Statistics II : Introduction to Inference

Module 7: Testing of Hypothesis

1 Some Basic Concepts and Definitions

Hypothesis testing deals with evaluating the feasibility of two competing statements about the underlying population based on a random sample drawn from the same.

Definition 1. *(Hypothesis, Null and Alternative) A statement or conjecture about the population or population parameters is called a hypothesis.*

The two contradictory statements (hypotheses) in a hypothesis testing problem are called the null hypothesis, denoted by H_0, and the alternative hypothesis, denoted by H_1.

Testing of the hypothesis mainly deals with accepting or rejecting the null hypothesis H_0 given the data, given H_1 as the alternative.

Example 1. Suppose $X_1, \cdots, X_n$ is a random sample from $N(\mu, 1)$ distribution, and we are interested in testing $H_0 : \mu = 0$ against $H_1 : \mu > 0$. While testing the hypothesis H_0 we will collect evidence from the data in support of H_0, given H_1 as the alternative. If the sample mean is much larger than *zero*, that is indicative of the fact that the population mean is also much larger than zero (recall that $\bar{X}_n$ is a consistent estimate of $E(X) = \mu$). In that case we reject H_0, otherwise, we accept H_0.

Note that, if the sample indicates that the population mean is much smaller than zero then also, in view of H_1, we accept H_0.

Definition 2. *(One-sided or Two-sided Alternatives) Suppose we are testing $H_0 : \theta = \theta_0$. The possible alternatives can be $H_{1,1} : \theta = \theta_1$ where $\theta_1 > \theta_0$, $H_{1,2} : \theta = \theta_1$ where $\theta_1 < \theta_0$, $H_{1,3} : \theta > \theta_0$, $H_{1,4} : \theta < \theta_0$, $H_{1,5} : \theta \neq \theta_0$, etc. The first four alternatives are one-sided, whereas $H_{1,5}$ is a two-sided alternative.*

Definition 3. *(Simple and Composite Hypotheses) Under a hypothesis H, if the population distribution is completely specified, then H is called a simple hypothesis, otherwise, it is called a composite hypothesis.*

Example 1. (continue) Suppose $X_1, \cdots, X_n$ is a random sample from $N(\mu, 1)$ distribution, and we are interested in testing $H_0 : \mu = 0$ against $H_1 : \mu > 0$. Here H_0 is simple, but H_1 is composite.

Definition 4. *(Hypothesis Test) A hypothesis test is a set of rules that indicates which sample values (realizations) lead to acceptance of H_0, and which sample values lead to rejection of H_0. A testing procedure partitions the sample space into two regions: one, called acceptance region, leads to acceptance of H_0, and the other, called critical region, leads to rejection of H_0. In other words, if the observed sample falls in the critical region then H_0 is rejected, otherwise, H_0 is accepted.*

Definition 5. *(Critical and Acceptance Regions) Let the support of $\mathbf{x}$ be $S_X \subseteq \mathbb{R}^n$. A subset C of S_X (or, $\mathbb{R}^n$) such that if the data $\mathbf{x} \in C$ then H_0 is rejected, is called the critical region. A subset A of S_X (or, $\mathbb{R}^n$) such that if the data $\mathbf{x} \in A$ then H_0 is accepted, is called the acceptance region. Note that $S_X \subseteq C \cup A$.*

Sometimes it is convenient to define the test in terms of a function from $\phi : S_X \to [0, 1]$, such that

$$\phi(\mathbf{x}) = 1 \quad \text{if} \quad \mathbf{x} \in C, \qquad \text{and} \qquad \phi(\mathbf{x}) = 0 \quad \text{if} \quad \mathbf{x} \in A.$$

Such a function is called a test function.

Definition 6. *(Test Function) Any function ϕ from $S_X \to [0, 1]$ is known a test function.*

One can interpret the test function as: $\phi(\mathbf{x}) = P(\text{Reject } H_0 \mid \mathbf{X} = \mathbf{x})$.

2 Errors in Testing and Their Probabilities

In a testing of hypothesis two possible decision can be taken: i. *accept* H_0, and ii. *reject* H_0. These two decisions may lead to two possible errors which are given below.

Definition 7. *(Type I and Type II Errors) One may reject the null hypothesis when it is indeed true, or one may accept the null hypothesis when it is indeed false. The first type of error is called Type I error, and the second type of error is called Type II error.*

True State $\rightarrow$ *Decision* $\downarrow$	H_0 *is true*	H_0 *is false*
Accept H_0	*Correct decision*	*Type II error*
Reject H_0	*Type I error*	*Correct decision*

Definition 8. *(Probabilities of Type I and Type II Errors, Power) The probability of type I error is* $P(\mathbf{X} \in C \mid H_0)$*, where* C *is the critical region. The probability of type II error is* $P(\mathbf{X} \in \bar{C} \mid H_1)$*. The probability of the complement of type II error is called power. Thus* power *is the probability of rejecting* H_0 *when it is indeed false.*

Example 1. (continue) In the above example, suppose we construct the following test procedure based on n samples: If the sample mean is greater than or equal to 6 then we reject H_0. Then the probability of type I error is $P\left[\bar{X}_n \geq 6 \mid \bar{X}_n \sim N(0, 1/n)\right]$.

Definition 9. *(Power Function) Suppose we want to test* $H_0 : \theta \in \Theta_0$ *against* $H_1 : \theta \in \Theta_1$*, where* $\Theta_0 \cup \Theta_1 = \Theta$*, based on a random sample* $\mathbf{X}$*. Suppose further that a test function* ϕ *is proposed for testing* H_0 *against* H_1*. Then the power function of the test* ϕ *is*

$$\beta_\phi(\theta) = E_\theta\left[\phi(\mathbf{X})\right].$$

The function $\beta_\phi(\theta)$ can be interpreted as the probability of rejecting H_0 given θ (marginalizing $\mathbf{x}$). When $\theta \in \Theta_0$, then $\beta_\phi(\theta)$ provides the probability of type I error at θ, and when $\theta \in \Theta_1$, $\beta_\phi(\theta)$ provides the complement of probability of type II error (power) at θ.

Remark 1. *Ideally, one would like to set a test procedure for minimizing the probabilities of both types of errors. However, in general, if one tends to minimize the probability of one error then the probability of the other error increases. Thus, in practice one bounds the maximum probability of type I error to a pre-assigned level* α *(which is small enough), and then minimizes the probability of type II error. The pre-assigned threshold of maximum probability of type I error* α *is called the level of significance, or just the level of the test.*

Definition 10. *(Level of Significance) Let* ϕ *be a test function for testing* $H_0 : \theta \in \Theta_0$ *against* $H_1 : \theta \in \Theta_1$*. Then* ϕ *is called a level-*α *test, or a test with level of significance* α *if*

$$E_\theta\left[\phi(\mathbf{X})\right] \leq \alpha, \quad \text{for all} \quad \theta \in \Theta_0, \quad \text{or, equivalently,} \quad \sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha. \tag{1}$$

Definition 11. *(Size of a Test) Let* ϕ *be a test function for testing* $H_0 : \theta \in \Theta_0$ *against* $H_1 : \theta \in \Theta_1$*. Then the size of* ϕ *is* $\sup_{\theta \in \Theta_0} \beta_\phi(\theta)$*.*

Example 2. Suppose $X_1, \ldots, X_{10}$ be a random sample of size 10 from `Bernoulli`(p), and consider the testing problem $H_0 : p = 0.5$ against $H_1 : p = 0.75$. One would reject H_0 if more number of heads appear. What would be a good test procedure? Consider the following test procedures and the corresponding probabilities of type I errors and powers.

Test	Procedure	P(Type I error)	Power = 1 - P(Type II error)
ϕ_1	Reject if 8 or more heads appear	$\sum_{x=8}^{10} \binom{10}{x} (0.5)^{10}$ ≈ 0.0547	$\sum_{x=8}^{10} \binom{10}{x} (0.75)^x (0.25)^{10-x}$ ≈ 0.5256
ϕ_2	Reject if 9 or more heads appear	$\sum_{x=9}^{10} \binom{10}{x} (0.5)^{10}$ ≈ 0.0107	$\sum_{x=9}^{10} \binom{10}{x} (0.75)^x (0.25)^{10-x}$ ≈ 0.2440
ϕ_3	Reject if 10 heads heads appear	$\binom{10}{10} (0.5)^{10}$ ≈ 0.001	$\binom{10}{10} (0.75)^{10}$ ≈ 0.0563

Remark 2. *Given the above remark, a usually recommended test procedure is as follows: Suppose we want to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ at level α. Then we will only consider tests satisfying the level-α conditions (1). Then among the tests satisfying (1), we will consider the test having the highest power.*

3 Uniformly Most Powerful Test

Definition 12. *(Most Powerful (MP) Test) Suppose we are interested in testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ at level α, and Φ_α is the class of tests satisfying the level-α condition (i.e., for any test $\phi \in \Phi_\alpha$, (1) is satisfied). A test $\phi_0 \in \Phi_\alpha$ is called most powerful test against an alternative $\theta_1 \in \Theta_1$ if*

$$\beta_{\phi_0}(\theta_1) \geq \beta_\phi(\theta_1) \quad \text{for all } \phi \in \Phi_\alpha.$$

Definition 13. *(Uniformly Most Powerful (UMP) Test) Suppose we are interested in testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ at level α, and Φ_α is the class of tests satisfying the level-α condition. A test $\phi_0 \in \Phi_\alpha$ is called uniformly most powerful test if*

$$\beta_{\phi_0}(\theta_1) \geq \beta_\phi(\theta_1) \quad \text{for all } \phi \in \Phi_\alpha, \quad \text{uniformly in } \theta_1 \in \Theta_1 \ (\text{i.e., for all } \theta_1 \in \Theta_1).$$

Example 2. (continue) Suppose we want to find the MP test for testing H_0 against H_1 in the last example at level 0.05. Of course the test ϕ_1 does not satisfy the level condition. Both ϕ_2 and ϕ_3 satisfy the level condition. However, ϕ_2 has higher power than ϕ_3. So, ϕ_2 should be preferred over ϕ_3. Can we construct a better test than ϕ_2?

Suppose I consider a test as follows:

- If the number of heads is 9 or more, then H_0 is rejected.
- If the number of heads is 8, then select a random number U from Uniform$(0,1)$. If the realized value of U, say u, satisfies $u < 0.85$, then reject H_0, otherwise accept H_0.
- If the number of heads is 7 or less, then accept H_0.

Does this test satisfy the level condition? What is the power of this test?

We may write the test in terms of a test function as follows:

$$\phi_4(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i = 9, 10, \\ 0.85 & \text{if } \sum_{i=1}^n x_i = 8 \\ 0 & \text{otherwise.} \end{cases}$$

Checking the level condition of ϕ_4: The probability of type I error is

$$\begin{aligned} P(\text{Reject } H_0 \mid H_0 \text{ is true}) &= P_{p=0.5}(\text{Reject } H_0) \\ &= P_{p=0.5}\left(\text{Reject } H_0 \mid \textstyle\sum_{i=1}^{10} X_i \in \{9,10\}\right) P_{p=0.5}\left(\textstyle\sum_{i=1}^{10} X_i \in \{9,10\}\right) \\ &\quad + P_{p=0.5}\left(\text{Reject } H_0 \mid \textstyle\sum_{i=1}^{10} X_i = 8\right) P_{p=0.5}\left(\textstyle\sum_{i=1}^{10} X_i = 8\right) \\ &\quad + P_{p=0.5}\left(\text{Reject } H_0 \mid \textstyle\sum_{i=1}^{10} X_i < 8\right) P_{p=0.5}\left(\textstyle\sum_{i=1}^{10} X_i < 8\right) \\ &= P_{p=0.5}\left(\textstyle\sum_{i=1}^{10} X_i \in \{9,10\}\right) + 0.85 \times P_{p=0.5}\left(\textstyle\sum_{i=1}^{10} X_i = 8\right). \end{aligned}$$

Thus, $P(\text{Type I error}) = 0.0481$, which also satisfies the level condition.

Next, consider the power of the test. A similar calculation would lead to

$$\text{Power} = 1 - P\left(\text{Type II error}\right) = P\left(\text{Reject } H_0 \mid H_1\right) = 0.4834,$$

which is much higher than the power of ϕ_2. Is ϕ_4 the MP test? No, because one may further adjust the test function for the case $\sum x_i = 8$ to get better power, at the cost of reduced probability of type I error. One may continue to make adjustments as long as the level condition is valid.

Definition 14. *(Randomized and Non-randomized Tests) A test function of the form $I_C(\mathbf{x})$ is called a non-randomized test. Any other test function (a function from the sample space S_x to $[0,1]$) corresponds to a randomized test.*

In the above example ϕ_1, ϕ_2 and ϕ_3 are non-randomized test, while ϕ_4 is a randomized test.

4 Neyman Pearson Lemma

The following theorem prescribes a method of obtaining the most powerful test for testing a simple versus simple hypothesis.

Theorem 1 (Neyman-Pearson Lemma)**.** *Consider the problem of testing simple vs simple hypotheses, $H_0 : \mathbf{X} \sim f_0(\mathbf{x})$ against $H_1 : \mathbf{X} \sim f_1(\mathbf{x})$, using a test ϕ that satisfies*

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } f_1(\mathbf{x}) > k f_0(\mathbf{x}), \\ 0 & \text{if } f_1(\mathbf{x}) < k f_0(\mathbf{x}), \end{cases} \tag{2}$$

for some $k \geq 0$, and

$$\alpha = E_{H_0}\left[\phi(\mathbf{X})\right]. \tag{3}$$

Any test that satisfies (2) and (3) is a MP level α test.

Example 3. Let X be a random variable with p.m.f. under H_0 and H_1 are as follows. Find an MP level $\alpha = 0.025$ test using Neyman-Pearson Lemma.

x	1	2	3	4	5	6
$f_0(x)$	0.01	0.01	0.01	0.01	0.01	0.95
$f_1(x)$	0.05	0.04	0.03	0.02	0.01	0.85
$\lambda(x)$	5	4	3	2	1	17/19

Example 4. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Test $H_0 : \mu = \mu_0, \sigma = \sigma_0$ against $H_1 : \mu = \mu_1, \sigma = \sigma_0$ $(\mu_1 > \mu_0)$.

Example 5. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Test $H_0 : \mu = \mu_0, \sigma = \sigma_0$ against $H_1 : \mu = \mu_0, \sigma = \sigma_1$ $(\sigma_1 < \sigma_0)$.

Example 6. Let $X_1, \ldots, X_n$ be a random sample from Poisson(λ). Test $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda = \lambda_1$ $(\lambda_1 < \lambda_0)$.

Remark 3. *In the above three examples, observe that the ratio $\lambda(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta_1)/f_{\mathbf{X}}(\mathbf{x}|\theta_0)$ turns out to be a function of a statistic $T(\mathbf{X})$, so that $\lambda(\mathbf{x}) > k$ is equivalent to $T(\mathbf{x}) > k_0$ or $T(\mathbf{x}) < k_0$ depending on the alternative hypothesis and the exact relation of T and λ.*

In particular, in Example 4, $\lambda(\mathbf{x}) > k$ is equivalent to $T(\mathbf{x}) = \sum_i x_i > k_0$; in Example 5, $\lambda(\mathbf{x}) > k$ is equivalent to $T(\mathbf{x}) = \sum_i (x_i - \mu_0)^2 < k_0$; and in Example 6, $\lambda(\mathbf{x}) > k$ is equivalent to $T(\mathbf{x}) = \sum_i x_i < k_0$.

Therefore, the MP test in Examples 4 can alternatively be expressed as

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } T(\mathbf{x}) > k_0, \\ 0 & \text{if } T(\mathbf{x}) < k_0, \end{cases}$$

for some $k_0 \geq 0$ where $T(\mathbf{x}) = \sum_i x_i$, and

$$\alpha = E_{H_0}\left[\phi(\mathbf{X})\right].$$

Thus, the test is essentially based on an appropriate statistic, $T(\mathbf{X})$, called test statistic.

Definition 15. *(Test Statistics) As Remark 3 suggests, typically, a hypothesis test is specified in terms of the values of a statistic $T(\mathbf{X})$. This statistic is called test statistic.*

5 Generalization of MP Test to Composite Hypotheses

In Example 4, observe that

I. If we replace μ_1 by any choice of μ, say μ_1' such that $\mu_1' > \mu_0$, then we would obtain the sample MP test. Therefore, the test obtained in Example 4 remains MP for any $\mu > \mu_0$ (To see this, just follow the procedure of obtaining the MP test for any two choices of $\mu_1 > \mu_0$. You will arrive at the same MP test).

   Thus, it is a UMP test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.

II. Suppose we want to generalize the testing problem in Example 4 as $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$. As discussed above for any μ_0', μ_1' such that $\mu_0' < \mu_1'$, the MP test must be of the form

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \ T(\mathbf{x}) > k_0, \\ 0 & \text{if } \ T(\mathbf{x}) < k_0, \end{cases}$$

   for some $k_0 \geq 0$ where $T(\mathbf{x}) = \sum_i x_i$.

   The particular choice of k_0 is obtained from the second condition (3), which requires *size* $= \alpha$, where α is the chosen level of significance. Now, for $H_0 : \mu \leq \mu_0$,

$$\textit{size} = \sup_{\mu:\mu\leq\mu_0} \beta_\phi(\mu) = \sup_{\mu:\mu\leq\mu_0} P_\mu(T(\mathbf{X}) > k_0).$$

   Observe that $\beta_\phi(\mu) = P_\mu(T(\mathbf{X}) > k_0)$ is an increasing function of μ. Therefore, *size* $= \beta_\phi(\mu_0)$, and consequently, the equation *size* $= \alpha$ leads to the same choice of k_0 as in Example 4.

   Thus, the same test which is UMP for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, is now UMP for testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$.

III. Following similar steps as above, verify that the MP test obtained in Example 5 is UMP for testing $H_0 : \sigma \geq \sigma_0$ against $H_1 : \sigma > \sigma_1$, and the MP test obtained in Example 6 is UMP for testing $H_0 : \lambda \geq \lambda_0$ against $H_1 : \lambda < \lambda_0$.

IV. Naturally, one would ask when such generalizations are possible. Such generalizations are possible for a special family of distributions, called families with Monotone likelihood ratio (MLR).

   Discussion on MLR is beyond the scope of this course. For sake of completeness the definition of an MLR family is provided below.

   Definition 16. *(Monotone Likelihood Ratio, MLR) A family of distributions with pdf/pmf* $\{f_{\mathbf{X}}(;\theta), \theta \in \Theta \subseteq \mathbb{R}\}$ *is said to have a monotone likelihood ratio (MLR) in a statistic* $T(\mathbf{x})$ *if for any* $\theta_1 < \theta_2$ *the ratio* $\lambda(\mathbf{x}, \theta_1, \theta_2) = f_{\mathbf{X}}(\mathbf{x}; \theta_1)/f_{\mathbf{X}}(\mathbf{x}; \theta_2)$ *is a monotone (non-increasing or non-decreasing) function of* $T(\mathbf{x})$ *for the set of values* $\mathbf{x}$ *for which at least one of* $f_{\mathbf{X}}(\mathbf{x}; \theta_1)$ *and* $f_{\mathbf{X}}(\mathbf{x}; \theta_2)$ *is positive.*

6 Unbiased Test

Let $X_1, \cdots, X_n$ be a random sample from $N(\mu, 1)$ distribution. Consider the problem of testing $H_0 : \mu = \mu_0$ against $\mu \neq \mu_0$ at level $\alpha > 0$.

Consider three possible tests for the above testing problem.

A.

$$\phi_A(\mathbf{x}) = \begin{cases} 1 & \text{if } \ \sum_i x_i \geq k_0, \\ 0 & \text{if } \ \sum_i x_i < k_0, \end{cases}$$

with k_0 such that $P_{\mu_0}(\sum_i X_i \geq k_0) = \alpha$.

B.

$$\phi_B(\mathbf{x}) = \begin{cases} 1 & \text{if } \ \sum_i x_i \leq k_0', \\ 0 & \text{if } \ \sum_i x_i > k_0', \end{cases}$$

with k_0 such that $P_{\mu_0}(\sum_i X_i \leq k_0') = \alpha$.

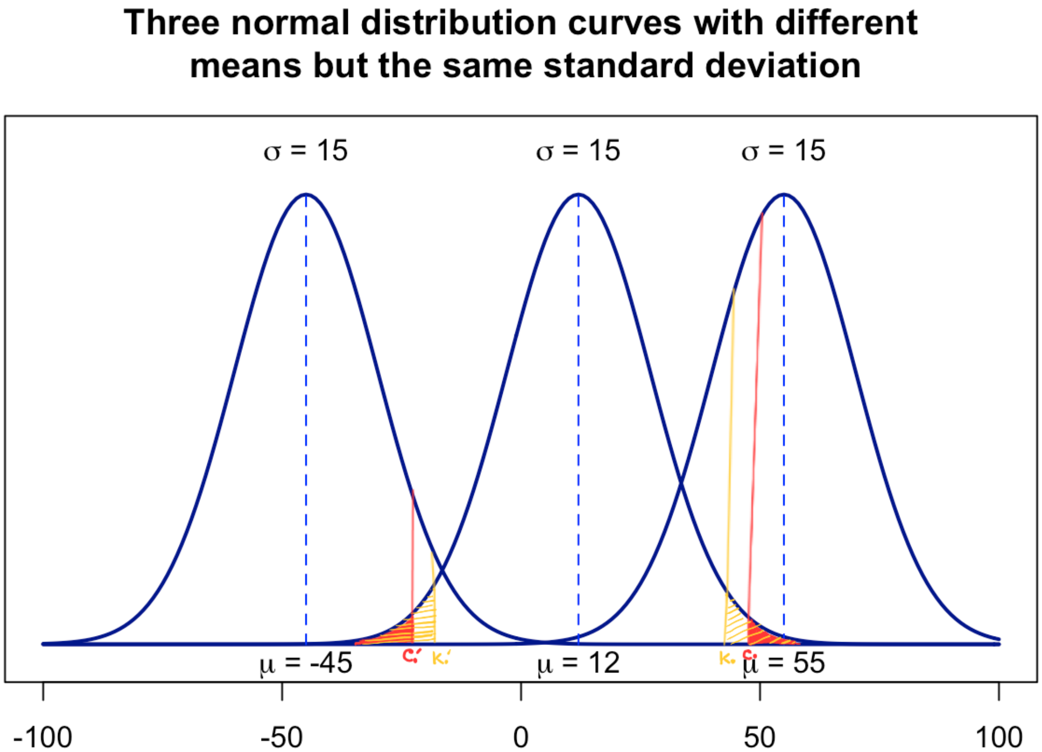

Figure 1: Rejection regions of ϕ_A, ϕ_B and ϕ_C.

C.

$$\phi_C(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_i x_i \leq c_0', \quad \text{or} \quad \sum_i x_i \geq c_0, \\ 0 & \text{otherwise,} \end{cases}$$

with k_0 such that $P_{\mu_0}(\sum_i X_i \leq c_0') + P_{\mu_0}(\sum_i X_i \geq c_0) = \alpha$.

Note that ϕ_A is the UMP test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, and ϕ_B is the UMP test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$. Therefore, ϕ_A achieves highest power among all the level-α tests for $\mu > \mu_0$, but shows bad power property in the region $\mu < \mu_0$. Similarly, ϕ_B achieves highest power among all the level-α tests for $\mu < \mu_0$, but shows bad power property in the region $\mu > \mu_0$. Thus, there does not exist any UMP test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

Now, consider ϕ_C. Although it seems reasonable to reject H_0 for high as well as low values of $\sum_i X_i$ (unlike in ϕ_A and ϕ_B), the power of ϕ_C is lower than ϕ_A for any $\mu > \mu_0$, and is lower than ϕ_B for any $\mu < \mu_0$ (see Figure 2). [To see this, recall that all the 3 tests need to meet the *size* $= \alpha$ condition. While ϕ_A and ϕ_B rejects for one-sided values of $\sum_i X_i$, ϕ_C considers a two-sided rejection region. Thus, $k_0' < c_0'$ and $k_0 > c_0$. Hence, ϕ_C must have strictly lower power than ϕ_A (or, ϕ_B) for $\mu > \mu_0$ (or, $\mu < \mu_0$). See Figure 1]

Note that for ϕ_A is not even acceptable tests as for any $\mu < \mu_0$,

$$P(\text{Reject } H_0 \text{ by } \phi_A \mid \mu_0) > P(\text{Reject } H_0 \text{ by } \phi_A \mid \mu),$$

i.e., probability of *'rejecting H_0 when H_0 is true'* is higher than probability of *'rejecting H_0 when H_0 is false'*, which is not desirable. Similarly, ϕ_B is not also acceptable.

For any *reasonable* test the probability of type I error must be lower than power. Thus, a *reasonable* test must satisfy the *size*<*power* property. This property is called *unbiasness of a test*.

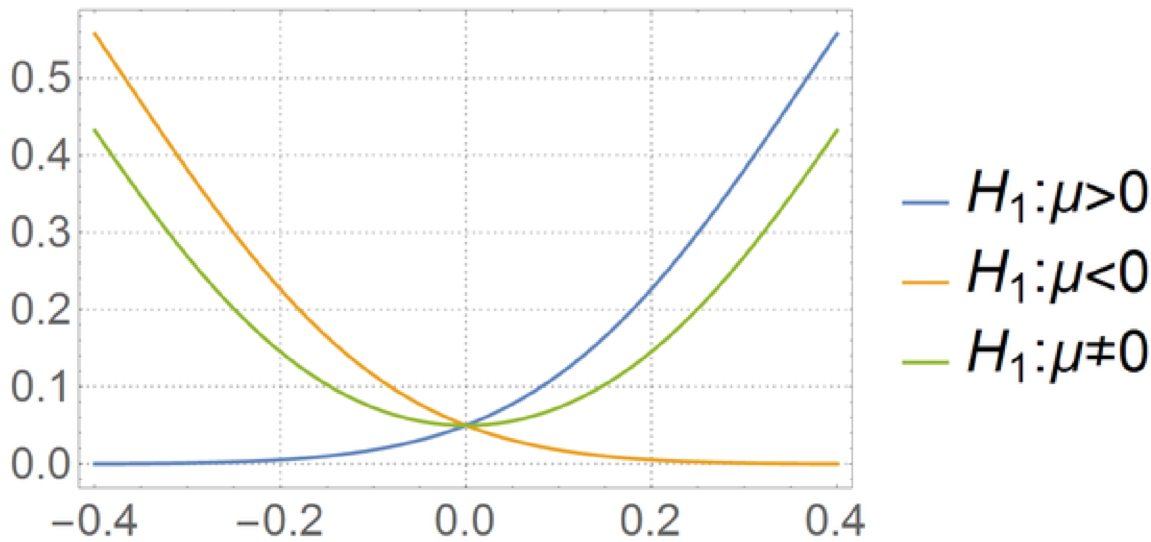

Figure 2: Power functions of ϕ_A, ϕ_B and ϕ_C.

Definition 17. *(Unbiased test) Consider the problem of testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. A test ϕ is called an unbiased test for testing H_0 against H_1 if for any $\theta' \in \Theta_0$ and $\theta'' \in \Theta_1$,*

$$\beta_\phi(\theta') \leq \beta_\phi(\theta''),$$

equivalently, if

$$\textit{size} = \sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \inf_{\theta \in \Theta_1} \beta_\phi(\theta) = \textit{power}.$$

Observe that ϕ_A and ϕ_B are not unbiased, while ϕ_C is unbiased. In fact, one can show that ϕ_C is the UMP test in the class of unbiased level α test, that is the UMPU (uniformly most powerful unbiased) level-α test.

7 p-value

Consider the problem of testing $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$, based on a random sample $X_1, \cdots, X_n$ from `normal`$(\mu, 1)$ distribution. From Section 5 it is evident that the test

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_i x_i \geq 1.67\sqrt{n}, \\ 0 & \text{otherwise}, \end{cases}$$

is a UMP test.

Let $n = 100$. Now, for a particular realization of $X_1, \cdots, X_n$, say $\{x_1, \ldots, x_n\}$, suppose you observe $\sum_i x_i = 17$, and for another set of realizations, say $\{y_1, \ldots, y_n\}$, you observe $\sum_i y_i = 25$. Obviously, the test ϕ would reject H_0 for both the realizations. However, one can clearly see that the second realization indicates a more extreme situation (compared to the first). Similarly, one would accept H_0 for both the realizations $\sum_i x_i = 15$ and $\sum_i y_i = 1$. However, the confidence level with which H_0 is accepted in the second case is much higher than that in the first case.

Thus, there should be a way to quantify the amount of confidence associated with the decision, and p-value serves the purpose.

Informally, p-value is the probability (under H_0) of observing a data which is at least as extreme as the realization in hand. What does 'extreme' mean? It depends on the experimental design, i.e., the alternative hypothesis. For the current example, extreme means observing higher values of $\sum_i X_i$. For the realizations $\sum_i x_i = 17$ and $\sum_i y_i = 25$, the associated p-values are

$$p(17) = P_{H_0}(\sum_i X_i \geq 17) = 0.0446, \quad \text{and} \quad p(25) = P_{H_0}(\sum_i X_i \geq 25) = 0.0062.$$

This clearly shows that the second realization is much less likely to observe under H_0 than the first realization, and hence indicates a more extreme situation.

Remark 4. *Observe that that if the level α is set to $\alpha \geq 0.0446$, then one would reject H_0 based on the realization $\sum_i x_i = 17$ at level α, and for an choice of $\alpha < 0.0446$ one would accept H_0. In that sense, $p(17)$ is the* smallest level at which H_0 is rejected by the test ϕ, given $\sum_i x_i = 17$.

Definition 18. *(p-value) The p-value is the probability of observing data at least as extreme as the realized value, assuming the null hypothesis to be true. The smaller the p-value, the more extreme the outcome and the stronger the evidence against H_0. It is defined as*

$$p = p(\mathbf{x}) := \inf\{\alpha : \quad \textit{given } \mathbf{x},\ H_0 \textit{ is rejected against } H_1 \textit{ at level } \alpha\}.$$

Remark 5. *For a fixed level of significance, say $\alpha = 0.05$, if the p-value corresponding to a realization, say $\mathbf{x}$, is less than or equal to α, i.e., $p(\mathbf{x}) \leq \alpha$, then H_0 is rejected based on that observation.*

Finally, consider the problem of testing $H_0 : \mu \geq 0$ against $H_1 : \mu < 0$, based on a random sample $X_1, \cdots, X_n$ from `normal`$(\mu, 1)$ distribution. From Section 5 it is evident that the test

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \ \sum_i x_i \leq -1.67\sqrt{n}, \\ 0 & \text{otherwise,} \end{cases}$$

is a UMP test.

Now, based on $n = 100$ observations, suppose we observe $\sum_i x_i = -1.05$. What will be the p-value associated with this realization? Observe that, in this case in view of the alternative hypothesis lower values of $\sum_i x_i$ indicates extremity. Thus, in this case

$$p(-1.05) = P_{H_0}(\sum_i X_i \leq -1.05) = 0.4582.$$

Consequently, H_0 is accepted at level $\alpha = 0.05$ based on this observation.