Statistical Hypothesis Testing

Motonobu Kanagawa

Introduction to Statistics, EURECOM

April 6, 2023

Outline

1. Introduction: The Lady Tasting Tea Experiment
2. Procedure of Statistical Hypothesis Testing
3. Type 1 Error, Type 2 Error and the Power of a Test
4. Test Statistics
5. *P*-Value
6. Neyman-Pearson Lemma and Likelihood Ratio Test
7. Conclusions and Further Reading

The Lady Tasting Tea Experiment [Fisher, 1937, Chapter II]

- There was a lady who claimed that she can distinguish the tastes of tea with milk made in the following two different ways:

Way M: Milk is first poured into the cup, and tea later.

Way T: Tea is first poured into the cup, and milk later.

- Ronald Fisher [Fisher, 1937] came up with an idea of testing her claim by a randomized experiment.

The Lady Tasting Tea Experiment [Fisher, 1937, Chapter II]

1) Let's make 8 cups of tea, of which

- 4 cups are made in Way M.
- 4 cups are made in Way T.

2) Shuffle the order of the 8 cups randomly:

- For instance, assume that as a result, the cups are ordered as:

M-M-T-M-T-T-T-M

- This information was not shared to the lady.

- She only knew that 4 of them were made in M; and the other 4 cups in T.

The Lady Tasting Tea Experiment [Fisher, 1937, Chapter II]

3) Ask the lady

- to taste the 8 cups of tea in the given order; and
- to pick up 4 cups of M from the 8 cups.

- In the end, the lady correctly identified all the 4 cups of M from the 8 cups (i.e., did no mistake).

- Fisher concluded that it is likely that she can distinguish the two ways of making tea.

- What was Fisher's reasoning?

Fisher's Reasoning

- In total, there are 70 different ways of choosing the 4 cups for M from the 8 cups

$$70 = \frac{8!}{4!4!} = \frac{8 \times 7 \times 6 \times 5 \times 4}{4 \times 3 \times 2 \times 1}$$

- Assume that the lady

- was not able to distinguish the tastes (= null hypothesis); and
- just did a random guess, picking one of the 70 ways randomly.

- **Under this assumption**, the probability of correctly identifying 4 cups of M from the 8 cups is $1/70 \approx 0.014$:

- This probability is very small, so we can conclude that

- It is unlikely that the lady is doing a random guess.
- i.e., the null hypothesis is unlikely to be true.

Fisher's Reasoning

- Assume instead a situation where the lady

  - correctly identified 3 M cups, but
  - wrongly chose 1 cup.

- There are 16 different ways of choosing 3 M cups correctly and one cup wrongly (**Exercise**: confirm this).

- Thus, under the null hypothesis (= the lady is doing a random guess),

  - the probability of correctly choosing 3 M cups and wrongly choosing 1 cup is $16/70 \approx 0.23$.

- This probability is "not very small," and therefore

  - we cannot deny the null hypothesis that the lady was doing a random guess.

Fisher's Reasoning

- This example illustrates the idea of statistical hypothesis testing and a randomized experiment.

- In this lecture, we'll learn basics of hypothesis testing.

- For reading, I recommend [Rao, 1973, Chapter 7].

Outline

1. Introduction: The Lady Tasting Tea Experiment
2. Procedure of Statistical Hypothesis Testing
3. Type 1 Error, Type 2 Error and the Power of a Test
4. Test Statistics
5. *P*-Value
6. Neyman-Pearson Lemma and Likelihood Ratio Test
7. Conclusions and Further Reading

Hypothesis Testing: Statistical Proof by Contradiction

Hypothesis testing may be understood as a statistics version of Proof by Contradiction:

Proof by Contradiction (Mathematics)

1. To prove a statement A, assume that A is not true;
2. Starting from the assumption, derive a statement B that produces a contradiction.
3. Conclude that the statement A is true.

Procedure of Testing: Step 1. Defining Hypotheses

- Hypothesis testing starts from defining a null hypothesis H_0 and an alternative hypothesis H_1

Null Hypothesis H_0

The hypothesis that you try to reject in the end.

Alternative Hypothesis H_1

The hypothesis that you try to "prove" (statistically).

Example (The lady tasting tea experiment)

- The null hypothesis H_0:
  — The lady cannot distinguish the tastes of tea of different kinds.
- The alternative hypothesis H_1:
  — The lady can distinguish the tastes of tea of different kinds.

Procedure of Testing: Step 1. Defining Hypotheses

- Let $(\Omega, \mathcal{F})$ be a measurable space, where

  - Ω is a sample space, consisting possible outcomes of the experiment.
  - $\mathcal{F}$ is a σ-algebra, i.e., a set of subsets of Ω for which probabilities can be defined.

- For the null H_0 and alternative hypotheses H_1, define the associated probability distributions P_0 and P_1 on $(\Omega, \mathcal{F})$:

Distributions under the Null and Alternative Hypotheses

- P_0 is the probability distribution on Ω when the null H_0 is true.
- P_1 is the probability distribution on Ω when the alternative H_1 is true.
- We may write P_0 and P_1 in the form of conditional distribution:

$$P(S \mid H_0) := P_0(S), \quad P(S \mid H_1) := P_1(S), \quad S \in \mathcal{F}.$$

Procedure of Testing: Step 1. Defining Hypotheses

Example (The lady tasting tea experiment)

- The sample space Ω consists of 70 different ways of choosing 4 cups of M from 8 cups:

$$\Omega := \{\omega_1, \omega_2, \ldots, \omega_{70}\},$$

where each $\omega_i \in \Omega$ represents one way of ordering, e.g.,

$$\omega_1 := \text{M-M-M-M-T-T-T-T}$$
$$\omega_2 := \text{M-M-M-T-M-T-T-T}$$
$$\ldots$$
$$\omega_{69} := \text{M-T-T-T-M-M-M-T}$$
$$\omega_{70} := \text{T-T-T-T-M-M-M-M}$$

Procedure of Testing: Step 1. Defining Hypotheses

Example (The lady tasting tea experiment)

- Under the null hypothesis H_0, the lady gives a random guess; therefore the distribution P_0 under the null is

$$P_0(\{\omega_1\}) = P_0(\{\omega_2\}) = \cdots = P_0(\{\omega_{70}\}) = 1/70.$$

- Under the alternative hypothesis H_1, let's **assume** that the lady can identify the correct 4 cups of M with probability 1:

$$P_1(\{\omega_{32}\}) = 1, \quad P(\{\omega_i\}) = 0 \text{ for all } i \neq 32,$$

where $\omega_{32} \in \Omega$ is the correct ordering:

$$\omega_{32} := \text{M-M-T-M-T-T-T-M}.$$

Procedure of Testing: Step 1. Defining Hypotheses

Example (The lady tasting tea experiment)

- Note that the way of defining P_1 is not unique: we may define, e.g.,

$$P_1(\{\omega_{32}\}) = 0.9, \quad P_1(\{\omega_i\}) = 0.1/69 \text{ for all } i \neq 32,$$

- This may represent another alternative hypothesis H_1' that
  - the lady can distinguish tastes of tea of different kinds
  - but may loose her tasting ability with probability 1/10.

Step 2: Defining Significance Level and Critical Region

- The next step is to decide the level of significance and the critical region for the test.

Significance Level

- Define a small constant $\alpha > 0$, called the level of significance (e.g., $\alpha = 0.05$ or $\alpha = 0.01$).

Step 2: Defining Significance Level and Critical Region

Critical Region

- Given a significance level $\alpha > 0$, determine a subset $S_\alpha \subset \Omega$ (such that $S_\alpha \in \mathcal{F}$), called the critical region, such that

1. the probability of S_α under the null H_0 is less than or equal to α:
$$P_0(S_\alpha) \leq \alpha;$$
2. the probability of S_α under the alternative H_1
$$P_1(S_\alpha)$$
becomes as large as possible.

Remark

- The second requirement is equivalent to choosing S_α so that $P_1(\Omega \backslash S_\alpha) = 1 - P_1(S_\alpha)$ becomes as small as possible .

Step 2: Defining Significance Level and Critical Region

Example (The lady tasting tea experiment)

- Let's define $\alpha := 0.05$ as our significance level.
- We may define the critical region S_α as the singleton set of ω_{32}:

$$S_\alpha := \{\omega_{32}\},$$

where $\omega_{32} :=$ M-M-T-M-T-T-T-M is the correct ordering of 8 cups.

- Then

1. The probability of S_α under the null H_0 (the lady cannot distinguish the tastes) is
$$P_0(S_\alpha) = 1/70 \approx 0.014 \leq 0.05 = \alpha.$$
2. The probability of S_α under the alternative H_1 (the lady can perfectly distinguish the tastes) is
$$P_1(S_\alpha) = 1.$$

Step 2: Defining Significance Level and Critical Region

Example (The lady tasting tea experiment)

- Note that $S_\alpha = \{\omega_{32}\}$ is not the only way of defining a critical region.

- For instance, we may define

$$S_\alpha := \{\omega_{31},\ \omega_{32},\ \omega_{33}\},$$

  where ω_{31} and ω_{33} are two ways of wrongly identifying one M cup as T.

$$\omega_{31} := \text{M-M-T-M-T-T-M-T},$$
$$\omega_{33} := \text{M-T-M-M-T-T-T-M}$$

- In this case,

$$P_0(S_\alpha) = 3/70 \approx 0.043 \leq 0.05 = \alpha,$$
$$P_1(S_\alpha) = P_1(\{\omega_{32}\}) + P_1(\{\omega_{31}, \omega_{33}\}) = 1 + 0 = 1.$$

Step 2: Defining Significance Level and Critical Region

Example (The lady tasting tea experiment)

- Or even we may define the critical region S_α for arbitrary $i = 1, 2, \ldots, 70$ with $i \neq 32$ such that

$$S_\alpha := \{\omega_i\}$$

- In this case, we have

$$P_0(S_\alpha) = 1/70 \approx 0.014 \leq 0.05 = \alpha,$$
$$P_1(S_\alpha) = 0.$$

- Since $P_1(S_\alpha) = 0$, this critical region S_α should not be chosen for our alternative hypothesis H_1.

Step 3: Obtain a Sample, and Make a Decision

- After deciding a significance level $\alpha > 0$ and a critical region $S_\alpha \subset \Omega$, make a statistical decision in the following way:

Statistical decision of whether rejecting H_0 or not

- Obtain a sample $\omega_e \in \Omega$ by performing an experiment.
  - If $\omega_e \in S_\alpha$, we reject the null hypothesis H_0.
  - If $\omega_e \notin S_\alpha$, we don't reject the null hypothesis H_0.
- We may say that the test is significant with level α.

Step 3: Obtain a Sample, and Make a Decision

Example (The lady tasting tea experiment)

- Let $\alpha := 0.05$ and $S_\alpha := \{\omega_{32}\}$.

- As a result of the experiment, the lady correctly identified the 4 M cups out of 8 cups, i.e.,

$$\omega_e = \text{M-M-T-M-T-T-T-M} = \omega_{32}.$$

- Thus we have $\omega_e \in S_\alpha$; and thus
- We reject the null hypothesis H_0 that the lady cannot distinguish the tastes of tea of different kinds.
- This test is significant with the level $\alpha = 0.05$.

Remarks on the Testing Procedure

- Ronald Fisher made the following remarks on the testing procedure.

[Fisher, 1937, Section 8]

- It should be noted that the null hypothesis is never proved or established, but is possibly **disproved**, in the course of experimentation.
- Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

- This means that $\omega_e \notin S_\alpha$ does not prove the null hypothesis H_0; we just don't reject the null H_0.

Remarks on the Testing Procedure

Example (The lady tasting tea experiment)

- Assume that the lady made one mistake: $\omega_e := \omega_{31} \neq \omega_{32}$.
- Then $\omega_e \notin S_\alpha = \{\omega_{32}\}$, and we don't reject the null H_0.
- But this **does not prove** the null hypothesis H_0 that the lady cannot distinguish the tastes of tea.

Outline

1. Introduction: The Lady Tasting Tea Experiment
2. Procedure of Statistical Hypothesis Testing
3. Type 1 Error, Type 2 Error and the Power of a Test
4. Test Statistics
5. *P*-Value
6. Neyman-Pearson Lemma and Likelihood Ratio Test
7. Conclusions and Further Reading

Type 1 Error and Type 2 Error

- In hypothesis testing, there are two kinds of errors: **Type 1** and **Type 2**.

Type 1 Error and Type 2 Error

- **Type 1 Error**:
  — Rejecting the null hypothesis H_0, when H_0 is true.
- **Type 2 Error**:
  — Not rejecting the null hypothesis H_0, when an alternative hypothesis H_1 is true.

Type 1 Error and Type 2 Error

Example (The lady tasting tea experiment)

- **Type 1 Error:**
  - Rejecting the null hypothesis H_0 that the lady is doing a random guess
  - when the lady is really doing a random guess (H_0 is true)
- **Type 2 Error:**
  - Not rejecting that the null hypothesis H_0 that the lady is doing a random guess
  - when the lady has the ability of distinguishing the tastes of tea (H_1 is true)

Type 1 Error and the Level of Significance

- Recall that

  - we reject the null H_0 when $\omega_e \in S_\alpha$;
  - we don't reject the null H_0 when $\omega_e \notin S_\alpha$ (i.e., when $\omega_e \in \Omega \backslash S_\alpha$).

- Thus, the probability of making the Type 1 error may be given by

$$P(S_\alpha \mid H_0) := P_0(S_\alpha) \leq \alpha,$$

where the inequality follows from the definition of critical region S_α.

- i.e., the level of the significance α is (the upper-bound of) the probability of making the Type 1 error.

Type 2 Error and Statistical Power

- On the other hand, the probability of making the Type 2 error is:

$$P_1(\Omega \backslash S_\alpha) = 1 - P_1(S_\alpha).$$

- Thus, the following ways of choosing a critical region S_α are equivalent:

1. $P_1(S_\alpha)$ is maximized.
2. $1 - P_1(S_\alpha)$ is minimized (probability of Type 2 error).

- This probability $P_1(S_\alpha)$ is called the power of the test.

Power of a Test, $P_1(S_\alpha)$

- The probability of rejecting the null hypothesis H_0, when the alternative hypothesis H_1 is true.

Recap: Critical Region

Critical Region

- Given a significance level $\alpha > 0$, determine a subset $S_\alpha \subset \Omega$ (such that $S_\alpha \in \mathcal{F}$), called the critical region, such that

1. the probability of S_α under the null H_0 is less than or equal to α:

$$P_0(S_\alpha) = \textbf{Probability of Type 1 Error} \leq \alpha;$$

2. the probability of S_α under the alternative H_1

$$P_1(S_\alpha) = \textbf{Power of the Test}$$

becomes as large as possible.

Remark

- The second requirement is equivalent to choosing S_α so that $P_1(\Omega \backslash S_\alpha) = \textbf{Prob. of Type 2 Error} = 1 - P_1(S_\alpha)$ becomes as small as possible .

Type 1 Error, Type 2 Error, and Power of a Test

- Relations between the Type 1 error, Type 2 error and the power of a test can be summarized as follows:

Reality \ Test	Not Reject H_0	Reject H_0
H_0 is true	(prob. $1-\alpha$)	Type 1 Error (prob. α)
H_1 is true	Type 2 Error (prob. β)	(**Power** $=$ prob. $1-\beta$)

Outline

1. Introduction: The Lady Tasting Tea Experiment
2. Procedure of Statistical Hypothesis Testing
3. Type 1 Error, Type 2 Error and the Power of a Test
4. **Test Statistics**
5. P-Value
6. Neyman-Pearson Lemma and Likelihood Ratio Test
7. Conclusions and Further Reading

Test Statistics

- In practice, the determination of a critical region S_α is done by defining a test statistic.

Test Statistics

- Let Ω be a sample space.
- A test statistic T is a (measurable) function from Ω to $\mathbb{R}$:

$$T : \Omega \to \mathbb{R}.$$

Remark

- Depending on the problem, we may define a different range for a statistic T.
- e.g., $T : \Omega \to \mathbb{Z}$ (where where $\mathbb{Z}$ is the set of all integers).

- A test statistic $T : \Omega \to \mathbb{R}$ summarizes characteristics of an experiment outcome $\omega_e \in \Omega$ into one dimensional value $T(\omega_e) \in \mathbb{R}$.

Test Statistics

- For any (measurable) subset $A \subset \mathbb{R}$, we can define the corresponding subset in Ω by the inverse map of T as

$$T^{-1}(A) := \{\omega \in \Omega \mid T(\omega) \in A\} \subset \Omega$$

- Therefore, we can define a critical region $S_\alpha \subset \Omega$ by defining a corresponding subset $I_\alpha \subset \mathbb{R}$ for T:

$$S_\alpha := T^{-1}(I_\alpha) = \{\omega \in \Omega \mid T(\omega) \in I_\alpha\} \subset \Omega$$

- We thus call I_α a critical region with significance level $\alpha > 0$, if it satisfies

$$P_{0,T}(I_\alpha) := P_0(T^{-1}(I_\alpha)) = P_0(S_\alpha) \leq \alpha,$$

- Here, $P_{0,T}$ is the probability distribution on $\mathbb{R}$, induced from the test statistic $T : \Omega \to \mathbb{R}$ and the distribution P_0 on Ω under the null H_0.

Hypothesis Testing with a Test Statistic

- Hypothesis testing of significance level $\alpha > 0$ can be carried out, with the test statistic T and the critical region $I_\alpha \subset \mathbb{R}$ in the following way:

Hypothesis Testing with a Test Statistic

- Let $\omega_e \in \Omega$ be the outcome of an experiment.
  - Reject the null hypothesis H_0, if $T(\omega_e) \in I_\alpha$;
  - Not reject the null hypothesis H_0, if $T(\omega_e) \notin I_\alpha$.

- The question is how to choose the critical region $I_\alpha \subset \mathbb{R}$.

- To this end, we need to consider the probabilities of Type 1 and 2 errors, and the power of the test.

- This requires considering the distributions of the test statistic T under the null H_0 and alternative H_1, respectively.

Probability Distributions of a Test Statistic

Distribution of T under the Null Hypothesis H_0

- Let $(\Omega, \mathcal{F}, P_0)$ be the probability space associated with the null hypothesis H_0.
- Under the null H_0, the test statistic $T : \Omega \to \mathbb{R}$ can be interpreted as a random variable in $\mathbb{R}$ induced from $(\Omega, \mathcal{F}, P_0)$:

$$T(\omega), \quad \omega \sim P_0$$

- Then the probability distribution of T under the null hypothesis H_0, denoted by $P_{0,T}$, is given by

$$P_{0,T}(A) := P_0(T^{-1}(A)) \ \text{ for any measurable } A \subset \mathbb{R}$$

Probability Distributions of a Test Statistic

Distribution of T under the Alternative Hypothesis H_1

- Let $(\Omega, \mathcal{F}, P_1)$ be the probability space associated with the alternative hypothesis H_1.
- Under the alternative H_1, the test statistic $T : \Omega \to \mathbb{R}$ can be interpreted as a random variable in $\mathbb{R}$ induced from $(\Omega, \mathcal{F}, P_1)$:

$$T(\omega), \quad \omega \sim P_1$$

- Then the probability distribution of T under the alternative hypothesis H_1, denoted by $P_{1,T}$, is given by

$$P_{1,T}(A) := P_1(T^{-1}(A)) \text{ for any measurable } A \subset \mathbb{R}$$

Type 1 Error, Type 2 Error, and Power

- Recall that the Type 1 and Typer 2 errors of a test are defined as:

- **Type 1 Error:** rejecting the null H_0 when H_0 is true;
- **Type 2 Error:** not rejecting the null H_0 when an alternative H_1 is true.

- Since the test rejects H_0 when $T(\omega_e) \in I_\alpha$, the probability of making the Type 1 Error is thus given by

$$P_{0,T}(I_\alpha) = P_0(T^{-1}(I_\alpha))$$

- Since the test does not reject H_0 when $T(\omega_e) \notin I_\alpha$, the probability of making the Type 2 Error is

$$P_{1,T}(\mathbb{R} \backslash I_\alpha) = 1 - P_{1,T}(I_\alpha)$$

- The **Test Power**, i.e., the probability of rejecting when H_1 is true, is thus

$$P_{1,T}(I_\alpha) = 1 - \text{Prob. Type 2 Error}$$

Test Statistics: How to Choose the Critical Region

- To summarize, the critical region $I_\alpha \subset \mathbb{R}$ should be chosen as follows:

Critical Region for a Test Statistic

- Let $T : \Omega \to \mathbb{R}$ be a test statistic.
- Given a significance level $\alpha > 0$, determine a subset $I_\alpha \subset \mathbb{R}$, called the critical region, such that

1. the probability of I_α under the null H_0 is less than or equal to α:

$$P_{0,T}(I_\alpha) := P_0(T^{-1}(I_\alpha)) = \textbf{Type 1 Error} \le \alpha;$$

2. the probability of I_α under the alternative H_1

$$P_{1,T}(I_\alpha) := P_1(T^{-1}(I_\alpha)) = \textbf{Power of the Test}$$

becomes as large as possible.

Example: Testing the Location of a Gaussian Mean

- Let p^* be an unknown probability density function on $\mathbb{R}$.

- Assume that we know/believe that p^* is Gaussian, with unknown mean $\mu \in \mathbb{R}$ and known variance $\sigma^2 > 0$:

$$p^*(x) = p_{\mathrm{gauss}}(x; \mu, \sigma_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

- Assume that we can perform an experiment to obtain an i.i.d. sample of size n from p^*:

$$x_1, \ldots, x_n \in \mathbb{R}$$

- Assume that we are interested in testing whether the unknown mean μ is equal to some specified value $\mu_0 \in \mathbb{R}$ or not.

- Thus, the null hypothesis H_0 and alternative hypothesis H_1 may be defined as

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

Example: Testing the Location of a Gaussian Mean

- For instance, assume that μ_0 is the average blood pressure of the whole French population.

- Assume that we are interested in the effect of a certain drug on the blood pressure.

- Let $\omega_e = (x_1, \ldots, x_n)$ be the blood pressures of n randomly selected French people, measured after each being treated the drug.

- By testing the null hypothesis $H_0 : \mu = \mu_0$, we could investigate whether the drug is effective in changing the blood pressure or not.

Example: Testing the Location of a Gaussian Mean

- We can define the sample space Ω as

$$\Omega := \mathbb{R}^n.$$

- Each $\omega := (x_1, \dots, x_n) \in \Omega$ represents a possible experiment outcome of n i.i.d. observations.

- Thus, the distribution P_0 on Ω under the null hypothesis H_0 is given by the density function $p_0 : \Omega \to \mathbb{R}$:

$$p_0(\omega) = \prod_{i=1}^{n} p_{\mathrm{gauss}}(x_i; \mu_0, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_i - \mu_0)^2}{2\sigma^2})$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{2\sigma^2}), \quad \omega := (x_1, \dots, x_n) \in \Omega.$$

Example: Testing the Location of a Gaussian Mean

- We can define a test statistic $T : \Omega \to \mathbb{R}$ as

$$T(\omega) := T((x_1, \ldots, x_n)) := \frac{\sqrt{n}}{\sigma} \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right),$$
$$\omega := (x_1, \ldots, x_n) \in \Omega := \mathbb{R}^n.$$

- Consider

$$\omega = (X_1, \ldots, X_n) \sim P_0 \quad (i.e., X_1, \ldots, X_n \sim p(x; \mu_0, \sigma^2), \quad i.i.d.)$$

as a random variable under the null hypothesis H_0.

- Then the distribution $P_{0,T}$ of the test statistic

$$T(\omega) = T((X_1, \ldots, X_n)) = \frac{\sqrt{n}}{\sigma} \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu_0 \right)$$

is Gaussian, with mean 0 and variance 1.

Example: Testing the Location of a Gaussian Mean

- In other words, the density function $p_{0,T}$ of the distribution $P_{0,T}$ of the test statistic T under the null hypothesis H_0 is

$$p_{0,T}(t) := p_{\text{gauss}}(t; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}), \quad t \in \mathbb{R}.$$

Exercise: Prove this.

Hint: First derive the probability distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$.

To this end, use the following facts (where $X \sim p_{\text{gauss}}(x; \mu_0, \sigma^2)$):

- The sum of Gaussian random variables is Gaussian.
- $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X] = \mu_0$
- $\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\mathbb{V}[X] = \frac{\sigma^2}{n}$.

Example: Testing the Location of a Gaussian Mean

- Thus, we may define a critical region I_α with significance level $\alpha > 0$

$$I_\alpha := (-\infty, -c_\alpha] \cup [c_\alpha, \infty) \subset \mathbb{R}$$

where c_α is a constant satisfying

$$P_{0,T}(I_\alpha) = \int_{-\infty}^{-c_\alpha} p_{0,T}(t)dt + \int_{c_\alpha}^{\infty} p_{0,T}(t)dt = \alpha.$$

- For instance, if $\alpha := 0.05$, we can take $c_\alpha \approx 1.96$.

Example: Testing the Location of a Gaussian Mean

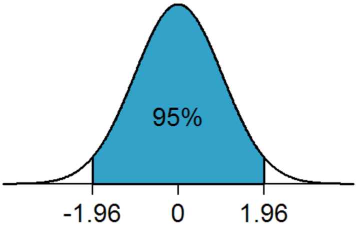


Figure 1: From Wikipedia "1.96"

- The tail regions are the critical region I_α with $\alpha = 0.05$.
- We reject the null hypothesis $H_0 : \mu = \mu_0$ if

$$T(\omega_e) > 1.96 \quad \text{or} \quad T(\omega_e) < -1.96$$

for an experiment outcome $\omega_e = (x_1, \ldots, x_n)$.

Test Statistics: Important Points

- A test statistic $T : \Omega \to \mathbb{R}$ summarizes characteristics of an experiment outcome $\omega_e \in \Omega$ into one dimensional value $T(\omega_e) \in \mathbb{R}$.

- This summary $T(\omega_e)$ should capture important characteristics of ω_e for testing the null hypothesis H_0 against an alternative H_1.

- At the same time, $T : \Omega \to \mathbb{R}$ should be designed so that the distribution $P_{0,T}$ under the null hypothesis H_0 is easy to compute.

  - This is needed to determine the critical region.

Outline

1. Introduction: The Lady Tasting Tea Experiment
2. Procedure of Statistical Hypothesis Testing
3. Type 1 Error, Type 2 Error and the Power of a Test
4. Test Statistics
5. P-Value
6. Neyman-Pearson Lemma and Likelihood Ratio Test
7. Conclusions and Further Reading

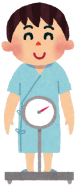
P-Value

- Hypothesis testing outputs binary decisions ("Reject" or "Not reject") with a pre-specified significance level $\alpha > 0$.

  - Recall that a lower value of α implies that the test is more significant, in the sense that the probability of Type 1 Error ($= \alpha$) is smaller.

- The *p*-value provides a continuous measure of statistical significance for an experimental outcome $\omega_e \in \Omega$ against the null hypothesis H_0.

  - A lower *p*-value indicates more that the null hypothesis H_0 fails to explain the characteristics of the observed outcome ω_e.



P-Value

Definition of *P*-Value [Lehmann and Romano, 2005, Section 3.3]

- For each $\alpha > 0$, let $S_\alpha \subset \Omega$ be the critical region for the null hypothesis H_0 such that

$$P_0(S_\alpha) = \alpha.$$

- Assume that the critical regions are nested:

$$S_\alpha \subset S_{\alpha'} \subset \Omega \quad \text{for all } 0 < \alpha < \alpha' < 1$$

- Then the *p*-value for an experimental outcome ω_e is defined by

$$p\text{-value} := \mathbf{p}(\omega_e) := \min_{\alpha > 0} \alpha \;\; \text{such that} \;\; \omega_e \in S_\alpha$$

- i.e., the minimum significance level α such that the critical region S_α contains the outcome ω_e.

P-Value

- Note that the p-value depends on

- The definition of the probability distribution P_0 under the null hypothesis H_0;
- The definition of the critical regions S_α, $0 < \alpha < 1$ (i.e., the test).

P-Values for a Test Statistic

- In practice, p-values are defined for a given test statistic T and the distribution P_0 under the null hypothesis H_0.

P-Values for a Test Statistic

- Let $T : \Omega \to \mathbb{R}$ be a test statistic with probability distribution $P_{0,T}$ under the null hypothesis H_0.
- For each $\alpha > 0$, let $I_\alpha \subset \mathbb{R}$ be the critical region such that

$$P_{0,T}(I_\alpha) = \alpha \quad \text{for all } 0 < \alpha < 1.$$

- Assume that the critical regions are nested:

$$I_\alpha \subset I_{\alpha'} \subset \mathbb{R}, \quad 0 < \alpha < \alpha' < 0.$$

- Then the p-value of an observed outcome $\omega_e \in \Omega$ is given by

$$p\text{-value} := \mathbf{p}(\omega_e) = \min_{\alpha > 0} \alpha \ \text{ such that } \ T(\omega_e) \in I_\alpha.$$

P-Values for a Test Statistic

- Since $I_\alpha \subset I_{\alpha'}$ for $\alpha < \alpha'$, we have

$$S_\alpha = \{\omega \in \Omega \mid T(\omega) \in I_\alpha\} \subset \{\omega \in \Omega \mid T(\omega) \in I_{\alpha'}\} = S_{\alpha'}$$

- Thus, I_α being nested implies S_α being nested:

$$I_\alpha \subset I_{\alpha'} \Longrightarrow S_\alpha \subset S_{\alpha'}, \quad 0 < \alpha < \alpha' < 1.$$

- Therefore the definition of the p-value for a test statistic $T : \Omega \to \mathbb{R}$ is consistent with the definition of the p-value with significant regions S_α in the original sample space Ω.

P-Values for a Test Statistic

According to the **American Statistical Association**'s Statement on *p*-Values [Wasserstein and Lazar, 2016, Section 2]:

- *Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data ... would be equal to or more extreme than its observed value.*

P-Values for a Test Statistic

- For instance, assume that the critical region I_α is given by

$$I_\alpha := [c_\alpha, \infty),$$

for constant c_α satisfying

$$c_{\alpha'} < c_\alpha \quad \text{for all } 0 < \alpha < \alpha' < 1$$

so that

$$I_\alpha = [c_\alpha, \infty) \subset [c_{\alpha'}, \infty) = I_{\alpha'}$$

- Then the *p*-value is given by

$$\mathbf{p}(\omega_e) = \min_{\alpha > 0} \alpha \quad \text{such that } T(\omega_e) \in [c_\alpha, \infty)$$

i.e., the minimum significance level α such that the critical region $[c_\alpha, \infty)$ contains the test statistic $T(\omega_e)$.

Illustration of *P*-Value

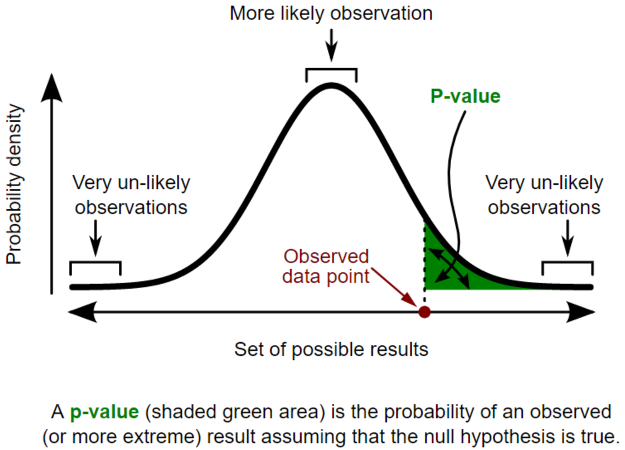


Figure 2: This figure illustrates the p-value for one-sided critical region of the form $[c_\alpha, \infty)$. From Wikipedia "p-value".

P-Value: Example of the Location Test of a Gaussian Mean

- Consider again the location test of a Gaussian mean.

- We constructed the two-sided critical regions I_α with a significance level $\alpha > 0$ as

$$I_\alpha := (-\infty, -c_\alpha] \cup [c_\alpha, \infty)$$

for a constant $c_\alpha > 0$ satisfying

$$P_{0,T}(I_\alpha) = \int_{-\infty}^{-c_\alpha} p_{0,T}(t)dt + \int_{c_\alpha}^{\infty} p_{0,T}(t)dt = \alpha.$$

- For instance, if $\alpha := 0.05$, we can take $c_\alpha \approx 1.96$.

P-Value: Example of the Location Test of a Gaussian Mean

- Assume that we obtained an experiment outcome $\omega_e := (x_1, \ldots, x_n) \in \Omega$ such that

$$T(\omega_e) = \frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu_0\right) = 2.24$$

- In this case, the *p*-value is given by

$$\begin{aligned}\mathbf{p}(\omega_e) &= \min_{\alpha>0} \alpha \quad \text{such that } T(\omega_e) = 2.24 \in (-\infty, -c_\alpha] \cup [c_\alpha, \infty) \\ &\approx 0.025.\end{aligned}$$

- Thus, the null hypothesis $H_0 : \mu = \mu_0$ would have been rejected if the significance level was set to $\alpha = 0.05$ (since $c_\alpha \approx 1.96$ for $\alpha = 0.05$).

P-Value: Example of the Location Test of a Gaussian Mean

Exercise:

- Derive p-values for the cases where, e.g.,

$T(\omega_e) = 1.26.$

$T(\omega_e) = 3.42.$

- You can for instance use the table from

`https://en.wikipedia.org/wiki/Standard_normal_table`

Interpretation and Use of P-Value

- P-values have been widely used in scientific literature.

- However, the interpretation and use of p-values involve a lot of controversy.

- Ronald Fisher, the advocate of p-values, explains that [Fisher, 1934, Section 20]:

- If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested.
- If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts.

- Here "P" is the p-value, and

- "the hypothesis tested" is the null hypothesis H_0.

Interpretation and Use of *P*-Value

- The **American Statistical Association**'s Statement on *p*-Values [Wasserstein and Lazar, 2016] explains that

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based **only on** whether a *p*-value passes a specific threshold.

Interpretation and Use of P-Value

4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

- The statement concludes that "*No single index should substitute for scientific reasoning*."

- See also e.g. [Berger and Sellke, 1987, McShane et al., 2019] and references therein.

Outline

1. Introduction: The Lady Tasting Tea Experiment
2. Procedure of Statistical Hypothesis Testing
3. Type 1 Error, Type 2 Error and the Power of a Test
4. Test Statistics
5. P-Value
6. Neyman-Pearson Lemma and Likelihood Ratio Test
7. Conclusions and Further Reading

What is the Most Powerful Test?

- So far we have not discussed how to construct a test statistic.

- A test statistic $T : \Omega \to \mathbb{R}$ and a critical region $I_\alpha \subset \mathbb{R}$ should be constructed so that

  - For a given $\alpha > 0$, the Type 1 Error probability is bounded by α
  $$P_{0,T}(I_\alpha) = P_0(T^{-1}(I_\alpha)) \leq \alpha.$$
  - The test power
  $$P_{1,T}(I_\alpha) = P_1(T^{-1}(I_\alpha))$$
  is as large as possible,

where P_0 and P_1 are the probability distributions on Ω under the null H_0 and alternative H_1 hypotheses, respectively.

- The question is how to construct a test statistic with a high test power.

What is the Most Powerful Test?

- One answer is provided by the Neyman-Pearson lemma [Neyman and Pearson, 1933].

- This lemma states that the likelihood ratio test statistic provides the most powerful test.

Likelihood Ratio Test

- Let P_0 and P_1 be the probability distributions on Ω under the null H_0 and alternative H_1 hypotheses, respectively.

- Assume P_0 and P_1 have density functions

$$p_0 : \Omega \to [0, \infty), \quad p_1 : \Omega \to [0, \infty)$$

with respect to a base measure ν (e.g., ν is the Lebesgue measure when $\Omega \subset \mathbb{R}^n$.)

- i.e., for any measurable subset $S \subset \Omega$, we have

$$P_0(S) = \int_S p_0(\omega) d\nu(\omega), \quad P_1(S) = \int_S p_1(\omega) d\nu(\omega).$$

Likelihood Ratio Test

- Define a test statistic $T : \Omega \to [0, \infty)$ by

$$T(\omega) := \frac{p_1(\omega)}{p_0(\omega)}, \quad \omega \in \Omega$$

- This is called the likelihood ratio test statistic.

- Define a test of the form

  - Reject the null hypothesis H_0, if $T(\omega_e) \geq c_\alpha$;
  - Not reject the null hypothesis H_0, if $T(\omega_e) < c_\alpha$,

where $c_\alpha \geq 0$ is defined so the Type 1 Error probability becomes $\alpha > 0$.

i.e., we define the critical region I_α for the test statistic T as

$$I_\alpha = [c_\alpha, \infty).$$

Neyman-Pearson Lemma

- The Neyman-Pearson Lemma states that

The likelihood ratio test is the most powerful test among all tests with the significance level α.

Neyman-Pearson Lemma

Neyman-Pearson Lemma [Neyman and Pearson, 1933]

- Define $\alpha > 0$ as the level of significance.
- Let $c_\alpha > 0$ be a constant such that the critical region defined by

$$S_\alpha^* := T^{-1}([c_\alpha, \infty)) = \left\{ \omega \in \Omega \mid T(\omega) := \frac{p_1(\omega)}{p_0(\omega)} \geq c_\alpha \right\}$$

  satisfies

$$P_{0,T}([c_\alpha, \infty)) := P_0(S_\alpha^*) = \alpha.$$

- Then the test based on S_α^* has the highest power among all tests with the significance level α;
- i.e., for all $S_\alpha \subset \Omega$ such that $P_0(S_\alpha) = \alpha$, we have

$$P_1(S_\alpha^*) \geq P_1(S_\alpha).$$

Neyman-Pearson Lemma: Proof

- Since $S_\alpha^* \cap S_\alpha \subset S_\alpha^*$, we have

$$P_0(S_\alpha^* \backslash (S_\alpha^* \cap S_\alpha)) = P_0(S_\alpha^*) - P_0(S_\alpha^* \cap S_\alpha) = \alpha - P_0(S_\alpha^* \cap S_\alpha).$$

- Similarly, since $S_\alpha^* \cap S_\alpha \subset S_\alpha$, we have

$$P_0(S_\alpha \backslash (S_\alpha^* \cap S_\alpha)) = P_0(S_\alpha) - P_0(S_\alpha^* \cap S_\alpha) = \alpha - P_0(S_\alpha^* \cap S_\alpha).$$

- Therefore

$$P_0(S_\alpha^* \backslash (S_\alpha^* \cap S_\alpha)) = P_0(S_\alpha \backslash (S_\alpha^* \cap S_\alpha)).$$

Neyman-Pearson Lemma: Proof

- Recall that

$$\frac{p_1(\omega)}{p_0(\omega)} \geq c_\alpha, \quad \forall \omega \in S_\alpha^*, \qquad \frac{p_1(\omega)}{p_0(\omega)} < c_\alpha, \quad \forall \omega \in \Omega \backslash S_\alpha^*$$

- Therefore,

$$p_1(\omega) \geq c_\alpha p_0(\omega), \quad \forall \omega \in S_\alpha^*.$$

- Thus, for any subset $S \subset S_\alpha^*$, we have

$$P_1(S) = \int_S p_1(\omega) d\nu(\omega) \geq \int_S c_\alpha p_0(\omega) d\nu(\omega) = c_\alpha P_0(S).$$

- On the other hand,

$$p_1(\omega) < c_\alpha p_0(\omega), \quad \forall \omega \in \Omega \backslash S_\alpha^*.$$

- Thus, for all $S' \subset \Omega \backslash S_\alpha^*$,

$$P_1(S') = \int_{S'} p_1(\omega) d\nu(\omega) < \int_{S'} c_\alpha p_0(\omega) d\nu(\omega) = c_\alpha P_0(S').$$

Neyman-Pearson Lemma: Proof

- Since

$$S := S_\alpha^* \backslash (S_\alpha^* \cap S_\alpha) \subset S_\alpha^*, \quad S' := S_\alpha \backslash (S_\alpha^* \cap S_\alpha) \subset \Omega \backslash S_\alpha^*,$$

and since

$$P_0(S_\alpha^* \backslash (S_\alpha^* \cap S_\alpha)) = P_0(S_\alpha \backslash (S_\alpha^* \cap S_\alpha)),$$

we have

$$\begin{aligned} P_1(S_\alpha^* \backslash (S_\alpha^* \cap S_\alpha)) &\geq c_\alpha P_0(S_\alpha^* \backslash (S_\alpha^* \cap S_\alpha)) \\ &= c_\alpha P_0(S_\alpha \backslash (S_\alpha^* \cap S_\alpha)) > P_1(S_\alpha \backslash (S_\alpha^* \cap S_\alpha)). \end{aligned}$$

Therefore

$$\begin{aligned} P_1(S_\alpha^*) &= P_1(S_\alpha^* \backslash (S_\alpha^* \cap S_\alpha)) + P_1((S_\alpha^* \cap S_\alpha)) \\ &> P_1(S_\alpha \backslash (S_\alpha^* \cap S_\alpha)) + P_1((S_\alpha^* \cap S_\alpha)) = P_1(S_\alpha). \end{aligned}$$

Thus the proof completes. $\square$

Example: Testing the Location of a Gaussian Mean

- Consider again testing the location of a Gaussian mean.

- Let p^* be an unknown probability density function on $\mathbb{R}$.

- Assume that we know/believe that p^* is Gaussian, with unknown mean $\mu \in \mathbb{R}$ and known variance $\sigma^2 > 0$:

$$p^*(x) = p_{\mathrm{gauss}}(x; \mu, \sigma_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

- Assume that we can perform an experiment to obtain an i.i.d. sample of size n from p^*:

$$x_1, \ldots, x_n \in \mathbb{R}$$

- Assume that we are interested in testing whether the unknown mean μ is equal to some specified value $\mu_0 \in \mathbb{R}$ or not.

Example: Testing the Location of a Gaussian Mean

- Thus, the null hypothesis H_0 is defined as

$$H_0 : \mu = \mu_0.$$

- For simplicity, we consider a simple alternative hypothesis H_1 where the unknown mean μ is another specified value $\mu_1 \neq \mu_0$:

$$H_1 : \mu = \mu_1.$$

Example: Testing the Location of a Gaussian Mean

- We can define the sample space Ω as

$$\Omega := \mathbb{R}^n.$$

- Each $\omega := (x_1, \ldots, x_n) \in \Omega$ represents a possible experiment outcome of n i.i.d. observations.

- Thus, the distribution P_0 on Ω under the null hypothesis H_0 is given by the density function $p_0 : \Omega \to \mathbb{R}$:

$$p_0(\omega) = \prod_{i=1}^{n} p_{\text{gauss}}(x_i; \mu_0, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_0)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{2\sigma^2}\right), \quad \omega := (x_1, \ldots, x_n) \in \Omega.$$

Example: Testing the Location of a Gaussian Mean

- Similarly, the density function p_1 of P_1 under the alternative is given by, for $\omega := (x_1, \dots, x_n)$,

$$p_1(\omega) = \prod_{i=1}^{n} p_{\text{gauss}}(x_i; \mu_1, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu_1)^2}{2\sigma^2})$$

- The likelihood ratio test statistic is thus given by, for $\omega := (x_1, \dots, x_n)$,

$$\begin{aligned} T(\omega) &:= \frac{p_1(\omega)}{p_0(\omega)} = \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu_1)^2 - \sum_{i=1}^{n}(x_i - \mu_0)^2}{2\sigma^2}) \\ &= \exp(-\frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\mu_1 + \mu_1^2) - \sum_{i=1}^{n}(x_i^2 - 2x_i\mu_0 + \mu_0^2)}{2\sigma^2}) \\ &= \exp(\frac{2(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i - n(\mu_1^2 - \mu_0^2)}{2\sigma^2}) \end{aligned}$$

Example: Testing the Location of a Gaussian Mean

- Therefore, the test is given by the critical region determined by the threshold

$$\exp\left(\frac{2(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i - n(\mu_1^2 - \mu_0^2)}{2\sigma^2}\right) \geq c_\alpha$$

where $c_\alpha \geq 0$ is such that we have $P_0(S_\alpha) = \alpha$ for the critical region

$$S_\alpha := \{\omega := (x_1, \ldots, x_n) \in \mathbb{R} \mid T(\omega) \geq c_\alpha\}$$

- Taking the logarithm in the both sides, we have

$$\frac{2(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i - n(\mu_1^2 - \mu_0^2)}{2\sigma^2} \geq \log(c_\alpha)$$

$$\iff \quad (\mu_1 - \mu_0)\frac{1}{n}\sum_{i=1}^{n} x_i \geq \frac{1}{2}\left(2\sigma^2 \log(c_\alpha) + (\mu_1^2 - \mu_0^2)\right)$$

Example: Testing the Location of a Gaussian Mean

$$(\mu_1 - \mu_0)\frac{1}{n}\sum_{i=1}^{n} x_i \geq \frac{1}{2}\left(2\sigma^2 \log(c_\alpha) + (\mu_1^2 - \mu_0^2)\right)$$

-Thus, if $(\mu_1 - \mu_0) > 0$ (i.e., $\mu_1 > \mu_0$), the rejection threshold is given by

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \frac{1}{2(\mu_1 - \mu_0)}\left(2\sigma^2 \log(c_\alpha) + (\mu_1^2 - \mu_0^2)\right) =: r_\alpha$$

- If $(\mu_1 - \mu_0) < 0$ (i.e., $\mu_1 < \mu_0$), the rejection threshold is given by

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \frac{1}{2(\mu_1 - \mu_0)}\left(2\sigma^2 \log(c_\alpha) + (\mu_1^2 - \mu_0^2)\right) =: \ell_\alpha$$

Example: Testing the Location of a Gaussian Mean

- Note that, under the null H_0 where $x_1, \ldots, x_n \sim p_{\mathrm{gauss}}(t; \mu_0, \sigma)$ (i.i.d.), we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i \sim p_{\mathrm{gauss}}(t; \mu_0, \sigma^2/n).$$

- Thus, we can derive the rejection threshold r_α

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq r_\alpha$$

directly as r_α satisfying

$$\text{Type 1 Error Probability} = \int_{r_\alpha}^{\infty} p_{\mathrm{gauss}}(t; \mu_0, \sigma^2/n) dt = \alpha.$$

- This shows that the rejection threshold r_α does not depends on the value of μ_1, as long as $\mu_1 > \mu_0$.

Example: Testing the Location of a Gaussian Mean

- This means that the likelihood ratio test is the uniformly most powerful for a composite alternative hypothesis

$$H_1 : \mu > \mu_0$$

- Similarly, if $\mu_1 < \mu_0$ we can derive the threshold ℓ_α as the one satisfying

$$\text{Type 1 Error Probability} = \int_{-\infty}^{\ell_\alpha} p_{\text{gauss}}(t; \mu_0, \sigma^2) dt = \alpha.$$

- This shows that the rejection threshold ℓ_α does not depends on the value of μ_1, as long as $\mu_1 < \mu_0$

- This means that the likelihood ratio test is the uniformly most powerful for a composite alternative hypothesis

$$H_1 : \mu < \mu_0$$

Example: Testing the Location of a Gaussian Mean

- However, this shows that there does not exist a uniformly most powerful test for a composite alternative hypothesis $H_1 : \mu \neq \mu_0$, i.e.,

$$H_1 : \mu < \mu_0 \quad \text{or} \quad \mu_0 < \mu$$

- This is because, when the true unknown mean μ satisfies $\mu > \mu_0$, then the test based on the right rejection threshold

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq r_\alpha$$

is the most powerful,

- while when the true unknown mean μ satisfies $\mu < \mu_0$, then the test based on the left rejection threshold

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \ell_\alpha$$

becomes the most powerful.

Important Points to Remember

- The likelihood ratio test depends on how we define an alternative hypothesis.

  - This is true for any test, because the test power (or the Type 2 error) is defined for a given alternative hypothesis.

- For a composite alternative hypothesis (where the alternative contains a variable parameter), there might be no uniformly most powerful test.

- Anyway, the likelihood ratio test and the Neyman-Pearson lemma provides a guideline to design a powerful test.

Outline

1. Introduction: The Lady Tasting Tea Experiment
2. Procedure of Statistical Hypothesis Testing
3. Type 1 Error, Type 2 Error and the Power of a Test
4. Test Statistics
5. *P*-Value
6. Neyman-Pearson Lemma and Likelihood Ratio Test
7. Conclusions and Further Reading

Some Key Points to Remember

- To design a test, we need to specify the distribution P_0 on the space Ω of experiment outcomes (or data) under the null hypothesis H_0.

- We should be careful that P_0 may be misspecified.

- For instance, consider the example of testing the location of a Gaussian mean.

- We assumed that the data $\omega = (x_1, \ldots, x_n)$ are i.i.d. with a Gaussian distribution with known variance $\sigma^2 > 0$.

  - The knowledge of the variance $\sigma^2 > 0$ is not available in practice, and we need to estimate it from data.
  - This requires modifying the testing procedure, and results in the Student t-test.

Some Key Points to Remember

- More generally, the Gaussian assumption itself may be misspecified.

- Under such a misspecification, the Type 1 Error probability

$$P_{0,T}(I_\alpha) = P_0(T^{-1}I_\alpha)$$

may be deviated from a desired level α of significance.

- Thus, in general we should define a null hypothesis H_0 with a weaker assumption about the data distribution P_0.

Some Key Points to Remember

- To derive a critical region $I_\alpha \subset \mathbb{R}$, we need to be able to calculate the probability of I_α under the null H_0

$$P_{0,T}(I_\alpha) = P_0(T^{-1}(I_\alpha)).$$

- This may not be easy in general, in particular when we pose a less restrictive assumption about P_0.

- A modern approach to this purpose is the bootstrap method, developed by Bradley Efron (See [Efron and Hastie, 2016, Section 10]).

  - This method uses Monte Carlo (or simulations) to approximate the distribution $P_{0,T}$ under the null.
  - The approach can be used for a wide range of problems and easy to implement.

Further Reading

- Again, I recommend you to have a look at [Rao, 1973].

- The following are recommendations for further reading.

Introduction to Hypothesis Testing and Design of Experiments

[Fisher, 1934, Fisher, 1937]

Introduction to the Neyman-Pearson Theory (or the Frequentist Theory)

[Neyman and Pearson, 1933]

About the Conflicts between the Fisher and Neyman-Pearson Theories

[Lehmann, 1993] [Efron and Hastie, 2016, Sections 2 and 4]

Further Reading

P-values and Statistical Significance

[Berger and Sellke, 1987] [Wasserstein and Lazar, 2016]
[McShane et al., 2019]

Connections between the Likelihood Ratio Test and the KL Divergence

[Rao, 1973, Section 7a. 3] [Eguchi and Copas, 2006]

Berger, J. O. and Sellke, T. (1987).
Testing a point null hypothesis: The irreconcilability of p values and evidence.
Journal of the American Statistical Association, 82(397):112–122.

Efron, B. and Hastie, T. (2016).
Computer Age Statistical Inference.
Cambridge University Press.

Eguchi, S. and Copas, J. (2006).
Interpreting kullback–leibler divergence with the neyman–pearson lemma.
Journal of Multivariate Analysis, 97(9):2034–2040.

Fisher, R. A. (1934).
Statistical Methods for Research Workers (Fifth Edition).
Oliver & Boyd (Edinburgh).

Fisher, R. A. (1937).
Design of Experiments (second edition).

Macmillan.

Lehmann, E. L. (1993).
The fisher, neyman-pearson theories of testing hypotheses: one theory or two?
Journal of the American statistical Association, 88(424):1242–1249.

Lehmann, E. L. and Romano, J. P. (2005).
Testing Statistical Hypotheses (Third Edition).
Springer Science & Business Media.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019).
Abandon statistical significance.
The American Statistician, 73(sup1):235–245.

Neyman, J. and Pearson, E. S. (1933).
On the problem of the most efficient tests of statistical hypotheses.

Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694-706):289–337.

Rao, C. R. (1973).
Linear Statistical Inference and Its Applications.
Wiley New York.

Wasserstein, R. L. and Lazar, N. A. (2016).
The ASA Statement on p-Values: Context, Process, and Purpose.
The American Statistician, 70(2).