Statistics 2 Unit 2

Kurt Hornik

Outline

- Estimation of parameters and fitting of probability distributions

Outline

- Estimation of parameters and fitting of probability distributions
  - The method of moments
  - The method of maximum likelihood

Example: Gamma distribution

For the Gamma (and hence in particular the exponential) distribution, there are two alternative parametrizations: In R (see also `http://en.wikipedia.org/wiki/Gamma_distribution`), the **shape** parameter α and the **scale** parameter s are used, with corresponding density:

$$f(t) = \frac{t^{\alpha-1}e^{-t/s}}{s^{\alpha}\Gamma(\alpha)}, \qquad t > 0.$$

In Rice, the **rate** parameter $\lambda = 1/s$ is used instead of the scale parameter s:

$$f(t) = \frac{\lambda^{\alpha}t^{\alpha-1}e^{-\lambda t}}{\Gamma(\alpha)}, \qquad t > 0.$$

Example: Gamma distribution

Of course one can simply use $s \longleftrightarrow 1/\lambda$ to move between the parametrizations.

When we refer to the parameters of the Gamma distribution, we shall always explicitly indicate whether the second parameter is scale or rate.

Example: Gamma distribution

Let us find the method of moment estimates for the shape parameter α and rate parameter λ of the Gamma distribution.

Substituting $u = \lambda t$ we find in general that

$$\begin{aligned}
\mu_k &= \int_0^\infty t^k \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}\, dt \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{u}{\lambda}\right)^{\alpha+k-1} e^{-u}\, \frac{du}{u} \\
&= \frac{1}{\Gamma(\alpha)\lambda^k} \int_0^\infty u^{\alpha+k-1} e^{-u}\, du \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\lambda^k}.
\end{aligned}$$

Example: Gamma distribution

Hence,

$$\mu_k = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\lambda^k} = \frac{\alpha \times \cdots \times (\alpha+k-1)}{\lambda^k}$$

and in particular,

$$\mu_1 = \frac{\alpha}{\lambda}, \qquad \mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

This expresses μ_1 and μ_2 as functions of α and λ. We need to invert this relation to express α and λ as functions of μ_1 and μ_2 (i.e., solve the system of 2 non-linear equations in 2 variables).

Example: Gamma distribution

From the second equation,

$$\mu_2 = \frac{\alpha}{\lambda}\left(\frac{\alpha}{\lambda} + \frac{1}{\lambda}\right) = \mu_1\left(\mu_1 + \frac{1}{\lambda}\right).$$

Thus,

$$\frac{\mu_2 - \mu_1^2}{\mu_1} = \frac{1}{\lambda} \Rightarrow \lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$$

and

$$\alpha = \mu_1\lambda = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

Example: Gamma distribution

This gives the MoM estimates

$$\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}, \qquad \hat{\lambda} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

As before, with

$$\hat{\mu}_1 = \bar{X}, \qquad \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \hat{\sigma}^2$$

we can write the MoM estimates as

$$\hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}, \qquad \hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2}.$$

Example: Gamma distribution

What about the sampling distribution of the estimate?

(Again, we estimate two parameters, so this is a bivariate distribution.)

Example: Gamma distribution

What about the sampling distribution of the estimate?

(Again, we estimate two parameters, so this is a bivariate distribution.)

Well, this is not “well known”: it does not have a name, and there are no ready-made dpqr functions for it.

What can we do?

Example: Gamma distribution

What about the sampling distribution of the estimate?

(Again, we estimate two parameters, so this is a bivariate distribution.)

Well, this is not “well known”: it does not have a name, and there are no ready-made dpqr functions for it.

What can we do?

Easy in theory: use simulation. We’re looking for the distribution of

$$\left(\frac{\bar{X}^2}{\hat{\sigma}^2}, \frac{\bar{X}}{\hat{\sigma}^2} \right)$$

where $X_1, \ldots, X_n$ are i.i.d. Gamma with shape α and rate λ.

So we could generate B such samples of size n, and approximate the underlying distribution by the empirical distribution.

Example: Gamma distribution

Ahem . . . but we don't know α and λ, so how can we simulate?

Well, we have the MoM estimates $\hat{\alpha}$ and $\hat{\lambda}$, so perform the simulation using these parameters.

This gives the following simple **bootstrap** procedure for approximating the sampling distribution (in general) and standard errors of the estimates (in particular).

Example: Gamma distribution

1. Draw B samples of size n from the Gamma distribution with shape and rate parameters $\hat{\alpha}$ and $\hat{\lambda}$.

Example: Gamma distribution

1. Draw B samples of size n from the Gamma distribution with shape and rate parameters $\hat{\alpha}$ and $\hat{\lambda}$.
2. In each bootstrap sample, estimate the parameters (using the method of moments), giving estimates α_b^* and λ_b^*.

Example: Gamma distribution

1. Draw B samples of size n from the Gamma distribution with shape and rate parameters $\hat{\alpha}$ and $\hat{\lambda}$.
2. In each bootstrap sample, estimate the parameters (using the method of moments), giving estimates α_b^* and λ_b^*.
3. Estimate standard errors as

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\alpha_b^* - \bar{\alpha})^2}$$

where $\bar{\alpha} = B^{-1}\sum_{b=1}^{B}\alpha_b^*$, and similarly for $s_{\hat{\lambda}}$.

Example: Gamma distribution

1. Draw B samples of size n from the Gamma distribution with shape and rate parameters $\hat{\alpha}$ and $\hat{\lambda}$.
2. In each bootstrap sample, estimate the parameters (using the method of moments), giving estimates α_b^* and λ_b^*.
3. Estimate standard errors as

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\alpha_b^* - \bar{\alpha})^2}$$

where $\bar{\alpha} = B^{-1}\sum_{b=1}^{B}\alpha_b^*$, and similarly for $s_{\hat{\lambda}}$.

(Alternatively, use the standard deviation of $\alpha_1^*, \ldots, \alpha_B^*$.)

Consistency

Of course, using these “plug-in” procedures where we substitute a parameter θ by an estimate $\hat{\theta}$ only makes sense if the latter is close to the former.

Definition (Consistency). *Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges to θ in probability as $n \to \infty$.*

Similarly, $\hat{\theta}_n$ is strongly consistent if $\hat{\theta}_n \to \theta$ almost surely as $n \to \infty$.

Consistency

Of course, using these “plug-in” procedures where we substitute a parameter θ by an estimate $\hat{\theta}$ only makes sense if the latter is close to the former.

Definition (Consistency). *Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges to θ in probability as $n \to \infty$.*

Similarly, $\hat{\theta}_n$ is strongly consistent if $\hat{\theta}_n \to \theta$ almost surely as $n \to \infty$.

Again, note that we do not know the underlying parameter θ!

Consistency

In the above cases, we can use LLNs to establish that sample moments converge to (population) moments.

Hence, if

$$\theta = (\theta_1, \ldots, \theta_m) = h(\mu_1, \ldots, \mu_m)$$

with h **continuous**, we will have

$$\hat{\theta}_{MoM} = h(\hat{\mu}_1, \ldots, \hat{\mu}_m) \to \theta$$

as $n \to \infty$, in probability or almost surely.

Consistency

Similarly, if the standard errors are of the form

$$\sigma_{\hat{\theta}} = h(\theta)/\sqrt{n},$$

and h is continuous, then for the plug-in estimate

$$s_{\hat{\theta}} = h(\hat{\theta})/\sqrt{n}$$

we will have

$$\sigma_{\hat{\theta}}/s_{\hat{\theta}} \to 1$$

as $n \to \infty$, in probability or almost surely.

Outline

- Estimation of parameters and fitting of probability distributions
  - The method of moments
  - The method of maximum likelihood

Motivation

Suppose you observe a sample $x_1, \ldots, x_n$ from a discrete distribution with unknown parameter θ.

Consider the probability mass function (called frequency function in Rice)

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | \theta),$$

where the "conditioning" on θ is used to explicitly indicate that the probability is computed for a specific value θ of the unknown parameter.

The above gives the "likelihood" (probability) of observing what you observed.

When estimating θ, would you rather take θ to make the above large or small?

Motivation

Well, the smaller, the more unlikely is observing what we observed.

In extremis, observing what we observed becomes impossible. Strange.

Motivation

Well, the smaller, the more unlikely is observing what we observed.

In extremis, observing what we observed becomes impossible. Strange.

So clearly, it makes much more sense to choose θ so that the likelihood is large, perhaps even as large as possible.

This is the principle of **maximum likelihood estimation** (MLE).

The method of maximum likelihood

In general, suppose random variables $X_1, \ldots, X_n$ have a joint density

$$f(x_1, \ldots, x_n|\theta).$$

(One can consider the usual densities as “with respect to Lebesgue measure” and probability mass functions as densities “with respect to counting measure”.)

The maximum likelihood estimate of θ is the (if unique) value of θ that maximizes $f(x_1, \ldots, x_n|\theta)$, thus making the observed $x_1, \ldots, x_n$ “most likely”.

The method of maximum likelihood

More formally, write

$$\text{lik}(\theta|x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta)$$

to express the fact that the likelihood function is a function of the unknown parameter for fixed observations $x_1, \ldots, x_n$.

Often (as in Rice) one simply writes $\text{lik}(\theta)$ omitting the dependence on the observations.

The MLE is obtained by maximizing $\text{lik}(\theta|x_1, \ldots, x_n)$ over θ, ideally finding

$$\hat{\theta}_{MLE}(x_1, \ldots, x_n) = \arg\max_{\theta} \text{lik}(\theta|x_1, \ldots, x_n).$$

The method of maximum likelihood

Typically, it is more convenient to work with the (natural) logarithm of the likelihood, the so-called **log-likelihood** $\ell = \log(\text{lik})$.

The method of maximum likelihood

Typically, it is more convenient to work with the (natural) logarithm of the likelihood, the so-called **log-likelihood** $\ell = \log(\text{lik})$.

As log is increasing, maximizing the likelihood is equivalent to maximizing the log-likelihood.

The method of maximum likelihood

Typically, it is more convenient to work with the (natural) logarithm of the likelihood, the so-called **log-likelihood** $\ell = \log(\text{lik})$.

As log is increasing, maximizing the likelihood is equivalent to maximizing the log-likelihood.

(Notation follows Rice. Not my favorite notation: I would write L and LL for likelihood and log-likelihood, respectively.)

The method of maximum likelihood

If the X_i are i.i.d., the joint density is the product of the marginal densities:

$$\text{lik}(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

and the log-likelihood becomes the sum of the marginal log-densities:

$$\ell(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} \log(f(x_i|\theta).$$

Example: Poisson distribution

If $X_1, \ldots, X_n$ are i.i.d. Poisson(λ), the log-likelihood is

$$\begin{aligned}
\ell(\lambda | x_1, \ldots, x_n) &= \sum_{i=1}^{n} \log\left(\mathbb{P}(X_i = x_i | \lambda)\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right) \\
&= \sum_{i=1}^{n} (x_i \log(\lambda) - \log(x_i!) - \lambda) \\
&= \log(\lambda) \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!).
\end{aligned}$$

Example: Poisson distribution

To maximize with respect to λ, compute the derivative and set it to zero:

$$\ell'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$$

from which the MLE for the sample $x_1, \ldots, x_n$ is obtained as

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

(One can easily verify that the critical point indeed gives the maximum.)

Example: Poisson distribution

To maximize with respect to λ, compute the derivative and set it to zero:

$$\ell'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$$

from which the MLE for the sample $x_1, \ldots, x_n$ is obtained as

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

(One can easily verify that the critical point indeed gives the maximum.)

The MLE agrees with the MoM estimate, and thus has the same sampling distribution.

Example: Normal distribution

If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$,

$$f(x_1, \ldots, x_n | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x_i - \mu)^2/(2\sigma^2))$$

The log-likelihood for $\theta = (\mu, \sigma^2)$ is thus

$$\begin{aligned} \ell(\mu, \sigma^2 | x_1, \ldots, x_n) &= \sum_{i=1}^{n} \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ &= -\frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \end{aligned}$$

Example: Normal distribution

To maximize with respect to μ and σ^2, compute the partial derivatives and set these to zero.

This first gives

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \times (-2) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$

and

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{2\sigma^2}\left(-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Example: Normal distribution

Setting the partials to zero then yields

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

(Again, one can easily verify that the critical point indeed gives the maximum.)

Example: Normal distribution

Setting the partials to zero then yields

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

(Again, one can easily verify that the critical point indeed gives the maximum.)

Again, the MLE agrees with the MoM estimate, and thus has the same sampling distribution.

Example: Normal distribution

Setting the partials to zero then yields

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2.$$

(Again, one can easily verify that the critical point indeed gives the maximum.)

Again, the MLE agrees with the MoM estimate, and thus has the same sampling distribution.

Note that the MLE of the variance again is not the sample variance.

Example: Gamma distribution

If $X \sim \text{Gamma}(\alpha, \text{rate} = \lambda)$, the density is

$$f(x|\alpha, \lambda) = \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \qquad x > 0.$$

Example: Gamma distribution

If $X \sim \text{Gamma}(\alpha, \text{rate} = \lambda)$, the density is

$$f(x|\alpha,\lambda) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \qquad x > 0.$$

Thus, the log-likelihood for observations $x_1, \ldots, x_n$ from from $X_1, \ldots, X_n$ i.i.d. Gamma$(\alpha, \text{rate} = \lambda)$ is

$$\begin{aligned}
&\ell(\alpha,\lambda|x_1,\ldots,x_n) \\
&= \sum_{i=1}^n \log\left(\frac{\lambda^\alpha x_i^{\alpha-1} e^{-\lambda x_i}}{\Gamma(\alpha)}\right) \\
&= n\alpha\log(\lambda) + (\alpha-1)\sum_{i=1}^n \log(x_i) - \lambda\sum_{i=1}^n x_i - n\log(\Gamma(\alpha)).
\end{aligned}$$

Example: Gamma distribution

To maximize with respect to α and λ, we could again try to compute the partial derivatives and set these to zero. For the partials, we get

$$\begin{aligned}\frac{\partial \ell}{\partial \alpha} &= \frac{\partial \ell}{\partial \alpha}\left(n\alpha \log(\lambda) + (\alpha - 1)\sum_{i=1}^{n} \log(x_i) - \lambda \sum_{i=1}^{n} x_i - n \log(\Gamma(\alpha))\right) \\ &= n \log(\lambda) + \sum_{i=1}^{n} \log(x_i) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\end{aligned}$$

and

$$\frac{\partial \ell}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i.$$

Example: Gamma distribution

Setting $\partial \ell / \partial \lambda = 0$ gives

$$\hat{\lambda} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n} x_i} = \frac{\hat{\alpha}}{\bar{x}}.$$

Substituting into $\partial \ell / \partial \alpha = 0$ gives

$$n \log\left(\frac{\hat{\alpha}}{\bar{x}}\right) + \sum_{i=1}^{n} \log(x_i) - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0$$

This is a non-linear equation for the MLE of α which we cannot solve "explicitly".

Example: Gamma distribution

With R, we can simply compute the MLEs via numerical optimization (remember the examples for Poisson and normal in the last unit of Computing).

To approximate the sampling distributions, we can again use the bootstrap.

Comparing with the approximation for the MoM estimates would show that the distributions for the MLE are substantially less dispersed.

Consistency of the MLE

Theorem (Consistency of the MLE). *Under appropriate smoothness conditions on f, the MLE from i.i.d. samples from f is consistent.*

Consistency of the MLE

Theorem (Consistency of the MLE). *Under appropriate smoothness conditions on f, the MLE from i.i.d. samples from f is consistent.*

Proof/sketch. Consider maximizing

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

If θ_0 is the underlying parameter, i.e., if $X_1, \ldots, X_n$ are i.i.d. with density $f(x|\theta_0)$, then as $n \to \infty$, by the law of large numbers:

$$\frac{\ell_n(\theta)}{n} \to \mathbb{E}_{\theta_0} \log(f(X_i|\theta)) = \int \log(f(x|\theta)) f(x|\theta_0)\, dx.$$

Consistency of the MLE

Suppose we can show that as $n \to \infty$, the θ maximizing $\ell_n(\theta)/n$ converges to the θ maximizing $\lim_{n\to\infty} \ell_n(\theta)/n$.

(This is far from being straightforward, and needs f to be nice enough.)

Consistency of the MLE

Suppose we can show that as $n \to \infty$, the θ maximizing $\ell_n(\theta)/n$ converges to the θ maximizing $\lim_{n\to\infty} \ell_n(\theta)/n$.

(This is far from being straightforward, and needs f to be nice enough.)

Then what remains to be shown is that the limit is maximized at θ_0.

Consistency of the MLE

Suppose we can show that as $n \to \infty$, the θ maximizing $\ell_n(\theta)/n$ converges to the θ maximizing $\lim_{n\to\infty} \ell_n(\theta)/n$.

(This is far from being straightforward, and needs f to be nice enough.)

Then what remains to be shown is that the limit is maximized at θ_0.

Consider the function

$$h(t) = t\log(t) - t + 1$$

for $t \geq 0$. Then

$$h'(t) = \log(t) + t \times \frac{1}{t} - 1 = \log(t), \qquad h''(t) = \frac{1}{t}$$

so that h has its minimum at $t = 1$ with value $h(1) = 0$.

Consistency of the MLE

Hence, for all $t \geq 0$,

$$h(t) = t\log(t) - t + 1 \geq 0$$

with equality iff $t = 1$, and thus for all $u, v \geq 0$,

$$vh\left(\frac{u}{v}\right) = v\left(\frac{u}{v}\log\left(\frac{u}{v}\right) - \frac{u}{v} + 1\right) = u\log\left(\frac{u}{v}\right) - u + v \geq 0$$

with equality iff $u = v$.

Consistency of the MLE

Hence, for all $t \geq 0$,

$$h(t) = t\log(t) - t + 1 \geq 0$$

with equality iff $t = 1$, and thus for all $u, v \geq 0$,

$$vh\left(\frac{u}{v}\right) = v\left(\frac{u}{v}\log\left(\frac{u}{v}\right) - \frac{u}{v} + 1\right) = u\log\left(\frac{u}{v}\right) - u + v \geq 0$$

with equality iff $u = v$.

Now take $u = f(x|\theta_0)$ and $v = f(x|\theta)$. Then for all x,

$$0 \leq f(x|\theta_0)\log\left(\frac{f(x|\theta_0)}{f(x|\theta)}\right) - f(x|\theta_0) + f(x|\theta).$$

Hence,

Consistency of the MLE

$$
\begin{aligned}
0 &\leq \int \left(f(x|\theta_0) \log\left(\frac{f(x|\theta_0)}{f(x|\theta)}\right) - f(x|\theta_0) + f(x|\theta) \right) dx \\
&= \int \log(f(x|\theta_0)) f(x|\theta_0)\, dx - \int \log(f(x|\theta)) f(x|\theta_0)\, dx \\
&\quad - \int f(x|\theta_0)\, dx + \int f(x|\theta)\, dx.
\end{aligned}
$$

As densities integrate to 1, this yields

$$
\int \log(f(x|\theta)) f(x|\theta_0)\, dx \leq \int \log(f(x|\theta_0)) f(x|\theta_0)\, dx
$$

with strict inequality unless the densities agree.

Consistency of the MLE

Hence, unless densities could agree for different parameters, the underlying parameter is the unique maximizer.

Fisher information

Suppose that f is nice enough (more below) and consider the random variable

$$s(\theta) = \nabla_\theta \log(f(X|\theta)) = \left[\frac{\partial \log(f(X|\theta))}{\partial \theta_j} \right]'.$$

This is the gradient of the log-density, which connaisseurs call the **score function**.

Fisher information

Clearly,

$$\int f(x|\theta)\,dx = 1 \quad \Longrightarrow \quad \frac{\partial}{\partial \theta_j} \int f(x|\theta)\,dx = 0$$

Now assume that we may change integration and differentiation (which in particular needs the support of f to not depend on θ), and remember that

$$\frac{\partial \log(f(x|\theta))}{\partial \theta_j} = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta_j} \quad \Longrightarrow \quad \frac{\partial f(x|\theta)}{\partial \theta_j} = \frac{\partial \log(f(x|\theta))}{\partial \theta_j} f(x|\theta).$$

Fisher information

Then,

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \theta_j} \int f(x|\theta)\, dx \\
&= \int \frac{\partial f(x|\theta)}{\partial \theta_j}\, dx \\
&= \int \frac{\partial \log(f(x|\theta))}{\partial \theta_j} f(x|\theta)\, dx \\
&= \mathbb{E}_\theta \frac{\partial \log(f(X|\theta))}{\partial \theta_j}.
\end{aligned}
$$

Thus, the score has mean zero:

$$\mathbb{E}_\theta(s(\theta)) = 0.$$

Fisher information

The covariance matrix of $s(\theta)$ is called the **Fisher information** matrix:

$$I(\theta) = \text{cov}_\theta(s(\theta)) = \mathbb{E}_\theta(s(\theta)s(\theta)').$$

Explicitly, the (j, k) element of $I(\theta)$ is

$$[I(\theta)]_{j,k} = \mathbb{E}_\theta\left(\frac{\partial \log(f(X|\theta))}{\partial \theta_j}\frac{\partial \log(f(X|\theta))}{\partial \theta_k}\right).$$

Fisher information

Under appropriate smoothness conditions on f, $I(\theta)$ may also be expressed as

$$I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log(f(X|\theta))}{\partial\theta\partial\theta'}\right) = -\mathbb{E}_\theta\left(H_\theta(\log(f(X|\theta)))\right)$$

where H_θ denotes the Hessian with respect to θ.

Explicitly, the (j,k) element of $I(\theta)$ is

$$[I(\theta)]_{j,k} = -\mathbb{E}_\theta\left(\frac{\partial^2 \log(f(X|\theta))}{\partial\theta_j\partial\theta_k}\right).$$

Fisher information

To see why, take the above

$$0 = \int \frac{\partial \log(f(x|\theta))}{\partial \theta_j} f(x|\theta)\, dx$$

and differentiate once more with respect to θ_k.

Fisher information

Then

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \theta_k}\left(\int \frac{\partial \log(f(x|\theta))}{\partial \theta_j} f(x|\theta)\, dx\right) \\
&= \int \frac{\partial}{\partial \theta_k}\left(\frac{\partial \log(f(x|\theta))}{\partial \theta_j} f(x|\theta)\right) dx \\
&= \int \left(\frac{\partial^2 \log(f(x|\theta))}{\partial \theta_j \partial \theta_k} f(x|\theta) + \frac{\partial \log(f(x|\theta))}{\partial \theta_j}\frac{\partial f(x|\theta)}{\partial \theta_k}\right) dx \\
&= \int \left(\frac{\partial^2 \log(f(x|\theta))}{\partial \theta_j \partial \theta_k} f(x|\theta) + \frac{\partial \log(f(x|\theta))}{\partial \theta_j}\frac{\partial \log(f(x|\theta))}{\partial \theta_k} f(x|\theta)\right) dx.
\end{aligned}
$$

Fisher information

I.e.,

$$\int \frac{\partial^2 \log(f(x|\theta))}{\partial\theta_j \partial\theta_k} f(x|\theta)\, dx = -\int \frac{\partial \log(f(x|\theta))}{\partial\theta_j} \frac{\partial \log(f(x|\theta))}{\partial\theta_k} f(x|\theta)\, dx$$

or equivalently,

$$\mathbb{E}_\theta\left(\frac{\partial^2 \log(f(X|\theta))}{\partial\theta_j \partial\theta_k}\right) = -\mathbb{E}_\theta\left(\frac{\partial \log(f(X|\theta))}{\partial\theta_j} \frac{\partial \log(f(X|\theta))}{\partial\theta_k}\right).$$

Asymptotic normality of the MLE

Theorem (Asymptotic normality of the MLE). *Under appropriate smoothness conditions on f, the MLE $\hat{\theta}$ from i.i.d. samples from f satisfies*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1}).$$

(The RHS is a multivariate normal distribution with mean zero and covariance the inverse of the Fisher information matrix $I(\theta_0)$.)

Asymptotic normality of the MLE

Proof/sketch. From a Taylor series expansion,

$$0 = \nabla \ell(\hat{\theta}) \approx \nabla_\theta \ell(\theta_0) + H_\theta \ell(\theta_0)(\hat{\theta} - \theta_0)$$

from which

$$\begin{aligned} H_\theta \ell(\theta_0)(\hat{\theta} - \theta_0) &\approx -\nabla_\theta \ell(\theta_0) \quad \Longrightarrow \\ \hat{\theta} - \theta_0 &\approx -(H_\theta \ell(\theta_0))^{-1} \nabla_\theta \ell(\theta_0). \end{aligned}$$

and thus

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\left(\frac{H_\theta \ell(\theta_0)}{n} \right)^{-1} \frac{\nabla_\theta \ell(\theta_0)}{\sqrt{n}}.$$

Asymptotic normality of the MLE

Proof/sketch. From a Taylor series expansion,

$$0 = \nabla \ell(\hat{\theta}) \approx \nabla_\theta \ell(\theta_0) + H_\theta \ell(\theta_0)(\hat{\theta} - \theta_0)$$

from which

$$\begin{aligned} H_\theta \ell(\theta_0)(\hat{\theta} - \theta_0) &\approx -\nabla_\theta \ell(\theta_0) \quad \Longrightarrow \\ \hat{\theta} - \theta_0 &\approx -(H_\theta \ell(\theta_0))^{-1} \nabla_\theta \ell(\theta_0). \end{aligned}$$

and thus

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\left(\frac{H_\theta \ell(\theta_0)}{n} \right)^{-1} \frac{\nabla_\theta \ell(\theta_0)}{\sqrt{n}}.$$

We now show that the 1st fraction satisfies an LLN and the 2nd a CLT.

Asymptotic normality of the MLE

As $X_1, \ldots, X_n$ are i.i.d.,

$$\ell(\theta) = \sum_{i=1}^{n} \log(f(X_i|\theta)).$$

By what we just established, if θ_0 is the underlying parameter,

- the random variables $\nabla_\theta \log(f(X_1|\theta_0)), \ldots, \nabla_\theta \log(f(X_n|\theta_0))$ are i.i.d. with mean zero and covariance matrix $I(\theta_0)$

Asymptotic normality of the MLE

As $X_1, \ldots, X_n$ are i.i.d.,

$$\ell(\theta) = \sum_{i=1}^{n} \log(f(X_i|\theta)).$$

By what we just established, if θ_0 is the underlying parameter,

- the random variables $\nabla_\theta \log(f(X_1|\theta_0)), \ldots, \nabla_\theta \log(f(X_n|\theta_0))$ are i.i.d. with mean zero and covariance matrix $I(\theta_0)$
- the random variables $H_\theta \log(f(X_1|\theta_0)), \ldots, H_\theta \log(f(X_n|\theta_0))$ are i.i.d. with mean $-I(\theta_0)$.

Asymptotic normality of the MLE

Hence, by the LLN (actually, the “obvious” multivariate generalization),

$$\frac{H_\theta \ell(\theta_0)}{n} = \frac{1}{n}\sum_{i=1}^{n} H_\theta \log(f(X_i|\theta_0)) \to \mathbb{E}_{\theta_0}\left(H_\theta \log(f(X|\theta_0))\right) = -I(\theta_0).$$

Asymptotic normality of the MLE

Hence, by the LLN (actually, the “obvious” multivariate generalization),

$$\frac{H_\theta \ell(\theta_0)}{n} = \frac{1}{n}\sum_{i=1}^{n} H_\theta \log(f(X_i|\theta_0)) \to \mathbb{E}_{\theta_0}\left(H_\theta \log(f(X|\theta_0))\right) = -I(\theta_0).$$

And by the CLT (actually, an “obvious” multivariate generalization),

$$\frac{\nabla_\theta \ell(\theta_0)}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta \log(f(X_i|\theta_0)) \xrightarrow{d} N(0, I(\theta_0)).$$

Asymptotic normality of the MLE

Combining both,

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\left(\frac{H_\theta \ell(\theta_0)}{n}\right)^{-1} \frac{\nabla_\theta \ell(\theta_0)}{\sqrt{n}} \xrightarrow{d} (I(\theta_0))^{-1} N(0, I(\theta_0)).$$

Asymptotic normality of the MLE

Combining both,

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\left(\frac{H_\theta \ell(\theta_0)}{n}\right)^{-1} \frac{\nabla_\theta \ell(\theta_0)}{\sqrt{n}} \xrightarrow{d} (I(\theta_0))^{-1} N(0, I(\theta_0)).$$

Now if A is a matrix and Y has a multivariate normal distribution with mean 0 and covariance Σ, then AY has a multivariate normal distribution with mean zero and covariance $A\Sigma A'$.

Thus, with $A = (I(\theta_0))^{-1}$ and $\Sigma = I(\theta_0)$,

$$\begin{aligned} (I(\theta_0))^{-1} N(0, I(\theta_0)) &\stackrel{d}{=} N(0, (I(\theta_0))^{-1} I(\theta_0) (I(\theta_0))^{-1}) \\ &= N(0, I(\theta_0)^{-1}). \end{aligned}$$

Asymptotic normality of the MLE

The lecture notes explicitly handle the case where we have a single parameter only. But this has always been confusing, so now we do the real thing(s).

If there is a single parameter, $I(\theta)$ is a number. In this case, we can also write the result as

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, 1).$$

(If one write $S^{1/2}$ for the symmetric square root of a positive definite symmetric matrix, one can also generally write

$$(nI(\theta_0))^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I_m)$$

with I_m the $m \times m$ identity matrix.)

Confidence intervals

A **confidence interval** for a population parameter θ is a random interval which contains θ with some specified (coverage) probability.

A $100(1-\alpha)$ percent confidence interval contains θ with probability (at least) $1-\alpha$; if we took many random samples and formed confidence intervals from each one, $100(1-\alpha)$ percent of these would contain θ.

Confidence intervals are frequently used in conjunction with point estimates to convey information about the uncertainty of the estimates.

Example: Normal distribution

The MLEs of μ and σ^2 from an i.i.d. normal sample are

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

A confidence interval for μ is based on the fact that

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

where S^2 is the sample variance and t_{n-1} the (Student) t distribution with $n-1$ degrees of freedom.

Example: Normal distribution

Write $Q_F(\alpha)$ for the α quantile of distribution F.

Then

$$\mathbb{P}\left(Q_{t_{n-1}}(\alpha/2) \leq \frac{\sqrt{n}(\bar{X}-\mu)}{S} \leq Q_{t_{n-1}}(1-\alpha/2) \right) = 1-\alpha$$

and rearranging and using the symmetry of the t distribution gives

$$\mathbb{P}\left(\bar{X} - \frac{S}{\sqrt{n}} Q_{t_{n-1}}(1-\alpha/2) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} Q_{t_{n-1}}(1-\alpha/2) \right) = 1-\alpha.$$

Example: Normal distribution

To obtain a confidence interval for σ^2, note that

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$$

where χ^2_{n-1} denotes the chi-squared distribution with $n-1$ degrees of freedom.

Thus,

$$\mathbb{P}\left(Q_{\chi^2_{n-1}}(\alpha/2) \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq Q_{\chi^2_{n-1}}(1-\alpha/2)\right) = 1-\alpha,$$

and rearranging gives

$$\mathbb{P}\left(\frac{n\hat{\sigma}^2}{Q_{\chi^2_{n-1}}(1-\alpha/2)} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{Q_{\chi^2_{n-1}}(\alpha/2)}\right) = 1-\alpha.$$

Approximate confidence intervals

Where exact intervals cannot be obtained, we can use the fact that in general, $\sqrt{n}(\hat{\theta} - \theta_0)$ approximately has an $N(0, (I(\theta_0))^{-1})$ distribution.

The unknown $I(\theta_0)$ can be approximated by the plug-in estimate $I(\hat{\theta})$.

If θ is a single parameter, we have

$$I(\hat{\theta})/I(\theta_0) \to 1$$

and hence

$$\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) = \sqrt{\frac{I(\hat{\theta})}{I(\theta_0)}}\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, 1).$$

Approximate confidence intervals

Therefore

$$\mathbb{P}\left(z_{\alpha/2} \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z_{1-\alpha/2}\right) \approx 1 - \alpha,$$

and hence an approximate $100(1-\alpha)$ percent confidence interval is

$$\hat{\theta} \pm z_{\alpha/2}/\sqrt{nI(\hat{\theta})}.$$

Example: Poisson distribution

The MLE of the parameter λ from a sample from a Poisson distribution is $\hat{\lambda} = \bar{X}$.

The sampling distribution is known, but depends on the unknown parameter.

Approximate confidence intervals can be obtained from the above.

We have

$$\log(f(x|\lambda)) = \log\left(\frac{\lambda^x}{x!}e^{-\lambda}\right) = x\log(\lambda) - \log(x!) - \lambda$$

so that the Fisher information is given by

$$\mathbb{E}_\lambda\left(\frac{\partial \log(f(x|\lambda))}{\partial\lambda}\right)^2 = \mathbb{E}_\lambda\left(\frac{X}{\lambda} - 1\right)^2 = \mathbb{E}_\lambda\frac{(X-\lambda)^2}{\lambda^2} = \frac{\mathrm{var}_\lambda(X)}{\lambda^2} = \frac{1}{\lambda}.$$

Example: Poisson distribution

Thus, an approximate $100(1-\alpha)$ percent confidence interval for λ is given by

$$\hat{\lambda} \pm z_{\alpha/2}/\sqrt{nI(\hat{\lambda})} = \hat{\lambda} \pm z_{\alpha/2}/\sqrt{n/\hat{\lambda}} = \bar{X} \pm z_{\alpha/2}\sqrt{\bar{X}/n}.$$

Bootstrap confidence intervals

If the distribution of $\Delta = \hat{\theta} - \theta_0$ was known, confidence intervals could be obtained via

$$\mathbb{P}(Q_\Delta(\alpha/2) \leq \hat{\theta} - \theta_0 \leq Q_\Delta(1-\alpha/2)) = 1-\alpha$$

as

$$\mathbb{P}(\hat{\theta} - Q_\Delta(1-\alpha/2) \leq \theta_0 \leq \hat{\theta} - Q_\Delta(\alpha/2)) = 1-\alpha.$$

Bootstrap confidence intervals

If the distribution of $\Delta = \hat{\theta} - \theta_0$ was known, confidence intervals could be obtained via

$$\mathbb{P}(Q_\Delta(\alpha/2) \leq \hat{\theta} - \theta_0 \leq Q_\Delta(1-\alpha/2)) = 1-\alpha$$

as

$$\mathbb{P}(\hat{\theta} - Q_\Delta(1-\alpha/2) \leq \theta_0 \leq \hat{\theta} - Q_\Delta(\alpha/2)) = 1-\alpha.$$

But since θ_0 is not known, we use $\hat{\theta}$ in its place.

Bootstrap confidence intervals

We generate B bootstrap samples from the distribution with value $\hat{\theta}$, and compute the respective MLEs θ_b^*.

The distribution of $\hat{\theta} - \theta_0$ is then approximated by that of $\theta^* - \hat{\theta}$ and the quantiles of this are used to form the approximate confidence interval.

I.e., the quantiles Q_Δ are approximated by the empirical quantiles of $(\theta_1^* - \hat{\theta}, \ldots, \theta_B^* - \hat{\theta})$.