MATH 189
Inference for the Mean: Preliminaries

Wenxin Zhou

UC San Diego

Time: 2:00—3:20 & 3:30—4:50pm TueThur

Location: CENTR 115

Inference for Multivariate Mean

- Statistical inference is the process of using sample data to analyze the properties of an underlying population probability distribution.

- In this lecture we consider the properties of the sample mean vector.

- We will also consider hypothesis testing problems on the population mean vector.

Linear Combinations of Random Variables

- In statistics, it is often of interest to investigate the linear combination of multiple random variables

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_p X_p = \sum_{j=1}^{p} c_j X_j = \boldsymbol{c}' \boldsymbol{X}.$$

- Here what we have is a set of coefficients c_1 through c_p that are multiplied by corresponding variables X_1 through X_p.

- The selection of the coefficients c_1 through c_p depend on the application of interest and the type of scientific questions we would like to address.

Example: USDA Women's Health Survey Data

- Suppose the variables in the dataset are:

  $X_1 =$ Calcium (mg), $X_2 =$ Iron (mg), $X_3 =$ Protein (g),
  $X_4 =$ Vitamin A (µg) and $X_5 =$ Vitamin C (mg).

- In addition to addressing questions about the individual nutritional component, we may wish to address questions about certain combinations of these components.

What is the total intake of vitamins A and C (in mg)?

Total intake of vitamins A and C (in mg)

X_1 = Calcium (mg), X_2 = Iron (mg), X_3 = Protein (g),
X_4 = Vitamin A (µg) and X_5 = Vitamin C (mg).

- Note that Vitamin A is measured in micrograms (μg), while Vitamin C is measured in milligrams (mg). 1 μg = 1 thousandth mg.
- So the total intake of the two vitamins, Y, can be expressed as

$$Y = 0.001X_4 + X_5.$$

- In this case, $c_1 = c_2 = c_3 = 0,\ c_4 = 0.001$ and $c_5 = 1$.

Example: Monthly Employment Data

- Suppose a dataset contains the following 6 variables about monthly employment:

$X_1 =$ # people laid off or fired, $X_2 =$ # of people resigned,
$X_3 =$ # of people retired, $X_4 =$ # of jobs created,
$X_5 =$ # of people hired, $X_6 =$ # of people entering the workforce

- We want to calculate the following variables as linear combinations of the above variables:
  - Net employment increase;
  - Net unemployment increase;
  - Unfilled jobs.

X_1 = # people laid off or fired, X_2 = # of people resigned,
X_3 = # of people retired, X_4 = # of jobs created,
X_5 = # of people hired, X_6 = # of people entering the workforce

- Net employment increase:

$$Y = X_5 - X_1 - X_2 - X_3$$

- Net unemployment increase:

$$Y = X_1 + X_2 + X_6 - X_5$$

- Unfilled jobs:

$$Y = X_4 - X_5$$

Descriptive Statistics of Linear Combination of Random Variables

- Linear combinations are functions of random quantities, and hence have population means and variances. Moreover, if we are looking at several linear combinations, they will have covariances and correlations as well.

- We are interested in knowing:
  1. What is the population mean of Y?
  2. What is the population variance of Y?
  3. What is the population covariance between two linear combinations Y_1 and Y_2?

Mean of Y

- The population mean of a linear combination is equal to the same linear combination of the population means of the component variables.

$$\text{If } Y = \sum_{j=1}^{p} c_j X_j, \text{ then } \mathbb{E}(Y) = \sum_{j=1}^{p} c_j \mu_j.$$

- We can estimate the population mean by replacing the population means with the corresponding sample means

$$\bar{Y} = \sum_{j=1}^{p} c_j \bar{X}_j.$$

Variance of Y

- The population variance of a linear combination is expressed as the following double sum over all pairs of variables

$$\mathrm{var}(Y) = \sum_{j=1}^{p} \sum_{k=1}^{p} c_j c_k \sigma_{jk} = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}.$$

- The population variance of Y can be estimated by the sample variance of Y

$$s_Y^2 = \sum_{j=1}^{p} \sum_{k=1}^{p} c_j c_k s_{jk} = \mathbf{c}'\mathbf{S}\mathbf{c}.$$

Covariance between Y_1 and Y_2

- Consider a pair of linear combinations

$$Y_1 = \sum_{j=1}^{p} c_j X_j \quad \text{and} \quad Y_2 = \sum_{k=1}^{p} d_k X_k$$

- The population covariance between Y_1 and Y_2 is obtained by summing over all pairs of variables

$$\mathrm{cov}(Y_1, Y_2) = \sigma_{Y_1,Y_2} = \sum_{j=1}^{p} \sum_{k=1}^{p} c_j d_k \sigma_{jk} = \mathbf{c}'\mathbf{\Sigma}\mathbf{d}.$$

- The population covariance can be estimated by the sample covariance

$$s_{Y_1,Y_2} = \sum_{j=1}^{p} \sum_{k=1}^{p} c_j d_k s_{jk} = \mathbf{c}'\mathbf{S}\mathbf{d}.$$

Correlation between Y_1 and Y_2

- Consider the pair of linear combinations

$$Y_1 = \sum_{j=1}^{p} c_j X_j \quad \text{and} \quad Y_2 = \sum_{k=1}^{p} d_k X_k$$

- The population correlation between Y_1 and Y_2 is defined as

$$\rho_{Y_1,Y_2} = \frac{\sigma_{Y_1,Y_2}}{\sigma_{Y_1}\sigma_{Y_2}}.$$

- The population correlation can be estimated by the sample correlation

$$r_{Y_1,Y_2} = \frac{s_{Y_1,Y_2}}{s_{Y_1}s_{Y_2}}.$$

Variance of Univariate Sample Mean

- As noted previously, sample mean $\bar{x}$ is also a random variable with a mean and a variance.
- We have discussed that the mean of sample mean $\mathbb{E}(\bar{x})$ equals the population mean μ.
- With some calculations, the variance of the sample mean, generated from independent samples of size n, is equal to the population variance, σ^2 divided by n.

$$\mathrm{var}(\bar{x}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(x_i) = \frac{\sigma^2}{n}.$$

Variance of Univariate Sample Mean (cont.)

- The population variance of sample mean is a function of unknown population parameter σ.
- To estimate the population variance of sample mean, we can replace the population parameter σ with sample standard deviation s,

$$\widehat{\mathrm{var}(\bar{x})} = \frac{s^2}{n}.$$

- The square root of this quantity is called the standard error of the mean

$$\mathrm{se}(\bar{x}) = \frac{s}{\sqrt{n}}.$$

Standard Error of Sample Mean

- Standard error of sample mean is a measure of the uncertainty of our estimate of the population mean.
- If the standard error is large, then we are ***less confident*** of our estimate of the mean.
- If the standard error is small, then we are ***more confident*** of our estimate of the mean.
- What is meant by large or small depends on the application at hand.
- In any case, the standard error is a decreasing function of sample size, the larger our sample is the more confident we can be of our estimate.

Variance of Sample Mean Vector

- In the multivariate setting, the sample mean is a random vector $\bar{\boldsymbol{x}}$.
- We have discussed that the mean of sample mean vector $\mathbb{E}(\bar{\boldsymbol{x}})$ equals the population mean vector $\boldsymbol{\mu}$ (unbiased).
- The population variance-covariance matrix of sample mean vector, generated from independent samples of size n, is

$$\operatorname{var}(\bar{\boldsymbol{x}}) = \frac{1}{n}\boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is the population variance-covariance matrix of $\boldsymbol{x}_i$.

Variance of Sample Mean Vector (cont.)

- The population variance-covariance matrix of sample mean vector is a function of $\boldsymbol{\Sigma}$.

- To estimate the population variance-covariance matrix of sample mean vector, we replace $\boldsymbol{\Sigma}$ with sample variance-covariance matrix $\mathbf{S}$:

$$\widehat{\mathrm{var}(\bar{\mathbf{x}})} = \frac{1}{n}\mathbf{S}.$$

Distribution of Univariate Sample Mean

- Suppose $x_1, x_2, \ldots, x_n$ are independently sampled from a normal distribution with mean μ and variance σ^2.
- In this case, the sample mean $\bar{x}$ is normally distributed as

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- This conclusion depends on the iid (independent and identically distributed) normal assumption.

- Can you see its connection to the unbiasedness and variance of the sample mean?

Distribution of Sample Mean Vector

- Suppose $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ are independently sampled from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.
- In this case, the sample mean $\overline{\boldsymbol{x}}$ follows a multivariate normal distribution:

$$\overline{\boldsymbol{x}} \sim N\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right).$$

- Again, the above argument depends on the iid normal assumption.

- Can you see its connection to the unbiasedness and variance-covariance matrix of sample mean vector?

What if the Population is Not Normal?

- The previous results depend on the assumption that the observation is sampled from a normal distribution.

- This can be an idealization from reality. The distribution of population is usually unknown to us, and deviates far away from normal.

- What is the distribution of sample mean or sample mean vector when the observations are NOT sampled from a normal distribution?

Central Limit Theorem (Univariate Case)

- If the observations $x_1, x_2, \ldots, x_n$ are independently and identically sampled from a population with mean μ and variance $\sigma^2 < \infty$, then, the sample mean, $\bar{x}$, is **approximately normally** distributed with mean μ and variance σ^2/n.

- In other words, if the above conditions are satisfied, the following linear transformation of sample mean converges to a normal distribution with mean zero and variance σ^2:

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2) \text{ as } n \to \infty.$$

How to Understand CLT?

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2) \text{ as } n \to \infty.$$

- The assumption that the population is normally distributed is removed.
- The sample mean is **approximately normally distributed.**
- The convergence rate is $1/\sqrt{n}$. The error between $\bar{x}$ and μ is a random variable whose mean is of order $1/\sqrt{n}$.
- The accuracy of normal approximation increases as the sample size n increases.

Central Limit Theorem (Multivariate Case)

- If p-dimensional observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ are independently and identically sampled from a population with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

- Then, the sample mean vector $\overline{\boldsymbol{x}}$ converges to a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}/n$:

$$\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$