Testing Statistical Hypothesis

November 15, 2023

Motivation

Suppose we have discovered a new drug that we believe will increase the rate of recovery from some disease over the recovery rate when an old established drug is applied.

- Our hypothesis is against the complementary (alternative) hypothesis that the new drug does not improve on the old drug.
- Suppose that we know from past experience that a fixed proportion $\theta_0 = 0.3$ recover from the disease with the old drug.
- What the complementary hypothesis means is that the chance that an individual randomly selected from the ill population will recover is the same or worse with the new drug than with the old drug.
- To investigate this question we would have to perform a random experiment and use the data in the support of one or the other claim.
- Most simply, we would sample n patients, administer the new drug, and then base our decision on the observed sample $X = (X_1, ..., X_n)$, where X_i is 1 if the ith patient recovers and 0 otherwise.

Two hypotheses

- Suppose we observe $S = \sum_{i=1}^{n} X_i$, the number of recoveries among the n randomly selected patients who have been administered the new drug.
- If we let θ be the probability that a patient to whom the new drug is administered recovers, then S has a $B(n, \theta)$ distribution.
- If we suppose the new drug is at least as effective as the old, then $\Theta = [0, 1]$, where θ_0 is the probability of recovery using the old drug.
- Now, one of the hypotheses is $\theta \in [0, \theta_0]$ and the other is that $\theta \in (\theta_0, 1]$.
- As it will be seen from our methodology, there is a need to differentiate the two hypotheses by deciding for which of the two we need a strong evidence from the data.
- This decision is subjective but it is very consequential.
- In the example, one could argue that the strong evidence from the data is needed for the claim that the new drug is better than the old one (one does not want to start a new production of a drug if it could turn out that the drug is not better than the old one).

The null versus alternative

- The hypothesis for which we want a strong evidence to go along with it is called the alternative hypothesis and denoted by H_a, K or H_1.
- The other hypothesis is referred to as the null hypothesis and is denoted by H or H_0.
- In the text, the convention of H vs. K has been chosen.
- We write $H : \theta \leq \theta_0$ vs $K : \theta > \theta_0$.
- After deciding on what we want to test for, and which of the hypotheses seeks a strong support from the data, we need to decide how to use the data for the purpose.
- In this example, it is reasonable to reject H is S is "much" bigger than what would be expected by chance if H is true.
- We base our decision if $S > k$, or not, for some value of k to be yet decided.

Formulation of the problem

- Suppose that we are going to observe the value of a random vector $\mathbf{X}$. Let $\mathcal{X}$ denote the set of possible values that $\mathbf{X}$ can take and, for $\mathbf{x} \in \mathcal{X}$, let $p(\mathbf{x}|\theta)$ denote the density (or probability mass function) of $\mathbf{X}$, where the parameter θ is some unknown element of the set Θ.
- The null hypothesis specifies that θ belongs to some subset Θ_0 of Θ. The question arises as to whether the observed data $\mathbf{x}$ is consistent with the hypothesis that $\theta \in \Theta_0$, often written as $H : \theta \in \Theta_0$.
- The null hypothesis is contrasted with the so-called alternative hypothesis $K : \theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$ for which we need a strong evidence from the data if we want to go along with it.
- The testing hypothesis is aiming at finding in the data $\mathbf{x}$ enough evidence to reject the null hypothesis:

  $$H : \theta \in \Theta_0,$$

  in favor of the alternative hypothesis

  $$K : \theta \in \Theta_1.$$

- Due to the focus on control of the error rate for rejecting H, the set up in the role of the hypotheses is not exchangeable.

Two types of error

In a hypothesis testing situation, two types of error are possible.

- The first type of error is to **reject the null hypothesis** $H : \theta \in \Theta_0$ as being inconsistent with the observed data $\mathbf{x}$ when, in fact, $\theta \in \Theta_0$ i.e. when, in fact, the null hypothesis happens to be true. This is referred to as **Type I Error**.
- The second type of error is to **fail to reject the null hypothesis**[1] $H : \theta \in \Theta_0$ as being inconsistent with the observed data $\mathbf{x}$ when, in fact, $\theta \in \Theta_1$ i.e. when, in fact, the null hypothesis happens to be false. This is referred to as **Type II Error**.

The goal is to propose a procedure that for given data $\mathbf{X} = \mathbf{x}$ would automatically point which of the hypothesis is more favorable.

- It must be done in such a way that chances of making Type I Error are some prescribed small $\alpha \in (0, 1)$ – the **significance level** of a test (it is also called the **size** of the test but it is not a common terminology although it is used in the text).
- For given data $\mathbf{x}$, we evaluate a statistic $T(\mathbf{x})$ that is called a *test statistic* and if it falls in a certain *critical region* R_α (often also called *rejection region*), we reject H in the favor of K. We demand that $T(\mathbf{x})$ and R_α are chosen in such a way that Type I Error is at most α, i.e. for $\theta \in \Theta_0$

$$P(T(\mathbf{X}) \in R_\alpha | \theta) \leq \alpha,$$

[1] Notice, the asymmetry is the language.

The *p*-value

- The test procedure can be identified with a test statistic $T(\mathbf{x})$ and a rejection region R_α.
- It is quite natural to expected that R_α is decreasing with α (it should be harder to reject H if error 1 is smaller).
- Thus for a given sample $\mathbf{x}$, there should be an $\hat{\alpha}$ such that for $\alpha > \hat{\alpha}$ we have $T(\mathbf{x}) \in R_\alpha$ and for $\alpha < \hat{\alpha}$ the test statistics $T(\mathbf{x})$ is outside R_α.
- The value $\hat{\alpha}$ is called the ***p*-value** for a given test.
- This is also named the **observed significant level**, is only dependent on the data and is independent of the choice of α.
- We observe that $\hat{\alpha} \geq \alpha$ means that we do not reject H and $\hat{\alpha} < \alpha$ means the we reject H.
- The p-value can be viewed as a version of the test statistics (remember that it does depend on the original choice of the test statistics T).
- We observe that if $P(T(\mathbf{X}) \in R_\alpha | \theta_0) = \alpha$, than it means that $P(\hat{\alpha} < \alpha | \theta_0) = \alpha$ and thus the distribution of the p value is uniform on $[0, 1]$ under $H : \theta = \theta_0$.

The power of a test

While the focus in setting a testing hypothesis problem is on Type I Error (controlled by the significance level), it is also important to have chances of Type II Error as small as possible.

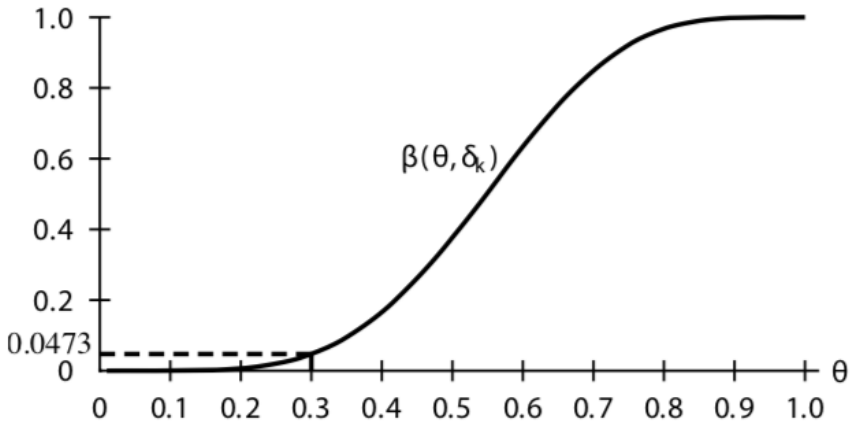
- For a given testing procedure smaller chances of Type I Error are at the cost of bigger chances of Type II Error.
- However, the chances of Type II Error can serve for comparison of testing procedures for which the significance level is the same.
- For this reason, the concept of the *power* of a test has been introduced.
- **The power of a test** is a function $\beta(\theta) = \beta(\theta, T)$ of $\theta \in \Theta_1$ and equals the probability of rejecting H while the true parameter is θ, i.e. under the alternative hypothesis
- Among two tests in the same problem and at the same significance level, the one with larger power for all $\theta \in \Theta_1$ is considered *uniformly* better.
- The power of a given procedure is increasing with the sample size of data, therefore it is often used to determine a sample size so that not rejecting H will represent a strong support for H, not only a lack of evidence for the alternative.

Comparing tests

- For a given test procedure δ consider for $\theta \in \Theta$:

$$\beta(\theta) = \beta(\theta, \delta) = P_\theta(\delta(X) = 1) = P_\theta(T(X) \in R_\alpha)$$

- $\theta \notin \Theta_0$: $\beta(\theta)$ is smaller than the significance level α.
- $\theta \notin \Theta_1$: $\beta(\theta)$ is the probability of not making of Type II Error: the bigger this probability the better.
- The test δ_1 is uniformly more powerful than δ_2, both on the same significance level, if $\beta(\theta, \delta_1) \leq \beta(\theta, \delta_2)$ for all $\theta \in \Theta_1$.



Typical graph of a power function

Testing for the mean

Suppose the data consist of a random sample $X_1, X_2, \ldots, X_n$ from a $\mathcal{N}(\theta, 1)$ density. Let $\Theta = (-\infty, \infty)$ and $\Theta_0 = (-\infty, 0]$ and consider testing $H : \theta \in \Theta_0$, i.e.

$$H : \theta \leq 0$$

The standard estimate of θ for this example is $\bar{X}$. The bigger the positive value of $\bar{X}$ that we observe the stronger is the evidence against the null hypothesis that $\theta \leq 0$, in favor of the alternative $\theta > 0$. So $T(\mathbf{X}) = \bar{X}$ is a sensible test statistics.

- How big does $\bar{X}$ have to be in order for us to reject H? In other words we want to determine the rejection region R_α.
- It is quite natural to consider $R_\alpha = [a_\alpha, \infty)$, so we reject H if $\bar{X}$ is too large, i.e. $\bar{X} \geq a_\alpha$. To determine a_α we recall that controlling Type I Error means that

$$\mathbb{P}(\bar{X} \geq a_\alpha | \theta) \leq \alpha, \ \theta \leq 0$$

and

$$\mathbb{P}(\bar{X} \geq a_\alpha | \theta) \leq \mathbb{P}(\bar{X} \geq a_\alpha | \theta = 0) = 1 - \Phi(a_\alpha \sqrt{n}),$$

from which we get that $a_\alpha = z_{1-\alpha}/\sqrt{n}$.

The form of the power function

Clearly in our case we have

$$\begin{aligned}\beta(\theta) &= P(\bar{X} \geq a_\alpha|\theta) \\ &= P\left((\bar{X}-\theta)\sqrt{n} \geq z_{1-\alpha/2} - \theta\sqrt{n}|\theta\right) \\ &= 1 - \Phi(z_{1-\alpha/2} - \theta\sqrt{n}).\end{aligned}$$

- We see clearly that $\beta(0) = \alpha$.
- The function is increasing to one when θ is increasing: *the power of the test to detect the bigger θ is increasing*.
- The function is increasing to one when n is increasing: *the power of the test to detect a given $\theta > 0$ is increasing with more data available.*

The optimal test

- Ideally, one would like to have the test that is **uniformly most powerful (UMP)**, i.e. a test δ such that for any other test $\tilde{\delta}$ on the same level
$$\beta(\theta, \delta) \geq \beta(\theta, \tilde{\delta}), \ \theta \in \Theta_1.$$
- In general, the problem may be not solvable but in one special case it has a complete solution.
- Namely if both Θ_0 and Θ_1 are made of one element.
- We call such hypotheses **simple**.
- The optimal test is based on the **likelihood ration**.

The likelihood ratio

- This test statistic is based on the idea that the log likelihood at θ_0 should bigger than the log likelihood at θ_1, if $H : \theta = \theta_0$ is correct.
- The test statistic is

$$T_1(\mathbf{x}) = \log \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} = \log L(\theta_1|\mathbf{x}) - \log L(\theta_0|\mathbf{x})$$

- Reject H if $T_1 > k$.
- As we will see by a proper choice of k this test is the most powerful no matter what a form of the distribution $p(x|\theta)$ of the data one considers.
- $k = k_\alpha = q_{1-\alpha}$, where q_p is the p-quantile of T_1 under H, since by the definition of a quantile we have

$$P_{\theta_0}(T_1 > k) \leq P_{\theta_0}(T_1 \geq q_{1-\alpha}) \leq 1 - (1 - \alpha) = \alpha$$
$$P_{\theta_0}(T_1 \leq k) \leq 1 - \alpha$$

- When the distribution of T_1 is continuous we simply have $P_{\theta_0}(T_1 > k) = 1 - \alpha$ and we assume that from now on.

Formulation of the lemma

The Neyman-Pearson lemma provides us with a way of finding most powerful tests. It demonstrates that the likelihood ratio test is the most powerful for the above problem. To avoid distracting technicalities of the non-continuous case we formulate and prove it for the continuous distribution case.

Lemma (The Neyman-Pearson lemma)

Let R_α be a subset of the sample space defined by

$$R_\alpha = \{\mathbf{x} : L(\theta_1|\mathbf{x})/L(\theta_0|\mathbf{x}) \geq k\}$$

where k is uniquely determined from the equality

$$\alpha = P\left(\mathbf{X} \in R_\alpha|\theta_0\right).$$

Then R_α defines the most powerful test at the significance level α for testing the simple hypothesis $H : \theta = \theta_0$ against the alternative simple hypothesis $K : \theta = \theta_1$.

The proof of the lemma

Proof.

We need to prove that if $\mathcal{A}$ is another critical region of size α (so the corresponding test is at the significance level α), then the power of the test associated with R_α is at least as great as the power of the test associated with $\mathcal{A}$, i.e.

$$P_{\theta_1}(\mathcal{A}) \leq P_{\theta_1}(R_\alpha).$$

We have for the set indicator function $\mathbf{1}_{\mathcal{A}}$:

$$\begin{aligned}
P_{\theta_1}(\mathcal{A}) &= E_{\theta_0}\left(\mathbf{1}_{\mathcal{A}}\frac{L_{\theta_1}}{L_{\theta_0}}\right) \\
&= E_{\theta_0}\left(\mathbf{1}_{R_\alpha}\frac{L_{\theta_1}}{L_{\theta_0}}\right) - E_{\theta_0}\left(\mathbf{1}_{\mathcal{A}^c\cap R_\alpha}\frac{L_{\theta_1}}{L_{\theta_0}}\right) + E_{\theta_0}\left(\mathbf{1}_{\mathcal{A}\cap R_\alpha^c}\frac{L_{\theta_1}}{L_{\theta_0}}\right) \\
&\leq P_{\theta_1}(R_\alpha) - kP_{\theta_0}\left(\mathcal{A}^c\cap R_\alpha\right) + kP_{\theta_0}\left(\mathcal{A}\cap R_\alpha^c\right) \\
&= P_{\theta_1}(R_\alpha) + k\left(P_{\theta_0}\left(\mathcal{A}\cap R_\alpha^c\right) + P_{\theta_0}\left(\mathcal{A}\cap R_\alpha\right) - P_{\theta_0}\left(\mathcal{A}^c\cap R_\alpha\right) - P_{\theta_0}\left(\mathcal{A}\cap R_\alpha\right)\right) \\
&= P_{\theta_1}(R_\alpha) + k\left(P_{\theta_0}\left(\mathcal{A}\right) - P_{\theta_0}\left(R_\alpha\right)\right) \\
&= P_{\theta_1}(R_\alpha)
\end{aligned}$$

Example

Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(0,1)$, and and we want to test $H : \theta = \theta_0$ versus $K : \theta = \theta_1$, where $\theta_1 > \theta_0$. We should reject H if $Z = \sqrt{n}(\bar{X} - \theta')$ is large, or equivalently if $\bar{X}$ is large. We can now use the Neyman-Pearson lemma to show that the test is "best". The likelihood function is

$$L(\theta) = (2\pi)^{-n/2} \exp\{-\sum_{i=1}^{n}(x_i - \theta)^2/2\}.$$

According to the Neyman-Pearson lemma, a best critical region is given by

$$\frac{1}{n} \ln[L(\theta_1)/L(\theta_0)] \geq k_2.$$

But

$$\begin{aligned} \frac{1}{n} \ln \frac{L(\theta_1)}{L(\theta_0)} &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{(x_i - \theta_0)^2}{2} - \frac{(x_i - \theta_1)^2}{2} \right) = \frac{1}{2n} \sum_{i=1}^{n} [2(\theta_1 - \theta_0)x_i + \theta_0^2 - \theta_1^2] \\ &= (\theta_1 - \theta_0)\bar{x} + \frac{1}{2}[\theta_0^2 - \theta_1^2]. \end{aligned}$$

So the best test rejects H_0 when $\bar{x} \geq k$, where k is a constant. But this is exactly the form of the rejection region for the proposed test. Therefore, it is the "best".