Multivariate Probability Distributions

Professor Yongmiao Hong
Cornell University
April 16, 2020

CONTENTS

Population and Random Sample

- **Statistical analysis** is based on outcomes of a large number of repeated **random experiments** of same or similar kind.

- Suppose a random variable X_i denotes the outcome of the i-th experiment. We then obtain a sequence of outcomes, $X_1, \cdots, X_n$, if n experiments are implemented.

- This sequence of outcomes then constitutes a so-called **random sample** from which one can make inference of the underlying **probability law** which has generated the observed data.

Population and Random Sample

Definition 1 (6.1). [Random Sample]

A random sample, denoted as $\mathbf{X}^n = (X_1, \cdots, X_n)$, is a sequence of n random variables $X_1, \cdots, X_n$.

A **realization** of the random sample $\mathbf{X}^n$, denoted as $\mathbf{x}^n = (x_1, \cdots, x_n)$, is called a **data** set generated from $\mathbf{X}^n$ or a sample point of $\mathbf{X}^n$.

A random sample $\mathbf{X}^n$ can generate many different data sets. The collection of all possible sample points of $\mathbf{X}^n$ constitutes the **sample space** of the random sample $\mathbf{X}^n$.

Population and Random Sample

Example 1 (6.1). [Throwing *n* Coins]

Let X_i denote the outcome of throwing the i-th coin, with $X_i = 1$ if the head shows up, and $X_i = 0$ if the tail shows up. Then $\mathbf{X}^n = (X_1, \cdots, X_n)'$ constitutes a random sample.
If we throw n coins, we will obtain a sequence of real numbers, such as

$$\mathbf{x}^n = (1, 1, 0, 0, 1, 0, \cdots, 1).$$

This sequence is a data set of size n from the random sample $\mathbf{X}^n$.

Population and Random Sample

Example 1 (6.1). [Throwing *n* Coins]

Obviously, if we throw the n coins again, we will get a different sequence, such as

$$\mathbf{x}^n = (1, 0, 0, 1, 1, 1, \cdots, 0).$$

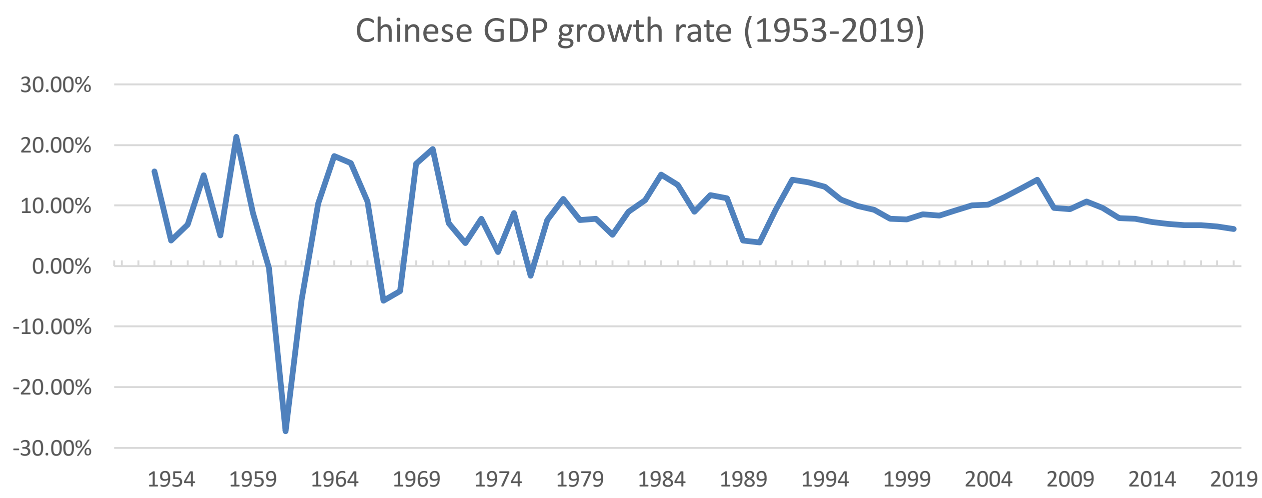
This is another data set from the random sample $\mathbf{X}^n$. Apparently, the random sample $\mathbf{X}^n$ can generate a total of 2^n different data sets, each with size n.

Population and Random Sample

Example 2 (6.2). [Chinese GDP Annual Growth

Let X_i denote the Chinese GDP growth rate in year i, from 1953 to 2019. Then $\mathbf{X}^n = (X_1, \cdots, X_n)'$ constitutes a random sample with sample size $n = 68$. The observed data $\mathbf{x}^n = (x_1, \cdots, x_n)'$, depicted in Figure 6.1, is a realization of $\mathbf{X}^n$.

Figure 6.1

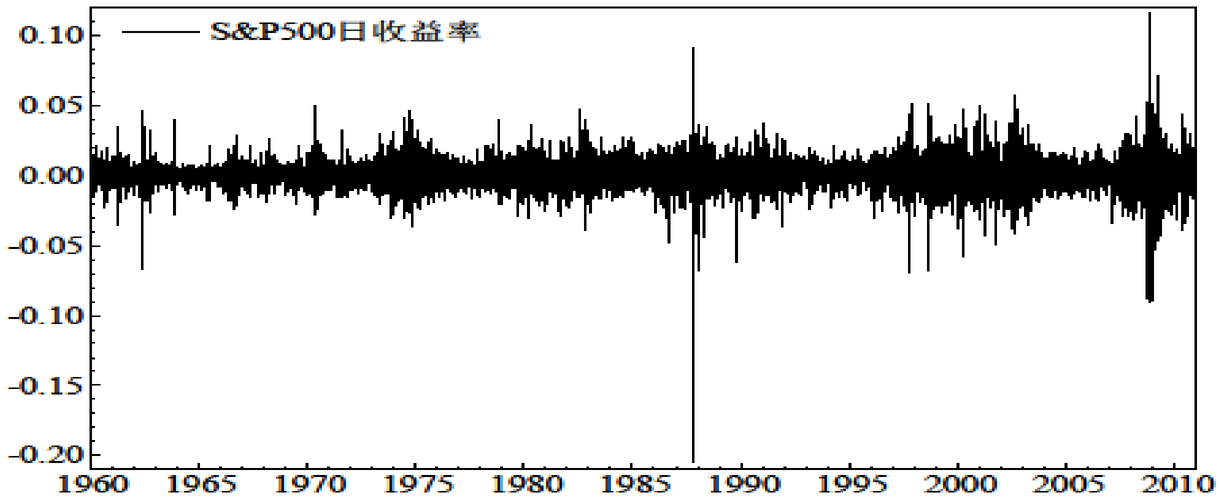


Population and Random Sample

Example 3 (6.3). [S&P 500 price index]

Let X_i be the return on S&P500 price index at day i, from January 4, 1960 to December 31, 2010. Then $\mathbf{X}^n = (X_1, \cdots, X_n)$ forms a random sample with size $n = 12839$.
The observed data set $\mathbf{x}^n = (x_1, \cdots, x_n)$, depicted in Figure 6.2, is a realization of the random sample $\mathbf{X}^n$.

Figure 6.2



Population and Random Sample

Remarks:

- While in theory a random sample $\mathbf{X}^n$ could generate many different data sets $\mathbf{x}^n$, each with size n, one may only observe or obtain one data set $\mathbf{x}^n$ in practice. This is the case with Examples 6.2 and 6.3 which are called time series data.

- For example, if we would like to obtain another data set for the Chinese GDP growth rate, we would have to let the Chinese economy repeat again back from 1953, and this is simply impossible due to the non-experimental nature of a real economy.

Population and Random Sample

- In statistical analysis, we still assume that the only observed data in Example 6.2 or Example 6.3 is one of many possible realizations from the random sample $\mathbf{X}^n$.

- For some random samples, the order of the random variables $X_1, \cdots, X_n$ in the sample, together with their realizations, may not be altered freely.

- An example is the **time series random sample** of Examples 6.2, where the random variables $X_1, \cdots, X_n$ are not jointly independent, and the behavior of X_i may depend on the previous outcomes $\{X_{i-1}, X_{i-2}, \cdots\}$. Such a dynamic structure could not be preserved if one altered the order of random variables and their realizations.

Population and Random Sample

- A **random sample** $\mathbf{X}^n$ can be viewed as a n-**dimensional random vector**, namely, $\mathbf{X}^n : S \to \mathbb{R}^n$, where S is the sample space of the underlying random experiment.

- The information of a random sample $\mathbf{X}^n$ is completely described by the joint PMF/PDF of the n random variables,

$$f_{\mathbf{X}^n}(\mathbf{x}^n) = \prod_{i=1}^{n} f_{X_i|\mathbf{X}^{i-1}}(x_i|\mathbf{x}^{i-1}),$$

  where, by convention, $f_{X_1|\mathbf{X}^0}(x_1|x^0) = f_{X_1}(x_1)$ is the marginal PMF/PDF of random variable X_1.

- The joint PMF/PDF can be used to calculate probabilities involving the random sample $\mathbf{X}^n$.

Population and Random Sample

- The above definition of a random sample covers both independent samples and time series samples:
  - For the former, $X_1, \cdots, X_n$ in the sample are jointly independent;
  - For the latter, $X_1, \cdots, X_n$ in the sample are not jointly independent.

Population and Random Sample

Definition 2 (6.2). [IID Random Sample]

The sequence $\{X_1, \cdots, X_n\}$ is called an independent and identically distributed (IID) **random sample** of size n from the **population distribution** $F_X(x)$ if:

(1) random variables $X_1, \cdots, X_n$ are mutually *independent*;

(2) each random variable X_i has the same marginal distribution $F_X(x)$.

Population and Random Sample

Question: What is the interpretation and implication of an IID random sample?

- Suppose we have a random experiment in which the variable of interest X has a probability distribution $F_X(x)$.
- Suppose the random experiment is repeated n times. Then we observe n outcomes for the variable of interest, denoted as $\mathbf{x}^n = (x_1, \cdots, x_n)$.

Population and Random Sample

Question: What is the interpretation and implication of an IID random sample?

- If we denote X_i as the variable of interest associated with the i-th experiment, then X_i has the probability distribution $F_X(x)$ and x_i can be viewed as a realization of X_i.

- **Identical distribution** for the X_i means repeated experiments of **same kind**, and **independence** means that experiments are implemented independently so that new information can be obtained from each experiment.

Population and Random Sample

- The main purpose of **statistical analysis** is to **infer population distribution** $F_X(x)$ based on **an observed data set** generated from a large number of repeated experiments of the same kind.

Population and Random Sample

Question: How to define the population if the random variables $X_1, \cdots, X_n$ in the sample are not identically distributed?

- The random variables $X_1, \cdots, X_n$ in a random sample may not have identical probability distributions, due to the existence of **heterogeneity** among economic agents or **structural changes** of economic relationships over time.

- Although each X_i has a different distribution, we may assume that they may still share certain **common features** (e.g., common parameter values) in their probability distributions, and these common features of distributions can be defined as the population.

Population and Random Sample

- **Inference of population based on a random sample** is the most important feature of statistical analysis.

Question:

- What are the requirements on the random sample?
- What is the best inference method given a random sample?
- What should be done if the random sample have certain drawbacks (e.g., sample selection bias, missing data, outliers, etc)?

Population and Random Sample

Question: How to summarize useful information in a data set $\mathbf{x}^n$? What is a tool to do so?

Population and Random Sample

Definition 3 (6.3). [Statistic]

Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be a random sample of size n from a population.
A *statistic* $T(\mathbf{X}^n) = T(X_1, \cdots, X_n)$ is a real-valued or vector-valued function of a random sample $\mathbf{X}^n$.

Population and Random Sample

Remarks:

- The function $T(\cdot)$ is a mapping from the n-dimensional sample space of $\mathbf{X}^n$ to a **low-dimensional** Euclidean space.
- A statistic $T(\mathbf{X}^n)$ **does not involve any unknown parameter**. It is entirely a function of random sample $\mathbf{X}^n$. Given any data set $\mathbf{x}^n$, we can obtain a real-valued number or vector for the statistic $T(\mathbf{X}^n)$.

Population and Random Sample

- A statistic $T(\mathbf{X}^n)$ can be used to **effectively summarize some features of data** (e.g., maximum and minimum values, median, mean, standard deviation, etc), to **estimate unknown parameters**, to conduct **hypothesis testing**, etc.

- **Interpretability** of statistics is very important!

Population and Random Sample

Example 4 (6.4)

Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be a random sample. Then the sample mean

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$$

and the sample variance

$$S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

are two statistics.

Population and Random Sample

Question: $\bar{X}_n$ and S_n^2 can be used to estimate μ_X and σ_X^2 of the population distribution $F_X(x)$. Why are $\bar{X}_n$ and S_n^2 "good" estimators of μ_X and σ_X^2 respectively?

- We will develop various concepts to measure the closeness of an estimator to the parameter of interest in Chapters 7 and 8.

Population and Random Sample

Example 5 (6.5)

Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be an IID random sample from the population $f(x, \theta)$, where θ is some unknown parameter. Then the logarithm of the joint PMF/PDF of $\mathbf{X}^n$

$$\begin{aligned}\hat{L}(\theta|\mathbf{X}^n) &= \ln \prod_{i=1}^{n} f(X_i, \theta) \\ &= \sum_{i=1}^{n} \ln f(X_i, \theta)\end{aligned}$$

is called the log-likelihood function of θ, conditional on the random sample $\mathbf{X}^n$.

Population and Random Sample

Example 5 (6.5)

Remarks

$\hat{L}(\theta|\mathbf{X}^n)$ depends on the random sample $\mathbf{X}^n$, but it is not a statistic, because it is a function of the unknown parameter θ.

Population and Random Sample

Definition 4 (6.4). [Sampling Distribution]

The probability distribution of a statistic $T(\mathbf{X}^n)$ is called the *sampling distribution* of $T(\mathbf{X}^n)$.

Remarks

- Since $T(\mathbf{X}^n)$ is a function of n random variables, $T(\mathbf{X}^n)$ itself is a low-diemensional random vector.
- The distribution of $T(\mathbf{X}^n)$ is called the sampling distribution because this distribution can be derived from the joint distribution of the variables $X_1, \cdots, X_n$ in the random sample.

Population and Random Sample

- The sampling distribution of $T(\mathbf{X}^n)$ is different from the population distribution $F_X(x)$. The latter is the marginal distribution of each X_i in an IID random sample $\mathbf{X}^n$.
- The sampling distribution of a statistic $T(\mathbf{X}^n)$ plays a vital role in statistical inference. For example, it is needed to obtain critical values when constructing a confidence interval estimator and a hypothesis test statistic.
- $T(\mathbf{X}^n)$ can be viewed as a **partition of** the **sample space** of $\mathbf{X}^n$. A random sample $\mathbf{X}^n$ can generate many data sets $\mathbf{x}^n$, each of which is called a sample point in the sample space of $\mathbf{X}^n$. Let

$$A(t) = \{\mathbf{x}^n : T(\mathbf{x}^n) = t\}$$

be the collection of all sample points $\mathbf{x}^n$ that satisfy the restriction $T(\mathbf{x}^n) = t$. Then a single value of $T(\mathbf{x}^n) = t$ summarizes all sample points in $A(t)$ which give the same value for $T(\mathbf{x}^n)$.

CONTENTS

Sampling Distribution of Sample Mean

Definition 5 (6.5). [Sample Mean]

Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is a random sample from a population with mean μ and variance σ^2. Then

$$T(\mathbf{X}^n) \equiv \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the sample mean for the random sample $\mathbf{X}^n$.

Sampling Distribution of Sample Mean

Remarks:

- The distribution of $\bar{X}_n$ is called the sampling distribution of $\bar{X}_n$.
- When one has only a single observed sample (i.e., data set) $\mathbf{x}^n$, the sample mean $\bar{x}_n$ does not appear random. However, if we realize that the observed sample $\mathbf{x}^n$ is only one of many possible samples that could have been drawn and each sample has a different sample mean, we can then see that the sample mean is in fact random.

Sampling Distribution of Sample Mean

Theorem 1 (6.1)

Suppose $\mathbf{X}^n$ is a random sample. Then

$$\bar{X}_n = \arg \min_{-\infty < a < \infty} \sum_{i=1}^{n} (X_i - a)^2.$$

Sampling Distribution of Sample Mean

Remarks:

- The objective function $\sum_{i=1}^{n}(X_i - a)^2$ is called the sum of squared residuals.
- The sample mean $\bar{X}_n$ is essentially the Ordinary Least Squares (OLS) estimator for a very simple linear regression model

$$X_i = a + \varepsilon_i,$$

where $\{\varepsilon_i\}$ is an IID sequence with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$.

Sampling Distribution of Sample Mean

- We shall investigate the following statistical properties of $\bar{X}_n$:
  - What is the mean of $\bar{X}_n$?
  - What is the variance of $\bar{X}_n$?
  - What is the sampling distribution of $\bar{X}_n$?

Sampling Distribution of Sample Mean

Theorem 2 (6.2)

Suppose $X_1, \cdots, X_n$ are a sequence of n identically distributed random variables with the same population mean μ. Then for all $n \geq 1$,

$$E(\bar{X}_n) = \mu.$$

Proof:

$$\begin{aligned} E(\bar{X}_n) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu. \end{aligned}$$

Sampling Distribution of Sample Mean

Remarks:

- This result does not require that the random variables $X_1, \cdots, X_n$ be mutually independent.
- $E(\bar{X}_n) = \mu$ implies that the sample mean estimator $\bar{X}_n$ does not make a systematic mistake in estimating the population mean μ. That is, for any given n, if one generates a large number of data sets $\mathbf{x}^n$, each of which gives a value $\bar{x}_n$ for $\bar{X}_n$, then the average of these sample mean values will be arbitrarily close to μ.

Sampling Distribution of Sample Mean

Theorem 3 (6.3)

Suppose $\mathbf{X}^n$ is an IID random sample from a population with mean μ and variance σ^2. Then for all $n \geq 1$,

$$\operatorname{var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Proof

Sampling Distribution of Sample Mean

Proof:

- When X and Y are mutually independent, we have

$$\begin{aligned} \text{var}(a + bX + cY) &= b^2\sigma_X^2 + c^2\sigma_Y^2 + 2bc \cdot \text{cov}(X, Y) \\ &= b^2\sigma_X^2 + c^2\sigma_Y^2. \end{aligned}$$

- Similarly, for an IID random sample $\mathbf{X}^n$, we have

$$\begin{aligned} \text{var}(\bar{X}_n) &= \text{var}\left(\sum_{i=1}^{n} n^{-1} X_i\right) \\ &= \sum_{i=1}^{n} n^{-2}\text{var}(X_i) \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Sampling Distribution of Sample Mean

Remarks:

- The variance σ^2/n of $\bar{X}_n$ is different from the population variance σ^2 of each random variable X_i.
- $\text{var}(\bar{X}_n) = \sigma^2/n$ implies that the dispersion of $\bar{X}_n$ from its center $E(\bar{X}_n)$ shrinks to zero as $n \to \infty$.
- Since $E(\bar{X}_n) = \mu$, we have the mean squared error of $\bar{X}_n$

$$\begin{aligned} E(\bar{X}_n - \mu)^2 &= \text{var}(\bar{X}_n) \\ &= \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty. \end{aligned}$$

Sampling Distribution of Sample Mean

Example 6 (6.6). [Idiosyncratic Risk Elimination via Diversification]

According to the standard capital asset pricing model (CAPM), the return of asset i over certain holding period is given as:

$$R_i = \alpha + \beta_i R_m + \varepsilon_i,$$

where α is a constant representing the return on the risk-free asset, R_m is the market risk factor common to all individual assets, β_i is a factor loading coefficient, and ε_i represents an idiosyncratic risk associated with asset i. It is further assumed that the sequence of $(\varepsilon_1, \cdots, \varepsilon_n)$ is IID with mean 0 and variance σ^2, and is uncorrelated with the market risk factor R_m. The risk of asset i, as measured by its variance, is given by

$$\mathrm{var}(R_i) = \beta_i^2 \mathrm{var}(R_m) + \sigma^2,$$

Sampling Distribution of Sample Mean

Example 6 (6.6). [Idiosyncratic Risk Elimination via Diversification]

where $\beta_i^2 \text{var}(R_m)$ is a systematic risk which cannot be avoided, and σ^2 is the idiosyncratic risk which can be eliminated by forming a portfolio with a large number of assets.

To see this, consider the return on an equal-weighting portfolio with n assets:

$$\begin{aligned}\bar{R}_n &= \sum_{i=1}^{n} \frac{1}{n} R_i \\ &= \alpha + \bar{\beta}_n R_m + \bar{\varepsilon}_n,\end{aligned}$$

Sampling Distribution of Sample Mean

Example 6 (6.6). [Idiosyncratic Risk Elimination via Diversification]

where the average beta $\bar{\beta}_n = n^{-1}\sum_{i=1}^{n}\beta_i \to \beta \neq 0$ as $n \to \infty$, and $\bar{\varepsilon}_n = n^{-1}\Sigma_{i=1}^{n}\varepsilon_i$ is the sample mean of the individual risk sample $(\varepsilon_1, \cdots, \varepsilon_n)$. It follows that

$$\begin{aligned} \mathrm{var}(\bar{R}_n) &= \bar{\beta}_n^2 \mathrm{var}(R_m) + \frac{\sigma^2}{n} \\ &\to \beta^2 \mathrm{var}(R_m) \text{ as } n \to \infty. \end{aligned}$$

Thus, the idiosyncratic risks associated with individual assets can be eliminated by including a very large number n of assets contained in the portfolio.

Sampling Distribution of Sample Mean

Theorem 4 (6.4)

Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an IID normally distributed random sample with population mean μ and population variance $\sigma^2 < \infty$. Define the standardized sample mean

$$
\begin{aligned}
Z_n &= \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\text{var}(\bar{X}_n)}} \\
&= \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \\
&= \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.
\end{aligned}
$$

Then

$$Z_n \sim N(0, 1) \text{ for all } n \geq 1.$$

Proof

Sampling Distribution of Sample Mean

Proof:

- Put $Y_i = (X_i - \mu)/\sigma$. Then $Y_i \sim N(0,1)$ and has MGF

$$M_{Y_i}(t) = e^{\frac{1}{2}t^2} \text{ for all } i.$$

- Now consider the MGF of $Z_n = n^{-1/2}\sum_{i=1}^{n} Y_i$:

$$\begin{aligned}
M_{Z_n}(t) &= E\left(e^{tZ_n}\right) \\
&= E\left(e^{tn^{-\frac{1}{2}}\Sigma_{i=1}^{n} Y_i}\right) \\
&= E\left(\prod_{i=1}^{n} e^{tn^{-\frac{1}{2}}Y_i}\right) \\
&= \prod_{i=1}^{n} E\left(e^{tn^{-\frac{1}{2}}Y_i}\right) \\
&= \prod_{i=1}^{n} M_{Y_i}\left(tn^{-\frac{1}{2}}\right) \\
&= \left[e^{\frac{1}{2}(tn^{-\frac{1}{2}})^2}\right]^n \\
&= e^{\frac{1}{2}t^2}.
\end{aligned}$$

It follows that $Z_n \sim N(0,1)$ for all $n \geq 1$.

Sampling Distribution of Sample Mean

Remarks:

- The sum of n independent normal random variables is still a normal variable. This is called the **reproductive property** of the normal distribution.
- When the random sample $\mathbf{X}^n$ is not from a normal population, $\bar{X}_n$ and Z_n no longer follow a normal distribution. For example, in Example 6.1, $n\bar{X}_n$ follows a Binomial(n, p) distribution for any given n.

CONTENTS

Sampling Distribution of Sample Variance

Question:
How to estimate $\sigma^2 = \text{var}(X_i)$?

- Recall the variance formula

$$\sigma^2 = E(X_i - \mu)^2,$$

one plausible estimator for σ^2 might be the sample average

$$n^{-1} \sum_{i=1}^{n} (X_i - \mu)^2.$$

- Since μ is unknown, we shall replace μ with the sample mean $\bar{X}_n$ and the average of $(X_i - \bar{X}_n)^2$:

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

- In fact, we will use the sample variance estimator

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Sampling Distribution of Sample Variance

Question: Why dividing by $n-1$?

- What is the mean of S_n^2?
- What is the variance of S_n^2?
- What is the sampling distribution of S_n^2?

Sampling Distribution of Sample Variance

Theorem 5 (6.5)

Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an IID random sample from a population with (μ, σ^2). Then for all $n > 1$,

$$E(S_n^2) = \sigma^2.$$

Proof

Sampling Distribution of Sample Variance

Proof:

- Using the formula $(a-b)^2 = a^2 - 2ab + b^2$, we have

$$\begin{aligned}
&\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\
= &\sum_{i=1}^{n}[(X_i - \mu) - (\bar{X}_n - \mu)]^2 \\
= &\sum_{i=1}^{n}(X_i - \mu)^2 - 2\sum_{i=1}^{n}(X_i - \mu)(\bar{X}_n - \mu) + \sum_{i=1}^{n}(\bar{X}_n - \mu)^2 \\
= &\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X}_n - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\bar{X}_n - \mu)^2 \\
= &\sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\bar{X}_n - \mu)^2 + n(\bar{X}_n - \mu)^2 \\
= &\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X}_n - \mu)^2,
\end{aligned}$$

To be Continued

Sampling Distribution of Sample Variance

Proof: where we have used the fact

$$\sum_{i=1}^{n}(X_i-\mu)=n(\bar{X}_n-\mu).$$

Taking the expectations for both sides, we have

$$\begin{aligned} E\sum_{i=1}^{n}(X_i-\bar{X}_n)^2 &= \sum_{i=1}^{n}E(X_i-\mu)^2-nE[(\bar{X}_n-\mu)^2] \\ &= n\sigma^2-n\cdot\frac{\sigma^2}{n}=(n-1)\sigma^2, \end{aligned}$$

where we have used the fact that $E(\bar{X}_n-\mu)^2=\sigma^2/n$.

- It follows that

$$\begin{aligned} E(S_n^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2\right] \\ &= \sigma^2. \end{aligned}$$

Sampling Distribution of Sample Variance

Remarks:

- It is important to assume independence among the n random variables $X_1, ..., X_n$ here, because we have used the fact that $E(\bar{X}_n - \mu)^2 = \sigma^2/n$.
- The reason of using $n-1$ instead of n is to ensure that S_n^2 is unbiased for σ^2.

Sampling Distribution of Sample Variance

Lemma 1 (6.6). [■-Distribution]

Let $Z_1, \cdots, Z_\nu$ be IID $N(0,1)$ random variables, where ν is a positive integer. Then

$$\sum_{i=1}^{\nu} Z_i^2 \sim \chi_\nu^2.$$

Proof

Sampling Distribution of Sample Variance

Proof:

- When $Z_i \sim N(0,1)$, we have $Z_i^2 \sim \chi_1^2$, whose MGF

$$M_{Z_i^2}(t) = (1-2t)^{-\frac{1}{2}}.$$

- Put $X = \Sigma_{i=1}^{\nu} Z_i^2$. Then given the independence among $Z_1, \cdots, Z_\nu$, we have

$$\begin{aligned} M_X(t) &= E\left(e^{t\Sigma_{i=1}^{\nu} Z_i^2}\right) \\ &= \prod_{i=1}^{\nu} E\left(e^{tZ_i^2}\right) \\ &= (1-2t)^{-\frac{\nu}{2}}. \end{aligned}$$

- It follows that $X \sim \chi_\nu^2$ by the uniqueness of the MGF. This is called the **reproductive property** of the χ^2 distribution.

Sampling Distribution of Sample Variance

Remarks:

- The χ^2_ν distribution has

$$E\left(\chi^2_\nu\right) = \nu$$

and

$$\operatorname{var}\left(\chi^2_\nu\right) = 2\nu.$$

Sampling Distribution of Sample Variance

Theorem 7 (6.7). [■ -Distribution]

Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an IID $N(\mu, \sigma^2)$ random sample. Then for each $n > 1$,

$$\begin{aligned} \frac{(n-1)S_n^2}{\sigma^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \\ &\sim \chi^2_{n-1}, \end{aligned}$$

where χ^2_{n-1} is a Chi-square distribution with $n-1$ degrees of freedom.

Proof

Sampling Distribution of Sample Variance

Proof:

- It is straightforward to establish the recursive relation

$$(n-1)S_n^2 = (n-2)S_{n-1}^2 + \frac{n-1}{n}(X_n - \bar{X}_{n-1})^2.$$

We shall show the theorem by induction:

- (1) We first consider $n = 2$, the minimum interger for S_n^2. We have

$$\begin{aligned} \frac{(2-1)S_2^2}{\sigma^2} &= \frac{1}{2\sigma^2}(X_2 - X_1)^2 \\ &= \left(\frac{X_2 - X_1}{\sqrt{2}\sigma}\right)^2 \\ &\sim \chi_1^2 \end{aligned}$$

because $(X_2 - X_1)/\sqrt{2}\sigma \sim N(0,1)$.

To be Continued

Sampling Distribution of Sample Variance

Proof:

- (2) Next, suppose for $n = \nu$, an arbitrary positive integer with $\nu > 1$, we have $(\nu-1)S_\nu^2/\sigma^2 \sim \chi^2_{\nu-1}$. Then we shall show that for $n = \nu+1$, $\nu S^2_{\nu+1}/\sigma^2 \sim \chi^2_\nu$.
- For $n = \nu + 1$, we have

$$\frac{\nu S^2_{\nu+1}}{\sigma^2} = \frac{(\nu-1)S_\nu^2}{\sigma^2} + \frac{\nu}{(\nu+1)\sigma^2}(X_{\nu+1} - \bar{X}_\nu)^2.$$

  - We now consider the second term. Since $X_{\nu+1} \sim N(\mu, \sigma^2)$, $\bar{X}_\nu \sim N(\mu, \frac{1}{\nu}\sigma^2)$, and $X_{\nu+1}$ and $\bar{X}_\nu$ are independent, we have

$$X_{\nu+1} - \bar{X}_\nu \sim N\left(0, \sigma^2 + \frac{\sigma^2}{\nu}\right)$$

To be Continued

Sampling Distribution of Sample Variance

Proof:

or equivalently

$$\sqrt{\frac{\nu}{(\nu+1)\sigma^2}}\,(X_{\nu+1}-\bar{X}_\nu) \sim N(0,1).$$

Hence, $\frac{\nu}{\nu+1}(X_{\nu+1}-\bar{X}_\nu)^2/\sigma^2 \sim \chi^2_1$.

- Suppose this term is independent of S^2_ν. Then, given $(\nu-1)S^2_\nu/\sigma^2 \sim \chi^2_{\nu-1}$ and the fact that the sum of two independent χ^2 random variables follow a χ^2 distribution, we have $\nu S^2_{\nu+1}/\sigma^2 \sim \chi^2_\nu$.
- The theorem will thus be proved provided the following result is shown:

Sampling Distribution of Sample Variance

Theorem 8 (6.8)

Suppose $\mathbf{X}^n$ is an IID $N(\mu, \sigma^2)$ random sample. Then for any $n > 1$, S_n^2 and $\bar{X}_n$ are mutually independent.

Sampling Distribution of Sample Variance

Lemma 2 (6.9)

Let $X_i \sim$ IID $N(\mu, \sigma^2)$, $i = 1, \cdots, n$. For constants a_{ij}, and b_{rj}, define

$$
\begin{aligned}
U_i &= \sum_{j=1}^{n} a_{ij} X_j, \ i = 1, \cdots, \nu, \\
V_r &= \sum_{j=1}^{n} b_{rj} X_j, \ r = 1, \cdots, m,
\end{aligned}
$$

where $\nu + m \leq n$. Then
(1) For each pair (i, r), the random variables U_i and V_r are independent if and only if $\text{cov}(U_i, V_r) = 0$.
(2) The random vectors $(U_1, \cdot\cdot\cdot, U_\nu)$ and $(V_1, \cdot\cdot\cdot, V_m)$ are independent if and only if U_i is independent of V_r for all pairs (i, r), where $i = 1, \cdot\cdot\cdot, \nu, r = 1, \cdot\cdot\cdot, m$.

Sampling Distribution of Sample Variance

Remarks:

- The U_i random variables and the V_r random variables follow a joint normal distribution.
- Under the joint normal distribution, the U_i random variables and the V_r random variables are independent if and only of their covariances are zero for all pairs of i, r.

Sampling Distribution of Sample Variance

Proof of Theorem 6.8:

- Note that $S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ is a function of n random variables $(X_1 - \bar{X}_n), \cdots, (X_n - \bar{X}_n)$. It suffices to show that $\bar{X}_n$ and $(X_1 - \bar{X}_n, \cdots, X_n - \bar{X}_n)$ are mutually independent.
- We apply Lemma 6.9. Put $U_1 = \bar{X}_n - \mu$, and $V_r = X_r - \bar{X}_n$, $r = 1, \cdots, n$. We first show that U_1 and V_r are mutually independent for all $r = 1, \cdots, n$.

To be Continued

Sampling Distribution of Sample Variance

Proof of Theorem 6.8:

- Because for any given $r = 1, \cdots, n$, we have

$$\begin{aligned} \operatorname{cov}(U_1, V_r) &= E(U_1 V_r) \\ &= E\left[(\bar{X}_n - \mu)(X_r - \mu))\right] - E\left(\bar{X}_n - \mu\right)^2 \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} \\ &= 0. \end{aligned}$$

It follows from Lemma 6.9(1) that U_1 and V_r are independent. We have immediately from Lemma 6.9(2) that U_1 and $(V_1, \cdots, V_n)$ are mutually independent.

To be Continued

Sampling Distribution of Sample Variance

Proof of Theorem 6.8:

- Now, put $g(U_1) = U_1 + \mu$, and $h(V_1, \cdots, V_n) = (n-1)^{-1}\sum_{r=1}^{n} V_r^2$. Then $g(U_1)$ and $h(V_1, \cdots, V_n)$ are independent, i.e. $\bar{X}_n$ and S_n^2 are independent.

Sampling Distribution of Sample Variance

Another Heuristic Proof of Theorem 6.8:

- Let $\mathbf{X} = (X_1, \cdots, X_n)'$ be a $n \times 1$ vector, $\mathbf{i} = (1, \cdots, 1)'$ be a $n \times 1$ vector of ones, and $\mathbf{I}$ be a $n \times n$ identity matrix, where $\mathbf{A}'$ denotes the transpose of a vector or matrix $\mathbf{A}$.
- Define a $n \times n$ matrix

$$\mathbf{M} = \mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}'.$$

Note that $\mathbf{M}^2 = \mathbf{M}$ and $\mathbf{M}' = \mathbf{M}$. Then we have

$$\begin{aligned} n\bar{X}_n &= \mathbf{i}'\mathbf{X}, \\ (n-1)S_n^2 &= (\mathbf{MX})'(\mathbf{MX}) \\ &= \mathbf{X}'\mathbf{M}^2\mathbf{X} \\ &= \mathbf{X}'\mathbf{MX}. \end{aligned}$$

To be Continued

Sampling Distribution of Sample Variance

Another Heuristic Proof of Theorem 6.8:

- To show that $\bar{X}_n$ and S_n^2 are independent, it suffices to show the random variable $\mathbf{i}'\mathbf{X}$ and the $n \times 1$ random vector $\mathbf{MX}$ are independent.
- Put

$$
\begin{aligned}
\mathbf{Z} &= \begin{pmatrix} \mathbf{i}'\mathbf{X} \\ \mathbf{MX} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{i}' \\ \mathbf{M} \end{pmatrix} \mathbf{X} \\
&= \mathbf{AX}, \text{ say},
\end{aligned}
$$

where $\mathbf{A}$ is a $(n+1) \times n$ matrix.

To be Continued

Sampling Distribution of Sample Variance

Another Heuristic Proof of Theorem 6.8:

- Because $\mathbf{Z}$ is a linear combination of $\mathbf{X}$, and $\mathbf{X} \sim N(0, \sigma^2\mathbf{I})$ is a vector of IID normal random variables, $\mathbf{Z}$ follows a multivariate normal distribution.
- Furthermore, the variance-covariance matrix between $\mathbf{i}'\mathbf{X}$ and $\mathbf{MX}$

$$\begin{aligned}
\text{cov}(\mathbf{i}'\mathbf{X}, \mathbf{MX}) &\equiv E\left\{[\mathbf{i}'\mathbf{X} - E(\mathbf{i}'\mathbf{X})]\,[\mathbf{MX} - E(\mathbf{MX})]'\right\} \\
&= E\left\{\mathbf{i}'\,[\mathbf{X} - E(\mathbf{X})]\,[\mathbf{X} - E(\mathbf{X})]'\,\mathbf{M}'\right\} \\
&= \mathbf{i}'E\left\{[\mathbf{X} - E(\mathbf{X})]\,[\mathbf{X} - E(\mathbf{X})]'\right\}\mathbf{M} \\
&= \mathbf{i}'\sigma^2\mathbf{IM} \\
&= \sigma^2\mathbf{i}'\mathbf{M} \\
&= \mathbf{0}
\end{aligned}$$

given $\mathbf{i}'\mathbf{M} = \mathbf{0}$.

To be Continued

Sampling Distribution of Sample Variance

Another Heuristic Proof of Theorem 6.8:

- Since $\mathbf{i}'\mathbf{X}$ and $\mathbf{MX}$ follow a joint normal distribution, and they are uncorrelated, it follows that $\mathbf{i}'\mathbf{X}$ and $\mathbf{MX}$ are mutually independent.

Sampling Distribution of Sample Variance

Remarks:

- Theorem 6.7 states that when $\{X_i\}_{i=1}^n$ is IID $N(\mu, \sigma^2)$, $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$, where $n-1$ is called the degrees of freedom. This is a concept associated with sums of squares.
- The random sample $\mathbf{X}^n = (X_1, \cdots, X_n)$ are n **linearly independent** observations, we now use them to estimate σ^2. If we knew μ, an estimator for σ^2 would be $n^{-1}\sum_{i=1}^n (X_i - \mu)^2$.
- Unfortunately we usually do not know the population mean μ. Therefore, we have to replace it with the sample mean $\bar{X}_n$ and use the estimator $S_n^2 = (n-1)^{-1}\sum_{t=1}^n (X_i - \bar{X}_n)^2$.

Sampling Distribution of Sample Variance

- Here, we have actually used the n actual observations $(X_1 - \bar{X}_n, \cdots, X_n - \bar{X}_n)$. These n observations are subject to one restriction

  $$\sum_{i=1}^{n}(X_i - \bar{X}_n) = 0.$$

  Thus, given the $n-1$ observations, we can always obtain the remaining one from the above restriction. In this sense, in estimating S_n^2, we lose one degree of freedom in the original sample due to the restriction. The sum of squares $\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ has only $n-1$ degrees of freedom.

Sampling Distribution of Sample Variance

- More generally, the number of degrees of freedom associated with a sum of squares is given by the number of observations used to compute the sum of squares minus the number of unknown parameters that have to be replaced by their sample estimates. The number of parameters replaced is equal to the number of restrictions placed on data used to form the sum of squares.

Sampling Distribution of Sample Variance

Theorem 10 (6.10)

Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an IID $N(\mu, \sigma^2)$ random sample. Then for all $n > 1$,

$$\text{var}(S_n^2) = \frac{2\sigma^4}{n-1}.$$

Proof

Sampling Distribution of Sample Variance

Proof:

- Because

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

and the variance of χ^2_{n-1} is $2(n-1)$, we have

$$\text{var}\left[\frac{(n-1)S_n^2}{\sigma^2}\right] = 2(n-1).$$

or

$$\left[\frac{(n-1)^2}{\sigma^4}\right] \cdot \text{var}(S_n^2) = 2(n-1).$$

Therefore, $\text{var}(S_n^2) = 2\sigma^4/(n-1)$.

Sampling Distribution of Sample Variance

Remark:

- $\text{var}(S_n^2) = 2\sigma^4/(n-1)$ and $E(S_n^2) = \sigma^2$ imply

$$\begin{aligned} \text{MSE}(S_n^2) &= E(S_n^2 - \sigma^2)^2 \\ &= \text{var}(S_n^2) \\ &= \frac{2\sigma^4}{n-1} \to 0 \text{ as } n \to \infty. \end{aligned}$$

CONTENTS

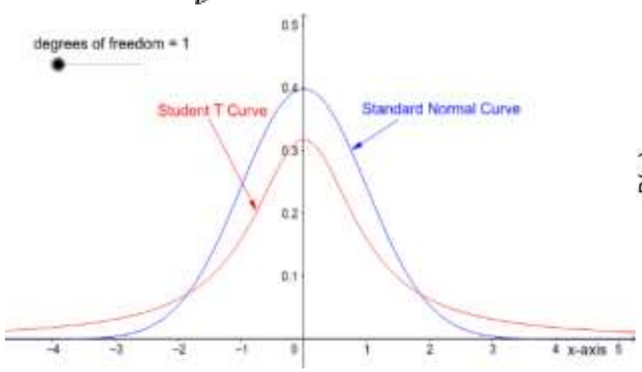
Student's *t*-Distribution

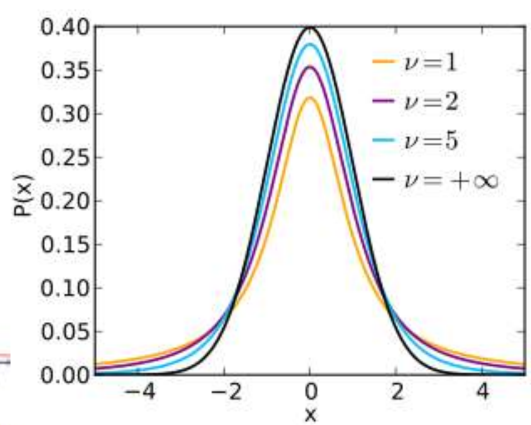
Definition 6 (6.6).[Student's *t*-Distribution]

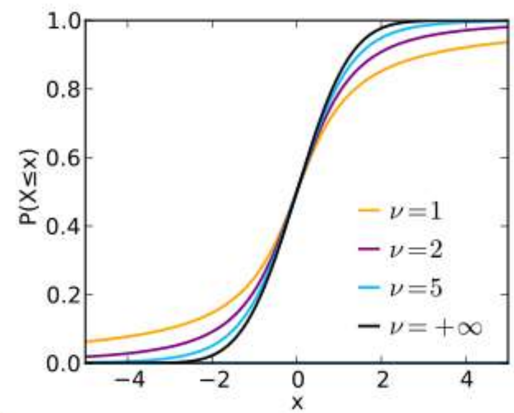
Let $U \sim N(0,1), V \sim \chi^2_\nu$, and U and V are independent. Then the random variable

$$
\begin{aligned}
T &= \frac{U}{\sqrt{V/\nu}} \\
&\sim \frac{N(0,1)}{\sqrt{\chi^2_\nu/\nu}}
\end{aligned}
$$

follows a Student's t distribution with ν degrees of freedom, denoted as $T \sim t_\nu$.







Student's *t*-Distribution

Remarks:

- The PDF of a Student's t_ν distribution is

$$f_T(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{(\nu\pi)^{1/2}} \frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}, \qquad -\infty < t < \infty.$$

- This could be obtained by first finding the PDF $f_{TR}(t,r)$ of the **bivariate transformation**

$$\begin{aligned} T &= U/\sqrt{V/\nu}, \\ R &= U, \end{aligned}$$

and then integrating out R.

Student’s *t*-Distribution

Lemma 3 (6.11).[Properties of the Student t_ν Distribution]

(1) The PDF of t_ν is symmetric about 0.

(2) t_ν has a heavier distributional tail than $N(0,1)$ (see Figure 6.5 below).

(3) Only the first $\nu - 1$ moments exist. In particular, the mean $\mu = 0$, and the variance $\sigma^2 = \nu/(\nu - 2)$ when $\nu > 2$. The MGF does not exist for any given ν.

(4) When $\nu = 1$, $t_1 \sim$ Cauchy (0,1).

(5) $t_\nu \to N(0,1)$ as $\nu \to \infty$.

Student's *t*-Distribution

Remarks:

- The convergence of t_ν to $N(0,1)$ can be seen from the limit

$$\begin{aligned}\lim_{\nu\to\infty} f_T(t) &= \lim_{\nu\to\infty}\sqrt{\frac{2}{\nu}}\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\lim_{\nu\to\infty}\frac{1}{(1+t^2/\nu)^{1/2}}\frac{1}{\sqrt{2\pi}}\lim_{\nu\to\infty}\frac{1}{(1+t^2/\nu)^{\nu/2}}\\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}\end{aligned}$$

by using the fact that $(1+a/\nu)^\nu \to e^a$ as $\nu\to\infty$. Here, as $\nu\to\infty$,

$$\sqrt{\frac{2}{\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\to 1.$$

Student's *t*-Distribution

- The Student t-distribution has classical importance in statistical inference:
  - When $\mathbf{X}^n$ is an IID $N(\mu, \sigma^2)$ random sample, we have for all $n \geq 1$,
$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$
  - However, since σ is unknown, we have to replace σ with an estimator, usually the sample standard deviation S_n.
  - Thus, the theory that follows leads to the exact distribution of
$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}.$$

Student's *t*-Distribution

Theorem 12 (6.12)

Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be an IID random sample from a $N(\mu, \sigma^2)$ distribution. Then for all $n > 1$, the standardized sample mean

$$\begin{aligned}\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} &= \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}/(n-1)}} \\ &\sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \\ &\sim t_{n-1},\end{aligned}$$

where t_{n-1} is the Student t-distribution with $n-1$ degrees of freedom.

Proof

Student's *t*-Distribution

Proof:

- Put $U = (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$, and $V = (n-1)S_n^2/\sigma^2$. Then $U \sim N(0,1)$ and $V \sim \chi^2_{n-1}$. Also, $\bar{X}_n$ and S_n^2 are independent.
- It follows that

$$\begin{aligned} \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} &= \frac{(\bar{X}_n - \mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)S_n^2/[\sigma^2(n-1)]}} \\ &\sim t_{n-1}. \end{aligned}$$

Student's *t*-Distribution

Example 7 (6.7).[Confidence Interval Estimation for Population Mean μ]

Suppose $\mathbf{X}^n = (X_1, \cdots, X_n)$ is an IID random sample from a $N(\mu, \sigma^2)$ population, where both μ and σ^2 are unknown. We are interested in constructing a confidence interval estimator for μ at the $(1-\alpha)100\%$ confidence level.

Solution

Student's *t*-Distribution

Solution

- Given $\alpha \in (0,1)$, a $(1-\alpha)100\%$-confidence interval estimator for μ is defined an random interval $[\hat{L}, \hat{U}]$ such that
$$P\left(\hat{L} < \mu < \hat{U}\right) = 1 - \alpha.$$
- To construct an interval estimator for μ when σ^2 is unknown, we define the upper-tailed critical value $C_{t_{n-1},\alpha}$ of a Student's t_{n-1} distribution by
$$P\left(t_{n-1} > C_{t_{n-1},\alpha}\right) = \alpha.$$

To be Continued

Student's *t*-Distribution

Solution

- By Theorem 6.12 and the symmetry of the Student-t distribution, we have

$$P\left[\left|\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n}\right| > C_{n-1,\frac{\alpha}{2}}\right] = \alpha$$

or equivalently,

$$P\left[\left|\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n}\right| \leq C_{n-1,\frac{\alpha}{2}}\right] = 1-\alpha.$$

- This yields a $(1-\alpha)100\%$ confidence interval estimator for μ when σ^2 is unknown:

$$P\left(\bar{X}_n - \frac{S_n}{\sqrt{n}}C_{t_{n-1},\frac{\alpha}{2}} < \mu < \bar{X}_n + \frac{S_n}{\sqrt{n}}C_{t_{n-1},\frac{\alpha}{2}}\right) = 1-\alpha.$$

To be Continued

Student's *t*-Distribution

Solution

- The random interval estimator

$$\left[\bar{X}_n - \frac{S_n}{\sqrt{n}} C_{t_{n-1},\frac{\alpha}{2}},\ \bar{X}_n + \frac{S_n}{\sqrt{n}} C_{t_{n-1},\frac{\alpha}{2}}\right]$$

  is computable when σ^2 is unknown.

- Note that the sampling distribution of

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

  plays a crucial role of determining the critical value $C_{t_{n-1},\alpha/2}$, and thus the confidence interval estimator.

Student's *t*-Distribution

Example 8 (6.8).[Hypothesis Testing on Population Mean: The *t*-test]

Suppose there is an IID $N(\mu, \sigma^2)$ random sample $\mathbf{X}^n = (X_1, \cdots, X_n)$ of size n, and we are interested in testing the hypothesis

$$\mathbb{H}_0 : \mu = \mu_0,$$

where μ_0 is a given (known) constant (e.g., $\mu_0 = 0$). How can we test this hypothesis?

Solution

Student's *t*-Distribution

Solution

- To test the hypothesis $\mathbb{H}_0 : \mu = \mu_0$, we consider the statistic

$$\bar{X}_n - \mu_0 = (\bar{X}_n - \mu) + (\mu - \mu_0).$$

  – When $\mathbb{H}_0 : \mu = \mu_0$, we have

$$\bar{X}_n - \mu_0 = \bar{X}_n - \mu \to 0 \text{ as } n \to \infty$$

  in terms of mean squared error. Therefore, $\bar{X}_n - \mu_0$ will be close to zero as $n \to \infty$.

To be Continued

Student's *t*-Distribution

Solution

– On the other hand, if $\mathbb{H}_0$ is false, i.e. $\mu \neq \mu_0$, then

$$\begin{aligned} \bar{X}_n - \mu_0 &= (\bar{X}_n - \mu) + (\mu - \mu_0) \\ &\to \mu - \mu_0 \neq 0 \text{ as } n \to \infty \end{aligned}$$

in terms of mean squared error. Therefore, a test for $\mathbb{H}_0$ can be based on the statistic $\bar{X}_n - \mu_0$:
(1) If $\bar{X}_n - \mu_0$ is sufficiently small, then $\mathbb{H}_0$ is true;
(2) Otherwise if $\bar{X}_n - \mu_0$ is sufficiently large in absolute value, then $\mathbb{H}_0$ is false.

Student's *t*-Distribution

Question:

How far away $\bar{X}_n - \mu_0$ is from zero will be considered as "sufficiently large" in absolute value?

Solution

Student's *t*-Distribution

Solution

- This is described by the sampling distribution of $\bar{X}_n - \mu_0$. From the sampling distribution of $\bar{X}_n - \mu_0$, we can find a threshold value called **critical value** to judge whether $\bar{X}_n - \mu_0$ is significantly large.
  - Suppose $\mathbf{X}^n \sim$ IID $N(\mu, \sigma^2)$. Then for each positive integer n,

$$\bar{X}_n - \mu \sim N\left(0, \frac{\sigma^2}{n}\right).$$

It follows that

To be Continued

Student's *t*-Distribution

Solution

$$\bar{X}_n - \mu_0 = (\bar{X}_n - \mu) + (\mu - \mu_0) \sim N\left(0, \frac{\sigma^2}{n}\right).$$

Therefore, the standardized random variable

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \sim N\left(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma}, 1\right).$$

To be Continued

Student's *t*-Distribution

Solution

When the hypothesis $\mathbb{H}_0$ holds,

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

which implies that $(\bar{X}_n - \mu_0)/(\sigma/\sqrt{n})$ will take small and finite values with very high probability.

To be Continued

Student's *t*-Distribution

Solution

– On the other hand, when $\mathbb{H}_0$ is false,

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \to \infty \text{ as } n \to \infty$$

with high probability. Therefore, we can test $\mathbb{H}_0$ by examining whether $(\bar{X}_n - \mu_0)/(\sigma/\sqrt{n})$ is large in absolute value.

To be Continued

Student's *t*-Distribution

Solution

- However, the quantity

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

is not a feasible statistic, because it involves the unknown parameter σ. We have to replace σ with an estimator for σ, say the sample standard deviation S_n. This leads use to consider the following feasible t-test statistic $T(\mathbf{X}^n) = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$.

To be Continued

Student's *t*-Distribution

Solution

- However, the distribution of $T(\mathbf{X}^n)$ is no longer $N(0,1)$; instead it becomes a Student t-distribution with $n-1$ degrees of freedom:
  - Under $\mathbb{H}_0 : \mu = \mu_0$,

$$T(\mathbf{X}^n) \sim t_{n-1}$$

for all $n > 1$. This follows because under $\mathbb{H}_0 : \mu = \mu_0$

To be Continued

Student's *t*-Distribution

Solution

$$
\begin{aligned}
T(\mathbf{X}^n) &= \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \\
&= \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \\
&\sim t_{n-1}.
\end{aligned}
$$

Thus, with very high probability, the t-test statistic $T(\mathbf{X}^n)$ will take small and finite values.

– On the other hand, when $\mathbb{H}_0 : \mu = \mu_0$ is false, i.e, when $\mu \neq \mu_0$, we have

$$
\begin{aligned}
T(\mathbf{X}^n) &= \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} + \frac{\sqrt{n}(\mu - \mu_0)}{S_n} \\
&\to \infty
\end{aligned}
$$

as $n \to \infty$ with high probability.

Student's *t*-Distribution

Decision Rule for the *T*-test Using Critical Values:

- Reject the hypothesis $\mathbb{H}_0 : \mu = \mu_0$ at the prespecified significance level $\alpha \in (0,1)$ if

$$|T(\mathbf{X}^n)| > C_{t_{n-1}, \frac{\alpha}{2}},$$

where $C_{t_{n-1}, \frac{\alpha}{2}}$ is the upper-tailed critical value of the Student t_{n-1} distribution at level $\frac{\alpha}{2}$, determined by $P(t_{n-1} > C_{t_{n-1}, \frac{\alpha}{2}}) = \frac{\alpha}{2}$.

- Accept the hypothesis $\mathbb{H}_0$ at the significance level α if $|T(\mathbf{X}^n)| \leq C_{t_{n-1}, \frac{\alpha}{2}}$.

Student's *t*-Distribution

Remarks:

- In testing $\mathbb{H}_0$ using an observed data generated from the random sample $\mathbf{X}^n$ of size n, there exist two type of errors:

- One possibility is that $\mathbb{H}_0$ is true but we reject it. This is possible because the test statistic $T(\mathbf{X}^n)$ follows a Student t_{n-1} distribution under $\mathbb{H}_0$, which has an unbounded support. Thus, there exists a small probability that $T(\mathbf{X}^n)$ can still take a larger value than the critical value under $\mathbb{H}_0$.This is the so-called **Type I error**. The significance level α controls Type I error. If

$$P\left[|T(\mathbf{X}^n)| > C_{t_{n-1},\frac{\alpha}{2}}|\mathbb{H}_0\right] = \alpha,$$

we call the decision rule a size α test or a test with size α.

Student's *t*-Distribution

Remarks:

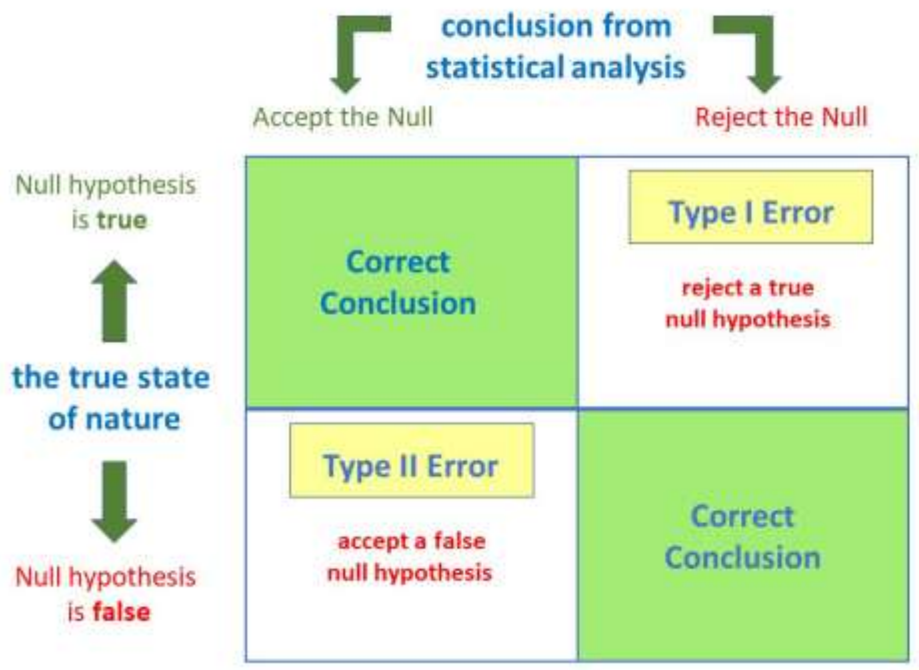
- On the other hand, the probability

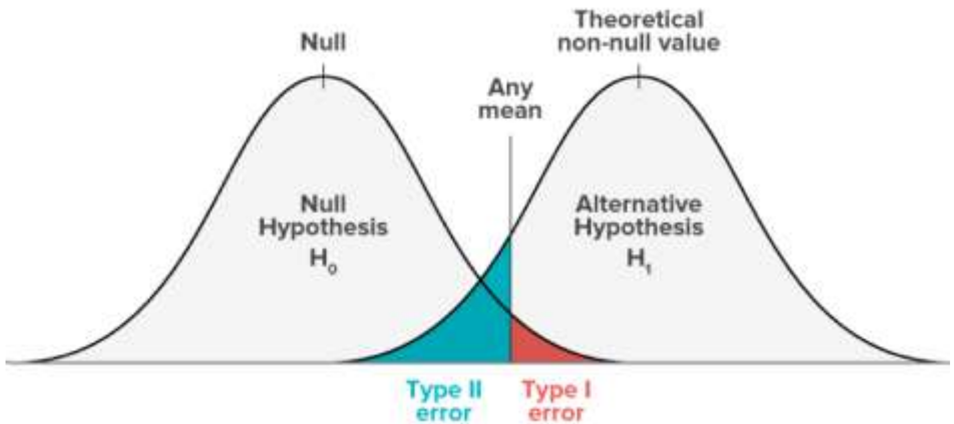
$$P\left[|T(\mathbf{X}^n)| > C_{t_{n-1},\frac{\alpha}{2}} | \mathbb{H}_0 \text{ is false}\right]$$

  is called the **power** function of the size-α t-test. When $P[|T(\mathbf{X}^n)| > C_{t_{n-1},\frac{\alpha}{2}} | \mathbb{H}_0 \text{ is false}] < 1$, there exists a possibility that one may accept $\mathbb{H}_0$ when it is false. This is called a **Type II error**.

- When n is finite, due to the nature of limited information offered by the random sample $\mathbf{X}^n$, both the Type I and Type II errors are unavoidable and there usually exists a tradeoff between them. In practice, one usually sets a level for the Type I error and then minimizes the Type II error.

Type I Error & Type II Error





Student's *t*-Distribution

Decision Rule for the *T*-test Using *P*-Values:

- Given any observed data set $\mathbf{x}^n$, we can compute a value (i.e., a realization)
$$T(\mathbf{x}^n) = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$$
for the t-test statistic $T(\mathbf{X}^n)$.

- Then the probability
$$\begin{aligned} p(\mathbf{x}^n) &= P\left(|T(\mathbf{X}^n)| > |T(\mathbf{x}^n)| \,|\, \mathbb{H}_0\right) \\ &= P\left(|t_{n-1}| > |T(\mathbf{x}^n)|\right) \end{aligned}$$
is called the P-value of the t-test statistic $T(\mathbf{X}^n)$ when a data set $\mathbf{x}^n$ is observed.

Student's *t*-Distribution

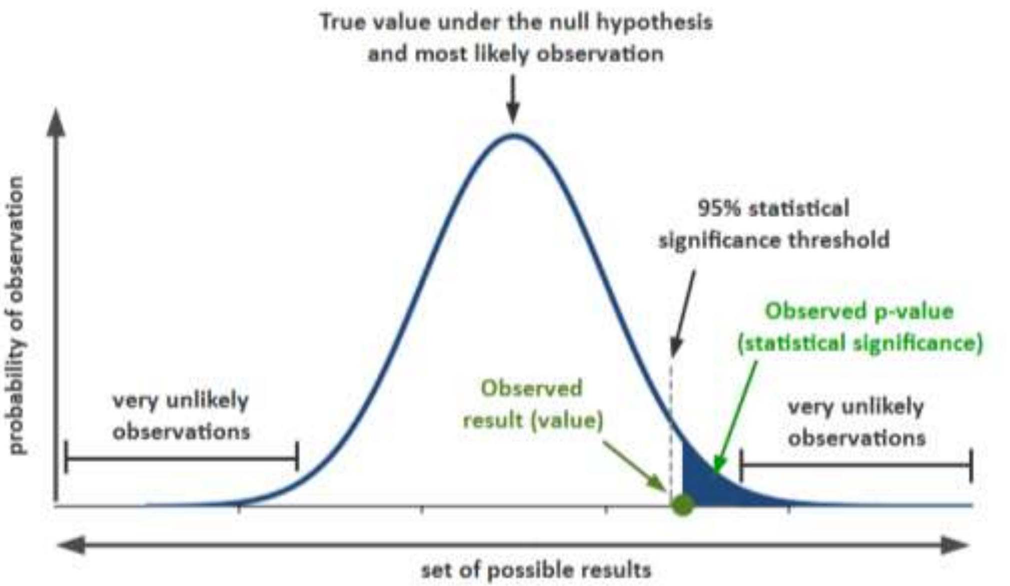
Decision Rule for the *T*-test Using *P*-Values:

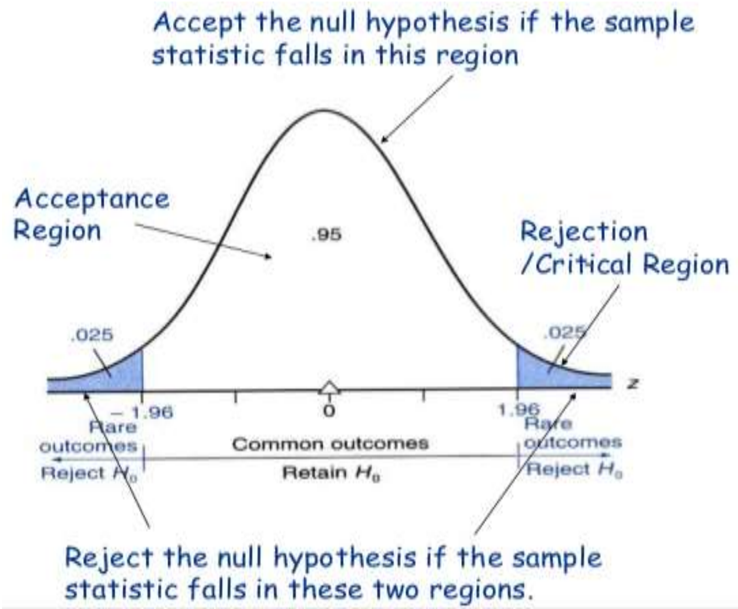
- **Interpretation** of P-value: it can be viewed as the probability that the t-test statistic $T(\mathbf{X}^n)$ is larger than the **observed value** $T(\mathbf{x}^n)$ when $\mathbb{H}_0$ holds.
- If the observed value $T(\mathbf{x}^n)$ is large, $p(\mathbf{x}^n)$ will be small. Thus, a small P-value is strong evidence against the null hypothesis $\mathbb{H}_0$, while a large P-value shows that the data are consistent with $\mathbb{H}_0$.

Student's *t*-Distribution

P-Values Based Decision Rule:

- Reject the hypothesis $\mathbb{H}_0$ at the significance level α if $p(\mathbf{x}^n) < \alpha$.
- Accept the hypothesis $\mathbb{H}_0$ at the significance level α if $p(\mathbf{x}^n) \geq \alpha$.





Student's *t*-Distribution

Remarks:

- The P-value is the smallest value of the significance level α for which $\mathbb{H}_0$ can be rejected. The P-value not only tells us whether $\mathbb{H}_0$ should be accepted or rejected at a given significance level, but also tells us whether the decision to accept or reject $\mathbb{H}_0$ is a close call.

- **Statistical verus Economic Significance:** A rejection of $\mathbb{H}_0$ based on either of the above decision rules is called a *statistically significant* effect. From a statistical perspective, for any deviation from $\mathbb{H}_0$ (i.e., any difference between $\mu - \mu_0$), no matter how small it is, a rejection decision will be made as long as the sample size n is sufficiently large.

Student's *t*-Distribution

Remarks:

- However, a small difference $\mu - \mu_0$ may not be important from an economic perspective. For example, one may be interested in whether the expected return (μ) on a mutual fund is practically significantly different from a pre-specified rate (μ_0) of return. The size of the difference $\mu - \mu_0$ should be large enough to consider an investment on the mutual fund, due to (e.g.,) existence of transaction costs. However, a statistic test like the t-test introduced above will reject any nonzero small difference $\mu - \mu_0$ as long as the sample size n is sufficiently large. In other words, an economically insignificant effect is likely to be statistically significant.

Student's *t*-Distribution

Question:
Is the P-value inference a scientific approach?

- Data snooping?

CONTENTS

Snedecor's *F* Distribution

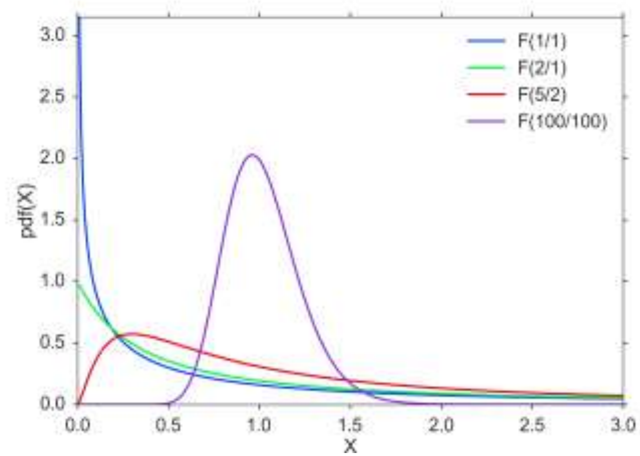
Definition 7 (6.7). [The *F* Distribution]

Let U and V be two independent Chi-square random variables with p and q degrees of freedom respectively. Then the random variable

$$F = \frac{U/p}{V/q} \sim F_{p,q}$$

follows a F distribution with p and q degrees of freedom.

What is the PDF of a $F_{p,q}$ distribution?



Snedecor's *F* Distribution

- The PDF is given by

$$f_F(x) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})}\left(\frac{p}{q}\right)^{p/2}\frac{x^{(p/2)-1}}{[1+(p/q)x]^{(p+q)/2}}, \qquad 0 < x < \infty.$$

- This PDF could be obtained by using the bivariate transformation

$$\begin{aligned} F &= (U/p)/(V/q), \\ G &= U, \end{aligned}$$

and then integrating out G.

Snedecor's *F* Distribution

Lemma 4 (6.13). [Properties of $F_{p,q}$ Distribution]

(1) If $X \sim F_{p,q}$, then $X^{-1} \sim F_{q,p}$;
(2) If $X \sim t_q$, then $X^2 \sim F_{1,q}$;
(3) If $q \to \infty$, then $p \cdot F_{p,q} \to \chi^2_p$.

Snedecor's *F* Distribution

Proof:

- Result (1) follows from the definition of a F random variable.
- For Result (2), recall that a t_q random variable is defined as

$$t_q \sim \frac{Z}{\sqrt{\chi_q^2/q}},$$

where $Z \sim N(0,1)$ and it is independent of χ_q^2. It follows that

$$t_q^2 \sim \frac{\chi_1^2/1}{\chi_q^2/q} \sim F_{1,q}.$$

Snedecor's *F* Distribution

Example 9 (6.9) [Hypothesis Testing on Equality of Population Variances]

Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be a random sample of size n from a $N(\mu_X, \sigma_X^2)$ population, and $\mathbf{Y}^m = (Y_1, \cdots, Y_m)$ be a random sample of size m from a $N(\mu_Y, \sigma_Y^2)$ population. Assume that $\mathbf{X}^n$ and $\mathbf{Y}^m$ are independent.

Suppose we are interested in comparing variability of the population, i.e. interested in testing whether $\mathbb{H}_0 : \sigma_X^2 = \sigma_Y^2$ holds. Then a test statistic can be based on the sample variance ratio

$$\frac{S_X^2}{S_Y^2}.$$

Snedecor's *F* Distribution

Since $S_X^2 \to \sigma_X^2$ as $n \to \infty$ in MSE, and $S_Y^2 \to \sigma_Y^2$ as $m \to \infty$ in MSE, we have

$$\frac{S_X^2}{S_Y^2} \to \frac{\sigma_X^2}{\sigma_Y^2} \text{ as } n, m \to \infty.$$

Under $\mathbb{H}_0 : \sigma_X^2 = \sigma_Y^2$, we have

$$\begin{aligned} \frac{S_X^2}{S_Y^2} &= \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \\ &\sim F_{n-1,m-1}. \end{aligned}$$

Snedecor's *F* Distribution

If $\mathbb{H}_0$ is false, and so $\sigma_X^2 \neq \sigma_Y^2$, then

$$\frac{S_X^2}{S_Y^2} \neq \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}$$

Therefore, by checking whether S_X^2/S_Y^2 follows the $F_{n-1,m-1}$ distribution, we can test whether the variances are equal. In particular, if one knows that $\sigma_X^2 > \sigma_Y^2$ under $\mathbb{H}_0$, then one can use the right-tailed critical value of $F_{n-1,m-1}$.

Snedecor's *F* Distribution

Remarks:

- Because the F-distribution is closely related to the ratio of sample variances, it is sometime called the variance ratio distribution.
- The F-test is an important testing principle in classical statistics and econometrics, where S_X^2 and S_Y^2 are generalized to the sums of squared residuals of a restricted regression model and an unrestricted regression model respectively.

CONTENTS

Sufficient Statistics

The KISS Principle:

Keep It Sophisticatedly Simple

Question

Suppose we are interested in making inference of parameter θ using a set of data generated from a random sample $\mathbf{X}^n$ from a population $f_X(x) = f(x, \theta)$. Under what conditions, can the information about θ that is contained in the random sample $\mathbf{X}^n$ be completely summarized by some low-dimensional function of $\mathbf{X}^n$, say, some statistic $T(\mathbf{X}^n)$?

Sufficient Statistics

- Suppose Person A observes a realiation $\mathbf{x}^n$ while Person B only observes the value of $t = T(\mathbf{x}^n)$. Generally, Person A knows better than Person B about the unknown parameter value of θ.

- However, there may exist situations in which Person B can do just as well as Person A. This occurs when the statistic $T(\mathbf{X}^n)$ summarizes all information about θ that is contained in $\mathbf{X}^n$, so that individual values of $\mathbf{x}^n$ are irrelevant in search for a good estimator of θ.

- A statistic $T(\mathbf{X}^n)$ that has this desired property is called a **sufficient statistic** for parameter θ. An important implication of a sufficient statistic for parameter θ is that one can then just keep the sufficient statistic $T(\mathbf{X}^n)$, which is low dimensional.

Sufficient Statistics

- For example, suppose the random sample $\mathbf{X}^n \sim$ IID $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2)$. Then for inference of θ, only the sample mean $\bar{X}_n$ and the sample variance S_n^2 should be retained, because they are sufficient statistics for (μ, σ^2).
- Sufficient statistic is an important method for **data reduction**.

Sufficient Statistics

Question

How can one check $(\bar{X}_n, S_n^2)$ are sufficient for $\theta = (\mu, \sigma^2)$ for a random sample $\mathbf{X}^n$ from a normal population? More generally, how can one find a sufficient statistic for parameter θ associated with a given population?

Sufficient Statistics

Definition 8 (6.8). [Sufficient Statistic]

Let $\mathbf{X}^n$ be a random sample from some population with parameter θ. A statistic $T(\mathbf{X}^n)$ is a sufficient statistic for θ if the conditional distribution of the sample $\mathbf{X}^n = \mathbf{x}^n$ given that the value of the statistic $T(\mathbf{X}^n) = T(\mathbf{x}^n)$ does not depend on θ; that is,

$$f_{\mathbf{X}^n|T(\mathbf{X}^n)}[\mathbf{x}^n|T(\mathbf{x}^n), \theta] = h(\mathbf{x}^n) \text{ for all possible } \theta,$$

where the left hand side is the conditional PMF/PDF of $\mathbf{X}^n = \mathbf{x}^n$ given $T(\mathbf{X}^n) = T(\mathbf{x}^n)$, which generally depends on θ. The right hand side $h(\mathbf{x}^n)$ does not depend on θ; it is a function of $\mathbf{x}^n$ only.

Sufficient Statistics

Remarks:

- Suppose $f_{\mathbf{X}^n|T(\mathbf{X}^n)}[\mathbf{x}^n|T(\mathbf{x}^n),\theta]$, the conditional probability of $\mathbf{X}^n = \mathbf{x}^n$ given $T(\mathbf{X}^n) = T(\mathbf{x}^n)$, does not depend on θ.

- Then all sample points $\{\mathbf{x}^n\}$ which yield the same value of $T(\mathbf{x}^n) = t$ for $T(\mathbf{X}^n)$, will be just equally likely for any value of θ. In other words, since the conditional distribution of $\mathbf{X}^n = \mathbf{x}^n$ given $T(\mathbf{X}^n) = T(\mathbf{x}^n)$ does not depend on θ, the data $\mathbf{x}^n$ beyond the value of $T(\mathbf{x}^n) = t$ does not provide any additional useful information about θ. All knowledge about θ that can be gained from the observed value $\mathbf{x}^n$ of the sample $\mathbf{X}^n$ can just as well be gained from the value of $T(\mathbf{x}^n)$ alone.

Sufficient Statistics

Sufficient Statistic in the Discrete Case:

- First of all, sufficiency implies that the conditional PMF

$$\begin{aligned} & f_{\mathbf{X}^n|T(\mathbf{X}^n)}[\mathbf{x}^n|T(\mathbf{x}^n),\theta] \\ \equiv\ & P_\theta[\mathbf{X}^n=\mathbf{x}^n|T(\mathbf{X}^n)=T(\mathbf{x}^n)] \\ =\ & h(\mathbf{x}^n) \end{aligned}$$

for all θ, where $P_\theta(\cdot)$ is the probability measure under the probability distribution of $\mathbf{X}^n$ which is usually indexed by θ.

Sufficient Statistics

- The full information of a random sample $\mathbf{X}^n$ is described by the joint probability of $\mathbf{X}^n = \mathbf{x}^n$, denoted by $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = P(\mathbf{X}^n = \mathbf{x}^n)$. This joint probability depends on θ in general. For example, when $\mathbf{X}^n$ is an IID random sample with population PMF $f(x, \theta)$. Then

$$f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Sufficient Statistics

- Because $\mathbf{X}^n = \mathbf{x}^n$ implies $T(\mathbf{X}^n) = T(\mathbf{x}^n)$ but not vice versa, we have the event $A = \{\mathbf{X}^n = \mathbf{x}^n\} \subseteq B = \{T(\mathbf{X}^n) = T(\mathbf{x}^n)\}$. Therefore, the joint PMF of the random sample $\mathbf{X}^n$

$$\begin{aligned} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) &= P(\mathbf{X}^n = \mathbf{x}^n) \\ &= P(A) \\ &= P(A \cap B) \\ &= P(A|B)P(B) \\ &= P[\mathbf{X}^n = \mathbf{x}^n | T(\mathbf{X}^n) = T(\mathbf{x}^n)]P[T(\mathbf{X}^n) = T(\mathbf{x}^n)] \\ &= h(\mathbf{x}^n) f_{T(\mathbf{X}^n)}[T(\mathbf{x}^n), \theta] \end{aligned}$$

by sufficiency, where $f_{T(\mathbf{X}^n)}[T(\mathbf{x}^n), \theta] \equiv P[T(\mathbf{X}^n) = T(\mathbf{x}^n)]$ depends on θ but $h(\mathbf{x}^n)$ does not depend on θ.

Sufficient Statistics

- Only the marginal probability $P[T(\mathbf{X}^n) = T(\mathbf{x}^n)]$ of the sufficient statistic $T(\mathbf{X}^n)$ is related to θ. Therefore, if we are interested in making inference of θ, then we can only retain the information of $T(\mathbf{X}^n)$.

Sufficient Statistics

- For example, the so-called maximum likelihood estimation (MLE) for θ, to be introduced in Chapter 8, is to maximize the objective function–the log-likelihood function

$$\ln f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = \ln h(\mathbf{x}^n) + \ln f_{T(\mathbf{X}^n)}[T(\mathbf{x}^n), \theta].$$

Because the first part is irrelevant to θ, we have

$$\arg\max_{\theta \in \Theta} \ln f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = \arg\max_{\theta \in \Theta} \ln f_{T(\mathbf{X}^n)}[T(\mathbf{x}^n), \theta],$$

where Θ is a parameter space. In other words, it suffices to maximize the log-likelihood function $\ln f_{T(\mathbf{X}^n)}[T(\mathbf{x}^n), \theta]$ of the sufficient statistic $T(\mathbf{X}^n)$ for MLE of θ.

Sufficient Statistics

Question

How can one check if a statistic $T(\mathbf{X}^n)$ is sufficient for parameter θ?

Sufficient Statistics

Theorem 10 (6.14). [Factorization Theorem]

Let $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$ denote the joint PDF (or PMF) of a random sample $\mathbf{X}^n$. A statistic $T(\mathbf{X}^n)$ is a sufficient statistic for θ if and only if there exist functions $g(t, \theta)$ and $h(\mathbf{x}^n)$ such that for any sample point $\{\mathbf{x}^n\}$ in the sample space of $\mathbf{X}^n$ and for any parameter value $\theta \in \Theta$,

$$f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = g[T(\mathbf{x}^n), \theta]h(\mathbf{x}^n),$$

where $g(t, \theta)$ depends on parameter θ but $h(\mathbf{x}^n)$ does not depend on parameter θ.

Sufficient Statistics

Proof:

We shall show only the discrete case, where $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = P(\mathbf{X}^n = \mathbf{x}^n)$.

- (1) [*Necessity*]: When $T(\mathbf{X}^n)$ is sufficient, noting that $\{\mathbf{X}^n = \mathbf{x}^n\} \subseteq \{T(\mathbf{X}^n) = T(\mathbf{x}^n)\}$, we have

$$\{\mathbf{X}^n = \mathbf{x}^n\} = \{\mathbf{X}^n = \mathbf{x}^n\} \cap \{T(\mathbf{X}^n) = T(\mathbf{x}^n)\}.$$

It follows that

To be Continued

Sufficient Statistics

It follows that

$$
\begin{aligned}
f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) &= P(\mathbf{X}^n = \mathbf{x}^n) \\
&= P[\mathbf{X}^n = \mathbf{x}^n, T(\mathbf{X}^n) = T(\mathbf{x}^n)] \\
&= P[\mathbf{X}^n = \mathbf{x}^n | T(\mathbf{X}^n) = T(\mathbf{x}^n)] P[T(\mathbf{X}^n) = T(\mathbf{x}^n)] \\
&= h(\mathbf{x}^n) P[T(\mathbf{X}^n) = T(\mathbf{x}^n)] \\
&= h(\mathbf{x}^n) g[T(\mathbf{x}^n), \theta],
\end{aligned}
$$

where $g[T(\mathbf{x}^n), \theta] = P[T(\mathbf{X}^n) = T(\mathbf{x}^n)]$ and $h(\mathbf{x}^n) = P[\mathbf{X}^n = \mathbf{x}^n | T(\mathbf{X}^n) = T(\mathbf{x}^n)]$.

To be Continued

Sufficient Statistics

- (2) [*Sufficiency*]: Now suppose we have

$$f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = g[T(\mathbf{x}^n), \theta]h(\mathbf{x}^n).$$

We shall show that the conditional probability $P[\mathbf{X}^n = \mathbf{x}^n | T(\mathbf{X}^n) = T(\mathbf{x}^n)]$ does not depend on θ.

Because

$$\{\mathbf{X}^n = \mathbf{x}^n\} = \{\mathbf{X}^n = \mathbf{x}^n\} \cap \{T(\mathbf{X}^n) = T(\mathbf{x}^n)\},$$

To be Continued

Sufficient Statistics

we have

$$
\begin{aligned}
P[\mathbf{X}^n &= \mathbf{x}^n | T(\mathbf{X}^n) = T(\mathbf{x}^n)] \\
&= \frac{P[\mathbf{X}^n = \mathbf{x}^n, T(\mathbf{X}^n) = T(\mathbf{x}^n)]}{P[T(\mathbf{X}^n) = T(\mathbf{x}^n)]} \\
&= \frac{P(\mathbf{X}^n = \mathbf{x}^n)}{P[T(\mathbf{X}^n) = T(\mathbf{x}^n)]} \\
&= \frac{g[T(\mathbf{x}^n), \theta]h(\mathbf{x}^n)}{P[T(\mathbf{X}^n) = T(\mathbf{x}^n)]}.
\end{aligned}
$$

To be Continued

Sufficient Statistics

We now consider the denominator:

$$
\begin{aligned}
P[T(\mathbf{X}^n) &= T(\mathbf{x}^n)] \\
&= \sum_{\{\mathbf{y}^n : T(\mathbf{y}^n) = T(\mathbf{x}^n)\}} f_{\mathbf{X}^n}(\mathbf{y}^n, \theta) \\
&= \sum_{\{\mathbf{y}^n : T(\mathbf{y}^n) = T(\mathbf{x}^n)\}} g[T(\mathbf{y}^n), \theta] h(\mathbf{y}^n) \\
&= \sum_{\{\mathbf{y}^n : T(\mathbf{y}^n) = T(\mathbf{x}^n)\}} g[T(\mathbf{x}^n), \theta] h(\mathbf{y}^n) \\
&= g[T(\mathbf{x}^n), \theta] \sum_{\{\mathbf{y}^n : T(\mathbf{y}^n) = T(\mathbf{x}^n)\}} h(\mathbf{y}^n)
\end{aligned}
$$

To be Continued

Sufficient Statistics

where the sum is taken over all possible sample points $\{\mathbf{y}^n\}$ in the sample space of $\mathbf{X}^n$ that yield the same value of $T(\mathbf{y}^n) = T(\mathbf{x}^n)$. It follows that the conditional probability

$$
\begin{aligned}
& P\left[\mathbf{X}^n = \mathbf{x}^n | T(\mathbf{X}^n) = T(\mathbf{x}^n)\right] \\
= \; & \frac{g[T(\mathbf{x}^n), \theta] h(\mathbf{x}^n)}{P[T(\mathbf{X}^n) = T(\mathbf{x}^n)]} \\
= \; & \frac{g[T(\mathbf{x}^n), \theta] h(\mathbf{x}^n)}{g[T(\mathbf{x}^n), \theta] \sum_{\{\mathbf{y}^n : T(\mathbf{y}^n) = T(\mathbf{x}^n)\}} h(\mathbf{y}^n)} \\
= \; & \frac{h(\mathbf{x}^n)}{\sum_{\{\mathbf{y}^n : T(\mathbf{y}^n) = T(\mathbf{x}^n)\}} h(\mathbf{y}^n)},
\end{aligned}
$$

which does not depend on θ.

Sufficient Statistics

Example 10 (6.10)

Suppose $\mathbf{X}^n \sim$ IID Bernoulli(θ), where $0 < \theta < 1$. Show that the sample proportion $T(\mathbf{X}^n) = n^{-1}\sum_{i=1}^{n} X_i$ is a sufficient statistic for θ. Note that $\theta = E(X_i)$.

Solution

- The PMF of a Bernoulli(θ) random variable X_i is

$$f(x_i, \theta) = \theta^{x_i}(1-\theta)^{1-x_i},$$

where x_i takes value 0 or 1.

To be Continued

Sufficient Statistics

- Suppose $\mathbf{x}^n$ is a realization (i.e., a data set) of the random sample $\mathbf{X}^n$. We have

$$\begin{aligned} P(\mathbf{X}^n &= \mathbf{x}^n) = \prod_{i=1}^{n} f(x_i, \theta) \\ &= \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{\Sigma_{i=1}^n x_i} (1-\theta)^{n-\Sigma_{i=1}^n x_i} \\ &= \theta^{nT(\mathbf{x}^n)} (1-\theta)^{n-nT(\mathbf{x}^n)} \\ &= g[T(\mathbf{x}^n), \theta] h(\mathbf{x}^n), \end{aligned}$$

where $T(\mathbf{X}^n) = n^{-1} \sum_{i=1}^n X_i$, $h(\mathbf{x}^n) = 1$, and $g[T(\mathbf{x}^n), \theta] = \theta^{nT(\mathbf{x}^n)} (1-\theta)^{n-nT(\mathbf{x}^n)}$.

Sufficient Statistics

Example 11 (6.11)

Let $\mathbf{X}^n \sim$ IID $N(\mu, \sigma^2)$, where σ^2 is a known value. Then show $T(\mathbf{X}^n) = \bar{X}_n$ is a sufficient statistic for μ.

Solution

- In this example, the (unknown) parameter $\theta = \mu$. Since σ^2 is a given (known) number, it is no longer a parameter.
- The joint PDF of $\mathbf{X}^n$

To be Continued

Sufficient Statistics

$$
\begin{aligned}
& f_{\mathbf{X}^n}(\mathbf{x}^n, \mu) \\
= \quad & \prod_{i=1}^{n} f(x_i, \theta) \\
= \quad & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
= \quad & \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\Sigma_{i=1}^{n}(x_i-\bar{x}_n+\bar{x}_n-\mu)^2}{2\sigma^2}} \\
= \quad & \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\Sigma_{i=1}^{n}(x_i-\bar{x}_n)^2+n(\bar{x}_n-\mu)^2}{2\sigma^2}} \\
= \quad & \left[\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\Sigma_{i=1}^{n}(x_i-\bar{x}_n)^2}{2\sigma^2}}\right] e^{-\frac{n(\bar{x}_n-\mu)^2}{2\sigma^2}} \\
= \quad & h(\mathbf{x}^n) g(\bar{x}_n, \mu),
\end{aligned}
$$

To be Continued

Sufficient Statistics

where

$$h(\mathbf{x}^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\Sigma_{i=1}^{n}(x_i - \bar{x}_n)^2}{2\sigma^2}},$$

$$g[T(\mathbf{x}^n), \theta] = e^{-\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2}}.$$

It follows that $T(\mathbf{X}^n) = \bar{X}_n$ is a sufficient statistic for μ.

Sufficient Statistics

Example 12 (6.12)

Let $\mathbf{X}^n \sim$ IID $N(\mu, \sigma^2)$, where μ, σ^2 are unknown parameters. Then $T(\mathbf{X}^n) = (\bar{X}_n, S_n^2)$ is a sufficient statistic for (μ, σ^2).

Solution

- In this example, the unknown parameter $\theta = (\mu, \sigma^2)$ is a two-dimensional vector.
- Because the joint PDF of the random sample $\mathbf{X}^n$

To be Continued

Sufficient Statistics

$$
\begin{aligned}
& f_{\mathbf{X}^n}(\mathbf{x}^n, \mu, \sigma^2) \\
= \; & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
= \; & \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}} \\
= \; & \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{(n-1)[(n-1)^{-1}\Sigma_{i=1}^{n}(x_i-\bar{x}_n)^2]}{2\sigma^2} - \frac{n(\bar{x}_n-\mu)^2}{2\sigma^2}} \\
= \; & \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{(n-1)s_n^2+n(\bar{x}_n-\mu)^2}{2\sigma^2}} \\
= \; & g[T(\mathbf{x}^n), \theta] h(\mathbf{x}^n),
\end{aligned}
$$

where $h(\mathbf{x}^n) = 1$ for all $\mathbf{x}^n$, it follows that the two-dimensional statistic $T(\mathbf{X}^n) = (\bar{X}_n, S_n^2)$ is a sufficient statistic for $\theta = (\mu, \sigma^2)$.

Sufficient Statistics

Remarks:

- For a normally distributed random sample $\mathbf{X}^n$ with unknown μ and σ^2, it suffices to summarize the data by reporting the sample mean and sample variance, because $(\bar{X}_n, S_n^2)$ is a sufficient statistic for (μ, σ^2).
- However, suppose it is not normal. Then $(\bar{X}_n, S_n^2)$ may not be sufficient statistics. In other words, a sufficient statistic $T(\mathbf{X}^n)$ is generally model-dependent or population distribution dependent.

Sufficient Statistics

Question

Can you provide an example of population distribution for which $(\bar{X}_n, S_n^2)$ are not sufficient statistics for $\theta = (\mu, \sigma^2)$?

Sufficient Statistics

Theorem 11 (6.15). [Invariance Principle]

If $T(\mathbf{X}^n)$ is a sufficient statistic for θ, then any 1–1 function $R(\mathbf{X}^n) = r[T(\mathbf{X}^n)]$ is also a sufficient statistic for θ, and a sufficient statistic for the transformed parameter $r(\theta)$.

Proof:

- Because $T(\mathbf{X}^n)$ is a sufficient statistic for θ, the joint PMF/PDF of the random sample $\mathbf{X}^n$

$$f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = g[T(\mathbf{x}^n), \theta]h(\mathbf{x}^n)$$

for some functions $g(\cdot, \cdot)$ and $h(\cdot)$.

To be Continued

Sufficient Statistics

Proof:

- Next, because the function $r(\cdot)$ is a 1–1 mapping, its inverse function $r^{-1}(\cdot)$ exists and $T(\mathbf{x}^n) = r^{-1}[R(\mathbf{x}^n)]$. It follows that

$$
\begin{aligned}
f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) &= g\{r^{-1}[R(\mathbf{x}^n)], \theta\}h(\mathbf{x}^n) \\
&= \tilde{g}[R(\mathbf{x}^n), \theta]h(\mathbf{x}^n)
\end{aligned}
$$

where $\tilde{g}(\cdot, \theta) = g[r^{-1}(\cdot), \theta]$ depends on parameter θ. Hence, $R(\mathbf{X}^n)$ is a sufficient statistic for θ by the definition of sufficient statistic.

To be Continued

Sufficient Statistics

Proof:

- Similarly, because $\theta = r^{-1}[r(\theta)] = r^{-1}(\beta)$, where $\beta = r(\theta)$ is a transformed parameter, we have

$$\begin{aligned} f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) &= g\{r^{-1}[R(\mathbf{x}^n)], r^{-1}(\beta)\}h(\mathbf{x}^n) \\ &= g^*[R(\mathbf{x}^n), \beta]h(\mathbf{x}^n), \end{aligned}$$

where the function $g^*(\cdot, \beta) = g[r^{-1}(\cdot), r^{-1}(\beta)]$ depends on parameter β. It follows that $R(\mathbf{X}^n)$ is also a sufficient statistic for β.

To be Continued

Sufficient Statistics

Definition 9 (6.9). [Exponential Family]

A family of probability distributions is called an exponential family if their population PMF/PDF can be expressed as

$$f(x,\theta) = h(x)c(\theta)e^{\Sigma_{j=1}^{k} w_j(\theta)t_j(x)}.$$

Remarks:

- Most important distributions introduced in Chapter 4—both discrete and continuous—belong to the exponential family.

Sufficient Statistics

- An example is the normal $N(\mu, \sigma^2)$ distribution, whose PDF

$$
\begin{aligned}
f(x,\theta) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}},
\end{aligned}
$$

where

$$
\begin{aligned}
h(x) &= 1, & w_2(\theta) &= \frac{\mu}{\sigma^2}, \\
c(\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}}, & t_1(x) &= x^2, \\
& & t_2(x) &= x. \\
w_1(\theta) &= -\frac{1}{2\sigma^2}, & &
\end{aligned}
$$

Sufficient Statistics

Theorem 12 (6.16).

Let $\mathbf{X}^n = (X_1, \cdots, X_n)$ be an IID random sample from the population $f(x, \theta)$. If

$$f(x,\theta) = h(x)c(\theta)e^{\Sigma_{j=1}^{k} w_j(\theta)t_j(x)},$$

then the $k \times 1$ statistic vector

$$T(\mathbf{X}^n) = \left[\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i)\right]$$

is a sufficient statistic for θ.

Proof: This is left as an exercise.

Sufficient Statistics

Remarks:

- It is always true that the random sample $\mathbf{X}^n$ itself is a sufficient statistic for θ.

  This is because we can always partition the joint PMF/PDF of $\mathbf{X}^n$ as

  $$f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = g[T(\mathbf{x}^n), \theta]h(\mathbf{x}^n),$$

  where $T(\mathbf{x}^n) = \mathbf{x}^n$, $h(\mathbf{x}^n) = 1$, and $g[T(\mathbf{x}^n), \theta] = f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$ for all $\mathbf{x}^n$. By the factorization theorem, $T(\mathbf{X}^n) = \mathbf{X}^n$ is always a sufficient statistic.

Sufficient Statistics

Question:

There may exist many sufficient statistics for the same parameter θ. Sufficient statistics for θ may differ from each other in the degree of summarizing the sample information. What is the most efficient way to summarize information of θ that is contained in a random sample $\mathbf{X}^n$?

Sufficient Statistics

Definition 10 (6.10). [Minimal Sufficient Statistic]

A sufficient statistic $T(\mathbf{X}^n)$ is called a minimal sufficient statistic for parameter θ if, for any other sufficient statistic $R(\mathbf{X}^n)$,the statistic $T(\mathbf{X}^n)$ is a function of $R(\mathbf{X}^n)$. That is, for any sufficient statistic $R(\mathbf{X}^n)$, there always exists some function $r(\cdot)$ such that $T(\mathbf{X}^n) = r[R(\mathbf{X}^n)]$.

Remarks:

- All sufficient statistics of θ contain all sample information that is relevant to θ, but the minimal sufficient statistic achieves the greatest possible summary of the data among all sufficient statistics for parameter θ. Why?

Sufficient Statistics

- Suppose $T(\mathbf{X}^n) = r[R(\mathbf{X}^n)]$, and $t = r(\tau)$. Define two subsets of sample points in the sample space of $\mathbf{X}^n$:

$$\begin{aligned} A_n(\tau) &= \{\mathbf{x}^n : R(\mathbf{x}^n) = \tau\}, \\ B_n(t) &= \{\mathbf{x}^n : T(\mathbf{x}^n) = t\} \\ &= \{\mathbf{x}^n : r[R(\mathbf{x}^n)] = r(\tau)\}. \end{aligned}$$

The first subset $A_n(t)$ is indexed by t and the second subset $B_n(\tau)$ is indexed by τ, where $t = r(\tau)$. Then $A_n(\tau) \subseteq B_n(t)$.

Sufficient Statistics

- Therefore, the sample information summarized by $T(\mathbf{x}^n) = t$ is a larger set than the sample information summarized by $R(\mathbf{x}^n) = \tau$. This implies that $T(\mathbf{X}^n)$ summarizes the larger information of the random sample $\mathbf{X}^n$ for parameter θ.

- A minimal sufficient statistic is not unique. Any 1–1 function of a minimal sufficient statistic is also a minimal sufficient statistic.

How can one find a minimal sufficient statistic?

Sufficient Statistics

Theorem 13 (6.17).

Let $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$ be the PMF/PDF of a random sample $\mathbf{X}^n$. Suppose there exists a function $T(\mathbf{X}^n)$ such that, for two sample points $\mathbf{x}^n$ and $\mathbf{y}^n$ in the sample space of $\mathbf{X}^n$, the ratio of joint PMF/PDF $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)/f_{\mathbf{X}^n}(\mathbf{y}^n, \theta)$ is constant as a function of θ (i.e. is independent of θ) if and only if $T(\mathbf{x}^n) = T(\mathbf{y}^n)$. Then $T(\mathbf{X}^n)$ is a minimal sufficient statistic for parameter θ.

Proof:

- (1) First we shall show that $T(\mathbf{X}^n)$ is a sufficient statistic for θ under the stated condition.

To be Continued

Sufficient Statistics

Proof:

- Define the partition sets of the sample space of $\mathbf{X}^n$ induced by $T(\mathbf{x}^n) = t$ for a given t as $A(t) = \{\mathbf{x}^n : T(\mathbf{x}^n) = t\}$. For each $A(t)$, we choose and fix one element $\mathbf{x}_t^n \in A(t)$. In other words, for any sample point $\mathbf{x}^n$ with $T(\mathbf{x}^n) = t$, let $\mathbf{x}_t^n$ be a fixed element that is in the same set $A(t)$ as $\mathbf{x}^n$.

- Since $\mathbf{x}^n$ and $\mathbf{x}_t^n$ are in the same set $A(t)$, we have $T(\mathbf{x}^n) = T(\mathbf{x}_t^n)$ and hence, $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)/f_{\mathbf{X}^n}(\mathbf{x}_t^n, \theta)$ is constant as a function of θ given the condition in the theorem. Thus, we can define a function $h(\mathbf{x}^n) = f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)/f_{\mathbf{X}^n}(\mathbf{x}_t^n, \theta)$, which does not depend on θ and is a function of $\mathbf{x}^n$ only.

To be Continued

Sufficient Statistics

Proof:

- Also, define a function $g(t,\theta) = f_{\mathbf{X}^n}(\mathbf{x}_t^n, \theta)$. Then we have

$$\begin{aligned}
& f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) \\
= \; & f_{\mathbf{X}^n}(\mathbf{x}_t^n, \theta) \frac{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)}{f_{\mathbf{X}^n}(\mathbf{x}_t^n, \theta)} \\
= \; & f_{\mathbf{X}^n}(\mathbf{x}_t^n, \theta) h(\mathbf{x}^n) \\
= \; & g(t,\theta) h(\mathbf{x}^n) \\
= \; & g[T(\mathbf{x}^n), \theta] h(\mathbf{x}^n),
\end{aligned}$$

where the last equality follows from $t = T(\mathbf{x}^n)$. Thus, by the factorization theorem, $T(\mathbf{X}^n)$ is a sufficient statistic for θ.

To be Continued

Sufficient Statistics

Proof:

- (2) Now we shall show that $T(\mathbf{X}^n)$ is minimal. Let $\tilde{T}(\mathbf{X}^n)$ be any other sufficient statistic for θ. By the factorization theorem, there exist functions $\tilde{g}(\cdot, \cdot)$ and $\tilde{h}(\cdot)$ such that $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = \tilde{g}[\tilde{T}(\mathbf{x}^n), \theta]\tilde{h}(\mathbf{x}^n)$.

- Let $\mathbf{x}^n$ and $\mathbf{y}^n$ be any two sample points in the sample space of $\mathbf{X}^n$ with $\tilde{T}(\mathbf{x}^n) = \tilde{T}(\mathbf{y}^n)$. Then

$$\begin{aligned}\frac{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)}{f_{\mathbf{X}^n}(\mathbf{y}^n, \theta)} &= \frac{\tilde{g}[\tilde{T}(\mathbf{x}^n), \theta]\tilde{h}(\mathbf{x}^n)}{\tilde{g}[\tilde{T}(\mathbf{y}^n), \theta]\tilde{h}(\mathbf{y}^n)} \\ &= \frac{\tilde{h}(\mathbf{x}^n)}{\tilde{h}(\mathbf{y}^n)},\end{aligned}$$

which does not deped on θ.

To be Continued

Sufficient Statistics

Proof:

- Since the ratio $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)/f_{\mathbf{X}^n}(\mathbf{y}^n, \theta)$ does not depend on θ, the conditions of the present theorem imply $T(\mathbf{x}^n) = T(\mathbf{y}^n)$. In other words, we have $\tilde{T}(\mathbf{x}^n) = \tilde{T}(\mathbf{y}^n)$ implies $T(\mathbf{x}^n) = T(\mathbf{y}^n)$. This means that for any given $\mathbf{x}^n$,

$$\left\{\mathbf{y}^n : \tilde{T}(\mathbf{y}^n) = \tilde{T}(\mathbf{x}^n)\right\} \subseteq \{\mathbf{y}^n : T(\mathbf{y}^n) = T(\mathbf{x}^n)\}.$$

Thus, $T(\mathbf{x}^n)$ is minimal.

Sufficient Statistics

Example 13 (6.13)

Let $\mathbf{X}^n$ be an IID random sample from a $N(\mu, \sigma^2)$ population with both μ and σ^2 unknown. Let $\mathbf{x}^n$ and $\mathbf{y}^n$ denote two sample points in the sample space of $\mathbf{X}^n$, and let $(\bar{x}_n, s_X^2)$ and $(\bar{y}_n, s_Y^2)$ be the sample means and sample variances corresponding to $\mathbf{x}^n$ and $\mathbf{y}^n$ respectively. Then,

$$\begin{aligned}\frac{f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)}{f_{\mathbf{X}^n}(\mathbf{y}^n, \theta)} &= \frac{(2\pi\sigma^2)^{-n/2} e^{-[n(\bar{x}_n - \mu)^2 + (n-1)s_X^2]/2\sigma^2}}{(2\pi\sigma^2)^{-n/2} e^{-[n(\bar{y}_n - \mu)^2 + (n-1)s_Y^2]/2\sigma^2}} \\ &= 1\end{aligned}$$

if and only if $(\bar{x}_n, s_X^2) = (\bar{y}_n, s_Y^2)$. Thus, $(\bar{X}_n, S_n^2)$ is a minimal sufficient statistic for (μ, σ^2).

CONTENTS

Conclusion

- The basic idea of statistical analysis is to use a subset or sample information to infer the knowledge of the data generating process.
- In this chapter, we have introduced some basic concepts and ideas of statistical theory, including the concepts of population, random sample, data set, statistic, parameter and statistical inference.
- We examine in detail the statistical properties of two important statistics—sample mean and sample variance estimators, establishing the finite sample distribution theory for them under the assumption of an IID normal random sample.

Conclusion

- This finite sample theory highlights the importance of the Student-t and F distributions in statistical inference.
- Finally, we introduce the concept of sufficient statistic and discuss its role in data reduction. The sufficiency principle best captures the essential idea of statistical analysis, namely, how to most efficiently summarize the observed data in inference of the population distribution or population parameter.

Conclusion

Question

- How useful is the sufficiency principle in Big data analysis?
- What are other methods/techniques for data reduction?
  - Principal component analysis (PCA)
  - Factor analysis

Thank You !