Visualising Distribution

NANYANG THE INNOVATIVE POLYTECHNIC School of Information Technology

Learning Outcomes

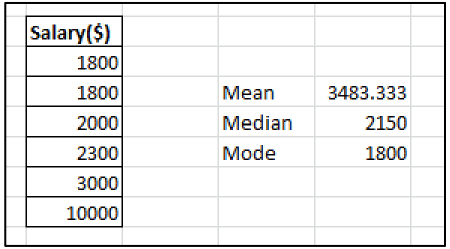
By the end of this lesson, you should be able to

- Identify the patterns of distribution
- Use and explain the various data comparing approaches
- Explain and apply the techniques and best practices used

Introduction to Distribution

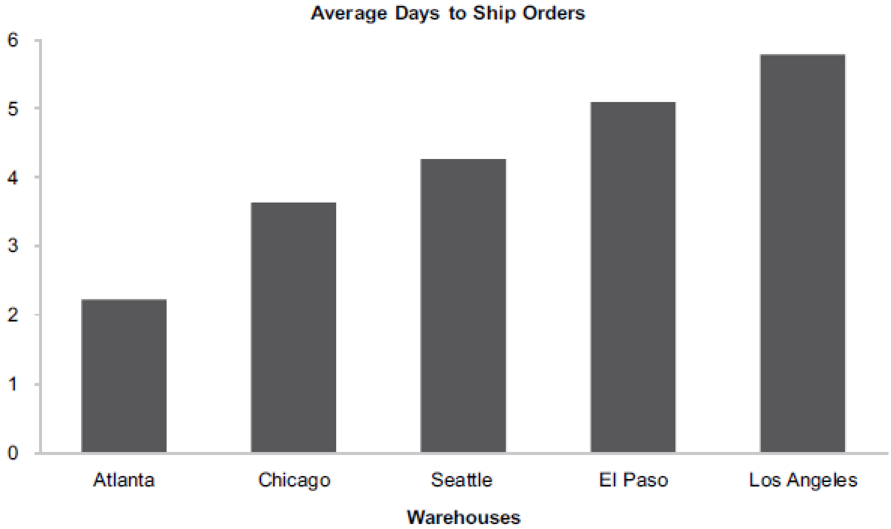
What is Distribution?

- Still remember mean, median and mode?
  - Mean – sum of all data points divided by the total number of points
  - Median – order data from least to greatest and mark the halfway point
  - Mode – number that occurs the most
  - They describe how parts of your data are distributed
- BUT you are not looking at the full distribution.



Salary($)			
1800			
1800		Mean	3483.333
2000		Median	2150
2300		Mode	1800
3000			
10000			

Introduction to Distribution

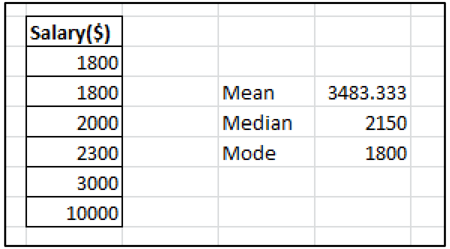


With an average shipment timeline of 4.2 days, the Seattle warehouse could be keeping some customers waiting 10 days or more, but this fact would remain hidden in the graph above.

Introduction to Distribution

What is Distribution?

- Examining sets of quantitative values to see how the values are distributed from lowest to highest
- Compare and contrast how multiple sets of values are distributed



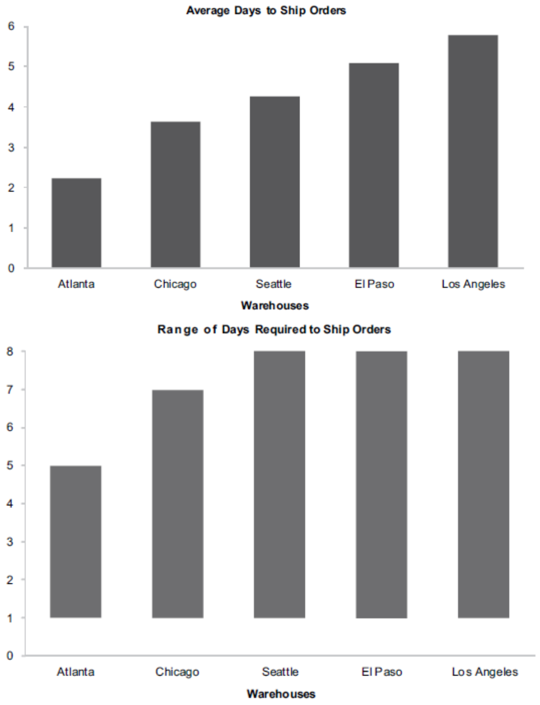
Salary($)			
1800			
1800		Mean	3483.333
2000		Median	2150
2300		Mode	1800
3000			
10000			

1800	2
2000	1
2300	1
3000	1
10000	1

Introduction to Distribution

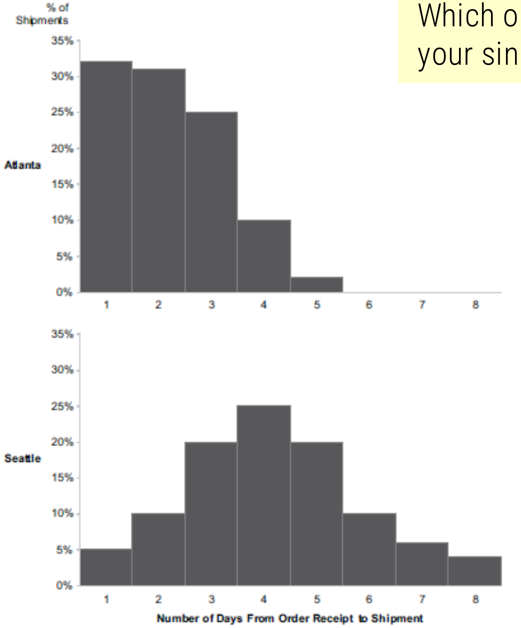
- Average days to Ship Orders tells us nothing about variability – Atlanta : 2.1 days; Seattle : 4.2 days

- Range of days tells us the spread but we have no idea of the distributions – Atlanta is 1-5 days, Seattle is 1-8 days



Introduction

- With distribution, we can tell that most orders are shipped from Atlanta on the same day, with a decreasing number of shipments as days increases
- Seattle has symmetrical distribution and greatest percentage on the fourth day



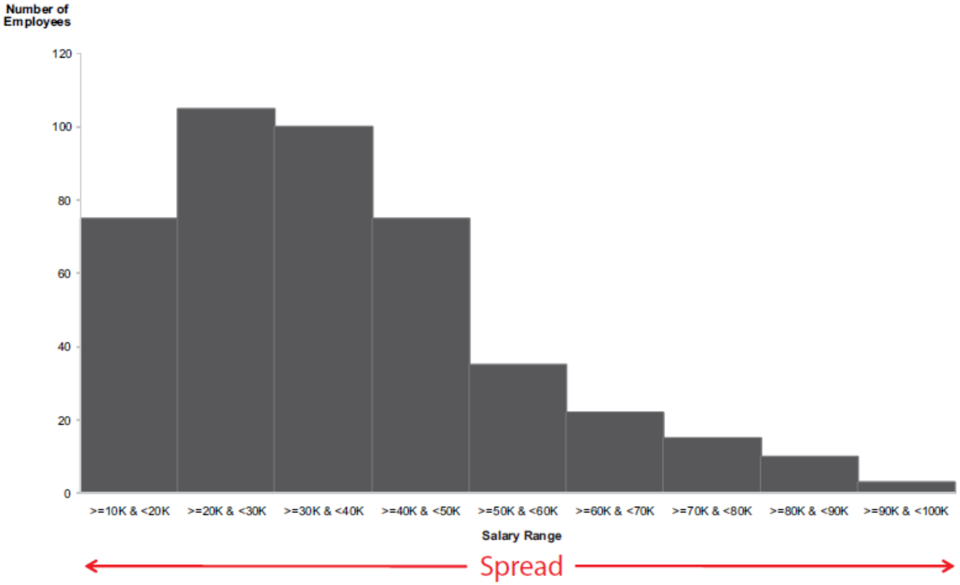
Which one will you choose for your singles' day's order?

Describing Distribution

- Visual Characteristic of Distributions
  - Spread
  - Center
  - Shape
- Statistical Summaries of Distributions
  - 3-value summary of distribution
  - 5-value summary of distribution

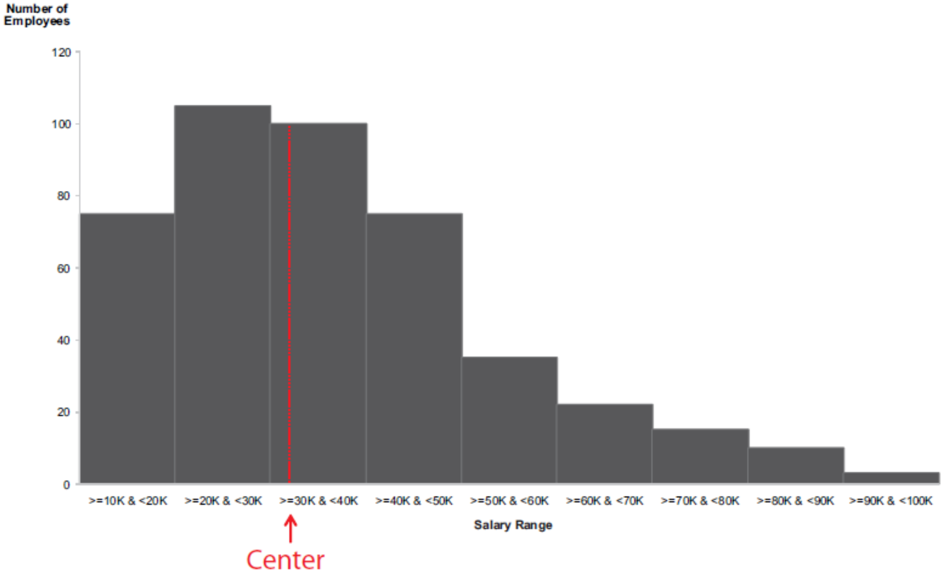
Visual Characteristic

- Spread – lowest value, highest value and distance between them



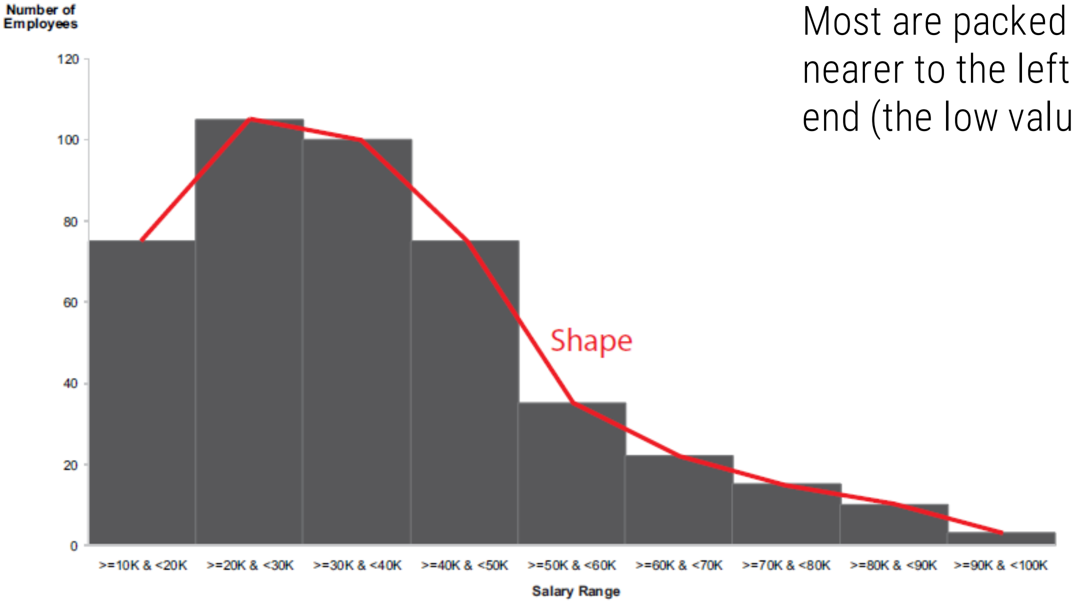
Visual Characteristic

- Center – estimate of middle set of values or value that is most typical



Visual Characteristic

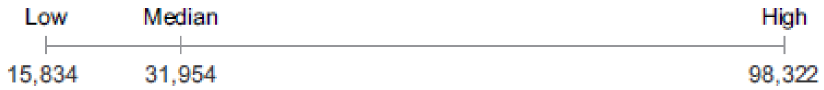
- Shape – where the values are located



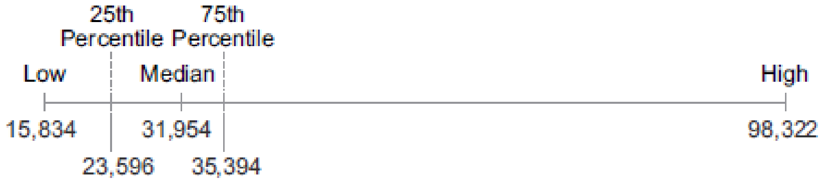
Most are packed nearer to the left end (the low values)

Statistical Summaries

- 3-value summary of distribution



- 5-value summary of distribution

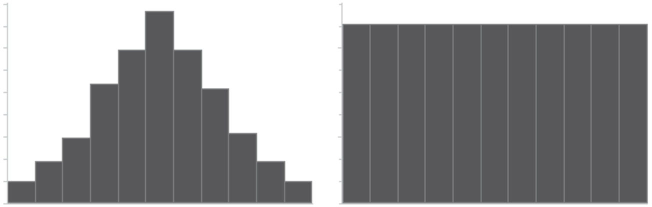


Distribution Patterns

- Shape
  - Curved or flat?
  - If curved, upward or downward?
  - If curved upward, single or multiple peaked?
  - If single peaked, symmetrical or skewed?
  - Concentrations?
  - Gaps?
- Outliers

Distribution Patterns

- Curved or flat?

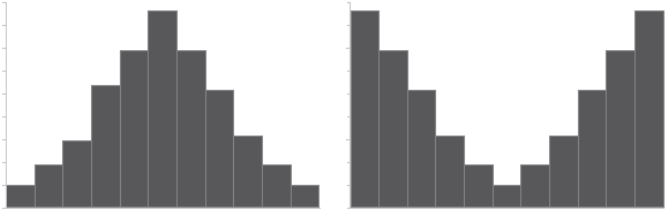


Curved

Flat

Upward : Number of items or frequency begins relatively low, increases to a peak and then decreases until relatively low. E.g., IQ distribution across population

- If curved, upward or downward?



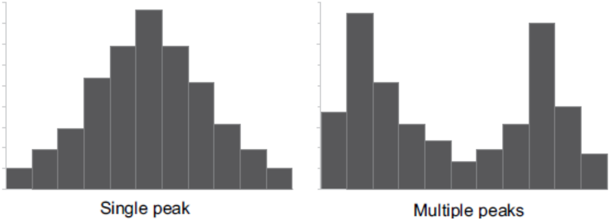
Upward

Downward

Downward – less common. E.g., amount of leisure time people enjoy throughout their lives ?

Distribution Patterns

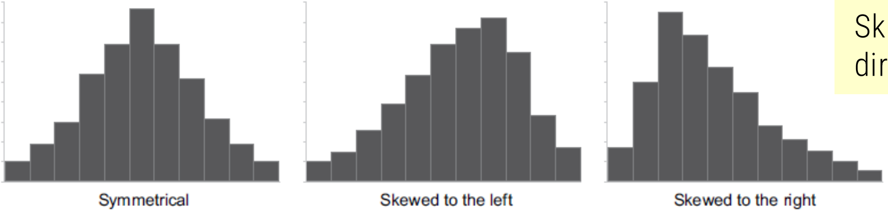
- If curved upward, one or two peaks?



Bi-modal

- If single peaked, symmetrical or skewed?

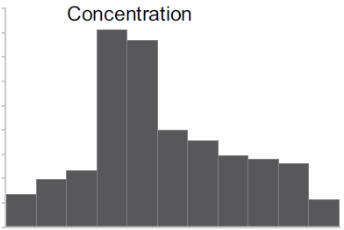
Normal distribution or bell-shaped

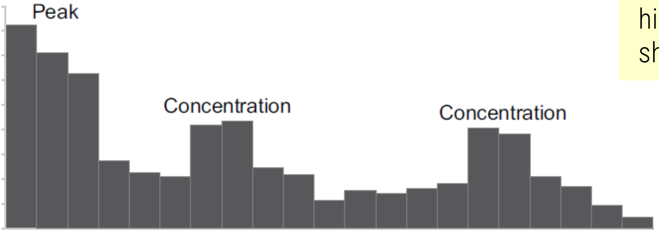


Skew refers to direction of the tail

Distribution Patterns

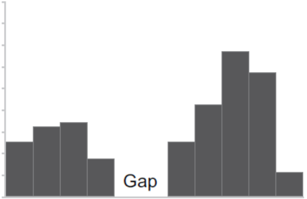
- Concentrations?





Predominant peak on the left, distribution is skewed to the right but there are also high concentration near middle and end – should investigate

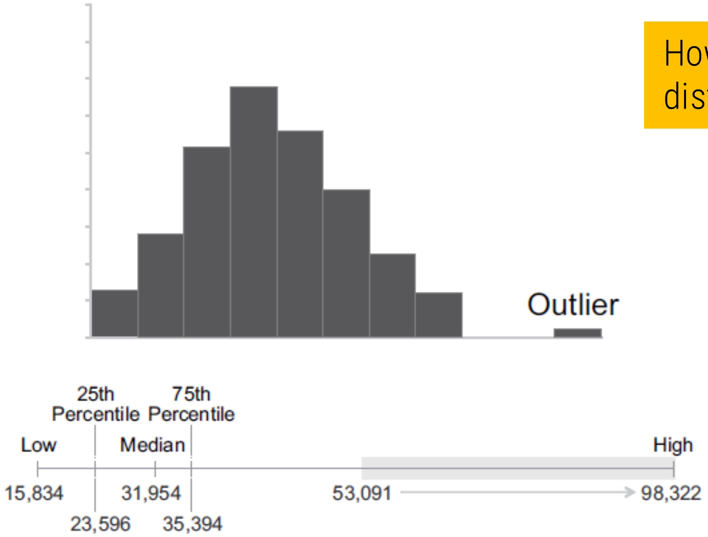
- Gaps?



No sales to the age groups near the middle of the distribution, why?

Distribution Patterns

Outliers



How to identify outliers in distribution?

Good rule of thumb : Taking the mid-spread , the distance between the 25th and 75th percentiles – multiplying it by 1.5 and adding it to 75th percentile to mark the upper threshold

Distribution Displays

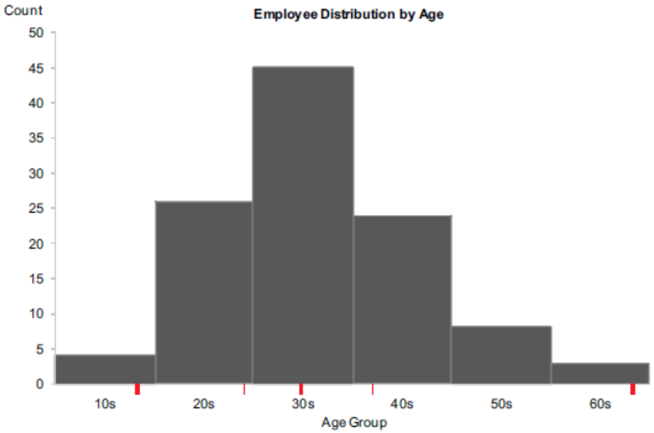
- Single Distribution Displays
  - Histograms
  - Frequency Polygons
  - Strip Plots
  - Stem-and-Leaf Plots

- Multiple Distribution Displays
  - Box Plots
  - Multiple strip plots
  - Frequency polygons
  - Distribution deviation graphs

Histograms

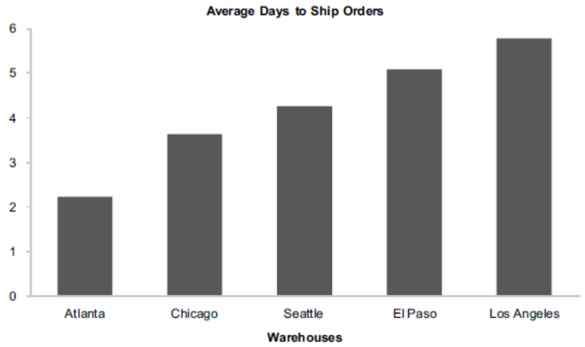
Single Distribution

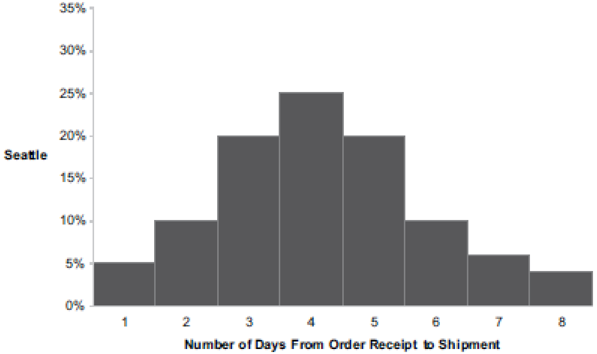
- When bars are used to display a distribution
- X-axis – categorical scale; Y-axis – quantitative scale



3 thick red lines – low, median and high values and 2 thin red lines – 25th and 75th percentiles.

Bar chart *vs* Histogram



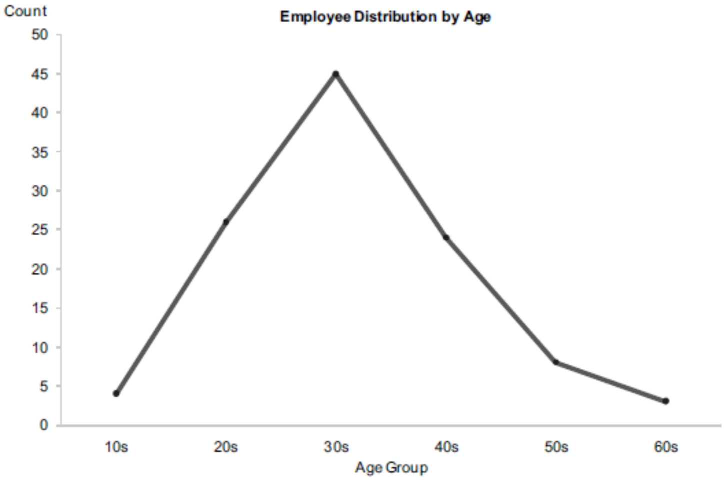


- Histograms are used to show distributions of variables while bar charts are used to compare variables. Histograms plot binned quantitative data while bar charts plot categorical data. Bars can be reordered in bar charts but not in histograms

Frequency Polygons

Single Distribution

- Essentially line chart on categorical scale to display a distribution.
- Benefit – focus on the shape



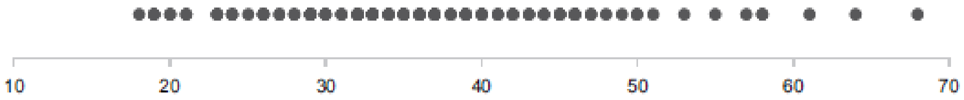
Does not support magnitude comparison between intervals as well as the histogram

Strip Plots

Single Distribution

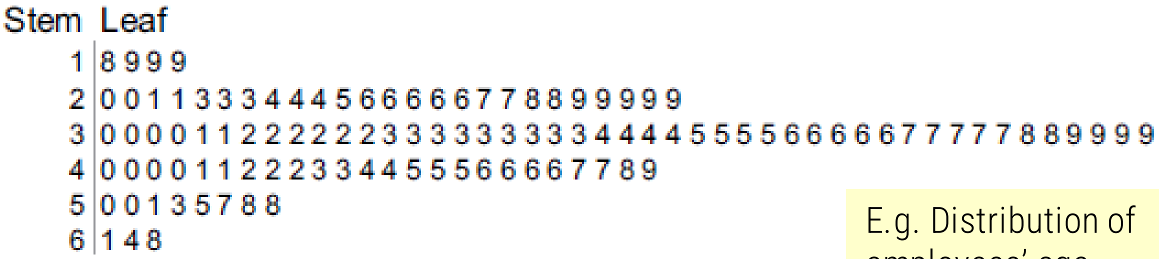
- One dimensional scatterplot

Can see low and high value but no shape distribution



Stem and Leaf Plots

Single Distribution



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

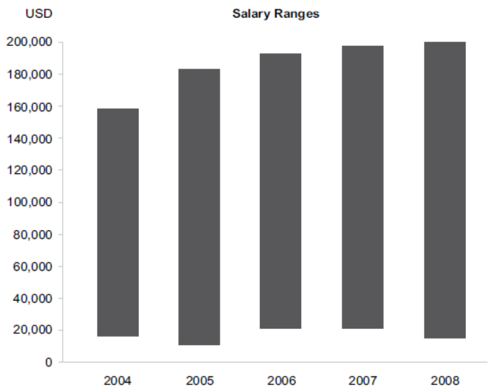
E.g. Distribution of employees’ age.
How to read?
Top row : 1 – 18; 3 - 19

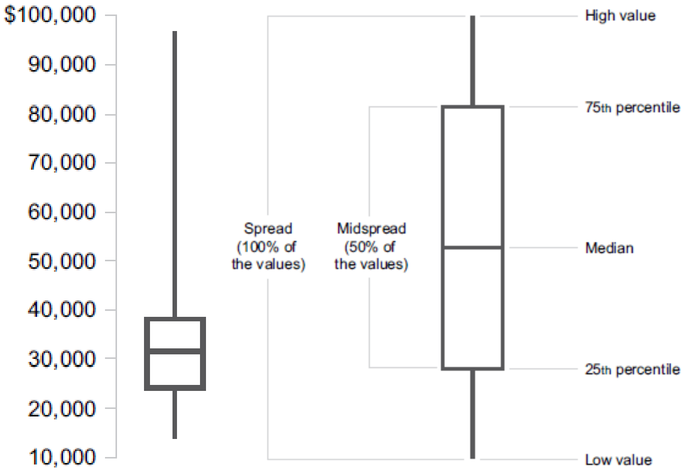
NANYANG THE INNOVATIVE POLYTECHNIC School of Information Technology

Box plots

Multiple Distribution

- Range bars are never adequate as they reveals the distribution’s spread while ignoring its center and shape.





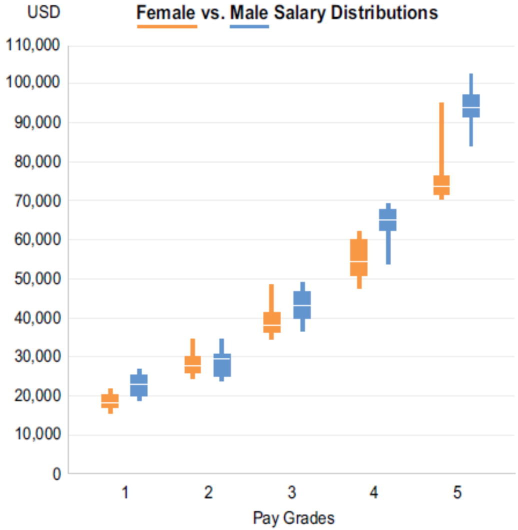
Box plots

Multiple Distribution

What does this plots tell about Female vs Male Salary Distribution?

- List the facts first

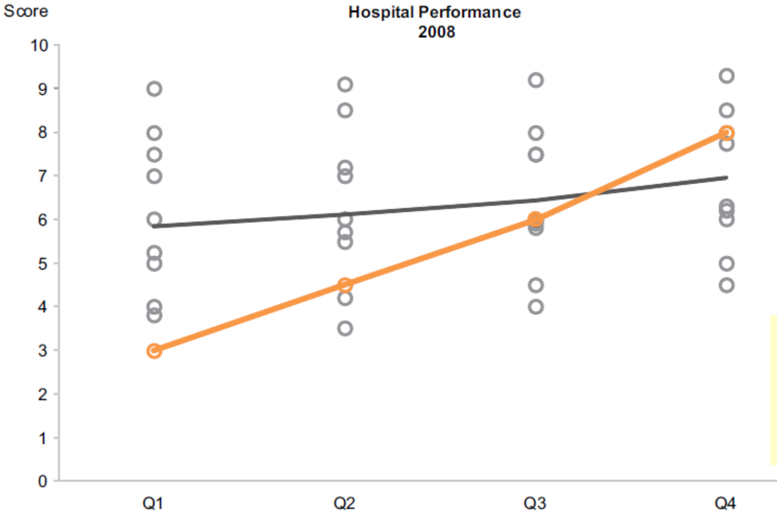
  - women are typically paid less than men in all salary grades
  - The disparity in salaries between men and women becomes increasingly greater as salaries increase
  - Salaries vary the most for women in the higher salary grades



Multiple Strip Plots

Multiple Distribution

- Health-care performances of 10 hospitals differed from one-another and changed through time.

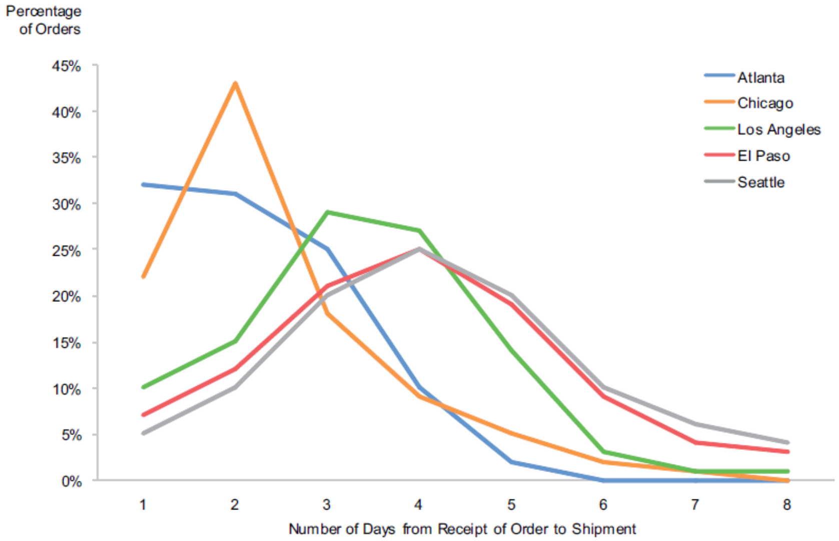


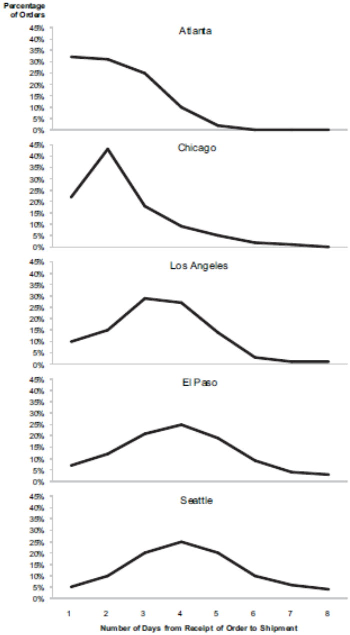
Ability to click on any dot and have the full series of values is nice feature

Frequency Polygons

Multiple Distribution

- Compare shapes of multiple distributions



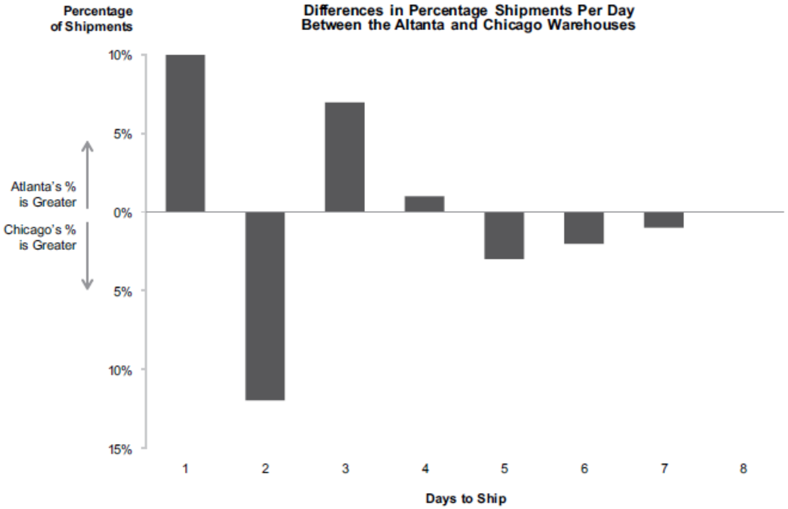


Trellis arrangement of line graphs

Distribution Deviation Graphs

Multiple Distribution

- When we want to focus on how two distributions differ, display the differences directly (can be in percentage)



There is no need to care about the shapes of the two distributions as we only care about how they differ

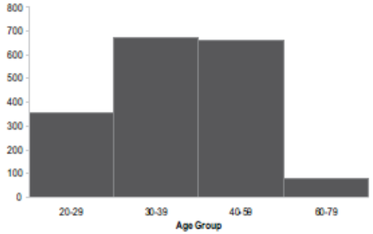
Distribution Analysis Techniques and Best Practices

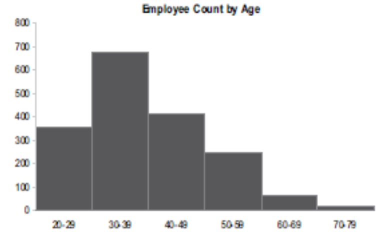
- Keeping the intervals consistent
- Selecting the best interval
- Using measures that are resistant to outliers

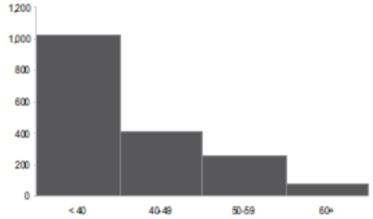
Keeping Intervals Consistent

- The size of the intervals along the categorical scale must be equals.

- Break this rule only when vast majority of values fall within a particular range

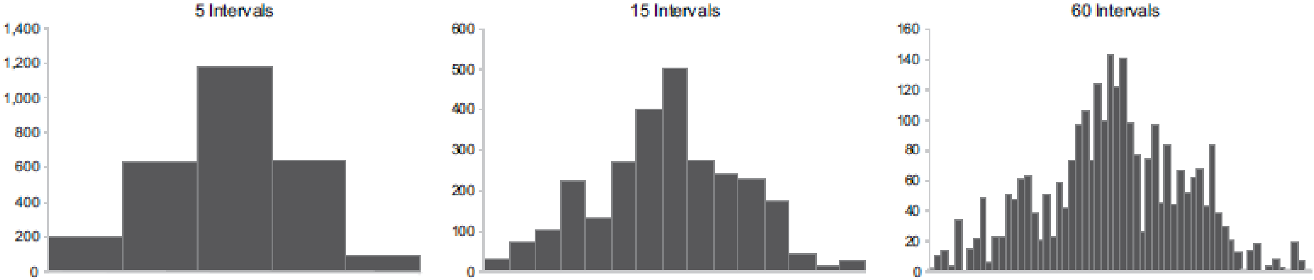






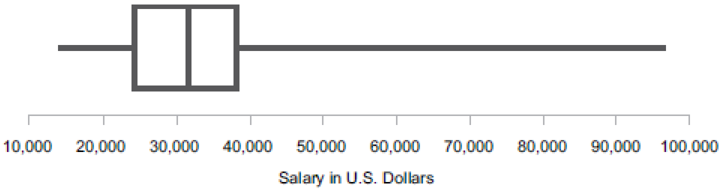
Selecting Best Interval

- Determine how large to make the interval or the number of intervals to use.
- Too many intervals → raggedly shaped distribution
- Too few intervals → too generalized, loss of variations



Resistant to Outliers

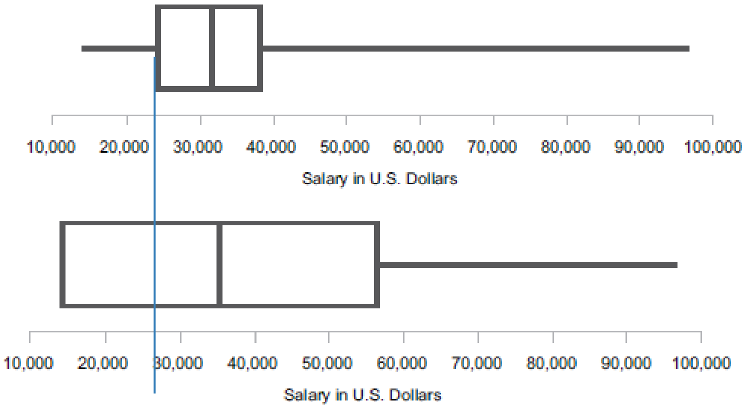
- Median and percentiles is recommended as it is more resistant than mean and standard deviation.

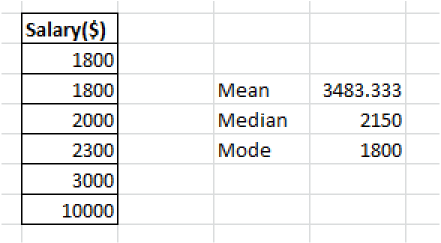


- The center is median and ends of box represent the 25th and 75th percentiles.
- With the extremely high salaries, long tail formed by the right whisker but didn’t influence the median

Resistant to Outliers

- In the second plot, the center represents the mean. The end of the box represent one standard deviation below and above the mean.
- Mean is higher susceptible to outliers. Single outlier can shift the mean significantly





Salary($)			
1800			
1800		Mean	3483.333
2000		Median	2150
2300		Mode	1800
3000			
10000			

How can mean be used?

Summary

- Visual Characteristics and Statistical Summaries of Distributions
- Distribution patterns
- How to display distribution data
- Techniques and best practices to consider for distribution analysis