Math 140
Introductory Statistics

Professor Bernardo Ábrego

Lecture 1

Statistical Work

- Data Exploration
  - Examination of data for *patterns*.
  - Tools: graphs, averages, etc.
- Inference (making inferences from data)
  - Definition: Deciding whether or not an observed feature of the data could reasonably be attributed to chance.

Data from Tables

Variables [columns]
Characteristics of each case.
It allows us to see the ***variability***

Cases [rows]
Subjects/objects
of statistical examination

Row	Job Title	Pay	…	RIF	Age
2	Engineering Clerk	H		0	25
3	Engineering Tech I	H		0	38
4	Engineering Tech II	H		0	56
…					
…					

In the example:
•Cases = individual Westvaco employees
•Variables = year of birth, job title, pay, etc.

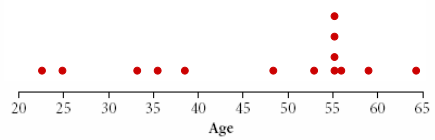
Understanding Variability

- To understand how the characteristics of the cases varies we look at their *distribution*.
- Distribution: What the values are and how often they occur (record of variability)
- How can we study the distribution?
  - By observing the values in each column of the table.
  - By graphing the values in a dot plot.

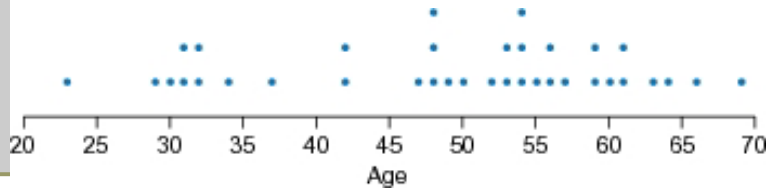
Dot Plots

- Each case is represented by a dot located according to the numerical value of the variable we are investigating.

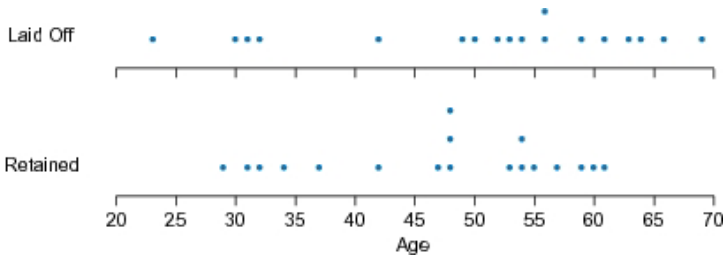
Row	Job Title	Age
1	Engineering Clerk	25
2	Engineering Tech II	38
3	Engineering Tech II	56
4	Secretary	48
5	Engineering Tech II	53
6	Engineering Tech II	55
7	Engineering Tech II	59
8	Parts Crib Attendant	22
9	Engineering Tech II	55
10	Engineering Tech II	64
11	Technical Secretary	55
12	Engineering Tech II	55
13	Engineering Tech II	33



Dot Plots (salaried workers)



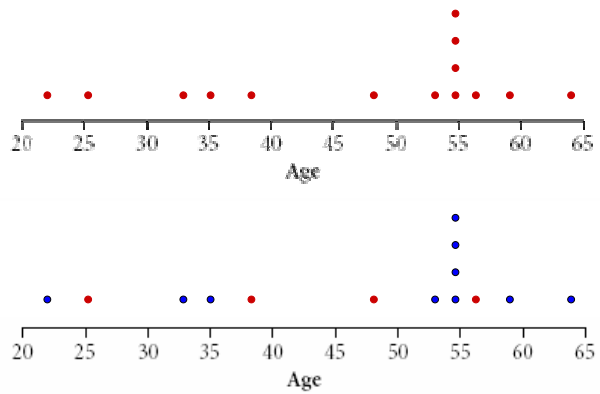
Laid Off vs. Retained



Discussion: Exploring the *Martin v. Westvaco* Data

- D1. Suppose you were on a jury in the *Martin v. Westvaco* case. How would you use the information in Display 1.1 (The table) to decide if Westvaco tended to lay off older workers (for whatever reason)?
- D2. Does the dot plot in Display 1.5 (next slide) shows a clear-cut case of age discrimination in layoffs of hourly workers at Westvaco, a possible case of age discrimination, or no discrimination?

Comparing dot plots



Display 1.3 Hourly workers: Ages of those laid off (open circles) and those retained (solid dots).

Using Tables to Compare

- The summary table shown here classifies salaried workers using two yes/no questions: Under 50? and Laid off ?

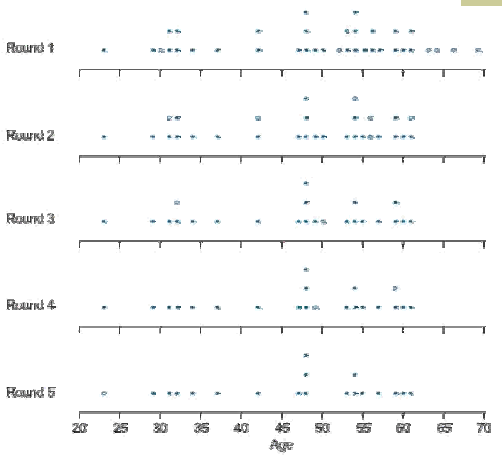
Laid Off?

Under 50?	Yes	No	Total	% Yes
Yes	6	10	16	44.4
No	12	8	20	55.5
Total	18	18	36	50

Discussion: Exploring the *Martin v. Westvaco* Data

- D3. The dot plots in Display 1.4 (next slide) show the ages of the salaried workers laid off and retained by round. For example, in the top dot plot, the open circles show the ages of the hourly workers laid off in Round 1 and the solid dots show the ages of the hourly workers whose jobs survived Round 1. Compare the round-by-round information you get in Display 1.4 with the summary for all rounds in Display 1.2.Which display provides stronger support for Martin's claim that Westvaco discriminated against older workers?

Round by Round

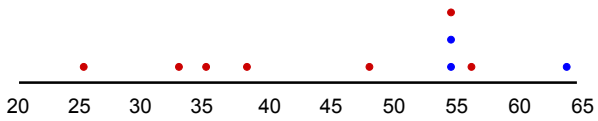


How to Analyze Patterns?

- Overall, the exploratory work we just did shows that older workers were more likely than younger ones to be laid off, and they were laid off earlier. One of the main arguments in the court case was about what those patterns mean:
  - Can we infer from them that Westvaco has some explaining to do?
  - Or are the patterns of the sort that might happen even if there was no discrimination?

Summary Statistic

- Consider as an example of our analysis Round 2 of the layoffs.



- To simplify the statistical analysis to come, it will help to "condense" the data into a single number, called a **summary statistic.** One possible summary statistic is the average, or mean, age of the three who lost their jobs:

$$\boxed{\text{average} = \frac{55+55+64}{3} = 58 \text{ years}}$$

Martin v. Westvaco

- ***Martin:*** Look at the pattern in the data. All three of the workers laid off were much older than the average age of all workers. That's evidence of age discrimination.
- ***Westvaco:*** Not so fast! You're looking at only ten people total, and only three positions were eliminated. Just one small change and the picture would be entirely different. For example, suppose it had been the 25-year-old instead of the 64-year-old who was laid off. Switch the 25 and the 64 and you get a totally different set of averages:
  - Actual data: 25 33 35 38 48 **55 55** 55 56 **64**
  - Altered data: **25** 33 35 38 48 **55 55** 55 56 64

  See! Just one small change and the average age of the three who were laid off is *lower* than the average age of the others.

	Laid Off	Retained
Actual data	58.0	41.4
Altered data	45.0	47.0

Martin v. Westvaco

- ***Martin:*** Not so fast, yourself! Of all the possible changes, you picked the one that is most favorable to your side. If you'd switched one of the 55-year-olds who got laid off with the 55-year-old who kept his or her job, the averages wouldn't change at all. Why not compare what actually happened with ***all*** the possibilities that might have happened?
- ***Westvaco:*** What do you mean?
- ***Martin:*** Start with the ten workers, treat them all alike, and pick three at random. Do this over and over, to see what typically happens, and compare the actual data with these results. Then we'll find out how likely it is that their average age would be 58 or more.

Discussion

- D5. If you pick three of the ten ages at random, do you think you are likely to get an average age of 58 or more?
- D6. If the probability of getting an average age of 58 or more turns out to be small, does this favor Martin or Westvaco?