \sim N(0, 1)$ is our universal reference

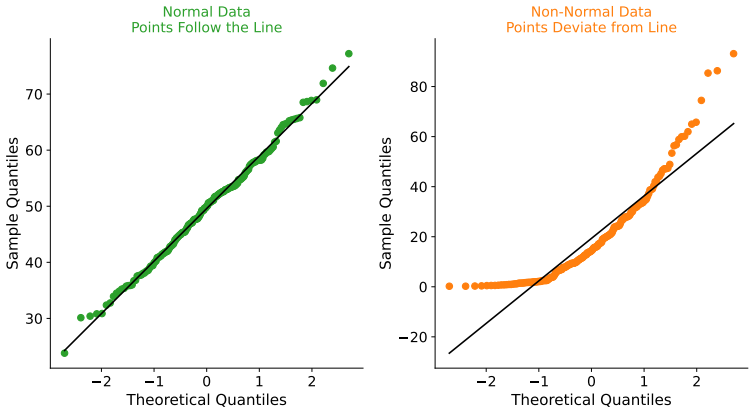
Appendix: Finding Quantiles



Key Insight: Formula: $X = \mu + z \cdot \sigma$ where z comes from the standard normal

Appendix: QQ-Plots for Normality

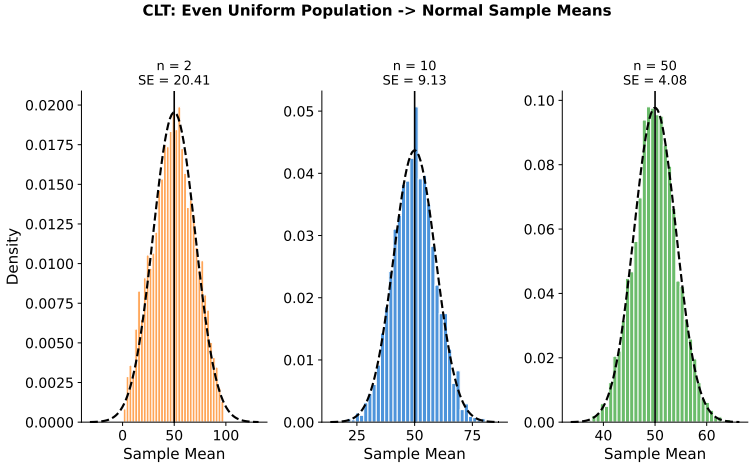
QQ-Plots: Visual Tool for Assessing Normality



QQ-Plot Interpretation: If data is normal, points should fall close to the diagonal line

Key Insight: Points on the diagonal line = data is approximately normal

Appendix: CLT Demonstration (Uniform)

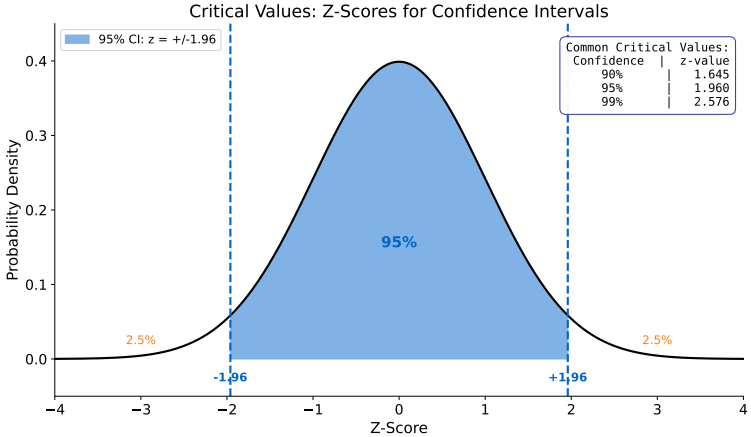


Key Insight: Even from a completely flat distribution, sample means become normal

Appendix: Customer Satisfaction Survey Example

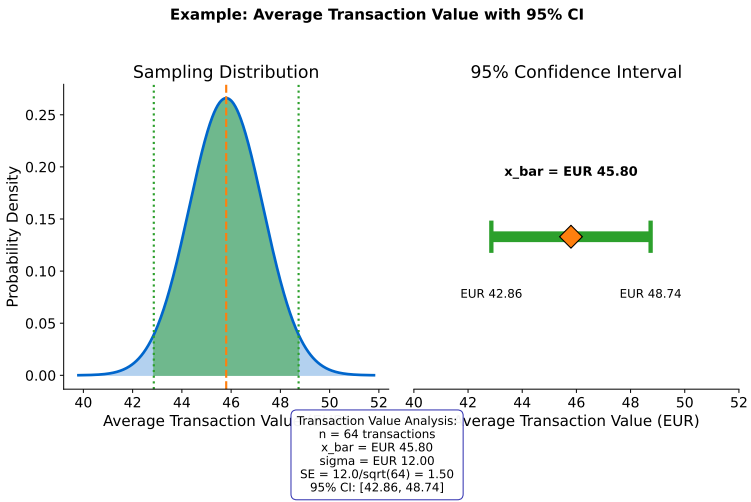
Key Insight: The CI gives us a range of plausible values for true satisfaction

Appendix: Critical Values Table



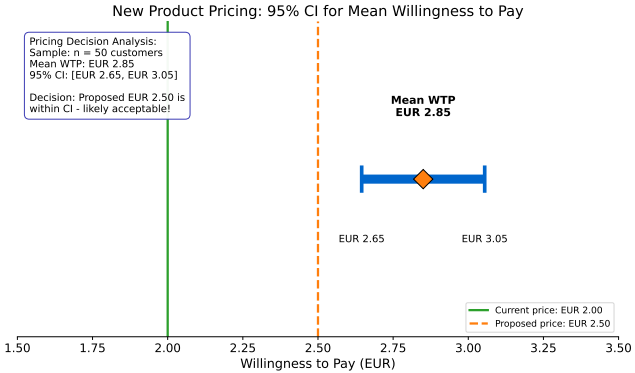
Key Insight: Higher confidence requires wider interval (larger critical value)

Appendix: Transaction Value CI Example



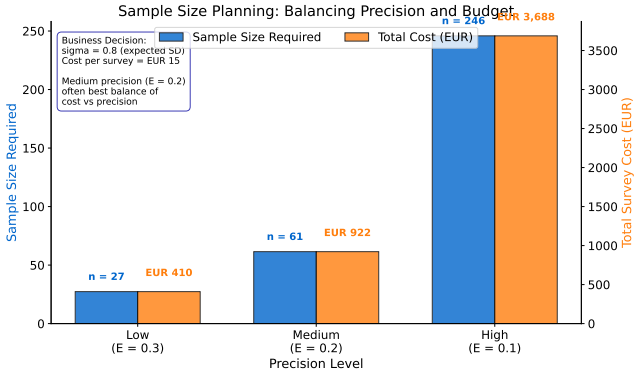
Key Insight: The CI tells management the likely range for true average transaction

Appendix: Pricing CI Example



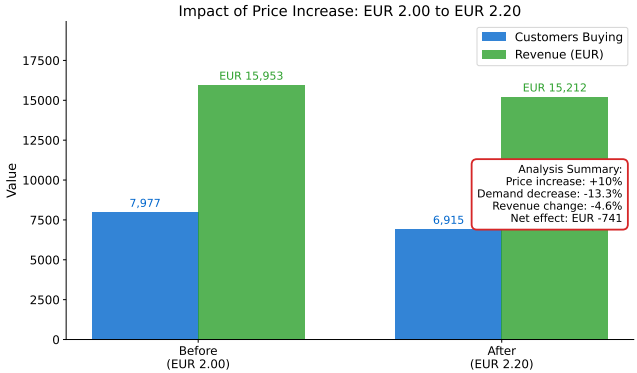
Key Insight: CI for WTP helps determine if proposed price is viable

Appendix: Marketing Sample Size Example



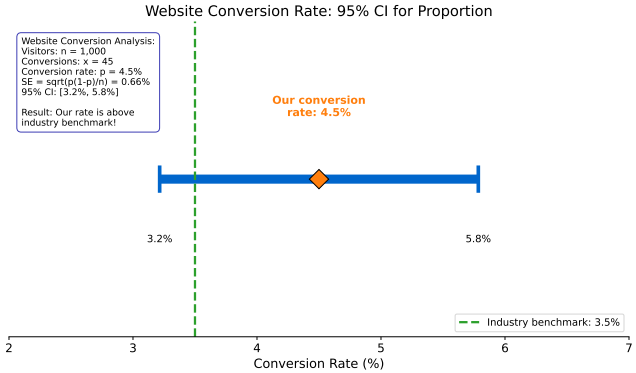
Key Insight: Balance precision needs against survey budget

Appendix: Chocolate Revenue Impact



Key Insight: Distribution analysis shows net revenue impact of price change

Appendix: Website Conversion Rate Example



Key Insight: CI for proportions uses $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$