MATH562 – Fall 2025

Problem Set: Week 5

*1. A location-scale model is of the form $y_0 = \eta + \tau\, v_0$, where v_0 has a known distribution, with location and scale parameters $\eta \in \mathbb{R}$ and $\tau > 0$. A location estimator $m(y)$ and scale estimator $s(y)$, based on a random sample $y = (y_1, \ldots, y_n)$ with the y_i iid. distributed as y_0, are said to be *equivariant* if they satisfy

$$\begin{aligned} m(a + b\,y) &= m(a + b\,y_1, \ldots, a + b\,y_n) = a + b\,m(y_1, \ldots, y_n) = a + b\,m(y), \\ s(a + b\,y) &= s(a + b\,y_1, \ldots, a + b\,y_n) = b\,s(y_1, \ldots, y_n) = b\,s(y), \quad a \in \mathbb{R}, b > 0. \end{aligned}$$

(a) Show that $q_1(y, \eta) = \{m(y) - \eta\}/s(y)$ and $q_2(y, \tau) = s(y)/\tau$ are pivots, and explain how to use them to construct confidence intervals for η and τ.

Solution: We have

$$m(y) = m(\eta + \tau v) = \eta + \tau m(v), \quad s(y) = s(\eta + \tau v) = \tau s(v),$$

where $v = (v_1, \ldots, v_n)$, so

$$q_1(y, \eta) = \frac{m(y) - \eta}{s(y)} = \frac{\eta + \tau m(v) - \eta}{\tau s(v)} = \frac{m(v)}{s(v)}, \quad q_2(y, \tau) = s(y)/\tau = \tau s(v)/\tau = s(v),$$

so both $q_1(y, \eta)$ and $q_2(y, \tau)$ are functions of the data y and parameters that have known distributions, as those of $m(v)/s(v)$ and $s(v)$ are both known (at least in principle). If $Q_1 = q_1(v, 0)$ and $Q_2 = q_2(v)$ have respective α quantiles $q_1'(\alpha, n)$ and $q_2'(\alpha, n)$, i.e., $\Pr\{Q_1 \le q_1'(\alpha, n)\} = \alpha$ and $\Pr\{Q_2 \le q_2'(\alpha, n)\} = \alpha$ for $\alpha \in (0, 1)$, then we can write

$$1 - 2\alpha = \Pr\{q_1'(\alpha, n) < Q_1 \le q_1'(1 - \alpha, n)\} = \Pr\left\{q_1'(\alpha, n) < \frac{m(Y) - \eta}{s(Y)} \le q_1'(1 - \alpha, n)\right\},$$

and rearrangement of the inequality in the right-hand probability shows that

$$L = m(Y) - s(Y)q_1'(1 - \alpha, n), \quad U = m(Y) - s(Y)q_1'(\alpha, n),$$

are the limits of a $(1 - 2\alpha)$ confidence interval for η. Likewise

$$1 - 2\alpha = \Pr\{q_2'(\alpha, n) < Q_2 \le q_2'(1 - \alpha, n)\} = \Pr\{q_2'(\alpha, n) < s(Y)/\tau \le q_2'(1 - \alpha, n)\},$$

and rearrangement of the inequality in the right-hand probability shows that

$$L = (s(Y)/q_2'(1 - \alpha, n), \quad U = s(Y)/q_2'(\alpha, n)$$

are the limits of a $(1 - 2\alpha)$ confidence interval for τ.

(b) Deduce that pivots can be formed by taking

(i) $m_1(y) = \bar{y}$ and $s_1(y) = \left\{\sum_j (y_j - \bar{y})^2\right\}^{1/2}$,

(ii) $m_2(y) = \text{median}(y)$ and $s_2(y) = \text{IQR}(y)$, the interquartile range,

whatever the distribution of v_0. Briefly discuss the corresponding confidence intervals.

Solution:

Clearly $m_1(y) = m_1(\eta + \tau v) = n^{-1}\sum_j(\eta + \tau v_j) = \eta + \tau\bar{v}$ and a similar calculation shows that $s_1(y) = \tau s_1(v)$, and likewise for $m_2(y)$ and $s_2(y)$, leading to pivots.

(i) corresponds to the t and χ^2 statistics used for inference on η and τ when $y_1, \ldots, y_n \overset{i.i.d.}{\sim} N(\eta, \tau^2)$. (ii) should give intervals that are highly robust to outliers.

(c) Provide a pivot that could be used to construct a prediction interval for y_0 based on $y_1, \ldots, y_n$.

Solution: As $\{y_0 - m(y)\}/s(y)$ is easily shown to be independent of the parameters, with a known distribution, it can be used to make prediction intervals for y_0.

2. Find an α-level confidence interval for Y_+, when $Y_1, \ldots, Y_n, Y_+ \overset{i.i.d.}{\sim} \exp(\lambda)$. How does the interval change as $n \to \infty$. *Hint: if* $E \sim \exp(1)$ *then* $2E \sim \chi^2_2$.

Solution:

As $\lambda Y_j \sim \exp(1)$ we have $2\lambda(Y_1 + \cdots + Y_n) \sim \chi^2_{2n}$ independent of $2\lambda Y_+ \sim \chi^2_2$, so

$$Q = \frac{2\lambda Y_+/2}{2\lambda(Y_1 + \cdots + Y_n)/(2n)} = \frac{Y_+}{\bar{Y}} \sim F_{2,2n}$$

is a pivot. Hence a $1 - \alpha_L - \alpha_U$ tolerance interval is obtained using

$$1 - \alpha_L - \alpha_U = \Pr(q_{\alpha_L} \le Y_+/\bar{Y} \le q_{1-\alpha_U}) = \Pr(\bar{Y} q_{\alpha_L} \le Y_+ \le \bar{Y} q_{1-\alpha_U}) = \Pr(L \le Y_+ \le U),$$

where q_α is the α quantile of the $F_{2,2n}$ distribution. say, As $n \to \infty$, the F quantiles converge to those of $\exp(1)$, corresponding to λ being known.

3. (a) An experiment consists of observing the number of success y_1 in a fixed number n_1 of independent Bernoulli trials with unknown success probability $\theta \in (0, 1)$. Show that the corresponding density is

$$f(y_1 \mid \theta) = \binom{n_1}{y_1} \theta^{y_1} (1-\theta)^{n_1 - y_1}, \quad y_1 \in \mathcal{S}_\infty = \{0, 1, \ldots, n_1\}.$$

If prior information on θ can be summarised by the *beta density*

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}, \quad 0 < \theta < 1, \quad a, b > 0,$$

show that the posterior density for θ given y_1 is

$$f(\theta \mid y_1) = \frac{\Gamma(n_1 + a + b)}{\Gamma(y_1 + a)\Gamma(n_1 - y_1 + b)} \theta^{y_1 + a - 1} (1-\theta)^{n_1 - y_1 + b - 1}, \quad 0 < \theta < 1.$$

Solution: According to Bayes' theorem

$$f(\theta \mid y_1) = \frac{f(y_1 \mid \theta) f(\theta)}{f(y_1)} \propto \theta^{y_1 + a - 1} (1-\theta)^{n_1 - y_1 + b - 1}, \quad 0 < \theta < 1,$$

where the constant of proportionality ensures that the right-hand side has unit integral. Since the beta density has unit integral for any $a, b > 0$, and since $y_1 + a, n_1 - y_1 + b > 0$, the constant of proportionality must be obtained by replacing a and b by $y_1 + a$ and $n_1 - y_1 + b$, and thus must equal $\Gamma(n_1 + a + b)/\{\Gamma(y_1 + a)\Gamma(n_1 - y_1 + b)\}$. This gives the stated posterior density.

Note that this argument avoids any need for integration, and that the constants in the densities cancel from the numerator and denominator of the posterior.

(b) Another experiment conducts independent Bernoulli trials until there are y_2 successes, at which point there have been n_2 trials. Show that the corresponding density is

$$f(n_2 \mid \theta) = \binom{n_2 - 1}{y_2 - 1} \theta^{y_2} (1-\theta)^{n_2 - y_2}, \quad n_2 \in \mathcal{S}_\in = \{y_2, y_2 + 1, \ldots\}.$$

Without doing any calculations, write down the posterior density for θ based on the prior in (a).

Solution: Here

$$f(n_2 \mid \theta)f(\theta) \propto \theta^{y_2+a-1}(1-\theta)^{n_2-y_2+b-1}, \quad 0<\theta<1,$$

and we see at once using the argument from (a) that

$$f(\theta \mid n_2) = \frac{\Gamma(n_2+a+b)}{\Gamma(y_2+a)\Gamma(n_2-y_2+b)}\theta^{y_2+a-1}(1-\theta)^{n_2-y_2+b-1}, \quad 0<\theta<1.$$

(c) Show that if $y_1 = y_2$ and $n_1 = n_2$, then Bayesian inferences based on either of the two experiments will be identical, i.e., they do not take into account the different reference sets $\mathcal{S}_\infty$ and $\mathcal{S}_\in$.

Solution: The two posterior densities will be the same, so any Bayesian inferences based on the two experiments will be identical.

(d) Consider testing the hypothesis that $\theta = \frac{1}{2}$ against the alternative that $\theta < \frac{1}{2}$. Explain why the respective significance levels for the experiments in (a) and (b) would be

$$\sum_{y=0}^{y_1}\binom{n_1}{y}2^{-n_1}, \qquad \sum_{n=n_2}^{\infty}\binom{n-1}{y_2-1}2^{-n},$$

and evaluate these when $n_1 = n_2 = 12$, $y_1 = y_2 = 3$. How does this compare with (c)?

Hint: Recall that for $\alpha > 0$ the gamma function is defined as $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}\, du$, and that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\Gamma(n+1) = n!$ for $n \in \{1, 2, \ldots\}$, and $\Gamma(1/2) = \sqrt{\pi}$.

Solution: In (a) the number of successes Y_1 will tend to be small if $\theta < \frac{1}{2}$, so the observed significance level is the binomial probability

$$\Pr_0(Y_1 \le y_1) = \sum_{y=0}^{y_1}\Pr_0(Y=y) = \sum_{y=0}^{y_1}\binom{n_1}{y}2^{-n_1},$$

where $\Pr_0$ denotes probability computed under the null hypothesis $\theta = \frac{1}{2}$. Similarly if $\theta < \frac{1}{2}$ then it will take longer to attain y_2 successes than if $\theta = \frac{1}{2}$, so we compute the negative binomial probability

$$\Pr_0(N_2 \ge n_2) = \sum_{n=n_2}^{\infty}\Pr_0(N_2 = n) = \sum_{n=n_2}^{\infty}\binom{n-1}{y_2-1}2^{-n}.$$

Running `pbinom` in `R` with inputs $x = 3$, `size`$= 12$ and `prob`$= 0.5$, we get 0.07299805 and for $x = 11$, `size`$= 3$ and `prob`$= 0.5$ and `lower.tail = F`, we get 0.006469727. Unlike with (c), these suggest quite different evidence against the null hypothesis, because they sum probabilities over two different reference sets.

4. Consider the shoe data example from the Slides (Part 1).

(a) Show that the average $\bar{D}$ has mean θ and variance $\sigma^2 = m^{-2}\sum_{j=1}^m c_j^2$.

Solution: We note that $\mathrm{E}(I_j) = 0$ and $\mathrm{Var}(I_j) = 1$, so $\mathrm{E}(\bar{D}) = \theta$, and as the I_j are independent,

$$\mathrm{Var}(\bar{D}) = m^{-2}\sum_{j=1}^m \mathrm{Var}(\theta + I_j c_j) = m^{-2}\sum_{j=1}^m c_j^2\mathrm{Var}(I_j) = m^{-2}\sum_{j=1}^m c_j^2 = \sigma^2.$$

The c_j are unknown and therefore so is σ^2, which must be estimated from the data $D_1, \ldots, D_m$.

(b) Show that S^2 has mean σ^2 and hence can be used to estimate σ^2.

Solution: To estimate σ^2, we use the problems for week 3 to write

$$S^2 = \frac{1}{m(m-1)}\sum_{j=1}^m (D_j - \bar{D})^2 = \frac{1}{2m^2(m-1)}\sum_{j,k=1}^m (D_j - D_k)^2 = \frac{1}{2m^2(m-1)}\sum_{j\neq k}^m (D_j - D_k)^2,$$

and note that as $D_j - D_k = I_j c_j - I_k c_k$, $\mathrm{E}(I_j) = 0$ and $\mathrm{Var}(I_j) = 1$, and the I_j are independent, the right-most expression has expectation

$$\frac{2(m-1)}{2m^2(m-1)}\sum_{j=1}^m c_j^2\,\mathrm{E}(I_j^2) - \frac{2}{2m^2(m-1)}\sum_{j\neq k} c_j c_k\,\mathrm{E}(I_j I_k) = \frac{1}{m^2}\sum_{j=1}^m c_j^2 = \sigma^2,$$

(c) Extend this discussion to a balanced design in which m is even, $I_j = \pm 1$ and $\sum_{j=1}^m I_j = 0$ but the allocation is otherwise completely at random. This ensures that materials A and B appear equally often on the left and right shoes.

Solution: To ease the notation, let $m = 2n$. Under this randomization scheme the number of possible allocations is $\binom{m}{n}$, which equals 252 when $m = 10$; this is appreciably lower than the number 1024 obtained before.

The expectations and variances of the I_j are the same as in (a), but if $j \neq k$ then by symmetry

$$\mathrm{Cov}(I_j, I_k) = \mathrm{E}(I_j I_k) = 2\Pr(I_j = I_k) - 2\Pr(I_j \neq I_k) = 2\frac{n(n-1)}{2n(2n-1)} - 2\frac{n^2}{2n(2n-1)} = -\frac{1}{m-1}.$$

Under this randomisation scheme, $\sum_{j=1}^m I_j = 0$, so $\bar{D} = \theta + m^{-1}\sum_{j=1}^m I_j(c_j - \bar{c})$. Obviously $\mathrm{E}(\bar{D}) = \theta$ and

$$\begin{aligned}
m^2\mathrm{Var}(\bar{D}) &= \sum_{j=1}^m (c_j - \bar{c})^2\mathrm{Var}(I_j) + \sum_{i\neq j}(c_i - \bar{c})(c_j - \bar{c})\mathrm{Cov}(I_i, I_j) \\
&= \sum_{j=1}^m (c_j - \bar{c})^2 - \frac{1}{m-1}\sum_{j=1}^m (c_j - \bar{c})\sum_{i\neq j}(c_i - \bar{c}) \\
&= \sum_{j=1}^m (c_j - \bar{c})^2 + \frac{1}{m-1}\sum_{j=1}^m (c_j - \bar{c})^2 \\
&= \frac{m}{m-1}\sum_{j=1}^m (c_j - \bar{c})^2,
\end{aligned}$$

where the step from the second to third lines used the fact that $\sum_{i=1}^m (c_i - \bar{c}) = 0$ implies that $\sum_{i\neq j}(c_i - \bar{c}) = -(c_j - \bar{c})$. Hence

$$\mathrm{Var}(\bar{D}) = \frac{1}{m(m-1)}\sum_{j=1}^m (c_j - \bar{c})^2 = \tau^2,$$

say; note that subtracting $\bar{c}$ from the c_j will mean that it is very likely that $\tau^2 < \sigma^2$.

To find an estimator of the unknown τ^2 we write $v_j = I_j c_j$ and note that

$$\sum_{j=1}^m (D_j - \bar{D})^2 = \sum_{j=1}^m (v_j - \bar{v})^2 = \sum_{j=1}^m v_j^2 - m\bar{v}^2 = \sum_{j=1}^m I_j^2 c_j^2 - \frac{1}{m}\sum_{i,j=1}^m I_j I_i c_j c_i$$

has expected value

$$\sum_{j=1}^m c_j^2 - \frac{1}{m}\left\{\sum_{j=1}^m c_j^2 + \sum_{i\neq j}^m c_i c_j \mathrm{Cov}(I_i, I_j)\right\} = \sum_{j=1}^m c_j^2 - \frac{1}{m}\left\{\sum_{j=1}^m c_j^2 - \frac{1}{m-1}\sum_{j=1}^m c_j(m\bar{c} - c_j)\right\}$$

which equals

$$\sum_{j=1}^{m}(c_j - \bar{c})^2 - \frac{1}{m}\sum_{j=1}^{m}(c_j - \bar{c})^2 - \frac{1}{m(m-1)}\sum_{j=1}^{m}(c_j - \bar{c})^2 = \frac{m-2}{m-1}\sum_{j=1}^{m}(c_j - \bar{c})^2.$$

Hence τ^2 is estimated by

$$\frac{1}{m(m-2)}\sum_{j=1}^{m}(D_j - \bar{D})^2,$$

which can be computed from the observed differences $D_1, \ldots, D_m$.