Session 2
ℚuantitative Analysis of Financial Markets
Statistics

Christopher Ting

http://www.mysmu.edu/faculty/christophert/

Christopher Ting

: christophert@smu.edu.sg

: 6828 0364

: LKCSB 5036

October 12, 2018

Broad Lesson Plan

Learning Objectives

- Discuss time-series and cross-sectional data and their differences.
- Understand the difference between price/index level and return.
- Recall the basics of probability concepts needed in statistical inference:
  - mean, variance, covariance, correlation
  - independence
  - normal, chi-square, Student's t, and F distributions
- Recall the basics of statistical concepts:
  - sample mean, sample variance
  - unbiased estimators
  - law of large number, central limit theorem
- Discuss and develop the framework of hypothesis tests.

Quotable

Economists like to deal with things that can be counted, quantified and computerised. There is nothing wrong in this, but it is a short step from this position to the serious error of believing that quantifiable variables are the only things that really matter. And they seem surprised and disappointed when their prescriptions for economic growth did not work in country after country.

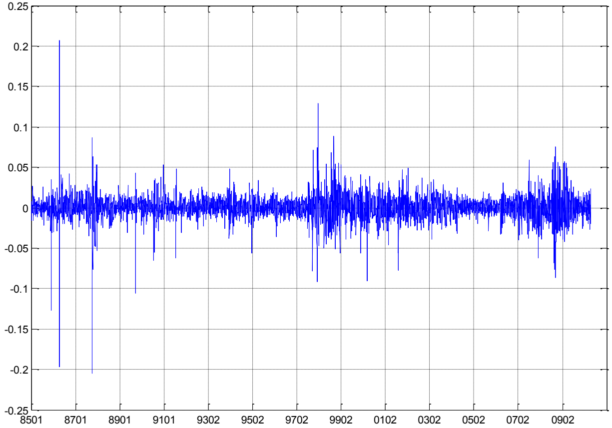
Dr Goh Keng Swee, November 1972

Market data are the outcomes of lots of people's decisions that involve theirs or their clients' money. These quantifiable variables must be taken seriously. But the econometric models and financial investment theories by which the data are analyzed are not the things that really matter to practitioners. Don't be surprised and disappointed that the models don't work time after time.

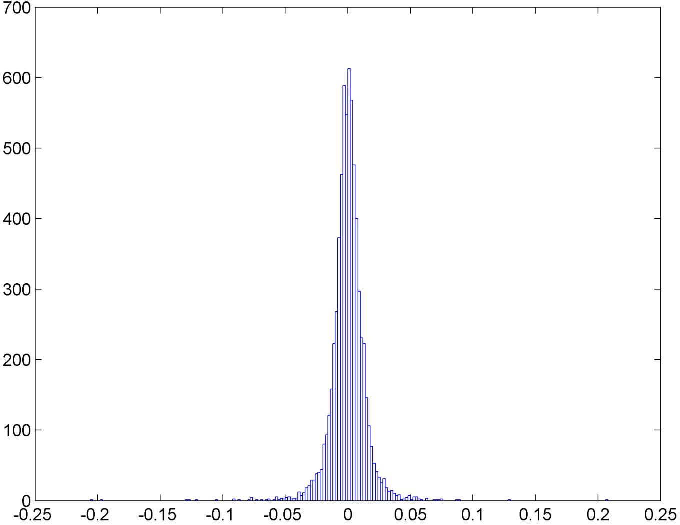
Straits Times Index and Major Events

SMU Classification: Restricted

Daily Return on Straits Times Index



Histogram of Daily Return on Straits Times Index



What is statistics?

- Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.
- Statistics is the art of making numerical conjectures about puzzling questions.
- Statistics is a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty.
- Statistics helps provide a systematic approach for obtaining reasoned answers together with some assessment of their reliability in situations where complete information is unobtainable or not available in a timely manner.
- Statistics is a body of methods for making wise decisions in the face of uncertainty.

What is statistics? (cont'd)

- Statistics is the art and science of gathering, analyzing, and making inferences from **data**.

- Statistics is the art of learning from **data**. It is concerned with the collection of **data**, its subsequent description, and its analysis, which often leads to drawing conclusions.

- Statistics is a set of concepts, rules, and methods for
  1. collecting **data**
  2. analyzing **data**
  3. drawing conclusions from **data**

What is statistics?

Statistics is
the study of algorithms
for **data analysis**.

Rudolf Beran
Statistical Science, Vol. 18, No. 2, Silver Anniversary of the Bootstrap (May, 2003), pp. 175-184.

Investment and Data

- Investment is the process of laying out funds in financial instruments and assets with the expectation of a profit.
- Before making an investment, financial data analysis is a crucial step.
- To scout for profitable opportunities (risk-adjusted), investment companies such as real-estate investment trusts, exchange-traded funds, mutual funds, and hedge funds perform in-depth analysis on all tradable financial securities and assets.

We don't start with models. We start with data. We don't have any preconceived notions. We look for things that can be replicated thousands of times.

James Simons

Time Series Data

◊ **Historical** observations of a financial variable

★ Prices

★ Trading volume

★ Financial indices

★ Economic indices

★ Insiders' trading activities

★ Investment companies' trading activities

★ Analysts' forecasts

★ Corporate earnings

★ Order flows

★ Money flows

Cross-Sectional Data

♦ Portfolios constructed based on securities' or assets' characteristics **at a given time**

⊛ Firm characteristics (e.g. market capitalization, growth versus value)

⊛ Risk profiles

⊛ Price characteristics (e.g. 52-week high versus 52-week low)

⊛ Physical characteristics (e.g. agricultural, metals)

⊛ Industry

⊛ Emerging versus developed markets

⊛ Country of domicile or geographic location

⊛ Funds' trading strategies (e.g. convertible arbitrage, event driven)

Framework

Definition 1

Statistical population is the set of all possible elements that are of interest for a statistical analysis.

Example 1

The **time series** of split-adjusted daily stock prices of Dell Inc. since IPO on June 22, 1988 till taken private on October 29, 2013.

Example 2

The **cross section** of daily returns of all component stocks of Nikkei 225 index on October 12, 2018.

Definition 2

A **statistical model** or a **data generating process** is a pair $(S, \mathcal{P})$, where S is the σ-algebra of a statistical population, i.e. the **sample space**, and $\mathcal{P}$ is a set of probability distributions on S.

Population versus Sample

- Random variable: X
- Mean of a statistical population: $\mathbb{E}(X) =: \mu$
- Variance of a statistical population: $\mathbb{V}(X) := \mathbb{E}\left((X-\mu)^2\right) =: \sigma^2$
- Sample of **size n taken randomly** from the population: $\{x_i\}_{i=1}^n$
  - What is the name of each x_i? Ans: ___________
  - Is each x_i known or unknown? Ans: ___________
- An example of sample average estimator: $\widehat{\mu} := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$
- An example of sample variance estimator:

$$\widehat{\sigma^2} := \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widehat{\mu}\right)^2$$

Example 3: Dow Jones Utility Average

Source: **Finance Yahoo!** (September 28, 2018)

Symbol	Company Name	Last Price	Change	% Change	Volume
NI	NiSource Inc.	24.92	0.14	0.56%	5,195,036
SO	The Southern Company	43.60	0.36	0.83%	7,748,750
CNP	CenterPoint Energy, Inc.	27.65	0.23	0.84%	16,733,274
AWK	American Water Works Company, Inc.	87.97	1.01	1.16%	698,492
NEE	NextEra Energy, Inc.	167.60	2.01	1.21%	2,356,243
ED	Consolidated Edison, Inc.	76.19	0.99	1.32%	3,163,334
DUK	Duke Energy Corporation	80.02	1.08	1.37%	4,540,880
EIX	Edison International	67.68	1.02	1.53%	1,840,841
AEP	American Electric Power Company, Inc.	70.88	1.12	1.61%	2,740,434
PCG	PG&E Corporation	46.01	0.74	1.63%	4,941,609
D	Dominion Energy, Inc.	70.28	1.14	1.65%	3,095,756
FE	FirstEnergy Corp.	37.17	0.64	1.75%	3,763,472
EXC	Exelon Corporation	43.66	0.85	1.99%	7,179,395
AES	The AES Corporation	14.00	0.30	2.19%	5,578,198
PEG	Public Service Enterprise Group Incorporated	52.79	1.45	2.82%	4,051,713

Expectation and Variance of Return

- Simple return over one period (eg. 5 minutes, one day, one week, one month)

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

- If P_t is observed at t, then the resulting R_t is said to be *ex post* return.
- If only P_{t-1} is known but P_t is not observed yet, then R_t is said to be *ex ante* return.
- The ex ante return R_t is a random variable.
- Expected value of R_t:

$$\mu := \mathbb{E}\left(R_t\right), \qquad \forall\, t$$

- Variance of R_t:

$$\sigma^2 := \mathbb{V}\left(R_t\right) = \mathbb{E}\Big(\left(R_t - \mu\right)^2\Big) = \mathbb{E}\left(R_t^2\right) - \mu^2, \qquad \forall\, t$$

Covariance and Correlation

- Consider the return on M1 $R_{X,t}$, and on Starhub $R_{Y,t}$. The respective mean and variance are $\mu_X, {\sigma_X}^2$ for M1 and $\mu_Y, {\sigma_Y}^2$ for Starhub.

- The covariance between $R_{X,t}$ and $R_{Y,t}$ is

$$\begin{aligned}\sigma_{XY} := \mathbb{C}\big(R_{X,t}\,,\,R_{Y,t}\big) &= \mathbb{E}\Big(\big(R_{X,t} - \mu_X\big)\big(R_{Y,t} - \mu_Y\big)\Big) \\ &= \mathbb{E}\big(R_{X,t}\,R_{Y,t}\big) - \mu_X\mu_Y\,.\end{aligned}$$

- The correlation between $R_{X,t}$ and $R_{Y,t}$ is

$$\rho_{XY} := \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Correlation: Normalized Covariance

- Normalization of covariance σ_{XX} gives rise to correlation, which is written as $\rho_{XY} := \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$.
- Correlation has the nice property that $-1 \leqq \rho \leqq 1$. If two variables have a correlation of +1 (-1), then we say they are **perfectly correlated** (anti-correlated).
- If one random variable causes the other random variable, or that both variables share a common underlying driver, then they are highly correlated.
- But high correlation does not necessarily imply causation of one variable on the other.
- If two variables are uncorrelated, it does not necessarily follow that they are unrelated.
- **So what does correlation tell you?**

Properties of Expectation and Variance Operators

Mean, Variance, and Covariance

Let a, b and c be constant. Let X and Y be two random variables, with means μ_X and μ_Y, respectively. Also, the corresponding variances are σ_X^2 and σ_Y^2. Then,

$$\mathbb{E}\left(aX + bY + c\right) = a\,\mathbb{E}\left(X\right) + b\,\mathbb{E}\left(Y\right)c \tag{1}$$

$$\mathbb{V}\left(X\right) = \mathbb{E}\left(X^2\right) - \mu_X^2 \tag{2}$$

$$\mathbb{V}\left(aX + b\right) = a^2\,\mathbb{V}\left(X\right) \tag{3}$$

$$\mathbb{V}\left(aX + bY + c\right) = a^2\,\mathbb{V}\left(X\right) + b^2\,\mathbb{V}\left(Y\right) + 2ab\,\mathbb{C}\left(X, Y\right) \tag{4}$$

More on Covariance

Definition 3: Covariance

Covariance is a generalized version of variance. It is defined as

$$\mathbb{C}(X, Y) \equiv \sigma_{XY} := \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right).$$

- Variance is a special case: $\mathbb{C}(X, X) = \sigma_{XX} = \mathbb{V}(X)$.
- Whereas variance is strictly positive, covariance can be positive, negative, and zero.
- If X and Y are independent, then it must be that $\mathbb{C}(X, Y) = 0$.
- If $\mathbb{C}(X, Y) = 0$, it is not necessarily true that X and Y are independent.

Class Exercises

1. $\sigma_{XY} = \mathbb{E}\left(XY\right) - \mu_X \mu_Y.$
2. $\mathbb{C}\left(X, Y\right) = \mathbb{C}\left(Y, X\right).$
3. $\mathbb{C}\left(X + Y, Z\right) = \mathbb{C}\left(X, Z\right) + \mathbb{C}\left(Y, Z\right).$

Linear Combination of Two Random Variables

Proposition 1

Suppose X and Y form a pair random variables with means $\mu_X := \mathbb{E}(X)$ and $\mu_Y := \mathbb{E}(Y)$, respectively. Also, suppose a and b are two constants. Then,

$$\mathbb{V}\left(aX + bY\right) = a^2\,\mathbb{V}\left(X\right) + b^2\,\mathbb{V}\left(Y\right) + 2ab\,\mathbb{C}\left(X, Y\right). \tag{5}$$

Proof

1. $\mathbb{V}\left(aX + bY\right) = \mathbb{E}\left((aX + bY)^2\right) - (a\mu_X + b\mu_Y)^2$.
2. Expanding the two quadratic term and collecting the expanded terms accordingly, we obtain

$$a^2\,\mathbb{E}\left(X^2\right) - a^2\mu_X^2 + b^2\,\mathbb{E}\left(Y^2\right) - b^2\mu_Y^2 + 2ab\,\mathbb{E}\left(XY\right) - 2ab\mu_X\mu_Y.$$

$$\Longrightarrow\ a^2\left(\mathbb{E}\left(X^2\right) - \mu_X^2\right) + b^2\left(\mathbb{E}\left(Y^2\right) - \mu_Y^2\right) + 2ab\left(\mathbb{E}\left(XY\right) - \mu_X\mu_Y\right).$$

Note: If X and Y are independent, then

$$\mathbb{V}\left(aX + bY\right) = a^2\,\mathbb{V}\left(X\right) + b^2\,\mathbb{V}\left(Y\right).$$

An Example and a Question

- The covariance between the return on gold and the return on silver is 0.04. The volatility of return on gold is 60%, and the volatility of return on silver is 30%. What is the correlation between gold return and silver return?

  Answer: ______________________

- How should the notion of co-volatility be defined?

Independently and Identically Distributed

- Suppose the random variables X_t for $t = 1, 2, \ldots, n$ are i.i.d.
- Law of Large Number (LLN)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} X_t \xrightarrow{\mathbb{P}} \mu = \mathbb{E}(X_t)$$

- Central Limit Theorem (CLT)
  For sufficiently large sample size n, given μ and σ,

$$Y := \frac{\frac{1}{n}\sum_{t=1}^{n} X_t - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\frac{1}{n}\sum_{t=1}^{n} X_t = \mu + \frac{\sigma}{\sqrt{n}} Y \overset{d}{\sim} N\left(\mu\,, \frac{\sigma^2}{n}\right) \tag{6}$$

Model 0

- Applying the theorem in Slide 20, we have

$$\mathbb{E}\left(Y\right) = 0, \qquad \mathbb{V}\left(Y\right) = 1$$

- From CLT's (6), we write $\mu = \frac{1}{n}\sum_{t=1}^{n} X_t - \frac{\sigma}{\sqrt{n}} Y$
- In reality, μ and σ are unknown. We replace μ by X, $\frac{\sigma}{\sqrt{n}}$ by ς, and let $Y := -Z$.
- Given $\{X_i\}_{i=1}^n$, we compute the sample mean (aka average). $\overline{X} := \frac{1}{n}\sum_{t=1}^{n} X_t$, we now introduce **Model 0**:

$$X = \overline{X} + \varsigma Z. \tag{7}$$

- Given the dateset $\{X_t\}_{t=1}^n$, a forecast of X is the sample mean!

$$\mathbb{E}\left(X \middle| \{X_t\}_{t=1}^n\right) = \overline{X}. \tag{8}$$

Normal Distribution

- A very common assumption of finance is that the returns are normally distributed.
$$r \overset{d}{\sim} N(\mu, \sigma^2).$$
- The probability density function $f(r)$ is
$$f(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{r-\mu}{\sigma}\right)^2 \right).$$
- The mean and variance are, respectively,
$$\begin{aligned} \mathbb{E}(r) &= \int_{-\infty}^{\infty} r\, f(r)\, dr = \mu; \\ \mathbb{V}(r) &= \int_{-\infty}^{\infty} (r-\mu)^2\, f(r)\, dr = \sigma^2. \end{aligned}$$

Standard Normal Distribution

- For convenience, define

$$z := \frac{r_t - \mu}{\sigma}, \qquad f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

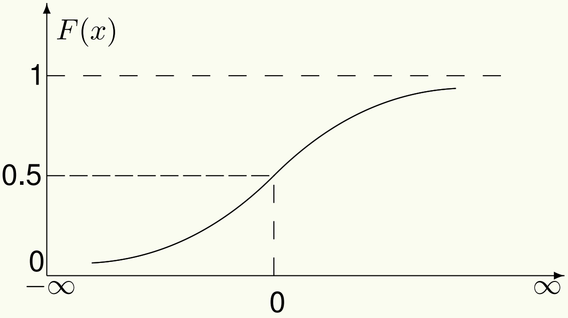
- The probability density function $f(z)$ is the well known bell-shaped curve with mean 0 and variance 1

Cumulative Distribution Function $F(x)$

- What is the probability that $z < -1.645$?

$$F(-1.645) := \mathbb{P}(z < -1.645) = \int_{-\infty}^{-1.645} f(z)\,dz = 0.05$$

- Thus there is 5% probability that $r_t < \mu - 1.645\sigma$



Most Popular Statistic: Sample Mean

- Given a set of past observations $\{R_1, R_2, \ldots, R_n\}$, how should one estimate μ and σ^2?
- Treat the observed value as a particular outcome or realization of the random variable R_t at each t:

$$\overline{R} = \frac{1}{n}\sum_{t=1}^{n} R_t$$

- The sample mean $\overline{R}$ itself is a random variable with mean and variance, assuming identical distribution,

$$\begin{aligned}\mathbb{E}(\overline{R}) &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}(R_t) = \mu \\ \mathbb{V}(\overline{R}) &= \frac{1}{n^2}\sum_{t=1}^{n}\mathbb{V}(R_t) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}\end{aligned}$$

- For the sample variance, independence is also assumed.

Estimation of Variance and t Statistic

- Unbiased sample variance is

$$s^2 = \frac{1}{n-1}\sum_{t=1}^{n}\left(R_t - \overline{R}\right)^2 \tag{9}$$

- The ratio of the sample variance with population variance

$$V := (n-1)\frac{s^2}{\sigma^2} \overset{d}{\sim} \chi^2_{n-1}$$

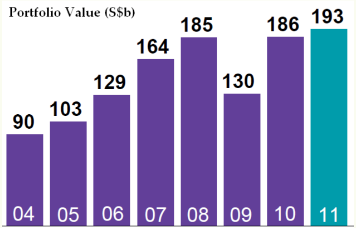
- Application

$$\overline{R} \overset{d}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{or} \qquad \frac{\overline{R}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\sqrt{n}\left(\overline{R}-\mu\right)}{\sigma} \overset{d}{\sim} N(0,1)$$

$$\Longrightarrow \qquad \frac{\frac{\sqrt{n}\left(\overline{R}-\mu\right)}{\sigma}}{\sqrt{\frac{V}{n-1}}} = \frac{\frac{\sqrt{n}\left(\overline{R}-\mu\right)}{\sigma}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\sqrt{n}\left(\overline{R}-\mu\right)}{s} \overset{d}{\sim} t_{n-1} \tag{10}$$

Case Study of Temasek's Performance

- The past observations of Temasek's portfolio value are



Source: Temasek Review 2011, Page 8

- Compute the annual portfolio returns.
- Compute the unbiased sample mean of annual portfolio return.
- Compute the unbiased sample variance of annual portfolio return.
- If $\mu = 7\%$, compute the t statistic.

Unbiasedness

- A statistic $\psi(\boldsymbol{X})$ is an unbiased estimator of θ if

$$\mathbb{E}\big(\psi(\boldsymbol{X})\big) = \theta\,.$$

- If $\mathbb{E}\big(\psi(\boldsymbol{X})\big) \neq \theta$, then the estimator is said to be biased,
- The bias is simply the difference:

$$\mathbb{E}\big(\psi(\boldsymbol{X})\big) - \theta\,.$$

- For convenience, we write $\widehat{\theta} := \psi(\boldsymbol{X})$

Bias of an Estimator

- **Definition 4**
  In statistics, the **bias** (or bias function) of an estimator $\widehat{\theta}$ is the difference between this estimator's expected value $\mathbb{E}\left(\widehat{\theta}\right)$ and the true value θ of the parameter being estimated.

$$\text{Bias} := \mathbb{E}\left(\widehat{\theta}\right) - \theta$$

Proposition 2

Sample mean is an unbiased estimator of population mean, i.e., $\mathbb{E}\left(\widehat{X}\right) = \mu.$

Proof:

$$\mathbb{E}\left(\widehat{X}\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(n\mu) = \mu$$

- **What are the assumptions required to prove the Proposition?**

Consistency

- In practice, unbiased estimators workable on small samples are rare.
- A sequence of estimators $\theta_n(\boldsymbol{X})$ of θ from sample $\boldsymbol{X}$ of size n is said to be a consistent estimator if

$$\theta_n \xrightarrow{\mathbb{P}} \theta \quad \text{as} \quad n \longrightarrow \infty \qquad \text{or} \qquad \operatorname{plim} \theta_n = \theta \,.$$

- That is, θ_n converges in probability to θ; for any arbitrary $\epsilon > 0$,

$$\lim_{n \longrightarrow \infty} \mathbb{P}\Big(\big| \theta_n - \theta \big| < \epsilon \Big) = 1 \,.$$

- Is a consistent estimator necessarily unbiased?

More on the Consistency of an Estimator

- **Definition 5**
  An estimator is said to be **consistent** if its difference with the true value θ. i.e., error, becomes smaller and insignificant, as the sample size grows larger and larger.

- The convergence is in probability, i.e., the absolute difference between the estimate and the true value mean being greater then some arbitrarily small margin ϵ has zero probability, as the sample size increases to ∞.

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{\theta} - \theta| > \epsilon\right) = 0.$$

- Implication: the more data you collect, a consistent estimator will be close to the real population parameter youâĂŹre trying to measure.

Class Participation

1. Consider another estimator of mean $\check{X} = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} X_i + \dfrac{X_n}{n}$
   (a) Is this estimator unbiased?
   (b) Is this estimator consistent?

2. Consider the estimator X_7 of μ.
   (a) Is this estimator unbiased?
   (b) Is this estimator consistent?

Is the Sample Variance Estimator Unbiased?

First we write

$$\frac{1}{n-1}\sum_{i=1}^{n}\big(X_i-\widehat{\mu}\big)^2=\frac{1}{n-1}\sum_{i=1}^{n}\Big(\big(X_i-\mu\big)-\big(\widehat{\mu}-\mu\big)\Big)^2$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}\Big(\big(X_i-\mu\big)^2-2\big(\widehat{\mu}-\mu\big)\big(X_i-\mu\big)+\big(\widehat{\mu}-\mu\big)^2\Big)$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}\big(X_i-\mu\big)^2-\frac{2}{n-1}\big(\widehat{\mu}-\mu\big)\sum_{i=1}^{n}\big(X_i-\mu\big)+\frac{1}{n-1}\big(\widehat{\mu}-\mu\big)^2\cdot n$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}\big(X_i-\mu\big)^2-\frac{2n}{n-1}\big(\widehat{\mu}-\mu\big)^2+\frac{n}{n-1}\big(\widehat{\mu}-\mu\big)^2$$

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

Is the Sample Variance Estimator Unbiased? (cont'd)

Taking expectation on the sample variance estimator,

$$\mathbb{E}\left(\widehat{\sigma}^2\right) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\mu\right)^2 - \frac{n}{n-1}\left(\widehat{\mu}-\mu\right)^2\right)$$

The first term is

$$\frac{1}{n-1}\mathbb{E}\left(\frac{n}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right) = \frac{n}{n-1}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right) = \frac{n}{n-1}\sigma^2$$

For the second term, we note that $\mathbb{E}\left(\left(\widehat{\mu}-\mu\right)^2\right) = \frac{\sigma^2}{n}$. It follows that

$$\mathbb{E}\left(\widehat{\sigma}^2\right) = \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\cdot\frac{1}{n}\sigma^2 = \frac{n-1}{n-1}\sigma^2 = \sigma^2.$$

Efficiency, BLUE

- ★ Given the dataset $\boldsymbol{X}$, an unbiased estimator $\psi_\star(\boldsymbol{X})$ of θ is said to be efficient if for any other unbiased estimator $\psi(\boldsymbol{X})$

$$\mathbb{V}\big(\psi_\star(\boldsymbol{X})\big) \leq \mathbb{V}\big(\psi(\boldsymbol{X})\big).$$

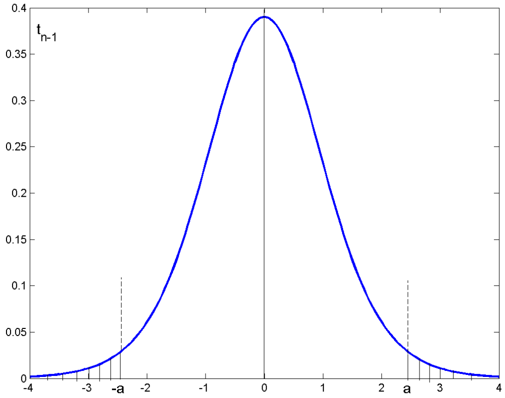
- ★ If ψ is an unbiased linear estimator and it has the minimum variance in the class of unbiased linear estimators, then ψ is said to be BLUE, best linear unbiased estimator.

Hypothesis and Test Statistic

- Suppose the population mean μ is, say, 7%. The null hypothesis is $H_0 = 7\%$. The alternative hypothesis is $H_A \neq 7\%$.
- A **statistical test** of the hypothesis is a decision rule that either rejects or does not reject the null H_0.
- Defined as $\{t_{n-1} < -a \text{ or } t_{n-1} > +a\}$, $a > 0$, the **critical region** is the set of values that leads to the rejection of H_0.
- The statistical rule on $H_0 : \mu = 7\%$, $H_A : \mu \neq 7\%$, is that if the t-distributed test statistic t_{n-1} (10) falls within the critical region, then H_0 is rejected. Otherwise H_0 cannot be rejected.
- Note that $\dfrac{s}{\sqrt{n}}$ is known as the ____________________ of the sample mean.

Illustration of Critical Regions

- The critical regions correspond to the shaded areas.
- The sum of the shaded area is the probability of rejecting H_0 when it is true. This probability is known as the **significance level**.



Two-Tail versus One-Tail

- If the hypotheses are, say, $H_0 : \mu = 7\%$ and $H_A : \mu \neq 7\%$, then the decision rule is based on **two-tail test**.

- The **critical region** comprises of the left and right tails of the t_{n-1} pdf.

- When the theory rules out, say, $\mu > 7\%$, the hypotheses become $H_0 : \mu = 7\%$ and $H_A : \mu < 7\%$, then the decision rule is based on **one-tail test**.

- The critical region is only the left side, for when $\mu < 7\%$, then $\left| t_{n-1} \right|$ will become larger. Thus at the one-tail 5% significance level, the critical region is $\{t_{n-1,95\%} > 1.671\}$ for $n = 61$, where $\{t_{n-1,95\%} > 1.671\}$ is the 95-th percentile of the t distribution.

P Value

- P value is the observed significance level, which is the probability of getting a value of the t statistic that is extreme or more extreme than the observed value of t statistic.

- Example
  - $H_0 = 7\%$ against $H_1 \neq 7\%$
  - $n = 25$
  - $\overline{R} = 10\%$
  - $s = 5\%$
  - $t = \sqrt{25}(10 - 7)/5 = 3$
  - The P value is $P = \mathbb{P}\big(|t_{24}| > 3\big)$.

Type I and Type II Errors

- If H_0 is true but is rejected, Type I error is committed.

- If H_0 is false but is “accepted,” Type II error is committed.

Result of the Test	Reality: H_0 is true	Reality: H_0 is false
Reject H_0	Type I error	Correct inference
Do not reject H_0	Correct inference	Type II error

Inference

- The probability of committing a Type I error when H_0 is true is called the ____________________.

- The probability of the population t-statistic exceeding the t-statistic obtained from the test sample is known as the p value.

- If the p value $<$ test significance level, reject H_0; otherwise H_0 cannot be rejected.

- In practice, the probability of Type I error is fixed and the significance level set at e.g. 10%, 5%, or 1%.

- Given that the null hypothesis is not true, the **power of a test** is the probability of not committing Type II error.

Confidence Interval

- Suppose data are randomly sampled from $X \overset{d}{\sim} N(\mu\,,\,\sigma^2)$ such that for an $a > 0$

$$\mathbb{P}\big(-a \leq t_{n-1} \leq +a\big) = 95\%.$$

- Given the formula for the t statistic, (10),

$$\mathbb{P}\left(-a \leq \frac{\sqrt{n}\big(\overline{X} - \mu\big)}{s} \leq a\right) = 0.95.$$

- Thus the probability of μ falling within the confidence interval is 95%:

$$\mathbb{P}\left(\overline{X} - a\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + a\frac{s}{\sqrt{n}}\right) = 95\%. \tag{11}$$

Connection with Model 0

- From (11), and given the critical value a for the t distribution, the lower bound is given by

$$\text{LB} := \overline{X} - \frac{s}{\sqrt{n}}a$$

and the upper bound by

$$\text{UB} := \overline{X} + \frac{s}{\sqrt{n}}a$$

- In summarizing the data, it is better to give a range rather than a point estimate. Therefore, given a sample $\{X_t\}_{t=1}^{n}$, the true value is between LB and UB with 95% confidence.
- Continuation from the case study of Temasek's performance (Slide 32) how would you summarize Temasek's 1-year return?

Application of Model 0: Forecasting

- Let $\widehat{X}_{n+1}$ denote a forecast of yet-to-be-observed X_{n+1} based on the sample $\{X_t\}_{t=1}^{n}$ observed up to period n.

- If X_{n+1} is assumed to be independently drawn from the same population as the sample, then the forecast that minimizes mean squared error is simply the sample mean $\overline{X}$, i.e.,

$$\widehat{X}_{n+1} = \overline{X}_n.$$

- We have obtained the point forecast.

- In the absence of other information, the sample mean is an unbiased forecast.

Forecast Error Under Model 0

- The standard error of the forecast (denoted as s_f) comprises the standard error of Model 0 and the standard error of the sample mean.
- Proof: In Slide 26 Model 0 is

$$X_{n+1} = \overline{X}_n + \varsigma Z.$$

Taking the variance operation on both sides and given i.i.d. assumption,

$$\mathbb{V}\left(X_{n+1}\right) = \mathbb{V}\left(\overline{X}_n\right) + \varsigma^2,$$

since $\mathbb{V}\left(Z\right) = 1$.
For any member of the population, the unbiased estimate of σ^2 is none other than the sample variance s^2! Therefore

$$s_f := \sqrt{\mathbb{V}\left(X_{n+1}\right)} = \sqrt{\frac{s^2}{n} + s^2} = s\sqrt{1 + \frac{1}{n}}. \tag{12}$$

- What is your forecast for Temasek's portfolio value in 2012?

Takeaways

- Standard normal and Student's t
- Population mean μ and variance σ^2, which are mostly unknown
- Unbiased sample mean $\overline{X}$ and variance s^2, given the sample
- **Model 0**: the unbiased point estimate, and a point forecast for the next outcome is the sample mean $\overline{X}$.
- The range of forecast under Model 0 is from LB to UB with a confidence level of, say 95%. In other words, there is a 5% chance that the actual value may fall outside the range.
- Hypothesis test, confidence interval, Type 1 and Type 2 errors
- Unbiasedness, consistency, efficiency, BLUE