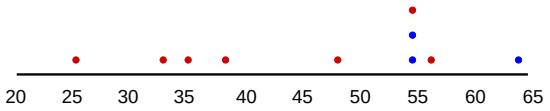
Math 140
Introductory Statistics

Professor Silvia Fernández

Lecture 2
Based on the book *Statistics in Action*
by A. Watkins, R. Scheaffer, and G. Cobb.

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:

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

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?

Martin v. Westvaco

- **Martin:** Look at the pattern in the data. All three of the workers laid off were much older than average.
- **Westvaco:** So what? You could get a result like that just by chance. If chance alone can account for the pattern, there's no reason to ask us for any other explanation.
- **Martin:** Of course you *could* get this result by chance. Th e question is whether it's easy or hard to do so. If it's easy to get an average as large as 58 by drawing at random, I'll agree that we can't rule out chance as one possible explanation. But if an average that large is really hard to get from random draws, we agree that it's not reasonable to say that chance alone accounts for the pattern. Right?
- **Westvaco:** Right

Martin v. Westvaco

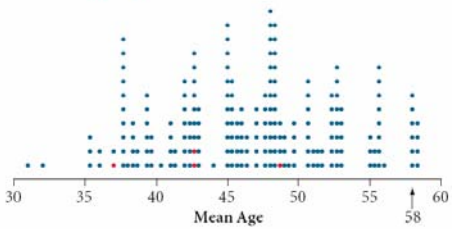
- **Martin:** Here are the results of my simulation. If you look at the three hourly workers laid off in Round 2, the probability of getting an average age of 58 or greater by chance alone is only 5%. And if you do the same computations for the entire engineering department, the probability is a lot lower, about 1%. What do you say to that?
- **Westvaco:** Well . . . I'll agree that it's really hard to get an average age that extreme simply by chance, but that by itself still doesn't prove discrimination.
- **Martin:** No, but I think it leaves you with some explaining to do!

Simulation

- In our example we can draw 3 of the 10 ages at random and compute the average. Then repeat this process a large number of times to see how likely would be to get 58 or more as the answer.
- Steps in a Simulation:
  - **Random model**: Create a model for the chance process (pieces of paper thoroughly mixed, sequence of random numbers, computer generated random numbers).
  - **Summary Statistic**: Calculate it (mean=average in our example)
  - **Repetition**: Repeat a large number of times (1000s)
  - **Display the distribution**: (Using a dot plot for example)
  - **Estimate the Probability**: (In our example the proportion of values that gave 58 or more)
  - **Reach a conclusion**: Interpret your results.

Simulation Martin Case: Round 2 - Hourly workers

										Average Age
25	33	35	38	48	55	55	55	56	64	42.7
25	33	35	38	48	55	55	55	56	64	48.0
25	33	35	38	48	55	55	55	56	64	42.7
25	33	35	38	48	55	55	55	56	64	37.0



Display 1.10 Results of 200 repetitions: the distribution of the average age of the three workers chosen for layoff by chance alone.

Discussion

- D7. Why must you estimate the probability of getting an average age of 58 *or greater* rather than the probability of getting an average age of 58?

Discussion

- D8. *How unlikely is "too unlikely"?* The probability in the previous activity is in fact exactly equal to 0.05. In a typical court case, a probability of 0.025 or less is required to serve as evidence of discrimination.
  - a. Did the Round 2 layoff s of hourly workers in the *Martin* case meet the court requirement?
  - b. If the probability in the *Martin* case had been 0.01 instead of 0.05, how would that have changed your conclusions? 0.10 instead of 0.05?

Inference

- Inference is a statistical procedure that involves deciding whether an event can reasonably be attributed to chance or whether you should look for some other explanation.
- In the *Martin* case we used **simulation** as a device for **inference** to determine whether the relatively high average age of the laid-off hourly employees in Round 2 could reasonably be due to chance.
- The probability was about 0.05, which was considered small enough to warrant asking for an explanation from Westvaco but not small enough to present in court as clear evidence of discrimination.

Practice

- P4. Suppose three workers were laid off from a set of ten whose ages were the same as those of the hourly workers in Round 2 in the *Martin* case. This time, however, the ages of those laid off were 48, 55, and 55.

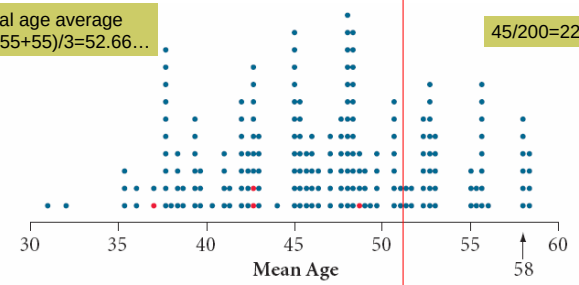
25 33 35 38 **48** 55 **55** **55** 56 64

- a. Use the dot plot in Display 1.10 on page 14 to estimate the probability of getting an average age as large as or larger than that of those laid off in this situation.
- b. What would your conclusion be if Westvaco had laid off workers of these three ages?

Average age of 3 workers out of 10

Actual age average (48+55+55)/3=52.66…

45/200=22.5%



Display 1.10 Results of 200 repetitions: the distribution of the average age of the three workers chosen for layoff by chance alone.

Practice

- At the beginning of Round 1, there were 14 hourly workers. Their ages were 22, 25, 33, 35, 38, 48, 53, 55, 55, 55, 55, 56, 59, and 64. After the layoffs were complete, the ages of those left were 25, 38, 48, and 56. Think about how you would repeat Activity 1.2a using these data.

a. What is the average age of the ten workers laid off?

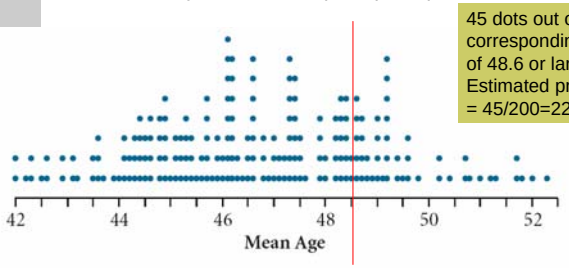
(22+33+35+53+55+55+55+55+59+64)/10=48.6

b. Describe a simulation for finding the distribution of the average age of ten workers laid off at random.

Step 1. Select 10 out of the 14 ages at random and find their average.
Step 2. Repeat step 1 many times. (For example, 200 times.)
Step 3. Create a dot plot containing the averages obtained from your repetitions.

c. The results of 200 repetitions from a simulation are shown in Display 1.11. Suppose 10 workers are picked at random for layoff from the 14 hourly workers. Make a rough estimate of the probability of getting, just by chance, the same or larger average age as that of the workers who actually were laid off (from part a).

45 dots out of 200 to the right, corresponding to an average of 48.6 or larger.
Estimated probability = 45/200=22.5%



Display 1.11 Results of 200 repetitions.

d. Does this analysis provide evidence in Martin's favor?

No, a probability of 22.5% is too large to be considered evidence that the actual average may not be due to chance.

Visualizing Distributions

- Recall the definition:

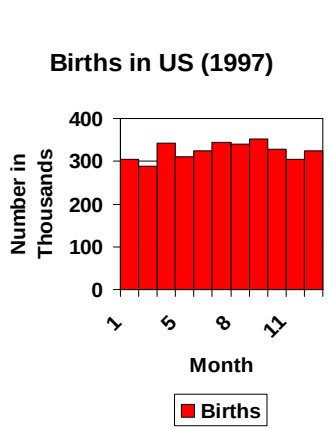
  The values of a summary statistic (e.g. the average age of the laid-off workers) and how often they occur.
- Four of the most common basic **shapes**:
  - Uniform or Rectangular
  - Normal
  - Skewed
  - Bimodal (Multimodal)

Uniform (or Rectangular) Distribution

- Each outcome occurs roughly the same number of times.
- Examples.
  - Number of U.S. births per month in a particular year (see Page 25)
  - Computer generated random numbers on a particular interval.
  - Number of times a fair die is rolled on a particular number.

Month	Births (in thousands)	Deaths (in thousands)
1	305	218
2	289	191
3	313	198
4	342	189
5	311	195
6	324	182
7	345	192
8	341	178
9	353	176
10	329	193
11	304	189
12	324	192

Uniform (or Rectangular) Distribution



Month	Births (in thousands)	Deaths (in thousands)
1	305	218
2	289	191
3	313	198
4	342	189
5	311	195
6	324	182
7	345	192
8	341	178
9	353	176
10	329	193
11	304	189
12	324	192

Normal Distributions

- These distributions arise from
  - Variations in measurements.

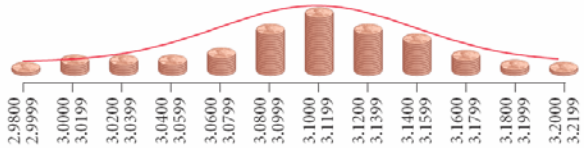
    (e.g. pennies example, see 2.3 page 31)
  - Natural variations in population sizes

    (e.g. weight of a set of people)
  - Variations in averages of random samples.

    (e.g. Average age of 3 workers out of 10, see 1.10 in page 14)

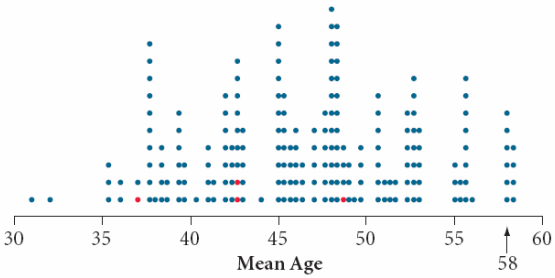
Pennies example

Pennies minted in the United States are supposed to weigh 3.110 g, but a tolerance of 0.130 g is allowed in either direction. Display 2.3 shows a plot of the weights of 100 pennies.



Display 2.3 Weights of pennies. [*Source:* W. J. Youden, *Experimentation and Measurement* (National Science Teachers Association, 1985), p. 108.]

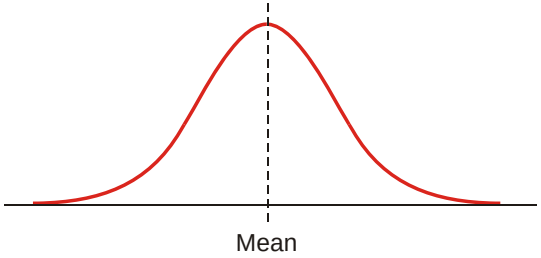
Average age of 3 workers out of 10



Display 1.10 Results of 200 repetitions: the distribution of the average age of the three workers chosen for layoff by chance alone.

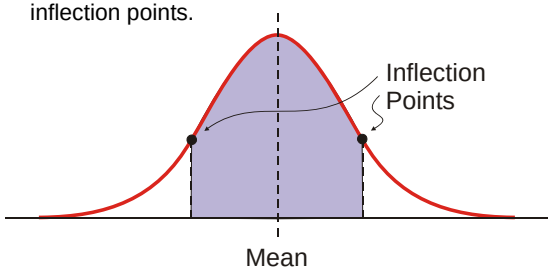
Normal Distributions

- Idealized shape shown below (see 2.4 page 32)
- Properties:
  - Single peak: The x-value of it is called the **mean**.
  - The mean tells us where is the **center** of the distribution.
  - The distribution is symmetric with respect to the mean.



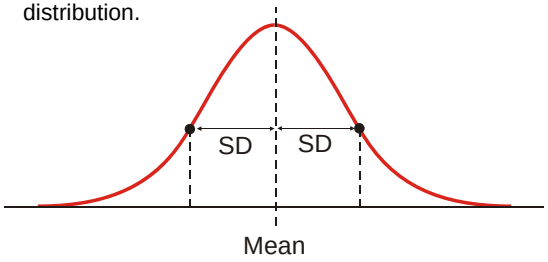
Normal Distributions

- Idealized shape shown below (see 2.4 page 32)
- Properties:
  - Inflection points: Where concavity changes.
  - Roughly 2/3 of the area below the curve is between the inflection points.



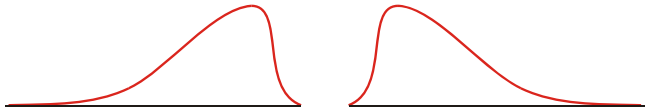
Normal Distributions

- Idealized shape shown below (see 2.4 page 32)
- Properties:
  - The distance between the mean and either of the inflection points is called the **standard deviation** (SD)
  - The standard deviation measures how **spread** is the distribution.



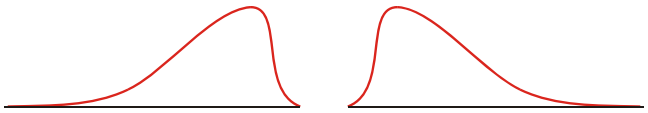
Skewed Distributions

- These are similar to the normal distributions but they are not symmetric. They have values bunching on one end and a long tail stretching in the other direction
- The tail tells you whether the distribution is **skewed left** or **skewed right**.

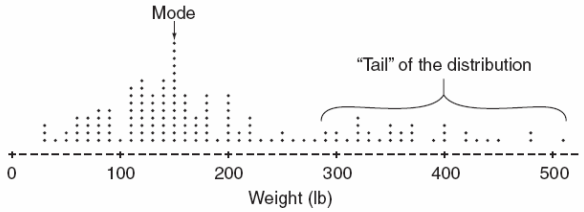


Skewed Distributions

- Skewed distributions often occur because of a "wall", that is, values that you cannot go below or above. Like zero for positive measurements, or 100 for percentages.
- To find out about **center** and **spread** it is useful to look at **quartiles**.



Example of a skewed right distribution

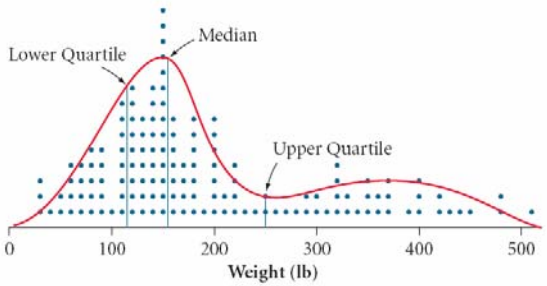


Display 2.6 Weights of bears in pounds. [*Source:* MINITAB data set from *MINITAB Handbook*, 3rd ed.]

Median and Quartiles

- Median: the value of the line dividing the number of values in equal halves. (Or the area under the curve in equal halves.)
- Repeat this process in each of the two halves to find the lower quartile (Q1) and the upper quartile (Q3).
- Q1, the median, and Q3 divide the number of values in **quarters**. The quartiles Q1 and Q3 enclose 50% of the values.

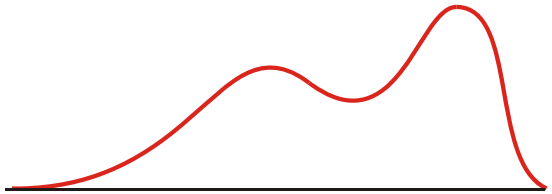
Visualizing Median and Quartiles



Display 2.8 Estimating center and spread for the weights of bears.

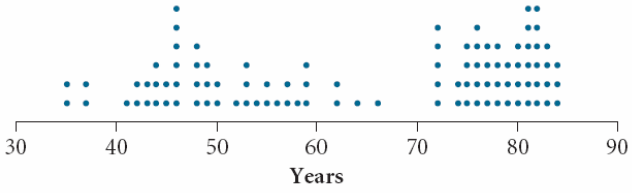
Bimodal Distributions.

- Previous distributions have had only one peak (**unimodal**) but some have two (**bimodal**) or even more (**multimodal**).



Bimodal Distribution

Example of a bimodal distribution



Display 2.9 Life expectancy of females by country on two continents. [*Source:* Population Reference Bureau, *World Population Data Sheet*, 2005.]