Matemàtiques (Veterinària)

Facultat de Veterinària

Universitat Autònoma de Barcelona

Curs 2005-06

Professorat de l'assignatura

	Classes	*Adreça electrònica*
David Marín	Teoria	davidmp@mat.uab.es
Noèlia Viles	Teoria	nviles@mat.uab.es
Anna Espinal	Problemes	aespinal@mat.uab.es
Carme Florit	Problemes	cflorit@mat.uab.es
Berta Barquero	Pràctiques	barquero@mat.uab.es
Omar El Idrissi	Pràctiques	idrissi@mat.uab.es
Antoni Ferragut	Pràctiques	ferragut@mat.uab.es
Javier Sánchez	Pràctiques	jsanchez@mat.uab.es
Isabel Serra	Pràctiques	iserra@mat.uab.es

Objectius de l'assignatura

L'assignatura de Matemàtiques és de tipus instrumental: el seu objectiu és introduir uns quants continguts dels camps del Càlcul de Probabilitats i de l'Estadística que s'utilitzen com a eines a altres assignatures de la llicenciatura, posant l'èmfasi en l'ús correcte d'aquestes eines en cada situació, així com en el fet d'extreure les conclusions adients.

Material de classe

Tot el material de classe es podrà trobar a l'intranet docent (Campus Virtual).

També trobareu material addicional a copisteria.

Temari

- **Tema 0: Introducció** Objecte d'estudi, conceptes bàsics i camps de l'Estadística.
- **Tema 1: Estadística descriptiva**
  - 1.1 Característiques numériques d'un conjunt de dades.
  - 1.2 Funcions estadístiques en la calculadora científica.
- **Tema 2: Probabilitats**
  - 2.1 Models de probabilístics.
  - 2.2 Probabilitat condicionada.
  - 2.3 Variables aleatòries.
    - 2.3.1. Variables aleatòries discretes.
    - 2.3.2. Variables aleatòries contínues.
    - 2.3.3. Independència de variables aleatòries.
- **Tema 3: Inferència estadística.**
  - 3.1. Conceptes bàsics.
  - 3.2. Distribucions mostrals.
  - 3.3. Intervals de confiança.
    - 3.3.1. Intervals de confiança per als paràmetres d'una població Normal.
    - 3.3.2. Intervals de confiança per a mostres grans.

- **Tema 4: Tests d'hipòtesis.**
  - 4.1. Definicions i exemples.
  - 4.2. Tests sobre els paràmetres d'una població Normal.
  - 4.3. Tests de comparació de dues poblacions Normals.
    - 4.3.1. Mostres independents.
    - 4.3.2. Dades aparellades.
  - 4.4. Tests per a mostres grans.
    - 4.4.1. Una sola població.
    - 4.4.2. Mostres independents.
    - 4.4.3. Dades aparellades.
  - 4.5. Tests de comparació per a mostres petites.
    - 4.5.1. Mostres independents.
    - 4.5.2. Dades aparellades.
  - 4.6. Tests d'ajustament i d'independència.

Bibliografia

El curs no seguirà cap llibre de text concret, però els continguts són pròxims als del llibre:

- Delgado, R.: *Apuntes de probabilidad y estadística,* UAB (2002).
- Delgado, R.: *Iniciación a la probabilidad y la estadística,* Materials UAB 153 (2004).

La bibliografia següent és adequada per a consultes puntuals o ampliació de coneixements, i està seleccionada entre els llibres d'estadística que es poden trobar a la Biblioteca de la Facultat de Veterinària.

- Daniel, W.W.: *Bioestadística.*
- Milton, J.S.: *Estadística para biología y ciencias de la salud (2a ed.)*
- Sokal, R.R.; Rohlf, F.J.: *Introducción a la bioestadística.*
- Dunn, O.J.: *Basic statistics: A primer for the biomedical sciences.*

Estructura de l'assignatura

Hi ha tres tipus de classes: de teoria, de problemes i de pràctiques. L'assistència a les classes de teoria i de problemes no és obligatòria; no obstant, és responsabilitat de l'estudiant estar assabentat dels continguts que s'hi han desenvolupat i dels anuncis o comentaris de tipus organitzatiu que s'hagin pogut fer a l'aula, i que sempre tindran prioritat sobre les informacions de taulers d'anuncis i pàgines d'internet.

A les classes de problemes es resoldran exercicis d'unes llistes que estaran disponibles amb antelació. Tot i que no és obligatori fer els problemes, es recomana fortament haver-los fet o pensat abans de la classe corresponent. Tingueu en compte que és fent problemes i a classe de problemes on solen sorgir els dubtes sobre la teoria.

L'assistència a les classes de pràctiques *és obligatòria.*

Avaluació de l'assignatura

- **Requisits:** Per tenir dret a presentar-se a l'examen és obligatòria l'assistència a les pràctiques. Es permet una sola absència no justificada. És obligatori també el lliurament a temps de tots els informes de pràctiques. Aquests informes són avaluats com a "Apte" o "No apte". Es pot tenir un "No apte" com a màxim per tenir dret a fer l'examen.

- **Examen:** La nota final de l'assignatura s'obté a partir de l'examen, si bé els informes lliurats es poden tenir en compte per a resoldre situacions dubtoses.

  L'examen consisteix en uns quants problemes de l'estil dels fets a classe de problemes o als exemples de les classes de teoria. L'examen conté també algunes qüestions relatives a les pràctiques (20% del valor total del examen).

  Per superar l'examen cal: Obtenir almenys el 50% de la puntuació màxima total, i obtenir almens un 50% de la puntuació màxima en la part relativa a pràctiques.

PRÀCTIQUES VETERINÀRIA CURS 2005-06.

SETMANA	PRÀCTICA	ENTREGA	CORRECCIÓ
1: 26/9-30/9			
2: 3/10-7/10			
3: 10/10-14/10	1 (8g)		
4: 17/10-21/10	1 (1g)		
5: 24/10-28/10	2(9g)	p1	
6: 31/10-4/11		p2	correcció p1
7: 7/11-11/11			correcció p2
8: 14/11-18/11	3(9g)		
9: 21/11-25/11		p3	
10: 28/11-2/12	4(9g)		correcció p3
11: 5/12-9/12	No hi ha classe		
12: 12/12-16/12	5(9g)	p4	
13: 19/12-21/12		p5	correcció p4
14: 9/1-13/1	6(9g)		correcció p5
15: 16/1-20/1	Fi de classes	p6	
16: 23/1-27/1			correcció p6 llista presentables
17: 30/1-3/2			
18: 6/2-10/2	Examen 10 febrer		

Tema 0: Introducció a l'estadística

Estadística: És la ciència que té per objecte la **recollida**, **resum** i **organització** de dades, i **treure conclusions** i **prendre decisions** a partir d'elles.

Població: És un grup d'**objectes** (en sentit ampli) que es vol estudiar estadísticament.

Mostra: És una **part** de la població, de la qual es recullen les dades.

Així doncs l'estadística comprèn:

- Mostreig: Recollida de dades.
- Estadística Descriptiva: Resum i organització de dades.
- Estadística Inferencial: Treure conclusions i prendre decisions.

Probabilitat: És la branca de les matemàtiques que modelitza l'atzar. Ens permetrà donar un sentit inequívoc a la part d'estadística inferencial.

Variable: És un símbol que pot prendre **un valor** dins d'un conjunt de **valors possibles**. Exemples:

- Resultat de llençar una moneda: $X \in \{cara, creu\}$.
- El sexe d'una persona escollida a l'atzar: $S \in \{dona, home\}$
- L'edat d'una persona (en anys sencers) $E \in \{0, 1, 2, \ldots, 150\}$ o bé $E \in \mathbb{N} = \{0, 1, 2, 3, \ldots\}$.
- L'edat **exacta** d'una persona: $Ex \in [0, 150]$ o bé $Ex \in \mathbb{R}$.

D'acord al conjunt de valors possibles, una variable s'en diu:

- Categòrica si pren valors **no numèrics**. Per exemple, X, S.
- Discreta si pren valors numèrics que es poden **contar**. Per exemple, E.
- Contínua si pren valors numèrics en un **continu**. Per exemple, Ex.

Tema 2: Probabilitats

2.1 Models probabilístics.

2.2 Probabilitat condicionada.

2.3 Variables aleatòries.

2.3.1. Variables aleatòries discretes.

2.3.2. Variables aleatòries contínues.

2.3.3. Indepèndencia de variables aleatòries.

2.1 Models probabilístics

Experiment aleatori: Fenomen físic en el qual intervé l'atzar: al repetir l'observació del fenonem s'obtenen resultats diferents. El resultat no es pot predir amb exactitud.

Exemples:

- Llançar una modena. Característica d'interès: Quina cara surt.
- Llançar un dau. Característica d'interès: Quina cara surt.
- Escollir a l'atzar un individu d'una població. Característica d'interès: Pes.
- No es un fenomen aleatori: Deixar un got de vidre a l'aire a veure si cau o no. Sabem amb certesa que caurà. Això és un fenomen **determinista**. En canvi si que és un experiment aleatori observar si el got es trenca o no.

Esdeveniment: Quelcom que es pot observat si ha passat o no al realitzar un experiment aleatori.

Com ara, al exemple del dau, podem observar si "ha sortit un nombre parell", "ha sortit un nombre més gran que 4" o bé "ha sortir un 1".

Esdeveniment elemental: Cada resultat possible de l'experiment. A l'exemple del dau, "sortir un 1" és un esdeveniment elemental.

Espai mostral Ω: És el **conjunt** de resultats possibles d'un experiment aleatori. A l'exemple del dau, $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Aleshores, un esdeveniment és qualsevol **subconjunt** de Ω. Per exemple, "Sortir un nombre parell" és el subconjunt $\{2, 4, 6\}$, "Sortir un nombre més gran que 4" és $\{5, 6\}$ i "Sortir un 1" és $\{1\}$.

Un cop identificats els esdeveniments amb conjunts, podem utilitzar el llenguatge i la notació dels conjunts per parlar d'esdeveniments.

En particular, si A i B són esdeveniments aleshores:

- La unió $A \cup B$ representa l'esdeveniment que succeeix quan passa A o bé passa B.
- La intersecció $A \cap B$ representa l'esdeveniment que succeeix quan passa A i a la vegada passa B.
- El complementari A^c representa l'esdeveniment que succeeix quan no passa A.
- Diferència $B \setminus A = B \cap A^c$ representa l'esdeveniment que succeeix quan passa B però no A.

Exemple del dau: $\Omega = \{1, 2, 3, 4, 5, 6\}$. Considerem els esdeveniments A:"Sortir senar" i B:"Sortir més petit que 4". Aleshores,

$$\begin{aligned} A &= \{1, 3, 5\}, \\ B &= \{1, 2, 3\}, \\ A \cup B &= \{1, 2, 3, 5\}, \\ A \cap B &= \{1, 3\}, \\ A^c &= \{2, 4, 6\}, \\ B \setminus A &= \{2\}, \\ B^c &= \{4, 5, 6\}, \\ A \setminus B &= \{5\}, \\ A \cap A^c &= \emptyset. \end{aligned}$$

El concepte de probabilitat s'introdueix per **quantificar** l'atzar.

Hi ha un esdeveniment que passarà amb tota seguretat: l'esdeveniment total Ω. A aquest esdeveniment li posem arbitrariament **probabilitat** 1.

Ara aquest 1 se l'han de repartir tots els resultats possibles. En l'exemple del dau $\Omega = \{1, 2, 3, 4, 5, 6\}$, i, si el dau és "equilibrat", tots els resultats tenen la mateixa probabilitat de sortir. Per tant, en aquest cas, posem diem que cada esdeveniment elemental $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$ i $\{6\}$ tenen probabilitat $\frac{1}{6}$.

L'objectiu és determinar la probabilitat de **tots** els esdeveniments. A l'exemple anterior, n'hi ha $2^6 = 64$ esdeveniments en total. Per exemple, la probabilitat de l'esdeveniment "Sortir senar" és la probabilitat de que surti un 1, un 3 o un 5, per tant, sumen

$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

Però com podem "saber" que les probabilitats de cadascú dels resultats són les mateixes?

Si realitzem un experiment aleatori un nombre $n \geq 1$ de vegades, aleshores podem contar el nombre $k \in \{0, \ldots, n\}$ de vegades s'ha produit un cert esdeveniment A. Aleshores definim:

La freqüència absoluta de A com k.

La freqüència relativa de A com $\frac{k}{n}$, que és un nombre entre 0 i 1.

Hi ha un resultat de la Teoria de la Probabilitat que diu que les freqüències relatives d'un esdeveniment A tendeixen a la probabilitat d'aquest esdeveniment quan el nombre n es fa gran.

Això ens permet **estimar** la probabilitat dels esdeveniments elementals utilitzant aquesta regla. En altres paraules, estaríem fent inferència estadística.

Definició de probabilitat

Sigui Ω el conjunt de resultats possibles d'un experiment aleatori i sigui $\mathcal{F}$ la col.lecció de **tots els esdeveniments**. Una **probabilitat** P és una aplicació

$$P : \mathcal{F} \to [0,1]$$

que verifica les dues propietats següents:

(1) $P(\Omega) = 1$,

(2) si $A, B \in \mathcal{F}$ i $A \cap B = \emptyset$ aleshores

$$P(A \cup B) = P(A) + P(B)$$

(i anàlogament si en lloc de dos són infinits conjunts disjunts).

La terna $(\Omega, \mathcal{F}, P)$ s'anomena model probabilístic o espai de probabilitat.

Observem que gràcies a la propietat (2), P ve "determinada" pel seu valor sobre els esdeveniments elementals com passava al exemple del dau.

Definició de probabilitat

(1) $P(\Omega) = 1$.

(2) Si $A \cap B = \emptyset$ aleshores $P(A \cup B) = P(A) + P(B)$.

Propietats de les probabilitats

Siguin $A, B \in \mathcal{F}$ esdeveniments.

(3) Si $A \subset B$ aleshores $P(B \setminus A) = P(B) - P(A)$.

(4) Si $A \subset B$ aleshores $P(A) \leq P(B)$.

(5) $P(A^c) = 1 - P(A)$.

(6) $P(A \cup B) + P(A \cap B) = P(A) + P(B)$.

Els problemes de probabilitat consisteixen en calcular la probabilitat d'alguns esdeveniments a partir de la probabilitat d'altres esdeveniments.

Exemple: Una cabina telefònica no està en molt bones condicions, i pot passar que:

R_1: Accepta monedes i dóna línia.

R_2: Accepta monedes i no dóna línia.

R_3: No accepta monedes i no dóna línia.

R_4: No accepta monedes i dóna línia.

S'observa que de cada 10 intents:

- en 8 accepta monedes,
- en 8 dóna línia,
- en 9 "fa alguna cosa".

Pregunta: Quina és la probabilitat que en un cert intent el telèfon funcioni correctament (situació R_1)?

Dades del model probabilístic $(\Omega, \mathcal{F}, P)$:

- $\Omega = \{R_1, R_2, R_3, R_4\}$,
- $P(\{R_1, R_4\}) = 0.8$,
- $P(\{R_1, R_2\}) = 0.8$,
- $P(\{R_1, R_2, R_4\}) = 0.9$.

La incògnita és $P(\{R_1\})$ que podem calcular fent

$$\begin{aligned} & P(\{R_1\}) = P\big(\{R_1, R_2\} \cap \{R_1, R_4\}\big) = \\ \overset{(6)}{=} \quad & P(\{R_1, R_2\}) + P(\{R_1, R_4\}) - P\big(\{R_1, R_2, R_4\}\big) \\ = \quad & 0.8 + 0.8 - 0.9 = 0.7. \end{aligned}$$

També podem calcular la probabilitat dels altres esdeveniments elementals:

$$\begin{aligned} P(\{R_2\}) & \overset{(3)}{=} P(\{R_1, R_2\}) - P(\{R_1\}) = 0.8 - 0.7 = 0.1 \\ P(\{R_4\}) & \overset{(3)}{=} P(\{R_1, R_4\}) - P(\{R_1\}) = 0.8 - 0.7 = 0.1 \\ P(\{R_3\}) & \overset{(5)}{=} 1 - P(\{R_1, R_2, R_4\}) = 1 - 0.9 = 0.1 \end{aligned}$$

A partir d'aquí podem calcular la probabilitat de qualsevol esdeveniment i per tant el model queda determinat al donar les probabilitats dels esdeveniments que venien en les dades del problema.

2.2 Probabilitat condicionada

Exemple introductori: En una capsa tenim cinc boles blanques, numerades del 1 al 5, i tres boles negres, numerades del 1 al 3.

Experiment aleatori: Treiem una bola a l'atzar.

Espai mostral: $\Omega = \{B_1, B_2, B_3, B_4, B_5, N_1, N_2, N_3\}$.

Probabilitat:
$P(\{B_1\}) = P(\{B_2\}) = \cdots = P(\{N_3\}) = \frac{1}{8}$
(a l'atzar vol dir que tots els esdeveniments elementals són equiprobables).

Pregunta: Quina és la probabilitat de l'esdeveniment A:"treure una bola amb un nombre parell"?

Resposta: $A = \{B_2, B_4, N_2\}$ i $P(A) = P(\{B_2\}) + P(\{B_4\}) + P(\{N_2\}) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$.

Pregunta: Quina és la probabilitat de treure un nombre parell **sabent** que la bola és blanca?

La informació addicional ens fa canviar el model. Ara els resultats possibles són només $B = \{B_1, B_2, B_3, B_4, B_5\}$ i per tant la probabilitat de parell és la de $\{B_2, B_4\}$, és a dir $\frac{2}{5}$.

Observació important: $\frac{2}{5} = \frac{2/8}{5/8} = \frac{P(A \cap B)}{P(B)}$.

Definició de probabilitat condicionada

Si A i B són dos esdeveniments es defineix la **probabilitat de A condicionada a B** com

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Fórmula de Bayes

Si A i B són esdeveniments aleshores

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

Més generalment, si $A, B_1, \ldots, B_n$ són esdeveniments **disjunts** tals que $B_1 \cup \cdots \cup B_n = \Omega$ aleshores

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n).$$

Problema 1: Tenim dues capses, la primera amb 3 boles negres i dues boles blanques i la segona amb dues boles de cada color.

Experiment aleatori: Treiem una bola a l'atzar de la primera capsa i la introduim en la segona. A continuació, treiem una bola a l'atzar de la segona capsa.

Pregunta: Quina és la probabilitat de que la segona bola sigui negra?

Model probabilístic: $\Omega = \{B, N\}$ no és adequat per calcular. En canvi
$\Omega = \{(B, B), (B, N), (N, B), (N, N)\}$ sí que ho és.
Introduim els següents esdeveniments no elementals:
"primera bola blanca"$= 1_B = \{(B, B), (B, N)\}$,
"primera bola negra"$= 1_N = \{(N, B), (N, N)\}$,
"segona bola blanca"$= 2_B = \{(B, B), (N, B)\}$,
"segona bola negra"$= 2_N = \{(B, N), (N, N)\}$.

Dades: Coneixem les probabilitats $P(1_B) = \frac{2}{5}$, $P(1_N) = \frac{3}{5}$, i les probabilitats condicionades següents: $P(2_N|1_B) = \frac{2}{5}$ i $P(2_N|1_N) = \frac{3}{5}$.

Resolució: Utilitzant la fórmula de Bayes tenim que

$$\begin{aligned} P(2_N) &= P(2_N \cap 1_N) + P(2_N \cap 1_B) \\ &= P(2_N|1_N)P(1_N) + P(2_N|1_B)P(1_B) \\ &= \frac{3}{5} \cdot \frac{3}{5} + \frac{2}{5} \cdot \frac{2}{5} = \frac{13}{25}. \end{aligned}$$

Problema 2 (examen de setembre 2005):

L'aparició de banyes a l'ovella és controlada per dos al.lels H i h. L'al.lel H per a la formació de banyes és dominant las mascles però recessiu a les femelles. L'al.lel h per l'absència de banyes és dominant a les femelles però és recessiu als mascles. Això fa que un mascle heterozigòtic Hh tingui banyes i en canvi una femella també heterozigòtica Hh no en tingui. Suposem precisament que aquests dos animals s'aparellen i que cada cria d'aquest encreuament té la mateixa probabilitat de ser mascle o femella. Calculeu:

1. La probabilitat que una cria sigui mascle i tingui banyes.
2. La probabilitat que una cria tingui banyes.

Resolució: Siguin B, M i F respectivament els esdeveniments la cria té banyes, és mascle i és femella. Les dades del problema són:

$$P(M) = P(F) = \frac{1}{2}, \quad P(B|M) = \frac{3}{4} \quad \text{i} \quad P(B|F) = \frac{1}{4}.$$

(a) $P(M \cap B) = P(B|M)P(M) = \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8}$.
(b) De la mateixa manera, tenim que
$P(F \cap B) = P(B|F)P(F) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$ i per tant,
$P(B) = P(B \cap M) + P(B \cap F) = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$.

Exercici:

En una població d'una certa espècie animal amb mascles i femelles a parts iguals, el 5% dels mascles i el 0.25% de les femelles tenen una certa característica genètica (són albins, per exemple). Escollim un individu a l'atzar.

1. Quina és la probabilitat que tingui aquesta característica?
2. I si hagués un 80% de femelles i un 20% de mascles?
3. En la mateixa situació que l'apartat (a), suposem ara que sabem que l'individu escollit a l'atzar té la característica anterior. Quina és la probabilitat que sigui femella?

Solució:

1. $P(C) = P(C|M)P(M) + P(C|F)P(F) = \frac{21}{800} = 0.02625$
2. $P(C) = P(C|M)P(M) + P(C|F)P(F) = \frac{3}{250} = 0.012$
3. $P(F|C) = \frac{P(F \cap C)}{P(C)} = \frac{P(C|F)P(F)}{P(C)} = \frac{1}{21} = 0.04762$

Exercici:

Respondre a les preguntes (a) i (c) del problema anterior suposant que la proporció de femelles és de 30%:

(a) $P(C) = P(C|M)P(M) + P(C|F)P(F) = \frac{5}{100} \cdot \frac{70}{100} + \frac{0.25}{100} \cdot \frac{30}{100} = \frac{715}{20000} = \frac{143}{4000} = 0.03575$

(c) $P(F|C) = \frac{P(F \cap C)}{P(C)} = \frac{P(C|F)P(F)}{P(C)} = \frac{\frac{0.25}{100} \cdot \frac{30}{100}}{\frac{143}{4000}} = \frac{3}{143} = 0.020979$

El problema de Monty Hall

Estàs en un concurs de televisió. En aquest concurs l'idea és guanyar com a premi un cotxe. El presentador del programa t'ensenya tres portes. Diu que hi ha un cotxe darrera d'una de les portes i que darrera les altres dues hi han carbasses. Et demana d'escollir una porta. Tu esculls una porta, que encara no s'obre. Aleshores, el presentador obre una de les portes que tu no has escollit i et mostra una carbassa (ja que ell sap què hi ha darrera cada porta). Aleshores et diu que tens una última oportinutat de canviar d'opinió abans que les portes s'obrin i aconsegueixis un cotxe o una carbassa. Et pregunta si vols canviar d'idea i escollir la porta sense obrir. ¿Què has de fer?

Solució al problema de Monty Hall

La probabilitat de que el cotxe estigui a la porta que hem triat al començament és $\frac{1}{3}$, en canvi la probabilitat de que estigui a l'altra porta és $\frac{2}{3}$. Així doncs, el més adequat és canviar l'elecció.

Raonant amb probabilitats condicionades:

Siguin A, B les dues portes que no hem triat i siguin L_A i L_B els esdeveniments: "el presentador obre la porta A/B". Siguin C_A i C_B els esdeveniments: "el cotxe es troba darrera la porta A/B". Aleshores, la probabilitat de que el cotxe no estigui darrera la porta que hem triat al començament és

$$P\Big((L_A \cap C_B) \cup (L_B \cap C_A)\Big)$$

$$= P(L_A \cap C_B) + P(L_B \cap C_A)$$

$$= P(L_A|C_B)P(C_B) + P(L_B|C_A)P(C_A)$$

$$= 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3}.$$

Un parell d'exemples més sobre el tema del condicionament:

Dau: Llancem un dau i considerem els esdeveniments
A: "sortir un nombre més gran que 2",
B: "sortir un nombre parell".
Calculeu $P(A)$, $P(B)$, $P(A \cap B)$, $P(A|B)$ i $P(B|A)$.

Solució: $P(A) = \frac{2}{3}$, $P(B) = \frac{1}{2}$, $P(A \cap B) = \frac{1}{3}$, $P(A|B) = \frac{2}{3}$ i $P(B|A) = \frac{1}{2}$.

Cartes: Considereu un joc de cartes amb 4 pals A,B,C,D i 12 cartes per pal.

1. Quina és la probabilitat de treure un 5?
2. Quina és la probabilitat de treure una carta del pal A?
3. Quina és la probabilitat de treure el 5 del pal A?
4. Quina és la probabilitat de que una carta sigui del pal A sabent que és un 5?
5. Quina és la probabilitat de que una carta sigui un 5 saben que és del pal A?

Definició de independència

Dos esdeveniments A i B són **independents** si el coneixement que ha succeït un d'ells no modifica la probabilitat que l'altre succeeixi, és a dir,

$$P(A|B) = P(A), \quad \text{o equivalentment,} \quad P(B|A) = P(B).$$

En a quest cas, $P(A) = P(A|B) = \frac{P(A\cap B)}{P(B)}$ i per tant,

$$(\star) \qquad P(A \cap B) = P(A)P(B)$$

La propietat $(\star)$ és molt més simètrica i equival al fet que A i B són independents; a més, generalitza trivialment al cas de més de dos esdeveniments: $A_1, \ldots, A_n$ són independents si

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n).$$

Propietat de la independència

Si A i B són esdeveniments independents aleshores també ho són A i B^c; A^c i B; A^c i B^c.

Problema-exemple: Dos caçadors, Quim i Ramon, disparen a un mateix objectiu. Quim sol encertar un 70% de les vegades i Ramon un 80%. Quina és la probabilitat que l'objectiu sigui tocat?

Solució: Considerem els esdeveniments Q: "Quim encerta" i R: "Ramon encerta". Aleshores sabem que $P(Q) = 0.7$ i $P(R) = 0.8$ i ens demanen $P(Q \cup R)$:

$$\begin{aligned} P(Q \cup R) &\overset{(6)}{=} P(Q) + P(R) - P(Q \cap R) \\ &= 0.7 + 0.8 - (0.7)(0.8) = 0.94, \end{aligned}$$

si utilitzem el fet de que el resultat d'un caçador no influeix en l'altre, és a dir, que Q i R són esdeveniments independents i per tant $P(Q \cap R) = P(Q)P(R)$.

Una altra manera: Calcular primer la probabilitat de que tots dos fallin imposant l'independència:

$$P(Q^c \cap R^c) = P(Q^c)P(R^c) = (0.3)(0.2) = 0.06$$

I després, utilitzant que $(A \cup B)^c = A^c \cap B^c$,

$$P(Q \cup R) = 1 - P(Q^c \cap R^c) = 1 - 0.06 = 0.94.$$

2.3 Variables aleatòries

Exemple introductori

Experiment aleatori: Llancem dos daus equilibrats i ens interessem a la suma X dels punts dels dos daus (per exemple si esteu jugant al monopoly).

Espai mostral de la suma és $\Omega_X = \{2, 3, \ldots, 12\}$, però posar-hi aquí la probabilitat P_X adequada no és fàcil, ja que no tots els resultats són equiprobables.

Espai mostral dels dos daus: Considerem el conjunt de resultats dels dos daus
$\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$ juntament a l'aplicació

$$\begin{array}{rcl} X : \Omega & \longrightarrow & \{2, 3, \ldots, 12\} \subset \mathbb{R} \\ (i,j) & \longmapsto & i + j \end{array}$$

Probabilitat en Ω: Com tots els 36 resultats diferents de Ω tenen la mateixa probabilitat, posem $P(\{(i,j)\}) = \frac{1}{36}$ per tot $1 \leq i, j \leq 6$.

Probabilitat en Ω_X:

Calculem la probabilitat de que la suma X dels resultats dels dos daus sigui 7:

$$P_X(\{7\}) = P\big(\{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}\big) = \frac{6}{36}.$$

Probabilitat en Ω_X:

Construïm la probabilitat en Ω_X

$$P_X : \Omega_X = \{2, \ldots, 12\} \to [0, 1]$$

a partir de la probabilitat en Ω:

$$P : \Omega = \{(i, j),\ 1 \leq i, j \leq 6\} \to [0, 1]$$

de la manera següent:

$$\begin{aligned}
P_X(\{2\}) &= P(\{(1,1)\}) = \frac{1}{36}, \\
P_X(\{3\}) &= P(\{(1,2),(2,1)\}) = \frac{2}{36}, \\
P_X(\{4\}) &= P(\{(1,3),(2,2),(3,1)\}) = \frac{3}{36}, \\
&\vdots \\
P_X(\{7\}) &= P(\{(6,1),(5,2),\ldots,(1,6)\}) = \frac{6}{36}, \\
P_X(\{8\}) &= P(\{(6,2),(5,3),\ldots,(2,6)\}) = \frac{5}{36}, \\
&\vdots \\
P_X(\{12\}) &= P(\{(6,6)\}) = \frac{1}{36}.
\end{aligned}$$

Definició de variable aleatòria

Si $(\Omega, \mathcal{F}, P)$ és un model probabilístic, una **variable aleatòria** és una aplicació

$$X : \Omega \to \mathbb{R},$$

és a dir, una mateixa característica numèrica de cada individu de la població Ω.

Exemple: Si Ω és una població de persones, podem prendre com a variable aleatòria $X : \Omega \to \mathbb{R}$ l'alçada de cada individu.

Definició de llei d'una variable aleatòria

Sigui $(\Omega, \mathcal{F}, P)$ un model probabilístic i $X : \Omega \to \mathbb{R}$ una variable aleatòria. La **llei** o **distribució** de X és la probabilitat P_X en $\mathbb{R}$ definida per

$$P_X(A) := \underbrace{P\big(\{\omega \in \Omega : X(\omega) \in A\}\big)}_{\text{per abreviar } P(X \in A)}, \quad \text{si } A \subset \mathbb{R}.$$

Exemple dels dos daus (continuació): Si $A = \{3, 4\}$,

$$\{X \in A\} = \{(1,2), (2,1), (1,3), (2,2), (3,1)\}.$$

Per tant, $P_X(A) = \frac{5}{36}$. També podem fer:

$$P_X(A) = P_X(\{3\}) + P_X(\{4\}) = \frac{2}{36} + \frac{3}{36} = \frac{5}{36}.$$

Propietats de les variables aleatòries

Les variables aleatòries d'un mateix espai probabilístic es poden **sumar** i **multiplicar**: Si $X, Y : \Omega \to \mathbb{R}$ aleshores $X + Y, X \cdot Y : \Omega \to \mathbb{R}$ vénen definides per

$$\begin{aligned}(X+Y)(\omega) &= X(\omega)+Y(\omega),\\ (X\cdot Y)(\omega) &= X(\omega)\cdot Y(\omega),\end{aligned}$$

per tot $\omega \in \Omega$.

Exemple dels dos daus (continuació): Llancem dos daus i considerem l'espai probabilístic[a] $(\Omega, \mathcal{F}, P)$, on $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$ i $P(\{i,j\}) = \frac{1}{36}$. Considerem les dues variables aleatòries X_1 i X_2 que donen el valor de la primera i segona tirada respectivament. Aleshores la variable aleatòria suma considerada anteriorment és $X = X_1 + X_2$.

[a] En aquest cas n'hi ha $2^{36} = 68\,719\,467\,436$ esdeveniments, gairebé 69.000 milions!

2.3.1. Variables aleatòries discretes

Definició de variable aleatòria discreta: Una variable aleatòria $X : \Omega \to \mathbb{R}$ és **discreta** si la seva imatge és un conjunt **numerable** de valors de $\mathbb{R}$.

Exemples clàssics de variables aleatòries discretes:

- Uniforme: Modela la situació en que tots els valors de la variable tenen les mateixes oportunitats d'obtenir-se. Si $X : \Omega \to \mathbb{R}$ només pren n valors diferents $a_1, \ldots, a_n \in \mathbb{R}$, la probabilitat de cada esdeveniment elemental és $P_X(\{a_i\}) = \frac{1}{n}$. Notació: $X \sim \mathrm{Unif}(a_1, \ldots, a_n)$.

- Binomial: Considerem un experiment aleatori i un cert esdeveniment A que té una probabilitat p. Repetim l'experiment n vegades i considerem $X =$quantitat de vegades que s'ha produït A en les n repeticions. Notació: $X \sim \mathrm{Bin}(n, p)$. La llei de X ve donada per

$$P_X(\{k\}) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

si $k \in \{0, \ldots, n\}$.

Calculem la llei de $X \sim \mathrm{Bin}(n, p)$ depenent de n:

Si $n = 1$, aleshores $P_X(\{1\}) = P(X = 1) = P(A) = p$ i $P_X(\{0\}) = P(A^c) = 1 - p$.

Si $n = 2$, aleshores $P_X(\{2\}) = P(\{(A, A)\}) = p^2$, $P_X(\{1\}) = P(\{(A, A^c), (A^c, A)\}) = 2p(1 - p)$ i $P_X(\{0\}) = P(\{(A^c, A^c)\}) = (1 - p)^2$.

Per n general, $P_X(\{k\})$ és la probabilitat de que A es produeix k vegades i A^c es produeix $n - k$ vegades: $p^k(1 - p)^{n-k}$ Però hem de tenir en compte les maneres de repartir les k vegades de A en les n repeticions:

$$\frac{n!}{k!(n-k)!} =: \binom{n}{k}.$$

Calculem la llei de $X \sim \mathrm{Bin}(n,p)$ depenent de n:

Si $n = 1$, aleshores $P_X(\{1\}) = P(X = 1) = P(A) = p$ i $P_X(\{0\}) = P(A^c) = 1 - p$.

Si $n = 2$, aleshores $P_X(\{2\}) = P(\{(A, A)\}) = p^2$, $P_X(\{1\}) = P(\{(A, A^c), (A^c, A)\}) = 2p(1 - p)$ i $P_X(\{0\}) = P(\{(A^c, A^c)\}) = (1 - p)^2$.

Per n general, $P_X(\{k\})$ és la probabilitat de que A es produeix k vegades i A^c es produeix $n - k$ vegades: $p^k(1 - p)^{n-k}$ Però hem de tenir en compte les maneres de repartir les k vegades de A en les n repeticions:

$$\frac{n!}{k!(n-k)!} =: \binom{n}{k}.$$

Exemple de variable binomial:

Repetim 10 vegades l'experiment d'agafar a l'atzar una vaca d'una població de 1000 vaques i considerem l'esdeveniment A:"la vaca té BSE[a]". Suposem que el 30% de les vaques tenen BSE. Sigui X la quantitat de vaques amb BSE que hem obtingut en les 10 repeticions. Llavors $X \sim \mathrm{Bin}(10, 0.3)$ i per exemple,

$$P(X = 4) = \binom{10}{4}(0.3)^4(0.7)^6 \simeq 0.2001.$$

[a] Bovine Spongiform Encephalopaty

Observacions sobre la binomial:

Els experiments han de ser idèntics i "independents" entre sí.

En particular, a l'exemple anterior, una mateixa vaca pot sortir dues o més vegades en la seqüència de 10 experiments.

Continuació de l'exemple anterior:

Si agafem les 10 vaques a la vegada el càlculs són més complicats. Fem-ho amb $n = 2$:

Si $X \sim \text{Bin}(2, 0.3)$, $P(X = 2) = (0.3)^2$. Si no,

$$P(M_1 \cap M_2) = P(M_2|M_1)P(M_1) = \frac{299}{999} \cdot (0.3) \neq (0.3)^2.$$

No dóna igual, però sí molt semblant. Podríem dir que la situació binomial "aproxima" l'altra. Això és perquè la població era gran (1000 vaques).

A la pràctica el paràmetre p és desconegut, i el que voldrem és precisament tenir una idea de quan val a partir de com són les vaques examinades: Inferència estadística.

Poisson

En un determinat espai lineal (temps, longitud) esperem que un determinat esdeveniment succeeixi $\lambda > 0$ vegades (per exemple, el nombre de trucades que es reben en un dia). Suposant que el resultat és equiprobable en el temps, si dividim l'espai inicial en n segments iguals, el nombre esperat de vegades en cada segment és $p_n := \frac{\lambda}{n}$. Així podem considerar la suma X_n de les vegades que succeeix l'esdeveniment donat en els n segments com a una variable binomial $X_n \sim \mathrm{Bin}(n, p_n)$. Si fem el límit quan el nombre n de segments tendeix a infinit obtenim

$$X = \lim_{n\to\infty} X_n \sim \mathrm{Poiss}(\lambda).$$

Podem deduir $P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ fent:

$$\begin{aligned}
\lim_{n\to\infty} P(X_n = k) &= \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k-1)}{k!} \frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k-1)}{n^k} \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{\frac{1}{-\lambda/n}\frac{-\lambda(n-k)}{n}} \\
&= \frac{\lambda^k}{k!}e^{-\lambda}.
\end{aligned}$$

Esperança per a una variable aleatòria discreta X:

$$E[X] := \sum_k k P_X(\{k\}),$$

on k varia en el conjunt de valors possibles de X.

L'esperança mesura el "valor mitjà" que pren una variable aleatòria.

Exemples:

Si $X \sim \mathrm{Unif}(1, \ldots, n)$ aleshores

$$E[X] = 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \cdots + n \cdot \frac{1}{n} = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}.$$

En particular, pel cas d'un dau equilibrat ($n = 6$), l'esperança és 3.5 no 3.

Càlcul de l'esperança d'una variable binomial:

Si $X \sim \text{Bin}(n,p)$ aleshores $E[X]$ és igual a

$$\begin{aligned}
\sum_{k=0}^{n} kP_X(\{k\}) &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\
&= \sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\
&= np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k} \\
&= np\sum_{k=1}^{n} \binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} \\
&= np\sum_{\ell=0}^{n-1} \binom{n-1}{\ell}p^{\ell}(1-p)^{n-1-\ell} \\
&= np\cdot 1.
\end{aligned}$$

Per tant, si $X \sim \text{Bin}(n,p)$, aleshores $E[X] = np$.

Càlcul de l'esperança d'una variable binomial:

Si $X \sim \mathrm{Bin}(n,p)$ aleshores $E[X]$ és igual a

$$\begin{aligned}
\sum_{k=0}^{n} kP_X(\{k\}) &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\
&= \sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\
&= np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k} \\
&= np\sum_{k=1}^{n} \binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} \\
&= np\sum_{\ell=0}^{n-1} \binom{n-1}{\ell}p^{\ell}(1-p)^{n-1-\ell} \\
&= np\cdot 1.
\end{aligned}$$

Per tant, si $X \sim \mathrm{Bin}(n,p)$, aleshores $E[X] = np$.

D'una manera similar, si $X \sim \mathrm{Poiss}(\lambda)$, $E[X] = \lambda$:

$$\sum_{k=0}^{\infty} kP_X(\{k\}) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} = \lambda.$$

Propietats de l'esperança:

Ja havíem observat que tenia sentit sumar i multiplicar dues variables aleatòries $X, Y : \Omega \to \mathbb{R}$.

És fàcil veure que:

(1) $E[X+Y] = E[X] + E[Y]$,

(2) $E[c \cdot X] = c \cdot E[X]$ on c és una constant.

També té sentit compondre una variable aleatòria $X : \Omega \to \mathbb{R}$ amb una funció $g : \mathbb{R} \to \mathbb{R}$: Com calculem $E[g(X)]$? Tenim dues possibilitats: o bé calculem primer la llei P_Y de $Y = g(X)$, o bé fem

$$E[g(X)] = \sum_k g(k) P_X(\{k\}).$$

Per exemple, si $g(x) = x^2$, $E[X^2] = \sum_k k^2 P_X(\{k\})$.

Definició de variància d'una variable aleatòria X:

$$\begin{aligned} Var[X] &:= E[(X - E[X])^2] \\ &= E[X^2 - 2E[X] \cdot X + E[X]^2] \\ &= E[X^2] - 2E[X] \cdot E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2. \end{aligned}$$

La variància mesura la dispersió dels valors d'una variable aleatòria.

Propietats de la variància:

- $Var[X] \geq 0$,
- $Var[c \cdot X] = c^2 \cdot Var[X]$ si c és una constant.

Exemples:

- Si $X \sim \text{Unif}(4, 6)$, $E[X] = 5$ i $Var[X] = 1$.
- Si $X \sim \text{Unif}(3, 7)$, $E[X] = 5$ i $Var[X] = 4$.

Definició de variància d'una variable aleatòria X:

$$\begin{aligned} Var[X] &:= E[(X - E[X])^2] \\ &= E[X^2 - 2E[X] \cdot X + E[X]^2] \\ &= E[X^2] - 2E[X] \cdot E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2. \end{aligned}$$

La variància mesura la dispersió dels valors d'una variable aleatòria.

Propietats de la variància:

- $Var[X] \geq 0$,
- $Var[c \cdot X] = c^2 \cdot Var[X]$ si c és una constant.

Exemples:

- Si $X \sim \text{Unif}(4, 6)$, $E[X] = 5$ i $Var[X] = 1$.
- Si $X \sim \text{Unif}(3, 7)$, $E[X] = 5$ i $Var[X] = 4$.
- Si $X \sim \text{Bin}(n, p)$, $E[X] = np$ i
$$Var[X] = np(1 - p).$$
- Si $X \sim \text{Poiss}(\lambda)$, $E[X] = \lambda$ i $Var[X] = \lambda$.

2.3.2. Variables aleatòries contínues

Definició de variable aleatòria contínua: Una variable aleatòria $X : \Omega \to \mathbb{R}$ és **contínua** si

$$P(a \leq X \leq b) = P(X \in [a,b]) = P_X([a,b])$$

es calcula com $\int_a^b f$ per una certa funció $f \geq 0$, que s'anomena **funció de densitat** de X.

Diferències amb el cas discret:

- $P(X = x) = P_X(\{x\}) = 0$, és a dir, tots els esdeveniments elementals tenen probabilitat zero i per tant no determinen la llei.
- $\mathbb{R} = \bigcup_x \{x\}$ és una unió disjunta però en canvi
  $$1 = P_X(\mathbb{R}) \neq \sum_x P_X(\{x\}) = \sum_x 0,$$
  ja que hi ha una infinitat **no numerable** de conjunts.
- La funció de densitat determina la probabilitat dels intervals. I la resta de conjunts? Apliquem la regla
  $$(2)\ \ A \cap B = \emptyset \ \Rightarrow P_X(A \cup B) = P_X(A) + P_X(B)$$

amb una quantitat infinita numerable de subconjunts.

Es pot calcular així la probabilitat de qualsevol subconjunt de $\mathbb{R}$? No. Quan posem en $\mathbb{R}$ una probabilitat P_X, aquesta no està definida sobre tots els subconjunts de $\mathbb{R}$, només sobre els intervals i unions disjuntes d'aquests. És a dir, en el model probabilístic $(\mathbb{R}, \mathcal{F}, P_X)$, la col.lecció d'esdeveniments $\mathcal{F}$ **no** està formada per **tots** els subconjunts de $\mathbb{R}$.

- Això a la pràctica no és cap problema, ja que només necessitem probabilitats d'intervals.

Exemples de variables aleatòries contínues:

- Uniforme: $X \sim \mathrm{Unif}([a,b])$ si

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si} \quad x \in [a,b] \\ 0 & \text{si} \quad x \notin [a,b] \end{cases}$$

Tots els intervals de la mateixa longitud tenen la mateixa probabilitat. Correspon a l'experiment aleatori (ideal) d'escollir un nombre a l'atzar a l'interval $[a,b]$.

- Exponencial: $X \sim \mathrm{Exp}(\lambda)$ si

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si} \quad x \geq 0 \\ 0 & \text{si} \quad x < 0 \end{cases}$$

Molt utilitzada com a model de supervivència. Modelitza el temps de vida de components elèctrics, electrònics i d'alguns organismes vius.

- Normal: $X \sim \mathrm{N}(\mu, \sigma^2)$ si

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

S'ajusta bé a les mesures que són suma de moltes petites causes, com ara l'alçada o el pes d'un ésser viu.

Tipificació de les variables normals:

Si $X \sim N(\mu, \sigma^2)$ aleshores $\frac{X-\mu}{\sigma} \sim N(0,1)$.

Els valors de $N(0,1)$ estan tabulats.

Tipificació de les variables normals:

Si $X \sim N(\mu, \sigma^2)$ aleshores $\frac{X-\mu}{\sigma} \sim N(0,1)$.

Els valors de $N(0,1)$ estan tabulats.

Exemple: Si l'alçada X d'una població segueix una distribució $N(170, 25)$ aleshores $Z = \frac{X-170}{5} \sim N(0,1)$.

- $P(X \leq 180) = P(\frac{X-170}{5} \leq \frac{180-170}{5}) = P(Z \leq 2)$ $= 0.9772$.
- $P(160 \leq X \leq 170) = P(\frac{160-170}{5} \leq \frac{X-170}{5} \leq \frac{170-170}{5})$ $= P(-2 \leq Z \leq 0) = P(Z \leq 0) - P(Z \leq -2) =$ $0.5 - (1 - P(Z \leq 2)) = 0.5 - (1 - 0.97725) = 0.47725$.
- Quin és x tal que $P(X \leq x) = 0.90$?

  $$0.90 = P(\frac{X-170}{5} \leq \frac{x-170}{5}) = P(Z \leq z) \to z = 1.28,$$

  on $z = \frac{x-170}{5}$. Per tant, $x = 5z + 170 = 176.40$.
- Quin és x tal que $P(X \geq x) = 0.90$?

  $$0.90 = P(\frac{X-170}{5} \geq \frac{x-170}{5}) = P(Z \geq z) = P(Z \leq -z)$$

  $\to -z = 1.28 \to z = -1.28$, on $z = \frac{x-170}{5}$. Per tant, $x = 5z + 170 = 163.60$.

Esperança de variables aleatòries contínues:

Si X és una variable aleatòria contínua amb densitat f, definim la seva esperança com

$$E[X] := \int_{-\infty}^{\infty} x \cdot f(x).$$

(És com en el cas discret, canviant el sumatori per integrals.)

També,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x).$$

La variància es defineix igual que en el cas discret:

$$Var[X] := E[(X - E[X])^2].$$

Exemples:

- Si $X \sim \text{Unif}([a,b])$, $E[X] = \frac{a+b}{2}$ i $Var[X] = \frac{(a-b)^2}{12}$.
- Si $X \sim \text{Exp}(\lambda)$, $E[X] = \frac{1}{\lambda}$ i $Var[X] = \frac{1}{\lambda^2}$.
- Si $X \sim N(\mu, \sigma^2)$, $E[X] = \mu$ i $Var[X] = \sigma^2$.

2.3.3. Independència de variables aleatòries

Definició d'independència

Dues variables aleatòries X, Y són **independents** si per a tots A i B intervals de $\mathbb{R}$,

$$P(\{X \in A\} \cap \{Y \in B\}) = P(\{X \in A\}) \cdot P(\{Y \in B\}).$$

(En el cas discret n'hi ha prou amb considerar A i B esdeveniments elementals.)

Propietats de la independència

Si X i Y són independents aleshores:

- $E[X \cdot Y] = E[X] \cdot E[Y]$,
- $Var[X + Y] = Var[X] + Var[Y]$.

Exemple: Recordem que si $X \sim \text{Bin}(n, p)$ aleshores X representa la quantitat de vegades que s'ha produït un esdeveniment de probabilitat p en n repeticions. Podem pensar que X és la suma de n variables aleatòries **independents** que prenen el valors 1 i 0 amb probabilitats p i $1 - p$ respectivament. Per tant,

$X = Y_1 + \cdots + Y_n$ amb $Y_j \sim \text{Bin}(1, p)$ **independents**.

Càlcul de l'esperança i la variància d'una binomial:

Si $X \sim \text{Bin}(n, p)$ aleshores,

$X = Y_1 + \cdots + Y_n$ amb $Y_j \sim \text{Bin}(1, p)$ **independents**.

Calculem $E[Y_j] = 1 \cdot p + 0 \cdot (1 - p) = p$ i

$$\begin{aligned} Var[Y_j] &= E[Y_j^2] - E[Y_j]^2 \\ &= 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p - p^2 \\ &= p(1 - p). \end{aligned}$$

Ara bé,

$$E[X] = E[Y_1] + \cdots + E[Y_n] = np.$$

D'altra banda, degut a la independència de les Y_j,

$$Var[X] = Var[Y_1] + \cdots + Var[Y_n] = np(1 - p).$$

Aproximació de la Binomial per la Normal:

Són evidents les dificultats de càlcul amb la llei binomial. Molts cop aquestes dificultats es poden evitat suposant que $X \sim \mathrm{Bin}(n,p)$ en lloc de ser binomial, segueix una distribució normal $N(np, np(1-p))$.

Això s'anomena aproximar la Binomial per la Normal, i es pot fer quan n és prou gran. Com de gran? Depèn de p. En general s'accepta la regla següent:

$\mathrm{Bin}(n,p)$ és aproximable per $N(np, np(1-p))$ sempre que $np \geq 5$ i $n(1-p) \geq 5$.

Aproximació de la Binomial per la Normal:

Són evidents les dificultats de càlcul amb la llei binomial. Molts cop aquestes dificultats es poden evitat suposant que $X \sim \mathrm{Bin}(n,p)$ en lloc de ser binomial, segueix una distribució normal $N(np, np(1-p))$.

Això s'anomena aproximar la Binomial per la Normal, i es pot fer quan n és prou gran. Com de gran? Depèn de p. En general s'accepta la regla següent:

$\mathrm{Bin}(n,p)$ és aproximable per $N(np, np(1-p))$ sempre que $np \geq 5$ i $n(1-p) \geq 5$.

Exemple: Sigui $X \sim \mathrm{Bin}(10, 0.5)$ (la condició per a tenir una bona aproximació es satisfà justet).

$$P(2 \leq X \leq 4) = P(X=2)+P(X=3)+P(X=4) = 0.3663$$

Ara aproximen per la normal $X \approx N(5, 2.5)$ i calculem

$$P(1.5 \leq X \leq 4.5) = P(\frac{1.5-5}{\sqrt{2.5}} \leq Z \leq \frac{4.5-5}{\sqrt{2.5}}) = 0.3613,$$

on $Z \sim N(0,1)$.

Tema 3: Inferència estadística

3.1. Conceptes bàsics.

3.2. Distribucions mostrals.

3.3. Intervals de confiança.

3.3.1. Intervals de confiança per als paràmetres d'una població Normal.

3.3.2. Intervals de confiança per a mostres grans.

3.1 Conceptes bàsics

- Estadística descriptiva: Obtenir valors i gràfics que representin les **dades**.
- Estadística inferencial: Deduir coses sobre una **població** a partir de la informació subministrada per una **mostra**.

Es podria dir que en un cert sentit, l'estadística inferencial va en direcció contrària al càlcul de probabilitats.

Exemple:

Població 1000 vaques	Població de 1000 vaques
Mostra de 10 vaques	Mostra de 10 vaques
Sabem: 30% amb BSE	Observem: 3 amb BSE
Volem: Probabilitat de 3 amb BSE	Volem: quina proporció de BSE?

La primera situació ja sabem cóm es resol, considerant $X \sim Bin(10, 0.3)$ i calculant $P(X = 3) = \binom{10}{3}(0.3)^3(0.7)^7 = 0.2668$.

L'objectiu d'aquest tema és tractar amb situacions com la segona.

Quina seria la resposta més raonable? Un 30%, i.e. 300 vaques.

Però gairebé segur que ens equivoquem. Com podem arreglar aquesta aparent contradicció?

Recordem que havíem introduït la noció de Població com un conjunt d'objectes que es vol estudiar estadísticament.

Parlar d'una població, en sentit estadístic, és equivalent a parlar d'una variable aleatòria

$$\begin{aligned} X : \Omega = \text{població} &\longrightarrow \mathbb{R} = \{\text{nombres reals}\} \\ \text{individu} &\longmapsto \text{característica numèrica} \end{aligned}$$

La pregunta és: Quina és la llei de X? Això és el que voldríem (idealment) esbrinar a partir d'una mostra.

Exemple 1:

$$\begin{aligned} X : \Omega = \text{població} &\longrightarrow \mathbb{R} = \{\text{nombres reals}\} \\ \text{vaca} &\longmapsto \begin{cases} 1 & \text{si té BSE} \\ 0 & \text{sinó} \end{cases} \end{aligned}$$

La llei de X és binomial $Bin(1, p)$, però no coneixem p!

Exemple 2:

$$\begin{aligned} X : \Omega = \text{població} &\longrightarrow \mathbb{R} = \{\text{nombres reals}\} \\ \text{persona} &\longmapsto \text{alçada.} \end{aligned}$$

La llei de X és normal $N(\mu, \sigma^2)$, però no coneixem μ ni σ^2!

Tenim uns "models probabilístics" que no estan acabats de concretar.

Per precisar-lo una mica més, volem determinar els paràmetres (o almenys dir alguna cosa sobre ells).

Tenim un conjunt de lleis possibles:

- Exemple 1: $\{Bin(1, p) \ : \ 0 \leq p \leq 1\}$;
- Exemple 2: $\{N(\mu, \sigma^2) \ : \ \mu \in \mathbb{R}, \ \sigma^2 \geq 0\}$;

que s'anomenen **models estadístics**.

També és possible estudiar la situació en què no tenim ni idea de la llei que segueix la població (el model estadístic consisteix en totes les lleis de probabilitat). En aquest curs no estudiarem aquesta situació perquè és molt més difícil de tractar.

Nosaltres sempre tindrem un model estadístic d'aquest estil: el tipus de llei està determinat, però hi haurà paràmetres desconeguts.

Definició de Mostra:

Donada una població (equivalentment, una variable aleatòria X), una **mostra de mida** n de la població (equivalentment, de la variable aleatòria X) és una seqüència de n variables aleatòries $X_1, \ldots, X_n$ **independents**, totes amb la **mateixa llei de** X.

Realització d'una mostra: és la seqüència concreta de nombres obtinguts $x_1, \ldots, x_n$ després de realitzar l'experiment aleatori n vegades.

Exemple (BSE): $X \sim Bin(1, p)$. Agafem una vaca a l'atzar i considerem

$$X_1 \sim Bin(1, p) : \begin{cases} 1 & \text{si té BSE} \\ 0 & \text{sinó} \end{cases}$$

la deixem al seu lloc i agafem una altra vaca a l'atzar i considerem

$$X_2 \sim Bin(1, p) : \begin{cases} 1 & \text{si té BSE} \\ 0 & \text{sinó} \end{cases}$$

i així fins arribar a la mida $n = 10$ de la mostra.

En una realització concreta d'aquest experiment, obtindrem 10 nombres $x_1, \ldots, x_{10}$ com ara

$$1, 0, 0, 0, 1, 0, 1, 0, 0, 1.$$

Per tant, pensarem en **agafar una mostra** com la repetició, en idèntiques condicions de l'experiment aleatori "agafar un individu a l'atzar". De aquí que les variables aleatòries de la mostra siguin independents i amb la mateixa llei. Si agaféssim les 10 vaques de cop no estaríem en aquestes condicions (tot i que els càlculs serien aproximadament els mateixos si la població es prou gran).

Normalment no seran necessaris tots els valors individuals d'una realització de la mostra; n'hi haurà prou amb uns pocs valors que s'obtenen a partir d'aquests.

Definició d'Estadístic: Qualsevol funció de la mostra.

Exemples importants d'estadístics:

Sigui $X_1, \ldots, X_n$ una mostra de mida n d'una certa variable aleatòria X.

- Mitjana mostral:

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

  ($\bar{X}$ és una **variable aleatòria**! En una realització concreta de la mostra $x_1, \ldots, x_n$ estaríem calculant la mitjana $\bar{x}$,en el sentit d'estadística descriptiva, d'uns nombres concrets).

- Variància mostral:

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

  (mateixos comentaris que per $\bar{X}$).

Exemples importants d'estadístics:

- Mitjana mostral:

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Variància mostral:

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

- Variància mostral corregida:

$$\hat{S}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Desviació tipus mostral:

$$S := \sqrt{S^2}$$

- Desviació tipus mostral corregida:

$$\hat{S} := \sqrt{\hat{S}^2}$$

Exemple il.lustratiu dels conceptes introduïts:

Població de vaques: $\Omega = \{V_1, V_2, V_3, V_4, V_5\}$.
Vaques amb BSE: V_1, V_4

Població de mostres de mida 2 de Ω:

$$\Omega^2 = \left\{ \begin{array}{cccc} (V_1, V_1), & (V_1, V_2), & \cdots & (V_1, V_5) \\ (V_2, V_1), & (V_2, V_2), & \cdots & (V_2, V_5) \\ \vdots & \vdots & \ddots & \vdots \\ (V_5, V_1), & (V_5, V_2), & \cdots & (V_5, V_5) \end{array} \right\}.$$

Considerem les variables aleatòries $X_1, X_2 : \Omega \to \mathbb{R}$ definides per

$$X_i : V_1 \mapsto 1,\ V_2 \mapsto 0,\ V_3 \mapsto 0,\ V_4 \mapsto 1,\ V_5 \mapsto 0.$$

Tenim que la llei de X_i ve donada per

$$\begin{aligned} P(X_i = 0) &= P(\{V_2, V_3, V_5\}) = \frac{3}{5}, \\ P(X_i = 1) &= P(\{V_1, V_4\}) = \frac{2}{5}. \end{aligned}$$

Per tant,

$$
\begin{aligned}
E[X_i] &= 0\cdot\frac{3}{5}+1\cdot\frac{2}{5}=\frac{2}{5},\\
Var[X_i] &= E[X_i^2]-E[X_i]^2 = 0^2\cdot\frac{3}{5}+1^2\cdot\frac{2}{5}-\left(\frac{2}{5}\right)^2\\
&= \frac{2}{5}-\frac{4}{25}=\frac{6}{25}=0.24.
\end{aligned}
$$

Observem que són nombres fixes.

Totes les possibles realitzacions d'una mostra de mida 2 són:

$$
(X_1,X_2)\sim\left\{\begin{array}{ccccc}
11 & 01 & 01 & 11 & 01\\
10 & 00 & 00 & 10 & 00\\
10 & 00 & 00 & 10 & 00\\
11 & 01 & 01 & 11 & 01\\
10 & 00 & 00 & 10 & 00
\end{array}\right\}.
$$

Considerem l'estadístic $\bar{X} = \frac{X_1+X_2}{2}$ com una variable aleatòria definida sobre Ω^2, la població de mostres de mida 2:

$$(\bar{X}(V_i, V_j))_{i,j} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

Clarament la llei de $\bar{X}$ ve donada per

$$P(\bar{X} = 0) = \frac{9}{25}, \ P(\bar{X} = \frac{1}{2}) = \frac{12}{25}, \ P(\bar{X} = 1) = \frac{4}{25}.$$

Per tant,

$$\begin{aligned} E[\bar{X}] &= 0 \cdot \frac{9}{25} + \frac{1}{2} \cdot \frac{12}{25} + 1 \cdot \frac{4}{25} = \frac{10}{25} = \frac{2}{5} = E[X_i], \\ Var[\bar{X}] &= 0^2 \cdot \frac{9}{25} + \left(\frac{1}{2}\right)^2 \cdot \frac{12}{25} + 1^2 \cdot \frac{4}{25} - \left(\frac{2}{5}\right)^2 = \frac{3}{25} \\ &= \frac{Var[X_i]}{2}. \end{aligned}$$

Els valors de $\bar{X}$ varien menys que els de X_i.

També podríem haver calculat la llei de $\bar{X}$ d'una altra manera: $X_1, X_2 \sim Bin(1, \frac{2}{5})$ independents, per tant $X_1 + X_2 \sim Bin(2, \frac{2}{5})$ i

$$
\begin{aligned}
P(\bar{X} = 0) &= P(X_1 + X_2 = 0) = \binom{2}{0}\left(\frac{2}{5}\right)^0\left(\frac{3}{5}\right)^2 \\
&= \frac{9}{25} \\
P(\bar{X} = \frac{1}{2}) &= P(X_1 + X_2 = 1) = \binom{2}{1}\left(\frac{2}{5}\right)^1\left(\frac{3}{5}\right)^1 \\
&= \frac{12}{25} \\
P(\bar{X} = 1) &= P(X_1 + X_2 = 2) = \binom{2}{2}\left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right)^0 \\
&= \frac{4}{25}
\end{aligned}
$$

Considerem ara l'estadístic

$$\begin{aligned} S^2 &= \frac{1}{2}\sum_{i=1}^{2}(X_i - \bar{X})^2 \\ &= \frac{1}{2}\left[\left(X_1 - \frac{X_1+X_2}{2}\right)^2 + \left(X_2 - \frac{X_1+X_2}{2}\right)^2\right] \end{aligned}$$

com a variable aleatòria definida sobre Ω^2:

$$(S^2(V_i, V_j))_{i,j} = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 \end{pmatrix}$$

Així la llei de S^2 ve donada per

$$P(S^2 = 0) = \frac{13}{25}, \; P(S^2 = \frac{1}{4}) = \frac{12}{25},$$

i per tant,

$$E[S^2] = 0 \cdot \frac{13}{25} + 1 \cdot \frac{12}{25} = \frac{3}{25} = \frac{Var[X_i]}{2},$$

$$Var[S^2] = 0^2 \cdot \frac{13}{25} + \left(\frac{1}{4}\right)^2 \frac{12}{25} - \left(\frac{3}{25}\right)^2 = \frac{39}{2500} = 0.0156$$

Per acabar aquest exemple, considerem l'estadístic

$$\hat{S}^2 = \frac{1}{2-1}\sum_{i=1}^{2}(X_i - \bar{X})^2 = 2S^2$$

com a variable aleatòria definida sobre Ω^2:

$$(\hat{S}^2(V_i, V_j))_{i,j} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

La llei de $\hat{S}^2$ ve donada per

$$P(\hat{S}^2 = 0) = \frac{13}{25}, \; P(\hat{S}^2 = \frac{1}{2}) = \frac{12}{25},$$

i per tant,

$$E[\hat{S}^2] = 0 \cdot \frac{13}{25} + \frac{1}{2} \cdot \frac{12}{25} = \frac{6}{25} = Var[X_i],$$

$$Var[\hat{S}^2] = 0^2 \cdot \frac{13}{25} + \left(\frac{1}{2}\right)^2 - \left(\frac{6}{25}\right)^2 = \frac{39}{625}.$$

3.2 Distribucions mostrals

Els estadístics són variables aleatòries, i per tant tenen una certa llei. Per fer inferència estadística farem servir els estadístics dels exemples anteriors (o derivats d'ells). Per tant, és interessant saber quina llei tenen. Aquestes lleis s'anomenen **distribucions mostrals**.

Comencem amb el cas particular de les lleis normals.

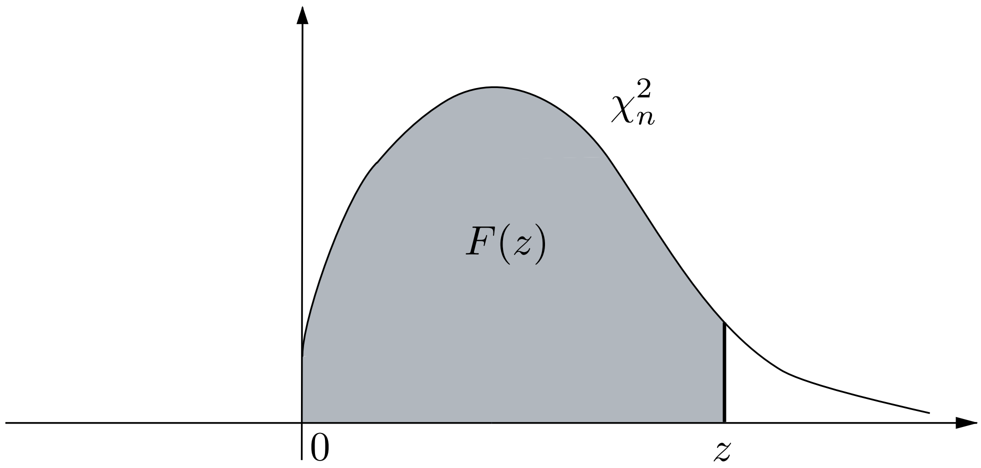
Teorema: Si $X_1, \ldots, X_n$ és una mostra de $X \sim N(\mu, \sigma^2)$ aleshores $\bar{X}$ i $\hat{S}^2$ són independents, i

(1) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$,

(2) $\frac{n-1}{\sigma^2}\hat{S}^2 \sim \chi^2(n-1)$ (khi-quadrat amb $n-1$ graus de llibertat),

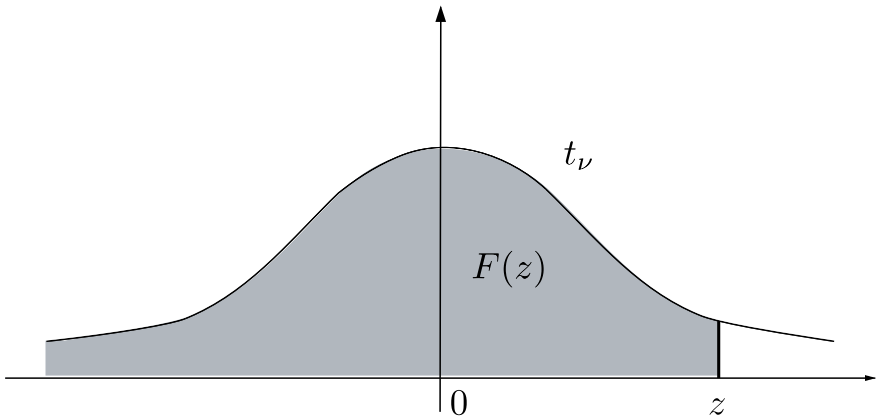
(3) $\frac{\bar{X}-\mu}{\frac{\hat{S}}{\sqrt{n}}} \sim t(n-1)$ (t de Student amb $n-1$ graus de llibertat).

La variable aleatòria $\chi^2(n)$ és contínua amb funció de densitat:



Si $X_1, \ldots, X_n \sim N(0,1)$ independents, aleshores $X_1^2 + \cdots + X_n^2 \sim \chi^2(n)$.

La variable aleatòria $t(n)$ és contínua amb funció de densitat semblant a la de la normal:



Com més gran és n, més s'assemblen: "$t(n) \xrightarrow{n\to\infty} N(0,1)$".

A més, si $X \sim N(0,1)$ i $Y \sim \chi^2(n)$ són independents, aleshores $\frac{X}{\sqrt{Y/n}} \sim t(n)$.

Com a conseqüència del teorema anterior, tenim que

$$E[\bar{X}] = \mu = E[X] \quad \text{i} \quad Var[\bar{X}] = \frac{\sigma^2}{n}.$$

Però això és compleix en **general**: Si $X_1, \ldots, X_n$ és una mostra de **qualsevol** variable aleatòria X, aleshores

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n} nE[X] = E[X] \\ Var[\bar{X}] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \\ &= \frac{1}{n^2} nVar[X] = \frac{Var[X]}{n}. \end{aligned}$$

La "població" de totes les possibles mitjanes mostrals $\bar{x}$ té per "valor central" $E[X]$.

Per tant, $\bar{X}$ ens dóna una "bona" estimació de:

- μ si $X \sim N(\mu, \sigma^2)$,
- λ si $X \sim Poiss(\lambda)$,
- p si $X \sim Bin(1, p)$ (en aquest últim cas, p coincideix amb la proporció d'individus que tenen la característica que mesura X).

Com de bona és aquesta estimació? Millor com més gran sigui la mostra, ja que $Var[\bar{X}] = \frac{Var[X]}{n}$.

Què hi ha de S^2?

$$\begin{aligned}
S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-E[X])-(\bar{X}-E[X])\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i-E[X])^2+(\bar{X}-E[X])^2\right. \\
&\quad \left.-2(X_i-E[X])(\bar{X}-E[X])\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(X_i-E[X])^2+n(\bar{X}-E[X])^2\right. \\
&\quad \left.-2(\bar{X}-E[X])\sum_{i=1}(X_i-E[X])\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(X_i-E[X])^2-n(\bar{X}-E[X])^2\right].
\end{aligned}$$

Què hi ha de S^2?

$$\begin{aligned}
S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-E[X])-(\bar{X}-E[X])\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i-E[X])^2+(\bar{X}-E[X])^2\right. \\
&\quad \left.-2(X_i-E[X])(\bar{X}-E[X])\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(X_i-E[X])^2+n(\bar{X}-E[X])^2\right. \\
&\quad \left.-2(\bar{X}-E[X])\sum_{i=1}(X_i-E[X])\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(X_i-E[X])^2-n(\bar{X}-E[X])^2\right].
\end{aligned}$$

$$\begin{aligned}
E[S^2] &= E\left[\frac{1}{n}\left(\sum_{i=1}^{n}(X_i-E[X_i])^2-n(\bar{X}-E[X])^2\right)\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}E[(X_i-E[X_i])^2]-nE[(\bar{X}-E[X])^2]\right) \\
&= \frac{1}{n}\left(Var[X_i]-nVar[\bar{X}]\right) \\
&= \frac{1}{n}(nVar[X]-Var[X]) = \frac{n-1}{n}Var[X].
\end{aligned}$$

Així doncs,

$$\begin{aligned} E[\hat{S}^2] &= E[\frac{n}{n-1}S^2] = \frac{n}{n-1}E[S^2] \\ &= \frac{n}{n-1}\frac{n-1}{n}Var[X] = Var[X]. \end{aligned}$$

Al igual que $\bar{X}$ "estimava bé" $E[X]$, ara veiem que

$\hat{S}^2$ "estima bé" $Var[X]$.

D'aquí la utilitat de la **variància mostral corregida**.

No escriurem la $Var[S^2]$ ó $Var[\hat{S}^2]$ perquè és una expressió complicada que no necessitarem.

Cas binomial

Recordem que $X \sim Bin(1,p)$ modelitza la situació: $X = 1$ si l'individu té una certa característica i $X = 0$ sinó, on p és la proporció d'individus amb aquesta característica.

Què podem dir sobre la llei d'algun estadístic en aquest cas?

Com que $n\bar{X} = \sum_{i=1}^{n} X_i \sim Bin(n,p)$, per n prou gran

$$n\bar{X} \approx N(np, np(1-p))$$

i per tant,

$$\frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{n\bar{X}-np}{\sqrt{np(1-p)}} \approx N(0,1).$$

3.3 Intervals de confiança

La inferència estadística té dues parts:

- Estimació de paràmetres
  - Estimació puntual mitjançant els estadístics de l'apartat anterior. Volem estadístics que tinguin per esperança el paràmetre que volem estimar. I entre aquests, volem els que tinguin variància mínima (observem que $Var[\bar{X}] \to 0$ quan $n \to \infty$). Exemple: usem $\bar{X}$ per estimar l'esperança de X i $\hat{S}^2$ per estimar la variància de X.
  - Estimació per intervals
- Tests d'hipòtesis

En l'estimació per intervals, en lloc de donar un valor concret del paràmetre, es dóna un interval en el que creiem que es troba.

Exemple: Suposem que ens demanin estimar el pes d'una persona només mirant-la. Podríem dir: $60Kg$ (estimació puntual). Si tenim bon ull, potser ens haurem aproximat molt, però segur que la nostra estimació és falsa, en el sentit que no haurem encertat **exactament**.

Suposem que ens permeten donar un interval de longitud 10 en lloc d'un sol valor. Llavors direm $[55, 65]$ i estarem molt més segurs que abans d'encertar. Tindrem més **confiança** en la nostra estimació.

Si ens permeten un interval de longitud 20, direm $[50, 70]$ i la nostra confiança augmentarà. Estarem molt més a prop de la certesa absoluta d'encertar.

Observem el compromís entre la **informació** que es dóna i el grau de **confiança** en ella.

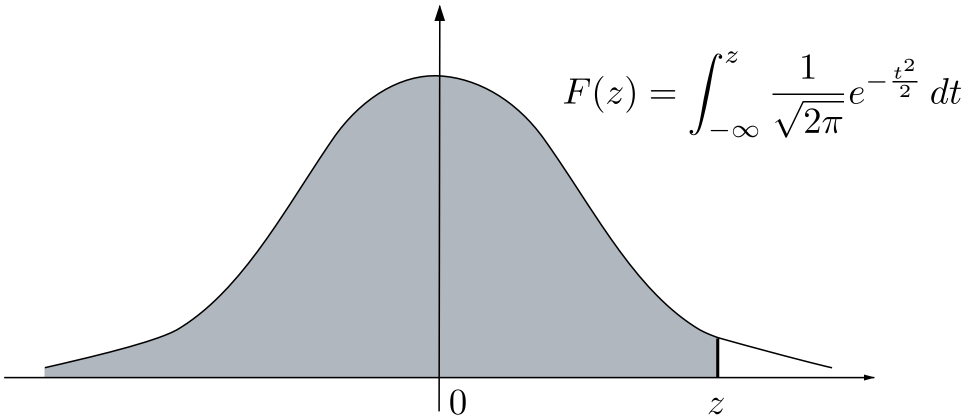
Exemple:

Es fa un examen de coneixements a estudiants de secundària cada any. Cada promoció pot venir amb diferents graus de preparació, però suposem que la experiència passada diu que les puntuacions segueixen una distribució normal amb variància $\sigma^2 = 4$.

Tot just acabada la prova, corregim els exàmens d'una petita mostra, per intentar estimar quina serà la mitjana μ de tota la població. Suposem que la mida de la mostra és $n = 10$ i la mitjana mostral és $\bar{x} = 5.4$ (aquesta seria la millor estimació puntual que podríem donar). Anem a fer una cosa més interessant:

Sabem que $X \sim N(\mu, \sigma^2)$, d'on $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ i per tant $Z := \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. Busquem un nombre a tal que

$$P(-a \leq Z \leq a) = 0.95$$



Això ho sabem fer. S'obté de les taules que $a = 1.96$.

Per tant, com $\sigma = 2$ i $n = 10$,

$$\begin{aligned}
0.95 &= P\left(-1.96 \leq \frac{\bar{X}-\mu}{2/\sqrt{10}} \leq 1.96\right) \\
&= P\left(-1.96\frac{2}{\sqrt{10}} \leq \bar{X}-\mu \leq 1.96\frac{2}{\sqrt{10}}\right) \\
&= P\left(-1.96\frac{2}{\sqrt{10}} - \bar{X} \leq -\mu \leq 1.96\frac{2}{\sqrt{10}} - \bar{X}\right) \\
&= P\left(1.96\frac{2}{\sqrt{10}} + \bar{X} \geq \mu \geq -1.96\frac{2}{\sqrt{10}} + \bar{X}\right) \\
&= P\left(\bar{X} - 1.96\frac{2}{\sqrt{10}} \leq \mu \leq \bar{X} + 1.96\frac{2}{\sqrt{10}}\right)
\end{aligned}$$

Amb la mostra concreta que tenim $\bar{x} = 5.4$ i obtenim l'interval $[4.1604, 6.6396]$. Diem que μ està en aquest interval amb una confiança del 0.95.

Interpretació:

NO és que $P(4.1604 \leq \mu \leq 6.6396) = 0.95$.

Això no té sentit, ja que μ és un nombre fix ben definit (la mitjana de tota la població que la coneixerem quan corregim tots els exàmens) i per tant, la probabilitat de que estigui a l'interval en qüestió és 1 o 0 depenent si hi és o no.

La interpretació correcta és: de tots els possibles intervals del tipus

$$[\bar{X} - 1.96\frac{2}{\sqrt{10}}, \bar{X} + 1.96\frac{2}{\sqrt{10}}]$$

(recordem que $\bar{X}$ és una variable aleatòria), el 95% d'ells contenen el nombre fix, però desconegut, μ. No sabem si l'interval concret $[4.1604, 6.6396]$ és d'aquests o bé del 5% restant.

Exemple (continuació):

Si volem estar-ne més segurs, podem prendre com nivell de confiança $NC = 0.99$ en lloc de 0.95. Si repetim els càlculs, $a = 2.58$ i

$$[\bar{x} - a\frac{2}{\sqrt{10}}, \bar{x} + a\frac{2}{\sqrt{10}}] = [3.7746, 7.0214].$$

Confiança més gran ⇒ Interval més gran ⇒ Menys informació.

Suposem que ens conformem amb una confiança de 0.90 alehores $a = 1.64$ i l'interval surt més petit:

$$[\bar{x} - a\frac{2}{\sqrt{10}}, \bar{x} + a\frac{2}{\sqrt{10}}] = [4.3628, 6.4372].$$

Confiança més petita ⇒ Interval més petit ⇒ Més informació.

Suposem $NC = 0.90$ però que tenim $n = 100$ en lloc de $n = 10$. Què passarà? L'interval es fa més petit:

$$[\bar{x} - 1.64\frac{2}{\sqrt{100}}, \bar{x} + 1.64\frac{2}{100}] = [5.072, 5.728]$$

Mostra més gran ⇒ Interval més petit ⇒ Més informació.

3.3.1 Intervals de confiança per als paràmetres de $X \sim N(\mu, \sigma^2)$

(a) Interval de confiança per μ si σ^2 és coneguda

És el que hem fet a l'exemple anterior.

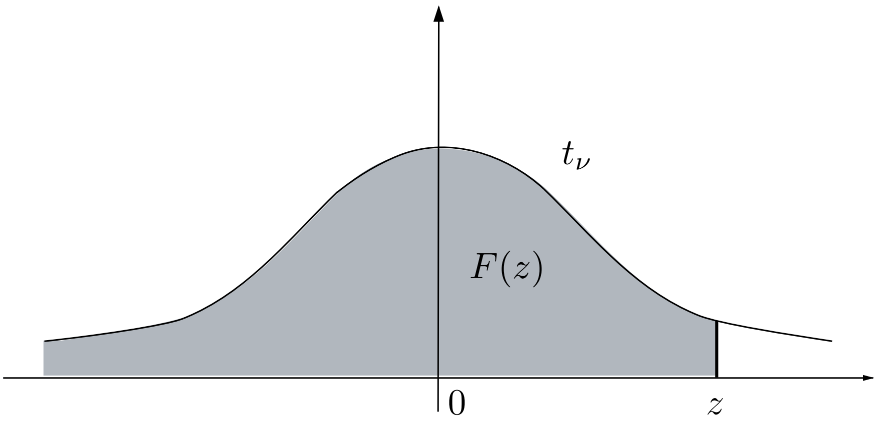
(b) Interval de confiança per μ si σ^2 és desconeguda

Ara no és possible usar $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ com abans perquè σ no és coneguda. Usarem

$$\frac{\bar{X}-\mu}{\frac{\hat{S}}{\sqrt{n}}} \sim t(n-1).$$

Exemple (continuació): Suposem ara que σ^2 és desconeguda, $n = 10$, $\bar{x} = 5.4$ i $\hat{s} = 2$. Aleshores $\frac{\bar{X}-\mu}{\hat{S}/\sqrt{10}} \sim t(9)$ i imposant

$$P\left(-a \leq \frac{\bar{X}-\mu}{\hat{S}/\sqrt{10}} \leq a\right) = 0.95.$$



la taula ens dóna el valor $a = 2.262$.

Per tant,

$$\begin{aligned} 0.95 &= P\left(-2.262 \leq \frac{\bar{X}-\mu}{\hat{S}/\sqrt{10}} \leq 2.262\right) \\ &= P\left(-2.262\frac{\hat{S}}{\sqrt{10}} - \bar{X} \leq -\mu \leq 2.262\frac{\hat{S}}{\sqrt{10}} - \bar{X}\right) \\ &= P\left(\bar{X} - 2.262\frac{\hat{S}}{\sqrt{10}} \leq \mu \leq \bar{X} + 2.262\frac{\hat{S}}{\sqrt{10}}\right). \end{aligned}$$

L'interval és

$$\left[\bar{X} - 2.262\frac{\hat{S}}{\sqrt{10}}, \bar{X} + 2.262\frac{\hat{S}}{\sqrt{10}}\right]$$

i utilitzant la realització de la mostra que tenim ($\bar{x} = 5.4$ i $\hat{s} = 2$) obtenim $[3.9694, 6.8306]$

Exemple: Tenim una població de rates de laboratori adultes. Volem estimar el seu pes mitjà amb un interval de confiança al nivell 0.95. Agafem una mostra de $n = 9$ rates i trobem $\bar{x} = 256.2$ grams i $\hat{s} = 8.23$ grams.

Resolució: Suposem que el pes és una variable aleatòria normal $N(\mu, \sigma^2)$. Volem un interval per a μ sense conèixer σ^2. Per això, calculem a tal que

$$P\left(-a \leq \frac{\bar{X} - \mu}{\hat{S}/\sqrt{9}} \leq a\right) = 0.95 \quad \xrightarrow{\text{Taula } t(8)} \quad a = 2.306$$

Per tant, l'interval de confiança és

$$\left[\bar{X} - 2.306\frac{\hat{S}}{\sqrt{9}}, \bar{X} + 2.306\frac{\hat{S}}{\sqrt{9}}\right]$$

i la realització concreta

$$[249.8739, 262.5261].$$

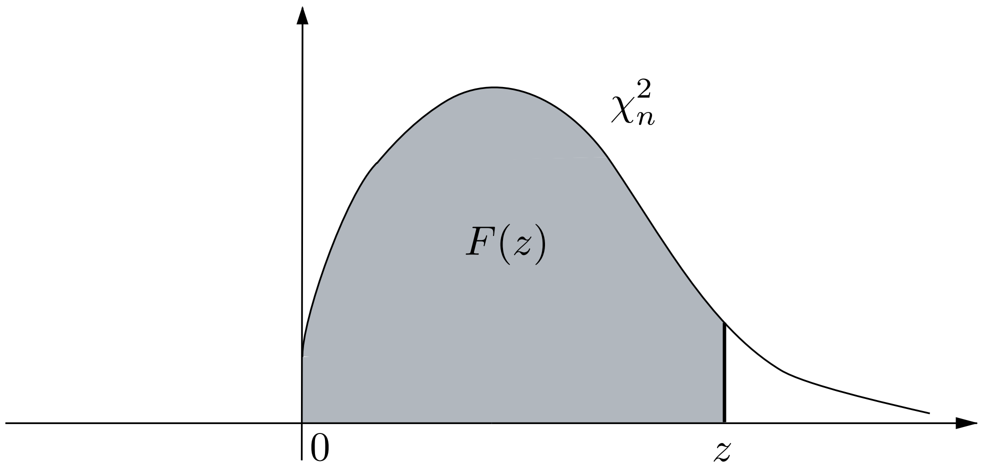
(c) Interval de confiança per σ^2 si μ és desconeguda

En aquest cas utilitzem que

$$\frac{n-1}{\sigma^2}\hat{S}^2 \sim \chi^2(n-1)$$

per buscar nombres a i b tals que

$$P\left(a \leq \frac{n-1}{\sigma^2}\hat{S}^2 \leq b\right) = NC = 0.95$$



Hi ha moltes maneres d'agafar a i b però ho farem deixant la mateixa probabilitat a banda i banda, i.e. de manera que

$$\begin{aligned} P(a \leq \chi^2(n-1)) &= \frac{1-NC}{2} = 0.025, \\ P(\chi^2(n-1) \leq b) &= \frac{1+NC}{2} = 0.975 \end{aligned}$$

Continuació de l'exemple anterior: Mostra de mida $n = 9$, $\bar{x} = 256.2$ grams i $\hat{s} = 8.23$ grams.

La taula ens proporciona $a = 2.1797$ i $b = 17.5345$.

Per tant,

$$\begin{aligned}
0.95 &= P\left(2.1797 \leq \frac{8}{\sigma^2}\hat{S}^2 \leq 17.5345\right) \\
&= P\left(\frac{2.1797}{8\hat{S}^2} \leq \frac{1}{\sigma^2} \leq \frac{17.5345}{8\hat{S}^2}\right) \\
&= P\left(\frac{8\hat{S}^2}{2.1797} \geq \sigma^2 \geq \frac{8\hat{S}^2}{17.5345}\right) \\
&= P\left(\frac{8\hat{S}^2}{17.5345}\sigma^2 \leq \sigma \leq \frac{8\hat{S}^2}{2.1797}\right)
\end{aligned}$$

I l'interval amb el 95% de confiança és

$$\left[\frac{8\hat{S}^2}{17.5345}\sigma^2, \frac{8\hat{S}^2}{2.1797}\right]$$

i la seva realització concreta a partir de les dades de la mostra:

$$[30.9, 248.59]$$

(d) Interval de confiança per σ^2 si μ és coneguda

Aquest cas no es sol donar, però ho fem per completar l'exposició. En el cas anterior, usàvem $\frac{n-1}{\sigma^2}\hat{S}^2 \sim \chi^2(n-1)$. Si a la fórmula

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

substituïm $\bar{X}$ per μ obtenim[a]

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

Amb la qual cosa surten uns intervals una mica més petits que si no usem μ.

[a]Una suma de quadrats de $N(0,1)$ independents.

Exemple anterior (continuació): Mostra de mida $n = 9$, $\hat{s} = 8.23$ grams, $NC = 0.95$. Suposem que $\mu = 262.2$ i $\sum_{i=1}^{9}(x_i - 256.2)^2 = 541.8632$.

Aleshores $\frac{1}{\sigma^2}\sum_{i=1}^{9}(X_i - 262.2)^2 \sim \chi^2(9)$. En aquest cas s'obté $a = 2.7004$, $b = 19.0228$, l'interval de confiança és

$$\left[\frac{\sum_{i=1}^{9}(X_{i-262.2})^2}{19.0228}, \frac{\sum_{i=1}^{9}(X_{i-262.2})^2}{2.7004}\right]$$

i la seva realització $[28.51, 200.87]$.

Observem la diferència d'amplades gairebé amb les mateixes dades: (c) 217.69 (d) 172.36.

Exercici: La variació en la potència d'una dosis fixada de medicament ha de ser petita. Volem estudiar la variància de la potència en un lot d'un cert medicament. Es pren una mostra de mida $n = 20$ i s'obté $\hat{s}^2 = 0.0018$. Es demana un interval amb nivell de confiança de 99% per a la variància de la potència, suposant normalitat.

Resposta: $[0.0009, 0.0050]$.

3.3.2 Intervals de confiança per a mostres grans ($n \geq 30$)

En aquest punt no suposarem que $X \sim N(\mu, \sigma^2)$. De fet, X podrà seguir qualsevol llei, però només farem intervals de confiança per a la mitjana poblacional $E[X]$.

Hi intervindrà la mitjana mostral $\bar{X}$, que sabem que és un "bon estimador" de $E[X]$.

Recordem que, en el cas $X \sim N(\mu, \sigma^2)$, teníem $\bar{X} \sim N(\mu, \sigma^2/n)$. En general,

$$\bar{X} = \frac{X_1}{n} + \frac{X_2}{n} + \cdots \frac{X_n}{n}.$$

Si n és gran, $\bar{X}$ és la suma de "petites contribucions" independents. Un bon model per aquesta situació (basat en un teorema important de teoria de probabilitats) és suposar que $\bar{X}$ és normal, encara que X no ho sigui. A més, sabem

$$E[\bar{X}] = E[X], \qquad Var[\bar{X}] = \frac{Var[X]}{n}.$$

Estem afirmant doncs, que per n gran podem aproximar

$$\bar{X} \approx N\left(E[X], \frac{Var[X]}{n}\right).$$

Si ara posem, per abreviar, $\mu := E[X]$ i $\sigma^2 := Var[X]$, tenim que

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0, 1).$$

Aquesta aproximació es sol admetre quan $n \geq 30$.

(a) Interval de confiança per μ si σ^2 és coneguda

Utilitzem

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

i procedim com al punt 3.3.1.(a).

(b) Interval de confiança per μ si σ^2 és desconeguda

Substituïm σ^2 pel seu "bon estimador" $\hat{S}^2$, usem

$$\frac{\bar{X}-\mu}{\hat{S}/\sqrt{n}} \approx N(0,1)$$

i procedim com al punt 3.3.1.(b). substituint $t(n-1)$ per $N(0,1)$.

Exemple (BSE):

Tenim una població de vaques. Volem estimar la proporció de vaques amb BSE, amb un nivell de confiança del 95% sabent que d'una mostra de mida $n = 30$, 9 tenen BSE.

Resolució: Considerem la variable aleatòria

$$X = \begin{cases} 1 & \text{si la vaca té BSE} \\ 0 & \text{si no} \end{cases}$$

Llavors, $X \sim Bin(1, p)$ on $p = E[X]$ és la proporció de vaques amb BSE. Tenim $n = 30$ i $\bar{x} = \frac{9}{30}$. Necessitem també $\hat{s}$, però ho podem deduir (en el cas binomial mitjana i variància estan lligades):

$$\begin{aligned} \hat{s}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \frac{1}{29} \left[9 \cdot \left(1 - \frac{9}{30}\right)^2 + 21 \cdot \left(0 - \frac{9}{30}\right)^2 \right] \\ &= 0.2172 \end{aligned}$$

Per tant, $\hat{s} = 0.4661$.

Aproximant $\frac{\bar{X}-p}{\frac{\hat{S}}{\sqrt{30}}}$ per una $N(0,1)$ tenim que

$$\begin{aligned}
0.95 &= P\left(-1.96 \leq \frac{\bar{X}-p}{\frac{\hat{S}}{\sqrt{30}}} \leq 1.96\right) \\
&= P\left(-1.96\frac{\hat{S}}{\sqrt{30}} - \bar{X} \leq -p \leq 1.96\frac{\hat{S}}{\sqrt{30}} - \bar{X}\right) \\
&= P\left(\bar{X} - 1.96\frac{\hat{S}}{\sqrt{30}} \leq p \leq \bar{X} + 1.96\frac{\hat{S}}{\sqrt{30}}\right)
\end{aligned}$$

Per tant, l'interval de confiança és

$$\left[\bar{X} - 1.96\frac{\hat{S}}{\sqrt{30}}, \bar{X} + 1.96\frac{\hat{S}}{\sqrt{30}}\right]$$

i una realització concreta és

$$[0.1332, 0.4668].$$

Podem dir: "p està entre el 13.32% i el 46.68%".

No diem massa. Doncs, això es tot el que es pot dir amb una mostra de mida 30.

Pregunta: Si volem reduir l'interval a la meitat de longitud, aproximadament, quina mida n de mostra necessitem?

Resposta:

$$\frac{1}{2}\left(1.96\frac{\hat{S}}{\sqrt{30}}\right) = 1.96\frac{\hat{S}}{\sqrt{n}} \Rightarrow n = 2^2 \cdot 30 = 120.$$

Observació: Diem "aproximadament" la meitat, perquè $\hat{S}$ canviarà una mica al canviar la mostra.

Tema 4: Tests d'hipòtesis

4.1 Definicions i exemples.

4.2 Tests sobre els paràmetres d'una població Normal.

4.3 Tests sobre els paràmetres de dues poblacions Normals.

4.3.1. Mostres independents.

4.3.2. Dades aparellades.

4.4 Tests per a mostres grans.

4.4.1. Una sola població.

4.4.2. Mostres independents.

4.4.3. Dades aparellades.

4.5 Tests per a mostes petites.

4.5.1. Mostres independents.

4.5.2. Dades aparellades.

4.6 Tests d'ajustament i independència.

4.1 Introducció

Recordem que la Inferència Estadística comprèn:

- Estimació de paràmetres:
  - Estimació puntual.
  - Estimació per intervals de confiança.
- Tests d'hipòtesis.

En els tests d'hipòtesis es tracta, no d'estimar un paràmetre desconegut de la població, sinó només de contestar amb SI o NO una pregunta sobre la població. Més precisament, decidir entre dues alternatives que es proposem com hipòtesis.

Exemple: Un cert medicament té un percentatge d'èxit reconegut del 84% per tractar una determinada malaltia. Un investigador creu haver trobat un remei que donarà un percentatge d'èxits més gran. Ho experimenta amb una mostra de pacients i obté una proporció d'èxit de 86%.

Preguntes: Hi ha prou evidència que el nou medicament és millor? És significativa la diferència observada? O bé, val la pena el cost de posar en marxa un nou medicament?

Exemple: Un cert medicament té un percentatge d'èxit reconegut del 84% per tractar una determinada malaltia. Un investigador creu haver trobat un remei que donarà un percentatge d'èxits més gran. Ho experimenta amb una mostra de pacients i obté una proporció d'èxit de 86%.

Preguntes: Hi ha prou evidència que el nou medicament és millor? És significativa la diferència observada? O bé, val la pena el cost de posar en marxa un nou medicament?

Necessitem construir una **regla de decisió** que ens digui quina decisió prendre en funció de la proporció obtinguda a la mostra. Tot i així, com que es tracta d'un procediment estadístic ens podem equivocar amb la decisió, i de dues maneres:

- Dient que el nou medicament és millor que l'antic, quan no ho és.
- Dient que el nou medicament no és millor, quan sí que ho és.

Anem a descriure bé les definicions dels conceptes que s'apunten a l'exemple.

Definicions

Un **test d'hipòtesis** consisteix en el planteig de dues **hipòtesis estadístiques** contradictòries i en una **regla de decisió** que permeti decidir, a partir de la informació subministrada per una mostra, quina de les hipòtesis és la correcta.

Les dues hipòtesis s'anomenen:

- H_0: Hipòtesi Nul.la
- H_1: Hipòtesi Alternativa

I no juguem un paper simètric:

H_0 no serà rebutjada a no ser que hi hagi una forta evidència en contra. Per tant, escollirem H_0 com la hipòtesi que resulti més perillós (o costós, etc...) rebutjar-la quan és certa. També quan impliqui **canvi en el coneixement establert**.

Exemple (medicament): Sigui X la variable aleatòria que val 1 si el nou medicament ha funcionat i 0 sinó. És a dir, $X \sim \text{Bin}(1, p)$ amb p la probabilitat d'èxit del nou medicament. Volem decidir si $p > 0.84$ o bé $p \leq 0.84$.

Plantejament de les hipòtesis:

$$\begin{cases} H_0: & p \leq 0.84 \\ H_1: & p > 0.84 \end{cases}$$

Errors que podem cometre:

- Error de tipus I: Rebutjar H_0 essent certa en realitat.
- Error de tipus II: Acceptar H_0 essent falsa en realitat.

Decisió\Realitat	H_0 certa	H_0 falsa
H_0 s'accepta	No Error	Error Tipus II
H_0 es rebutja	Error Tipus I	No Error

El que s'intenta sempre és controlar l'error de tipus I, en el sentit que la probabilitat de cometre'l sigui petita.

La probabilitat de cometre l'error de tipus I la fixem nosaltres i s'anomena **nivell de significació**.

Exemple complet:

Suposem que la mitjana de la nota de selectivitat a Catalunya és 5 i que en una certa comarca, una mostra de $n = 10$ té una mitjana $\bar{x} = 5.4$. Suposem que els resultats en aquesta comarca segueixen una llei $N(\mu, \sigma^2)$ i se sap que $\sigma^2 = 4$.

Pregunta: Podem concloure que en nivell dels estudiants en aquesta comarca és superior a la mitjana de Catalunya, amb un nivell de significació 0.05?

Resolució: Sabem $X \sim N(\mu, 4)$. Plantejament de les hipòtesis:

$$\begin{cases} H_0 : & \mu \leq 5 \\ H_1 : & \mu > 5 \end{cases}$$

Volem fer càlculs suposant H_0 certa. Però en H_0, μ pot ser qualsevol nombre fins a 5 i per fer els càlculs necessitem fixar un valor concret. Quin agafem? El valor límit $\mu = 5$ que és el que més s'assembla a $\bar{x} = 5.4$.

Si $\mu = 5$, llavors $Z = \frac{\bar{X}-5}{2/\sqrt{10}} \sim N(0, 1)$.

Sota H_0, la probabilitat de que Z sigui gran ha de ser petita. Com de petita? El nivell de significació NS:

$$P(Z \geq a) = NS = 0.05 \quad \xrightarrow{\text{Taula}} \quad a = 1.64.$$

Si H_0 és certa aleshores, amb una probabilitat molt alta (0.95), el valor $\frac{\bar{x}-5}{2/\sqrt{10}}$ calculat a partir de la mostra hauria de ser més petit que $a = 1.64$. Si no ho és, vol dir que no ens creiem H_0.

La regla de decisió és donar per a Z:

- Regió d'acceptació de H_0: $(-\infty, 1.64]$
- Regió de rebuig de H_0: $(1.64, \infty)$

En el nostre cas, $\bar{x} = 5.4$ i $\frac{\bar{x}-5}{2/\sqrt{10}} = 0.6325$ que està a la regió d'acceptació. Per tant, ens quedem amb H_0: No hi ha prou evidència per a la hipòtesi que els estudiants d'aquesta comarca són millors (amb una probabilitat d'equivocar-nos de 0.05).

$$P(Z \geq a) = NS = 0.05 \quad \xrightarrow{\text{Taula}} \quad a = 1.64.$$

Si H_0 és certa aleshores, amb una probabilitat molt alta (0.95), el valor $\frac{\bar{x}-5}{2/\sqrt{10}}$ calculat a partir de la mostra hauria de ser més petit que $a = 1.64$. Si no ho és, vol dir que no ens creiem H_0.

La regla de decisió és donar per a Z:

- Regió d'acceptació de H_0: $(-\infty, 1.64]$
- Regió de rebuig de H_0: $(1.64, \infty)$

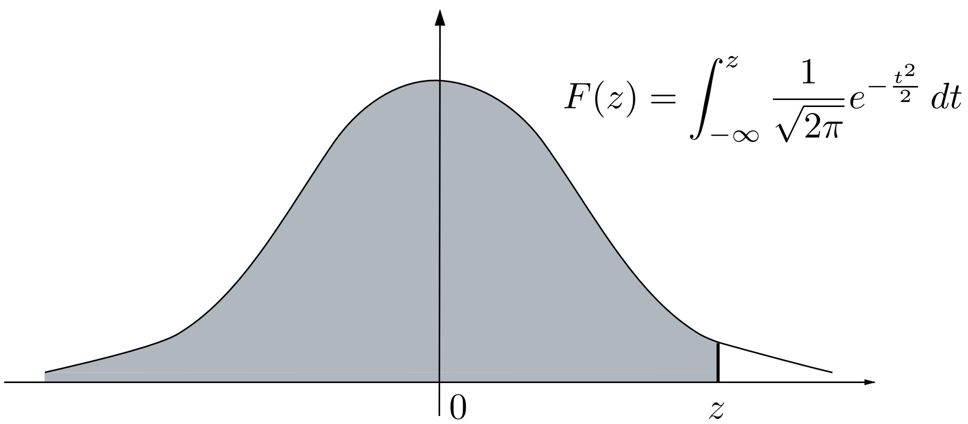
En el nostre cas, $\bar{x} = 5.4$ i $\frac{\bar{x}-5}{2/\sqrt{10}} = 0.6325$ que està a la regió d'acceptació. Per tant, ens quedem amb H_0: No hi ha prou evidència per a la hipòtesi que els estudiants d'aquesta comarca són millors (amb una probabilitat d'equivocar-nos de 0.05).

Si la mostra fos de mida $n = 100$, amb el mateix nivell de significació, les regions d'acceptació i rebuig serien les mateixes, però l'estadístic seria diferent $Z = \frac{\bar{X}-5}{2/\sqrt{100}} \sim N(0,1)$, així com el valor associat a la mostra: $\frac{\bar{x}-5}{2/\sqrt{100}} = 2$ que ara està a la regió de rebuig.

Amb una mostra de mida $n = 100$, rebutjaríem H_0 i podríem afirmar que aquests estudiants són millors (amb una probabilitat d'equivocar-nos de 0.05).

Continuació de l'exemple: Si la mostra fos de mida $n = 100$ però volem disminuir fins al 0.01 la probabilitat d'equivocar-nos al rebutjar la hipòtesi nul.la (i.e. prendre com a nivell de significació $NS = 0.01$) aleshores:

$$P\left(\frac{\bar{X}-5}{2/\sqrt{100}} \geq z\right) = 0.01 \quad \xrightarrow{\text{Taula}} \quad z = 2.33.$$



Per tant, la regió de rebuig en aquest cas és $(2.33, \infty)$. Com que el valor de l'estadístic segons la nostra mostra $\frac{\bar{x}-5}{2/\sqrt{100}} = 2$ cau a la regió d'acceptació, ara no podríem rebutjar H_0.

Una pregunta natural és: Amb les dades d'una mostra donada, a partir de quin nivell de significació es rebutja la hipòtesi nul.la?

En el nostre cas tenim $n = 100$, $\bar{x} = 5.4$, $\frac{\bar{x}-5}{2/\sqrt{100}} = 2$. La probabilitat

$$p := P\left(\frac{\bar{X}-5}{2/\sqrt{100}} \geq 2\right) = 0.02275$$

s'anomena el **p-valor** de la mostra.

És el que donen els programes informàtics quan se'ls hi demana fer un test. Llavors, fixat un nivell de significació NS la regla de decisió és:

- Si $p < NS$ llavors REBUTJEM H_0.
- Si $p > NS$ llavors ACCEPTEM H_0.

Continuació de l'exemple: Tornem a la mostra de mida $n = 10$, però suposem ara que $\bar{x} = 4.6$.
Aleshores és lògic preguntar-se si els estudiants d'aquesta comarca són pitjors. Plantejament del test:

$$\begin{cases} H_0 : & \mu \geq 5 \\ H_1 : & \mu < 5 \end{cases}$$

Comencem igual, si $\mu = 5$, $\frac{\bar{X}-5}{2/\sqrt{10}} \sim N(0,1)$. Però ara, sota H_0, és "poc" probable que $\bar{X}$ sigui "petit":

$$P\left(\frac{\bar{X}-5}{2/\sqrt{10}} \leq a\right) = NS = 0.05 \quad \xrightarrow{\text{Taula}} \quad a = -1.64.$$

Per tant,

- Regió d'acceptació: $[-1.64, \infty)$.
- Regió de rebuig: $(-\infty, -1.64)$.

El valor de l'estadístic per a la mostra que tenim és $\frac{\bar{x}-5}{2/\sqrt{10}} = 0.6325$ que cau a la regió d'acceptació. Per tant, en aquest cas acceptem H_0.

Metodologia pels Tests d'Hipòtesis:

- Plantejament:

$$\begin{cases} H_0 : \ldots \\ H_1 : \ldots (>(\text{resp. } <)) \end{cases}$$

- Estadístic i distribució:

$$T \sim \text{Llei}$$

  depenent de cada situació.

- Determinació de la regió de rebuig de H_0:

$$P(T \underset{(<)}{>} a) = NS$$

  Regió de rebuig: $R = (a, +\infty)$ (resp. $R = (-\infty, a)$).

- Càlcul t de l'estadístic T per a la mostra i conclussió:

$$\text{si } t \in R \text{ aleshores rebutjem} H_0.$$

4.2 Tests d'hipòtesis sobre els paràmetres de $X \sim N(\mu, \sigma^2)$

(a) Test sobre μ si σ^2 és coneguda

És el que hem fet a l'exemple anterior, usant que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

4.2 Tests d'hipòtesis sobre els paràmetres de $X \sim N(\mu, \sigma^2)$

(a) Test sobre μ si σ^2 és coneguda

És el que hem fet a l'exemple anterior, usant que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

(b) Test sobre μ si σ^2 és desconeguda

En aquest cas usarem que

$$\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim t(n-1).$$

Exemple: Una empresa que ven rates de laboratori ens assegura que el pes mitjà és 262 grams. Agafem una mostra de mida $n = 9$ i ens surt $\bar{x} = 256.2$ i $\hat{s} = 8.23$.

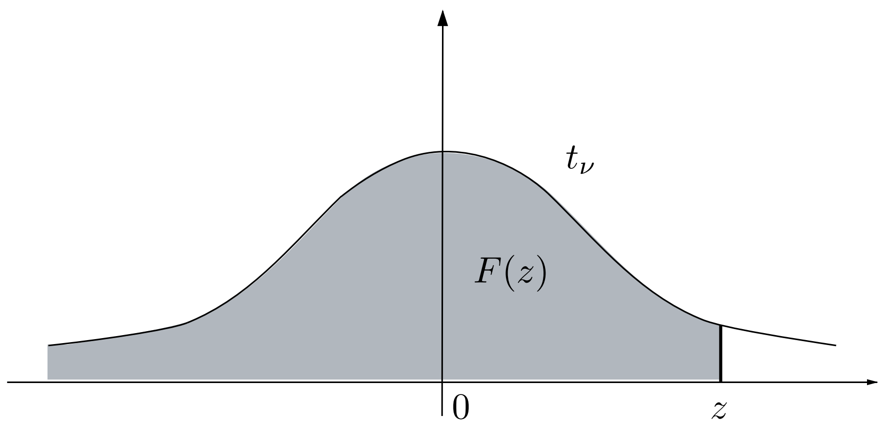
Pregunta: Ens enganya el venedor?

Resolució: $X \sim N(\mu, \sigma^2)$. Plantejament del test:

$$\begin{cases} H_0 : \quad \mu \geq 262 \text{ (no ens enganya)} \\ H_1 : \quad \mu < 262 \text{ (ens enganya)} \end{cases}$$

Agafem un nivell de significació $NS = 0.05$. Sota H_0 (amb $\mu = 262$), tenim que $\frac{\bar{X}-262}{\hat{S}/\sqrt{9}} \sim t(8)$ i

$$P\left(\frac{\bar{X} - 262}{\hat{S}/\sqrt{9}} \leq a\right) = 0.05 \quad \xrightarrow{\text{Taula}} \quad a = -1.860.$$



Per tant, la regió de rebuig és $(-\infty, -1.860)$.

El valor de l'estadístic per a la nostra mostra és $\frac{\bar{x}-262}{\hat{s}/\sqrt{9}} = -2.1142$ que cau a la regió de rebuig $(-\infty, -1.860)$. Per tant, amb una probabilitat d'equivocar-nos del 0.05 hem de concloure que ens enganya.

Pregunta: I si el nivell de significació volem que sigui 0.01?

Resposta: Aleshores $a = -2.896$ i el valor de l'estadístic -2.1142 cau dins de la regió d'acceptació $[-2.896, \infty)$ en aquest cas. Per tant, no podem afirmar que ens enganya amb una probabilitat d'equivocar-nos del 0.01.

Pregunta: Quin és el p-valor?

Resposta: Utilitzant el MINITAB,

$$p = P\left(\frac{\bar{X}-262}{\hat{S}/\sqrt{9}} \leq -2.1142\right) = 0.0337.$$

(c) Test sobre σ^2 si μ és desconeguda

En aquest cas usarem que

$$\frac{(n-1)}{\sigma^2}\hat{S}^2 \sim \chi^2(n-1).$$

Exemple: La potència d'un medicament $X \sim N(\mu, \sigma^2)$. Una mostra de mida $n = 20$ dóna $\hat{s}^2 = 0.0018$.

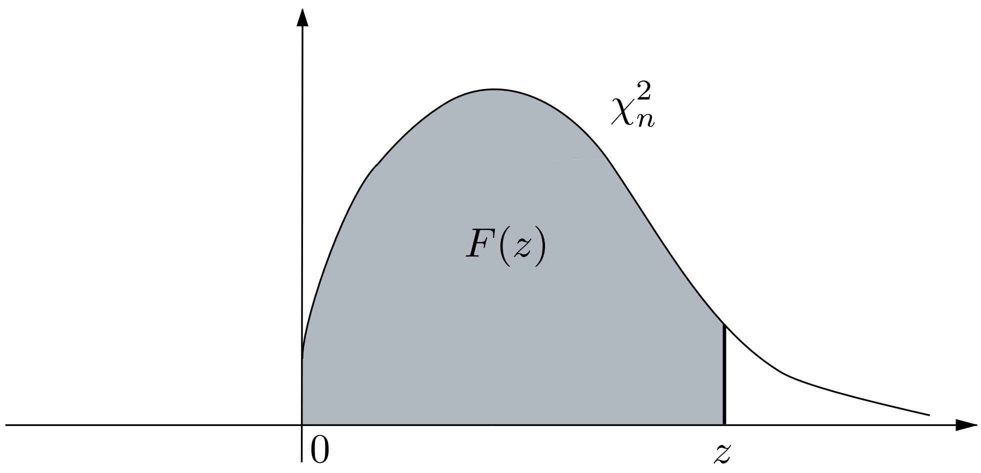
Qüestió: Volem saber si podem afirmar que $\sigma^2 < 0.004$ amb un nivell de significació de $NS = 0.10$.

Plantejament del test:

$$\begin{cases} H_0: & \sigma^2 \geq 0.004 \\ H_1: & \sigma^2 < 0.004 \end{cases}$$

Resolució: Sota H_0 (amb $\sigma^2 = 0.004$), $\frac{19}{0.004}\hat{S}^2 \sim \chi^2(19)$ és "poc" probable que sigui "petit":

$$P\left(\frac{19}{0.004}\hat{S}^2 \leq a\right) = NS = 0.10 \quad \xrightarrow{\text{Taula}} \quad a = 11.7.$$



Per tant, la regió de rebuig és $[0, 11.7)$.

El valor de l'estadístic per a la nostra mostra és $\frac{19}{0.004}\hat{s}^2 = 8.55$, que cau a la regió de rebuig. Per tant, podem afirmar que $\sigma^2 < 0.004$ amb un nivell de significació de 0.05.

Si fem els càlculs amb $NS = 0.05$ obtenim $a = 10.12$ en lloc de 11.7, però el valor de l'estadístic 8.55 continua a la nova regió de rebuig $[0, 10.12)$.

Si busquem el p-valor amb el MINITAB ens dona

$$P\left(\frac{19}{0.004}\hat{S}^2 \leq 8.55\right) = 0.0198.$$

(d) Test sobre σ^2 si μ és coneguda

En aquest cas s'utilitza que

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

Exemple anterior: Suposem que se sap $\mu = 1.5$, aleshores busquem $a = 12.4$ a la taula de $\chi^2(20)$ per tal que

$$P\left(\frac{1}{0.004}\sum_{i=1}^{20}(X_i-1.5)^2 \leq a\right) = NS = 0.10.$$

Per tant, la regió de rebuig és $[0, 12.4)$. (Una mica més gran que la que obteníem abans $[0, 11.7)$ sense utilitzar la informació $\mu = 1.5$.)

4.3 Tests de comparació de dues poblacions normals

La situació ara és que tenim:

- una mostra de mida n de $X \sim N(\mu_1, \sigma_1^2)$,
- una mostra de mida m de $Y \sim N(\mu_2, \sigma_2^2)$.

Volem fer tests del tipus

$$\begin{cases} H_0 : \mu_1 \leq \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases} \quad \text{o} \quad \begin{cases} H_0 : \sigma_1^2 \leq \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{cases}$$

Hi ha dues situacions que es tracten de manera diferent:

(1) Les dues mostres són **independents** entre sí (per exemple si provenen de poblacions independents, com ara gossos entrenats i gossos normals).

(2) Les dues mostres són de **dades aparellades** (per exemple, els mateixos gossos abans i després de entrenar-se).

4.3.1 Mostres independents

Exemple: Siguin X i Y les variables aleatòries que donen els ritmes cardíacs (en pulsacions per minut) de gossos normals i entrenats, de les quals tenim mostres de mides $n = 11$ i $m = 10$ respectivament:

X	65	70	75	72	74	73	72	78	68	73	72
Y	71	65	69	64	63	60	61	62	65	60	-

Suposem que les lleis de Xi Y són normals de paràmetres (μ_1, σ_1^2) i (μ_2, σ_2^2).

Obtenim $\bar{x} = 72$ i $\bar{y} = 64$ i per tant sospitem que $\mu_1 > \mu_2$. Volem concloure si n'hi ha prou evidència amb un nivell de significació de 0.05.

Plantejament del test:

$$\begin{cases} H_0 : \mu_1 \leq \mu_2 & (\Leftrightarrow \mu_1 - \mu_2 \leq 0) \\ H_1 : \mu_1 > \mu_2 & (\Leftrightarrow \mu_1 - \mu_2 > 0) \end{cases}$$

(a) Test sobre les μ si σ_1^2 i σ_2^2 són conegudes

A l'exemple anterior, suposem que sabem que $\sigma_1^2 = 13$ i $\sigma_2^2 = 14$.

Recordem que $\bar{X} \sim N(\mu_1, \sigma_1^2/n)$ i $\bar{Y} \sim N(\mu_2, \sigma_2^2/m)$.

Usarem que la suma o resta de variables normals independents és una altra variable normal. Tenint en compte que $Var[X - Y] = Var[X] + Var[Y]$:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right),$$

i per tant, $\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \sim N(0,1)$.

Sota H_0 (amb $\mu_1 - \mu_2 = 0$), $\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{13}{11}+\frac{14}{10}}} \sim N(0,1)$,

$$P\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{13}{11} + \frac{14}{10}}} > a\right) = 0.05 \quad \xrightarrow{\text{Taula}} \quad a = 1.64$$

Per tant, la regió de rebuig per H_0 és $(1.64, \infty)$. El valor de l'estadístic per a les nostres mostres és $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{13}{11}+\frac{14}{10}}} = 4.9788$ que cau a la regió de rebuig.

Per tant, podem concloure que els gossos entrenats tenen el pols més lent.

Variant de l'exemple anterior: Suposem ara que volem saber si $\mu_1 - \mu_2 > 6$. N'hi ha prou evidència amb el mateix nivell de significació del 5%?

Plantejament del test:

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 6 \\ H_1 : \mu_1 - \mu_2 > 6 \end{cases}$$

Resolució: Tot es fa igual, excepte que ara

$$\frac{\bar{X} - \bar{Y} - 6}{\sqrt{\frac{13}{11} + \frac{14}{10}}} \sim N(0, 1).$$

El valor de $a = 1.64$ continua sent el mateix, però ara l'estadístic dóna

$$\frac{\bar{x} - \bar{y} - 6}{\sqrt{\frac{13}{11} + \frac{14}{10}}} = 1.2447$$

que cau a la regió d'acceptació $(-\infty, 1.64]$.

Per tant, hem d'acceptar H_0 i no podem concloure que $\mu_1 > \mu_2 + 6$ amb una probabilitat d'equivocar-nos de 0.05.

(b) Test sobre les μ si σ_1^2 i σ_2^2 són desconegudes

En aquest cas, usarem que aproximadament,

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\hat{S}_X^2}{n}+\frac{\hat{S}_Y^2}{m}}} \sim t\,(N)$$

on N és el enter més pròxim al nombre següent:

$$T := \left[\frac{1}{n-1}\left(\frac{\frac{\hat{s}_x^2}{n}}{\frac{\hat{s}_x^2}{n}+\frac{\hat{s}_y^2}{m}}\right)^2+\frac{1}{m-1}\left(\frac{\frac{\hat{s}_y^2}{m}}{\frac{\hat{s}_x^2}{n}+\frac{\hat{s}_y^2}{m}}\right)^2\right]^{-1}$$

Exemple: Si tenim que $\hat{s}_x^2 = 13.3333$ i $\hat{s}_y^2 = 13.556$, llavors el valor de $T = 18.7782$ s'ha d'arrodonir a l'enter més proper $N = 19$. Per tant, en aquest cas, la llei és $t(19)$.

$$P\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\hat{S}_X^2}{11}+\frac{\hat{S}_Y^2}{10}}} > a\right) = 0.05 \quad \xrightarrow{\text{Taula}} \quad a = 1.729.$$

Per tant, la regió d'acceptació de H_0 és $(-\infty, 1.729]$ i el valor de l'estadístic dóna $\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\hat{s}_x^2}{11}+\frac{\hat{s}_y^2}{10}}} = 4.9925$ que és a la regió de rebuig de H_0.

(c) Test sobre les σ^2 si μ_1 i μ_2 són desconegudes

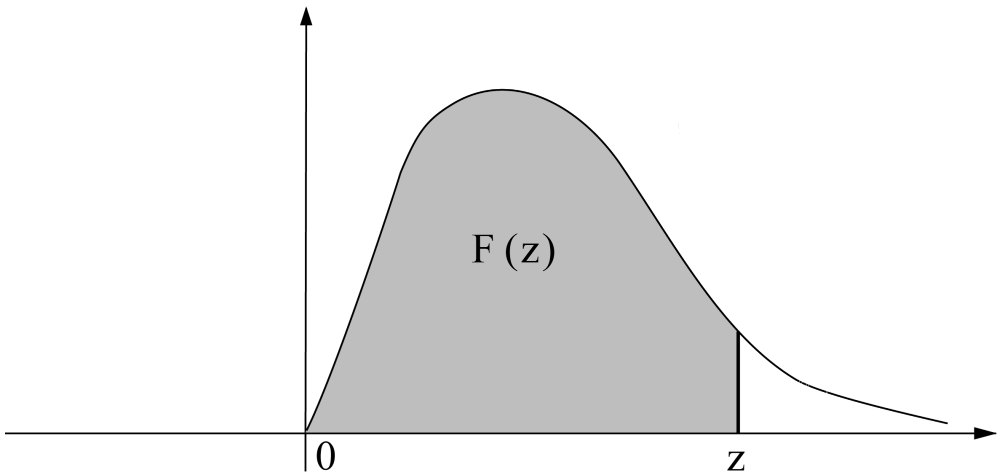
Utilitzarem que si

- $X_1, \ldots, X_n$ és una mostra de $X \sim N(\mu_1, \sigma^2)$
- $Y_1, \ldots, Y_m$ és una mostra de $Y \sim N(\mu_2, \sigma^2)$

independents entre sí, aleshores el quocient de les variàncies mostrals corregides

$$\frac{\hat{S}_X^2}{\hat{S}_Y^2} \sim F(n-1, m-1)$$

segueix una distribució F de Fisher-Snédecor amb $(n-1, m-1)$ graus de llibertat.



Exemple: Utilizant les dades del exemple del ritme cardíac dels gossos normals ($X \sim N(\mu_1, \sigma_1^2)$) i entrenats ($Y \sim N(\mu_2, \sigma_2^2)$). Suposem que volem saber si podem afirmar, amb un nivell de significació de 0.10, que $\sigma_1^2 < \sigma_2^2$.

Plantejament del test:

$$\begin{cases} H_0 : \sigma_1^2 \geq \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{cases}$$

Usarem que sota H_0 (amb $\sigma_1^2 = \sigma_2^2$), $\frac{\hat{S}_X^2}{\hat{S}_Y^2} \sim F(10, 9)$ i que és "poc" probable que l'estadístic sigui "petit":

$$P\left(\frac{\hat{S}_X^2}{\hat{S}_Y^2} \leq a\right) = 0.10 \quad \xrightarrow{\text{Taula}} \quad a = ?$$

Equivalentment, $\frac{\hat{S}_Y^2}{\hat{S}_X^2} \sim F(9, 10)$ i

$$P\left(\frac{\hat{S}_Y^2}{\hat{S}_X^2} \geq a\right) = 0.10 \quad \xrightarrow{\text{Taula}} \quad a = 2.35$$

Per tant, la regió de rebuig de H_0 per l'estadístic $\frac{\hat{S}_Y^2}{\hat{S}_X^2}$ és $(2.35, \infty)$. El valor de l'estadístic de la mostra que tenim és $\frac{\hat{s}_Y^2}{\hat{s}_X^2} = 1.0167$ que és a la regió d'acceptació.

(d) Test sobre les σ^2 si μ_1 i μ_2 són conegudes

Utilitzarem que si

- $X_1, \ldots, X_n$ és una mostra de $X \sim N(\mu_1, \sigma^2)$
- $Y_1, \ldots, Y_m$ és una mostra de $Y \sim N(\mu_2, \sigma^2)$

independents entre sí, aleshores

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_1)^2}{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \mu_2)^2} \sim F(n, m).$$

És a dir, sempre farem els càlculs sota H_0 amb

$$\sigma_1^2 = \sigma_2^2.$$

4.3.2 Dades aparellades

(a)-(b) Tests sobre les μ

Considerem

- $X_1, \ldots, X_n$ una mostra de mida n de $X \sim N(\mu_1, \sigma_1^2)$,
- $Y_1, \ldots, Y_n$ una mostra de mida n de $Y \sim N(\mu_2, \sigma_2^2)$.

Aleshores $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$ és una mostra de

$$Z = X - Y \sim N(\mu_1 - \mu_2, ?)$$

i per tant,

$$\frac{\bar{Z} - (\mu_1 - \mu_2)}{\frac{\hat{S}_Z}{\sqrt{n}}} \sim t(n-1).$$

Exemple: Considerem el ritme cardíac en pulsacions per minut d'una mostra de $n = 10$ gossos abans ($X \sim N(\mu_1, \sigma_1^2)$) i després ($Y \sim N(\mu_2, \sigma_2^2)$) d'entrenar-se:

X	65	70	75	72	74	73	72	78	68	73
Y	71	65	69	64	63	60	61	62	65	60
Z	-6	5	6	8	11	13	11	16	3	13

Sospitem que $\mu_1 > \mu_2$. Fem un test amb un nivell de significació de 0.01.

Plantejament del test:

$$\begin{cases} H_0 : \mu_1 \leq \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$$

Resolució: Sota H_0 (amb $\mu_1 = \mu_2$), $\frac{\bar{Z}}{\hat{S}_Z/\sqrt{10}} \sim t(9)$ i és "poc" probable que l'estadístic sigui "gran":

$$P\left(\frac{\bar{Z}}{\hat{S}_Z/\sqrt{10}} > a\right) = 0.01 \quad \xrightarrow{\text{Taula}} \quad a = 2.821.$$

Per tant, la regió de rebuig és $(2.821, \infty)$. El valor de l'estadístic per a la mostra és $\frac{\bar{Z}}{\hat{s}_Z/\sqrt{10}} = 3.967$ que és a la regió de rebuig de H_0. Per tant, podem afirmar que $\mu_1 > \mu_2$ amb una probabilitat d'equivocar-nos de 0.01. (Amb $NS = 0.001$, $a = 4.781$ i no rebutjaríem H_0.)

(c)-(d) Tests sobre les σ^2

En aquest cas no hi ha cap manera raonable de fer una comparació de variàncies. Hi ha tests aproximats per a mostres grans (tests de raó de versemblances asimptòtic), però tampoc els veurem.

4.4. Tests per a mostres grans ($n \geq 30$)

Valen totes les consideracions que vam fer al principi de la secció 3.3.2: Intervals de confiança per a mostres grans.

En particular, si notem $\mu = E[X]$, $\sigma^2 = Var[X]$, tenim que aproximadament

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n}).$$

4.4.1. Una sola població

(a) Test sobre μ si σ^2 és coneguda

És fa igual que en 4.2.(a) usant que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

(b) Test sobre μ si σ^2 és desconeguda

És fa igual que en 4.2.(b) usant que

$$\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \approx N(0, 1)$$

en lloc de $t(n-1)$.

4.4.2. Dues mostres independents

Siguin X, Y variables aleatòries amb qualsevol distribució i

- $X_1, \ldots, X_n$ mostra de X, amb $n \geq 30$,
- $Y_1, \ldots, Y_m$ mostra de Y, amb $m \geq 30$,

independents entre sí.

Farem el següent conveni de notació:

$$\begin{array}{lll} \mu_1 := E[X] & \qquad & \sigma_1^2 := Var[X] \\ \mu_2 := E[Y] & \qquad & \sigma_2^2 := Var[Y] \end{array}$$

(a) Test sobre les μ si les σ^2 són conegudes

Volem testar:

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases} \quad \text{o bé} \quad \begin{cases} H_0 : \mu_1 - \mu_2 \geq 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{cases}$$

Sabem:

$$\left.\begin{array}{l} \bar{X} \approx N\left(\mu_1, \frac{\sigma_1^2}{n}\right) \\ \bar{Y} \approx N\left(\mu_2, \frac{\sigma_2^2}{m}\right) \end{array}\right\} \Rightarrow \bar{X} - \bar{Y} \approx N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Per tant,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \approx N(0,1)$$

i procedim com a la secció 4.3.1. (a).

(a) Test sobre les μ si les σ^2 són conegudes

Volem testar:

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases} \quad \text{o bé} \quad \begin{cases} H_0 : \mu_1 - \mu_2 \geq 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{cases}$$

Sabem:

$$\left.\begin{array}{l} \bar{X} \approx N\left(\mu_1, \frac{\sigma_1^2}{n}\right) \\ \bar{Y} \approx N\left(\mu_2, \frac{\sigma_2^2}{m}\right) \end{array}\right\} \Rightarrow \bar{X} - \bar{Y} \approx N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Per tant,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \approx N(0, 1)$$

i procedim com a la secció 4.3.1. (a).

I si el test fos

$$\begin{cases} H_0 : \mu_1 - \mu_2 \leq 4 \\ H_1 : \mu_1 - \mu_2 > 4 \end{cases}$$

usaríem que $\frac{\bar{X} - \bar{Y} - 4}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \approx N(0, 1)$.

(b) Test sobre les μ si les σ^2 són desconegudes

S'utilitza que, sota H_0 (amb $\mu_1 - \mu_2 = 0$)

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{S}_X^2}{n} + \frac{\hat{S}_Y^2}{m}}} \approx N(0,1)$$

i es procedeix com a la secció 4.3.1.(b) però amb $N(0,1)$ en lloc de $t(\cdot)$.

4.4.3. Dades aparellades

Siguin X, Y variables aleatòries amb qualsevol distribució i

- $X_1, \ldots, X_n$ mostra de X,
- $Y_1, \ldots, Y_n$ mostra de Y,

i.e. $n = m \geq 30$.

(a)-(b) Test sobre les μ

Si definim $Z := X - Y$ aleshores $E[Z] = \mu_1 - \mu_2$ però $Var[Z]$ és desconeguda. Per tant, utilitzem que

$$\frac{\bar{Z} - (\mu_1 - \mu_2)}{\frac{\hat{S}_Z}{\sqrt{n}}} \approx N(0,1)$$

i procedim com a la secció 4.3.2. (a)-(b) però amb $N(0,1)$ en lloc de $t(n-1)$.

Observació: Si n és gran aleshores $t(n) \approx N(0,1)$ i els resultats no diferirien massa.

El cas binomial

Si $X \sim B(1,p)$, aleshores p és la mitjana poblacional (esperança) i la variància poblacional és $p(1-p)$ que queda totalment determinada per l'esperança. Això fa que, quan es fa un test sobre p, hem de considerar que sota la hipòtesi nul.la, la variància és **coneguda**.

Exemple: Tenim una població de vaques. En una mostra de mida $n = 100$ d'una certa explotació, hem trobat 2 amb una certa malaltia. La reglamentació europea diu que, quan es detecta aquest malaltia cal sacrificar tots els animals de l'explotació, llevat que hi hagi evidència estadística que la proporció de malalts és inferior al 5%, amb un nivell de significació de 0.01.

Pregunta: S'han de sacrificar les vaques en aquest cas?

Resolució: Considerem X la variable aleatòria definida per

$$X = \begin{cases} 1, & \text{si la vaca està malalta,} \\ 0, & \text{si no.} \end{cases}$$

Aleshores $X \sim Bin(1, p)$ i

$$E[X] = p, \quad Var[X] = p(1-p), \quad \bar{x} = 0.02.$$

Plantejament del test:

$$\begin{cases} H_0 : p \geq 0.05 \\ H_1 : p < 0.05 \end{cases}$$

Sota H_0 (amb $p = 0.05$), sabem que $\sum_{i=1}^{n} X_i \sim Bin(n, p) \approx N(np, np(1-p))$ i per tant

$$\bar{X} \approx N\left(p, \frac{p(1-p)}{n}\right) \Longrightarrow \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1).$$

Observem que $\frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\bar{X}-p}{\frac{\sigma}{\sqrt{n}}}$ i estem al cas (a): la variància σ^2 és coneguda.

Sota H_0 (amb $p = 0.05$ i $\sigma = \sqrt{p(1-p)} = 0.2179$),

$$P\left(\frac{\bar{X} - 0.05}{\frac{0.2179}{\sqrt{100}}} < a\right) = 0.01 \quad \xrightarrow{\text{Taula}} \quad a = -2.33.$$

Per tant, la regió de rebuig és $(-\infty, -2.33)$ i l'estadístic $\frac{\bar{x}-0.05}{\frac{0.2179}{\sqrt{100}}} = -1.3768$ cau a la regió d'acceptació. Per tant, acceptem H_0 i hauríem de sacrificar l'explotació.

Pregunta: A partir de quin nivell de significació hauríem pogut salvar les vaques? És a dir, quin és el p-valor de la mostra?

$$P(N(0,1) < 1.38) = 1 - 0.91465 = 0.08535.$$

4.5. Tests de comparació de mitjanes per a mostres petites

Veurem algunes coses que es poden fer en cas que la població no sigui normal i la mostra no és prou gran per aplicar les idees vistes anteriorment.

4.5.1. Mostres independents

Exemple: Hem fet un examen a dos grups d'estudiants (per exemple assajant dos mètodes pedagògics diferents) i les notes són:

- $X : 5.5, 3.5, 4.5, 1, 6.5, 2.5, 2, 1.5, 0.5$ (mida $n = 9$).
- $Y : 4, 10, 8.5, 3, 6, 8, 7, 5$ (mida $m = 8$),

Pregunta: Sembla que són millors les notes de la població Y. Ho podem assegurar amb un nivell de significació de 0.05?

Plantejament del test: Si posem $\mu_1 = E[X]$ i $\mu_2 = E[Y]$,

$$\begin{cases} H_0 : \mu_1 \geq \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

Ordenem les notes així:

X: 0.5; 1; 1.5; 3; 2.5		3.5		4.5		5.5		6.5	
Y:	3		4		5		6		7; 8; 8.5; 10

$$\begin{cases} H_0 : \mu_1 \geq \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

X: 0.5; 1; 1.5; 3; 2.5		3.5		4.5		5.5		6.5	
Y:	3		4		5		6		7; 8; 8.5; 10

Considerem l'estadístic $U(n,m)$ següent: quantitat de valors de X més petits que el primer valor de Y $+$ quantitat de valors de X més petits que el segon valor de Y $+ \cdots +$ quantitat de valors de X més petits que l'últim valor de Y.

Raonem que rebutjarem H_0 quan $U(n,m)$ sigui "molt gran". La llei de $U(n,m)$ està tabulada (sota la hipòtesi $\mu_1 = \mu_2$) per a valors petits de n i m (amb $n \geq m$):

$$P(U(9,8) \geq a) = 0.05 \quad \xrightarrow{\text{Taula}} \quad a = 54.$$

Per tant la regió de rebuig és $(54, \infty)$.

El valor concret de l'estadístic per a la nostra mostra és

$$5 + 6 + 7 + 8 + 9 + 9 + 9 + 9 = 62$$

que cau a la regió de rebuig. Per tant, deduïm que el grup X és millor.

X: 0.5; 1; 1.5; 3; 2.5		3.5		4.5		5.5		6.5	
Y:	3		4		5		6		7; 8; 8.5; 10

Pregunta: Què passa si hi ha empats?

- Dins d'una mateixa mostra no té importància.
- Si hi ha empats entre les dues mostres podem:
  - trencar tots els empats de manera que el test ens quedi conservador (per exemple, si el 3.5 fos 3 ho ordenaríem exactament com ho hem fet);
  - trencar el primer així, i després anar alternant.

4.5.2. Dades aparellades

Exemple: Es vol comparar el rendiment de dos fertilitzants diferents per al cultiu d'un cert producte. S'agafen 10 camps de cultiu que poden estar allunyats i ser diferents quant a característiques del sòl, etc. Cada camp es divideix en dues parcel.les i es posa un fertilitzant diferent en cadascuna. S'obtenen els resultats següents:

X	46	110	70	54	60	120	82	76	37	28
Y	42	87	75	50	48	108	80	67	40	25
T	1	1	0	1	1	1	1	1	0	1

on $T_i = 1$ si $X_i > Y_i$ i $T_i = 0$ si $X_i < Y_i$. En cas de empat $(X_i = Y_i)$ treiem aquest valor de les mostres.

Sembla que X és millor que Y perquè guanya en 8 dels 10 camps. Volem saber si ho podem afirmar amb diversos nivells de significació. Tenim que $T \sim Bin(1, p)$

Plantejament del test:

$$\begin{cases} H_0 : p \leq \frac{1}{2} \text{ (i.e. } X \text{ no és millor que } Y) \\ H_1 : p > \frac{1}{2} \text{ (i.e. } X \text{ és millor que } Y) \end{cases}$$

En el fons estem realitzant un test sobre el paràmetre d'una binomial. Ja n'hem fet un però era per a mostres grans i utilitzàvem l'aproximació per la normal. Ara no podem fer-ho així.

Sota H_0 (amb $p = \frac{1}{2}$), $S := \sum_{i=1}^{10} T_i \sim Bin(10, \frac{1}{2})$ és "poc" probable que sigui "gran":

$$P(S \geq a) = NS$$

En lloc de fixar el nivell de significació NS i buscar a, anem a trobar el p-valor:

$$\begin{aligned} P(S \geq 8) &= P(S = 8) + P(S = 9) + P(S = 10) \\ &= \left[\binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right]\left(\frac{1}{2}\right)^{10} \\ &= 0.0439 + 0.0098 + 0.0010 \\ &= 0.0547 \end{aligned}$$

- Si $NS < 0.0547$ acceptem H_0.
- Si $NS > 0.0547$ rebutjem H_0.

4.6. Tests d'ajustament i independència

Hem vist tests d'hipòtesis per decidir si el valor d'un paràmetre estava en un lloc o en un altre.

Típicament,

$$\begin{cases} H_0 : \mu \leq 5 \\ H_1 : \mu > 5 \end{cases}$$

i coses per l'estil.

També podem fer tests per saber si dues variables (poblacions) són independents o no, o si una població segueix o no una llei concreta.

Exemple: Volem saber si un dau és equilibrat o no. Si X és la variable que representa el resultat de llançar el dau, llavors "dau equilibrat" vol dir

$$P(X = i) = \frac{1}{6}, \qquad i = 1, \ldots, 6.$$

Llancem n vegades el dau i considerem el nombre n_i de vegades que ha sortit i. Si n és gran i el dau és equilibrat, resulta que

$$\sum_{i=1}^{6} \frac{(n_i - n\frac{1}{6})^2}{n\frac{1}{6}} \approx \chi^2(5).$$

Tests d'ajustament

Plantejament del test:

$$\begin{cases} H_0 : \text{Dau equilibrat} \\ H_1 : \text{Dau no equilibrat} \end{cases}$$

Un valor gran de l'estadístic donarà suport a la hipòtesi alternativa. Imposem que sota H_0,

$$P\left(\sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} > a\right) = NS = 0.05 \xrightarrow{\text{Taula}} a = 11.1$$

Per tant, la regió de rebuig és $(11.1, \infty)$. Suposem $n = 100$, $n_1 = 20$, $n_2 = 24$, $n_3 = 15$, $n_4 = 21$, $n_5 = 12$, $n_6 = 8$. El valor de l'estadístic és

$$\frac{(20 - 100\frac{1}{6})^2}{100\frac{1}{6}} + \cdots + \frac{(8 - 100\frac{1}{6})^2}{100\frac{1}{6}} = 11$$

que cau a la regió d'acceptació. Per tant, acceptem que el dau és equilibrat (per poc). Si pugem una mica el nivell de significació NS segur que rebutjem.

En general, aquest test s'utilitza sobre variables discretes X que prenen un nombre finit de valors $a_1, \ldots, a_m$, posant com a H_0 una llei concreta de la variable X:

$$H_0 : P(X = a_i) = p_i, \quad i = 1, \ldots, m;$$

usant que sota H_0,

$$\sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} \approx \chi^2(m-1),$$

on l'aproximació es considera bona quan $np_i \geq 5$ per tot $i = 1, \ldots, m$.

Observacions:

- És el primer cop que posem una hipòtesi nul.la tan concreta. Acceptar-la vol dir que no tenim prou evidència en contra seva; això no vol dir exactament que sigui certa. Per exemple, si canviem "una mica" els $\frac{1}{6}$ d'abans és possible que el test segueixi donant l'acceptació de H_0.

Observacions (continuació):

- Per exemple si posem

$$H_0 : p_1 = \frac{20}{100},\ p_2 = \frac{24}{100},\ p_3 = \frac{15}{100},\ p_4 = \frac{21}{100},\ p_5 = \frac{12}{100},\ p_6 = \frac{8}{100}$$

  l'estadístic dóna **zero** i per tant, és segur que acceptem H_0. Però quina és la llei que realment segueix el dau? No ho sabem.

- En canvi, si el resultat del test és rebutjar H_0, aleshores sí que podem estar bastant segurs (amb una probabilitat d'equivocar-nos de $NS = 0.05$).

- Si volem que la hipòtesi de ser equilibrat sigui la alternativa (per exemple, als tests de bioequivalència), la situació és molt més difícil i no la tractarem.

- Aquest mètode també es pot aplicar al cas de variables contínues fent-ne una discretització mitjançant un histograma, però tampoc tractarem aquesta situació.

Exemple:

Es creu que el 20% d'una població té una certa malaltia genètica. Nosaltres pensem que el percentatge és més alt, perquè en una mostra de mida $n = 50$ tenim 14 malalts. Hi ha prou evidència, amb un 0.10 de nivell de significació?

Solució 1: Test d'ajustament. Sigui p_1 la probabilitat de tenir la malaltia i $p_2 = 1 - p_1$.

$$\begin{cases} H_0 : p_1 = \frac{20}{100},\ p_2 = \frac{80}{100} \\ H_1 : p_1 \neq \frac{20}{100},\ p_2 \neq \frac{80}{100} \end{cases}$$

Sota H_0,

$$P\left(\sum_{i=1}^{2} \frac{(n_i - 50p_i)^2}{50p_i} \geq a\right) = 0.10 \xrightarrow{\text{Taula}} a = 2.71$$

i per tant, la regió de rebuig és $(2.71, \infty)$. El valor de l'estadístic és

$$\frac{(14-10)^2}{10} + \frac{(36-40)^2}{40} = 2$$

que cau a la regió d'acceptació. No hi ha prou evidència.

Solució 2: Test de comparació, per a una mostra gran, del paràmetre p d'una variable binomial $Bin(1,p)$:

$$\begin{cases} H_0 : p \leq \frac{20}{100} \\ H_1 : p > \frac{20}{100} \end{cases}$$

Sota H_0 amb $p = \frac{20}{100}$, la variància queda fixada $\sigma^2 = \frac{20}{100}\frac{80}{100} = \frac{4}{25}$. Per tant, $\frac{\bar{X}-\frac{20}{100}}{\frac{2/5}{\sqrt{50}}} \approx N(0,1)$ i

$$P\left(\frac{\bar{X}-\frac{20}{100}}{\frac{2/5}{\sqrt{50}}} \geq a\right) = 0.10 \quad \xrightarrow{\text{Taula}} \quad a = 1.28$$

La regió de rebuig és $(1.28, \infty)$ i el valor de l'estadístic és $\frac{\bar{x}-\frac{20}{100}}{\frac{2}{5\sqrt{50}}} = 1.4142$ que cau a la regió de rebuig. Per tant, en aquest cas rebutjem H_0.

Aquesta diferència entre els resultats és conseqüència dels plantejaments: ajustament versus comparació.

En aquest cas, no és adequat fer-ho de la primera manera. És millor la segona.

Tests d'independència

Anem ara a aplicar idees semblants per fer tests d'independència entre dues variables discretes que prenen un nombre finit de valors.

Exemple: Hi ha relació entre el color dels ulls i el grau de miopia? En una mostra de $n = 500$ persones s'ha obtingut la taula següent:

Miopia i\Color j	1. Blau	2. Verd	3. Marró	Total i
1. Gens	60	88	154	302
2. Poca	27	46	75	148
3. Molta	11	15	24	50
Total j	98	149	253	500

Suposarem que no hi ha relació, i mirarem si hi ha prou evidència en contra:

$$\begin{cases} H_0 : \text{Color i miopia són independents} \\ H_1 : \text{Color i miopia NO són independents} \end{cases}$$

Considerem:

- n_{ij} = el nombre d'individus a la casella del color i i grau de miopia j;
- n_i = el total d'individus amb color i;
- n_j = el total d'individus amb grau de miopia j.

Sota H_0 (independència),

$$T := \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{\left(n_{ij}-\frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}} \approx \chi^2(4),$$

$P(\{\text{Color } i\} \cap \{\text{Miopia } j\}) = P(\{\text{Color } i\}) \cdot P(\{\text{Miopia } j\})$.

Com que les freqüències relatives aproximen les probabilitats,

$$\frac{n_{ij}}{n} \approx \frac{n_i}{n} \cdot \frac{n_j}{n}$$

sota H_0, $n_{ij} - \frac{n_i n_j}{n}$ i T haurien de ser "petits":

$$P(T \geq a) = 0.05 \quad \xrightarrow{\text{Taula}} \quad a = 9.49$$

Per tant, la regió de rebuig és $(9.49, \infty)$ i el valor de l'estadístic

$$\frac{\left(60 - \frac{98\cdot 302}{500}\right)^2}{\frac{98\cdot 302}{500}} + \cdots + \frac{\left(24 - \frac{253\cdot 50}{500}\right)^2}{\frac{253\cdot 50}{500}} = 0.4996$$

cau a la regió d'acceptació. Per tant, acceptem la independència.

Tests d'independència

En general, tenim

- una variable X que pren valors $a_1, \ldots, a_m$,
- una variable Y que pren valors $b_1, \ldots, b_k$.

Plantejament del test:

$$\begin{cases} H_0 : X \text{ i } Y \text{ són independents} \\ H_1 : X \text{ i } Y \text{ no són independents} \end{cases}$$

Resolució:

Sota H_0,

$$T := \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}} \approx \chi^2\big((m-1)(k-1)\big),$$

$$P(T \geq a) = NS \quad \xrightarrow{\text{Taula}} \quad a$$

ens determina la regió de rebuig (a, ∞).