Inferential Statistics
L5 - Hypothesis Testing

Erlis Ruli (erlis.ruli@unipd.it)

Department of Statistics, University of Padova

Contents

Problem statement

Suppose that the average energy consumption of our population of WMs mounting a standard motor is μ_0.

It's claimed that NG1 family motors would lead to more efficient WMs, i.e. would lead to average consumption μ, with $\mu < \mu_0$.

There are two possibilities:

- the claim is false, so $\mu \geq \mu_0$; this is called Null Hypothesis ("null" because it adds nothing to the current state of art)
- the claim is true, so $\mu < \mu_0$; it's called Alternative Hypothesis.

Problem statement (cont'd)

Concretely, suppose that $\mu_0 = 20$.

We equip 10 WM's with the NG1 motor and measure their E consumption.

Let these energy values be
19.1, 20.6, 17.3, 21.1, 19.5, 19.5, 21.4, 19.1, 20.5, 19.5.

Their average is 19.76. But $19.76 < \mu_0 = 20$, so the NG1 motor seems to lead on average to more efficient WMs!

Is that really so? Couldn't this be due to a pure luck?

A Simple formulation

Suppose the sample above is a realisation of the iid random sample $Y_1, \ldots, Y_n$ with $Y_i \sim N(\mu, 5)$.

This is a reasonable assumption given that overall energy consumption of a WM is the sum of the consumptions due to the various components of a WM (motor, resistor, etc.).

For this specific example we have (from L4) that

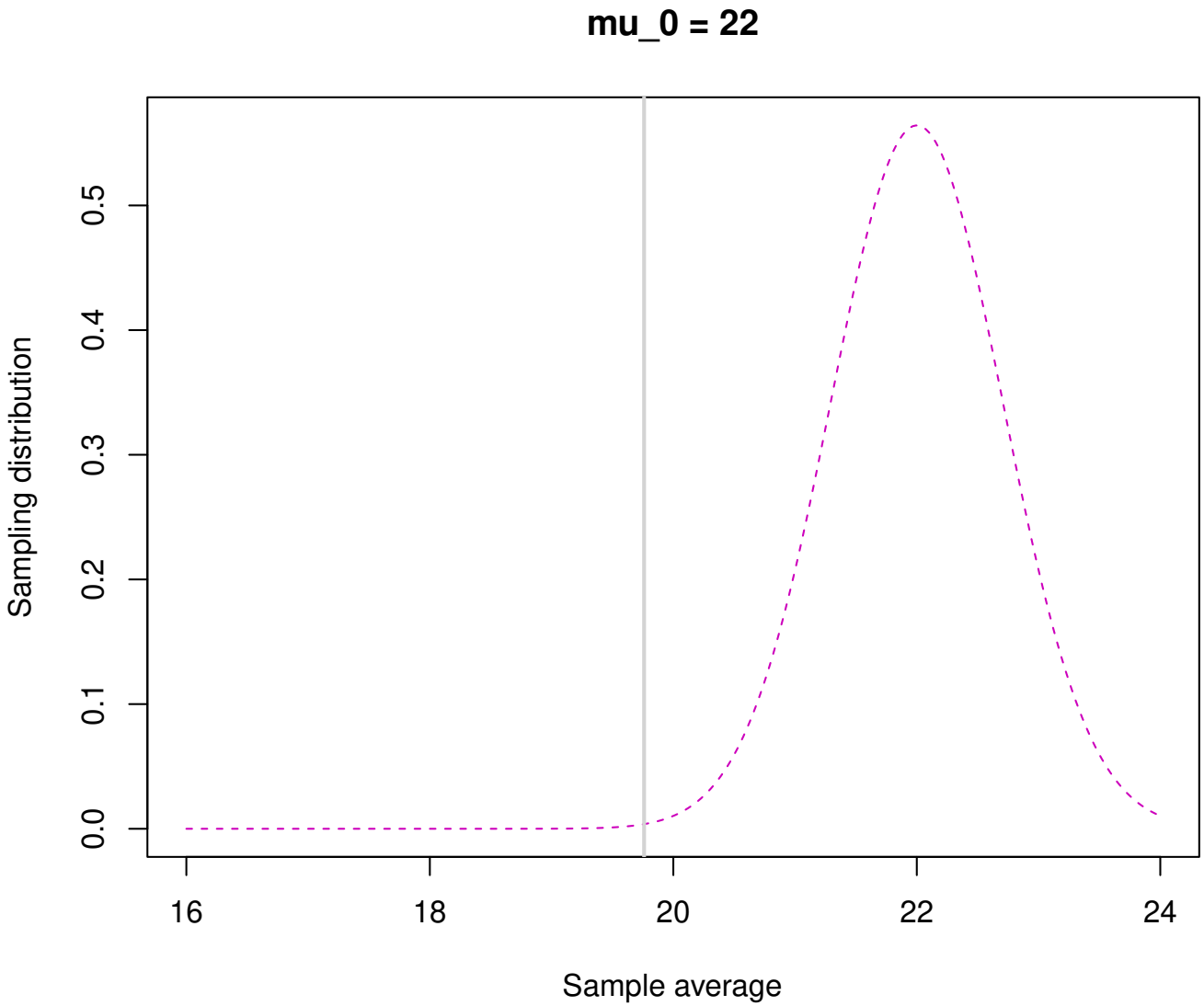
$$\frac{\sqrt{n}(\overline{Y}-\mu)}{\sqrt{5}} \sim N(0,1),$$

and thus

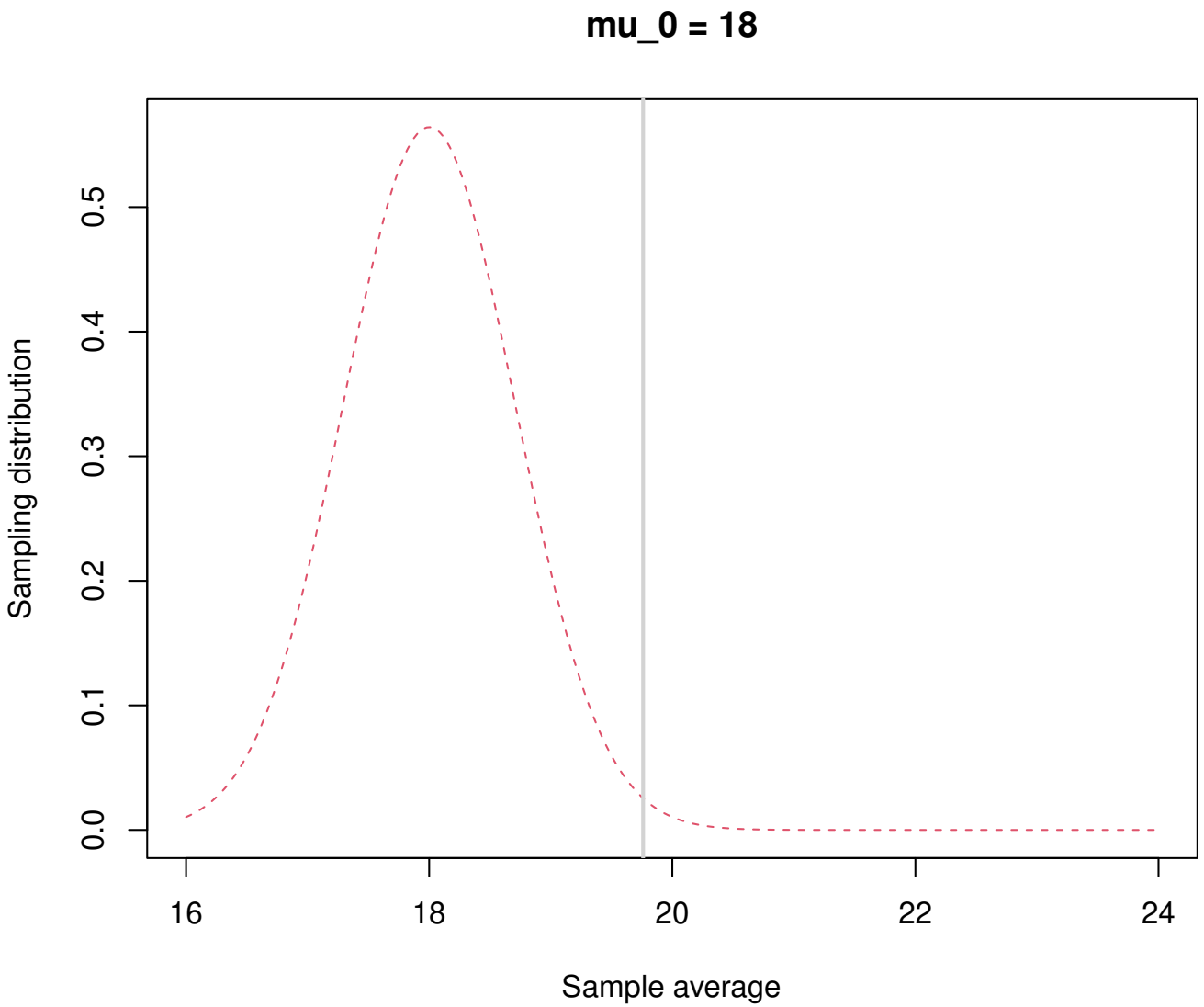
$$\overline{Y} \sim N(\mu, 0.5).$$

The following figure shows this distribution for several values of μ.

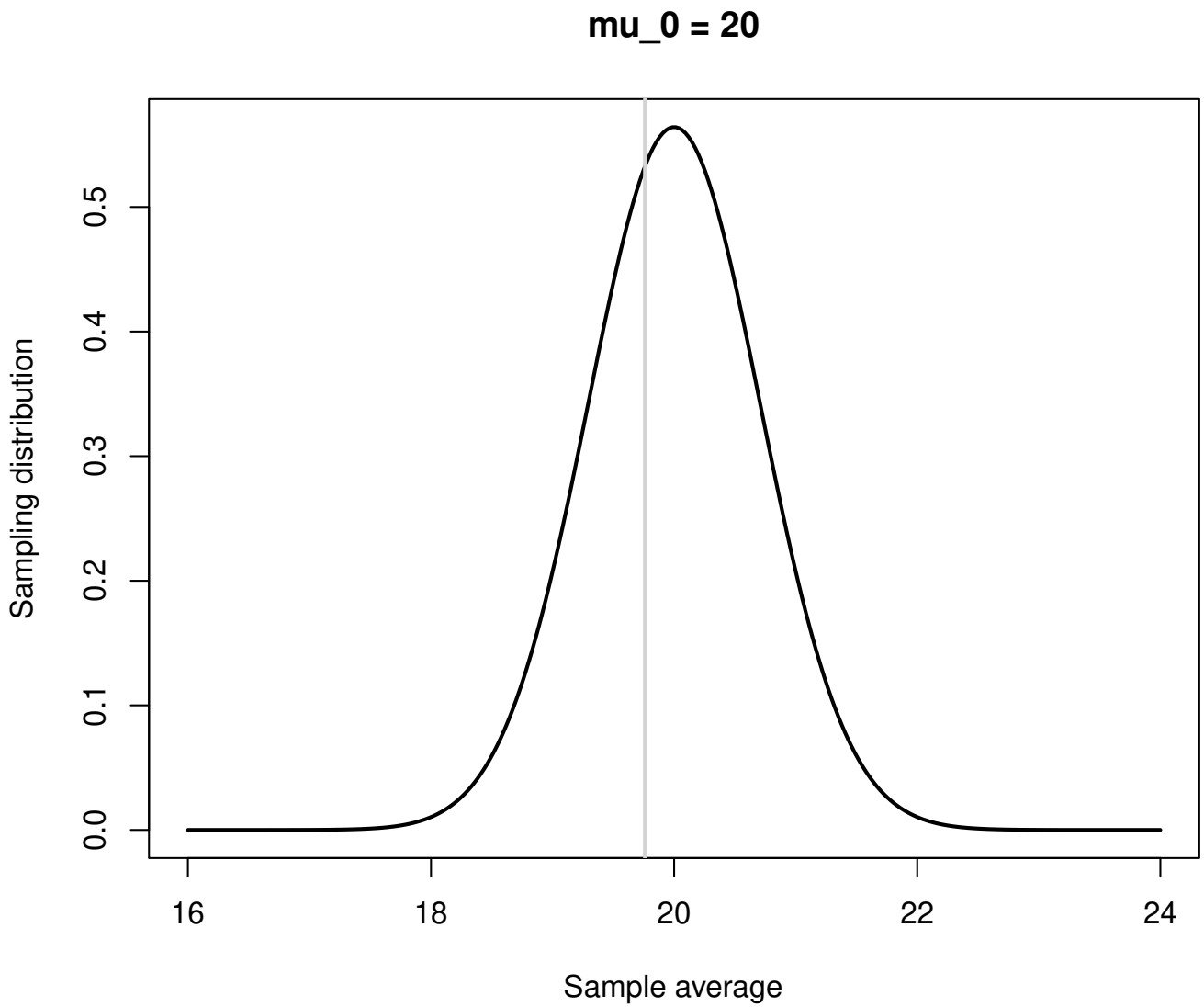
Sampling distribution of $\overline{Y}$



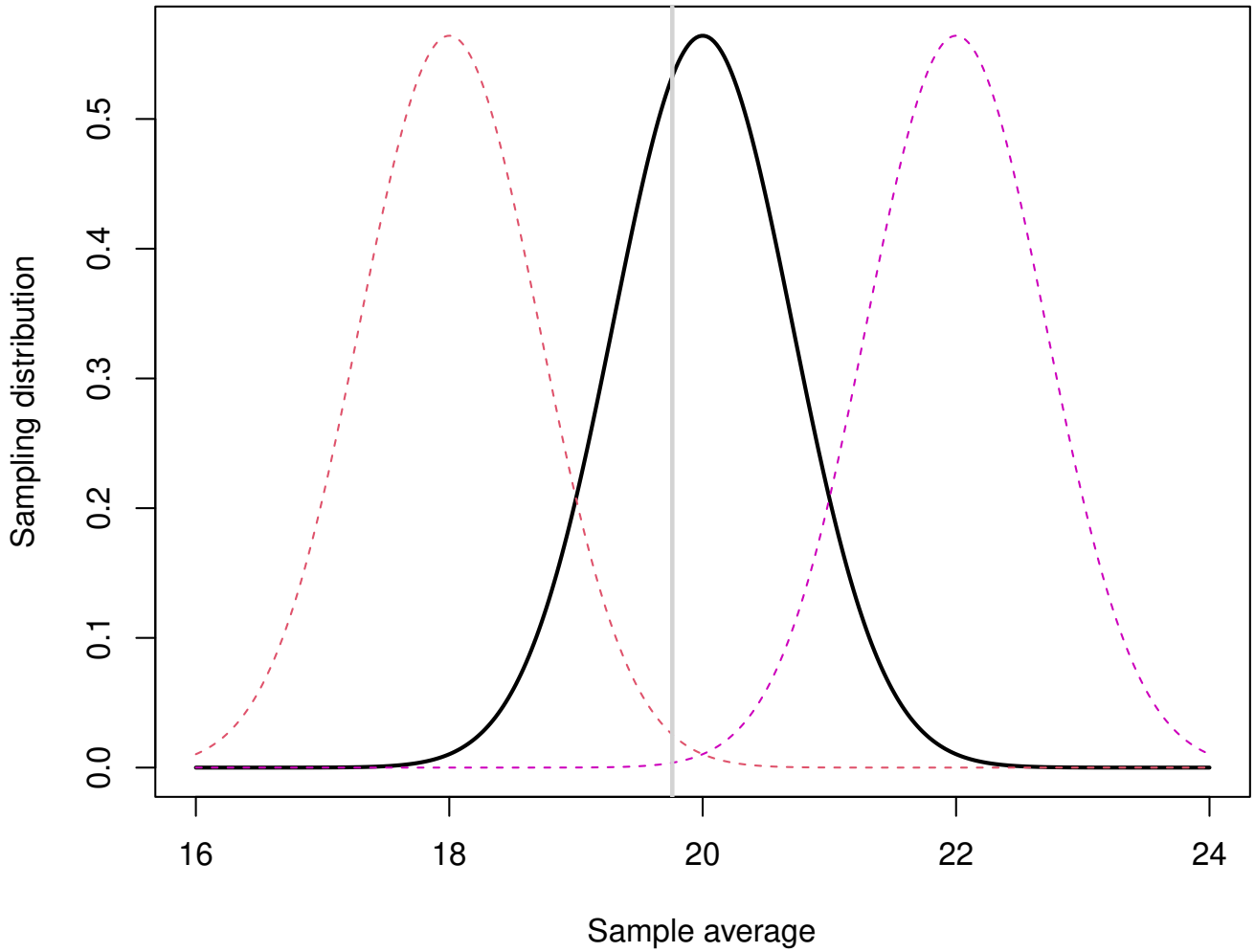
Sampling distribution of $\overline{Y}$



Sampling distribution of $\overline{Y}$



Sampling distributions of $\overline{Y}$



Test hypothesis and decision

Suppose that we judge surprising all values that under the sampling distribution are very unlikely to happen, say all those values with prob less than 0.01.

Under the sampling distribution with $\mu_0 = 20$, this value is 19.42, since $P_{\mu_0}(\overline{Y} \leq 19.42) = 0.01$.

We got a decision rule: a sample average < 19.42 is deemed surprising and should make us suspect about the worthiness of the null hypothesis, so we reject the null hypothesis. Otherwise we do not reject.

In the case above, the observed sample average was $\overline{y} = 19.76 > 19.42$, according to the rule, we should not be surprised and thus do not reject the null hypothesis.

Terminology

In the problem above we tested the null hypothesis $H_0 : \mu \geq \mu_0$ against the alternative $H_1 : \mu < \mu_0$.

In other situations we may be interested in testing

$$H_0 : \mu \leq \mu_0 \text{ against } H_1 : \mu > \mu_0 \qquad (*)$$

or

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0 \qquad (**)$$

Hypotheses s.t. (*) are called one-tailed, and (**) are called two-tailed.

That was too simple

In our motivating example we assumed population variance was known ($\sigma^2 = 5$). In practice, this assumption is not realistic and must be relaxed.

Furthermore, we assumed an iid random sample with Y_i following a normal distribution. But, in many problems, the normal distribution is not suitable.

The point is that, relaxing $\sigma^2 = 5$ or the distributional assumptions makes the above test useless; thus we have to look elsewhere. In other words, we need to know how to build a test for a given problem at hand.

In the rest of this lecture we will discuss two popular methods (both frequentist-parametric) and we will provide criteria for evaluating their performance.

Setting the scene

Let $Y_1, \ldots, Y_n$ be an iid random sample with $Y_i \sim F_\theta$, where $\theta \in \Theta$ is the unknown parameter with parameter space Θ.

We assume that F_θ has pdf f, indexed by the same parameter θ.

The iid assumption can be relaxed, but let's keep it simple for the moment.

We denote by $\mathcal{Y}$ the range of Y_i and by $\mathcal{Y}^n = \mathcal{Y} \times \mathcal{Y} \times \cdots \times \mathcal{Y}$, the Cartesian product of $\mathcal{Y}$ n times.

Setting the scene (cont'd)

Performing a statistical test essentially entails

building a decision rule from a sample to decide if reject

$$H_0 : \theta \in \Theta_0 \text{ in favour of } H_1 : \theta \in \Theta_0^c,$$

where $\Theta_0 \subset \Theta$.

Specifically, a test is a binary decision rule operating on a subset $R \subset \mathcal{Y}^n$ as follows:

reject H_0 if the observed sample $\mathbf{y} = (y_1, \ldots, y_n) \in R$, and accept H_0 otherwise.

Methods for computing statistical tests

A statistical test is typically defined on the basis of a test statistic $T(\mathbf{Y}) = T(Y_1, \ldots, Y_n)$ which is a function of the sample and closely related to a statistic seen in L4.

The test statistic used for computing the statistical test determines the nature and the name of the statistical test itself.

Likelihood Ratio Tests

The likelihood ratio test statistic for $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{y}) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}.$$

A likelihood ratio test (LRT) is any test that has rejection region

$$R = \{\mathbf{y} : \lambda(\mathbf{y}) < c\},$$

where c is any number s.t. $0 \leq c \leq 1$.

Recalling that $\widehat{\theta}$ is the MLE of θ and denoting by $\widehat{\theta}_0$ the constrained MLE of θ when the parameter space is Θ_0, then the LRT statistic is

$$\lambda(\mathbf{y}) = \frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}.$$

Example 1

Let $Y_i \sim N(\theta, 5)$, be an iid sample of size n and consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Here θ_0 is a number fixed prior to the experiment.

Under H_0 we have only one possible value for θ, thus the numerator of $\lambda(\mathbf{y})$ is $L(\theta_0)$. On the other hand the (unrestricted) MLE for θ is $\widehat{\theta} = \overline{y}$, the LRT statistic is

$$\begin{aligned}\lambda(\mathbf{y}) &= \frac{(10\pi)^{-n/2}\exp[-\sum_i (y_i-\theta_0)^2/10]}{(10\pi)^{-n/2}\exp[-\sum_i (y_i-\overline{y})^2/10]} \\ &= \exp[-n(\overline{y}-\theta_0)^2/10].\end{aligned}$$

The LRT is thus a test that rejects H_0 for small values of $\lambda(\mathbf{y})$, and the rejection region can be written as

$$\left\{\mathbf{y} : |\overline{y}-\theta_0| \geq \sqrt{-10(\log c)/n}, \right\}$$

for some $c \in (0, 1]$

Example 2

Let all be as in the previous example but $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$. These are the hypotheses we wanted to test in the motivating example in slide 3.

Discarding some additive constants, the log-likelihood function is

$$\begin{aligned} \ell(\theta) &= -\sum_{i=1}^{n}(y_i - \theta)^2/10 \\ &= -\sum_{i}(y_i - \overline{y})^2/10 - n(\overline{y} - \theta)^2/10, \end{aligned}$$

quadratic and concave. Thus, under H_0, $\sup \ell(\theta)$ is attained at

(I) θ_0 if $\theta_0 > \overline{y}$

(II) $\overline{y}$ if $\theta_0 \leq \overline{y}$.

The denominator of the LRT statistic is as before. So the likelihood ratio test statistic is

Example 2 (cont'd)

$$\lambda(\mathbf{y}) = \begin{cases} \exp[-n(\overline{y} - \theta_0)^2/10] & \text{if } \theta_0 > \overline{y} \\ 1 & \textit{otherwise} \end{cases}$$

and the rejection region for the LRT test is

$$R = \begin{cases} \{\mathbf{y} : \overline{y} < \theta_0 - \sqrt{-10(\log c)/n}\} & \text{if } \theta_0 > \overline{y} \\ \emptyset & \text{otherwise.} \end{cases}$$

Thus the test rejects if the sample average is lower than some threshold on the left of θ_0, precisely as we saw in slides 5-9.

We'll discuss the choice of c in the next section, but if take $c = 0.0977$, and go back to the example in slides 5-7 where $\widehat{\theta}_0 = 20$, $n = 10$, we have that

$$\overline{y} = 19.41 \not< \theta_0 - \sqrt{-10(\log c)/n} = 18.47$$

We cannot reject H_0, e.g.
there is no evidence about the efficiency claim of NG1.

Four possibilities

A decision may be wrong, indeed, there are four possibilities

Decision	**Truth**	
	$\theta \in \Theta_0$ (H_0 is true)	$\theta \notin \Theta_0$ (H_1 is true)
Reject H_0	type I error	ok
Don't reject H_0	ok	type II error

In a single test, we may either get a correct (ok) or a wrong decision. In the latter case, we could make a type I or type II error.

The size of type I error is defined by

$$\alpha' = \sup_{\theta \in \Theta_0} P(\text{reject } H_0 | H_0 \text{ is true}).$$

and the size of type II error is defined by

$$\beta(\theta) = 1 - P(\text{reject } H_0 | H_1 \text{ is true}) \ \forall \theta \in \Theta_0^c.$$

No free meal

Ideally, we'd like error-free decision, i.e.

$$\alpha' + \beta(\theta) = 0, \forall\, \theta,$$

but this is impossible.

Indeed, for $\alpha' = 0$ we must never reject H_0. But then, if H_1 is true, we make a type II error for sure, so $\beta(\theta) = 1$.

On the other hand, for $\beta(\theta) = 0$ we have to always reject H_0. But so doing and when H_0 is true, we make a type I error for sure, so $\alpha' = 1$.

Choosing the threshold

The current practice is to fix α at some small value (e.g. 0.01, or 0.05) and make sure that

$$\alpha' \leq \alpha,$$

without worrying about the value of $\beta(\theta)$.

The rationale for such a choice is that, often, making a type I error is more dangerous than making a type II error.

Fixing α entails fixing the amount of type I error, which entails fixing the size of the rejection region R, or the value of c in the case of LRT.

Example 3

Consider again Example 2 and let $\alpha = .01$. The rejection region is

$$R = \{\mathbf{y} : \overline{y} < \theta_0 - \sqrt{-10(\log c)/n}\}.$$

First note that

$$\overline{Y} \sim N(\theta_0, 10/n),$$

thus

$$\begin{aligned} P_{\theta_0}(\mathbf{Y} \in R) &= P_{\theta_0}\left(\overline{Y} < \theta_0 - \sqrt{-10(\log c)/n}\right) \\ &= P(Z \leq -\sqrt{-\log c}) \leq \alpha \end{aligned}$$

Solving the last inequality, gives $z_\alpha = -2.326 = -\sqrt{-\log c}$, so $c = .0977$; here z_α denotes the quantile of level α of the standard normal distribution.

The level and the size of a test

Sometimes we may want to distinguish between cases when the inequality on the size of type I error can be reached from cases in which it cannot be reached.

Indeed

- if $\alpha' = \alpha$ the test is called a test of size α.
- If $\alpha' \leq \alpha$, we call it a test of level α.

Often, especially when F_θ is discrete, a test of level α is the best result we can get.

However, in most practical cases, we can only compute tests of size α as $n \to \infty$; these are called asymptotic tests.

Example 4 (A Poisson test)

Let $Y_1, \ldots, Y_n$ be an iid random sample from the Poi(θ), with θ unknown. Suppose we want to test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. The LRT statistic is

$$\lambda(\mathbf{y}) = \frac{e^{-n\theta_0}\theta_0^{\sum_i y_i}\prod_i (y_i!)^{-1}}{e^{-n\overline{y}}\overline{y}^{\sum_i y_i}\prod_i (y_i!)^{-1}}$$
$$\exp\left(-\sum_i y_i - n\theta_0\right)(\theta_0/\overline{y})^{\sum_i y_i}.$$

In this case it is not possible to determine c exactly since $P_{\theta_0}(\mathbf{Y} \in R)$ is not computable analytically. But we can build a different test as follows.

First, note that under H_0, $n\overline{Y} \sim$ Poi($n\theta_0$). Thus, if we fix a large threshold based on this distribution, we may judge "suspicious" values of $\overline{y}$ above this threshold.

A Poisson test (cont'd)

The simplest method is to set as a threshold a quantile of $\text{Poi}(n\theta_0)$ of level $\leq \alpha$. The test is thus:
reject H_0 if $n\overline{y}$ is greater than the threshold.

Concretely, let $\theta_0 = 1$, $\alpha = .05$ and let the observed sample be $0, 0, 3, 5, 7$. We see that $n\overline{y} = 15$.

Because $\sum_i Y_i \sim \text{Poi}(5)$, a threshold is 9 (We will see in the next section why choosing any value < 9, is not ok). Because $15 > 9$, we reject H_0.

Note that, in this particular example $\alpha' = .031 < .05$ and thus the test is only of level .05.

It is possible to improve this test to have size α through a technique known as randomisation; but we won't see randomised tests in this course.

Back to LRT

In many cases the LRT statistic doesn't have a known distribution, but it has a limiting distribution as n diverges. This is indeed, the reason why the LRT is so widely used in practice.

Theorem 5

Suppose $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$ *and let* $\Theta_0 \subset \Theta$ *s.t.*

$$\Theta_0 = \{\theta : \theta_{q+1} = \theta_{0,q+1}, \theta_{q+2} = \theta_{0,q+2}, \ldots, \theta_r = \theta_{0,r}\}.$$

Under $H_0 : \theta \in \Theta_0$ *and suitable regularity conditions,*

$$-2 \log \lambda(\mathbf{Y}) \xrightarrow{d} \chi^2_{r-q} \quad \text{as } n \to \infty.$$

The degrees of freedom in the limiting distribution are $r - q = dim(\Theta) - dim(\Theta_0)$*;* $dim(S)$ *denotes the dimension of the space* S*.*

Few notes before going on

Strictly speaking, this is an asymptotic result, i.e. valid only in the limit as $n \to \infty$ so this LRT test is guaranteed to have size α in the limit.

In practice we work with a finite n. Nevertheless, if n is "high enough", $-2\log\lambda$ will have a distribution close to χ^2_{r-q}. Thus, in practice, we read "$\xrightarrow{d}$" as "$\dot{\sim}$".

The distance between the distributions $-2\log\lambda$ and χ^2_{r-q} for a fixed n, depends on many factors (number of parameters to estimate, degree of dependence in the sample, etc.) Roughly speaking, the larger n/r the smaller is this distance.

Example 6 (LRT in full action)

Back to Example 4 and using the above theorem, for large n

$$P(\lambda(\mathbf{Y}) < c) \doteq P_{\theta_0}(-2\log\lambda(\mathbf{Y}) < -2\log(c)).$$

Equating the last probability with α we have

$$-2\log c = \chi^2_{1,1-\alpha} \implies c = \exp\left(-\tfrac{1}{2}\chi^2_{1,1-\alpha}\right),$$

where $\chi^2_{1,1-\alpha}$ denotes the quantile of level $1-\alpha$ of the χ^2_1 distribution. For $\alpha = 0.05$, $\chi^2_{1,1-\alpha} = 3.84$, so $c = 0.1465$.

Because (check!) $\lambda(\mathbf{y}) = 0.0015 < c$, we reject H_0 (again).

The Wald test

Is useful for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, when there is an estimator for $\widehat{\theta}$ that has (at least a limiting) normal distribution.

We saw in L4, that, under regularity conditions, the MLE has a limiting normal distribution. Indeed, the Wald test is typically used in conjunction with the MLE.

Formally, for a scalar parameter, recall that the MLE is s.t.

$$\widehat{\theta} \dot{\sim} N\left(\theta, I_n(\widehat{\theta})^{-1}\right).$$

The Wald test of approximate level α is then to reject $H_0 : \theta = \theta_0$ if

$$|\text{Wald test statistic}| = |W| = \left|\frac{\widehat{\theta}-\theta_0}{\widehat{\text{se}}}\right| \geq z_{1-\alpha/2},$$

where $\widehat{\text{se}} = \sqrt{1/I_n(\widehat{\theta})}$ is the estimated standard error of $\widehat{\theta}$; the asymptotically equivalent version with J in place of I may be used.

Example of Wald test, use the previous example

Example 7

An IT-alert message test was sent to some of the residents in regione Veneto on 21st September 2023. It was claimed that the message was sent to roughly 1/3 of the population. Suppose we take random sample of adults in the regione Veneto and ask them if they received the message or not.

So let $X_1, \ldots, X_n$ be an iid sample with $X_i \sim \text{Ber}(\theta)$ ($X_i = 1$ if message received). We wish to test $H_0 : \theta = 1/3$ vs $H_1 : \theta \neq 1/3$.

The MLE of θ is $\widehat{\theta} = \sum_i X_i/n$. A Wald test of approximate level $\alpha = 0.05$ has $\widehat{\text{se}} = \sqrt{\widehat{\theta}(1-\widehat{\theta})/n}$, thus

$$W = \frac{\sqrt{n}(\widehat{\theta} - 1/3)}{\sqrt{\widehat{\theta}(1-\widehat{\theta})}},$$

and we reject H_0 provided $|W| > 1.96$. (Exercise. Conduct a small survey and use the data to test the above hypothesis.)

Wald with mulitvariate Θ

Let now θ be a $p \times 1$ parameter and the model F_θ is regular and s.t. the MLE has a normal limiting distribution, i.e.

$$\widehat{\theta} \dot{\sim} N_p(\theta, \widehat{I}^{-1}).$$

Furthermore, let θ_i denotes the ith component of θ and $\widehat{\theta}_i$ denotes the ith component of $\widehat{\theta}$.

To test the hypotheses $H_0 : \theta_i = \theta_{i0}$ against $H_1 : \theta_i \neq \theta_{i0}$, at the level α, where θ_{i0} is a scalar fixed before observing the data, by the Wald test is to
reject H_0 if $\left|\frac{\widehat{\theta}_i - \widehat{\theta}_{i0}}{\widehat{\text{se}}_i}\right| > z_{1-\alpha/2}$;
here $\widehat{\text{se}}_i = \sqrt{\widehat{I}^{ii}}$, or its equivalent version based on $\widehat{J}$.

Example 8

Suppose we to compare the performance of two ML classification algorithms. Algo1 was run on a test set of m observations and Algo2 was run on a test set of size n.

Let X be the number of misclassifications with Algo1 and Y those with Algo2. Assuming the sample is independent, then $X_1, \ldots, X_m$ are iid Ber(θ_1) and $Y_1, \ldots, Y_n$ are iid Ber(θ_2).

The hypotheses of interest are then $H_0 : \theta_1 = \theta_2$ vs $H_1 : \theta_1 \neq \theta_2$. Let $\delta = \theta_1 - \theta_2$, then above hypotheses translate to $H_0 : \delta = 0$ vs $H_1 : \delta \neq 0$. Let's apply a Wald test to δ. First, we estimate θ_1, θ_2 by MLE and then estimate δ using the equivariance principle by $\widehat{\delta} = \widehat{\theta}_1 - \widehat{\theta}_2$.

By the properties of the MLE, we have that $\widehat{\delta}$ is normally distributed and its standard deviation is approximately

$$\widehat{\text{se}}(\widehat{\delta}) = \sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{m} + \frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n}}$$

Example 8 (cont'd)

The Wald test of level α is then to reject H_0 if $|\widehat{\delta}/\widehat{\text{se}}(\widehat{\delta})| > z_{1-\alpha/2}$.

In a practical experiment conducted on $m = 20, n = 30$ with Algo1 and Algo2 we obtained 5 and 10 misclassifications. Are the two algorithms performing equally good (or bad)?

In this case the observed Wald statistic is $w = -0.64$.

Since $w < 1.96$ we do not reject H_0 and conclude that there is no evidence to support hat the two algorithms have different classification performance.

Example 9

Consider $X_1, \ldots, X_n$ an iid random sample with $X_i \sim N(\mu, \sigma^2)$, where σ^2 is known and suppose we wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Let's compute α' and the power of the LRT.

The LRT at significance level α is to reject H_0 if $|\sqrt{n}(\overline{X} - \mu_0)/\sigma| > z_{1-\alpha/2}$. Now

$$\begin{aligned}\alpha' = \inf_{\mu \leq \mu_0} P_\mu(\mathbf{X} \in R) &= \inf_{\mu \leq \mu_0} P_\mu(|\sqrt{n}(\overline{X} - \mu)/\sigma| > z_{1-\alpha/2}) \\ &= P(|Z| > z_{1-\alpha/2}) = \alpha.\end{aligned}$$

On the other hand, for any $\mu > \mu_0$

$$\begin{aligned}\beta(\mu) = P_\mu(\mathbf{X} \notin R) &= P_\mu(|\sqrt{n}(\overline{X} - \mu_0)/\sigma| \leq z_{1-\alpha/2}) \\ &= \Phi(z_{1-\alpha/2} + \sqrt{n}(\mu_0 - \mu)/\sigma) - \Phi(-z_{1-\alpha/2} + \sqrt{n}(\mu_0 - \mu)/\sigma)\end{aligned}$$

Example 9 (cont'd)

The power of the test is $1 - \gamma(\mu)$ and depends on (i) n, (ii) σ^2, and (iii) μ, besides α.

The larger n, the larger the power, indeed $\lim_{n\to\infty} \gamma(\theta) = 1$.

The power is also larger when μ_0 is distant from μ.

The practical importance of this is that the larger the sample size the more likely it is to reject H_0 correctly.

Pearson's χ^2 test

Let $(X_1, \ldots, X_k) \sim \text{Mn}(n, \theta_1, \ldots, \theta_k)$, then $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$, where $\widehat{\theta}_i = X_i/n$, where $n = \sum_i X_i$. Suppose we want to test

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

where $\theta_0 = (\theta_{01}, \ldots, \theta_{0k})$ is a fixed vector prior to observing the data.

Consider the statistic

$$T = \sum_{i=1}^{k} \frac{(X_i - n\theta_{0i})^2}{n\theta_{0i}}.$$

It can be shown that $T \xrightarrow{d} \chi^2_{k-1}$ as $n \to \infty$. The Person χ^2 test of approximate size α is then to
reject H_0 if the observed value of T is greater than $\chi^2_{k-1,1-\alpha}$.

Example 10

It is conjectured that in the human population, 48% have blood type 0, 38% have type A, 10% have type B and 7% have type AB. To test this hypothesis at a level $\alpha = 0.05$ a sample of n people is taken, where n_1 have type 0, n_2 have type A, n_3 have type B and n_4 have type AB.

The observed test statistic is

$$T^{obs} = \frac{(n_1 - 0.48n)^2}{0.48n} + \frac{(n_2 - 0.38n)^2}{0.38n} + \frac{(n_3 - 0.10n)^2}{0.10n} + \frac{(n_4 - 0.07n)^2}{0.07n}.$$

We reject H_0 if $T^{obs} > 7.815$.

An alternative view

Summing up, a test statistics is performed in three steps:

(i) identify a test statistic for the parameter of interest and build a suitable rejection region R;

(ii) compute the test statistic at the observed sample to get the observed test statistic, say T^{obs};

(iii) if T^{obs} follows in R, reject H_0 otherwise accept it.

Many scholars do not find this binary choice very informative; they prefer to compute a p-value and take an action based on this.

The p-value

The p-value is defined as the smallest α that leads to reject H_0. More formally, for every $\alpha \in (0, 1)$, if R_α is a rejection region of size α, then

$$p\text{-value} = \inf\{\alpha : T(X_1, \ldots, X_n) \in R_\alpha\}.$$

A 'small' p-value indicates that H_0 is not compatible with the data and it must be rejected.

A p-value is "small" when lower than α, the size of type I error.

Concretely, if $T_n = T(X_1, \ldots, X_n)$ is a test statistic with observed value $T^{obs} = T(x_1, \ldots, x_n)$ and a size α test has rejection region of the form

$$\{X_1, \ldots, X_n : T(X_1, \ldots, X_n) \geq c\},$$

then the p-value is defined by

$$\sup_{\theta \in \Theta_0} P_\theta \left(T(X_1, \ldots, X_n) \geq T^{obs} \right).$$

On the other hand, if the rejection region is of the form

$$\{X_1, \ldots, X_n : T(X_1, \ldots, X_n) \leq c\},$$

then the p-value is defined by

$$\sup_{\theta \in \Theta_0} P_\theta \left(T(X_1, \ldots, X_n) \leq T^{obs} \right).$$

Finally, if the rejection region is of the form

$$\{X_1, \ldots, X_n : T(X_1, \ldots, X_n) \geq c_1\} \cup \{X_1, \ldots, X_n : T(X_1, \ldots, X_n) \leq c_2\},$$

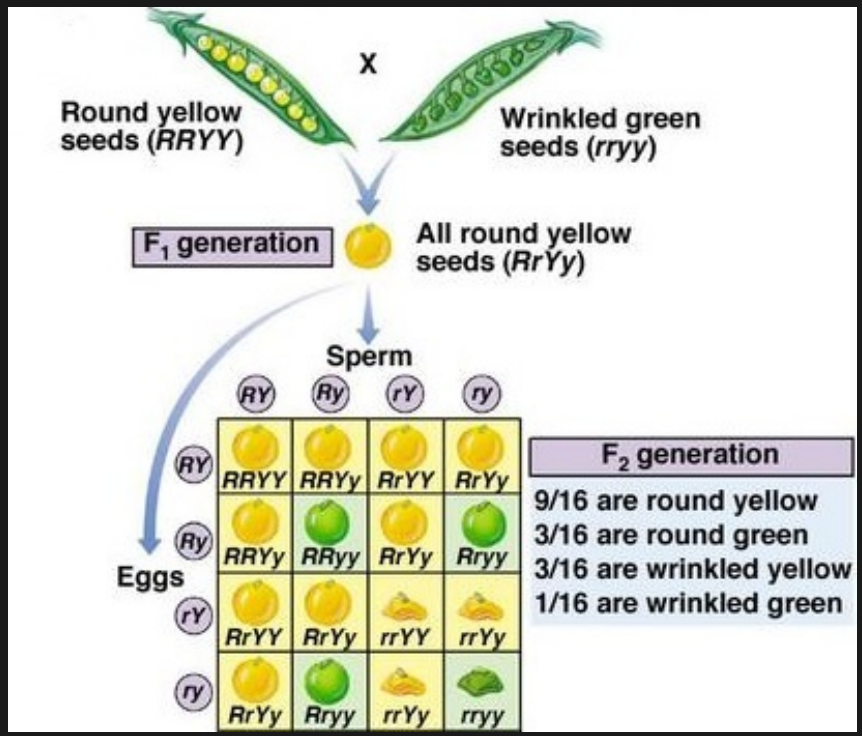
then

$$p\text{-value} = 2 \min \left(\sup_{\theta \in \Theta_0} P_\theta(T(X_1, \ldots, X_n) \leq T^{obs}), \sup_{\theta \in \Theta_0} P_\theta(T(X_1, \ldots, X_n) \geq T^{obs}) \right)$$

If $\Theta_0 = \{\theta_0\}$ then replace $\sup_{\theta \in \Theta_0} P_\theta$ by P_{θ_0}.

Example 11

Consider Mendel's experiment on peas, where round yellow seeds are breed with wrinkled green seeds. There are four types of progeny: round yellow, wrinkled yellow, round green, wrinkled green.



Example 11 (cont'd)

Let (Y_1, Y_2, Y_3, Y_4) be vector with the numbers seeds of the four types. Then follows a multinomial distribution with parameter $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$.

Mendel's theory of inheritance predicts

$$\theta = \theta_0 = \left(\tfrac{9}{16}, \tfrac{3}{16}, \tfrac{3}{16}, \tfrac{1}{16}\right).$$

In $n = 556$ his trials he obtained $\mathbf{y} = (315, 101, 108, 32)$.

Let's test if his theory is supported by the data using the LRT test. Let

$$H_0 : \theta = \theta_0 \text{ and } H_1 : \theta \neq \theta_0,$$

so at the observed data

$$\lambda(\mathbf{y}) = -2 \log\left(\frac{L(\widehat{\theta_0})}{L(\widehat{\theta})}\right) = 0.48,$$

and p-value $= P(\chi^2_3 \geq 0.48) = 0.92$. (There are four parameters, but only three are free to vary and under H_0 they are all fixed, thus df=3)

A testing problem can often be solved by means of several tests. The following two criteria are (the most widely) used for choosing the best one:

- prefer tests with smallest type I error, e.g. smallest α' but still $\alpha' \leq \alpha$;
- prefer tests with smallest type II error $\beta(\theta)$, or equivalently, prefer tests with highest power

$$\gamma(\theta) = 1 - \beta(\theta), \ \forall \theta \in \Theta_0^c.$$

Due to the trade-off between the type I and type II errors, a test with α' much smaller than α will most likely have small power.

In some particular cases (simple null vs simple alternative) it is possible to show that a test based on the LRT statistic is the most powerful among all test; this result is known as the Newman-Pearson Lemma.

In general, this optimality result is hard or impossible to apply. However,

it turns out that the LRT test and the Wald test tend to have decent power when n is sufficiently larger than the number of parameters.

To tell which test outperforms which in a given problem one often has to resort Monte Carlo methods as analytical calculations are typically impossible.

Example 12 (The t-test)

Let $X_1, \ldots, X_n$ be an iid random sample from $N(\mu, \sigma^2)$, with both parameters unknown; thus $\theta = (\mu, \sigma^2)$. We wish to test the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ via the LRT test.

Under H_0 we have that

$$\sup_{\theta \in \Theta_0} L(\theta) = \frac{\exp\left[-\frac{1}{2\widehat{\sigma^2_{\mu_0}}} \sum_{i=1}^n (X_i - \mu_0)^2\right]}{(2\pi)^{n/2} \widehat{\sigma^2_{\mu_0}}^{n/2}},$$

where $\widehat{\sigma^2_{\mu_0}} = \sum_{i=1}^n (X_i - \mu_0)^2/n$. Under H_1 we have

$$\sup_{\theta \in \Theta} L(\theta) = \frac{e^{-n/2}}{(2\pi)^{n/2}} \left[\frac{\sum_i (X_i - \overline{X})^2}{n}\right]^{-n/2}.$$

The LRT test is thus

Example 12 (cont'd)

$$\frac{\sum_i (X_i-\mu_0)^2}{\sum_i (X_i-\overline{X})^2} \geq b \iff \frac{n(\overline{X}-\mu_0)^2}{\sum_i (X_i-\overline{X})^2} \geq b-1$$
$$\iff \frac{n(\overline{X}-\mu_0)^2(n-1)}{\sum_i (X_i-\overline{X})^2} \geq (b-1)(n-1)$$
$$\iff \frac{n(\overline{X}-\mu_0)^2}{S^2} = T^2 \geq (b-1)(n-1) = d$$

The rejection region for the LRT is of the type

$$\{\mathbf{X} : T_n^2 \geq d\} \equiv \{\mathbf{X} : |T_n| \geq \sqrt{d} = a\}$$

In order to define a size α test we have to find a such that

$$\inf_{\theta \in \Theta_0} P_\theta(|T_n| \geq a) \leq \alpha.$$

But $T_n \sim t_{n-1}$, thus choosing $a = t_{n-1,1-\alpha/2}$ fills the bill.

Example 12 (cont'd)

The test is thus:

Reject H_0 if $|T_{obs}|$ is greater than the quantile of level $1-\alpha/2$ of the t_{n-1} distribution,

where $T^{obs} = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ is the observed t-statistic. This is also known as the t-test.

The p-value for this test is $2\,\mathbf{min}\left(P(t_{n-1} > T^{obs}), P(t_{n-1} < T^{obs})\right)$.

As a numerical example, suppose that an observed sample of $n = 10$ WM's lead to $\overline{x} = 201$ and $s^2 = 5^2$ and suppose we wish to test $H_0 : \mu = 200$ against $H_1 : \mu \neq 200$ at the level $\alpha = .05$. Then

$$T_n^{obs} = \frac{\sqrt{10}(201-200)}{5} = .632.$$

Since $t_{9,0.975} = 2.26 \not< 632$, we do not reject H_0 at level $\alpha = .05$. The p-value is $2P(t_9 > .632) = .543$, which suggests no evidence against H_0.