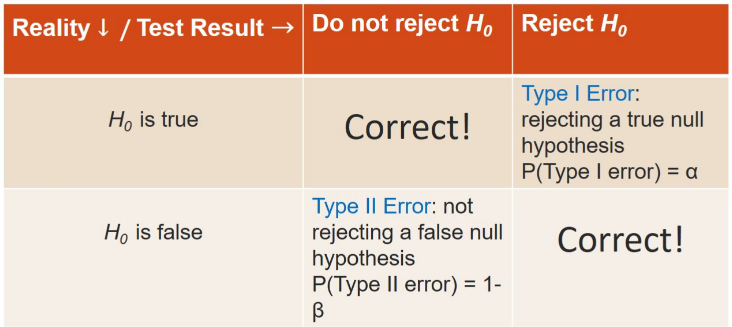
6 Elements of hypothesis testing

[CB8.3,BD4.2-4.3]

6.1 Introduction

Hypothesis testing begins with an assumption, called a hypothesis, that we make about a population parameter.
The bottom line in hypothesis testing is when we ask whether a population like we think this one is would be likely to produce a sample like the one we are looking at.

In hypothesis testing, we must state the assumed or hypothesized distribution of the population before we begin sampling.
The assumption we wish to test is called the **null hypothesis** **(H)**. In parametric inference this will be in terms of a finite number of parameters.
Whenever we reject the hypothesis, the conclusion we draw is called **alternative hypothesis** **(K)**.
Note: Null hypotheses are either rejected, or else there is insufficient evidence to reject them. (i.e., we don't accept null hypotheses.)



Reality ↓ / Test Result →	Do not reject H_0	Reject H_0
H_0 is true	Correct!	Type I Error: rejecting a true null hypothesis P(Type I error) = α
H_0 is false	Type II Error: not rejecting a false null hypothesis P(Type II error) = 1-β	Correct!

Type I Error: rejecting a true null hypothesis.
Max value of P(Type I error)=α Significance level of the test
Type II Error: not rejecting a false null hypothesis.
P(Type II error) = 1-β=1-Power of the test

Definition 1 *The* **power** *of a test ϕ against the alternative θ is the probability of rejecting H when θ is true and is denoted by $\beta(\theta, \phi)$.*

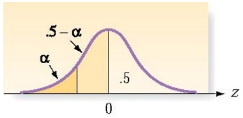
Example 1: The null hypothesis is that the battery has an average life of 300 days, with the alternative hypothesis being that the battery life is more than 300 days. You are the quality control engineer for the battery manufacturer.
(a) Would you rather make a Type I or a Type II error?

(b) Based on your answer to part (a), should you use a high or a low significance level?

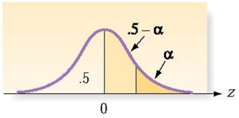
Testing procedure: Fix H, K, α. Obtain sample. Calculate a test statistic based on the sample. If the test statistic has a low probability (fixed at α) when H is true, then H is rejected. Otherwise H is not rejected.

One and two sided alternative hypotheses. The null hypothesis is usually stated as an equality. The alternative hypothesis can be either an equality or an inequality.

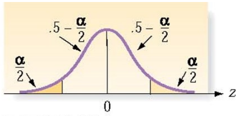
One and two tailed tests. Depending on the type of the alternative the rejection region can be right-tailed, left-tailed or two-tailed.



a. Form of H_a: <



b. Form of H_a: >



c. Form of H_a: ≠

Example 2: A drug will be released in the market only if it's efficacy is more than 30%. What are the null and alternative hypotheses? Which is appropriate, a one-tailed or a two-tailed test?

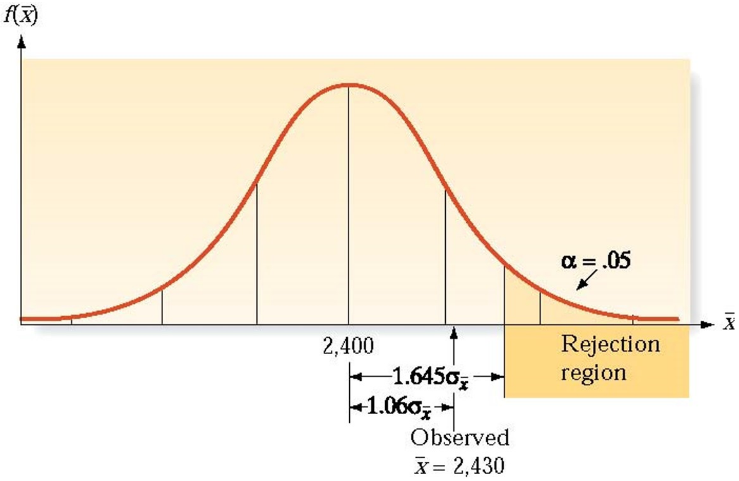


Figure 1: Example 3

Example 3: In figure 6.1, we have a sample of size n from a normal population with unknown mean μ. H: $\mu = 2400$, K: $\mu > 2400$. At $\alpha = 0.05$, we reject H for values of $\bar{X} > 2400 + 1.645\sigma_{\bar{X}}$. In this case, the observed value of $\bar{X}$ is 2430, which is $1.06\sigma_{\bar{X}}$. Hence we fail to reject H.

Parametric set-up: Data: $X \in \mathcal{X}, X \sim P_\theta, H : \theta \in \Theta_0, K : \theta \in \Theta_1$. If Θ_0 consists of a single point, we call it a **simple null**, otherwise, a **composite**

null. Similarly, with K.

In example 2, $H : \theta = \theta_0, K = \theta > \theta_0$, then we have a simple null vs a composite alternative. If we allow $H : \theta \leq \theta_0$, then we have a composite null. In most cases (Monotone Likelihood Ratio situations), the solutions to both problems are the same. In this example with $H : \theta = \theta_0$, it is reasonable to reject H if $X=$ number of cases in which the drug is effective in n trials, is "much" larger than what would be expected by chance if H is true and the value of θ is θ_0.
Thus, we reject H if X exceeds or equals some integer, say k.

Critical region or rejection region denotes the values of the test statistic X for which we reject H. In this example the critical region C is $\{X : X > k\}$. This is equivalent to specifying a test function $\phi : \mathcal{X} \to \{0, 1\}$, where 1 denotes rejection.

Thus $P(typeIerror) = P_{\theta=\theta_0}(X \geq k)$
and $P(typeIIerror) = P_\theta(X < k), \theta < \theta_0$.
k is called the critical value.
The power is obtained as $\sum_{i=k}^{n} \binom{n}{k} \theta^i (1-\theta)^{n-i}$. A plot of the function for $n = 10, \theta_0 = 0.3, k = 6$ is in figure 4.1.1 below (taken from BD).

Note that in this example the power at $\theta = \theta_1 > 0.3$ is the probability that the level 0.05 test will detect an improvement of the recovery rate from 0.3 to θ_1. When θ_1 is 0.5, a 67% improvement, this probability is only .3770. What is needed to improve on this situation is a larger sample size n. One of the most important uses of power is in the selection of sample sizes to achieve reasonable chances of detecting interesting alternatives

Also, the power function is increasing. It follows that the level and size of the test are unchanged if instead of $\Theta_0 = \{\theta_0\}$ we used $\Theta_0 = [0, \theta_0]$. That is,

$$\alpha(k) = sup\{P_\theta(X \geq k) : \theta < \theta_0\} = P_{\theta_0}(X \geq k).$$

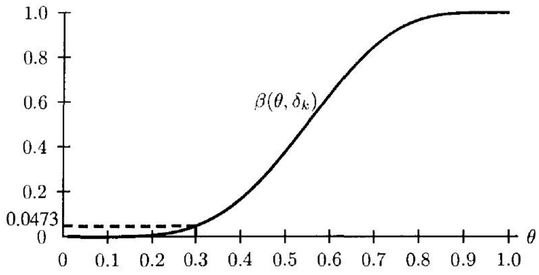


Figure 4.1.1. Power function of the level 0.05 one-sided test δ_k of $H : \theta = 0.3$ versus $K : \theta > 0.3$ for the $\mathcal{B}(10, \theta)$ family of distributions. The power is plotted as a function of θ, $k = 6$ and the size is 0.0473.

6.2 Neyman-Pearson Theory

We start with the problem of testing a simple hypothesis $H : \theta = \theta_0$ versus a simple alternative $K : \theta = \theta_1$. Simple **likelihood ratio statistic** is defined by

$$L(x, \theta_0, \theta_1) = \frac{p(x, \theta_0)}{p(x, \theta_1)}$$

where $p(x, \theta)$ is the density (pdf) or frequency (pmf) function of the random vector X.
We call ϕ_k a likelihood ratio or Neyman-Pearson (NP) test (function) if for some $0 < k < \infty$ we can write the test function ϕ_k as

$$\phi_k(x) = \begin{cases} 1 & \text{if} \quad L(x, \theta_0, \theta_1) < k \\ 0 & \text{if} \quad L(x, \theta_0, \theta_1) > k \end{cases} \tag{1}$$

with $\phi_k(x)$ any value in (0,1) if equality occurs.

Because we want results valid for all possible test sizes α in $[0, 1]$, we consider randomized tests ϕ, which are tests that may take values in $(0, 1)$. If $0 < \phi(x) < 1$ for the observation vector x, the interpretation is that we toss a coin with probability of heads $\phi(x)$ and reject H iff the coin shows heads.

Theorem 1 *(Neyman-Pearson Lemma)*

1. *If $\alpha > 0$ and ϕ_k is a size α likelihood ratio test, then ϕ_k is MP in the class of level α tests.*
2. *For each $0 < \alpha < 1$ there exists an MP size α likelihood ratio test provided that randomization is permitted, $0 < \phi(x) < l$,for some x.*
3. *If ϕ is an MP level α test, then it must be a level α likelihood ratio test; that is, there exists k such that $P_\theta(\phi_k(x) \neq \phi(x), L(X, \theta_0, \theta_1 \neq k) = 0$ for $\theta = \theta_0$ and $\theta = \theta_1$*

It follows from the Neyman-Pearson lemma that an MP test has power at least as large as its level; that is,

Corollary 1 *If ϕ is an MP level α test, then $E_{\theta_1}\phi(x) > \alpha$ with equality iff $p(x, \theta_0) = p(x, \theta_1) \forall x$.*

In example 2, suppose the alternative is $\theta_1 = 0.5$. As before $\theta_0 = 0.3$. Thus we have a simple null vs simple alternative situation where the model is $X \sim Bin(n, \theta)$. The likelihood ration is

$$L(X, \theta_0, \theta_1) = \frac{\binom{n}{X}(0.3)^X(0.7)^{n-X}}{\binom{n}{X}(0.5)^X(0.5)^{n-X}} = (3/7)^X(7/5)^n$$

$L < k$ is equivalent to $X > (n \log(7/5) - k)/\log(7/3) = k_1$ (say). So the test that rejects the null hypothesis for large values of X is MP in the class of level α tests by NP lemma.

In order to determine the test explicitly given $\alpha = 0.05$ and n=10, we find the highest k_1 such that $P(X > k_1) < \alpha$.
From R, 1-pbinom(4,10,0.3)=0.1502683=$P(X > 4)$
and 1-pbinom(5,10,0.3)=0.04734899=$P(X > 5)$.
So, k_1=5 and a=$(\alpha - P(X > k))/P(X = 5)$=0.02575813. So the test function is

$$\phi(x) = \begin{cases} 1 & \text{if} X > 5 \\ 0.02575813 & \text{if} X = 5 \\ 0 & \text{if} X < 5 \end{cases} \tag{2}$$

That is, reject H if $X > 5$ and with probability a if $X = 5$.
For $\theta = 0.5$, the power is

$$\begin{aligned} \beta(\theta, \phi) &= P(X < 5) + aP(X = 5) \\ &= 1 - pbinom(5, 10, 0.5) + a * dbinom(5, 10, 0.5) \\ &= 0.383292 \end{aligned}$$

6.3 UMP tests and MLR families

Now we want to consider the case of composite null $H : \theta \in \Theta_0$ vs composite alternative $K : \theta \in \Theta_1$

Definition 2 *A level α test $\phi*$ is uniformly most powerful (UMP) for H vs K if*

$$\beta(\theta, \phi*) \geq \beta(\theta, \phi) \forall \theta \in \Theta_1$$

for any other level α test ϕ.

Definition 3 *The family of models $\{P_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}$ is said to be a monotone likelihood ratio (MLR) family if for $\theta_1 < \theta_2$ the distributions P_{θ_0} and P_{θ_2} are distinct and the ratio $p(x, \theta_2)/p(x, \theta_1)$ is an increasing function of a statistic $T(x)$.*

In example 2, for $\theta_1 < \theta_2$,

$$p(x, \theta_2)/p(x, \theta_1) = \frac{\theta_2^X (1 - \theta_2)^{n-X}}{\theta_1^X (1 - \theta_1)^{n-X}} = \left(\frac{\theta_2(1 - \theta_1)}{\theta_1(1 - \theta_2)} \right)^X \left(\frac{1 - \theta_2}{1 - \theta_1} \right)^n$$

is increasing in X and the model is MLR in $T(X) = X$.

Result: Consider the one-parameter exponential family model

$$p(x, \theta) = h(x) exp\{\eta(\theta) T(x) - B(\theta)\}.$$

If $\eta(\theta)$ is strictly increasing in $\theta \in \Theta$, then this family is MLR. Example 2 is of this form with $T(x) = x$ and $\eta(\theta) = \log(\theta/(1 - \theta))$.

Define the Neyman-Pearson (NP) test function

$$\delta_t(x) = \begin{cases} 1 & \text{if} \quad T(x) > t \\ 0 & \text{if} \quad T(x) < t \end{cases} \tag{3}$$

with $\delta_t(x)$ any value in (0,1) if $T(x) = t$. Consider the problem of testing $H : \theta = \theta_0$ versus $K : \theta = \theta_1$ with $\theta_0 < \theta_1$. If $\{P_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}$, is an MLR family in $T(x)$, then $L(x, \theta_0, \theta_1) = g(T(x))$ for some increasing function g. Thus, δ_t equals the likelihood ratio test $\phi_{g(t)}$ and is MP. Because δ_t does not depend on θ_1 it is UMP at level $\alpha := E_{\theta_0}\delta_t(X)$ for testing $H : \theta = \theta_0$ versus $K : \theta > \theta_0$.

Theorem 2 *Suppose $\{P_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}$ is an MLR family in $T(x)$. Then*

1. *For each $t \in (0, \infty)$, the power function $\beta(\theta) = E_\theta \delta_t(X)$ is increasing in θ.*
2. *If $E_{\theta_0}\delta_t(X) = \alpha > 0$, then δ_t is UMP level α for testing $H : \theta \leq \theta_0$ versus $K : \theta > \theta_1$ for $\theta_1 > \theta_0$.*

6.4 Unbiased tests

Definition 4 *A test ϕ is unbiased if $\beta_\phi(\theta) \geq \alpha$ for all $\theta \in \Theta_1$ and $\beta_\phi(\theta) \leq \alpha$ for all $\theta \in \Theta_0$.*

Remark: If ϕ is a UMP level α test, then ϕ is unbiased. Proof: compare ϕ with the trivial test function $\tilde{\phi} \equiv alpha$.

Definition 5 *A uniformly most powerful unbiased level α test is a test $\tilde{\phi}$ for which $E_\theta \tilde{\phi} \geq E_\theta \phi$ for all $\theta \in \Theta_1$ and for all unbiased level α tests ϕ.*

That is, $\tilde{\phi}$ is uniformly (for all $\theta \in \Theta_1$) most powerful ($E_\theta \tilde{\phi} \geq E_\theta \phi$) among all unbiased tests ϕ.

Theorem 3 *Consider testing $H : \theta = \theta_0$ versus $K : \theta \neq \theta_0$ in a one parameter exponential family with natural parameter θ and natural sufficient statistic T. The test ϕ with $E_{\theta_0}\phi(T) = \alpha$ given by*

$$\phi(T(x)) = \begin{cases} 1 & \text{if} \quad T(x) > c_2 \quad \text{or} \quad T(x) < c_1 \\ 0 & \text{if} \quad c_1 < T(x) < c_2 \end{cases} \tag{4}$$

with $\phi(T(x))$ any value in $\gamma \in (0, 1)$ if $T(x) = c_i, i = 1, 2$. is UMPU for H versus K.