Math 170S

Lecture Notes Section 6.1 [*][†]

Descriptive statistics

Instructor: Swee Hong Chan

NOTE: The notes is a summary for materials discussed in the class and is not supposed to substitute the textbook. Materials that appear in the textbook but do not appear in the lecture notes might still be tested. Please send me an email if you find typos.

[*] Version date: Sunday 4th October, 2020, 13:47.

[†] This notes is based on Hanbaek Lyu's and Liza Rebrova's notes from the previous quarter, and I would like to thank them for their generosity. "*Nanos gigantum humeris insidentes* (I am but a dwarf standing on the shoulders of giants)".

1 What is Statistics?

Statistics is studying how to figure out what is inside a dish by tasting the food.

- **Known:** The dish is egg fried rice.
- **Problem:** Find how much MSG is put into the egg fried rice.
- **Input**: Finish one plate of the egg fried rice, then two plates, then three plates, ...

2 What is Statistics? (real)

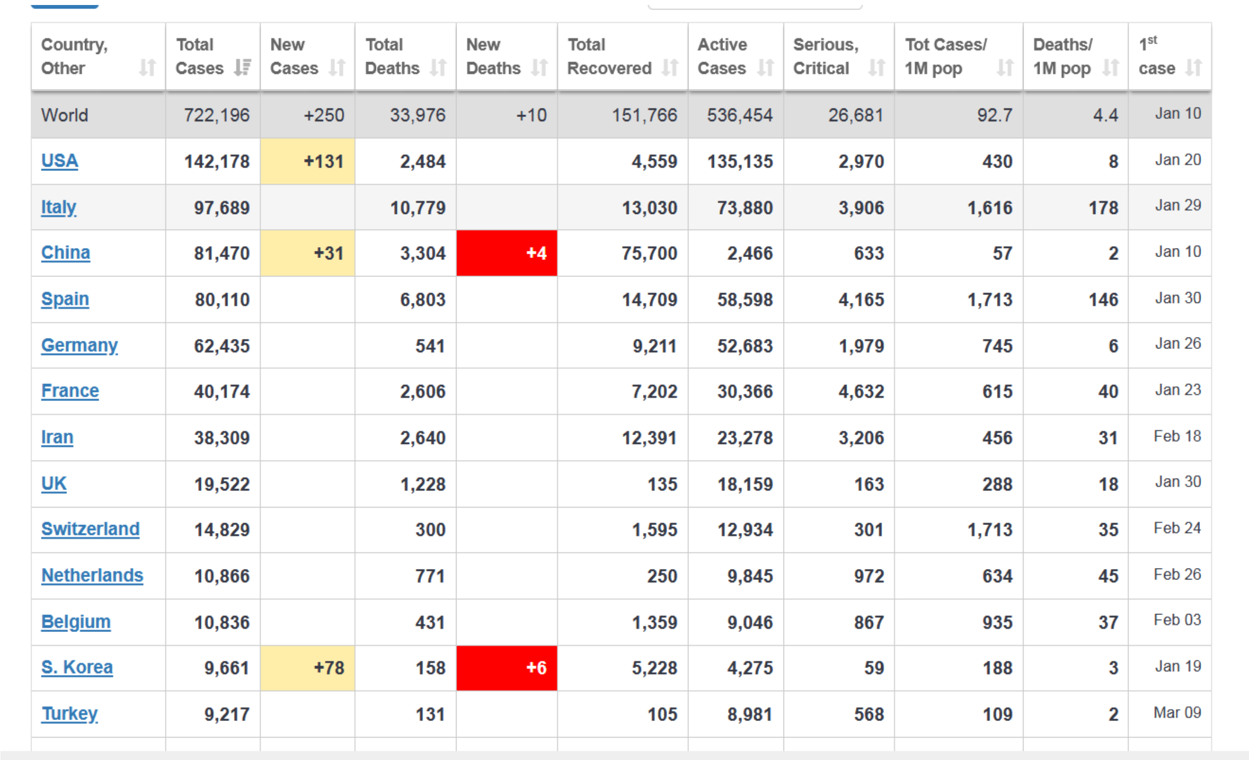
Statistics is studying how to figure out the parameters of a (probabilistic) model by examining the data.

- **Known:** The model is some random variable X (e.g., uniform, normal).
- **Problem**: Find the parameters of X (e.g., mean, variance).
- **Input**: Sample value $x_1, \ldots, x_n$, where x_i is the outcome of the i-th experiment.

Note that here $x_1, x_2, \ldots, x_n$ are not necessarily distinct.

3 Example: Covid-19

You are being tasked to predict how far Covid-19 will spread, give the following data.



Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	1st case
World	722,196	+250	33,976	+10	151,766	536,454	26,681	92.7	4.4	Jan 10
USA	142,178	+131	2,484		4,559	135,135	2,970	430	8	Jan 20
Italy	97,689		10,779		13,030	73,880	3,906	1,616	178	Jan 29
China	81,470	+31	3,304	+4	75,700	2,466	633	57	2	Jan 10
Spain	80,110		6,803		14,709	58,598	4,165	1,713	146	Jan 30
Germany	62,435		541		9,211	52,683	1,979	745	6	Jan 26
France	40,174		2,606		7,202	30,366	4,632	615	40	Jan 23
Iran	38,309		2,640		12,391	23,278	3,206	456	31	Feb 18
UK	19,522		1,228		135	18,159	163	288	18	Jan 30
Switzerland	14,829		300		1,595	12,934	301	1,713	35	Feb 24
Netherlands	10,866		771		250	9,845	972	634	45	Feb 26
Belgium	10,836		431		1,359	9,046	867	935	37	Feb 03
S. Korea	9,661	+78	158	+6	5,228	4,275	59	188	3	Jan 19
Turkey	9,217		131		105	8,981	568	109	2	Mar 09

Figure 1: The number of COVID-19 cases around the world on March 29th, taken from `www.worldometer.com`.

The goal is to be able to draw the following figure.

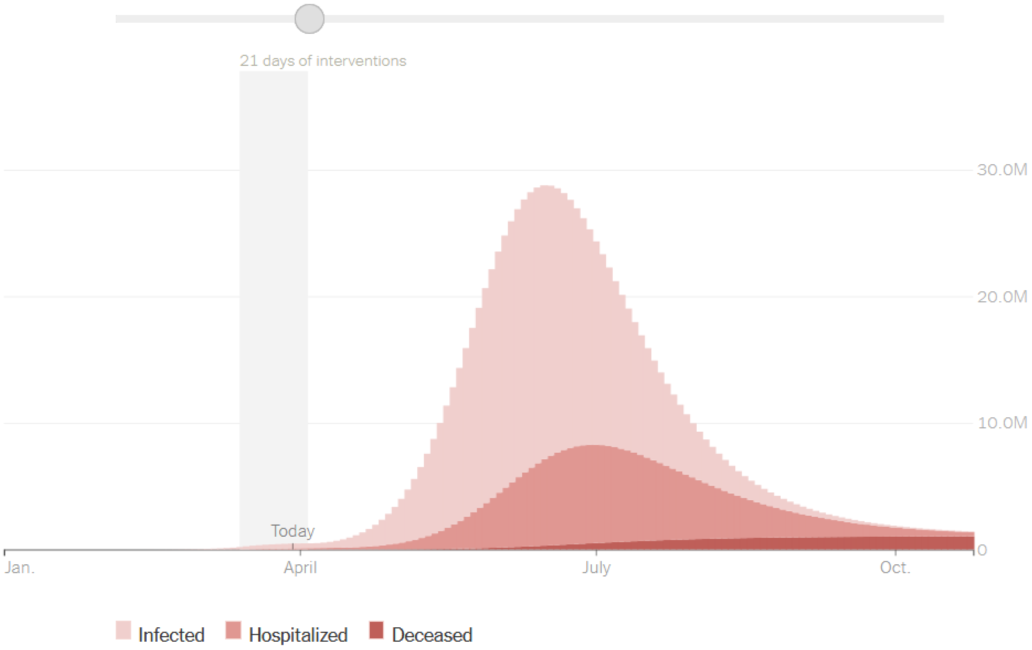


Figure 2: Prediction of the casualties of the COVID-19 outbreak, taken from New York Times.

In our setting, your task is of the following form:

- **Known:** A normal random variable $N(\mu, \sigma^2)$ with unknown mean μ and variance σ^2.
- **Problem**: Find the mean and variance.
- **Input**: The total number of casualties so far.

4 Example: Casinos

The friendly instructor was hired by the casino to check if a player had cheated in the coin-flipping game.

- **Known:** A Bernoulli random variable $B(p)$ with unknown success probability p.
- **Problem**: Find the value of p.
- **Input**: Toss the coin 20 times.

The frieanly instructor flipped the coin 20 times:

Tail, Tail, Tail, Head, Tail, Head, Tail, Tail, Tail, Head,

Tail, Tail, Head, Head, Tail, Head, Tail, Tail, Tail, Head.

He got 7 Heads and 13 Tails.

Thus, the instructor estimated that $p = \frac{7}{20}$, i.e.,

- The coin-flip will come out Head with probability roughly $\frac{7}{20} = 35\%$.
- The coin-flip will come out Tail with probability roughly $\frac{13}{20} = 65\%$.

So this is not a fair coin (i.e., $p = 1/2$), so there is some cheating going on here[1].

[1] In fact, the instructor actually knows that, by the central limit theorem, the probability of cheating going on is roughly 90.1%.

5 Empirial distribution

The probability distribution from the previous example is called the empirical distribution.

Definition 1. Suppose that you have performed n random experiments with outcomes $x_1, x_2, \ldots x_n$; these are called **samples**.
The **empirical distribution** is the probability distribution on real numbers given by

$$P(\text{outcome is } s) = \frac{\#\text{ of } x_i\text{'s equal to } s}{n},$$

where s is any real number.

6 Sample mean

Two important parameters of a random variable is its mean and variance. Here is our 'best guess' for the mean of the (unknown) random variable.

Definition 2. Let X be a random variable with sample values $x_1, x_2, \ldots, x_n$. The corresponding **sample mean** is

$$\overline{\mathrm{x}} := \frac{1}{n}\sum_{i=1}^{n} x_i \ = \ \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

In the casinos example, the sample mean is

$$\overline{\mathrm{x}} = \frac{1}{20}(7 \times 1 + 13 \times 0) = \frac{7}{20},$$

where Head is 1 and Tail is 0.

7 Variance of the empirical distribution

Definition 3. The **variance of the empirical distribution** is

$$\begin{aligned} v :=& \frac{1}{\mathbf{n}} \sum_{i=1}^{n} (x_i - \overline{\mathrm{x}})^2 \\ =& \frac{1}{n} \left[(x_1 - \overline{\mathrm{x}})^2 + (x_2 - \overline{\mathrm{x}})^2 + \ldots + (x_n - \overline{\mathrm{x}})^2 \right]. \end{aligned}$$

In the casino example, the variance of the empirical distribution is

$$\begin{aligned} v =& \frac{1}{20} \left[7 \times (1 - \frac{7}{20})^2 + 13 \times (0 - \frac{7}{20})^2 \right] \\ =& \frac{1}{20} \left[\frac{7 \times 13^2 + 13 \times 7^2}{20^2} \right] = \frac{91}{400}. \end{aligned}$$

8 Sample variance

Definition 4. The **sample variance** is

$$
\begin{aligned}
s^2 &:= \frac{1}{\mathbf{n\text{-}1}} \sum_{i=1}^{n} (x_i - \overline{\mathrm{x}})^2 \\
&= \frac{1}{n-1} \left[(x_1 - \overline{\mathrm{x}})^2 + (x_2 - \overline{\mathrm{x}})^2 + \ldots + (x_n - \overline{\mathrm{x}})^2 \right].
\end{aligned}
$$

In the casino example, the sample variance is

$$
\begin{aligned}
s^2 &= \frac{1}{19} \left[7 \times (1 - \frac{7}{20})^2 + 13 \times (0 - \frac{7}{20})^2 \right] \\
&= \frac{1}{19} \left[\frac{7 \times 13^2 + 13 \times 7^2}{20^2} \right] = \frac{91}{380}.
\end{aligned}
$$

Remark 5. Why $n - 1$ instead of n? It turns out that sample variance is (surprisingly) a better estimate for the variance of X. It is because there is an "inherent bias" that comes with sampling, so we need to use a smaller denominator to cancel this inherent bias.

9 Sample standard deviation

Definition 6. The **sample standard deviation** is

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

the (positive) square root of the sample variance.

In the casino example, the sample standard deviation is

$$s = \sqrt{\frac{91}{380}}.$$

10 Equivalent formulas

Exercise 7. Show that v and s^2 can also be computed by the formulas

$$v = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - \overline{\mathrm{x}}^2;$$

$$s^2 = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n-1} - \frac{n}{n-1}\,\overline{\mathrm{x}}^2\,.$$