Prophetic Statistics

Haiyue Liu

Dr. Hubert Bray, Duke University

Mathematics of the Universe

July 24, 2017

Abstracts:

This paper includes a variety of information about statistics, a branch of mathematics. You can learn the important facts of statistics, as well as the applications of statistics through the paper. Some real analytical cases are also contained.

Introduction:

Our contemporary society features a boom of data. Whenever you browse the Internet, navigate your car, upload raw data or make deals in the stock market, you are producing big data. Moreover, these data are so valuable that they can enable companies to make tremendous profits from public fashion trends, enable governments to decrease rates of crime, enable the fire department to prevent forest fire, and enable us to **predict** lots of things in every walk of life.

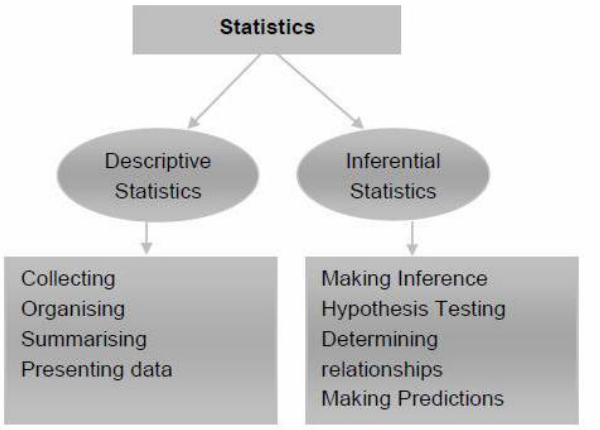
[1]

[1]https://media.licdn.com/mpr/mpr/AAEAAQAAAAAAAAitAAAAJDAzZTkwNjZjLWFjNjktNDEwNy04YTA3LTJiZWEyMmMxNmFkOA.jpg

However, confronted with such a huge database, it is too complex for us to analyse it without a set of systematic methods. Therefore, **statistics**, a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data on the basis of probability, flourishes.[2]

Statistics in stock market [3]

In statistics, we collect and scrutinize every data sample in a set of items from which samples can be drawn. The results have two kinds: **inferential statistics** and **descriptive statistics**. The former deduce properties of an underlying probability distribution by analysis of data, while the latter quantitatively describe or summarize features of a collection of information.[4] They can be pretty correct. But, despite all its advantages, statistics is concerned with the use of data in the context of uncertainty and decision making in the face of uncertainty.[5] So, we should remain a bit skeptical of the statistics itself.



---

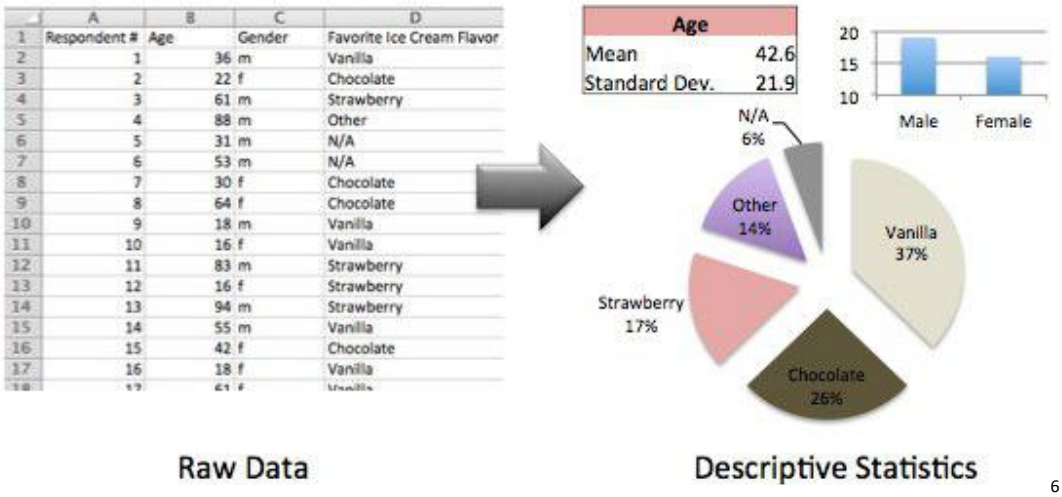
[2] Dodge, Y. (2006) *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-920613-9

[3] https://media.licdn.com/mpr/mpr/shrinknp_800_800/AAEAAQAAAAAAAAN0AAAAJDdlYmQ5NTlmLTFiMWEtNDhiZS1hMmExLTFmNGE0NjBmZDVhNg.jpg

[4] Mann, Prem S. (1995). *Introductory Statistics* (2nd ed.). Wiley. ISBN 0-471-31009-3.

[5] Chance, Beth L.; Rossman, Allan J. (2005). "Preface". *Investigating Statistical Concepts, Applications, and Methods*

From raw data to descriptive statistics



[6]

History of Statistics:

Before solving practical cases, let's learn some important developments of statistics:

The birth of statistics is often dated to 1662, when **John Graunt** developed early human statistical and census methods that provided a framework for modern demography. He produced the first **life table**, giving probabilities of survival to each age.[7]

Table 1. Life table for the total population: United States, 2009

Age (years)	Probability of dying between ages x and $x+1$ q_x	Number surviving to age x l_x	Number dying between ages x and $x+1$ d_x	Person-years lived between ages x and $x+1$ L_x	Total number of person-years lived above age x T_x	Expectation of life at age x e_x
0–1	0.006372	100,000	637	99,444	7,846,926	78.5
1–2	0.000407	99,363	40	99,343	7,747,481	78.0
2–3	0.000274	99,322	27	99,309	7,648,139	77.0
3–4	0.000209	99,295	21	99,285	7,548,830	76.0
4–5	0.000160	99,274	16	99,266	7,449,545	75.0
5–6	0.000150	99,259	15	99,251	7,350,279	74.1
6–7	0.000135	99,244	13	99,237	7,251,028	73.1
7–8	0.000122	99,230	12	99,224	7,151,791	72.1
8–9	0.000109	99,218	11	99,213	7,052,566	71.1
9–10	0.000095	99,207	9	99,203	6,953,354	70.1
10–11	0.000087	99,198	9	99,194	6,854,151	69.1
11–12	0.000093	99,189	9	99,185	6,754,957	68.1
12–13	0.000127	99,180	13	99,174	6,655,773	67.1
13–14	0.000193	99,167	19	99,158	6,556,599	66.1
14–15	0.000279	99,148	28	99,134	6,457,441	65.1
15–16	0.000370	99,121	37	99,102	6,358,307	64.1
16–17	0.000454	99,084	45	99,061	6,259,205	63.2
17–18	0.000537	99,039	53	99,012	6,160,143	62.2
18–19	0.000615	98,986	61	98,955	6,061,131	61.2
19–20	0.000691	98,925	68	98,891	5,962,175	60.3

[8]

[6] http://www.mymarketresearchmethods.com/wp-content/uploads/2011/11/Descriptive-Statistics.jpg

[7] Willcox, Walter (1938) *The Founder of Statistics*. Review of the International Statistical Institute.

[8] https://www.mathworks.com/help/finance/life_table_1.png

And then, the approach of statistics extended from governance to more fields, whose mathematical foundations heavily drew on the new probability theory at that time. For example, the well-known **Bayes' theorem**:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

where $P(B) \neq 0$, and $P(A \mid B)$ means conditional probability, the probability of observing event A given that event B is true.

In 1795, **Gauss** discovered the **normal distribution**, which was significant in statistics, because it is often used in the natural and social sciences to represent real-valued random variables whose distributions are not known.[9] Below is the $N(\mu, \sigma^2)$ distribution:

$$f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

where μ is the expectation of the distribution, σ is the standard deviation, and σ^2 is the variance.

Then, **Thorvald Nicolai Thiele**, in 1889, introduced the **likelihood function**.[10] This function has a very simple basic form:

$$\mathcal{L}(\theta \mid x) = P(x \mid \theta)$$

And let's see a easy example of it:

Suppose that a coin is tossed $n = 10$ times and that $s = 4$ heads are observed. With no knowledge whatsoever concerning the probability of getting a head on a single toss, the appropriate statistical

---

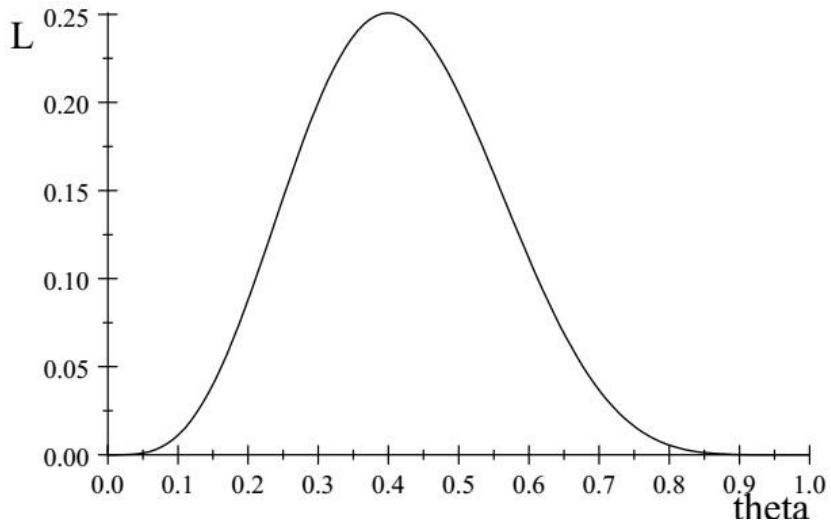
[9] *Normal Distribution*, Gale Encyclopedia of Psychology

[10] Anders Hald (1998). *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.

model for the data is the Binomial (10, θ) model with $\theta \in \Omega = [0,1]$.[11] The likelihood function is given by

$$L(\theta \mid 4) = \binom{10}{4}\theta^4(1-\theta)^6$$

which is plotted as



This likelihood peaks at $\theta = 0.4$ and takes the value 0.2508 there. Find θ not 0.5? This is right, because it's based on the observed data. However, the sample is far from enough, so that it defies common sense. Remember that a sample of proper **size** is the key to success in statistics!

Thiele

As time goes by, the statistical methods have been improved and enriched. The use of modern computers has expedited large-scale statistical computations, and has also made possible

[11] Probability and Statistics The Science of Uncertainty 2nd Edition

new methods that are impractical to perform manually. Statistics continues to be an area of active research, especially on the problem of how to analyze Big data.[12]

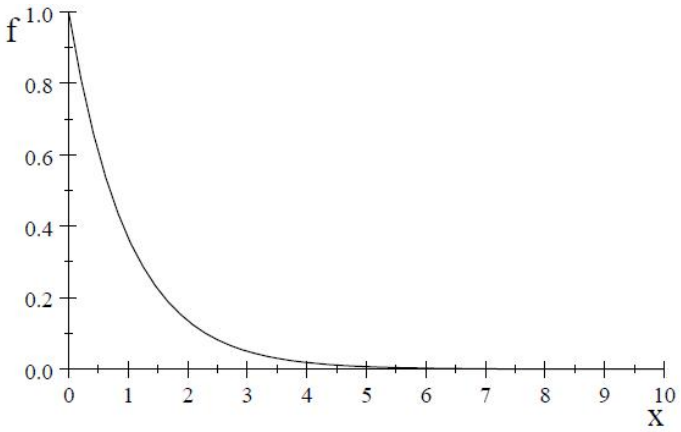
Cases of Statistical Analysis:

Here are some cases shedding light on the functions of statistics. In this paper, the process of collecting data is ignored, since it is onerous; we primarily focus on the **inferential statistics**.

Case 1 : First, I'll give a case as enlightenment.

Consider the life-length X in years of a machine where it is known that $X \sim \text{Exp}(1)$.

i.e. $f_X(x) = e^{-x},\ x \geq 0.$



Plot of the Exponential(1) density f. [13]

We get several questions below, and each one of them means something.

1) X lives in $(0,c)$ of probability 95%, solving for c.

$$0.95 = \int_0^c e^{-x}\,dx = 1 - e^{-c}$$

$$c = -\ln(0.05) = 2.9957.$$

---

[12] "Science in a Complex World - Big Data: Opportunity or Threat?". *Santa Fe Institute*.

[13] Probability and Statistics The Science of Uncertainty 2nd Edition

This interval gives a reasonable range of probable life-lengths for the new machine. This piece of information is of great importance to both customers and producers.

2) The probability of $X > 5$?

$$P(X > 5) = \int_5^{\infty} e^{-x}\, dx = e^{-5} = 0.0067$$

Quite low, isn't it? This agrees with the conclusion from Q1. Five years is not a plausible life-length for a newly purchased machine.

3) The expectation of X?

Well, the expectation, or **expected value,** of X is given by:

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

Then substitute the unknown function with our $f_X(x)$:

$$E(X) = \int_0^{\infty} xe^{-x} dx = 1.$$

So, on average, this kind of machine works for 1 year, which is like the "use by" date on food. You may consider repairing or replacing the old machine after using it for 1 year.

4) Suppose we bought a machine that has already been used for one year, what is the expected life length of it?

This is a very interesting question, which actually asks:

$$E(X|X > 1) = ?$$

We have

$$P(X > x) = \int_x^{\infty} e^{-x} dx = e^{-x}.$$

Employ **Bayes' theorem**:

$$P(X > x|X > 1) = \frac{P(\{X > x\} \cap \{X > 1\})}{P(X > 1)} = \frac{P(X > x)}{P(X > 1)} = \frac{e^{-x}}{e^{-1}} = e^{-(x-1)}.$$

$$f_{X|X>1}(x) = e^{-(x-1)}.$$

Harness expectation as has been mentioned above:

$$E(X|X > 1) = \int_1^{\infty} xe^{-(x-1)}dx = 2.$$

Curiously, we find that the machine that has been used for 1 year is expected to work for **another** year! And this really make sense, because those machines that have survived the first year must enjoy good quality, so that they can last longer than average machines.

5) Suppose we bought a machine that has already been used for one year, then its life-length X lives in $(1,c)$ of probability 95%, solving for c.

$$0.95 = \int_1^c e^{-(x-1)}\,dx = e(e^{-1} - e^{-c}),$$

$$c = -\ln\left(e^{-1} - 0.95e^{-1}\right) = 3.9957.$$

Compared with Q1, we can more directly learn the difference between “inferior” machines and “superior” machines.

Conclusion:

Statistics can really tell future; however, with things changing so fast, the statistical model should be improved over time. This is an era of big data, boom of information, and the grail holding the statistics will not tarnish.

References:

https://media.licdn.com/mpr/mpr/AAEAAQAAAAAAAAitAAAAJDAzZTkwNjZjLWFjNjktNDEwNy04YTA3LTJiZWEyMmMxNmFkOA.jpg

Dodge, Y. (2006) The Oxford Dictionary of Statistical Terms, OUP. ISBN 0-19-920613-9

https://media.licdn.com/mpr/mpr/shrinknp_800_800/AAEAAQAAAAAAAAN0AAAAJDdlYmQ5NTlmLTFiMWEtNDhiZS1hMmExLTFmNGE0NjBmZDVhNg.jpg

Mann, Prem S. (1995). Introductory Statistics (2nd ed.). Wiley. ISBN 0-471-31009-3.

Chance, Beth L.; Rossman, Allan J. (2005). "Preface". Investigating Statistical Concepts, Applications, and Methods

http://www.mymarketresearchmethods.com/wp-content/uploads/2011/11/Descriptive-Statistics.jpg

Willcox, Walter (1938) The Founder of Statistics. Review of the International Statistical Institute.

https://www.mathworks.com/help/finance/life_table_1.png

Normal Distribution, Gale Encyclopedia of Psychology

Anders Hald (1998). A History of Mathematical Statistics from 1750 to 1930. New York: Wiley.

Probability and Statistics The Science of Uncertainty 2nd Edition

"Science in a Complex World - Big Data: Opportunity or Threat?". Santa Fe Institute.