Mathematical statistics

February 18^{th}, 2019

Lecture 4: Statistics and sampling distribution

Topics

Week 1	Probability reviews
Week 2	**Chapter 6: Statistics and Sampling Distributions**
Week 4	Chapter 7: Point Estimation
Week 7	Chapter 8: Confidence Intervals
Week 10	Chapter 9: Test of Hypothesis
Week 14	Regression

Descriptive statistics

Pictorial methods

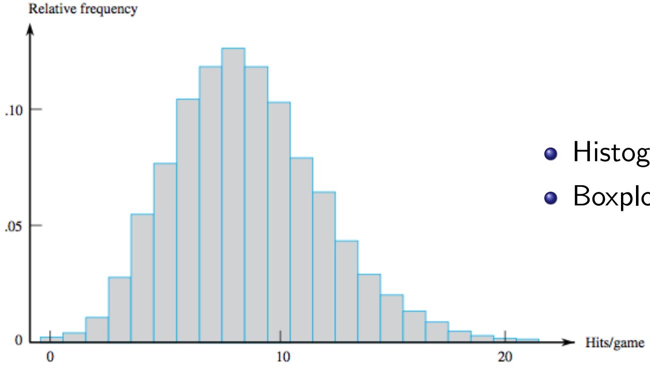


Figure 1.6 Histogram of number of hits per nine-inning game

- Histograms
- Boxplots

1.3: Measures of locations

- The Mean
- The Median
- Trimmed Means

Measures of locations: mean

The **sample mean** $\bar{x}$ of observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measures of locations: median

Step 1: ordering the observations from smallest to largest

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\dfrac{n+1}{2}\right)^{\text{th}} \text{ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\dfrac{n}{2}\right)^{\text{th}} \text{ and } \left(\dfrac{n}{2}+1\right)^{\text{th}} \text{ ordered values} \end{cases}$$

Median is not affected by outliers

Measures of locations: trimmed mean

- A α% trimmed mean is computed by:
  - eliminating the smallest α% and the largest α% of the sample
  - averaging what remains
- $\alpha = 0 \rightarrow$ the mean
- $\alpha \approx 50 \rightarrow$ the median

Measures of variability: deviations from the mean

The **sample variance**, denoted by s^2, is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

The **sample standard deviation**, denoted by s, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

Measures of variability: deviations from the mean

The **sample variance**, denoted by s^2, is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

The **sample standard deviation**, denoted by s, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

- Why squared? Because it is easier to do math with x^2 than $|x|$
- Why $(n-1)$? Because that makes s^2 an *unbiased estimator* of the population variance (Chapter 7)

Computing formula for s^2

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Proof Because $\bar{x} = \sum x_i/n$, $n\bar{x}^2 = \left(\sum x_i\right)^2/n$. Then,

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x} \sum x_i + \sum (\bar{x})^2$$

$$= \sum x_i^2 - 2\bar{x} \cdot n\bar{x} + n(\bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$$

Properties of the sample standard deviation

Let $x_1, x_2, \ldots, x_n$ be a sample and c be a constant.

1. If $y_1 = x_1 + c, y_2 = x_2 + c, \ldots, y_n = x_n + c$, then $s_y^2 = s_x^2$, and

2. If $y_1 = cx_1, \ldots, y_n = cx_n$, then $s_y^2 = c^2 s_x^2$, $s_y = |c| s_x$,

where s_x^2 is the sample variance of the x's and s_y^2 is the sample variance of the y's.

Order the n observations from smallest to largest and separate the smallest half from the largest half; the median $\tilde{x}$ is included in both halves if n is odd. Then the **lower fourth** is the median of the smallest half and the **upper fourth** is the median of the largest half. A measure of spread that is resistant to outliers is the **fourth spread** f_s, given by

$$f_s = \text{upper fourth} - \text{lower fourth}$$

Boxplots

40 52 55 60 70 75 85 85 90 90 92 94 94 95 98 100 115 125 125

The five-number summary is as follows:

smallest $x_i = 40$ lower fourth $= 72.5$ $\tilde{x} = 90$ upper fourth $= 96.5$
largest $x_i = 125$

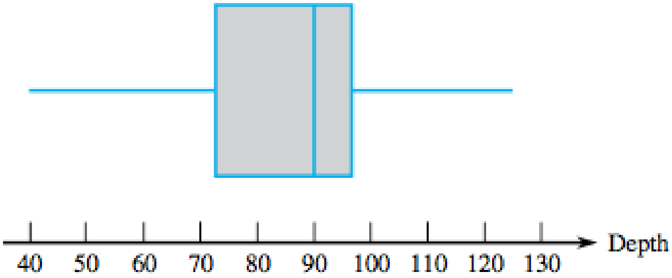
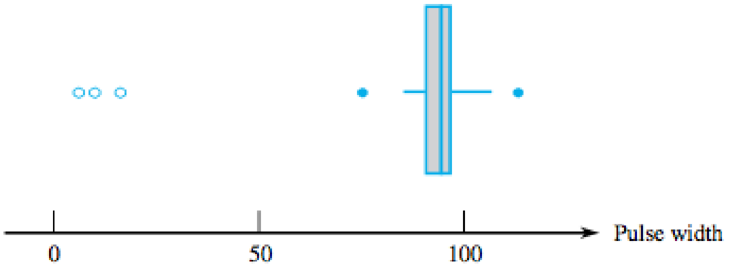


Figure 1.17 A boxplot of the corrosion data

Boxplot with outliers

Any observation farther than $1.5f_s$ from the closest fourth is an **outlier**. An outlier is **extreme** if it is more than $3f_s$ from the nearest fourth, and it is **mild** otherwise.



Overview

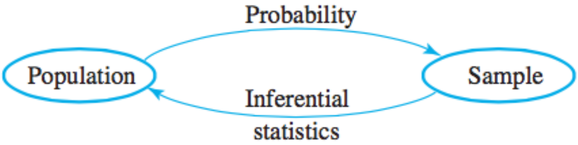
6.1 Statistics and their distributions

6.2 The distribution of the sample mean

6.3 The distribution of a linear combination

Order $6.1 \rightarrow 6.3 \rightarrow 6.2$

Random sample



Definition

The random variables $X_1, X_2, ..., X_n$ are said to form a (simple) random sample of size n if

1. the X_i's are independent random variables
2. every X_i has the same probability distribution

Recap: Independent random variables

Definition

Two random variables X and Y are said to be independent if for every pair of x and y values,

$$P(X = x, Y = y) = P_X(x) \cdot P_Y(y) \qquad \text{if the variables are discrete}$$

or

$$f(x, y) = f_X(x) \cdot f_Y(y) \qquad \text{if the variables are continuous}$$

Property

If X and Y are independent, then for any functions g and h

$$E[g(X) \cdot h(Y)] = E[g(X)] \cdot E[h(Y)]$$

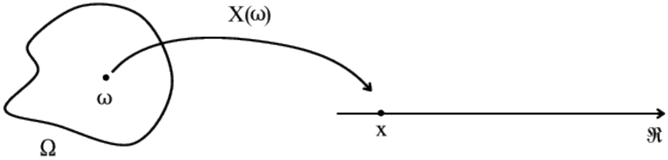
Statistics

Definition

A statistic is any quantity whose value can be calculated from sample data

- prior to obtaining data, there is uncertainty as to what value of any particular statistic will result $\rightarrow$ a statistic is a random variable
- the probability distribution of a statistic is referred to as its *sampling distribution*

Random variables



- random variables are used to model uncertainties
- Notations:
  - random variables are denoted by uppercase letters (e.g., X);
  - the calculated/observed values of the random variables are denoted by lowercase letters (e.g., x)

Example of a statistic

- Let $X_1, X_2, \ldots, X_n$ be a random sample of size n
- The sample mean of $X_1, X_2, \ldots, X_n$, defined by

$$\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n},$$

  is a statistic
- When the values of $x_1, x_2, \ldots, x_n$ are collected,

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n},$$

  is a realization of the statistic $\bar{X}$

Example of a statistic

- Let $X_1, X_2, \ldots, X_n$ be a random sample of size n
- The random variable

$$T = X_1 + 2X_2 + 3X_5$$

  is a statistic
- When the values of $x_1, x_2, \ldots, x_n$ are collected,

$$t = x_1 + 2x_2 + 3x_5,$$

  is a realization of the statistic T

Questions for this chapter

Given statistic T computed from sample $X_1, X_2, \ldots, X_n$

- Question 1: If we **know** the distribution of X_i's, can we obtain the distribution of T?
- Question 2: If we **don't know** the distribution of X_i's, can we still obtain/approximate the distribution of T?

Real questions: If T is a linear combination of X_i's, can we

- compute the distribution of T in some easy cases?
- compute the expected value and variance of T?

Questions for this section

Real questions: If $T = X_1 + X_2$

- compute the distribution of T in some easy cases
- compute the expected value and variance of T

Example 1

Problem

Consider the distribution P

x	10	15	20
$p(x)$	0.2	0.3	0.5

Let $\{X_1, X_2\}$ *be a random sample of size 2 from* P*, and* $T = X_1 + X_2$.

1. *Compute* $P[T = 40]$

Example 1

Problem

Consider the distribution P

x	10	15	20
$p(x)$	0.2	0.3	0.5

Let $\{X_1, X_2\}$ *be a random sample of size 2 from* P, *and* $T = X_1 + X_2$.

1. *Compute* $P[T = 40]$
2. *Derive the probability mass function of* T

Example 1

Problem

Consider the distribution P

x	10	15	20
$p(x)$	0.2	0.3	0.5

Let $\{X_1, X_2\}$ *be a random sample of size 2 from P, and* $T = X_1 + X_2$.

1. *Compute* $P[T = 100]$
2. *Derive the probability mass function of T*
3. *Compute the expected value and the standard deviation of T*