Statistical Inference

Testing of Statistical Hypotheses

Stanislav Katina[1]

[1]Institute of Mathematics and Statistics, Masaryk University
Honorary Research Fellow, The University of Glasgow

December 11, 2018

Testing of Statistical Hypotheses

Null and alternative hypothesis

- a **'hypothesis'** is a theory which is assumed to be true unless evidence is obtained which indicates otherwise
- **'null'** means 'nothing' and the term **'null hypothesis'** (H_0) means a 'theory of no change' – that is 'no change' from what would be expected from past experience
- **'alternative hypothesis'** (H_1) means a 'theory of change' – that is 'change' from what would be expected from past experience
- the procedure which is used to decide between these two opposite theories is called **'hypothesis test'** or sometimes **'significance test'**
- **one-tail test** – test in which the alternative hypothesis proposes a change in parameter in only one direction – increase or decrease
- **two-tail test**– test in which the alternative hypothesis suggests a difference in parameter in either direction

Testing of Statistical Hypotheses

Test statistic, rejection and acceptance region, critical value and quantile

- the **test statistic** is calculated from the sample – its value is used to decide whether the null hypothesis should be rejected
- the **rejection** (or **critical**) **region** gives the values of the test statistic for which the null hypothesis is rejected
- the **acceptance region** gives the values of the test statistic for which the null hypothesis is not rejected
- the boundary value(s) of the rejection region is (are) called the **critical value**(**s**) or **quantile**(**s**)
- the **significance level** α of a test gives the probability of the test statistic falling in the rejection region when null hypothesis is true

Testing of Statistical Hypotheses

Hypothesis testing procedure

- a **hypothesis** is a statement about a population parameter base on a sample from this population
- H_0 and H_1 are two complementary hypotheses in a hypothesis testing problem
- a **hypothesis testing procedure** or **hypothesis test** is a rule that specifies – for which sample values the decision is made to accept null hypothesis as true – and for which sample values H_0 is rejected
- the subset of sample space for which H_0 will be rejected is called **rejection region** (**critical region**)
- the complement of the rejection region is called the **acceptance region**

Testing of Statistical Hypotheses
Four possibilities

Four choices:

- A H_0 is true – our decision is to reject H_0
- B H_0 is true – our decision is not to reject H_0
- C H_1 is true – our decision is not to reject H_0
- D H_1 is true – our decision is to reject H_0

Decision-reality table:

decision/reality	H_0 is true	H_0 is not true
to reject H_0	Type I error	true decision
not to reject H_0	true decision	Type II error

Testing of Statistical Hypotheses
Four possibilities

Four choices:

- A) $\Pr(\text{A}) = \Pr(\text{Type I error}) \leq \alpha$ [significance level]
- B) $\Pr(\text{B}) \geq 1 - \alpha$ [coverage probability, confidence coefficient (level)]
- C) $\Pr(\text{C}) = \Pr(\text{Type II error}) = \beta$
- D) $\Pr(\text{D}) = 1 - \beta$ [power]

Four choices (formalised):

- A) $1 - \alpha \leq \Pr(\text{don't reject } H_0 | H_0 \text{ is true})$
- B) $\alpha \geq \Pr(\text{CHPD}) = \Pr(\text{reject } H_0 | H_0 \text{ is true})$
- C) $\beta = \Pr(\text{CHDD}) = \Pr(\text{don't reject } H_0 | H_0 \text{ isn't true})$
- D) $1 - \beta = \Pr(\text{reject } H_0 | H_0 \text{ isn't true})$

Testing of Statistical Hypotheses
Empirical $100 \times (1-\alpha)\%$ confidence intervals for parameter θ

Relationship of confidence interval and statistical test

- Empirical $100(1-\alpha)\%$ confidence interval (CI) for parameter θ
- α-level hypothesis test about θ

Three types of intervals:

- $\Pr(LB(X) < \theta < UB(X)) = 1 - \alpha$ (**two-tailed** CI)
- $\Pr(\theta < UB^*(X)) = 1 - \alpha$ (**one-tailed** (**right-tailed**) CI)
- $\Pr(LB_*(X) < \theta) = 1 - \alpha$ (**one-tailed** (**left-tailed**) CI)

Testing of Statistical Hypotheses
Acceptance region

Definition (Acceptance region of H_0)

Let X be a random variable with certain distribution (probabilistic model) dependent on parameter $\theta \in \Theta$, $g(\theta)$ is parametric function. We are testing null hypothesis $H_{01} : g(\theta) = g(\theta_0)$ against two-sided alternative $H_{11} : g(\theta) \neq g(\theta_0)$. Let (LB, UB) be interval estimate of parametric function $g(\theta)$ with coverage probability $1 - \alpha$. Then

$$\mathcal{A}_{\text{CI},1} = \{LB, UB; g(\theta_0) \in (LB, UB)\}$$

is **acceptance region of a test H_{01} against H_{11} on significance level** α. If we are testing $H_{02} : g(\theta) \leq g(\theta_0)$ against one-sided (right) alternative $H_{12} : g(\theta) > g(\theta_0)$ and if LB_* be lower estimate of $g(\theta)$ with coverage probability $1 - \alpha$, then

$$\mathcal{A}_{\text{CI},2} = \{LB_*; LB_* < g(\theta_0)\}$$

is **acceptance region of a test H_{02} against H_{12} on significance level** α. If we are testing $H_{03} : g(\theta) \geq g(\theta_0)$ against one-sided (left) alternative $H_{13} : g(\theta) < g(\theta_0)$ and if UB^* is upper estimate of $g(\theta)$ with coverage probability $1 - \alpha$, then

$$\mathcal{A}_{\text{CI},3} = \{UB^*; UB^* > g(\theta_0)\}$$

is **acceptance region of a test H_{03} against H_{13} on significance level** α.

Definition (Rejection (critical) region of H_0)

Let X be a random variable with certain distribution (probabilistic model) dependent on parameter $\theta \in \Theta$, $g(\theta)$ is parametric function. We are testing null hypothesis $H_{01}: g(\theta) = g(\theta_0)$ against two-sided alternative $H_{11}: g(\theta) \neq g(\theta_0)$. Let (LB, UB) be interval estimate of parametric function $g(\theta)$ with coverage probability $1-\alpha$. Then

$$\mathcal{W}_{\mathrm{CI},1} = \{LB, UB; g(\theta_0) \notin (LB, UB)\}$$

is **critical region of a test H_{01} against H_{11} on significance level** α. If we are testing $H_{02}: g(\theta) \leq g(\theta_0)$ against one-sided (right) alternative $H_{12}: g(\theta) > g(\theta_0)$ and if LB_* be lower estimate of $g(\theta)$ with coverage probability $1-\alpha$, then

$$\mathcal{W}_{\mathrm{CI},2} = \{LB_*; LB_* \geq g(\theta_0)\}$$

is **critical region of a test H_{02} against H_{12} on significance level** α. If we are testing $H_{03}: g(\theta) \geq g(\theta_0)$ against one-sided (left) alternative $H_{13}: g(\theta) < g(\theta_0)$ and if UB^* is upper estimate of $g(\theta)$ with coverage probability $1-\alpha$, then

$$\mathcal{W}_{\mathrm{CI},3} = \{UB^*; UB^* \leq g(\theta_0)\}$$

is **critical region of a test H_{03} against H_{13} on significance level** α.

Definition (Test criterion)

A **test criterion** is a test statistic $T = T_0 = T_0(X_1, X_2, \ldots, X_n)$, with known asymptotic distribution if H_0 is true. The set of possible values of T_0 is divided to two subsets, i.e. **acceptance region** H_0 (notation $\mathcal{A}$) and **critical region** H_0 (notation $\mathcal{W}$). These two regions are divided by **critical values** $t_{\alpha/2}$ and $t_{1-\alpha/2}$, resp. t_α and $t_{1-\alpha}$ (for particular H_0 and H_1) of the distribution of test statistics T_0 (if H_0 is true).

Definition (Confidence interval)

A **confidence interval** (CI) is a type of interval estimate of a population parameter θ. It is an observed, often called ***empirical***, interval (i.e., it is calculated from the observations) that includes the value of an unobservable parameter θ if the experiment is repeated. The frequency that observed interval contains the parameter is determined by the **confidence coefficient** $1-\alpha$ (i.e. **confidence level**, **coverage probability**).

Step 1 define the null and alternative hypothesis (H_0 and H_1)

Step 2 decide on a significance level $\alpha = 0.1, 0.05, 0.01$

Step 3 calculate the test statistic (test criterion) T_0

Step 4 determine the critical value(s)

Step 5 decide on the outcome of the test (reject/don't reject H_0) depending on one of the following ways:

- base on critical region $\mathcal{W} = \mathcal{W}_T$ (observed test statistic $t_0 = t_{\mathrm{obs}}$ and critical values $t_{\alpha/2}$ and $t_{1-\alpha/2}$, resp. t_α and $t_{1-\alpha}$),
- base on critical region $\mathcal{W}_{\mathrm{IS}}$, i.e. empirical confidence interval (and $g(\theta_0)$),
- base on p-value.

Step 6 state the conclusion in words

Definition (Testing based on critical region $\mathcal{W}$)

Rejecting H_0. If observed test statistic (realisation of test statistic) t_0 of test statistic T_0 is within a critical region $\mathcal{W}$ (equivalently is not from an acceptance region $\mathcal{A}$), H_0 is rejected at a significance level α, i.e. we do have sufficiently enough evidence to reject H_0.
Not rejecting H_0. If observed test statistic t_0 of test statistic T_0 is within an acceptance region $\mathcal{A}$ (equivalently, it is not from a critical region $\mathcal{W}$), H_0 is not rejected at a significance level α, i.e. we don't have sufficiently enough evidence to reject H_0.

Let $t_{\min}$ be the smallest possible value of a test criteria T_0 and $t_{\max}$ be the highest possible value of a test criteriaT_0, then

1. **two-sided alternative** – critical region $\mathcal{W}_1 = (t_{\min}, t_{1-\alpha/2}\rangle \cup \langle t_{\alpha/2}, t_{\max})$,
2. **one-sided (right) alternative** – critical region $\mathcal{W}_2 = \langle t_\alpha, t_{\max})$,
3. **one-sided (left) alternative** – critical region $\mathcal{W}_3 = (t_{\min}, t_{1-\alpha}\rangle$.

Testing of Statistical Hypotheses

To carry out a hypothesis test – based on CI

Definition (Testing based on CI)

Rejecting H_0: If $g(\theta) = g(\theta_0)$ is within CI (H_0 is valid), H_0 is rejected at the significance level α, i.e. we do have sufficiently enough evidence to reject H_0.
Not rejecting H_0: If $g(\theta) = g(\theta_0)$ is not within CI (H_0 is valid), H_0 isn't rejected at a significance level α, i.e. we don't have sufficiently enough evidence to reject H_0.

Relationship of confidence interval and statistical test

- hypothesis testing $\equiv$ CIs
- α-level hypothesis test $\equiv 100(1-\alpha)\%$ CI
- **one-tail test** $\equiv$ one-sided CI (left-sided CI $\equiv$ right-sided alternative, right-sided CI $\equiv$ left-sided alternative
- **two-tail test** $\equiv$ two-sided CI
- parameter(s) $\in$ CI $\equiv$ not reject H_0
- parameter(s) $\notin$ CI $\equiv$ reject H_0

Testing of Statistical Hypotheses

To carry out a hypothesis test – based on p-value (observed significance level)

Definition (Testing based on p-value)

Minimal significance level α (for some test statistic T_0), base on which $H_{02} : g(\theta) \leq g(\theta_0)$ is rejected (tested against $H_{12} : g(\theta) > g(\theta_0)$), is called **observed significance level** or **p-value**, i.e.

$$\text{p-value} = \alpha_{\text{obs}} = \sup_{\theta \in \Theta_0} \Pr\left(T(X_1, X_2, \ldots, X_n) \geq T(x_1, x_2, \ldots, x_n); \theta\right).$$

This could be written less formally as p-value $=$ Pr(any test statistics equal or greater than observed $|H_0$ is true).

The closer α_{obs} is to zero, the smaller is the probability that any test statistic $T(X_1, X_2, \ldots, X_n)$ produces a p-value (under H_0) equal to or smaller than that observed, while the probability is higher under H_1. Therefore, p-value could be understood as an indicator of credibility of H_0.

Testing of Statistical Hypotheses

To carry out a hypothesis test – based on p-value (observed significance level)

- Usually, if $\alpha_{\text{obs}} < \alpha = 0.05$, there is sufficiently enough evidence to reject H_0 and the result of a test **is statistically significant**.
- While $\alpha_{\text{obs}} > \alpha = 0.1$, there is sufficiently enough evidence to reject H_0 and the result of a test **is not statistically significant**.
- The values between 0.05 and 0.1 should be taken as reference points in a broad sense. As α_{obs} gets closer to either boundary point of the interval $\langle 0.05, 0.1 \rangle$, so this is taken as increasing evidence for one or other alternative.
- Situation with $\alpha_{\text{obs}} \in \langle 0.05, 0.1)$ are usually most difficult to handle and the result is here **marginally statistically significant**.

Testing of Statistical Hypotheses

To carry out a hypothesis test – based on p-value (observed significance level)

Wording of the results of a statistical test:

range for p-value	stars of significance	wording of the result
$\langle 0, 0.001)$	***	extremely highly statistically significant
$\langle 0.001, 0.01)$	**	high statistically significant
$\langle 0.01, 0.05)$	*	statistically significant
$\langle 0.05, 0.1)$	.	marginally statistically significant
$\langle 0.1, 1\rangle$		non-significant

Interpretation of p-values:

- p-value < 0.001: the ***prevalence*** of an estimated effect is smaller than one to one thousand (the ***odds*** of estimated effect is smaller than $1:999$), if an effect is not present in a population (the presence of such an effect is **highly improbable**, if an effect is not present in a population – and – the presence of such an effect is **highly probable**, if an effect is present in a population)
- p-value < 0.01: the ***prevalence*** of an estimated effect is smaller than one to one hundred (the ***odds*** of estimated effect is smaller than $1:99$), if an effect is not present in a population (the presence of such an effect is **very improbable**, if an effect is not present in a population – and – the presence of such an effect is **very probable**, if an effect is present in a population)
- p-value < 0.05: the ***prevalence*** of an estimated effect is smaller than one to one hundred (the ***odds*** of estimated effect is smaller than $5:95$ or $1:19$), if an effect is not present in a population (the presence of such an effect is **sufficiently improbable**, if an effect is not present in a population – and – the presence of such an effect is **sufficiently probable**, if an effect is present in a population)
- p-value ≥ 0.05: the prevalence of an estimated effect is five to one hundred or greater (5 % or more);
- p-value $= k, k \in \langle 0.05, 1\rangle$: the prevalence of an estimated effect is $100 \times k$ to one hundred ($100 \times k$ % or more).

How is the p-value (mostly) calculated?

1. **two-sided alternative** – p-value $= 2\min(\Pr(T_0 \leq t_0|H_0), \Pr(T_0 \geq t_0|H_0))$, e.g. for normal and Student distribution of test statistic (symmetric distributions) and for χ^2_{df} and F_{df_1,df_2} distribution of test statistic (asymmetric distributions) or p-value $= \min(\Pr(T_0 \leq t_0|H_0), \Pr(T_0 \geq t_0|H_0))$, e.g. for χ^2_{df} and F_{df_1,df_2} distribution of test statistic (asymmetric distributions)
2. **one-sided (right) alternative** – p-value $= \Pr(T_0 \geq t_0|H_0)$
3. **one-sided (left) alternative** – p-value $= \Pr(T_0 \leq t_0|H_0)$

- distinction between 'rejecting H_0' and 'accepting H_1'
- '*rejecting* H_0' – nothing implies about what state the experimenter *is* accepting, only that the state defined by H_0 is being rejected
- distinction between 'accepting H_0' and 'not rejecting H_0'
- 'accepting H_0' – the experimenter is willing to assert the state of nature specified by H_0
- '*not rejecting* H_0' – the experimenter really does not believe H_0 but does not have the evidence to reject it

Definition (Conservative and liberal test)

A test with **actual/observed significance level** smaller than **nominal significance level** α, is called **conservative** (the test should theoretically be "rejecting quickly" H_0, but, in reality, it is the opposite, i.e. the test is "rejecting slowly").
A test with **actual/observed significance level** greater than **nominal significance level** α, is called **liberal** (the test should theoretically be "rejecting slowly" H_0, but, in reality, it is the opposite, i.e. the test "rejecting quickly").

Definition (Conservative and liberal CI)

CI with **actual/real coverage probability** greater than **nominal coverage probability** $1-\alpha$, is called **conservative** (i.e. the probability that θ_0 is within CI is greater that expected).
CI with **actual/real coverage probability** smaller than **nominal coverage probability** $1-\alpha$, is called **liberal** (i.e. the probability that θ_0 is within CI is smaller that expected).

Two types of hypotheses:

1. **simple hypothesis** – $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, then **simple likelihood ratio** is equal to

$$\lambda(\mathbf{x}) = \lambda = \frac{L(\theta_0|\mathbf{x})}{\sup_{\theta\in\Theta} L(\theta|\mathbf{x})} = \frac{L(\theta_0|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})},$$

where $\lambda(\mathbf{x}) = \mathcal{L}(\theta_0|\mathbf{x})$ is test statistic and $L(\theta|\mathbf{x})$ is continuous for all x.

2. **composite hypothesis** – $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, then **generalised likelihood ratio** is equal to

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

Subsets of Θ, Θ_0 and Θ_1, remain the same after monotone transformation of $\lambda(\mathbf{x})$, i.e. the statistical tests before and after transformation are equivalent. Therefore, **likelihood ratio test statistic** is equal to

$$U_{\mathrm{LR}} = -2\ln\lambda(\mathbf{X}).$$

Its realisation, **observed likelihood ratio test statistic**, is equal to $u_{\mathrm{LR}} = -2\ln\lambda(\mathbf{x})$, where $u_{\mathrm{LR}} \in (0,\infty)$.

After applying Taylor series of $l(\theta_0)$ about $\widehat{\theta}$,

$$U_{\mathrm{LR}} = -2(l(\theta_0|\mathbf{X}) - l(\widehat{\theta}|\mathbf{X})) \approx -2\left((\theta_0-\widehat{\theta})S(\widehat{\theta}) - \frac{1}{2}(\theta_0-\widehat{\theta})^2\mathcal{I}(\widehat{\theta})\right),$$

where $S(\widehat{\theta}) = 0$. Under H_0, **Wald test statistic** U_{W}, is defined as follows

$$U_{\mathrm{LR}} \approx n(\theta_0-\widehat{\theta})^2\frac{\mathcal{I}(\theta_0)}{n} \approx n(\theta_0-\widehat{\theta})^2 i(\theta_0) \overset{H_0}{\approx} n(\theta_0-\widehat{\theta})^2 i(\widehat{\theta}) = U_{\mathrm{W}},$$

where $\frac{1}{n}\mathcal{I}(\widehat{\theta}) \xrightarrow{\mathcal{P}} i(\theta_0)$; its realisation, **observed Wald test statistic** is u_W. Under H_0, **Score test statistic** U_{S}, is defined as follows

$$U_{\mathrm{LR}} \approx n(\theta_0-\widehat{\theta})^2 i(\theta_0) \overset{H_0}{\approx} \frac{(S(\theta_0))^2}{n\, i(\theta_0)} = U_{\mathrm{S}},$$

where $\sqrt{n}(\widehat{\theta}-\theta_0) \overset{H_0}{\approx} S(\theta_0)/(\sqrt{n}(i(\theta_0)))$; its realisation, **observed Score test statistic** is u_S.

Geometrical interpretation:

1. U_{LR} – is measuring properly standardised difference between log-likelihoods in $\widehat{\theta}$ and θ_0 (i.e. in direction of y axis)
2. U_W – is measuring properly standardised absolute value of a difference of $\widehat{\theta}$ a θ_0 (in direction of x axis)
3. U_S – is measuring properly standardised slope of log-ratio in θ_0

Example (normal distribution)

Let $X \sim N(\mu, \sigma^2)$, where σ^2 is known, $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, where $\boldsymbol{\theta}_0 = (\mu_0, \sigma^2)^T$. Then

1. $U_{\text{LR}} = -2(l(\theta_0|\mathbf{X}) - l(\widehat{\theta}|\mathbf{X})) = -\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2 + \sum_{i=1}^n (X_i - \mu_0)^2/\sigma^2 = n\frac{(\overline{X}-\mu_0)^2}{\sigma^2}$,
2. $U_{\text{W}} = (\overline{X} - \mu_0)^2 \mathcal{I}(\overline{x}) = n\frac{(\overline{X}-\mu_0)^2}{\sigma^2}$,
3. $U_{\text{S}} = \frac{(S(\mu_0))^2}{\mathcal{I}(\mu_0)} = \frac{(n(\overline{X}-\mu_0)/\sigma^2)^2}{n/\sigma^2} = n\frac{(\overline{X}-\mu_0)^2}{\sigma^2}$.

All three test statistics are equal, i.e. $U_{\text{LR}} = U_{\text{W}} = U_{\text{S}}$.

Let θ be a scalar. **Null hypothesis** $H_0 : \theta = \theta_0$ and **alternative hypothesis** $H_1 : \theta \neq \theta_0$, where θ_0 is a scalar from H_0. Let $\widehat{\theta}$ be the maximal likelihood estimate of θ. Let $\widehat{Var}[\widehat{\theta}]$ be the variance of $\widehat{\theta}$.

Then three test statistics are defined as follows:

1. $U_{\text{LR}} = -2(l(\theta_0|\mathbf{X}) - l(\widehat{\theta}|\mathbf{X})) \overset{\mathcal{D}}{\sim} \chi^2_1$,
2. $U_{\text{W}} = (\widehat{\theta} - \theta_0)^2 \mathcal{I}(\widehat{\theta}) \overset{\mathcal{D}}{\sim} \chi^2_1$ and equivalently $U_{\text{W}}^{1/2} = Z_{\text{W}} \overset{\mathcal{D}}{\sim} N(0,1)$,
3. $U_{\text{S}} = \frac{(S(\theta_0))^2}{\mathcal{I}(\theta_0)} \overset{\mathcal{D}}{\sim} \chi^2_1$ and equivalently $U_{\text{S}}^{1/2} = Z_{\text{S}} \overset{\mathcal{D}}{\sim} N(0,1)$.

Let $\boldsymbol{\theta}$ be a vector of all parameters of length k. **Null hypothesis** $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and **alternative hypothesis** $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is a vector of parameters from H_0. Let $\widehat{\boldsymbol{\theta}}$ be the maximal likelihood estimate of $\boldsymbol{\theta}$. Let $\widehat{Var}[\widehat{\boldsymbol{\theta}}]$ be the covariance matrix.

Then three test statistics are defined as follows:

1. $U_{\text{LR}} = -2(l(\boldsymbol{\theta}_0|\mathbf{X}) - l(\widehat{\boldsymbol{\theta}}|\mathbf{X})) \overset{\mathcal{D}}{\sim} \chi^2_k$,
2. $U_{\text{W}} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathcal{I}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{\mathcal{D}}{\sim} \chi^2_k$,
3. $U_{\text{S}} = (S(\boldsymbol{\theta}_0))^T (\mathcal{I}(\boldsymbol{\theta}_0))^{-1} S(\boldsymbol{\theta}_0) \overset{\mathcal{D}}{\sim} \chi^2_k$.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T$, where $\boldsymbol{\theta}$ is a vector of all parameters of length k. Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be subsets of parameters of length k_1 and k_2, where $k_1 + k_2 = k$. **Null hypothesis** $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$ and **alternative hypothesis** $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is a vector of parameters from H_0. Let $\widehat{\boldsymbol{\theta}}$ be maximal likelihood estimate of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_{2|0}$ be maximal likelihood estimate of $\boldsymbol{\theta}_2$ if H_0 is true, i.e. $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$. Then $\widehat{\boldsymbol{\theta}}_0 = (\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}_{2|0})^T$. Let $\widehat{Var}_{11}[\widehat{\boldsymbol{\theta}}]$ be a submatrix of the covariance matrix $\widehat{Var}[\widehat{\boldsymbol{\theta}}]$ corresponding to $\boldsymbol{\theta}_1$.

Then three test statistics are defined as follows:

1. $U_{\text{LR}} = -2(l(\widehat{\boldsymbol{\theta}}_0|\mathbf{X}) - l(\widehat{\boldsymbol{\theta}}|\mathbf{X})) \overset{\mathcal{D}}{\sim} \chi^2_{k_1}$,
2. $U_{\text{W}} = (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0)^T \mathcal{I}_{11}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0) \overset{\mathcal{D}}{\sim} \chi^2_{k_1}$,
3. $U_{\text{S}} = (S(\boldsymbol{\theta}_0))^T (\mathcal{I}_{11}(\widehat{\boldsymbol{\theta}}_0))^{-1} S(\boldsymbol{\theta}_0) \overset{\mathcal{D}}{\sim} \chi^2_{k_1}$.

There is a relationship between likelihood ratio test statistic for subset of parameters and **profile likelihood function**:

$$L_P(\boldsymbol{\theta}_1|\mathbf{x}) = \max_{\forall \boldsymbol{\theta}_2} L(\boldsymbol{\theta}|\mathbf{x}) = L((\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_{2|0})^T|\mathbf{x})$$

or **logarithm of profile likelihood function**

$$l_P(\boldsymbol{\theta}_1|\mathbf{x}) = l((\boldsymbol{\theta}_1, \widehat{\boldsymbol{\theta}}_{2|0})^T|\mathbf{x}).$$

Likelihood ratio test statistic is defined as:

$$u_{\mathrm{LR}} = -2\ln \mathcal{L}_P(\boldsymbol{\theta}_1|\mathbf{x}) = -2\left(l_P(\boldsymbol{\theta}_1|\mathbf{x}) - l_P(\widehat{\boldsymbol{\theta}}_1|\mathbf{x})\right),$$

where $\widehat{\boldsymbol{\theta}}_1$ is maximal likelihood estimate of $\boldsymbol{\theta}_1$ with respect to $\mathcal{L}_P(\boldsymbol{\theta}_1|\mathbf{x})$. U_{LR} is also called **generalised likelihood ratio statistic**.

Additionally

$$L_P(\widehat{\boldsymbol{\theta}}_1|\mathbf{x}) = \max_{\forall \boldsymbol{\theta}_1}\left\{\max_{\forall \boldsymbol{\theta}_2} L(\boldsymbol{\theta}|\mathbf{x})\right\} = \max_{\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2} L((\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T|\mathbf{x}).$$

Having $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$ a $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$, then

$$L_P(\boldsymbol{\theta}_0|\mathbf{x}) = \max_{\forall \boldsymbol{\theta}_2} L((\boldsymbol{\theta}_0, \boldsymbol{\theta}_2)^T|\mathbf{x}) = \max_{H_0} L((\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T|\mathbf{x})$$

and

$$u_{\mathrm{LR}} = -2\ln \frac{\max_{H_0} L((\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T|\mathbf{x})}{\max_{\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2} L((\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T|\mathbf{x})} = -2\ln \frac{L_P(\boldsymbol{\theta}_0|\mathbf{x})}{L_P(\widehat{\boldsymbol{\theta}}_1|\mathbf{x})}.$$

Quadratic approximation of relative profile log-likelihood is defined as:

$$\ln \mathcal{L}_P(\boldsymbol{\theta}_1|\mathbf{x}) \approx -\frac{1}{2}\left(\boldsymbol{\theta}_1 - \widehat{\boldsymbol{\theta}}_1\right)^T (\mathcal{I}^{11}(\boldsymbol{\theta}))^{-1}\left(\boldsymbol{\theta}_1 - \widehat{\boldsymbol{\theta}}_1\right),$$

and **quadratic approximation of generalised likelihood ratio statistic** $-2\ln \mathcal{L}_P(\boldsymbol{\theta}_1|\mathbf{x})$ is defined as:

$$u_{\mathrm{LR}} \approx u_{\mathrm{W}} = \left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\right)^T (\mathcal{I}^{11}(\widehat{\boldsymbol{\theta}}))^{-1}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\right).$$

Marginal distribution of $\boldsymbol{\theta}_1$ if H_0 is true is defined as $\widehat{\boldsymbol{\theta}}_1 \sim N_{k_1}(\boldsymbol{\theta}_0, I^{11}(\boldsymbol{\theta}))$.

If θ is a scalar, three confidence intervals are defined as follows:

1. **empirical likelihood ratio** $(1-\alpha) \times 100\%$ **CI for** θ is defined as

$$\mathcal{CS}_{1-a} = \left\{\theta : U_{\mathrm{LR}}(\theta) < \chi_1^2(\alpha)\right\},$$

   where $U_{\mathrm{LR}}(\theta) = -2\ln \frac{L(\theta|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}$.
2. **empirical Wald** $(1-\alpha) \times 100\%$ **CI for** θ is defined based on a pivot (pivotal statistics)$T_{\mathrm{piv}} = U_{\mathrm{W}}(\theta)$
3. **empirical Score** $(1-\alpha) \times 100\%$ **CI for** θ is defined based on a pivot $T_{\mathrm{piv}} = U_{\mathrm{S}}(\theta)$

If $\boldsymbol{\theta}$ is a vector, CIs can be generalized to **confidence set** $\mathcal{CS}_{1-a}$.

- If $k = 2$, $\mathcal{CS}_{1-a}$ is an **confidence ellipse**.
- If $k > 2$, $\mathcal{CS}_{1-a}$ is an **confidence ellipsoid**.

Additionally, if $k = 1$, $\mathcal{CS}_{1-a}$ is an **confidence interval**.

Wald empirical $(1-\alpha)\times 100\%$ **CI for** θ is defined as

$$(l,u)=\left(\widehat{\theta}_L,\widehat{\theta}_U\right)=\left(\widehat{\theta}-t_{\alpha/2}\widehat{SD[\widehat{\theta}]},\widehat{\theta}+t_{\alpha/2}\widehat{SD[\widehat{\theta}]}\right),$$

where the critical value $t_{\alpha/2}$ depends on the choice of $\widehat{\theta}$.

Likelihood ratio empirical $(1-\alpha)\times 100\%$ **CI for** θ is defined by its lower and upper bounds as $k\%$ cut-offs of standardized relative log-likelihood as follows

$$\Pr\left(\frac{L(\theta|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}>c_\alpha\right)=\Pr\left(-2\ln\frac{L(\theta|\mathbf{x})}{L(\widehat{\theta}|\mathbf{x})}<-2\ln c_\alpha\right)=1-\alpha,$$

where $c_\alpha=e^{-\frac{1}{2}\chi_1^2(\alpha)}$. Then

- if $1-\alpha=0.95$, then $c_\alpha=0.1465001\doteq 0.15$ (15% cut-off),
- if $1-\alpha=0.90$, then $c_\alpha=0.2585227\doteq 0.26$ (26% cut-off),
- if $1-\alpha=0.99$, then $c_\alpha=0.0362452\doteq 0.04$ (4% cut-off).

Bisection method
Let $\theta_{01},\theta_{02}\in\langle\theta_L,\theta_U\rangle$ and $f(\theta_{01})f(\theta_{02})<0$, $f(\cdot)$ is continuous with at least one root within the interval $\langle\theta_{01},\theta_{02}\rangle$, where

$$f(\theta)=-2\ln\mathcal{L}(\theta|\mathbf{x})-\chi_1^2(\alpha)=0.$$

If the first derivative of $f(\cdot)$ is having constant sign, then exactly one root $\theta^*\in\langle\theta_{01},\theta_{02}\rangle$ of $f(\theta)=0$ exists.

The iterative process is defined as follows:

1. initialisation step – starting point $\theta^{(0)}=(\theta_{01}+\theta_{02})/2$ and $i=1$,
2. updating equations – substitution of the boundaries θ_{01} and θ_{02} is defined as

$$\langle\theta_{i1},\theta_{i2}\rangle=\begin{cases}\left\langle\theta_{i-1,1},\theta^{(i-1)}\right\rangle, & \text{if } f(\theta_{i-1,1})f(\theta^{(i-1)})<0\\ \left\langle\theta^{(i-1)},\theta_{i-1,2}\right\rangle, & \text{if } f(\theta_{i-1,1})f(\theta^{(i-1)})>0\end{cases},$$

if $f(\theta^{(i-1)})=0$, then *end*, if not,

3. calculate the mid-point $\theta^{(i)}=(\theta_{i1}+\theta_{i2})/2$,
4. stopping rule (with the **threshold** ϵ is sufficiently small) based on

- relative convergence criteria

$$\frac{\left|\theta^{(i)}-\theta^{(i-1)}\right|}{\left|\theta^{(i-1)}\right|}<\epsilon,$$

- absolute convergence criteria

$$\left|\theta^{(i)}-\theta^{(i-1)}\right|<\epsilon,$$

- or often also based on

$$\left|f(\theta^{(i)})\right|<\epsilon.$$

Modifications are based on **bracketing methods**, i.e. bounding the root within a sequence of intervals.

Brent method (**Brent-Dekker method**) – the combination of bisection method with inverse interpolation. If the interpolation is linear, then it is **secant method**, where the updating equations are modified as follows

$$\theta^{(i)}=\begin{cases}\theta^{(i-1)}-\frac{\theta^{(i-1)}-\theta^{(i-2)}}{f(\theta^{(i-1)})-f(\theta^{(i-2)})}f(\theta^{(i-1)}), & \text{if } f(\theta^{(i-1)})\neq f(\theta^{(i-2)})\\ (\theta_{i1}+\theta_{i2})/2, & \text{otherwise}\end{cases},$$

where the approximation of the first derivative $f'(\theta^{(i-1)})\approx\frac{f(\theta^{(i-1)})-f(\theta^{(i-2)})}{\theta^{(i-1)}-\theta^{(i-2)}}$. If $f(\theta)$ is twice differentiable, then $f(\theta)$ has single root ($f'(\theta)\neq 0$ for all $\theta\in\langle\theta_L,\theta_U\rangle$).

Geometrical interpretation: $\theta^{(i)}$ is the crossing point of secant through the points $[\theta^{(i-1)}, f(\theta^{(i-1)})]$ and $[\theta^{(i-2)}, f(\theta^{(i-2)})]$, and x axis.

In R:

- `uniroot(f, interval,tol,...)`
- during the search for lower and upper boundary of $100 \times (1-\alpha)\%$ for θ, the R-function `uniroot()` should be used twice as follows
  1. for lower bound – starting interval is defined as $\left\langle \theta_L, \widehat{\theta} \right\rangle$,
  2. for upper bound – starting interval is defined as $\left\langle \widehat{\theta}, \theta_U \right\rangle$.

  Then the solutions are $\widehat{\theta}_L$ and $\widehat{\theta}_U$ (`root`).

Example (Brent-Dekker method)

Let $X \sim Bin(N,p)$, where $N = 10$ and $n = x = 8$. Estimate the boundaries of empirical $100\times(1-\alpha)\%$ CI for (1) p and (2) log odds $\ln \frac{p}{1-p}$. The empirical CI are of the two types (A) likelihood and (B) Wald. Draw the log-likelihood function and its quadratic approximation with the lower and upper boundary of CI.

Solution (partial)

Empirical Wald $100 \times (1-\alpha)\%$ CI for p:

$\widehat{p} = \frac{8}{10} = 0.8$; $\widehat{SD[\widehat{p}]} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{N}} = 0.13$.

$(l,u) = (\widehat{p}_l, \widehat{p}_u) = \left(\widehat{p} - u_{\alpha/2}\widehat{SD[\widehat{p}]}, \widehat{p} + u_{\alpha/2}\widehat{SD[\widehat{p}]}\right) = (0.55, 1.05)$.

Empirical Likelihood $100 \times (1-\alpha)\%$ CI for p:

$\mathcal{CS}_{1-\alpha} = \left\{p : -2\ln\frac{L(p|\mathbf{x})}{L(\widehat{p}|\mathbf{x})} \leq 3.84\right\}$, where

$(l,u) = (\widehat{p}_L, \widehat{p}_U) = (0.50, 0.96)$,

Wald empirical $100 \times (1-\alpha)\%$ CI for $g(p)$:

$g(\widehat{p}) = \ln\frac{\widehat{p}}{1-\widehat{p}} = \ln\frac{0.8}{0.2} = 1.39$; $\frac{\partial}{\partial p} g(p) = \frac{1}{p} + \frac{1}{1-p}$; $\widehat{SD[g(\widehat{p})]} =$

$\widehat{SD[\widehat{p}]}\left(\frac{1}{\widehat{p}} + \frac{1}{1-\widehat{p}}\right) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{N}}\left(\frac{1}{\widehat{p}} + \frac{1}{1-\widehat{p}}\right) = \sqrt{\frac{1}{n} + \frac{1}{N-n}} = 0.79$.

Then $(l_g, u_g) = (g(\widehat{p}_L), g(\widehat{p}_U)) = (-0.16, 2.94)$ and back-transformed $(l,u) = (\widehat{p}_L, \widehat{p}_U) = (0.46, 0.95)$.


```
x <- 8; N <- 10
probs <- seq(0.4,.99,length=1000)
like  <- dbinom(8,10,probs)
rellike <- like/max(like)
relloglike    <- -2*log(rellike)
cutoff <- exp(-1/2*qchisq(0.95,df=1)) #0.1465001
likeCI.p <- range(probs[rellike>cutoff]) #0.5009910 0.9634234
cutoff <- qchisq(0.95,df=1) #3.841459
likeCI.p <- range(probs[relloglike<cutoff]) #0.500991 0.9634234
p.hat <- x/N
i.hat <- N/p.hat/(1-p.hat)
loglikeapprox <- -i.hat/2*(probs-p.hat)^2
ra <- range(log(rellike))
waldCI.p <- p.hat + c(-1,1)*qnorm(0.975)*sqrt(1/i.hat)
waldCI.p # 0.552082 1.047918
gprobs <- log(probs)-log(1-probs)
gp.hat <- log(p.hat)-log(1-p.hat)
i.hat <- x*(N-x)/N
lgp <- -i.hat/2*(gprobs-gp.hat)^2
x <- (gp.hat+c(-1,1)*qnorm(0.975)*sqrt(1/i.hat)) #-0.1632 2.9358
waldCI.gp <- exp(x)/(1+exp(x))
waldCI.gp # 0.4592920 0.9495872
```

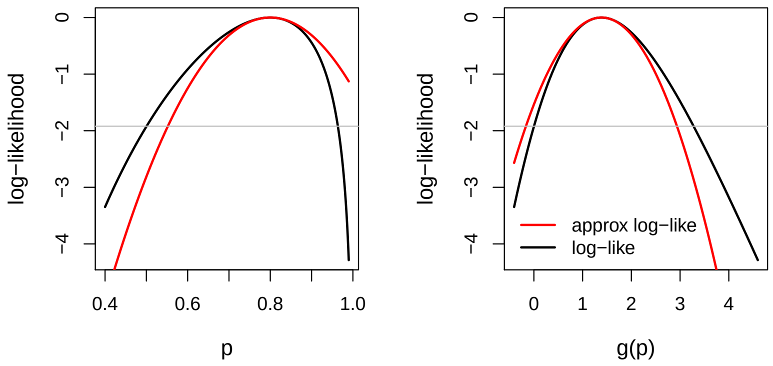

Figure: Log-likelihood of p and its quadratic approximation