MDL exercises, tenth handout

(due May 11th, 14:00) (there are 3 pages)

1. **[Bad Optional Continuation, with p-values]** Consider a standard p-value based null hypothesis test. According to the null hypothesis $\mathcal{H}_0 = \{P_0\}$, the data $X_1, X_2, \ldots$ are i.i.d. normally distributed with mean 0 and variance 1. According to the alternative, $\mathcal{H}_1 = \{P_\mu : \mu \in \mathbb{R}, \mu \neq 0\}$ they are i.i.d. normal, with mean $\mu \neq 0$ and variance 1. You take the standard level of significance $\alpha = 0.05$. For a test based on n data points, you use as your test statistic the standard Z-score $Z_n = (\sum_{i=1}^n X_i)/\sqrt{n}$. This has a $N(0,1)$ distribution under the null hypothesis. You do a two-sided test, so that the p-value based on observing a particular value z_n for the Z-score is given by $p(z_n) := P_0(|Z_n| \geq |z_n|)$. We write $\bar{p}_n$ for the random variable corresponding to p, i.e. if $Z^n = z^n$ then $\bar{p}_n = p(z^n)$. Note that we deal with a *strict* p-value, i.e. for $0 \leq \alpha \leq 1$, $P_0(\bar{p}_n \leq \alpha) = \alpha$.

   a) **[1 point]** You first perform a hypothesis test based on $n = 50$ data points. Suppose that the null is true. What is the probability that you reject the null? (this probability is called the Type-I error)

      Solution:

      By definition, the Type-I error is equal to the level of significance $\alpha = 0.05$.

   b) **[1 point]** Now suppose that in practice you observe, after 50 data points, a p-value of $\bar{p}_{50} = 0.10$. Not enough to reject, but promising! So you ask your boss if there is money and time for additional analysis. Your boss says yes, so you gather an additional 50 data points. You then analyze the data as if you had originally planned to observe 100 data points, i.e. you calculate the p-value $p(z_{100})$ and reject if this p-value is ≤ 0.05. Explain why this approach is problematic (you may use the statement that you are asked to prove in the next question in your answer to this question).

      Solution:

      The p-value does not only depend on the observed data, but also on the premise of the experiment that was performed (or the stopping rule used). The p-value $p(z_{100})$ belongs to the experiment, where you immediately observe 100 data points. To calculate the p-value for the altered experiment, we must take into account the probability that we continue after 50 data points. This will be exemplified in the next question.

   c) **[2 points]** Suppose you follow the following protocol: if, after 50 data points, the resulting p-value $p(z_{50})$ is no greater than $\alpha = 0.05$, you reject the null. If it is larger than 0.1, you accept the null and you

stop. If it is between 0.05 and 0.10, you gather an additional 50 data points. You then re-calculate the p-value $p(z_{100})$. If $p(z_{100})$ is smaller than 0.05, you reject the null after all, otherwise you accept the null and you stop.
Suppose that the null hypothesis is true. Show that with this protocol α', the actual probability of rejection under P_0, is at least 0.056. (HINT: use the fact that for all n, $\bar{p}_n \leq 0.1$ iff $|z_n| \geq 1.64$).

Solution:

The probability of rejection after the first 50 data points is 0.05 (see (a)). Let us denote by G the event that we gather additional data points. Then:

$$\begin{aligned}\mathbb{P}(G) &= \mathbb{P}(0.05 \leq p(z_{50}) \leq 0.10)\\ &= \mathbb{P}(p(z_{50}) \leq 0.10) - \mathbb{P}(p(z_{50}) \leq 0.05)\\ &= 0.10 - 0.05 = 0.05.\end{aligned}$$

By the hint and because of symmetry, we have:

$$\begin{aligned}\mathbb{P}(z_{50} \geq 1.64|G) = 0.5\\ \mathbb{P}(z_{50} \leq -1.64|G) = 0.5.\end{aligned}$$

Now we see

$$\begin{aligned}\mathbb{P}\left(p(z_{100}) \leq 0.05|z_{50} \geq 1.64\right) &= \mathbb{P}\left(z_{100} \geq |1.96||z_{50} \geq 1.64\right)\\ &\geq \mathbb{P}\left(z_{100} \geq 1.96|z_{50} \geq 1.64\right)\\ &= \mathbb{P}\left(\frac{\sum_{i=1}^{50} X_i + \sum_{i=51}^{100} X_i}{\sqrt{100}} \geq 1.96 \,\middle|\, z_{50} \geq 1.64\right)\\ &= \mathbb{P}\left(\frac{\sum_{i=51}^{100} X_i}{\sqrt{100}} \geq 1.96 - \frac{1.64}{\sqrt{2}}\right)\\ &= \mathbb{P}\left(\frac{\sum_{i=51}^{100} X_i}{\sqrt{50}} \geq \sqrt{2}\left(1.96 - \frac{1.64}{\sqrt{2}}\right)\right)\\ &\approx \mathbb{P}\left(\frac{\sum_{i=51}^{100} X_i}{\sqrt{50}} \geq 1.13186\right)\\ &\approx 0.128,\end{aligned}$$

where we have used that $z_{50} \geq 1.64 \Leftrightarrow \frac{\sum_{i=1}^{50} X_i}{\sqrt{100}} \geq \frac{1.64}{\sqrt{2}}$. Using symmetry once more, we find

$$\begin{aligned}\mathbb{P}(\text{rejection}) &= \mathbb{P}(z_{50} \leq 0.05) + \mathbb{P}(G)\mathbb{P}\left(|z_{100}| \geq |1.96|| \, G\right) \\ &= 0.05 + \mathbb{P}(G)(\mathbb{P}(|z_{100}| \geq |1.96||G, z_{50} \geq 1.64)\mathbb{P}(z_{50} \geq 1.64|G) \\ &\quad + \mathbb{P}(\mathbb{P}(|z_{100}| \geq 1.64|G, z_{50} \leq -1.64)\mathbb{P}(z_{50} \leq -1.64|G)) \\ &\geq 0.05 + 0.05(0.128 \cdot 0.5 + 0.128 \cdot 0.5) \\ &= 0.0564.\end{aligned}$$

d) **[1 point]** Consider the situation of question a) again. After observing $p(z^{50}) = 0.08$ with 50 data points you would really like to continue with 50 more data points. In light of the answer of the previous question, you decide that this may be a good idea after all, as long as you report as your employed significance level α' instead of α. Explain why this is not a good solution to the problem.

Solution:

In general, significance levels should be determined before the experiment begins. Altering it on basis of the data is not good practice.

Alternatively: one cannot be entirely sure about the threshold at which we would have not gathered any additional data points. All we know is that after seeing $p(z_{50}) = 0.08$, we gathered additional points. It could be that after seeing $p(z_{50}) = 0.11$, we would have also gathered additional data points. Therefore, we can't determine the significance level as in the previous question, as we do not know the exact decision boundary.

e) **[2 points]** Consider the following exaggerated version of the previous 'optional continuation' protocol: you start with $n = 50$ data points. If $\bar{p}_n \leq 0.05$ you reject the null and you stop. Otherwise, you observe an additional data point X_{n+1}; if $\bar{p}_{n+1} \leq 0.05$, you reject the null and you stop. Otherwise, you observe an additional point X_{n+2}. If $\bar{p}_{n+2} \leq 0.05$, you reject the null and you stop. Otherwise you continue, observe X_{n+3}, reject if ..., and so on. Suppose that the null hypothesis is true. Show that with this procedure, you will then almost surely stop at some finite n and reject the null (HINT: use the 'law of the iterated logarithm').

Solution:

Rephrasing the law of iterated logarithm, we see

$$\limsup_{n\to\infty} \frac{Z_n}{\sqrt{2\log\log n}} = 1 \text{ a.s.}$$

So there almost surely exist $n \geq 50$ for which $\frac{Z_n}{\sqrt{2\log\log n}}$ gets arbitrarily close to 1. In such cases

$$Z_n \approx \sqrt{2\log\log n} \overset{(n\geq 50)}{\geq} 1.64,$$

so the null is rejected almost surely.

2. **[Good Optional Continuation, with S-Values]** Now consider the following statistic for the problem above, whose definition depends on a choice for a prior density w on μ:

$$S_{n,w}(X_1, \ldots, X_n) := \frac{\int_{\mu \in \mathbb{R}} p_\mu(X^n) w(\mu) d\mu}{p_0(X^n)},$$

where p_μ is the density of $X_1, \ldots, X_m$ under P_μ, and P_μ is as in the previous exercise.

   a) **[2 points]** Show that, for arbitrary prior density w and arbitrary n, $S_{n,w}(X^n)$ is an S-value.
   Let us denote $p_1(X^n) := \int_{\mu \in \mathbb{R}} p_\mu(X^n) w(\mu) d\mu$. Then we see

$$\begin{aligned} \mathbb{E}_{P_0}[S_{n,w}] &= \mathbb{E}_{P_0}\left[\frac{p_1(X^n)}{p_0(X^n)}\right] \\ &= \int_{x^n} p_0(x^n) \frac{p_1(x^n)}{p_0(x^n)} dx^n \\ &= \int_{x^n} p_1(x^n) dx^n = 1. \end{aligned}$$

   b) Consider the protocol of Exercise 1c). According to S-value theory, we can safely (in terms of Type-I error) output as evidence in this protocol

$$S'(X^{100}) := \begin{cases} S_{50,w}(X^{50}) & \text{if we stop at } n = 50 \\ S_{50,w}(X^{50}) \cdot S_{50,w}(X_{51}, \ldots, X_{100}) & \text{if we stop at } n = 100 \end{cases}$$

   Show (i) **[2 points]** that $S'(X^{100})$ is an S-value and (ii) **[1 point]** explain why this implies that $P_0(S'(X^{100}) \geq 1/\alpha) \leq \alpha$. (HINT for (i): let $\tau \in \{50, 100\}$ be the random variable indicating at what time we stop. Let $\mathbf{1}_{\tau=j}$ be the indicator function which takes value 1 if $\tau = j$ and 0 otherwise. Use that:

$$\mathbf{E}_{P_0}[S'(X^{100})] = \mathbf{E}_{P_0}[\mathbf{1}_{\tau=50} S'(X^{100}) + \mathbf{1}_{\tau=100} S'(X^{100})].$$

)
   Solution:
   Define τ as in the hint (note that τ depends on the first 50 data points, so the notation $1_{\tau(X^{50})=j}$ is slightly more appropriate). Then

$$\begin{aligned} \mathbb{E}_{P_0}[S'(X^{100})] &= \mathbb{E}_{P_0}[S'(X^{100}) \mathbf{1}_{\tau(X^{50})=50} + S'(X^{100}) \mathbf{1}_{\tau(X^{50})=100}] \\ &= \mathbb{E}_{P_0}[S_{50,w}(X^{50}) \mathbf{1}_{\tau(X^{50})=50}] \\ &+ \mathbb{E}_{P_0}[S_{50,w}(X^{50}) S_{50,w}(X_{51}, \ldots, X_{100}) \mathbf{1}_{\tau(X^{50})=100}]. \end{aligned}$$

For the first part, we see:

$$
\begin{aligned}
\mathbb{E}_{P_0}[S_{50,w}(X^{50})\mathbf{1}_{\tau(X^{50})=50}] &= \mathbb{E}_{P_0}\left[\frac{p_1(X^{50})}{p_0(X^{50})}\mathbf{1}_{\tau(X^{50})=50}\right] \\
&= \int_{x^{50}} p_0(x^{50})\frac{p_1(x^{50})}{p_0(x^{50})}\mathbf{1}_{\tau(X^{50})=50} dx^{50} \\
&= \int_{x^{50}} p_1(x^{50})\mathbf{1}_{\tau(X^{50})=50} dx^{50} \\
&= P_1(\tau(X^{50}) = 50).
\end{aligned}
$$

For the second part:

$$
\begin{aligned}
&\mathbb{E}_{P_0}[S_{50,w}(X^{50})S_{50,w}(X_{51},\ldots,X_{100})\mathbf{1}_{\tau(X^{50})=100}] \\
&\quad = \mathbb{E}\left[\frac{p_1(X^{50})}{p_0(X^{50})}\frac{p_1(X_{51},\ldots,X_{100})}{p_0(X_{51},\ldots,X_{100}}\mathbf{1}_{\tau(X^{50})=100}\right] \\
&\quad = \int_{x^{100}} p_0(x^{100})\frac{p_1(x^{50})}{p_0(x^{50})}\frac{p_1(x_{51},\ldots,x_{100})}{p_0(x_{51},\ldots,x_{100})}\mathbf{1}_{\tau(X^{50})=100} dx^{100} \\
&\quad = \int_{x^{100}} p_1(x^{50})p_1(x_{51},\ldots,x_{100})\mathbf{1}_{\tau(X^{50})=100} dx^{100} \\
&\quad = \int_{x^{50}} p_1(x^{50})\mathbf{1}_{\tau(X^{50})=100} \int_{x_{51},\ldots x_{100}} p_1(x_{51},\ldots,x_{100}) dx^{100} \\
&\quad = \int_{x^{50}} p_1(x^{50})\mathbf{1}_{\tau(X^{50})=100} dx^{50} \\
&\quad = P_1(\mathbf{1}_{\tau(X^{50})=100}).
\end{aligned}
$$

Putting everything together, we see

$$
\mathbb{E}_{P_0}[S'(X^{100})] = P_1(\tau(X^{50} = 50) + P_1(\tau(X^{50}) = 100) = 1.
$$

(ii) follows from Markov's inequality.

3. One possible interpretation of an S-value is as 'an inverse p-value equiped with a prior over how extreme it is'. For simplicity we only consider this interpretation in a special case in which the outcome space $\mathcal{X} = \mathbb{N}_0$ is countable and that there is just one outcome. Let $p(x)$ be a p-value for some given null distribution P_0, and let $\bar{p}$ be the corresponding random variable, i.e. if $X = x$ then $\bar{p} := p(x)$. Since $\mathcal{X}$ is countable, we do not require p to be strict, i.e. we have, for $0 \leq \alpha \leq 1$, $P_0(\bar{p} \leq \alpha) \leq \alpha$, with the rightmost inequality being strict for some α. For concreteness, you can think of the case that the null distribution P_0 expresses that X is distributed according to a Poisson distribution with given mean value parameter μ_0, and $p(x) := P_0(X \geq x)$; but your derivation should hold for general p-values on countable $\mathcal{X}$.

a) **[2 points]** Take an arbitrary probability mass function π on the values that $\bar{p}$ can take. Show that $S(\bar{p}) := \pi(\bar{p}) \cdot 1/\bar{p}$ is an S-value, i.e. $\mathbf{E}_{\bar{P}_0}[S(\bar{p})] \leq 1$.

Solution:

$$
\begin{aligned}
\mathbb{E}_{P_0}[S(\bar{p})] &= \mathbb{E}_{P_0}[\pi(\bar{p})1/\bar{p}] \\
&= \sum_{\bar{p}'} \mathbb{P}(\bar{p} = \bar{p}')\pi(\bar{p})1/\bar{p}' \\
&\leq \sum_{\bar{p}'} \mathbb{P}(\bar{p} \leq \bar{p}')\pi(\bar{p})1/\bar{p}' \\
&\leq \sum_{\bar{p}'} \bar{p}'\pi(\bar{p})1/\bar{p}' \\
&= \sum_{\bar{p}'} \pi(\bar{p}) = 1.
\end{aligned}
$$

b) **[2 points]** Suppose that P_0 has infinite support, i.e. there exist infinitely many $x \in \mathbb{N}$ such that $P_0(X = x) > 0$. For bonus points, you may show that for every $c > 0$, $S'(\bar{p}) := c/\bar{p}$ is *not* an S-value, i.e. $\mathbf{E}_{\bar{P}_0}[S'(\bar{p})] = \infty$. (This means that we really need to 'downweight' $1/\bar{p}$ for small $\bar{p}$ to make it an S-value). (HINT: first consider the case where P_0 is a distribution such that $\bar{p} \in \{1, 1/2, 1/4, 1/8, \ldots\}$) (NOTE: this bonus exercise is not so easy so don't spend too much time on it).

Solution: Let $A = \{p(x) : P_0(X = x) > 0\}$ be the countable support of $p(X)$, and order the elements of A in decreasing order as $\alpha_1, \alpha_2, \ldots$ (in the case of the HINT, we have $\alpha_j = 2^{-j+1}$). The statement is actually only correct if the p-value is strict on A, i.e. for all $\alpha \in A$ we have $P_0(p(X) \leq \alpha) = \alpha$. In the case referred to in the hint, we have

$$
\begin{aligned}
\mathbb{E}_{P_0}\left[\frac{1}{p(X)}\right] &= \sum_j P(p(X) = \alpha_j)\frac{1}{\alpha_j} \\
&= \sum_j \frac{P(p(X) \leq \alpha_j) - P(p(X) \leq \alpha_{j+1})}{\alpha_j} \\
&= \sum_j \frac{\alpha_j - \alpha_{j+1}}{\alpha_j} \\
&= \sum_j \left(1 - \frac{\alpha_{j+1}}{\alpha_j}\right).
\end{aligned}
$$

In the case of the hint, this expectation is clearly infinite ($\alpha_{j+1}/\alpha_j = 1/2$). For the general case, take a sequence $1 = j_1 < j_2 < j_3 < \ldots$

such that for all i, we have $\alpha_{j_i+1} \leq \alpha_{j_i}/2$. We then have

$$\begin{aligned}
\mathbb{E}_{P_0}\left[\frac{1}{p(X)}\right] &= \sum_i \sum_{j:j=j_i,j_i+1,\ldots j_{i+1}-1} P(p(X)=\alpha_j)\frac{1}{\alpha_j} \\
&\geq \sum_i \sum_{j:j=j_i,j_i+1,\ldots j_{i+1}-1} \frac{P(p(X)=\alpha_j)}{\alpha_{j_i}} \\
&= \sum_i \frac{P(p(X)\leq \alpha_{j_i}) - P(p(X)\leq \alpha_{j_{i+1}})}{\alpha_{j_i}} \\
&= \sum_j \left(1 - \frac{\alpha_{j_i+1}}{\alpha_j}\right)
\end{aligned}$$

and again the result follows.