Assignment sheet 7

October 31, 2018

Assignment 1. The p-value, as a function of the sample, is a random variable taking values in $[0, 1]$. We will see the distribution of this random variable using simulations. Let $X_1, \dots, X_n \sim N(\mu, 1)$ be independent, and consider testing the null hypothesis $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$.
(a) What is the distribution of $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$?
(b) Using part (a), find a number v_α such that the test function $\mathbf{1}\{|\overline{X}| > v_\alpha\}$ has significance level α.
(c) Find a formula, as explicit as possible, to the p-value of the test, as a function of the sample $X_1, \dots, X_n$.
(d) (Optional) Use the formula in (c) to empirically find the distribution of the p-value under the null : fix n, generate $X_1, \dots, X_n$, and compute the p-value. Repeat this `REP` times. Store all the p-values in a numerical vector `p` of length `REP`. Use the command `hist(p)` to plot a histogram of `p`. What do you observe ? What happens when you change n ? What happens to this distribution under the alternative ?

Assignment 2. (a) Let $X_1, X_2, \dots, X_n$ be an i.i.d. sample from the $N(\theta, 1)$ distribution. Find $100(1 - \alpha)\%$ confidence regions for θ by
(i) inverting the likelihood ratio test, and
(ii) inverting the asymptotic test obtained using the approximation $\sqrt{n}(\overline{X} - \theta) \xrightarrow{d} N(0, 1)$, i.e. the test from the previous assignment.
(iii) Are the two confidence intervals the same ?
(b) Let $X_1, X_2, \dots, X_n$ e an i.i.d. sample from the $Ber(p)$ distribution. Find $100(1 - \alpha)\%$ confidence regions for p by
(i) inverting the asymptotic likelihood ratio test obtained using Wilks' theorem,
(ii) inverting the Wald test, and
(iii) inverting the asymptotic test obtained using the approximation $\sqrt{n}(\overline{X} - p)/\sqrt{p(1 - p)} \xrightarrow{d} N(0, 1)$.
(iv) Are these confidence intervals the same ?

Assignment 3. Let $X_1, X_2, \dots, X_n$ be an i.i.d. sample from a continuous distribution with continuous density function f. Let $h > 0$ and define a partition $\{I_j\}_{j\in\mathbb{Z}}$ of $\mathbb{R}$, where $I_j = [\kappa + (j - 1)h, \kappa + jh)$ for a fixed real number κ.
The *histogram* of $X_1, X_2, \dots, X_n$ with *bin-width* h and *origin* κ is defined as the function $x \mapsto \text{Hist}_{X_1,X_2,\dots,X_n}(x)$, where

$$\text{Hist}_{X_1,X_2,\dots,X_n}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}(X_i \in I_j) \quad \text{if } x \in I_j$$

for each $j \in \mathbb{Z}$.
(a) Show that $\int_{-\infty}^{\infty} \text{Hist}_{X_1,X_2,\dots,X_n}(x)dx = 1$.
(b) Find the distribution of $nh\text{Hist}_{X_1,X_2,\dots,X_n}(x)$ for each x. Hence, find its mean and variance.
(c) What happens to $\mathbb{E}[\text{Hist}_{X_1,X_2,\dots,X_n}(x)]$ when $h \to 0$.
(Note : This limit indicates what $\text{Hist}_{X_1,X_2,\dots,X_n}(x)$ is estimating for each x for sufficiently small h.)
(d) Using part (b), find $\mathbb{E}\{[\text{Hist}_{X_1,X_2,\dots,X_n}(x) - f(x)]^2\}$.
(e) What happens to the value of the mean squared error in part (e) when $h \to 0$ and $nh \to \infty$?
(f) Interpret the limits $h \to 0$ and $nh \to \infty$.
(g) Is $\text{Hist}_{X_1,X_2,\dots,X_n}(x)$ a consistent estimator of $f(x)$? Justify your answer.

Assignment 4. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$ are unknown parameters. In this problem, we will find different types of confidence regions for (μ, σ^2).
(a) Find $100(1-\alpha)\%$ confidence interval for μ (when σ is known) and σ^2 (when μ is unknown) separately using $\overline{X}$ and S^2. Denote these by $R_{1,\alpha}(\mathbf{X})$ and $R_{2,\alpha}(\mathbf{X})$.
(b) Is the region $R_{1,\alpha}(\mathbf{X}) \times R_{2,\alpha}(\mathbf{X})$ a $100(1-\alpha)\%$ confidence region for (μ, σ^2) ? Otherwise, find β (depending on α) such that $R_{1,\beta}(\mathbf{X}) \times R_{2,\beta}(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence region for (μ, σ^2) using the Bonferroni method.
(c) Use the independence of $\overline{X}$ and S^2 to find β such that $R_{1,\beta}(\mathbf{X}) \times R_{2,\beta}(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence region for (μ, σ^2). Denote this region by $R_A(\mathbf{X})$.
(Note : $R_A(\mathbf{X})$ is called the Mood exact region.)
(d) Which one of the above confidence regions obtained in (b) and (c) do you think is preferable ? Justify your answer.
(e) Write down the likelihood ratio statistic for testing the hypothesis $H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$ vs $H_1 : \mu \neq \mu_0, \sigma^2 \neq \sigma_0^2$.
(f) Use Wilks' theorem and the expression of the likelihood ratio statistic to derive an asymptotic $100(1-\alpha)\%$ confidence region for (μ, σ^2). Denote this region by $R_B(\mathbf{X})$.
FACT : It can be proved that $\sqrt{n}\{(\overline{X}, S^2)^\top - (\mu, \sigma^2)^\top\}$ converges in distribution to $(Z_1, Z_2)^\top$, where $Z_1 \sim N(0, \sigma^2)$, $Z^2 \sim N(0, 2\sigma^4)$, and they are independent.
(g) Use the above fact to find the asymptotic distribution of $U_n = n(\overline{X} - \mu)^2/\sigma^2 + n(S^2 - \sigma^2)/(2\sigma^4)$.
(h) Use part (g) to find an asymptotic $100(1-\alpha)\%$ confidence region for (μ, σ^2). Denote this region by $R_C(\mathbf{X})$.
(i) What is the asymptotic distribution of $V_n = n(\overline{X} - \mu)^2/S^2 + n(S^2 - \sigma^2)/(2S^4)$?
(j) Use part (i) to find an asymptotic $100(1-\alpha)\%$ confidence region for (μ, σ^2). Denote this region by $R_D(\mathbf{X})$.
(k) (optional) Suppose that $n = 10$, $\overline{x} = 0$, $s^2 = 1$ and $\alpha = 0.05$. Write a code in `R` to understand how the above four 95% confidence regions, namely, $R_A(\mathbf{X}), R_B(\mathbf{X}), R_C(\mathbf{X})$ and $R_D(\mathbf{X})$ look like.
(*Hint : Each confidence region will be a set of the form* $\{(\mu, \sigma^2) : H(\mu, \sigma^2) \leq h\}$, *where* H *is a real-valued function and* h *is a real number. You will get the function* H *after you simplify and put the values of* n, $\overline{x}$, s^2 *and* α. *To draw this set, you can use the following code*)

```
f <- function(a,b) H(a,b)
mu_vals <- seq(from=-1,to=1,length=100)
sig_vals <- seq(from=0.5,to=1.5,length=100)
z <- outer(mu_vals,sig_vals,f)
contour(mu_vals,sig_vals,z,levels=h,drawlabels=FALSE)
abline(h=1,v=0,col="red")
```

What do you observe ? What do you observe if you take $n = 25$ and $n = 100$?
(Note : It can be proved that the likelihood based region $R_B(\mathbf{X})$ asymptotically has the smallest expected area.)

Assignment 5. (Optional) In this exercise we will explore how, in testing many hypotheses simultaneously, compiling a list on tests based on a small p-values cut-off might result in many false positives with high probability.

Mars Reconnaissance Orbiter is a NASA orbiter that aims to prove that water persisted on the surface on Mars for long period of times. Assume Europe sent its own orbiter to check for presence of liquid water on Mars. The orbiter will snap a picture at 100 pre-sampled location. Under the null, each sampled location is assumed to have a probability of 5% to host water. To have more certainty, the orbiter will gravitate around the planet until it has taken 200 pictures of each location.

(i). Run the following code and comment it.

```
set.seed(25102017)
positions <- 100;
trials <- 200;
true.p <- 0.05;
alpha <- 0.01;
nrep <- 1000


p <- matrix(replicate(positions*nrep,prop.test(rbinom(1,trials,true.p),
              trials,true.p)$p.value),nrep)
dim(p)
mean(apply(p,1,min)<alpha)
```

(ii). What happens if $\alpha = 0.05$? Comment the result.

(iii). Use the function `p.adjust` to adjust the p-values using Bonferroni, Holms and Hochberg's correction.

(iv). What happens when you change the numbers ? For example, if trials "small" ? How could you overcome this ?