Econ 325/327

Notes on Power of Test

By Hiro Kasahara

Type I error, Type II error, and Power of Test

When we test the null hypothesis given a test statistic, we control Type I error by setting the significance level α. Therefore, by construction, the probability of making Type I error is the same across different test statistics, i.e.,

$$\Pr(\text{Type I error}) = \Pr(\text{Reject } H_0 | \ H_0 \text{ is true}) = \alpha.$$

However, in general, the probability of making Type II error,

$$\Pr(\text{Type II error}) = \Pr(\text{Not Reject } H_0 | \ H_0 \text{ is false}),$$

is different across different test statistics. The power of test is defined as

$$\text{Power} = 1\text{-}\Pr(\text{Type II error}) = 1\text{-}\Pr(\text{Not Reject } H_0 | \ H_0 \text{ is false}).$$

Ideally, we would like to have small probabilities of both Type I error and Type II error but there is a trade off between making Type I error and making Type II error.

To understand the trade-off between Type I error and Type II error, consider the following example. In our justice system, a person on trial is assumed to be innocent until proven guilty. So, we set the null hypothesis to be a person to be innocent. Then, on trial, we evaluate the evidence and ask if there is strong enough evidence against the presumption that a person is innocent. But, how strong the evidence has to be to give a guilty verdict? In hypothesis testing, we make a probability of putting an innocent person in jail (Type I error) to be small: this probability is called the significance level α. Decreasing the value of α by demanding stronger evidence for guilty verdict is a good thing if a person on trial is in fact innocent (decreasing the probability of Type I error). However, demanding stronger evidence for guilty verdict may increase the probability of giving an innocent verdict to a guilty person (increasing the probability of Type II error).

Example

Let $\{X_1, X_2, ..., X_n\}$ be $n = 25$ observations, each of which is randomly drawn from normal distribution with mean μ and variance σ^2. The value of μ is not known while σ^2 is known and equal to 100. We are interested in testing the null hypothesis $H_0 : \mu \geq 5$ against the alternative hypothesis $H_1 : \mu < 5$. Consider hypothesis testing based on the following two different test statistics: (i) sample mean $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ and (ii) mean of the first four observations, $\hat{X} = (1/4)(X_1 + X_2 + X_3 + X_4)$. Suppose that the realized values of $\bar{X}$ and $\hat{X}$ are given by $\bar{X} = 2.0$ and $\hat{X} = 1.0$, respectively.

1. Test the null hypothesis $H_0 : \mu \geq 5$ against $H_1 : \mu < 5$ *using the test statistic* $\bar{X}$ at the significance level $\alpha = 0.1$.

   Answer: Under H_0, $\bar{X} \sim N(5, \sigma^2/n)$ with $\sigma^2/n = 100/25 = 4$. The critical value at $\alpha = 0.1$ is given by $5 - 1.28(\sigma/\sqrt{n}) = 5 - 1.28 \times 2 = 2.44$ and the rejection region is $(-\infty, 2.44)$. Since the realized value of $\bar{X} = 2.0$ is less than 2.44, we reject H_0.

2. Compute the p-value for testing the null hypothesis $H_0 : \mu \geq 5$ against $H_1 : \mu < 5$ *using the test statistic* $\bar{X}$.

   Answer: the p-value is defined by the smallest significance level at which H_0 is rejected when the realized value of $\bar{X} = 2.0$. p-value $= (\bar{X} - 5)/(\sigma/\sqrt{n}) \leq (2-5)/2) = \Pr(Z \leq -1.5) = 0.0668$.

3. Test the null hypothesis $H_0 : \mu \geq 5$ against $H_1 : \mu < 5$ *using the test statistic* $\hat{X}$ at the significance level $\alpha = 0.1$.

   Answer: Under H_0, $\hat{X} \sim N(5, \sigma^2/n)$ with $\sigma^2/n = 100/4 = 25$. The critical value is given by $5 - 1.28(\sigma/\sqrt{n}) = 5 - 1.28 \times 5 = -1.40$ and the rejection region is $(-\infty, -1.40)$. Since the realized value of $\hat{X} = 1.0$ is larger than -1.40, we do not reject H_0.

4. Compute (i) the power of test using *using the test statistic* $\bar{X}$ when the true value of μ is equal to 0 and (ii) the power of test *using the test statistic* $\hat{X}$ when the true value of μ is equal to 0. Based on the power comparison, which test statistics, $\bar{X}$ or $\hat{X}$, do you recommend using for hypothesis testing?

   Answer: (i) Power $= Pr(\bar{X} \leq 2.44 | \mu = 0) = Pr((\bar{X} - 0)/2 \leq (2.44 - 0)/2 | \mu = 0) = \Pr(Z \leq (2.44 - 0)/2) = \Pr(Z \leq 1.22) = 0.8888$. (ii) Power $= Pr(\hat{X} \leq -1.40 | \mu = 0) = Pr((\hat{X} - 0)/5 \leq (-1.40 - 0)/5 | \mu = 0) = \Pr(Z \leq (-1.40 - 0)/5) = \Pr(Z \leq -0.28) = 1 - 0.6103 = 0.3897$. The test using $\bar{X}$ is more powerful than the test using $\hat{X}$ and, thus, the test using $\bar{X}$ is recommended.

5. Compute the power of test using $\bar{X}$ when the true value of μ is 2, 4, and 9 for each of cases. What would happen to the power of test as the value of μ approaches 5 from below?

   Answer: The power of test decreases to $\alpha = 0.1$ as the true value of μ approaches 5.

6. Suppose that we have the sample size of $n = 100$, 10000, and 1000000. Consider testing $H_0 : \mu = 5$ against $H_1 : \mu < 5$ based on the sample mean $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. As the sample size increases, what would happen to the power of test.

   Answer: The power of test increases to 1.

In the above example, we consider testing a specific null hypothesis $H_0 : \mu = 5$ against $H_1 : \mu < 5$ using two different test statistics $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ for $n = 25$ and $\hat{X} = (1/4)(X_1 + X_2 + X_3 + X_4)$. It turns out that the hypothesis testing based on $\bar{X}$ leads to rejecting H_0 while the hypothesis testing based on $\hat{X}$ leads to not rejecting H_0. Which result should we trust more? The answer is the test statistic that gives the higher power.[1]

[1]Using the analogy in the trial example, having different test statistics is like having different prosecutors with varying ability. Different prosecutors have access to the same data but they summarize evidence

We find that the test based on $\bar{X}$ has the higher power than the test based on $\hat{X}$ but what makes the test based on $\bar{X}$ more powerful than the test based on $\hat{X}$? Both $\bar{X}$ and $\hat{X}$ are unbiased estimators of μ but $\bar{X}$ has a lower variance than $\hat{X}$. The test statistic that gives a lower variance has a higher power.[2] To better understand why the test statistic that gives lower variance leads to higher power, we provide more discussion on the power of test next.

Power of Test

Consider a generalized version of the above example as follows. Suppose we have $\{X_1, X_2, ..., X_n\}$, where $X_i \sim N(\mu, \sigma^2)$ with σ^2 is known. We test $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$, where μ_0 is constant. If we test H_0 based on $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ at the significance level α, then

$$\text{we reject } H_0 \text{ if } \quad \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha, \tag{1}$$

where z_α is defined by $\Pr(Z \geq z_\alpha) = \alpha$.

Suppose that the null hypothesis is false and the true value of μ is equal to μ_1 which is strictly smaller than μ_0 so that $\mu_1 < \mu_0$. We are interested in computing the power of test. In the above example, we set $n = 4$ or 25, $\mu_0 = 5$, $\mu_1 = 0$, and $\sigma^2 = 100$ but we may repeat the hypothesis test for any value of n, μ_0, μ_1, and σ^2. We analyze the power of test without specifying the value of n, μ_0, and σ^2 and then ask how changing the value of n, μ_1, and σ^2 will affect the power of test.

If the true value of μ is μ_1 instead of μ_0, $\bar{X}_n$ is normally distributed with mean μ_1 and variance σ^2/n so that $\frac{\bar{X}_n - \mu_1}{\sigma/\sqrt{n}} \sim N(0,1)$. Because we follow the decision rule given by (1), we have

$$\begin{aligned}
\text{Power} &= \Pr\left(\text{Reject } H_0 | \mu = \mu_1\right) \\
&= \Pr\left(\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha \middle| \frac{\bar{X}_n - \mu_1}{\sigma/\sqrt{n}} \sim N(0,1)\right) \\
&= \Pr\left(\frac{\bar{X}_n - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha \middle| \frac{\bar{X}_n - \mu_1}{\sigma/\sqrt{n}} \sim N(0,1)\right) \\
&= \Pr\left(Z \leq -z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right), \quad Z \sim N(0,1)
\end{aligned}$$

differently. In this analogy, the probability of putting an innocent person in jail (Type I error) is the same regardless of how evidence is summarized, but the probability of failing to put a guilty person in jail (Type II error) would be different depending on how evidence is summarized. A "good" prosecutor ($\bar{X}$) may summarize evidence better than a "bad" prosecutor ($\hat{X}$). Both "good" and "bad" prosecutors will have the same probability of putting an innocent person in jail but a "good" prosecutor $\bar{X}$, has a higher probability of putting a guilty person in jail (i.e., a higher power) than a "bad" prosecutor, $\hat{X}$.

[2]More generally, we can consider a class of test statistics of the form

$$\tilde{X} = \sum_{i=1}^n w_i X_i \quad \text{with } \sum_{i=1}^n w_i = 1.$$

Putting $w_i = 1/n$ leads to the lowest variance of $\tilde{X}$ and, therefore, the test based on $\bar{X}$ is the most powerful test.

Because z_α only depends on the choice of α (e.g., if $\alpha = 0.05$, then $z_\alpha = 1.645$), the power is determined by $\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}$. Note that $\mu_0 - \mu_1 > 0$ and so $\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} > 0$. Therefore, the area for Z defined by $\{Z \leq -z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\}$ is larger than the area defined by $\{Z \leq -z_\alpha\}$ because we are adding a positive number to the right hand side of inequality. It follows that

$$\text{Power} = \Pr\left(Z \leq -z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) > \Pr\left(Z \leq -z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) = \alpha. \tag{2}$$

Therefore, the power of test is larger than the significance level.

Further, the area $\{Z \leq -z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\}$ decreases as the value of $\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}$ decreases. Therefore,

$$\text{Power is a decreasing function of } \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}.$$

A few comments:

- If we view the power as a function of μ_1, the power as a function of μ_1 is called **power function**. The power decreases as μ_1 approaches μ_0.
- Because $\mu_0 - \mu_1 > 0$, the power is decreasing in $\sigma/\sqrt{n}$. Note that $\sigma/\sqrt{n}$ is the variance of $\bar{X}_n$. Therefore, we have multiple test statistics with different variances (provided that they are unbiased estimators), then the test statistic that has lower variance leads to higher power.
- In particular, the variance of $\bar{X}_n$ decreases as the sample size increases. Therefore, the power increases as the sample size increases.

Equation (2) also makes it clear what will happen to the power of test as $\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}$ approaches to 0:

$$\lim_{\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} \to 0} \Pr\left(Z \leq -z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) = \Pr\left(Z \leq -z_\alpha\right) = \alpha.$$

Because $\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} \to 0$ when $\mu_1 \to \mu_0$, the power decreases to the significance level as μ_1 approaches μ_0, which is the answer to question 5 in the above example. This is intuitive. As the value of μ_1 gets closer to μ_0, it gets more difficult to figure out whether the data is generated under $\mu = \mu_0$ or $\mu = \mu_1$ and, in the limit, there is no difference between μ_0 and μ_1.

We may also analyze the effect of the sample size on the power. Because we may write $\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}$ as $\sqrt{n} \times \frac{\mu_0-\mu_1}{\sigma}$, and because $\frac{\mu_0-\mu_1}{\sigma} > 0$, the value of $\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}$ increases to ∞ as n increases to ∞. Therefore, the power increases to 1 as the sample size increases to ∞ as

$$\lim_{n\to\infty} \Pr\left(Z \leq -z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) = \Pr\left(Z \leq \infty\right) = 1.$$

This is the answer to question 6. As the sample size increases to ∞, the variance of $\bar{X}_n$ approaches zero. Therefore, $\bar{X}_n$ contains the precise information on the population mean when the sample size approaches ∞ and, hence, we can correctly reject H_0 if H_0 is false when the sample size is infinity.