Statistical inference
Part 1
Ideas of inference.[*]

Michael Goldstein
Durham University
APTS December 2023

---

[*] This course is built on the work of previous presenters, in particular Jonty Rougier and Simon Shaw. However, any errors are all mine.

Outline.

In this introductory section, we will introduce some of the basic ideas that we will be exploring during this series.

We will do this by example, to recall and refresh basic notation, definitions and methods.

On the way, we will introduce some of the general questions that we will be discussing.

What is statistical inference?

In statistical inference experimental or observational data are modelled as the observed values of random variables, to provide a framework from which inductive conclusions may be drawn about the mechanism giving rise to the data.

[Essentials of Statistical Inference, Young and Smith, Cambridge (2005)]

What is statistical inference?

In statistical inference experimental or observational data are modelled as the observed values of random variables, to provide a framework from which inductive conclusions may be drawn about the mechanism giving rise to the data.

[Essentials of Statistical Inference, Young and Smith, Cambridge (2005)]

What is statistical inference?

ChatGPT

Statistical inference is the process of drawing conclusions or making predictions about a population based on a sample of data from that population.

Statistical inference

The whole of this appendix, and indeed the whole book, is concerned with statistical inference.

The object is to provide ideas and methods for the critical analysis and, as far as feasible, the interpretation of empirical data arising from a single experimental or observational study or from a collection of broadly similar studies bearing on a common target.

[Principles of Statistical Inference, Cox, Cambridge. (2006)]

Statistical versus scientific inference

The extremely challenging issues of scientific inference may be regarded as those of synthesising very different kinds of conclusions if possible into a coherent whole or theory and of placing specific analyses and conclusions within that framework.

This process is surely subject to uncertainties beyond those of the component pieces of information and, like statistical inference, has the features of demanding stringent evaluation of the consistency of information with proposed explanations.

The use, if any, in this process of simple quantitative notions of probability and their numerical assessment is unclear and certainly outside the scope of the present discussion.

[Principles of Statistical Inference, Cox, Cambridge.]

Statistical inference

This course will be largely guided by the traditional view of statistical inference

[because lots of interesting, important and essential results have been developed within this viewpoint].

However, we must also keep an eye on the wider context of scientific inference

[because this is where many important problems are that we may need to address.]

Statistical models

Much of the theory of statistical inference relates to the analysis of statistical models.

In the simplest version of a statistical model, we have observed data $\underline{x}$ (typically vector).

$\underline{x}$ is considered to be the realisation of random vector $\underline{X}$.

The probability distribution of $\underline{X}$ depends on a (possibly vector) parameter θ. We write this distribution as $f(\underline{x}|\theta)$.

In the simplest case, we know everything about $f(.|\theta)$ except for the “true but unknown” value of θ.

For example, $\underline{x}$ might be an independent sample from a population whose possible distributions are indexed by parameter θ.

Questions

There are two basic types of question we must consider.

(i) Inside the model:
assuming the truth of the model, what information does $\underline{x}$ provide about θ?

(ii) Outside the model:

where did the model come from?

how reliable is it?

how does the within model inference relate to real world uncertainty?

Spinning coins

We introduce many of the core ideas by considering a simple example of a statistical model and associated data.

Suppose that we spin a coin n times and observe each outcome as heads or tails.

Notation Tosses $\underline{X} = (X_1, X_2, ..., X_n)$

where $X_i = 1$ if heads, $X_i = 0$ otherwise.

Observed sample is $\underline{X} = \underline{x} = (x_1, x_2, ..., x_n)$

We want to estimate θ, the probability of heads.

Questions The probability of heads means exactly what?

Does it have a real world meaning or is it simply a modelling construct?

The probability of heads

User

What does it mean to say that the probability of a coin landing heads is 1/2?

ChatGPT

Saying that the probability of a coin landing heads is 1/2 means that, in a fair and unbiased coin toss, there is an equal chance of the coin landing heads (H) as there is of it landing tails (T). The probability of 1/2 represents a 50% chance that the outcome will be heads.

Chatting to Chat GPT

User

Is the probability that the coin lands heads a property of the coin or of our knowledge about the coin?

ChatGPT

The probability that a coin lands heads is a property of our knowledge about the coin and the way we model and understand the coin-tossing process. It is not an inherent property of the physical coin itself.

... In summary, the probability of a coin landing heads is a subjective assessment based on our knowledge, assumptions, and modeling of the coin-tossing process.

It represents our best estimate of the likelihood of the coin landing heads under the given circumstances.

Chatting to Chat GPT

User

Is there a true likelihood of the coin landing heads?

ChatGPT

The concept of a ”true likelihood” of a coin landing heads is a philosophical and epistemological question that depends on how one interprets probability and the nature of reality.

Independent outcomes

In many cases we will consider, the observations $\underline{X} = (X_1, \ldots, X_n)$ are independent given θ.

We say that $(X_1, \ldots, X_n)$ is an independent sample of size n.

If all observations are drawn from the same distribution $f(x|\theta)$ then we call $\underline{X}$ an iid (independent and identically distributed) sample of n from $f(.|\theta)$.

In such cases $f(\underline{x}|\theta)$ factorises as

$$f(\underline{x}|\theta) = \mathbb{P}((X_1 = x_1), \ldots, (X_n = x_n)|\theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Coin spinning

In the coin spinning example, for a single spin, $X = 0$ or $X = 1$ and

$$f(x|\theta) = \theta^x (1-\theta)^{1-x}$$

so

$$f(\underline{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^k (1-\theta)^{n-k}$$

where $k = \sum_{i=1}^{n} x_i$ is the number of heads observed in the n spins.

Sufficiency

We say that statistic $T(x_1, ..., x_n)$ (possibly a vector) is **sufficient** for parameter θ if we can factorise the joint probability function as

$$\mathbb{P}((X_1 = x_1), ..., (X_n = x_n)|\theta) = g(T(x_1, ..., x_n), \theta)h(x_1, ..., x_n)$$

In our coin spinning experiment, we have

$$f(\underline{x}|\theta) = \theta^k(1-\theta)^{n-k}$$

Therefore, we see that k, or equivalently

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is sufficient for our sampling problem.

Binomial sampling

Compare our formulation to estimating the parameter θ of a binomial distribution.

Here Y is the number of successes in n independent trials each with probability θ of success so that

$$P(Y = k|\theta) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}$$

We write this as $Y \sim Bi(n, \theta)$

The only thing that we appear to lose is our ability to test the validity of the model by examining the pattern of 1’s and 0’s in $\underline{x}$.

Questions

Why and when should our inference only depend on the sufficient statistics?

When can we build a statistical model on the sufficient statistics alone?

Estimating θ

Let

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

be the sample average or, equivalently in our case, the proportion of heads in the sample.

The observed value is

$$\overline{X} = \overline{x} = \frac{x_1 + ... + x_n}{n}$$

.

Let's use $\overline{x}$ as an estimator of θ.

Questions

How do we choose good estimators?

And what does good mean?

Properties of our estimator: mean

Let’s find the mean and variance of our estimator.

$$\begin{aligned} \mathrm{E}(\overline{X}) &= \mathrm{E}(\frac{1}{n}(X_1 + X_2 + ... + X_n)) \\ &= \frac{1}{n}(\mathrm{E}(X_1) + \mathrm{E}(X_2) + ... + \mathrm{E}(X_n)) \end{aligned}$$

as expectation is linear:

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$$

Bias

For each X_i, if the true parameter value is θ,

$$\mathrm{E}(X_i) = 1 \times P(X_i = 1) + 0 \times P(X_i = 0) = \theta$$

so

$$\mathrm{E}(\overline{X}) = \frac{n\theta}{n} = \theta$$

We say that our estimator $\overline{X}$ is **unbiased**:

the average of our estimator, for many repetitions of the experiment, “tends to the true value”.

Question Is this property important for our single experiment?

Properties of our estimator:variance

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}(\frac{1}{n}(X_1 + X_2 + ... + X_n)) \\
&= \frac{1}{n^2}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + ... + \mathrm{Var}(X_n)) \\
&= \frac{\mathrm{Var}(X)}{n}
\end{aligned}$$

as the variance of the sum of independent random quantities is the sum of the variances of the individual quantities and $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

The general form is

$$\mathrm{Var}(\sum_i Y_i) = \sum_i \mathrm{Var}(Y_i) + 2\sum_{i<j} \mathsf{Cov}(Y_i, Y_j)$$

Variance

$$\begin{aligned}\mathrm{Var}(X_i) &= \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2 \\ &= \theta - \theta^2\end{aligned}$$

(as $X_i = X_i^2$, so $\mathrm{E}(X_i) = \mathrm{E}(X_i^2)$)

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\theta(1-\theta)}{n}$$

Properties of our estimator: large sample distribution

The proportion of heads in n spins is

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

The central limit theorem (CLT) says that, under weak conditions, the probability distribution of the sum of a sequence of n independent, identically distributed random quantities tends to a normal distribution as n increases.

Therefore, by the central limit theorem, the large sample distribution of $\overline{X}$ is approximately Gaussian.

Properties of our estimator: large sample distribution

For every sample size n,

$$\mathrm{E}(\overline{X}) = \theta, \;\; \mathrm{Var}(\overline{X}) = \frac{\theta(1-\theta)}{n}$$

Therefore,

$$\overline{X} \sim N(\theta, \frac{\theta(1-\theta)}{n})$$

(approximately, for large n).

Questions

When will “good” estimators be approximately Gaussian, for large n?

For large n, is there a simple general way to identify roughly what the variance of a “good” estimator will be, and will it be at least approximately unbiased?

Large sample confidence interval

As $\overline{X} \sim N(\theta, \frac{\theta(1-\theta)}{n})$ (approximately), we have, for any α that

$$1-\alpha \approx P(-z_{\alpha/2} \leq \frac{\overline{X}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \leq z_{\alpha/2})$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ value of a standard normal distribution.
Rearranging, we have

$$1-\alpha \approx P(\overline{X} - z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}} \leq \theta \leq \overline{X} + z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}})$$

We substitute $\overline{X}$ as an approximation for θ in the variance estimate to give

$$\overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$$

as an approximate large sample $(1-\alpha)$ confidence interval for θ.

Confidence intervals

$$\overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$$

is an approximate large sample $(1-\alpha)$ confidence interval for θ.

This means that, for any value θ, the chance of drawing a sample $\underline{X}$ for which the above interval contains θ is $(1-\alpha)$ (approximately).

Therefore, the properties of a confidence interval derive from the process of generating samples and creating the corresponding confidence intervals.

Question Is there a general way of constructing (large sample) confidence intervals with good properties?

And what does “good” mean”?

Observed confidence intervals

Question How does the confidence property relate to what happens when we make a particular sample $\underline{X} = \underline{x}$ and create a particular interval

$$\overline{x} \pm z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$$

Note, in particular, that it does not mean that there is a probability of $1-\alpha$ that the true value of θ is within the actual interval obtained from our sample.

We can easily create confidence intervals which cannot contain the true parameter values - for example, they might be empty.

Comment:Reliability of this approximation

$$\overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$$

Question Is our particular approximation to the binomial confidence interval reliable?

Reasonably, yes. But not for θ near zero or one.

There is an extensive literature on improving the approximation for this particular problem.

See, for example, the Wikipedia page on the "Binomial proportion confidence interval".

Simulation experiments

In this course, we will discuss various large sample approximations.

Simulation experiments give a simple way to explore the reliability of such approximations.

There are two basic types of simulation.

Firstly, “in model” simulations.

In the spinning coins example, for various choices of n and θ generate many samples $\underline{x}$ and count the proportion of the samples for which our suggested interval contains the corresponding value of θ.

You could develop your own rule of thumb for how big n needs to be, for different choices of θ, for the result to be reliable.

For example, a common rule of thumb is to consider the normal approximation to the binomial to be reasonable when $n\theta$ and $n(1-\theta)$ are both greater than 5. How does this translate into accuracy of the derived confidence intervals?

Out of model uncertainty simulations

Also important are “out of model” simulations.

By this, consider the most important ways in which the model might fail to represent the random mechanisms generating the data.

Then build a representation of such alternative forms, say $f^*(x|\theta)$, embodying the alternative random mechanisms and evaluate their effect by simulation experiments.

For example, in the coin spinning experiment, we might allow the probability of heads on spin $i+1$ to depend on the outcome of the spin i.

Or we might consider that the probability of heads θ is not constant but instead varies (randomly or systematically) across the sequence of spins.

Out of model simulation experiments

Evaluations of $f^*(x|\theta)$ reveal how much structural error the model can support.

This may lead us to add some extra error to our model for $f(x|\theta)$ or to remodel the problem or to be happy with the original model.

This will depend on the intended purpose of the model.

For example, do we want to learn about θ or to make forecasts for future outcomes?

How good are the forecasts from f when compared with f^*?

Such considerations are strongly context specific and require careful thought (which is a good thing!)

Stopping rules

So far, all of our analyses have been based on assuming that the sample size n is fixed and known in advance. Suppose that this is not true.

Compare the following scenarios

(i) I spin the coin 20 times and see 5 heads (number of tosses fixed in advance).
(ii) I decide to keep tossing until I have seen 5 heads. I toss the coin 20 times. (In this case, the number of tosses is the random variable.)
(iii) I use a random stopping mechanism (independent of p) which stops after 20 tosses. I have seen 5 heads.
(iv) The experiment stops after seeing 5 heads in 20 tosses, due to a random event that was unforeseen and with unknown probability distribution (independent of p).

In each case, I have tossed the coin 20 times and seen 5 heads.

Question Should my inference be the same or different for each case?

Bayesian formulation

In the Bayesian formulation, the parameters of the probability model (and any other unknown quantities) are viewed as random quantities.

We assign a prior distribution, $p(\theta)$, for the parameter θ.

We update the prior distribution for θ to the posterior distribution for θ given data $\underline{x}$ using Bayes theorem

$$p(\theta|\underline{x}) = \frac{p(\underline{x}|\theta)p(\theta)}{p(\underline{x})} \propto p(\underline{x}|\theta)p(\theta)$$

where we find $p(\underline{x})$ as

$$p(\underline{x}) = \int p(\underline{x}|\theta)p(\theta)d\theta$$

General questions

In the Bayes approach, all of our inferences about θ are contained in $p(\theta|\underline{x})$

Questions

What does $p(\theta)$ mean?

Why is $p(\theta|\underline{x})$ the inference from the data?

[And are the “in model” and “out of model” answers different?]

The beta distribution

For a simple Bayesian analysis of the coin spinning experiment, we will use the convenience of a beta distribution prior.

We say that w has a **Beta distribution** with parameters $\alpha > 0$ and $\beta > 0$,

if the probability density function of w is given by

$$p(w) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} w^{\alpha-1}(1-w)^{\beta-1}$$

for $w \in [0,1]$. We write this as $w \sim \text{Be}(\alpha, \beta)$.

Note that, if $w \sim \text{Be}(\alpha, \beta)$ with $\alpha = \beta = 1$, then

$$p(w) = 1, \ \ 0 \leq w \leq 1,$$

i.e. w has a uniform distribution on [0,1].

The gamma function

The Gamma function Γ is defined, for any real number $z > 0$ as

$$\Gamma(z) = \int_0^\infty t^{z-1} \exp(-t) dt$$

The Gamma function satisfies the following properties:

- if n is integer, then

$$\Gamma(n) = (n-1)!$$

- for every z,

$$\Gamma(z+1) = z\Gamma(z)$$

- more generally, if n is a positive integer and $z > 0$ then

$$\Gamma(z+n) = z(z+1)\cdots(z+n-1)\Gamma(z)$$

- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Properties of the beta distribution

Suppose that $w \sim \text{Be}(\alpha, \beta)$

[1] The expectation of w is

$$E[w] = \frac{\alpha}{\alpha + \beta}$$

[2] The variance of w is

$$\text{Var}[w] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

[3] If $w \sim \text{Be}(\gamma, \delta)$, and if γ, δ are large, then unless $\gamma >> \delta$ or $\delta >> \gamma$, approximately

$$w \sim N(\frac{\gamma}{\gamma + \delta}, \frac{\gamma\delta}{(\gamma + \delta)^2(\gamma + \delta + 1)})$$

Binomial sampling

For binomial sampling, the likelihood is given by

$$P(X = k \mid w) = \binom{n}{k} w^k (1 - w)^{n-k} = C' w^k (1 - w)^{n-k}$$

for $k \in [0..n]$.

Viewed as a function of w, the binomial coefficient $C' = \binom{n}{k}$ is a multiplicative constant.

Therefore, the likelihood for binomial sampling is of the form

$$P(X = k \mid w) \propto w^k (1 - w)^{n-k}$$

.

Form of posterior distribution

Suppose that the prior distribution p on $W = [0, 1]$ is a beta distribution of the form

$$P(w) = C'' w^{\alpha-1}(1 - w)^{\beta-1} \propto w^{\alpha-1}(1 - w)^{\beta-1}$$

where C'' is a multiplicative constant.

Applying Bayes theorem, the posterior is then of the form:

$$P(w \mid X = k) \propto w^{\alpha-1}(1-w)^{\beta-1} \times w^{k}(1-w)^{n-k} \propto w^{\alpha+k-1}(1-w)^{\beta+(n-k)-1}$$

Hence, the posterior follows the same parametric form as the prior.

In other words, the passage from prior to posterior only involves a change in the hyperparameters with no additional calculation.

Properties of beta binomial sampling

Therefore, if

$$X \mid w \sim \mathsf{Bi}(n, w)$$

and

$$w \sim \mathsf{Be}(\alpha, \beta)$$

then

$$w \mid X = k \sim \mathsf{Be}(\alpha + k, \beta + n - k)$$

Conjugate Family: General Definition

Let $X_1, \ldots, X_n$ be an independent sample of size n, each with likelihood $P(x \mid w)$.

A **conjugate family** for sampling from $P(x \mid w)$ is a set $\mathcal{M}$ of distributions with the following property:

if the prior for w is any member of $\mathcal{M}$, then for any sample size n and any sample values $\{X_i = x_i\}_{i \in [1..n]}$, the posterior distribution for w is also a member of $\mathcal{M}$.

For example, the beta distributions are a conjugate family for binomial sampling.

Comments

Many sampling problems have natural conjugate families

Conjugate families contain a wide variety of probability distributions, so you often find a distribution that provides a good approximation to your prior knowledge within a conjugate family.

Also helpful for exploring the inferences over a wide range of differing prior judgements.

You can use mixtures of conjugate priors for more complicated, eg multimodal shapes.

We will find them particularly useful when we look at decision procedures, where we need to make large numbers of updates a priori.

Limiting properties of beta binomial sampling

In our sampling problem, with prior Be(α, β) and observation k successes in n spins, our posterior for θ is Be(γ, δ) where

$$\gamma = \alpha + k, \delta = \beta + n - k$$

For large k, n, $\mathrm{E}(w|k), \mathrm{Var}(w|k)$ are

$$\frac{\gamma}{\gamma + \delta} = \frac{\alpha + k}{\alpha + \beta + n} \approx \frac{k}{n}$$

$$\frac{\gamma\delta}{(\gamma + \delta)^2(\gamma + \delta + 1)}) = \frac{(\alpha + k)(\beta + n - k)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \approx \frac{k(n - k)}{n^3}$$

Normal approximation

If

$$w \sim \text{Be}(\gamma, \delta)$$

,
where γ, δ large, then approximately

$$w \sim N(\frac{\gamma}{\gamma+\delta}, \frac{\gamma\delta}{(\gamma+\delta)^2(\gamma+\delta+1)})$$

so, approximately,

$$w \sim N(\overline{x}, \frac{\overline{x}(1-\overline{x})}{n})$$

(as $\overline{x} = \frac{k}{n}$.)

Large sample credible interval

A $(1-\alpha)$ level **credible interval** for a parameter θ, given data $\underline{x}$ is one for which the probability that θ is in the interval given probability distribution $p(\theta|\underline{x})$ is $(1-\alpha)$.

We have shown that the posterior distribution for θ for large n is approximately

$$N(\overline{x}, \frac{\overline{x}(1-\overline{x})}{n})$$

Therefore, the central $(1-\alpha)$ credible interval for θ is approximately

$$\overline{x} \pm z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$$

Questions

When do posterior distributions tend to normality?

Is there a general way of assessing the mean and variance of the normal approximation that does not require the conjugate form?

Discussion

We have shown that the approximate large sample $(1-\alpha)$ confidence interval for the binomial parameter θ is

$$\overline{x} \pm z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$$

We have also shown that the approximate large sample $(1-\alpha)$ credible interval for θ, for any Beta prior, is

$$\overline{x} \pm z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$$

Question

The confidence interval and the credible interval are answering completely different questions.

Why are they the same, and how general is this equivalence?