50.034 - Introduction to Probability and Statistics

Week 11 – Lecture 20

January–May Term, 2019

Lecture 20 Assigned Readings: 9.2, 9.3

Outline of Lecture

- Likelihood ratio statistic
- Likelihood ratio test
- Linear combinations of error probabilities
- Likelihood ratio
- Neyman–Pearson lemma
- Uniformly most powerful test

Recall: Null hypothesis and alternative hypothesis

Main idea of hypothesis testing: To decide whether a specific hypothesis H_0 is to be **rejected** or **not rejected**, where the decision depends on the observed values obtained.

- This specific hypothesis H_0 is called the null hypothesis.
- The name "null hypothesis" comes about because we want to **nullify** the null hypothesis.

For hypothesis testing to make sense, we need a statistical model consisting of observable R.V.'s $X_1, \ldots, X_n$ with some unknown parameter θ. As part of this statistical model, we have to specify the parameter space Ω of θ.

- A null hypothesis H_0 is of the form "$\theta \in \Omega_0$", where $\Omega_0 \subseteq \Omega$ is some subset.
  - We can decide what hypothesis we want to test, so we can decide what we want Ω_0 to be.
- Let Ω_1 be the complement set of Ω_0 in Ω. The hypothesis H_1 defined by "$\theta \in \Omega_1$" is called the alternative hypothesis.

Recall: Steps for Hypothesis Testing

Model set-up: Let $X_1, \ldots, X_n$ be observable R.V.'s with unknown parameter θ. Let Ω be the parameter space of θ.

- Goal: Perform hypothesis testing on the parameter θ.

1. Specify some **null hypothesis** $H_0 : \theta \in \Omega_0$.
   - $\Omega_0 \subseteq \Omega$ is a subset chosen based on your specific application.
   - You wish to test whether the "true" value of θ is not in Ω_0.
2. Specify some **test statistic** $T = T(X_1, \ldots, X_n)$.
   - Your final decision will depend on the observed value of T.
3. Specify some **rejection region** $R \subseteq \mathbb{R}$.
   - This represents the region for where to reject H_0.
   - Note: R can be different from the complement of Ω_0.
4. Collect experimental evidence
   - Get observed values $X_1 = x_1, \ldots, X_n = x_n$.
5. Final decision: To reject or not to reject?
   - "Reject H_0" if $T(x_1, \ldots, x_n) \in R$.
   - "Do not reject H_0" if $T(x_1, \ldots, x_n) \notin R$.

The entire test procedure is collectively called a hypothesis test.

Functions of estimators as test statistic

Important Note: In a hypothesis test, you are free to choose your test statistic T and the corresponding rejection region R.

- T is a statistic of the observable R.V.'s $X_1, \ldots, X_n$ of your statistical model.
  - **Recall:** A statistic is a function of observable R.V.'s.
- You do not necessarily have to use functions of the the sample mean as your statistic T.
  - We learned that estimators are statistics. Thus, we could use functions of estimators as test statistics in hypothesis testing.

One very useful test statistic for hypothesis testing is called the **likelihood ratio statistic**, which uses the notion of the likelihood function, and is closely related to maximum likelihood estimators.

- If θ is a parameter of the observable R.V.'s $X_1, \ldots, X_n$, then the likelihood function of θ is defined using the exact same expression for the joint condition pmf/pdf (either $p_n(\mathbf{x}|\theta)$ or $f_n(\mathbf{x}|\theta)$), but treated as a function only in terms of θ.

Likelihood ratio statistic

Let θ be the parameter of some observable R.V.'s $X_1, \ldots, X_n$, and let Ω be the parameter space of θ.

- **Recall:** A maximum likelihood estimator of θ is a statistic $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ that maps each possible vector of observed values $\mathbf{x} = (x_1, \ldots, x_n)$ for $(X_1, \ldots, X_n)$ to some value $\theta_0 \in \Omega$ that maximizes the likelihood function of θ.
- In other words, if we substitute $\theta = \hat{\theta}(\mathbf{x})$ in the likelihood function (either $p_n(\mathbf{x}|\theta)$ or $f_n(\mathbf{x}|\theta)$), then we get

$$p_n(\mathbf{x}|\hat{\theta}(\mathbf{x})) = \sup_{\theta \in \Omega} p_n(\mathbf{x}|\theta) \qquad \text{or} \qquad f_n(\mathbf{x}|\hat{\theta}(\mathbf{x})) = \sup_{\theta \in \Omega} f_n(\mathbf{x}|\theta).$$

Definition: Given a subset $\Omega_0 \subseteq \Omega$, the likelihood ratio statistic associated to Ω_0 is the statistic $\Lambda = \Lambda(X_1, \ldots, X_n)$ defined by

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Omega_0} p_n(\mathbf{x}|\theta)}{\sup_{\theta \in \Omega} p_n(\mathbf{x}|\theta)} \qquad \text{or} \qquad \Lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Omega_0} f_n(\mathbf{x}|\theta)}{\sup_{\theta \in \Omega} f_n(\mathbf{x}|\theta)}$$

for each possible vector of observed values $\mathbf{x} = (x_1, \ldots, x_n)$.

Example 1

Coin Toss Experiment: Toss a coin 10 times.

- Let $X_1, \ldots, X_{10}$ be iid Bernoulli R.V.'s with unknown parameter θ.
  - $X_i = 1$ if the i-th coin toss is heads; $X_i = 0$ otherwise.
  - Let $[0, 1]$ be the parameter space of θ.

Consider the **null hypothesis** $H_0 : p = 0.3$.

- Given any vector $\mathbf{x} = (x_1, \ldots, x_n)$ of observed values, the likelihood function of θ is

$$p_{10}(\mathbf{x}|\theta) = \theta^{(x_1+\cdots+x_{10})}(1-\theta)^{10-(x_1+\cdots+x_{10})}.$$

- Thus the **likelihood ratio statistic** $\Lambda = \Lambda(X_1, \ldots, X_{10})$ associated to $\Omega_0 = \{0.3\}$ is given by

$$\Lambda(\mathbf{x}) = \frac{0.3^{(x_1+\cdots+x_{10})}0.7^{10-(x_1+\cdots+x_{10})}}{\sup_{\theta\in[0,1]} \theta^{(x_1+\cdots+x_{10})}(1-\theta)^{10-(x_1+\cdots+x_{10})}}$$

for each possible vector $\mathbf{x} = (x_1, \ldots, x_n)$.

Example 1 (continued)

For convenience, let $y = x_1 + \cdots + x_n$, and let $Y = X_1 + \cdots + X_{10}$.

- Note: y is a given integer satisfying $0 \leq y \leq 10$.

Using some calculus, we check that $\theta^y(1-\theta)^{10-y}$ is maximized at $\theta = \frac{y}{10}$ (as we let θ vary over the interval $[0,1]$).

Hence, the denominator in $\Lambda(\mathbf{x})$ is

$$\sup_{\theta \in [0,1]} \theta^y(1-\theta)^{10-y} = \frac{y^y(10-y)^{10-y}}{10^{10}},$$

and we conclude that the **likelihood ratio statistic** is

$$\Lambda(X_1, \ldots, X_{10}) = \left(\frac{3}{Y}\right)^Y \left(\frac{7}{10-Y}\right)^{(10-Y)}.$$

Likelihood ratio test

Let $\mathcal{H}$ be a hypothesis test with the null hypothesis $H_0 : \theta \in \Omega_0$, where θ is a parameter with parameter space Ω, and $\Omega_0 \subseteq \Omega$ is some given subset.

Definition If $\mathcal{H}$ has the **likelihood ratio statistic** associated to Ω_0 as its test statistic, then we say that $\mathcal{H}$ is a likelihood ratio test.

- Typically, the corresponding rejection region is an interval of the form $(-\infty, k]$ for some real number k.
- Generally, k is chosen so that $\mathcal{H}$ has some desired significance level α_0.

Interpretation: Given a likelihood ratio test, the null hypothesis $H_0 : \theta \in \Omega_0$ is rejected if the maximum value of the likelihood function over Ω_0 is "small" as compared to the maximum value of the likelihood function over the entire parameter space Ω.

- The precise meaning of "small" is determined by the rejection region chosen.

Recall: Power function and significance level

Let $\mathcal{H}$ be a hypothesis test with the null hypothesis $H_0 : \theta \in \Omega_0$, where θ is a parameter with parameter space Ω.
Suppose T is the test statistic, and let R be the rejection region.

- The power function of $\mathcal{H}$ is a function $\pi : \Omega \to \mathbb{R}$ defined by

$$\pi(\omega) = \Pr(T \in R | \theta = \omega)$$

  for every possible value $\omega \in \Omega$.
  - **Interpretation:** $\pi(\omega)$ is the probability that we will **reject** the null hypothesis H_0, given that the "true" value of θ equals ω.
  - Ideally, we wish that $\pi(\omega)$ is "very small" for all $\omega \in \Omega_0$, and $\pi(\omega)$ is "not too large" for all $\omega \in \Omega_1 = \Omega \backslash \Omega_0$.

Definition: We say that $\mathcal{H}$ a level α_0 test, or equivalently, that $\mathcal{H}$ has a significance level of α_0, if $\pi(\omega) \leq \alpha_0$ for all $\omega \in \Omega_0$. The smallest possible significance level for $\mathcal{H}$ is called the size of $\mathcal{H}$.

- **Interpretation:** "$\mathcal{H}$ is a level α_0 test" is exactly the same as "the probability that a type I error occurs for $\mathcal{H}$ is at most α_0."

Example 2

(Continuation of Example 1)
Consider a random sample $\{X_1, \ldots, X_{10}\}$ of observable Bernoulli R.V.'s with parameter θ, where the parameter space of θ is $[0, 1]$. Let $\mathcal{H}$ be a **likelihood ratio test** with null hypothesis $H_0 : \theta = 0.3$.

Suppose the rejection region of $\mathcal{H}$ is $(-\infty, c]$, where c is some constant to be determined.

Determine a value of c that maximizes the power of $\mathcal{H}$ at significance level 0.05.

Example 2 - Solution

From Example 1, we know that the likelihood ratio statistic Λ is

$$\Lambda(X_1, \ldots, X_{10}) = \left(\frac{3}{Y}\right)^{Y} \left(\frac{7}{10 - Y}\right)^{(10-Y)},$$

where $Y = X_1 + \cdots + X_{10}$. Define another function

$$\Lambda_Y(y) = \left(\tfrac{3}{y}\right)^{y} \left(\tfrac{7}{10-y}\right)^{(10-y)},$$

and note that $\Lambda(x_1, \ldots, x_{10}) = \Lambda_Y(x_1 + \cdots + x_{10})$ for all possible vectors $(x_1, \ldots, x_{10})$ of observed values for $(X_1, \ldots, X_{10})$.

If $H_0 : \theta = 0.3$ is true, then the pmf of Y (given $\theta = 0.3$) is

$$p_Y(y) = \binom{10}{y} 0.3^y 0.7^{(10-y)}.$$

Hence we can compute the following values:

y	0	1	2	3	4	5	6	$\cdots$
$\Lambda_Y(y)$	0.028	0.312	0.773	1.000	0.797	0.418	0.147	$\cdots$
$p_Y(y)$	0.028	0.121	0.233	0.267	0.200	0.103	0.037	$\cdots$

Example 2 - Solution (continued)

Here are all possible values for $y = 0, 1, \ldots, 10$:

y	0	1	2	3	4	5	6	$\cdots$
$\Lambda_Y(y)$	0.028	0.312	0.773	1.000	0.797	0.418	0.147	$\cdots$
$p_Y(y)$	0.028	0.121	0.233	0.267	0.200	0.103	0.037	$\cdots$

y	$\cdots$	7	8	9	10
$\Lambda_Y(y)$	$\cdots$	0.034	0.005	3×10^{-4}	6×10^{-6}
$p_Y(y)$	$\cdots$	0.009	0.001	1×10^{-4}	6×10^{-6}

Given any possible vector $\mathbf{x} = (x_1, \ldots, x_n)$ of observed values for $(X_1, \ldots, X_{10})$, let $y = x_1 + \cdots + x_{10}$. It follows from definition that we reject the null hypothesis H_0 if $\Lambda(\mathbf{x}) = \Lambda_Y(y) \leq c$.

We also want $\mathcal{H}$ to have a significance level of 0.05.

- This means that $\Pr(\Lambda \leq c) \leq 0.05$.
- For example, if $c = 0.005$, then $\Lambda(\mathbf{x}) \leq c = 0.005$ exactly for those vectors $\mathbf{x}$ whose corresponding y-value equals 8, 9 or 10, since $\Lambda_Y(10) < \Lambda_Y(9) < \Lambda_Y(8) = 0.005$.
  - Remember, $\Lambda(x_1, \ldots, x_{10}) = \Lambda_Y(x_1 + \cdots + x_{10})$.

Example 2 - Solution (continued)

y	0	1	2	3	4	5	6	$\cdots$
$\Lambda_Y(y)$	0.028	0.312	0.773	1.000	0.797	0.418	0.147	$\cdots$
$p_Y(y)$	0.028	0.121	0.233	0.267	0.200	0.103	0.037	$\cdots$

y	$\cdots$	7	8	9	10
$\Lambda_Y(y)$	$\cdots$	0.034	0.005	3×10^{-4}	6×10^{-6}
$p_Y(y)$	$\cdots$	0.009	0.001	1×10^{-4}	6×10^{-6}

In this example where $c = 0.005$, setting the rejection region to be $R = (-\infty, 0.005]$ would mean that the significance level of $\mathcal{H}$ is

$$\begin{aligned}\Pr(\Lambda \leq 0.005 | \theta = 0.3) &= \Pr(Y = 10 \text{ or } 9 \text{ or } 8) \\ &= p_Y(10) + p_Y(9) + p_Y(8) \\ &= 6 \times 10^{-6} + 1 \times 10^{-4} + 0.001 \\ &\approx 0.001106 < 0.05\end{aligned}$$

As we increase c, the power of $\mathcal{H}$ increases. We could still increase the value of c and still get a significance level ≤ 0.05.

- We want to maximize the power of $\mathcal{H}$ while still keeping to the constraint that the significance level must be ≤ 0.05.

Example 2 - Solution (continued)

y	0	1	2	3	4	5	6	$\cdots$
$\Lambda_Y(y)$	0.028	0.312	0.773	1.000	0.797	0.418	0.147	$\cdots$
$p_Y(y)$	0.028	0.121	0.233	0.267	0.200	0.103	0.037	$\cdots$

y	$\cdots$	7	8	9	10
$\Lambda_Y(y)$	$\cdots$	0.034	0.005	3×10^{-4}	6×10^{-6}
$p_Y(y)$	$\cdots$	0.009	0.001	1×10^{-4}	6×10^{-6}

We check that

$$\begin{array}{cccccccc} \Lambda_Y(10) < & \Lambda_Y(9) < & \Lambda_Y(8) < & \Lambda_Y(0) < & \Lambda_Y(7) < & \Lambda_Y(6) < & \cdots \\ 6 \times 10^{-6} < & 3 \times 10^{-4} < & 0.005 < & 0.028 < & 0.034 < & 0.147 < & \cdots \end{array}$$

$$\begin{aligned} \Pr(\Lambda \leq 0.034 | \theta = 0.3) &= p_Y(10) + p_Y(9) + p_Y(8) + p_Y(0) + p_Y(7) \\ &\approx 0.038106 < 0.05; \\ \Pr(\Lambda \leq 0.147 | \theta = 0.3) &= p_Y(10) + p_Y(9) + p_Y(8) + p_Y(0) + p_Y(7) + p_Y(6) \\ &\approx 0.075106 > 0.05. \end{aligned}$$

Thus, we can set c to be any value satisfying $0.034 \leq c < 0.147$ to maximize the power of $\mathcal{H}$ at significance level 0.05.

Hypothesis testing beyond a single parameter

Suppose $X_1, \ldots, X_n$ are observable R.V.'s (not necessarily iid), whose joint distribution is unknown, but you have narrowed down the possibilities to just two possible joint distributions $\mathcal{D}_1$ and $\mathcal{D}_2$. How do you test which of $\mathcal{D}_1, \mathcal{D}_2$ is the actual distribution?

- e.g. $\mathcal{D}_1$ could be a distribution where $X_1, \ldots, X_n$ are iid, and $\mathcal{D}_2$ could be a distribution where $X_1, \ldots, X_n$ are dependent.
  - If you could test between $\mathcal{D}_1$ and $\mathcal{D}_2$, then you could check if $\{X_1, \ldots, X_n\}$ is a random sample or not.

So far, we have looked at examples of hypothesis tests, where in the null hypothesis $H_0 : \theta \in \Omega_0$, we assumed that θ is a single parameter of the observable R.V.'s $X_1, \ldots, X_n$. More generally, hypothesis testing still works if θ is a vector instead.

- θ could be a **vector of parameters** that completely describes the joint distribution of $X_1, \ldots, X_n$.
- Even more generally, θ could be any vector of R.V.'s.
  - These R.V.'s should be latent. (If they were observable, then we can just do experiments to observe their values directly.)

Testing simple hypotheses

Recall: Any hypothesis $H_i : \theta \in \Omega_i$ is called simple if Ω_i contains exactly one value, and is called composite if Ω_i contains more than one value.

- This definition still applies to the case when θ is a vector.
- e.g. if $\theta = (\mu, \sigma^2)$ is a pair of parameters representing the mean and variance of a normal distribution, and if $\Omega_0 = \{(0, 1)\}$, then the hypothesis $H_0 : \theta \in \Omega_0$ is **simple**.
  - This is because Ω_0 contains exactly one value, the pair $(0, 1)$.
  - We can think of $(0, 1)$ as a **single** point on the xy-plane.

The simplest kind of hypothesis testing is one where both the null hypothesis H_0 and the alternative hypothesis H_1 are simple.

$$H_0 : \theta = \theta_0 \quad \text{(null hypothesis)};$$
$$H_1 : \theta = \theta_1 \quad \text{(alternative hypothesis)};$$

- θ_0 and θ_1 could be real numbers, or they could be real vectors.

Error probabilities

Suppose $\mathcal{H}$ is a hypothesis test with the following hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{(null hypothesis)};$$
$$H_1 : \theta = \theta_1 \quad \text{(alternative hypothesis)};$$

Recall: There are two types of errors in hypothesis testing.

- A type I error occurs if H_0 is **true** but we **reject** H_0.
  - The type I error probability of $\mathcal{H}$ is $\Pr(\text{reject } H_0 | \theta = \theta_0)$.
- A type II error occurs if H_0 is **false** but we **do not reject** H_0.
  - The type II error probability of $\mathcal{H}$ is $\Pr(\text{do not reject } H_0 | \theta = \theta_1)$.
- Following the notation in the course textbook, we denote these two error probabilities by $\alpha(\mathcal{H})$ and $\beta(\mathcal{H})$ respectively.

Ideally, we want both $\alpha(\mathcal{H})$ and $\beta(\mathcal{H})$ to be small.

- Usually, $\alpha(\mathcal{H})$ and $\beta(\mathcal{H})$ are not equally important.
  - e.g. we may impose more importance on $\alpha(\mathcal{H})$ being small, rather than on $\beta(\mathcal{H})$ being small.
- **Question:** How do we express this idea mathematically, that $\alpha(\mathcal{H})$ is more important than $\beta(\mathcal{H})$?

Linear combinations of error probabilities

Let $\mathcal{H}$ be a hypothesis test with the following hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{(null hypothesis)};$$
$$H_1 : \theta = \theta_1 \quad \text{(alternative hypothesis)};$$

Recall: The error probabilities are denoted by:

- Type I error probability: $\alpha(\mathcal{H}) = \Pr(\text{reject } H_0 | \theta = \theta_0)$.
- Type II error probability: $\beta(\mathcal{H}) = \Pr(\text{do not reject } H_0 | \theta = \theta_1)$.

Mathematically, we fix some positive constants a and b, and we want to find a hypothesis test $\mathcal{H}$ (e.g. by choosing the rejection region suitably) such that $a\alpha(\mathcal{H}) + b\beta(\mathcal{H})$ is minimized.

- If $a > b$, then $\alpha(\mathcal{H})$ is effectively "more important" than $\beta(\mathcal{H})$ when trying to minimize $a\alpha(\mathcal{H}) + b\beta(\mathcal{H})$.
  - e.g. if we want to minimize $3\alpha(\mathcal{H}) + 2\beta(\mathcal{H})$, and we have two tests $\mathcal{H}_1, \mathcal{H}_2$ (out of many tests) such that $(\alpha(\mathcal{H}_1), \beta(\mathcal{H}_1)) = (0.1, 0.6)$ and $(\alpha(\mathcal{H}_2), \beta(\mathcal{H}_2)) = (0.6, 0.1)$, then $3\alpha(\mathcal{H}) + 2\beta(\mathcal{H})$ is NOT minimized at $\mathcal{H} = \mathcal{H}_2$.

Minimizing linear combinations of error probabilities

Let $\mathcal{H}$ be a hypothesis test with the following hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{(null hypothesis)};$$
$$H_1 : \theta = \theta_1 \quad \text{(alternative hypothesis)};$$

where θ is a random vector of parameters for the joint distribution of $X_1, \ldots, X_n$, and θ_0, θ_1 are given constant real vectors.

- Given some vector $\mathbf{x}$ of observed values, let $\mathcal{L}(\theta|\mathbf{x})$ be the likelihood function of θ (i.e. $\mathcal{L}(\theta|\mathbf{x}) = p_n(\mathbf{x}|\theta)$ in the discrete case, or $\mathcal{L}(\theta|\mathbf{x}) = f_n(\mathbf{x}|\theta)$ in the continuous case).

Theorem: Let a, b be positive constants. Suppose we reject H_0 if $a\mathcal{L}(\theta_0|\mathbf{x}) < b\mathcal{L}(\theta_1|\mathbf{x})$, and we don't reject H_0 if $a\mathcal{L}(\theta_0|\mathbf{x}) > b\mathcal{L}(\theta_1|\mathbf{x})$. Then for every possible hypothesis test $\mathcal{H}'$ with the same H_0, H_1,

$$a\alpha(\mathcal{H}) + b\beta(\mathcal{H}) \leq a\alpha(\mathcal{H}') + b\beta(\mathcal{H}').$$

Technicality: When $a\mathcal{L}(\theta_0|\mathbf{x}) = b\mathcal{L}(\theta_1|\mathbf{x})$, it does not matter whether H_0 is rejected or not rejected. Either way, the theorem is still true.

Minimizing linear combinations (continued)

Let $\mathcal{H}$ be a hypothesis test with null hypothesis $H_0 : \theta = \theta_0$ and alternative hypothesis $H_1 : \theta = \theta_1$, where θ is a random vector of parameters for the joint distribution of $X_1, \ldots, X_n$. Let $\mathcal{L}(\theta|\mathbf{x})$ be the likelihood function of θ (given the vector $\mathbf{x}$ of observed values).

Reformulation of previous theorem: Let a, b be positive constants. Suppose the test statistic Λ of $\mathcal{H}$ is defined by

$$\Lambda(\mathbf{x}) = \frac{\mathcal{L}(\theta_1|\mathbf{x})}{\mathcal{L}(\theta_0|\mathbf{x})}$$

for each possible vector $\mathbf{x}$ of observed values, and let the rejection region of $\mathcal{H}$ be either $(\frac{a}{b}, \infty)$ or $[\frac{a}{b}, \infty)$. Then every test $\mathcal{H}'$ with the same H_0, H_1 must satisfy

$$a\alpha(\mathcal{H}) + b\beta(\mathcal{H}) \leq a\alpha(\mathcal{H}') + b\beta(\mathcal{H}').$$

- The ratio $\frac{\mathcal{L}(\theta_1|\mathbf{x})}{\mathcal{L}(\theta_0|\mathbf{x})}$ is called the likelihood ratio of $\mathbf{x}$.
- **Note:** The likelihood ratio is NOT the same as the likelihood ratio statistic (introduced earlier in today's lecture).

Interpretation of likelihood ratio

Consider a hypothesis test $\mathcal{H}$ with null hypothesis $H_0 : \theta = \theta_0$ and alternative hypothesis $H_1 : \theta = \theta_1$, where θ is a random vector of parameters for the joint distribution of $X_1, \ldots, X_n$.

- $\mathcal{L}(\theta|\mathbf{x})$ is the likelihood function of θ (given $\mathbf{x}$).
- $\frac{\mathcal{L}(\theta_1|\mathbf{x})}{\mathcal{L}(\theta_0|\mathbf{x})}$ is the likelihood ratio of $\mathbf{x}$.

Interpretation: To minimize $a\alpha(\mathcal{H}) + b\beta(\mathcal{H})$, it suffices to choose the rejection region of $\mathcal{H}$ so that H_0 is rejected if the likelihood ratio of the observed vector $\mathbf{x}$ exceeds $\frac{a}{b}$, and H_0 is not rejected if the likelihood ratio does not exceed $\frac{a}{b}$. The previous theorem says that such a test is **optimal** (i.e. "best possible"), in the sense that it is impossible to get a smaller value for $a\alpha(\mathcal{H}) + b\beta(\mathcal{H})$.

- **Important Note:** It is very unfortunate that the likelihood ratio is NOT the observed value of the likelihood ratio statistic.
  - Likelihood ratio and likelihood ratio statistic are somewhat related, since both involve ratios of likelihood functions, but they are NOT directly related!

Neyman–Pearson lemma

Let $\mathcal{H}$ be a hypothesis test described as follows:

- Null hypothesis $H_0 : \theta = \theta_0$; alternative hypothesis $H_1 : \theta = \theta_1$.
- θ is a random vector of parameters.
- Given $\mathbf{x}$, reject H_0 if the likelihood ratio of $\mathbf{x}$ exceeds k (fixed positive constant), and do not reject H_0 otherwise.

Neyman–Pearson lemma: (**Consequence of previous theorem**) If $\mathcal{H}'$ is another hypothesis test with the same hypotheses H_0 and H_1, but with a smaller type I error probability, i.e. $\alpha(\mathcal{H}') < \alpha(\mathcal{H})$, then its type II error probability must be larger, i.e. $\beta(\mathcal{H}') > \beta(\mathcal{H})$, or equivalently, $\mathcal{H}'$ must have a smaller **power**.

- **Recall:** "$\mathcal{H}$ has power β_0" means exactly "$\beta(\mathcal{H}) \leq 1 - \beta_0$".
- $\mathcal{H}$ is called most powerful at significance level α_0 if the power of $\mathcal{H}$ is maximum among all level α_0 tests. (with the same hypotheses)

Spelling error in course textbook: "Nayman" in textbook should be "Neyman".

Interpretation: This lemma says that we can always find a most powerful test at any given significance level α_0 to be a test that uses likelihood ratio (as described in the previous theorem).

Testing composite hypotheses

Let $X_1, \ldots, X_n$ be observable R.V.'s, whose joint distribution is parametrized by an unknown vector θ of parameters. You have narrowed down the possibilities for θ to several vectors.

- For each pair of possible vectors θ_0, θ_1, we can do a hypothesis test with simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$.
- Given some fixed α_0, Neyman–Pearson's lemma tells us we can find a most powerful level α_0 test $\mathcal{H}$ using likelihood ratio.
  - $\mathcal{H}$ most powerful $\Rightarrow \beta(\mathcal{H}) \leq \beta(\mathcal{H}')$ for all level α_0 tests $\mathcal{H}'$ with the same null hypothesis $H_0 : \theta = \theta_0$ and alternative hypothesis $H_1 : \theta = \theta_1$.

Question: Can we find a test $\mathcal{H}$ that still remains most powerful as we vary the alternative hypothesis over all possible vectors θ_1?

Answer: In general, no! It is not always possible to find a test $\mathcal{H}$ that is simultaneously most powerful.

- However, if we can find such a test $\mathcal{H}$ that so happens to be simultaneously most powerful, then we say that $\mathcal{H}$ is a uniformly most powerful (UMP) test.
  - See next slide for a more precise definition.

Uniformly most powerful test

Let $\mathcal{H}$ be a hypothesis test with the following hypotheses:

$$H_0 : \theta \in \Omega_0 \quad \text{(null hypothesis)};$$
$$H_1 : \theta \in \Omega_1 \quad \text{(alternative hypothesis)};$$

where θ is a random vector of parameters for the joint distribution of some observable R.V.'s $X_1, \ldots, X_n$. Assume that H_1 is composite. Let $T_{\mathcal{H}}$ be the test statistic of $\mathcal{H}$. Let $R_{\mathcal{H}}$ be the rejection region of $\mathcal{H}$.

- The power function of $\mathcal{H}$ is $\pi(\omega|\mathcal{H}) = \Pr(T_{\mathcal{H}} \in R_{\mathcal{H}} | \theta = \omega)$.
- The significance level of $\mathcal{H}$ is α_0 if $\pi(\omega|\mathcal{H}) \leq \alpha_0$ for all $\omega \in \Omega_0$.
- $\mathcal{H}$ has power β_0 if $\pi(\omega|\mathcal{H}) \geq \beta_0$ for all $\omega \in \Omega_1$.

Definition: We say that $\mathcal{H}$ is a uniformly most powerful (UMP) test at significance level α_0 if the statement

$$\text{“}\pi(\omega|\mathcal{H}) \geq \pi(\omega|\mathcal{H}') \quad \text{for every } \omega \in \Omega_1\text{”}$$

is true for **all** level α_0 tests $\mathcal{H}'$ with the same null hypothesis H_0 and alternative hypothesis H_1.

Summary

- Likelihood ratio statistic
- Likelihood ratio test
- Linear combinations of error probabilities
- Likelihood ratio
- Neyman–Pearson lemma
- Uniformly most powerful test

Reminders:

There is **mini-quiz 4** (15mins) next week during Cohort Class.

- Final mini-quiz! Tested on all materials from Lectures 15–20 and Cohort classes weeks 9–11. Today's lecture is Lecture 20.

Make-up class for **next week**'s Friday's Cohort Class

- Originally on 19th April (Good Friday).
- Make-up: On 17th April (Wednesday), 2–4pm, CC14 (2.507).
  - So your mini-quiz 4 will be on Wednesday!
- Next Thursday's cohort classes are on as usual.
- This week's cohort classes are on as usual.