STA 114: Statistics

Notes 18. 2-sided & 1-sided ML tests, Fixed level testing, P-values

Two-sided ML tests for normal models

In last lecture we saw that for $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, σ^2 fixed, a size α ML test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is given by

$$\textit{reject } H_0 \text{ if and only if } \mu_0 \notin \bar{x} \mp z(\alpha)\frac{\sigma}{\sqrt{n}}, \text{ i.e., } \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} > z(\alpha).$$

What about a size α ML test for the same hypotheses when σ is not assumed known? We need a bit of a care here, because H_0 is no longer a point null! It contains all (μ, σ^2) for which $\mu = \mu_0$ and $\sigma^2 > 0$ is arbitrary. And so size calculation actually requires taking a maximum of the power function over a non-singleton set.

However, it is not difficult to derive the form of an ML test. This is because an ML test rejects H_0 if and only if $L_x^*(\mu_0) \geq k \max_{\mu \in (-\infty, \infty)} L_x^*(\mu)$ where $L_x^*(\mu) = \max_{\sigma^2 > 0} L_x(\mu, \sigma^2)$ is the profile likelihood in μ. So the ML test rejects H_0 if and only if μ_0 does not belong to the corresponding profile ML interval for μ. We know the form of these intervals: $B_c(x) = \bar{x} \mp c s_x/\sqrt{n}$ with a flat coverage $2\Phi_{n-1}(c) - 1$ at every (μ, σ^2). Therefore an ML test is of the form:

$$\textit{reject } H_0 \text{ if and only if } \mu_0 \notin \bar{x} \mp c\frac{s_x}{\sqrt{n}} \text{ i.e., } \frac{|\bar{x} - \mu_0|}{s_x/\sqrt{n}} > c$$

with size $= \max_{\sigma^2 > 0}\{1 - \gamma((\mu_0, \sigma^2); B_c)\} = 2\{1 - \Phi_{n-1}(c)\}$. Again, with $c = z_{n-1}(\alpha)$, the corresponding ML test has size α. In summary, for $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, (μ, σ^2) unknown, and size α ML test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is given by:

$$\textit{reject } H_0 \text{ if and only if } \mu_0 \notin \bar{x} \mp z_{n-1}(\alpha)\frac{s_x}{\sqrt{n}} \text{ i.e., } \frac{|\bar{x} - \mu_0|}{s_x/\sqrt{n}} > z_{n-1}(\alpha)$$

By a similar argument we can say that for $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma^2)$, $Y_1, \cdots, Y_m \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma^2)$, a size α ML test for $H_0 : \mu_1 - \mu_2 = \eta_0$ against $H_1 : \mu_1 - \mu_2 \neq \eta_0$ is given by

$$\textit{reject } H_0 \text{ if and only if } \eta_0 \notin (\bar{x} - \bar{y}) \mp z_{n+m-w}(\alpha)\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

$$\text{i.e., } \frac{|(\bar{x} - \bar{y}) - \eta_0|}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}} > z_{n+m-2}(\alpha)$$

One-sided ML tests

In the drug study example, suppose a currently available soporific drug is known to give an hour's of extra sleep on average. Then it is more reasonable to test whether, with the new drug, we have $H_0 : \mu \leq 1$ against $H_1 : \mu > 1$. Such hypotheses are called one-sided, as they only care about one side of an existing standard, rather than exact match with the current standard. To distinguish from this case, the hypotheses $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$ are called two-sided.

Deriving a size α ML test for a pair of one-sided hypotheses is usually easy if we know the size α ML tests for the two sided version. To set ideas, let's look at $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, where σ is fixed and we're interested in $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_1$. We know the ML intervals of μ are of the form $\bar{x} \mp c\sigma/\sqrt{n}$. The corresponding ML test rejects H_0 when no $\mu \leq \mu_0$ is within the ML interval, which happens if and only if $\mu_0 < \bar{x} - c\sigma/\sqrt{n}$.

Call this test $\delta_c(x)$, we'll calculate its size. For any $\mu \leq \mu_0$,

$$\pi(\mu; \delta_c) = P_{[X|\mu]}(\mu_0 < \bar{X} - c\sigma/\sqrt{n}) \leq P_{[X|\mu]}(\mu < \bar{X} - c\sigma/\sqrt{n}) = 1 - \Phi(c).$$

Therefore, $\alpha(\delta_c) = \max_{\mu \leq \mu_0} \pi(\mu; \delta_c) = 1 - \Phi(c)$. Because $c \geq 0$, the size is never larger than $1/2$. So, for any $\alpha \in (0, 1/2]$ a size α ML test for $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ is given by

$$\textit{reject } H_0 \text{ if and only if } \mu_0 < \bar{x} - z(2\alpha)\frac{\sigma}{\sqrt{n}}, \text{ i.e., } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z(2\alpha).$$

Note the use of $z(2\alpha)$ instead of $z(\alpha)$. No size α ML test exists for $\alpha > 1/2$.

By symmetry, if we instead wanted to test $H_0 : \mu \geq \mu_0$ against $H_1 : \mu < \mu_0$ then a size α ML test (for $\alpha \leq 1/2$) would be given by

$$\textit{reject } H_0 \text{ if and only if } \mu_0 > \bar{x} + z(2\alpha)\frac{\sigma}{\sqrt{n}}, \text{ i.e., } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z(2\alpha).$$

Similar arguments lead us to the results on Table 1.

Fixed level testing

Neyman and Pearson advocated the following approach to carry out testing. Once you have set up the model and the hypotheses, choose a small $\alpha \in (0, 1)$, usually 1%, 5% or 10%. Next choose a size α test with good power at the alternatives (usually an ML test if possible), and then accept or reject H_0 based on this test. If α was chosen 5% and the corresponding test returns *reject* H_0, then we say the null hypothesis is rejected at 5% significance level [although a more correct description would be to add the phrase "based on ML tests", etc.]. If the test returns *accept* H_0, then we say we failed to reject the null hypothesis at 5% level of significance.

The purpose of choosing α beforehand is that you're stating upfront how conservative you are about H_0 (smaller α means more conservative). Choosing the level equal 5% means that you're willing to entertain an erroneous rejections of H_0 at most in 5% cases.

P-value

Consider the sleep study example where for $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ our observed data had $n = 10$, $\bar{x} = 2.33$ and $s_x = 2$. Suppose we want to test $H_0 : \mu = 0$ against $\mu \neq 0$ at 5% level. So we check whether 0 belongs to the interval $\bar{x} \mp z_{n-1}(.05)s_x/\sqrt{n} = [0.899, 3.761]$, which it doesn't, so we reject H_0 at 5% significance level.

Now suppose our data instead had $\bar{x} = 1.44$, then this interval would be $[0.009, 2.871]$ and we would still reject H_0 at 5%, but only marginally so. Although we make the same decisions in either case, we do it with very different levels of assurance. In the second case we were very close to taking the other decision (accept H_0).

To reflect the strength of assurance in our decision, Fisher recommended reporting the p-value, which is the smallest size α test which rejects H_0 based on the observed data. The smaller the p-value, the more assurance we have against H_0.

To understand why Fisher recommended this, consider the following. Suppose you have an infinite number of testers, each using a different size α (ML) test. Together, they cover the whole range $\alpha \in (0, 1)$. The testers with smaller α are more conservative about H_0, they need to see more evidence against H_0 to reject it. Next you show your recorded data to all testers and each take a decision to reject/accept H_0. The most liberal testers, those with α very close to 1, would be quick to report *reject* H_0 while the most conservative ones will stick to *accept* H_0. In between, there's a point of switch, a value $\alpha_0(x)$ so that all testers with $\alpha \geq \alpha_0(x)$ have rejected H_0 and all testers with $\alpha < \alpha_0(x)$ have failed to reject H_0. This switch point is the p-value. The smaller the switch point, the more compelling the evidence against H_0 has been (converting more conservatives).

Keep in mind that the p-value is subjective to the collection of tests (with size covering the whole range) you decided to use. This is why we'd talk about "p-value based on ML tests" or "p-value based on median tests", etc. (see HW9). Also remember that if you find the ML based p-value to be 0.04, then the size 5% ML test would reject H_0, as would any other size α ML test with $\alpha \geq 0.04$. On the other hand, a size 1% ML test, and any other size α ML test with $\alpha < 0.04$ would accept H_0.

In our sleep study example, to calculate p-value, we simply find the α for which 0 is just on the border of the interval $\bar{x} \mp z(\alpha)\sigma/\sqrt{n}$. With $\bar{x} = 2.33$ we get p-value = 0.005, while with $\bar{x} = 1.44$ we would have p-value = 0.049. So we have more compelling evidence against H_0 in the first case than in the second.

The border matching trick is universal – it applies to all tests listed on Table 1, including the one-sided ones. For example, if we were testing $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$ with $n = 10$, $\bar{x} = 2.33$ and $s_x = 2$, then we would find α such that $2.33 - z(2\alpha) \times 2/\sqrt{10} = 0$ which gives $\alpha = 0.0025$. Hence ML test based p-value for these one sided hypotheses is 0.0025.

Model	Hypotheses	Size α ML test rejects H_0 if and only if
$X = (X_1, \cdots, X_n)$	$H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0$	$\mu_0 \notin \bar{x} \mp z(\alpha)\frac{\sigma}{\sqrt{n}}$
$X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$	$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$	$\mu_0 < \bar{x} - z(2\alpha)\frac{\sigma}{\sqrt{n}}$
σ known	$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$	$\mu_0 > \bar{x} + z(2\alpha)\frac{\sigma}{\sqrt{n}}$
$X = (X_1, \cdots, X_n)$	$H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0$	$\mu_0 \notin \bar{x} \mp z_{n-1}(\alpha)\frac{s_x}{\sqrt{n}}$
$X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$	$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$	$\mu_0 < \bar{x} - z_{n-1}(2\alpha)\frac{s_x}{\sqrt{n}}$
	$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$	$\mu_0 > \bar{x} + z_{n-1}(2\alpha)\frac{s_x}{\sqrt{n}}$
$X = (X_1, \cdots, X_n)$,$Y = (Y_1, \cdots, Y_m)$	$H_0 : \mu_1 - \mu_2 = \eta_0,\ H_1 : \mu_1 - \mu_2 \neq \eta_0$	$\eta_0 \notin (\bar{x} - \bar{y}) \mp z_{n+m-2}(\alpha)\sqrt{(\frac{1}{n} + \frac{1}{m})\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
$X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $Y_j \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma^2)$	$H_0 : \mu_1 - \mu_2 \leq \eta_0,\ H_1 : \mu_1 - \mu_2 > \eta_0$	$\eta_0 < (\bar{x} - \bar{y}) - z_{n+m-2}(2\alpha)\sqrt{(\frac{1}{n} + \frac{1}{m})\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
X_i's, Y_j's independent	$H_0 : \mu_1 - \mu_2 \geq \eta_0,\ H_1 : \mu_1 - \mu_2 < \eta_0$	$\eta_0 > (\bar{x} - \bar{y}) + z_{n+m-2}(2\alpha)\sqrt{(\frac{1}{n} + \frac{1}{m})\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$
$X = (X_1, \cdots, X_n)$,$Y = (Y_1, \cdots, Y_m)$	$H_0 : \mu_1 - \mu_2 = \eta_0,\ H_1 : \mu_1 - \mu_2 \neq \eta_0$	$\eta_0 \notin (\bar{x} - \bar{y}) \mp z_{r(x,y)}(\alpha)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$
$X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma_1^2)$, $Y_j \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma_2^2)$	$H_0 : \mu_1 - \mu_2 \leq \eta_0,\ H_1 : \mu_1 - \mu_2 > \eta_0$	$\eta_0 < (\bar{x} - \bar{y}) - z_{r(x,y)}(2\alpha)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$
X_i's, Y_j's independent	$H_0 : \mu_1 - \mu_2 \geq \eta_0,\ H_1 : \mu_1 - \mu_2 < \eta_0$	$\eta_0 > (\bar{x} - \bar{y}) + z_{r(x,y)}(2\alpha)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$
Regular models $X \sim f(x\|\theta)$	$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$	$\theta_0 \notin \hat{\theta}_{\text{MLE}}(x) \mp z(\alpha)/\sqrt{I_x}$
(binomial, Poisson, exponential)	$H_0 : \theta \leq \theta_0, H_1 : \theta > \theta_0$	$\theta_0 < \hat{\theta}_{\text{MLE}}(x) - z(2\alpha)/\sqrt{I_x}$
θ scalar	$H_0 : \theta \geq \theta_0, H_1 : \theta < \theta_0$	$\theta_0 > \hat{\theta}_{\text{MLE}}(x) + z(2\alpha)/\sqrt{I_x}$

Table 1: 2- and 1-sided ML tests of size α. For one sided tests, $\alpha \leq 1/2$. For the fourth model, $r(x, y)$ refers to Welch's approximation. Size calculations for fourth and fifth models are approximate.