Chapter 9. Hypothesis Testing (I)

Hypothesis Testing, together with statistical estimation, are the two most frequently used statistical inference methods. It addresses a different type of practical problems from statistical estimation.

9.1 Basic idea, p-values

Based on the data, a (statistical) test is to make a binary decision on a well-defined hypothesis, denoted as H_0:

$$\text{Reject } H_0 \quad \text{or} \quad \text{Not reject } H_0$$

Consider a simple experiment: toss a coin n times.

Let $X_1, \cdots, X_n$ be the outcomes: Head – $X_i = 1$, Tail – $X_i = 0$

Probability distribution: $P(X_i = 1) = \pi = 1 - P(X_i = 0)$, $\pi \in (0,\ 1)$

Estimation: $\widehat{\pi} = \bar{X} = (X_1 + \cdots + X_n)/n$.

Test: to assess if a hypothesis such as "*a fair coin*" is true or not, which may be formally represented as

$$H_0: \ \pi = 0.5.$$

$\boxed{\text{The answer cannot be resulted from the estimator } \widehat{\pi}}$

If $\widehat{\pi} = 0.9$, H_0 is unlikely to be true

If $\widehat{\pi} = 0.45$, H_0 may be true (and also may be untrue)

If $\widehat{\pi} = 0.7$, what to do then?

A customer complaint: the amount of coffee in a Hilltop coffee bottle is less than the advertised weight 3 pounds.

Sample 20 bottles, yielding the average 2.897

Is this sufficient to substantiate the complaint?

Again statistical estimation cannot provide a satisfactory answer, due to random fluctuation among different samples

We cast the problem into a hypothesis testing problem:

Let the weight of coffee be a normal random variable $X \sim N(\mu, \sigma^2)$. We need to test the hypothesis $\mu < 3$. In fact, we use the data to test the hypothesis

$$H_0: \ \mu = 3 \quad (\text{or } H_0: \ \mu \geq 3)$$

If we could reject H_0, the customer complaint will be vindicated.

Suppose one is interested in estimating the mean income of a community. Suppose the income population is normal $N(\mu,\ 25)$ and a random sample of $n = 25$ observations is taken, yielding the sample mean $\bar{X} = 17$.

Three expert economists give their own opinions as follows:

- Mr A claims the mean income $\mu = 16$
- Mr B claims the mean income $\mu = 15$
- Mr C claims the mean income $\mu = 14$

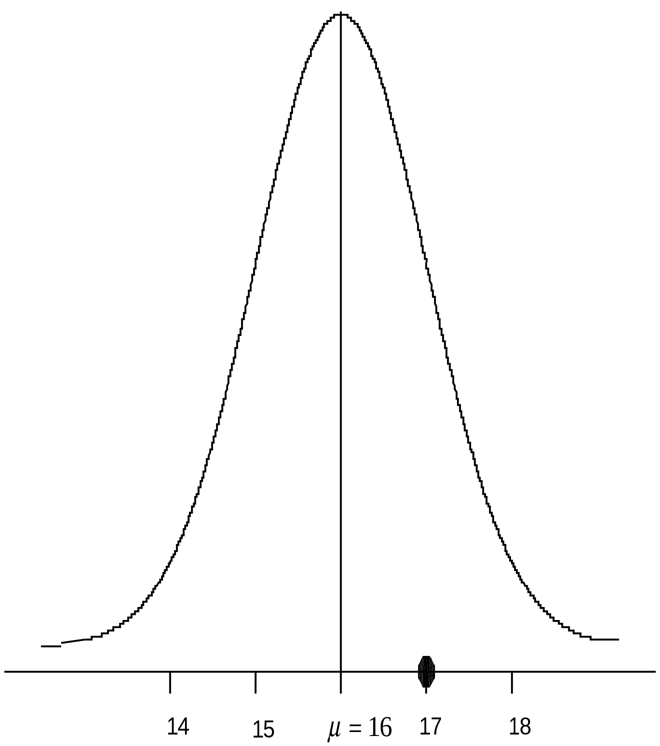
How would you assess those experts' statements?

Note. $\bar{X} \sim N(\mu, \sigma^2/n) = N(\mu, 1)$ — we assess the statements based on this distribution.

If Mr A's claim were correct, $\bar{X} \sim N(16,\ 1)$.

The observed value $\bar{X} = 17$ is *one standard deviation* away from μ, and may be regarded as a *typical observation* from the distribution.

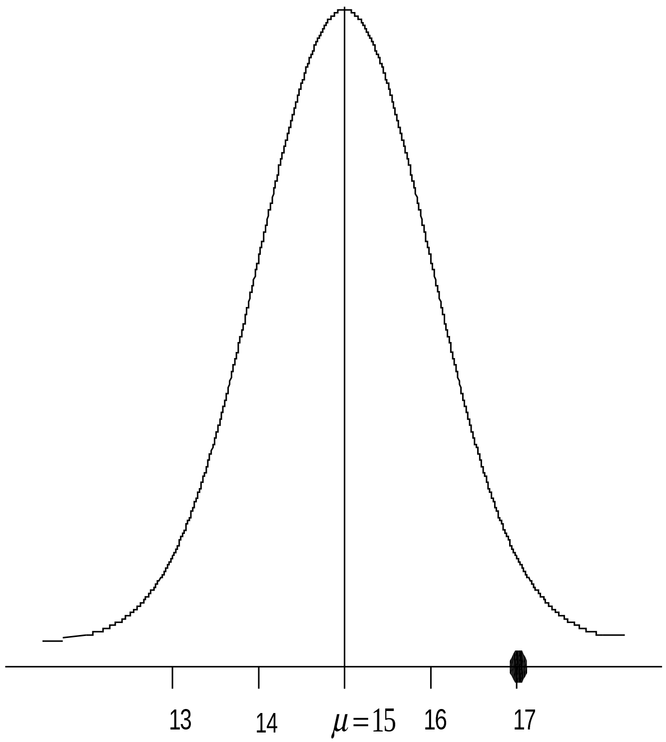
Little inconsistency between the claim and the data evidence.



If Mr B's claim were correct, $\bar{X} \sim N(15,\ 1)$.

The observed value $\bar{X} = 17$ begins to look *a bit extreme*, as it is *two standard deviation* away from μ.

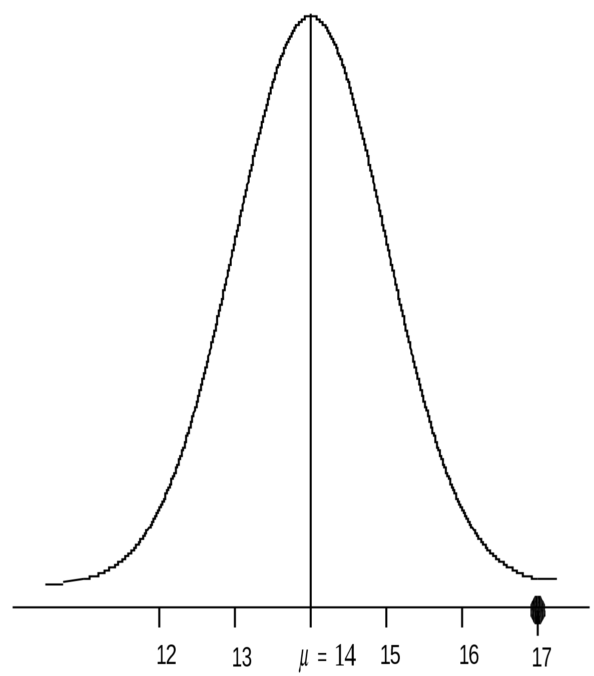
Inconsistency between the claim and the data evidence.



If Mr C's claim were correct, $\bar{X} \sim N(14,\ 1)$.

The observed value $\bar{X} = 17$ is *extreme* indeed, as it is *three standard deviation* away from μ.

Strong inconsistency between the claim and the data evidence.



A measure of the discrepancy between the hypothesised (claimed) value for μ and the observed value $\bar{X} = x$ is the probability of observing $\bar{X} = x$ or more extreme values. This probability is called the p-value. That is

- under $H_0: \ \mu = 16$,
$$P(\bar{X} \geq 17) + P(\bar{X} \leq 15) = P(|\bar{X} - 16| \geq 1) = 0.317$$
- under $H_0: \ \mu = 15$,
$$P(\bar{X} \geq 17) + P(\bar{X} \leq 13) = P(|\bar{X} - 15| \geq 2) = 0.046$$
- under $H_0: \ \mu = 14$,
$$P(\bar{X} \geq 17) + P(\bar{X} \leq 11) = P(|\bar{X} - 14| \geq 3) = 0.003$$

In summary, we reject the hypothesis $\mu = 15$ or $\mu = 14$, as, for example, if the hypothesis $\mu = 14$ is true, the probability of observing $\bar{X} = 17$ or more extreme values is merely 0.003. We are comfortable with this decision, as *a small probability event would not occur in a single experiment*.

On the other hand, we cannot reject the hypothesis $\mu = 16$.

But this does not imply that this hypothesis is necessarily true, as, for example, $\mu = 17$ or 18 are at least as likely as $\mu = 16$.

Not Reject $\neq$ Accept

A statistical test is incapable to accept a hypothesis.

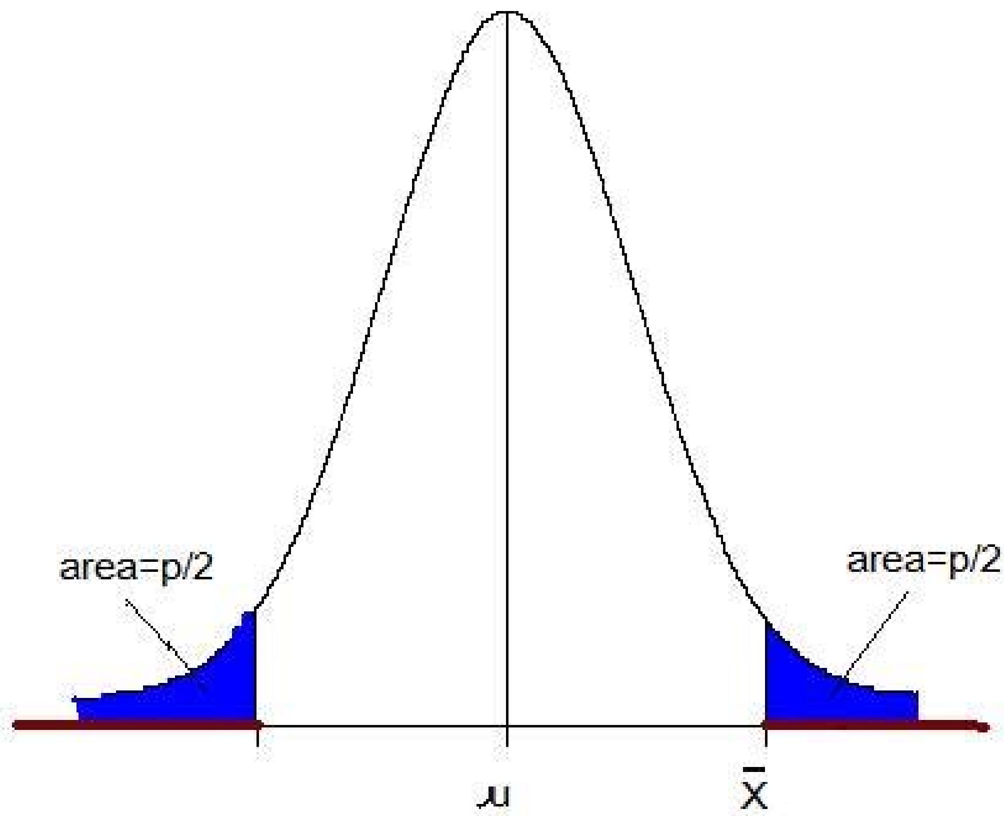
p-value: the probability of the event that a test statistic takes the observed value or more extreme (i.e. more unlikely) values under H_0

It is a measure of the discrepancy between a hypothesis and data.

p-value small: hypothesis is not supported by data

p-value large: hypothesis is not inconsistent with data

p-value may be seen as a risk measure of rejecting hypothesis H_0



General setting of hypothesis test

Let $\{X_1, \cdots, X_n\}$ be a random sample from a distribution $F(\cdot, \theta)$. We are interested in testing the hypotheses

$$H_0: \theta = \theta_0 \qquad \text{vs} \qquad H_1: \theta \in \Theta_1,$$

where θ_0 is a fixed value, Θ_1 is a set, and $\theta_0 \notin \Theta_1$.

- H_0 is called a null hypothesis
- H_1 is called an alternative hypothesis

Significance level α: a small number between 0 and 1 selected subjectively.

Often we choose $\alpha = 0.1, \ 0.05$ or 0.01, i.e. tests are often conducted as the significance levels 10%, 5% or 1%.

Decision: **Reject H_0 if p-value $\leq \alpha$**

Statistical testing procedure:

Step 1. Find a test statistic $T = T(X_1, \cdots, X_n)$. Denote T_0 the value of T with the given sample of observations.

Step 2. Compute the p-value, i.e.

$$p = P_{\theta_0}(\, T = T_0 \text{ or more extreme values}\,),$$

where P_{θ_0} denotes the probability distribution with $\theta = \theta_0$.

Step 3. If $p \leq \alpha$, reject H_0. Otherwise, H_0 is not rejected.

Remarks. 1. The alternative hypothesis H_1 is helpful to identify powerful test statistic T.

2. The significance level α controls how small is small for p-values.

3. “More extreme values" refers to those more unlikely values (than T_0) under H_0 in favour of H_1.

Example 1. Let $X_1, \cdots, X_{20}$, taking values either 1 or 0, be the outcomes of an experiment of tossing a coin 20 times, i.e.

$$P(X_i = 1) = \pi = 1 - P(X_i = 0), \quad \pi \in (0, 1).$$

We are interested in testing

$$H_0 : \pi = 0.5 \qquad \text{against} \qquad H_1 : \pi \neq 0.5.$$

Suppose there are 17 X_i's taking value 1, and 3 taking value 0. Will you reject the null hypothesis at the significance level 5%?

Let $Y = X_1 + \cdots + X_{20}$. Then $Y \sim Bin(20, \pi)$. We use Y as the test statistic.

With the given sample, we observe $Y = 17$. What are the more extreme values for Y if H_0 is true?

Under H_0, $EY = n\pi_0 = 10$. Hence 3 is as extreme as 17, and the more extreme values are

$$18,\ 19,\ 20,\ \text{ and }\ 0,\ 1,\ 2.$$

Thus the p-value is

$$\begin{aligned}
& \left(\sum_{i=0}^{3} + \sum_{i=17}^{20}\right) P_{H_0}(Y = i) \\
= & \left(\sum_{i=0}^{3} + \sum_{i=17}^{20}\right) \frac{20!}{i!(20-i)!}(0.5)^i (1-0.5)^{20-i} \\
= & \ 2 \times (0.5)^{20} \sum_{i=0}^{3} \frac{20!}{i!(20-i)!} \\
= & \ 2 \times (0.5)^{20} \times \{1 + 20 + 20 \times 19/2 + 20 \times 19 \times 18/(2 \times 3)\} \\
= & \ 0.0026.
\end{aligned}$$

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

$$H_0 : \pi = 0.5 \qquad \text{against} \qquad H_1 : \pi > 0.5.$$

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 $\sum_{17 \le i \le 20} P_{H_0}(Y = i) = 0.0013$. Now the evidence against H_0 is even stronger.

On the other hand, if we test

$$H_0 : \pi = 0.5 \qquad \text{against} \qquad H_1 : \pi < 0.5.$$

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 now is $\sum_{i \leq 17} P_{H_0}(Y = i) = 1 - 0.0013 = 0.9987$.

Remark. We only reject H_0 if there is significance evidence in favour of H_1.

Two types of errors

Statistical tests are often associated with two kinds of errors, which are displayed in the table below.

		Decision Made	
		H_0 not rejected	H_0 rejected
True State of Nature	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

Remarks. 1. Ideally we would like to have a test that minimises the probabilities of making both types of errors, which unfortunately is not feasible.

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

3. We do not have an explicit control on the probability of Type II error. For a given significance level α, we choose a test statistic such that, hopefully, the probability of Type II error is small.

4. **Power**. The power function of the test is defined as

$$\beta(\theta) = P_\theta\{ H_0 \text{ is rejected}\}, \quad \theta \in \Theta_1,$$

i.e. $\beta(\theta) = 1-$ Probability of Type II error.

5. **Asymmetry**: null hypothesis H_0 and alternative hypothesis H_1 are not treated equally in a statistical test. The choice of H_0 is based on the subject matter concerned and/or technical convenience.

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

9.2 The Wald test

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

$$(\widehat{\theta} - \theta)/\mathrm{SE}(\widehat{\theta}) \xrightarrow{D} N(0,1), \qquad \text{as } n \to \infty.$$

Then under H_0, $(\widehat{\theta} - \theta_0)/\mathrm{SE}(\widehat{\theta}) \sim N(0,1)$ approximately.

The Wald test at the significance levet α: Let $T = (\widehat{\theta} - \theta_0)/\text{SE}(\widehat{\theta})$ be the test statistic. We reject H_0 against

$$H_1 : \theta \neq \theta_0 \text{ if } |T| > z_{\alpha/2} \text{ (i.e. the } p\text{-value} < \alpha\text{), or}$$
$$H_1 : \theta > \theta_0 \text{ if } T > z_{\alpha} \text{ (i.e. the } p\text{-value} < \alpha\text{), or}$$
$$H_1 : \theta < \theta_0 \text{ if } T < -z_{\alpha} \text{ (i.e. the } p\text{-value} < \alpha\text{),}$$

where z_α is the top-α point of $N(0, 1)$, i.e. $P\{N(0, 1) > z_\alpha\} = \alpha$.

Remark. Since the Wald test is based on the asymptotic normality, it only works for reasonably large n.

Example 2. To deal with the customer complaint that the amount of coffee in a Hilltop coffee bottle is less than the advertised 3 pounds, 20 bottles were weighed, yielding observations

2.82, 3.01, 3.11, 2.71, 2.93, 2.68, 3.02, 3.01, 2.93, 2.56,
2.78, 3.01, 3.09, 2.94, 2.82, 2.81, 3.05, 3.01, 2.85, 2.79

The sample mean and standard deviation:

$$\bar{X} = 2.897, \qquad S = 0.148$$

Hence $\mathrm{SE}(\bar{X}) = 0.148/\sqrt{20} = 0.033$. By the CLT, $(\bar{X} - \mu)/\mathrm{SE}(\bar{X}) \xrightarrow{D} N(0, 1)$.

To test $H_0 : \mu = 3$ vs $H_1 : \mu < 3$, we apply the Wald test with $T = (\bar{X} - 3)/SE(\bar{X}) = -3.121 < -z_{0.01} = -2.326$. Hence we reject $H_0 : \mu = 3$ at the 1% significance level.

We conclude that there is significant evidence which supports the claim that the coffee in a Hilltop coffee bottle is less than 3 pounds.

9.3 χ^2-distribution and t-distribution

χ^2-Distributions

Background. χ^2-distribution is one of the important distributions in statistics. It is closely linked with normal, t- and F-distributions. Inference for variance parameter σ^2 relies on χ^2-distributions. More importantly most goodness-of-fit tests are based on χ^2-distributions.

Definition. Let $X_1, \cdots, X_k$ be independent $N(0, 1)$ r.v.s. Let

$$Z = X_1^2 + \cdots + X_k^2 = \sum_{i=1}^{k} X_i^2.$$

The distribution of Z is called the χ^2-distribution with k degrees of freedom, denoted by $\chi^2(k)$ or χ_k^2.

We list some properties of the distribution χ_k^2 as follows.

1. χ_k^2 is a continuous distribution on $[0, \infty)$.

2. **Mean**: $EZ = kE(X_1^2) = k$.

3. **Variance**: $\text{Var}(Z) = 2k$.

   Due to the independence among X_i's,

$$\text{Var}(Z) = k\text{Var}(X_1^2) = k[E(X_1^4) - \{E(X_1^2)\}^2] = k\{E(X_1^4) - 1\}.$$

$$
\begin{aligned}
E(X_1^4) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4 e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^3 e^{-x^2/2}d(x^2/2) \\
&= -\frac{x^3}{\sqrt{2\pi}}e^{-x^2/2}\Big|_{-\infty}^{\infty} + \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2}dx = \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2}dx \\
&= \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-x^2/2}d(x^2/2) = -\frac{3x}{\sqrt{2\pi}}e^{-x^2/2}\Big|_{-\infty}^{\infty} + \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}dx \\
&= \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}dx \;=\; 3.
\end{aligned}
$$

4. If $Z_1 \sim \chi_k^2$, $Z_2 \sim \chi_p^2$, and Z_1 and Z_2 are independent, then $Z_1 + Z_2 \sim \chi_{k+p}^2$.

According to the definition, we may write

$$Z_1 = \sum_{i=1}^{k} X_i^2, \qquad Z_2 = \sum_{j=k+1}^{k+p} X_j^2,$$

where all X_i's are independent $N(0,1)$ r.v.s. Hence

$$Z_1 + Z_2 = \sum_{i=1}^{k+p} X_i^2 \sim \chi^2_{k+p}.$$

5. The probability density function of χ_k^2 is

$$f(x) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} & x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where

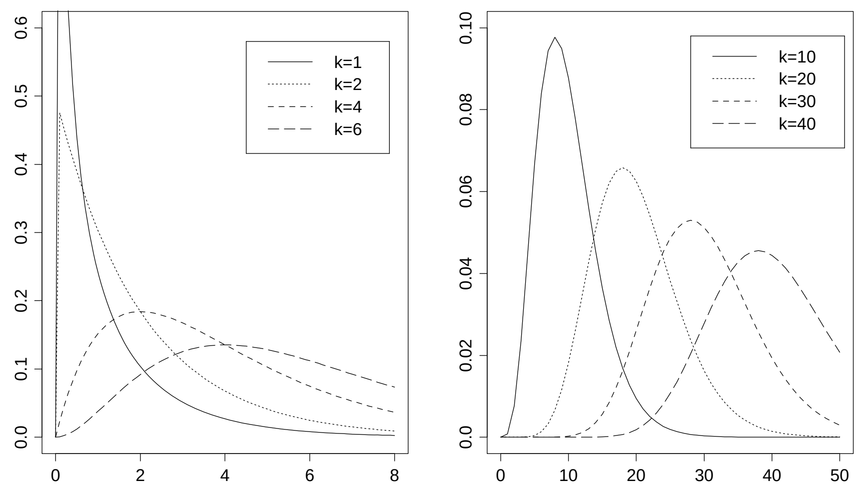
$$\Gamma(y) = \int_0^\infty u^{y-1} e^{-u} du.$$

For any integer k, $\Gamma(k) = (k-1)!$.

Hence χ_2^2 is the exponential distribution with mean 2, as its pdf is

$$f(x) = \begin{cases} \frac{1}{2} e^{-x/2} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Probability density functions of χ^2_k-distributions



6. The values of distribution functions of χ_k^2-distributions (for different k) have been tabulated, and can also be easily obtained from statistical packages such as R.

Let $Y_1, \cdots, Y_n$ be independent $N(\mu, \sigma^2)$ r.v.s. Then

$$(Y_i - \mu)/\sigma \sim N(0, 1).$$

Hence

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 \sim \chi^2_n.$$

Note that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \frac{n}{\sigma^2}(\bar{Y} - \mu)^2. \tag{1}$$

Since $\bar{Y} \sim N(\mu, \sigma^2/n)$, $\frac{n}{\sigma^2}(\bar{Y} - \mu)^2 \sim \chi^2_1$. It may be proved that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2_{n-1}.$$

Thus decomposition (1) may be formally written as

$$\chi^2_n = \chi^2_{n-1} + \chi^2_1.$$

Confidence Interval for σ^2

Let $\{X_1, \cdots, X_n\}$ be a random sample from population $N(\mu, \sigma^2)$.

Let $M = \sum_{i=1}^n (X_i - \bar{X})^2$. Then $M/\sigma^2 \sim \chi^2_{n-1}$.

For any given small $\alpha \in (0, 1)$, we may find $0 < K_1 < K_2$ such that

$$P(\chi^2_{n-1} < K_1) = P(\chi^2_{n-1} > K_2) = \alpha/2,$$

where χ^2_{n-1} stands for a r.v. with χ^2_{n-1}-distribution. Then

$$1 - \alpha = P(K_1 < M/\sigma^2 < K_2) = P(M/K_2 < \sigma^2 < M/K_1)$$

Hence an $100(1-\alpha)\%$ confidence interval for σ^2 is

$$(M/K_2,\ M/K_1).$$

Suppose $n = 15$ and the sample variance $S^2 = 24.5$. Let $\alpha = 0.05$.

From a table of χ^2-distributions, we may find

$$P(\chi^2_{14} < 5.629) = P(\chi^2_{14} > 26.119) = 0.025.$$

Hence a 95% confidence interval for σ^2 is

$$\begin{aligned}&(M/26.119,\ M/5.629) = (14S^2/26.119,\ 14S^2/5.629)\\ &= (0.536S^2,\ 2.487S^2) = (13.132,\ 60.934).\end{aligned}$$

In the above calculation, we have used the formula

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2 = \frac{1}{n-1}M = M/14.$$

Student's t-distribution

Background. Another important distribution in statistics

- The t-test is perhaps the most frequently used statistical test in application.
- Confidence intervals for normal mean with unknown variance may be *accurately* constructed based on t-distribution.

Historical note. The t-distribution was first studied by W.S. Gosset (1876-1937), who worked as a statistician for Guinness, writing under the pen-name 'Student'.

Definition. Suppose $X \sim N(0,1)$ and $Z \sim \chi_k^2$, and X and Z are independent. Then the distribution of the random variable

$$T = X \Big/ \sqrt{Z/k}$$

is called the t-distribution with k degrees of freedom, denoted by t_k or $t(k)$.

We now list some properties of the t_k distribution below.

1. t_k is a continuous and symmetric distribution on $(-\infty, \infty)$.

   (T and $-T$ share the same distribution.)

2. $E(T) = 0$ provided $E|T| < \infty$.

3. **Heavy tails**. If $T \sim t_k$, $E\{|T|^k\} = \infty$. For $X \sim N(\mu, \sigma^2)$, $E\{|X|^p\} < \infty$ for any $p > 0$. Therefore, t-distributions have heavier tails. This is a useful properties in modelling abnormal phenomena in financial or insurance data.

   Note. $E\{|T|^{k-\varepsilon}\} < \infty$ for any small constant $\varepsilon > 0$.

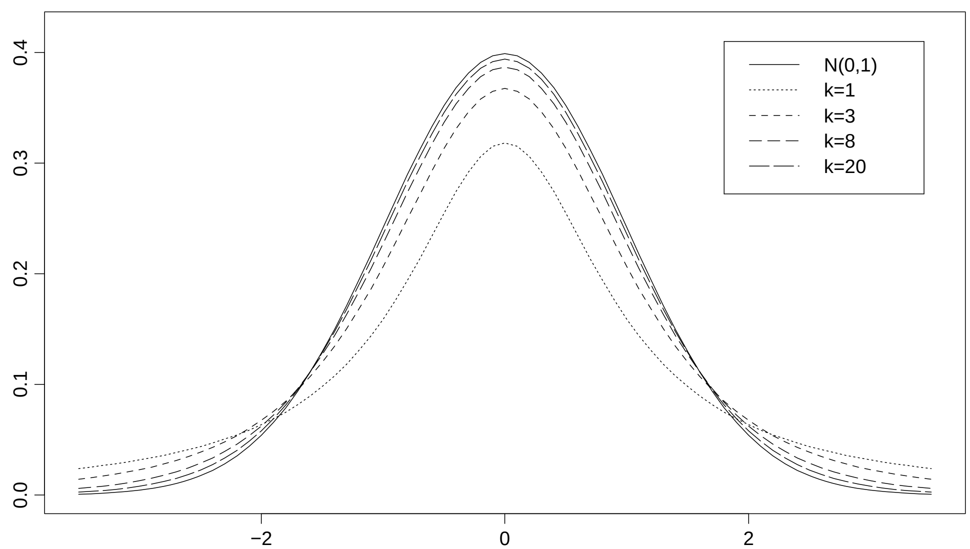
4. As $k \to \infty$, the distribution of t_k converges to the distribution of $N(0, 1)$.

   For $Z \sim \chi^2_k$, $Z = X_1^2 + \cdots + X_k^2$, where $X_1, \cdots, X_k$ are i.i.d. $N(0, 1)$. By the LLN, $Z/k \to E(X_1^2) = 1$. Thus $T = X/\sqrt{Z/k} \to X \sim N(0, 1)$.

5. The probability density function of t_k:

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}\big(1 + \frac{x^2}{k}\big)^{-\frac{k+1}{2}}.$$

Probability density functions of t_k-distributions



An important property of normal samples

Theorem. Let $\{X_1, \cdots, X_n\}$ be a sample from $N(\mu,\ \sigma^2)$. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \mathrm{SE}(\bar{X}) = \frac{S}{\sqrt{n}}.$$

Then

(i) $\bar{X} \sim N(\mu,\ \sigma^2/n)$, and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$,
(ii) $\bar{X}$ and S^2 are independent, and therefore

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} = \frac{\bar{X} - \mu}{\mathrm{SE}(\bar{X})} \sim t_{n-1}.$$

The t-interval — an accurate $(1-\alpha)$ confidence interval for μ:

$$\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = (\bar{X} - t_{\alpha/2,n-1}\cdot \mathrm{SE}(\bar{X}),\ \bar{X} + t_{\alpha/2,n-1}\cdot \mathrm{SE}(\bar{X})),$$

where $t_{\alpha/2,n-1}$ is a constant such that $P(t_{n-1} > t_{\alpha/2,n-1}) = \alpha/2$.

Proof of Theorem. Let $Y_i = (X_i - \mu)/\sigma$. Then $\bar{Y} = (\bar{X} - \mu)/\sigma$, and

$$S_y^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}) = S^2/\sigma^2.$$

Hence we only need to show that (a) $(n-1)S_y^2 \sim \chi^2_{n-1}$, and (b) $\bar{Y}$ and S_y^2 are independent.

As $\mathbf{Y} \equiv (Y_1, \cdots, Y_n)' \sim N(0, \mathbf{I}_n)$, it also holds that

$$\mathbf{Z} \equiv (Z_1, \cdots, Z_n)' \equiv \mathbf{\Gamma}\mathbf{Y} \sim N(0, \mathbf{I}_n) \quad \text{for any orthogonal } \mathbf{\Gamma}.$$

Let $(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}})$ be the first row of $\mathbf{\Gamma}$. Then $Z_1 = \sqrt{n}\bar{Y}$. Hence

$$(n-1)S_y^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Y}^2 = \sum_{i=2}^{n} Z_i^2 \sim \chi_{n-1}^2,$$

and it is independent of $Z_1 = \sqrt{n}\bar{Y}$. $\square$

The t-distributions with different degrees of freedom have been tabulated in all statistical tables.

The table below lists some values of C_α defined by the equation

$$P(t_k > C_\alpha) = \alpha$$

	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.005$
$k = 1$	6.314	12.706	63.657
$k = 2$	2.593	4.303	9.925
$k = 3$	2.353	3.182	5.841
$k = 10$	1.812	2.228	3.169
$k = 20$	1.725	2.086	2.845
$k = 120$	1.658	1.980	2.617
$\cdots$		$\cdots$	
$N(0, 1)$	1,645	1.960	2.576

Remark. When $k \geq 120$, $t_k \approx N(0, 1)$.

9.4 t-tests – one of the most frequently used tests in practice.

9.4.1 Tests for normal means – One-sample problems

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

$$H_0 : \mu = \mu_0 \qquad \text{against} \qquad H_1 : \mu \neq \mu_0,$$

where μ_0 is known.

The famous t-statistic:

$$T = \sqrt{n}(\bar{X} - \mu_0)/S = \sqrt{n}(\bar{X} - \mu_0) \Big/ \Big\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\Big\}^{1/2} = \frac{\bar{X} - \mu_0}{\mathrm{SE}(\bar{X})},$$

where $\bar{X} = n^{-1}\sum_i X_i$ and $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$. Note that under hypothesis H_0,

$$\sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0, 1), \qquad (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

Therefore

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} = \frac{\sqrt{n}(\bar{X} - \mu_0)/\sigma}{\sqrt{S^2/\sigma^2}} \sim t_{n-1} \quad \text{under } H_0.$$

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

Remark. $H_0 : \mu = \mu_0$ is rejected against $H_1 : \mu \neq \mu_0$ at the α significance level iff μ_0 lies outside the $(1 - \alpha)$ t-interval $\bar{X} \pm t_{\alpha/2,n-1}\text{SE}(\bar{X})$.

Example 2. (Continue) We use t-test to re-examine this data set. Recall

$$n = 20, \quad \bar{X} = 2.897, \quad S = 0.148, \quad \mathrm{SE}(\bar{X}) = 0.033,$$

we are interested in testing hypotheses

$$H_0 : \mu = 3, \qquad H_1 : \mu < 3.$$

We reject H_0 at the level α if $T < -t_{\alpha,19}$. Since $T = (\bar{X} - 3)/\mathrm{SE}(\bar{X}) = -3.121 < -t_{0.01,19} = -2.539$, we reject the null hypothesis $H_0 : \mu = 3$ at 1% significance level.

9.4.2 Tests for normal means – two-sample problems

Available two independent samples: $\{X_1, \cdots, X_{n_x}\}$ from $N(\mu_x, \sigma_x^2)$ and $\{Y_1, \cdots, Y_{n_y}\}$ from $N(\mu_y, \sigma_y^2)$. We are interested in testing

$$H_0 : \mu_x - \mu_y = \delta \quad \text{against} \quad H_1 : \mu_x - \mu_y \neq \delta \text{ (or } \mu_x - \mu_y > \delta \text{ etc)},$$

where δ is a known constant. Let

$$\bar{X} = \frac{1}{n_x}\sum_{i=1}^{n_x} X_i, \quad S_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}(X_i - \bar{X})^2,$$

$$\bar{Y} = \frac{1}{n_y}\sum_{i=1}^{n_y} Y_i, \quad S_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}(Y_i - \bar{Y})^2.$$

Then

$$\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \ \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}), \quad (n_x - 1)\frac{S_x^2}{\sigma_x^2} + (n_y - 1)\frac{S_y^2}{\sigma_y^2} \sim \chi^2_{n_x+n_y-2}.$$

With an addition assumption $\sigma_x^2 = \sigma_y^2$, it holds that

$$\sqrt{\frac{n_x + n_y - 2}{1/n_x + 1/n_y}} \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}} \sim t_{n_x+n_y-2}$$

Define a t-statistic

$$T = \sqrt{\frac{n_x + n_y - 2}{1/n_x + 1/n_y}} \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}}$$

The null hypothesis $H_0 : \mu_x - \mu_y = \delta$ is rejected against

$H_1 : \mu_x - \mu_y \neq \delta$ if $|T| > t_{\alpha/2,\, n_x+n_y-2}$, or
$H_1 : \mu_x - \mu_y > \delta$ if $T > t_{\alpha,\, n_x+n_y-2}$, or
$H_1 : \mu_x - \mu_y < \delta$ if $T < -t_{\alpha,\, n_x+n_y-2}$,

where $t_{\alpha,\, k}$ is the top-α point of the t_k-distribution.

Example 3. Two types of razor, A and B, were compared using 100 men in an experiment. Each man shaved one side, chosen at random, of his face using one razor and the other side using the other razor. The times taken to shave, X_i and Y_i minutes, $i = 1, \cdots, 100$, corresponding to the razors A and B respectively, were recorded, yielding

$$\bar{X} = 2.84, \quad S_x^2 = 0.48, \quad \bar{Y} = 3.02, \quad S_y^2 = 0.42.$$

Also available is the sample variance of the differences $Z_i \equiv X_i - Y_i$: $S_z^2 = 0.6$.

Test, at the 5% significance level, if the two razors lead to different shaving times. State clearly the assumptions used in the test.

Assumption. Suppose $\{X_i\}$ and $\{Y_i\}$ are two samples from, respectively, $N(\mu_x, \ \sigma_x^2)$ and $N(\mu_y, \ \sigma_y^2)$.

The problem requires to test hypotheses

$$H_0 : \mu_x = \mu_y \qquad \text{vs} \qquad H_1 : \mu_x \neq \mu_y.$$

There are three approaches: a pairwise comparison method, two two-sample comparisons based on different assumptions. Since the data are recorded pairwisely, the pairwise comparison is most relevant and effective to analyse this data

Method I: Pairwise comparison — one sample t-test

Note $Z_i = X_i - Y_i \sim N(\mu_z,\ \sigma_z^2)$ with $\mu_z = \mu_x - \mu_y$. We test

$$H_0 : \mu_z = 0 \qquad \text{vs} \qquad H_1 : \mu_z \neq 0.$$

This is the standard one-sample t-test,

$$\sqrt{n}\frac{\bar{Z} - \mu_z}{S_z} = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_z/\sqrt{n}} \sim t_{n-1}.$$

H_0 is rejected if $|T| > t_{0.025,\,99} = 1.98$, where

$$T = \sqrt{n}\bar{Z} \,/\, S_z = \sqrt{100}(\bar{X} - \bar{Y})/S_z.$$

With the given data, we observe $T = 10(2.84 - 3.02)/\sqrt{0.6} = -2.327$. Hence we reject the hypothesis that the two razors lead to the same shaving time.

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

$$\bar{X} - \bar{Y} \pm t_{0.025,\, n-1} S_z / \sqrt{n} = -0.18 \pm 0.154 = (-0.334,\ -0.026).$$

Remark. (i) Zero is not in the confidence interval for $\mu_x - \mu_y$.

(ii) $t_{0.025,\,99} = 1.98$ is pretty close to $z_{0.025} = 1.96$. **Indeed when n is large, the t-test and the Wald test are almost the same.**

Method II: Two sample t-test with equal but unknown variance

Additional assumption: two samples are independent, $\sigma_x^2 = \sigma_y^2$.

Now $\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y,\ \sigma_x^2/50)$, $99(S_x^2 + S_y^2)/\sigma_x^2 \sim \chi^2_{198}$. Hence

$$\frac{\sqrt{50}\{\bar{X} - \bar{Y} - (\mu_x - \mu_y)\}}{\sqrt{99(S_x^2 + S_y^2)/198}} = 10 \times \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_x^2 + S_y^2}} \sim t_{198}$$

Hence we reject H_0 if $|T| > t_{0.025,\,198} = 1.97$ where

$$T = 10(\bar{X} - \bar{Y})\Big/\sqrt{S_x^2 + S_y^2}.$$

For the given data, $T = -1.897$. Hence we <u>cannot</u> reject H_0.

A 95% confidence interval for $\mu_x - \mu_y$ contains 0:

$$(\bar{X} - \bar{Y}) \pm \frac{t_{0.025,\,198}}{10}\sqrt{S_x^2 + S_y^2} = -0.18 \pm 0.1870 = (-0.367,\ 0.007),$$

Method III: The Wald test — The normality assumption is not required. But the two samples are assumed to be independent. Note

$$\text{SE}(\bar{X} - \bar{Y}) = \sqrt{S_x^2/n_1 + S_y^2/n_2}.$$

Hence it holds approximately that

$$\{\bar{X} - \bar{Y} - (\mu_x - \mu_y)\}/\text{SE}(\bar{X} - \bar{Y}) \sim N(0,\ 1).$$

Hence, we reject H_0 when $|T| > 1.96$ at the 95% significance level, where

$$T = (\bar{X} - \bar{Y})/\sqrt{S_x^2/100 + S_y^2/100}.$$

For the given data, $T = -0.18/\sqrt{0.009} = -1.9$. Hence we <u>cannot</u> reject H_0.

An approximate 95% confidence interval for $\mu_x - \mu_y$ is

$$\bar{X} - \bar{Y} \pm 1.96 \times \sqrt{S_x^2/100 + S_y^2/100} = -0.18 \pm 0.186 = (-0.366,\ 0.006).$$

The value 0 is contained in the interval now.

Remarks. (i) Different methods lead to different but *not contradictory* conclusions, as

$$\text{Not reject} \neq \text{Accept}$$

(ii) The pairwise comparison is intuitively most relevant, and leads to most conclusive inference (i.e. rejection). It also produces the shortest confidence interval.
(iii) Methods II and III ignore the pairing of the data, and therefore fail to take into account the variation due to the different individuals. Consequently the inference is less conclusive and less accurate.
(iv) A general observation: H_0 is rejected iff the hypothesized value by H_0 is not in the corresponding confidence interval.
(v) It is much more challenging to compare two normal means with unknown and different variances, which is not discussed in this course. On the other hand, the Wald test provides an easy alternative when both n_x and n_y are large.

9.4.3. t-tests with *R*

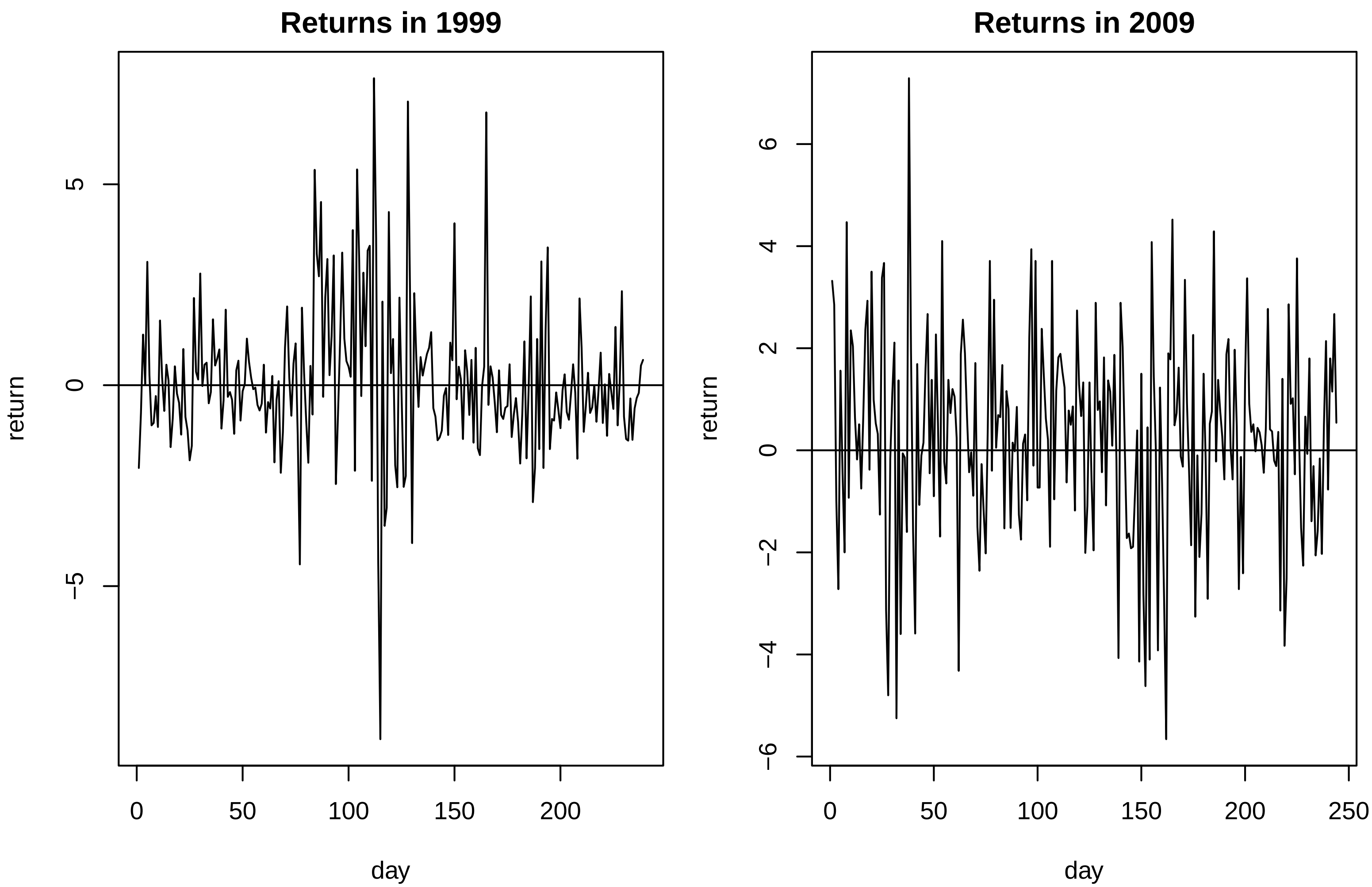
The *R*-function `t.test` performs one-sample, or two-sample t-tests with one-sided or two-sided alternatives. We illustrate it via an example.

Example 4. The daily returns of the Shanghai Stock Exchange Composite Index in 1999 and 2009: two subsets of the data analysed in Example 1 of Chapter 8.

(i) First we extract the two subsets and conduct some preliminary data analysis.

```
> x <- read.table("shanghaiSECI.txt", skip=3, header=T)
> y <- x[,4]*100   # daily returns in percentages
> y1999 <- y[1005:1243] # extract daily returns in 1999
> y2009 <- y[3415:3658] # extract daily returns in 2009
> par(mar=c(4,4,2,1),mfrow=c(1,2))
```

```r
> plot(y1999, type='l', xlab='day', ylab='return',
     main='Returns in 1999')
> plot(y2009, type='l', xlab='day', ylab='return',
     main='Returns in 2009')
> length(y1999); length(y2009)
[1] 239  # sample size of returns in 1999
[1] 244  # sample size of returns in 2009
> summary(y1999)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.8100 -0.8800 -0.1300  0.1037  0.7400  7.6400
> summary(y2009)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.6600 -0.8150  0.3650  0.2561  1.4780  7.2900
> var(y1999); var(y2009)
[1] 3.493598
[1] 3.922712
```


Returns in 1999
return
5
0
−5
0
50
100
150
200
day
Returns in 2009
return
6
4
2
0
−2
−4
−6
0
50
100
150
200
250
day

(ii) *One sample t-test*. Let X_i denote the returns in 1999, and Y_i denote the returns in 2009. Then $n_x = 239$, $n_y = 244$, and

$$\bar{X} = 0.1037, \quad \bar{Y} = 0.2561, \quad S_x^2 = 3.4936, \quad S_y^2 = 3.9227.$$

We test $H_0 : \mu_x = 0$ vs $H_1 : \mu_x > 0$ first.

```
> t.test(y1999, alternative='greater', mu=0, conf.level=0.95)
     # use alternative='two.sided' for two-sided alternative
        One Sample t-test
data:  y1999
t = 0.8576, df = 238, p-value = 0.196
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 -0.09596304         Inf
sample estimates:
mean of x
 0.103682
```

Since the p-value is 0.196, we cannot reject $H_0 : \mu_x = 0$, i.e. the returns in 1999 are not significantly different from 0.

Corresponding the one-sided alternative, R also gives a corresponding one-sided confidence interval for μ: $(-0.096, \infty)$, which contains 0. (Note that the setting indicates that we believe μ is either 0 or positive. Therefore reasonable confidence intervals are in the form (a, ∞).)

Now we test $H_0 : \mu_y = 0$ vs $H_1 : \mu_y > 0$.

```
> t.test(y2009, alternative='greater', mu=0, conf.level=0.99)
        One Sample t-test
data:  y2009
t = 2.0202, df = 243, p-value = 0.02223
alternative hypothesis: true mean is greater than 0
99 percent confidence interval:
 -0.04077725         Inf
sample estimates:
mean of x
0.2561475
```

For the returns in 2009, the p-value of the t-test is 0.022. Hence we reject $H_0 : \mu_y = 0$ at the 5% significance level, but cannot reject H_0 at the 1%

level. We conclude that there exists evidence indicating that the returns in 2009 tend to greater than 0, although the evidence is not overwhelming.

Remark. With the sample sizes over 200, the above t-tests yield practically the same results as the Wald test.

(iii) *Two-sample t-tests*. We now test $H_0 : \mu_x - \mu_y = 0$ against $H_1 : \mu_x - \mu_y \neq 0$ or $H_1 : \mu_x - \mu_y < 0$.

```
> t.test(y1999, y2009, mu=0, alternative='two.sided', var.equal=T)
      # without flag "var.equal=T", the Welch–Satterthwaite approximate
      # test will be used instead
        Two Sample t-test
data:  y1999 and y2009
t = -0.8697, df = 481, p-value = 0.3849
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 -0.4969197  0.1919887
sample estimates:
mean of x mean of y
0.1036820 0.2561475

> t.test(y1999, y2009, mu=0, alternative='less', var.equal=T)
        Two Sample t-test
data:  y1999 and y2009
t = -0.8697, df = 481, p-value = 0.1924
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1364386
sample estimates:
mean of x mean of y
0.1036820 0.2561475
```

Both the tests indicate that there is no significant evidence against the hypothesis that the average returns in the two years are the same.

9.5 Most Powerful Tests and Neyman-Pearson Lemma

Ideally we would choose, among those tests of size α, the test which minimises the probability of Type II error, i.e. that maximises the power $\beta(\theta)$ over $\theta \in \Theta_1$. If such a test exists, it is called *the most powerful test* (MPT).

Neyman-Pearson Lemma. If a test of size α for

$$H_0: \ \theta = \theta_0 \quad \text{against} \quad H_1: \ \theta = \theta_1$$

rejects H_0 when

$$L(\theta_1; \mathbf{x}) > K L(\theta_0; \mathbf{x}),$$

and does not reject H_0 when

$$L(\theta_1; \mathbf{x}) < K L(\theta_0; \mathbf{x}),$$

then it is a most powerful test of size α, where $K > 0$ is a constant.

Note. Both H_0 and H_1 are simple hypotheses.

Example 5 Let $X_1, X_2, \ldots, X_n$ be a sample from $N(\mu, 1)$. To test

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu = 5,$$

the likelihood ratio is

$$\text{LR} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^n (X_i - 5)^2/2\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^n X_i^2/2\right)} \propto \exp(5n\bar{X}).$$

Thus $LR > K$ is equivalent to $\bar{X} > K_1$, K_1 is determined by the size of the test. Thus the MPT of size α rejects H_0 iff $\bar{X} > z_\alpha/\sqrt{n}$, where z_α is a top-α point of $N(0, 1)$.

Question: If we change the alternative hypothesis to $H_1 : \mu = 10$, what is the MPT then?

Uniformly Most Powerful Tests

Suppose that the MPT for testing

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

does not change its form for all $\theta_1 \in \Theta_1$. Then it is the *Uniformly Most Powerful Test* (UMPT) for testing

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

Note. Typically, $\Theta_1 = (-\infty, \theta_0)$ or $\Theta_1 = (\theta_0, \infty)$.

Example 5 (continue). For

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu > 0,$$

the UMPT of size α rejects H_0 iff $\bar{X} > z_\alpha$.

A more general case

Let $\mathbf{X} = (X_1, \cdots, X_n)^T$ be random variables with joint pdf $f(\mathbf{x}, \theta)$. We test the hypotheses

$$H_0: \ \theta \leq \theta_0 \quad \text{vs} \quad H_1: \ \theta > \theta_0. \tag{2}$$

Denoted by $T \equiv T(\mathbf{X})$ the MPT of size α for simple hypotheses

$$H_0: \ \theta = \theta_0 \quad \text{vs} \quad H_1: \ \theta = \theta_1,$$

exists, where $\theta_1 > \theta_0$.

Then T is the UMPT of the same size α for hypotheses (2) provided that

(i) T remains unchaged for all values of $\theta_1 > \theta_0$, and

(ii) $P_\theta(T \text{ rejects } H_0) \leq P_{\theta_0}(T \text{ rejects } H_0) = \alpha$ for all $\theta < \theta_0$.

Note. For hypotheses $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$, the UMPT may be obtained in the similar manner.

Example 6. Let $(X_1, \cdots, X_n)$ be a random sample from an exponential distribution with mean $1/\lambda$. We are interested in testing

$$H_0: \lambda \leq \lambda_0 \quad \text{vs} \quad H_1: \lambda > \lambda_0.$$

For

$$H_0: \lambda = \lambda_0 \quad \text{vs} \quad H_1: \lambda = \lambda_1,$$

the MPT rejects H_0 iff $\sum_{i=1}^n X_i \leq K$ for any $\lambda_1 > \lambda_0$, where K is determined by $P_{\lambda_0}\{\sum_{i=1}^n X_i < K\} = \alpha$.

It is easy to verify that for $\lambda < \lambda_0$, $P_{\lambda}\{\sum_{i=1}^n X_i < K\} < \alpha$.

Hence the MPT for the simple null hypothesis against simple alternative is also the UMPT for the composite hypotheses.