STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 18

Testing using p-value

What is a *p-value*?

The p-value for a test

- The *p-value* for a test is a *statistic*, i.e. a function of the data $\mathbf{X}$ (and therefore is a random variable).
- It is defined to be *the smallest α level at which observing $\mathbf{X}$ will lead to a rejection of the test*. (Note the dependence on $\mathbf{X}$.)
- In other words, it is the smallest α level for which the corresponding rejection region of the test covers $\mathbf{X}$.

$$p(\mathbf{X}) = \inf\{\alpha : \mathbf{X} \in \mathscr{R}(\alpha)\}.$$

- (Draw a figure.)

- So for sensible tests, the smaller the p-value, the *stronger* the evidence *against* H_0.
- For example, for the LR test, what is the meaning of the p-value? (Think about the gold miner analogy.)
- It is the minimal "budget" you need for considering purchasing the land that covers the observed data.
- Depending on the context, people use "*p-value*" to refer to either the *p-value statistic* or the observed value of the p-value statistic.

Lot testing example

- We observe $X_1, X_2, \ldots, X_n$ i.i.d. Exponential(λ).
- Consider testing

$$H_0 : \lambda = 1.0 \quad \text{vs} \quad H_1 : \lambda > 1.0.$$

The UMP level-α test rejects when

$$\bar{X} < F^{-1}(\alpha)$$

where F is the cdf of Gamma$(n, n\lambda)$. So the rejection region is

$$\mathscr{R}(\alpha) = \{(x_1, x_2, \ldots, x_n) : \bar{x} < F^{-1}(\alpha).\}$$

What is the p-value?

- If we observe $X_1, X_2, \dots, X_n$, what is the p-value?
- By definition of the p-value, we know it is

$$\begin{aligned} p(\mathbf{X}) &= \inf\{\alpha : (X_1, X_2, \dots, X_n) \in \mathscr{R}(\alpha)\} \\ &= \inf\{\alpha : \bar{X} < F^{-1}(\alpha)\} \\ &= \inf\{\alpha : \alpha > F(\bar{X})\} \\ &= F(\bar{X}). \end{aligned}$$

- For example, if $\bar{X} = 0.7$ and $n = 20$, then

$$p(\mathbf{x}) = F(0.7) \approx 0.077.$$

A heuristic interpretation of the p-value

启发式，探索式

- The p value is the “size” of the rejection (as measured by the Type I error) that “barely covers” X.
- This rejection region contains the data values that provide at least as much evidence in favor of the alternative as the data $\mathbf{X}$.
- For generalized LR tests, this rejection region is the region that contains all data values that correspond to a Λ value no larger than the observed Λ value.
- The gold miner analogy.
- So the p-value can be intuitively thought of as the probability to get more “extreme” values of the data under the null.

- Back to the lot testing example, the p-value is given by

$$p(\mathbf{X}) = F(\bar{X}).$$

- This is indeed the chance of observing more "extreme" values under H_0. (Draw a figure.)

Example: Finding the p-value of a two-sided t-test

- Suppose our data are i.i.d. from $\mathrm{N}(\mu, \sigma^2)$ where both the mean μ and the variance σ^2 are unknown.
- Consider testing

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_1$$

at level α.

- We have seen that the level α LR test leads to the t test which rejects when
$$|T| > C$$
with $T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{s}$, and $C = F^{-1}_{t_{n-1}}(1 - \frac{\alpha}{2})$.
- In other words, the rejection region, which depends on the level α, is

$$\mathscr{R}(\alpha) = \left\{ (x_1, x_2, \ldots, x_n) : \frac{\sqrt{n}|\bar{x} - \mu_0|}{s} > F^{-1}_{t_{n-1}}(1 - \frac{\alpha}{2}) \right\}.$$

- So the p-value is

$$
\begin{aligned}
p(\mathbf{X}) &= \inf\{\alpha : \mathbf{X} \in \mathscr{R}(\alpha)\} \\
&= \inf\left\{\alpha : |T| > F_{t_{n-1}}^{-1}(1 - \frac{\alpha}{2})\right\} \\
&= \inf\left\{\alpha : F_{t_{n-1}}(|T|) > 1 - \frac{\alpha}{2}\right\} \\
&= \inf\left\{\alpha : \alpha > 2\left(1 - F_{t_{n-1}}(|T|)\right)\right\} \\
&= 2\left(1 - F_{t_{n-1}}(|T|)\right).
\end{aligned}
$$

  (Draw a figure.)

- For example, if $\mu_0 = 1$, $n = 20$, $\bar{X} = 2.2$, $s^2 = 2.3$, then the p-value is

$$
2\left(1 - F_{t_{n-1}}\left(\frac{\sqrt{10} \times |2.2 - 1|}{\sqrt{2.3}}\right)\right) = 0.0216.
$$

- This is indeed the probability to observe a value "more extreme" than $\mathbf{X}$ under H_0. (Draw a figure.)

Testing using p-values

- The null is rejected with the test at level α, *if and only if* the p-value is no larger than α. Why? Let's consider both directions.
  - Note that $p(\mathbf{X}) = \inf\{\alpha : \mathbf{X} \in \mathscr{R}(\alpha)\}$ means that

    $$p(\mathbf{X}) \leq \alpha_1 \Leftrightarrow \alpha_1 \in \{\alpha : \mathbf{X} \in \mathscr{R}(\alpha)\} \Leftrightarrow \mathbf{X} \in \mathscr{R}(\alpha_1).$$

    Equivalently,

    $$p(\mathbf{X}) > \alpha_1 \Leftrightarrow \alpha_1 \notin \{\alpha : \mathbf{X} \in \mathscr{R}(\alpha)\} \Leftrightarrow \mathbf{X} \notin \mathscr{R}(\alpha_1).$$
  - The gold miner analogy.

The p-value as a measure of evidence against the null

- The p-value, however, contains more information than whether the test is rejected or not at a given level, say, $\alpha = .05$.
- It measures the strength of evidence against the null hypothesis.
- E.g., $p \approx .048$ compared to $p \approx .00001$. Both suggest that the null hypothesis is rejected at level α, but $p \approx .00001$ provides much stronger evidence against H_0.
- So in the previous lot testing example and two-sample t-test example, do we reject H_0 at level 0.1, 0.05, 0.01?
- The p-value gives more information about the evidence from the data than just a reject/accept decision for a particular test level.

A problem with the p-value

- Note that it says nothing about how much the data supports the alternative in *absolute* terms.
- In the gold miner analogy: It is the total price for purchasing all the land that are more precious (i.e., with higher gold-to-dollar ratio) than a given point (the observed data), but we have no idea how much gold it contains, which may be very little!

Sampling distribution of the p-value under H_0

- We have seen that the p-value is a statistic, and thus it also has a sampling distribution.
- Question: If we repeat the experiment many many times, what is the distribution of the p-value that we observe under H_0?

- Let us try to find the cdf of this sampling distribution.
- For $\alpha \in [0,1]$, we have by the definition of the p-value

$$P(p(\mathbf{X}) \leq \alpha | H_0) = P(\mathbf{X} \in \mathscr{R}(\alpha) | H_0).$$

- But $P(\mathbf{X} \in \mathscr{R}(\alpha) | H_0)$ is exactly the Type I error corresponding to the rejection region $R(\alpha)$, which by definition is α.
- Therefore

$$P(p(\mathbf{X}) \leq \alpha | H_0) = \alpha \quad \text{for } \alpha \in [0,1].$$

  What distribution has this cdf?

p-value and the non-reproducibility of published science

- Under H_0, $p(\mathbf{X})$ has a standard uniform distribution!
- So under the null, there is 5% chance that the p-value will be no more than 5%!
- Why are most published science nonreproducible?
- What does a p-value of 0.9999 mean?

The multiple testing problem

- If we test many null hypotheses, even if all of them are true, by chance we may have some small p-values.
- In such cases, the p-values lose their nominal meaning and must be "corrected".

$$\begin{aligned} P(\min\{U_1, U_2, \dots, U_T\} \leq \alpha) &= 1 - P(U_1 > \alpha) \times \cdots \times P(U_T > \alpha) \\ &= 1 - (1-\alpha)^T, \end{aligned}$$

which increases to 1 as T goes to infinity!