STATISTICAL LABORATORY

Lecture 3: Exploratory Data Analysis (Continued)

Yiu Lim LUI, Allen

Lisbon School of Economics & Management
Universidade de Lisboa

allenlui@iseg.ulisboa.pt

Applied Mathematics for Economics and Management
1st Year / 1st Semester – 2026/2027

Announcement: Venue change

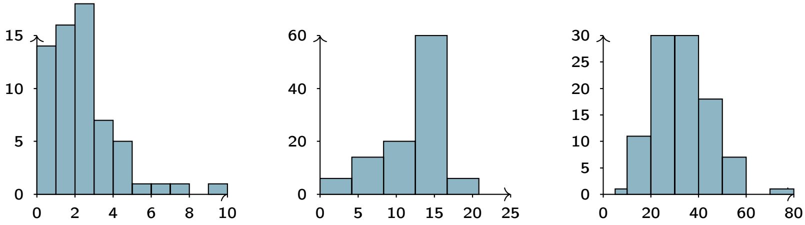
1. Tomorrow (Wednesday, September 23rd), our practice session will be held in room 202 Francesinhas 2.
2. Starting next week, we will return to room 109 Francesinhas 2.
3. The written test will take place during the 12th week (the fourth week of November, likely on the 24th or 25th). I will confirm the exact date as soon as possible and let you know.

Program

1. Fundamental Concepts of Statistics
2. Exploratory Data Analysis
3. Organizing and Summarizing Data
4. Association and Relationships Between Variables
5. Index Numbers
6. Time Series Analysis

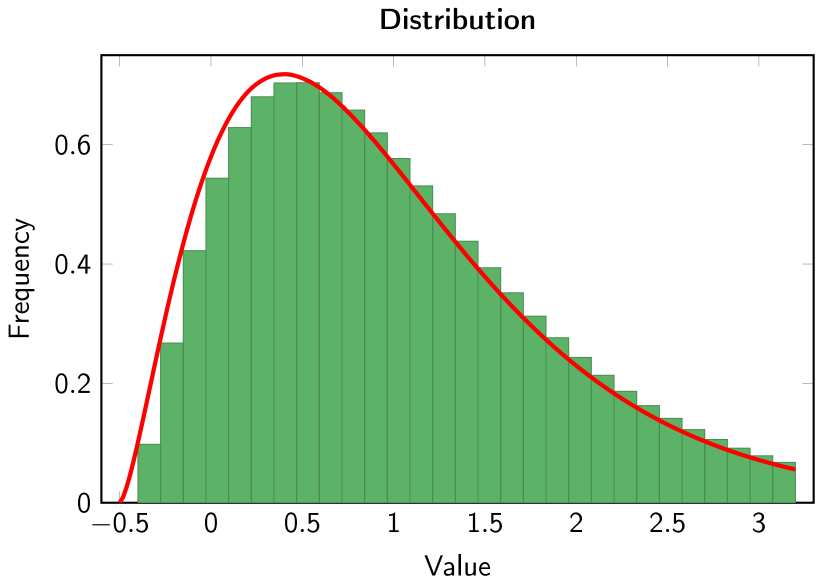
Skewness in Histograms

Is the histogram right skewed, left skewed, or symmetric?

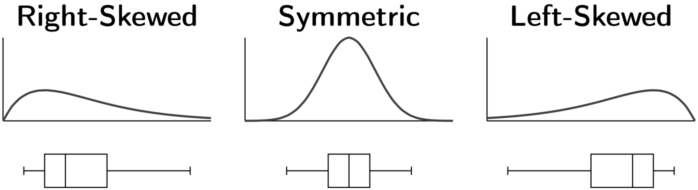


- **Right-skewed (positively skewed) Distribution:**
  - Tail extends to the right (higher values). Most data concentrated on the left. Skewness > 0.
- **Left-skewed (negatively skewed) Distribution:**
  - Tail extends to the left (lower values). Most data concentrated on the right. Skewness < 0.
- **Symmetric Distribution:**
  - Tails roughly equal on both sides. Data evenly distributed around the center. Skewness ≈ 0.

Skewness in Histograms

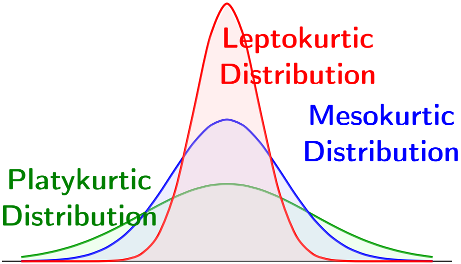


Skewness in Boxplots



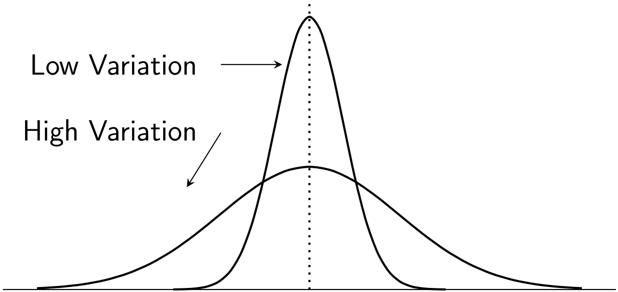
- Boxplots also clearly reflect distribution skewness (Right-skewed, Symmetric, Left-skewed).
- Note: The complete boxplot anatomy will be discussed in subsequent slides.

Kurtosis Visualization



- **Leptokurtic Distribution** (high peak, heavy tails):
  - Most data concentrated near the mean.
  - More extreme values (outliers) likely.
- **Platykurtic Distribution** (low peak, light tails):
  - Data more evenly spread.
  - Fewer extreme values.
- **Mesokurtic Distribution** (moderate peak, normal tails):
  - Moderate concentration of data around the mean.

Variation Visualization



- **Low Variation:** Narrow, tall curve. Data values cluster tightly around the center.
- **High Variation:** Wide, flat curve. Data values are widely dispersed.

Stem-and-Leaf Diagram: Example

- A simple way to see distribution details in a data set.
- **Method:** Separate the sorted data series into leading digits (the stem) and the trailing digits (the leaves).

Data in ordered array:

$$21, 24, 24, 26, 27, 27, 30, 32, 38, 41$$

Completed stem-and-leaf diagram:

Stem	Leaves
2	1 4 4 6 7 7
3	0 2 8
4	1

Exercise 1.32

1.32 Consider the following data:

17, 62, 15, 65, 28, 51, 24, 65, 39, 41, 35, 15, 39, 32, 36, 37, 40, 21, 44, 37, 59, 13, 44, 56, 12

Tasks:

1. Construct a stem-and-leaf display.

Exercise 1.32 d): Solution

Steps for Stem-and-Leaf Solution:

- **Step 1 - Data:** Sorted raw values.
- **Step 2 - Determine stems & leaves:** Tens digit $\rightarrow$ stem, Units digit $\rightarrow$ leaf.
- **Step 3 & 4 - Assign and Order Leaves:**
  - Stem 1: $2, 3, 5, 5, 5, 7$ (Freq: 6)
  - Stem 2: $1, 4, 8$ (Freq: 3)
  - Stem 3: $2, 5, 6, 7, 7, 9, 9$ (Freq: 7)
  - Stem 4: $0, 1, 4, 4$ (Freq: 4)
  - Stem 5: $1, 4, 6, 9, 9$ (Freq: 5)
  - Stem 6: $2, 4, 5, 5$ (Freq: 4)

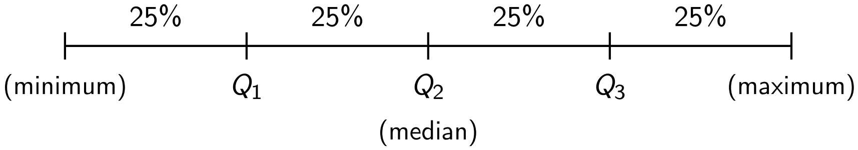
Exercise 1.32 d): Solution

- **Step 5 - Final stem-and-leaf plot with frequencies:**

Stem	Leaf	Frequency
1	2 3 5 5 5 7	6
2	1 4 8	3
3	2 5 6 7 7 9 9	7
4	0 1 4 4	4
5	1 4 6 9 9	5
6	2 4 5 5	4

- **Step 6 - Interpretation:** Most data are concentrated in the 30s. Frequencies highlight dominant stems.

Quantiles: Definition



- **What are Quantiles?** Values that divide a dataset into equal parts.
- **Special cases:**
  - **Quartiles (Q_1, Q_2, Q_3):** Divide data into 4 equal parts (Median $= Q_2$).
  - **Deciles ($D_1, \ldots, D_{10}$):** Divide data into 10 equal parts (Median $= D_5$).
  - **Percentiles ($P_1, \ldots, P_{100}$):** Divide data into 100 equal parts (Median $= P_{50}$).
- **Order-p quantile (q_p):** With $0 < p < 1$, the observed value dividing sample data into values below and above q_p.
  - If $p = 0.9$: You are looking for the value ($q_{0.9}$) where 90% of the data is below it, leaving only the top 10% above.

Percentiles and Quartiles

- To find percentiles and quartiles, data must first be arranged in order from the smallest to the largest values.
- P**th percentile:** A value such that **approximately** $P\%$ of the observations are at or below that number.
  - Percentiles separate large ordered data sets into 100 intervals.
  - The 50th percentile is the median.
- **Position formula:**

$$P\text{th percentile} = \text{value located in the } \left(\frac{P}{100}\right)(n+1)\text{th ordered position}$$

- **Quartiles:** Descriptive measures that separate large data sets into four quarters.
  - **First quartile (Q_1 / 25th percentile):** Separates approximately the smallest 25% of the data from the remainder.
  - **Second quartile (Q_2 / 50th percentile):** The median.

Newbold et al. (2013)

Percentiles and Quartiles

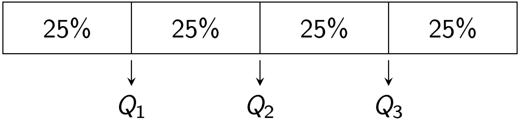
- Percentiles and Quartiles indicate the position of a value relative to the entire set of data.
- Generally used to describe large data sets.
- Example: An IQ score at the 90^{th} percentile means that 10% of the population has a higher IQ score and 90% have a lower IQ score.

$$P^{\text{th}} \text{ percentile} = \text{value located in the } \left(\frac{P}{100}\right)(n+1)^{\text{th}} \text{ ordered position}$$

Newbold et al. (2013)

Quartiles

- Quartiles split the ranked data into 4 segments with an equal number of values per segment (note that the widths of the segments may be different).



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger.
- Q_2 is the same as the median (50% are smaller, 50% are larger).
- Only 25% of the observations are greater than the third quartile.

Newbold et al. (2013)

Quartile Formulas

Find a quartile by determining the value in the appropriate position in the ranked data, where $(n+1)p$:

- First quartile position:

$$\boxed{Q_1 = 0.25(n+1)}$$

- Second quartile position (the median position):

$$\boxed{Q_2 = 0.50(n+1)}$$

- Third quartile position:

$$\boxed{Q_3 = 0.75(n+1)}$$

where n is the number of observed values.

Newbold et al. (2013)

Steps to Calculate an Order-p Quantile (q_p)

1. **Order the sample:** Arrange data in ascending order: $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.
2. **Determine the quantile position:** Compute $(n+1)p$. Let $(n+1)p = r + a$, where r is the integer part and a is the fractional part.
3. **Calculate the quantile:**
   - If $(n+1)p$ is an integer: $q_p = x_{(r)}$
   - If $(n+1)p$ is not an integer: $q_p = (1-a)x_{(r)} + a\,x_{(r+1)}$ (set it to a value between $x_{(r)}$ and $x_{(r+1)}$)

Note: Different software packages (SPSS, Excel) may use alternative formulas yielding minor discrepancies. Choose one method and apply consistently.

Compare the two formulas

- Suppose that $p = 0.9$. $q_{0.9}$ and 90th percentile share the same meaning ($p = P/100$).
- So the percentile formula is

$$\frac{90}{100} \times (n+1) = 0.9 \times (n+1). \tag{1}$$

- If $0.9 \times (n+1)$ is an integer and say it equals to $r = 50$, the 90th percentile is $x_{(50)}$.
- Note that the notation $x_{(r)}$ denotes the r-th value value in the ascending-ordered dataset. (So in general $x_{50} \neq x_{(50)}$).
- However, if the position $r = 50.1$ (not an integer), we take an interpolated value between $x_{(50)}$ and $x_{(51)}$.

Steps to Calculate an Order-p Quantile (q_p) with Example

- **Step 1 – Order the sample**
  - Data: $12, 15, 7, 20, 18$
  - Arrange in ascending order:

$$x_{(1)} = 7,\ x_{(2)} = 12,\ x_{(3)} = 15,\ x_{(4)} = 18,\ x_{(5)} = 20$$

- **Step 2 – Determine the quantile position**
  - Suppose we want the **0.4 quantile** ($p = 0.4$, i.e., the 40th percentile).
  - Compute $(n+1)p = (5+1) \cdot 0.4 = 2.4$
  - Here, $r = 2$ (integer part), $a = 0.4$ (fractional part)
- **Step 3 – Calculate the quantile**
  - Since $(n+1)p$ is not an integer, interpolate:

$$q_{0.4} = (1 - 0.4)x_{(2)} + 0.4\,x_{(3)} = 0.6 \cdot 12 + 0.4 \cdot 15 = 13.2$$

Percentiles and Quartiles

- Indicate the position of a value relative to the entire dataset.
- Generally used to describe large datasets.
- **Pth Percentile formula:** Value located in the position:

$$\left(\frac{P}{100}\right)(n+1)$$

- **Quartiles:** Split ranked data into 4 segments ($Q_1 = 25$th percentile, $Q_2 =$ Median/50th percentile, $Q_3 = 75$th percentile).

Interquartile Range (IQR)

- Measures the spread in the middle 50
- Eliminates extreme high- and low-valued observations.
- **Formula:**

$$\text{IQR} = Q_3 - Q_1$$

- **Note on terminology:** While IQR strictly denotes the numerical difference $Q_3 - Q_1$, the interval $[Q_1, Q_3]$ is sometimes informally described as the interquartile interval containing the middle 50% of observations.

Interquartile Range (IQR): Example

Imagine comparing the daily high temperatures in April for two cities in Portugal:

- Lisbon (Low IQR): Because of the moderating effect of the Atlantic Ocean, the middle 50% of April days in Lisbon might have highs between 18°C and 21°C. The IQR is just 3°C. This means the core spring weather is quite consistent.
- Évora (High IQR): The middle 50% of April days might range from 14°C all the way to 26°C. The IQR is 12°C. This higher IQR shows that the spring weather in Évora is much more variable and spread out.
- If a heat wave occurs in April for 1-2 days, it will not affect the value of IQR, since we only care about the middle 50% of the data.

Five-Number Summary

The five-number summary consists of five descriptive measures arranged in order:

1. Minimum
2. First Quartile (Q_1)
3. Median (Q_2)
4. Third Quartile (Q_3)
5. Maximum

$$\text{minimum} \leq Q_1 \leq \text{median} \leq Q_3 \leq \text{maximum}$$

Exercise 2.6

2.6 During the last 3 years Consolidated Oil Company expanded stations into convenience food stores (CFSs). Daily sales (in hundreds of dollars) from a random sample of 10 weekdays are:

$$6,\ 8,\ 10,\ 12,\ 14,\ 9,\ 11,\ 7,\ 13,\ 11$$

Tasks:

1. Find the five-number summary.

Exercise 2.6): Solution

Step 1 - Order the sample:

$$x_{(1)} = 6,\ x_{(2)} = 7,\ x_{(3)} = 8,\ x_{(4)} = 9,\ x_{(5)} = 10,$$
$$x_{(6)} = 11,\ x_{(7)} = 11,\ x_{(8)} = 12,\ x_{(9)} = 13,\ x_{(10)} = 14$$

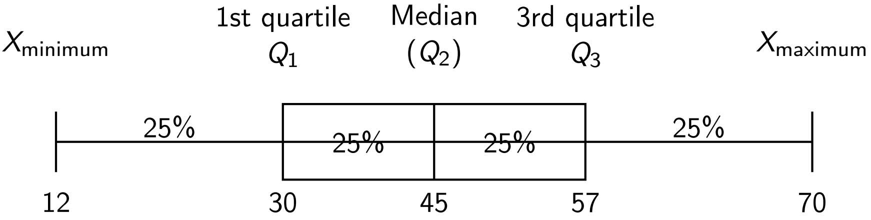
Step 2 - Quantile positions ($n = 10$):

- **Minimum:** 6
- **Maximum:** 14
- Q_1**:** $(10+1)(0.25) = 2.75 \implies Q_1 = 0.25(7) + 0.75(8) = 7.75$
- **Median (Q_2):** $(10+1)(0.5) = 5.5 \implies \text{Median} = 0.5(10) + 0.5(11) = 10.5$
- Q_3**:** $(10+1)(0.75) = 8.25 \implies Q_3 = 0.75(12) + 0.25(13) = 12.25$

Step 3 - Five-number summary:

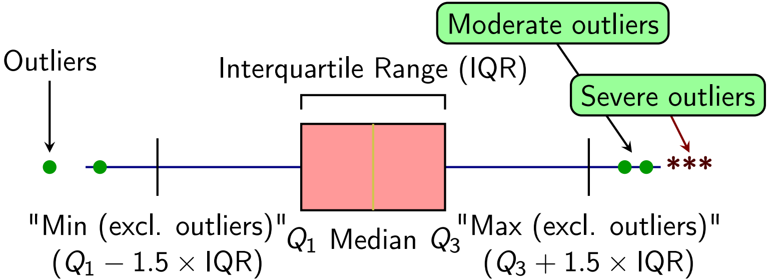
$$\text{Min} = 6, \quad Q_1 = 7.75, \quad \text{Median} = 10.5, \quad Q_3 = 12.25, \quad \text{Max} = 14$$

Box-and-Whisker Plot / Boxplot



- Visualizes the five-number summary graphically.
- Box spans from Q_1 to Q_3 (containing the middle 50% of data), divided by the median line (Q_2).
- Whiskers extend to minimum and maximum values (or specific fence boundaries).
- Can be oriented horizontally or vertically.

Boxplots and Outliers



- **Interquartile Range:** $\text{IQR} = Q_3 - Q_1$
- **Inner Fences (Moderate outlier thresholds):**

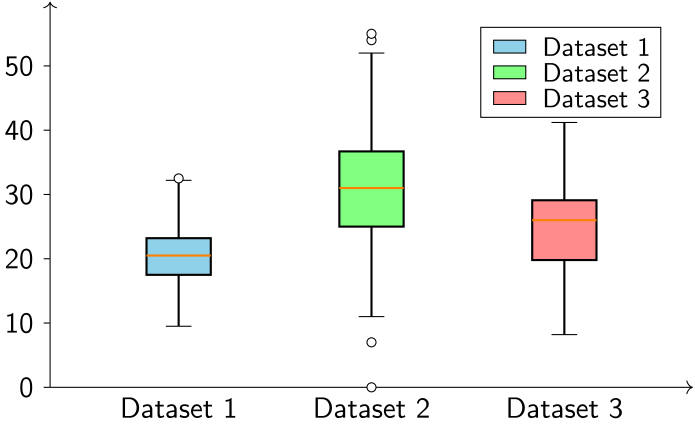
$$Q_1 - 1.5 \times \text{IQR} \quad \text{and} \quad Q_3 + 1.5 \times \text{IQR}$$

  Points beyond these are **moderate outliers** (marked with circles).
- **Outer Fences (Severe outlier thresholds):**

$$Q_1 - 3 \times \text{IQR} \quad \text{and} \quad Q_3 + 3 \times \text{IQR}$$

  Points beyond these are **severe outliers** (marked with asterisks).

Comparing Data Sets



- Use measures of central tendency, dispersion, and shape.
- Visualize with histograms, boxplots, and stem-and-leaf plots.
- Identify patterns, similarities, and differences between data sets.

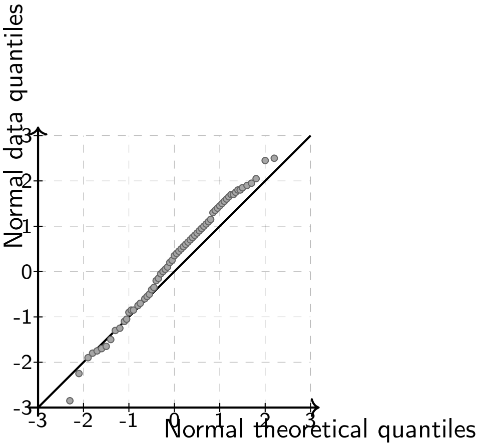
Quantile-Quantile (Q-Q) Plots: Definition & Purpose

- **Definition:** A graphical tool used to compare the distribution of two datasets or check if a dataset follows a theoretical distribution.
- **Purpose:**
  - If data and reference distributions share the same shape, points form a straight line.
  - Deviations reveal structural differences.
- **Common Applications:**
  - Normality checks (comparing sample quantiles to a normal distribution).
  - Distribution comparisons.
  - Outlier detection.

Q-Q Plots: Construction & Interpretation

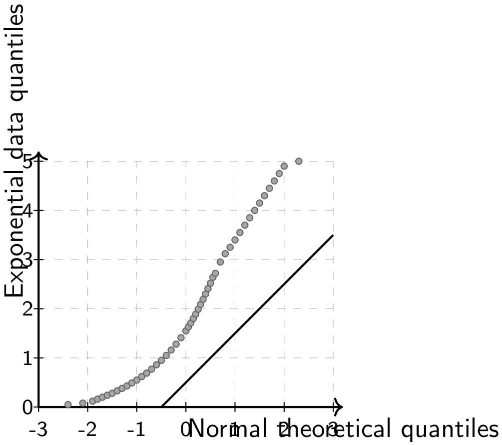
- **Construction Steps:**
  1. Sort the observed data in ascending order.
  2. Calculate corresponding theoretical quantiles.
  3. Plot observed quantiles (y-axis) against theoretical quantiles (x-axis).
- **Interpretation:**
  - Points close to a straight line $\rightarrow$ data follows the reference distribution.
  - Nonlinear patterns or shifts $\rightarrow$ departures from normality, presence of skewness, or differences in mean/variance.

Q-Q plots Examples



- A normal Q–Q plot comparing randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis.
- The linearity of the points suggests that the data are normally distributed.

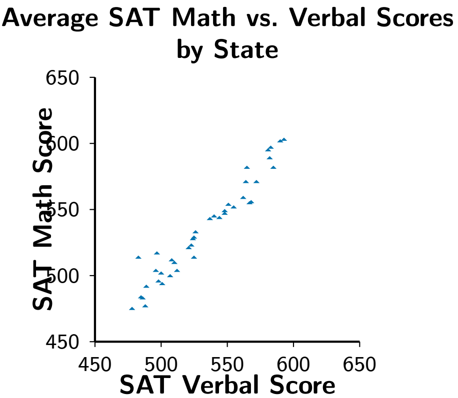
Normal Q–Q Plot for Exponential Data



- A normal Q–Q plot of randomly generated, independent standard exponential data ($X \sim \mathrm{Exp}(1)$).
- The points follow a strongly nonlinear pattern, suggesting non-normality, while the offset indicates a non-zero mean.

Scatter Diagram / Scatterplot

Average SAT scores by state: 1998		
	Verbal	Math
Alabama	562	558
Alaska	521	520
Arizona	525	528
Arkansas	568	555
California	497	516
Colorado	537	542
Connecticut	510	509
Delaware	501	493
D.C.	488	476
Florida	500	501
Georgia	486	482
Hawaii	483	513
. . .		
W.Va.	525	513
Wis.	581	594
Wyo.	548	546

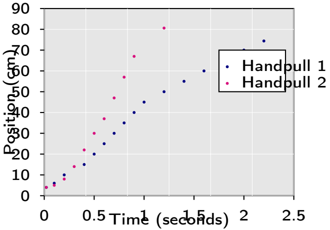


- Used for paired observations taken from two numerical variables.
- One variable is measured on the vertical axis and the other on the horizontal axis.
- Helps identify bivariate relationships, correlations, and clusters (e.g., SAT Math vs. Verbal scores by state).

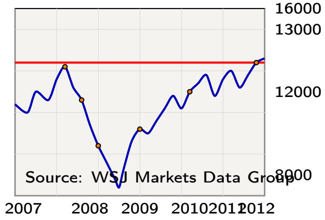
Time Plot

- **Definition:** A graphical representation displaying the temporal sequence or ordering of events.
- **Purpose:**
  - Show data or events in chronological order.
  - Identify trends, patterns, or anomalies over time.
- **Key Features:**
  - Time/sequence on the x-axis, observations on the y-axis.
  - Often displayed as connected dots, especially in financial publications like *The Wall Street Journal*.

Time Plot Examples



The following graph shows a physics-related time plot with the position vs. time for two spark tapes pulled through a spark timer at different constant speeds.



While a time plot can resemble a scatter plot, with a series of dots, you will often see these plots with the dots connected, especially in financial publications like The Wall Street Journal.

Frequency table for discrete and continuous variable

Frequency table for discrete variable: example

1. Discrete Variable Example

Suppose we collected $n = 15$ observations of a discrete variable x:

$$x = (2, 3, 3, 4, 2, 5, 3, 4, 2, 5, 4, 3, 2, 4, 3)$$

Step 1: Identify distinct values

$$x'_1 = 2, \quad x'_2 = 3, \quad x'_3 = 4, \quad x'_4 = 5$$

Step 2: Count frequencies

Absolute Frequencies

$$n_1 = 4, \quad n_2 = 5, \quad n_3 = 4, \quad n_4 = 2$$

Step 3: Compute relative frequencies

Relative Frequencies

$$f_j = \frac{n_j}{n} \implies f_1 = 0.267, \quad f_2 = 0.333, \quad f_3 = 0.267, \quad f_4 = 0.133$$

Step 4: Determine range

$$x_{\min} = 2, \quad x_{\max} = 5, \quad \text{Range} = 3$$

Frequency table for discrete variable: example

Step 5: Table (Frequency Distributions)

x'_j	Absolute Frequencies (n_j)	Relative Frequencies (f_j)
2	4	0.267
3	5	0.333
4	4	0.267
5	2	0.133
Total	**15**	**1.0**

Sample Size

$n = 15$

$$\sum_{j=1}^{m} n_j = n \qquad \sum_{j=1}^{m} f_j = 1$$

Frequency table for discrete variable: example

Suppose we collected $n = 12$ observations of a continuous variable x:

$$x = (1.2, 2.3, 1.8, 2.0, 3.1, 2.7, 3.5, 1.5, 2.8, 3.0, 1.9, 2.5)$$

Step 1: Determine range and class width

$$x_{\min} = 1.2, \quad x_{\max} = 3.5, \quad \text{Range} = 3.5 - 1.2 = 2.3$$

Using Sturges' rule, $m = 4$ classes:

$$\text{Class width} = \frac{\text{Range}}{m} = \frac{2.3}{4} \approx 0.575 \approx 0.6$$

Frequency table for discrete variable: example

Step 2: Construct classes $(I_j = [l_{j-1}, l_j],\ j = 1, 2, \ldots, m)$

$I_j \cap I_k = \emptyset$ (non-overlap) and
$D \subset \bigcup_{j=1}^{m} I_j$ (inclusive)
$D = [x_{\min}, x_{\max}]$

Start at 1.2:

1. $[1.2, 1.8]$
2. $(1.8, 2.4]$
3. $(2.4, 3.0]$
4. $(3.0, 3.6]$

Step 3: Find midpoints

$$x'_1 = \frac{1.2 + 1.8}{2} = 1.5, \quad x'_2 = \frac{1.8 + 2.4}{2} = 2.1, \quad x'_3 = \frac{2.4 + 3.0}{2} = 2.7,$$

$$x'_4 = \frac{3.0+3.6}{2} = 3.3$$

Frequency table for discrete variable: example

Step 4: Count frequencies

- Class $[1.2, 1.8]$: $1.2, 1.5, 1.8 \implies n_1 = 3$
- Class $]1.8, 2.4]$: $1.9, 2.0, 2.3 \implies n_2 = 3$
- Class $]2.4, 3.0]$: $2.5, 2.7, 3.0 \implies n_3 = 3$
- Class $]3.0, 3.6]$: $3.1, 3.5 \implies n_4 = 2$

Step 5: Compute relative frequencies

$$f_1 = \frac{3}{12} = 0.25, \quad f_2 = 0.25, \quad f_3 = 0.25, \quad f_4 = \frac{2}{12} \approx 0.167$$

Step 6: Table (Frequency Distribution with Totals)

Class I_j	Midpoint x'_j	n_j	f_j
$[1.2, 1.8]$	1.5	3	0.25
$]1.8, 2.4]$	2.1	3	0.25
$]2.4, 3.0]$	2.7	3	0.25
$]3.0, 3.6]$	3.3	2	0.167
Total	–	**12**	**1.0**

Cumulative frequency function

Cumulative Absolute Frequency Function

$$N(x) = \begin{cases} 0 & x < x'_1 \\ n_1 & x'_1 \leq x < x'_2 \\ n_1 + n_2 & x'_2 \leq x < x'_3 \\ \dots & \dots \\ n_1 + n_2 + \cdots + n_{m-1} & x'_{m-1} \leq x < x'_m \\ n_1 + n_2 + \cdots + n_m = n & x \geq x'_m \end{cases}$$

- $N(x)$ is a function whose domain is the real line, with $N(-\infty) = 0$, $N(+\infty) = n$ and $0 \leq N(x) \leq n$.
- It is a **non-decreasing function and right-continuous**.
- It is a **step function** where each jump corresponds to an absolute frequency. For example: $N(x'_3) - N(x'_2) = (n_1+n_2+n_3)-(n_1+n_2) = n_3$.

Cumulative Relative Frequency Function

$$F^*(x) = \begin{cases} 0 & x < x_1' \\ f_1 & x_1' \leq x < x_2' \\ f_1 + f_2 & x_2' \leq x < x_3' \\ \dots & \dots \\ f_1 + f_2 + \dots + f_{m-1} & x_{m-1}' \leq x < x_m' \\ f_1 + f_2 + \dots + f_m = 1 & x \geq x_m' \end{cases}$$

- $F^*(x)$ has properties similar to $N(x)$, except with respect to the codomain: $0 \leq F^*(x) \leq 1$.
- $F^*(-\infty) = 0$ and $F^*(+\infty) = 1$.
- It represents the proportion of observations less than or equal to x.

Cumulative frequency function for discrete variable: Example (Data)

Sample information:

- Observations: $x = (2, 3, 3, 4, 2, 5, 3, 4, 2, 5, 4, 3, 2, 4, 3)$
- Number of observations: $n = 15$
- Sample range: Sample range $= x_{\text{max}} - x_{\text{min}} = 5 - 2 = 3$

Frequency Distribution Table (with cumulative frequencies)

x'_j	n_j	f_j	$N(x'_j)$	$F^*(x'_j)$
2	4	0.267	4	0.267
3	5	0.333	9	0.600
4	4	0.267	13	0.867
5	2	0.133	15	1.0
Total	**15**	**1.0**	–	–

Cumulative frequency function for discrete variable: Example (Functions)

Cumulative Absolute Frequency Function

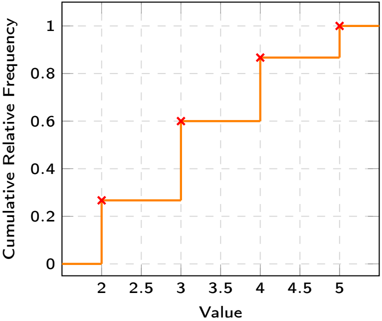
$$N(x) = \begin{cases} 0, & x < 2 \\ 4, & 2 \leq x < 3 \\ 9, & 3 \leq x < 4 \\ 13, & 4 \leq x < 5 \\ 15, & x \geq 5 \end{cases}$$

Cumulative Relative Frequency Function

$$F^*(x) = \begin{cases} 0, & x < 2 \\ 0.267, & 2 \leq x < 3 \\ 0.600, & 3 \leq x < 4 \\ 0.867, & 4 \leq x < 5 \\ 1, & x \geq 5 \end{cases}$$

Cumulative frequency graph for a discrete variable: example

- Here is the graph of the cumulative relative frequency function (ogive) for your sample: Values: $2, 3, 4, 5$, Frequencies: $4, 5, 4, 5$.
- The red marks indicate the cumulative proportions at each value.
- This shows how the cumulative relative frequency increases step by step until it reaches 1 (100%).



Cumulative Relative Frequency Function

$$F^*(l_0) = 0$$

$$F^*(l_1) = f_1$$

$$F^*(l_2) = f_1 + f_2$$

$$\dots$$

$$F^*(l_{m-1}) = f_1 + f_2 + \cdots + f_{m-1}$$

$$F^*(l_m) = f_1 + f_2 + \cdots + f_m = 1$$

It is assumed that the frequencies are **uniformly distributed within each class** and that the cumulative frequencies refer to the **class boundaries**.

Frequency cumulative function for continuous variable: Example (Data)

Sample information:

- Observations: $x = (1.2, 2.3, 1.8, 2.0, 3.1, 2.7, 3.5, 1.5, 2.8, 3.0, 1.9, 2.5)$
- Number of observations: $n = 12$
- Sample range: Sample range $= x_{\text{max}} - x_{\text{min}} = 3.5 - 1.2 = 2.3$
- Classes (Sturges, width ≈ 0.6): $[1.2, 1.8],]1.8, 2.4],]2.4, 3.0],]3.0, 3.6]$
- Midpoints: $1.5, 2.1, 2.7, 3.3$

Frequency Distribution Table (with cumulative frequencies)

Class	l_j	n_j	f_j	$N(x_j)$	$F^*(x_j)$
$[1.2, 1.8]$	1.2	3	0.25	3	0.25
$]1.8, 2.4]$	1.8	3	0.25	6	0.50
$]2.4, 3.0]$	2.4	3	0.25	9	0.75
$]3.0, 3.6]$	3.0	3	0.25	12	1.00
Total	–	**12**	**1.00**	–	–

Frequency cumulative function for continuous variable: Example (Functions)

Cumulative Absolute Frequency Function

$$N(x) = \begin{cases} 0, & x < 1.2 \\ 3, & 1.2 \leq x < 1.8 \\ 6, & 1.8 \leq x < 2.4 \\ 9, & 2.4 \leq x < 3.0 \\ 12, & x \geq 3.0 \end{cases}$$

Cumulative Relative Frequency Function

$$F^*(x) = \begin{cases} 0, & x < 1.2 \\ 0.25, & 1.2 \leq x < 1.8 \\ 0.50, & 1.8 \leq x < 2.4 \\ 0.75, & 2.4 \leq x < 3.0 \\ 1.0, & x \geq 3.0 \end{cases}$$

Cumulative Frequency Graph for a Continuous Variable: Properties

Ogive (Cumulative Frequency Graph)

Where does it start?

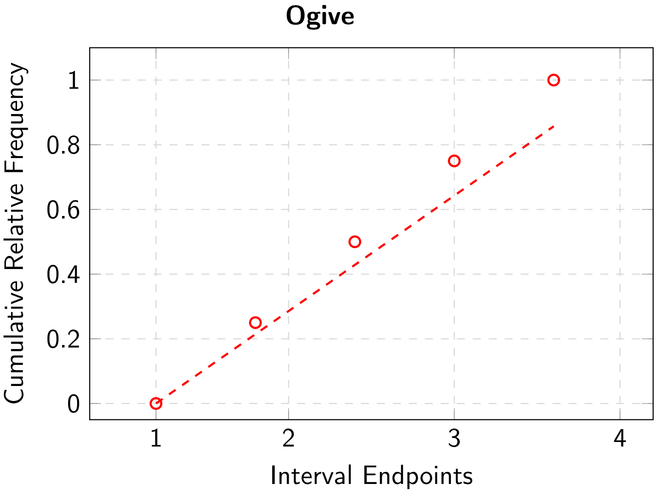
- It starts at the lower limit of the first class, with cumulative frequency $= 0$.
- The graph is built using the upper class limits on the x-axis and the cumulative frequencies (or percentages) on the y-axis.

What is it used for?

- To show how data accumulate across classes.
- To identify the median, quartiles, deciles, and percentiles.
- To compare cumulative distributions.

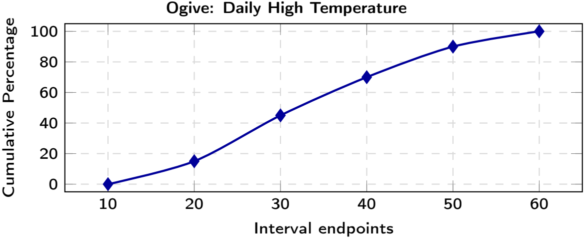
In this case, the ogive appears as a straight line rather than a curve, because the absolute and relative frequencies have the same values for all classes.

Cumulative Frequency Graph for a Continuous Variable: Example



Cumulative Relative Frequency Function

Interval	Upper endpoint	Cumulative Percentage
Less than 10	10	0
10 but less than 20	20	15
20 but less than 30	30	45
30 but less than 40	40	70
40 but less than 50	50	90
50 but less than 60	60	100



Newbold et al. (2013)

THANKS!

Questions?