Knowledge Discovery & Data Mining
— Data Exploration: Data Visualization—

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

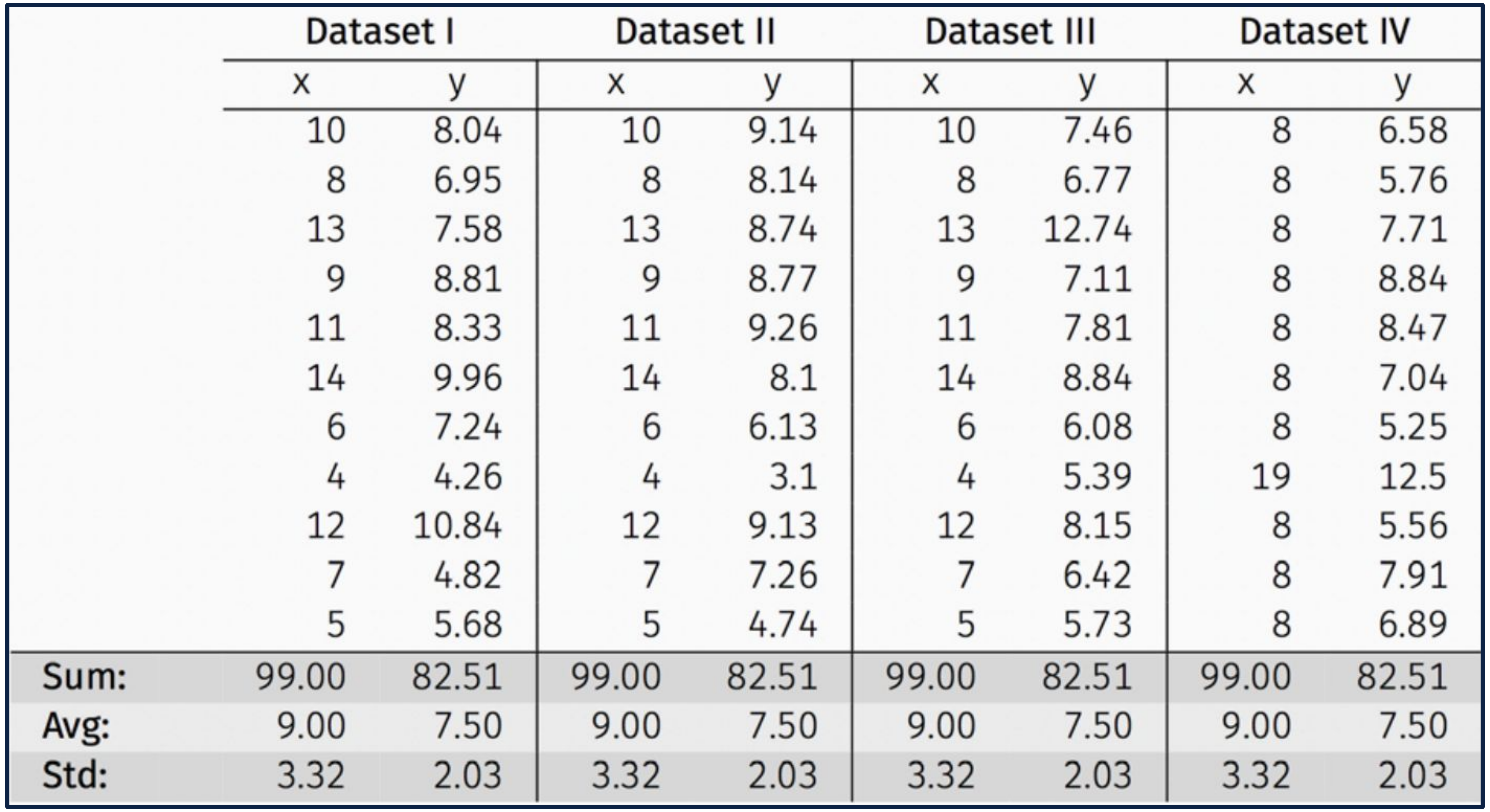
Based on the original version by Professor Yizhou Sun

Outline

- Data Visualization
  - Quantile-Quantile (Q-Q) plot
  - Histograms
  - Pie chart
  - Scatter plots

Anscombe's Quartet

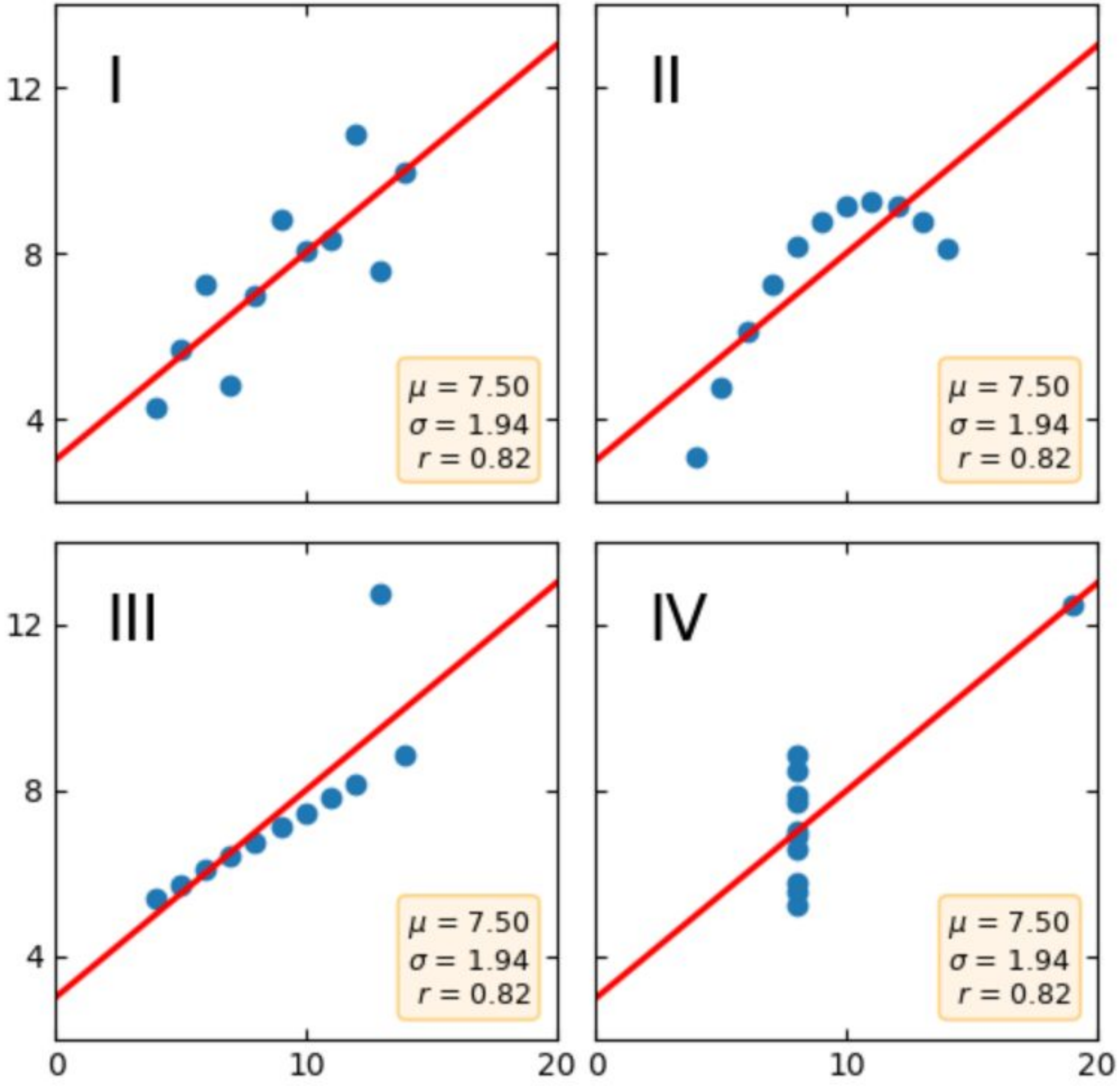
- The following four data sets comprise the Anscombe’s Quartet; all four sets of data have identical simple summary statistics.



	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

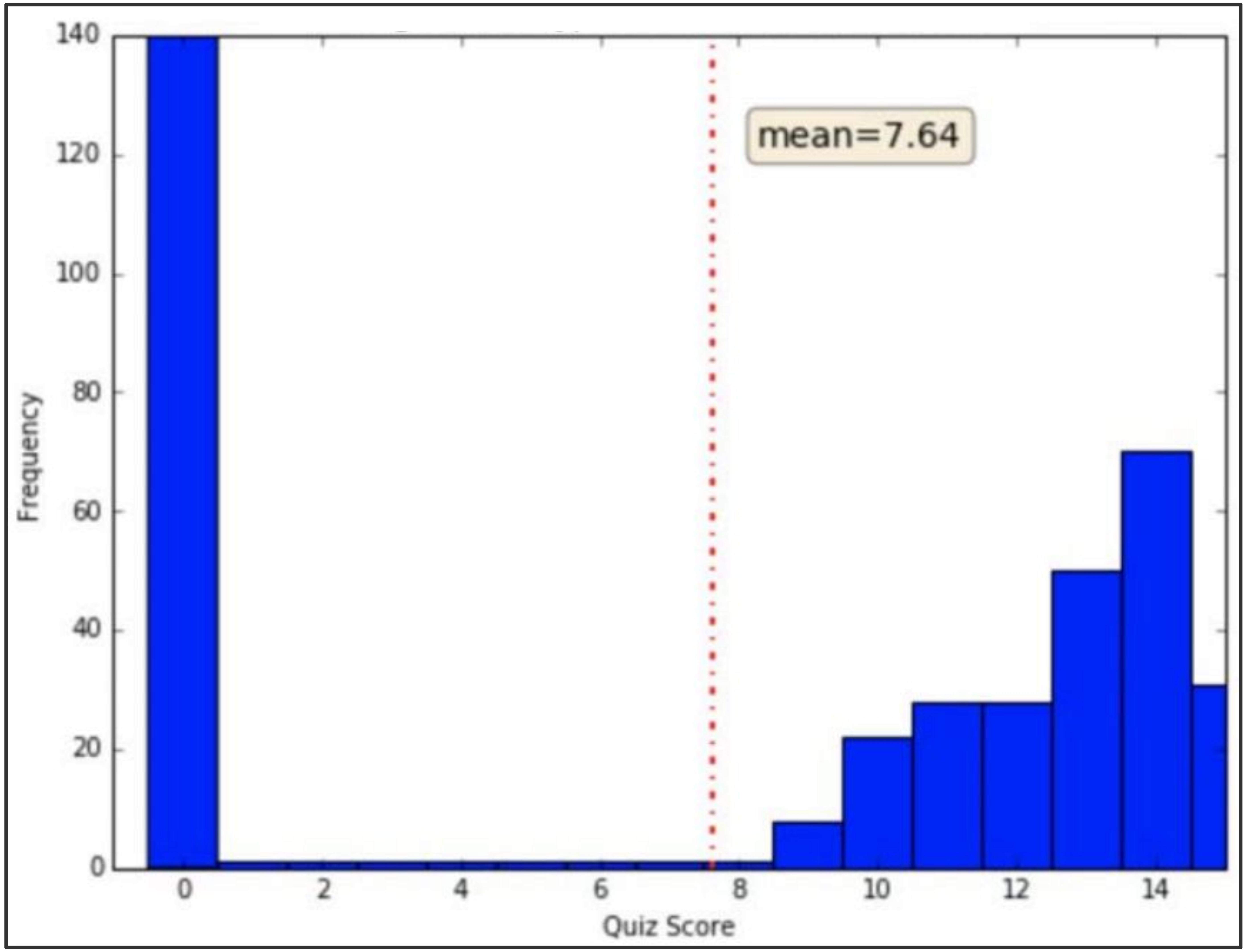
Anscombe's Quartet

- Summary statistics clearly don’t tell the story of how they differ.
- A picture can be worth a thousand words.



More Visualization Motivation

- If I tell you that the average score for a Homework is: 7.64/15 = 50.9%, what does that suggest?



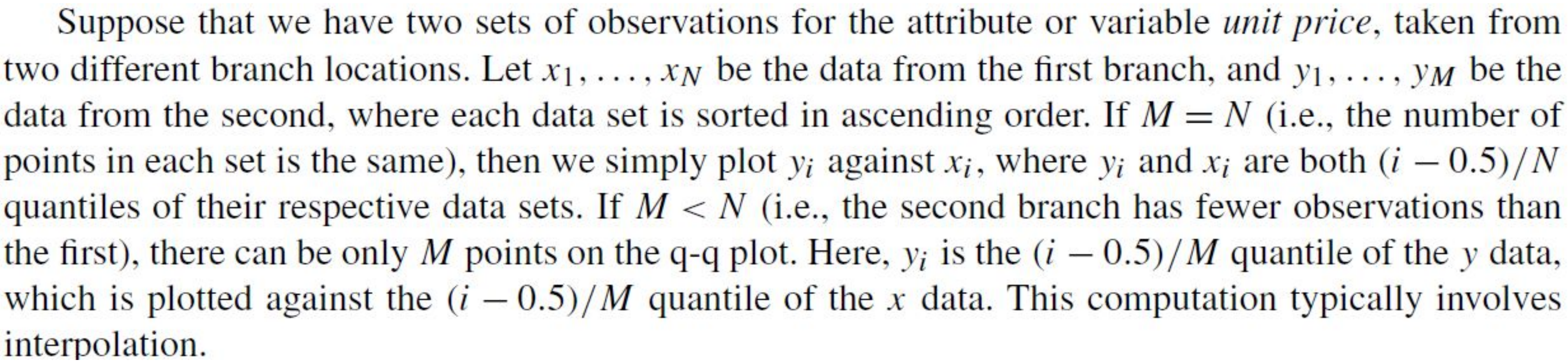
- And what does this graph suggest?

Types of Visualizations

- What do you want your visualization to show about your data?
  - **Distribution:** how a variable or variables in the dataset distribute over a range of possible values.
  - **Relationship:** how the values of multiple variables in the dataset relate
  - **Composition:** how the dataset breaks down into subgroups
  - **Comparison:** how trends in multiple variable or datasets compare

Quantile-Quantile (Q-Q) plot

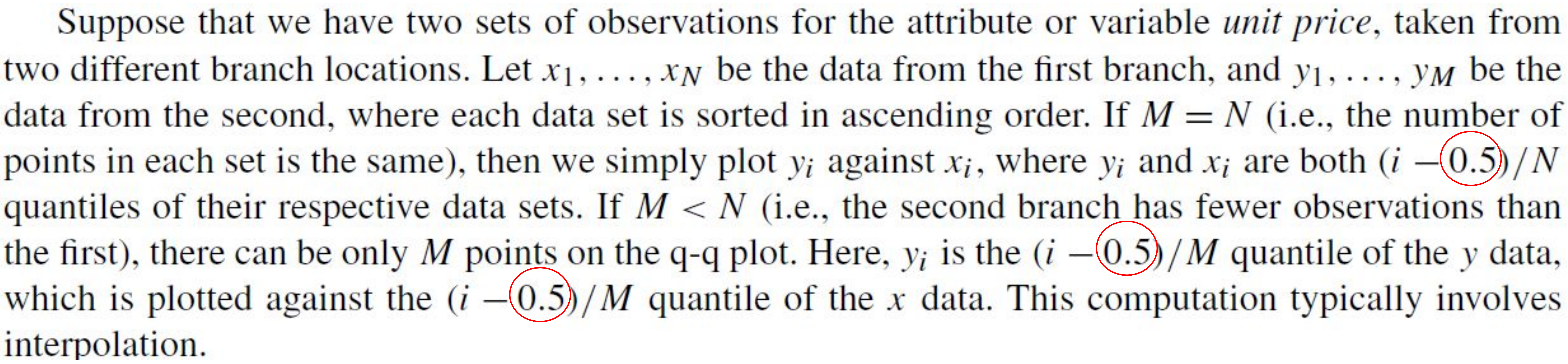
Quantile-Quantile plot: or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another.



Suppose that we have two sets of observations for the attribute or variable *unit price*, taken from two different branch locations. Let $x_1, \ldots, x_N$ be the data from the first branch, and $y_1, \ldots, y_M$ be the data from the second, where each data set is sorted in ascending order. If $M = N$ (i.e., the number of points in each set is the same), then we simply plot y_i against x_i, where y_i and x_i are both $(i - 0.5)/N$ quantiles of their respective data sets. If $M < N$ (i.e., the second branch has fewer observations than the first), there can be only M points on the q-q plot. Here, y_i is the $(i - 0.5)/M$ quantile of the y data, which is plotted against the $(i - 0.5)/M$ quantile of the x data. This computation typically involves interpolation.

Quantile-Quantile (Q-Q) plot

Quantile-Quantile plot: or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

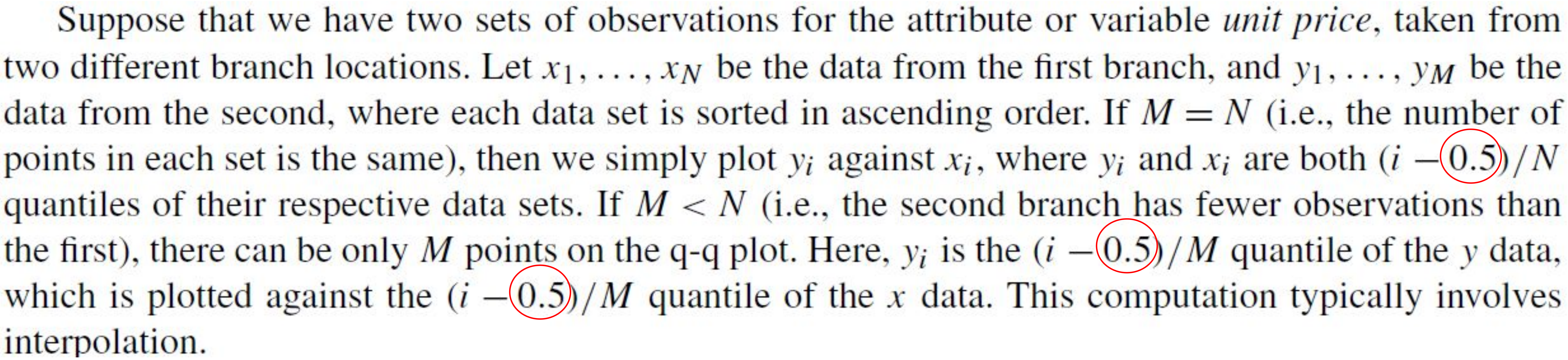


Suppose that we have two sets of observations for the attribute or variable *unit price*, taken from two different branch locations. Let $x_1, \ldots, x_N$ be the data from the first branch, and $y_1, \ldots, y_M$ be the data from the second, where each data set is sorted in ascending order. If $M = N$ (i.e., the number of points in each set is the same), then we simply plot y_i against x_i, where y_i and x_i are both $(i - 0.5)/N$ quantiles of their respective data sets. If $M < N$ (i.e., the second branch has fewer observations than the first), there can be only M points on the q-q plot. Here, y_i is the $(i - 0.5)/M$ quantile of the y data, which is plotted against the $(i - 0.5)/M$ quantile of the x data. This computation typically involves interpolation.

Why is the subtraction of 0.5 needed when calculating the quantiles?

Quantile-Quantile (Q-Q) plot

Quantile-Quantile plot: or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another.



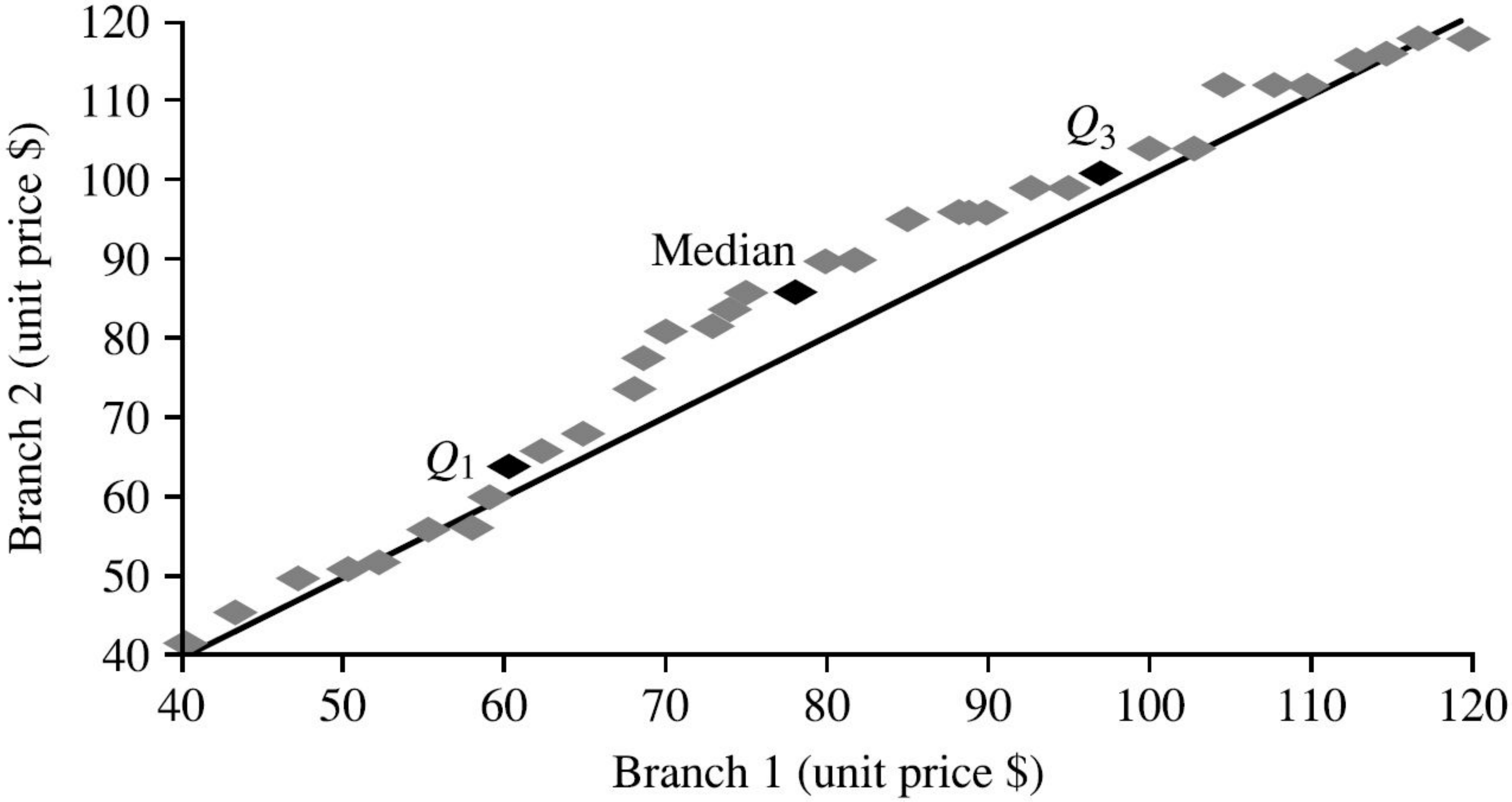
Suppose that we have two sets of observations for the attribute or variable *unit price*, taken from two different branch locations. Let $x_1, \ldots, x_N$ be the data from the first branch, and $y_1, \ldots, y_M$ be the data from the second, where each data set is sorted in ascending order. If $M = N$ (i.e., the number of points in each set is the same), then we simply plot y_i against x_i, where y_i and x_i are both $(i - 0.5)/N$ quantiles of their respective data sets. If $M < N$ (i.e., the second branch has fewer observations than the first), there can be only M points on the q-q plot. Here, y_i is the $(i - 0.5)/M$ quantile of the y data, which is plotted against the $(i - 0.5)/M$ quantile of the x data. This computation typically involves interpolation.

Why is the subtraction of 0.5 needed when calculating the quantiles?

center the data

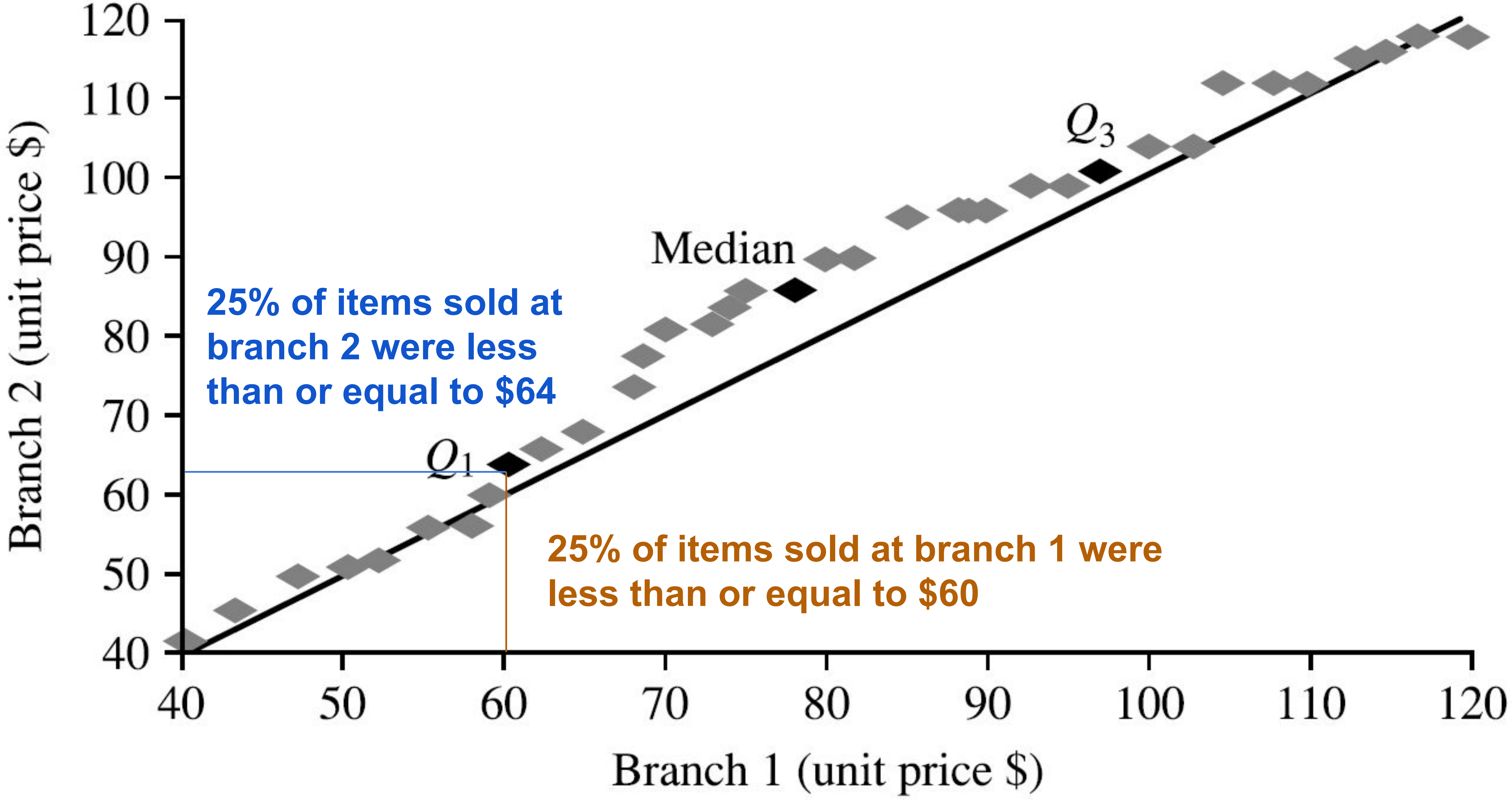
Quantile-Quantile (Q-Q) plot

Example. The following figure shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile.



Quantile-Quantile (Q-Q) plot

Example. The following figure shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile.

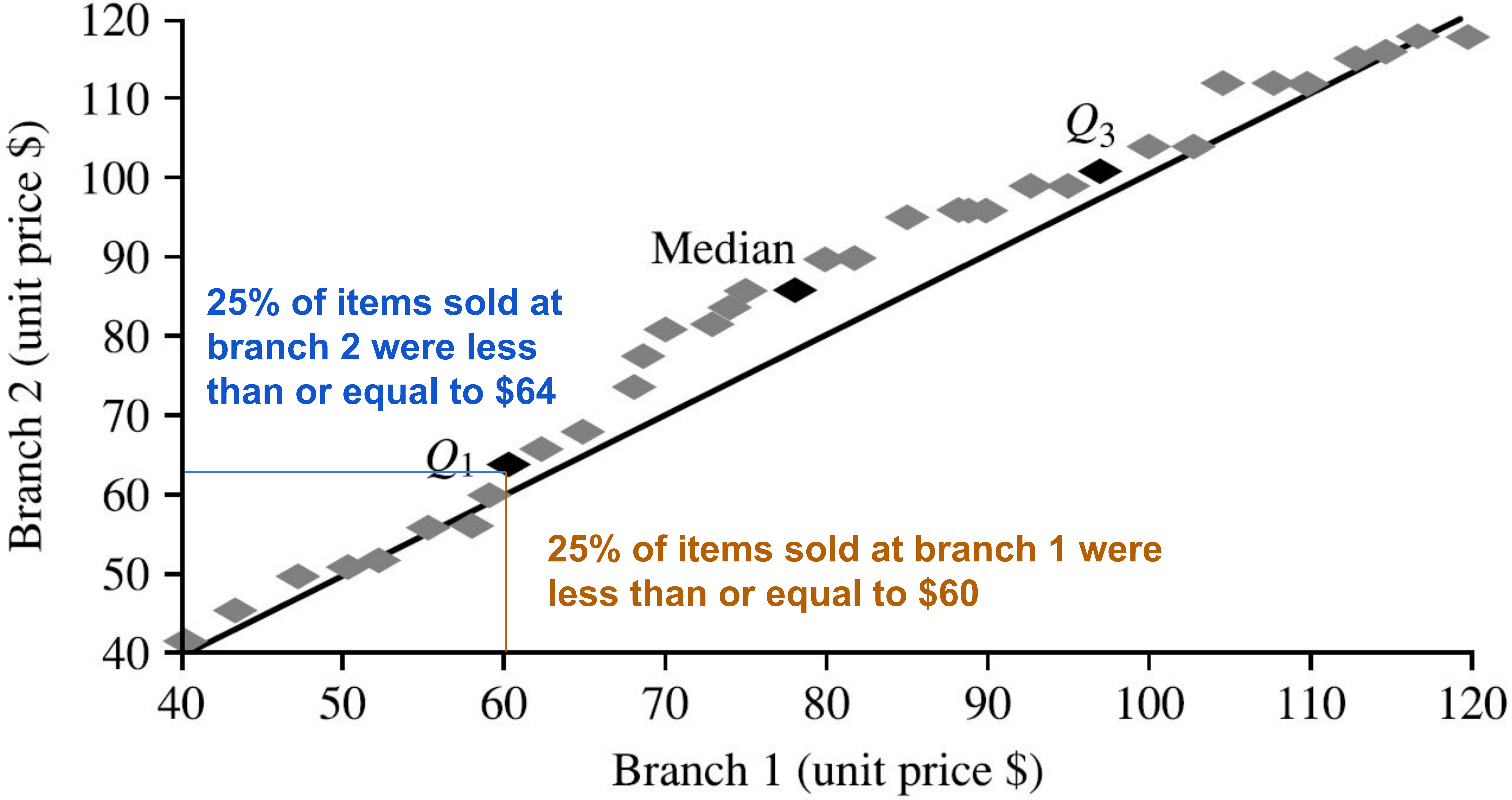


25% of items sold at branch 2 were less than or equal to $64

25% of items sold at branch 1 were less than or equal to $60

Quantile-Quantile (Q-Q) plot

Example. The following figure shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. **Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.**

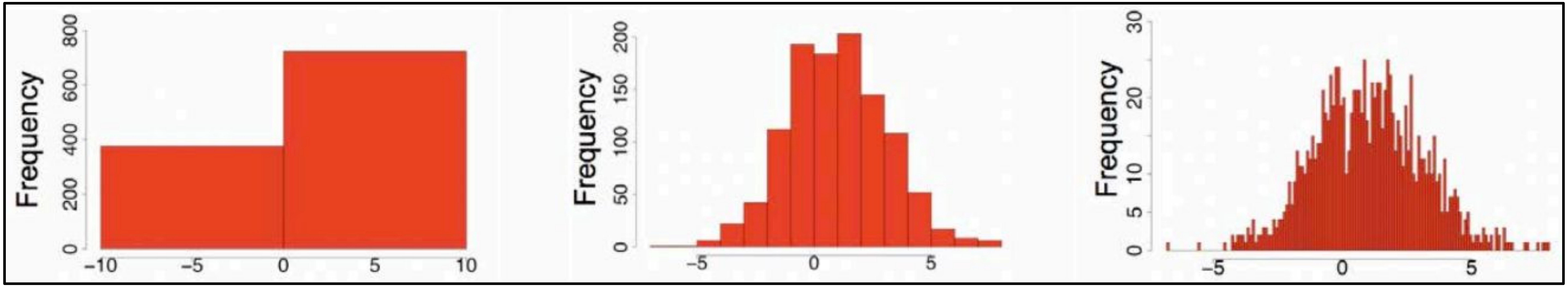


25% of items sold at branch 2 were less than or equal to $64

25% of items sold at branch 1 were less than or equal to $60

Histograms to visualize distribution

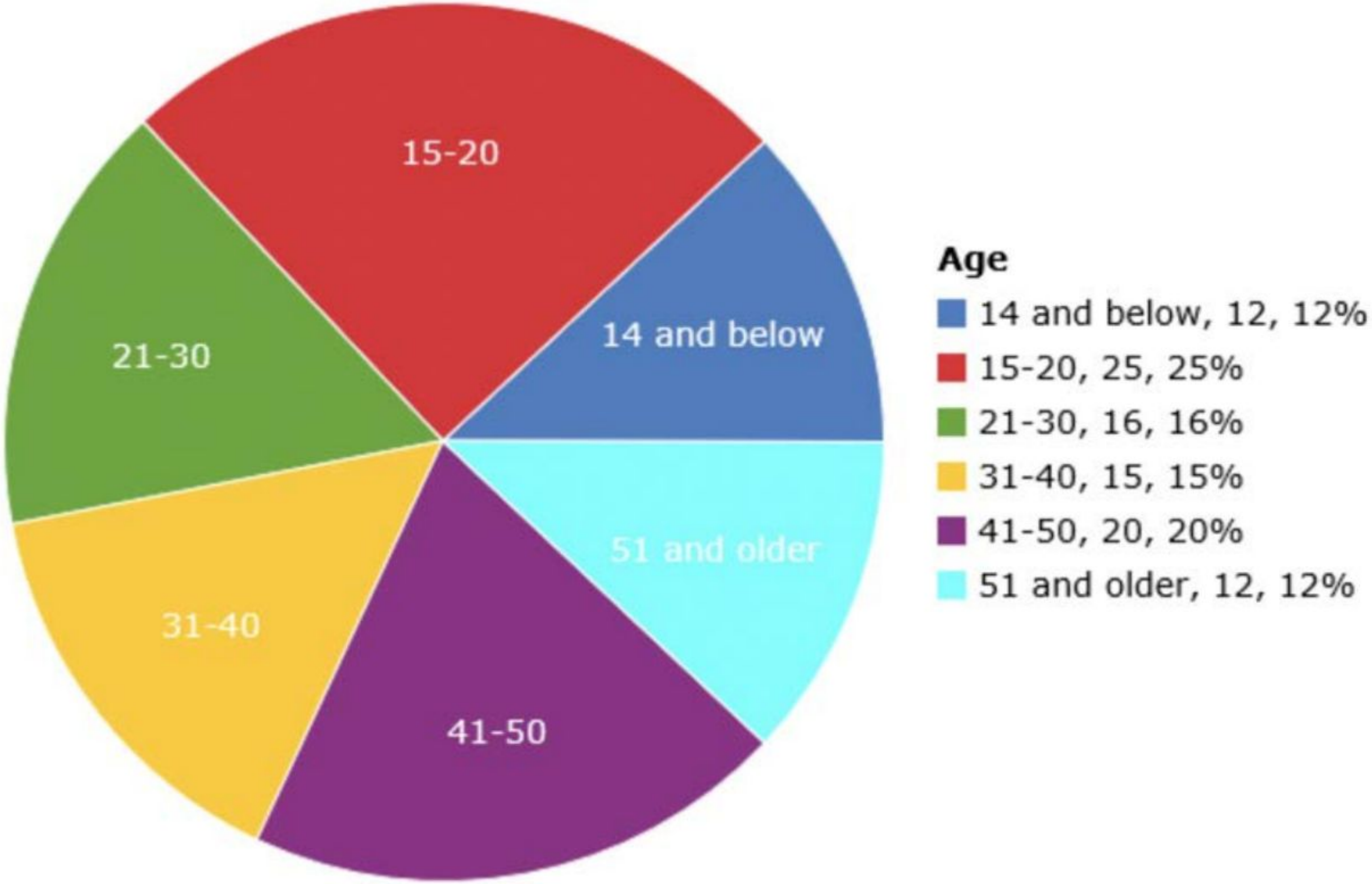
- A **histogram** is a way to visualize how 1- dimensional data is distributed across certain values.



- Note: Trends in histograms are sensitive to number of bins.

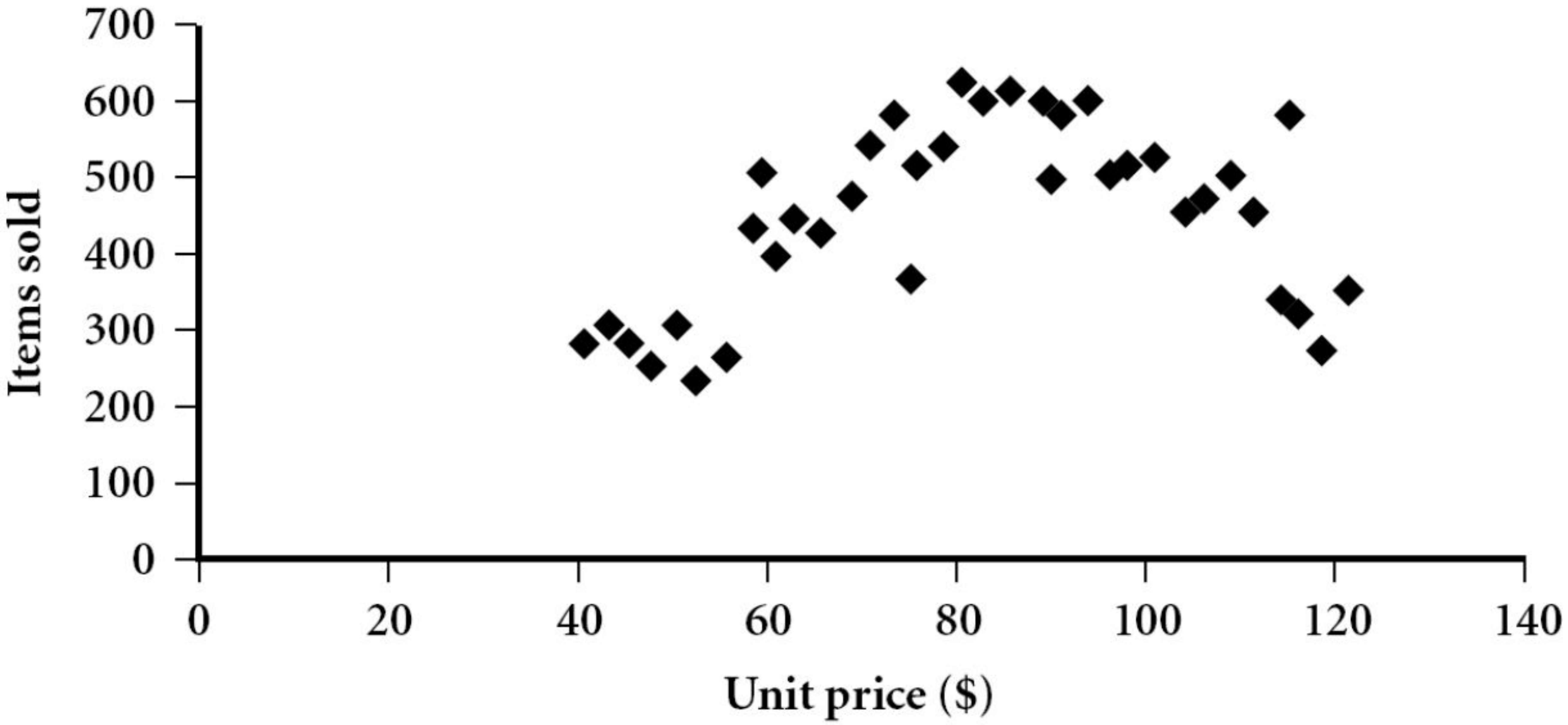
Pie chart for a categorical variable

- A **pie chart** is a way to visualize the static composition (aka, distribution) of a variable (or single group).



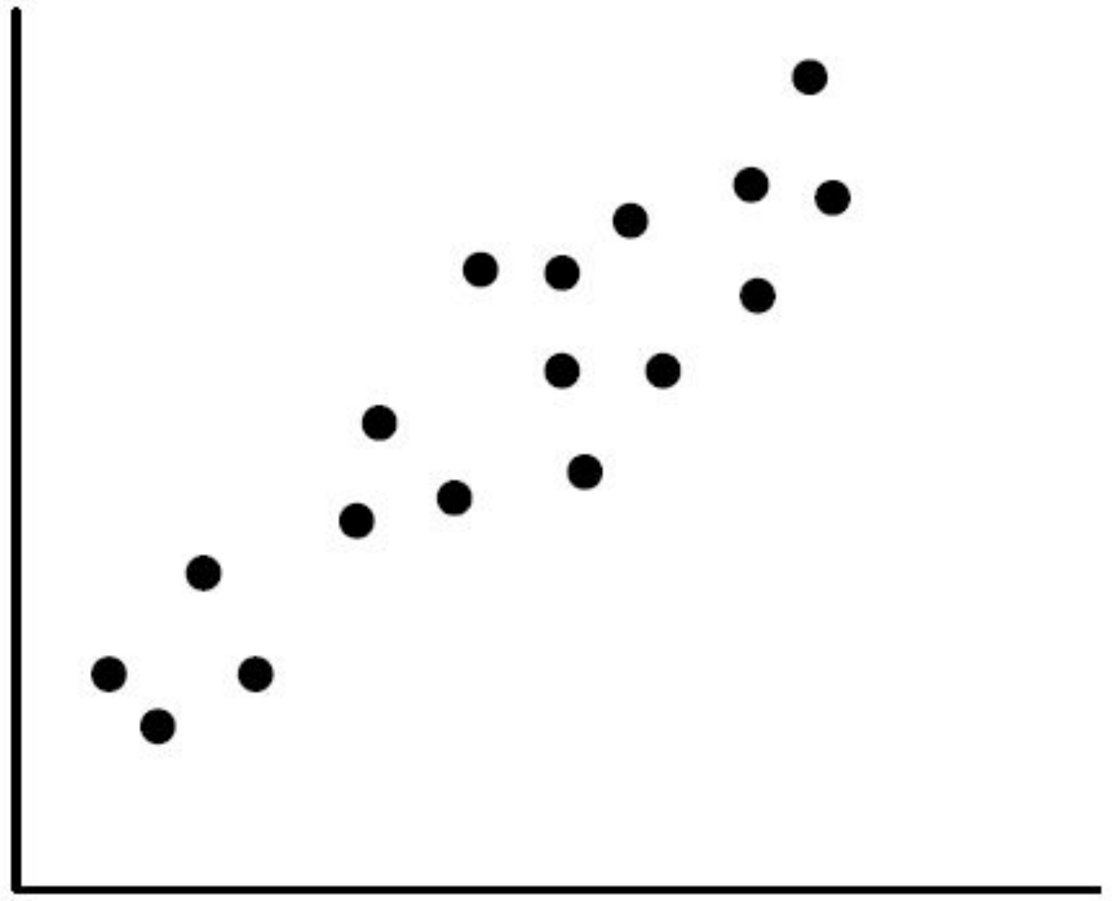
Scatter plots to visualize relationships

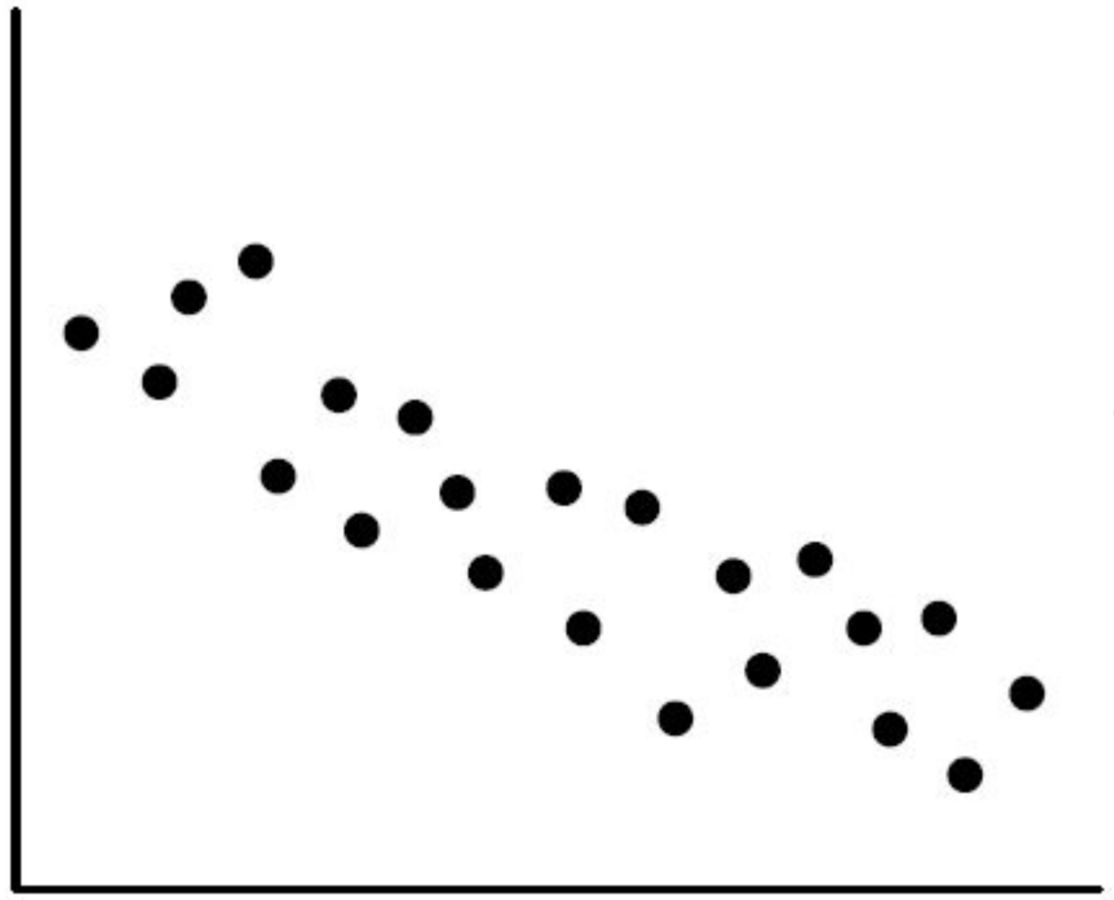
- A **scatter plot** is a way to visualize the relationship between two different attributes of multi-dimensional data.
  - Provides a first look at bivariate data to see clusters of points, outliers, or to explore the possibility of correlation relationships
  - Each pair of values is treated as a pair of coordinates and plotted as points in the plane



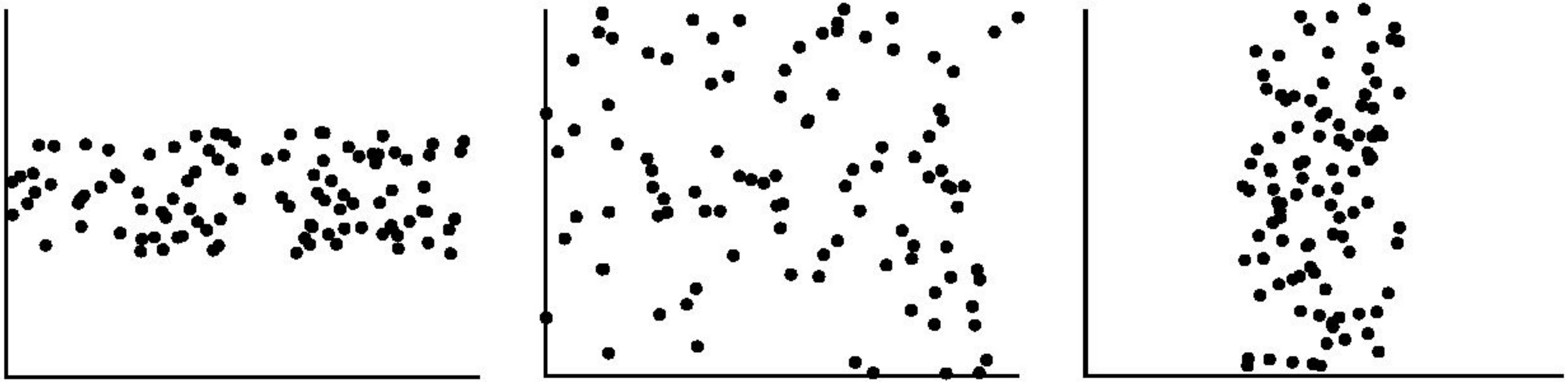
Positively and Negatively Correlated Data

Positive correlation





Negative correlation



No observed correlation

Summary

- Data Visualization
    - Quantile-Quantile (Q-Q) plot
    - Histograms
    - Pie chart
    - Scatter plots