5 Hypothesis Testing

5.1 Hypothesis

In hypothesis testing, we would like to assess a hypothesis, a statement about the population.

- "the drug has no effect"
- "$p = 0.5$"

But our sample is random. Hypothesis testing helps us to answer

Can we distrust the hypothesis from a random sample?

The hypotheses are subsets of $\mathcal{P}$ by dividing Θ into two parts Θ_0 and Θ_1 such that

$$\Theta = \Theta_0 \cup \Theta_1 \quad \text{and} \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

1. The null hypothesis, denoted by H_0, means that $\mathcal{P}_0 = \{\mathrm{P}_\theta : \theta \in \Theta_0\}$. We write $H_0 : \theta \in \Theta_0$.
2. The alternative hypothesis, denoted by H_1 or H_A, means that $\mathcal{P}_1 = \{\mathrm{P}_\theta : \theta \in \Theta_1\}$. We write $H_1 : \theta \in \Theta_1$.

We can work with

- Simple hypothesis, e.g., $H : \theta = \theta_0$, just one point.
- Composite hypothesis, e.g., $H : \theta \leq 0$ or $H : \theta^2 = 1$ or $H : \theta_1 - \theta_2 = 0$, not simple hypothesis.
- One-sided hypothesis, e.g., $H : \theta \geq 0$.
- Two-sided hypothesis, e.g., $H : \theta \neq 0$ or $H : \theta \leq 1$ or $\theta \geq 3$.

We need to quantify the evidence against H_0 using a test statistic T. If there is sufficient evidence against H_0, then we reject the H_0. Otherwise, we cannot reject H_0. We **DO NOT** say we accept H_0.

5.2 Power Function

Definition 1. A nonrandomized test ϕ is a statistic from the sample space $\mathcal{X}$ to $\{0, 1\}$:

$$\phi(x) = \begin{cases} 1, & \text{if } x \in C_1, \text{ reject } H_0, \\ 0, & \text{if } x \in C_0, \text{ do not reject } H_0, \end{cases}$$

where $x = (x_1, ..., x_n)$, $\mathcal{X} = C_1 \cup C_0$ and $C_1 \cap C_0 = \emptyset$. Here C_1 is the rejection region (aka critical region), and C_0 is the acceptance region.

Due to randomness in our data, we may commit two types of errors.

		Decision	
		Cannot reject H_0	Reject H_0
Truth	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

Definition 2 (Power function)**.** For the test problem $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. The power function of the test ϕ is defined by

$$\pi(\theta) = \mathrm{E}[\phi(X) \mid \theta], \quad \theta \in \Theta,$$

where $X = (X_1, ..., X_n)$.

In the nonrandomized test,

$$\pi(\theta) = \mathrm{E}[\phi(X) \mid \theta] = \mathrm{P}(\phi = 1 \mid \theta) = \mathrm{P}(X \in C_1 \mid \theta), \quad \theta \in \Theta,$$

the probability of rejecting H_0.

- For $\theta \in \Theta_0$, then $\pi(\theta)$ is the type I error probability. We want it to be low.
- For $\theta \in \Theta_1$, then $1 - \pi(\theta)$ is the type II error probability. We want it to be high.

The trade-off between two errors is that it is not possible to decrease both towards zero.

- In practice, it is common to restrict the probability of type I error to be low, e.g., no higher than a prespecified level, the significance level.
- We then seek tests that have low probability of type II error among the tests with well-controlled probability of type I error.

Example 1 (Normal-test)**.** Let $X_1, ..., X_n$ be a random sample from $N(\theta, \sigma^2)$ with σ^2 known. We would like to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Suppose that we reject H_0 if

$$\frac{|\bar{X} - \theta_0|}{\sigma/\sqrt{n}} > c, \text{ for some constant } c.$$

The power function is

$$\pi(\theta) = \mathrm{P}\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} < -c - \frac{\theta - \theta_0}{\sigma/\sqrt{n}} \text{ or } \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c - \frac{\theta - \theta_0}{\sigma/\sqrt{n}} \mid \theta\right),$$

where $\frac{\bar{X}-\theta}{\sigma/\sqrt{n}} \sim N(0,1)$. If we want the type I error probability to be a pre-specified value $\alpha \in (0,1)$, we need to find the value c such that

$$\alpha = \mathrm{P}\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} < -c \text{ or } \frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c \mid \theta_0\right).$$

Example 2. Let $X \sim \text{Binomial}(6, \theta)$. We want to test $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$. If X is too large, we should reject H_0. Under H_0, the probabilities are

X	0	1	2	3	4	5	6
$p(x; \theta = 1/2)$	0.02	0.09	0.23	0.31	0.23	0.09	0.02

If we want the type I error probability to be $\alpha = 0.02$, then $C_1 = \{x = 6\}$. But it is not possible to have a nonrandomized test where the type I error probability is 0.05.

Definition 3 (Randomized test)**.** A randomized test ϕ us a statistic from the sample space $\mathcal{X}$ to $[0,1]$:

$$\phi(x) = \begin{cases} 1, & \text{if } x \in C_1, \text{ reject } H_0, \\ \gamma, & \text{if } x \in C_=, \text{ reject } H_0 \text{ with probability } \gamma, \\ 0, & \text{if } x \in C_0, \text{ do not reject } H_0, \end{cases}$$

where $\mathcal{X} = C_1 \cup C_= \cup C_0$ and C_1, $C_=$, and C_0 are mutually disjoint. The power function of a randomized test is

$$\pi(\theta) = \mathrm{E}[\phi(X) \mid \theta] = \mathrm{P}(X \in C_1 \mid \theta) + \gamma \mathrm{P}(X \in C_= \mid \theta),$$

for $\theta \in \Theta$, still the probability of rejecting H_0.

Example 3. Let $X \sim \text{Binomial}(6, \theta)$. We want to test $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$. Solving

$$p(6; \theta = 1/2) + \gamma \times p(5; \theta = 1/2) = 0.05$$

produces the desired type I error probability if we let $\gamma = 1/3$.

5.3 Neyman-Pearson Test

A hypothesis is simple if it completely specifies the distribution. Consider simple hypotheses

$$H_0 : \mathrm{P}_0 \quad \text{versus} \quad H_1 : \mathrm{P}_1.$$

Definition 4 (UMP)**.** A test ϕ^* is called the uniformly most powerful test (UMP test) of size α if and only if $\mathrm{E}[\phi^*(X) \mid H_0] = \alpha$ and $\mathrm{E}[\phi^*(X) \mid H_1] \geq \mathrm{E}[\phi(X) \mid H_1]$ for all tests with $\mathrm{E}[\phi(X) \mid H_0] \leq \alpha$.

Definition 5 (Neyman-Pearson test)**.** Consider testing $H_0 : \mathrm{P}_0$ versus $H_1 : \mathrm{P}_1$. For $k \geq 0$ and $\gamma \in [0, 1]$, the Neyman-Pearson test is

$$\phi(x) = \begin{cases} 1, & \text{if } p_0(x) < kp_1(x), \\ \gamma, & \text{if } p_0(x) = kp_1(x), \\ 0, & \text{if } p_0(x) > kp_1(x), \end{cases}$$

where p_0 and p_1 are the probability functions related to P_0 and P_1, respectively.

Theorem 1 (Neyman-Pearson test as UMP)**.** *Consider testing $H_0 : \mathrm{P}_0$ versus $H_1 : \mathrm{P}_1$. For all $\alpha \in [0, 1]$, there exist $\gamma(\alpha)$ and $k(\alpha)$ such that the Neyman-Pearson test with $\gamma(\alpha)$ and $k(\alpha)$ has size α and is UMP of size α. Especially, for $\alpha > 0$,*

1. *if there exits a $k(\alpha)$ such that*

$$P_0(p_0(\boldsymbol{X}) < k(\alpha)p_1(\boldsymbol{X})) = \alpha,$$

   then $\gamma(\alpha) = 0$. Otherwise, $k(\alpha)$ satisfies

$$P_0(p_0(\boldsymbol{X}) < k(\alpha)p_1(\boldsymbol{X})) < \alpha < P_0(p_0(\boldsymbol{X}) \leq k(\alpha)p_1(\boldsymbol{X}))$$

   and $\gamma(\alpha)$ is determined by

$$P_0(p_0(\boldsymbol{X}) < k(\alpha)p_1(\boldsymbol{X})) + \gamma(\alpha)P_0(p_0(\boldsymbol{X}) = k(\alpha)p_1(\boldsymbol{X})) = \alpha.$$

2. *If ϕ is a UMP test of size α, then ϕ is the Neyman-Pearson test except on a zero measure set.*

3. *The power of the Neyman-Pearson test is larger than α.*

Proof. We only show the first result. In order for the test to have size α, we need to have

$$\alpha = \mathrm{E}[\phi(X) \mid H_0] = \mathrm{P}_0(p_0(X) < kp_1(X)) + \gamma\mathrm{P}_0(p_0(X) = kp_1(X)).$$

We need to show that this test has the highest power: for any test ψ with $\mathrm{E}[\psi(X) \mid H_0] \leq \alpha$, we have $\mathrm{E}[\phi(X) \mid H_1] \geq \mathrm{E}[\psi(X) \mid H_1]$.

1. If $\phi(x) - \psi(x) > 0$, then we must have $\phi(x) > 0$ (since $\phi = \{1, \gamma > 0, 0\}$ and $\psi \in [0, 1]$) and $p_0(\boldsymbol{x}) \leq kp_1(\boldsymbol{x})$.

2. If $\phi(x) - \psi(x) < 0$, then we must have $\phi(x) < 1$ (since $\phi = \{1, \gamma > 0, 0\}$ and $\psi \in [0, 1]$) and $p_0(\boldsymbol{x}) \geq kp_1(\boldsymbol{x})$.

In all cases, we have $[\phi(x) - \psi(x)][kp_1(x) - p_0(x)] \geq 0$. Hence,

$$\begin{aligned} \int [\phi(x) - \psi(x)][kp_1(x) - p_0(x)]\, d\mu(x) &\geq 0 \\ &\Downarrow \\ k\int [\phi(x) - \psi(x)]p_1(x)\, d\mu(x) &\geq \int [\phi(x) - \psi(x)]p_0(\boldsymbol{x})\, d\mu(x) \\ &\Downarrow \\ k\mathrm{E}[\phi(x) - \psi(x) \mid H_1] &\geq \mathrm{E}[\phi(x) \mid H_0] - \mathrm{E}[\psi(x) \mid H_0] \\ &\geq \alpha - \alpha = 0. \end{aligned}$$

$\square$

Example 4. Let $X \sim \text{Binomial}(6, \theta)$. We want to test $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$. Under H_0, the probabilities are

X	0	1	2	3	4	5	6
$\frac{p(x;\theta=1/2)}{p(x;\theta=3/4)}$	64.00	21.33	7.11	2.37	0.79	0.26	0.09
$p(x;\theta=1/2)$	0.02	0.09	0.23	0.31	0.23	0.09	0.02

- If we want a UMP test of size $\alpha = 0.02$, then $C_1 = \{x = 6\}$.
- If we want a UMP test of size $\alpha = 0.05$, we solve $p(6; \theta = 1/2) + \gamma \times p(5; \theta = 1/2) = 0.05$ and obtain $\gamma = 1/3$.

In Neyman-Pearson Lemma for testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, we reject H_0 if

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})} < k, \\ \gamma, & \text{if } \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})} = k, \\ 0, & \text{if } \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})} > k. \end{cases}$$

It is equivalent to compare the likelihood ratio

$$\frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})} \quad \text{or} \quad \frac{p_0(\boldsymbol{x})}{\max\{p_0(\boldsymbol{x}), p_1(\boldsymbol{x})\}}$$

with k. In statistics, a common name for a Radon-Nikodym density is likelihood ratio.

Definition 6. For $0 \le \alpha \le 1$, a test with power function $\pi(\theta)$ is called

- a size α test if $\sup_{\theta \in \Theta_0} \pi(\theta) = \alpha$,
- a level α test if $\sup_{\theta \in \Theta_0} \pi(\theta) \le \alpha$.

Typically, we specify a small $\alpha = 0.01$, 0.05 and 0.1 (small values), meaning that we do not reject H_0 unless the evidence is strong. For this reason, H_1 is also called the research hypothesis. We usually put the hypothesis that we would like to *verify* in H_1.

Example 5. Let $X_1, ..., X_n$ be a random sample from $N(\theta, \sigma^2)$, where σ^2 is known. We want to test $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$. We reject H_0 if

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c,$$

for some constant c. The power function is

$$\pi(\theta) = \mathrm{P}\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c - \frac{\theta - \theta_0}{\sigma/\sqrt{n}} \mid \theta\right) = \mathrm{P}\left(N(0,1) > c - \frac{\theta - \theta_0}{\sigma/\sqrt{n}} \mid \theta\right).$$

In order to have a size α test, we need to have

$$\alpha = \sup_{\theta \le \theta_0} \mathrm{P}\left(N(0,1) > c - \frac{\theta - \theta_0}{\sigma/\sqrt{n}} \mid \theta\right) = \mathrm{P}(N(0,1) > c \mid \theta).$$

Definition 7 (UMP Level α Test)**.** A test ϕ^* is a UMP level α test if and only if

$$\sup_{\theta \in \Theta_0} \mathrm{E}[\phi^*(X) \mid \theta] = \alpha,$$

and $\mathrm{E}[\phi^*(X) \mid \theta] \ge \mathrm{E}[\phi(X) \mid \theta]$ for all $\theta \in \Theta_1$ and for all tests with $\sup_{\theta \in \Theta_0} \mathrm{E}[\phi(X) \mid \theta] \le \alpha$.

5.4 Test and Confidence Interval

A random set is a $1-\alpha$ confidence region for a parameter θ if

$$\mathrm{P}\left\{\theta \in S\left(X\right) \mid \theta\right\} \quad \geq \quad 1-\alpha, \quad \forall \theta \in \Theta.$$

Theorem 2 (Duality between test and confidence set)**.** *Suppose that $A\left(X, \theta_0\right)$ is the acceptance region of a nonrandomized level α test with $H_0 : \theta = \theta_0$ such that*

$$P\left\{X \in A\left(X, \theta\right) \mid \theta\right\} \quad \geq \quad 1-\alpha, \quad \forall \theta \in \Theta.$$

For each x, define $C\left(x\right) = \left\{\theta;\ x \in A\left(X, \theta\right)\right\}$. Then, the random set $C\left(X\right)$ is a $1-\alpha$ confidence set. Conversely, let $C\left(X\right)$ be a $1-\alpha$ confidence set. For any $\theta_0 \in \Theta$, define $A\left(X, \theta_0\right) = \left\{\theta_0 : \theta_0 \in C\left(X\right)\right\}$. Then $A\left(X, \theta_0\right)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Example 6. Let X_1, ..., X_n be a random sample from $N\left(\theta, \sigma^2\right)$ with σ^2 known. The UMP level α test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$ rejects H_0 if $\bar{X} < \theta_0 - \sigma z_{1-\alpha}/\sqrt{n}$. The acceptance region is $\left\{x :\ \bar{X} \geq \theta_0 - \sigma z_{1-\alpha}/\sqrt{n}\right\}$. The confidence interval is $\{\theta :\ \bar{x} \geq \theta - \sigma z_{1-\alpha}/\sqrt{n}\}$.

5.5 Likelihood Ratio Test

Definition 8 (LRT)**.** The likelihood ratio test (LRT) statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ is

$$\lambda\left(x\right) \quad = \quad \frac{\sup\limits_{\theta \in \Theta_0} L\left(\theta \mid x\right)}{\sup\limits_{\theta \in \Theta} L\left(\theta \mid x\right)} = \frac{L\left(\hat{\theta}_0 \mid x\right)}{L\left(\hat{\theta} \mid x\right)},$$

where $\hat{\theta}_0$ is the MLE when $\theta \in \Theta_0$, and $\hat{\theta}$ is the MLE when $\theta \in \Theta$. An LRT rejects H_0 if $\lambda\left(x\right)$ is too small.

Example 7. Let X_1, X_2, ..., X_n be random sample from $N\left(\theta, 1\right)$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_1$. The likelihood is

$$L\left(\theta; x\right) \quad = \quad \left(2\pi\right)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(x_i - \theta\right)^2\right].$$

When $\theta \in \Theta_0$, $\hat{\theta}_0 = \theta_0$. When $\theta \in \Theta$, $\hat{\theta} = \bar{X}$. The LRT statistic is

$$\begin{aligned} \lambda\left(x\right) &= \frac{L\left(\hat{\theta}_0 \mid x\right)}{L\left(\hat{\theta} \mid x\right)} = \frac{\left(2\pi\right)^{-n/2}\exp\left[-\sum_{i=1}^{n}\left(x_i - \theta_0\right)^2/2\right]}{\left(2\pi\right)^{-n/2}\exp\left[-\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2/2\right]} \\ &= \exp\left[-n\left(\bar{x} - \theta_0\right)^2/2\right]. \end{aligned}$$

We reject H_0,if $\lambda\left(x\right) = \exp\left[-n\left(\bar{x} - \theta_0\right)^2/2\right] \leq c$, for some c.

Example 8. Let X_1, X_2, ..., X_n be i.i.d. from the exponential distribution with pdf

$$p\left(x \mid \theta\right) = \begin{cases} \exp\left\{-\left(x - \theta\right)\right\}, & x \geq \theta \\ 0, & x < \theta. \end{cases}$$

Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. The likelihood function is

$$L\left(\theta | x\right) \quad = \quad \exp\left[-\sum_{i=1}^{n} x_i + n\theta\right] I\left(\theta \leq x_{(1)}\right),$$

which is an increasing function of θ if $\theta \leq x_{(1)}$. Otherwise, $L(\theta|x) = 0$. So $\hat{\theta} = X_{(1)}$ and $\hat{\theta}_0 = \begin{cases} X_{(1)}, & X_{(1)} \leq \theta_0, \\ \theta_0, & X_{(1)} > \theta_0. \end{cases}$

The likelihood function under H_0 is

$$L\left(\hat{\theta}_0|x\right) = \begin{cases} \exp\left[-\sum_{i=1}^{n} x_i + nx_{(1)}\right], & x_{(1)} \leq \theta_0, \\ \exp\left[-\sum_{i=1}^{n} x_i + n\theta_0\right], & x_{(1)} > \theta_0. \end{cases}$$

Over the entire parameter space, $L\left(\hat{\theta}|\boldsymbol{x}\right)$ is $\exp\left[-\sum_{i=1}^{n} x_i + nx_{(1)}\right]$. The ratio is

$$\lambda(x) = \frac{L\left(\hat{\theta}_0 \mid x\right)}{L\left(\hat{\theta} \mid x\right)} = \begin{cases} 1 & x_{(1)} \leq \theta_0, \\ \exp\left[n\theta_0 - nx_{(1)}\right] & x_{(1)} > \theta_0. \end{cases}$$

We reject H_0 if $\lambda(x) \leq c$. $\lambda(x)$ is a non-increasing function of $x_{(1)}$. Hence, we reject H_0 if $x_{(1)}$ is too big, say $\left\{x : x_{(1)} \geq k\right\}$.

Theorem 3. *Suppose that X_1, ..., X_n are iid following P_θ, where $\theta \in \Theta$ and Θ is an open set in $\mathbb{R}^p$. Suppose that, ,*

1. *$\frac{\partial \log p(x|\theta)}{\partial \theta}$ is well defined,*
2. *the true value θ_0 is the solution to $E\left[\frac{\partial \log p(x|\theta)}{\partial \theta} \mid \theta\right] = 0$,*
3. *the MLE is the solution to $\frac{\partial \log L(\theta;x)}{\partial \theta} = 0$ and is a consistent estimator of θ_0,*
4. *$E\left[\left\|\frac{\partial \log p(x|\theta)}{\partial \theta}\right\|^2 \mid \theta_0\right] < \infty$, where $\|\|$ is the Euclidean norm,*
5. *$\frac{\partial^2 \log p(x|\theta)}{\partial \theta \partial \theta^T}$ is nonsingular and $E\left[\left\|\frac{\partial^2 \log p(x|\theta)}{\partial \theta \partial \theta^T}\right\| \mid \theta_0\right] < \infty$,*
6. *the model is identified, i.e., $p(x \mid \theta_1) = p(x \mid \theta_2)$ almost everywhere under μ implies $\theta_1 = \theta_2$,*
7. *we can interchange the order of integration and differentiation such that $Var\left[\frac{\partial \log p(x|\theta)}{\partial \theta} \mid \theta\right] = -E\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta \partial \theta^T} \mid \theta\right]$.*

Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. Suppose that the true value θ_0 satisfies H_0. Then,

$$-2 \log \lambda(x) \quad \xrightarrow{d} \quad \chi_v^2,$$

where the degrees of freedom v is the difference in the number of free parameters.