# Econ 6190: Econometrics I
# Introduction to Statistical Inference

Chen Qiu

Cornell Economics

2024 Fall

# Contents

# Reference

- Hansen Ch. 5 and 6
- Casella and Berger, Ch. 6

# 1. Sampling Model

# Motivation

- Economists often collect data that consist of some observations on variables of interest

Table: Some Observations from March 2009 Current Population Survey

| Observation | Wage | Education |
|---|---|---|
| 1 | 37.93 | 18 |
| 2 | 40.87 | 18 |
| 3 | 14.18 | 13 |
| 4 | 16.83 | 16 |
| 5 | 33.17 | 16 |
| 6 | 29.81 | 18 |
| 7 | 54.62 | 16 |
| 8 | 43.08 | 18 |
| 9 | 14.42 | 12 |
| 10 | 14.90 | 16 |
| 11 | 21.63 | 18 |
| 12 | 11.09 | 16 |
| 13 | 10.00 | 13 |
| 14 | 31.73 | 14 |
| 15 | 11.06 | 12 |
| 16 | 18.75 | 16 |
| 17 | 27.35 | 14 |
| 18 | 24.04 | 16 |
| 19 | 36.06 | 18 |
| 20 | 23.08 | 16 |

- The statistical view of the table:

  a **random sample** from a large **population**, from which we can learn about the wages/education of the population

# The population

- **Definition**: Let $X$ be a random vector of interest. The distribution of $X$, denoted as $F$, is called **population distribution**, or **population**
- We have $n$ repeated observations made from $X$

$$\{X_1, X_2 \ldots X_n\},$$

  which we call a sample or data
- What we observe for $X_1$ is an realization of the random vector $X_1$
- Notation: Capital $X$ refers to a random variable; lowercase $x$ refers to a realization of variable $X$
- We need to model how these observations are collected

# The random sampling model

- **Definition**: The collection of random vectors $\{X_1, X_2 \ldots X_n\}$ are called **a random sample of size n from population** $F$ if $\{X_1 \ldots X_n\}$ are
  - **mutually independent**
  - have the **same marginal** distribution $F$
- Alternatively, we say $\{X_1 \ldots X_n\}$ are **independent and identically distributed (iid)** random vectors

- Because of the random sampling scheme, the joint pdf/pmf of $\{X_1 \ldots X_n\}$ is given by

$$\begin{aligned}\underbrace{f(x_1, x_2 \ldots x_n)}_{\text{joint pdf/pmf}} &= \underbrace{f(x_1)}_{\text{marginal pdf/pmf of } X_1} \cdot f(x_2) \cdots f(x_n) \\ &= \underbrace{\prod_{i=1}^{n} f(x_i)}_{\substack{\text{because of random sampling,} \\ \text{all marginal distributions are the same}}}\end{aligned}$$

- If $f(\cdot)$ is known, we can use the joint pdf/pmf of the random sample to calculate any probability events about the random sample

## Example: exponential distribution

- Let $\{X_1 \dots X_n\}$ be a random sample from the exponential distribution with parameter $\beta$:

$$f(x \mid \beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{1}{\beta} x}, & x \geq 0, \\ 0, & x < 0 \end{cases}$$

- Then, the joint pdf of $\{X_1 \dots X_n\}$ is

$$\begin{aligned} f(x_1, \dots x_n) &= \prod_{i=1}^{n} f(x_i \mid \beta) \\ &= \begin{cases} \left(\frac{1}{\beta}\right)^n e^{-\frac{1}{\beta} \sum_{i=1}^{n} x_i}, & x_i \geq 0, \text{ for all } i = 1, \dots n, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

- We may calculate

$$
\begin{aligned}
& P\{X_1 > 2, \ldots, X_n > 2\} \\
= & \int_2^\infty \cdots \int_2^\infty f(x_1, \ldots x_n) dx_1 \ldots dx_n \\
= & \int_2^\infty \cdots \int_2^\infty \left(\frac{1}{\beta}\right)^n e^{-\frac{1}{\beta}\sum_{i=1}^n x_i} dx_1 \ldots dx_n \\
= & e^{-2/\beta} \int_2^\infty \cdots \int_2^\infty \left(\frac{1}{\beta}\right)^{n-1} e^{-\frac{1}{\beta}\sum_{i=2}^n x_i} dx_2 \ldots dx_n (\text{integrate out } x_1) \\
& \vdots \\
= & \left(e^{-2/\beta}\right)^n = e^{-2n/\beta}
\end{aligned}
$$

- Alternatively, we may also calculate

$$
\begin{aligned}
& P\{X_1 > 2, \dots, X_n > 2\} \\
=& P\{X_1 > 2\} \cdot \ldots \cdot P\{X_n > 2\} \\
=& [P\{X_1 > 2\}]^n \\
=& \left(e^{-2/\beta}\right)^n = e^{-2n/\beta}
\end{aligned}
$$

- In general, calculation of such probabilities for any random sample may be difficult, even if the population distribution is known

# Statistics, parameters and estimators

- A **parameter** $\theta$ is any function of the population $F$
- A **statistic** is a function of sample $\{X_i : i = 1, \ldots, n\}$, say $T(X_1, \ldots X_n)$ for a real or vector valued function $T$
- A statistic is a random vector. Its distribution is called **sampling distribution**
  - Sampling distribution of $T(X_1, \ldots X_n)$ can be quite tractable if $\{X_1, \ldots X_n\}$ is a random sample
- An **estimator** $\hat{\theta}$ for a parameter $\theta$ is a **statistic** intended as a guess about $\theta$
  - $\hat{\theta}$ is an **estimate** when it is a specific (or realized) value calculated in a specific sample

# Example 1: Judging whether I have a fair coin

- I want to figure out whether I have a fair coin by flipping it 10 times and recording 0 for each tail and 1 for each head
- Sample: $\mathbf{X} = (X_1, X_2 \dots X_n)$, where $X_i$ is the result of $i$-th experiment
- Note $X_i \sim$ i.i.d. *Bernoulli*$(p)$. That is, the pmf of each $X_i$ is $f(x_i) = p^{x_i}(1-p)^{1-x_i}$
- The pmf of $\mathbf{X}$ is $f_{\mathbf{X}}(x_1, x_2 \dots x_n) = \prod_{i=1}^{10} p^{x_i}(1-p)^{1-x_i}$, known up to $p$
- The goal is to make some judgment about $p$
- A statistic is any function of $\mathbf{X}$, e.g.,

$$Y_1 = \{\text{number of heads}\} = \sum_{i=1}^{n} X_i$$

$$\begin{aligned} Y_2 &= \{\text{the order number of the first experiment resulting in heads, with 0 if no heads}\} \\ &= X_1 + 2X_2(1-X_1) + 3X_3(1-X_2)(1-X_1) + \dots \end{aligned}$$

- For example, if we observe a sample $\{0, 1, 1, 0, 0, 0, 1, 0, 1, 1\}$, $Y_1 = 5,\ Y_2 = 2$.
- Notice both $Y_1$ and $Y_2$ are random variables and have a distribution that depends on $p$
- For example, in this example, $Y_1$ follows a binomial distribution with parameter $(n, p)$

$$P\{Y_1 = k\} = \left( \begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^k, k = 1, \ldots, n,$$

where

$$\left( \begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

## Example 2: Estimate average income of a worker

- Suppose you want to estimate the average income of a worker aged between 25 and 65 who resides in Ithaca
- A sample of $n$ workers: $\mathbf{X} = \{X_1, X_2 \ldots X_n\}$, where $X_i \sim$ i.i.d. $F(\cdot)$, and $F(\cdot)$ is the unknown distribution of income
- The distribution of $\mathbf{X}$: $F_{\mathbf{X}}(x_1, x_2 \ldots x_n) = \prod_{i=1}^{n} F(x_i)$
- The parameter of interest is $\mu = \int u dF(u)$, the mean of the unknown income distribution
- A statistic is any function of $\mathbf{X}$, e.g.,

$$Y_1 = \frac{1}{n} \sum_{i=1}^{n} X_i \text{ (average).}$$

$$Y_2 = \text{average of 80\% of middle values (trimmed mean)}$$

- The distribution of $Y_1$ and $Y_2$ can be difficult to characterize

# The goal of this course

- Based on observed random sample/data $\{X_1 \dots X_n\}$, construct a "good" statistic to learn about the population parameter of interest $\theta$
- Here, "good" means "good statistical property". $\Rightarrow$ Requires careful evaluation of the sampling uncertainty (the underlying randomness of our data)$\Rightarrow$ Need to study the sampling distribution of any statistic
- Three approaches: Finite sample approach, asymptotic approach, and bootstrap

# Alternative sampling models

- **i.n.i.d.** sampling: each $X_i$ is independent but not necessarily identically distributed, i.e., $X_i$ is drawn from heterogeneous population $F_i$

- Bootstrap **with replacement**
  - a finite population of $N$ values $\{x_1, \ldots x_N\}$
  - Each $X_i, i = 1 \ldots n$, is drawn from the $N$ values with equal probability (think of drawing numbers from a hat)
  - Then, each $X_i$ is a **discrete** random variable that takes on values $\{x_1, \ldots x_N\}$ with equal probability $1/N$

$$P\{X_i = x_k\} = \frac{1}{N}, k = 1 \ldots N$$

  - The joint pmf of $\{X_1, X_2 \ldots X_n\}$ is

$$P\{X_1 = t_1, \ldots, X_n = t_n\} = \left(\frac{1}{N}\right)^n, t_j \in \{x_1, \ldots x_N\}, j = 1 \ldots n.$$

- Bootstrap **without replacement**
    - a finite population of $N$ values $\{x_1, \ldots x_N\}$
    - $X_1$ is drawn from the $N$ values with equal probability $\frac{1}{N}$. Record $X_1 = x_1$
    - $X_2$ is drawn from remaining $N-1$ values equal probability $\frac{1}{N-1}$. Record $X_1 = x_2$
    - $\ldots$
- With bootstrap **without replacement**, the sample we get

$$\{X_1 \ldots X_n\}$$

**does not** satisfy i.i.d assumption.

## Useful result

In bootstrap without replacement,

$$\{X_1 \dots X_n\}$$

are NOT independently distributed. However, they are identically distributed.

- Proof

# 2. Some Common Statistics

# Sample mean and sample variance

- We now define three statistics that are often used and provide goos summaries of the random sample

- The **sample mean** is the arithmetic average of the values in a random sample

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The **sample variance** is the statistic defined by

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

The **sample standard deviation** is the statistic defined by

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

# Properties of sample mean and sample statistics

- $\bar{X}$ and $s^2$ are themselves random variables
- We start by deriving some basic algebraic properties of the sample mean and variance

**Theorem**

The following are true:

- $\min_a \sum_{i=1}^n (X_i - a)^2 = \sum_{i=1}^n (X_i - \bar{X})^2$
- $(n-1)s^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\left(\bar{X}\right)^2$

# Proof

# Useful results

- We now begin our study of sampling distributions by considering their moments. The following result will be useful.

**Theorem**

Let $\{X_1, \ldots X_n\}$ be a random sample from a population. Let $g(x)$ be a function such that $\mathbb{E}g(X_1)$ and $\text{var}(X_1)$ exist. Then:

1. $\mathbb{E}\left[\sum_{i=1}^{n} g(X_i)\right] = n\mathbb{E}g(X_1)$;
2. $Var(\sum_{i=1}^{n} g(X_i)) = nVar(g(X_1))$

# Proof

# Moments of sample mean and variance

### Theorem

Let $\{X_1, \dots X_n\}$ be a random sample from a population with mean $\mu$ and variance $\sigma^2$, then:

1. $\mathbb{E}[\bar{X}] = \mu$,
2. $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$,
3. $\mathbb{E}[s^2] = \sigma^2$.

# Proof

- To prove (1), directly use the linearity of expectations and iid assumption
- To prove (2), note

$$
\begin{aligned}
\mathrm{var}[\bar{X}] &= \mathrm{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}\left[X_i\right] && \text{(by mutual independence)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}\left[X\right] && \text{(by identical distribution)} \\
&= \frac{1}{n}\mathrm{var}\left[X\right] = \frac{\sigma^2}{n}
\end{aligned}
$$

- Thus, the variance of sample mean declines with sample size at rate $\frac{1}{n}$

- To show (3), by the previous theorem,

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\left(\bar{X}_n\right)^2\right]$$

- Thus,

$$\begin{aligned}
\mathbb{E}\left[s^2\right] &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - n\mathbb{E}\left[\left(\bar{X}_n\right)^2\right]\right] \\
&= \frac{1}{n-1}\left[n\mathbb{E}\left[X_1^2\right] - n\mathbb{E}\left[\left(\bar{X}_n\right)^2\right]\right] \\
&= \frac{1}{n-1}\left[n\left(\mu^2+\sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] \\
&= \sigma^2,
\end{aligned}$$

where we have used

$$\begin{aligned}
\mathbb{E}\left[X_1^2\right] &= Var\left(X_1\right) + \left(\mathbb{E}\left[X_1\right]\right)^2, \\
\mathbb{E}\left[\left(\bar{X}_n\right)^2\right] &= Var\left(\bar{X}_n\right) + \left(\mathbb{E}\left[\bar{X}_n\right]\right)^2.
\end{aligned}$$

# 3. Sampling from Normal Distribution

# Motivation

- In order to make statistical inference, we often need to know the distribution of a statistics
- The most widely used statistical model assumes samples are drawn from a normal distribution
- In this section, we study the properties of common statistics when observations are normally distributed
- This also leads us to many well-known sampling distributions

# Normal sampling model

- Let $\{X_1, X_2, \ldots X_n\}$ be a random sample from a normal distribution $\mathrm{N}(\mu, \sigma^2)$. This is called a **normal sampling model**
- The normal sampling model has many attractive and tractable properties, since $\{X_1, X_2, \ldots X_n\}$ follows a multivariate normal distribution with positive-definite and diagonal covariance matrix
- Before studying sampling distribution under the normal sampling model, we first introduce the univariate and multivariate normal distributions.

# Univariate normal

- A random variable $Z$ has the standard normal distribution, written as $Z \sim \mathrm{N}(0,1)$, if it has the density
$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \; x \in \mathbb{R}.$$
- The cdf of a standard normal does not have a closed form but is written as
$$\Phi(x) = \int_{-\infty}^{x} \phi(u) du.$$
- Note key properties of $\phi(\cdot)$ and $\Phi(\cdot)$
  - $\int_{-\infty}^{\infty} \phi(x) dx = 1$ (a pdf must integrate to 1)
  - $\phi(x) = \phi(-x)$, and $\Phi(-x) = 1 - \Phi(x)$ (due to symmetry of $\phi(\cdot)$ around 0)

- If $Z \sim \mathrm{N}(0, 1)$, and $X = \mu + \sigma Z$ for $\mu \in \mathbb{R}$ and $\sigma \geq 0$, then $X$ has the normal distribution, written as $X \sim \mathrm{N}(\mu, \sigma^2)$.
- If $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\sigma > 0$, then $X$ has the density

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \; x \in \mathbb{R}.$$

# Moments of normal distribution

- All positive integer moments of the standard normal distribution are finite. This is because the tails of the density decline exponentially.
- If $Z \sim \mathrm{N}(0, 1)$, then $\mathbb{E}[Z] = 0$, $Var(Z) = 1$.
- For any positive integer $m$,

$$\mathbb{E}[Z^m] = \begin{cases} 0, & m \text{ odd}, \\ 2^{-\frac{m}{2}} \frac{m!}{(m/2)!} & m \text{ even}. \end{cases}$$

# Quantiles of standard normal

- The normal distribution is commonly used for statistical inference. Its quantiles are used for hypothesis testing and confidence interval construction

Figure: Normal probabilities and quantiles

| | $\mathbb{P}[Z \leq x]$ | $\mathbb{P}[Z > x]$ | $\mathbb{P}[\lvert Z\rvert > x]$ |
|---|---|---|---|
| $x = 0.00$ | 0.50 | 0.50 | 1.00 |
| $x = 1.00$ | 0.84 | 0.16 | 0.32 |
| $x = 1.65$ | 0.950 | 0.050 | 0.100 |
| $x = 1.96$ | 0.975 | 0.025 | 0.050 |
| $x = 2.00$ | 0.977 | 0.023 | 0.046 |
| $x = 2.33$ | 0.990 | 0.010 | 0.020 |
| $x = 2.58$ | 0.995 | 0.005 | 0.010 |

- Historically, statistical and econometrics textbooks would include extensive tables of normal (and other) quantiles. This is unnecessary today since these calculations are embedded in statistical software.

# Multivariate standard normal

- Let $\{Z_1, Z_2, \dots Z_m\}$ be iid standard normal. Therefore, the joint pdf of $\{Z_1, Z_2, \dots Z_m\}$ equals

$$\begin{aligned} f(z_1, \dots z_m) &= \prod_{i=1}^{m} f(z_i) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) \\ &= \frac{1}{(2\pi)^{\frac{m}{2}}} \exp\left(-\frac{\sum_{i=1}^{m} z_i^2}{2}\right) \\ &= \frac{1}{(2\pi)^{\frac{m}{2}}} \exp\left(-\frac{\mathbf{z}'\mathbf{z}}{2}\right), \end{aligned}$$

  where $\mathbf{z} = (z_1, z_2 \dots z_m)'$.

- The above density is called multivariate standard normal density

- **Definition**: An $m$ dimensional vector $\mathbf{Z}$ has the **multivariate standard normal distribution**, written $\mathbf{Z} \sim \mathrm{N}(0, I_m)$ if it has joint pdf

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{\frac{m}{2}}} \exp\left(-\frac{\mathbf{z}'\mathbf{z}}{2}\right)$$

- It is the joint pdf of $m$ independently and identically distributed standard normal random variables
- The mean of $\mathbf{Z}$ is $\mathbb{E}[\mathbf{Z}] = 0$, and the covariance matrix of $\mathbf{Z}$ is $\mathrm{var}(Z) = I_m$
- Since we have now introduced a vector of random variables, we next review some useful matrix-based notations.

# Expectation and covariance

- **Definition**: The expectation of $\mathbf{X} \in \mathbb{R}^m$ is the vector of expectations of its elements

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_m] \end{pmatrix}$$

- **Definition**: The $m \times m$ covariance matrix of $\mathbf{X} \in \mathbb{R}^m$ is

$$\begin{aligned} \Sigma = \mathrm{var}(\mathbf{X}) &= \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])\,(\mathbf{X} - \mathbb{E}[\mathbf{X}])'\right] \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{pmatrix} \end{aligned}$$

where on the diagonal $\sigma_j^2 = \mathrm{var}(X_j), j = 1 \ldots m$, and on the off-diagonal $\sigma_{ij} = \mathrm{cov}(X_i, X_j), i \neq j$

# Property of $\Sigma$

- **Theorem**: $\Sigma = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])'\right]$ is
  - symmetric: $\Sigma = \Sigma'$
  - positive semi-definite: for any vector $a \neq 0$, $a'\Sigma a \geq 0$
- Proof: Symmetry holds because $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$. For positive semi-definiteness,

$$\begin{aligned} a'\Sigma a &= a'\mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])'\right] a \\ &= \mathbb{E}\left[a'(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])' a\right] \\ &= \mathbb{E}\left\{\left[a'(\mathbf{X} - \mathbb{E}[\mathbf{X}])\right]^2\right\} \geq 0 \end{aligned}$$

since $\left[a'(\mathbf{X} - \mathbb{E}[\mathbf{X}])\right]^2 \geq 0$

# Property of expectation and covariance

- **Theorem**: If $\mathbf{X} \in \mathbb{R}^m$ has expectation $\mu$ and covariance matrix $\Sigma$, and $\mathbf{A}$ is $q \times m$, then $\mathbf{AX}$ is a random vector with mean $\mathbf{A}\mu$ and covariance $\mathbf{A}\Sigma\mathbf{A}'$

- Proof:

$$
\begin{aligned}
\mathbb{E}[\mathbf{AX}] &= \mathbf{A}\mathbb{E}[\mathbf{X}] = \mathbf{A}\mu \\
\mathrm{var}[\mathbf{AX}] &= \mathbb{E}\left[(\mathbf{AX} - \mathbb{E}[\mathbf{AX}])(\mathbf{AX} - \mathbb{E}[\mathbf{AX}])'\right] \\
&= \mathbb{E}\left[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}]))'\right] \\
&= \mathbb{E}\left[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])'\mathbf{A}'\right] \\
&= \mathbf{A}\mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])'\right]\mathbf{A}' \\
&= \mathbf{A}\Sigma\mathbf{A}'
\end{aligned}
$$

# Multivariate normal

- **Definition**: If $\mathbf{Z} \sim \mathrm{N}(0, I_m)$ and $\mathbf{X} = \mu + \mathbf{BZ}$ for $q \times m$ $\mathbf{B}$, then $\mathbf{X}$ has the multivariate normal distribution, written $\mathbf{X} \sim \mathrm{N}(\mu, \Sigma)$, with $q \times 1$ mean vector $\mu$ and $q \times q$ covariance matrix $\Sigma = \mathbf{BB}'$

- If $\mathbf{X} \sim \mathrm{N}(\mu, \Sigma)$ where $\Sigma$ is invertible, then $\mathbf{X}$ has pdf

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{m}{2}}(\det\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)}{2}\right)$$

- The mean of $\mathbf{X}$ is $\mathbb{E}[\mathbf{X}] = \mu$, the covariance matrix of $\mathbf{X}$ is $Var(\mathbf{X}) = \Sigma$.

# Property of multivariate normal

- **Theorem**: If $X$ and $Y$ are multivariate normal with $\text{cov}(X, Y) = 0$, then $X$ and $Y$ are independent
- **Theorem**: If $\mathbf{X} \sim \mathrm{N}(\mu, \Sigma)$ then

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X} \sim \mathrm{N}(\mathbf{a} + \mathbf{B}\mu, \mathbf{B}\Sigma\mathbf{B}')$$

- In words: if **X** is multivariate (jointly) normal, then any linear combination of **X** is also multivariate (jointly) normal
- However, note the following statement is WRONG:
    - Wrong statement: If $X$ and $Y$ are both normal, then $X + Y$ are also normal

- **Theorem**: If $(X, Y)$ are multivariate normal

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathrm{N}\left( \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} \right)$$

  with $\Sigma_{YY} > 0$ and $\Sigma_{XX} > 0$, then the conditional distributions $Y \mid X$ and $X \mid Y$ are also normal

$$\begin{aligned} Y \mid X &\sim \mathrm{N}\left(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right) \\ X \mid Y &\sim \mathrm{N}\left(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y), \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\right). \end{aligned}$$

# In summary

- Multivariate normal distribution has many attractive properties. The most important insight is:
  - If a random vector $\mathbf{X}$ has a multivariate normal distribution, then any of their **marginal** and **conditional** distributions are also multivariate normal
- We are now ready to study the sampling distribution of key statistics under the normal sampling model

# Sampling distribution under normal sampling model

- **Theorem**: if $X_i$, $i = 1 \dots n$ are i.i.d $\mathrm{N}(\mu, \sigma^2)$, then

$$\bar{X}_n \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

- **Proof**: use the fact that a linear combination of multivariate normal random variables is still normal

# Sampling distribution of sample variance

- Recall sample variance is

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

- To study its distribution under normal sampling, introduce the notion of $\chi_r^2$ **distribution**

- **Definition**: Let $\{Z_1, Z_2 \ldots Z_r\}$ be $r > 0$ i.i.d $\mathrm{N}(0, 1)$ random variables. Then $\sum_{i=1}^{r} Z_i^2$ follows a **chi square distribution with degrees of freedom** $r$, written as $\chi_r^2$

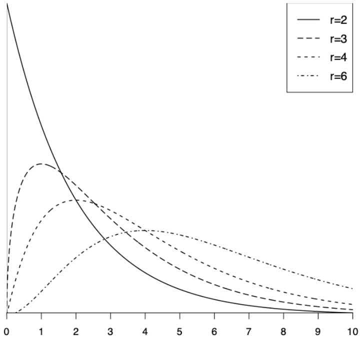



Figure: Chi-Square Densities

- **Theorem**: if $X_i$, $i = 1 \dots n$ are i.i.d $\mathrm{N}(\mu, \sigma^2)$, then

  1. $\bar{X}_n$ and $s^2$ are independent;
  2. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

# Proof of statement ❶

- Define residual $\hat{e}_i = X_i - \bar{X}_n$, $i = 1 \dots n$
- Note $\hat{e}_i$ is a linear combination of $X_1, \dots, X_n$, which are multivariate normal. So $\hat{e}_i$ is also normal
- Also $\mathbb{E}[\hat{e}_i] = \mathbb{E}[X_i] - \mathbb{E}[\bar{X}_n] = \mu - \mu = 0$, and

$$\begin{aligned}
\text{cov}(\hat{e}_i, \bar{X}_n) &= \mathbb{E}\left[\hat{e}_i \left(\bar{X}_n - \mu\right)\right] \\
&= \mathbb{E}\left[\left(X_i - \mu + \mu - \bar{X}_n\right)\left(\bar{X}_n - \mu\right)\right] \\
&= \mathbb{E}\left[\left(X_i - \mu\right)\left(\bar{X}_n - \mu\right)\right] - \mathbb{E}\left[\left(\bar{X}_n - \mu\right)^2\right] \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0
\end{aligned}$$

- Since $\hat{e}_i$ and $\bar{X}_n$ are jointly normal, uncorrelatedness means independence
- Thus, any function of $\hat{e}_i$ (including $s^2$) and $\bar{X}_n$ are also independent

# Proof of statement ❷

- We now show $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
- Write $s_n^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ and use proof by induction
- First verify that (left for homework)

$$(n-1)s_n^2 = (n-2)s_{n-1}^2 + \frac{n-1}{n}(X_n - \bar{X}_{n-1})^2 \qquad (1)$$

- Consider $n = 2$. Define $0 \cdot s_1^2 = 0$, so that we have

$$s_2^2 = (X_2 - \bar{X}_1)^2 = \frac{1}{2}(X_2 - X_1)^2$$

- Since $\frac{X_2 - X_1}{\sqrt{2\sigma^2}} \sim \mathrm{N}(0,1)$, we have

$$\frac{s_2^2}{\sigma^2} = \frac{1}{2\sigma^2}(X_2 - \bar{X}_1)^2 = \left(\frac{X_2 - X_1}{\sqrt{2\sigma^2}}\right)^2 \sim \chi^2_1$$

- Suppose when $n = k$, $k \geq 1$, $\frac{(k-1)s_k^2}{\sigma^2} \sim \chi^2_{k-1}$
- Then for $n = k+1$, we have from (1)

$$ks_{k+1}^2 = (k-1)s_k^2 + \frac{k}{k+1}(X_{k+1} - \bar{X}_k)^2$$

- Note we assumed $\frac{(k-1)s_k^2}{\sigma^2} \sim \chi^2_{k-1}$
- Proof is done if we can establish

$$(\blacktriangle) \quad \frac{k}{(k+1)\,\sigma^2}(X_{k+1} - \bar{X}_k)^2 \sim \chi^2_1$$

$$(\blacktriangledown) \quad \frac{k}{(k+1)\,\sigma^2}(X_{k+1} - \bar{X}_k)^2 \text{ is independent of } s_k^2$$

- $(\blacktriangle)$ follows from $X_{k+1} - \bar{X}_k \sim \mathrm{N}(0, \frac{k+1}{k}\sigma^2)$
- $(\blacktriangledown)$ follows from statement ❶ and $X_{k+1}$ independent of $s_k^2$

# Studentized t ratio

- We know if $\{X_1, \dots X_n\}$ are i.i.d $\mathrm{N}(\mu, \sigma^2)$, then

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0,1) \tag{2}$$

- If $\sigma$ is known, (2) can be used for inference on $\mu$
- Usually $\sigma$ is unknown. Replacing $\sigma$ with $s$, it is natural to consider distribution of $\frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}}$
- Note

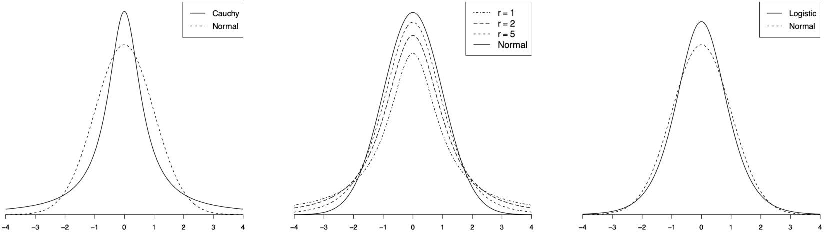
$$\frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} = \frac{\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\mathrm{N}(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{(n-1)}}}$$

Moreover, $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ is independent of $\sqrt{\frac{s^2}{\sigma^2}}$

- **Definition**: Let $Z \sim \mathrm{N}(0,1)$ and $Q \sim \chi^2_r$ be independent. Then $T = \frac{Z}{\sqrt{Q/r}}$ has a **Student's t distribution with** $r$ **degrees of freedom**, written as $T \sim t_r$
- **Theorem**: if $X_i$, $i = 1 \ldots n$ are i.i.d $\mathrm{N}(\mu, \sigma^2)$, then

$$\frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

# Student t distribution




(a) Cauchy and Normal

(b) Student t

(c) Logistic

Figure: Normal, Cauchy, Student t, and Logistic Densities

# Some facts about t distribution

- The pdf of $t_r$ distribution is symmetric around 0
- The pdf of $t_r$ distribution has heavier tails than $\mathrm{N}(0, 1)$
- Only the first $r - 1$ moment exists (vs. all moments of $\mathrm{N}(0, 1)$ exists)
- As $r \to \infty$, $t_r$ distribution is approaching to $\mathrm{N}(0, 1)$

# Motivation for F distribution

- Variability comparison of two independent populations $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\mathrm{N}(\mu_Y, \sigma_Y^2)$
- One ideal ratio is $\frac{\sigma_X^2}{\sigma_Y^2}$
- Information about the aforementioned ratio is contained in $\frac{s_X^2}{s_Y^2}$
- Since ${}^{(n-1)s_X^2}\!/_{\sigma_X^2} \sim \chi^2_{n-1}$, ${}^{(m-1)s_Y^2}\!/_{\sigma_Y^2} \sim \chi^2_{m-1}$

$$\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)}$$

# F distribution

- **Definition**: Let $Q_p \sim \chi^2_p$ and $Q_q \sim \chi^2_q$ be independent. Then $\frac{Q_p/p}{Q_q/q}$ follows an $F$ **distribution with** $p$ **and** $q$ **degrees of freedom**, written as

$$\frac{Q_p/p}{Q_q/q} \sim F_{p,q}$$

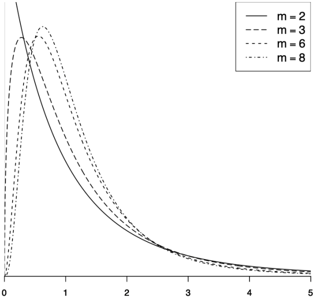



Figure: $F(m, r)$ Distribution Densities with $r = 10$

- **Theorem**: Let $\{X_1, \ldots, X_n\}$ be a random sample from $\mathrm{N}(\mu_X, \sigma_X^2)$ population. Let $\{Y_1, \ldots, Y_m\}$ be a random sample from an independent $\mathrm{N}(\mu_Y, \sigma_Y^2)$ population. Then

$$\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}$$

- Some facts about F distribution
  - If $X \sim F_{m,r}$, then $\frac{1}{X} \sim F_{r,m}$
  - If $X \sim t_q$, then $X^2 \sim F_{1,q}$

# 4. Sufficient Statistics

# Introduction

- Suppose we want to use a sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ to learn about a parameter of interest $\theta$
- All the information we can use is from $\mathbf{X}$
- However, $\mathbf{X}$ is a long list of vectors that can be hard to interpret
- As one data reduction technique, the concept of sufficient statistics allows to separate information from $\mathbf{X}$ into two parts: one part containing all useful information about $\theta$ and the other containing no useful information

# Sufficient statistics

- **Definition**: A statistic $T(\mathbf{X})$ is sufficient for $\theta$ if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ does not depend on $\theta$
- A sufficient statistic $T(\mathbf{X})$ contains all useful information about $\theta$ in the following sense
    - Experimenter 1 is provided with $\mathbf{X}$ and can learn about $\theta$ from pair $(\mathbf{X}, T(\mathbf{X}))$
    - Experimenter 2 is not provided with $\mathbf{X}$, but only $T(\mathbf{X})$
    - Since $T(\mathbf{X})$ is a sufficient statistics, the conditional distribution of $X$ given $T(\mathbf{X})$ is known to Experimenter 2
    - Experimenter 2 can back out the joint distribution of $(\mathbf{X}, T(\mathbf{X}))$ without knowing $\mathbf{X}$
    - Thus, Experimenter 2 has as much information as Experimenter 1

- **Theorem**: If $p(\mathbf{x}|\theta)$ is the joint pdf or pmf of $\mathbf{X}$ and $q(t|\theta)$ is the pdf or pmf of a statistic $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for $\theta$ if

  $$\frac{p(\mathbf{x}|\theta)}{q(t|\theta)} \text{ does not depend on } \theta \text{ for all } \mathbf{x} \text{ in the sample space.}$$

- Proof

# Example: Normal sufficient statistic with known variance

- Let $\{X_1 \dots X_n\}$ be iid $\mathrm{N}(\mu, \sigma^2)$ where $\sigma^2$ known
- We show that sample mean $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for $\mu$
- Note the joint pdf of the sample $\mathbf{X}$ is

$$
\begin{aligned}
f(\mathbf{x}|\mu) &= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\bar{x}+\bar{x}-\mu)^2}{2\sigma^2}\right) \\
&= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

where the last equality holds since the cross-product term $\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) = (\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x}) = 0$

- Recall in a normal sampling model $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$. It follows

$$\frac{p(\mathbf{x}|\theta)}{q(t|\theta)} = \frac{\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)}{(2\pi\sigma^2/n)^{-\frac{1}{2}} \exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)}$$

$$= n^{-\frac{1}{2}} \left(2\pi\sigma^2\right)^{-\frac{n-1}{2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2\sigma^2}\right),$$

which does not depend on $\mu$.

# Factorization Theorem

- It may be unwise to use the definition of a sufficient statistic to find a sufficient statistic for a particular parameter
- The following theorem allows find a sufficient statistic more conveniently
- **Theorem** (Factorization Theorem): Let $f(\mathbf{x}|\theta)$ be the joint pdf or pmf of $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for $\theta$ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and for all parameter points $\theta$

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}). \tag{3}$$

# Proof for Factorization Theorem

- We give a proof only for discrete distributions
- Only if: Suppose $T(\mathbf{X})$ is a sufficient statistic. Choose

$$g(t|\theta) = P_\theta\{T(\mathbf{X}) = t\}$$
$$h(\mathbf{x}) = P\{\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})\}.$$

Since $T(\mathbf{X})$ is sufficient, $h(\mathbf{x})$ does not depend on $\theta$. For this choice, we have

$$\begin{aligned} f(\mathbf{x}|\theta) &= P_\theta\{\mathbf{X} = \mathbf{x}\} \\ &= P_\theta\{\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x})\} \\ &= P_\theta\{T(\mathbf{X}) = T(\mathbf{x})\}P\{\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})\} \\ &= g(T(\mathbf{x})|\theta)h(\mathbf{x}) \end{aligned}$$

so the only if part is established

- For the if part, suppose factorization (3) exists
- Let $q(t|\theta)$ be the pmf of $T(\mathbf{X})$. To show $T(\mathbf{X})$ is sufficient, it suffices to examine the ratio $\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$ for each $\mathbf{x}$
- Define $A_{T(\mathbf{x})} = \{\mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})\}$. Then

$$
\begin{aligned}
\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} && \text{(apply (3))} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} f(\mathbf{x}|\theta)} && \text{(by definition of pmf)} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} && \text{(apply (3))} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\sum_{A_{T(\mathbf{x})}} h(\mathbf{y})} && (T \text{ is a constant on } A_{T(\mathbf{x})}) \\
&= \frac{h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} h(\mathbf{y})}
\end{aligned}
$$

which does not depend on $\theta$

# Example: Normal sufficient statistic with unknown variance

- Let $\{X_1 \ldots X_n\}$ be iid $\mathrm{N}(\mu, \sigma^2)$ where $\sigma^2$ **unknown**. Thus, the parameter is $\theta = (\mu, \sigma^2)$
- Note we already know

$$f(\mathbf{x}|\theta) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right),$$

  which depends on $\mathbf{x}$ only through $T_1(\mathbf{x}) = \bar{x}$, and $T_2(\mathbf{x}) = s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$
- We can define $h(\mathbf{x}) = 1$ and

$$g(t|\theta) = g(t_1, t_2|\mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)t_2 + n(t_1 - \mu)^2}{2\sigma^2}\right)$$

- Thus $f(\mathbf{x}|\theta) = g(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu, \sigma^2)h(\mathbf{x})$. By the Factorization Theorem,

$$T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\bar{X}, s^2)$$

  is a sufficient statistic for $(\mu, \sigma^2)$ in this normal model

## Example: discrete uniform distribution

- Let $\{X_1, \ldots, X_n\}$ be a random sample from the discrete uniform distribution on $\{1, 2 \ldots \theta\}$. That is, the pmf for $X_i$ is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & x = 1, 2 \ldots \theta, \\ 0, & \text{otherwise.} \end{cases}$$

  Show that $\max_i X_i$ is a sufficient statistic for $\theta$.

- Proof

# Refinement of sufficient statistic

- It should be obvious that each problem has numerous sufficient statistic. For example:
  - In the previous normal model with unknown variance, $\left(\bar{X}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$ is also a sufficient statistic
  - it is always true that the complete sample, $\mathbf{X}$, is sufficient statistic, as for all $\mathbf{x}$

$$f(\mathbf{x}|\theta) = f(T(\mathbf{X})|\theta)h(\mathbf{x}), \text{ by letting } T(\mathbf{X}) = \mathbf{x}, h(\mathbf{x}) = 1.$$

  - Also, any one-to-one function of a sufficient statistic is a sufficient statistic (exercise)
- Is there one sufficient statistic better than another?

# Minimal sufficient statistic

- **Definition**: A sufficient statistic $T^*(\mathbf{X})$ is a minimal sufficient statistic if for any sufficient statistic $T(\mathbf{X})$, there exists some function such that

$$T^*(\mathbf{X}) = r(T(\mathbf{X})).$$

- The above definition implies that, for any sufficient statistic $T(\mathbf{X})$, if $T(\mathbf{x}) = T(\mathbf{y})$, then $T^*(\mathbf{x}) = T^*(\mathbf{y})$
- Intuitively, the minimal sufficient statistic achieves the greatest data reduction without a loss of information about parameters

# Finding a minimal sufficient statistic

- **Theorem**: Let $f(\mathbf{x}|\theta)$ be the joint pdf or pmf of $\mathbf{X}$. Suppose there exists a $T(\mathbf{X})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio

  $$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} \text{ does not depend on } \theta \text{ if and only if } T(\mathbf{x}) = T(\mathbf{y}).$$

  Then $T(\mathbf{X})$ is a minimal sufficient statistic
- We leave this statement unproven here
- Note minimal sufficient statistic is also not unique

# Example: Normal minimal sufficient statistic

- Consider the previous example where $\{X_1 \dots X_n\}$ is iid $\mathrm{N}(\mu, \sigma^2)$ with $\sigma^2$ **unknown**
- Let $\mathbf{x}$ and $\mathbf{y}$be two sample points, and let $(\bar{x}, s_{\mathbf{x}}^2)$ and $(\bar{y}, s_{\mathbf{y}}^2)$ be the sample means and variances corresponding two the $\mathbf{x}$ and $\mathbf{y}$ samples, respectively
- It follows

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)s_{\mathbf{x}}^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)}{\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)s_{\mathbf{y}}^2 + n(\bar{y}-\mu)^2}{2\sigma^2}\right)}$$
$$= \exp\left(\frac{(n-1)(s_{\mathbf{y}}^2 - s_{\mathbf{x}}^2) + n(\bar{y}^2 - \bar{x}^2) + 2n\mu(\bar{x} - \bar{y})}{2\sigma^2}\right).$$

This ratio is a constant not depending on $(\mu, \sigma^2)$ if and only if $\bar{x} = \bar{y}$ and $s_{\mathbf{y}}^2 = s_{\mathbf{x}}^2$. Thus, $(\bar{X}, s^2)$ is a minimal sufficient statistic

# 4. Examples of Estimators and Measures of Their Quality

# Estimators and some examples

- An **estimator** $\hat{\theta}$ for a parameter $\theta$ is a also a **statistic**, intended as a guess about $\theta$
  - $\hat{\theta}$ is an **estimate** when it is a specific (or realized) value calculated in a specific sample
- Let population parameter be $\mu = \mathbb{E}[X]$
  - The **sample mean** is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Let population parameter be $\theta = \mathbb{E}[g(X)]$ for some known function $g$
  - An estimator is the sample mean of $g(X_i)$: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}[g(X_i)]$
- Let population parameter be $\beta = h(\mathbb{E}[g(X)])$ for some known functions $g$ and $h$
  - A plug-in estimator for $\beta$ is $\hat{\beta} = h(\hat{\theta}) = h\left(\frac{1}{n}\sum_{i=1}^{n}[g(X_i)]\right)$

# Quality of an estimator: estimation bias

- **Definition**: The **bias** of an estimator $\hat{\theta}$ of a parameter $\theta$ is

$$\text{bias}[\hat{\theta}] = \mathbb{E}[\hat{\theta}] - \theta$$

  - An estimator is is **unbiased** if the bias is zero
  - Bias depends on the population distribution $F$

- Let $\mathscr{F}$ be a collection of possible distributions

- An estimator $\hat{\theta}$ of a parameter $\theta$ is **unbiased in $\mathscr{F}$** if $\text{bias}[\hat{\theta}] = 0$ for every $F \in \mathscr{F}$

- **Theorem**: $\bar{X}$ is unbiased for $\mu = \mathbb{E}[X]$ if $\mathbb{E}|X| < \infty$
  - Sample mean is an unbiased estimator for population mean as long as population mean is finite

# Quality of an estimator: sampling variance

- **Definition**: The **variance** of an estimator $\hat{\theta}$, also called **sampling variance**, is $\mathrm{var}[\hat{\theta}]$
- We already know that If $\mathbb{E}X^2 < \infty$, then $\mathrm{var}[\bar{X}] = \frac{\sigma^2}{n}$, where $\sigma^2 = \mathrm{var}(X)$
- Therefore, the variance of $\bar{X}$ declines with sample size at rate $\frac{1}{n}$

# Estimation of sampling variance

- Sampling variance is the variance of an estimator and thus usually unknown!
- To estimate $\text{var}[\bar{X}_n]$, we need an estimator for

$$\sigma^2 = \text{var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

- The **plug-in** estimator for $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\bar{X}_n\right)^2$$

- **Theorem**: If $\sigma^2 < \infty$, then $\mathbb{E}[\hat{\sigma}^2] = (1 - \frac{1}{n})\sigma^2$ (proof left as homework).
- Question: is there an unbiased estimator for $\sigma^2$?

# Standard error

- **Definition**: The **standard error** of an estimator $\hat{\theta}$ for parameter $\theta$ is

$$se(\hat{\theta}) = \hat{V}^{1/2}, \text{ where } \hat{V} \text{ is an estimator for } V = \text{var}[\hat{\theta}]$$

- Standard error can be interpreted as an estimator for $V^{1/2}$, the **standard deviation** of $\hat{\theta}$
- Standard error is usually a biased estimator of $V^{1/2}$
- Example:
  - sample mean $\bar{X}_n$ is an estimator for $\mu$
  - the exact variance of $\bar{X}_n$ is $\frac{\sigma^2}{n}$
  - if we estimate $\sigma^2$ by the plug-in estimator $\hat{\sigma}^2$
  - the standard error of $\bar{X}_n$ is $\sqrt{\frac{\hat{\sigma}^2}{n}}$

## Quality of an estimator: mean square error

- A standard measure of estimation quality is mean square error (MSE)

- **Definition**: The **mean square error** of an estimator $\hat{\theta}$ for $\theta$ is

$$\text{mse}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

- **Theorem**: For any estimator with a finite variance

$$\text{mse}(\hat{\theta}) = \text{var}(\hat{\theta}) + (\text{bias}[\hat{\theta}])^2$$

- **Proof**: start from

$$\begin{aligned} \text{mse}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \end{aligned}$$

  and apply standard algebra

- An estimator with smaller MSE is considered to be better, or more efficient

# Best unbiased estimator

- **Among a class of unbiased estimators**, the one with the lowest sampling variance also has the smallest MSE
- This motivates finding the best unbiased estimator for estimating parameter $\theta$
- **Theorem**: If $\sigma^2 < \infty$, the sample mean $\bar{X}_n$ has the lowest variance among all **linear unbiased estimators** of $\mu$

# Proof

- Consider a class of linear estimators

$$\tilde{\mu} = \sum_{i=1}^{n} w_i X_i$$

with some weights $\{w_1, \ldots w_n\}$

- Unbiasedness requires

$$\mu = \mathbb{E}\tilde{\mu} = \sum_{i=1}^{n} w_i \mathbb{E}[X_i] = \sum_{i=1}^{n} w_i \mu$$

which holds if and only if

$$\sum_{i=1}^{n} w_i = 1$$

- The variance of $\tilde{\mu}$ is

$$\mathrm{var}(\tilde{\mu}) = \mathrm{var}\left(\sum_{i=1}^{n} w_i X_i\right) \overset{\text{(independence)}}{=} \sum_{i=1}^{n} w_i^2 \mathrm{var}\left(X_i\right) = \sigma^2 \sum_{i=1}^{n} w_i^2$$

- Hence the best unbiased linear estimator solves

$$\min_{w_1 \ldots w_n} \sum_{i=1}^{n} w_i^2, \text{ s.t.} \sum_{i=1}^{n} w_i = 1$$

which has an Lagrangian

$$L(w_1, \ldots w_n) = \sum_{i=1}^{n} w_i^2 - \lambda \left( \sum_{i=1}^{n} w_i - 1 \right)$$

- FOC with respect to $w_i, i = 1 \ldots n$ is

$$2w_i - \lambda = 0 \Rightarrow w_i = \frac{\lambda}{2}$$

implying $w_i = \frac{1}{n}$ in order to satisfy $\sum_{i=1}^{n} w_i = 1$. Conclusion follows

- In fact, we have a much stronger statement
- **Theorem**: If $\sigma^2 < \infty$, the sample mean $\bar{X}_n$ has the lowest variance among all **unbiased estimators** of $\mu$

# Multivariate means

- Let $X \in \mathbb{R}^m$ be a random vector and $\mu = \mathbb{E}[X]$ be its mean. The sample mean estimator for $\mu$ is

$$\begin{aligned}\bar{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \begin{pmatrix} \bar{X}_{1n} \\ \bar{X}_{2n} \\ \vdots \\ \bar{X}_{mn} \end{pmatrix}\end{aligned}$$

- Most properties of the univariate sample mean extend to the multivariate mean

- The multivariate mean is unbiased for the population expectation: $\mathbb{E}\left[\bar{X}_n\right] = \mu$

- The exact covariance matrix of $\bar{X}_n$ is

$$Var\left(\bar{X}_n\right) = \mathbb{E}\left[\left(\bar{X}_n - \mathbb{E}(\bar{X}_n)\right)\left(\bar{X}_n - \mathbb{E}(\bar{X}_n)\right)'\right]$$
$$= \frac{1}{n} Var(X) = \frac{\Sigma}{n}$$

- The MSE matrix of $\bar{X}_n$ is

$$MSE\left(\bar{X}_n\right) = \mathbb{E}\left[\left(\bar{X}_n - \mu\right)\left(\bar{X}_n - \mu\right)'\right] = \frac{\Sigma}{n}$$

- $\bar{X}_n$ is the best unbiased estimator for $\mu$

- An unbiased covariance matrix estimator is

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(X_i - \bar{X}_n\right)\left(X_i - \bar{X}_n\right)'\right]$$

# Connection between efficiency and sufficient statistics

- Suppose we have a random sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ from a distribution $F_\theta$, where $\theta \in \mathbb{R}^k$ is the parameter of interest
- Let $\widehat{\theta} := \widehat{\theta}(\mathbf{X})$ be a candidate estimator for $\theta$ that we, as researchers, think is "good" (e.g., it has some desirable MSE properties)
- Suppose we also know that $T(\mathbf{X})$ is a sufficient statistics for $\theta$
- Question: Can we do better than $\widehat{\theta}$?

# Rao-Blackwell Theorem

## Rao-Blackwell Theorem

Under the setup from last slide, let

$$\tilde{\theta}(\mathbf{X}) := \mathbb{E}\left[\widehat{\theta}(\mathbf{X}) \mid T(\mathbf{X})\right].$$

Then,

1. $MSE(\tilde{\theta}(\mathbf{X})) \leq MSE(\widehat{\theta}(\mathbf{X}))$
2. If $\widehat{\theta}(\mathbf{X})$ is an unbiased estimator, so is $\tilde{\theta}(\mathbf{X})$

# Proof