STAT3001/STATM012/STATG012

Statistical Inference

Course Lecturer: Maria De Iorio

m.deiorio@ucl.ac.uk

Department of Statistical Science

Session 2015-16

Aim of the Course

Statistical inference means drawing conclusions about a population based on data. There are many contexts in which inference is desirable, and there are many approaches to performing inference.

Data (and statistical inference) play a major part in many fields: finance, medicine, epidemiology, economics, accounting, actuarial science, atmospheric science, biology, education, environmental science, epidemiology genetics, manufacturing, marketing, pharmaceutical industry, sport, psychology, sociology and on and on ...

This course is designed to provide a "clear" introduction to philosophy and methods of statistical inference.

Two main paradigms of statistical reasoning will be covered:

- Frequentist approach
- Bayesian approach

Prerequisites....

You need to know the basics:

- **how to differentiate...**

$$\frac{\partial}{\partial p} \log \left(\frac{p}{1-p} \right)^n = \qquad ?$$

Prerequisites....

You need to know the basics:

- **how to differentiate...**

$$\frac{\partial}{\partial p} \log \left(\frac{p}{1-p} \right)^n = \frac{n}{p} + \frac{n}{1-p}$$

Prerequisites....

You need to know the basics:

- how to differentiate...

$$\frac{\partial}{\partial p}\log\left(\frac{p}{1-p}\right)^{n} = \frac{n}{p} + \frac{n}{1-p}$$

- how to integrate....

$$\int_0^\infty x^{\alpha-1}e^{-\beta x}\mathbf{d}x = \qquad ?$$

Prerequisites....

You need to know the basics:

- how to differentiate...

$$\frac{\partial}{\partial p}\log\left(\frac{p}{1-p}\right)^{n} = \frac{n}{p} + \frac{n}{1-p}$$

- how to integrate....

$$\int_0^\infty x^{\alpha-1}e^{-\beta x}\mathbf{d}x = \frac{\Gamma(\alpha)}{\beta^{\alpha}}$$

Prerequisites....

You need to know the basics:

- how to differentiate...

$$\frac{\partial}{\partial p}\log\left(\frac{p}{1-p}\right)^{n} = \frac{n}{p} + \frac{n}{1-p}$$

- how to integrate....

$$\int_0^\infty x^{\alpha-1}e^{-\beta x}\mathbf{d}x = \frac{\Gamma(\alpha)}{\beta^{\alpha}}$$

- basic probability....., conditional probabilities.....

$$\mathbf{Var}(aX + bY) = \qquad ?$$

Prerequisites....

You need to know the basics:

- how to differentiate...

$$\frac{\partial}{\partial p}\log\left(\frac{p}{1-p}\right)^n = \frac{n}{p} + \frac{n}{1-p}$$

- how to integrate....

$$\int_0^\infty x^{\alpha-1}e^{-\beta x}\mathbf{d}x = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

- basic probability.....

$$\mathsf{Var}(aX + bY) = a^2\mathsf{Var}(X) + b^2\mathsf{Var}(Y) + 2ab\mathsf{Cov}(X, Y)$$

Prerequisites....

You need to know the basics:

- how to differentiate...

$$\frac{\partial}{\partial p}\log\left(\frac{p}{1-p}\right)^{n} = \frac{n}{p} + \frac{n}{1-p}$$

- how to integrate....

$$\int_0^\infty x^{\alpha-1}e^{-\beta x}\mathbf{d}x = \frac{\Gamma(\alpha)}{\beta^{\alpha}}$$

- basic probability.....

$$\mathsf{Var}(aX+bY) = a^2\mathsf{Var}(X) + b^2\mathsf{Var}(Y) + 2ab\mathsf{Cov}(X,Y)$$

- Probability distribution functions.....Conditional probabilities.....

1 Course Information

- No lecture notes
- Tutorials starts in Week 7 (attendance is compulsory)
- Extra Revision Exercises
- 2 workshops
- ICA IS CLOSED BOOK (will include 1 exercise from tutorials)

What is statistical inference?

Descriptive statistics: **graphs, tables, means, variances** *etc*.

Inferential statistics: **practical** *implications* **of the data, via some use of probability theory (***eg***. tests, confidence intervals, hpd regions).**

Example

$x =$ **amount of fertilizer,** $y =$ **yield of crop,** n **pairs of readings.**

1. Descriptive

What is statistical inference?

Descriptive statistics: **graphs, tables, means, variances** *etc*.

Inferential statistics: **practical** *implications* **of the data, via some use of probability theory (***eg***. tests, confidence intervals, hpd regions).**

Example

$x =$ **amount of fertilizer,** $y =$ **yield of crop,** n **pairs of readings.**

1. Descriptive

Plot y **against** x**; calculate correlation coefficient; fit a straight line** $y = a + bx$ *etc*. **No use of probability theory, no assessment of** *accuracy* **of estimates** *etc*.

2. Inferential

Probability model	Eg. $y_i \sim N(\alpha + \beta x_i, \sigma^2)$
(how data were generated)	& $y_1, \ldots y_n$ independent
$\Downarrow$ (deduction)	
Behaviour of potential data	Joint prob distribution of $y_1, \ldots y_n$ given the model and parameters α, β, σ^2
$\Downarrow$	
Sample data (+ possibly other information)	(induction) $\Rightarrow$ Implications for α, β, σ^2

2. Inferential

Probability model (how data were generated)	Eg. $y_i \sim N(\alpha + \beta x_i, \sigma^2)$ & $y_1, \ldots y_n$ independent
$\Downarrow$ (deduction)	
Behaviour of potential data	Joint prob distribution of $y_1, \ldots y_n$ given the model and parameters α, β, σ^2
$\Downarrow$	
Sample data (+ possibly other information)	(induction) $\Rightarrow$ Implications for α, β, σ^2

Statistical inference is concerned with the final inductive step.

In practice the procedure is an iterative one, since the sample data may lead us to reflect on the model itself, and we cycle through the above stages.

Key to statistical thinking is to understand variation

- variation in the population
- sampling variability

You should not too easily jump from conclusions about samples to conclusions about a population.

Statistical inference involves the selection of a probabilistic model to resemble the process you wish to investigate, investigation of that models behaviour, and interpretation of the results.

Outline of the Course

- Frequentist and Bayesian approaches to statistical inference
- Summary statistics, sampling distributions
- Sufficiency, likelihood, and information
- Asymptotic properties of estimators
- Bayesian inference
- Hypothesis testing
- Likelihood ratio tests, application to linear models

1 Approaches to Statistical Inference

In this course we consider some of the theoretical ideas underlying the methodology of simple statistical inference.

We shall explore and compare some alternative approaches to inference.

1.1 General Set up and definitions

Consider a population composed of individuals (for example, people,animals.....) presenting each a certain characteristic.

Our aim is to say something about this characteristic (its distribution in the population).

In most situations, it is not feasible to enumerate and examine all the elements of the population (e.g. population too large, difficult to enumerate, ...) and measure the characteristic of interest.

Thus, we assume that we have at our disposal only a random sample of size n extracted from the population.

We will denote the random sample with $(X_1, \ldots, X_n)$.

From observations $\{x_1, x_2, \ldots, x_n\}$ collected on our sample, we then want to infer likely values for the characteristic under study.

Recap:

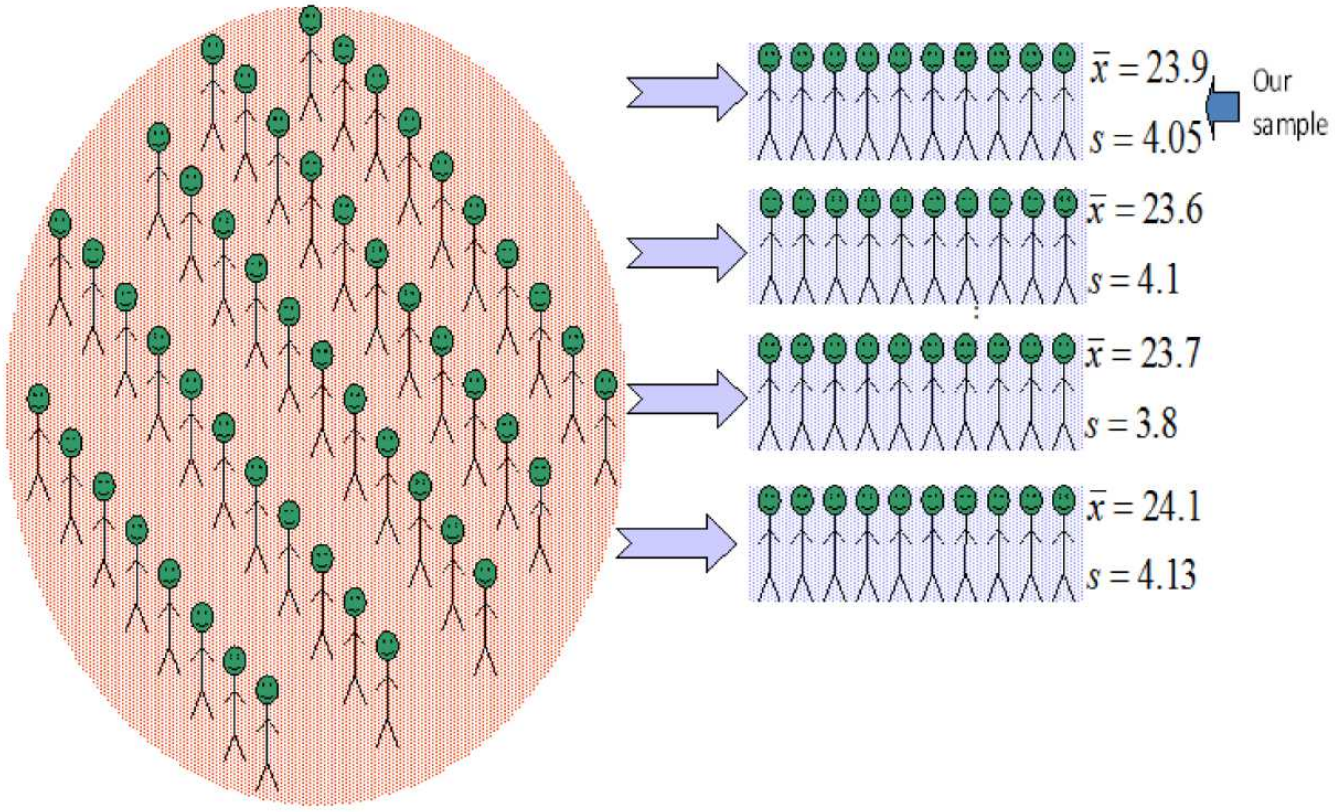
- Population: The totality of subjects under study

- Sample: The subset of the population actually studied.

Of course, the sample has inherent variability: another sample of the same size would not give exactly the same values for the $\{x_i\}'$s, but we can take this into account in our inference.

Sampling variability

If we repeatedly choose samples from the same population, our estimates (e.g. for the the mean) will take different values in different samples



A bit more formalism....

Statistical inference problem:

Data (observations) have been generated in accordance of some unknown probability distribution.

We analyse the data to infer the unknown distribution.

In this course, we assume the distribution that generate the data is completely known except for some values of one or more parameters.

Example: Suppose that the distribution of the weights of the individuals in a certain population is known to be a Normal distribution with mean μ and variance σ^2, but the values of μ and σ^2 are unknown.

μ and σ^2 are the parameters we need to estimate so that the distribution of weight in the population is completely known.

If we observe the weights of the individuals in a random sample selected from the population, then from these observations and any other available information, we can make inference about the values of μ and σ^2.

Some definitions

Let X be a random variable, with distribution dependent on unknown parameters θ:

$$X \sim p_{\theta}(x)$$

Parameter space Θ: set of all possible values of θ.

Let $(X_1, \ldots, X_n)$ be a random sample from X

Let $\{x_1, x_2, \ldots, x_n\}$ be a sample of observations of X.

Random Samples

The random variables $X_1, \ldots, X_n$ are called a random sample of size n from the population $p_\theta(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same $p_\theta(x)$.

We say that $X_1, \ldots, X_n$ are independent and identically distributed (iid) random variables with pdf or pmf $p_\theta(x)$.

- The random sampling model describes an experimental situation in which the variable of interest has a probability distribution described by $p_\theta(x)$

- if we have only one observation X, then the probabilities regarding X can be calculated using $p_\theta(x)$

- Usually $n > 1$. Then each X_i is an observation on the same variable X, with marginal distribution $p_\theta(x)$

- Since the X_i are iid, the joint pdf/pmf of the random sample $X_1, \ldots, X_n$ is

$$p_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(x_i)$$

- The joint pdf/pmf can be used to calculate probabilities involving the sample.

In this course we only consider pdf/pmf that are members of a certain parametric family (e.g. Normal, Exponential, Bernoulli).

Definition

A parametric statistical model consists of the observation of a random variable X, distributed according to $p_\theta(x)$, where only the parameter θ is unknown and belongs to a vector space Θ.

We assume that the population we are observing is a member of a parametric family but the true parameter value θ is unknown,

By considering different possible values of θ, we can study how a random sample would behave for different populations (in our case different values of θ).

Example: Let $X_1, \ldots, X_n$ be a random sample from an Exponential population with parameter λ. For example, the X_i might correspond to failure time of a particular device.
The joint pdf of the sample is

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{\lambda(x_i)} = \prod_{i=1}^{n} \lambda e^{-x_i \lambda} = \lambda^n e^{-\lambda \sum x_i}$$

The joint pdf can be used to answer questions about the sample.

What is the probability that all the devices last more than 2 time units?

What is the probability that all the devices last more than 2 time units?

$$\Pr(X_1 > 2, \cdots, X_n > 2) = \prod_i \Pr(X_i > 2) = \prod_i e^{-\lambda 2} = e^{-2n\lambda}$$

If λ is small compared to n, this probability is close to 1.

You can consider the behaviour of the sample for different λ.

How can we estimate θ using the random sample?

Clearly, our estimate will be a function $\phi(x_1, \ldots, x_n)$ of the sample

The random variable $T_n = \phi(X_1, X_2, \ldots, X_n)$ is called an estimator of θ.

From each observed sample $(x_1, x_2, \ldots, x_n)$, we get an estimate $\hat{\theta}_n$ of θ:

$$\hat{\theta}_n = \phi(x_1, x_2, \ldots, x_n).$$

Example: Estimating the mean

$T_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$ is an estimator of the arithmetic mean

$\hat{\theta}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}$ is its observed value from a sample.

Statistic

Given a random sample $(X_1, \ldots, X_n)$ from $X \sim p_\theta(x)$, any real-valued function $T = T(X_1, \ldots, X_n)$ of the observation in the random sample is called a statistic

Examples: sample mean, sample variance, sample maximum.

In any estimation problem, an estimator of θ is a statistic whose value can regarded as an estimate of the value of θ

Since sample measurements are observed values of random variables, the value of a sample statistic will vary in a random manner from sample to sample.

Important: $T = T(X_1, \ldots, X_n)$ is a random variable as it is a function of $(X_1, \ldots, X_n)$. The distribution of T can be derived from the joint distribution of $(X_1, \ldots, X_n)$ and, in general, it will depend on θ.

The distribution of $T = T(X_1, \ldots, X_n)$ is called sampling distribution

In practice, the sampling ditribution of a statistics is obtained mathematically or by simulations.

Common Sampling Distributions

Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Normal(μ, σ^2) be a random sample from a Normal distribution. Then

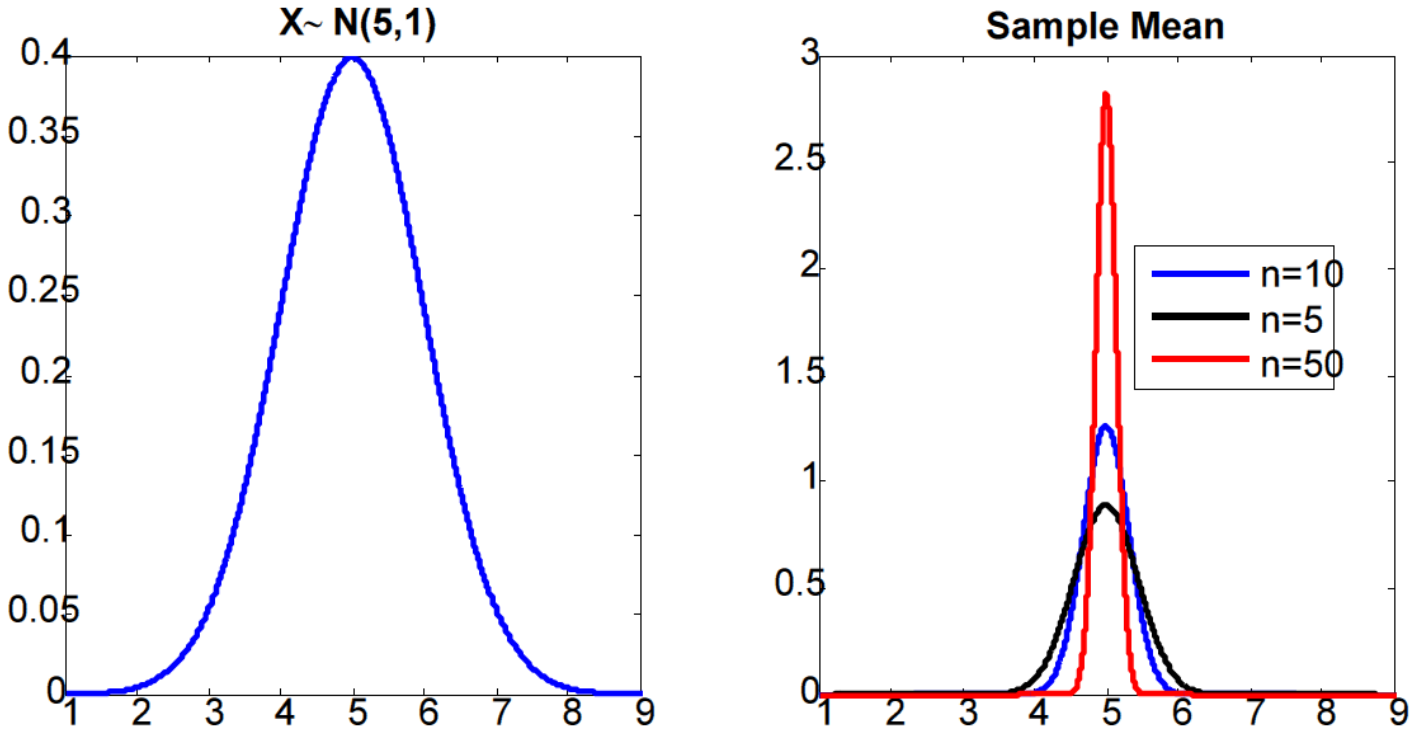
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$\bar{X}$ and S^2 are independent, with $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Distribution of the Sample Mean



The sampling distribution provides information on the behaviour of $\bar{X}$ in repeated sampling. For example, if $n = 10$, the probability of obtaining a sample of 10 obs and obtain a value of $\bar{x}$ in the interval $2 \leq \bar{x} \leq 6$ will be 0.99.

1.2 The Decision Problem

Statistical inference can be thought of in terms of decision theory

Statistical decision theory is an approach to decision-making in the presence of uncertainty. A general decision problem has the following basic elements:

- a parameter space, Θ, consisting of all the possible states of nature, θ ;
- a set, D, of all available decisions (or actions), d ;
- a loss function, $l(\theta, d)$, specifying the loss incurred if a decision d is made when the true state of nature is θ (where a negative loss is a gain).

The true state of nature is unknown.

A statistical decision problem has an additional element:

the statistician can perform an experiment in order to obtain further information about that state.

We assume:

- statistician can collect observation x (generally a vector) $\in$ sample space $\mathcal{X}$,

- $\mathcal{X}$ consists of all possible data values that could be sampled.

- X denotes the *random variable* corresponding to the observed value x.

- The probability distribution (density or mass function), $p(x;\theta)$, of X depends on θ (the true state of nature)

- $\Longrightarrow$ observation x provides information about the true state of nature.

- The statistical decision function, $\delta(x)$, a mapping from the sample space $\mathcal{X}$ to the decision space D, specifies the decision d to be taken for given observed data x.

- In our case, the decision is the estimate of θ and $\delta(X)$ is a statistic.

The decision $\delta(x)$ depends on the data

$$\Downarrow$$

- the loss incurred by taking decision $\delta(X)$ when the true state is θ is a function of the observed data

- $l(\theta, \delta(X))$ is a random variable.

The loss function can be averaged over all possible outcomes of the experiment to give the risk function, $R(\theta, \delta)$:

$$R(\theta, \delta) = \int_{\mathcal{X}} l(\theta, \delta(x))\, p(x; \theta)\ \mathrm{d}x = \mathrm{E}\{l(\theta, \delta(X))\}$$

for continuous X; the integral is replaced by a sum for discrete X.

In general,

- the unknown state of nature can be described by a vector.
- it will be assumed that the distribution of X has a simple parametric form

Examples:

- X_i are independent Poisson variables with common mean μ, and the decision required is an estimate of this mean (thus here $\theta = \mu$).
- the common distribution of each X_i is either Poisson with mean μ or geometric with mean μ. In this case, the unknown state of nature can be described by a vector $\theta = (i, \mu)$, where $i = 1$ if the distribution is Poisson and $i = 2$ if it is geometric.

The structure of the decision space clearly depends on the type of inference to be made:

- For *point estimation* of the parameter θ, D contains the same set of points as the parameter space Θ and each decision corresponds to the choice of a value for the parameter θ.
  $\Longrightarrow$ $\delta(X)$ is an estimator of θ, with corresponding estimate $d = \delta(x)$.

- For the construction of *confidence intervals*, D consists of ordered pairs of points from the parameter space.

- For *hypothesis testing*, the decision space has just two elements, corresponding to acceptance, or not, of the null hypothesis.

An appropriate loss function also needs to be specified:

- In *point estimation* of a single unknown parameter, θ, a quadratic loss function is often used in which $l(\theta, d)$ is taken to be proportional to the square of the estimation error (squared error loss):

$$l(\theta, d) \propto (d - \theta)^2$$

  The corresponding risk function is then the familiar mean squared error (mse$(\delta(X); \theta)$).

- For a vector θ, a weighted sum of squares

$$(d - \theta)^T W (d - \theta)$$

  is often used.

1.2 Criteria for Choosing the Decision Function

- If the true value of the parameter were known (say, $\theta = \theta_0$) then we would simply choose the decision d to minimise the loss $l(\theta_0, d)$.

- For example, if we know $\theta = \theta_0$ then we can estimate θ perfectly so that the mean squared error is zero.

- The inference problem arises because of our uncertainty about the true value of θ and the need to use the information provided by the sampled data x.

- There are then a number of possible approaches to determining $\delta(x)$. We consider three important methods.

The Frequentist Approach

- This approach is also referred to as the classical or sampling theory approach
- The true parameter value θ is regarded as *fixed but unknown*
- The approach involves concepts such unbiased estimators in point estimation and significance levels in hypothesis testing.
- One main characteristic is that the properties of the chosen decision function (e.g. the point estimator, the critical region of the test) depend on the full sample space $\mathcal{X}$.
- For example, if we say that a sample mean, $\overline{X}$, is unbiased for a parameter μ, then this means that the expected value of $\overline{X}$, taken over the whole sample space $\mathcal{X}$ is equal to μ.

The Likelihood Approach

- This approach focuses on the probability of the data actually observed and does not take the rest of the sample space into account.

- It concentrates attention on the sampling distribution $p(x;\theta)$ as a function of θ over the whole parameter space Θ, but only at the single observed data point x.

- The likelihood function $\mathcal{L}(\theta;x)$, or simply $\mathcal{L}(\theta)$, is any function proportional to $p(x;\theta)$, regarded as a function of θ.

The Bayesian Approach

- In many ways this is close to the likelihood approach, as it also focuses on the observed data and ignores unobserved points in the sample space $\mathcal{X}$.

- However, the unknown state of nature, the parameter θ, is regarded as random rather than fixed, and is assigned a probability distribution over the parameter space that reflects a degree of belief about the value of the parameter.

The ideas are straightforward:

- before any data are observed, θ is assumed to have a prior distribution $\pi(\theta)$ which expresses a prior degree of belief about the true value of the parameter before the experiment is performed.

- the data x then provide information about θ that can be used to update the prior degree of belief about the true value of θ

- ... to give a posterior distribution, or degree of belief, $\pi(\theta \mid x)$, via the likelihood.

The mechanism for this updating is Bayes' Theorem and is based on the simple formula:

$$\pi(\theta \mid x) \propto p(x \mid \theta)\, \pi(\theta) = \mathcal{L}(\theta; x)\, \pi(\theta)\,,$$

where we have written $p(x \mid \theta)$ rather than $p(x; \theta)$ to indicate that, in the Bayesian framework where θ is random rather than fixed, this is regarded as a conditional probability distribution.

The symbol $\propto$ in the above equation is to be interpreted as regarding both sides of the equation as functions of θ; in other words, the constant of proportionality will generally be a function of x:

$$\pi(\theta \mid x) = \frac{\mathcal{L}(\theta; x)\, \pi(\theta)}{p(x)}$$

$$p(x) = \int_{\Theta} p(x \mid \theta)\, \pi(\theta) \mathrm{d}\theta = \int_{\Theta} \mathcal{L}(\theta; x)\, \pi(\theta) \mathrm{d}\theta.$$

- The Bayesian method, therefore, combines information about θ from two sources: prior information, represented by $\pi(\theta)$, and information from the data, represented by $\mathcal{L}(\theta; x)$.

- In contrast to the other two methods, it treats the unknown parameter as a random variable rather than a fixed but unknown constant.

- The frequentist and likelihood approaches do not rule out the inclusion of other information about θ but do not have a formal mechanism for doing so.

Definition

A Bayesian statistical model is made of a parametric statistical model, $p(x \mid \theta)$, and a prior on the parameters, $\pi(\theta)$.

In a Bayesian analysis, the risk can be averaged over values of θ in the parameter space to give the **Bayes risk**, $R(\delta)$:

$$\begin{aligned} R(\delta) &= \int_{\Theta} R(\theta, \delta)\pi(\theta)\mathrm{d}\theta \\ &= \int_{\Theta}\int_{\mathcal{X}} l(\theta, \delta(x))\, p(x \mid \theta)\pi(\theta)\mathrm{d}x\mathrm{d}\theta \\ &= \int_{\mathcal{X}} \underbrace{\left\{\int_{\Theta} l(\theta, \delta(x))\, \pi(\theta \mid x)\mathrm{d}\theta\right\}}_{\text{posterior expected loss}} p(x)\mathrm{d}x \end{aligned}$$

$\Longrightarrow$ the Bayes risk can also be interpreted as the average (over the sample space) of the **posterior expected loss** of making decision $\delta(x)$ given the observation x, where the latter expectation is taken with respect to the posterior distribution of the parameter θ given the data x.

Notation: In general, we distinguish the random variable from its value, using capital letters for the former and lower case letters for the latter. Thus X denotes a random variable (a mapping from a (measurable) space to the sample space $\mathcal{X}$).

However, for simplicity we will use θ to denote the state nature, regardless of whether it is fixed or random, and let Θ denote the parameter space.

Example 1.1 This is a simple and somewhat contrived example to emphasize the differences between the above methods.

Suppose we define a random variable R as follows: we toss a fair coin

- if it results in a tail then we set $R = 0$
- if it results in a head then we draw 2 balls from an urn and set R equal to one more than the number of red balls in the sample.
- We assume that the urn has a very large number of balls, a proportion, θ, of which are red.

The probability distribution of R is

r	0	1	2	3
p_r	$\frac{1}{2}$	$\frac{1}{2}(1-\theta)^2$	$\theta(1-\theta)$	$\frac{1}{2}\theta^2$

The problem is to estimate θ from a random sample of 10 observations.

Observations x: six 0s, three 1s and one 2

Frequentist Approach: Unbiased Estimator

We find that

$$\mathsf{E}(R) = \frac{1}{2}(1 + 2\theta)$$

so that $R - \frac{1}{2}$ is an unbiased estimator for θ.

The sample mean is $\frac{1}{2}$ and so the unbiased estimate of θ is 0.

This is clearly not sensible as there is definitely at least 1 red ball in the urn.

It is also clear that the observations of 0 (when the coin gives tails) tell us nothing about the composition of balls in the urn and yet the unbiased estimator makes use of these observations.

Likelihood Approach: Maximum Likelihood Estimator

The likelihood is proportional to the joint probability of the observations:

$$\mathcal{L}(\theta; x) = \left(\frac{1}{2}\right)^6 \left[\frac{1}{2}(1-\theta)^2\right]^3 [\theta(1-\theta)] \propto \theta(1-\theta)^7$$

This is maximised when $\theta = 1/8$ so the maximum likelihood estimate is $1/8 = 0.125$.

Bayesian Approach: Bayes Estimator

We need a prior distribution for θ. The parameter space is $[0, 1]$ and we assume a uniform prior distribution so that $\pi(\theta) = 1, 0 \leq \theta \leq 1$,

We shall take the mean of the posterior distribution as the estimate of θ.

The posterior distribution of θ is proportional to the product of the likelihood and the prior:

$$\pi(\theta \mid x) \propto \theta(1-\theta)^7, \qquad 0 \leq \theta \leq 1$$

We recognise the kernel of Beta$(2, 8)$ distribution. The mean of this distribution is 0.2, and thus the Bayes estimate of θ is 0.2.

Comment: It is clear that when the coin shows tails, the value of R does not provide any information on the proportion of balls in the urn, as the actual data values observed are not dependent on the value of θ. The frequentist approach includes these data whereas the likelihood and Bayesian methods do not. The Bayes estimator takes the prior information on θ into account and, in this particular example, this results in a Bayes estimate that is larger than the maximum likelihood estimate.

If we use the frequentist approach only on those data (4 values) where the experiment gives a head (i.e. conditionally on H), then $R \mid H \sim 1 + \text{Bin}(2, \theta)$ and $\text{E}(R \mid H) = 1 + 2\theta$. The sample mean for R given H is $\frac{3\times 1+2}{4} = 1.25$. Thus an estimate of θ is $\frac{1.25-1}{2} = \frac{1}{8}$, so in this case the frequentist and likelihood approaches give the same estimate.

Review of the most common probability distributions which should be already familiar.

Bernoulli trials

Binomial and Negative Binomial Distribution

Geometric and Hypergeometric Distribution

Multinomial Distribution

Poisson Distribution

Gamma and Exponential Distribution

Beta Distribution

Normal Distribution

χ^2 Distribution

Student's t Distribution

F Distribution

2 Review of Sampling Distributions

Statistical inference rests on the observed outcome of a statistical experiment.

We assume the existence of a random variable, X, with probability distribution $p(x;\theta)$.

In this course we shall not distinguish notationally between discrete and continuous random variables, so that $p(x;\theta)$ will denote either the density of X or its probability mass function, as appropriate.

2.1 Bernoulli trials and some related distributions

A simple statistical experiment might consist of a sequence of independent Bernoulli trials, where θ is the (unknown) common probability of success.

The binomial distribution Let X be the number of successes (1s) obtained in a fixed, known, number n of trials. Then X has the binomial distribution with index n and parameter θ, $0 < \theta < 1$, where

$$\mathbf{P}(X = r) = p(r;\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}, r = 0, 1 \ldots, n$$

and $p(r;\theta) = 0$ otherwise.

The expected value of X is $n\theta$ and the variance is $n\theta(1-\theta)$.

- X can be regarded as the sum of n indicator (zero/one) variables (*Bernoulli* variables)

- By the Central Limit Theorem, as $n \to \infty$,

  the distribution of $(X - n\theta)/\sqrt{n\theta(1-\theta)} \to$ Standard Normal

- Inference procedures for the proportion X/n are usually based on this normal approximation.

- The sum of independent binomial variables having the same value of θ is again binomial.

The negative binomial distribution X represents the number of Bernoulli trials that must be observed until a fixed number k of successes is obtained.

Then X has a negative binomial distribution where

$$\mathsf{P}(X = r) = p(r;\theta) = \binom{r-1}{k-1}\theta^k(1-\theta)^{r-k}, r = k, k+1, \ldots$$

and $p(r;\theta) = 0$ otherwise.

The expected value of X is k/θ, and the variance is $k(1-\theta)/\theta^2$.

- **Note that we may also consider the number Y of *failures* to the kth success.**

  Then $Y = X - k$ takes values in $\{0, 1, \ldots\}$ with probabilities $\mathsf{P}(Y = r; \theta) = \mathsf{P}(X = r + k; \theta)$, so that

$$\mathsf{E}(Y) = \mathsf{E}(X - k) = \frac{k(1-\theta)}{\theta}, \text{ and } \mathsf{var}(Y) = \mathsf{var}(X).$$

  The distribution of Y is also called the negative binomial distribution.

- **The sum of independent negative binomial variables having the same value of θ is again negative binomial (*why is this intuitively clear?*).**

The geometric distribution

This is the special case of the negative binomial distribution with $k = 1$. Thus

$$\mathsf{P}(X = r) = p(r; \theta) = \theta(1 - \theta)^{r-1}, r = 1, 2, \ldots.$$

with $\mathsf{E}(X) = 1/\theta$ and $\mathsf{var}(X) = (1 - \theta)/\theta^2$.

The hypergeometric distribution

Suppose that there a sample of size n is to be randomly chosen without replacement from an urn containing N balls (or items), R of which are white ('success') and $N - R$ are black ('failure').

In this case the probability of a success changes trial by trial, according to which items have already been sampled.

Note that a sequence of Bernoulli trial can be thought of sampling from an infinite population, or to a finite population sampled with replacement.

Let X be the number of successes in the sample. Then X has the hypergeometric distribution, which has two parameters, N and R, and where

$$\mathsf{P}(X = r) = p(r; \theta) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}, r = 0, \ldots, \min(n, R),$$

and $p(r; \theta) = 0$ otherwise.

The expected value of X is $\frac{nR}{N}$ and the variance $n\left(\frac{R}{N}\right)\left(\frac{N-R}{N}\right)\left(\frac{N-n}{N-1}\right)$.

Set $\theta = R/N$ and compare the results for sampling without replacement with those obtained for the binomial distribution (sampling with replacement)

$$\Downarrow$$

- **E$(X) = n\theta$ in both cases,**

- **the hypergeometric distribution has variance $n\theta(1-\theta)(1-n/N)/(1-1/N)$, compared to $n\theta(1-\theta)$for the binomial distribution.**

If θ and n are fixed while N tends infinity (so that R is also increasing to infinity) then we see that the hypergeometric variance tends to that of the binomial distribution.

In fact it is intuitively obvious in this situation that the removal of a single item from the pool makes a negligible difference to the probabilities of selecting the two types of object, and that the hypergeometric distribution tends to the binomial distribution. This can be confirmed algebraically.

The multinomial distribution Suppose that a sequence of n independent and identical experiments is performed and that each experiment can result in any one of m possible outcomes/categories, with

$$\text{P(item is of type } i) = p_i, \quad i = 1, \ldots, m\,,$$

where $\sum_{i=1}^{m} p_i = 1$.

- The parameter θ is a vector, $\theta = (p_1, \ldots, p_m)$
- the parameter space is $\Theta = \{p_i, i = 1, \ldots, m : 0 \leq p_i \leq 1, \sum_{i=1}^{m} p_i = 1\}$.
- Let X_i be the number of type i items in the sample (so that $\sum_{i=1}^{m} X_i = n$).

Then the multinomial distribution with index n and parameters $p_1, \ldots, p_m$ is

$$\mathbf{P}(X_i = n_i, i = 1, \ldots, m) = \begin{cases} \frac{n!}{n_1! \ldots n_m!} p_1^{n_1} \ldots p_m^{n_m}, & n_1 + \ldots + n_m = n \\ 0 & \text{otherwise.} \end{cases}$$

It is really a distribution of $m - 1$ variables, since $\sum_{i=1}^{m} X_i = n$.

The special case when $m = 2$ is just the binomial distribution *i.e.* $X_1 \sim \text{Bin}(n, p_1)$, which has mean np_1 and variance $np_1(1 - p_1)$.

The marginal distribution of X_i is binomial, with index n and parameter p_i.

Similarly, the conditional distribution of X_i given $X_j = n_j$ is binomial with index $n - n_j$ and probability $\frac{p_i}{1-p_j}$.

2.2 Distributions related to Poisson processes

Another simple statistical experiment is to observe some aspect of a homogeneous Poisson process, in which point events occur singly and at a constant rate λ in time. In such a process, events occurring in disjoint time intervals are independent.

The Poisson distribution Let X be the number of point events observed in a fixed time period, T say. Then X has the Poisson distribution with mean $\theta = \lambda T$, where

$$\mathbf{P}(X = r) = p(r; \theta) = \frac{\theta^r e^{-\theta}}{r!}, r = 0, 1, \ldots$$

and $p(r; \theta) = 0$ otherwise. The variance of X is also θ, so that the Poisson distribution has equal mean and variance.

Recall that the binomial distribution, Bin(n, θ), has mean $n\theta$ and variance $n\theta(1-\theta)$. If we let $n \to \infty$ in such a way that the mean $n\theta = \mu$ is fixed, then the variance, $\mu(1-\mu/n)$, is asymptotically equal to the mean μ. A stronger result is that the binomial distribution tends to the Poisson distribution with mean μ under these assumptions.

Recap on the gamma and beta functions

The gamma function is defined as follows:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \text{ for } \alpha > 0.$$

By a simple change of variable, it is clear that

$$\int_0^\infty \lambda^\alpha x^{\alpha-1} e^{-\lambda x} dx = \Gamma(\alpha), \text{ for } \alpha > 0.$$

- Integrating by parts and assuming that $\alpha > 1$, we can show that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.
- setting $\alpha = n$, an integer, and iterating, that $\Gamma(n) = (n-1)!$
- $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$.

The beta function $B(\alpha, \beta)$ is defined by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

It can be shown that $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. If α and β are integers then this can also be written as $B(\alpha, \beta) = (\alpha-1)!(\beta-1)!/(\alpha+\beta-1)!$.

The exponential distribution Suppose now that the time elapsing until the next event of a Poisson process is observed. Then X has the exponential distribution with parameter $\theta = \lambda$, $\lambda > 0$, so that

$$p(x;\theta) = \theta e^{-\theta x}, \text{ for } x \geq 0$$

and $p(x;\theta) = 0$ otherwise.

The expected value of X is $1/\theta$ while the variance is $1/\theta^2$ (so the mean and standard deviation are equal for the exponential distribution).

The gamma distribution Let X be the time until kth subsequent point events occur. Then X has a gamma distribution with parameter $\theta = (k, \lambda)$ where k is the index of the distribution and λ is the scale parameter. In the most general form of the gamma distribution, the index, α say, does not have to be an integer but can take any positive value. The density then has the form

$$p(x;\theta) = \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \text{ for } x \geq 0$$

and $p(x;\theta) = 0$ otherwise. When α takes an integer value, k say, we will often write $(k-1)!$ rather than $\Gamma(k)$.

The expected value of X is α/λ and the variance is α/λ^2.

The transformed variable $Z = \lambda X$ also has a gamma distribution, with index α, but with a unit scale parameter (*check!*).

The exponential distribution is the special case of the gamma distribution when $k = 1$. By the Poisson process construction, since the intervals of a Poisson process are independent, it is clear that X can be regarded as the sum of k independent, exponential random variables. By the Central Limit Theorem, it follows that the gamma distribution becomes asymptotically normal when the index tends to infinity.

More generally, the sum of independent gamma variables having a common scale parameter itself has a gamma distribution with the same parameter, and its index is the sum of those of the component variables.

Gamma distributions (which are always positively skewed) are often used to model non-negative variables. The family of distributions has a flexible shape and is mathematically tractable. In a Bayesian framework, the gamma distribution is often used as a prior distribution for a non-negative parameter.

The beta distribution

The variable X has the beta distribution with parameter $\theta = (\alpha, \beta)$, $(\alpha, \beta > 0)$ if it has density

$$p(x;\theta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \text{ for } 0 \leq x \leq 1$$

and $p(x;\theta) = 0$ otherwise.

The expected value of X is $\frac{\alpha}{\alpha+\beta}$ while the variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Beta distributions are often used to model variables constrained to take positive, bounded values. The family of distributions has a flexible shape (it is symmetric when $\alpha = \beta$, positively skewed when $\alpha < \beta$ and negatively skewed when $\alpha > \beta$) and is mathematically tractable.

In a Bayesian framework, the beta distribution is often used as a prior distribution for a parameter which is a probability (e.g. for Bernoulli trials).

2.3 Distributions relating to normal variables

The normal distribution The random variable X has the normal distribution with parameter $\theta = (\mu, \sigma^2)$, $-\infty < \mu < \infty, \sigma \geq 0$, if

$$p(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

The mean of this distribution is μ and the variance is σ^2.

The transformed variable $Z = (X - \mu)/\sigma$ has the standard normal distribution, which has zero mean and unit variance.

The density function is symmetric and 'bell shaped'.

- Sums of normal variables are again normally distributed (they do not have to be independent) and, more generally, linear functions of normally distributed random variables are normally distributed.

- If $X_1, \ldots, X_n$ are i.i.d. normal variables, with common mean μ and variance σ^2, then the sample mean, $\overline{X} = (X_1 + \cdots + X_n)/n$, is again normally distributed with mean μ and variance σ^2/n, and the sampling distribution of $\sqrt{n}(\overline{X} - \mu)/\sigma$, which is used to test hypotheses about μ when σ^2 is known, is standard normal.

This distribution plays an extremely important role in statistics and is often used as the error distribution in statistical modelling. Much of the justification for its use comes from the Central Limit Theorem—under fairly general conditions, the sampling distributions of sample means and totals can be well-approximated by the normal distribution provided that the sample size is large, regardless of the distribution of the basic population from which the sample is drawn.

Many characteristics of natural populations are approximately normally distributed, essentially because a large number of factors have contributed to them in an approximately linear way.

The χ^2 distribution

- If X has the standard normal distribution, then X^2 has the gamma distribution with index $\frac{1}{2}$ and parameter $\frac{1}{2}$.

- If $X_1, \ldots, X_n$ are independent, standard normal variables, then $X_1^2 + \ldots + X_n^2$ has the gamma distribution with index $n/2$ and parameter $\frac{1}{2}$. This distribution is also known as the χ^2 distribution with n degrees of freedom.

- If $X_1, \ldots, X_n$ are independent normal variables, with common mean μ and variance σ^2, then the sampling distribution of $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2$ is χ^2 with n degrees of freedom. This result is used in testing hypotheses about σ^2 when μ is known.

- The distribution of $\sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2$ can be shown to be the same as that of a sum of squares of $n-1$ standard normal variables and thus has a χ^2 distribution with $n-1$ degrees of freedom.

- It follows that if S^2 denotes the sample variance, $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$, then $(n-1)S^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom. This statistic is often used in testing hypotheses about σ^2 when μ is unknown.

- It can be shown that $\overline{X}$ and S^2 are independent variables. This has a important consequences for estimation and hypothesis testing.

- The χ^2 distribution is also the appropriate null distribution in goodness-of-fit tests and tests for association in contingency tables.

Student's t distribution

If Z has a standard normal distribution, V has a χ^2 distribution with ν degrees of freedom, and Z and V are independent variables then

$$T = \frac{Z}{\sqrt{V/\nu}}$$

has the Student t distribution with ν degrees of freedom with probability density function

$$f_T(t) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

for $-\infty < t < \infty$.

The distribution of T is symmetric about 0, so that $\mathrm{E}(T) = 0$ (for $\nu > 1$). For $\nu > 2$, the variance of T can be shown to be $\nu/(\nu-2)$.

The Student t distribution with $n-1$ d.f. is the sampling distribution for $\sqrt{n}(\overline{X}-\mu)/S$, where $\overline{X}$ and S^2 are the mean and variance of a normal sample as defined above. This statistic is used in testing hypotheses about μ when the population variance σ^2 is unknown and must be estimated.

As the degrees of freedom ν goes to infinity, the t distribution tends to the standard normal distribution. The approximation is very good even for quite moderate values of ν.

The F distribution

If V_1 and V_2 are independent χ^2 variables with ν_1 and ν_2 degrees of freedom respectively, then the random variable

$$W = \frac{V_1/\nu_1}{V_2/\nu_2}$$

has the F distribution with (ν_1, ν_2) degrees of freedom with probability density function

$$f_W(w) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} w^{\nu_2/2-1}(1+\frac{\nu_1 w}{\nu_2})^{-(\nu_1+\nu_2)/2}$$

for $0 < w < \infty$.

It follows from the definition that $1/W$ has the F distribution with (ν_2, ν_1) degrees of freedom.

Notice that if T has the t distribution with ν degrees of freedom then

$$T^2 = \frac{Z^2}{V/\nu}$$

T^2 has the F distribution with $(1, \nu)$ df, since Z^2 has the χ^2 distribution with 1 df.

Since $\mathsf{E}(T) = 0$, $\mathsf{E}(T^2) = \mathsf{var}(T) = \nu/(\nu - 2)$ for $\nu > 2$.

More generally, it can be shown that, for all ν_1 and for $\nu_2 > 2$, $\mathsf{E}(W) = \nu_2/(\nu_2 - 2)$.

- If S_1^2 and S_2^2 are the sample variances for independent samples of sizes n_1 and n_2 from two distinct populations of independent and identically distributed normal variables with variances σ_1^2, σ_2^2 then the ratio of S_1^2/σ_1^2 to S_2^2/σ_2^2 has the F distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom, and can be used to test hypotheses about the variance ratio σ_1^2/σ_2^2.

- More commonly, the F distribution is used for making inferences about the appropriateness of models and whether certain variables should be included in the model specification; these inferences are all made via tests involving ratios of sample variances.