ECON 57 - Lecture 3

Statistical Thinking & Densities

Augusto Gonzalez-Bonorino

Pomona College

Fall 2023

Table of Contents

1. Statistical Thinking
2. Frequency Distributions
    - Discrete Density
    - Continuous Density
3. Cumulative Frequency Function
4. Visualizing Frequency Tables and Data
    - Bar Charts, Histograms, Pie Charts, Box Plots, & Scatterplots
5. Conceptual Exercises
6. Mathematical Exercises

Statistical Thinking

Statistical thinking involves understanding the underlying distributions, analyzing data patterns, and drawing meaningful conclusions from data.

Key Concepts:

- **Distributions and statistics**: Understanding the relationship between data distributions and summary statistics.
- **Hypothesis testing**: Assessing the validity of claims based on sample data.
- **Data visualization**: Utilizing graphs and charts to communicate data insights effectively.
- **Good questions**: Carefully defining the question(s) to be answered is, arguably, the most important component of any study.

Statistical Thinking

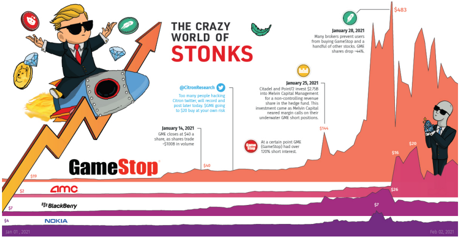
Best Practices for Daily Life

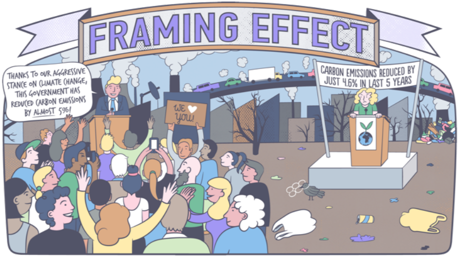
- **Pay attention to the context**: Always consider the context while interpreting the data. The same data can be interpreted differently in different scenarios. Thus, it is important not just to understand the numbers but the story behind the numbers.
- **Practice Skepticism**: Don't take the numbers at face value. Always be on the lookout for biases, manipulations or misinterpretations in the data.
- **Logical Reasoning**: Use your logical reasoning skills to draw conclusions from your data analysis. Question whether your interpretation makes sense in the real-world context.
- **Ethical Considerations**: Respect confidentiality and privacy while dealing with data. Ensure your conclusions are not biased or discriminatory.

Statistical Thinking

Some common biases

- Herd mentality
- Loss aversion
- Framing effect
- Confirmation bias
- Anchoring effect
- Availability heuristics




FRAMING EFFECT
THANKS TO OUR AGGRESSIVE STANCE ON CLIMATE CHANGE, THIS GOVERNMENT HAS REDUCED CARBON EMISSIONS BY ALMOST 5%!
CARBON EMISSIONS REDUCED BY JUST 4.6% IN LAST 5 YEARS
WE ♥ YOU!

Frequency Distributions

Introduction

- The *frequency* of a value is the number of times it occurs in a dataset.
- A *frequency distribution* is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.
  - **Ungrouped frequency distribution** := # of observations of each value of a variable.
  - **Grouped frequency distribution** := # of observations of each class interval of a variable.
  - **Relative frequency distribution** := The proportion of observations of each value or class interval of a variable. Useful for comparing frequencies.
  - **Cumulative frequency distribution** := The sum of the frequencies less than or equal to each value or class interval of a variable. Useful for understanding how often observations fall below certain values.

Frequency Distributions

Frequency/Density Function

For the type of data we will be focusing on in this course (i.e., numerical), we define a function $Fr(.)$ called the **frequency function**. It takes in two parameters:

1. The variable X under study
2. The subset B we want to check for

Therefore, $Fr(X \in B)$ returns the relative frequency with which the variable X takes a value included in $B \subset \mathbb{R}$.

Frequency/Density Function

Discrete case

The discrete frequency function is employed for discrete random variables, where the variable can only take specific, isolated values with associated probabilities.

Mathematically, the discrete density function can be denoted as $p(x)$ and satisfies the following properties:

$$\sum_{\text{all } x} p(x) = 1; \; 0 < p(x) \leq 1, \forall x$$

x_i	$p(x_i)$
2	0.1
3	0.4
5	0.3
6	0.2

Table: Frequency Table for x

$$Fr(x = 2) = 0.1$$
$$Fr(x = 3) = 0.4$$
$$Fr(x \in \{2, 6\}) = 0.1 + 0.2 = 0.3$$
$$Fr(x \in [4, 10]) = 0.3 + 0.2 = 0.5$$

Frequency/Density Function

Continuous case

A continuous density function $f(x)$ is used when dealing with continuous random variables. It represents the relative likelihood of observing a specific value within an interval.

$$f(x) = \begin{cases} c_1 & \text{if } x_1 \leq x < x_2 \\ c_2 & \text{if } x_2 \leq x \leq x_3 \\ \dots & \dots \\ c_k & \text{if } x_k \leq x \leq x_{k+1} \\ 0 & \text{otherwise} \end{cases}$$

Properties:

- $f(x) \geq 0, \forall x$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- For $-\infty \leq x_a \leq x_b \leq \infty$ where $a < b$, we have $Fr(x_a \leq x \leq x_b) = \int_{x_a}^{x_b} f(x)dx$
- $c_i :=$ density $= \frac{p_i}{x_{i+1} - x_i}, i : 1 \dots k$

Frequency/Density Function

Example (discrete): Questionnaire responses

Let's apply these concepts to the data collected from the questionnaire.

Frequency/Density Function

Example (continuous): Income distribution

Income Intervals	Frequency	p_i	δ_i	d_i
\$10,000 - \$20,000	8	$\frac{8}{100} = 0.08$	\$10,000	$\frac{0.08}{10{,}000}$
\$20,000 - \$30,000	12	$\frac{12}{100} = 0.12$	\$10,000	$\frac{0.12}{10{,}000}$
\$30,000 - \$40,000	20	$\frac{20}{100} = 0.20$	\$10,000	$\frac{0.20}{10{,}000}$
\$40,000 - \$50,000	25	$\frac{25}{100} = 0.25$	\$10,000	$\frac{0.25}{10{,}000}$
\$50,000 - \$60,000	15	$\frac{15}{100} = 0.15$	\$10,000	$\frac{0.15}{10{,}000}$
\$60,000 - \$70,000	9	$\frac{9}{100} = 0.09$	\$10,000	$\frac{0.09}{10{,}000}$
\$70,000 - \$80,000	6	$\frac{6}{100} = 0.06$	\$10,000	$\frac{0.06}{10{,}000}$
\$80,000 - \$90,000	3	$\frac{3}{100} = 0.03$	\$10,000	$\frac{0.03}{10{,}000}$
\$90,000 - \$100,000	2	$\frac{2}{100} = 0.02$	\$10,000	$\frac{0.02}{10{,}000}$

Table: Income Intervals, Frequencies, Relative Frequencies (p_i), Interval Length (δ_i), and Density (d_i)

Cumulative Frequency Function

The cumulative frequency function, also known as the cumulative distribution function (CDF), provides valuable insights into the probability of a random variable being less than or equal to a given value.
For a continuous random variable X, the cumulative distribution function $F(x)$ is defined as:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

For a discrete random variable X, the cumulative distribution function $F(x)$ is defined as the sum of probabilities up to x:

$$F(x) = \sum_{t \leq x} p(t)$$

Cumulative Frequency Function

Properties of a CDF for discrete variables

1. F(x) always starts from 0 (on the left of x) and always reaches 1 (on the right of x)
2. F(x) is monotonic and non-decreasing
3. is a step function
4. F(x) has discontinuity points
5. is continuous from the right at the discontinuity points (as many as the number of data points)
6. F(x) is bounded from above and below
7. The height of each step is equal to the corresponding relative frequency
8. There exists a one-to-one correspondence between a discrete variable and its cumulative frequency distribution function.

Cumulative Frequency Function

Properties of a CDF for continuous variables

1. F(x) is monotonic and non-decreasing
2. F(x) is continuous (i.e., no discontinuity points)
3. $\lim_{x->-\infty} F(x) = 0$; $\lim_{x->\infty} F(x) = 1$
4. The slope of F(x) in each interval represents the density of that interval.
5. Since F(x) is continuous, we can take its derivative to obtain the density function. That is, $F'(x) = f(x) \forall x$

Cumulative Frequency Function

Example: Daily stock close prices

AAPL daily close prices 3 months

Visualizing Frequency Tables

The importance of visualization

Visualizations are powerful tools for summarizing and understanding data. There are many benefits of exploring your data visually prior to any statistical analysis.

- **Identifying Patterns and Trends:** Visualizations allow you to quickly identify patterns, trends, and anomalies in your data that might not be immediately apparent from the raw numbers. For example, a bar chart can help you visualize which categories have higher or lower frequencies, making it easier to detect outliers or unusual data points.
- **Enhancing Interpretation:** Visualizations provide a more intuitive way to interpret data. Instead of working with abstract numbers, you can see the data in a graphical format that's easier to understand and explain to others. This is particularly important when communicating results to non-technical stakeholders.

Visualizing Frequency Tables

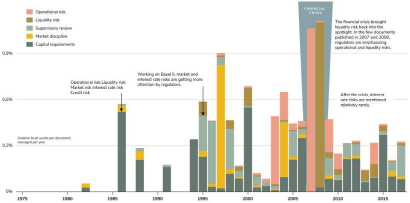
The importance of visualization

- **Comparing Categories:** Visualizations make it simple to compare the frequencies of different categories side by side. This can help you make informed decisions and draw meaningful insights from the data. For instance, a bar chart can show you the relative sizes of various categories, allowing you to identify which categories are dominant and which are less common.
- **Storytelling:** Visualizations can turn data into a compelling story. By carefully selecting the type of visualization and customizing its design, you can convey specific messages and narratives about your data. This is especially useful when presenting your findings to an audience.

Visualizing Frequency Tables

Bar Chart

A bar chart displays the distribution of categorical data using rectangular bars. Each bar's height corresponds to the frequency of the category it represents.



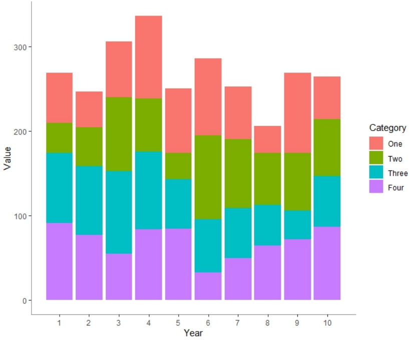
Operational risk Liquidity risk Market risk Interest rate risk Credit risk

Working on Basel II, market and interest rate risks are getting more attention by regulators.

Relative to all words per document, averaged per year

The financial crisis brought liquidity risk back into the spotlight. In the few documents published in 2007 and 2008, regulators are emphasizing operational and liquidity risks.

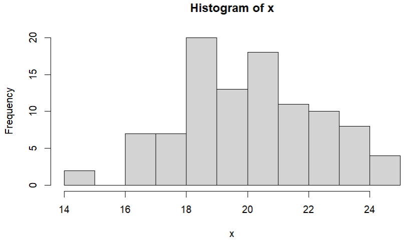
After the crisis, interest rate risks are mentioned relatively rarely.

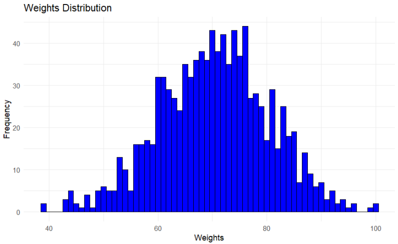


Visualizing Data

Histograms

Apart from frequency tables, we can visualize continuous data using histograms.
A histogram divides the range of continuous data into intervals called bins and represents the frequency of observations falling within each bin.





Bar charts and histograms help visualize the distribution of data and identify patterns.

Visualizing Data

Pie Charts

Pie charts are used to represent the proportion or percentage distribution of categorical data.

Each category is represented as a slice of the pie, and the size of each slice corresponds to the proportion of data in that category. They are useful for visualizing the relative sizes of different categories in the whole.

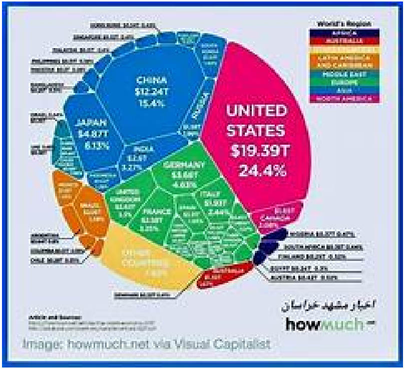
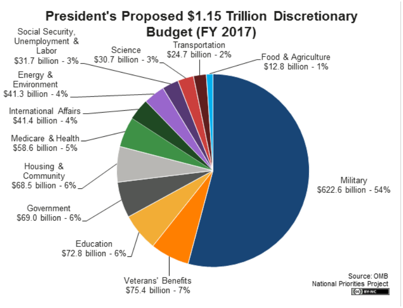
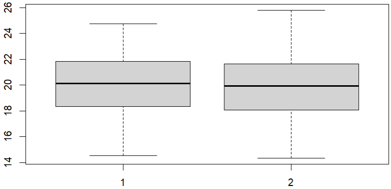


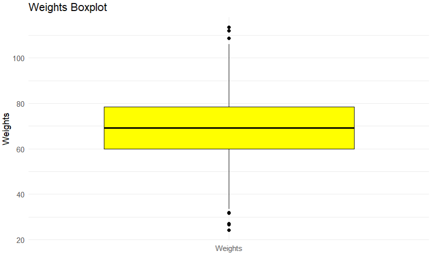
Image: howmuch.net via Visual Capitalist



Visualizing Data
Box Plots

Box plots, also known as box-and-whisker plots, provide a graphical summary of the distribution of continuous data.
The box represents the interquartile range (IQR), with the median marked by a line inside the box. The whiskers extend to the minimum and maximum data points within a certain range (usually 1.5 times the IQR).



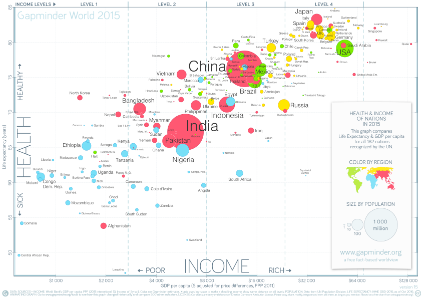


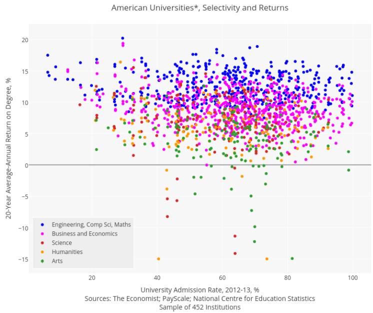
Box plots help identify outliers and compare distributions between different groups.

Visualizing Data

Scatterplots

Scatterplots are used to visualize the relationship between two continuous variables. Each data point is represented as a point on the plot, with one variable on the x-axis and the other on the y-axis.





Sources: The Economist; PayScale; National Centre for Education Statistics
Sample of 452 Institutions

Scatterplots help identify patterns, trends, and the strength of the relationship between the two variables. They play a key role in regression analysis and thus econometrics.

Conceptual Exercises

To solidify our understanding of statistical concepts, let's work on some conceptual exercises.

Exercise 1

Exercise 1: Imagine you have survey data on the ages of 50 participants. You create a frequency table with age intervals and their corresponding frequencies. Explain how you would calculate the relative frequency for each interval and why it's important to use relative frequencies when comparing different datasets.

Solution to Exercise 1

Exercise 1: Imagine you have survey data on the ages of 50 participants. You create a frequency table with age intervals and their corresponding frequencies. Explain how you would calculate the relative frequency for each interval and why it's important to use relative frequencies when comparing different datasets.

Solution to Exercise 1: To calculate the relative frequency for each interval, divide the frequency of that interval by the total number of participants. Relative frequencies allow for easy comparison between datasets of different sizes, as they give a proportion instead of an absolute count. This normalization is essential for comparing distributions across different sample sizes.

Exercise 2

Exercise 2: Consider a dataset of exam scores with a frequency distribution. How does the shape of the distribution affect the interpretation of the dataset? Provide examples of symmetric and skewed distributions and explain how their shapes impact the analysis.

Solution to Exercise 2

Exercise 2: Consider a dataset of exam scores with a frequency distribution. How does the shape of the distribution affect the interpretation of the dataset? Provide examples of symmetric and skewed distributions and explain how their shapes impact the analysis.

Solution to Exercise 2: The shape of the distribution provides insights into the dataset's central tendency and variability. A symmetric distribution (e.g., normal distribution) suggests balanced data, where the mean and median are close. Skewed distributions (e.g., positively skewed) indicate an imbalance, with the tail stretching towards higher values. In skewed distributions, the mean is pulled in the direction of the tail, while the median remains more resistant to outliers.

Exercise 3

Exercise 3: A frequency distribution shows that a certain age group has a mode of 35 years. Explain the significance of this mode in the context of the data and how it relates to other measures of central tendency.

Solution to Exercise 3

Exercise 3: A frequency distribution shows that a certain age group has a mode of 35 years. Explain the significance of this mode in the context of the data and how it relates to other measures of central tendency.

Solution to Exercise 3: The mode is the most frequent value in the dataset. In this context, the mode of 35 years indicates that a significant number of individuals in the age group are around 35 years old. This can be useful for identifying common ages within the group. The mode, along with the mean and median, provides different perspectives on the central tendency of the data.

Mathematical Exercises

Let's work through some mathematical exercises to reinforce our understanding of probability density functions and cumulative distribution functions.

Exercise 1

Exercise 1: Consider a continuous random variable X with probability density function $f(x) = 2x$ for $0 \leq x \leq 1$ and $f(x) = 0$ otherwise. Calculate the cumulative distribution function $F(x)$ for this random variable.

Solution to Exercise 1

Solution to Exercise 1: The cumulative distribution function $F(x)$ can be calculated by integrating the probability density function $f(x)$:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

For $0 \leq x \leq 1$, we have:

$$F(x) = \int_{0}^{x} 2t\, dt = x^2$$

And for $x > 1$, $F(x) = 1$ since the probability is 1 for the entire distribution.

Exercise 2

Exercise 2: Given a probability density function $f(x) = \frac{1}{3}x^2$ for $0 \leq x \leq 3$ and $f(x) = 0$ otherwise, find the probability that X lies between 1 and 2, i.e., $P(1 \leq X \leq 2)$.

Solution to Exercise 2

Solution to Exercise 2: To find the probability $P(1 \leq X \leq 2)$, we need to integrate the probability density function $f(x)$ over the interval $[1, 2]$:

$$P(1 \leq X \leq 2) = \int_1^2 \frac{1}{3}x^2 \, dx$$

Evaluating the integral gives:

$$P(1 \leq X \leq 2) = \frac{1}{3}\left(\frac{2^3}{3} - \frac{1^3}{3}\right) = \frac{7}{27}$$

Exercise 3

Exercise 3: Consider a discrete random variable Y with probability mass function $p(y) = \frac{1}{6}$ for $y = 1, 2, 3, 4, 5, 6$. Calculate the cumulative distribution function $F(y)$ for this random variable.

Solution to Exercise 3

Solution to Exercise 3: The cumulative distribution function $F(y)$ for a discrete random variable can be calculated by summing the probabilities up to the value y:

$$F(y) = \sum_{t \leq y} p(t)$$

For $y = 1$, $F(1) = p(1) = \frac{1}{6}$. For $y = 2$, $F(2) = p(1) + p(2) = \frac{1}{3}$, and so on. The cumulative distribution function is a step function that increases by $\frac{1}{6}$ at each value of y.

Exercise 4

Exercise 4: Based on the properties of a continuous density function $f(x)$, explain why the integral of $f(x)$ over its entire domain equals 1. How does this property relate to the concept of probability?

Solution to Exercise 4

Solution to Exercise 4: The integral of a continuous density function $f(x)$ over its entire domain equals 1 due to the normalization property. This property ensures that the total area under the density curve represents the entire probability space. When integrating $f(x)$ over its domain, we're summing up the probabilities of all possible outcomes, and this sum should equal 1 since we're considering all possible events. This property ensures that probabilities are appropriately scaled and can be interpreted in terms of likelihoods within the given interval.