SM-4331 Advanced Statistics
Chapter 4 (Hypothesis Testing)

Dr Haziq Jamil
FOS M1.09

Universiti Brunei Darussalam

Semester II 2019/20

Outline

Introduction

Statistics provide the framework for discerning 'credible truth' by means of 'statistical significance'. That is, in the absence of absolute truth, rely on probabilistic statements conceived from observation of evidence (data) in order to make conclusions.

Motivating example

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. How do we know she's telling the truth?

Introduction (cont.)

- Set up an experiment where eight cups of milk tea were prepared, four of which had the milk poured in first, and the remaining had tea poured in first.
- Offer the cups of tea to the lady at random.
- She then tells us which four cups had the milk poured in first, and which four cups had the tea poured in first.
- Her possible answers are: 0, 1, 2, 3, or 4 correct (since the lady knows there are four of each type)

Introduction (cont.)

Of course, *whatever her answer is*, we can never know for certain whether she is telling the truth about her tea-tasting ability.

At best, *assuming* that she is completely guessing her answers, the probability that she gets

- **4 correct** is $1/\binom{8}{4} \approx 0.014$.
- **3 correct** is $\binom{4}{3}\binom{4}{1}/\binom{8}{4} \approx 0.229$.
- **2 correct** is $\binom{4}{2}\binom{4}{2}/\binom{8}{4} \approx 0.514$.
- **1 correct** is $\binom{4}{1}\binom{4}{3}/\binom{8}{4} \approx 0.229$.
- **0 correct** is one minus the sum of all the above probabilities, which is 0.014.

Introduction (cont.)

Let $X \in \{0, \ldots, 4\}$ denote the number of correct answers. Let's rephrase: What is $\mathrm{P}(X \geq x)$, the probability of observing a result *equal to, or better than*

- **0 correct**? This is equal to 1.
- **1 correct**? $\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X = 0) = 1 - 0.014 = 0.986$.
- **2 correct**? $\mathrm{P}(X \geq 2) = 1 - \mathrm{P}(X = 0, 1) = 0.472$.
- **3 correct**? $\mathrm{P}(X \geq 3) = 1 - \mathrm{P}(X = 0, 1, 2) = 0.243$.
- **4 correct**? $\mathrm{P}(X \geq 4) = 1 - \mathrm{P}(X = 0, 1, 2, 3) = 0.014$.

These are the so-called p-values for this experiment. We use these probabilities to gauge the likelihood that the lady's claim is true.

p-values

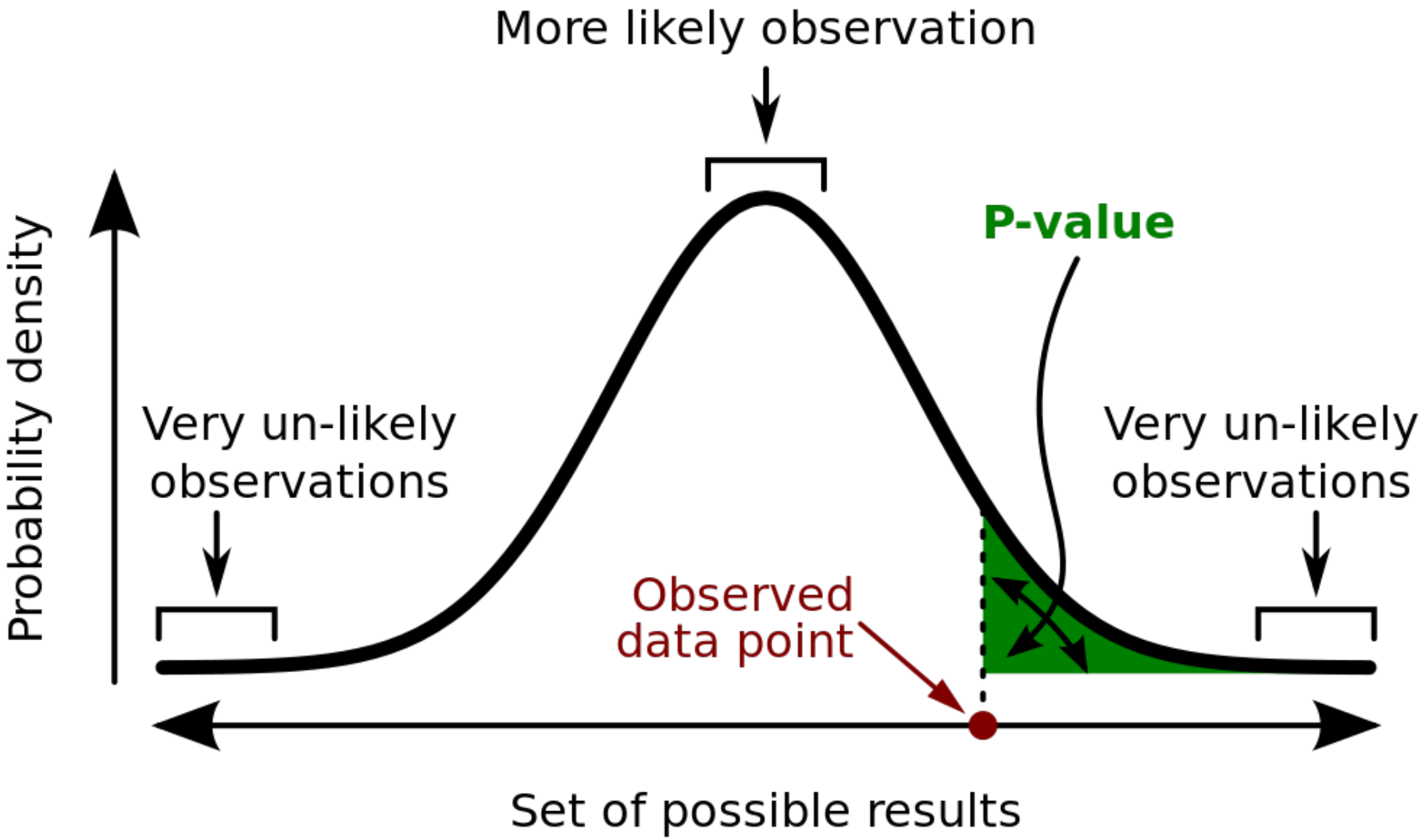
Definition 1 (p-values)

In statistical hypothesis testing, the p-value (or probability value) is the probability of obtaining test results *at least as extreme as* the results actually observed during the test, assuming that the null hypothesis is correct.

Depending how we interpret "a result (X) at least as extreme than the value observed (x)", we could mean

- $\{X \geq x\}$ (right-tail event);
- $\{X \leq x\}$ (left-tail event); or
- an event involving both left- and right-tail events.

p-values (cont.)



Significance levels

How do we then make a decision based on these p-values?

Definition 2 (Statistical significance)

In statistical hypothesis testing, a result has statistical significance when it is very unlikely to have occurred given the null hypothesis. More precisely, a study's defined significance level, denoted by α, is the probability of the study rejecting the null hypothesis, given that the null hypothesis were assumed to be true.

A result is deemed to be **statistically significant** when $p < \alpha$. The significance level for a study is chosen before data collection, and is typically set to 10%, 5%, or 1%.

General setting of a hypothesis test

Let $X_1, \ldots, X_n$ be a random sample from a distribution $f(x|\theta)$. We are interested in testing the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta \in \Theta_1$$

where θ_0 is a fixed value, Θ_1 is a set, and $\theta_0 \notin \Theta_1$. H_0 is called the **null hypothesis**, and H_1 is called the **alternative hypothesis**.

Based on the data, a statistical test is therefore making a binary decision whether or not to

- Reject H_0 (if p-value $\leq \alpha$), or to
- Not reject H_0 (if p-value $> \alpha$).

Remark

Not reject $\neq$ **Accept**. A statistical test is incapable to accept a hypothesis. A large p-value is indicative of a lack of evidence to reject the null hypothesis.

Statistical testing procedure

Step 1. Decide on the null and alternative hypotheses (H_0 and H_1), and fix a significance level α.

Step 2. Find a test statistic $T = T(X_1, \ldots, X_n)$. Denote T_0 the value of T with the given sample of observations.

Step 3. Compute the p-value, i.e.

$$p = \mathrm{P}(T = T_0 \text{ or more extreme values} | H_0)$$

where this probability is calculated assuming the null hypothesis is true (i.e. using the pdf $f(x|\theta_0)$). Here, "more extreme values" refers to those more unlikely values (than T_0) under H_0 in favour of H_1.

Step 4. If $p \leq \alpha$, reject H_0. Otherwise, H_0 is not rejected.

Example

Example 3

Let $X_1, \ldots, X_{20} \in \{T, H\}$ be the outcomes of an experiment of tossing a coin 20 times, i.e.

$$\mathrm{P}(X = H) = \pi = 1 - \mathrm{P}(X = T), \quad \pi \in (0, 1).$$

Let $Y = [X_1 = H] + \cdots + [X_{20} = H]$. Then $Y \sim \mathrm{Bin}(20, \pi)$, and an estimate of π is $\hat{\pi} = Y/20$. We would like to assess whether or not the hypothesis that "the coin is fair" is true. That is,

$$H_0 : \ \pi = 0.5 \quad \text{v.s.} \quad H_1 : \ \pi \neq 0.5$$

Example (cont.)

Example 3

Remark: **The answer cannot be resulted from the estimator $\hat{\pi}$, for**

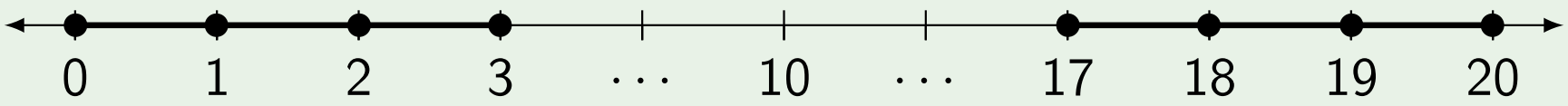
- if $\hat{\pi} = 0.9$, then H_0 is unlikely to be true.
- if $\hat{\pi} = 0.45$, then H_0 is may be true (and also may be untrue).
- if $\hat{\pi} = 0.7$, then what?

Furthermore, realise that $\hat{\pi} = \bar{X}$ is a random variable, and therefore subject to random fluctuation among different samples. We must cast this into a statistical testing framework.

Example (cont.)

Example 3

Let Y be the test statistic, and suppose we observe $Y = 17$. What are the more extreme values for Y if H_0 is true? Under H_0, $\mathrm{E}(Y) = np = 10$.



If H_0 is true, then

- The values $Y = \{17, 18, 19, 20\}$ are as *extreme or more extreme* than the observed $Y = 17$.
- Also, by symmetry, the value $Y = 3$ is as extreme as $Y = 17$, and thus the $Y = \{0, 1, 2, 3\}$ are also *extreme or more extreme* than the observed $Y = 17$.

Example (cont.)

Example 3

Thus, the p-value is

$$
\begin{aligned}
p &= \sum_{i=0}^{3} \mathrm{P}(Y = i|\pi = 0.5) + \sum_{i=17}^{20} \mathrm{P}(Y = i|\pi = 0.5) \\
&= \sum_{i \in \{0,\ldots,3,17,\ldots,20\}} \frac{20!}{i!(20-i)!} 0.5^i (1-0.5)^{20-i} \\
&= 2 \times \sum_{i=0}^{3} \frac{20!}{i!(20-i)!} 0.5^{20} \quad \text{by symmetry} \\
&= 2 \times 0.5^{20} \times \{1 + 20 + (20 \cdot 19)/2 + (20 \cdot 19 \cdot 18)/(2 \cdot 3)\} \\
&= 0.0026
\end{aligned}
$$

Hence, we reject the hypothesis of a fair coin at the significance level 1%.

Impact of H_1

In the above example, if we test

$$H_0 : \pi = 0.5 \quad \text{v.s.} \quad H_1 : \pi > 0.5$$

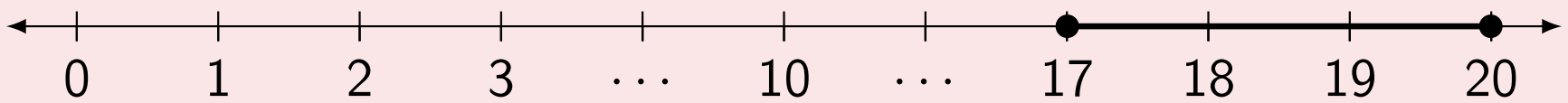
then we should only reject H_0 if there is strong evidence against H_0 in favour of H_1. Having observed $Y = 17$, the more extreme values are 18, 19 and 20. Therefore, the p-value is

$$p = \sum_{i \geq 17} \mathrm{P}(Y = i | \pi = 0.5) = 0.0013,$$

and the evidence against H_0 is even stronger.

Remark

This is a right-tail event.



Impact of H_1 (cont.)

On the other hand, if we test

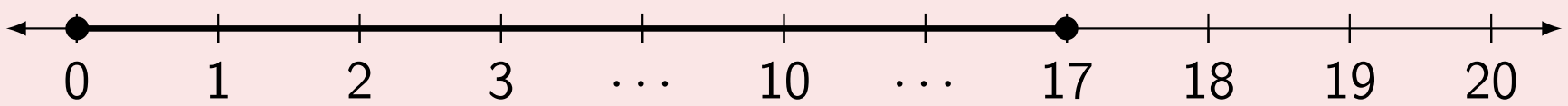
$$H_0 : \pi = 0.5 \quad \text{v.s.} \quad H_1 : \pi < 0.5$$

then the observation $Y = 17$ is more in favour of H_0 rather than H_1 now. We cannot reject H_0, as the p-value is now

$$p = \sum_{i \leq 17} \mathrm{P}(Y = i | \pi = 0.5) = 0.9987.$$

Remark

This is a left-tail event.



Type I and II errors

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = 1 - α)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = 1 - β)

Type I and II errors (cont.)

Definition 4 (Type I error)

The Type I error (false positive) is defined to be

$$\mathrm{P}(\text{Reject } H_0 | H_0 \text{ is true}).$$

The probability of making a Type I error is the p-value, and is not greater than α, the significance level. Hence, it is under control.

Definition 5 (Type II error)

The Type II error (false negative) is defined to be

$$\mathrm{P}(\text{Fail to reject } H_0 | H_0 \text{ is false}) =: \beta.$$

Unfortunately, we do not have explicit control on the probability of making a Type II error.

Power of a test

Definition 6 (Power of a test)

The power function of a test is defined as the probability of rejecting the null hypothesis correctly in favour of the alternative,

$$B(\theta) = \mathrm{P}(\text{Reject } H_0|\theta), \quad \theta \in \Theta_1$$

That is, $B(\theta) = 1 - \beta$ (one minus the probability of making a Type II error).

Of course, as statistical power increases, the probability of making a Type II error decreases.

Remark

It is more conclusive to end a test with H_0 rejected, as the decision "Not reject" does not imply that H_0 is accepted.

The Wald test

Suppose we would like to test $H_0 : \theta = \theta_0$, and $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is an estimator and is asymptotically normal, i.e. as $n \to \infty$,

$$T_\theta(X_1, \ldots, X_n) := \frac{\hat{\theta} - \theta}{\mathrm{SE}(\hat{\theta})} \xrightarrow{\mathrm{D}} \mathrm{N}(0, 1).$$

Then, under H_0, $\hat{\theta} \approx \mathrm{N}(\theta_0, \widehat{\mathrm{Var}}(\hat{\theta}))$.

Definition 7 (The Wald test)

The Wald test at the significance level α uses the test statistic T_{θ_0} to reject H_0 in favour of

- $H_1 : \theta \neq \theta_0$ if $|T_{\theta_0}| > z(\alpha/2)$
- $H_1 : \theta > \theta_0$ if $T_{\theta_0} > z(\alpha)$
- $H_1 : \theta < \theta_0$ if $T_{\theta_0} < -z(\alpha)$

where $z(\alpha)$ is the top-α point of $\mathrm{N}(0, 1)$.

The Wald test (cont.)

Example 8

To deal with a coffee shop's customer complaint that the amount of chilled coffee in their bottled drinks is less than the advertised 300ml, 20 bottles were decanted and the coffee measured, yielding X_i as follows:

282	301	311	271	293	268	302	301	293	256
278	301	309	294	282	281	305	301	285	279

The sample mean and the standard deviation are

$$\bar{X} = 289.7 \text{ ml} \qquad s = 14.8.$$

By the CLT,

$$\bar{X} \xrightarrow{\mathrm{D}} \mathrm{N}\left(\mu, \frac{14.8^2}{20}\right)$$

The Wald test (cont.)

Example 8

To test

$$H_0 : \mu = 300 \quad \text{v.s.} \quad H_1 : \mu < 300$$

we apply the Wald test with

$$T = \frac{\bar{X} - 300}{14.8/\sqrt{20}} = -3.121.$$

Comparing this against $-z(0.01) = -2.326$, we find that $T < z(0.01)$, hence we reject $H_0 : \mu = 300$ at the 1% significance level.

Conclusion: There is significant evidence which supports the claim that the bottled coffee is less than the advertised value of 300ml.

Likelihood ratio tests

One of the most popular ways of constructing tests when both null and alternative hypotheses are composite (i.e. not a single point).

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim f(\mathbf{x}|\boldsymbol{\theta})$. Consider hypotheses

$$H_0 : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \quad \text{v.s.} \quad H_1 : \ \boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$$

The likelihood ratio test will reject H_0 for the large values of the statistic

$$\text{LR} = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}|\mathbf{X})}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta}|\mathbf{X})} = \frac{f(\hat{\boldsymbol{\theta}}|\mathbf{X})}{f(\tilde{\boldsymbol{\theta}}|\mathbf{X})}$$

where $\hat{\boldsymbol{\theta}}$ is the (unconstrained) ML estimate, and $\tilde{\boldsymbol{\theta}}$ is the constrained ML estimate under H_0.

Likelihood ratio tests (cont.)

Remark

- It is easy to see that LR ≥ 1.
- The exact sampling distribution of LR is usually unknown, except in a few special cases.

Likelihood ratio tests (cont.)

Example 9 (One-sample t-test)

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$. We are interested in testing hypotheses

$$H_0 : \ \mu = \mu_0 \quad \text{v.s.} \quad H_1 : \ \mu \neq \mu_0,$$

where μ_0 is given, and σ^2 is unknown and is a nuisance parameter. Recall the likelihood function as being

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right\},$$

and maximising this without restriction yields

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Likelihood ratio tests (cont.)

Example 9

On the other hand, under H_0, μ is fixed at μ_0, while the constrained MLE for σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2.$$

The LR statistic (after simplification) is then

$$\text{LR} = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{L(\mu_0, \tilde{\sigma}^2)} = \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)^{n/2}.$$

Since $\tilde{\sigma}^2 = \hat{\sigma}^2 + (\bar{X} - \mu_0)^2$, it holds that $\tilde{\sigma}^2/\hat{\sigma}^2 = 1 + T^2/(n-1)$, where

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \Big/ n}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

Likelihood ratio tests (cont.)

Example 9

Under H_0, we know that $T \sim t_{n-1}$. Given that the LR test statistic is a monotonic function in T, we can use the distribution of T to make a decision. That is, the LR test will reject H_0 at the α significance if and only if $|T| > t_{n-1}(\alpha/2)$, the top $\alpha/2$ point of the t_{n-1} distribution.

Remark

Later in the chapter we will look at t-test for means, and you will find the exact same test statistic being used. In this special case, the LR test is equivalent to the t-test.

Asymptotic distribution of LRT statistic

Sometimes, we won't know for certain the sampling distribution of LR. However, we can rely on this asymptotic result:

Theorem 10 (Wilk's theorem)

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim f(\mathbf{x}|\boldsymbol{\theta})$ be a random sample, and consider testing the hypothesis

$$H_0 : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \quad \text{v.s.} \quad H_1 : \ \boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0.$$

The distribution of

$$D = -2 \log \left[\frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta}|\mathbf{X})}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}|\mathbf{X})} \right] \xrightarrow{\text{D}} \chi^2_{d-d_0}$$

as $n \to \infty$, where $\dim(\boldsymbol{\Theta}) = d$ and $\dim(\boldsymbol{\Theta}_0) = d_0$.

Asymptotic distribution of LRT statistic (cont.)

Example 11

Let $X_1, \ldots, X_n$ be independent, and $X_i \sim \mathrm{N}(\mu_i, 1)$. Consider the null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_n.$$

The likelihood function is

$$L(\mu_1, \ldots, \mu_n) = C \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu_i)^2\right\},$$

for some constant C which is independent of the μ_is. Then, the unconstrained MLE are $\hat{\mu}_i = X_i$, while the constrained MLE is $\tilde{\mu} = \bar{X}$. Hence,

$$\mathrm{LR} = \frac{L(\hat{\mu}_1, \ldots, \hat{\mu}_n)}{L(\tilde{\mu}, \ldots, \tilde{\mu})} = \exp\left\{\frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}$$

Asymptotic distribution of LRT statistic (cont.)

Example 11

Therefore,

$$D = -2 \log \frac{L(\tilde{\mu}, \ldots, \tilde{\mu})}{L(\hat{\mu}_1, \ldots, \hat{\mu}_n)} = 2 \log \mathrm{LR} = \sum_{i=1}^{n} (X_i - \bar{X})^2 \xrightarrow{\mathrm{D}} \chi^2_{n-1}$$

as $n \to \infty$ by Wilk's theorem. So we reject the null hypothesis in favour of the alternative for large values of D (compared to χ^2_{n-1} distribution).

It turns out that D has an exact χ^2_{n-1} distribution since $D/(n-1) = s^2$, and we saw in Chapter 3 that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$.

Recall: The score function and its properties

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}$. Then, the log-likelihood function is given by

$$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$

Assume that the first and second derivatives of the log-likelihood function exists. These are given by

$$S(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i|\theta) \quad \text{and} \quad S'(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta).$$

Under certain regularity conditions, the Fisher information is then defined to be

$$\mathcal{I}(\theta) = -\,\mathrm{E}\left(S'(\theta)\right).$$

Recall: The score function and its properties (cont.)

We are also aware that Fisher information is additive, that is

$$\mathcal{I}(\theta) = \mathcal{I}_1(\theta) + \cdots + \mathcal{I}_n(\theta),$$

where $\mathcal{I}_i(\theta) = -\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\right]$ is the unit Fisher information. Due to iid, we have that $\mathcal{I}(\theta) = n\mathcal{I}_1(\theta)$.

We also previously showed the following properties to be true (the expectation and variance of the score function):

- $\mathrm{E}\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right] = 0$.
- $\mathrm{Var}\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right] = \mathcal{I}_i(\theta)$.

This seems to suggest that the score function is a random variable with some distribution to it.

The score test

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, and define a new random variable (the score)

$$Y_i = \frac{\partial}{\partial \theta} \log f(X_i|\theta).$$

Immediately by the CLT we obtain

$$\bar{Y} \xrightarrow{\text{D}} \text{N}\left(\overbrace{\text{E}\left[\frac{\partial}{\partial \theta} \log f(X_i|\theta) \right]}^{0}, \overbrace{\text{Var}\left[\frac{\partial}{\partial \theta} \log f(X_i|\theta) \right] / n}^{\mathcal{I}_1 / n} \right)$$

as $n \to \infty$. In other words,

$$S(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i|\theta) = n\bar{Y} \xrightarrow{\text{D}} \text{N}(0, \overbrace{n\mathcal{I}_1}^{\mathcal{I}(\theta)}).$$

The score test (cont.)

We can use the above result for testing

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta \neq \theta_0.$$

Definition 12 (The score test)

Let $X_1, \ldots, X_n$ be a sample, and consider testing the hypothesis as above. Define the score test statistic to be

$$Z = \frac{S(\theta_0)}{\sqrt{\mathcal{I}(\theta_0)}},$$

where $S(\theta_0) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta_0)$ and $\mathcal{I}(\theta_0) = -\,\mathrm{E}(S'(\theta_0))$. Then we reject H_0 in favour of H_1 for large values of Z, i.e. when $|Z| > z(\alpha/2)$ at the α level significance.

The score test (cont.)

Example 13

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ where σ^2 is known. Using standard results we can show that

$$S(\mu) = \frac{n(\bar{X} - \mu)}{\sigma^2} \quad \text{and} \quad \mathcal{I}(\mu) = n/\sigma^2$$

The score test rejects $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu \neq \mu_0$ if

$$|Z| = \left| \frac{S(\mu)}{\sqrt{\mathcal{I}(\mu)}} \right| = \left| \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right| > z(\alpha/2),$$

which turns out to be the standard Z-test.

The score test (cont.)

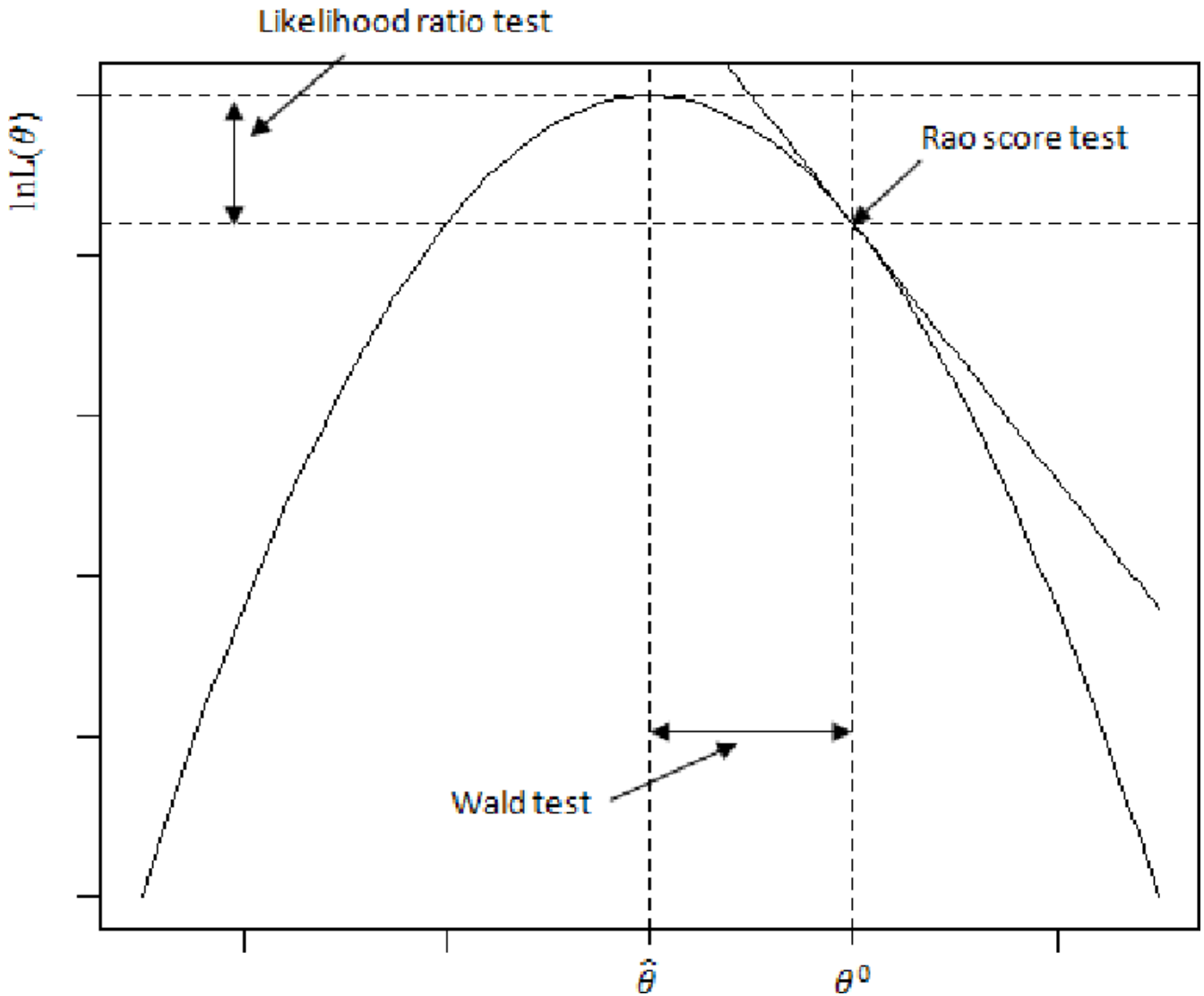
Remark

The score test

- is sometimes known as Rao's test.
- is intuitively clear: As $S(\theta)$ deviates away from zero (its expectation), then θ is also far away from the MLE $\hat{\theta}$.
- works whether the parameter is a scalar θ or a vector $\boldsymbol{\theta}$. In the vector case, $S(\boldsymbol{\theta}) \in \mathbb{R}^p$ is a vector and $\mathcal{I}(\boldsymbol{\theta}) \in \mathbb{R}^{p \times p}$ is a matrix, and the test statistic is

$$T^2 = S(\boldsymbol{\theta})^\top \mathcal{I}(\boldsymbol{\theta})^{-1} S(\boldsymbol{\theta}) \xrightarrow{\mathrm{D}} \chi^2_p.$$

Comparison of asymptotic tests



Tests may give different results in finite samples, but they are all equivalent asymptotically.

One-sample t-test

Let $X_1, \ldots, X_n$ be a sample from $\mathrm{N}(\mu, \sigma^2)$, where both μ and $\sigma^2 > 0$ are unknown. Test the hypothesis that

$$H_0 : \ \mu = \mu_0 \quad \text{v.s.} \quad H_1 : \ \mu \neq \mu_0$$

for some known μ_0. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Definition 14 (The t-test)

Under H_0, the test statistic defined by

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

is distributed according to t_{n-1}. Hence, we reject H_0 if $|T| > t_{n-1}(\alpha/2)$, where α is the significance level of the test, and $t_k(\alpha)$ is the top-α point of the t_k-distribution.

One-sample t-test (cont.)

Remark

There are three different but equivalent ways to reject the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$:

- Reject H_0 if $|T| > t_{n-1}(\alpha/2)$.
- Reject H_0 if $p = \mathrm{P}(|T| > t_{n-1}(\alpha/2)) < \alpha$.
- Reject H_0 if $\mu_0 \notin \{\bar{X} \pm t_{n-1}(\alpha/2) \cdot \mathrm{SE}(\bar{X})\}$.

Remark

Of course we can also test the null hypothesis against other alternatives, such as

- $H_1 : \mu > \mu_0$ (reject H_0 if $T > t_{n-1}(\alpha)$)
- $H_1 : \mu < \mu_0$ (reject H_0 if $T < -t_{n-1}(\alpha)$)

One-sample t-test (cont.)

Example 15

Going back to the coffee data (Example 8), we can use the t-test instead of the Wald test, with the additional assumption that $X_i \sim \mathrm{N}(\mu, \sigma^2)$, where the true mean and variance are unknown.

In testing

$$H_0 : \ \mu = 300 \quad \text{v.s.} \quad H_1 : \ \mu < 300$$

we use the test statistic

$$T = \frac{\bar{X} - 300}{14.8/\sqrt{20}} = -3.121$$

which under H_0 is distributed according to t_{19}. Since T is less than $-t_{19}(0.01) = -2.539$, we reject the null hypothesis in favour of the alternative.

Two-sample t-test

Suppose we have two independent samples. Let

- $X_1, \ldots, X_n$ be a sample from $\mathrm{N}(\mu_x, \sigma_x^2)$; and
- $Y_1, \ldots, Y_m$ be a sample from $\mathrm{N}(\mu_y, \sigma_y^2)$

All means and variances, μ_x, μ_y, σ_x^2, and σ_y^2, are unknown. We are interested in testing

$$H_0 : \ \mu_x - \mu_y = \delta \quad \text{v.s.} \quad H_1 : \ \mu_x - \mu_y \neq \delta$$

for some known constant δ. Define, as usual, the sample means and variances as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i \qquad s_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$$

Two-sample t-test (cont.)

Using the distributional properties of Chapter 3, we know that

$$\bar{X} \sim \mathrm{N}(\mu_x, \sigma_x^2/n) \qquad \bar{Y} \sim \mathrm{N}(\mu_y, \sigma_y^2/m)$$

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

and

$$\frac{n-1}{\sigma_x^2} s_x^2 \sim \chi_{n-1}^2 \qquad \frac{m-1}{\sigma_y^2} s_y^2 \sim \chi_{m-1}^2$$

$$\frac{n-1}{\sigma_x^2} s_x^2 + \frac{m-1}{\sigma_y^2} s_y^2 \sim \chi_{n+m-2}^2$$

Two-sample t-test (cont.)

Let's build a test statistic based on these distributions Define T_1 and T_2 as

$$T_1 = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$
$$= \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sigma\sqrt{1/n + 1/m}} \quad \text{assume } \sigma_x = \sigma_y = \sigma$$

and

$$T_2 = \frac{n-1}{\sigma_x^2}s_x^2 + \frac{m-1}{\sigma_y^2}s_y^2$$
$$= \frac{1}{\sigma^2}\left\{(n-1)s_x^2 + (m-1)s_y^2\right\} \quad \text{assume } \sigma_x = \sigma_y = \sigma$$

Two-sample t-test (cont.)

Of course we know that $T_1 \sim \mathrm{N}(0,1)$ and $T_2 \sim \chi^2_{n+m-2}$. Then, with the additional assumption that $\sigma_x = \sigma_y = \sigma$,

$$
\begin{aligned}
T &= \frac{T_1}{\sqrt{T_2/(n+m-2)}} \\
&= \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\cancel{\sigma}\sqrt{1/n + 1/m}} \times \cancel{\sigma}\sqrt{\frac{n+m-2}{(n-1)s_x^2 + (m-1)s_y^2}} \\
&= \sqrt{\frac{n+m-2}{1/n+1/m}} \cdot \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sim t_{n+m-2}
\end{aligned}
$$

Two-sample t-test (cont.)

Definition 16 (Two-sample t-test)

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, X_m \overset{\text{iid}}{\sim} \mathrm{N}(\mu_y, \sigma_y^2)$ independently of each other. Further assume that $\sigma_x = \sigma_y$. Under H_0, the test statistic defined by

$$T = \sqrt{\frac{n+m-2}{1/n+1/m}} \cdot \frac{\bar{X} - \bar{Y} - (\overbrace{\mu_x - \mu_y}^{\delta})}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}}$$

follows a t_{n+m-2} distribution. we reject H_0 if $|T| > t_{n+m-2}(\alpha/2)$, where α is the significance level of the test, and $t_k(\alpha)$ is the top-α point of the t_k-distribution.

Two-sample t-test (cont.)

Remark

- Note that we can also do one-sided tests.
- It is much more challenging to compare two normal means with unknown and **different** variances, and this is beyond the scope of this course–refer to Welch's t-test.
- On the other hand, the Wald test provides an easy alternative when both n_x and n_y are large (see next example).

Two-sample t-test (cont.)

Example 17

A company invents a new type of mechanical keyboard which they feel makes typing more effortless. They wish to take out an advertisement to claim that people are able to type faster using their new keyboard (X) rather than their main competitor's keyboard (Y). Before making such a bold claim, they conduct an experiment on 100 individuals chosen at random, whereby each individual was asked to write a 120-word paragraph, and their typing time measured, using the two kinds of keyboard X and Y. This experiment yielded the following results

$$\bar{X} = 2.84 \qquad \bar{Y} = 3.02$$
$$s_x^2 = 0.48 \qquad s_y^2 = 0.42$$

Test, at the 5% significance level, if the two keyboard lead to different typing times.

Two-sample t-test (cont.)

Example 17

Assume that the two samples $X_1, \ldots, X_{100} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_{100} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_y, \sigma_y^2)$ are independent of each other. The problem requires to test the hypothesis

$$H_0 : \ \mu_x = \mu_y \quad \text{v.s.} \quad H_1 : \ \mu_x \neq \mu_y.$$

If we assume additionally that $\sigma_x = \sigma_y$, then we can use the two-sample t-test test statistic, which under H_0 takes value

$$T = \sqrt{\frac{198}{1/100 + 1/100}} \cdot \frac{2.84 - 3.02 - (\overbrace{\mu_x - \mu_y}^{0})}{\sqrt{(99)0.48 + (99)0.42}} = -1.897$$

which is then compared against $t_{198}(0.025) = 1.97$. Since $|T| \not> t_{198}(0.025)$, we are unable to reject H_0 in favour of H_1.

Two-sample t-test (cont.)

Example 17

Moreover, a 95% confidence interval for $\mu_x - \mu_y$ is

$$(\bar{X} - \bar{Y}) \pm t_{198}(0.025)\frac{\sqrt{s_x^2 + s_y^2}}{10} = -0.18 \pm 0.187$$
$$= (-0.367, 0.007).$$

Notice that this interval contains 0.

Two-sample t-test (cont.)

Example 17

Suppose the normality assumption for the X_i and Y_i do not hold. In this case, we can use a Wald test, which does not require normality. Note that

$$\mathrm{SE}(\bar{X} - \bar{Y}) = \sqrt{s_x^2/100 + s_y^2/100},$$

and thus the Wald test statistic under H_0 is

$$T = \frac{\bar{X} - \bar{Y}}{\mathrm{SE}(\bar{X} - \bar{Y})} \sim \mathrm{N}(0, 1),$$

and thus we reject H_0 when $|T| > 1.96$ at the 5% significance level. For the given data, $T = -1.9$, so we are again unable to reject H_0 in favour of H_1.

Two-sample t-test (cont.)

Example 17

We are also able to invert the Wald statistic to obtain a 95% confidence interval for $\mu_x - \mu_y$:

$$\begin{aligned}(\bar{X} - \bar{Y}) \pm z(0.025)\mathrm{SE}(\bar{X} - \bar{Y}) &= -0.18 \pm 0.186 \\ &= (-0.366, 0.006),\end{aligned}$$

which we see 0 is not included as well.

Remark

Since two observations were collected from each individual, perhaps the independence assumption of the two samples is not a very good one. The next test we shall look at is the paired t-test.

Paired t-test

Consider again we two samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu_y, \sigma_y^2)$. Suppose now we are able to pair the two samples together to create a sample of matched pairs $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

We are interested in testing

$$H_0 : \ \mu_x - \mu_y = \delta \quad \text{v.s.} \quad H_1 : \ \mu_x - \mu_y \neq \delta$$

for some known value δ, but all means and variances μ_x, μ_y, σ_x^2, and σ_y^2 are unknown.

A typical situation would be when an individual or unit has been tested twice, and the interest was to see whether a "before/after" effect exists.

Paired t-test (cont.)

Define $Z_i = X_i - Y_i$. Then with $\mu_z = \mu_x - \mu_y$ and $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$, $Z_i \sim \mathrm{N}(\mu_z, \sigma_z^2)$, and the above hypothesis is equivalent to testing

$$H_0 : \ \mu_z = \delta \quad \text{v.s.} \quad H_1 : \ \mu_z \neq \delta$$

using the standard one-sample t-test (Definition 14). We also require the following estimators:

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \quad = \bar{X} - \bar{Y}$$

and

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

Paired t-test (cont.)

Example 18

Back to the keyboard example. It is more suitable to think of the two typing speeds as paired samples, as they come from the same individual. We must still assume that both samples X_i and Y_i come from normal distributions. Additionally, suppose we have $s_z^2 = 0.6$, the sample variance of the differences $Z_i := X_i - Y_i \sim \mathrm{N}(\mu_z, \sigma_z^2)$. We test

$$H_0 : \ \mu_z = 0 \quad \text{v.s.} \quad H_1 : \ \mu_z \neq 0$$

using the test statistic

$$T = \frac{\overbrace{\bar{Z}}^{\bar{X}-\bar{Y}}}{\mathrm{SE}(\bar{Z})} = \frac{\bar{X} - \bar{Y}}{s_z/\sqrt{n}} = -2.327.$$

This is then compared against $t_{100-1}(0.025) = 1.98$. Since $|T| > t_{100-1}(0.025)$, we are able to reject the H_0 using the paired t-test.

Paired t-test (cont.)

Example 18

A 95% confidence interval for $\mu_x - \mu_y$ is

$$\begin{aligned}\bar{Z} \pm t_{99}(0.025)\mathrm{SE}(\bar{Z}) &= (2.84 - 3.02) \pm 1.98 \times \frac{0.6}{\sqrt{100}} \\ &= -0.18 \pm 0.154 \\ &= (-0.334, -0.026).\end{aligned}$$

Zero is not included in this confidence interval.

Remark

For this keyboard example, different methods lead to different but not contradictory conclusions, since **Not reject** $\neq$ **Accept**.

Pearson's χ^2-test

Goodness of fit tests are used to test if a given distribution fits the data. Many measures and test statistics exist. In the context of discrete data, we shall look at the χ^2-test.

Let $\{X_1, \ldots, X_n\}$ be a random sample from a discrete distribution of m categories, $1, \ldots, m$. Denote the probability function

$$p_j = \mathrm{P}(X_i = j) \geq 0$$

for $j = 1, \ldots, m$ such that $\sum_{j=1}^{m} p_j = 1$.

Denote the number of times category j occurs by

$$O_j = \sum_{i=1}^{n} [X_i = j].$$

Obviously, $n = \sum_{j=1}^{m} O_j$.

Pearson's χ^2-test (cont.)

Tabulate the data as follows:

Category	1	2	$\cdots$	m
Frequency	O_1	O_2	$\cdots$	O_m

We are interesting in testing the null hypothesis

$$H_0 : \ p_j = p_j(\theta), \quad j = 1, \ldots, m,$$

where the function forms of $p_j(\theta)$ are known, but the parameter θ is unknown. Examples include

- Binomial distribution: θ is success probability.
- Poisson distribution: θ is the rate.
- Etc.

The alternative hypothesis is H_1 : Data does not follow p_j.

Pearson's χ^2-test (cont.)

We first have to estimate θ by, for example, its ML estimate $\hat{\theta}$. Then, under H_0, the **expected frequencies** are (for $j = 1, \ldots, m$)

$$E_j = np_j(\hat{\theta}).$$

Listing them together with the observed frequencies, we have

Category	1	2	$\cdots$	m
Observed frequencies	O_1	O_2	$\cdots$	O_m
Expected frequencies	E_1	E_2	$\cdots$	E_m

If H_0 is true, we expect $O_j \approx E_j = np_j$, since the LLN states

$$\frac{O_j}{n} = \frac{1}{n}\sum_{i=1}^{n}[X_i = j] \xrightarrow{\mathrm{P}} \mathrm{E}[X_i = j] = p_j.$$

Pearson's χ^2-test (cont.)

Definition 19 (Pearson's χ^2-test)

Under H_0, the test statistic

$$T = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \xrightarrow{\text{D}} \chi^2_{m-1-p}$$

where p is the number of components of θ.

Remark

- It is important that $E_j \geq 5$. This may be achieved by combining together categories with smaller expected frequencies.
- When p_j are completely specified (i.e. known) under H_0, $p = 0$.

Pearson's χ^2-test (cont.)

Example 20

A supermarket recorded the numbers of arrivals over 100 one-minute intervals. The data were summarised as follows:

No. of arrivals	0	1	2	3	4	5	6	7
Frequency	13	29	32	20	4	1	0	1

Do the data match a Poisson distribution?

This requires us to test the following null hypothesis:

$$H_0 : p_j = \lambda^j e^{-\lambda}/j!, \quad j = 0, 1, \ldots$$

The ML estimate for λ is in fact the sample mean $\hat{\lambda} = \bar{X}$.

Pearson's χ^2-test (cont.)

Example 20

The sample mean is calculated as

$$\bar{X} = \frac{0 \times 13 + 1 \times 29 + 2 \times 32 + \cdots + 0 \times 6 + 1 \times 7}{100} = 1.81.$$

So with $\hat{\lambda} = 1.81$, the expected frequencies are calculated as

$$E_j = np_j(\hat{\lambda}) = 100 \times \frac{e^{-1.81}1.81^j}{j!}.$$

We also combine the last four categories to ensure $E_j \geq 5$.

Pearson's χ^2-test (cont.)

Example 20

Append the new information to the table above:

No. of arrivals	0	1	2	3	≥ 4	Total
O_j	13	29	32	20	6	100
p_j	0.164	0.296	0.268	0.162	0.110	1
E_j	16.4	29.6	26.8	16.2	11.0	100
$(O_j - E_j)^2/E_j$	0.705	0.012	1.010	0.891	2.273	4.891

Under H_0, we find that $T = \sum_j (O_j - E_j)^2/E_j \sim \chi^2_{5-1-1} = \chi^2_3$. Since $T = 4.891 < \chi^2_3(0.1) = 6.25$, we cannot reject the assumption that the data follow a Poisson distribution.

Pearson's χ^2-test (cont.)

Remark

The Pearson's χ^2 goodness of fit test is widely used in practice. However, we should bear in mind that when H_0 cannot be rejected, we are *not* in a position to conclude that the assumed distribution is true, as **Not reject** $\neq$ **Accept**.

Test of independence

Let (X, Y) be two discrete random variables, where X has r categories and Y has c categories. Let

$$p_{ij} = \mathrm{P}(X = i, Y = j) \geq 0, \quad i = 1, \ldots, r, j = 1, \ldots, c.$$

Note that $\sum_{i,j} p_{ij} = 1$. Also let

$$p_{i\cdot} = \mathrm{P}(X = i) = \sum_{j=1}^{c} \mathrm{P}(X = i, Y = j) = \sum_{j=1}^{c} p_{ij}$$

$$p_{\cdot j} = \mathrm{P}(Y = j) = \sum_{i=1}^{r} \mathrm{P}(X = i, Y = j) = \sum_{i=1}^{r} p_{ij}$$

be the marginal probabilities for X and Y respectively.

Recall that X and Y are independent iff $p_{ij} = p_{i\cdot} p_{\cdot j}$.

Test of independence (cont.)

Suppose we have n pairs of observations from (X, Y). The data are presented in a **contingency table** below.

	$Y = 1$	$Y = 2$	$\cdots$	$Y = c$
$X = 1$	O_{11}	O_{12}	$\cdots$	O_{1c}
$X = 2$	O_{21}	O_{22}	$\cdots$	O_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$X = r$	O_{r1}	O_{r2}	$\cdots$	O_{rc}

Here, O_{ij} represents the number of (X, Y) pairs equal to (i, j).

Test of independence (cont.)

It is often useful to add the totals into the table as well

	$Y = 1$	$Y = 2$	$\cdots$	$Y = c$	Total
$X = 1$	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1\cdot}$
$X = 2$	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X = r$	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r\cdot}$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot c}$	n

where

$$O_{i\cdot} = \sum_{j=1}^{c} O_{ij} \quad \text{and} \quad O_{\cdot j} = \sum_{i=1}^{r} O_{ij}.$$

Test of independence (cont.)

We are interested in testing the independence

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1, \ldots, r, j = 1, \ldots, c.$$

Under H_0, a natural estimator for p_{ij} is

$$\tilde{p}_{ij} = \hat{p}_{i\cdot} \hat{p}_{\cdot j} = \frac{O_{i\cdot}}{n} \frac{O_{\cdot j}}{n}$$

Hence, the expected frequencies at the (i, j)th cell is

$$E_{ij} = n\tilde{p}_{ij} = \frac{O_{i\cdot} O_{\cdot j}}{n}.$$

If H_0 is true, then we again expect $O_{ij} \approx E_{ij}$.

Test of independence (cont.)

Definition 21 (χ^2 test of independence)

Under H_0, the goodness of fit test statistic is defined as

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

follows a $\chi^2_{(r-1)(c-1)}$ distribution.

Remark

For testing independence, it always holds that

$$O_{i\cdot} - E_{i\cdot} = 0 \quad \text{and} \quad O_{\cdot j} - E_{\cdot j} = 0$$

which is useful for checking for calculation errors.

Test of independence (cont.)

Example 22

The table below lists the counts on the sugar preferences and gender of 150 randomly selected bubble tea drinkers.

	Full sugar	Half sugar	No sugar	Total
Male	20	40	20	80
Female	30	30	10	70
Total	50	70	30	150

Is sugar preference independent of gender? Test this hypothesis at the 5% significance level.

Test of independence (cont.)

Example 22

First, calculate the expected frequencies

$$E_{11} = \frac{80 \cdot 50}{150} = 26.67, \; E_{12} = \frac{80 \cdot 70}{150} = 37.33, \; E_{13} = \frac{80 \cdot 30}{150} = 16$$
$$E_{21} = \frac{70 \cdot 50}{150} = 23.33, \; E_{22} = \frac{70 \cdot 70}{150} = 32.67, \; E_{23} = \frac{70 \cdot 30}{150} = 14.$$

Then, calculate $(O_{ij} - E_{ij})^2/E_{ij}$.

$(O_{ij} - E_{ij})^2/E_{ij}$	Full sugar	Half sugar	No sugar
Male	1.668	0.191	1.000
Female	1.907	0.218	1.142

Under the null hypothesis of independence, $T = \sum_{ij}(O_{ij} - E_{ij})^2/E_{ij} \sim \chi^2_2$. As $T = 6.126 > \chi^2_2(0.05) = 5.991$, reject the null hypothesis at the 5% level. Evidently sugar preference and gender are not independent.

Tests for $r \times c$ tables–a general description

In general, we may test for different types of structure in an $r \times c$ table, for example, the symmetry ($p_{ij} = p_{ji}$).

The key is to compute the expected frequencies E_{ij} under the null hypothesis H_0. Under H_0, the test statistic is

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{k-p},$$

where

- $k =$ no. of "free" counts amount O_{ij}.
- $p =$ no. of estimated "free" parameters.

We reject H_0 at the α% significance level if $T > \chi^2_{k-p}(\alpha)$.

Other tests

- One more important test which involves group means is the ANOVA. We will try to motivate this from a linear regression standpoint in the next chapter.
- All of these tests are parametric, i.e. it must be assumed that the data follows some probability distribution.
- However, non-parametric tests exists as well, and these are briefly mentioned:
  - Wilcoxon signed-rank test and Mann-Whitney U test in place of t-tests (also used when data is non-normal).
  - Kruskal-Wallis test and Friedman test for ANOVA (also used when data is non-normal).
  - Fisher exact test in place of χ^2-tests (used when cell counts are small).
  - Permutation test (used when sampling distribution of test statistic is unknown).

References I

Benhamou, E. and V. Melot (2018). "Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation". *arXiv preprint arXiv:1808.09171*.

Casella, G. and R. L. Berger (2002). *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury. ISBN: 978-0-534-24312-8.

Pawitan, Y. (2001). *In All Likelihood*. Statistical Modelling and Inference Using Likelihood. Oxford University Press. ISBN: 978-0-19-850765-9.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.