1. INTRODUCTION

1.1 What is Statistics?

STATISTICS: a subject dealing with the **collection** and the **use** of data in order to (help to) answer the questions like:

- Do stock markets rise and fall randomly?
- Is global warming really happening?
- Does a certain new drug prolong life for AIDS sufferers?
- Are GCSE and A level examinations standards declining?
- Is the national lottery making us a nation of compulsive gamblers?
- Is the gap between rich and poor widening in Britain?
- Do Persil adverts really make us want to buy Persil?
- .....

Those questions are difficult to study in laboratory, and admit no self-evident axioms

Typically *Data* are subject to **uncertainty** — *random variables*

Statistics: **Let the Data Speak**!

- making guesses about the process generating the data – *Estimation*
- testing the guesses – *Testing hypotheses*
- forecasting future – *Prediction*

Note. Statistics also deals with *experiment design* which will be untouched in this course.

Statistics offers quantitative approaches to solve practical problems. Therefore, **a good knowledge** on the practical world behind data is essential for decent statistics practice.

No unique statistical solution for most practical problems! Therefore

Statistics is also an art

Some guidelines for learning/applying statistics:

- Understand what data say in each specific context. All the methods are just tools to help understand data
- Concentrate on what to do and why, rather than concrete calculation and graphing
- It may take a while to catch the essences of statistics – Keep thinking!

What is the difference between *probability* **and** *statistical inference***?**

Probability — a mathematical subject

Statistics — an applied oriented subject: inference based on data plus 'assumptions'

Note. The assumptions imposed in statistical inference are typically based on probability theory.

Consider a simple experiment: a coin is tossed n times.

Define a random variable X = the number of 'heads'.

Probability Questions:

- What is $P(X = 1)$, $P(X \leq 1)$, $P(X = x)$ or $P(X \leq x)$?
- What is $E(X)$, $\mathrm{Var}(X)$, $E(X^m)$ or $E(e^{tX})$?

Probability Answers

If we **assume** that the tosses are independent with a constant probability of success π, then

- $P(X = 1) = n\pi(1-\pi)^{n-1}$,
  $P(X \leq 1) = (1-\pi)^{n-1}(1-\pi+n\pi)$,
  $P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$, etc
- $E(X) = n\pi$, $\mathrm{Var}(X) = n\pi(1-\pi)$, etc

Statistical Questions

- What is the value of π?
- Is there any evidence that the coin is 'biased' i.e. $\pi \neq 0.5$?
- Are the tosses independent?

Statistical Answers

- Estimate π by the estimator x/n. Give an interval in which π may be expected to fall.
- Reject the Null Hypothesis that $\pi = 0.5$ if x is too far away from $n/2$
- Reject the model of independent trials if there is too much regularity in the observations - e.g. HTHTHTHTHTHTHTHT....

1.2 Terminology and Notation

The classic statistical inference is based on a **sample** (i.e. a set of **data**)

$$\mathbf{X} = (X_1, \cdots, X_n)^T$$

and the **assumption** that $\{X_i\}$ i.i.d. with

$$X_i \sim F_X(x;\theta), \quad \theta \in \Theta,$$

where $F_X(\cdot;\theta)$ is a probability distribution and is known up to an *unknown* **parameter** θ, Θ is the **parameter space**.

Population: $\{F_X(\cdot;\theta),\ \theta \in \Theta\}$, which is assumed to contain the **true** distribution of X_i

Sample space: consisting of all possible values of $\mathbf{X}$

Statistic: any (known) function of $\mathbf{X}$ only

Dual identity of a sample:

- a set of observed real numbers in practice, and
- a set of i.i.d. r.v.s in theoretical exploration.

Goal of statistical inference in this course: estimating the true value of θ or testing the hypotheses about θ

Note. A sample or a random sample refers to a set of i.i.d. observations.

Example 1. Let $X_1, \cdots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$. Then $(X_1, \cdots, X_n)$ is a sample, R^n is the sample space, and

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \qquad \Theta = R \times R^+.$$

The population is

$$\{N(\mu, \sigma^2),\ (\mu, \sigma^2) \in \Theta\}.$$

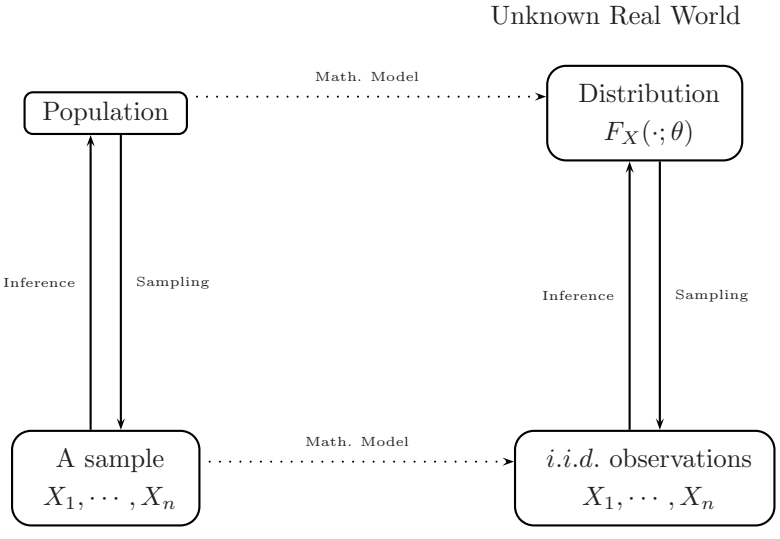
Both sample mean $\bar{X} = n^{-1} \sum_{1 \le i \le n} X_i$ and sample variance $S = \frac{1}{n-1} \sum_{1 \le i \le n} (X_i - \bar{X})^2$ are statistics. But $\bar{X} - \mu$ is not!

Note. The above setting is referred as a **parametric model** since the population is known up to unknown parameters.

In contrast, a **nonparametric model** specifies that the true distribution belongs to a class which cannot be specified by a finite number of parameters only. For example, a population may be

$$\{\text{all continuous density functions in one variable}\}.$$

We deal with parametric models first.



1.3 Statistics from Normal samples

In this section we always assume that $\mathbf{X} = (X_1, \cdots, X_n)^\tau$ be a random sample from $N(0,1)$.

Then the two summary statistics are sample mean and sample variance:

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_i - \bar{X})^2.$$

Theorem 1. The following statements hold.

(i) $\bar{X} \sim N(0, 1/n)$,
(ii) $(n-1)S^2 \sim \chi^2_{n-1}$,
(iii) $\bar{X}$ and S^2 are independent.

Note. The distribution of $\sum_{1\le j\le n} X_i^2$ is called **the χ^2-square distribution with n degrees of freedom**, denoted as χ^2_n or $\chi^2(n)$.

Proof. We make a linear transformation from $X_1, \ldots, X_n$ to $\mathbf{Z} = (Z_1, \ldots, Z_n)^\tau$ in such a way that the Z_i's are still i.i.d. $N(0,1)$. This can be achieved by any transformation $\mathbf{Z} = \mathbf{HX}$ with $\mathbf{H}$ being a $n \times n$ orthogonal matrix (i.e. $\mathbf{H}^\tau\mathbf{H} = \mathbf{I}_n$). Let the first row of $\mathbf{H}$ be

$$n^{-1/2}(1, 1, \ldots, 1).$$

Then $Z_1 = \sqrt{n}\bar{X}$. Hence (i) holds. Since

$$\sum_{j=1}^{n} Z_j^2 = \mathbf{X}^\tau\mathbf{H}^\tau\mathbf{H}\mathbf{X} = \mathbf{X}^\tau\mathbf{X} = \sum_{j=1}^{n} X_j^2,$$

it holds that

$$\sum_{j=2}^{n} Z_j^2 = \sum_{j=1}^{n} X_j^2 - Z_1^2 = \sum_{j=1}^{n} X_j^2 - n\bar{X}^2 = (n-1)S^2.$$

Therefore (ii) and (iii) also hold.

Remark. The above proof indicates the following decomposition

$$\begin{array}{ccccc} \sum_{i=1}^n X_i^2 & = & \sum_{i=1}^n (X_i - \bar{X})^2 & + & n\bar{X}^2 \\ \chi^2_n & & \chi^2_{(n-1)} & & \chi^2_{(1)}, \end{array}$$

where the two χ^2 random variables on the RHS are independent.

Example 1. Let $Y_1, \cdots, Y_n$ iid from $N(\mu, \sigma^2)$. Define

$$X_i = (Y_i - \mu)/\sigma.$$

Then $\{X_i\}$ i.i.d. $N(0,1)$, and

$$\bar{Y} = \sigma\bar{X} + \mu, \quad S_y^2 = \sigma^2 S_x^2.$$

Hence

(i) $\bar{Y} \sim N(\mu, \sigma^2/n)$,
(ii) $(n-1)S_y^2/\sigma^2 \sim \chi^2_{n-1}$,
(iii) $\bar{Y}$ and S_y^2 are independent.

Further,

$$\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{\sigma^2} + n\frac{(\bar{Y} - \mu)^2}{\sigma^2}.$$

F distribution

Definition. Let $Y \sim \chi^2_n$ and $X \sim \chi^2_m$, and X and Y are independent. Then the random variable

$$U \equiv \frac{X/m}{Y/n} = \frac{n}{m}\frac{X}{Y}$$

has an F-distribution with degrees of freedom (m, n). We write $U \sim F(m, n)$ or $U \sim F_{m,n}$.

Obviously, $1/U \sim F_{n,m}$.

A formal notation: $F_{m,n} = \frac{\chi^2_m/m}{\chi^2_n/n}$.

Typical use of F distribution

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two indep samples from $N(\mu_x, \sigma^2_x)$ and $N(\mu_y, \sigma^2_y)$ respectively. Then

$$\frac{\frac{1}{(m-1)\sigma^2_x}\sum_{i=1}^{m}(X_i - \bar{X})^2}{\frac{1}{(n-1)\sigma^2_y}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \sim F_{(m-1),(n-1)}.$$

Student's t-distribution

Definition. Let $Z \sim N(0, 1)$ and $U \sim \chi^2_{(k)}$, and Z and U be independent. Then the random variable

$$T = \frac{Z}{[\frac{1}{k}U]^{\frac{1}{2}}},$$

has a t-distribution with k degrees of freedom. We write $T \sim t_k$ or $T \sim t(k)$.

Obviously, $T^2 \sim F_{1,k}$.

Also T and $-T$ have the same distribution, i.e. the pdf is symmetric. In fact $ET = 0$ for $k > 1$.

A formal notation: $t_k = \frac{N(0,1)}{\{\chi^2_k/k\}^{1/2}}$.

Typical use of t distribution

If $Y_1, Y_2, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, then

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

and

$$(n-1)S^2_y/\sigma^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

and these two are independent. It follows that

$$\frac{\bar{Y} - \mu}{\sqrt{S^2_y/n}} \sim t_{n-1}.$$