Lecture 9: Comparison of hypothesis tests

9.1 Synopsis of Hypothesis Testing

Remark 9.1 Suppose our data come from $f_X(x \mid \theta)$, $\theta \in \Omega$ and we want to test the following hypothesis:

$$H_0 : \theta \in \omega_0 \quad \text{vs.} \quad H_1 : \theta \in \omega_1 \quad (\omega_0 \cap \omega_1 = \emptyset, \omega_0 \cup \omega_1 = \Omega).$$

Our *decision* to either reject H_0 or fail to reject H_0 is based upon the space of the sample $\mathcal{D}$ where $\mathcal{D} := \text{space}\{(X_1, \ldots, X_n)\}$. In particular, we reject H_0 if $(X_1, \ldots, X_n) \in C$ where $C \subseteq \mathcal{D}$ is called the *rejection/critical region.* We say C is of *size* α if

$$\max_{\theta \in \omega_0} P_\theta[(X_1, \ldots, X_n) \in C] = \alpha.$$

Next, the *significance level* is the maximum allowable probability of committing a *Type I error* and is defined *a priori.* A test has significance level α if the size is less than or equal to α. Then, among the critical regions that are size α, we want to then minimize the probability of committing a *Type II Error* (i.e., accept H_0 when H_1 is true). Minimizing Type II Error is equivalent to maximizing *power*, where the *power function* of a critical region C is

$$\gamma_C(\theta) = P_\theta[(X_1, \ldots, X_n) \in C], \quad \theta \in \omega_1.$$

Example 9.1 Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where $\sigma^2 > 0$ is known. We want to test the following hypothesis

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

at a significance level $\alpha = 0.05$. Here are all of the possible ways to perform this test given an observation $(x_1, \ldots, x_n)$.

(i) Define my rejection region C to be

$$C := \{(X_1, \ldots, X_n) \in \mathcal{D} \mid \bar{X} \leq c\}$$

where c is chosen to guarantee C is size α. In particular, we find c as follows:

$$P_{\mu_0}(\bar{X} \leq c) = P_{\mu_0}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha \implies \frac{c - \mu_0}{\sigma/\sqrt{n}} = z_{0.05} \implies c = \mu_0 + \frac{\sigma}{\sqrt{n}} \cdot z_{0.05}$$

where $z_{0.05}$ is the $\alpha = 0.05$ quantile of the standard normal distribution. Now, when we observe $(x_1, \ldots, x_n)$, we reject H_0 if $\bar{x} \leq \mu_0 + \frac{\sigma}{\sqrt{n}} \cdot z_{0.05}$.

(ii) To make a decision on the hypotheses, we can compute an *observed significance level* (i.e., *p-value*) for our observations $(x_1, \ldots, x_n)$. We still consider the same rejection rule as in part (i), that is, "reject H_0 in favor of H_1 if $\bar{X} \leq c$." However, instead of determining the value of c which determines the size of our test, we can ask "Is $\bar{x}$ sufficiently small to reject H_0 in favor of H_1?" In particular, we compute the probability [under H_0] that our test statistics is *as or more* extreme than our observation,

$$\text{p-value} = P_{\mu_0}(\bar{X} \leq \bar{x}).$$

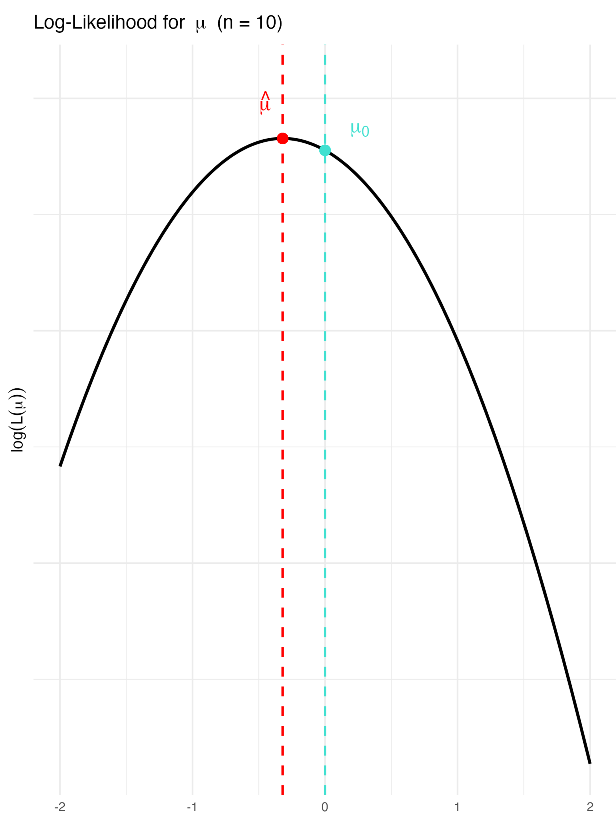
This *observed significance level* (i.e., p-value) is then compared to a *nominal* α level and we reject if p-value $\leq \alpha$.

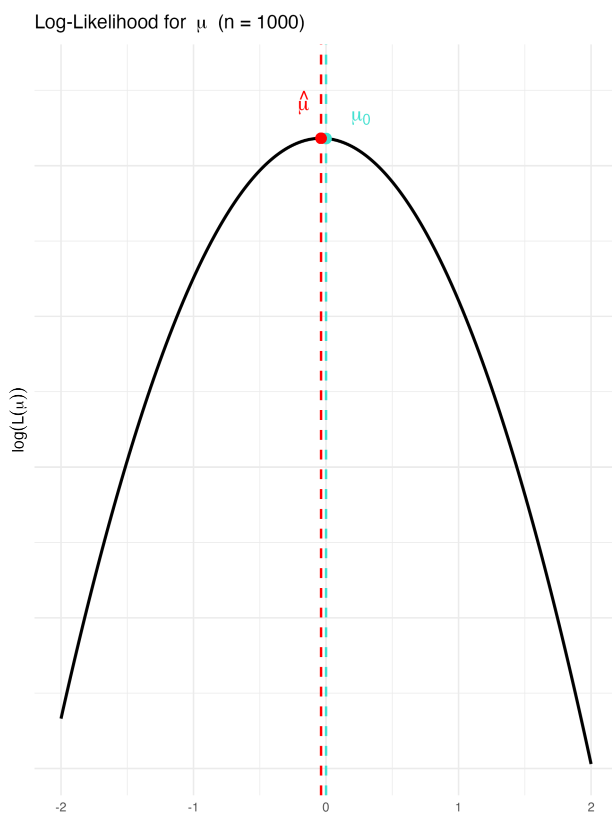
(iii) Instead of focusing on $\bar{X}$ as the test statistic, instead we can compare the likelihood under H_0 to the likelihood across the entire parameter space and conduct a *likelihood ratio test* (LRT). Specifically, we compute

$$\Lambda = \frac{L(\mu_0 \mid x_1, \ldots, x_n)}{L(\hat{\mu}^{MLE} \mid x_1, \ldots, x_n)}$$

and the decision rule becomes "reject H_0 in favor of H_1 if $\Lambda \leq c$" where $c \in [0, 1]$ is chosen such that $\alpha = P_{\mu_0}(\Lambda \leq c)$.

Remark 9.2 Further insight into why $\Lambda \leq 1$ for the LRT. Consider the following simulation study where $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. Assume the true value, $\mu_0 = 0$. We will compare the log-likelihood function $l(\mu \mid x_1, \ldots, x_n)$ for both $n = 10$ and $n = 1000$.





9.2 Comparing Testing Methods

Let's consider the case of a *simple hypothesis* test for now, where both the null and alternative hypotheses are composed of a single probability distribution. This is the easiest way to understand the criteria for comparing testing methods. Consider a case where we test the following hypothesis using a random sample $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, where $\theta \in \Omega = \{\theta_0, \theta_1\}$

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

Let C denote the rejection region and $\boldsymbol{X} = (X_1, \ldots, X_n)$. The error probabilities associated with the test method are defined in two ways.

- Type I Error Probability $= P_{\theta_0}(\boldsymbol{X} \in C)$
- Type II Error Probability $= P_{\theta_1}(\boldsymbol{X} \notin C) = 1 - P_{\theta_1}(\boldsymbol{X} \in C)$

Definition 9.1 Let C denote a subset of the sample space $\mathcal{D}$. We say C is a **best critical region** of size α for testing simple hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ if

(a) $P_{\theta_0}[(X_1, \ldots, X_n) \in C] = \alpha$
(b) If $P_{\theta_0}[(X_1, \ldots, X_n) \in A] = \alpha$ $(A \subseteq \mathcal{D})$, then $P_{\theta_1}[(X_1, \ldots, X_n) \in C] \geq P_{\theta_1}[(X_1, \ldots, X_n) \in A]$.

Example 9.2 Suppose $X \sim \text{Binomial}(n = 5, p = \theta)$ and $f_X(x \mid \theta)$ denotes the pmf. We are interesting in testing $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta = \frac{3}{4}$. The following table provides the pmf values under each hypothesis as well as their ratio. We will use X as the test statistic (i.e., the rejection region is of the form $X \in C$) at a significance level of $\alpha = \frac{1}{32}$. We want to find a best critical region of size $\alpha = \frac{1}{32}$.

x	0	1	2	3	4	5
$f(x;1/2)$	1/32	5/32	10/32	10/32	5/32	1/32
$f(x;3/4)$	1/1024	15/1024	90/1024	270/1024	405/1024	243/1024
$\frac{f(x;1/2)}{f(x;3/4)}$	32/1	32/3	32/9	32/27	32/81	32/243

Solution. First, let's find subsets of the sample space that are level $\alpha = \frac{1}{32}$. There exists two possible choices, $A_1 = \{x \mid x = 0\}$ and $A_2 = \{x \mid x = 5\}$ because

$$P(X \in A_1 \mid H_0) = P_{\{\theta=1/2\}}(X \in A_1) = \frac{1}{32}, \quad \text{and} \quad P(X \in A_2 \mid H_0) = P_{\{\theta=1/2\}}(X \in A_2) = \frac{1}{32}.$$

A_1 and A_2 are all possible subsets of $\mathcal{D}$ with level α. In order to determine which A_1 or A_2 is the *best critical region*, we need to notice

$$\frac{243}{1024} = P_{\{\theta=3/4\}}(X \in A_2) > P_{\{\theta=3/4\}}(X \in A_1) = \frac{1}{1024},$$

which implies that the best critical region is A_2. ■

Remark 9.3 Why did we include the ratio of densities in Example 9.2? We saw that the best critical region was found by including the point (or points) at which $f(x \mid \theta = \frac{1}{2})$ is *small* in comparison to $f(x \mid \theta = \frac{3}{4})$. We see that for our best critical region A_2, the ratio $\frac{f(x;1/2)}{f(x;3/4)}$ is minimized at $x = 5$ (i.e., A_2). Thus, looking at this ratio provides a precise tool to find a best critical region C for any given size α.

For example, now take $\alpha = \frac{6}{32}$. [Left as a homework problem to find the best critical region.]

Remark 9.4 A hypothesis of the form $H : \theta = \theta_0$ completely specifies the underlying distribution and is called a **simple** hypothesis. A hypothesis of the form, for example, $H : \theta < \theta_0$ is a **composite** hypothesis because they are *composed* of many simple hypotheses (and hence do not completely specify the distribution).

Theorem 9.1 ***(Neyman-Pearson Theorem):*** *Let $X_1, \dots, X_n \overset{iid}{\sim} f_X(x \mid \theta)$. The likelihood of $X_1, \dots, X_n$ is*

$$L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} f_X(x_i \mid \theta), \quad \boldsymbol{x} = (x_1, \dots, x_n)^T.$$

Let θ_0 and θ_1 denote fixed values of θ where $\theta \in \Omega = \{\theta_0, \theta_1\}$, and let $k > 0$. Let C be a subset of the sample space $\mathcal{D}$, where C has the following properties:

(a) $\frac{L(\theta_0|\boldsymbol{x})}{L(\theta_1|\boldsymbol{x})} \le k$ *for each* $\boldsymbol{x} \in C$.
(b) $\frac{L(\theta_0|\boldsymbol{x})}{L(\theta_1|\boldsymbol{x})} \ge k$ *for each* $\boldsymbol{x} \in C^c = \mathcal{D} \setminus C$.
(c) $\alpha = P_{H_0}[\boldsymbol{X} \in C]$.

Then C is a best critical region of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$

Example 9.3 Let $X_1, \dots, X_n \overset{iid}{\sim} f_X(x \mid \theta)$ where

$$f_X(x \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\theta)^2}{2}\right), \quad x \in (-\infty, \infty).$$

Determine the best critical region of size α for testing $H_0 : \theta = 0$ versus $H_1 : \theta = 1$.
Solution. Following the steps of Theorem 9.1, we first look at the ratio $\frac{L(\theta_0|\boldsymbol{x})}{L(\theta_1|\boldsymbol{x})}$.

$$\begin{aligned}\frac{L(\theta_0 \mid \boldsymbol{x})}{L(\theta_1 \mid \boldsymbol{x})} &= \frac{L(0 \mid \boldsymbol{x})}{L(1 \mid \boldsymbol{x})} \\ &= \frac{\left(1/\sqrt{2\pi}\right)^n \exp\left(-\sum_{i=1}^n x_i^2/2\right)}{\left(1/\sqrt{2\pi}\right)^n \exp\left(-\sum_{i=1}^n (x_i-1)^2/2\right)} \\ &= \exp\left(\frac{n}{2} - \sum_{i=1}^{n} x_i\right)\end{aligned}$$

Let $k > 0$. Recall that the goal is to determine C (i.e., what $(x_1, \ldots, x_n)$ satisfy the three conditions in Theorem 9.1, and then these $(x_1, \ldots, x_n)$ make up our C). Let's figure out a way to characterize the set of $(x_1, \ldots, x_n)$ that satisfy condition (a) of Theorem 9.1.

$$\exp\left(\frac{n}{2} - \sum_{i=1}^{n} x_i\right) \leq k$$
$$\frac{n}{2} - \sum_{i=1}^{n} x_i \leq \log(k)$$
$$\frac{n}{2} - \log(k) \leq \sum_{i=1}^{n} x_i$$

In this case, a best critical region is the set $C = \left\{(x_1, \ldots, x_n) \mid \sum_{i=1}^{n} x_i \geq c, \text{ where } c = \frac{n}{2} - \log(k)\right\}$. Then the value of c can be chosen to achieve the desired size α as in part (c) of Theorem 9.1. Notice that it would be equivalent to write the critical region as $C = \left\{(x_1, \ldots, x_n) \mid \bar{x}_i \geq c^*, \text{ where } c^* = \frac{1}{2} - \frac{\log(k)}{n}\right\}$. Therefore, given observations $(x_1^*, \ldots, x_n^*)$, H_0 would be rejected at significance level α if $\bar{x}^* \geq c^*$.
■

Remark 9.5 The simple hypotheses above only depend on *one* parameter, but the procedures above work for any finite number of parameters. The only requirement is that the hypotheses H_0 and H_1 be *simple*, in that they completely specify the distributions. Furthermore, the simple hypotheses H_0 and H_1 do **not** need to be hypotheses about the parameters of a distribution, **nor** do the random variables $X_1, \ldots, X_n$ need to be independent. That is, if H_0 is the simple hypothesis that the joint pdf or pmf is $g(x_1, \ldots, x_n)$ and H_1 is the simple hypothesis that the joint pdf or pmf is $h(x_1, \ldots, x_n)$, then C is a best critical region of size α if, for $k > 0$,

(a) $\frac{g(x_1,\ldots,x_n)}{h(x_1,\ldots,x_n)} \leq k$, for $(x_1, \ldots, x_n) \in C$.

(b) $\frac{g(x_1,\ldots,x_n)}{h(x_1,\ldots,x_n)} \geq k$, for $(x_1, \ldots, x_n) \in C^c = \mathcal{D} \setminus C$.

(c) $\alpha = P_{H_0}[(X_1, \ldots, X_n) \in C]$

Definition 9.2 The critical region C is a **uniformly most powerful (UMP) critical region** of size α for testing a *simple* hypothesis H_0 again a *composite* hypothesis H_1 if the set C is a best critical region of size α for testing H_0 against *each simple* hypothesis in H_1. A test defined by this critical region C is called a **UMP test**, with significance level α, for testing the simple hypothesis H_0 against the alternative composite hypothesis H_1.

Example 9.4 Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$ with θ unknown. Find the UMP test of the simple hypothesis $H_0 : \theta = \theta_0$ versus the composite hypothesis $H_1 : \theta > \theta_0$.
Solution. In order to follow the steps of Theorem 9.1, let $\theta_1 > \theta_0$ be parameter value in the

"alternative" parameter space. Then, for $k > 0$, we can determine a best rejection region C by

$$
\begin{aligned}
\frac{L(\theta_0 \mid \boldsymbol{x})}{L(\theta_1 \mid \boldsymbol{x})} &\leq k \\
\frac{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_0)^2\right)}{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right)} &\leq k \\
\exp\left\{-(\theta_1 - \theta_0)\sum_{i=1}^{n} x_i + \frac{n}{2}\left[\theta_1^2 - \theta_0^2\right]\right\} &\leq k \\
(\theta_1 - \theta_0)\sum_{i=1}^{n} x_i &\geq \frac{n}{2}\left[\theta_1^2 - \theta_0^2\right] - \log k \\
\sum_{i=1}^{n} x_i &\geq \frac{n}{2}[\theta_1 + \theta_0] - \frac{\log k}{\theta_1 - \theta_0}.
\end{aligned}
$$

Therefore,

$$
C := \left\{(x_1, \ldots, x_n) \mid \sum_{i=1}^{n} x_i \geq \frac{n}{2}[\theta_1 + \theta_0] - \frac{\log k}{\theta_1 - \theta_0}\right\}
$$

and we can find a value for k such that C is of size α and therefore is a best critical region of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ for $\theta_1 > \theta_0$. Since C is well defined for all $\theta_1 > \theta_0$ (i.e., for all simple hypotheses composing the composite alternative hypothesis), C is a UMP critical region of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. ■