A Very Brief Summary of Statistical Inference, and Examples

Trinity Term 2009 Prof. Gesine Reinert

Our standard situation is that we have data $\mathbf{x} = x_1, x_2, \ldots, x_n$, which we view as realisations of random variables $X_1, X_2, \ldots, X_n$ with a distribution (model) $f(x_1, x_2, \ldots, x_n; \theta)$, where θ is unknown. In *frequentist analysis*, $\theta \in \Theta$ is an unknown constant.

1. Likelihood and Sufficiency

(Fisherian) Likelihood approach: We define the likelihood of θ given the data as

$$L(\theta) = L(\theta, \mathbf{x}) = f(x_1, x_2, \ldots, x_n; \theta).$$

Often: $X_1, X_2, \ldots, X_n$ are independent, identically distributed (*i.i.d.*); then $L(\theta, \mathbf{x}) = \prod_{i=1}^n f(x_i, \theta)$.

We summarize information about θ: find a *minimal sufficient statistic* $t(\mathbf{x})$; from the Factorization Theorem: $T = t(\mathbf{X})$ is sufficient for θ if and only if there exists functions $g(t, \theta)$ and $h(\mathbf{x})$ such that for all $\mathbf{x}$ and θ

$$f(\mathbf{x}, \theta) = g(t(\mathbf{x}), \theta) h(\mathbf{x}).$$

Moveover $T = t(\mathbf{X})$ is minimal sufficient when it holds that

$$\frac{f(\mathbf{x}, \theta)}{f(\mathbf{y}, \theta)} \text{ is constant in } \theta \iff t(\mathbf{x}) = t(\mathbf{y})$$

Example
Let $X_1, \ldots, X_n$ be a random sample from the truncated exponential distribution, where

$$f_{X_i}(x_i) = e^{\theta - x_i},\ x_i > \theta$$

or, using the indicator function notation,

$$f_{X_i}(x_i) = e^{\theta - x_i} I_{(\theta, \infty)}(x_i).$$

Show that $Y_1 = \min(X_i)$ is sufficient for θ.

Let $T = T(X_1, \ldots, X_n) = Y_1$. We need the pdf $f_T(t)$ of the smallest order statistic. Calculate the cdf for X_i,

$$F(x) = \int_\theta^x e^{\theta - z} dz = e^\theta [e^{-\theta} - e^{-x}] = 1 - e^{\theta - x}.$$

Now

$$P(T > t) = \prod_{i=1}^n (1 - F(t)) = (1 - F(t))^n.$$

Differentiating gives that $f_T(t)$ equals

$$n[1-F(t)]^{n-1}f(t) = ne^{(\theta-t)(n-1)} \times e^{\theta-t} = ne^{n(\theta-t)}, t > \theta.$$

So the conditional density of $X_1, \ldots X_n$ given $T = t$ is

$$\frac{e^{\theta-x_1}e^{\theta-x_2}\ldots e^{\theta-x_n}}{ne^{n(\theta-t)}} = \frac{e^{-\sum x_i}}{ne^{-nt}},\ x_i \geq t,\ i = 1, \ldots n,$$

which does not depend on θ for each fixed $t = \min(x_i)$. Note that since $x_i \geq t, i = 1, \ldots, n$, neither the expression nor the range space depends on θ, so the first order statistic, $X_{(1)} = min(X_i)$, is a sufficient statistic for θ.

Alternatively, use the Factorization Theorem:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} e^{\theta-x_i} I_{(\theta,\infty)}(x_i) \\ &= I_{(\theta,\infty)}(x_{(1)}) \prod_{i=1}^{n} e^{\theta-x_i} \\ &= e^{n\theta} I_{(\theta,\infty)}(x_{(1)}) e^{-\sum_{i=1}^{n} x_i}. \end{aligned}$$

With

$$g(t,\theta) = e^{n\theta} I_{(\theta,\infty)}(t) \text{ and } h(\mathbf{x}) = e^{-\sum_{i=1}^{n} x_i}$$

we see that $X_{(1)}$ is minimal sufficient.

2. Point Estimation

Estimate θ by a function $t(x_1, \ldots, x_n)$ of the data; often by maximum-likelihood:

$$\hat{\theta} = arg\max_{\theta} L(\theta),$$

or by method of moments. Neither are unbiased in general, but the m.l.e. is asymptotically unbiased and asymptotically efficient; under some regularity assumptions,

$$\hat{\theta} \approx \mathcal{N}(\theta, I_n^{-1}(\theta)),$$

where

$$I_n(\theta) = \mathbf{E}\left[\left(\frac{\partial \ell(\theta, \mathbf{X})}{\partial \theta}\right)^2\right]$$

is the Fisher information (matrix). Under more regularity,

$$I_n(\theta) = -\mathbf{E}\left(\frac{\partial^2 \ell(\theta, \mathbf{X})}{\partial \theta^2}\right).$$

If $\mathbf{x}$ is a random sample, then $I_n(\theta) = nI_1(\theta)$; often we abbreviate $I_1(\theta) = I(\theta)$.

The m.l.e. is a function of a sufficient statistic; recall that, for scalar θ, there is a nice theory using the Cramer-Rao lower bound and the Rao-Blackwell theorem on how to obtain minimum variance unbiased estimators based on a sufficient statistic and an unbiased estimator. The mle possesses the *invariance property:* The m.l.e. of a function $\phi(\theta)$ is $\phi(\hat{\theta})$.

Example
Suppose $X_1, X_2, \ldots, X_n$ random sample from Gamma distribution with density

$$f(x; c, \beta) = \frac{x^{c-1}}{\Gamma(c)\beta^c} e^{-\frac{x}{\beta}}, \quad x > 0,$$

where $c > 0, \beta > 0$; $\theta = (c, \beta)$;

$$\ell(\theta) \quad = \quad -n \log \Gamma(c) - nc \log \beta + (c-1) \log \prod x_i - \frac{1}{\beta} \sum x_i$$

Put $D_1(c) = \frac{\partial}{\partial c} \log \Gamma(c)$, then

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= -\frac{nc}{\beta} + \frac{1}{\beta^2} \sum x_i \\ \frac{\partial \ell}{\partial c} &= -nD_1(c) - n \log \beta + \log(\prod x_i) \end{aligned}$$

Setting these equal to zero yields

$$\hat{\beta} = \frac{\overline{x}}{\hat{c}},$$

where $\hat{c}$ solves

$$D_1(\hat{c}) - \log(\hat{c}) = \log([\prod x_i]^{1/n} / \overline{x}).$$

(We could calculate that sufficient statistics is indeed $(\sum x_i, [\prod x_i)$, and is minimal sufficient)
We need to check the second derivatives to assure that we have a maximum; we leave this out here for time reasons. We calculate the Fisher information: Put $D_2(c) = \frac{\partial^2}{\partial c^2} \log \Gamma(c)$,

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta^2} &= \frac{nc}{\beta^2} - \frac{2}{\beta^3} \sum x_i \\ \frac{\partial^2 \ell}{\partial \beta \partial c} &= -\frac{n}{\beta} \\ \frac{\partial^2 \ell}{\partial c^2} &= -nD_2(c). \end{aligned}$$

We use that $EX_i = c\beta$ to obtain

$$I_n(\theta) = n \begin{pmatrix} \frac{c}{\beta^2} & \frac{1}{\beta} \\ \frac{1}{\beta} & D_2(c) \end{pmatrix}.$$

Note that if c was large, then by the CLT each observation X_j is approximately normal, because, in distribution, $X_j = Y_1 + Y_2 + \cdots + Y_k$, with $Y_i, i = 1, \ldots, k$, being i.i.d. $Gamma(c/k, \beta)$.

Recall also that there is an iterative method to compute m.l.e.s, related to the Newton-Raphson method.

3. Hypothesis Testing

For simple null hypothesis and simple alternative, the Neyman-Pearson Lemma says that the most powerful tests are likelihood-ratio tests. These can sometimes be generalized to one-sided alternatives in such a way as to yield uniformly most powerful tests.

Example
Suppose as above that $X_1, \ldots, X_n$ is a random sample from the truncated exponential distribution, where

$$f_{X_i}(x_i; \theta) = e^{\theta - x_i},\ x_i > \theta.$$

Find a UMP test of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$:

Let $\theta_1 > \theta_0$, then the LR is

$$\frac{f(\mathbf{x}, \theta_1)}{f(\mathbf{x}, \theta_0)} = e^{n(\theta_1 - \theta_0)} I_{(\theta_1, \infty)}(x_{(1)}),$$

which increases with $x_{(1)}$, so we reject H_0 if the smallest observation, $T = X_{(1)}$, is large. Under H_0, we have calculated that the pdf of f_T is

$$n e^{n(\theta_0 - y_1)}, y_1 > \theta_0.$$

So for $t > \theta$, under H_0,

$$P(T > t) = \int_t^\infty n e^{n(\theta_0 - s)} ds = e^{n(\theta_0 - t)}.$$

For a test at level α, choose t such that

$$t = \theta_0 - \frac{\ln \alpha}{n}.$$

The LR test rejects H_0 if $X_{(1)} > \theta_0 - \frac{\ln \alpha}{n}$. The test is the same for all $\theta > \theta_0$, so it is UMP.

For a general null hypothesis and a general alternative, $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1 = \Theta \backslash \Theta_0$, the (generalized) LR test uses the likelihood ratio statistic

$$T = \frac{\max\limits_{\theta \in \Theta} L(\theta; \mathbf{X})}{\max\limits_{\theta \in \Theta_0} L(\theta; \mathbf{X})}.$$

We reject H_0 for large values of T; we use the chisquare asymptotics $2\log T \approx \chi^2_p$, where $p = dim\Theta - dim\Theta_0$ (which is valid for nested models only). Alternatively, we could use score tests, which are based on the *score function* $\partial\ell/\partial\theta$, with asymptotics in terms of normal distribution $\mathcal{N}(0, I_n(\theta))$. Or, we could use a Wald test, which is based on the asymptotic normality of the m.l.e.; often $\hat{\theta} \approx \mathcal{N}\big(\theta, I_n^{-1}(\theta)\big)$ if θ is the true parameter. Recall: In a random sample, $I_n(\theta) = nI_1(\theta)$.

An important example is Pearson's Chisquare test, which we derived as a score test, and we saw that it is asymptotically equivalent to the generalized likelihood ratio test

Example
Let $X_1, \ldots, X_n$ be a random sample from a geometric distribution with parameter p;

$$P(X_i = k) = p(1-p)^k, \quad k = 0, 1, 2, \ldots.$$

Then

$$L(p) = p^n(1-p)^{\sum(x_i)}$$

and

$$\ell(p) = n\ln p + n\overline{x}\ln(1-p),$$

so that

$$\partial\ell/\partial p(p) = n\left(\frac{1}{p} - \frac{\overline{x}}{1-p}\right)$$

and

$$\partial^2\ell/\partial p^2(p) = n\left(-\frac{1}{p^2} - \frac{\overline{x}}{(1-p)^2}\right).$$

We know that $\mathbf{E}X_1 = (1-p)/p$. Calculate the information

$$I(p) = \frac{n}{p^2(1-p)}.$$

Suppose $n = 20$, $\overline{x} = 3$, $H_0 : p_0 = 0.15$ and $H_1 : p_0 \neq 0.15$. Then $\partial\ell/\partial p(0.15) = -62.7$ and $I(0.15) = 1045.8$. The test statistic is then

$$Z = -62.7/\sqrt{1045.8} = 1.9388.$$

Compare to 1.96 for a test at level $\alpha = 0.05$: do not reject H_0.

Example continued.
For a generalized LRT the test statistic is based on $2\log[L(\hat{\theta})/L(\theta_0)] \approx \chi^2_1$, and the test statistic is thus

$$2(\ell(\hat{\theta}) - \ell(\theta_0) = 2n(\ln(\hat{\theta}) - \ln(\theta_0) + \overline{x}(\ln(1-\hat{\theta}) - \ln(1-\theta_0)).$$

We calculate that the m.l.e. is

$$\hat{\theta} = \frac{1}{1+\overline{x}}.$$

Suppose again $n = 20$, $\overline{x} = 3$, $H_0 : \theta_0 = 0.15$ and $H_1 : \theta_0 \neq 0.15$. The m.l.e. is $\hat{\theta} = 0.25$, and $\ell(0.25) = -44.98$; $\ell(0.15) = -47.69$. Calculate

$$\chi^2 = 2(47.69 - 44.98) = 5.4$$

and compare to chisquare distribution with 1 degree of freedom: 3.84 at 5 percent level, so reject H_0.

4. Confidence Regions

If we can find a *pivot*, a function $t(\mathbf{X}, \theta)$ of a sufficient statistics whose distribution does not depend on θ, then we can find confidence regions in a straightforward manner. Otherwise we may have to resort to approximate confidence regions, for example using the approximate normality of the m.l.e. Recall that confidence intervals are equivalent to hypothesis tests with simple null hypothesis and one- or two-sided alternatives

Not to forget about:

Profile likelihood

Often $\theta = (\psi, \lambda)$ where ψ contains the parameters of interest; then we may base our inference on the profile likelihood for ψ,

$$L_P(\psi) = L(\psi, \hat{\lambda}_\psi).$$

Again we can use a (generalized) LRT or score test; if ψ scalar, $H_0 : \psi = \psi_0$, $H_1^+ : \psi > \psi_0$, $H_1^- : \psi < \psi_0$, use test statistic

$$T = \frac{\partial \ell(\psi_0, \hat{\lambda}_0; \mathbf{X})}{\partial \psi},$$

where $\hat{\lambda}_0$ is the MLE for λ when H_0 true. Large positive values of T indicate H_1^+; large negative values indicate H_1^-, and

$$T \approx \ell'_\psi - I^{-1}_{\lambda,\lambda} I_{\psi,\lambda} \ell'_\lambda \approx N(0, 1/I^{\psi,\psi}),$$

where $I^{\psi,\psi} = (I_{\psi,\psi} - I^2_{\psi,\lambda} I^{-1}_{\lambda,\lambda})^{-1}$ is the top left element of I^{-1}. We estimate the parameters by substituting the null hypothesis values; calculate the practical standardized form of T as

$$Z = \frac{T}{\sqrt{\mathrm{Var}(T)}} \approx \ell'_\psi(\psi, \hat{\lambda}_\psi)[I^{\psi,\psi}(\psi, \hat{\lambda}_\psi)]^{1/2},$$

which is approximately standard normal.

Bias and variance approximations: the delta method

The delta method is useful if we cannot calculate mean and variance directly. Suppose $T = g(S)$ where $ES = \beta$ and $\operatorname{Var} S = V$. Taylor expansion gives

$$T = g(S) \approx g(\beta) + (S - \beta)g'(\beta).$$

Taking the mean and variance of the r.h.s.:

$$ET \approx g(\beta), \qquad \operatorname{Var} T \approx [g'(\beta)]^2 V.$$

This also works for vectors S, β, with T still a scalar. If $\big(g'(\beta)\big)_i = \partial g / \partial \beta_i$, and $g''(\beta)$ the matrix of second derivatives, then

$$\operatorname{Var} T \approx [g'(\beta)]^T V g'(\beta)$$

and

$$ET \approx g(\beta) + \frac{1}{2} \operatorname{trace}[g''(\beta)V].$$

Exponential family

For distributions in the exponential family, many calculations have been standardized, see the lecture notes.