STK 1110, Autumn 2016, Oblig II: Some notes

Nils Lid Hjort, November 2016

Exercise 1

(c) Multiplying out leads to

$$Q=\sum_{i=1}^n \frac{X_i^2-2\widehat\mu X_i+\widehat\mu^2}{\sigma_i^2}=\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}-A\widehat\mu^2.$$

Note that

$$\mathrm{E}\,\widehat\mu=\widetilde\mu \quad\text{and}\quad \mathrm{Var}\,\widehat\mu=1/A.$$

Hence, using similar algebra,

$$\mathrm{E}\,Q=\sum_{i=1}^n\frac{\mu_i^2+\sigma_i^2}{\sigma_i^2}-A(\widetilde\mu^2+1/A)=n-1+\sum_{i=1}^n\frac{\mu_i^2}{\sigma_i^2}-A\widetilde\mu^2=n-1+\sum_{i=1}^n\frac{(\mu_i-\widetilde\mu)^2}{\sigma_i^2}.$$

(d) (i) With all $\sigma_i=\sigma$: Then $\widehat\mu=\bar X$, and

$$Q=\sum_{i=1}^n\frac{(X_i-\bar X)^2}{\sigma^2}\quad\text{has mean}\quad n-1+\sum_{i=1}^n\frac{(\mu_i-\bar\mu)^2}{\sigma^2}.$$

This is more or less the result of (b). (ii) With all $\mu_i=\mu$: Then $\mathrm{E}\,Q=n-1$, a generalisation of the classical result to the case of unequal σ_i.

Exercise 2

(a) This has to do with sufficiency – $(\bar x_i,\widehat\sigma_i)$ contains all of the essential information for further analysis of data from student i. Given $(\bar x_i,\widehat\sigma_i)$, we can create an artificial dataset, for student i, with these values, and the same information content.

(b) We find $\widehat\rho=0.2833$, via `cor(xbarN,xbarE)`. On the scale of $\zeta=A(\rho)$, the 95% interval is $A(\widehat\rho)\pm 1.96/\sqrt{n-3}$, found to be $[0.0403,0.5422]$. Transforming back, after having found the inverse function to be

$$A^{-1}(x)=\frac{\exp(2x)-1}{\exp(2x)+1},$$

we find the 95% on the ρ scale to be

$$[A^{-1}(0.0403),A^{-1}(0.5422)]=[0.0403,0.4947].$$

This concerns the correlation for observed pairs (x,y) – as partly opposed to the 'real underlying correlation', say $\mathrm{corr}(x_0,y_0)$, where x_0 is the average length of all words in all Norwegian books owned by student i, and analogously for her or his English books.

- We learn that a student having Norwegian books with long words tends to have English books with long words too, and vice versa. The reasons for this interesting finding are not clear, but it's interesting to do a bit of speculation – some readers prefer longer-worded books, others might like shorter-worded literature. We're also reminded that the students were not instructed to choose books from their bookshelves in a totally random fashion, so there's a limit to how far we should stretch our imagination here.

(c) Using the usual two-sample test for equality of means, we may compute

$$t_1 = \frac{\bar{x}_N - \bar{x}_E}{(\widehat{v}_N^2/n + \widehat{v}_E^2/n)^{1/2}} = \frac{0.1134}{0.0843} = 1.345,$$

where $\widehat{v}_N$ and $\widehat{v}_E$ are the empirical standard deviations for the word length averages $x_{i,N}$ and $x_{i.E}$, respectively. Under the null hypothesis H_0 that the overall means μ_N and μ_E are equal, the t_1 is approximately normal. Observing $t_1 = 1.345$ is not clashing with that hypothesis.

- Since there is correlation present (somewhat surprisingly), one may argue that the denominator used for t_1 is not correct, and that it would be safer to go to the differences $z_i = x_{i.N} - x_{i,E}$, and test whether the population mean δ of these is equal to zero or not. This leads to

$$t_2 = \frac{\bar{x}_N - \bar{x}_E}{(\widehat{v}^2/n)^{1/2}} = \frac{0.1134}{0.0718} = 1.581,$$

where $\widehat{v}$ is the empirical standard deviation of the z_i. Again, under H_0 of no difference in population means, which is the same as $\delta = 0$, the t_2 has an approximately standard normal distribution. Observing $t_2 = 1.581$ is not quite managing to push us away from thinking that H_0 is ok. With 35 more students in the course, and the variances and covariance of $x_{i,N}$ and $x_{i,E}$ staying the same, the difference of 0.113 between $\bar{x}_N$ and $\bar{x}_E$ would be significant at the 0.05 level.

(d) The observed empirical standard variances have distributions of the type

$$\widehat{\sigma}_{i,N}^2 \sim \sigma_{i,N}^2 \frac{\chi^2_{m-1}}{m-1} \quad \text{and} \quad \widehat{\sigma}_{i,E}^2 \sim \sigma_{i,E}^2 \frac{\chi^2_{m-1}}{m-1},$$

where $m = 100$ is the number of words scrutinised for each case. As explained in the exercise text, these distributions are approximately normal (since $m-1$ is as big as 99). We may hence use the apparatus above, mutatis mutandis, to test the hypothesis H_0' that the population parameters σ_N and σ_E are equal. Just as above we may form

$$t_3 = \frac{\bar{\sigma}_N - \bar{\sigma}_E}{(\widehat{w}_N^2/n + \widehat{w}_E^2/n)^{1/2}} = \frac{0.3295}{0.0849} = 3.879,$$

where $\bar{\sigma}_N = (1/n)\sum_{i=1}^n \widehat{\sigma}_{i,N}$ and similarly for $\bar{\sigma}_E$, and with $\widehat{w}_N^2$ and $\widehat{w}_E^2$ being the empirical variances for these quantities. Under H_0', the distribution of t_3 ought to be approximately standard normal. But 3.879 is a far too big value for a $\mathrm{N}(0,1)$. Hence H_0' is to be rejected.

- The same conclusion is arrived at if we compute t_4, just as with t_2, with a more careful denominator, to take care of the possibility of correlation; indeed I find $t_4 = 0.3259/0.0772 = 4.267$. So Norwegian prose have words with more spread in their distribution than has English.

(e) We have

$$\bar{x}_i \,|\, \mu_i \sim \mathrm{N}(\mu_i, \kappa_i^2) \quad \text{and} \quad \mu_i \sim \mathrm{N}(\mu_0, \tau^2),$$

which also may be written

$$\bar{x}_i = \mu_0 + \delta_i + \varepsilon_i,$$

where $\delta_i \sim \mathrm{N}(0, \tau^2)$ and $\varepsilon_i \sim \mathrm{N}(0, \kappa_i^2)$ (and where $\kappa_i^2 = \sigma_i^2/m$). Hence

$$\bar{x}_i \sim \mathrm{N}(\mu_0, \tau^2 + \kappa_i^2) \quad \text{for } i = 1, \ldots, n.$$

The variation in the $\bar{x}_i$ is being decomposed into two parts; its variation as an estimator of the given student's μ_i, and the variation of the μ_i in the population of students (and their bookshelves).

(i) Suppose the σ_i are equal (which they are, to a not unreasonable degree of approximation). For the Norwegian data, we find

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{x}_i - \bar{x})^2 = 0.2793,$$

estimating the variance of $\bar{x}_i$, i.e. $\tau^2 + \kappa^2$. But we may clearly estimate the κ part separately, and then subtract, to find an estimate of τ. The natural estimate of a common κ is via the average of the $\widehat{\kappa}_i^2 = \widehat{\sigma}_i^2/m$:

$$\widehat{\kappa}^2 = (1/n) \sum_{i=1}^{n} \widehat{\kappa}_i^2 = 0.0745 = 0.2729^2.$$

This leads by subtraction to

$$\widehat{\tau}_N = (S^2 - \widehat{\kappa}^2)^{1/2} = 0.4526,$$

the estimated standard deviation of the distribution of average word-lengths across students. For the English dataset, entirely similar calculations lead to

$$\widehat{\tau}_E = (S_E^2 - \widehat{\kappa}_E^2)^{1/2} = (0.1758 - 0.2370^2)^{1/2} = 0.3459.$$

Note that we actually managed to estimate τ, the standard deviation of a certain distribution of some μ_i, without actually seeing these μ_i; it was sufficient to have estimates of these μ_i, and then subtract $\widehat{\kappa}^2$ from the observed estimate of $\tau^2 + \kappa^2$.

- To put up a formal test for $H_0\colon \tau = 0$, we may use $F = S^2/\widehat{\kappa}^2$, computed to be 3.751 for the Norwegian dataset. Here

$$F = \frac{S^2}{\widehat{\kappa}^2} \sim \frac{\tau^2 + \kappa^2}{\kappa^2} \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{n(m-1)}/\{n(m-1)\}} = (1 + \tau^2/\kappa^2) F_{63,6336}.$$

Under $\tau = 0$, the F has a Fisher distribution with degrees of freedom $(63, 6336)$, which, incidentally, is almost the same as a $\chi^2_{63}/63$. But the chance of having an F as big as 3.751, for such a variable, is ridiculously small. Hence $\tau = 0$ is soundly rejected. The same goes for the English case, for which $F = 3.130$.

– When the σ_i are taken common, across students, we may estimate this common value, and hence $\kappa = \sigma/\sqrt{m}$, as above, with very high precision. Hence we may construct a confidence interval for $\tau^2+\kappa^2$, from $S^2 \sim (\tau^2+\kappa^2)\chi^2_{63}/63$, and then subtract our way to a confidence interval for τ. For the Norwegian set, we start from $\Pr(a \le \chi^2_{n-1} \le b) = 0.95$ and

$$\frac{a}{n-1} \le \frac{S^2}{\tau^2+\kappa^2} \le \frac{b}{n-1},$$

where a and b are the 0.025 and 0.975 quantiles of the χ^2_{n-1}, to find

$$S^2\frac{n-1}{a} \ge \tau^2+\kappa^2 \ge S^2\frac{n-1}{b},$$

with interval $[0.2027, 0.4097]$ for $\tau^2+\kappa^2$. With $\widehat\kappa = 0.2729$ we may subtract and find

$$[(0.2027-\widehat\kappa^2)^{1/2}, (0.4097-\widehat\kappa^2)^{1/2}] = [0.3588, 0.4790]$$

as the 95% confidence interval for $\tau = \tau_N$. For the English dataset, using entirely analogous methods, I find $[0.2672, 0.4491]$.

(ii) Assume next that the σ_i and hence the $\kappa_i = \kappa_i/\sqrt{m}$ are not equal enough to be taken common. Several options may then be pursued regarding constructing an estimate of τ, partly using variations on the themes of Exercise 1. The simplest of these is to stick to S^2, and to show and then utilise the following result:

$$\mathrm{E}\,S^2 = \mathrm{E}\frac{1}{n-1}\sum_{i=1}^n(\bar x_i-\bar x)^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\,\bar x_i = \frac{1}{n}\sum_{i=1}^n(\tau^2+\kappa_i^2) = \tau^2+\bar\kappa^2,$$

in which $\bar\kappa^2 = (1/n)\sum_{i=1}^n \kappa_i^2$. But this leads to precisely the same estimator $\widehat\tau = (S^2-\widehat\kappa^2)^{1/2}$ as above, i.e. o.4526 for Norwegian and 0.3459 for English. The partly cosmetic difference is that we now view $\widehat\kappa^2 = (1/n)\sum_{i=1}^n \widehat\kappa_i^2$ estimator as an estimator of the average variance parameter $\bar\kappa^2$, rather than of a common κ^2.

– Setting up a clear formal test for the hypothesis $\tau = 0$ is now somewhat more complicated, but one may again start with $F = S/\widehat\kappa^2$, and reject when F is large enough (and indeed $F = 3.751$ turns out to be more than large enough). The p-value is $\Pr(F \ge 3.751 \,|\, \tau = 0)$, and this may be computed via simulation.

Exercise 3

(a) Integration gives

$$\mathrm{E}\,X = \frac{b^a}{\Gamma(a)}\int_0^\infty x^a\exp(-bx)\,\mathrm{d}x = \frac{b^a}{\Gamma(a)}\frac{\Gamma(a+1)}{b^{a+1}} = \frac{a}{b},$$

and similarly

$$\mathrm{E}\,X^2 = \frac{b^a}{\Gamma(a)}\int_0^\infty x^{a+1}\exp(-bx)\,\mathrm{d}x = \frac{b^a}{\Gamma(a)}\frac{\Gamma(a+2)}{b^{a+2}} = \frac{a(a+1)}{b^2},$$

which leads to the desired variance formula a/b^2.

(b) Solving

$$\bar{X} = \frac{\widehat{a}}{\widehat{b}} \quad \text{og} \quad S^2 = \frac{\widehat{a}}{\widehat{b}^2},$$

leads to the moment estimators

$$\widehat{a} = \frac{\bar{X}^2}{S^2} \quad \text{and} \quad \widehat{b} = \frac{\bar{X}}{S^2}.$$

Note that $\bar{X} \to_{\rm pr} a/b$ and $S^2 \to_{\rm pr} a/b^2$, via the law of large numbers, etc., which then implies $\widehat{a} \to_{\rm pr} a$ and $\widehat{b} \to_{\rm pr} b$, i.e. these are consistent estimators. – The general aspect at work here is that if $\widehat{a}, \widehat{b}, \widehat{c}$ are consistent estimators for a, b, c (converging in probability, to these three quantities), then $h(\widehat{a}, \widehat{b}, \widehat{c})$ is also consistent for $h(a, b, c)$, provided the h transform is continuous.

(c) For the 26 women and 127 men, we compute empirical means and variances, use the above, and find

$$(\widehat{a}_m, \widehat{b}_m) = (3.7827, 0.0967) \quad \text{and} \quad (\widehat{a}_w, \widehat{b}_w) = (2.3312, 0.0718).$$

(d) This also leads to parametric median estimates

```
qhatm = qgamma(0.50,ahatm,bhatm) # 35.719
qhatw = qgamma(0.50,ahatw,bhatw) # 27.956
```

(e) Bootstrapping, as more or less laid out in the exercise text, yields approximate 90% confidence intervals

$$[32.695, 38.842] \text{ for men} \quad \text{and} \quad [21.728, 34.995] \text{ for women}.$$

These are fairly symmetrical around the associated point estimates (though the bootstrapping method is meant to work also in more dire straits, sometimes with additional tweaking; see e.g. Schweder and Hjort, *Confidence, Likelihood, Probability*, Cambridge University Press, 2016). The interval for the men is rather tighter than that for the women, since there is more data information for the men.

(f) Go confidently in the direction of your dreams.