DASC7011 Statistical Inference for Data Science

Chapter 2 Method of Moments Estimation

Department of Statistics and Actuarial Science
Department of Computer Science
The University of Hong Kong

Aug. 2024

Contents

§2.1 Method-of-Moments Estimation

2.1.1. Moments

- Let $X \sim F$ be a random variable with distribution F.
- The n-th **raw moment** (crude moment, moment about zero) of X (or F) is defined as (if exist):

$$\mu_n = \mathbb{E}(X^n), \qquad n \geq 1.$$

- The first (order) raw moment $\mu_1 = \mathbb{E}(X)$ is usually referred to as the **mean** or **expectation** of X, and denoted as μ.
- **Hausdorff moment problem**: For a distribution of mass or probability on a bounded interval, the collection of all the moments uniquely determines the distribution.

Central moments

- Suppose $X \sim F$ has a finite mean $\mu < \infty$.
- The n-th **central moment** of X is defined as:

$$\sigma_n = \mathbb{E}(X - \mu)^n, \qquad\qquad n \geq 2.$$

- The second (order) central moment σ_2, if exists, is called the **variance** of X, and denoted as σ^2.
- The non-negative square root of σ^2, denoted as σ, is called the **standard deviation** of X.

Standardized moments: skewness

- Suppose $X \sim F$ has a finite variance $\sigma^2 < \infty$, and hence has a finite mean $\mu < \infty$.
- The n-th **standardized moment**:

$$\eta_n = \frac{\sigma_n}{\sigma^n}, \qquad n \geq 3.$$

- The 3rd standardized moment η_3 is called the **skewness**.

$$\eta = \eta_3 = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \mathbb{E}\left(\frac{X-\mu}{\sigma}\right)^3.$$

  - Skewness is a measure of the asymmetry of the distribution F. We say F is positively or negatively skewed if $\eta > 0$ of < 0.
  - If X is normal, then $\eta = 0$.

Standardized moments: kurtosis

- The 4th standardized moment η_4 is called the **kurtosis**.

$$\kappa = \eta_4 = \frac{\mathbb{E}(X-\mu)^4}{\sigma^4} = \mathbb{E}\left(\frac{X-\mu}{\sigma}\right)^4.$$

  - Kurtosis is a measure of the "tailedness" of the distribution F.
  - If X is normal, then $\kappa = 3$.
  - $\kappa > 3$ indicates that the distribution F has *fatter* tail(s) than normal.
  - We call $(\kappa - 3)$ the **excess kurtosis**.

Covariance and correlation

- When a random vector, say (X, Y) is considered, two more important moments are frequently studied.
- Suppose X has finite mean μ_x and variance σ_x^2, and Y has finite mean μ_y and variance σ_y^2.
- The *mixed 2nd central moment*

$$\sigma_{x,y} = \mathbb{E}[(X - \mu_x)(Y - \mu_y)]$$

  is called the **covariance** between X and Y.
- The *mixed 2nd standardized moment*

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$$

  is called the **correlation**, or **linear correlation coefficient**, between X and Y.

2.1.2. Sample moments

- Moments are expectations, or functions of expectations, of X, the **population**.
- Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a **sample** from a population $X \sim F$.
- **The principal statistical idea**: estimating the expectation $\mathbb{E}[f(X)]$ by the average

$$\overline{f(X)} = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \tag{2.1}$$

  where $f(\cdot)$ is a (general) function.
- Applying this *averaging algorithm* to moments, we obtain **sample moments**.

Sample moments

- **Sample raw moments**:

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad k = 1, 2, \cdots$$

- **Sample central moments** when μ is known:

$$\widehat{\delta}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^k, \qquad k = 2, 3, \cdots$$

- **Sample central moments** when μ is unknown:

$$\widehat{\sigma}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_1)^k, \qquad k = 2, 3, \cdots$$

Sample mean and sample 2nd central moment

- We call the sample 1st raw moment

$$\overline{X}_n = \widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

the **sample mean** of X.

- The **sample 2nd central moment** of X is defined as

$$W_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \qquad \mu \text{ is known,}$$

$$T_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad \mu \text{ is unknown,}$$

where the suffix n in $\overline{X}_n$ is dropped for simplicity.

Sample variance

- Generally, the population mean μ would be unknown.
- We call the *bias-corrected* sample 2nd central moment

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

the **sample variance** of X.

- The non-negative square root of S_n^2,

$$S_n = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{1/2},$$

is called the **sample standard deviation**.

Sample skewness and sample kurtosis

- **Sample skewness**:

$$\widehat{\eta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\widehat{\sigma}_2^{3/2}} \quad \text{or} \quad \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{S^3}.$$

- **Sample kurtosis**:

$$\widehat{\kappa} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\widehat{\sigma}_2^{2}} \quad \text{or} \quad \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{S^4}.$$

- **Remark**: There are several slightly different definitions of sample skewness and sample kurtosis in literature, regarding the unbiasedness and/or efficiency.

Sample covariance and sample correlation

- Let $\{(X_i, Y_i) : 1 \leq i \leq n\}$ be a random sample from a bivariate population (X, Y), $\overline{X}$ and $\overline{Y}$ be the sample means, S_x^2 and S_y^2 be the sample variances, respectively.

- The **sample covariance** is defined as

$$S_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

- The **sample correlation** is defined as

$$\widehat{\rho}_{x,y} = \frac{S_{x,y}}{S_x S_y}.$$

It is also known as the **Pearson correlation coefficient** in literature, and denoted as $r_{x,y}$.

Sample moments in R

- **Example 2.1**: We use two generated samples with size $n = 100$, $\{X_i\} \overset{i.i.d.}{\sim} N(5,4)$ and $\{Y_i\} \overset{i.i.d.}{\sim} t(4)$, for illustration.

- Built-in functions of sample moments in R.

Moment	$\overline{X}$	S_x^2	S_x	$S_{x,y}$	$r_{x,y}$	$\widehat{\eta}_x$	$\widehat{\kappa}_x - 3$
R Function	`mean`	`var`	`sd`	`cov`	`cor`	`skewness`*	`kurtosis`*
Estimate	4.596	4.159	2.039	0.052	0.020	-0.039	-0.459

* In R package `e1071`.

- You can calculate them using your own functions/R expressions, to check the definitions of built-in formulas.

- In my attached program, both build-in and self-defined functions give the same estimates (estimated results). □

2.1.3. Method-of-Moments Estimation (MME)

- It is noticeable that the skewness is NOT an *original moment* (an expectation) by definition, but a (rational) function of two moments σ_3 and σ_2. In the meantime, we defined the sample skewness as (the same) function of corresponding sample moments $\widehat{\sigma}_3$ and S^2 (or $\widehat{\sigma}_2$).

- This is an immediate generalization of the principal statical idea, or the averaging algorithm.

- The idea behind: there exist certain mathematic relationship(s) between the target character (skewness, to be estimated) and moments.

- This leads to the **Method-of-Moments Estimation** (**MME**), or the MME algorithm.

Method-of-Moments Estimation (MME)

- Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a sample from a parametric (not a must) model $X \sim \mathcal{M}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are unknown parameter(s).

- The moments m_k's, if exist, will generally depend on the parameter(s) $\boldsymbol{\theta}$. If possible, they might be expressed as functions of $\boldsymbol{\theta}$, either explicit or implicit,

$$m_k = f_k(\boldsymbol{\theta}), \qquad 1 \leq k \leq p,$$

  for some $p \geq 1$. These functions are also known as *moment conditions* or *moment equations*.

- Suppose we can solve these moment equations for parameter(s) to obtain

$$\boldsymbol{\theta} = g(m_1, m_2, \cdots, m_p).$$

Method-of-Moments Estimation (MME)

- Replacing m_k's in the solution with corresponding sample moments $\widehat{m}_k$'s, we obtain the **method-of-moments estimate(s)** (**MME**) of parameter(s):

$$\widehat{\boldsymbol{\theta}} = g(\widehat{m}_1, \widehat{m}_2, \cdots, \widehat{m}_p). \tag{2.2}$$

- By definition, all sample moments in subsection 2.1.2 are MMEs of corresponding population moments.
- **Remark**: moment equations are not always solvable for parameters $\boldsymbol{\theta}$ (either existence or uniqueness). The **generalized method-of-moments estimation (GMME)** method provides a generalization of MME for this purpose.
  - GMME applies a statistical algorithm called *least-squares* to the estimation, and will be introduced in the next chapter.

Example 2.2: MMEs are not unique

- **Example 2.2**: Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample from an exponential distribution (population) with unknown rate λ. The pdf is $f(x) = \lambda e^{-\lambda x}$ for all $x > 0$.

- $\widehat{\lambda}_1 = \overline{X}^{-1}$ is one possible MME of λ since $\mathbb{E}(X) = 1/\lambda$.

- $\widehat{\lambda}_2 = S^{-1}$ is another valid MME of λ since $\mathrm{Var}(X) = 1/\lambda^2$, where S is the sample standard deviation of $\boldsymbol{X}$.

- **Remarks**: MMEs of parameters are not unique. Generally, we prefer using lower order moment(s) in finding MME(s) due to its/their better properties (convergence under weaker conditions).
    - We prefer $\widehat{\lambda}_1 = \overline{X}^{-1}$ in this example. $\square$

Example 2.3: MME for AR models

- **Example 2.3**: Consider a sample $\{X_t : 1 \leq t \leq T\}$ from an AR(1) model

$$X_t = \phi_0 + \phi_1 X_{t-1} + a_t, \qquad \{a_t\} \overset{i.i.d.}{\sim} (0, \sigma_a^2),$$

  where ϕ_0, $-1 < \phi_1 < 1$ and $\sigma_a^2 > 0$ are three parameters to be estimated.

- Under condition $-1 < \phi_1 < 1$, the model is *second order stationary*: the 1st and 2nd order moments are stationary (invariant) against time drift/transformation.

Example 2.3: MME for AR models

- Three moment conditions are:

$$\begin{aligned}
\mu_x &= \mathbb{E}(X_t) = \frac{\phi_0}{1-\phi_1}, \\
\rho_1 &= \mathrm{Corr}(X_t, X_{t-1}) = \phi_1, \\
\gamma_0 &= \mathrm{Var}(X_t) = \frac{\sigma_a^2}{1-\phi_1^2}.
\end{aligned}$$

- Solving these, we have

$$\phi_1 = \rho_1, \qquad \phi_0 = \mu_x(1-\rho_1), \qquad \sigma_a^2 = \gamma_0(1-\rho_1^2).$$

Example 2.3: MME for AR models

- Let $\overline{X}$ and S^2 be the sample mean and sample variance of $\{X_t\}$, respectively. Moreover, defined the *sample autocorrelation* at lag 1 as

$$\begin{aligned} r_1 &= \widehat{\mathrm{Corr}}(X_t, X_{t-1}) \\ &= \frac{\sum_{t=2}^{T}(X_t - \overline{X})(X_{t-1} - \overline{X})}{\sum_{t=1}^{T}(X_t - \overline{X})^2}. \end{aligned}$$

- Then, the MMEs are

$$\widehat{\phi}_1 = r_1, \qquad \widehat{\phi}_0 = \overline{X}(1 - r_1), \qquad \widehat{\sigma}_a^2 = S^2(1 - r^2).$$

□

Example 2.4: MME for Gamma distribution

- **Example 2.4**: Let $\{X_i : 1 \le i \le n\}$ be a random sample from a Gamma population $X \sim \Gamma(\alpha, \beta)$ with pdf

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0,$$

  where $\alpha > 0$ and $\beta > 0$ are the *shape* and *rate* parameters.

- There is no explicit formula for the maximum likelihood estimator (MLE, will be introduced in Chapter 4) due to the complexity of Gamma function $\Gamma(\alpha)$.

- On the other hand, the MMEs can be easily obtained.

Example 2.4: MME for Gamma distribution

- The Gamma function is defined as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \mathrm{d}t, \qquad \alpha > 0,$$

and has property $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

- Then,

$$\begin{aligned}
\mathbb{E}(X) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\beta x} x \mathrm{d}x \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty (\beta x)^\alpha \beta^{-\alpha-1} e^{-\beta x} \mathrm{d}(\beta x) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \beta^{-\alpha-1} \Gamma(\alpha+1) = \frac{\alpha}{\beta}.
\end{aligned}$$

Example 2.4: MME for Gamma distribution

- Similarly, it can be found that

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \frac{\alpha}{\beta^2}.$$

- These solve for

$$\alpha = \frac{[\mathbb{E}(X)]^2}{\mathrm{Var}(X)}, \qquad \beta = \frac{\mathbb{E}(X)}{\mathrm{Var}(X)}.$$

- Therefore, the MMES are

$$\widehat{\alpha} = \frac{(\overline{X})^2}{S^2}, \qquad \widehat{\beta} = \frac{\overline{X}}{S^2},$$

where $\overline{X}$ and S^2 are the sample mean and sample variance of $\{X_i\}$, respectively. $\square$

Example 2.5: MME for rate/probability

- **Example 2.5**: Let X denote the age of a general people in Hong Kong. We are interested in the proportion (rate/probability) of elder people, say, $p = \mathbb{P}(X \geq 65)$.

- Notice that $p = \mathbb{E}[1(X \geq 65)]$, where $1(A)$ stands for the indicator function of event A. The MME can be applied to estimate this rate.

- Suppose a random sample $\{X_i : 1 \leq i \leq n\}$ is drawn from the population X. Then

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \geq 65).$$

$\square$

§2.2 Statistical Inferences Based on MMEs

- Properties (mean, variance, bias and distribution) of sample moments, as estimates of population moments, are studied.
- Let $0 < \alpha < 1$ be a constant. Through out this section, all confidence intervals are constructed at the $(1 - \alpha)$-level of confidence, and all tests are conducted at the α-level of significance.

Properties of sample mean

2.2.1. Statistical inference on one mean.

- Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean of a *random sample* $\{X_i : 1 \le i \le n\}$ from a population $X \sim F$ with finite mean μ and variance σ^2.

- $\mathbb{E}(\overline{X}_n) = \mu$ and $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$. Hence, $\overline{X}_n$ is an unbiased and consistent estimator of μ.

- If $X \sim N(\mu, \sigma^2)$ is normal, then

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right). \qquad (2.3)$$

- For a general population F, by CLT, $\overline{X}$ is asymptotically normal,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1) \qquad \text{for large } n. \qquad (2.4)$$

(a) σ^2 is known.

(i) If $X \sim N(\mu, \sigma^2)$ is normal, by (2.3),

$$\mathbb{P}\left(-Z_{\alpha/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1 - \alpha,$$
$$\Leftrightarrow \mathbb{P}\left(\overline{X}_n - Z_{\alpha/2}\sigma/\sqrt{n} < \mu < \overline{X}_n + Z_{\alpha/2}\sigma/\sqrt{n}\right) = 1 - \alpha.$$

Therefore, the confidence of μ is

$$\left(\overline{X}_n - Z_{\alpha/2}\sigma/\sqrt{n},\ \overline{X}_n + Z_{\alpha/2}\sigma/\sqrt{n}\right). \tag{2.5}$$

Cf. Example 1.2.

(ii) For general F, by the asymptotical distribution in Eq. (2.4), the CI in (2.5) is approximately valid for large n.

CIs of one population mean

(b) σ^2 is unknown.

(i) If $X \sim N(\mu, \sigma^2)$ is normal, then $\overline{X}_n$ and S_n^2 are independent. Moreover,

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1), \quad \text{and} \quad \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1). \tag{2.6}$$

Therefore, the exact confidence interval of μ becomes

$$\left(\overline{X}_n - t_{\alpha/2}(n-1)S_n/\sqrt{n},\ \overline{X}_n + t_{\alpha/2}(n-1)S_n/\sqrt{n}\right). \tag{2.7}$$

(ii) For general F, by the asymptotical distribution in (2.4), and the fact that $t(n-1)$ is very closed to $N(0,1)$ when n is large, the approximate CI is

$$\left(\overline{X}_n - Z_{\alpha/2}S_n/\sqrt{n},\ \overline{X}_n + Z_{\alpha/2}S_n/\sqrt{n}\right). \tag{2.8}$$

Testing one population mean

- For illustration, consider testing $H_0: \ \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$ for some constant μ_0.

(a) σ^2 is known.

(i) When $X \sim N(\mu, \sigma^2)$ is normal, the test statistic is

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \text{under } H_0.$$

Therefore, the rejection region is

$$R_\alpha = \{|Z| > Z_{\alpha/2}\},$$

where $Z_{\alpha/2}$ is the upper quantile of standard normal at the $\frac{\alpha}{2}$ level.

(ii) For general F, the above results are approximately true for large n.

Testing one population mean

(b) σ^2 is unknown.

(i) When $X \sim N(\mu, \sigma^2)$ is normal, the test statistic is

$$T = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1) \qquad \text{under } H_0,$$

where $t(n-1)$ is the student-t distribution with $(n-1)$ degrees of freedom. Therefore, the rejection region is

$$R_\alpha = \left\{|T| > t_{\alpha/2}(n-1)\right\},$$

where $t_{\alpha/2}(n-1)$ is the upper quantile of $t(n-1)$ distribution at the $\frac{\alpha}{2}$ level.

(ii) For general F, the approximate (for large n) rejection region is

$$R_\alpha = \left\{|T| > Z_{\alpha/2}\right\}.$$

Example 2.6: Inferences on one mean

- **Example 2.6.** Consider the sample $\{x_i : 1 \le i \le 100\}$ generated from a normal $N(5, 2^2)$ population in Example 2.1, for which we have $\overline{X} = 4.596$ and $S = 2.039$. We are constructing the 95% confidence intervals, and testing $H_0 : \mu = 4$ at the 5% significance level.

(i) Assuming $\sigma = 2$ is known. There is no build-in R function for this case.

  - Using Formula (2.5), the 95% CI is calculated to be (4.204, 4.988).
  - The test statistic is $z = 2.982$. H_0 is rejected since $|z| > Z_{0.025} = 1.96$. Moreover, the p-value is (where $Z \sim N(0,1)$)

$$p = 2 \times \mathbb{P}(|Z| > |z|) = 0.00286.$$

Example 2.6: Inferences on one mean

(ii) Manual calculation assuming σ is unknown.

- Using Formula (2.7), the 95% CI is calculated to be (4.192, 5.001).
- The test statistic is $t = 2.924$. H_0 is rejected since $|z| > t_{0.025}(99) = 1.984$. The p-value is (where $T \sim t(99)$)

$$p = 2 \times \mathbb{P}(|T| > |t|) = 0.00428.$$

(iii) The build-in R function `t.test()` can be used to do the inferences in case (ii) including both the test and CI.

```
# (iii) Using t.test()
t.test(x, alternative = "two.sided", mu=mu0)
t.test(x, alternative = "less", mu=mu0)
t.test(x, alternative = "greater", mu=mu0)
```

Example 2.6: Inferences on one mean

- The first line tests H_0 against $H_1 : \mu \neq 4$. It gives the same results as in (ii).
- The last two lines do the one-sided test. Refer to the following output.

```
> t.test(x, alternative = "greater", mu=mu0)
        One Sample t-test
data:  x
t = 2.9245, df = 99, p-value = 0.002139
alternative hypothesis: true mean is greater than 4
95 percent confidence interval:
 4.257796      Inf
sample estimates:
mean of x
 4.596415
```

□

Properties of MMEs of variance

2.2.2. Statistical inference on variances.

- The sample variance S_n^2 is an unbiased estimator of the population variance σ^2, and

$$\mathrm{Var}(S_n^2) = \frac{1}{n}\left(\sigma_4 - \frac{n-3}{n-1}\sigma^4\right).$$

Notice that $\lim_{n\to\infty} \mathrm{Var}(S_n^2) = 0$. Hence, $\{S_n^2\}$ is consistent.

- If the population $X \sim N(\mu, \sigma^2)$ is normal, then $\overline{X}_n$ and S_n^2 are independent, and

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1). \tag{2.9}$$

(We repeat this once again.)

Properties of MMEs of variance

- T_n^2 is biased with

$$\mathbb{E}(T_n^2) = \frac{(n-1)\sigma^2}{n}, \qquad \text{bias}(T_n^2) = -\frac{\sigma^2}{n}, \qquad n \in \mathbb{N}.$$

- The mean squared error (MSE) of T_n^2 is

$$\text{mse}(T_n^2) = \frac{1}{n^3}\Big[(n-1)^2\sigma_4 - (n^2 - 5n + 3)\sigma^4\Big], \qquad n \in \mathbb{N}.$$

So, $\{T_n^2\}$ is consistent.

- When μ is known, $\mathbb{E}(W_n^2) = \sigma^2$, so W_n^2 is unbiased. Moreover,

$$\text{Var}(W_n^2) = \frac{1}{n}(\sigma_4 - \sigma^4), \qquad n \in \mathbb{N}.$$

So $\{W_n^2\}$ is consistent.

Properties of MMEs of variance

- It is noticeable that $\mathrm{Var}(W_n^2) < \mathrm{Var}(S_n^2)$ for all $n \geq 3$, and hence W_n^2 is preferred when μ is known.

- W_n^2 and S_n^2 are asymptotically equivalent since

$$\frac{\mathrm{Var}(W_n^2)}{\mathrm{Var}(S_n^2)} \to 1 \qquad \text{as } n \to \infty.$$

- There is no simple, general relationship between $\mathrm{mse}(T_n^2)$ and $\mathrm{mse}(S_n^2)$ or between $\mathrm{mse}(T_n^2)$ and $\mathrm{mse}(W_n^2)$, but the asymptotic relationship is simple. As $n \to \infty$,

$$\frac{\mathrm{mse}(T_n^2)}{\mathrm{mse}(W_n^2)} \to 1, \qquad \text{and} \qquad \frac{\mathrm{mse}(T_n^2)}{\mathrm{mse}(S_n^2)} \to 1.$$

Properties of MMEs of variance

- Comparisons become interesting if the population is normal, $X \sim N(\mu, \sigma^2)$.
- Mean squared errors of S_n^2 and T_n^2.
  - $\mathrm{mse}(T_n^2) = \frac{2n-1}{n^2}\sigma^4$.
  - $\mathrm{mse}(S_n^2) = \frac{2}{n-1}\sigma^4$.
  - $\mathrm{mse}(T_n^2) < \mathrm{mse}(S_n^2)$ for $n \geq 2$.
- Mean squared errors of W_n^2 and T_n^2.
  - $\mathrm{mse}(W_n^2) = \frac{2}{n}\sigma^4$.
  - $\mathrm{mse}(T_n^2) < \mathrm{mse}(W_n^2)$ for $n \geq 2$.

Properties of MMEs of variance

- When $X \sim N(\mu, \sigma^2)$ is normal with known μ,

$$\frac{nW_n^2}{\sigma^2} \sim \chi^2(n), \qquad n \geq 1. \tag{2.10}$$

- Asymptotically (for large n),

$$\frac{S_n^2 - \sigma^2}{\sqrt{2}\sigma^2/\sqrt{n-1}} \approx \frac{T_n^2 - \sigma^2}{\sqrt{2}\sigma^2/\sqrt{n}} \approx \frac{W_n^2 - \sigma^2}{\sqrt{2}\sigma^2/\sqrt{n}} \dot{\sim} N(0,1). \tag{2.11}$$

- Confidence intervals are not difficult to be obtained based on distributional properties in Eq.s (2.9) to (2.11).

Properties of sample standard deviation

- Assume $X \sim N(\mu, \sigma^2)$ is normal.
- Let

$$a_n = \frac{\sqrt{2}\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(\frac{n}{2})}, \qquad n \geq 1.$$

  Then, $0 < a_n < 1$ and $a_n \uparrow 1$ as $n \to \infty$.

- Means, biases, variances and mean squared errors of S and W are summarized in the following table.

	Mean	Bias	Variance	MSE
S	$a_{n-1}\sigma$	$(a_{n-1} - 1)\sigma$	$(1 - a_{n-1}^2)\sigma^2$	$2(1 - a_{n-1})\sigma^2$
W	$a_n\sigma$	$(a_n - 1)\sigma$	$(1 - a_n^2)\sigma^2$	$2(1 - a_n)\sigma^2$

Testing one variance

- Assume μ is unknown. Consider testing $H_0: \ \sigma^2 \leq \sigma_0^2$ against $H_a: \ \sigma^2 > \sigma_0^2$ for some constant σ_0^2.

(a) If $X \sim N(\mu, \sigma^2)$ is normal, by (2.9), the test statistic is

$$\chi^2 = \frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi^2(n-1) \qquad \text{if } \sigma^2 = \sigma_0^2.$$

H_0 is rejected at the α-level of significance if $\chi^2 > \chi^2_\alpha(n-1)$, the α-th upper quantile of the $\chi^2(n-1)$ distribution.

(b) For general F and large n, define

$$Z = \frac{S_n^2 - \sigma_0^2}{\sqrt{2}\sigma_0^2/\sqrt{n-1}} \dot{\sim} N(0,1) \qquad \text{if } \sigma^2 = \sigma_0^2.$$

Reject H_0 if $Z > Z_\alpha$. (Seldom used.)

Comparing two population variances

- Let $\{X_i : 1 \leq x \leq n_1\}$ and $\{Y_j : 1 \leq j \leq n_2\}$ be random samples from populations $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$, respectively. Assume both means are unknown.

- Testing $H_0 : \sigma_x \leq \sigma_y$ against $H_a : \sigma_x > \sigma_y$.

- By (2.9), the test statistic:

$$F = \frac{S_x^2}{S_y^2} = \frac{\frac{(n_1-1)S_x^2}{\sigma_x^2}/(n_1 - 1)}{\frac{(n_2-1)S_y^2}{\sigma_y^2}/(n_2 - 1)} \sim F(n_1 - 1,\, n_2 - 1) \qquad \text{if } \sigma_x^2 = \sigma_y^2.$$

- Reject H_0 in favor of H_a if $F > F_\alpha(n_1 - 1,\, n_2 - 1)$, the α-th upper quantile of the F-distributions with degrees of freedom $(n_1 - 1)$ and $(n_2 - 1)$.

Example 2.7. Inferences on variances

- **Example 2.7.** Consider two samples: $\{X_i : 1 \leq i \leq 100\}$ from $X \sim N(5, 4)$ as in Example 2.1, and $\{Y_j : 1 \leq j \leq 200\}$ from $Y \sim N(0, 6)$.
- A different seed is used in the *random number generator* for Y to ensure the independence between two samples.

```
# Generate samples from Y~N(0,6)
n2 <- 200
set.seed(2024)
y <- rnorm(n2, 0, sqrt(6))
```

```
> cor(x,y[1:100])
[1] -0.07510114
> cor(x,y[101:200])
[1] -0.05394609
```

Example 2.7. Inferences on variances

(i) Test $H_0 : \sigma_x^2 = 4$ vs $H_a : \sigma_x^2 \neq 4$.

- The R function EnvStats::varTest: *One-Sample Chi-Squared Test on Variance*, is used.

```
> varTest(x, alternative = "two.sided", sigma.squared = 4)
Results of Hypothesis Test
--------------------------
Null Hypothesis:                 variance = 4
Alternative Hypothesis:          True variance is not equal to 4
Test Name:                       Chi-Squared Test on Variance
Estimated Parameter(s):          variance = 4.159123
Data:                            x
Test Statistic:                  Chi-Squared = 102.9383
Test Statistic Parameter:        df = 99
P-value:                         0.7463159
95% Confidence Interval:         LCL = 3.206251
                                 UCL = 5.612692
```

Example 2.7. Inferences on variances

(ii) Test $H_0 : \sigma_x^2 \geq \sigma_y^2$ vs $H_a : \sigma_x^2 < \sigma_y^2$.

- The R function var.test: *F Test to Compare Two Variances*, is used.

```
> var.test(x, y, ratio = 1, alternative = "less")

        F test to compare two variances

data:  x and y
F = 0.65959, num df = 99, denom df = 199, p-value =
0.01051
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
 0.0000000 0.8861131
sample estimates:
ratio of variances
         0.6595878
```

□

2.2.3. Statistical inference on two means.

- Let $\{X_i : 1 \le i \le n_1\}$ and $\{Y_j : 1 \le j \le n_2\}$ be random samples from populations $X \sim F_x(\mu_x, \sigma_x^2)$ and $Y \sim F_y(\mu_y, \sigma_y^2)$, respectively. Assume both variances are unknown.
- Denote the sample means, sample variances, and sample covariance as $\overline{X}$, $\overline{Y}$, S_x^2, S_y^2, and $S_{x,y}$, respectively.
- Practically it is very common to do statistical inference on both population means μ_x and μ_y, e.g., in clinical trials.
- We consider testing $H_0: \ \mu_x \ge \mu_y$ against $H_a: \ \mu_x < \mu_y$ for illustration.
- The test depends on whether $\sigma_x^2 = \sigma_y^2$ or not, and on whether X and Y are independent.

The two-samples t-test

(a) X and Y are independently normal with equal variance $\sigma_x^2 = \sigma_y^2 = \sigma^2$.

- The difference in sample means $\overline{X} - \overline{Y}$ is normal with variance $\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2} = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$, in which σ^2 can be estimated by the pooled estimator

$$S_p^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2 \right).$$

- Define the test statistic as

$$t = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \qquad \text{under } H_0. \tag{2.12}$$

- Reject H_0 if $t < -t_\alpha(n_1 + n_2 - 2)$.

The Welch's t-test

(b) X and Y are independently normal with unequal $\sigma_x^2 \neq \sigma_y^2$.

- $\overline{X} - \overline{Y}$ is normal, and its variance is estimated by $\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}$.
- The t-test statistic is

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}} \sim t(v) \qquad \text{under } H_0, \tag{2.13}$$

  where the degrees of freedom is a real number instead of an integer,

$$v \approx \frac{\left(\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}\right)^2}{\frac{S_x^4}{n_1^2(n_1-1)} + \frac{S_y^4}{n_2^2(n_2-1)}}.$$

- Reject H_0 if $t < -t_\alpha(v)$.

Two-samples t-test or Welch's t-test?

- Welch's t-test and Student's t-test give identical results when the two samples have equal variances and sample sizes. The power of Welch's t-test comes close to that of Student's t-test, even when the population variances are equal and sample sizes are *balanced*.[1].

- Welch's t-test is more robust than two-samples Student's t-test and maintains type I error rates close to nominal for unequal variances and for unequal sample sizes under normality.

- It is *not recommended* to pre-test for equal variances and then choose between two-samples t-test or Welch's t-test.[2]

- When sample sizes are large, both tests are approximately valid for non-normal populations.

---

[1] Wikipedia

[2] Zimmerman, D.W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, **57 (Pt 1)**: 173-181.

The paired-samples t-test

(c) If $\{(X_i, Y_i) : 1 \leq i \leq n\}$'s are paired observations.

- This arises when we are investigating the impacts of different treatments (conditions) on the same (batch of) individuals.
- Assumption: $\{D_i \widehat{=} X_i - Y_i : 1 \leq i \leq n\}$ are i.i.d. normal $N(\mu_d, \sigma_d^2)$.
- Testing $H_0 : \mu_d = 0$ vs $H_a : \mu_d \neq 0$ assuming σ_d^2 is unknown.
- The problem reduces to testing hypothesis on one mean using a single sample $\{D_i\}$. The one-sample t-test is applied. Refer to subsection 2.2.1.
- We also call this the **paired-samples t-test**.

Example 2.8. Inferences on two means

- **Example 2.8.** Samples $\{X_i\}$ and $\{Y_j\}$ in Example 2.7 will be used for cases (a) and (b), the independent samples t-tests assuming equal variance or not.

- A pair of samples $\{X_i, Y_i) : 1 \leq i \leq 100\}$ are generated for case (c), the paired t-test. X and Y are not independent.

```
> set.seed(7011)
> X <- rt(n,5)
> Y <- runif(n,-0.8,1)
> cor(X,Y)
[1] -0.09167719
```

- The R function `t.test` can be applied for all 3 cases.

Example 2.8. Inferences on two means

(a) Two-samples t-test assuming equal variance.

```
> t.test(x, y, alternative = "two.sided", var.equal = T)

        Two Sample t-test

data:  x and y
t = 15.67, df = 298, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 3.968702 5.108677
sample estimates:
 mean of x  mean of y
4.59641471 0.05772509
```

Example 2.8. Inferences on two means

(b) Welch's t-test assuming unequal variance.

```
> t.test(x, y, alternative = "two.sided", var.equal = F)

        Welch Two Sample t-test

data:  x and y
t = 16.785, df = 237.96, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 4.005994 5.071385
sample estimates:
 mean of x  mean of y
4.59641471 0.05772509
```

Example 2.8. Inferences on two means

(c) Paired samples t-test.

```
> t.test(X, Y, alternative = "two.sided", paired = T)

        Paired t-test

data:  X and Y
t = -2.5966, df = 99, p-value = 0.01085
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -0.61160806 -0.08176225
sample estimates:
mean of the differences
             -0.3466852
```

□

2.2.4. The Binomial test.

- If the population distribution $F = \mathrm{Bin}(1, p)$, the Bernoulli (bi-point) distribution, the MME of the success or failure rate p is the sample proportion (also the sample mean)

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- Define $Y_n = \sum_{i=1}^{n} X_i$, then Y_n is Binomial $\mathrm{Bin}(n, p)$, and

$$\mathbb{E}(Y_n) = np, \qquad \mathrm{Var}(Y_n) = np(1-p).$$

- Since $\widehat{p} = \frac{Y_n}{n}$,

$$\mathbb{E}(\widehat{p}) = p, \qquad \mathrm{Var}(\widehat{p}) = p(1-p)/n.$$

Hence, $\widehat{p}$ is unbiased and consistent.

The exact test and CI

- The exact distribution of Y_n is usually used to do the *exact test* on p, especially when n is not too large. Such a test is referred to as the **Binomial test**.
- Suppose we want to test $H_0 : p = p_0$ against $H_a : p \neq p_0$ for some $0 < p_0 < 1$.
- Let

$$\begin{aligned} y_1 &= \max\{k : \mathbb{P}(Y_n \leq k | H_0) \leq \alpha/2\}, \\ y_2 &= \min\{k : \mathbb{P}(Y_n > k | H_0) \leq \alpha/2\}. \end{aligned}$$

  Then, H_0 is rejected if the observed $y_n \leq y_1$ or $\geq y_2$.
- By the duality between hypothesis test and confidence interval, the $(1-\alpha)$-th exact CI of p is the collection of all p_0's such that we fail to reject the null hypothesis $H_0 : p = p_0$.

The approximate test

- When the sample size n is large ($n \geq 40$ is suggested), by CLT,

$$Z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \dot{\sim} N(0,1).$$

This asymptotical distribution is usually used to do approximate inferences on rate p, especially hypothesis tests.

- To test $H_0 : p = p_0$ against $H_a : p \neq p_0$, define

$$Z_0 = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \dot{\sim} N(0,1) \qquad \text{under } H_0.$$

H_0 is rejected at the α-level of significance if $|Z_0| > Z_{\alpha/2}$.

The approximate CI

- By the approximate distribution of Z, for any $0 < \alpha < 1$,

$$|Z| < Z_{\alpha/2} \quad \Longleftrightarrow \quad (\widehat{p} - p)^2 < c_\alpha p(1-p),$$

where $c_\alpha = \frac{Z^2_{\alpha/2}}{n}$.

- Solving this inequality for p gives the analytical formula of the approximate CI $\widehat{p}_1$, $\widehat{p}_2)$, where $\widehat{p}_1 < \widehat{p}_2$ are two roots of the quadratic equation $(\widehat{p} - p)^2 = c_\alpha p(1-p)$, i.e.,

$$\widehat{p}_{1,2} = \frac{2\widehat{p} + c_\alpha \mp \sqrt{c_\alpha^2 + 4c_\alpha \widehat{p}(1-\widehat{p})}}{2(1+c_\alpha)}.$$

- Numerically, the approximate CI of p is usually constructed by the duality between hypothesis test and confidence interval.

Example 2.9. Binomial test

- **Example 2.9.** Consider a random sample of size $n = 50$ from a Bin(1, 0.4) population, generated by the following DGP.

```r
n <- 50
set.seed(7011)
X <- rbinom(n, 1, 0.4)
Y <- sum(X)              # Number of successes
X1 <- c(Y, n-Y)          # Numbers of successes/failures
p0 <- 0.5
```

  - The number of successes is defined as `Y`, and `X1` is a vector (of length 2) of both numbers of successes and failures.
  - R function `binom.text()` provides two equivalent ways to do the test using `Y` and `X1`, respectively.
  - We are testing $H_0 : p = p_0 = 0.5$ against $H_0 : p \neq p_0$.

Example 2.9. Binomial test

- R commands (two ways) are

```
binom.test(x = Y, n = n, p = p0)
binom.test(x = X1, p = p0)
```

- Both ways give exactly the same numerical results.

```
data: Y and n       or       X1
number of successes = 16, number of trials = 50,
p-value = 0.01535
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.1952042 0.4669938
sample estimates:
probability of success 0.32
```

Example 2.9. Binomial test

- The approximate test and corresponding analytical CI are calculated as follows.

```
> # Approximate Z-test
> Z <- (mean(X)-p0)/sqrt(p0*(1-p0)/n); Z
[1] -2.545584
> p.val <- pnorm(Z)*2; p.val
[1] 0.0109095
> # Approximate and analytic CI
> c <- qnorm(0.025)^2/n
> p.hat <- mean(X)
> p1 <- (2*p.hat+c-sqrt(c^2+4*c*p.hat*(1-p.hat)))/(2*(1+c))
> p2 <- (2*p.hat+c+sqrt(c^2+4*c*p.hat*(1-p.hat)))/(2*(1+c))
> c(p1, p2)
[1] 0.2075822 0.4581030
```

- The approximations are not too bad since $n = 50 > 40$. $\square$

Properties of sample skewness and kurtosis

2.2.5. The Jarque-Bera test.

- There is a very popular *normality test* in econometrics called the **Jarque-Bera test**. It mainly uses the following properties of sample skewness and sample kurtosis from normal populations.

- Assume $X \sim N(\mu, \sigma^2)$ is normal. For large n,

$$\widehat{\eta} \,\dot{\sim}\, N\left(0, \frac{6}{n}\right). \tag{2.14}$$

- Assume $X \sim N(\mu, \sigma^2)$ is normal. For large n,

$$\widehat{\kappa} \,\dot{\sim}\, N\left(3, \frac{24}{n}\right). \tag{2.15}$$

The Jarque-Bera test

- Normality test: "$H_0 : X$ is normal" vs "H_a : not H_0".
- The Jarque-Bera test statistics is defined as

$$JB = n\left(\frac{\widehat{\eta}^2}{6} + \frac{(\widehat{\kappa}-3)^2}{24}\right). \tag{2.16}$$

- Under H_0, $JB \sim \chi^2(2)$.
- Reject H_0 if $JB > \chi^2_\alpha(2)$.
- Inferences *purely on* higher order moments such as skewness and kurtosis are very seldom seen.

Example 2.10. Jarque-Bera test

- **Example 2.10.** Jarque-Bera test.

```
> require(tseries)
> x <- rnorm(100, 0, 1)
> jarque.bera.test(x)

	Jarque Bera Test

data:  x
X-squared = 0.32747, df = 2, p-value = 0.849

> y <- rcauchy(100, 0, 1)
> jarque.bera.test(y)

	Jarque Bera Test

data:  y
X-squared = 28840, df = 2, p-value < 2.2e-16
```

□