Session 3
Statistics

Christopher Ting

https://cting.neocities.org/

Hiroshima University

✉: cting@hiroshima-u.ac.jp

Table of Contents

What is science?

Definition 1.1 (Science).

According to **Britannica**, **science** is system of knowledge that is concerned with the physical world and its phenomena and that entails unbiased observations and systematic experimentation.

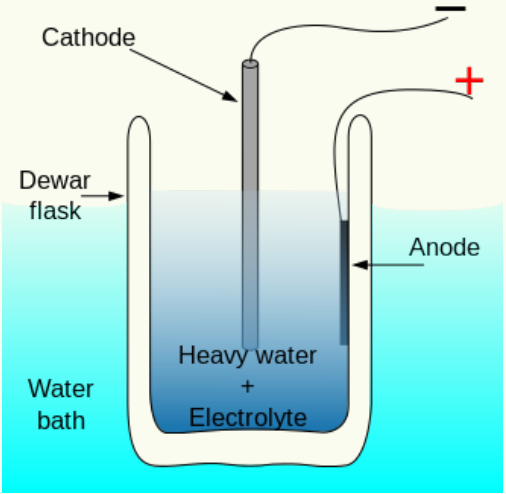
- In general, **science** involves a pursuit of knowledge covering general truths or the **operations** of fundamental laws.
- Physical sciences: astronomy, chemistry, earth science, physics
- Biological sciences: biology, medicine
- Social sciences: anthropology, economics, finance
- Computer sciences: AI, computer security, **data science**, internet, web

What qualifies as science?

- Objectivity: Unbiased reporting of observations, experimental results
- Testability: Hypotheses can be tested by experiments
- Repeatability: Experiments and observations can be repeated
- Reproducibility: Experiments can be reproduced by another team of scientists
- Cause and effect can be repeatably observed at present.
- A claim must be based on evidence.

Is there a negative example of science?

- Cold fusion: cold version of the nuclear fusion happening in the Sun
- Fleischmann–Pons experiment (1989)
- Heavy water (deuterium) and palladium metal as cathode in electrolysis
- The temperature rose suddenly to about 50 ℃ without changes in the input power.
- But this result is not repeatable.
- https://en.wikipedia.org/wiki/Cold_fusion



What is data science?

IBM

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning.

Amazon

Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

Where is the science in data science?

- Objectivity: Unbiased handling of data, absolutely no tempering; unbiased reporting of data analysis results
- Testability: Hypotheses of application domain can be tested by data analyses
- Repeatability: Similar results can be repeated in data not used in estimating the model
- Reproducibility: Same results and insight can be reproduced by other data scientists, another programming language, and different computer platforms
- Cause and effect can be repeatably observed at present by bootstrapping data.

What is cross section

Definition 2.1 (cross section).

A **cross section** is a collection of **observations** at a particular point of time for the purpose of finding out the properties that are common among the observations collected.

Example 1: World Federation of Exchanges

Founded in 1961, the **World Federation of Exchanges** (WFE) is the the global industry group for exchanges and clearing houses around the world. The signature stock exchanges of almost every country are their members. WFE classifies the world into three regions of Americas, Asia-Pacific, and Europe-Africa-Middle East (EAME). Across these three regions, the relevant data of interest to the operators of security exchanges are captured in Table 1.

An example of cross section data

Table 1: Cross-sectional data of world exchanges across three regions as at end of 2019.

Attributes/Region	Americas	Asia-Pacific	EAME
Number of exchanges	16	22	49
Number of listed companies	10,857	32,044	14,889
Total Market capitalization (USD millions)	42,008,600	28,934,050	21,976,959
Total Value of share trading (USD millions)	50,398,246	31,260,089	11,133,684
Number of trades (in thousands)	7,206,295	15,317,659	1,985,968

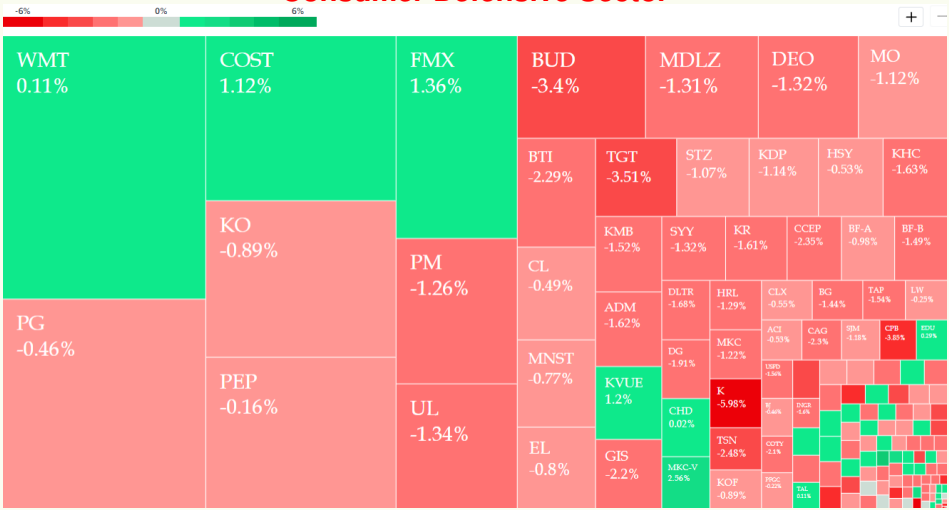
Source: **World Federation of Exchanges**

What are the sectors? (as at end of December 2020. Data source: yahoo!finance)

Sector	Number of stocks listed on NYSE and Nasdaq
Basic Materials	164
Communication Services	175
Consumer Cyclical	379
Consumer Defensive	151
Energy Services	205
Financial Services	1,417
Healthcare	444
Industrials Services	448
Real Estate	381
Technology Services	426
Utilities	113
Total	3,992

Heat Map (as at October 2, 2023. Data source: yahoo!finance)

Consumer Defensive Sector



Currency Cross Rates (as at 13:18 of October 2, 2023. Data source: TradingView)

	EUR	USD	AUD	GBP	NZD	CAD	CHF	JPY	HKD	SGD	CNY
EUR		1.04629	1.65409	0.86677	1.76854	1.43356	0.96149	156.788	8.19135	1.43834	7.6394
USD	0.9555		1.5809	0.82846	1.6907	1.3699	0.91898	149.847	7.8287	1.37465	7.3005
AUD	0.6043	0.63252		0.52365	1.06949	0.8667	0.58133	94.78	4.947	0.86954	4.6263
GBP	1.1531	1.2069	1.90816		2.04079	1.65376	1.10915	180.868	9.451	1.65927	8.8146
NZD	0.5647	0.59134	0.93474	0.48979		0.81035	0.54348	88.62	4.6339	0.81256	4.318
CAD	0.6974	0.7294	1.1536	0.6042	1.2322		0.67069	109.347	5.71321	1.0031	5.3297
CHF	1.0395	1.0877	1.7202	0.9011	1.8396	1.4897		163.042	8.5184	1.49568	7.9441
JPY	0.00638	0.00667	0.01055	0.00552	0.01127	0.00914	0.00613		0.05205	0.00913	0.04868
HKD	0.12201	0.12768	0.20193	0.10574	0.21555	0.17491	0.11738	19.118		0.17552	0.9321
SGD	0.695	0.7271	1.14774	0.60232	1.2249	-	0.6685	108.94	5.694		5.3112
CNY	0.13084	0.13693	0.212	0.1134	0.227	0.1872	0.1255	20.505	1.072	0.1879	

Population

Definition 3.1 (Population).

The **population** in statistics is the set of all members that share a common characteristic or a set of common features. It can also be defined as a group of all objects or events that share something in common.

- All the common stocks listed on an exchange as at end of December 2022
- All the common stocks delisted from an exchange as at end of December 2022
- All the cryptocurrencies in the world as at end of December 2022
- Companies that went public in 2022 in the United States.
- All the ETF providers in the world in 2022.

Sample

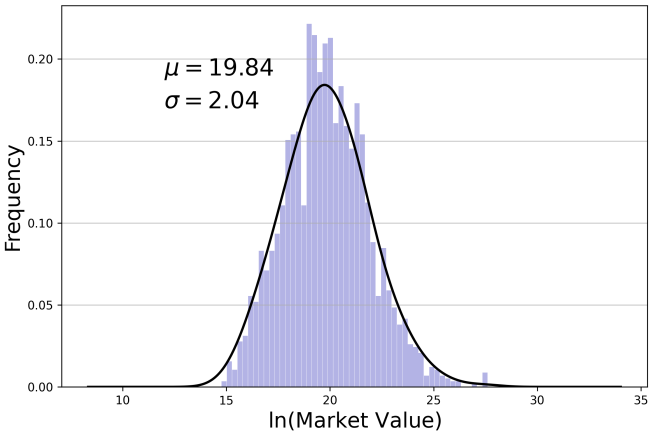
Definition 3.2.

A **sample** is a subset of population, which is drawn randomly or “blindly,”, i.e., without preconceived bias of one member against the other member to be chosen for constituting the sample.

- Sometimes it is impossible for every member of the population to be present.
- It is costly to estimate with the entire population.
- There is only population but many samples.
- Samples enable repeatable and reproducible tests.

Basic Data Visualization (as at May 24, 2019. Data source: Macrotrends)

- Take all stocks listed on Nasdaq.
- Convert **market capitalization** to the natural log scale.
- Present the distribution of log market value as a **histogram**.
- Obtain and plot the **kernel density estimation**.



Population Mean and Variance

Definition 3.3.

The **population average** μ is defined as the sum of all values divided by the total number N values in the summation. Given N values of $x_1, x_2, \ldots, x_N$, the **population mean** μ is given by

$$\mu := \mathbb{E}(x) = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{1}$$

Definition 3.4.

The **population variance** σ^2 is defined as the sum of squared deviations from the population average divided by the total number N. More precisely,

$$\sigma^2 := \mathbb{V}(x) \equiv \mathbb{E}\left((x-\mu)^2\right) = \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \mu\right)^2. \tag{2}$$

Population Coariance

Definition 3.5.

Consider two populations labeled respectively by their random variables x and y. Their population means are denoted by μ_x and μ_y. Both populations have equal number of constituents N. The **population covariance** is defined as

$$\sigma_{xy} := \mathbb{C}\left(x, y\right) \equiv \mathbb{E}\left((x-\mu_x)(y-\mu_y)\right) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y). \tag{3}$$

- In the special case where $x = y$, implying that $\mu_x = \mu_y =: \mu$, we obtain the **variance**, i.e.,
$$\mathbb{C}\left(x, x\right) \equiv \mathbb{E}\left((x-\mu)^2\right) = \sigma^2.$$
- In general, covariance captures how two random variables co-vary.

Properties of Variance

Proposition 1

Suppose x and y form a pair of random variables with means $\mu_x := \mathbb{E}(x)$ and $\mu_y := \mathbb{E}(y)$, respectively. Then

$$\mathbb{V}\left(x\right) = \mathbb{E}\left(x^2\right) - \mathbb{E}(x)^2, \tag{4}$$

and

$$\mathbb{C}\left(x, y\right) = \mathbb{E}\left(xy\right) - \mathbb{E}(x)\,\mathbb{E}(y). \tag{5}$$

- If $\mathbb{E}(x) = 0$, then $\mathbb{E}\left(x^2\right) = \mathbb{V}\left(x\right)$.
- If $\mathbb{E}(x) = \mathbb{E}(y) = 0$, then $\mathbb{E}\left(xy\right) = \mathbb{C}\left(x, y\right)$.

Proof of Proposition 1

Proof.

We shall prove (5) only and treat (4) as a corollary, since $\mathbb{C}(x,x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2$. Quadratic expansion produces

$$\begin{aligned}
\mathbb{C}\left(x,y\right) &= \mathbb{E}\left((x-\mu_x)(y-\mu_y)\right) \\
&= \mathbb{E}\left(xy - \mu_y x - \mu_x y + \mu_x\mu_y\right) \\
&= \mathbb{E}(xy) - \mu_y\,\mathbb{E}(x) - \mu_x\,\mathbb{E}(y) + \mu_x\mu_y \\
&= \mathbb{E}(xy) - \mu_y\mu_x - \mu_x\mu_y + \mu_x\mu_y = \mathbb{E}(xy) - \mu_x\mu_y \\
&= \mathbb{E}\left(xy\right) - \mathbb{E}(x)\,\mathbb{E}(y).
\end{aligned}$$

□

Variance of a Linear Combination of Two Variables

Proposition 2

Suppose a and b are two constants. Given the same setting of Proposition 1,

$$\mathbb{V}\left(ax+by\right)=a^2\,\mathbb{V}\left(x\right)+b^2\,\mathbb{V}\left(y\right)+2ab\,\mathbb{C}\left(x,y\right). \tag{6}$$

- Equation (6) is analogous to quadratic expansion

$$(ax+by)^2=a^2x^2+b^2y^2+2abxy.$$

- If $\mathbb{E}(x)=\mathbb{E}(y)=0$, we can apply the expectation operator on both sides of the above equation to obtain

$$\mathbb{E}\left((ax+by)^2\right)=a^2\,\mathbb{E}\left(x^2\right)+b^2\,\mathbb{E}\left(y^2\right)+2ab\,\mathbb{E}\left(xy\right).$$

- Since $\mathbb{E}(ax+by)=a\,\mathbb{E}(x)+b\,\mathbb{E}(y)=0$, by Proposition 1, we obtain $\mathbb{E}\left((ax+by)^2\right)=\mathbb{V}\left(ax+by\right)$. Likewise, $\mathbb{E}\left(xy\right)=\mathbb{C}\left(x,y\right)$.

Proof of Proposition 2

Proof.

Let $z = ax + by$. It follow from (4) that $\mathbb{V}(z) = \mathbb{E}\left(z^2\right) - \left(\mathbb{E}(z)\right)^2$.
Consequently,

$$\mathbb{V}\left(ax + by\right) = \mathbb{E}\left((ax + by)^2\right) - \left(\mathbb{E}(a\mu_x + b\mu_y)\right)^2.$$

Expanding the two quadratic term and collecting the expanded terms accordingly, we obtain

$$\begin{aligned}\mathbb{V}\left(ax + by\right) &= a^2\,\mathbb{E}\left(x^2\right) - a^2\mu_x^2 + b^2\,\mathbb{E}\left(y^2\right) - b^2\mu_y^2 + 2ab\,\mathbb{E}\left(xy\right) - 2ab\mu_x\mu_y \\ &= a^2\left(\mathbb{E}\left(x^2\right) - \mu_x^2\right) + b^2\left(\mathbb{E}\left(y^2\right) - \mu_y^2\right) + 2ab\left(\mathbb{E}\left(xy\right) - \mu_x\mu_y\right).\end{aligned}$$

Applying (4), the first two terms are $a^2\,\mathbb{V}(x)$ and $b^2\,\mathbb{V}(y)$, respectively. Applying (5), we recognize that the last term is $2ab\,\mathbb{C}(x, y)$. □

Sample Mean

Definition 3.6.

The **sample average** $\overline{x}$ is the sum of all the values divided by the total number n of values in the summation. Given a sample of randomly selected observations, $x_1, x_2, \ldots, x_n$, and by definition, $n < N$, the sample average is calculated with a subset of the population:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{7}$$

Sample Variance

Definition 3.7.

The **sample variance** s^2 is defined as the sum of squared deviations from the sample average divided by $n-1$. Given a sample of randomly selected observations $x_1, x_2, \ldots, x_n$, the sample variance is obtained as follows:

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2. \tag{8}$$

- It is important to recognize that both the sample mean $\overline{x}$ and the sample variance s^2 are random.
- Their randomness is driven by **random sampling**.

Variance of a Linear Combination of Two Variables

Proposition 3

If each data point of the sample is taken randomly from a population, then the **variance** $\mathbb{V}\left(\overline{x}\right)$ of **sample mean** is given by

$$\mathbb{E}\left(\left(\overline{x}-\mu\right)^2\right) \equiv \mathbb{V}\left(\overline{x}\right) = \frac{\sigma^2}{n}. \tag{9}$$

- The collection of sample averages has a lower variance than the population variance σ^2. In fact, it is n times smaller.
- Averaging a group of numbers essentially is to obtain a number “in the middle” to serve as a single representative of the group.
- The average provides a smoothing of the group by a single number in the middle.
- Every sample average is already a “smoothed” representative of the sample. Therefore, it has smaller than the raw sample.

Proof of Proposition 3

- The sample average is a linear combination of randomly taken n observations from the same population.
- Applying (6), we obtain

$$\mathbb{V}\left(\overline{x}\right) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left(x_i\right) + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathbb{C}\left(x_i, x_j\right).$$

- For any pair of observations randomly taken from the population, they should have no covariance by definition of randomness. Therefore,

$$\mathbb{V}\left(\overline{x}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + 0 = \frac{1}{n^2}n\sigma^2$$

$$= \frac{\sigma^2}{n}.$$

An Example of Empirical Test

- How true is Proposition 3?
- Altogether, there are 2,249 observations of non-zero market values (or capitalizations). We then divide exactly the population of stocks into 173 sampes, with each sample having 13 market values in natural log.
- In other words, every stock is in one and only one of the 173 samples, from which we obtain 173 sample averages.
- The average of these 173 sample averages matches exactly the population mean $\mu = 19.84$, as it should.
- We then compute the variance of the sample averages and it turns out to be 0.3343.

An Example of Empirical Test (つづき)

- Now, the variance σ^2 of the population of log market values can be computed by (2), and it is 4.1684.
- Since $n = 13$, according to Proposition 3, the variance of the sample averages should be
$$\frac{4.1684}{13} = 0.3206.$$
- This 0.3206 is the “theoretical"" value.
- The difference between these two variances is $0.3343 - 0.3206 = 0.0137$, only about 4.3% ($= 0.0137/0.3206$).

What is meant by unbiased?

Definition 3.8.

Suppose that $x_1, x_2, \ldots, x_n$ is a **random sample** drawn form the population. An estimator $\widehat{\mu}$ (as a function of the random sample) of μ is said to be **unbiased** if

$$\mathbb{E}\left(\widehat{\mu}\right) = \mu.$$

- Intuitively, this definition says that the average (more precisely the **expected value**) of the averages is the true value.

Proposition 4

The **sample average** $\overline{x}$ (7), as an estimator of population mean μ, is unbiased.

Proof of Proposition 4

Proof.

By simple summation and the linear property of $\mathbb{E}(\cdot)$,

$$\begin{aligned}\mathbb{E}\left(\overline{x}\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\,\mathbb{E}\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu \\ &= \mu\,.\end{aligned}$$

□

- We have assumed that summation and expectation operation are interchangeable.

A Useful Lemma

Lemma 3.9.

$$\sum_{i=1}^{n}(x_i - \mu) = n\big(\overline{x} - \mu\big).$$

Proof.

From the definition of sample average, we have $\sum_{i=1}^{n} x_i = n\overline{x}$. Consequently, noting that μ is a constant,

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \mu) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu. = n\overline{x} - \mu\sum_{i=1}^{n} 1 \\ &= n\overline{x} - n\mu\,.\end{aligned}$$

□

Unbiased sample variance s^2

Proposition 5

The sample variance s^2 in (8), as an estimator of population variance σ^2, is unbiased.

Proof.

We need to establish that

$$\mathbb{E}\left(s^2\right) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right) = \sigma^2.$$

We focus on the sum of squared deviations from the sample average.

Proof of Proposition 5 (cont'd)

Algebraically, in light of Lemma 3.9,

$$\begin{aligned}
\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 &= \sum_{i=1}^{n} \Big((x_i - \mu) - (\overline{x} - \mu) \Big)^2 \\
&= \sum_{i=1}^{n} \Big((x_i - \mu)^2 - 2(\overline{x} - \mu)(x_i - \mu) + (\overline{x} - \mu)^2 \Big) \\
&= \sum_{i=1}^{n} (x_i - \mu)^2 - 2(\overline{x} - \mu) \sum_{i=1}^{n} (x_i - \mu) + (\overline{x} - \mu)^2 \sum_{i=1}^{n} 1 \\
&= \sum_{i=1}^{n} (x_i - \mu)^2 - 2(\overline{x} - \mu) \cdot n(\overline{x} - \mu) + (\overline{x} - \mu)^2 \cdot n.
\end{aligned}$$

Proof of Proposition 5 (cont'd)

It follows that

$$\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\sum_{i=1}^{n}(x_i-\mu)^2-2n(\overline{x}-\mu)^2+n(\overline{x}-\mu)^2=\sum_{i=1}^{n}(x_i-\mu)^2-n(\overline{x}-\mu)^2.$$

Applying the expectation operator on both sides of the equation, we obtain, in light of (9),

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)&=\mathbb{E}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)-n\,\mathbb{E}\left((\overline{x}-\mu)^2\right)\\&=\sum_{i=1}^{n}\mathbb{E}\left(\left(x_i-\mu\right)^2\right)-n\frac{\sigma^2}{n}\\&=\sum_{i=1}^{n}\sigma^2-\sigma^2=\sigma^2\sum_{i=1}^{n}1-\sigma^2=n\sigma^2-\sigma^2=(n-1)\sigma^2.\end{aligned}$$

Proof of Proposition 5 (cont'd)

Therefore, we can conclude that

$$\sigma^2 = \frac{1}{n-1}\,\mathbb{E}\left(\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right) = \mathbb{E}\left(s^2\right).$$

Since $\mathbb{E}\left(s^2\right) = \sigma^2$, we have demonstrated that the sample variance s^2 (8) is indeed an unbiased estimator of σ^2.

□

- In summary, we have found that the **sample** mean defined by (3.6) and the **sample variance** defined by (8) are both unbiased.

Dispersion

- Suppose we define the **dispersion** or deviation from the sample average as ϵ_i for a particular observation i, i.e., $\epsilon_i := x_i - \overline{x}$.
- If $\overline{x}$ is unbiased, then $\mathbb{E}\left(\overline{x}\right) = \mu$.
- Now, since x_i for $i = 1, 2, \ldots, n$ are randomly taken from the same population with which $\overline{x}$ is computed, $\mathbb{E}\left(x_i\right) = \mu$ for every i.
- It follows that $\mathbb{E}\left(\epsilon_i\right) = \mathbb{E}\left(x_i\right) - \mathbb{E}\left(\overline{x}\right) = 0$.

Model 0

Definition 3.10.

A **cross-sectional sample** y_i, which is yet to be drawn from the population, can be modeled as a **random variable** as follows:

$$y_i = \overline{y} + \epsilon_i.$$

Here, $\overline{y}$ is the **sample mean**, and ϵ_i is "**noise**" with the property that $\mathbb{E}\left(\epsilon_i\right) = 0$, and $\mathbb{C}\left(\epsilon_i, \epsilon_j\right) = \sigma_\epsilon^2$ if $i = j$ and zero otherwise.

- Obviously, $\mathbb{E}(y_i) = \mathbb{E}\left(\overline{y}\right) + \mathbb{E}(\epsilon_i) = \mu$.
- Therefore, an intuitive and practical interpretation of this model is that the **unbiased prediction** for y_i is the **sample average** $\overline{y}$,

Variance of Noise

Proposition 6

Let σ^2 be the **population variance** of y_i and n the sample size. Then the **variance of noise** is given by

$$\sigma_\epsilon^2 = \sigma^2.$$

Proof.

The noise can be written as $\epsilon_i = y_i - \overline{y}$. Accordingly,

$$\mathbb{V}\left(\epsilon_i\right) = \mathbb{V}\left(y_i\right) + \mathbb{V}\left(\overline{y}\right) - 2\,\mathbb{C}\left(y_i, \overline{y}\right).$$

Now, the sample average $\overline{y}$ is computed based on the samples drawn randomly from the population. In the context of the model, the sample average is a constant. Therefore, $\mathbb{V}\left(\overline{y}\right) = 0$ and $\mathbb{C}\left(y_i, \overline{y}\right) = 0$. It follows that $\mathbb{V}\left(\epsilon_i\right) = \sigma^2 + 0 - 0 = \sigma^2$. □

Remarks

- This proposition suggests that the variance σ_ϵ^2 of the noise is necessarily a constant, which equals the variance of the population.
- In fact, it is a part of the statement of identical distribution: $\mathbb{V}\left(y_i\right) = \mathbb{V}\left(\epsilon_i\right) = \sigma^2$, $i = 1, 2, \ldots, N$.
- As another remark, $\epsilon_i = y_i - \overline{y}$ may be interpreted as observations that have been **de-meaned**.
- In other words, $\left\{\epsilon_i\right\}_{i=1}^n$ is a sample of n cross-sectional observations such that the sample mean is zero.
- In more advanced statistical analysis, **de-meaning** the data is an important procedure before feeding them to the algorithmic engine.

Introduction

- A **statistical test** provides an algorithm for making quantitative decisions given a collection of data and a statistical model.
- What is scientific about statistics is that a conjecture is put forth and a test is proposed to determine whether the conjecture can be rejected or not.
- The conjecture is called the **null hypothesis**, which carries the nuance of nothing interesting.
- The intent is to determine whether there is enough evidence to “reject” the null hypothesis about the process that presumably generates the data.
- If the null hypothesis cannot be rejected, we will take it as if the null hypothesis is true.

Very Important Principle

- It is important to stress that scientists should not yield to the temptation to unscientifically reject the null hypothesis when the test returns a “disappointing” result.
- Very possibly, we do not yet have enough data to “prove” our claim.

Standard Error

Definition 4.1.

The square root of the variance of an estimator $\widehat{\theta}$ is called the **standard error**. We denote it by $\text{se}\left(\widehat{\theta}\right)$.

- According to Proposition 3, the variance of the sample average estimator $\overline{y}$ is $\dfrac{\sigma^2}{n}$.
- Therefore, the standard error is

$$\text{se}\left(\overline{y}\right) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

z Score

Definition 4.2.

Suppose the sample average $\overline{y}$ is computed from a sample of n observations. The ratio

$$\widehat{z} := \frac{\overline{y} - \mu}{\text{se}(\overline{y})} = \frac{\overline{y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n} \cdot \frac{\overline{y} - \mu}{\sigma}. \tag{10}$$

is called the ***z* score**.

- Note that z score's mean is 0 and its variance is 1.
- The proof is simple:

$$\mathbb{E}\left(\widehat{z}\right) = \sqrt{n}\ \mathbb{E}\left(\frac{\overline{y} - \mu}{\sigma}\right) = \sqrt{n}\ \left(\frac{\mathbb{E}(\overline{y}) - \mu}{\sigma}\right) = 0.$$

$$\mathbb{V}\left(\widehat{z}\right) = \frac{n}{\sigma^2}\,\mathbb{V}(\overline{y} - \mu) = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1.$$

Discussion

- How close is the estimate $\overline{y}$ to the hypothesized value of μ?
- The difference $\overline{y} - \mu$ alone cannot answer this question.
- If the variance is large, a numerically big difference may be considered to be close.
- Conversely, if the variance is very small, even a numerically small difference may be considered to be far apart.
- If the absolute value of the z score is larger than some number, then we can say with certain **level of confidence** that the difference is **statistically significant**.
- More cautiously, the difference is not compatible with the hypothesis of no difference.

Example of z Score

- Take 13 samples from the log market values of Nasdaq's listed stocks:

17.277176	20.860215	17.636295	18.883911	20.861618	21.548366	20.041751
21.598480	19.580890	23.194863	18.980868	21.373855	18.852528	

- Calculated sample average is 20.05.
- In Slide 15, we have $\mu = 19.84$ and $\sigma = 2.04$.
- The z **score** with sample size of $n = 13$ is calculated according to (10) as follows:

$$\widehat{z} = \sqrt{13} \cdot \frac{20.05 - 19.84}{2.04} = 0.37\,.$$

- Statistically, the result suggests that the difference is about 0.37 standard deviations away from mean 0.

Hypotheses

Definition 4.3.

The **null hypothesis** H_0 is the statement about a population parameter. The **alternative hypothesis** H_a is a statement that directly contradicts the null hypothesis.

Definition 4.4.

If the null hypothesis is such that the population statistic $\theta = 0$ while the alternative hypothesis is either $\theta > 0$ or $\theta < 0$, then the test is said to be **one-tail**.

Definition 4.5.

When the null hypothesis is such that the population statistic $\theta = 0$ while the alternative hypothesis is $\theta \neq 0$, then the test is said to be **two-tail**.

Level of significance and p Value

Definition 4.6.

The **level of significance** α refers to the likelihood, i.e., probability of wrongly rejecting the null hypothesis when it is actually true.

Definition 4.7.

A ***p*-value** is the probability that a statistical summary of the data (such as the z score as a random variable from random sampling) would be equal to or more extreme than its estimated value $\widehat{z}$ in magnitude.

$$p\text{-value} := \mathbb{P}\left(|z| > \widehat{z}\right).$$

Insight of α and p

- The significance level α is the acceptable probability of being wrong about the null hypothesis.
- By convention, acceptable probability is usually set at 5% or 1%.
- The p value is the probability of getting an extreme z value that is more significant than the estimated value $\widehat{z}$.
- Thus, if $\widehat{z}$ is significant, then the chance of finding another z in the distribution more significant than it is very slim.
- So p becomes a lot smaller, and so we should reject the null hypothesis that $\widehat{z} = 0$.
- How small should p be to reject the null hypothesis?
- Answer: $p \leq \alpha$.

Principles and Caveats When Using the p value

1. The p-value method provides one approach to summarizing the incompatibility between a particular set of data (sample) and a proposed model for the data (null hypothesis).
2. A p-value does not measure the probability that the null hypothesis is true, or the probability that the data were produced by random chance alone. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.
3. Scientific conclusions and business or policy decisions should not be based only on whether it passes a specific threshold. Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making.

Principles and Caveats When Using the p value (つづき)

4. Proper inference requires full reporting and transparency.
5. A small p-value, or statistical significance, does not measure the size of an effect or the importance of a result. A test maybe statistically significant with a small p value. But the object (e.g., profit) of test may not be financially significant.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Level of significance and p Value

Definition 4.8.

Suppose the level of significance is set to a particular value denoted by α. The **one-tail critical value** is a positive value c_1 with respect to α such that

$$\text{either} \qquad \mathbb{P}\left(z < -c_1\right) = \alpha \qquad \text{or} \qquad \mathbb{P}\left(z > c_1\right) = \alpha$$

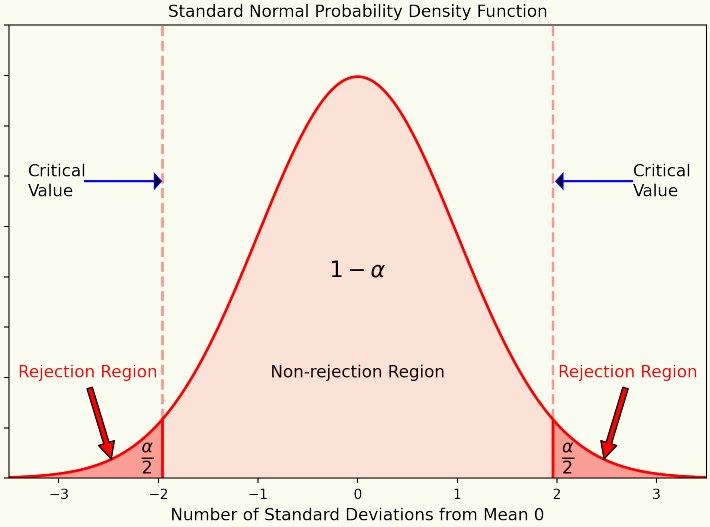
where z is the **standard normal random variable**.

Definition 4.9.

The **two-tail critical value** is a positive value c_2 with respect to α such that

$$\mathbb{P}\left(z < -c_2\right) = \frac{\alpha}{2} \qquad \text{and} \qquad \mathbb{P}\left(z > c_2\right) = \frac{\alpha}{2}.$$

Standard Normal PDF



Standard Normal Cumulative Distribution Function

Definition 4.10.

The **standard normal cumulative distribution function** $\Phi(x)$ is defined as

$$\Phi(x) := \mathbb{P}(X \leqq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}z^2} dz. \tag{11}$$

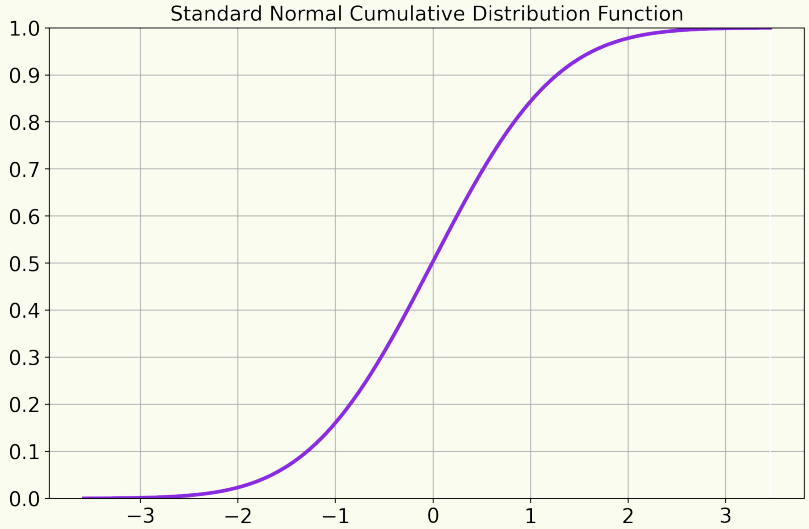
It is the area under the bell-shape curve in Slide 51 from negative infinity to an arbitrary real number x.

- Noting that $\mathbb{P}(X \leqq x) + \mathbb{P}(X > x) = 1 \quad \Longrightarrow \quad \mathbb{P}(X > x) = 1 - \Phi(x)$, the one-tail critical value defined in Definition 4.8 can be rewritten in terms of $\Phi(x)$ as

$$\text{either} \quad \mathbb{P}(z < -c_1) = \Phi(-c_1) = \alpha \quad \text{or} \quad \mathbb{P}(z > c_1) = 1 - \Phi(c_1) = \alpha.$$

- The same expressions apply for the two-tail critical value c_2 as well.

Standard Normal CDF



Non-Standard Normal

- Consider a non-standard normally distributed variable X with mean μ and variance σ^2, for which
$$X \sim N(\mu, \sigma^2).$$
- For example, given a constant value A, we want to find out about how the probability $\mathbb{P}(X \leqq A)$ is connected to the standard normal cumulative distribution function.
- Suppose we subtract the mean μ from both sides of the inequality $X \leqq A$ and then divide the difference by σ. Since σ is necessarily positive, the inequality direction is preserved:
$$\frac{X-\mu}{\sigma} \leqq \frac{A-\mu}{\sigma}.$$
- In other words, we have performed a linear transformation of X to $z = \dfrac{X-\mu}{\sigma}$ and $x = \dfrac{A-\mu}{\sigma}$. Consequently, $\mathbb{P}\left(X \leqq A\right) = \Phi\left(\dfrac{A-\mu}{\sigma}\right)$.

Confidence Interval

Definition 4.11.

In statistics, a **confidence interval** (CI) is a type of interval estimate that might contain the true value of an unknown population parameter. Computed from the statistics of the observed data, the interval has an associated confidence level $1-\alpha$, which quantifies the level of confidence that the parameter lies in the interval.

Proposition 7

Let θ be a population statistic and its **point estimate** is $\widehat{\theta}$ from a sample of observations. At the level of $1-\alpha$, the **two-tail confidence interval** is given by

$$\widehat{\theta} - \text{se}\left(\widehat{\theta}\right) c_2 < \theta < \widehat{\theta} + \text{se}\left(\widehat{\theta}\right) c_2,$$

where c_2 is the **two-tail critical value** of the distribution of the test statistic $\widehat{z}$, which is the z score computed from data.

Proof of Proposition 7

- At the $1-\alpha$ confidence interval, the test statistic such as the z score lies within the non-rejection part of the area under the curve, as in Figure 1. Hence,

$$-c_2 < \widehat{z} < c_2. \tag{12}$$

- Given that

$$\widehat{z} = \frac{\widehat{\theta} - \theta}{\text{se}\left(\widehat{\theta}\right)},$$

we examine the right inequality in (12) first:

$$\frac{\widehat{\theta} - \theta}{\text{se}\left(\widehat{\theta}\right)} < c_2.$$

Proof of Proposition 7 (つづき)

- Multiplying both sides by $-\text{se}\left(\widehat{\theta}\right)$ leads to

$$-\widehat{\theta} + \theta > -\text{se}\left(\widehat{\theta}\right) c_2,$$

which results in $\theta > \widehat{\theta} - \text{se}\left(\widehat{\theta}\right) c_2$, equivalently, $\widehat{\theta} - \text{se}\left(\widehat{\theta}\right) c_2 < \theta$.

- Likewise, for the inequality on the left in (12), i.e.,

$$-c_2 < \frac{\widehat{\theta} - \theta}{\text{se}\left(\widehat{\theta}\right)},$$

similar algebraic moves result in $\theta < \widehat{\theta} + \text{se}\left(\widehat{\theta}\right) c_2$.

□

Example

- In the earlier example, we have computed the sample average and obtained the value of is 20.05.
- The population standard deviation σ is 2.04. It follows that the standard error is

$$\frac{2.04}{\sqrt{13}} = 0.5658.$$

- Let's set the level of significance α at 5%.
- The cutoff on the right side that gives an area of 2.5% under the pdf of the standard normal distribution is 1.960. Thus, $c_2 = 1.960$.

Example (つづき)

- A $1-\alpha$ **confidence interval** for the mean of an assumed normal population is

$$20.05 - 1.960 \cdot 0.5658 < \mu < 20.05 + 1.960 \cdot 0.5658,$$

i.e.,

$$18.94 < \mu < 21.16.$$

- Indeed, the population mean 19.84 is in this confidence interval.

Takeaways

- Population in statistics is everything whereas sample is a subset of the population.
- Cross section is a snapshot of all members of a collection (e.g., currencies).
- Time series, on the other hand, refers to the change of an asset's value.
- Sample mean, sample variance, and sample covariance are estimators.
- If the expected value of an estimator is equal to the population's value, then the estimator is said to be unbiased.
- De-meaning is an important data pre-processing before feeding them to the algorithm.

Takeaways (つづき)

- z score is a statistic for measuring the significance of a null hypothesis.
- The level of significance refers to the likelihood, i.e., probability of wrongly rejecting the null hypothesis when it is actually true.
- There are principles and caveats when we use the p value.
- You need the probability distribution for identifying the critical values.
- Instead of the point estimate, it is more meaningful to provide also the confidence level withe the critical value and the standard error.