Common Fallacies in Applying Hypothesis Testing

Sanjeev G.
SPC Lab.,
Dept. of ECE,
IISc.

8 January 2011.

Agenda

- "While the value of this paper to an expert on hypothesis testing is limited, the points it raises are important for practitioners."
- Fallacies in applying single hypothesis tests.
- Fallacies in applying Neyman-Pearson tests.
- Not Rejected $\neq$ Accepted to be true.
- Conclusions.

Misuse of statistics

- Meta-criticism.
- Philosophical criticism.
- Pedagogic criticism.
- Practical criticism.
- Bayesian criticism.

Single HT and Cournot's principle

- Say, $x_1, x_2, \cdots, x_n$ are the observations with $p(x;\theta)$. Single hypothesis : $\mathcal{H} : \theta = \theta_0$.
- "Fate" of this HT is not either rejection or acceptance.
- Cournot's principle - Assuming the theory is incorrect, a verification of any of its bold predictions would be extremely improbable; given such verifications, the underlying assumption that it is incorrect should be abandoned.
- Test 1 : Obtain $\widehat{\theta}$, find (θ_l, θ_u) such that $P\{\widehat{\theta} \in (\theta_l, \theta_u)|\mathcal{H}\} = 1 - \alpha$, for a small α.
- Test 2 : Find $(\widehat{\theta_l}, \widehat{\theta_u})$, such that $P\{(\widehat{\theta_l}, \widehat{\theta_u}) \ni \theta_0|\mathcal{H}\} = 1 - \alpha$, for a small α.

Fisherian Hypothesis Testing

- A statistically significant result is the one which is unlikely to have occurred by chance.
- p-value is the probability of observing data atleast as extreme as that observed, given that the null hypothesis is true.
- Fisherian p-values are philosophically different than the NP type I error!
- Argument from ignorance - If something has not been disproved $\neq$ it is proved to be accepted.
- $\downarrow$ the p-value, less likely the result is if null hypothesis is true, and more "significant" the result is.

Fisherian Hypothesis Testing

- A statistically significant result is the one which is unlikely to have occurred by chance.
- p-value is the probability of observing data atleast as extreme as that observed, given that the null hypothesis is true.
- Fisherian p-values are philosophically different than the NP type I error!
- Argument from ignorance - If something has not been disproved $\neq$ it is proved to be accepted.
- $\downarrow$ the p-value, less likely the result is if null hypothesis is true, and more "significant" the result is.

Fisherian Hypothesis Testing

- A statistically significant result is the one which is unlikely to have occurred by chance.
- p-value is the probability of observing data atleast as extreme as that observed, given that the null hypothesis is true.
- Fisherian p-values are philosophically different than the NP type I error!
- Argument from ignorance - If something has not been disproved $\neq$ it is proved to be accepted.
- $\downarrow$ the p-value, less likely the result is if null hypothesis is true, and more "significant" the result is.

Fisherian Hypothesis Testing

- A statistically significant result is the one which is unlikely to have occurred by chance.
- p-value is the probability of observing data atleast as extreme as that observed, given that the null hypothesis is true.
- Fisherian p-values are philosophically different than the NP type I error!
- Argument from ignorance - If something has not been disproved $\neq$ it is proved to be accepted.
- $\downarrow$ the p-value, less likely the result is if null hypothesis is true, and more "significant" the result is.

Fisherian Hypothesis Testing

- A statistically significant result is the one which is unlikely to have occurred by chance.
- p-value is the probability of observing data atleast as extreme as that observed, given that the null hypothesis is true.
- Fisherian p-values are philosophically different than the NP type I error!
- Argument from ignorance - If something has not been disproved $\neq$ it is proved to be accepted.
- $\downarrow$ the p-value, less likely the result is if null hypothesis is true, and more “significant” the result is.

Single HT and Cournot's principle - revisited

- Say, $x_1, x_2, \cdots, x_n$ are the observations with $p(x;\theta)$. Single hypothesis : $\mathcal{H} : \theta = \theta_0$.
- "Fate" of this HT is not either rejection or acceptance.
- Cournot's principle - Assuming the theory is incorrect, a verification of any of its bold predictions would be extremely improbable; given such verifications, the underlying assumption that it is incorrect should be abandoned.
- Test 1 : Obtain $\widehat{\theta}$, find (θ_l, θ_u) such that $P\{\widehat{\theta} \in (\theta_l, \theta_u)|\mathcal{H}\} = 1 - \alpha$, for a small α.
- Test 2 : Find $(\widehat{\theta_l}, \widehat{\theta_u})$, such that $P\{(\widehat{\theta_l}, \widehat{\theta_u}) \ni \theta_0|\mathcal{H}\} = 1 - \alpha$, for a small α.

Correct test

- For a small (but not very small value) of α,
  - $\widehat{\theta} \notin (\theta_l, \theta_u)$ refutes $\mathcal{H}$ on solid grounds;
  - $\widehat{\theta} \in (\theta_l, \theta_u)$ weakly supports $\mathcal{H}$ on shaky ground; but the evidence is non conclusive.
- For an extremely small α,
  - $\widehat{\theta} \notin (\theta_l, \theta_u)$, reject $\mathcal{H}$;
  - $\widehat{\theta} \in (\theta_l, \theta_u)$, make no decision since it only supports $\mathcal{H}$ extremely weakly on very shaky ground.

Neyman Pearson Hypothesis Testing

- $\mathcal{H}_0 : \theta \in \Theta_0$ vs. $\mathcal{H}_1 : \theta \in \Theta_1$.
- **Type I error** : $P\{\text{accept } \mathcal{H}_1 | \mathcal{H}_0 \text{ true}\}$,
  Type II error : $P\{\text{accept } \mathcal{H}_0 | \mathcal{H}_1 \text{ true}\}$, and
  Power : $P\{\text{accept } \mathcal{H}_1 | \mathcal{H}_1 \text{ true}\}$.
- The analogous terminology in signal detection theory are **false alarm**, **miss**, and **detection** probabilities.

Guidelines

- The type I error should be assigned to the one with a more serious consequence.
- The null hypothesis should be the one whose rejection is of most interest.
- The generally accepted hypothesis should be designated as the null one.
- The null hypothesis should be the one under which the distribution of the test statistic is known better.

Examples

- Criminal-court example.
- Drug-development example.
- Scientific theory example.
- Detection of HIV by T cell count.
- Spectrum Sensing in CRs.

Differences between Single and NP HTs

- NP HT presumes that one and only one of the hypotheses should be rejected.
- In single HT, the question is whether the hypothesis is rejected. NP HT concerns with which hypothesis is better accepted.
- A single HT can be formulated as a NP test. However, the standard NP procedures cannot be applied because the distribution under $\mathcal{H}_1$ is not known. Therefore Fisherian formulation is more appropriate than its NP counterpart.
- In single HT, no conclusion can be drawn if $\mathcal{H}_0$ is not rejected. NP in such scenario "accepts" $\mathcal{H}_1$, in the sense that it minimizes an optimization criterion.

Not Rejected $\neq$ Accepted

- The truth may not be either of the hypotheses!
- In formulations like the GLRT, Bayes' test (prior pdf is averaged over the random parameters), the surrogate simple hypothesis is not even neccessarily a member of the original composite hypothesis.
- A not rejected hypothesis is accepted only when (a) No evidence disproves it, and (b) Its bold predictions are verified.
- The authors claim that for NP HT, there is no analogous to what Cournot's principle is to the Single HT.

Conclusions

- Inappropriate acceptance of single hypothesis when it is not rejected based on the observations.
- Blind assignment of the null hypothesis in NP testing.
- The two classes of tests work on two distinct principles.
- The truth could be none of the hypotheses!

Reference

- X. Rong Li, and Xiao-Bai Li, Common Fallacies in Applying Hypothesis Testing, *11th International Conference on Information Fusion, 2008.*