MATH 450: Mathematical statistics

September 10th, 2020

Lecture 4: Statistics and their distributions

Topics

Week 2	**Chapter 6: Statistics and Sampling Distributions**
Week 4	Chapter 7: Point Estimation
Week 6	Chapter 8: Confidence Intervals
Week 9	Chapter 9: Test of Hypothesis
Week 11	Chapter 10: Two-sample inference
Week 12	Regression

Overview

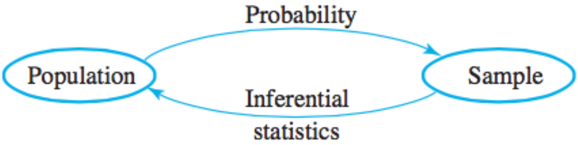
6.1 Statistics and their distributions

6.2 The distribution of the sample mean

6.3 The distribution of a linear combination

Order $6.1 \to 6.3 \to 6.2$

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,

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

or

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

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*
- Notations:
  - random variables are denoted by uppercase letters (e.g., X, $\bar{X}$);
  - the calculated/observed values of the random variables are denoted by lowercase letters (e.g., x, $\bar{x}$)

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

Real questions: If

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n,$$

can we

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

Focus:

$$T = X_1 + X_2$$

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

Example 2

Problem

Let $\{X_1, X_2\}$ *be a random sample of size 2 from the exponential distribution with parameter* λ

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

and $T = X_1 + X_2$.
What is the distribution of T*?*

For continuous random variable:

$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f(x)\, dx$$

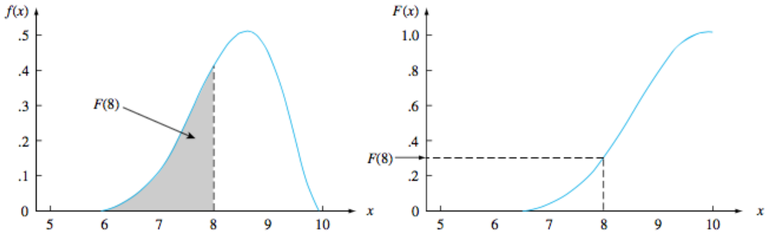


Figure 4.5 A pdf and associated cdf

Moreover:

$$f(x) = F'(x)$$

Example 2

Problem

Let $\{X_1, X_2\}$ be a random sample of size 2 from the exponential distribution with parameter λ

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

and $T = X_1 + X_2$.

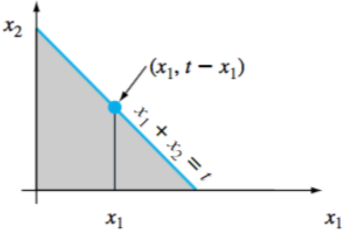
1. *Compute the cumulative density function (cdf) of T*

Example 2

$$F_{T_o}(t) = P(X_1 + X_2 \le t) = \iint_{\{(x_1, x_2): x_1 + x_2 \le t\}} f(x_1, x_2)\, dx_1\, dx_2$$

$$= \int_0^t \int_0^{t-x_1} \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2}\, dx_2\, dx_1 = \int_0^t (\lambda e^{-\lambda x_1} - \lambda e^{-\lambda t})\, dx_1$$

$$= 1 - e^{-\lambda t} - \lambda t e^{-\lambda t}$$



Example 2b

Problem

Let $\{X_1, X_2\}$ *be a random sample of size 2 from the exponential distribution with parameter* $\lambda = 2$

$$f(x) = \begin{cases} 2e^{-2x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

and $T = X_1 + X_2$.

1. *Compute the cumulative density function (cdf) of* T
2. *Compute the probability density function (pdf) of* T

Summary

1. If the distribution and the statistic T is simple, try to construct the pmf of the statistic (as in Example 1)
2. If the probability density function $f_X(x)$ of X's is known, the
   - try to represent/compute the cumulative distribution (cdf) of T

$$\mathbb{P}[T \leq t]$$

   - take the derivative of the function (with respect to t)

Example 1*

Problem

Consider the distribution P

x	40	45	50
$p(x)$	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. *Derive the probability mass function of* T
2. *Compute the expected value and the standard deviation of* T

Example 2*

Problem

Let $\{X_1, X_2\}$ *be a random sample of size 2 from the exponential distribution with parameter* λ

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

and $T = X_1 + 2X_2$.

1. *Compute the cumulative density function (cdf) of* T
2. *Compute the probability density function (pdf) of* T

Linear combination of normal random variables

Linear combination of normal random variables

Theorem

Let $X_1, X_2, \ldots, X_n$ be independent normal random variables (with possibly different means and/or variances). Then

$$T = a_1X_1 + a_2X_2 + \ldots a_nX_n$$

also follows the normal distribution.

What are the mean and the standard deviation of T?

- $E(T) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)$
- $\sigma_T^2 = a_1^2\sigma_{X_1}^2 + a_2^2\sigma_{X_2}^2 + \ldots + a_n^2\sigma_{X_n}^2$