Chapter 7. Introduction to Statistical Inference

7.1. What is Statistics: a scientific subject on collecting and analyzing data.

collecting: designing experiments/questionnaires, designing sampling schemes, administration of data collection

analyzing: estimation, testing and forecasting

Statistics is an application-oriented subject, is particularly useful or helpful in answering questions such as:

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

What to learn in Statistics: **basic ideas**, methods (including computation) and theory.

Some guidelines for learning/applying statistics:

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

7.2 Population, Sample and Parametric Models

Two practical situations:

- A new type of tyre was designed to increase the lifetime. The manufacturer tested 120 new tyres and obtained the average lifetime (over those 120 tyres) 35,391 miles. So it claims that the mean lifetime of the new tyres is 35,391 miles.

- A newspaper sampled 1000 potential voters, and 350 of them were Democratic party supporters. It claims that the proportion of the Democratic voters in the whole Country is 350/1000=35%.

In both cases, the conclusion is drawn on a **population** (i.e. all the objects concerned) based on the information from a **sample** (i.e. a subset of population).

In the first case, it is impossible to measure the whole population. For the second case, it is not economic to measure the whole population. Therefore, **errors are inevitable**!

Population is an entire set of the objects concerned, and those objects are typically represented by some numbers. We do not know the entire population in practice.

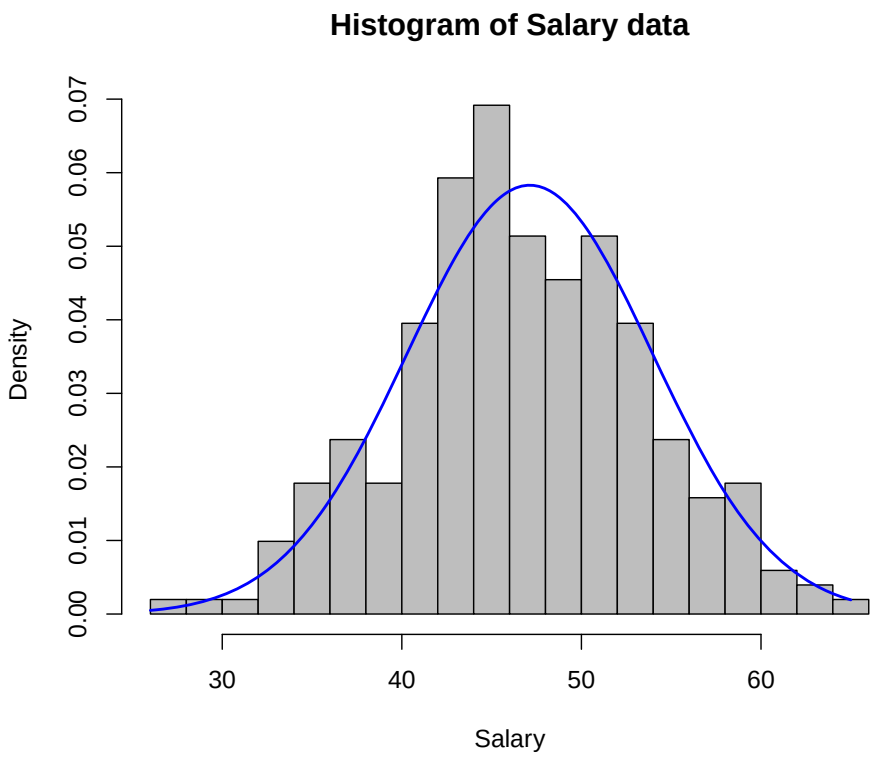
For the tyre example, the population consists of the lifetimes of all the tyres, including those to be produced in the future.

For the opinion poll examples, the population consists of many '1' and '0', where each '1' represents a voter for Democratic party, and each '0' represents a voter for other parties.

A **Sample** is a (randomly) selected subset of a population, and is a set of known data in practice.

Population is unknown. We represent a population by a probability distribution.

```
> jobs <- read.table("Jobs.txt", header=T, row.name=1)
> dim(jobs)
[1] 253   5
> mean(jobs[,4]); var(jobs[,4])
[1] 47.12648
[1] 46.83315
> hist(jobs[,4], probability=T, nclass=15, xlab="Salary",
            main="Histogram of Salary data")
> range(jobs[,4])
[1] 26 65
> x <- seq(26, 65, 0.1)
> lines(x, dnorm(x, 47.12648, sqrt(46.83315)))
      # superimpose the PDF of N(47.12648, 46.83315)
```



The blue curve is the PDF of $N(\bar{X}_n, S_n^2)$.

$n = 253$, $\bar{X}_n = 47.126$, and $S_n^2 = 46.833$, $S_n = 6.843$.

$\bar{X}_n \pm S_n = (40.283,\ 53.969)$

171 points are in this interval:

$171/253 = 67.58\%$.

$\bar{X}_n \pm 1.96 S_n = (33.714,\ 60.538)$

242 points are in this interval:

$242/253 = 95.65\%$

Suggesting $N(\bar{X}_n, S_n^2)$!!!

Parametric Models. For a given problem, we typically assume a population to be a probability distribution $F(\cdot; \theta)$, where the form of distribution F is known (such as normal, Poisson etc), and θ denotes some unknown characteristics (such as mean, variance etc) and is called a parameter. Such an assumed distribution is often called a parametric model.

For the tyre lifetime example, the population may be assumed to be $N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$, where μ is the 'true' lifetime. Let

$$X = \text{the lifetime of a tyre.}$$

Then $X \sim N(\mu, \sigma^2)$.

For the opinion poll example, the population is a Bernoulli distribution:

$$P(X = 1) = P(\text{ a Democratic voter }) = \pi,$$

$$P(X = 0) = P(\text{ a Republican voter }) = 1 - \pi,$$

where

π = the proportion of Democratic supporters in the USA
= the probability of a voter to be a Democratic supporter

A Sample: a set of data or random variables? – A Duality

A sample of size n: $\{X_1, \cdots, X_n\}$, is also called a random sample. It consists of n concrete numbers in a practical problem.

The word 'random' captures the character that samples (of the same size) taken by different people or at different times may be different, as they are different subsets of a population.

Furthermore, a sample is also viewed as n independent and identically distributed (i.i.d.) random variables, when we assess the performance of a statistical method.

For the tyre lifetime example, the sample (of size $n = 120$) used gives the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = 35,391$$

A different sample may give a different sample mean, say, 36,721.

Question: Is the sample mean $\bar{X}$ a good estimator for the unknown 'true' lifetime μ?

Obviously we cannot use the concrete number 35,391 to assess how good this estimator is, as a different sample may give a different average value, say, 36,721.

Key idea: By treating $X_1, \cdots, X_n$ as random variables, $\bar{X}$ is also a random variable. If the distribution of $\bar{X}$ concentrates closely around (unknown) μ, $\bar{X}$ is a good estimator for μ.

Statistic. Any known function of a random sample is called a statistic.

Statistic is used for statistical inference such as estimation, testing etc.

Example. Let $X_1, \cdots, X_n$ be a sample from population $N(\mu, \sigma^2)$. Then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad X_1 + X_n^2, \quad \sin(X_3) + 6$$

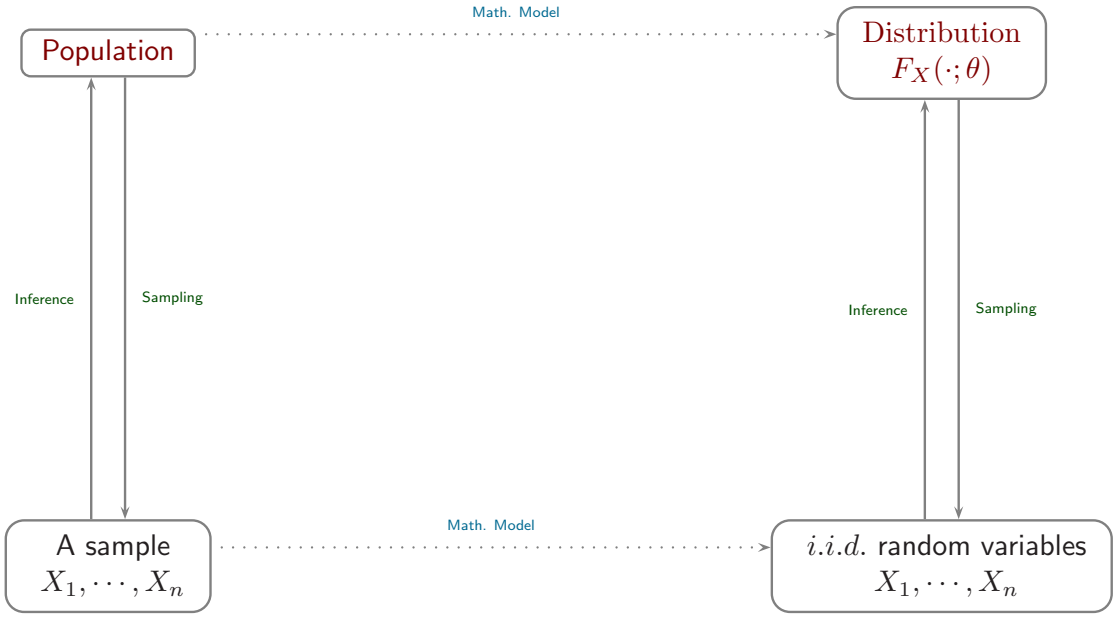
are all statistics. But

$$(X_1 - \mu)/\sigma$$

is not a statistic, as it depends on unknown quantities μ and σ^2.

Note. It is often to denote a random sample as $x_1, \cdots, x_n$, indicating they are n concrete numbers. They are seen as a realization or an instance of n i.i.d. random variables $X_1, \cdots, X_n$. But we do not make this difference, as it makes statements laboursome from time to time.

Unknown Real World



Known Information/Data

θ is called a **parameter**.

A known function of $X_1, \cdots, X_n$ is called a **statistic**.

Difference between Probability and Statistics

Probability: a mathematical subject

Statistics: an application oriented subject (which uses Probability heavily)

Example. Let

$$X = \text{No. of StatsI lectures attended by a student}$$

Then $X \sim \text{Bin}(17, p)$, i.e.

$$P(X = k) = \frac{17!}{k!(17-k)!} p^k (1-p)^{17-k}, \qquad k = 0, 1, \cdots, 17.$$

Probability questions: treating p as known

- what is E(X)? (the average lectures attended)
- what is $P(X \geq 14)$? (the proportion of the students attending at least 14 lectures)
- what is $P(X \leq 8)$? (the proportion of the students attending fewer than the half of lectures)

Statistics questions:

- what is p? (the average attendance rate)
- Is p not smaller than 0.9?
- Is p smaller than 0.5?

7.3 Fundamental concepts in statistical inference

Let $X_1, \cdots, X_n$ be a sample from a population $F(\cdot, \theta)$. Most inference problems can be identified as one of the three types: *(point) estimation, confidence sets* and *hypothesis testing* for parameter θ.

7.3.1 Point estimation: Provide a single "best guess" of θ, based on observations $X_1, \cdots, X_n$. Formally we may write

$$\widehat{\theta} \equiv \widehat{\theta}_n = g(X_1, \cdots, X_n)$$

as a point estimator for θ, where $g(X_1, \cdots, X_n)$ is a statistic.

For example, a natural point estimator for the mean $\mu = EX_1$ is the sample mean $\widehat{\mu} = \bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

Remark. Parameters to be estimated are unknown constants. Their estimators are viewed as r.v.s, although in practice $\widehat{\theta}$, $\widehat{\mu}$ admit some concrete values.

A good estimator should make $|\widehat{\theta} - \theta|$ as small as possible. However

(i) θ is unknown,
(ii) the value of $\widehat{\theta}$ changes with the sample observed.

Hence we seek for an estimator $\widehat{\theta}$ which makes the MSE as small as possible **for all possible values of** θ.

The **mean square error** of the estimator $\widehat{\theta}$ is defined as

$$\text{MSE}_\theta(\widehat{\theta}) = E_\theta\{(\widehat{\theta} - \theta)^2\} = \{\text{Bias}_\theta(\widehat{\theta})\}^2 + \text{Var}_\theta(\widehat{\theta}), \tag{1}$$

where $\text{Bias}_\theta(\widehat{\theta}) = E_\theta(\widehat{\theta}) - \theta$ is called the bias.

When $\text{Bias}_\theta(\widehat{\theta}) = 0$ for all possible values of θ, $\widehat{\theta}$ is called an *unbiased estimator*.

Note. The subscript 'θ' in E_θ etc indicates that the expectation etc are taken with θ being the true value.

The standard error of the estimator $\widehat{\theta}$: $\text{SE}(\widehat{\theta}) = \{\text{Var}_{\widehat{\theta}}(\widehat{\theta})\}^{1/2}$

Note. The standard deviation of $\widehat{\theta}$ $\{\text{Var}_{\theta}(\widehat{\theta})\}^{1/2}$ may depend on the unknown θ. The standard error $\text{SE}(\widehat{\theta})$ is known, and is an estimator for the standard deviation of $\widehat{\theta}$.

Example 1. Let $Y_1, \cdots, Y_n$ be a sample from Bernoulli(p). Let $\widehat{p} \equiv \widehat{p}_n = \bar{Y}_n = \sum_i Y_i / n$. Then

$$E(\widehat{p}) = \frac{1}{n}\sum_{i=1}^{n} EY_i = p, \quad \text{Var}(\widehat{p}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(Y_i) = \frac{p(1-p)}{n}.$$

Therefore $\widehat{p}_n$ is an unbiased estimator for p with the standard deviation $\sqrt{p(1-p)/n}$, and $\text{SE}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

For example, if $n = 10$ and $\bar{Y}_n = 0.3$, we have $\widehat{p} = 0.3$ and $\text{SE}(\widehat{p}) = 0.1449$ while the standard deviation of $\widehat{p}$ is $\sqrt{p(1-p)/10}$ unknown.

Proof of (1).

$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= E[\{(\widehat{\theta} - E\widehat{\theta}) + (E\widehat{\theta} - \theta)\}^2] \\
&= E\{(\widehat{\theta} - E\widehat{\theta})^2\} + (E\widehat{\theta} - \theta)^2 + 2(E\widehat{\theta} - \theta)E(\widehat{\theta} - E\widehat{\theta}) \\
&= \text{Var}(\widehat{\theta}) + \text{Bias}(\widehat{\theta})^2 + 0.
\end{aligned}$$

Consistency. $\widehat{\theta}_n$ is a consistent estimator for θ if $\widehat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$.

The consistency is a natural condition for a reasonable estimator, as $\widehat{\theta}_n$ should converge to θ if we have infinity amount of information. Therefore a non-consistent estimator should not be used in practice!

If $\text{MSE}(\widehat{\theta}_n) \to 0$, $\widehat{\theta} \xrightarrow{m.s.} \theta$. Therefore $\widehat{\theta} \xrightarrow{P} \theta$, i.e. $\widehat{\theta}$ is a consistent estimator for θ.

In Example 1 above, $\text{MSE}(\widehat{p}) = \text{Var}(\widehat{p}) = p(1-p)/n \to 0$. Hence $\widehat{p}$ is consistent.

Asymptotic Normality. An estimator $\widehat{\theta}_n$ is asymptotically normal if

$$(\widehat{\theta} - \theta)/\mathrm{SE}(\widehat{\theta}) \xrightarrow{D} N(0, 1).$$

Remark. Many good estimators such as MLE, LSE and MME are asymptotically normal under some mild conditions, such as finite moments and smooth likelihood function (as function of parameters)

7.3.2 Confidence sets

A point estimator is simple to construct and to use. But it is not very informative. For example, it does not reflect the uncertainty in the estimation.

Confidence Interval is the most commonly used confidence set, is more informative than a point estimator.

Example 2. Let us start with a simple example. A random sample $X_1, \cdots, X_n$ are drawn from $N(\mu, 1)$. Then $\sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$. Hence

$$P(-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96) = 0.95,$$

or

$$P(\bar{X} - 1.96/\sqrt{n} < \mu < \bar{X} + 1.96/\sqrt{n}) = 0.95.$$

So a 95% confidence interval for μ is

$$(\bar{X} - 1.96/\sqrt{n},\ \bar{X} + 1.96/\sqrt{n}).$$

Suppose $n = 4$, $\bar{X} = 2.25$. Then a 95% C.I. is $(2.25 - 0.98, 2.25 + 0.98) = (1.27, 3.23)$.

Question: what is $P(1.27 < \mu < 3.23)$? — Note μ is a unknown constant!

Answer: $(1.27, 3.23)$ is one instance of the **random interval** $(\bar{X} - 0.98,\ \bar{X} + 0.98)$ which covers μ with probability 0.95.

If one draw 10,000 samples, with size $n = 4$ each, to construct 10,000 intervals of the form $(\bar{X} - 0.98,\ \bar{X} + 0.98)$, about 9,500 intervals cover the true value of μ.

Definition. If $L \equiv L(X_1, \cdots, X_n)$ and $U \equiv U(X_1, \cdots, X_n)$ are two statistics for which

$$P\{L(X_1, \cdots, X_n) < \theta < U(X_1, \cdots, X_n)\} = 1 - \alpha,$$

$(L,\ U)$ is called a $100(1-\alpha)\%$ *confidence interval* for θ.

Remark. $1 - \alpha$ is called the confidence level, which is usually set at 0.90, 0.95 or 0.99. Naturally for given α, we shall search for the interval with the shortest length $U - L$, which gives the most accurate estimation.

Approximate confidence interval based on an asymptotically normal estimator: If $(\widehat{\theta} - \theta)/\mathrm{SE}(\widehat{\theta}) \xrightarrow{D} N(0,1)$. Then $\widehat{\theta} \pm Z_{\alpha/2}\mathrm{SE}(\widehat{\theta})$ is an approximate $1-\alpha$ confidence interval for θ, where Z_α is the top-α point of $N(0,1)$, i.e. $P\{N(0,1) > Z_\alpha\} = \alpha$.

For $\alpha = 0.05,\ Z_{\alpha/2} = 1.96 \approx 2$, one of the most used 95% confidence interval is

$$\widehat{\theta} \pm 2 \times \mathrm{SE}(\widehat{\theta}) = \left(\widehat{\theta} - 2 \times \mathrm{SE}(\widehat{\theta}),\ \widehat{\theta} + 2 \times \mathrm{SE}(\widehat{\theta})\right).$$

Example 1 (continue). Let $Y_1, \cdots, Y_n$ be a sample from Bernoulli(p). Let $\widehat{p} \equiv \widehat{p}_n = \bar{Y}_n = \sum_i Y_i / n$. By the CLT, $(\widehat{p}_n - p)/\mathrm{SE}(\widehat{p}_n) \sim N(0,1)$ asymptotically. Hence an approximate $1-\alpha$ confidence interval for p is

$$\widehat{p}_n \pm Z_{\alpha/2}\mathrm{SE}(\widehat{p}_n) = \widehat{p}_n \pm Z_{\alpha/2}\sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}.$$

For example, if $n = 10,\ \widehat{p}_n = 0.3$, an approximate 95% confidence interval for p is

$$0.3 \pm 2\sqrt{0.3(1-0.3)/10} = 0.3 \pm 0.145 = (0.155,\ 0.445).$$

However if now $n = 100$ and $\widehat{p}_n = 0.3$, an approximate 95% confidence interval for p is

$$0.3 \pm 2\sqrt{0.3(1-0.3)/100} = 0.3 \pm 0.046 = (0.254,\ 0.346).$$

Remark. The point estimator $\widehat{p}_n$ unchanged with $n = 10$ or 100. However the confidence interval is much shorter when $n = 100$, giving much more accurate estimation.

7.3.3 Hypothesis testing: We start with some default statement – called a null hypothesis denoted by H_0. We ask if the date provide significant evidence to reject the null hypothesis.

For example we may test if a coin is fair by using the hypothesis $H_0 : p = 0.5$.

Remark. (i) Estimation and testing address different needs in practice.

(ii) A statistical test often takes binary decision: '*reject* H_0' or '*not reject* H_0'. However technically a testing problem is more complex that an estimation problem.

7.4 Nonparametric models and Empirical distribution functions

The outcome of a statistical inference depends on two factors: *data* and *assumption*.

The data are objective, while the assumption is more subjective. We would like to let *data speak* as much as possible.

The classical statistical inference is typically based on an assumption of a parametric model: $X_1, \cdots, X_n$ is a sample from the distribution $F(\cdot, \theta)$, where the form of the distribution F is known (such as Normal, Exponential etc), and the parameter θ is unknown. The inference is on either estimation or testing of parameter θ.

Nonparametric model: let $X_1, \cdots, X_n$ is a sample from a distribution F belong to a class of distributions $\mathcal{F}$. For example, $\mathcal{F}$ may consist of all continuous distributions on $(-\infty, \infty)$. The statistical inference is either to estimate or to test some characteristics of F, or F itself.

Parametric models limit the tasks of statistical inference. It facilitates more efficient inference **if** the assume parametric form is correct. Nonparametric models impose less model-bias, however its statistical inference is more challenging.

Empirical Distribution Functions. Let $X_1, \cdots, X_n$ be a sample from CDF F. A natural estimator for F is defined as

$$\widehat{F}(x) = \frac{\text{No. of } X_i's \text{ not greater than } x}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x),$$

where $I(A) = 1$ if A occurs, and o otherwise. $\widehat{F}(\cdot)$ is a well-defined CDF, and is called the empirical distribution function.

Note $I(X_i \leq x)$ is a sequence of Bernoulli r.v.s with $p = F(x)$. Hence,

$$E\{\widehat{F}(x)\} = F(x), \quad \text{Var}\{\widehat{F}(x)\} = F(x)\{1 - F(x)\}/n,$$

and $\widehat{F}(x) \xrightarrow{m.s.} F(x)$. In fact it also holds that

$$\sup_x |\widehat{F}(x) - F(x)| \xrightarrow{P} 0, \quad P\{\sup_x |\widehat{F}(x) - F(x)| > \epsilon\} \leq 2e^{-2n\epsilon^2}.$$