Statistics 2 Unit 3

Kurt Hornik

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION AMBA ACCREDITED

Outline

- Estimation of parameters and fitting of probability distributions
- Testing hypotheses and assessing goodness of fit

Outline

- Estimation of parameters and fitting of probability distributions
  - The Bayesian Approach to Parameter Estimation
  - Efficiency
  - Sufficiency
- Testing hypotheses and assessing goodness of fit

Big picture

In the Bayesian approach, the unknown parameter θ is treated as a random variable with "prior" distribution $f_\Theta(\theta)$ representing what we know about the parameter before observing data.

(For now, we write Θ for the random variable corresponding to the parameter θ.)

I.e., uncertainty about parameters is also modeled probabilistically.

(Very nice idea, but often the priors have parameters (so-called hyperparameters) which are also unknown but no longer modeled probabilistically.)

For a given value $\Theta = \theta$, the data have probability distribution $f_{X|\Theta}(x|\theta)$.

(We used to write $f(x|\theta)$: the subscripts now indicate the corresponding random variables.)

Big picture

If Θ has a continuous distribution, the joint distribution of X and Θ is

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta) f_{\Theta}(\theta).$$

Big picture

If Θ has a continuous distribution, the joint distribution of X and Θ is

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta) f_{\Theta}(\theta).$$

The marginal distribution of X is

$$f_X(x) = \int f_{X,\Theta}(x,\theta) d\theta = \int f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)\, d\theta.$$

Big picture

If Θ has a continuous distribution, the joint distribution of X and Θ is

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta) f_{\Theta}(\theta).$$

The marginal distribution of X is

$$f_X(x) = \int f_{X,\Theta}(x,\theta) d\theta = \int f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)\, d\theta.$$

Finally, the distribution of Θ given the data, the so-called **posterior distribution**, is

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X,\Theta}(x,\theta)}{\int f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)\, d\theta} = \frac{f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)}{\int f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)\, d\theta}.$$

Big picture

(This is a bit awkward: in the denominator, θ is integrated out.)

Note that $f_{X|\Theta}(x|\theta)$ is the likelihood, and by the above (the denominator is the marginal density of x and hence a constant for fixed/given x)

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) f_{\Theta}(\theta).$$

This is useful if we can recognize the posterior from the numerator: we then do not need to compute the denominator (as we already know it), see below.

Big picture

In the above, X and x can also be vectors. Alternatively,

$$f_{\Theta|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) \propto f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta) \times f_\Theta(\theta)$$

and as usual, if $X_1,\ldots,X_n$ are i.i.d. given θ,

$$f_{\Theta|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) \propto f_{X_1|\Theta}(x_1|\theta) \times \cdots \times f_{X_n|\Theta}(x_n|\theta) \times f_\Theta(\theta).$$

After observing $x_1,\ldots,x_n$, the posterior contains all available information about the parameter, and inference is therefore always based on the posterior ("likelihood principle").

Example: Poisson distribution

Suppose that given $\Lambda = \lambda$, $X_1, \ldots, X_n$ are i.i.d. Poisson(λ), with Λ having a prior density $f_\Lambda(\lambda)$.

Then

$$f_{X_1,\ldots,X_n|\Lambda}(x_1,\ldots,x_n|\lambda) = \prod_{i=1}^{n} f_{X_i|\Lambda}(x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!} e^{-n\lambda}.$$

The posterior is thus (terms which only depend on the x_i cancel out)

$$f_{\Lambda|X_1,\ldots,X_n}(\lambda|x_1,\ldots,x_n) = \frac{\lambda^{\sum_i x_i} e^{-n\lambda} f_\Lambda(\lambda)}{\int \lambda^{\sum_i x_i} e^{-n\lambda} f_\Lambda(\lambda) d\lambda}.$$

To evaluate this, one needs to specify the prior, and carry out the integration in the denominator.

Example: Poisson distribution

Suppose we take the prior as Gamma(α, rate $= \nu$) (we usually write λ for the rate parameter, but that is already taken):

$$f_\Lambda(\lambda) = \frac{\nu^\alpha \lambda^{\alpha-1} e^{-\nu\lambda}}{\Gamma(\alpha)}$$

Then (canceling out constants)

$$f_{\Lambda|X_1,\ldots,X_n}(\lambda|x_1,\ldots,x_n) = \frac{\lambda^{\sum_i x_i+\alpha-1} e^{-(n+\nu)\lambda}}{\int \lambda^{\sum_i x_i+\alpha-1} e^{-(n+\nu)\lambda} d\lambda}.$$

Without computing the integral, we can see that the posterior is Gamma($\sum_i x_i + \alpha$, rate $= n + \nu$)!

Example: Poisson distribution

In the Bayesian paradigm, all information about Λ is contained in the posterior.

We can estimate the parameter e.g. by the mean or mode (**posterior mean** and **posterior mode**, respectively) of this distribution.

For a Gamma distribution with shape α and rate ν these are α/ν and $(\alpha-1)/\nu$, giving the estimates

$$\frac{\sum_i x_i + \alpha}{n+\nu}, \qquad \frac{\sum_i x_i + \alpha - 1}{n+\nu}.$$

Example: Poisson distribution

The Bayesian analogue to the confidence interval is the interval from the $\alpha/2$ to the $1-\alpha/2$ quantile of the posterior (a $1-\alpha$ **credible interval**).

Alternatively, the **high posterior density (HPD) interval** is obtained as a level set

$$\{\lambda : f_{\Lambda|X_1,\ldots,X_n}(\lambda|x_1,\ldots x_n) \geq c\}$$

with c chosen to achieve posterior coverage probability $1-\alpha$.

Example: Poisson distribution

One could choose other priors, e.g., a uniform prior on $[0, 100]$.

Then

$$f_{\Lambda|X_1,\ldots,X_n}(\lambda|x_1,\ldots,x_n) = \frac{\lambda^{\sum_i x_i} e^{-n\lambda}}{\int_0^{100} \lambda^{\sum_i x_i} e^{-n\lambda} d\lambda}, \qquad 0 \leq \lambda \leq 100.$$

In this case, the denominator has to be integrated numerically (note the relation to the distribution function of the Gamma distribution).

Example: Normal distribution

One conveniently reparametrizes the normal, replacing σ^2 by the **precision** $\xi = 1/\sigma^2$.

Writing θ instead of μ (so that we can write Θ for the corresponding random variable),

$$f_{X|\Theta,\Xi}(x|\theta,\xi) = \sqrt{\frac{\xi}{2\pi}} e^{-\xi(x-\theta)^2/2}.$$

Rice covers several cases (unknown mean and known variance, known mean and unknown variance, unknown mean and unknown variance).

For the last, one possibly model is to specify independent priors for Θ and Ξ as

$$\Theta \sim N(\theta_0, \xi_{\text{prior}}^{-1}), \qquad \Xi \sim \text{Gamma}(\alpha, \text{rate} = \lambda).$$

Example: Normal distribution

Then (if the X_i are i.i.d. as usual),

$$
\begin{aligned}
& f_{\Theta,\Xi|X_1,\ldots,X_n}(\theta,\xi|x_1,\ldots,x_n) \\
& \quad \propto \quad f_{X_1,\ldots,X_n|\Theta,\Xi}(x_1,\ldots,x_n|\theta,\xi) f_\Theta(\theta) f_\Xi(\xi) \\
& \quad \propto \quad \exp\left(-\frac{\xi}{2}\sum_i (x_i-\theta)^2\right) \exp\left(-\frac{\xi_{\text{prior}}}{2}(\theta-\theta_0)^2\right) \xi^{n/2+\alpha-1} e^{-\lambda\xi}.
\end{aligned}
$$

which looks rather “messy”.

If the priors are quite flat (i.e., α, λ and ξ_{prior} are small), we get (approximately)

$$
f_{\Theta,\Xi|X_1,\ldots,X_n}(\theta,\xi|x_1,\ldots,x_n) \propto \exp\left(-\frac{\xi}{2}\sum_i (x_i-\theta)^2\right) \xi^{n/2-1}.
$$

Example: Normal distribution

The marginal posterior of Θ is obtained by integrating out ξ as

$$f_{\Theta|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) \propto \left(\sum(x_i-\theta)^2\right)^{-n/2}$$

from which after some algebra it can be shown that under the marginal posterior,

$$\sqrt{n}\frac{\Theta-\bar{x}}{s} \sim t_{n-1}$$

corresponding to the result from maximum likelihood analysis.

More on priors

We saw that for the Poisson distribution, using a Gamma prior gave a Gamma posterior: in general, such priors (families of priors G for which when the data distribution is in a family H, then the posterior again is in G) are called **conjugate priors** (to the family of data distributions).

In many applications, it is desirable to use flat or “non-informative” priors—but this hard to make precise.

In the Poisson case with Gamma priors, these are flat when α and ν are small. But taking limits gives

$$f_\Lambda(\lambda) \propto \lambda^{-1}, \qquad \lambda > 0$$

which is not a valid density!

Such priors are called **improper priors**, and may result in proper or improper posteriors.

More on priors

E.g., in the Poisson case, using the improper prior $f_\Lambda(\lambda) \propto \lambda^{-1}$ results in the posterior

$$f_{\Lambda|X_1,\ldots,X_n}(\lambda|x_1,\ldots,x_n) \propto \lambda^{\sum x_i - 1} e^{-n\lambda}$$

which is proper iff $\sum_i x_i > 0$.

In which case it is a Gamma distribution with shape $\sum_i x_i$ and rate n, as obtained by taking limits in the posterior.

More on priors

E.g., in the normal case with unknown mean and precision, one can take

$$f_\Theta(\theta) \propto 1, \qquad f_\Xi(\xi) \propto \xi^{-1},$$

This gives the joint posterior

$$\begin{aligned} & f_{\Theta,\Xi|X_1,\ldots,X_n}(\theta, \xi | x_1, \ldots, x_n) \\ & \quad \propto \quad \xi^{n/2-1} \exp\left(-\frac{\xi}{2} \sum_i (x_i - \theta)^2 \right) \\ & \quad \propto \quad \xi^{n/2-1} \exp\left(-\frac{\xi}{2}(n-1)s^2 \right) \exp\left(-\frac{n\xi}{2}(\theta - \bar{x})^2 \right). \end{aligned}$$

Conditional on ξ, Θ is normal with mean $\bar{x}$ and precision $n\xi$.

Computational aspects

Bayesian inference typically requires considerable computational power, e.g., for computing the normalizing constants.

In high dimensional problems, difficulties arise, and one can use sophisticated methods such as **Gibbs sampling**.

Consider inference for a normal with unknown mean and variance and an improper prior ($\alpha \to 0$, $\lambda \to 0$, $\xi_{\text{prior}} \to 0$. Then (as before)

$$f_{\Theta,\Xi|X_1,\ldots,X_n}(\theta,\xi|x_1,\ldots,x_n) \propto \xi^{n/2-1}\exp\left(-\frac{\xi}{2}(n-1)s^2\right)\exp\left(-\frac{n\xi}{2}(\theta-\bar{x})^2\right).$$

To study the posterior by Monte Carlo, one would draw many pairs (θ_k, ξ_k) from this joint density—but how?

Computational aspects

Gibbs sampling alternates between simulating from the conditional distribution of one parameter given the others.

In our case, we note that

- given ξ, Θ is normal with mean $\bar{x}$ and precision $n\xi$
- given θ, Ξ has a Gamma distribution.

Computational aspects

One would then proceed as follows:

1. Choose an initial value θ_0, e.g., $\bar{x}$.

Computational aspects

One would then proceed as follows:

1. Choose an initial value θ_0, e.g., $\bar{x}$.
2. Generate ξ_0 from a Gamma density with parameters $n/2$ and $n(\theta_0 - \bar{x})^2/2$ (which will not work, as the latter is zero, so one really needs another initial value).

Computational aspects

One would then proceed as follows:

1. Choose an initial value θ_0, e.g., $\bar{x}$.
2. Generate ξ_0 from a Gamma density with parameters $n/2$ and $n(\theta_0 - \bar{x})^2/2$ (which will not work, as the latter is zero, so one really needs another initial value).
3. Generate θ_1 from a normal distribution with mean $\bar{x}$ and precision $n\xi_0$.

Computational aspects

One would then proceed as follows:

1. Choose an initial value θ_0, e.g., $\bar{x}$.
2. Generate ξ_0 from a Gamma density with parameters $n/2$ and $n(\theta_0 - \bar{x})^2/2$ (which will not work, as the latter is zero, so one really needs another initial value).
3. Generate θ_1 from a normal distribution with mean $\bar{x}$ and precision $n\xi_0$.
4. Generate ξ_1 from a Gamma density with parameters $n/2$ and $n(\theta_1 - \bar{x})^2/2$.

Computational aspects

One would then proceed as follows:

1. Choose an initial value θ_0, e.g., $\bar{x}$.
2. Generate ξ_0 from a Gamma density with parameters $n/2$ and $n(\theta_0 - \bar{x})^2/2$ (which will not work, as the latter is zero, so one really needs another initial value).
3. Generate θ_1 from a normal distribution with mean $\bar{x}$ and precision $n\xi_0$.
4. Generate ξ_1 from a Gamma density with parameters $n/2$ and $n(\theta_1 - \bar{x})^2/2$.
5. etc.

After a “burn-in” period of a several hundred steps, one obtains pairs which approximately have the posterior distribution (but are not independent of one another).

Outline

- Estimation of parameters and fitting of probability distributions
  - The Bayesian Approach to Parameter Estimation
  - Efficiency
  - Sufficiency
- Testing hypotheses and assessing goodness of fit

Efficiency

Given a variety of possible parameter estimates, which one should we use?

Ideally, the one whose sampling distribution was most concentrated about the underlying value.

One possible concentration measure is the mean squared error

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \theta)^2 = \mathrm{var}_\theta(\hat{\theta}) + (\mathbb{E}_\theta(\hat{\theta}) - \theta)^2.$$

Clearly, the above implicitly assumes that the parameter is real-valued. In the vector-valued case, we could use

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2.$$

Let's KISS and (mostly) do real-valued in this section.

Efficiency

Note: this is a function of the underlying parameter θ, although this is not made explicit by the notation.

Efficiency

Note: this is a function of the underlying parameter θ, although this is not made explicit by the notation.

Note: not a good measure for skewed or multi-modal distributions!

Reasonable for distributions which are approximately normal (such as the sampling distributions for MLEs from large enough samples).

Efficiency

Remember: we say that an estimate $\hat{\theta}$ is **unbiased** if

$$\mathbb{E}_\theta(\hat{\theta}) = \theta.$$

For unbiased estimates, the mean squared error equals the variance, and hence comparison of MSEs reduces to comparing the variances or standard errors, respectively.

For two unbiased estimates $\hat{\theta}$ and $\tilde{\theta}$, the (relative) **efficiency** of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined as

$$\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{var}_\theta(\tilde{\theta})}{\text{var}_\theta(\hat{\theta})}.$$

(Again, this is a function of θ.)

Cramér-Rao Inequality

Theorem (Cramér-Rao inequality). *Let $X_1, \ldots, X_n$ be i.i.d. with density function $f(x|\theta)$. Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimate of the real-valued θ. Then under suitable smoothness assumptions on $f(x|\theta)$,*

$$\operatorname{var}_\theta(T) \geq \frac{1}{nI(\theta)}.$$

Proof. Let

$$Z = \sum_{i=1}^{n} \frac{\partial \log(f(X_i|\theta))}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{f(X_i|\theta)} \frac{\partial f(X_i|\theta)}{\partial \theta}$$

We already know that $\mathbb{E}_\theta(Z) = 0$ and $\operatorname{var}_\theta(Z) = nI(\theta)$.

Cramér-Rao Inequality

Next,

$$\log(g)' = \frac{g'}{g} \Rightarrow g' = g \times \log(g)'.$$

If $g = g_1 \times \cdots \times g_n$,

$$\begin{aligned}
(g_1 \times \cdots \times g_n)' &= (g_1 \times \cdots \times g_n) \times \log(g_1 \times \cdots \times g_n)' \\
&= (g_1 \times \cdots \times g_n) \times (\log(g_1)' + \cdots + \log(g_n)') \\
&= (g_1 \times \cdots \times g_n) \times \left(\frac{g_1'}{g_1} + \cdots + \frac{g_n'}{g_n} \right).
\end{aligned}$$

(Product rule for differentiation of a product with arbitrarily many factors.)

Cramér-Rao Inequality

Since Z has mean zero,

$$\begin{aligned}
\mathrm{cov}_\theta(T,Z) &= \mathbb{E}_\theta(TZ) \\
&= \int \cdot \int t(x_1,\ldots,x_n)\left(\sum_{i=1}^{n}\frac{1}{f(x_i|\theta)}\frac{\partial f(x_i|\theta)}{\partial\theta}\right)\prod_{j=1}^{n} f(x_j|\theta)dx_j \\
&= \int \cdot \int t(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i|\theta)dx_i \\
&= \frac{\partial}{\partial\theta}\int \cdot \int t(x_1,\ldots,x_n)\prod_{i=1}^{n} f(x_i|\theta)dx_i \\
&= \frac{\partial}{\partial\theta}\mathbb{E}_\theta(T).
\end{aligned}$$

Cramér-Rao Inequality

Thus if T is unbiased,

$$\mathrm{cov}_\theta(T, Z) = \frac{\partial}{\partial \theta}\mathbb{E}_\theta(T) = \frac{\partial}{\partial \theta}\theta = 1.$$

Using the Cauchy-Schwarz inequality,

$$\mathrm{var}_\theta(T)\mathrm{var}_\theta(Z) \geq \mathrm{cov}_\theta(T, Z)^2 = 1$$

from which

$$\mathrm{var}_\theta(T) \geq \frac{1}{\mathrm{var}_\theta(Z)} = \frac{1}{nI(\theta)}.$$

Yes!

Cramér-Rao Inequality

If θ is vector-valued, it still holds that if T is unbiased,

$$\mathrm{var}_\theta(T) \geq (nI(\theta))^{-1}$$

where the inequality is now understood with respect to the half-order on symmetric non-negative definite matrices, i.e.,

$\mathrm{var}_\theta(T) - (nI(\theta))^{-1}$ is non-negative definite.

Example: Poisson distribution

We know that $I(\lambda) = 1/\lambda$.

Hence, for any unbiased estimator of λ,

$$\mathrm{var}_\lambda(T) \geq \lambda/n.$$

On the other hand, the MLE $\bar{X} = S/n$ is unbiased with variance λ/n, hence attains the bound. Hence, it is “most efficient” in the sense of having the smallest possible variance (MSE) among all unbiased estimators!

We say that the MLE is a MVUE (minimum variance unbiased estimator).

MLE in general

We have shown that for large enough i.i.d. samples, the MLE is approximately $N(0, (nI(\theta))^{-1}$, so that

- it is asymptotically unbiased
- it asymptotically attains the Cramér-Rao bound.

Thus (with a bit of hand-waiving), it is asymptotically efficient!

A bit more convincingly, the bias-corrected MLE asymptotically attains the Cramér-Rao bound, and hence is asymptotically efficient.

(Traditional statistical inference loves the notion of unbiasedness.)

Outline

- Estimation of parameters and fitting of probability distributions
  - The Bayesian Approach to Parameter Estimation
  - Efficiency
  - Sufficiency
- Testing hypotheses and assessing goodness of fit

Motivation

The notion of sufficiency arises as an attempt to answer the following question:

> *for a sample $X_1, \ldots, X_n$ from the density $f(x|\theta)$, is there a statistic $T = t(X_1, \ldots, X_n)$ which contains all information in the sample about θ?*

Think of Bernoulli experiments: we have the feeling that only the number of successes matters.

Motivation

The notion of sufficiency arises as an attempt to answer the following question:

> *for a sample $X_1, \ldots, X_n$ from the density $f(x|\theta)$, is there a statistic $T = t(X_1, \ldots, X_n)$ which contains all information in the sample about θ?*

Think of Bernoulli experiments: we have the feeling that only the number of successes matters.

The official definition is:

Definition. *A statistic $T = t(X_1, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ does not depend on θ, for any value of t.*

Example: Bernoulli experiment

Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli random variables with success probability $\mathbb{P}_\theta(X = 1) = \theta$, and let $T = X_1 + \cdots + X_n$.

Thus if $x_i \in \{0, 1\}$, we can readily verify that

$$\mathbb{P}_\theta(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$

(this is very useful to remember!)

We know that T has a binomial distribution with parameters n and θ.

Example: Bernoulli experiment

Thus if $t = x_1 + \cdots + x_n$ with all $x_i \in \{0, 1\}$,

$$\begin{aligned}
&\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n | T = t) \\
&= \frac{\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)}{\mathbb{P}_\theta(T = t)} \\
&= \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\mathbb{P}_\theta(T = t)} \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{1}{\binom{n}{t}}.
\end{aligned}$$

Example: Bernoulli experiment

We see that indeed,

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{1}{\binom{n}{t}}$$

does not depend on θ!

So (by definition), $T = X_1 + \cdots + X_n$ is sufficient for the Bernoulli experiment $(X_1, \ldots, X_n)$ (as it should be).

Factorization theorem

Theorem. *A necessary and sufficient condition for $t(X_1, \ldots, X_n)$ to be sufficient for a parameter θ is that the joint probability function factors in the form*

$$f(x_1, \ldots, x_n|\theta) = g(t(x_1, \ldots, x_n), \theta)h(x_1, \ldots, x_n).$$

In words: sufficiency if and only if the joint density can be written as the product of a function of $t(x_1, \ldots, x_n)$ and θ, and a function which depends on $x_1, \ldots, x_n$ but not θ.

Factorization theorem

Proof. We give a proof for the discrete case.

Let $X = (X_1, \ldots, X_n)$ and $x = (x_1, \ldots, x_n)$.

Suppose the pmf factors as given in the theorem. I.e.,

$$\mathbb{P}_\theta(X = x) = g(t(x), \theta)h(x).$$

Then

$$\begin{aligned}
\mathbb{P}_\theta(T = t) &= \sum_{x:t(x)=t} \mathbb{P}_\theta(X = x) \\
&= \sum_{x:t(x)=t} g(t(x), \theta)h(x) \\
&= g(t, \theta) \sum_{x:t(x)=t} h(x).
\end{aligned}$$

Factorization theorem

Hence, if $t = t(x)$,

$$\begin{aligned}
\mathbb{P}_\theta(X = x | T = t) &= \frac{\mathbb{P}_\theta(X = x, T = t)}{\mathbb{P}_\theta(T = t)} \\
&= \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T = t)} \\
&= \frac{g(t, \theta) h(x)}{g(t, \theta) \sum_{x: t(x) = t} h(x)} \\
&= \frac{h(x)}{\sum_{x: t(x) = t} h(x)}
\end{aligned}$$

does not depend on θ, as was to be shown.

Factorization theorem

Conversely, suppose the conditional distribution of X given T does not depend on θ.

Clearly,

$$\mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(T = t)\mathbb{P}_\theta(X = x|T = t) = g(t, \theta)h(x)$$

where

$$g(t, \theta) := \mathbb{P}_\theta(T = t)$$

and by assumption,

$$h(x) := \mathbb{P}_\theta(X = x|T = t)$$

does not depend on θ. □

Factorization theorem

The factorization theorem (and in fact, also the definition) implies that sufficient statistics are unique only up to invertible transformations.

If s is invertible and $t(X)$ is sufficient,

$$\begin{aligned} f(x|\theta) &= g(t(x), \theta)h(x) \\ &= g(s^{-1}(s(t(x)), \theta)h(x) \\ &= g_s(s(t(x)), \theta)h(x), \end{aligned}$$

where $g_s(u, \theta) := g(s^{-1}(u), \theta)$.

Hence, $s(t(X))$ is sufficient too.

Example: Bernoulli experiment

We have

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)} \end{aligned}$$

So writing $t = \sum_i x_i$,

$$f(x_1, \ldots, x_n|\theta) = \theta^t(1-\theta)^{n-t} = \left(\frac{\theta}{1-\theta}\right)^t (1-\theta)^n$$

which gives $g(t, \theta)$, and we can take $h(x_1, \ldots, x_n) = 1$.

Example: Normal distribution

For a random sample from the normal distribution with unknown mean and variance, we have

$$
\begin{aligned}
& f(x_1, \ldots, x_n | \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) \\
&= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right) \\
&= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right) \right).
\end{aligned}
$$

Example: Normal distribution

Clearly, this depends on $x_1, \ldots, x_n$ only through $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$.

Hence,

$$T = t(X_1, \ldots, X_n) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)$$

is sufficient for $\theta = (\mu, \sigma^2)$.

Now clearly $\sum_{i=1}^n x_i = n\bar{x}$ and we established that

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2.$$

So $(\bar{X}, \hat{\sigma}^2)$ and $(\bar{X}, S^2)$ are sufficient too.

Sufficiency and MLE

Theorem. *If T is sufficient for θ, the MLE of θ is a function of T.*

Proof. Because

$$f(x_1, \ldots, x_n|\theta) = g(t(x_1, \ldots, x_n), \theta)h(x_1, \ldots, x_n).$$

the MLE is found by maximizing $g(t(x_1, \ldots, x_n), \theta)$, i.e., a function of $t(x_1, \ldots, x_n)$.

Rao-Blackwell theorem

Theorem (Rao-Blackwell theorem). *Let $\hat{\theta}$ be an estimate of θ with $\mathbb{E}_\theta(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ, and let $\tilde{\theta} = \mathbb{E}(\hat{\theta}|T)$ (which does not depend on θ).*

Then, for all θ,

$$\mathbb{E}_\theta((\tilde{\theta} - \theta)^2) \leq \mathbb{E}_\theta((\hat{\theta} - \theta)^2)$$

and the inequality is strict unless $\tilde{\theta} = \hat{\theta}$ (almost surely under $\mathbb{P}_\theta$).

Proof. By the theorem of iterated conditional expectation,

$$\mathbb{E}_\theta(\tilde{\theta}) = \mathbb{E}_\theta(\mathbb{E}(\hat{\theta}|T)) = \mathbb{E}_\theta(\hat{\theta}).$$

Thus, to compare the MSEs we only need to compare the variances.

Rao-Blackwell theorem

Now using a result on conditional expectations,

$$\begin{aligned} \mathrm{var}_\theta(\hat{\theta}) &= \mathrm{var}_\theta(\mathbb{E}(\hat{\theta}|T)) + \mathbb{E}_\theta(\mathrm{var}(\hat{\theta}|T)) \\ &= \mathrm{var}_\theta(\tilde{\theta}) + \mathbb{E}_\theta(\mathrm{var}(\hat{\theta}|T)). \end{aligned}$$

Thus, $\mathrm{var}_\theta(\hat{\theta}) > \mathrm{var}_\theta(\tilde{\theta})$ unless $\mathbb{E}_\theta(\mathrm{var}(\hat{\theta}|T)) = 0$, in which case $\hat{\theta}$ must be a function of T, which would imply $\hat{\theta} = \tilde{\theta}$. Done!

Rao-Blackwell theorem

Now using a result on conditional expectations,

$$\begin{aligned} \mathrm{var}_\theta(\hat{\theta}) &= \mathrm{var}_\theta(\mathbb{E}(\hat{\theta}|T)) + \mathbb{E}_\theta(\mathrm{var}(\hat{\theta}|T)) \\ &= \mathrm{var}_\theta(\tilde{\theta}) + \mathbb{E}_\theta(\mathrm{var}(\hat{\theta}|T)). \end{aligned}$$

Thus, $\mathrm{var}_\theta(\hat{\theta}) > \mathrm{var}_\theta(\tilde{\theta})$ unless $\mathbb{E}_\theta(\mathrm{var}(\hat{\theta}|T)) = 0$, in which case $\hat{\theta}$ must be a function of T, which would imply $\hat{\theta} = \tilde{\theta}$. Done!

The Rao-Blackwell theorem gives a strong rationale for basing estimators on sufficient statistics if they exist: if they are not functions of the sufficient statistics, their variance can be reduced without changing their bias.

Outline

- Estimation of parameters and fitting of probability distributions
- Testing hypotheses and assessing goodness of fit

Outline

- Estimation of parameters and fitting of probability distributions
- Testing hypotheses and assessing goodness of fit
  - Introduction
  - The Neyman-Pearson paradigm

Introduction

Suppose we have two coins: with H denoting “head” (traditionally, the head of the monarch, now the nice graphic; conversely, “tail” shows the denomination)

$$P_0(H) = 0.5, \qquad P_1(H) = 0.7.$$

Suppose one of these coins is chosen, tossed 10 times, and the number of heads reported, without telling which coin was chosen.

How should we decide which one it was?

Introduction

Suppose we have two coins: with H denoting “head” (traditionally, the head of the monarch, now the nice graphic; conversely, “tail” shows the denomination)

$$P_0(H) = 0.5, \qquad P_1(H) = 0.7.$$

Suppose one of these coins is chosen, tossed 10 times, and the number of heads reported, without telling which coin was chosen.

How should we decide which one it was?

Natural idea: find out which coin makes the observations more likely.

Introduction

Suppose we have two coins: with H denoting “head” (traditionally, the head of the monarch, now the nice graphic; conversely, “tail” shows the denomination)

$$P_0(H) = 0.5, \qquad P_1(H) = 0.7.$$

Suppose one of these coins is chosen, tossed 10 times, and the number of heads reported, without telling which coin was chosen.

How should we decide which one it was?

Natural idea: find out which coin makes the observations more likely.

Technically, we specify two **hypotheses**:

$$H_0 : \text{coin 0 was tossed}, \qquad H_1 : \text{coin 1 was tossed}.$$

Introduction

If we observed 2 heads, the **likelihood ratio** $P_0(2)/P_1(2)$ is

```
R> dbinom(2, 10, 0.5) / dbinom(2, 10, 0.7)
```

```
[1] 30.37623
```

(as the number of heads is binomial with $n = 10$ and probability 0.5 or 0.7, respectively).

This strongly favors coin 0, so we would decide for H_0.

Introduction

If we observed 2 heads, the **likelihood ratio** $P_0(2)/P_1(2)$ is

```
R> dbinom(2, 10, 0.5) / dbinom(2, 10, 0.7)
```

```
[1] 30.37623
```

(as the number of heads is binomial with $n = 10$ and probability 0.5 or 0.7, respectively).

This strongly favors coin 0, so we would decide for H_0.

If we observed 8 heads,

```
R> dbinom(8, 10, 0.5) / dbinom(8, 10, 0.7)
```

```
[1] 0.1882232
```

would favor coin 1, so we would decide for H_1.

Introduction

If we have prior “beliefs” about the hypotheses, we can easily extend the above idea to a Bayesian approach:

We need to specify prior probabilities $\mathbb{P}(H_0)$ and $\mathbb{P}(H_1)$.

In the “basic” case of no a priori preference for either hypothesis,

$$\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2.$$

Introduction

After observing the data we can compute the posterior probabilities

$$\mathbb{P}(H_0|x) = \frac{\mathbb{P}(H_0, x)}{\mathbb{P}(x)} = \frac{\mathbb{P}(x|H_0)\mathbb{P}(H_0)}{\mathbb{P}(x)}, \qquad \mathbb{P}(H_1|x) = \frac{\mathbb{P}(x|H_1)\mathbb{P}(H_1)}{\mathbb{P}(x)}.$$

The corresponding ratio of posterior probabilities is

$$\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)} = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \frac{\mathbb{P}(x|H_0)}{\mathbb{P}(x|H_1)}.$$

I.e., the ratio of posteriors is the product of the ratio of the priors and the likelihood ratio.

Introduction

How to decide?

Introduction

How to decide? Reasonably, choose the hypothesis **with higher posterior probability**.

I.e., choose H_0 if

$$\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)} = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \frac{\mathbb{P}(x|H_0)}{\mathbb{P}(x|H_1)} > 1 \quad \Leftrightarrow \frac{\mathbb{P}(x|H_0)}{\mathbb{P}(x|H_1)} > \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)}.$$

(Clearly, it is not clear what to do when the posterior probabilities are the same. More on this later.)

Introduction

I.e., we get decision rules of the form

$$\text{likelihood ratio} = \frac{\mathbb{P}(x|H_0)}{\mathbb{P}(x|H_1)} > c$$

where the **critical value** c depends upon the prior probabilities.

Introduction

In our case, the likelihood ratios for the possible values $x = 0, \ldots, 10$ are

```
R> x <- 0 : 10
R> dbinom(x, 10, 0.5) / dbinom(x, 10, 0.7)
```

```
 [1] 165.38171688  70.87787866  30.37623371  13.01838588   5.57930823
 [6]   2.39113210   1.02477090   0.43918753   0.18822323   0.08066710
[11]   0.03457161
```

If e.g. $c = 1$, $\mathbb{P}(H_0) = \mathbb{P}(H_1)$, and we choose H_0 as long as $X \leq 6$.

If e.g. $c = 0.1$, $\mathbb{P}(H_0) = 10\,\mathbb{P}(H_1)$, and we choose H_0 as long as $X \leq 8$.

Introduction

When deciding for H_0 or H_1, we can make two errors:

- choose H_1 when H_0 is "true"
- choose H_0 when H_1 is "true".

Introduction

If $c = 1$, the corresponding error probabilities are

$$\mathbb{P}(\text{choose } H_1|H_0) = \mathbb{P}(X > 6|H_0), \qquad \mathbb{P}(\text{choose } H_0|H_1) = \mathbb{P}(X \leq 6|H_1)$$

with corresponding values

```
R> pbinom(6, 10, 0.5, lower.tail = FALSE)
[1] 0.171875
R> pbinom(6, 10, 0.7)
[1] 0.3503893
```

respectively.

Introduction

If $c = 10$, the corresponding error probabilities are

$$\mathbb{P}(\text{choose } H_1|H_0) = \mathbb{P}(X > 8|H_0), \qquad \mathbb{P}(\text{choose } H_0|H_1) = \mathbb{P}(X \leq 8|H_1)$$

with corresponding values

```
R> pbinom(8, 10, 0.5, lower.tail = FALSE)
[1] 0.01074219
R> pbinom(8, 10, 0.7)
[1] 0.8506917
```

respectively.

Introduction

In our introductory example, both hypotheses completely specified the probability distribution of the data (number of heads) as binomial with parameters 10 and 0.5 or 0.7, respectively: such hypotheses are called **simple** hypotheses.

Hypotheses which are not simple are called **composite**.

Introduction

In our introductory example, both hypotheses completely specified the probability distribution of the data (number of heads) as binomial with parameters 10 and 0.5 or 0.7, respectively: such hypotheses are called **simple** hypotheses.

Hypotheses which are not simple are called **composite**.

What we've seen is that for choosing between two simple hypotheses, it is reasonable to look at the likelihood ratio $\mathbb{P}(x|H_0)/\mathbb{P}(x|H_1)$ and use decision rules of the form

- choose H_0 if the likelihood ratio is large (enough)
- choose H_1 if the likelihood ratio is small (enough)

This can be generalized to Bayesian hypothesis testing.

Outline

- Estimation of parameters and fitting of probability distributions
- Testing hypotheses and assessing goodness of fit
  - Introduction
  - The Neyman-Pearson paradigm

The Neyman-Pearson paradigm

The Neyman and Pearson approach to hypothesis testing is also formulated in the framework of (binary) decision problems.

However, it bypasses the necessity of specifying prior probabilities, and introduces a fundamental asymmetry between the two hypotheses, now referred to as

- the **null hypothesis** H_0
- the **alternative hypothesis** H_A.

The decisions now become

accept H_0 (“choose H_0”), **reject** H_0 (“choose H_A”).

The Neyman-Pearson paradigm

Terminology:

- Rejecting H_0 when it is true is a **type I error**.

The Neyman-Pearson paradigm

Terminology:

- Rejecting H_0 when it is true is a **type I error**.
- Probability of a type I error: **size** of the test, often denoted by α.

The Neyman-Pearson paradigm

Terminology:

- Rejecting H_0 when it is true is a **type I error**.
- Probability of a type I error: **size** of the test, often denoted by α.
- Accepting H_0 when it is false is a **type II error**.

The Neyman-Pearson paradigm

Terminology:

- Rejecting H_0 when it is true is a **type I error**.
- Probability of a type I error: **size** of the test, often denoted by α.
- Accepting H_0 when it is false is a **type II error**.
- Probability of a type II error is typically denoted by β.

The Neyman-Pearson paradigm

Terminology:

- Rejecting H_0 when it is true is a **type I error**.
- Probability of a type I error: **size** of the test, often denoted by α.
- Accepting H_0 when it is false is a **type II error**.
- Probability of a type II error is typically denoted by β.
- The probability of rejecting H_0 when it is false: **power** of the test, equals $1-\beta$.

The Neyman-Pearson paradigm

Terminology:

- Rejecting H_0 when it is true is a **type I error**.
- Probability of a type I error: **size** of the test, often denoted by α.
- Accepting H_0 when it is false is a **type II error**.
- Probability of a type II error is typically denoted by β.
- The probability of rejecting H_0 when it is false: **power** of the test, equals $1-\beta$.

The Neyman-Pearson paradigm

Terminology:

- Testing is based on a **test statistic** (e.g., the likelihood ratio) computed from the data.

The Neyman-Pearson paradigm

Terminology:

- Testing is based on a **test statistic** (e.g., the likelihood ratio) computed from the data.
- Sets of values leading to acceptance or rejection of H_0: **acceptance region** and **rejection region**, respectively.

The Neyman-Pearson paradigm

Terminology:

- Testing is based on a **test statistic** (e.g., the likelihood ratio) computed from the data.
- Sets of values leading to acceptance or rejection of H_0: **acceptance region** and **rejection region**, respectively.
- Probability distribution of the test statistic when H_0 is true: **null distribution**.