

Math 140

Introductory Statistics

Professor B. Ábrego

Lecture 6

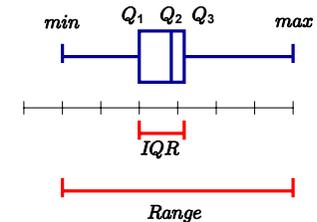
Sections 2.3, 2.4, and 2.5

Box Plots

- Example: Mammal speeds, 11, 12, 20, 25, 30, 30, 30, 32, 35, 39, 40, 40, 40, 42, 45, 48, 50, 70.
- A **Box Plot** is a *graphical display* of a five-point summary.

- $min = 11$
- $Q_1 = 30$
- Median = 37
- $Q_3 = 42$
- $max = 70$.

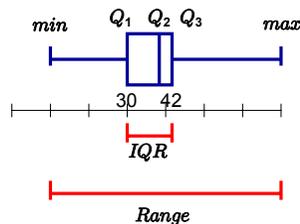
- Range = $70 - 11 = 59$
- $IQR = 42 - 30 = 12$



Modified Box Plots

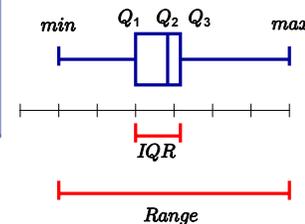
- Example: Mammal speeds, 11, 12, 20, 25, 30, 30, 30, 32, 35, 39, 40, 40, 40, 42, 45, 48, 50, 70.
- A **Modified Box Plot** also takes into account the *outliers*.
- An *outlier* is a value which is more than 1.5 times the *IQR* from the nearest quartile.

- Example:
 - $(1.5)IQR = 18$
 - $11 < 30 - 18$, so 11 is an outlier.
 - $70 > 42 + 18$, so 70 is an outlier.



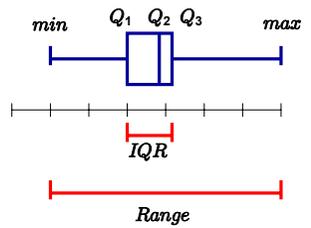
Modified Box Plots

- Example: Mammal speeds, 11, 12, 20, 25, 30, 30, 30, 32, 35, 39, 40, 40, 40, 42, 45, 48, 50, 70.
- Example:
 - $(1.5)IQR = 18$
 - $11 < 30 - 18$, so 11 is an outlier.
 - $70 > 42 + 18$, so 70 is an outlier.

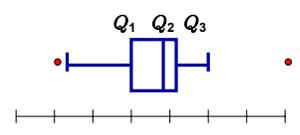


Modified Box Plots

- Example: Mammal speeds, 11, 12, 20, 25, 30, 30, 30, 32, 35, 39, 40, 40, 40, 42, 45, 48, 50, 70.



- Example:
 - $(1.5)IQR = 18$
 - $11 < 30 - 18$, so 11 is an outlier.
 - $70 > 42 + 18$, so 70 is an outlier.



Box Plots (Modified)

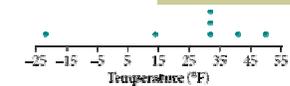
- Box Plots and Modified Box Plots are useful when plotting a single quantitative variable and:
 - want to compare shape, center, and spread of two or more distributions.
 - The distribution has a large number of values
 - Individual values do not need to be identified.
 - (Modified) We want to identify outliers.

Section 2.4 Recentering and Rescaling

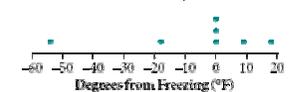
- Recentering** a data set (adding the same number c to all the values in the set)
 - Shape or spread do not change.
 - It slides the entire distribution by the amount c , adding c to the median and the mean.
- Rescaling** a data set (multiplying all the values in the set by the same positive number d)
 - Basic shape doesn't change.
 - It stretches or shrinks the distribution, multiplying the spread (IQR or standard deviation) by d and multiplying the center (median or mean) by d .

Example

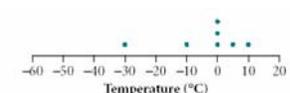
City	Country	Temperature (°F)
Addis Ababa	Ethiopia	32
Algiers	Algeria	32
Bangkok	Thailand	50
Madrid	Spain	14
Nairobi	Kenya	41
Brazilia	Brazil	32
Warsaw	Poland	-22



Display 2.63 Dot plot for record low temperatures in degrees Fahrenheit for seven capitals.



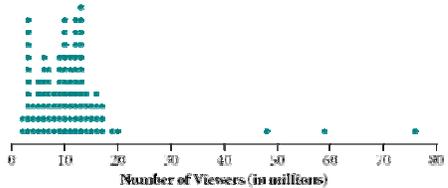
Display 2.64 Dot plot of the number of degrees Fahrenheit above or below freezing for record low temperatures for the seven capitals.



Display 2.65 Dot plot for record low temperatures in degrees Celsius for the seven capitals.

The Influence of Outliers

- A summary statistic is
 - **resistant to outliers** if it does not change very much when an outlier is removed.
 - **sensitive to outliers** if the summary statistic is greatly affected by the removal of outliers.

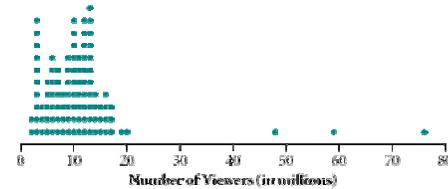


Display 2.66 Number of viewers of prime-time television shows in a particular week.

Example

Variable	N	Mean	Median	StDev
Ratings	101	11.187	10.150	9.896
Variable	Min	Max	Q1	Q3
Ratings	2.320	76.260	6.160	12.855

Display 2.67 Printout of summary statistics for number of viewers.



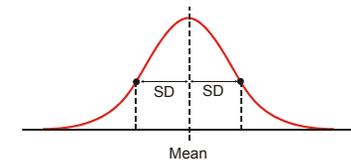
Display 2.66 Number of viewers of prime-time television shows in a particular week.

Percentiles and CRF plots

- You are responsible to read through this and understand the concepts of **percentile**, and **cumulative relative frequency plot**.

2.5 The Normal Distribution

- Shape



- Center: Mean

$$\bar{x} = \frac{\text{sum of values}}{\text{number of values}} = \frac{\sum x}{n}$$

- Spread: Standard Deviation

$$\sigma_{n-1} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Applications of the Normal Distribution

- The normal distribution tells us how:
 - Variability in measures behaves.
 - Variability in population behaves.
 - Averages and some other summary statistics behave when you repeat a random process.
- Nice property: A normal distribution is determined by its **mean** and **standard deviation**!
(If you know mean and SD you know everything)

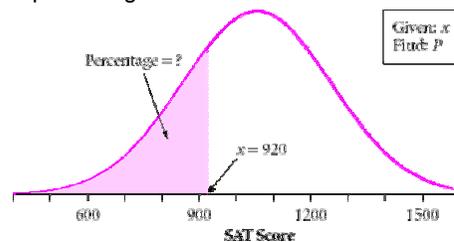
The Two Main Problems.

- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.
1. What percentage of scores were 920 or below?
(Unknown percentage problem)
 2. What SAT score separates the lowest 25% of the SAT scores from the rest?
(Unknown value problem)

Unknown percentage problem.

- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

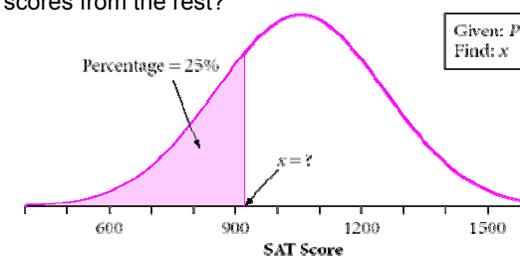
1. What percentage of scores were 920 or below?



Unknown value problem.

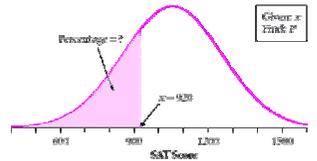
- The distribution of the SAT scores for the University of Washington was roughly normal in shape, with mean 1055 and standard deviation 200.

2. What SAT score separates the lowest 25% of the SAT scores from the rest?

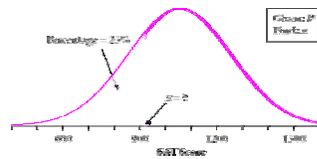


Which one is it?

1. Unknown percentage problem.
Given x , Find P .

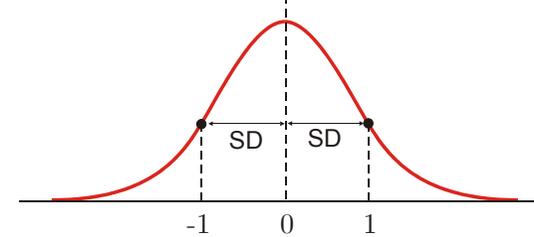


2. Unknown value problem.
Given P , Find x .



The Standard Normal Distribution.

- It is the normal distribution with **Mean = 0**, and **standard deviation = 1**.
The area under the curve equals 1 (or 100%)



The Standard Normal Distribution.

- It is the normal distribution with Mean = 0, and standard deviation = 1.
The area under the curve equals 1 (or 100%)
- The Standard Normal Distribution is important because any normal distribution can be **recentered** and/or **rescaled** to the standard normal distribution. This process is called **standardizing** or **converting to standard units**.
- Also, the two main problems can be easily solved in the Standard Normal Distribution with the help of **tables** or a **calculator**.

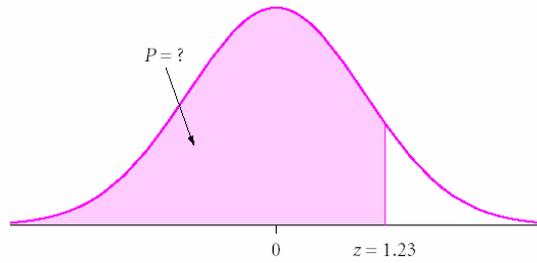
The Two Main Problems in the Standard Normal Distribution.

Unknown Percentage. (Given z , Find P)

- With Table A. Last page in the book
 - Use the units and the first decimal to locate the row and the closest hundredths digits to locate the column. The number found is the percentage of the number of values **below** z .
- With Calculator
 - Enter `normalcdf(-99999, z)` to get the percentage of the number of values **below** z .

Example

Find the percentage, P , of values below $z = 1.23$.



Display 2.77 The percentage of values below $z = 1.23$.