Stats Recap & Bayesian Thinking

Augusto Gonzalez Bonorino

August 23, 2023

Table of Contents

Recap of Descriptive Statistics and Distributions

Population vs Sample:

- Population: Entire set of elements under study.
- Sample: Subset of population used for analysis.

Random Sampling:

- Simple Random Sampling: Each member equally likely to be chosen.
- Stratified Sampling: Dividing population into subgroups.

Measures of Centrality:

- Mean:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Median: Middle value of ordered data.
- Mode: Most frequent value.

Measures of Dispersion:

- Variance:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

- Standard Deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

Probability Density Function (PDF):

- Describes the likelihood of a continuous random variable taking a particular value.
- Denoted by $f(x)$, where $f(x) \geq 0$ for all values of x.
- The area under the PDF curve over a range equals the probability of the variable falling within that range.

Cumulative Distribution Function (CDF):

- Describes the probability of a random variable being less than or equal to a certain value.
- Denoted by $F(x)$, where $0 \leq F(x) \leq 1$.

Recap of Descriptive Statistics and Distributions

Exercise 1: Calculate the mean, median, and standard deviation of the following dataset:

$$X = \{15, 18, 20, 22, 24, 25, 28, 30, 32, 35\}$$

Exercise 2 (Economic Application): Suppose the dataseta X_A and X_B represent the monthly sales (in thousands of dollars) of two competing firms, A and B, for the past year:

$$X = \{A : 100, 120, 110, 130, 105, 125, 135, 140, 115, 150\}$$

$$X = \{B : 90, 85, 100, 95, 80, 105, 110, 120, 85, 95\}$$

Calculate the coefficient of variation for each firm's sales to compare their variability.

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

- Mean ($\bar{x}$):

$$\bar{x} = \frac{15 + 18 + 20 + 22 + 24 + 25 + 28 + 30 + 32 + 35}{10} = 25.9$$

- Median (Middle value): Since there are 10 values, the median is the average of the 5th and 6th values:

$$\text{Median} = \frac{24 + 25}{2} = 24.5$$

- Standard Deviation (σ): First, calculate the squared differences from the mean and their sum:

$$(15 - 25.9)^2 + (18 - 25.9)^2 + \ldots + (35 - 25.9)^2 = 273.2$$

Then divide by the number of values and take the square root:

$$\sigma = \sqrt{\frac{273.2}{10}} = 5.23$$

Recap of Descriptive Statistics and Distributions

Exercise 2 Solution

Let's calculate the coefficient of variation for each firm's sales:

Firm A:

- Mean (M_A):

$$M_A = 123.5$$

- Standard Deviation (σ_A):

$$\sigma_A = 15.28$$

- Coefficient of Variation (A):

$$\text{CV(A)} = \frac{\sigma_A}{M_A} \times 100$$

Firm B:

- Mean (M_B):

$$M_B = 93.5$$

- Standard Deviation (σ_B):

$$\sigma_B = 11.65$$

- Coefficient of Variation (B):

$$\text{CV(B)} = \frac{\sigma_B}{M_B} \times 100$$

The coefficient of variation allows us to compare the variability of the sales of the two firms.

Review of PDF and CDF

Exercise 1

Consider a continuous random variable X with the following PDF:

$$f(x) = \begin{cases} 0.5x & \text{for } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the probability that X takes a value between 1 and 1.5.

To find the probability that X takes a value between 1 and 1.5, we integrate the PDF over that range:

$$P(1 \leq X \leq 1.5) = \int_{1}^{1.5} 0.5x \, dx$$

Solving the integral:

$$P(1 \leq X \leq 1.5) = \left[0.25x^2\right]_1^{1.5} = 0.25 \times (1.5^2 - 1^2) = 0.5625$$

For a certain random variable Y, the CDF is given by:

$$F(y) = \begin{cases} 0 & \text{for } y < 0 \\ 0.2y^2 & \text{for } 0 \leq y < 2 \\ 1 & \text{for } y \geq 2 \end{cases}$$

Find the probability that Y is greater than 1.

Review of PDF and CDF

Exercise 2 Solution

The probability that Y is greater than 1 is the complement of the probability that Y is less than or equal to 1:

$$P(Y > 1) = 1 - P(Y \leq 1) = 1 - F(1)$$

Substitute the value of $F(1)$ from the CDF:

$$P(Y > 1) = 1 - 0.2 \times 1^2 = 1 - 0.2 = 0.8$$

Frequentist vs Bayesian Statistics

- Frequentist thinking: Probability as long-term relative frequency. Focuses on data and long-run behavior (i.e., expected behavior).
- Main ideas of frequentist statistics: point estimation and confidence intervals.
- Bayesian thinking: Probability as degree of belief. Incorporates prior information and updates beliefs.
- Main ideas of Bayesian statistics: posterior distribution and credible intervals.

Frequentist vs Bayesian Thinking

	Frequentist Approach	Bayesian Approach
Pros	Objective, data-driven	Incorporates prior knowledge
	Well-established methods	Updates beliefs with new data
	Consistent in large samples	Flexible for complex models
Cons	Limited incorporation of prior info	Subjectivity in prior selection
	Doesn't handle small samples well	Can be computationally intensive
	No direct measure of belief	Interpretation of priors can vary

Table: Comparison of Frequentist and Bayesian Approaches

Conceptual Introduction to Statistical and Causal Inference

Descriptive Inference:

- Summarizing data.
- Example: Calculating the mean income in a city.

Predictive Inference:

- Making predictions based on data patterns.
- Example: Forecasting next month's sales.

Inferential Inference:

- Drawing conclusions about populations from samples.
- Example: Estimating the average age of all voters from a sample.

Causal Inference:

- Establishing cause-and-effect relationships.
- Example: An economist wants to study the impact of education level on individuals' lifetime earnings. This involves identifying the causal relationship between education and earnings while controlling for other factors.

Challenges of Causal Inference

Challenges:

- Confounding Variables: Uncontrolled factors affecting the outcome.
- Selection Bias: Non-random assignment to groups.
- Reverse Causation: Confusing cause and effect.
- Ecological Fallacy: Making assumptions about individuals from group data.

Example: Studying the impact of a new policy on income inequality while accounting for other economic factors.

Counterfactuals:

- Hypothetical scenarios to compare with observed reality.
- Central to causal reasoning.

Economic Applications

- Bayesian econometric models: Estimating demand elasticity using Bayesian regression.
- Decision theory in economics using Bayesian approach (Asymmetric or Imperfect information games).
- Example: Estimating consumer demand for a new product with Bayesian methods.
- Example: Bayesian analysis of investment decisions considering uncertainty in the market.