Confidence intervals and Hypothesis testing

Kang Li

StatB/E
Week 4, Wed

Contents

Contents

Confidence intervals: motivation

Previously we learned estimators, in particular the maximum likelihood estimator $\hat{\theta}(x)$, that takes the observation x and gives an estimate for the parameter θ.

But how close is the estimate to the true parameter?

Confidence intervals are methods to evaluate the uncertainty of estimates.

For example, along with the estimate $\hat{\theta}(x)$ we also provide an interval $I(x)$ to show how confident we are that the true parameter is covered by $I(x)$.

A narrow $I(x)$ means a more accurate estimate; a wide $I(x)$ implies more uncertainty.

Confidence intervals

In our statistical model, the random variable X denotes the measurements and x is the observation (a realization of X).

The true parameter in the model is denoted by θ^* and it is unknown. Recall we use MLE $\hat{\theta}(x)$ to estimate the unknown θ^*.

Given the observation $x \in E$, we want to construct an interval $I(x)$ such that

$$P(\theta^* \in I(X)) \geq \alpha.$$

We call the interval a α**-confidence interval** for θ^*.

A common choice of α is 0.95, and we will have a 95%-confidence interval.

Remark

- Confidence intervals are random - view it as a function taking the random variable X, $I(X)$
- Given an observation x, the constructed confidence interval $I(x)$ is not random - it is a realization of $I(X)$.
- We cannot say there is α probability that $I(x)$ covers θ^*.
- The probability that $I(X)$ covers θ^* is at least α:

$$P(\theta^* \in I(X)) \geq \alpha.$$

- Compare with this example: say we have a random variable $Y \in \mathbb{R}$ and $P(Y > 0) = 0.8$. We now have a realization $y = 1$.
  - We cannot say the probability of $y = 1 > 0$ is 0.8, but it is true that the probability of $Y > 0$ is 0.8.

Example: Standard error of the mean

Consider n IID random variables

$$X = \{X_1, \ldots, X_n\}$$

following a normal distribution $N(\mu, \sigma)$. We have a realization (observation) of X, denoted by

$$x = \{x_1, \ldots, x_n\}.$$

We have know from the previous lecture that the MLE for μ is given by the sample mean $\bar{x}$:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Here let us construct a 95%-confidence interval for μ.

Example: Standard error of the mean

Recall the following result we verified (by simulation) in an exercise.

Result

Let $\overline{X}$ denote the sample mean of X and $\widehat{SEM}$ the sample standard error of the mean. The statistic given by

$$t = \frac{\overline{X} - \mu}{\widehat{SEM}}$$

follows a t-distribution with $n - 1$ degrees of freedom.

The t-distribution is a symmetric distribution around 0. Thus, we have

$$P(z_{0.025} \leq t \leq z_{0.975}) = 0.95.$$

where $z_{0.025}$ and $z_{0.975}$ denote the 0.025 and 0.975-quantiles of the t-distribution with $n - 1$ degrees of freedom.

Example: Standard error of the mean

We have:

$$P(z_{0.025} \leq t \leq z_{0.975}) = 0.95.$$

Therefore,

$$P(z_{0.025} \leq \frac{\overline{X} - \mu}{\widehat{SEM}} \leq z_{0.975}) = 0.95$$

$$P(z_{0.025}\widehat{SEM} - \overline{X} \leq -\mu \leq z_{0.975}\widehat{SEM} - \overline{X}) = 0.95$$

$$P(\overline{X} - z_{0.975}\widehat{SEM} \leq \mu \leq \overline{X} - z_{0.025}\widehat{SEM}) = 0.95$$

$$P(\overline{X} - z_{0.975}\widehat{SEM} \leq \mu \leq \overline{X} + z_{0.975}\widehat{SEM}) = 0.95$$

The 95%-confidence interval for μ is given by

$$I(X) = [\overline{X} - z_{0.975}\widehat{SEM}, \quad \overline{X} + z_{0.975}\widehat{SEM}]$$

Example: Standard error of the mean

The 95%-confidence interval for μ is given by

$$I(X) = [\overline{X} - z_{0.975}\widehat{SEM}, \quad \overline{X} + z_{0.975}\widehat{SEM}]$$

Given a data set x, which is a realization of X, the realization of $I(X)$ is then

$$I(x) = [\overline{x} - z_{0.975}\widehat{sem}, \quad \overline{x} + z_{0.975}\widehat{sem}],$$

where $\overline{x}$ and $\widehat{sem}$ are realizations of $\overline{X}$ and $\widehat{SEM}$, evaluated given x.

Remark

Let us emphasize this again:

- We cannot say the probability of the deterministic interval

$$I(x) = [\overline{x} - z_{0.975}\widehat{sem}, \quad \overline{x} + z_{0.975}\widehat{sem}],$$

  covering μ is 95%. Once $I(x)$ is constructed, it either contains x or not.

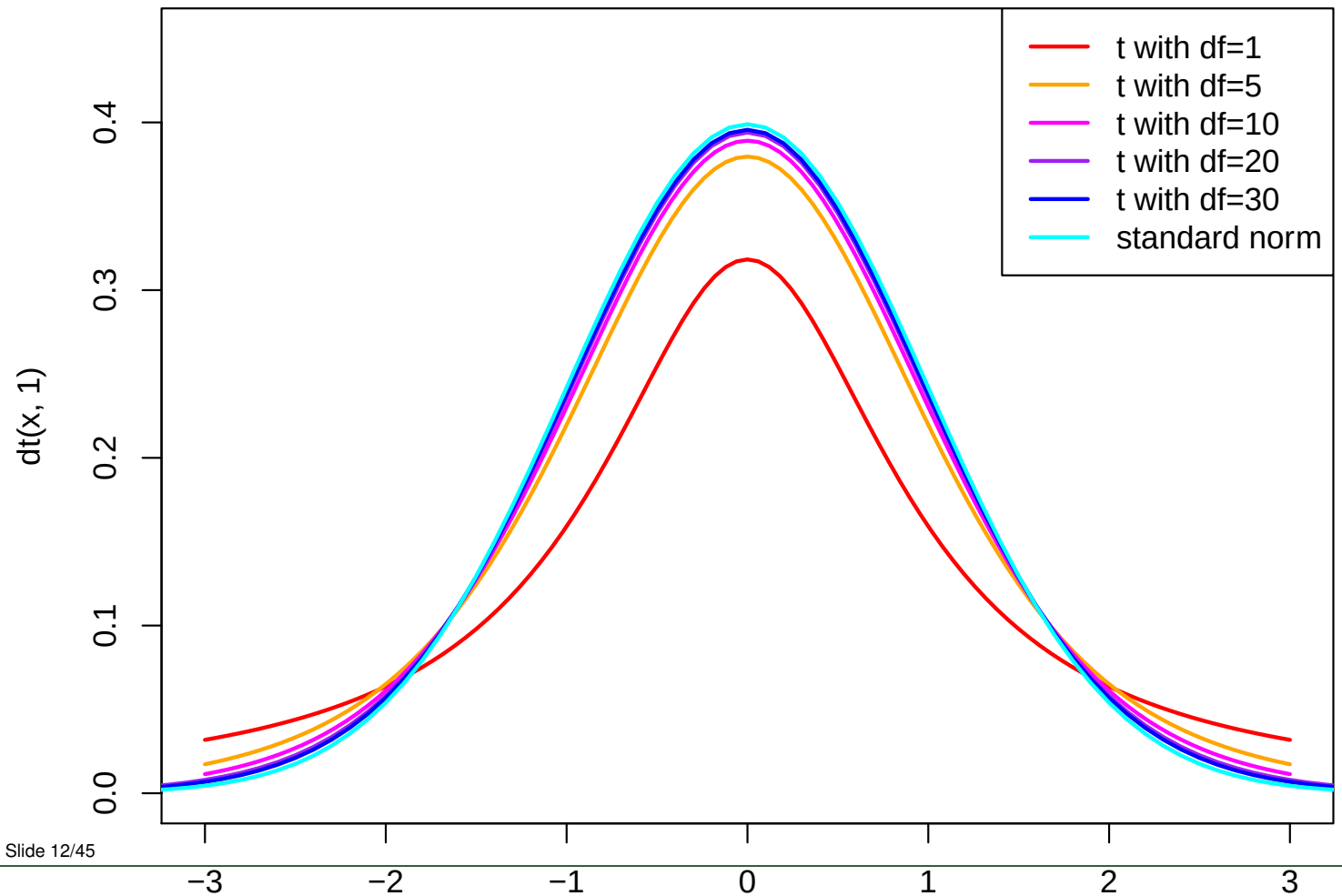
- We can indeed say the probability of the random interval

$$I(X) = [\overline{X} - z_{0.975}\widehat{SEM}, \quad \overline{X} + z_{0.975}\widehat{SEM}]$$

  covering μ is 95%.

More about the t-distribution

Convergence of the t-distribution with the degrees of freedom n to the standard normal distribution $N(0,1)$ as $n \to \infty$.



More about the t-distribution

$N(0,1)$ is considered a sufficient approximate to the t-distribution if the degrees of freedom is greater than 20.

For the 95%-confidence interval of μ in our previous example, if the sample size is larget than 20, we can **approximate** the confidence interval by

$$I(X) = [\overline{X} - q_{0.975}\widehat{SEM}, \quad \overline{X} + q_{0.975}\widehat{SEM}]$$

where $q_{0.975} \approx 1.96$ is the 0.975-quantile of $N(0,1)$.

Confidence interval for general MLE

Not required, but good to know:

Suppose $\hat{\theta}(x)$ is the maximum likelihood estimate for θ, the standard deviation (or standard error) of $\theta(X)$ can be estimated as

$$1/\sqrt{i(\hat{\theta}(x))},$$

where

$$i(\hat{\theta}(x)) = \frac{d^2 l_x(\hat{\theta}(x))}{d\theta^2}$$

is called the **observed Fisher information**.

The confidence interval for θ can be estimated as

$$[\hat{\theta}(x) - q_\alpha/\sqrt{i(\hat{\theta}(x))}, \quad \hat{\theta}(x) + q_\alpha/\sqrt{i(\hat{\theta}(x))}]$$

using normal quantiles q_α.

Contents

Hypothesis testing: motivation

In many cases, we need to say *yes* or *no* to a scientific hypothesis.

Can we say that two given groups of experimental measurements have the mean?

For the neural ISI data, is it fine to use the exponential model instead of the gamma model?

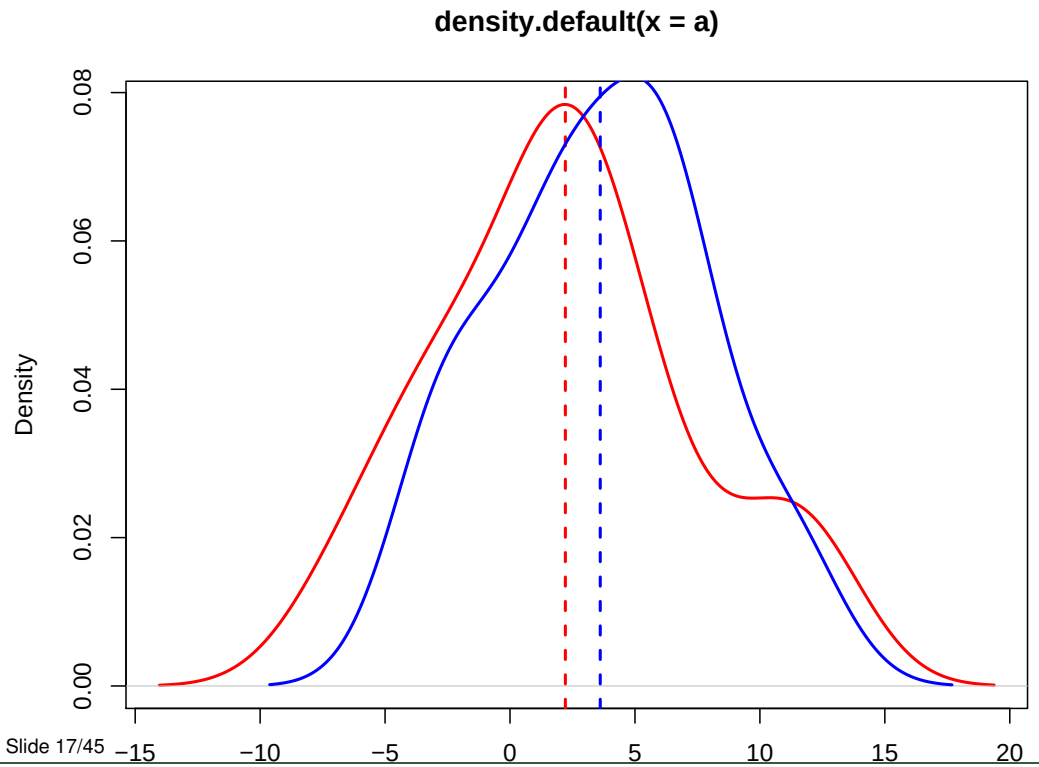
We want a hypothesis testing procedure to get the answer (yes or no).

Tell if two data sets have the same mean

Given two data sets:

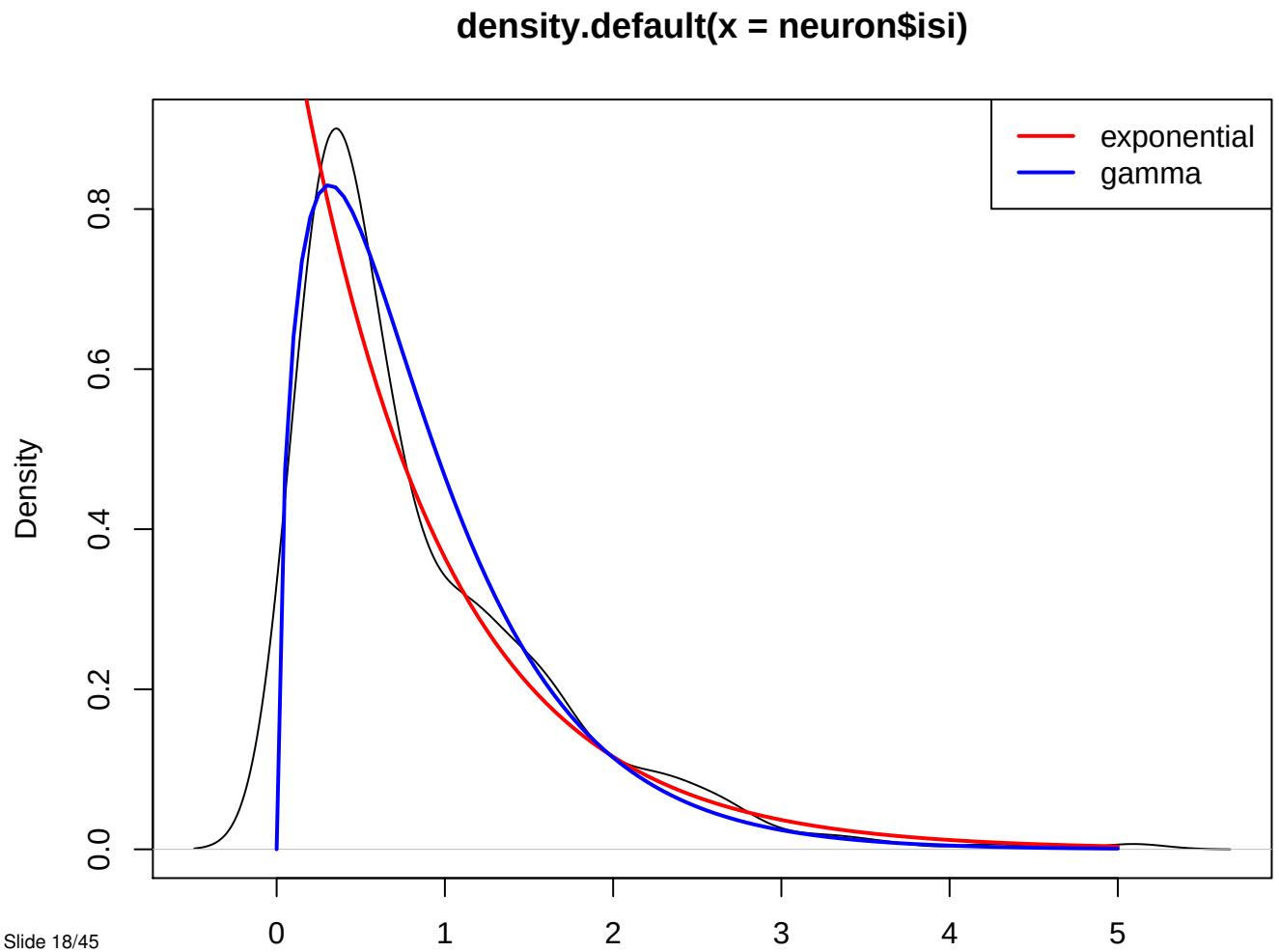
```
x1 = 4.9621214, -1.5449573, -5.0876427,  4.2827459, ...
x2 = -3.400786,  8.0611346,  5.3715170,  2.4292400, ...
```

We can plot their kernel density estimation:



Neural ISI data

Do we really need a more complicated gamma model with two parameters, or is the simpler exponential model with one parameter sufficient?



The essence of hypothesis testing

We have two statistical models, $P_{\theta,\theta\in\Theta_0}$ and $P_{\theta,\theta\in\Theta}$, with the former being a special case (nested model) of the latter.

The parameter space of the nested model, Θ_0, is a subset of the parameter space Θ of the other complete model:

$$\Theta_0 \subseteq \Theta.$$

Our **hypothesis** is $\theta \in \Theta_0$, i.e. the nested model is sufficient to describe the data. We call it the **null hypothesis** and write:

$$H_0 : \theta \in \Theta_0.$$

The **alternative hypothesis** is that we have to use the full parameter space Θ to describe the data.
A **statistical test** tells whether to **reject** or **accept** the null hypothesis.

Examples

For the example with two data sets where we want to decide whether they have equal mean

- Our complete model could be that the two groups follow two normal distributions with different means, $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, with a total of three parameters $\{\mu_1, \mu_2, \sigma\}$.
- The nested model is that the two means are equal, $\mu_1 = \mu_2$, such that all data follow the same distribution $N(\mu, \sigma)$, with two parameters $\{\mu, \sigma\}$.
- The null hypothesis is then $H_0 : \mu_1 = \mu_2$.
- We ask for a statistical test which tells if we should reject or accept H_0, i.e. if the nested model is sufficient.

Examples

For the neural ISI data,

- The complete model is a gamma distribution with parameters α, β.
- The nested model is an exponential distribution with parameter λ.
  Note that an exponential distribution with parameter λ is equivalent to a gamma distribution with parameters $\alpha = 1, \beta = \lambda$.
- The null hypothesis is $H_0 : \alpha = 1$.
- We again ask for a statistical test to tell whether to reject H_0.

The essence of hypothesis testing (continued)

We use a statistic $t(X)$, a random variable computed from the data X.

Under the null hypothesis H_0, we derive the theoretical distribution of $t(X)$.

If the null hypothesis is true, then $t(X)$ follows this distribution.

We check how well the realization of $t(X)$ for observation x, i.e. $t(x)$, follows the theoretical distribution of $t(X)$.

Examples

Assume the two groups of data X_1 and X_2 have the size of n and m, respectively. The empirical means are $\overline{X}_1$ and $\overline{X}_2$. The empirical standard deviation is $\hat{\sigma}$ (we assume equal variance).

- We use the statistic

$$t(X) = \frac{\sqrt{\frac{n+m}{nm}}(\overline{X}_1 - \overline{X}_2)}{\hat{\sigma}}.$$

- A theoretical result states that, under H_0, $t(X)$ follows a t-distribution with degrees of freedom $n+m-2$.
- We compute the realization $t(x)$ using observation x, and check if it follows the t-distribution. But How?

The essence of hypothesis testing (continued)

To check the observed statistic $t(X)$, we evaluate how ”extreme” it is.

We calculate the probability of observing a statistic more or as extreme as the currently observed $t(x)$. We call this probability the **p-value**.

We define a rejection level α, such that if the p-value is less than α we say the p-value is **significant** and we reject the null hypothesis. The statistical test is then called a **α-level test**.

Intuition: If the p-value is a very small value, say 0.02, that means the probability of observing a statistic more extreme than the currently observed statistic is 0.02. $\rightarrow$ The currently observed one is extreme enough $\rightarrow$ There must be something wrong! $\rightarrow$ The null hypothesis should be rejected!

p-value

p-value: The probability of observing a statistic $t(x)$ more or as extreme as the currently observed $t(x)$ computed from the observation x.

Two-sided test: If the theoretical distribution for $t(X)$ under the null hypothesis is symmetric around 0 and large absolute values are extreme, e.g. the t-distribution, then the p-value is given by

$$\begin{aligned} & P(t(X) \geq |t(x)| \cup t(X) \leq -|t(x)|) \\ =& F(-|t(x)|) + 1 - F(|t(x)|) \\ =& 2F(-|t(x)|). \end{aligned}$$

One-sided test: If the theoretical distribution for $t(X)$ under the null hypothesis is asymmetric and large values are extreme, then the p-value is given by

$$P(t(X) \geq t(x)) = 1 - F(t(x)).$$

Summarizing hypothesis testing

- Two statistical models, one being a nested model of the other.
- A null hypothesis H_0 stating the nested model is sufficient.
- A test statistic $t(X)$, a quantity computed from data.
- A good approximation of the theoretical distribution of $t(X)$ under the null hypothesis.
- A p-value telling how extreme the observed statistic $t(x)$ is.
- A significance level to reject the null hypothesis.

Hypothesis testing is essentially a method for **model selection** in statistical inference, telling if a nested, special model is sufficient for a complete model.

Common statistical tests

- One sample t-test
- Two sample t-test
- Pearson's Chi-squared test
- Likelihood ratio test

One-sample t-test

Give a data set $x = \{x_1, \ldots, x_n\}$, we want to test if the mean equals a certain value μ_0.

We consider n IID random variables $X = \{X_1, \ldots, X_n\}$ following a normal distribution, and x is a realization of X.

The full statistical model states X follows a normal distribution with unknown mean, $N(\mu, \sigma)$, with two parameters.

The nested model states X follows a normal distribution whose mean μ_0 is known, $N(\mu_0, \sigma)$, with one parameter.

The null hypothesis is $H_0 : \mu = \mu_0$.

One-sample t-test

Under the null hypothesis $\mu = \mu_0$, we have the following result.

Result

The statistic

$$t(X) = \frac{\overline{X} - \mu_0}{\widehat{SEM}}$$

follows a t-distribution with $n - 1$ degrees of freedom.

Given observation $x = \{x_1, \ldots, x_n\}$, the p-value for the one-sample t-test is

$$2F(-|t(x)|),$$

where F is the CDF of the t-distribution with $n - 1$ degrees of freedom.

Two-sample t-test with equal variance

Give two data sets $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_m\}$, we want to test if the two data sets have the same mean.

We consider n IID random variables $X = \{X_1, \ldots, X_n\}$ following a normal distribution, and m IID random variables $Y = \{Y_1, \ldots, Y_m\}$ also following a normal distribution. x and y are realizations of X and Y.

The full statistical model states X and Y follow two normal distributions with unequal mean but equal variance, $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, with three parameters.

The nested model states X and Y follow the same normal distribution $N(\mu, \sigma)$, with two parameters.

The null hypothesis is $H_0 : \mu_1 = \mu_2$.

Two-sample t-test

Under the null hypothesis $\mu_1 = \mu_2$, we have the following result.

Result

The statistic

$$t(X,Y) = \frac{\sqrt{\frac{n+m}{nm}}(\overline{X} - \overline{Y})}{\hat{\sigma}}.$$

follows a t-distribution with degrees of freedom $n+m-2$.

Given observation $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$, the p-value for the two-sample t-test with equal variance is

$$2F(-|t(x,y)|),$$

where F is the CDF of the t-distribution with $n+m-2$ degrees of freedom.

t-tests in R

```
# simulate data
x = rnorm(20, 2, 4)
y = rnorm(40, 2.5, 4.5)

# one-sample t-test for H0: mu=4
t.test(x, mu=4)
# one-sample t-test for H0: mu=5
t.test(x, mu=5)

# two-sample t-test for H0: mu1=mu2
t.test(x, y, var.equal=TRUE)
```

Pearson's Chi-squared (χ^2) test

The χ^2 test checks whether or not two groups of data follow the same frequency pattern. For example, we have the following contingency table showing experimental results of two loaded dice:

	Die 1	Die 2
1	23	33
2	26	27
3	13	23
4	23	39
5	20	30
6	31	39

Do the two loaded dice have the same frequency pattern?

Pearson's Chi-squared (χ^2) test

The full model consists of two empirical distributions for the two groups, with $(6-1) \times 2$ parameters representing probabilities of the two dice.

The nested model consists of one distribution for both groups, with $(6-1)$ parameters.

The null hypothesis is that one distribution is sufficient to explain both groups.

A statistic is computed from the data, which, under the null hypothesis, follows a χ^2 distribution. (Details omitted here)

Pearson's Chi-squared (χ^2) test in R

```
# simulate data from binomial distribution
# x and y have different frequency patterns
x = rbinom(3, 100, c(0.2, 0.3, 0.4))
y = rbinom(3, 50, c(0.5, 0.3, 0.5))

# chisq test
# takes a matrix as the argument
chisq.test(cbind(x,y))
```

Likelihood ratio test

A generic statistical test to compare two statistical models.

We have two statistical models with parameter space Θ_0 and Θ, the former being a nested model of the latter.

The null hypothesis, $H_0 : \theta \in \Theta_0$, states that the nested model is sufficient.

The likelihood ratio test tells if the nested model is sufficient.

Likelihood ratio test

Given the data x, we conduct MLE for both models, and obtain the maximized likelihood value for the two models:

$$L_1 = \max_{\theta \in \Theta_0} L_x(\theta), \qquad L_2 = \max_{\theta \in \Theta} L_x(\theta).$$

They are given by the likelihood functions evaluated at the maximum likelihood estimates.

Likelihood ratio statistic

The statistic given by

$$q(x) = \frac{L_1}{L_2},$$

i.e. the quotient of the two maximized likelihood values, is called the likelihood ratio test statistic.

Since L_1 is from the nested model and therefore $L_1 \leq L_2$, we have $q(x) \in (0, 1]$. **Small** values of $q(x)$ are extreme.

Likelihood ratio test

Distribution of $q(X)$

Suppose Θ is d-dimensional (number of parameter is d) and Θ_0 is d_0-dimensional, then the distribution of

$$-2\log(q(X))$$

can be approximated by a χ^2 distribution with $d - d_0$ degrees of freedom. Large values are extreme.

Given a data set x, the p-value is

$$1 - F(-2\log(q(x))),$$

with F being the CDF of the corresponding χ^2 distribution.

Remark

- The likelihood ratio test is a generic statistical test that will work as long as we can perform maximum likelihood estimation for the two nested models.
- All other statistical tests can be performed using the likelihood ratio test.
- However, the χ^2 distribution for the likelihood ratio statistic is really an **approximate** result.
- Use likelihood ratio test if no other specific tests are available.

Likelihood ratio test and one-sample t-test

```
x = rnorm(20, 2, 4); mu = 3
# fit the nested model with one parameter
mllk = function(param) -sum(log(dnorm(x, 3, param)))
l1 = optimize(mllk, c(0, 10))$objective

# fit the full model with two parameters
mllk = function(params) -sum(log(dnorm(x, params[1],
    params[2])))
l2 = optim(c(1,1), mllk)$value

# p-value for the likelihood ratio test
1 - pchisq(2*(l1-l2), df=1)   # example: 0.66
# compare with the one-sample t-test
t.test(x, mu=mu)              # example: 0.76

# The difference comes from
#  1) Approximation of the likelihood ratio test.
#  2) The MLE for variance is biased.
```

Contents

Motivation

In hypothesis testing, we calculate the p-value, and reject the null hypothesis according to a significance level.

The p-value is a probability, which is used to answer a binary (yes or no) question.

As a rule of thumb, the significance level in scientific research is 0.05.

We need some methods to describe the accuracy of a statistical test.

Accuracy of a statistical test

	Reject H_0	Accept H_0
H_0 true	V	U
H_0 false	S	T

$V, S, U, T \in \mathbb{N}_0$ are numbers of statistical tests.

The situation of V when H_0 is true, i.e. rejecting a true null-hypothesis, is called the **type I error**.

The situation of T when H_0 is false, i.e. accepting a false null-hypothesis, is called the **type II error**.

We want to maximize S and U.

Power of a statistical test

- $P(\text{Reject } H_0 | H_0 \text{ is true}) = P(\text{type I error}) = \alpha$, the significance level of the statistical test, e.g. 0.05
- $P(\text{Accept } H_0 | H_0 \text{ is false}) = P(\text{type II error}) = \beta$, not easy to obtain for a general test
- $1 - \beta$, i.e. the probability of rejecting H_0 if H_0 is false, is called the **power** of a statistical test.
- The probability $P(H_0 \text{ is true} | \text{reject } H_0)$ is called the **false discovery rate**.

Contents