Exercises and Lecture Notes
STK 4180, Autumn 2020

Version 0.31, 29-Sep-2020

Nils Lid Hjort

Department of Mathematics, University of Oslo

Abstract

These are Exercises and Lecture Notes for the course Confidence, Likelihood, Probability, STK 4180 (Master level) or STK 9180 (PhD level), for the autumn semester 2020. I'll add on more exercises as the course progresses.

1. The probability transform

Some of the following facts are related to various operations for confidence distributions and confidence curves.

(a) Suppose X has a continuous and increasing cumulative distribution function F, i.e. $F(x) = \Pr\{X \leq x\}$. Show that $U = F(X)$ is uniform on the unit interval. Any continuously distributed random variable can hence be transformed to uniformity, via this *probability transform*.

(b) Show that also $U_2 = 1 - F(X)$ and $U_3 = |1 - 2\,F(X)|$ have uniform distributions.

(c) Simulate a million copies of $x_i \sim \mathrm{N}(0,1)$, and check the histogram of $\Gamma_1(x_i^2)$, where Γ_ν is the cumulative distribution function of a χ^2_ν. Comment on what you find.

(d) Suppose $\widehat{\theta}$ is an estimator for the real parameter θ, based on data y, with some continuous distribution function $K_\theta(x) = \Pr_\theta\{\widehat{\theta} \leq x\}$; we are in particular assuming that the distribution of $\widehat{\theta}$ depends only on θ, not on other aspects of the underlying model employed. Consider the construction

$$C(\theta, y_{\rm obs}) = \Pr_\theta\{\widehat{\theta} \geq \widehat{\theta}_{\rm obs}\} = 1 - K_\theta(\widehat{\theta}_{\rm obs}),$$

a curve that can be computed and plotted post-data, where $\widehat{\theta}_{\rm obs} = \widehat{\theta}(y_{\rm obs})$ is the observed estimate. Show that it has the property that the random $C(\theta, Y)$ is uniformly distributed, for each fixed θ.

2. CD and cc for the normal standard deviation

Read Cunen and Hjort's *Confidence Curves for Dummies* (2020), a FocuStat Blog Post. Then do the details, regarding mathematics and implementation, for their introductory meant-to-be-simple example:

"Here's a simple example. You observe the data points 4.09, 6.37, 6.87, 7.86, 8.28, 13.13 from a normal distribution and wish to assess the underlying spread parameter, the famous standard deviation σ. We'll now introduce you to as many as two (2) curves: the confidence curve $\mathrm{cc}(\sigma)$ and the confidence distribution $C(\sigma)$. They're close cousins, actually, and it's not the case that both curves need to be displayed for each new statistical application."

(a) Here you might start with the classic fact concerning the empirical variance that $\widehat{\sigma}^2 \sim \sigma^2\chi^2_m/m$, where $m = n-1$, with n the sample size. Then deduce that

$$C(\sigma, y_{\rm obs}) = \Pr_\sigma\{\widehat{\sigma} \geq \widehat{\sigma}_{\rm obs}\} = 1 - \Gamma_m(m\widehat{\sigma}^2_{\rm obs}/\sigma^2).$$

Here $y_{\rm obs}$ represents the observed data, and $\widehat{\sigma}_{\rm obs}$ the observed point estimate. Show that $C(\sigma, Y) \sim$ unif, where Y represents a random data set $Y_1, \ldots, Y_n$, from the σ in question. In particular, the distribution of $C(\sigma.Y)$ does not depend on σ.

(b) Reproduce versions of Cunen and Hjort's Figures A and B, with the confidence curve $\mathrm{cc}(\sigma)$, the CD $C(\sigma)$, the median confidence estimate, etc.

(c) Compute also the *confidence density* $c(\sigma, y_{\rm obs})$ associated with the CD. Compute also its mode, say σ^*, and briefly assess its properties as an estimator of σ.

(d) A Bayesian approach to the same problem, i.d. finding a posterior distribution for σ, is to start with a prior $\pi(\sigma)$ and then compute $\pi(\sigma \,|\, y_{\rm obs}) \propto \pi(\sigma) g(\widehat{\sigma}, \sigma)$, where $g(\widehat{\sigma}, \sigma)$ is the likelihood, here the density function for $\widehat{\sigma}$ as a function of σ. When does such a Bayesian approach agree with the confidence density?

(e) Suppose there are two independent normal samples, with standard deviations σ_1 and σ_2. Construct a CD for $\rho = \sigma_1/\sigma_2$. Invent a second simple small dataset, to complement the first dataset given above, and then compute and display the confidence curve $\mathrm{cc}(\rho, \mathrm{data})$.

3. An often (but not always) useful CD construction

In Exercise 1 we saw that the simple construction $C(\theta, y) = \Pr_\theta\{\widehat{\theta} \geq \widehat{\theta}_{\rm obs}\}$ gives a CD, in the case of one-dimensional setups with a well-defined estimator $\widehat{\theta}$.

(a) More generally, assume $Y_1, \ldots, Y_n$ come from some distribution, depending on a single parameter θ, and that Z is a statistic with distribution stochastically increasing in θ. Then study $C(\theta, y) = \Pr_\theta\{Z \geq z_{\rm obs}\}$. Show that this is a bona fide CD.

(b) Show also that the construction works, if there are other parameters at play too, as long as the distribution of the chosen Z only depends on θ. Go through the details for the case of the Y_i being $\mathrm{N}(\mu, \sigma^2)$, with $Z = \sum_{i=1}^n (Y_i - \bar{Y})^2$, and also for $Z' = \sum_{i=1}^n |Y_i - M_n|$, where M_n is the empirical median. Compute, display, compare both CDs, based on Z and on Z', for the simple dataset of Exercise 2 (with $n = 6$). For the Z case, there is a formula, but for the Z' case you would need simulation, for a grid of σ values.

(c) For a normal sample from $\mathrm{N}(\mu, \sigma)$, we see that several $\Pr_{\mu,\sigma}\{Z \geq z_{\rm obs}\}$ schemes work, in that the Z in question has a distribution depending on σ, but not μ. Attempt to work with $C^*(\mu, y) = \Pr_{\mu,\sigma}\{\bar{Y} \geq \bar{y}_{\rm obs}\}$ ... and show that it will not really work (unless σ is known).

(d) But of course there *are* natural CD constructions for μ here. What is needed is a *pivot*, say $A = \text{piv}(\mu, y)$, a function binding the focus parameter and data together in a way which makes its distribution not depend on the parameters. Study indeed

$$t_n = t_n(\mu, Y) = \frac{\bar{Y} - \mu}{\widehat{\sigma}/\sqrt{n}},$$

with $\widehat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ the classical empirical variance. Pretend that you in all your cleverness have not seen this t_n before, and are unaware of its relation to a t distribution – but show that the distribution of t_n, call it K_n, does not depend on (μ, σ).

(e) Then show that $C(\mu, y_{\text{obs}}) = K_n(t_n(\mu, y_{\text{obs}}))$ is a CD for μ. Even if you do not see the connection to the classic t of Student (1908), you may still carry through this, by simulating $B = 10^5$ realisations of t_n, and use

$$C(\mu, y_{\text{obs}}) = K_n^*(t_n(\mu, y_{\text{obs}})) = \frac{1}{B} \sum_{j=1}^B I\{t_{n,j} \le t_n(\mu, y_{\text{obs}})\}.$$

But show that by all means K_n is a t_m, with $m = n - 1$, so the canonical CD for μ is and remains $C(\mu, y_{\text{obs}}) = G_m(\sqrt{n}(\mu - \bar{y}_{\text{obs}})/\widehat{\sigma}_{\text{obs}})$, with G_m the cdf for the t_m.

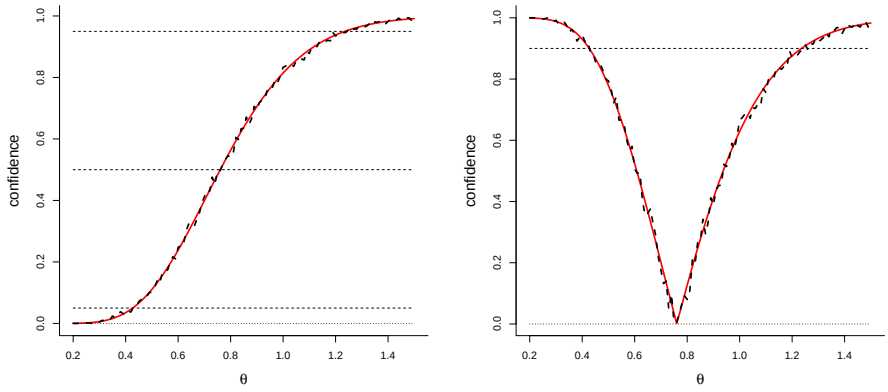

Figure 0.1: Left panel: confidence distribution $C(\theta)$, via simulations (black and wiggly curve) and via exact calculations (red and smooth curve); right panel: the two versions of the associated confidence curve $\text{cc}(\theta)$.

4. A skewed distribution on the unit interval

Consider the model $F(y, \theta) = y^\theta$ for observations on $[0, 1]$, where θ is an unknown positive parameter.

(a) Write down the log-likelihood function and find a formula for the maximum likelihood estimator $\widehat{\theta}$.

(b) Use theory of CLP, Chapter 2, to write down a normal approximation to the distribution of $\widehat{\theta}$.

(c) Consider the data set

```
0.013 0.054 0.234 0.286 0.332 0.507 0.703 0.763 0.772 0.920
```

Estimate θ and compute the confidence distribution $C(\theta) = \Pr_\theta\{\widehat{\theta} \geq \widehat{\theta}_{\rm obs}\}$, along with the confidence curve $\mathrm{cc}(\theta) = |1 - 2\,C(\theta)|$, (i) using simulations, (ii) using exact probability calculus. Reproduce a version of Figure 0.1.

(d) Supplement these two curves with approximations based (i) on the normal approximation for $\widehat{\theta}$ and (ii) on the chi-squared approximation for the deviance.

5. The children of Odin

As we know, Odin had six male offspring – Thor, Balder, Vitharr, Váli, Heimdallr, Bragi – with the sources saying nothing about daughters. So how many children is it likely that he had, in total? With N the number of children, and y the number of boys, we assume $y \,|\, N \sim \mathrm{Bin}(N, p)$, with $p = 0.514$ (a good point estimate for today's overall figure for human reproduction). So the data is that $y = 6$, and we can attempt confidence inference for N. The questions below expand on those given in CLP, Example 3.11.

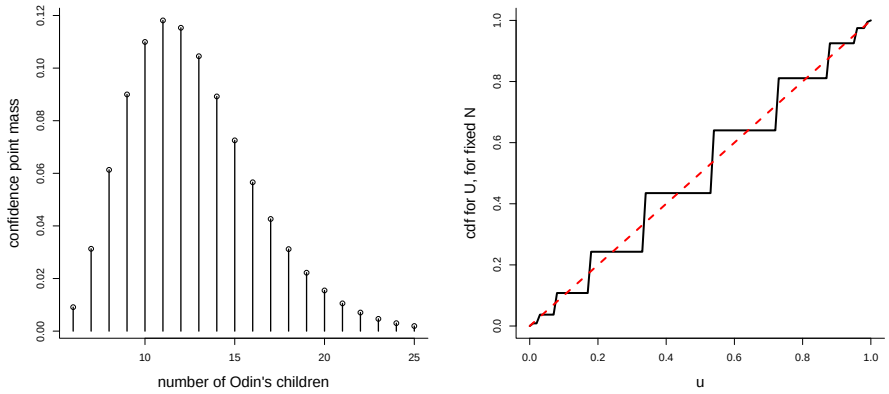

Figure 0.2: Left panel: confidence point masses $c(N, y)$, for $N \geq 6$; right panel: for a fixed $N = 14$ (in this example), the empirical cumulative distribution for $U = C(N, Y)$, with $Y \sim \mathrm{Bin}(N, p)$.

(a) A natural construction for a CD is

$$C(N, y) = \Pr_N\{Y > y\} + \tfrac{1}{2}\Pr_N\{Y = y\},$$

a version of the general method of Exercise 1(e), but with so-called half-correction for the discreteness. Compute and display this CD, and take differences to compute also the confidence point masses, $c(N, y)$. Construct a version of Figure 0.2, left panel.

(b) For the sport of it, carry out a Bayesian analysis too. Start with a reasonable prior $\pi(N)$ (formed before you read in a book that $y = 6$), compute the posterior distribution $\pi(N \,|\, y = 6)$, and compare with the CD analysis.

(c) A CD $C(\theta, y)$, for a parameter θ based on data y, should ideally have the uniformity property that $U = C(\theta_0, Y)$ has the uniform distribution, for any fixed θ_0, with Y a random dataset drawn from the model at that position in the parameter space. This is not quite possible here, since the situation is discrete, with not many values to attain for y. Construct a verion of Figure 0.2, right panel; here I took $N_0 = 14$, simulated say 10^4 realisations of $U = C(N_0, Y)$, and computed the empirical distribution function $\Pr\{U \leq u\}$. Comment on your findings.

(d) Find or dream up another situation (not necessarily with full data) where the model above might be used, i.e. p is known, but the binomial N is unknown.

6. Guess my range

I've simulated these points in my computer, from a uniform distribution over $[a, b]$, and I've ordered them, for simplicity. But I won't tell you the values I used for a or b, or indeed the range $\delta = b - a$. Your task will be to make inference about the δ – with a CD, a cc, and a median confidence estimate.

```
4.712 6.412 7.043 7.141 7.245 7.379 7.602 8.417 8.671 8.702
```

(a) With $Y_1, \ldots, Y_n$ from the uniform on $[a, b]$, explain that one may write $Y_i = a + (b - a)U_i$, with the U_i from the standard uniform over the unit interval. Deduce that

$$R_n = Y_{(n)} - Y_{(1)} = \delta R_{n,0}, \quad \text{with } R_{n,0} = U_{(n)} - U_{(1)},$$

relating the range of data naturally to the range of a uniform sample.

(b) Explain that R_n/δ is a pivot (as defined in Exercise 3). Simulate say 10^5 realisations in your computer from this distribution, say G_n.

(c) Show that

$$C(\delta, y) = \Pr_{a,b}\{R_n \geq R_{n,\text{obs}}\} = 1 - G_n(R_{n,\text{obs}}/\delta) \quad \text{for } \delta > R_{n,\text{obs}}$$

is a CD for δ. Compute it, using your simulations from G_n, and display as many as three curves: the CD $C(\delta, y_{\text{obs}})$; the $\text{cc}(\delta, y) = |1 - 2\,C(\delta, y_{\text{obs}})|$; and the confidence density $c(\delta, y_{\text{obs}})$. Also find the median confidence and maximum confidence point estimates Comment on your findings.

7. High drama: The cooling of newborns

Rather briefly, about the dramatic background for the after all simple exercise to follow, is at follows; read the FocuStat Blog Post Hjort (2017) for context and various details. – Without going into the drastic physiological details, in some rare cases newborns are being critically deprived of oxygen to the brain as a consequence of a difficult birth. Pioneering research, involving in particular Professor of Systems Physiology and Neonatal Neuroscience Marianne Thoresen from the Universities of Oslo and Bristol, has demonstrated that a form of cooling, where the little body has its temperature lowered to 33 degrees Celsius during a certain period just after birth, can save its life, and with no loss of later mental or motoric capacities. There is ongoing research and controversy, however, regarding the time window where the cooling operation is helpful, or useless, or too late.

In a thorough and important study, A. Laptook and a long list of coauthors published a paper in *Journal of the American Medical Association* (2017), involving

$$y_0 \sim \text{Bin}(m_0, p_0), \quad \text{for a group of noncooled newborns,}$$
$$y_1 \sim \text{Bin}(m_1, p_1), \quad \text{for a group of cooled newborns.}$$

The event in question is death or disability (with a precise definition of disability, assessed when the child is about 18 months old). Each case involved oxygen deprivation during birth, and where the cooling action, if taken, was initiated inside the time window 6 hours to 24 hours after birth (as opposed to starting earlier, which has been the general recommendation, so far).

So the article, with its partly controversial conclusions, is essentially about something as simple as comparing two binomial probabilities. Laptook et al. use Bayesian methods to tentatively argue that $p_1 < p_0$. In two articles in *Acta Paediatrica*, Walløe, Thoresen, Hjort (2019a, 2019b) have contested these findings; we argue rather that there is no significant difference between the two probabilities whatsoever (and that some of the conclusions phrased in Laptook et al. could lead to dangerous practices, having to do with advice for doctors regarding *the time window* for treating the neonates).

(a) The data are as follows: $y_0 = 22$ with $m_0 = 79$ for the noncooled group; $y_1 = 19$ with $m_1 = 78$ for the cooled group. Compute and display confidence curves $\text{cc}(p_0)$ and $\text{cc}(p_1)$ in a diagram. Find also 90 percent confidence intervals for the two. You may use the simple normal approximation, but it is slightly more precise to use

$$\text{cc}(p_0) = \Gamma_1(D_0(p_0)) \quad \text{and} \quad \text{cc}(p_1) = \Gamma_1(D_1(p_1)),$$

with the deviance functions, $D_0(p_0) = 2\{\ell_{0,\max} - \ell_0(p_0)\}$, etc. Comment on your findings.

(b) There are several ways in which to compare p_0 and p_1, and the medical world a typical choice would be

$$\text{odds ratio} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \quad \text{or} \quad \text{log-odds} = \log\frac{p_1}{1-p_1} - \log\frac{p_0}{1-p_0}.$$

The Laptook et al. paper focuses rather on the relative risk parameter $\rho = p_1/p_0$. Compute the log-likelihood profile function

$$\ell_{\text{prof}}(\rho) = \max\{\ell_0(p_0) + \ell_1(p_1) \colon p_1/p_0 = \rho\},$$

and then the deviance function $D(\rho) = 2\{\ell_{\text{prof},\max} - \ell_{\text{prof}}(\rho)\}$. Conclude by displaying the confidence curve $\text{cc}(\rho) = \Gamma_1(D(\rho))$, and comment on your findings.

8. A CD for the ratio of two exponential parameters

Consider independent samples from two exponential distributions, say $X_1, \ldots, X_m$ i.i.d. from $a\exp(-ax)$ and $Y_1, \ldots, Y_n$ i.i.d. from $b\exp(-by)$. Construct exact CDs and confidence curves for a and for b, separately, and then for the ratio $\rho = a/b$.

For a horrifying application of this machinery, check Hjort (2018), a FocuStat Blog Post on statistical sightings of better angels. We might come back to the underlying dataset later in this course, pertaining to the number of battle deaths in the last 95 great inter-state wars, from 1823 onwards.

9. Behaviour of maximum likelihood estimators

Here we work with the simple i.i.d. framework, where things are most easily examined. The main methods, tools, and results generalise to e.g. regression setups, with more work and attention to details, and with e.g. the Lindeberg theorem used instead of the simpler Central Limit Theorem.

So consider $Y_1, Y_2, \ldots$ being i.i.d. from a parametric model $f(y,\theta)$, with θ_0 the true parameter, of dimension say p. The model is assumed to have smooth derivatives of first and second order. The score function $u(y,\theta) = \partial \log f(y,\theta)/\partial\theta$ has mean zero at the model, $\mathrm{E}_\theta\, u(Y,\theta) = 0$. Its variance matrix is the Fisher information matrix

$$J(\theta) = \mathrm{Var}\, u(Y,\theta) = -\mathrm{E}_\theta \frac{\partial^2 \log f(Y,\theta)}{\partial\theta\partial\theta^{\mathrm{t}}},$$

assumed finite and continuous in a neighbourhood around θ_0. The log-likelihood function, after having observed n datapoints, is $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i,\theta)$. Our main object of study is the maximum likelihood estimator $\widehat{\theta}_n$, the maximiser of $\ell_n(\theta)$.

(a) Consider the random function $A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0)$, with s of the same dimension p as θ. Show via Taylor expansion that

$$A_n(s) = U_n^{\mathrm{t}} s - \tfrac{1}{2} s^{\mathrm{t}} J_n s + \varepsilon_n(s),$$

where

$$\begin{aligned} U_n &= n^{-1/2}\ell_n'(\theta_0) = n^{-1/2}\sum_{i=1}^n u(Y_i,\theta_0), \\ J_n &= -n^{-1}\ell_n''(\theta_0) = -n^{-1}\sum_{i=1}^n \frac{\partial^2 \log f(Y_i,\theta_0)}{\partial\theta\partial\theta^{\mathrm{t}}}, \end{aligned}$$

and $\varepsilon_n(s) \to_{\mathrm{pr}} 0$.

(b) Show that $U_n \to_d U \sim \mathrm{N}_p(0,J)$, and that $J_n \to_{\mathrm{pr}} J$, where $J = J(\theta_0)$ is the Fisher information matrix at the true parameter value.

(c) Show that $A_n(s) \to_d A(s) = U^{\mathrm{t}} s - \frac{1}{2} s^{\mathrm{t}} J s$, for each s. Actually, there is uniform convergence inside each ball $\|s\| \le c$.

(d) Argue that $\mathrm{argmax}(A_n) \to_d \mathrm{argmax}(A)$, modulo weak regularity conditions, and show from this the fundamental result about ML estimators under model conditions, that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d J^{-1}U \sim \mathrm{N}_p(0, J^{-1}).$$

(e) A very useful representation for the ML estimator, from these arguments, is that

$$\widehat{\theta}_n = \theta_0 + n^{-1}\sum_{i=1}^n J^{-1}u(Y_i,\theta_0) + \delta_n, \tag{0.1}$$

with $\delta_n \to_{\mathrm{pr}} 0$. One also says that the ML estimator has influence function $\mathrm{IF}(y) = J^{-1}u(y,\theta_0)$. Actually, we also have the alternative representation

$$\widehat{\theta}_n = \theta_0 + J_n^{-1} n^{-1}\sum_{i=1}^n u(Y_i,\theta_0) + \delta_n',$$

with $\delta_n' \to_{\mathrm{pr}} 0$. This holds both with $J_n = -n^{-1}\ell_n''(\theta_0)$ and $\widehat{J}_n = -n^{-1}\ell_n''(\widehat{\theta}_n)$. Both representations are useful, though it's (0.1) which is cleanest and often easiest to use.

(f) Similarly argue that $\max(A_n) \to_d \max(A)$, under weak regularity assumptions, and hence show that

$$\Delta_n = 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \to_d U^{\rm t} J^{-1} U \sim \chi^2_p.$$

This may be used to test the null hypothesis that θ is equal to some specified value. This result is one of several theorems carrying the 'Wilks Theorem' name. See Hjort and Pollard (1993) for sets of precise regularity conditions, also for regression setups.

10. A general Wilks Theorem for testing submodels

Material on Wilks Theorems is not 'naturally completed' before we also come to and include the lifting from dimension 1 to dimension k, so to speak. The basic story is simple to summarise, though not necessarily easy to prove with all the required steps, also since there are different versions and setups. The main story, at any rate, is as follows. Suppose we have n observations from a model $f(y,\theta)$, perhaps with regression parameters etc. Here θ is 'the full parameter vector', belonging to a parameter region Ω, in say p-dimensional space. Then there's a well defined log-likelihood function, say $\ell_n(\theta) = \sum_{i=1}^n \log f_i(y_i,\theta)$. Suppose one is interested in testing whether $\theta \in \Omega_0$ a subset of lower dimension $k < p$; perhaps this corresponds to having $\theta_j = 0$ for $p-k$ of the components. Then we may define and compute

$$\ell_{\max,\text{wide}} = \max\{\ell_n(\theta)\colon \theta \in \Omega\}, \quad \ell_{\max,\text{narr}} = \max\{\ell_n(\theta)\colon \theta \in \Omega_0\},$$

the maximised log-likelihood values under the full model and under the hypothesis H_0 that θ lies in this smaller space. Maxing over a bigger space yields a bigger number than maxing the same function over a smaller space. The splendidly useful Wilks Theorem, going back to Wilks (1938), says that under H_0 conditions,

$$\Delta_n = 2(\ell_{\max,\text{wide}} - \ell_{\max,\text{narr}}) \to_d \chi^2_q.$$

This is often presented, and made easier to remember and to use, by 'counting the degrees of freedom' as the dimension a priori minus the dimension under the hypothesis.

(a) Assume the H_0 in question is the simple one of $\theta = \theta_0$, so Ω_0 is a single point, of dimension zero. Verify that the Wilks theorem then is the same as what we've seen earlier, e.g. from Exercise [xx nemlig xx].

(b) Assume next that H_0 corresponds to $\phi = h(\theta) = \phi_0$, with $h(\theta)$ a smooth one-dimensional function. Note that saying $h(\theta) = \phi_0$ amounts to characterising a $(p-1)$-dimensional subspace of Ω. Verify that the general Wilks theorem above then corresponds to what we've worked with in the previous few exercises, with the deviance function, its limiting χ^2_1 distribution at the hypothesised value, etc. So in a certain usefulness-sense, this particular version of the general Wilks Theorem is the most useful one for the CLP course, with the wide model having dimension p, the narrow model dimension $p-1$, and hence with the deviance function having a χ^2_1 limit. That's how and why the recipe $\text{cc}(\phi) = \Gamma_1(D(\phi))$ works.

(c) [xx some examples, where we find Δ_n explicitly, perhaps also its distribution for finite n. (i) $Y \sim \text{Bin}(n,p)$, testing $p = p_0$. Write out the Δ_n and check that it tends to the χ^2_1 under $p = p_0$. (ii) Y_1, Y_2 are independent binomials (n,p_1) and (n,p_2). Write out the Δ_n for the

hypothesis $p_1 = p_2$. (iii) for $\mu = \mu_0$ in the normal. (iv) for $\sigma = \sigma_0$ in the normal. (v) for $(\mu, \sigma) = (\mu_0, \sigma_0)$ in the normal. (vi) for $\mu_1 = \mu_2$ for two normal samples. prove that the logLR is a function of the t-test. xx (vii) for testing $p = p_0$ in a multinomial situation. prove that the logLR test becomes asymptotically equivalent to the Pearson test of [xx nemlig xx]. xx]

(d) [xx separate point, some details to work through; to be polished. xx] the t test, seeing that the likelihood ratio test is a clear function of the t. sample 1, $\mathrm{N}(\xi_1, \sigma^2)$, with $\bar y_1$ and $Q_1 = \sum_{i=1}^{n_1}(y_{i,1} - \bar y_1)^2$; sample 2, $\mathrm{N}(\xi_2, \sigma^2)$, with $\bar y_2$ and $Q_2 = \sum_{i=1}^{n_2}(y_{i,2} - \bar y_2)^2$. write $N = n_1 + n_2$ and also $p_1 = n_1/N$ and $p_2 = n_2/N$. with $\widehat\delta = \bar y_2 - \bar y_1$, the t statistic is

$$t = \frac{\widehat\delta}{\widehat\sigma(1/n_1 + 1/n_2)^{1/2}} = N^{1/2}(p_1p_2)^{1/2}\frac{\widehat\delta}{\widehat\sigma}.$$

log-likelihood of combined sample is

$$\ell = -N\log\sigma - \tfrac12(1/\sigma^2)\{Q_1 + Q_2 + n_1(\bar y_1 - \xi_1)^2 + n_2(\bar y_2 - \xi_2)^2\}.$$

show that ML for σ^2 under wide and narrow model are

$$\widehat\sigma^2 = N^{-1}(Q_1 + Q_2) \quad \text{and} \quad \widehat\sigma_0^2 = N^{-1}(Q_1 + Q_2 + Np_1p_2\widehat\delta^2).$$

show that the exact log-likelihood-ratio test becomes

$$\Delta = 2(\ell_{\max} - \ell_{\max,H_0}) = N\log\frac{\widehat\sigma_0^2}{\widehat\sigma^2} = N\log(1 + N^{-1}t^2).$$

As we can see, for growing $N = n_1 + n_2$, this is close to t^2, which by the general theory is close to χ^2_1 under the $\mu_1 = \mu_2$ null hypothesis. The main point is however that the Δ is a sound easy function of the famous t, and we should continue to use t (as we've done, since 1908).

11. Proving the Wilks Theorem

Suppose $Y_1, \ldots, Y_n$ are i.i.d., where two models are considered: a narrow one, namely $f_0(y, \theta)$, with θ of dimension p; and a wide one, namely $f(y, \theta, \gamma)$, needing a further parameter vector γ of dimension q. We need the narrow model to be inside the wide one, so we assume that there is a γ_0 for which $f_0(y, \theta) = f(y, \theta, \gamma_0)$. We assume that γ_0 is an inner point in its parameter domain. We wish to construct a test for the hypothesis H_0 that the narrow model holds, and this is equivalent to testing $\gamma = \gamma_0$.

Let $\ell_n(\theta, \gamma)$ be the log-likelihood function for the wide model, which also means $\ell_n(\theta, \gamma_0)$ is the log-likelihood function in the narrow model. Let $(\widehat\theta, \widehat\gamma)$ be ML estimates in the wide model and $(\widetilde\theta, \gamma_0)$ ML estimates in the narrow model. Assuming that H_0 is in force, with density $f(y, \theta_0, \gamma_0)$ for the appropriate θ_0, we already know the principal answers regarding limit distributions for $\sqrt n(\widehat\theta - \theta_0, \widehat\gamma - \gamma_0)$ and $\sqrt n(\widetilde\theta - \theta_0)$ separately, but now we need to study them jointly, which calls for accurate representations and for linear matrix algebra to sort things out. Let the $(p+q)\times(p+q)$ information matrix $J = J(\theta_0, \gamma_0)$ and its inverse be partitioned into blocks:

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \quad \text{and} \quad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}.$$

(a) Here and in later settings the matrix $Q = J^{11}$ serves a special role. Show via matrix manipulations of $JJ^{-1} = I = J^{-1}J$ that $Q = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, similarly that $J^{00} = (J_{00} - J_{01}J_{11}^{-1}J_{10})^{-1}$, and that $J^{01} = -J_{00}^{-1}J_{01}J^{11}$.

(b) Show that there is simultaneous convergence in distribution

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta} - \theta_0) \\ \sqrt{n}(\widehat{\gamma} - \gamma_0) \end{pmatrix} \to_d \begin{pmatrix} A \\ B \end{pmatrix} = J^{-1} \begin{pmatrix} U \\ V \end{pmatrix} \quad \text{and} \quad \sqrt{n}(\widetilde{\theta} - \theta_0) \to_d C = J_{00}^{-1}U,$$

where

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathrm{N}_{p+q}(0, J) \quad \text{and hence} \quad \begin{pmatrix} A \\ B \end{pmatrix} = J^{-1} \begin{pmatrix} U \\ V \end{pmatrix} \sim \mathrm{N}_{p+q}(0, J^{-1}).$$

Show in particular that $B_n = \sqrt{n}(\widehat{\gamma} - \gamma_0) \to_d B \sim \mathrm{N}_q(0, Q)$ under the narrow model.

(c) Do Taylor expansion around $(\widehat{\theta}, \widehat{\gamma})$ to show that

$$\begin{aligned} \max_{\text{wide}} \ell - \max_{\text{narr}} \ell_n &= \sum_{i=1}^{n} \{\log f(Y_i, \widehat{\theta}, \widehat{\gamma}) - \log f(Y_i, \widetilde{\theta}, \gamma_0)\} \\ &= \tfrac{1}{2} n \begin{pmatrix} \widehat{\theta} - \widetilde{\theta} \\ \widehat{\gamma} - \gamma_0 \end{pmatrix}^{\mathrm{t}} J_n^* \begin{pmatrix} \widehat{\theta} - \widetilde{\theta} \\ \widehat{\gamma} - \gamma_0 \end{pmatrix} + \varepsilon_n, \end{aligned}$$

where the J_n^* matrix tends to J in probability and $\varepsilon_n \to_{\mathrm{pr}} 0$. Hence conclude that

$$\Delta_n = 2(\ell_{\max,\text{wide}} - \ell_{\max,\text{narr}}) \to_d \Delta = \begin{pmatrix} A - C \\ B \end{pmatrix}^{\mathrm{t}} \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \begin{pmatrix} A - C \\ B \end{pmatrix},$$

provided the narrow model holds, i.e. under H_0.

(d) It remains to establish that the limiting variable Δ has the advertised nice chi squared distribution. This is not obvious from its expression above – but do it by first discovering $A - C = -J_{00}^{-1}J_{01}B$ and then plugging in to simplify the expression for Δ. The result is $\Delta = B^{\mathrm{t}}Q^{-1}B$, which is a χ_q^2. – A rephrasing of this important result is as follows: If $\mathcal{M}_0$ is a model contained in a bigger $\mathcal{M}_1$ model, then twice the difference of maximised log-likelihoods, which is also by definition the *deviance distance* from the narrow model to the wider model, goes under the narrow model conditions to χ^2_{df}, with $\mathrm{df} = \dim(\mathcal{M}_1) - \dim(\mathcal{M}_0)$.

(e) xx should do local power too, under $f_n(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n})$. i think that $B_n = \sqrt{n}(\widehat{\gamma} - \gamma_0) \to_d B \sim \mathrm{N}_q(\delta, Q)$, and that

$$\Delta_n \to_d \Delta = B^{\mathrm{t}}Q^{-1}B \sim \chi_q^2(\delta^{\mathrm{t}}Q^{-1}\delta).$$

also, of separate interest: the log-LR Δ_n test is asymptotically equivalent to $\Delta_n' = B_n^{\mathrm{t}}\widehat{Q}^{-1}B_n = n(\widehat{\gamma} - \gamma_0)^{\mathrm{t}}\widehat{Q}^{-1}(\widehat{\gamma} - \gamma_0) \to_d B^{\mathrm{t}}Q^{-1}B$. write this out. xx

(f) [xx extension to regression models. xx]

12. CDs and posterior distributions with boundary constraints

Here we learn about construction of CDs when there is a boundary condition on the focus parameter. This is sometimes an easy task, involving a natural positive post-data probability on the boundary point. We also compare with Bayesian procedures. Matters may of course be extended and generalised in several directions here, but for simplicity and conciseness we study a very simple prototype situation: y is $\mathrm{N}(\theta, 1)$, and $\theta \geq 0$ a priori.

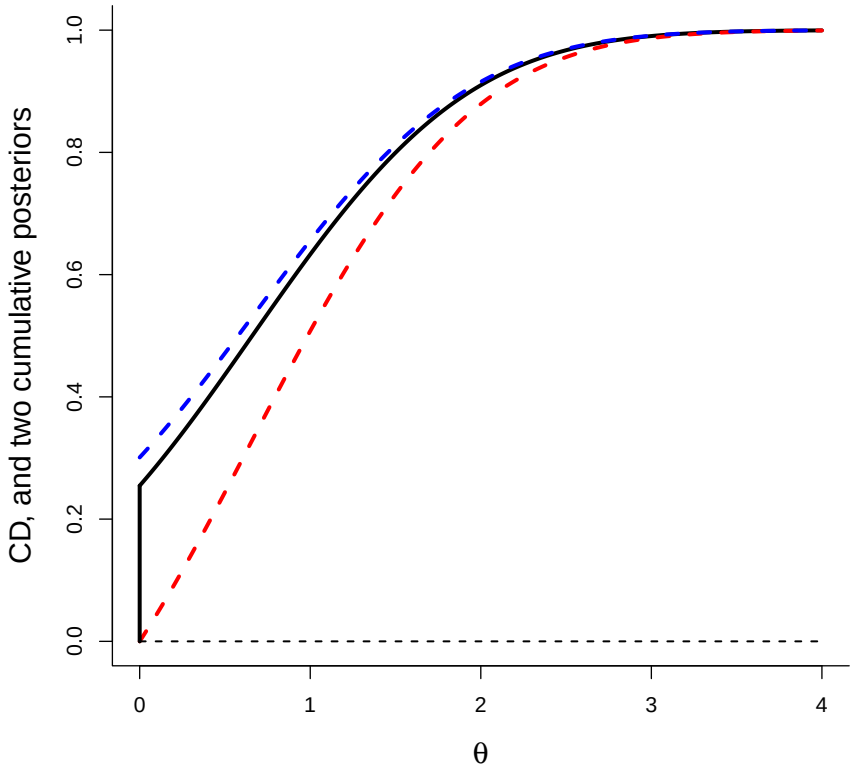

Figure 0.3: With $y_{\text{obs}} = 0.66$ for the $\text{N}(\theta, 1)$ model, the black curve is the natural CD, with positive point mass 0.255 at zero. The red and the blue curves are Bayesian posterior distributions, for the flat prior on the halfline, and for the mixture prior with $\frac{1}{2}$ at zero and $\frac{1}{2}$ flat on the halfline, respectively.

(a) Before we come to the parameter constraint, we deal with the more normal situation where there is no a priori constraint. The classical CD is then $C(\theta, y) = \Phi(\theta - y)$. Show that the Bayesian starting with a flat prior for θ finds the posterior distribution $\theta \,|\, y \sim \text{N}(y, 1)$, with cumulative $B(\theta \,|\, y) = \Phi(\theta - y)$, i.e. identical to the canonical CD. – The point below will partly be that this is *not* the same for the constrained problem.

(b) For the remaining points here, assume indeed that $\theta \geq 0$ a priori. Argue that the canonical CD should be $C(\theta, y) = \Phi(\theta - y)$ for $\theta \geq 0$. Its point mass at zero is $\Phi(-y)$. Graph the CD, for the three cases y_{obs} equal to $-0.22, 0.66, 1.99$.

(c) One Bayesian approach in this situation, where $\theta \geq 0$ a priori, is to let θ be flat on $[0, \infty)$. Show that then

$$\theta \,|\, y \sim \frac{\phi(\theta - y)}{\int_0^\infty \phi(\theta - y)\,\text{d}\theta} = \frac{\phi(\theta - y)}{\Phi(y)} \quad \text{for } \theta \geq 0,$$

and that the cumulative posterior distribution becomes

$$B(\theta \,|\, y) = \frac{\Phi(\theta - y) - \Phi(-y)}{1 - \Phi(-y)} = \frac{\Phi(\theta - y) - \Phi(-y)}{\Phi(y)} \quad \text{for } \theta \geq 0.$$

For the three cases of y_{obs} given above, graph the CD along with the Bayesian $B(\theta \,|\, y_{\text{obs}})$, and comment on what you find.

(d) There's a notable discrepancy between the frequentist Schweder-Hjort CD and the Bayesian posterior distribution associated with a flat prior on the $[0, \infty)$ interval, in cases where the $y_{\rm obs}$ is close to, or perhaps even to the left of, the boundary point. Read Schweder's FocuStat Blog Post (2017), where he dares to very much disagree with Nobel Prize Winner no wait a second I mean the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel Winner Professor Christopher Sims. It would have been better for Sims, in his chosen example featuring Bayesian methodology, to not use flat priors on positive halflines, but to allow pointmasses at zero too.

(e) In general terms, for the case of $y \,|\, \theta \sim {\rm N}(\theta, 1)$, let θ have the mixture prior distribution $p_0\pi_0 + p_1\pi_1$, with the sub-priors π_0 and π_1 having their individual posteriors $\pi_0(\theta \,|\, y)$ and $\pi_1(\theta \,|\, y)$. Show that the posterior has a natural mixture form,

$$\theta \,|\, y \sim p_0^*(y)\pi_0(\theta \,|\, y) + p_1^*(y)\pi_1(\theta \,|\, y),$$

where

$$p_0(y) = \frac{p_0 f_0(y)}{p_0 f_0(y) + p_1 f_1(y)} \quad \text{and} \quad p_1(y) = \frac{p_1 f_1(y)}{p_0 f_0(y) + p_1 f_1(y)},$$

and with $f_0(y) = \int \phi(y-\theta)\pi_0(\theta)\,{\rm d}\theta$ and $f_1(y) = \int \phi(y-\theta)\pi_1(\theta)\,{\rm d}\theta$ the marginal densities following from the two priors. (This structure generalises to general mixture priors in general models, though that does not concern us just now.)

(f) For the prior $p_0\pi_0 + p_1\pi_1$, with π_0 a unit pointmass at zero and π_1 a flat prior on the halfline, show that $f_0(y) = \phi(y)$ and $f_1(y) = \Phi(y)$. With a 50-50 mixture, show hence that

$$p_0(y) = \frac{\phi(y)}{\phi(y) + \Phi(y)} \quad \text{and} \quad p_1(y) = \frac{\Phi(y)}{\phi(y) + \Phi(y)}.$$

Draw curves of these two posterior probabilities, one for the zero-point and the other for the halfline-based part, as y goes from say -5 to 5. Show that the posterior cumulative distribution becomes

$$B^*(\theta \,|\, y) = p_0(y) + p_1(y)B(\theta \,|\, y) \quad \text{for } \theta \geq 0.$$

In particular, there's a pointmass $p_0(y)$ at zero. Construct a version of Figure 0.3.

(g) Show that there is no choice of (p_0, p_1) which makes the Bayesian cumulative posterior $B^*(\theta \,|\, y)$ agree with the CD $C(\theta, y)$. Devise a method for selection (p_0, p_1) such that the distance between $B^*(\theta \,|\, y)$ and $C(\theta, y)$ is small, for a relevant range of θ and possible observed $y_{\rm obs}$.

(h) Generalise the formulae above to the case of $y_1, \ldots, y_n$ i.i.d. ${\rm N}(\theta, \sigma^2)$, with known σ.

13. How many farmed salmon are escaping into Norwegian rivers?

Check Morgenbladet for key words like 'oppdrettslaks forskere industri' to get an idea of the High Temperature in various debates, or is it quarrels, in this zillion-dollar industry, with much at stake. We're not really going into this here, but I show a few neutral ingredients which do have some relation to certain complicated questions the salmon researchers are interested in.

Substantial amounts of farmed salmon escape and are found in 'the wild', e.g. the classic Norwegian wild-salmon rivers. One wishes to estimate

$$p = \Pr(A),$$

the proportion of farmed escapees in a river. Cathcing m salmon and finding y of these are from the farmed population gives information on a different probability p', not p itself, however.

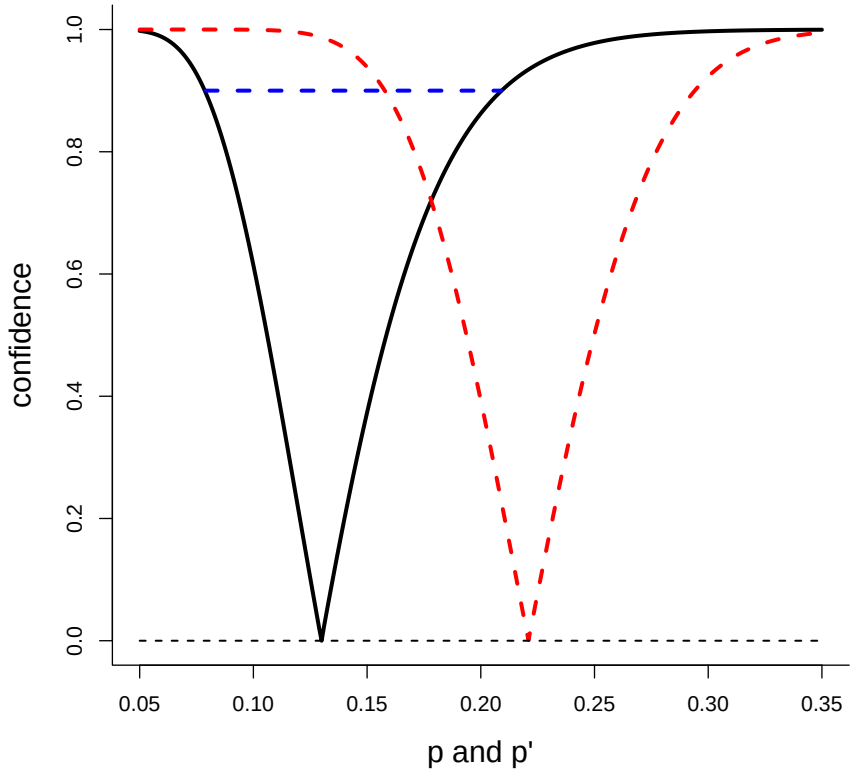

Figure 0.4: The information on p' is that $y = 22$ from $\text{Bin}(100, p')$, giving the $\text{cc}(p')$ in red slanted curve. The supplementary information on the ratio parameter ϕ is approximated with a Gamma $(18, 9)$. This leads via the log-likelihood profile recipe to the $\text{cc}(p)$ on the left, with 90 percent confidence interval $[0.079, 0.209]$.

(a) Show that the p' in question is

$$\begin{aligned} p' &= \Pr(\text{farmed} \,|\, \text{caught}) \\ &= \frac{p \Pr(\text{caught} \,|\, \text{farmed})}{p \Pr(\text{caught} \,|\, \text{farmed}) + (1-p) \Pr(\text{caught} \,|\, \text{wild})} \\ &= \frac{p\,\phi}{p\,\phi + 1 - p} = h(p, \phi), \end{aligned}$$

with

$$\phi = \frac{\Pr(\text{caught} \,|\, \text{farmed})}{\Pr(\text{caught} \,|\, \text{wild})}.$$

Only if both wild and farmed salmon have the same eagerness to being caught, by the methods employed by the salmon researchers, is p' the same as p.

(b) As an illustration, suppose that $y = 22$ farmed salmon are caught from a binomial $(100, p')$, and that and independent experiment on catchability leads to a log-likelihood component $\ell_0(\phi)$, well described by a Gamma density with parameters $(18, 9)$, with mean 2.00 and standard deviation 0.47. Compute and display the profiled log-likelihood function

$$\ell_{\rm prof}(p) = \max\{\ell_{\rm bin}(h(p,\phi)) + \ell_0(\phi)\colon \text{all } \phi\},$$

with $\ell_{\rm bin}(p')$ the usual binomial log-likelihood.

(c) Reconstruct a version of Figure 0.4.

(d) Hvorfor skapte Gud torsken? Han kunne ikke gjøre alle til laks. The salmonists I was briefly in contact with in 2017 had such data $y_j \sim \text{Bin}(m_j, p_j)$ for several Norwegian lakseelver. For illustration, suppose such data for three rivers amount to 22, 33, 11, with salmon sample sizes 100, 150, 60. Assume that the same catchability ratio ϕ is at work, for all rivers. Carry out the relevant profiling from the combined data information

$$\sum_{j=1}^{3} \ell_{{\rm bin},j}(h(p_j,\phi)) + \ell_0(\phi)$$

to find confidence curves $\text{cc}_j(p_j)$ for the three rivers. Exhibit point estimates and 90 percent confidence intervals.

(e) Use your imagination to set up a scenario where there is easy binomial information on some $p' = \Pr(A')$, but where interest lies in a different $p = \Pr(A)$, and where there is a link function $p' = h(p,\phi)$. One example, incidentally, is from CLP Exercise 4.2, with ψ the proportion of students having cheated on an exam in their student lives, and where $p' = 2/3 - 1/3p$.

14.

Here's a dataset, indirectly famous: Sir David Cox, now 96 years old, som heier på Aston Villa, he once told be, at CAS, is super-famous; his most super-famous paper is the 1972 one, where he invents the Cox regression model and method for survival data, changing the world, etc.; and there he used a simplified version of this particular dataset. The version I give below is a fuller version of what Cox described; I've found it on the the net and organised a bit further. I do not go into all the details, but the crucial variable is t, the number of weeeks in remission, for 42 leukemia patients, belonging to group 0 (placebo) or group 1 (treatment). The δ is 1 for failure (relapse) and 0 for censored; for patient with id number 41, for example, we have $t = 6$ and $\delta = 0$, which means he or she is still in remission, no relapse has taken place yet.

	id	t	delta	group
1	1	1	1	0
2	20	1	1	0
3	7	2	1	0
4	11	2	1	0
5	3	3	1	0
6	13	4	1	0
7	19	4	1	0
8	12	5	1	0
9	17	5	1	0
10	27	6	1	1

11	35	6	1	1
12	38	6	1	1
13	41	6	0	1
14	23	7	1	1
15	5	8	1	0
16	9	8	1	0
17	15	8	1	0
18	21	8	1	0
19	40	9	0	1
20	22	10	1	1
21	42	10	0	1
22	8	11	1	0
23	18	11	1	0
24	32	11	0	1
25	4	12	1	0
26	10	12	1	0
27	39	13	1	1
28	14	15	1	0
29	28	16	1	1
30	6	17	1	0
31	36	17	0	1
32	34	19	0	1
33	33	20	0	1
34	2	22	1	0
35	26	22	1	1
36	16	23	1	0
37	25	23	1	1
38	31	25	0	1
39	24	32	0	1
40	30	32	0	1
41	29	34	0	1
42	37	35	0	1

(a) Your first job is to fit the simple model where the hazard rates are $h_i(t) = \theta$ for group 0 and $h_i(t) = \theta\gamma$ for group 1 – constant hazard rates is equivalent to the variables being exponentially distributed. This needs the log-likelihood function to be written down carefully. Find confidence curves for θ and γ. Is γ likely to be close to 1?

(b) Then generalise to a more complicated model, with hazard rates

$$h_i(t) = \theta t^b \text{ for group 0}, \quad h_i(t) = \theta t^b \gamma \text{ for group 1}.$$

Find confidence intervals for θ, b, and γ. Is b sufficiently different from zero? Is the difference between the two groups significant?

15. Estimating x when you've only got a proxy y

There are many versions of the following situation, and also many methods for dealing with the implied problems. Statistical key words include 'measurement error' and 'proxy' and 'partial information'. I'm not sure if what I outline here is a so-called new take on it all.

Suppose one is very interested in estimating a variable x, perhaps for each of 1000 people, but this is either too complicated or too expensive, so one measures a simpler y instead, seen as a

proxy for x. Assume for simplicity that

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}),$$

a binormal with correlation ρ between them.

(a) Show that $x \,|\, y \sim \mathrm{N}(\rho x, 1-\rho^2)$. So matters are fine and simple if ρ is known. For illustration, suppose $y_{\rm obs} = 1.99$, for one of our 1000 humans, and that $\rho = 0.66$. Then give the natural estimate of x, along with a confidence curve.

(b) Suppose however that the correlation ρ is not known, but that a perhaps costly separate experiment has been carried out, leading to $\widehat{\rho} \sim \mathrm{N}(\rho, \kappa^2)$, with $\widehat{\rho} = 0.66$ and $\kappa = 0.12$. So conceptually we're two steps away from the real x: we've observed a proxy y, with another proxy, the estimated correlation, for the real correlation. Translated to somewhat more common terms, we have data, namely $(y, \widehat{\rho})$, and unknown parameters (x, ρ). Show that the log-likelihood for x, conditionally on having observed y, becomes

$$\ell(x, \rho \,|\, y) = -\tfrac{1}{2}\log(1-\rho^2) - \tfrac{1}{2}\frac{(x-\rho y)^2}{1-\rho^2} - \tfrac{1}{2}\frac{(\rho - \widehat{\rho})^2}{\kappa^2}.$$

(c) For the illustration with $y_{\rm obs} = 1.99$, $\widehat{\rho} = 0.66$ and $\kappa = 0.12$, compute and display the log-profile-likelihood

$$\ell_{\rm prof}(x \,|\, y_{\rm obs}) = \max\{\ell(x, \rho \,|\, y_{\rm obs}) \colon \text{all } \rho\}.$$

Construct a confidence curve $\mathrm{cc}(x \,|\, y_{\rm obs})$ from this. How much wider is a 90 percent confidence interval for x now, compared with the simpler situation where $\rho = 0.66$ is known (corresponding to $\kappa = 0$)?

(d) Attempt to make a little machine that for observed $y_1, \ldots, y_{1000}$ creates estimated values $x_1, \ldots, x_{1000}$, along with lower and upper endpoints of 90 percent confidence intervals.

(e) Try to formalise a setup which is perhaps more realistic than the simpler one above.

References

Cunen, C. and Hjort, N.L. (2020). Confidence Curves for Dummies. FocuStat Blog Post, April 2020.

Cunen, C. and Hjort, N.L. (2020). Combining information across diverse sources: the II-CC-FF paradigm. Submitted for publication.

Cunen, C., Hjort, N.L., and Nygård, H. (2020). Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *Journal of Peace Research* **57**, 221–234.

Hermansen, G., Stoltenberg, E.Aa., and Cunen, C. (2017). Bokmelding: Confidence, Likelihood, Probability: Statistical Inference With Confidence Distributions (Schweder og Hjort, CUP, 2016). FocuStat Blog Post, November 2017.

Hjort, N.L. (2017). Cooling of Newborns and the Difference Between 0.244 and 0.278. FocuStat Blog Post, December 2017.

Hjort, N.L. (2018). Towards a More Peaceful World [insert '!' or '?' here]. FocuStat Blog Post, January 2018.

Hjort, N.L. and Pollard, D.B. (1993). Asymptotics for ninimisers of convex processes. Statistical Research Report, Department of Mathematics, University of Oslo.

Hjort, N.L. and Schweder, T. (2018). Confidence distributions and related themes. [General introduction article to a Special Issue of the Journal of Statistical Planning and Inference dedicated to this topic, with eleven articles, and with Hjort and Schweder as guest editors; vol. 195, 1-13.

Laptook, A. et al. (2017). Effect of therapeutic hypothermia initiated after 6 hours of age on death and disability among newborns with hypoxic-ischemic encephalopathy: A randomized clinical trial. *Journal of the American Medical Association* **318**, 1550–1560.

Schweder, T. (2017). Bayesian Analysis: Always and Everywhere? Confidence curves for dummies. FocuStat Blog Post, November 2017.

Student (1908). The probable error of a mean. *Biometrika* **6**, 1–25.

Schweder, T. and Hjort, N.L. (2016). *Confidence, Likelihood, Probability*. Cambridge University Press, Cambridge.

Walløe, L., Hjort, N.L., and Thoresen, M. (2019a). Major concerns about late hypothermia study. *Acta Paediatrica* **108**, 588–589.

Walløe, L., Hjort, N.L., and Thoresen, M. (2019b). Why results from Bayesian statistical analyses of clinical trials with a strong prior and small sample sizes may be misleading: The case of the NICHD Neonatal Research Network Late Hypothermia Trial. *Acta Paediatrica* **108**, 1190–1191.

Wilks (1938). [xx fill in. xx]