Elements of Hypothesis Testing

Definition. *Given* $X_1, \cdots, X_n \sim f_\theta$ *Let* $p_\triangle = p_{tr} - p_{pl}$.

1. ***Null hypothesis*** $\mathcal{H}_0$ *is a simple hypothesis associated with a contradiction to a theory one would like to prove.*
2. ***Alternative Hypothesis*** *is a hypothesis (often composite) associated with a theory one would like to prove.*
3. ***Type I Error*** *incorrectly rejecting* $\mathcal{H}_0$ *(i.e. a false discovery / false positive)*
4. ***Type II Error*** *incorrectly retaining (not rejecting)* $\mathcal{H}_0$ *when we should (i.e. false negative)*
5. ***Parameter Space*** Θ *is holds all possible value of* θ *(in this case,* $\Theta = [-1, 1]$*) where competing hypothesis are of form*
$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$
6. ***Simple Hypothesis*** *is any hypothesis which specifies the population distribution completely.*
$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$
where $\Theta = \{\theta_0, \theta_1\}$.
7. ***Composite Hypothesis*** *If a hypothesis does not completely specify the probability distribution. For example, when deciding if a particular model fits the data. The alternative hypothesis is any distribution that is not the initially intended one; hence alternative hypothesis not specified*
8. *A* ***Statistical test*** *is a procedure whose inputs are samples and whose result is a hypothesis, i.e. it is a data driven probabilitstic decision rule with regard to* $\mathcal{H}_0$
9. ***p-value*** *The probability, assuming the* $\mathcal{H}_0$ *is true, of observing a result at least as extreme as the test statistic.*

Definition. *Suppose we test simple hypothesis*
$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

1. ***Size of a test*** *is the probability of incorrectly rejecting the* $\mathcal{H}_0$*, or equivalently the probability of Type I error (false positive). For composite hypotheses this is the supremum of the probability of rejecting the* $\mathcal{H}_0$ *over all cases covered by the null hypothesis.*

2. ***Significance level of a test*** *α is the upper bound imposed on the size of a test. Testing H_0 at significance level α means testing H_0 with a test whose size does not exceed α.*

$$\alpha = \mathbb{P}(rejecting \quad \mathcal{H}_0 | \theta = \theta_0)$$

3. ***Power of a test*** *is the probability of correctly rejecting the $\mathcal{H}_0$, equivalently the probability of NOT making a type II error, $\pi = 1 - \beta$ (false negative)*

$$\beta = \mathbb{P}(not \ \ rejecting \ \ \mathcal{H}_0 | \theta = \theta_1)$$

Note α and β are tradeoffs. Minimizing α is given priority to minimizing β.

Definition. ***Test statistic*** *$T(\underline{X})$ is a sample statistic whose distribution under $\mathcal{H}_0$ is known, which allows p-value to be calculated. It is selected or defined in such a way as to quantify, within observed data, behaviours that would distinguish the null from the alternative hypothesis.*

Remark. In determining if a coin is fair, the likelihood ratio of, or equivalently the number of heads, is called the test statistic. The probability distribution of number of heads when coin is fair ($\mathcal{H}_0$) is $Binom(10, 0.5)$ assuming 10 flips.

Definition. ***Rejection Region*** *The set of values of the test statistic that leads to rejection of $\mathcal{H}_0$ is called the rejection region. Statistical test that is based on test statistic $T(\underline{X})$ with rejection region $\mathcal{C} \subseteq \mathbb{R}^n$ is of the form*

$$reject \quad \mathcal{H}_0 \quad if \quad T(\underline{x}) \in \mathcal{C}$$

($\underline{x}$ is a sample statistic calculated from observation)

Remark. Given $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, with $\mathcal{H}_0 : \mu = \mu_0; \mathcal{H}_1 : \mu = \mu_1 > \mu_0$. Here we proposed test is to reject $\mathcal{H}_0$ if $\overline{X} \geq c$, for some c. Test statistic $\overline{X}$ has the following distribution under $\mathcal{H}_0$

$$\overline{X} \overset{\mathcal{H}_0}{\sim} \mathcal{N}(\mu_0, \frac{\sigma^2}{n})$$

And we want to find c such that Type I error of the test is the specified α (such that our test is restricted to a confidence level of α)

$$\mathcal{C} = \{(x_1, \cdots, x_n) \in \mathbb{R}^n : \overline{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\}$$

So the test **reject $\mathcal{H}_0$ if** $\overline{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$ has a confidence level α

Definition. ***Likelihood Ratio statistic*** *Given simple hypothesis*

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

based on $X_1, \cdots, X_n \sim f_\theta$. *The* ***Likelihood Ratio*** *is a statistic given by*

$$\lambda(\underline{x}) = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{f(x_1, \cdots, x_n|\theta_1)}{f(x_1, \cdots, x_n|\theta_0)}$$

which represents how likely $\mathcal{H}_1$ *is true compared to* $\mathcal{H}_0$

Definition. ***Likelihood Ratio Test (LRT)*** *A statistical test based on the region*

$$\mathcal{C} = \{\underline{x} \in \mathbb{R}^n : \lambda(\underline{x}) = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} \geq c\}$$

for some c *satisfying* $\mathbb{P}(\lambda(\underline{x}) \geq c|\theta = \theta_0) = \alpha$ *is called the the likelihood ratio test at significance level* α. *Intuitively, LRT rejects* $\mathcal{H}_0$ *if* $\lambda(\underline{x})$ *is big enough. How big is big enough depends on the significance level of the test, i.e., on what probability of Type I error is considered tolerable*

Remark. As an example, for $X_1, \cdots, X_n \overset{\mathcal{H}_0}{\sim} \mathcal{N}(\mu, \sigma^2)$ **reject** $\mathcal{H}_0$ **if** $\overline{x} \geq c$ is a likelihood ratio test. ($\underline{x}$, the test statistic is the result from evaluating $\lambda(\underline{x}) = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\lambda_0)}$) Now we choose c such that $\mathbb{P}(\overline{X} \geq c|\mu = \mu_0) = \alpha$, which gives

$$c = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$$

The strategy is to evaluate the likelihood ratio. Knowning null distribution of test statistic ($\overline{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ in this case), makes it possible to choose a critical level c such that the test yields a desired confidence level of α. As we will see, the resulting test is also the most power test at significance level α

LRT is often used for goodness-of-fit test to see if a particular distribution model fits the data.

Definition. ***Most Powerful Test*** *Suppose observed* $X_1, \cdots, X_n \sim f_\theta$ *and consider the simple hypothesis* $\mathcal{H}_0 : \theta = \theta_0$ *vs* $\mathcal{H}_1 : \theta = \theta_1$. *We say that the most powerful (MP) test at level* α *if*

1. *the significant level of the test is* α
2. *no other test at level* α *has a smaller* β

In other words, for a given size or significance level, the test with the greatest power for a given value of the parameter(s) being tested, contained in the alternative hypothesis.

Theorem. ***Heyman-Pearson Lemma*** *When performing a hypothesis test between two simple hypotheses* $\mathcal{H}_0 : \theta = \theta_0; \mathcal{H}_1 : \theta = \theta_1$, *the likelihood-ratio test based on rejection region*

$$\mathcal{C} = \{\underline{x} \in \mathbb{R}^n : \lambda(\underline{x}) = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\lambda_0)} \geq c\} \qquad \textit{where} \qquad \mathbb{P}(\lambda(\underline{x}) \geq c|\theta = \theta_0) = \alpha$$

is the most powerful test at significance level α *for a threshold* c

Proof. Proof consists of considering

$$\alpha = \mathbb{P}(\underline{X} \in \mathcal{C}|\theta = \theta_0) = \int_{\mathcal{C}} f(\underline{(x)}|\theta_0) d\underline{x} = \int_{\mathcal{C}} \mathcal{L}(\theta_0)$$

over disjoint sets $\mathcal{C} = (\mathcal{C} \cap \mathcal{D}) \cup (\mathcal{C} \cap \overline{\mathcal{D}})$ where $\mathcal{D}$ is the rejection region of an alternative test. Here we prove that the power π for LRT is higher than that of the alternative test π'. We do this by recognizing that

$$\pi = \mathbb{P}(\underline{x} \in \mathcal{C}|\theta = \theta_1) = \int_{\mathcal{C}} \mathcal{L}(\theta_1) d\underline{x} \geq \int_{\mathcal{D}} \mathcal{L}(\theta_1) d\underline{x} = \mathbb{P}(\underline{x} \in \mathcal{D}|\theta = \theta_1) = \pi'$$

where the inequality is derived from $\dfrac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} \geq c$ over $\mathcal{C}$ and $\dfrac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} < c$ over $\mathcal{D} \subseteq \overline{\mathcal{C}}$

□

Remark. **Interpretation** The strength of Neyman-Pearson approach is that only distribution only null hypothesis is needed in order to construct a test. However, the theory requires specifying the significance level, α, in advance of analyzing the data, but gives no buidance about how to make this choice. Another criticism is that it is built on the assumption that one must either reject of not reject a hypothesis. So say we made observation and find $t(\underline{x}) = 0.41$. the null hypothesis would have been rejected with $\alpha = 0.5$, if one were to reject / not reject. The *evidence* of whether or not rejection is made is summarized as p-value, which is the smallest significance level at which the null hypothesis would be rejected. (i.e. p-value $= 0.04$). In essence, *p-value is a summarized evidence against null hypothesis.* The smaller the p-value the stronger the evidence against null.

Remark. **Choise of Null**

1. It is conventional to choose simpler of the hypothesis as the null, i.e. distribution under which is easy to characterize.
2. Choose null where incorrectly rejecting it would cause graver issue than the other hypothesis being null.
3. Contextually, null hypothesis is often a simple explanation that must be discredited in order to demonstrate the presence of some physical phenomenon or effect.