Curs 16

Cristian Niculescu

1 Testarea semnificaţiei ipotezei 0 I

1.1 Scopurile învăţării

1. Să ştie definiţiile termenilor de testare a semnificaţiei: NHST, ipoteză 0, ipoteză alternativă, ipoteză simplă, ipoteză compusă, nivel de semnificaţie, putere.
2. Să poată face un test de semnificaţie pentru date Bernoulli şi binomiale.
3. Să poată calcula o p-valoare pentru o ipoteză normală şi s-o utilizeze într-un test de semnificaţie.

1.2 Introducere

Statistica frecvenţionistă este adesea aplicată în cadrul testării semnificaţiei ipotezei 0 (NHST). Paradigma Neyman-Pearson se concentrează pe o ipoteză numită ipoteza 0. Sunt şi alte paradigme pentru testarea ipotezelor, dar Neyman-Pearson este cea mai uzuală. Spus simplu, această metodă întreabă dacă datele sunt în afara regiunii unde ne-am aştepta să le vedem sub ipoteza 0. Dacă este aşa, respingem ipoteza 0 în favoarea unei a 2-a ipoteze, numită ipoteza alternativă.
Calculele făcute aici necesită toate funcţia de verosimilitate. Sunt 2 diferenţe majore între ce vom face aici şi ce am făcut la actualizarea Bayesiană:
1. Dovezile datelor vor fi considerate doar prin funcţia de verosimilitate, nu vor mai fi ponderate de convingerile noastre a priori.
2. Vom avea nevoie de o noţiune de date extreme, de exemplu 95 de aversuri din 100 de aruncări ale unei monede sau o muscă efemeră care trăieşte o lună.

1.2.1 Exemple motivante

Exemplul 1. Presupunem că vrem să decidem dacă o monedă este corectă. Dacă obţinem 85 de aversuri din 100 de aruncări, aţi considera că moneda este probabil incorectă? Dar la 60 de aversuri? Sau la 52? Majoritatea

oamenilor ar ghici că 85 de aversuri este o dovadă puternică pentru faptul că moneda este incorectă în timp ce 52 de aversuri nu este deloc dovadă. 60 de aversuri este mai puţin clar. Testarea semnificaţiei ipotezei 0 (NHST) este o abordare frecvenţionistă pentru a gândi cantitativ despre aceste întrebări.
Exemplul 2. Presupunem că vrem să comparăm un tratament medical nou cu un placebo sau cu standardul curent de îngrijire. Ce fel de dovezi v-ar convinge că noul tratament este mai bun decât placebo sau standardul curent? Din nou, NHST este un cadru cantitativ pentru a răspunde acestei întrebări.

1.3 Testarea semnificaţiei

Listăm ingredientele pentru NHST.

1.3.1 Ingrediente

H_0: ipoteza 0. Aceasta este presupunerea implicită pentru modelul care generează datele.
H_A: ipoteza alternativă. Dacă respingem ipoteza 0, acceptăm această alternativă ca cea mai bună explicaţie pentru date.
X: statistica testului. O calculăm din date.
Repartiţia 0: repartiţia de probabilitate a lui X presupunând H_0.
Regiunea de respingere: dacă X este în regiunea de respingere, respingem H_0 în favoarea lui H_A.
Regiunea de nerespingere: complementara regiunii de respingere. Dacă X este în această regiune, nu respingem H_0. Spunem "nu respingem" mai degrabă decât "acceptăm" deoarece cel mai bun lucru pe care-l putem spune este că datele nu susţin respingerea lui H_0.
Ipoteza 0 H_0 şi ipoteza alternativă H_A joacă roluri diferite. Respingem H_0 doar dacă avem destule dovezi împotriva ei.

1.4 Terminologia NHST

În această secţiune vom folosi un exemplu extins pentru a introduce şi explora terminologia utilizată în testarea semnificaţiei ipotezei 0 (NHST).
Exemplul 3. Pentru a testa dacă o monedă este corectă o aruncăm de 10 ori. Dacă obţinem un număr neaşteptat de mare sau de mic de aversuri, vom suspecta că moneda nu este corectă. Precizăm aceasta în limbajul NHST setând ingredientele. Fie θ probabilitatea aversului la aruncarea monedei.
1. Ipoteza 0: $H_0 =$ "moneda este corectă", i.e. $\theta = 0.5$.
2. Ipoteza alternativă: $H_A =$ "moneda nu este corectă", i.e. $\theta \neq 0.5$.

3. Statistica testului: X = numărul de aversuri în 10 aruncări.
4. Repartiţia 0: Aceasta este funcţia de probabilitate bazată pe ipoteza 0

$$p(x|\theta = 0.5) \sim \text{binomială}(10, 0.5).$$

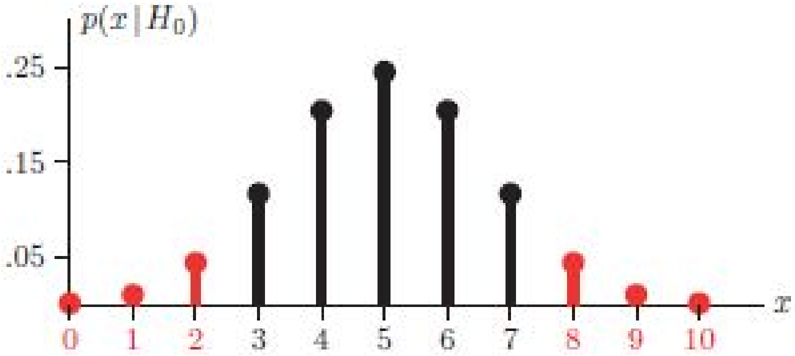
Iată tabelul de probabilitate pentru repartiţia 0:

x	0	1	2	3	4	5	6	7	8	9	10
$p(x \mid H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

5. Regiunea de respingere: sub ipoteza 0 ne aşteptăm să obţinem circa 5 aversuri în 10 aruncări. Vom respinge H_0 dacă numărul de aversuri este mult mai mic sau mai mare ca 5. Fie regiunea de respingere $\{0, 1, 2, 8, 9, 10\}$. Adică, dacă numărul de aversuri din 10 aruncări este această regiune, vom respinge ipoteza că moneda este corectă în favoarea ipotezei că nu este corectă.
Putem rezuma toate acestea în graficul şi tabelul de probabilitate de mai jos. Regiunea de respingere constă din valorile lui x în roşu. Probabilităţile corespunzătoare sunt pe fond roşu. Arătăm de asemenea repartiţia 0 ca o reprezentare cu valorile lui x de respingere în roşu.

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

Regiunea de respingere şi probabilităţile 0 ca un tabel pentru exemplul 3.



Regiunea de respingere şi probabilităţile 0 ca o reprezentare pentru exemplul 3.

Observaţii pentru exemplul 3:
1. Ipoteza 0 este implicit precautǎ: nu vom pretinde că moneda este incorectă dacă nu avem argumente convingătoare.
2. Regiunea de respingere constă din date care sunt extreme sub ipoteza 0. Adică, constă din rezultatele care sunt în coada repartiţiei 0, departe de centrul de probabilitate mare. Cât de departe, depinde de α, nivelul de semnificaţie al testului.

3. Dacă obţinem 3 aversuri în 10 aruncări, atunci statistica testului este în regiunea de nerespingere. Limbajul ştiinţific uzual ar fi să spunem că datele "nu sprijină respingerea ipotezei 0". Chiar dacă avem 5 aversuri, nu vom pretinde că datele demonstrează că ipoteza 0 este adevărată.
Întrebare: Dacă avem o monedă corectă, care este probabilitatea că vom decide incorect că este incorectă?
Răspuns: Ipoteza 0 este că moneda este corectă. Se cere probabilitatea ca datele dintr-o monedă corectă să fie în regiunea de respingere. Adică, probabilitatea că vom obţine 0, 1, 2, 8, 9 sau 10 aversuri în 10 aruncări. Aceasta este suma probabilităţilor în roşu. Adică,

$$P(\text{respingerea lui } H_0 | H_0 \text{ este adevărată}) = 0.11.$$

1.4.1 Ipoteze simple şi compuse

Definiţie: ipoteză simplă: O ipoteză simplă este una pentru care putem specifica complet repartiţia ei. O ipoteză simplă tipică este că parametrul de interes ia o anumită valoare.
Definiţie: Ipoteze compuse: Dacă repartiţia nu poate fi specificată complet, spunem că ipoteza este compusă. O ipoteză compusă tipică este că parametrul de interes se află într-un domeniu de valori.
În exemplul 3, ipoteza 0 este că $\theta = 0.5$, deci repartiţia 0 este binomială$(10, 0.5)$. Deoarece repartiţia 0 este complet specificată, H_0 este simplă. Ipoteza alternativă este că $\theta \neq 0.5$. În realitate sunt multe ipoteze în una: θ ar putea fi 0.51, 0.7, 0.99, etc. Deoarece repartiţia alternativă binomială$(10, \theta)$ nu este complet specificată, H_A este compusă.
Exemplul 4. Presupunem că avem datele $x_1, ..., x_n$. Presupunem că ipotezele noastre sunt
H_0: datele sunt extrase din $N(0, 1)$
H_A: datele sunt extrase din $N(1, 1)$.
Acestea sunt ambele ipoteze simple - fiecare ipoteză specifică complet o repartiţie.
Exemplul 5. **(Ipoteze compuse.)** Acum presupunem că ipotezele noastre sunt
H_0: datele sunt extrase dintr-o repartiţie Poisson cu parametru necunoscut.
H_A: datele nu sunt extrase dintr-o repartiţie Poisson.
Acestea sunt ambele ipoteze compuse, deoarece ele nu specifică complet repartiţia.
Exemplul 6. Într-un experiment de percepţie extrasenzorială (ESP), unui subiect i se cere să identifice culorile (pică, cupă, romb, treflă) a 100 de cărţi trase (cu înlocuire) dintr-un pachet de cărţi de joc. Fie T, numărul

succeselor. Ipoteza 0 (simplă) că subiectul nu are ESP este dată de

$$H_0 : T \sim \text{binomial}(100, 0.25).$$

Ipoteza alternativă (compusă) că subiectul are ESP este dată de

$$H_A : T \sim \text{binomial}(100, p) \text{ cu } p > 0.25.$$

O altă ipoteză alternativă (compusă) este că se întâmplă ceva în afara şansei pure, i.e. subiectul are ESP sau anti-ESP. Aceasta este dată de

$$H_A : T \sim \text{binomial}(100, p) \text{ cu } p \neq 0.25.$$

Valorile $p < 0.25$ reprezintă ipotezele că subiectul are un fel de anti-ESP.

1.4.2 Tipuri de eroare

Sunt 2 tipuri de erori pe care le putem face. Putem să respingem incorect ipoteza 0 când ea este adevărată sau putem să eşuăm incorect s-o respingem când este falsă. Acestea sunt numite erori de tipul I şi de tipul II. Rezumăm aceasta în următorul tabel.

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Don't reject' H_0	correct decision	Type II error

Tipul I: respingere falsă a lui H_0.
Tipul II: nerespingere falsă ("acceptare" falsă) a lui H_0.

1.4.3 Nivel de semnificaţie şi putere

Nivelul de semnificaţie şi puterea sunt folosite pentru a cuantifica calitatea testului de semnificaţie. Ideal, un test de semnificaţie n-ar face erori. Adică, n-ar respinge H_0 când H_0 era adevărată şi ar respinge H_0 în favoarea lui H_A când H_A era adevărată. Sunt cu totul 4 probabilităţi importante corespunzătoare tabelului 2×2 de mai sus:

$$\begin{array}{ll} P(\text{respingem } H_0 | H_0) & P(\text{respingem } H_0 | H_A) \\ P(\text{nu respingem } H_0 | H_0) & P(\text{nu respingem } H_0 | H_A). \end{array}$$

Cele 2 probabilităţi pe care ne concentrăm sunt:

$$\begin{aligned} \text{Nivelul de semnificaţie} &= P(\text{respingem } H_0 | H_0) \\ &= \text{ probabilitatea să respingem incorect } H_0 \\ &= P(\text{eroarea de tipul I}). \end{aligned}$$

$$\begin{aligned} \text{Puterea} = &\ \text{probabilitatea să respingem corect } H_0 \\ &= P(\text{respingem } H_0 | H_A) \\ &= 1 - P(\text{eroarea de tipul II}). \end{aligned}$$

Ideal, un test al unei ipoteze ar trebui să aibă un nivel de semnificaţie mic (aproape de 0) şi o putere mare (aproape de 1).

Analogii

1. Gândiţi H_0 ca ipoteza "nu se întâmplă nimic demn de remarcat", i.e. "moneda este corectă", "tratamentul nu este mai bun ca placebo" etc. Şi gândiţi H_A ca opusa: "ceva interesant se întâmplă". Atunci puterea este probabilitatea de a detecta ceva interesant când este prezent şi nivelul de semnificaţie este probabilitatea de a pretinde greşit că ceva interesant a apărut.

2. În SUA, inculpaţii sunt presupuşi nevinovaţi până sunt demonstraţi vinovaţi dincolo de un orice dubiu rezonabil. Putem formula aceasta în termeni de NHST ca

H_0: acuzatul este nevinovat (implicit)

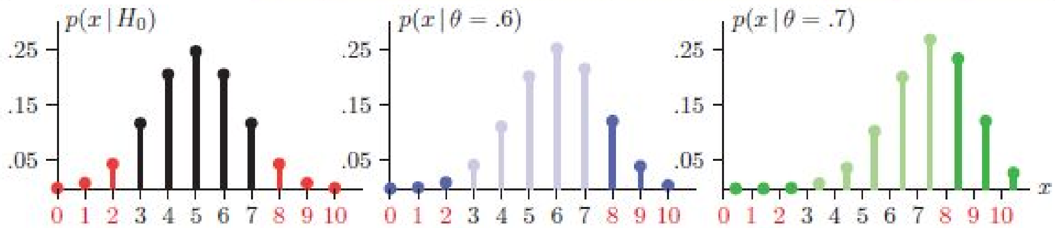
H_A: acuzatul este vinovat.

Nivelul de semnificaţie este probabilitatea de a găsi vinovată o persoană nevinovată. Puterea este probabilitatea de a găsi corect vinovată o parte vinovată. "Dincolo de orice dubiu rezonabil" înseamnă că ar trebui să cerem ca nivelul de semnificaţie să fie foarte mic.

Ipoteze compuse

H_A este compusă în exemplul 3, deci puterea este diferită pentru diferite valori ale lui θ. Extindem tabelul anterior de probabilităţi pentru a include unele valori alternative ale lui θ. Facem acelaşi lucru cu reprezentările. Ca întotdeauna în jocul NHST, ne uităm la verosimilităţi: probabilitatea datelor cunoscând ipoteza.

x	0	1	2	3	4	5	6	7	8	9	10
$H_0: p(x\|\theta = 0.5)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
$H_A: p(x\|\theta = 0.6)$	.000	.002	.011	.042	.111	.201	.251	.215	.121	.040	.006
$H_A: p(x\|\theta = 0.7)$	.000	.0001	.001	.009	.037	.103	.200	.267	.233	.121	.028



Regiunea de respingere şi probabilităţi 0 şi alternative pentru exemplul 3.

Folosim tabelul de probabilităţi pentru a calcula nivelul de semnificaţie şi puterea acestui test.

$$\begin{aligned}
\text{Nivelul de semnificaţie} &= \text{probabilitatea să respingem } H_0 \text{ când } H_0 \text{ este adevărată} \\
&= \text{ prob. că stat. testului e în regiunea de respingere când } H_0 \text{ e adev.} \\
&= \text{ prob. că stat. testului e în regiunea de respingere pe linia } H_0 \text{ a tab.} \\
&= \text{ suma valorilor căsuţelor roşii din linia } \theta = 0.5 \\
&= 0.11.
\end{aligned}$$

$$\begin{aligned}
\text{Puterea când } \theta = 0.6 &= \text{probabilitatea să respingem } H_0 \text{ când } \theta = 0.6 \\
&= \text{ prob. că stat. testului e în regiunea de respingere când } \theta = 0.6 \\
&= \text{ prob. că stat. test. e în regiunea de respingere pe linia } \theta = 0.6 \text{ a tab.} \\
&= \text{ suma valorilor căsuţelor albastru închis din linia } \theta = 0.6 \\
&= 0.18.
\end{aligned}$$

$$\begin{aligned}
\text{Puterea când } \theta = 0.7 &= \text{probabilitatea să respingem } H_0 \text{ când } \theta = 0.7 \\
&= \text{ prob. că stat. testului e în regiunea de respingere când } \theta = 0.7 \\
&= \text{ prob. că stat. test. e în regiunea de respingere pe linia } \theta = 0.7 \text{ a tab.} \\
&= \text{ suma valorilor căsuţelor verde închis din linia } \theta = 0.7 \\
&= 0.384.
\end{aligned}$$

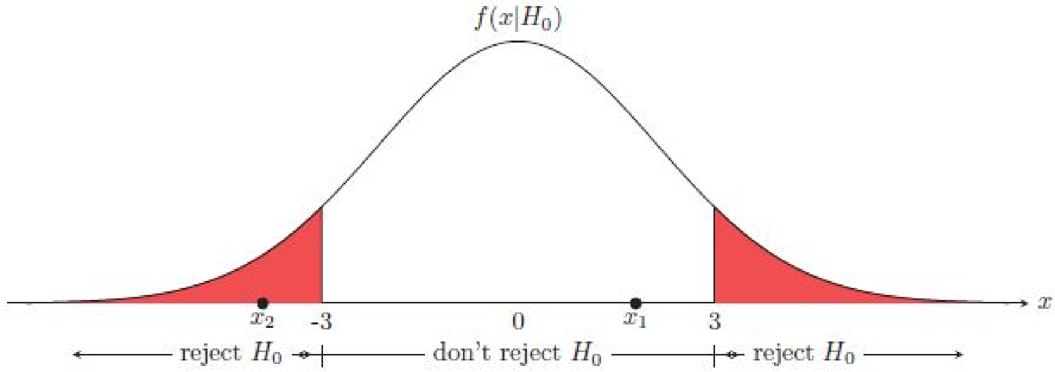
Vedem că puterea este mai mare pentru $\theta = 0.7$ decât pentru $\theta = 0.6$. Aceasta nu este surprinzător deoarece ne aşteptăm să fie mai uşor să recunoaştem că o monedă cu $\theta = 0.7$ este incorectă decât este să recunoaştem că o monedă cu $\theta = 0.6$ este incorectă. Tipic, obţinem putere mai mare când ipoteza alternativă este mai departe de ipoteza 0. În exemplul 3, ar fi foarte greu să distingem o monedă corectă de una cu $\theta = 0.51$.

1.4.4 Schiţe conceptuale

Ilustrăm noţiunile de ipoteză 0, regiune de respingere şi putere cu schiţe ale pdf-urilor pentru ipotezele 0 şi alternative.

Repartiţia 0: regiuni de respingere şi nerespingere

Prima diagramă de mai jos ilustrează o repartiţie 0 cu regiunile de respingere şi nerespingere. Sunt arătate de asemenea 2 posibile statistici ale testului: x_1 şi x_2 ("reject" = "respingem", "don't reject"="nu respingem").



Statistica x_1 a testului este în regiunea de nerespingere. Deci, dacă datele noastre au produs statisica testului x_1, atunci nu vom respinge ipoteza 0 H_0. Pe de altă parte, statistica testului x_2 este în regiunea de respingere, deci, dacă datele noastre au produs x_2, atunci vom respinge ipoteza 0 în favoarea ipotezei alternative.

Sunt câteva lucruri de observat în această imagine.

1. Regiunea de respingere constă din valorile departe de centrul repartiţiei 0.

2. Regiunea de respingere este bilaterală.

3. Ipoteza alternativă nu este menţionată. Respingem sau nu respingem H_0 bazaţi numai pe verosimilitatea $f(x|H_0)$. Ipoteza alternativă H_A ar trebui considerată când alegem o regiune de respingere, dar formal nu joacă niciun rol în respingerea sau nerespingerea lui H_0.

4. Uneori numim regiunea de nerespingere regiunea de acceptare. Aceasta este tehnic incorect deoarece niciodată nu acceptăm cu adevărat ipoteza 0. Sau respingem, sau spunem că datele nu sprijină respingerea lui H_0. Aceasta este adesea rezumat prin afirmaţia: nu putem niciodată demonstra ipoteza 0.

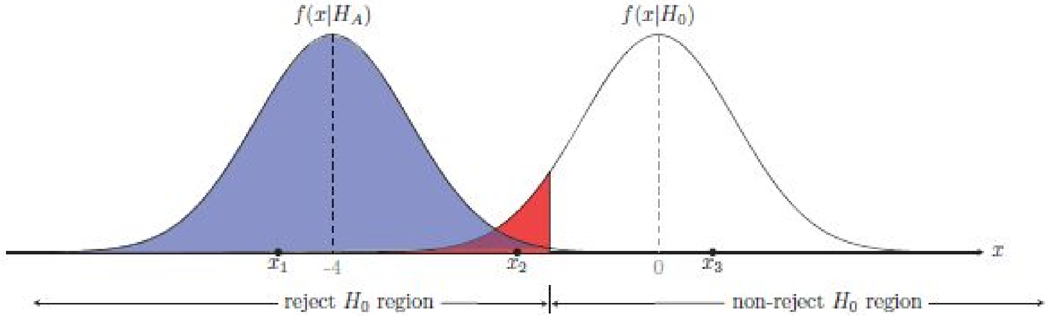
Teste de putere mare sau mică Următoarele 2 figuri arată teste de putere mare, respectiv mică.

Aria umbrită de sub $f(x|H_0)$ reprezintă nivelul de semnificaţie. Reamintim că nivelul de semnificaţie este:

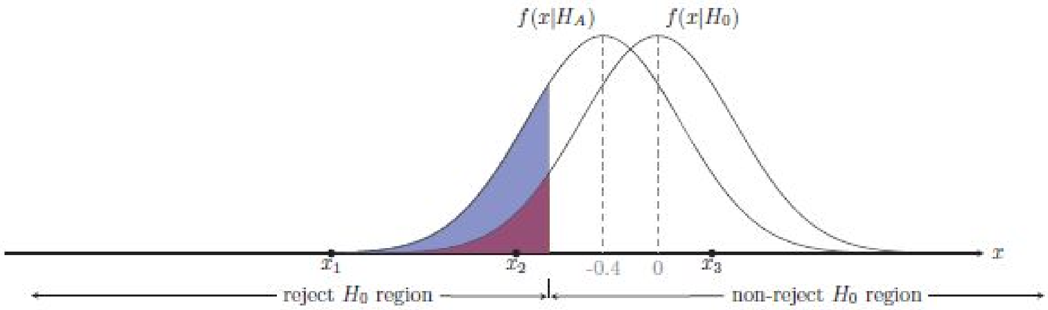
probabilitatea respingerii false a ipotezei 0 când ea este adevărată;

probabilitatea ca statistica testului să cadă în regiunea de respingere chiar dacă H_0 este adevărată.

Analog, regiunea umbrită de sub $f(x|H_A)$ reprezintă puterea, i.e. probabilitatea că statistica testului este în regiunea de respingere a lui H_0 când H_A este adevărată. Ambele teste au acelaşi nivel de semnificaţie, dar dacă $f(x|H_A)$ are o suprapunere considerabilă cu $f(x|H_0)$, puterea este mult mai mică.



Test de putere mare



Test de putere mică

În ambele teste, repartiţiile 0 sunt normale standard. Repartiţia 0, regiunea de respingere şi nivelul de semnificaţie sunt aceleaşi. (Nivelul de semnificaţie este aria roşie/mov de sub $f(x|H_0)$ şi de deasupra regiunii de respingere.) În figura de deasupra vedem că mediile celor 2 repartiţii sunt la distanţă de 4 deviaţii standard una faţă de cealaltă. Doarece ariile de sub densităţi au o foarte mică suprapunere, testul are putere mare. Adică, dacă datele x sunt extrase din H_A, vor fi aproape sigur în regiunea de respingere a lui H_0. De exemplu x_3 ar fi un rezultat foarte surprinzător pentru repartiţia H_A.

În figura de jos vedem că mediile sunt la doar 0.4 deviaţii standard una de alta. Deoarece ariile de sub densităţi au multă suprapunere, testul are putere mică. Adică, dacă datele x sunt extrase din H_A, este foarte probabil să se afle în regiunea de nerespingere a lui H_0. De exemplu, x_3 n-ar fi un rezultat foarte surprinzător pentru repartiţia H_A.

Tipic, putem creşte puterea unui test crescând numărul de date şi, prin aceasta, scăzând dispersia repartiţiilor 0 şi alternativă. În proiectarea experimentului este important să determinăm dinainte numărul de încercări sau de subiecţi de care avem nevoie pentru a atinge o putere dorită.

Exemplul 7. Presupunem că un medicament pentru o boală este comparat cu un placebo. Alegem ipotezele 0 şi alternativă astfel:

H_0 = medicamentul nu este mai bun ca placebo;

H_A = medicamentul este mai bun ca placebo.

Puterea testului ipotezei este probabilitatea ca testul să concluzioneze că medicamentul este mai bun, dacă este într-adevăr mai bun. Nivelul de semnificaţie este probabilitatea ca testul să concluzioneze că medicamentul este mai bun, când de fapt nu este mai bun.

1.5 Proiectarea unui test al ipotezei

Formal, tot ce necesită un test al ipotezei este H_0, H_A, o statistică a testului şi o regiune de respingere. În practică proiectarea este adesea făcută folosind paşii următori:

1. Alege ipoteza 0 H_0.

Adesea alegem H_0 să fie simplă. Sau, adesea alegem H_0 să fie cea mai simplă şi cea mai prudentă explicaţie, i.e. medicamentul nu are efect, fără ESP, moneda este corectă.

2. Decide dacă H_A este unilaterală sau bilaterală.

În exemplul 3 am vrut să ştim dacă moneda a fost incorectă. O monedă incorectă poate fi deplasată pentru sau împotriva aversului, deci $H_A : \theta \neq 0.5$ este o ipoteză bilaterală. Dacă ne pasă doar dacă moneda este deplasată sau nu pentru avers am putea folosi o ipoteză unilaterală $H_A : \theta > 0.5$.

3. Alege o statistică a testului.

De exemplu, media de selecţie, totalul de seleţie sau dispersia de selecţie. Unele statistici standard sunt z, t şi χ^2. Repartiţiile care merg cu aceste statistici sunt totdeauna condiţionate de ipoteza 0, adică vom calcula verosimilităţi ca $f(z|H_0)$.

4. Alege un nivel de semnificaţie şi determină regiunea de respingere.

Vom folosi de obicei α pentru a nota nivelul de semnificaţie. Paradigma Neyman-Pearson este de a alege α înainte. Valori tipice sunt 0.1, 0.05, 0.01. Reamintim că nivelul de semnificaţie este probabilitatea unei erori de tip I, i.e. respingerea incorectă a ipotezei 0 când ea este adevărată. Valoarea pe care o alegem va depinde de consecinţele unei erori de tipul I. Odată ce nivelul de semnificaţie este ales, putem determina regiunea de respingere în coada (cozile) repartiţiei 0. În exemplul 3, H_A este bilaterală, deci regiunea de respingere este împărţită între cele 2 cozi ale repartiţiei 0. Această repartiţie este dată în următorul tabel:

x	0	1	2	3	4	5	6	7	8	9	10
$p(x\|H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001

Dacă punem $\alpha = 0.05$, atunci regiunea de respingere trebuie să conţină cel mult 0.05 probabilitate. Pentru o regiune bilaterală, obţinem

$$\{0, 1, 9, 10\}.$$

Dacă punem $\alpha = 0.01$, regiunea de respingere este

$$\{0, 10\}.$$

Presupunem că schimbăm H_A în ”moneda este deplasată în favoarea aversului”. Acum avem o ipoteză unilaterală $\theta > 0.5$. Regiunea noastră de respingere va fi acum în coada dreaptă, deoarece nu vrem să respingem H_0 în favoarea lui H_A dacă obţinem un număr mic de aversuri. Acum, dacă $\alpha = 0.05$, regiunea de respingere este domeniul unilateral

$$\{9, 10\}.$$

Dacă punem $\alpha = 0.01$, regiunea de respingere este

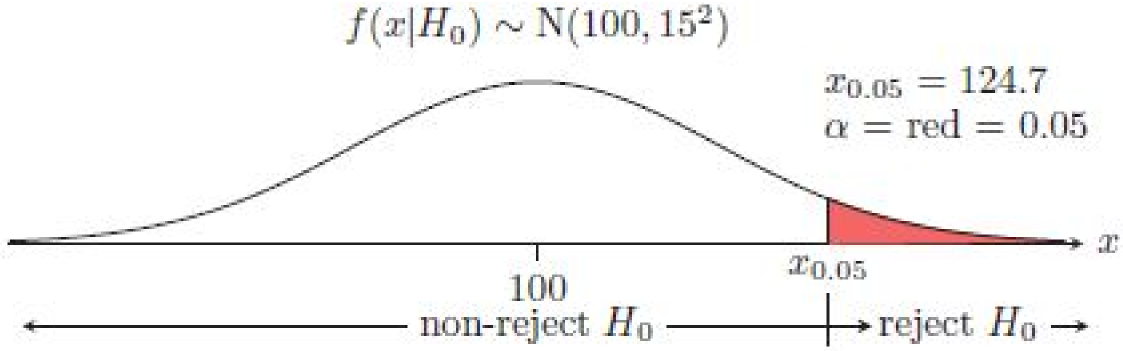
$$\{10\}.$$

5. Determină puterea (puterile)
După cum am văzut în exemplul 3, odată ce regiunea de respingere este stabilită, putem determina puterea testului la diverse valori ale ipotezei alternative.
Exemplul 8. (Consecinţele semnificaţiei.) Dacă $\alpha = 0.1$ aşteptăm o rată a erorii de tipul I de 10%. Adică, aşteptăm să respingem ipoteza 0 în 10% din acele experimente unde ipoteza H_0 este adevărată. Faptul dacă 0.1 este un nivel de semnificaţie rezonabil depinde de deciziile care vor fi făcute folosindu-l.
De exemplu, dacă faci un experiment pentru a determina dacă ciocolata ta are mai mult de 72% cacao, atunci o rată a erorii de tipul I de 10% este probabil în regulă. Adică, a considera fals că o ciocolată 72% are mai mult de 72%, este probabil acceptabil. Pe de altă parte, dacă laboratorul tău criminalistic identifică amprente pentru un proces de crimă, atunci o rată de 10% a erorii de tipul I, i.e. a pretinde greşit că amprentele găsite la locul crimei au aparţinut cuiva care în realitate este nevinovat, este categoric inacceptabilă.
Semnificaţia pentru o ipoteză 0 compusă. Dacă H_0 este compusă, atunci $P(\text{eroarea de tipul I})$ depinde de care membru al lui H_0 este adevărat. În acest caz nivelul de semnificaţie este definit ca maximul acestor probabilităţi.

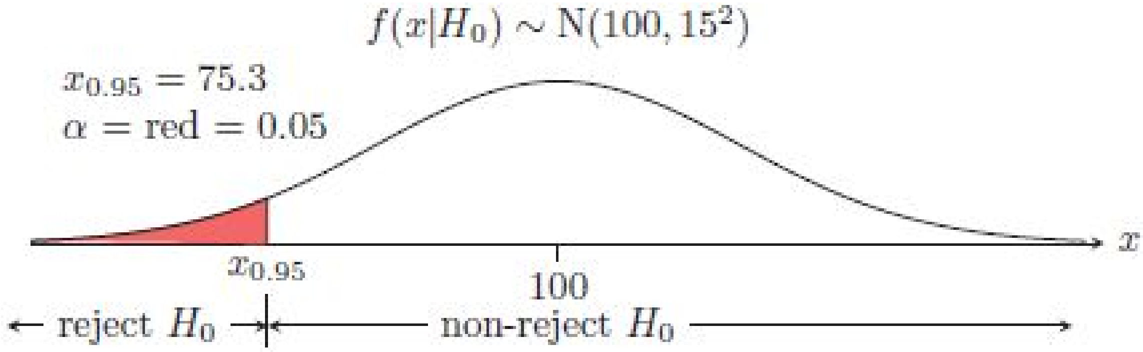
1.6 Valori critice

Valorile critice sunt ca cuantilele cu excepţia faptului că se referă la probabilitatea de la dreapta valorii, în loc de cea de la stânga.

Exemplul 9. Folosiţi R pentru a afla valoarea critică 0.05 pentru repartiţia normală standard.
Răspuns. Notăm această valoare critică cu $z_{0.05}$. Valoarea critică $z_{0.05}$ este chiar 0.95 cuantila, i. e. are 5% probabilitate la dreapta ei şi de aceea 95% probabilitate la stânga. O calculăm cu funcţia R qnorm: `qnorm(0.95,0,1)` sau `qnorm(0.95)` şi obţinem 1.644854.
Într-un test de semnificaţie tipic, regiunea de respingere constă din una sau ambele cozi ale repartiţiei 0. Valoarea care marchează începutul regiunii de respingere este o valoare critică.
Exemplul 10. Valori critice şi regiuni de respingere. Presupunem că statistica x a testului nostru are repartiţia 0 $N(100, 15^2)$, i. e. $f(x|H_0) \sim N(100, 15^2)$. Presupunem de asemenea că regiunea noastră de respingere este în dreapta şi avem un nivel de semnificaţie de 0.05. Aflaţi valoarea critică şi schiţaţi repartiţia 0 şi regiunea de respingere.
Răspuns. Notaţia folosită pentru valoarea critică cu coada dreaptă conţinând 0.05 probabilitate este $x_{0.05}$. Valoarea critică $x_{0.05}$ este chiar 0.95 cuantila, i. e. are 5% probabilitate la dreapta ei şi de aceea 95% probabilitate la stânga. O calculăm cu funcţia R qnorm: `qnorm(0.95,100,15)` şi obţinem $124.6728 \approx 124.7$. Aceasta este arătată în figura de mai jos.

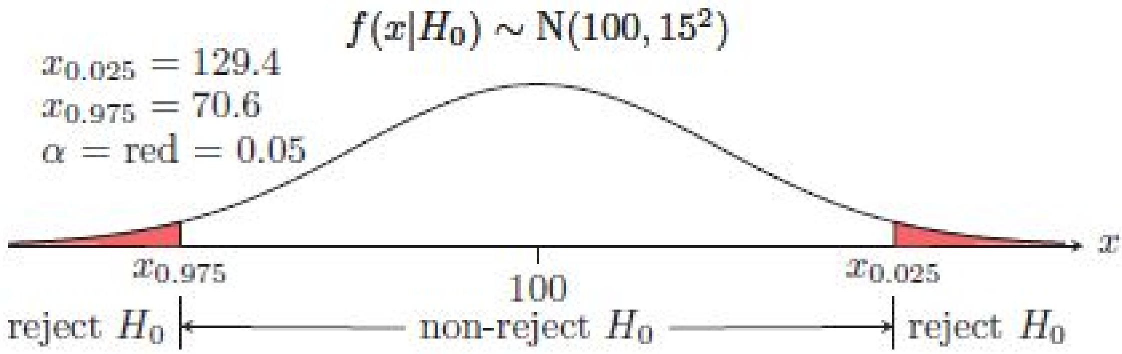


Exemplul 11. Valori critice şi regiuni de respingere. Repetaţi exemplul precedent pentru o regiune de respingere în stânga cu nivelul de semnificaţie 0.05.
Răspuns. În acest caz valoarea critică are 0.05 probabilitate la stânga şi de aceea 0.95 probabilitate la dreapta. Deci o notăm cu $x_{0.95}$. Deoarece este 0.05 cuantila, o calculăm în R cu `qnorm(0.05,100,15)` şi obţinem $75.3272 \approx 75.3$.



Exemplul 12. Valori critice. Repetaţi exemplul anterior pentru o regiune de respingere bilaterală. Puneţi jumătate din semnificaţie în fiecare coadă.

Răspuns. Pentru a avea un total de semnificaţie de 0.05 punem 0.025 în fiecare coadă. Adică, coada stângă începe la $x_{0.975} = q_{0.025}$ şi coada dreaptă începe la $x_{0.025} = q_{0.975}$. Calculăm aceste valori cu `qnorm(0.025,100,15)` şi `qnorm(0.975,100,15)`. Valorile aproximative sunt arătate în figura de mai jos.



1.7 Valori p

În practică, oamenii adesea specifică nivelul de semnificaţie şi fac testul de semnificaţie folosind valori p.

Dacă valoarea p este mai mică decât nivelul de semnificaţie α, atunci respingem H_0. Altfel nu respingem H_0.

Definiţie. p-valoarea este probabilitatea, presupunând ipoteza 0, de a vedea datele cel puţin la fel de extreme ca datele experimentale. Ce înseamnă "cel puţin la fel de extreme" depinde de proiectarea experimentului.

Ilustrăm definiţia şi utilizarea valorilor p cu un exemplu unilateral. Acest exemplu introduce de asemenea testul z. Toate acestea înseamnă că statistica testului nostru este normală standard (sau aproximativ normală standard).

Exemplul 13. testul z pentru ipoteze normale
IQ-ul este repartizat normal în populaţie conform unei repartiţii $N(100, 15^2)$. Presupunem că cei mai mulţi dintre studenţii FMI au IQ-ul peste medie, deci vom formula următoarele ipoteze:

$$\begin{aligned} H_0 &= \text{IQ-urile studenţilor FMI sunt repartizate identic cu cele ale populaţiei generale} \\ &= \text{IQ-urile studenţilor din FMI au o repartiţie } N(100, 15^2). \end{aligned}$$

$$\begin{aligned} H_A &= \text{IQ-urile studenţilor FMI tind să fie mai mari decât cele ale populaţiei generale} \\ &= \text{media IQ-urilor studenţilor din FMI este mai mare ca 100.} \end{aligned}$$

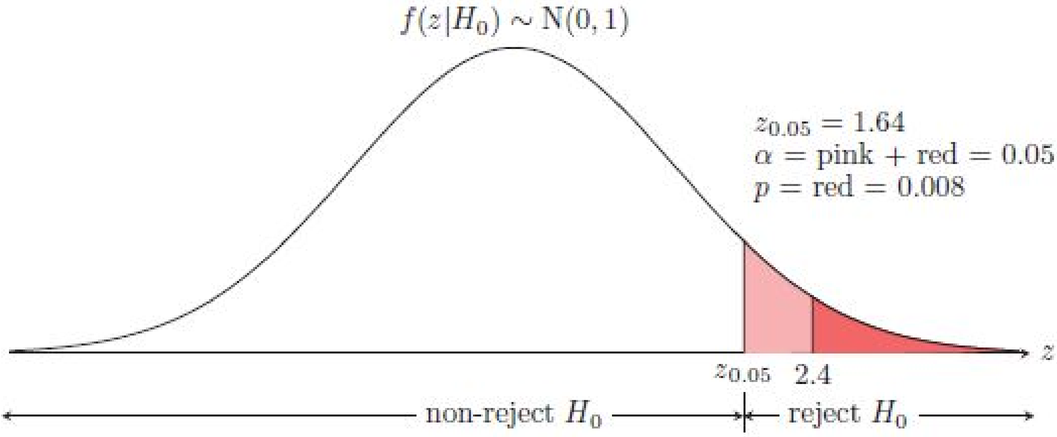
Observaţi că H_A este unilaterală.
Presupunem că testăm 9 studenţi şi aflăm că au un IQ mediu de $\overline{x} = 112$. Putem respinge H_0 la nivelul de semnificaţie $\alpha = 0.05$?
Răspuns. Pentru a calcula p, întâi standardizăm datele: Sub ipoteza 0, $\overline{x} \sim N(100, 15^2/9)$ şi de aceea

$$z = \frac{\overline{x} - 100}{15/\sqrt{9}} = \frac{36}{15} = 2.4 \sim N(0, 1).$$

Adică, repartiţia 0 pentru z este normală standard. Numim z o statistică z, o vom utiliza ca statistica testului nostru.
Pentru o ipoteză alternativă la dreapta fraza "datele cel puţin la fel de extreme" este o coadă unilaterală la dreapta lui z. Valoarea p este atunci

$$p = P(Z \geq 2.4) = \texttt{1-pnorm(2.4,0,1)=0.008197536}.$$

Deoarece $p \leq \alpha$ respingem ipoteza 0. Formulăm concluzia noastră astfel: Respingem ipoteza 0 în favoarea ipotezei că studenţii FMI au IQ-uri mai mari în medie. Am făcut aceasta la nivelul de semnificaţie 0.05 cu o valoare p de 0.008.
Observaţii: 1. Media $\overline{x} = 112$ este aleatoare: dacă facem experimentul din nou am putea obţine o valoare diferită pentru $\overline{x}$.
2. Am fi putut folosi statistica $\overline{x}$ direct. Standardizarea este preferabilă deoarece, cu practica, vom simţi bine sensul diferitelor valori z.
Justificarea respingerii lui H_0 când $p \leq \alpha$ este dată în următoarea figură.



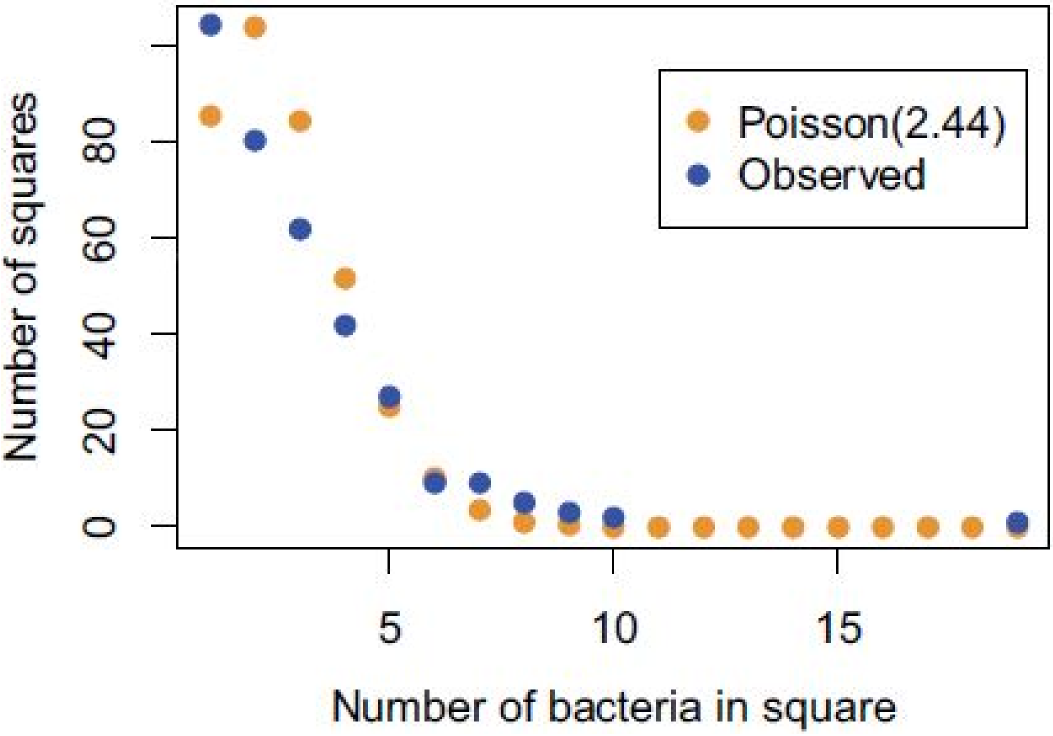
În acest exemplu $\alpha = 0.05, z_{0.05} = 1.64$ şi regiunea de respingere este domeniul de la dreapta lui $z_{0.05}$. De asemenea, $z = 2.4$ şi valoarea p este probabilitatea de la dreapta lui z. Figura ilustrează că
$z = 2.4$ este în regiunea de respingere;
aceasta este acelaşi lucru cu z este la dreapta lui $z_{0.05}$;
aceasta este acelaşi lucru cu probabilitatea de la dreapta lui z este mai mică decât 0.05,
ceea ce înseamnă $p < 0.05$.

1.8 Mai multe exemple

Testarea ipotezelor este larg utilizată în statistica deductivă.
Exemplul 14. Statistica χ^2 şi bunătatea potrivirii.
Pentru a testa nivelul contaminării bacteriene, a fost vărsat lapte peste o reţea cu 400 de pătrate. Cantitatea de bacterii din fiecare pătrat a fost numărată. Rezumăm datele în tabelul de mai jos. Ultima linie a tabelului dă numărul de pătrate diferite care aveau o cantitate dată de bacterii.

Amount of bacteria	0	1	2	3	4	5	6	7	8	9	10	19
Number of squares	56	104	80	62	42	27	9	9	5	3	2	1

Cantitatea medie de bacterii pe pătrat este 2.44. Deoarece repartiţia Poisson(λ) este folosită pentru a modela numărările evenimentelor relativ rare şi parametrul λ este media repartiţiei, decidem să vedem dacă aceste date ar putea veni dintr-o repartiţie Poisson. Pentru a face asta, întâi comparăm grafic frecvenţele observate cu cele aşteptate din Poisson(2.44).



Facem un test al ipotezei cu
H_0: datele vin dintr-o repartiţie Poisson(2.44).
H_A: datele vin dintr-o repartiţie diferită.
Folosim o statistică χ^2, numită aşa deoarece ea are aproximativ o repartiţie χ^2. Pentru a calcula X^2 întâi combinăm ultimele câteva celule în tabel astfel încât numărul aşteptat minim este în jur de 5 (o regulă a degetului mare generală în acest joc.)
Numărul aşteptat de pătrate cu o anumită cantitate de bacterii vine din considerarea a 400 de încercări dintr-o repartiţie Poisson(2.44), de exemplu, cu $l = 2.44$ numărul aşteptat de pătrate cu 3 bacterii este $400 \cdot e^{-l}\frac{l^3}{3!} = 84.4$.
Statistica χ^2 este $\sum \frac{(O_i - E_i)^2}{E_i}$, unde O_i este numărul observat şi E_i este numărul aşteptat.

Number per square	0	1	2	3	4	5	6	> 6
Observed	56	104	80	62	42	27	9	20
Expected	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
Component of X^2	12.8	4.2	5.5	6.0	1.7	0.14	0.15	44.5

Adunând obţinem $X^2 = 74.99$.
Doarece media (2.44) şi numărul total de încercări (400) sunt fixate, cele 8 celule au doar 6 grade de libertate. Deci, presupunând H_0, statistica noastră X^2 are (aproximativ) o repartiţie χ^2_6. Folosind această repartiţie,

$P(X^2 > 74.99) = 0$. Astfel, respingem decisiv ipoteza 0 în favoarea ipotezei alternative că repartiţia nu este Poisson(2.44).
Pentru a analiza mai departe, privim la componentele individuale ale lui X^2. Sunt contribuţii mari în coada distribuţiei, deci acolo potrivirea nu merge.
Exemplul 15. Testul t al lui Student.
Presupunem că vrem să comparăm un tratament medical pentru creşterea speranţei de viaţă cu un placebo. Dăm la n oameni tratamentul şi la m oameni placebo. Fie $X_1, ..., X_n$ numerele de ani pe care oamenii îi trăiesc după primirea tratamentului. Analog, fie $Y_1, ..., Y_m$ numerele de ani pe care oamenii îi trăiesc după primirea placebo. Fie $\overline{X}$ şi $\overline{Y}$ mediile de selecţie. Vrem să ştim dacă diferenţa dintre $\overline{X}$ şi $\overline{Y}$ este semnificativă statistic. Formulăm aceasta ca un test al ipotezei. Fie μ_X şi μ_Y mediile (necunoscute).

$$H_0 : \mu_X = \mu_Y,\ H_A : \mu_X \neq \mu_Y.$$

Cu anumite presupuneri şi o formulă adecvată pentru eroarea standard unificată s_p statistica testului $t = \frac{\overline{X}-\overline{Y}}{s_p}$ are o repartiţie t cu $n + m - 2$ grade de libertate. Deci regiunea noastră de respingere este determinată de un prag t_0 cu $P(t > t_0) = \alpha$.

2 Testarea semnificaţiei ipotezei 0 II

2.1 Scopurile învăţării

1. Să poată să listeze paşii comuni tuturor testelor semnificaţiei ipotezei 0.
2. Să poată defini şi calcula probabilitatea erorilor de tipul I şi II.
3. Să poată să caute şi să aplice t-teste pentru unul sau 2 eşantioane.

2.2 Introducere

Fiecare test face unele presupuneri despre date - adesea că sunt provenite dintr-o repartiţie normală. Toate testele urmează acelaşi tipar. Doar calculul statisticii testului şi tipul repartiţiei 0 se schimbă.

2.3 Configurarea şi folosirea unui test de semnificaţie

Există o mulţime standard de paşi care se fac pentru a configura şi a utiliza un test al semnificaţiei ipotezei 0.
1. Proiectează un experiment pentru a colecta datele şi a alege o statistică x a testului pentru a fi calculată din date. Cerinţa cheie aici este să ştim repartiţia 0 $f(x|H_0)$. Pentru a calcula puterea, trebuie să cunoaştem şi

repartiţia alternativă $f(x|H_A)$.
2. Decide dacă testul este unilateral sau bilateral bazat pe H_A şi forma repartiţiei 0.
3. Alege un nivel de semnificaţie α pentru respingerea ipotezei 0. Dacă se poate, calculează puterea corespunzătoare a testului.
4. Foloseşte experimentul pentru a colecta datele $x_1, x_2, ..., x_n$.
5. Calculează statistica testului x.
6. Calculează valoarea p corespunzătoare lui x folosind repartiţia 0.
7. Dacă $p < \alpha$, respinge ipoteza 0 în favoarea ipotezei alternative.

Observaţii.
1. În loc de a alege un nivel de semnificaţie, putem alege o regiune de respingere şi să respingem H_0 dacă x este în această regiune. Nivelul de semnificaţie corespunzător este atunci probabilitatea ca x să fie în regiunea de respingere.
2. Ipoteza 0 este adesea numită "ipoteza precaută". Cu cât punem nivelul de semnificaţie mai mic, cu atât mai multe "dovezi" vor fi necesare înainte de a respinge ipoteza noastră precaută în favoarea unei alternative. Este o practică standard a publica valoarea p însăşi, astfel încât alţii să poată trage concluziile lor proprii.
3. **Un punct cheie de confuzie:** Un nivel de semnificaţie de 0.05 nu înseamnă că testul greşeşte doar în 5% din cazuri. Înseamnă că dacă ipoteza 0 este adevărată, atunci probabilitatea ca testul să o respingă din greşeală este 5%. Puterea testului măsoară acurateţea testului când ipoteza alternativă este adevărată. Anume, puterea testului este probabilitatea de a respinge ipoteza 0 dacă ipoteza alternativă este adevărată. De aceea probabilitatea de a eşua fals să respingem ipoteza 0 este 1 minus puterea.

Erori. Putem rezuma aceste 2 tipuri de erori şi probabilităţile lor după cum urmează:

$$\begin{aligned}
\text{Eroarea de tipul I} &= \text{ respingerea lui } H_0 \text{ când } H_0 \text{ este adevărată.}\\
\text{Eroarea de tipul II} &= \text{ nerespingerea lui } H_0 \text{ când } H_A \text{ este adevărată.}\\
P(\text{eroarea de tipul I}) &= \text{ probabilitatea de a respinge fals } H_0\\
&= P(\text{statistica testului este în regiunea de respingere}|H_0)\\
&= \text{ nivelul de semnificaţie al testului}\\
P(\text{eroarea de tipul II}) &= \text{ probabilitatea de a nu respinge fals } H_0\\
&= P(\text{statistica testului este în regiunea de acceptare}|H_A)\\
&= 1 - \text{ putere.}
\end{aligned}$$

Analogii utile. În termeni de testare medicală pentru o boală, o eroare de tipul I este un rezultat fals pozitiv şi o eroare de tipul II este un rezultat fals negativ. În termenii unui proces juridic, o eroare de tipul I este condamnarea unui nevinovat şi eroarea de tipul II este achitarea unui vinovat.

2.4 Înţelegerea unui test de semnificaţie

Întrebări de pus:
1. Cum au colectat datele? Care este schema experimentului?
2. Care sunt ipotezele 0 şi alternativă?
3. Ce tip de test de semnificaţie a fost folosit?
Se potrivesc datele lor cu criteriile necesare pentru a utiliza acest tip de test? Cât de robust este testul la deviaţii de la aceste criterii?
4. De exemplu, unele teste compară 2 grupe de date presupunând că grupele sunt provenite din repartiţii care au aceeaşi dispersie. Acest fapt trebuie verificat înainte de aplicarea testului. Adesea verificarea este făcută folosind alt test de semnificaţie proiectat pentru a compara dispersiile a 2 grupuri de date.
5. Cum este calculată valoarea p?
Un test de semnificaţie vine cu o statistică a testului şi o repartiţie 0. În cele mai multe teste valoarea p este

$$p = P(\text{datele cel puţin la fel de extreme ca ce am obţinut}|H_0)$$

Ce înseamnă "datele cel puţin la fel de extreme ca datele pe care le-am văzut"? I.e., este testul unilateral sau bilateral?
6. Care este nivelul de semnificaţie pentru acest test? Dacă $p < \alpha$, atunci experimentatorul va respinge H_0 în favoarea H_A.

2.5 Teste t

Multe teste de semnificaţie presupun că datele provin dintr-o repartiţie normală, deci înainte de a folosi un astfel de test trebuie să examinăm datele pentru a vedea dacă ipoteza normalităţii este rezonabilă. Reprezentarea unei histograme este un start bun. Ca şi testul z, testele t cu o selecţie şi cu 2 selecţii pe care le considerăm mai jos pleacă de la această presupunere de normalitate.

2.5.1 Test z

Recapitulăm testul z.
Date: presupunem $x_1, x_2, ..., x_n \sim N(\mu, \sigma^2)$, unde μ este necunoscută şi σ este cunoscută.
Ipoteza 0: $\mu = \mu_0$ pentru o anumită valoare specifică μ_0.
Statistica testului: $z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ = media standardizată.

Repartiţia 0: $f(z|H_0)$ este pdf a $Z \sim N(0,1)$.
Valoare p unilaterală (partea dreaptă): $p = P(Z > z|H_0)$.
Valoare p unilaterală (partea stângă): $p = P(Z < z|H_0)$.
Valoare p bilaterală: $p = P(|Z| > |z||H_0)$.
Exemplul 1. Presupunem că avem datele care au o repartiţie normală de medie necunoscută μ şi dispersie cunoscută 4. Fie ipoteza 0: $\mu = 0$. Fie ipoteza alternativă $H_A : \mu > 0$. Presupunem că obţinem următoarele date:

$$1, 2, 3, 6, -1.$$

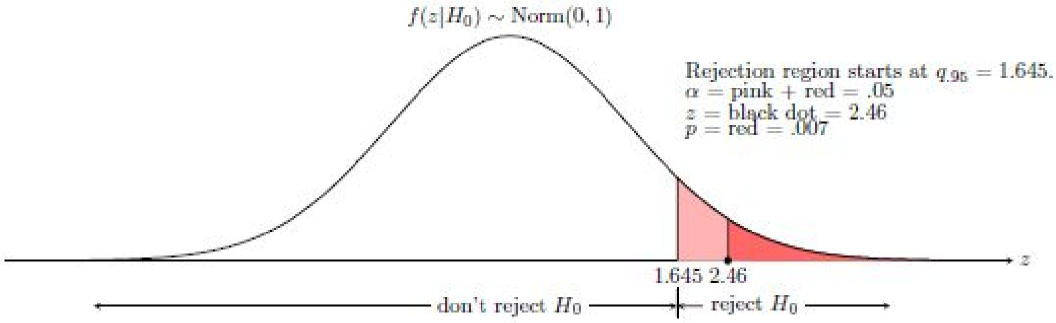
La un nivel de semnificaţie de $\alpha = 0.05$ ar trebui să respingem ipoteza 0?
Răspuns. Sunt 5 date cu media $\overline{x} = 2.2$. Deoarece avem date normale cu o dispersie cunoscută ar trebui să folosim un z test. Statistica noastră z este

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.2 - 0}{2/\sqrt{5}} = 2.46.$$

Testul nostru este unilateral deoarece ipoteza alternativă este unilaterală. Deci (folosind `R`) valoarea p a noastră este

$$p = P(Z > z) = P(Z > 2.46) = 1 - \texttt{pnorm}(2.46) \approx 0.007.$$

Deoarece $p < 0.05$, respingem ipoteza 0 în favoarea ipotezei alternative $\mu > 0$. Putem vizualiza testul astfel:



2.5.2 Repartiţia t a lui Student

"Student" este pseudonimul utilizat de William Gosset care a descris primul acest test şi această repartiţie. Vezi `http://en.wikipedia.org/wiki/Student's_t-test`.
Repartiţia t este simetrică şi are formă de clopot ca repartiţia normală. Are un parametru *df* care stă pentru grade de libertate. Pentru *df* mic repartiţia t are mai multă probabilitate în cozile ei ca repartiţia normală standard. Când *df* creşte, $t(df)$ devine din ce în ce mai asemănătoare cu repartiţia normală

standard.
Iată o aplicaţie care arată $t(df)$ şi o compară cu repartiţia normală standard: `http://mathlets.org/mathlets/t-distribution`.
Ca de obicei în `R`, funcţiile `pt`,`dt`, `qt`, `rt` corespund la cdf, pdf, cuantile, respectiv generarea unui eşantion aleator dintr-o repartiţie t. Reamintim că puteţi tasta `?dt` în `RStudio` pentru a vedea fişierul de ajutor care specifică parametrii lui `dt`. De exemplu, `pt(1.65,3)` calculează probabilitatea ca x să fie mai mic sau egal cu 1.65 dat fiind că x provine dintr-o repartiţie t cu 3 grade de libertate, i.e. $P(x \leq 1.65)$ dat fiind că $x \sim t(3)$.

2.5.3 Testul t cu o selecţie

Pentru testul z, am presupus că dispersia repartiţiei datelor este cunoscută. Totuşi, adesea nu ştim σ şi trebuie s-o estimăm din date. În aceste cazuri, folosim un test t cu o selecţie în locul unui test z şi media Studentizată în locul mediei standardizate.
Date: presupunem $x_1, x_2, ..., x_n \sim N(\mu, \sigma^2)$, unde atât μ cât şi σ sunt necunoscute.
Ipoteza 0: $\mu = \mu_0$ pentru o anumită valoare specifică μ_0.
Statistica testului:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}},$$

unde

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Aici t este numită media Studentizată şi s^2 este numită dispersia de selecţie. Ultima este o estimare a adevăratei dispersii σ^2.
Repartiţia 0: $f(t|H_0)$ este pdf pentru $T \sim t(n-1)$, repartiţia t cu $n-1$ grade de libertate.*
Valoare p unilaterală (partea dreaptă): $p = P(T > t|H_0)$.
Valoare p unilaterală (partea stângă): $p = P(T < t|H_0)$.
Valoare p bilaterală: $p = P(|T| > |t|)$.
*Există o teoremă (nu o presupunere) că dacă datele sunt normale cu media μ_0, atunci media Studentizată are o repartiţie t. Puteţi vedea demonstraţia în `http://en.wikipedia.org/wiki/Student's_t-distribution#Derivation`.
Exemplul 2. Presupunem acum că în exemplul precedent dispersia este necunoscută. Adică, avem date care au o repartiţie normală cu medie necunoscută μ şi dispersie necunoscută σ^2. Presupunem că obţinem aceleaşi date:

$$1, 2, 3, 6, -1.$$

Ca mai sus, fie $H_0 : \mu = 0$ şi $H_A : \mu > 0$. La nivelul de semnificaţie de $\alpha = 0.05$ ar trebui să respingem ipoteza 0?
Răspuns. Sunt 5 date cu media $\overline{x} = 2.2$. Deoarece avem date normale cu medie şi dispersie necunoscute ar trebui să folosim un t test pentru o selecţie. Calculând dispersia de selecţie obţinem

$$s^2 = \frac{1}{4}((1-2.2)^2 + (2-2.2)^2 + (3-2.2)^2 + (6-2.2)^2 + (-1-2.2)^2) = 6.7.$$

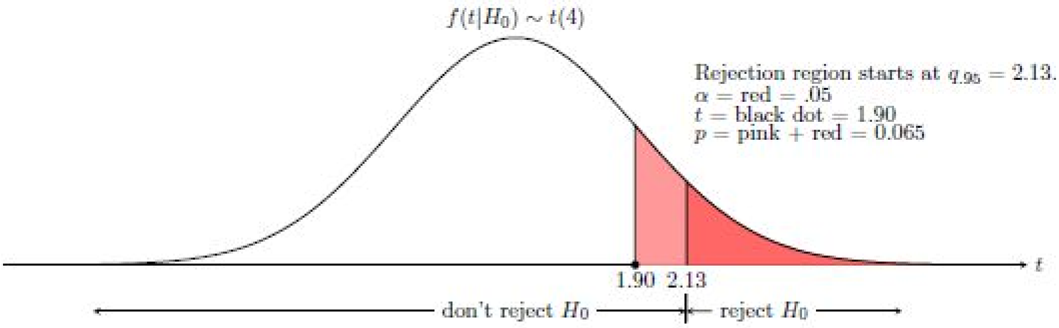
(Puteam folosi şi comenzile în `R`
`x=c(1,2,3,6,-1)`
`var(x)`.)
Statistica noastră t este

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{2.2 - 0}{\sqrt{6.7}/\sqrt{5}} = 1.900511.$$

Testul nostru este unilateral deoarece ipoteza alternativă este unilaterală. Deci (folosind `R`), valoarea p este

$$p = P(T > t) = P(T > 1.900511) = \mathtt{1 - pt(1.900511,4) = 0.06508106}.$$

Deoarece $p > 0.05$, nu respingem ipoteza 0. Putem vizualiza testul astfel:



2.5.4 Testul t pentru 2 selecţii cu dispersii egale

Considerăm în continuare cazul comparării mediilor a 2 selecţii. De exemplu, putem fi interesaţi în compararea eficienţelor medii ale 2 tratamente medicale.
Date: Presupunem că avem 2 mulţimi de date provenite din repartiţii normale

$$x_1, x_2, ..., x_n \sim N(\mu_1, \sigma^2)$$
$$y_1, y_2, ..., y_m \sim N(\mu_2, \sigma^2),$$

unde mediile μ_1 şi μ_2 şi dispersia σ^2 sunt toate necunoscute. Presupunem că cele 2 repartiţii au aceeaşi dispersie. Sunt n date în primul grup şi m date în al 2-lea.
Ipoteza 0: $\mu_1 = \mu_2$ (valorile lui μ_1 şi μ_2 nu sunt specificate).
Statistica testului:

$$t = \frac{\overline{x} - \overline{y}}{s_p},$$

unde s_p^2 este dispersia combinată

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right).$$

Aici s_x^2 şi s_y^2 sunt dispersiile de selecţie ale x_i-urilor şi respectiv y_j-urilor.
Repartiţia 0: $f(t|H_0)$ este pdf pentru $T \sim t(n+m-2)$.
Valoare p unilaterală (partea dreaptă): $p = P(T > t|H_0)$.
Valoare p unilaterală (partea stângă): $p = P(T < t|H_0)$.
Valoare p bilaterală: $p = P(|T| > |t|)$.
Observaţia 1. Unii autori folosesc o notaţie diferită. Ei definesc dispersia combinată ca

$$s^2_{p\text{-alţi-autori}} = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

şi ce am numit dispersia combinată ei arată că este dispersia estimată a lui $\overline{x} - \overline{y}$. Adică,

$$s_p^2 = s^2_{p\text{-alţi-autori}} \cdot \left(\frac{1}{n} + \frac{1}{m}\right) \approx s^2_{\overline{x}-\overline{y}}.$$

Observaţia 2. Există o versiune a t-testului cu 2 selecţii care permite celor 2 grupe să aibă dispersii diferite. În acest caz statistica testului este mai complicată, dar `R` o tratează la fel de uşor.
Exemplul 3. Următoarele date vin dintr-un studiu real în care 1408 femei au fost internate la o maternitate pentru (i) motive medicale sau (ii) prin internare de urgenţă. Durata sarcinii este măsurată în săptămâni complete de la începutul ultimului ciclu. Putem rezuma datele astfel:
Medicale: 775 de observaţii cu $\overline{x}_M = 39.08$ şi $s_M^2 = 7.77$.
Urgenţe: 633 de observaţii cu $\overline{x}_E = 39.6$ şi $s_E^2 = 4.95$.
Configuraţi şi folosiţi un test t cu 2 selecţii pentru a investiga dacă durata medie diferă pentru cele 2 grupuri.
Ce presupuneri aţi făcut?
Răspuns. Dispersia combinată pentru aceste date este

$$s_p^2 = \frac{774 \cdot 7.77 + 632 \cdot 4.95}{1406}\left(\frac{1}{775} + \frac{1}{633}\right) = 0.01866256.$$

Statistica t pentru repartiţia 0 este

$$\frac{\overline{x}_M - \overline{x}_E}{s_P} = -3.806429.$$

Avem 1406 grade de libertate. Folosind `R` pentru a calcula valoarea p bilaterală, obţinem

$$p = P(|T| > |t|) = 2 * \texttt{pt}(-3.806429, 1406) = 0.000147048.$$

p este foarte mică, mult mai mică decât $\alpha = 0.05$ sau $\alpha = 0.01$. De aceea respingem ipoteza 0 în favoarea alternativei că există o diferenţă între duratele medii.
În loc să calculăm valoarea p bilaterală exact folosind o repartiţie t am fi putut observa că, cu 1406 de grade de libertate, repartiţia t este în esenţă normală standard şi 3.806429 este aproape 4 deviaţii standard. Deci

$$p = P(|T| \geq 3.806429) \approx P(|Z| \geq 3.806429) < 0.001.$$

Am presupus că datele au fost normale şi că cele 2 grupe au avut dispersii egale. Dată fiind diferenţa mare dintre dispersiile de selecţie, această presupunere nu poate fi garantată.
De fapt, sunt alte teste de semnificaţie care testează dacă datele sunt aproximativ normale şi dacă cele 2 grupe au aceeaşi dispersie. În practică se pot aplica acestea întâi pentru a determina dacă un test t este potrivit în primul rând.
`http://en.wikipedia.org/wiki/Normality_test`
`http://en.wikipedia.org/wiki/F-test_of_equality_of_variances`