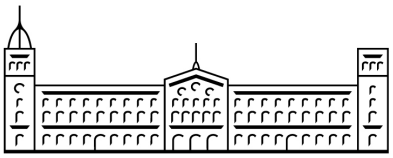

Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques
Universitat de Barcelona

REVISANT A MENDEL

Autora: Elisenda Mauri Baños

Directora: Dra. Olga Julià

Realitzat a: Departament de Probabilitat, Lògica i Estadística

Barcelona, 27 de juny de 2016

Abstract

Mendelian experiments served to prove and establish fix values that would eventually become Laws. Using the statistical and genetic resources, Mendel's Laws are proven and verified in an empirical sense as the driving force of the genetic inheritance behaviour in the 1900s.

Initially, we are presenting a theoretical introduction, which is essential to understand what comes after and to develop this thesis further, and then we proceed to explain Mendel's 1865 two-part paper *Versuche über Panzenhybriden*, therefore acquiring the required tools to analyse the data and produce a thorough study.

Adjustment test will be done taking into account Statistics and the results established by Pearson, we produce tests of adjustment to the multinomial model, comparing the hypothesis between the practical data and the theoretical obtained by Mendel. Practical data is extracted from the paper, while theoretical data is established by Genetics.

We be able to accomplish this through the R software, that allow us to get the test results and create the graphs.

My dissertation's objective is to demonstrate the veracity of Gregor Mendel's proposals through Statistics and Mathematics.

Resum

Els experiments mendelians van servir per comprovar i establir unes constants que acabarien, amb el temps, sent formulades com a Lleis. Fent ús dels recursos estadístics i genètics, les LLeis de Mendel es proven i es comproven en ple sentit empíric com el motor del comportament de l'herència genètica en el 1900.

En primer lloc exposem la part teòrica necessària per al desenvolupament del treball i després expliquem el que Mendel va deixar per escrit en el seu article *Versuche über Panzenhybriden* en el 1865, de tal manera que a partir de la teoria i amb les dades mendelianes tenim totes les eines necessàries per realitzar l'estudi.

A partir de l'Estadística i dels resultats establerts per Pearson farem tests d'ajustament del model multinomial fent contrasts d'hipòtesis entre les dades pràctiques i teòriques obtingudes per Mendel. Extraiem les dades pràctiques de l'article, mentre que les dades teòriques són les establertes per la Genètica.

Ho portem a la pràctica mitjançant el programa R, que ens permet obtenir els resultats dels tests i realitzar les gràfiques.

L'objectiu del meu treball és mostrar com a través de l'Estadística i les Matemàtiques es constata la veracitat de les propostes de Gregor Mendel.

Índex

1 Introducció

1.1 Biografia

Johann Mendel va néixer el 20 de juliol de 1822 a Heinzendorf, localitat llavors en territori de l'Imperi Austrohongarés i que en l'actualitat s'anomena Hyncice i està situada en el districte Novy Jicin de l'actual República de Txèquia. Va morir el 6 de juny de 1884. Als vint-i-un anys, el 9 d'octubre de l'any 1843, va entrar al convent dels frares agustins de Brno, en aquell moment anomenada Brünn, on va prendre el nom de pare Gregor. Va ser professor de ciències naturals a l'escola primària de Brno.

Així fou com Johann Mendel va passar a ser Gregor Mendel, que és el nom com se'l coneixerà com a naturalista.
Va passar a la posteritat a partir dels seus treballs d'hibridació de plantes, en concret experiments d'hibridació cultivant pèsols. Aquests treballs foren llegits en les reunions del 8 de febrer i del 8 de març de 1865 en l'Associació d'Història Natural de Brno i es van publicar posteriorment amb el títol: *Versuche über Plflanzenhybriden*, Experiments d'Hibridació en Plantes. Estem doncs situats en l'any 1866 quan els seus treballs són presentats a la comunitat científica internacional, sobre la que de fet no va tenir gaire, per no dir gens, impacte.

1.2 Context i desenvolupament científic

Ens podem preguntar com era doncs la comunitat científica en aquell moment de mitjans de la segona dècada de la segona meitat del segle XIX. Justament és a partir del segle XIX quan es produeix la separació definitiva entre la filosofia i les ciències exactes i experimentals, de manera que la ciència passa a ser allò que correspon a l'activitat planificada per conèixer la realitat, emprant l'anomenat mètode científic, entès com un conjunt de procediments per estudiar fenòmens, aconseguir coneixements i arribar a corroborar o modificar postulats, principis o coneixements previs. El mètode científic es basa en l'evidència empírica, observable i mesurable i ha d'establir les seves conclusions utilitzant els principis del raonament lògic.

Els procediments emprats pel mètode científic permeten obtenir un conjunt de dades a partir de l'observació i de l'experimentació i a partir d'elles arribar a formular teories. Els científics proposen hipòtesis que expliquen els fenòmens i després mitjançant l'experimentació intenten corroborar o desmentir les seves hipòtesis. Els passos de la seva experimentació han de ser repetibles, per descartar l'error o la confusió. A partir de les hipòtesis corroborades es poden formular les teories que les expliquin, les relacionin i alhora puguin ser el punt de partida de noves hipòtesis. Totes les ciències experimentals han de seguir 5 passos per arribar als coneixements vàlids que els permetin establir les seves teories: observació, formulació d'hipòtesis, experimentació, establiment de conclusions i elaboració de la teoria. Les ciències experimentals han de ser el més objectives possibles i alhora han de conservar totes les dades recollides o produïdes de manera que, emprant la metodologia utilitza-

da, sigui possible reproduir i verificar tantes vegades com es vulgui o interessi les hipòtesis plantejades i les teories establertes.

Mendel va seguir aquests passos, el seu treball s'ajusta a allò que en la seva època s'entén pròpiament com a ciència, és a dir, segueix els passos de: observació, hipòtesi, experimentació, conclusions i teoria.

1.3 Context i desenvolupament genètic

La ciència moderna de la Genètica s'originà a partir de les observacions i els treballs fets per Gregor Mendel, que es centraran a establir de manera detallada els mecanismes que regeixen l'herència en les plantes. Podem referir-nos als científics que de manera més directa van influir en el seus treballs. Aquests van ser: Karl Friedrich von Gaertner (1772-1850), Joseph Gottlieb Kölreuter (1733-1806) i Karl Wilhelm von Nägeli (1817-1891). Els tres son botànics, els dos primers alemanys i el tercer suís. Els alemanys es van distingir en els estudis dels híbrids vegetals, mentre que el suís va centrar els seus treballs en la descripció i classificació científica en botànica i va ser el descobridor dels cromosomes. En la seva obra, Mendel; cita en moltes ocasions els treballs de Gaertner. Cal dir que Nägeli i Gaertner van ser contemporanis de Mendel però mai van entendre la transcendència dels treballs mendelians. Mendel va treballar amb mongeteres i va determinar els models que regien l'herència de determinats caràcters, no tots, d'aquests vegetals, fent-ne descripcions matemàtiques i suggerint a partir dels seus treballs que l'herència era diferenciada d'un organisme a un altre i que els patrons que la regien podien ser explicats a partir de regles i ràtios bàsics. La transcendència dels seus treballs no va ser entesa en tota la seva dimensió fins a principis del segle XX, quan Mendel ja era mort, i quan altres científics que treballaven en les mateixes qüestions van redescobrir les seves investigacions. Els redescobridors de les investigacions, treballs i teories de Mendel són: William Bateson (1861-1926), Hugo Marie de Vries (1848-1935), Carl Erich Correns (1864-1933) i Erich von Tschermak Seysenegg (1871-1962). Aquests caràcters concrets o unitats bàsiques de l'herència, en que Mendel va treballar, seran denominats; gens, més endavant per Wilhelm Johannsen (1857-1927). Johannsen serà també el creador dels termes fenotip i genotip.

1.4 Redescobridors de Mendel

William Bateson fou el primer en utilitzar el terme genètica per descriure els estudis sobre l'heretabilitat i l'herència biològica. Va utilitzar per primer cop el terme en una carta personal datada del 1905 i en públic l'any següent a Londres en la tercera Conferència Internacional sobre Hibridació en Plantes, tres anys abans que Wilhelm Johannsen utilitzés el mot gen per a descriure les unitats d'informació de l'herència. Li devem a Bateson ser el popularitzador principal de les idees de Mendel. Cal recordar per això que l'adjectiu genètic ja era utilitzat amb anterioritat i va ser Charles Darwin qui el va utilitzar amb significació biològica a la seva obra: L'Origen de les Espècies.

Hugo Marie de Vries va ser un botànic neerlandès que al 1889 va fer servir per primera vegada el terme *pangen*, concepte que Mendel havia descrit molts anys abans com les unitats d'informació bàsica de l'herència. El terme *pangen* serà reemplaçat anys més tard definitivament pel terme gen de Johannsen.
Els seus treballs es van desenvolupar amb la planta denominada *oenothera lamarckiana* i el van portar a les mateixes conclusions a les que trenta anys abans havia arribat Gregor Mendel amb les mongeteres. De Vries, malgrat tenir notícies dels treballs de Mendel, no el va citar. Més tard es veuria obligat a rectificar, degut a les crítiques que pel seu oblit va rebre de Carl Correns, i reconèixer la prioritat de Mendel en l'establiment de les conclusions a les que ell, anys més tard, havia arribat.

Carl Erich Correns va ser un botànic i genetista alemany, notable sobretot pel seu descobriment independent dels principis de l'herència. Va treballar amb plantes del gènere *Hieracium* amb les que també havia treballat Mendel, tot i que no va adonar-se dels resultats ja obtinguts pel seu predecessor. Un cop reconeguda l'aportació anterior de Mendel, Correns va continuar els seus treballs per ampliar i aprofundir els treballs *mendelians*.

Erich von Tschermak-Seysenegg fou un agrònom austríac i al 1900 publica el seu treball on reconeix l'aportació prèvia i coincident amb les conclusions dels seus treballs, feta per Mendel 35 anys abans.

Els seus treballs, les relacions i els debats que es van suscitar entre ells quatre i d'altres biòlegs i científics, no només van servir per col·locar a Gregor Mendel en el lloc que li corresponia com a pare de la Genètica moderna, sinó que servirien també per concretar les seves conclusions en el que van passar a ser les **Lleis de Mendel**.

1.5 Context i desenvolupament estadístic aplicat a la genètica

Fins aquí hem vist un seguit de científics que d'una manera o altra, directa o indirectament i en el terreny de la Genètica, la Biologia, la Zoologia podem relacionar amb els treballs i les teories de Mendel; són els que es denominen *mendelians*. Hi va haver també un altre grup de científics que es van centrar en els aspectes pràctics d'aplicar els seus treballs en Matemàtiques i Estadística a altres ciències com la Biologia, la Genètica i la Zoologia; són els que es denominen *biomètrics*. *Mendelians* i *biomètrics* encetaran una polèmica que no s'acabarà fins a l'adopció de la síntesis moderna de la Teoria de l'Evolució. La polèmica es va encetar, a partir del redescobriment de l'obra de Mendel, entre els *biomètrics* Weldon i Pearson i el *mendelià* Baterson i va afectar en gran mesura els plantejaments ja establerts de la teoria de l'evolució i la importància i incidència que hi podia tenir el mètode estadístic.

Karl Pearson, més enllà de la polèmica amb els *mendelians*, va destacar per les seves grans aportacions en el camp de l'Estadística matemàtica. Va ser el creador del primer departament d'Estadística del món, a la University College de Londres. La seva investigació va destacar en l'aplicació dels mètodes estadístics a la Biologia i altres ciències. Pearson i Weldon, juntament amb Francis Galton, van ser els fundadors de la revista *Biometrika* (1901), una de les més importants en Estadística

i que encara es publica en l'actualitat. L'obra de Pearson sustenta molts dels mètodes estadístics clàssics que són avui en dia d'ús comú. Entre aquests podem mencionar: coeficient de correlació, mètode dels moments, fonaments de la teoria dels tests d'hipòtesi, p-valor i la teoria de la decisió estadística, prova de khi quadrat i anàlisi de components principals.

Ronald Aylmer Fisher fou un biòleg i estadístic que es va distingir per l'ús de les Matemàtiques per relacionar les Lleis de Mendel i la selecció natural plantejada per Charles Darwin. Aconseguiria reconciliar les Lleis de Mendel amb l'evolució gradual (Gradualisme), com l'enfocament *biomètric* podia conciliar-se amb l'enfocament *mendelià*. El seu treball serviria per avançar en una nova síntesi del Darwinisme que en l'actualitat es coneix com la síntesis evolutiva moderna. En el camp de l'Estadística, en el seu cas aplicada a la Biologia, les seves aportacions foren també molt importants. Se'l considera el pare de l'Estadística moderna. D'igual manera en el terreny de la Genètica les seves aportacions varen ser cabdals. Podem citar les següents: Anàlisis de la variància, la genètica de les poblacions aplicada als seus estudis de cultius entre d'altres.

L'interès en la Genètica i l'evolució es va suscitar en Fisher a la Universitat de Cambridge, a partir de la lectura d'un seguit d'articles de Karl Pearson. Els *darwinistes* estaven dividits entre els que creien que l'evolució funcionava a partir d'una successió gradual de petits canvis (Gradualisme) front els que creien que, al contrari, l'evolució es produeix a partir de canvis profunds en un moment puntual (Saltacionisme o Puntualisme). El corrent dominant era la dels *mendelians*, més partidaris del Gradualisme. Fisher no trigaria gaire en acceptar que el *mendelisme* era el principal mecanisme de l'herència i juntament amb Sewall Wright i J. B. S. Haldane seran els fundadors de l'anomenada Genètica de poblacions que acabarà conciliant la metodologia *biomètrica* amb la genètica *mendeliana*, primera fase de la Síntesis moderna de l'Evolució. Va fer aportacions originals en el camp de la inferència estadística així com en el disseny d'experiments que van suposar una gran ajuda als investigadors en Biologia i en Agronomia.

Veiem així doncs com els extraordinaris treballs de Gregor Mendel, incompresos i no valorats quan els va realitzar per la comunitat científica contemporània, van ser de fet una de les portes que obren el camí cap a la ciència moderna, que aconseguirà les bases en la segona meitat del segle XIX per assolir el seu estatus definitiu durant el segle XX. Serà sobretot després de les dues guerres mundials quan es podran dedicar recursos i esforços per potenciar la investigació i la ciència.

2 Variables aleatòries

En aquest capítol oferim un recull de definicions necessàries per al desenvolupament del treball. Primer començarem definint el concepte d'espai de probabilitats.

Definició 2.1. *Un espai de probabilitats és una terna* $(\Omega, \mathscr{A}, P)$ *tal que*

- Ω *és el conjunt de possibles resultats, anomenat espai mostral,*
- $\mathscr{A} \subseteq \mathscr{P}(\Omega)$ *és la família d'esdeveniments amb estructura de* σ*-àlgebra,*
- P *és una aplicació* $P\colon \mathscr{A} \longrightarrow [0,1]$ *anomenada probabilitat, que compleix:* $P(\Omega) = 1$ *i* σ*-additivitat, és a dir, per a qualssevol* $\{A_i, i \geq 1\} \subseteq \mathscr{A}$ $(A_i \cap A_j = \varnothing,$ $i \neq j)$

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Definició 2.2. *Una variable aleatòria és una funció mesurable que fa correspondre a cada element de l'espai mostral* Ω *un valor numèric. Per tant, és una aplicació* $X : \Omega \to \mathbb{R}$ *tal que*

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathscr{A}, \qquad \forall B \in \mathscr{B}(\mathbb{R}),$$

on $\mathscr{B}(\mathbb{R})$ *significa els borelians de* $\mathbb{R}$*.*

Definició 2.3. *La llei d'una variable aleàtoria* X*, la denotem per* P_X*, és la probabilitat sobre* $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ *definida per*

$$P_X(B) = P(X^{-1}(B)), \qquad \forall B \in \mathscr{B}(\mathbb{R}).$$

Una eina fonamental per treballar les lleis de les variables aleatòries és la funció de distribució.

Definició 2.4. *La funció de distribució associada a una variabla aleàtoria* X *es defineix com*

$$\begin{array}{rccl} F\colon & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & F(x) = P(X \leq x). \end{array}$$

Aquesta funció determina la llei de la variable aleatòria.

Proposició 2.1. *Tota funció de distribució* F *associada a una variable aleàtoria compleix les tres propietats següents:*

- *és creixent,*
- *és contínua per la dreta,*
- *satisfà*

$$\lim_{x\to\infty} F(x) = 1 \quad i \quad \lim_{x\to-\infty} F(x) = 0.$$

A més a més, tota funció $F : \mathbb{R} \to [0,1]$ que compleix aquestes tres propietats és la funció de distribució d'una llei de probabilitat.
Hi ha dues classes importants de variables aleatòries, segons el tipus de llei, les variables discretes i les absolutament contínues.

2.1 Variables aleatòries discretes

Definició 2.1.1. *Una variable aleàtoria X direm que és discreta si el conjunt $X(\Omega)$ és finit o numerable. Aquest conjunt el denotarem per $\mathscr{N}$ i serà representat per $\mathscr{N} = \{x_i, i \in I\}$,$I \subseteq \mathbb{N}$.*

La llei d'una variable aleàtoria discreta queda determinada pels valors

$$P_X(\{x_i\}) = P(X = x_i), \qquad \forall i \in I,$$

a l'aplicació

$$\begin{array}{rccl} p\colon & \mathscr{N} & \longrightarrow & [0,1] \\ & x_i & \longmapsto & p(x_i) = P_X(\{x_i\}), \end{array}$$

se l'anomena funció de massa de probabilitat.

Definició 2.1.2. *Una variable aleatòria discreta X amb funció de massa de probabilitat p té esperança finita si i només si*

$$\sum_{i\in I} |x_i| P(X = x_i) < \infty$$

i en aquest cas

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i).$$

2.2 Variables aleatòries absolutament contínues

Per definir les variables aleatòries absolutament contínues cal introduir el concepte de densitat.

Definició 2.2.1. *Direm que $f : \mathbb{R} \to \mathbb{R}$ és una funció de densitat si satisfà:*

- $f \geq 0$,
- *f és integrable en el sentit de Riemann en $\mathbb{R}$,*
- $\int_{-\infty}^{\infty} f(x)dx = 1$.

Definició 2.2.2. *Una variable aleatòria X és absolutament contínua si existeix una funció de densitat f tal que la seva funció de distribució F es pot escriure com*

$$F(x) = \int_{-\infty}^{x} f(y)dy, \qquad \forall x \in \mathbb{R}.$$

En aquest cas diem que la llei de X és absolutament contínua.

Definició 2.2.3. *Una variable aleàtoria X absolutament contínua amb densitat f té esperança finita si i només si*

$$\int_{-\infty}^{\infty} |x| f(x)dx < \infty$$

i en aquest cas

$$E(X) = \int_{-\infty}^{\infty} x f(x)dx.$$

2.3 Moments

Les definicions que donem en aquest apartat no depenen del tipus de variable aleatòria.

Definició 2.3.1. *Una variable aleatòria X té moment d'ordre $k > 0$, $k \in \mathbb{N}$, si X^k té esperança finita.*

Definició 2.3.2. *La variància d'una variable aleatòria X amb moment de segon ordre es defineix com*

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

L'arrel quadrada de la variància es coneix com la desviació típica

$$\sigma_x := \sqrt{Var(X)}.$$

Definició 2.3.3. *Sigui X una variable aleatòria. A la funció*

$$m_X(s) = E(e^{sX})$$

se l'anomena funció generatriu de moments de la variable aleatòria X i està definida per tots el valors de s sempre que $E(e^{sX}) < \infty$.

Observació 2.3.1. L'esperança $E(e^{sX})$ és finita almenys en un entorn del 0 i determina la probabilitat de la variable aleatòria.

2.4 Vectors aleatoris

Un vector aletaori n-dimensional és una aplicació $X\colon\ \Omega \longrightarrow \mathbb{R}$ amb $X = (X_1, \ldots, X_n)$, on cadascun dels components $X_i\colon\ \Omega \longrightarrow \mathbb{R}$, $\forall i = 1, \ldots, n$, és una variable aleatòria. Podem definir la llei d'un vector aleatori igual que hem fet per a variables aleatòries així F_X és la funció de distribució del vector X, és a dir, $F_X(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$.

2.5 Independència

Un concepte fonamental quan treballem amb variables aleatòries és el d'independència. Molts experiments consisteixen en repetir una mesura de forma independent, el que intuïtivament vol dir que els diferents resultats no s'afecten entre ells.

Definició 2.5.1. *Direm que les variables aleatòries $X_1, \ldots, X_n$ són independents si, per a tots $B_1, \ldots, B_n \in \mathscr{B}(\mathbb{R})$, es compleix*

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i).$$

Proposició 2.5.1. *Sigui $X = (X_1, \ldots, X_n)$ un vector aleatori n-dimensional. Les variables aleatòries $X_1, \ldots, X_n$ són independents, si i només si*

$$F_X(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \qquad \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Proposició 2.5.2. *Siguin X i Y dues variables aleatòries independents amb esperança finita. Aleshores, la variables aleatòria producte XY té esperança finita i es compleix:*

$$E(XY) = E(X)E(Y).$$

3 Distribució Normal

La distribució Normal també anomenada distribució gaussiana, en homenatge al matemàtic alemany Carl Friederich Gauss (1777-1855), és sense cap mena de dubte la llei més utilitzada en l'Estadística. Això és perquè apareix en molts fenòmens naturals, com en quasi totes les característiques biològiques. El nom de llei Normal es deu al fet que es pensava que era el patró natural de les distribucions.

3.1 Definicions i propietats

Definició 3.1.1. *Direm que una variable aleatòria X segueix una llei Normal o gaussiana amb paràmetres $\mu \in \mathbb{R}$ i σ^2, $\sigma > 0$, és a dir, $X \sim N(\mu, \sigma^2)$ si X té funció de densitat*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \qquad \forall x \in \mathbb{R}.$$

Propietats 3.1.1.

1. La distribució Normal queda determinada pels dos paràmetres μ i σ^2. Si $X \sim N(\mu, \sigma^2)$, aleshores:
   - $\mu = E(X)$,
   - $\sigma^2 = Var(X)$.
2. La funció de distribució de la llei Normal no té una fórmula explícita, $\forall x \in \mathbb{R}$ es té
$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy.$$
3. $f_X(\mu + x) = f_X(\mu - x)$, d'on es dedueix que la funció de densitat és simètrica respecte a μ.
4. La funció de densitat pren el valor màxim en el punt $x = \mu$ i el màxim és igual a $\frac{1}{\sqrt{2\pi\sigma^2}}$.
5. La funció generadora de moments ve determinada per $m_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

Així es té que com més petit sigui el valor de σ^2, l'àrea sota la funció de densitat està més concentrada al voltant del valor de μ.

Proposició 3.1.1. *Si $X \sim N(\mu, \sigma^2)$ i α, $\beta \in \mathbb{R}$, aleshores la variable aleatòria*

$$\alpha X + \beta \sim N(\alpha\mu + \beta, \alpha^2\sigma^2).$$

Demostració 3.1.1. *S'utilitzarà el fet que la funció generatriu de moments caracteritza la probabilitat de la variable. Es té que:*

$$\begin{gathered} m_{\alpha X+\beta}(t) = E(e^{t(\alpha X+\beta)}) = \\ E(e^{\alpha t X})e^{\beta t} = m_X(\alpha t)e^{\beta t} = \\ e^{\mu\alpha t+\sigma^2\alpha^2\frac{t^2}{2}}e^{\beta t} = e^{(\alpha\mu+\beta)t+\sigma^2\alpha^2\frac{t^2}{2}}. \end{gathered}$$

la qual és la funció generatriu de moments d'una distribució $N(\alpha\mu+\beta, \alpha^2\sigma^2)$, i el resultat queda demostrat.

Definició 3.1.2. *Si $X \sim N(\mu, \sigma^2)$ amb $\mu = 0$ i $\sigma = 1$ aleshores es diu que X té distribució Normal estàndard, la notem per Z i té funció de densitat*

$$f_Z(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}, \quad \forall x \in \mathbb{R}.$$

Donada una variable aleatòria amb llei Normal, la podem convertir en una variable aleatòria amb llei Normal estàndard per mitjà d'una transformació lineal que és fonamental en la teoria relacionada amb les distribucions normals. Utilitzant la proposició anterior, cal trobar els valors α i β pels quals la variable transformada tingui distribució Normal estàndard. Equival a solucionar les següents igualtats per α i β:

$$\begin{cases} \alpha\mu + \beta = 0 \\ \alpha^2\sigma^2 = 1 \end{cases}$$

d'on s'obtenen dues solucions $\alpha_1 = \frac{1}{\sigma}$, $\beta_1 = -\frac{\mu}{\sigma}$ i $\alpha_2 = -\frac{1}{\sigma}$, $\beta_2 = \frac{\mu}{\sigma}$. D'aquesta manera, hem trobat dues variables $Z_1 = \frac{X-\mu}{\sigma}$ i $Z_2 = \frac{\mu-X}{\sigma}$ amb distribució Normal estàndard. Normalment s'utilitza la transformació donada per la primera solució, i és coneguda com l'estandarització, i la variable $Z = Z_1 = \frac{X-\mu}{\sigma}$ es coneix com la variable X estandaritzada.

Exemple 3.1.1. Donem vàries gràfiques de la corresponent funció de densitat d'una $\mathrm{N}(\mu, \sigma^2)$. Utilitzem les comandes de R que podem veure en l'annex.

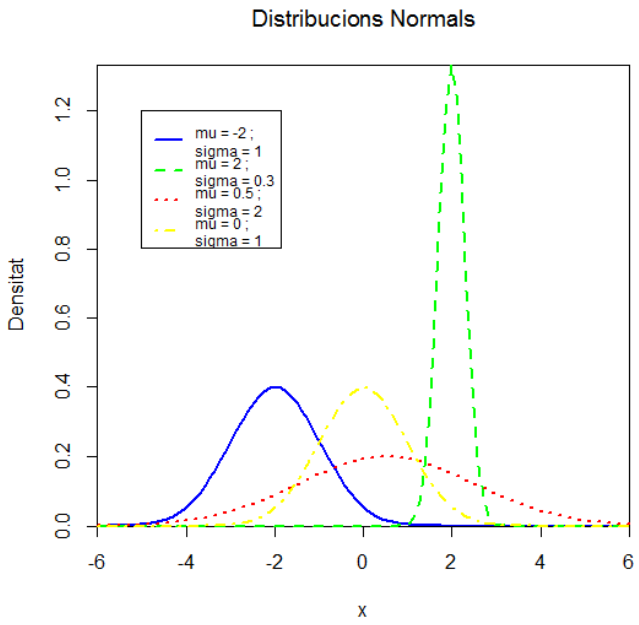

Figura 1: Gràfiques $\mathrm{N}(\mu, \sigma^2)$.

Un dels avantatges de la distribució Normal és que la suma de variables aleatòries normals sempre segueix una llei Normal, exceptuant el cas degenerat. En particular:

Proposició 3.1.2. *Siguin $X_1, \ldots, X_n$ variables aleatòries independents, on $X_i \sim N(\mu_i, \sigma_i^2)$ $\forall i = 1, \ldots, n$ aleshores la variable*

$$\sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2).$$

Proposició 3.1.3. *Siguin $X_1, \ldots, X_n$ variables aleatòries, on $X_i \sim N(\mu, \sigma^2)$ $\forall i = 1, \ldots, n$, aleshores la mitjana empírica té llei Normal*

$$\bar{X}_n \sim N(\mu, \sigma^2/n).$$

Aquest resultat és molt interessant i ens servirà per a poder construir els intervals de confiança que veurem més endavant.

3.2 Teorema del Límit Central

El Teorema del Límit Central que enunciem a continuació ens serà útil per estudiar el comportament de la suma de n variables aleatòries independents i idènticament distribuïdes per a valors de n grans independentment de la distribució de les variables aleatòries sempre i quan existeixi el moment de segon ordre, $E(X^2) < \infty$.

Teorema 3.2.0.1. ***Teorema del Límit Central*** *Sigui $X_1, \ldots, X_n$ una successió de variables aleatòries d'una distribució amb esperança μ finita i variància σ^2, $0 < \sigma^2 < \infty$, aleshores*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{\mathscr{L}} N(0,1)$$

On la convergència en Llei es refereix a la convergència de les funcions de distribució. A la pràctica vol dir que, per n prou gran (en molts casos $n > 30$ és suficient),

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

Per tant, distingim entre els dos casos següents:

- **Cas de σ coneguda**

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$$

- **Cas de σ desconeguda** Cas en què el valor de σ no és conegut, no podem utilitzar els resultats anteriors com hem fet abans, ja que ens apareixeria σ i no sabem quant val. Aleshores el que farem serà substituir el valor de σ per una estimació. L'estimació que farem servir és la desviació estàndard mostral

$$S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$$

  Així es demostra que:

$$\frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}} \sim t_{n-1}$$

Observem que ja no és una llei Normal sinó una llei t-Student amb $n-1$ graus de llibertat, però no aprofundirzarem més ja que no tenim necessitat per a aquest treball.

4 Distribució χ^2

Karl Pearson a partir del 1890 va desenvolupar els mètodes estadístics dirigits a l'anàlisi de qüestions biològiques, la majoria d'ells editats en la revista *Biometrika*, entre ells la distribució χ^2. És una de les distribucions més utilitzades en tots els camps, un dels seus usos més comuns és per provar si unes mesures obtingudes segueixen una distribució esperada.

4.1 Definicions i propietats

Definició 4.1.1. χ^2 ***de Pearson*** *Una variable aleatòria X té distribució khi quadrat amb n graus de llibertat, $X \sim \chi^2_n$ amb $n \in \mathbb{Z}^+$, si té com a funció de densitat*

$$f_X(x) = \frac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}, \qquad \forall x \in (0,\infty)$$

on $\Gamma(\cdot)$ és la funció Gamma. Recordem:

- $\forall a > 0$ es té $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$,
- $\forall n \in \mathbb{N}$ es té $\Gamma(n) = (n-1)!$.

Aquesta distribució és un cas particular de la distribució Gamma, però està íntimament relacionada amb la distribució Normal, tal i com es mostra en el següent resultat.

Proposició 4.1.1. *Si $Z_1, \ldots, Z_n$ són variables aleatòries independents i idènticament distribuïdes amb llei Normal estàndard, aleshores la variable $\sum_{i=1}^{n} Z_i^2$ té distribució χ^2_n amb n graus de llibertat.*

$$Z_1^2 + \ldots + Z_n^2 \sim \chi^2_n$$

Propietats 4.1.1. *Si X és una variable aleatòria amb distribució χ^2_n amb n graus de llibertat, aleshores:*

- $E(X) = n$,
- $Var(X) = 2n$,
- *no és simètrica,*
- *només pren valors positius i*
- $m_X(t) = (\frac{1}{1-2t})^{n/2}$, *per a $t < 1/2$, per altres valors de t no existeix.*

Proposició 4.1.2. *Siguin $X_1, \ldots, X_m$ variables aleatòries independents amb distribució $\chi^2_{n_i}$ per $i = 1, \ldots, m$ aleshores la variable $X = \sum_{i=1}^{m} X_i \sim \chi^2_k$ amb $k = \sum_{i=1}^{m} n_i$ graus de llibertat.*

Demostració 4.1.1. *Utilitzant la funció generatriu de moments i la independència de les variables aleatòries:*

$$m_X(t) = E(\exp(t\sum_{i=1}^{m} X_i)) = \prod_{i=1}^{m} E(e^{tX_i})$$

$$= \prod_{i=1}^{m} m_{X_i}(t) = \prod_{i=1}^{m} \left(\frac{1}{1-2t}\right)^{n_i/2} = \left(\frac{1}{1-2t}\right)^{\sum_{i=1}^{m} n_i/2}$$

la qual correspon a la funció generatriu de moments d'una distribució χ^2_k *amb* $k = \sum_{i=1}^{m} n_i$ *graus de llibertat, i el resultat queda demostrat.*

L'anterior resultat estableix que la suma de variables independents amb distribució khi quadrat segueix tenint distribució khi quadrat.

Exemple 4.1.1. Donem gràfiques de la funció de densitat de la distribució χ^2_n amb diferents graus de llibertat. Les comandes de R es poden trobar a l'annex.

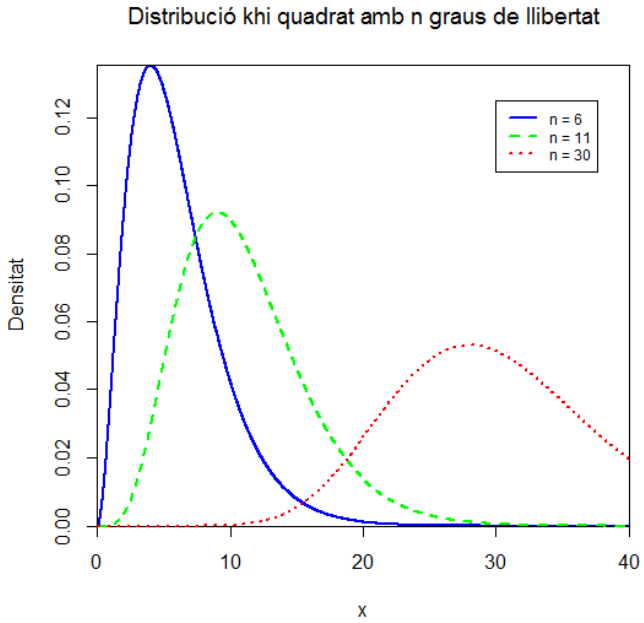

Figura 2: Gràfiques χ^2_n.

5 Intervals de confiança

5.1 Intervals de confiança per a una població Normal

Introduïm el càlcul d'intervals de confiança. L'objectiu és construir a partir de la nostra mostra un interval amb una probabilitat bastant alta anomenada nivell de confiança, de manera que el valor real del paràmetre que estem estimant quedi dins el nostre interval.

Suposarem que tenim una mostra aleatòria $X_1, \ldots, X_n$ amb distribució $N(\mu, \sigma^2)$.

5.1.1 Intervals de confiança per a μ, on σ és coneguda

Es tracta de construir interval de confiança per a μ suposant que el valor de σ és conegut amb un nivell de confiança γ. Els passos a seguir per calcular-lo són:

- Utilitzant que

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

- Calculem el valor crític u_γ tal que

$$P(-u_\gamma \leq Z \leq u_\gamma) = \gamma \Rightarrow P(Z \leq u_\gamma) = 1 - \frac{1-\gamma}{2}$$

  Per trobar u_γ en R fariem: $u_\gamma = \text{qnorm}\left(1 - \frac{1-\gamma}{2}\right)$.

- L'interval serà:

$$\left[\bar{X}_n - u_\gamma \frac{\sigma}{\sqrt{n}} \ , \ \bar{X}_n + u_\gamma \frac{\sigma}{\sqrt{n}}\right].$$

5.1.2 Intervals de confiança per a μ, on σ és desconeguda

Es tracta de construir un interval de confiança per a μ suposant que el valor de σ és desconegut amb un nivell de confiança γ. Els passos a seguir per calcular-lo són:

- Utilitzant la desviació estàndard mostral en lloc de σ; en aquest cas ens queda que

$$T_{n-1} = \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}} \sim t_{n-1}$$

  on recordem que $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

- Calculem el valor crític $t_{n-1,\gamma}$ tal que

$$P(-t_{n-1,\gamma} \leq T_{n-1} \leq t_{n-1,\gamma}) = \gamma \Rightarrow P(T_{n-1} \leq t_{n-1,\gamma}) = 1 - \frac{1-\gamma}{2}$$

  Per trobar $t_{n-1,\gamma}$ en R fariem: $t_{n-1,\gamma} = \text{qt}\left(1 - \frac{1-\gamma}{2}, n-1\right)$.

- L'interval serà:

$$\left[\bar{X}_n - t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}} \ , \ \bar{X}_n + t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}}\right].$$

5.1.3 Interval de confiança per a la variància σ^2

Per construir un interval de confiança per a σ^2 amb un nivell de confiança γ, utilitzarem un dels resultats anteriors: si $Z_1, \ldots, Z_n$ són variables aleatòries independents i idènticaments distribuïdes amb distribució Normal estàndard, aleshores

$$Z_1^2 + \ldots + Z_n^2 \sim \chi_n^2.$$

Proposició 5.1.3.1. *Siguin $X_1, \ldots, X_n$ variables aleatòries amb $X_i \sim N(\mu, \sigma^2)$ aleshores*

$$\frac{nS_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Ara, doncs, els passos que seguim per a calcular un interval de confiança per a σ^2 són:

- Utilitzant el resultat de la proposició anterior tenim

$$\frac{nS_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

- Calculem el valors crítics $u_{n-1,\gamma}$ i $v_{n-1,\gamma}$ tals que

$$P(u_{n-1,\gamma} \leq \chi_{n-1}^2 \leq v_{n-1,\gamma}) = \gamma$$

  Si ho volem fer en R:

$$u_{n-1,\gamma} = \text{qchisq}\left(\frac{1-\gamma}{2}, n-1\right) \text{ i } v_{n-1,\gamma} = \text{qchisq}\left(\frac{1-\gamma}{2}, n-1\right).$$

- L'interval serà

$$\left[\frac{nS_n^2}{v_{n-1,\gamma}} \ , \ \frac{nS_n^2}{u_{n-1,\gamma}}\right].$$

5.2 Intervals de confiança per una proporció

Suposem que estem interessats en la proporció en que es presenta una característica dels elements de la mostra. Tindrem doncs $X_1, \ldots, X_n$ variables aleatòries binàries independents i idènticament distribuïdes amb $P(X_i = 1) = p$, és a dir, $X_i = 1$ si l'element i té la característica i, $X_i = 0$ si no la té.

Es tracta de construir un interval de confiança per una proporció p amb un nivell de confiança γ. Els passos que seguim per calcular-lo són:

- Utilitzant el Teorema del Límit Central ja que $E(X_i) = p$ i $Var(X_i) = p(p-1)$:

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} = \frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{\mathscr{L}} N(0,1)$$

  Això ens permet aproximar $\bar{X}_n$ per una normal $N\left(p, \frac{p(1-p)}{n}\right)$. Estem en un cas similar al interval de confiança per μ.

- Calculem els valors crítics u_γ tal que

$$P(-u_\gamma \leq Z \leq u_\gamma) = \gamma \Rightarrow P(Z \leq u_\gamma) = 1 - \frac{1-\gamma}{2};$$

  en aquest cas, en R fariem: $u_\gamma = qnorm\left(1 - \frac{1-\gamma}{2}\right)$.

- L'interval serà:

$$\left[\bar{X}_n - u_\gamma \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \ , \ \bar{X}_n + u_\gamma \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

6 Teoria general dels tests d'hipòtesis

En tot procés científic, l'estudi, el tractament i la manipulació de les observacions empíriques porten al investigador a formular un model o teoria que expliqui el fenomen observat. Els tests d'hipòtesis ens permenten contestar a la pregunta de si la mostra o conjunt d'observacions corrobora el model o la teoria.

La idea dels tests d'hipòtesis consisteix en formular una certa hipòtesi sobre els paràmetres de la distribució i utilitzar les observacions que tenim, que usualment és l'única informació de la què disposem, per decidir entre la hipòtesi que hem formulat o la seva complementària.

6.1 Definicions

Definició 6.1.1. *Considerem un model estadístic paramètric* $(\Omega, \mathscr{A}, \{P_\theta, \theta \in \Theta\})$. *Plantejar una hipòtesi sobre el paràmetre és equivalent a fer una partició de l'espai de paràmetres*

$$\Theta = \Theta_0 \cup \Theta_1 \ , \ \Theta_0 \cap \Theta_1 = \varnothing$$

i demanar-se si θ *pertany o no a* Θ_0.

Habitualment, l'afirmació o conjectura que volem confirmar, és a dir, la qüestió que ens estem plantejant, serà la hipòtesi alternativa, i l'altra serà la nul·la. Per tant, anomenarem hipòtesi nul·la

$$H_0 = \theta \in \Theta_0$$

i hipòtesi alternativa, la complementària

$$H_1 = \theta \in \Theta_1 = \Theta \setminus \Theta_0.$$

Quan un dels dos conjunts, Θ_0 *o* Θ_1, *consta d'un sol element direm que la hipòtesi corresponent és simple. Altrament, direm que és composta.*

Definició 6.1.2. *La* ***hipòtesi nul·la****, que anomenarem* H_0, *és la hipòtesi conservadora. Serà la hipòtesi que mantindrem, llevat que les dades ens indiquin el contrari. L'objectiu del test és rebutjar o no aquesta hipòtesi.*

Definició 6.1.3. *La* ***hipòtesi alternativa*** *que anomenarem* H_1 *és la complementària de* H_0 *i serà amb la que ens quedarem si rebutgem* H_0.

6.2 Errors i nivell de significació

La hipòtesi nul·la i alternativa defineixen un problema de test d'hipòtesis. Ara s'ha de fixar una regla de decisió tal que a partir de les observacions puguem decidir si acceptem o rebutgem H_0. Així un test d'hipòtesis φ ve donat per una partició del conjunt Ω,

$$\Omega = A_0 \cup A_1 \ , \ A_0 \cap A_1 = \varnothing, \quad A_0 \in \mathscr{A},$$

tal que

$$\begin{array}{rccc} \varphi\colon & \Omega & \longrightarrow & \{0,1\} \\ & x & \longmapsto & \mathbf{1}_{A_1}(x). \end{array}$$

La regla de decisió que en resulta és la següent:

- si les observacions $x \in A_0$, és a dir $\varphi(x) = 0$, acceptarem H_0,
- si les observacions $x \in A_1$, és a dir $\varphi(x) = 1$, rebutjarem H_0.

D'acord amb aquesta regla de decisió, anomenarem regió d'acceptació A_0 i regió crítica o de rebuig A_1.

Hem de resoldre el problema de com determinar la regió d'acceptació, és a dir, necessitem algun tipus d'eina que ens permeti escollir una regió d'acceptació i no altres. Per això treballarem amb les probabilitats de prendre decisions incorrectes que comporta cada regla de decisió. Quan prenguem una decisió podem cometre dos tipus d'error:

- podem rebutjar la hipòtesi nul·la malgrat que sigui certa,
- podem acceptar la hipòtesi nul·la malgrat que sigui falsa.

Definirem així:

- L'error de tipus I, consisteix en la probabilitat de rebutjar H_0 tot i ser certa i que definirem com

$$\alpha_1(\theta) = P_\theta(A_1) = E_\theta[\varphi(x)], \quad \forall \theta \in \Theta_0.$$

- L'error de tipus II, consisteix en la probabilitat d'acceptar H_0 tot i ser falsa i que definirem com

$$\alpha_2(\theta) = P_\theta(A_0) = 1 - E_\theta[\varphi(x)], \quad \forall \theta \in \Theta_1.$$

La taula següent representa les diferents possibilitats que tenim:

	Acceptar H_0	Rebutjar H_0
H_0 certa	Decisió correcta	Error tipus I
H_0 falsa	Error de tipus II	Decisió correcta

Aquests dos errors poden expressar-se en termes de la funció de la potència

$$\begin{array}{rccl} \Phi\colon & \Theta & \longrightarrow & [0,1] \\ & \theta & \longmapsto & E_\theta[\varphi(x)] = P_\theta(A_1) \end{array}$$

i d'aquesta manera

$$\begin{aligned} \alpha_1(\theta) &= \Phi(\theta), \quad \theta \in \Theta_0, \\ \alpha_2(\theta) &= 1 - \Phi(\theta), \quad \theta \in \Theta_1. \end{aligned}$$

La restricció de Φ sobre $\theta \in \Theta_1$ s'anomena potència del test φ i representa la probabilitat de rebutjar la hipòtesi nul·la quan és falsa, ja que el valor correcte del paràmetre és $\theta \in \Theta_1$.

Observació 6.2.1. El **nivell de significació α** fixat intenta garantir que sigui molt poc freqüent rebutjar una hipòtesi nul·la correcta. En canvi, la preocupació perquè acceptem una H_0 falsa és menor. Així, quan el resultat d'un test ens diu que hem d'acceptar H_0, ho hem d'interpretar en el sentit que les observacions no ens han donat prou evidències per rebutjar-la, mentre que si la rebutgem és força segur que H_0 sigui falsa i per tant la hipòtesi alternativa certa. Tot això fa que els valors del nivell de significació més utilitzats siguin $\alpha = 0.1$; 0.05 i 0.01, essent $\alpha = 0.05$ el més famós de tots.

L'error màxim de tipus I s'anomena nivell de significació. Sigui $\alpha \in (0,1)$, direm que el test φ té nivell de significació α si

$$\sup_{\theta \in \Theta_0} \alpha_1(\theta) \leq \alpha.$$

Òbviament ens interessa prendre decisions que minimitzin els errors de tipus I i II. Ara bé, en general, no serà possible controlar simultàniament tots dos errors, ja que per disminuir el I ens cal augmentar A_0, o disminuir A_1, i això comporta habitualment l'augment de l'error de tipus II.

Com s'ha dit anteriorment, hem de fixar una estratègia per determinar un test que controli els dos erros. La proposta d'estratègia que presentem està basada en les idees desenvolupades per Neyman i Pearson als anys trenta del segle XX i podem resumir-la com segueix: dels dos errors, el de tipus I serà el que tindrà pitjor conseqüències, i entre totes les decisions possibles, considerarem únicament les que tinguin aquest error fitat per una quantitat fixada a priori; després, dins d'aquesta classe de tests amb error de tipus I fitat, provarem de trobar el que minimitzi l'error de tipus II (i maximitzi la potència del test).

Designarem $\mathscr{D}_\alpha$ el conjunt de tots els tests de nivell de significació α. Alesho-res, donats dos tests φ i φ' de $\mathscr{D}_\alpha$ direm que φ és més potent que φ' si

$$\Phi_{\varphi'}(\theta) \leq \Phi_\varphi(\theta), \quad \forall \theta \in \Theta_1.$$

Finalment, fixat un nivell de significació α, direm que un test de la classe $\mathscr{D}_\alpha$ és un test uniformement de màxima potència (UMP) si és més potent que qualsevol altre test de $\mathscr{D}_\alpha$.

Observació 6.2.2. En aquesta estratègia les dues hipòtesis no són tractades de la mateixa manera i, per tant, el tractament no és simètric. Les dues hipòtesis no són intercanviables.

Observació 6.2.3. Els tests d'hipòtesis solen ser conservadors a favor de H_0.

6.3 Funció de versemblança, Teorema de Neyman Pearson i p-valor

El concepte de funció de versemblança que presentem a continuació va ser introduït per Fisher i, encara que es pot presentar de forma molt general, aquí donem només la definició per als dos casos més usuals.

Definició 6.3.1. *L'aplicació* $L\colon\ \Omega\times\Theta \longrightarrow \mathbb{R}^+$ *definida de la forma:*

- *Si* P_θ *és una llei discreta:*

$$L(x,\theta) = P_\theta(x)$$

- P_θ *és una llei absolutament contínua amb densitat* f_θ *si:*

$$L(x,\theta) = f_\theta$$

Considerarem un model estadístic paramètric $(\Omega, \mathscr{A}, \{P_\theta, \theta \in \Theta\})$ amb funció de versemblança associada $L(x,\theta)$.

Un dels problemes més interessant i a la vegada més difícils de la teoria dels tests d'hipòtesis és determinar l'existència de tests UMP per a diferents tipus d'-hipòtesis. El resultat següent és conegut com a lema (fonamental) de Neyman i Pearson.

Teorema 6.3.0.1. *Fixem* $\alpha \in (0,1)$ *i assumim que existeix una constant positiva* c_α *tal que la regió* $A_0 = \{x; L(x,\theta_1) \le c_\alpha L(x,\theta_0)\}$ *satisfà* $P_{\theta_0}(A_0^c) = \alpha$. *Aleshores, el test* $\varphi(x) = \mathbf{1}_{A_0^c}$ *és UMP al nivell de significació* α *per contrastar*

$$H_0 : \theta = \theta_0 \quad \textit{contra} \quad H_1 : \theta = \theta_1.$$

Demostració 6.3.1. *Òbviament* $\varphi(x) = \mathbf{1}_{A_0^c}$ *és un test de significació* α. *Haurem de comprovar que és un test UMP. Sigui* $\varphi' \in \mathscr{D}_\alpha$ *qualsevol, això vol dir que si* B_0 *és la regió d'acceptació de* φ', $P_{\theta_0}(B_0^c) \le \alpha$. *Volem veure que*

$$P_{\theta_1}(B_0^c) \le P_{\theta_1}(A_0^c). \qquad (1)$$

Podem descompondre A_0 i B_0 en unions disjuntes de la forma:

$$A_0 = (A_0 \cap B_0^c) \uplus (A_0 \cap B_0) = A \uplus B$$

$$B_0 = (B_0 \cap A_0^c) \uplus (B_0 \cap A_0) = A' \uplus B$$

amb $A = A_0 \cap B_0^c$, $B = A_0 \cap B_0$ i $A' = B_0 \cap A_0^c$. Aleshores provar (1) és equivalent a demostrar

$$P_{\theta_1}(A') \geq P_{\theta_1}(A).$$

Per construcció de A_0 sabem que $P_{\theta_0}(A_0) = 1 - \alpha$ i com que B_0 és una regió d'acceptació de $\varphi' \in \mathscr{D}_\alpha$, tenim que $P_{\theta_0}(B_0) \geq 1 - \alpha$. Observem que

$$P_{\theta_0}(A') + P_{\theta_0}(B) = P_{\theta_0}(B_0) \geq P_{\theta_0}(A_0) = P_{\theta_0}(A) + P_{\theta_0}(B),$$

i, per tant,

$$P_{\theta_0}(A') \geq P_{\theta_0}(A). \qquad (2)$$

Ara, d'una banda, en el conjunt A' tenim que $L(x, \theta_1) > c_\alpha L(x, \theta_0)$, i suposant per exemple que P_{θ_0} i P_{θ_1} són absolutament contínues, tenim

$$P_{\theta_1}(A') = \int_{A'} L(x, \theta_1)dx \geq c_\alpha \int_{A'} L(x, \theta_0)dx = c_\alpha P_{\theta_0}(A'), \qquad (3)$$

aquesta desigualtat és vàlida també per a probabilitats discretes ja que només cal canviar les integrals per sumes. D'altra banda, en A es compleix $L(x, \theta_1) \leq c_\alpha L(x, \theta_0)$ i per tant,

$$P_{\theta_1}(A) \leq c_\alpha P_{\theta_0}(A). \qquad (4)$$

Aleshores ajuntant (2), (3) i (4) obtenim que:

$$P_{\theta_1}(A') \geq c_\alpha P_{\theta_0}(A') \geq c_\alpha P_{\theta_0}(A) \geq P_{\theta_1}(A),$$

que és el que volíem comprovar.

Definició 6.3.2. *El p-valor és la probabilitat d'obtenir un valor tan extrem com el que hem obtingut amb la nostra mostra si la hipòtesi H_0 és certa.*

A partir d'aquesta definició, podem deduir que si el p-valor és petit voldrà dir que el valor que hem observat és poc probable, i per tant apunta a que la hipòtesi nul·la és falsa i la decisió que prenem és rebutjar la hipòtesi nul·la. La conclusió és: si el p-valor és inferior al nivell de significació rebutgem H_0. La decisió que prendrem serà:

- Si p-valor$\leq \alpha$, rebutgem la hipòtesi nul·la.
- Si p-valor$> \alpha$, no tenim prou evidències per a rebutjar la hipòtesi nul·la.

6.3.1 Test i raó de versemblança

Per a les situacions què no podem utilitzar el teorema de Neyman i Pearson necessitem disposar d'algun mètode de constrast general que ens permeti construir tests amb bones propietats per realitzar constrastos per a qualsevol tipus d'hipòtesis. Suposem el model estadístic paramètric amb paràmetre desconegut θ, n observacions $x_1, \ldots, x_n$ d'una variable aleatòria $X = (X_1, \ldots, X_n)$: $\tilde{\Omega} \longrightarrow \Omega$ amb llei P_θ, i funció de versemblança $L(x, \theta)$.
Donada una partició de l'espai de paràmetres $\Theta = \Theta_0 \cup \Theta_1$, anomenarem raó de versemblança al quocient

$$\lambda(x) = \frac{L(x, \Theta_0)}{L(x, \Theta)},$$

on

$$\begin{aligned} L(x, \Theta_0) &= \sup_{\theta \in \Theta_0} L(x, \theta), \\ L(x, \Theta) &= \sup_{\theta \in \Theta} L(x, \theta). \end{aligned}$$

Fixem-nos que $\lambda(x) \in [0, 1]$.

Definició 6.3.1.1. *Anomenarem test de la raó de versemblança al nivell de significació $\alpha \in (0, 1)$ per contrastar les hipòtesis*

$$H_0 : \theta \in \Theta_0 \quad \text{contra} \quad H_1 : \theta \in \Theta_1.$$

el que té per regió d'acceptació

$$A_0 = \{x; \lambda(x) \geq c_\alpha\},$$

on c_α ($c_\alpha \leq 1$) es determina de manera que $P_\theta(A_0^c) \leq \alpha$, $\forall \theta \in \Theta_0$.

Òbviament ens interessa els que tinguin la regió A_0 tan petita com sigui possible.

Teorema 6.3.1.1. *Suposem $\Omega \subseteq \mathbb{R}^k$ i l'espai de paràmetre Θ és un interval obert de $\mathbb{R}^k$. Fixem un valor $\theta_0 \in \Theta$. Assumim que el model estadístic associat a una observació satisfà les condicions de regularitat habituals, veure per exemple* [2]. *Suposem també que per a cada $n \geq 1$ existeix una única solució de l'equació de versemblança.*
Aleshores quan contrastem

$$H_0 : \theta = \theta_0 \quad \textit{contra} \quad H_1 : \theta \neq \theta_0,$$

tenim que sota la hipòtesi nul·la

$$-2 \ln \lambda_n(x) \xrightarrow[n \to \infty]{\mathscr{L}} \chi^2_k$$

on $\lambda_n(x)$ és la raó de versemblança per a una mostra de mida n.

7 El model multinomial

Suposem que el fenomen aleatori on es poden donar m resultats possibles amb probabilitats respectives $p_1, \ldots, p_m$, tals que $\forall i : p_i > 0$ i $\sum_{i=1}^{m} p_i = 1$.

Sigui X un vector aleatori m-dimensional que ens indica el resultat obtingut amb un 1 a la posició corresponent. Aleshores:

$$\begin{aligned} P(X = (1, 0, \ldots, 0)) &= p_1, \\ P(X = (0, 1, \ldots, 0)) &= p_2, \\ &\ldots \\ P(X = (0, 0, \ldots, 1)) &= p_m. \end{aligned}$$

La distribució de X s'anomena Bernoulli m-dimensional de paràmetres $p_1, \ldots, p_m$ respectivament. Notem que les probabilitats les podem escriure de la forma

$$P(X = (x_1, \ldots, x_m)) = \prod_{i=1}^{m} p_i^{x_i}.$$

Sigui ara el model corresponent a n observacions independents i idènticament distribuïdes amb distribució Bernoulli m-dimensional de paràmetres $p_1, \ldots, p_m$. Aleshores:

$$P(X_1 = (x_{11}, \ldots, x_{1m}), X_2 = (x_{21}, \ldots, x_{2m}), \ldots, X_n = (x_{n1}, \ldots, x_{nm})) = \prod_{j=1}^{n} \prod_{i=1}^{m} p_i^{x_{ji}} = \prod_{i=1}^{m} p_i^{\sum_{j=1}^{n} x_{ji}}.$$

Donada una mostra x (que consistirà en n-vectors m-dimensionals) la funció de versemblança ve donada per:

$$L(x; p_1, \ldots, p_m) = P(X_1 = (x_{11}, \ldots, x_{1m}), \ldots, X_n = (x_{n1}, \ldots, x_{nm})) = \prod_{j=1}^{n} \prod_{i=1}^{m} p_i^{x_{ji}} = \prod_{i=1}^{m} p_i^{\sum_{j=1}^{n} x_{ji}}.$$

Notem que $N_i(x) = \sum_{j=1}^{n} x_{ji}$ compta el nombre de vegades que ha passat el resultat i–èssim en els n experiments. És fàcil veure que

$$P(N_1 = n_1, \ldots, N_m = n_m) = \frac{n!}{n_1! n_2! \ldots n_m!} p_1^{n_1} p_2^{n_2} \ldots p_m^{n_m},$$

aquesta distribució es coneix com distribució multinomial. Observem que l'espai de paràmetres del model multinomial és

$$\Theta = \{(p_1, p_2, \ldots p_m), p_i > 0; \sum_{i=1}^{m} p_i = 1\}$$

de manera que $dim(\Theta) = m - 1$.

7.1 Test d'ajustament del model multinomial

Karl Pearson en el 1900 va publicar la prova de la χ^2, una mesura de bonança d'una certa distribució en ajustar-se a un grup determinat de dades. La prova permet determinar, entre altres coses, si dos caràcters hereditaris eren transmesos de forma dependent o independent i també donar una mesura de l'ajustament entre una distribució teòrica i una experimental.

En un model multinomial volem fer el contrast

$$H_0 : (p_1, \ldots, p_m) = (p_{01}, \ldots, p_{0m}) \text{ contra } H_1 : (p_1, \ldots, p_m) \neq (p_{01}, \ldots, p_{0m})$$

Construïm el test de raó de versemblança, en aquest cas

$$\lambda_n(x) = \frac{L(x; p_0)}{\sup_{p \in \Theta} L(x, p)} = \frac{L(x; p_0)}{L(x, \hat{p})}$$

on utilitzem la notació vectorial: $p = (p_1, p_2, \ldots, p_m)$ i $\hat{p}$ és l'estimador de màxima versemblança. Recordem que la funció de versemblança és:

$$L(x; p) = \prod_{i=1}^{m} p_i^{n_i}.$$

Observació 7.1.1. Usualment en l'Estadística per trobar el màxim utilitzem el logaritme de la versemblança.

Prenent com a variables lliures, $p_1, p_2, \ldots, p_{m-1}$ i com $p_m = 1 - p_1 - p_2 \ldots - p_{m-1}$, les condicions d'extrem són

$$\partial_{p_i} \ln L(x; p) = \frac{n_i}{p_i} - \frac{n_m}{p_m} = 0, \quad i = 1, \ldots, m$$

això dóna

$$\hat{p_i} = \frac{n_i}{n}, \quad i = 1, \ldots, m$$

Per la Llei del Grans Nombres tenim

$$\frac{n_i}{n} \xrightarrow{c.s} p_i > 0$$

podem suposar que tots els n_i són positius, aleshores $L(x; p) = \prod_{i=1}^{m} p_i^{n_i}$ serà zero si algun p_i és zero, així el màxim estarà dins l'interior de Θ i correspondrà a la única solució que hem trobat. Tenim aleshores:

$$\lambda_n(x) = \frac{\prod_{i=1}^{m} p_{0i}^{n_i}}{\prod_{i=1}^{m} \hat{p_i}^{n_i}} = \prod_{i=1}^{m} \left(\frac{p_{0i}}{\hat{p_i}}\right)^{n_i} = \prod_{i=1}^{m} \left(\frac{p_{0i}}{\hat{p_i}}\right)^{n\hat{p_i}}$$

i sota H_0

$$W(x) = -2\ln\lambda_n(x) = -2n\sum_{i=1}^{m}\hat{p}_i\ln\frac{p_{0i}}{\hat{p}_i} \sim \chi^2_{m-1}$$

ja que $dim(\Theta) = m-1$ i $dim(\Theta_0) = 0$. La regió crítica de nivell α serà

$$A_1 = \{x, W(x) \geq \chi^2_{m-1;\alpha}\}$$

A la pràctica però s'utilitza un altre estadístic per construir la regió crítica.

Definició 7.1.1. *Definim* ***L'Estadístic de Pearson*** *com:*

$$D_n(x) = n\sum_{i=1}^{m}p_{0i}\Big(\frac{\hat{p}_i}{p_{0i}} - 1\Big)^2$$

També es pot expressar com:

$$D_n(x) = n\sum_{i=1}^{m}p_{0i}\Big(\frac{\hat{p}_i}{p_{0i}} - 1\Big)^2 = n\sum_{i=1}^{m}\frac{(\hat{p}_i - p_{0i})^2}{p_{0i}} = \sum_{i=1}^{m}\frac{(n\hat{p}_i - np_{0i})^2}{np_{0i}} = \sum_{i=1}^{m}\frac{(n_i - np_{0i})^2}{np_{0i}} = \sum_{i=1}^{m}\frac{(O_i - E_i)^2}{E_i},$$

on O_i és la freqüència observada i E_i l'esperada sota H_0.

Proposició 7.1.1. *La regió crítica de nivell α serà:*

$$A_1 = \{x, D_n(x) \geq k\}, \qquad on \quad P(\chi^2_{m-1} \leq k) = \alpha$$

i defineix un test que tindrà un nivel (aproximat) α. L'aproximació serà bona si $np_i \geq 5\ \forall i$.

Demostració 7.1.1. *Veiem que $W(x)$ i $D_n(x)$ tenen el mateix comportament asimptòtic sota H_0. Això demostrarà que les regions crítiques són equivalents.*

$$W(x) = 2n\sum_{i=1}^{m}\hat{p}_i\ln\frac{\hat{p}_i}{p_{0i}} = 2n\sum_{i=1}^{m}p_{0i}\frac{\hat{p}_i}{p_{0i}}\ln\frac{\hat{p}_i}{p_{0i}}$$

sabem que $\dfrac{\hat{p}_i}{p_{0i}} \xrightarrow[n\to\infty]{c.s} 1$ i si desenvolupem la funció $u\ln u$ al voltant del valor $u = 1$, tenim

$$u\ln u = u - 1 + \frac{1}{2}(u-1)^2 + \frac{1}{6}\frac{1}{u^*}(u-1)^3$$

amb $u^* \in (1, u)$. Tenim així

$$W(x) = 2n\sum_{i=1}^{m}p_{0i}\Big(\frac{\hat{p}_i}{p_{0i}} - 1\Big) + n\sum_{i=1}^{m}p_{0i}\Big(\frac{\hat{p}_i}{p_{0i}} - 1\Big)^2 + \frac{n}{3}\sum_{i=1}^{m}p_{0i}\frac{1}{u_i^*}\Big(\frac{\hat{p}_i}{p_{0i}} - 1\Big)^3,$$

on els $u_i^* \in (1, \frac{\hat{p}_i}{p_{0i}})$. *El primer sumant s'anul·la*

$$\sum_{i=1}^{m} p_{0i}\Big(\frac{\hat{p}_i}{p_{0i}} - 1\Big) = \sum_{i=1}^{m}(\hat{p}_i - p_{0i}) = \sum_{i=1}^{m}\hat{p}_i - \sum_{i=1}^{m} p_{0i} = 1 - 1 = 0,$$

el tercer sumant se'n va 0 *quan la* $n \to \infty$*:*

$$\frac{n}{3}\sum_{i=1}^{m} p_{0i}\frac{1}{u_i^*}\left(\frac{\hat{p}_i}{p_{0i}} - 1\right)^3 = \frac{1}{3\sqrt{n}}\sum_{i=1}^{m}\frac{(1-p_{0i})^{\frac{3}{2}}}{\sqrt{p_{0i}}u_i^*}\left(\frac{n_i - np_{0i}}{\sqrt{np_{0i}(1-p_{0i})}}\right)^3 \xrightarrow[n\to\infty]{P} 0$$

ja que

$$u_i^* \overset{c.s}{\to} 1 \qquad i \qquad \frac{n_i - np_{0i}}{\sqrt{np_{0i}(1-p_{0i})}} \xrightarrow{\mathscr{L}} N(0,1)$$

En definitiva $W(x)$ *i*

$$D_n(x) := n\sum_{i=1}^{m} p_{0i}\Big(\frac{\hat{p}_i}{p_{0i}} - 1\Big)^2$$

tenen el mateix comportament asimptòtic, s'apropen a una X^2_{m-1}.

8 Dades de Mendel

Gregor Mendel va publicar l'any 1865 l'article *Versuche über Pflanzenhybriden*, Experiments d'hibridació en plantes. Intentem reproduir el que Mendel va deixar per escrit en el seu article.
Gregor Mendel a Carl Nägeli durant l'abril del 1867, *In 1859 I obtained a very fertile descendent with large, tasty seeds from a first generation hybrid. Since in the following year, its progeny retained the desirable characteristics and were uniform, the variety was cultivated in our vegetable garden, and many plants were raised every year up to 1865*

8.1 Procediment dels experiments

Mendel realitzava sempre el mateix esquema de creuaments, creuava dues varietats o línies pures que diferien en un o més caràcters, utilitzant per a la fecundació la llavor i el pol·len de la planta. A aquests individus els anomenava parentals. Així obtenia la primera generació filial F_1, seguidament els autofecundava i obtenia la segona generació filial F_2, que també anomena com la primera generació dels híbrids. Finalment repetia els procés d'autofecundació de les plantes de F_2 i aconseguia la tercera generació filial F_3 també anomenada la segona generació dels híbrids. El creuament inicial el va dur a terme en les dues direccions possibles, és a dir, va realitzar creuaments recíprocs.

Els caràcters que es transmetien quasi o sense canvis en la hibridació per a les generacions posteriors els anomenem *dominants*, en canvi els *recessius* són aquells caràcters que sembla que desapareixin en la descendència dels híbrids però que tornen a reaparèixer sense canvis.

8.1.1 Selecció de les plantes experimentals

Mendel explica que per a la realització dels experiments és important la selecció del material per al propòsit pel qual s'utilitza. Així la selecció del grup de plantes s'ha de fer conseqüent ja que des de el principi s'han d'evitar els riscs de resultats qüestionables, per tant, les plantes escollides han de posseir necessàriament:

- Caràcters constants diferenciables,
- els híbrids d'aquestes plantes durant els períodes de floració han d'estar protegits de la influència del pol·len extern, o bé directament, o bé mitjançant mètodes de protecció i
- els híbrids i els seus descendents no haurien de patir alteracions de fertilitat en les successives generacions.

Va optar després d'alguns experiments per l'espècie *Pisium Sativum* ja que reunien les condicions necessàries explicades per a dur a terme els experiments. Les formes completament diferenciables d'aquest gènere fa que tinguin caràcters constants i fàcilment recognoscibles. A més, els creuaments obtenen perfectament descendents

fèrtils. Mendel explica també els avantatges, com la facilitat de cultiu en el camp obert i en testos, el seu curt període de creixament i la fecundació principal quasi sempre té èxit.

En total 34 varietats de pèsols es van posar a judici per als experiments durant dos anys, dues generacions successives per autofecundació, per comprovar que totes produïen descendència constant. Una d'elles no va mostrar les característiques necessàries per tal de dur a terme els experiments. Així que per la fertilització i el cultiu en va utilitzar 22 que mostraven constància sense cap excepció. Les varietats utilitzades eren línies pures, d'individus o grups individus que descendien d'ells per autofecundació sent homozigots per a tots els seus caràcters, formades per individus idèntics pels caràcters analitzats.

8.1.2 Propietats del *Pisium Sativum*

Aquesta espècie va ser la escollida per les següents raons:

- Eren barats i fàcils d'obtenir en el mercat,
- ocupaven poc espai i tenien un temps de generació relativament curt,
- produïen molts descendents,
- existien varietats diferents que mostraven diferents colors, formes i mides, presentaven així una gran variabilitat genètica i
- era una espècie autògama, és a dir, la reproducció sexual consisteix en la fusió de gamets masculins i femenins produïts pel mateix individu gràcies a la pol·linització.

8.1.3 Caràcters diferenciadors analitzats del *Pisium Sativum*

Caràcter	Dominant	Recessiu
Forma de les llavors madures	Rodona	Arrugada
Color de la llavor	Groga	Verda
Color de les flors	Lila	Blanca
Forma de les beines madures	Llisa	Arrugada
Color de les beines inmadures	Verda	Groga
Posició de les flors	Axial	Terminal
Longitud de la tija	Llarga	Petita

Podem observar els caràcters en la taula següent:

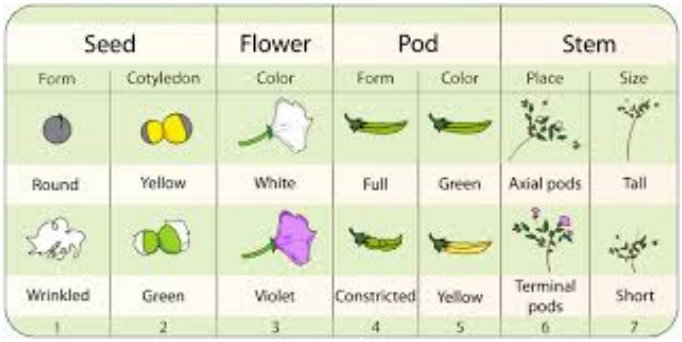

Figura 3: Caràcters.

8.1.4 Esquema dels creuaments

Mendel va realitzar 7 experiments, un per cada caràcter i va anar analitzant les generacions successives en relació als caràcters observats.

Observem els creuaments realitzats per Mendel per un únic caràcter.

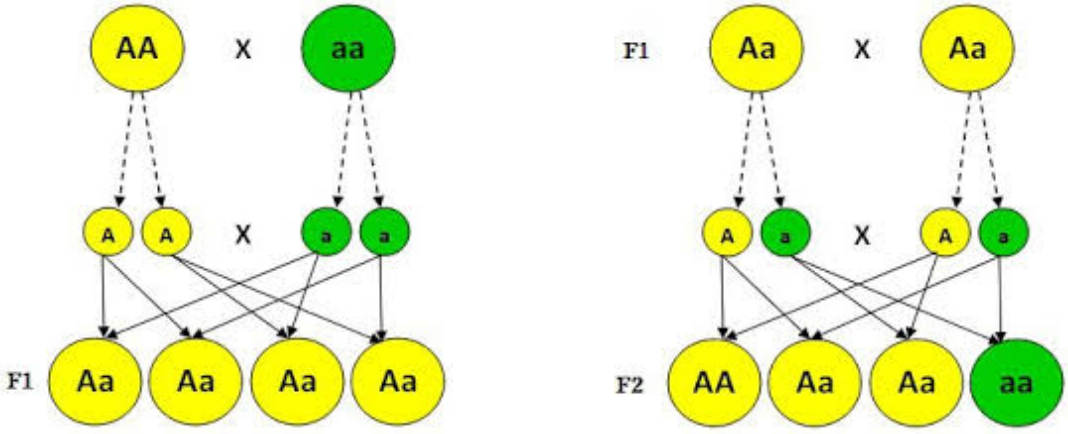

Figura 4: Creuaments.

8.2 Dades dels experiments

Per raons de brevetat, Mendel suposa que cada planta a cada generació aporta 4 llavors, ho veiem en la següent taula:

Generació	AA	Aa	aa	A :	Aa :	a
1	1	2	1	1 :	2 :	1
2	6	4	6	3 :	2 :	3
3	28	8	28	7 :	2 :	7
4	120	16	120	15 :	2 :	15
5	496	32	496	31 :	2 :	31
...	...	...	...	...	...	...
n	$\frac{2^n(2^n-1)}{2}$	2^n	$\frac{2^n(2^n-1)}{2}$	2^n-1 :	2 :	2^n-1

Donat que Mendel no deixa per escrit en el seu article algunes de les caselles de la generació n-èssima, presentem la demostració:

Demostració 8.2.1. *Comprovem els resultats de la taula per la generació n-èssima. Ho veiem per inducció.*

- *Per $n = 1$, comprovem la quantitat de descendents en la segona generació:*
  - $4(\frac{2^1(2^1-1)}{2}) + 2^1 = 6$,
  - $2(2^1) = 4$.
- *Suposem cert fins a $n-1$, veiem si es compleix per n, en efecte:*
  - $4(\frac{2^{n-1}(2^{n-1}-1)}{2}) + 2^{n-1} = 4(2^{n-2}(2^{n-1}-1)) + 2^{n-1} = 2^n(2^{n-1}-1) + 2^{n-1} = 2^{2n-1} - 2^n + 2^{n-1} = 2^{n-1}(2^n - 2 + 1) = 2^{n-1}(2^n - 1) = \frac{2^n(2^n-1)}{2}$,
  - $2(2^{n-1}) = 2^n$.

Observació 8.2.1. En general, si $k \in \mathbb{N}$ representa el nombre de caràcters estudiats, en relació al genotip, aleshores:

- 2^k dóna el nombre de descendents que permaneixen constants,
- 3^k dóna el nombre de possibilitats de la descendència,
- 4^k dóna el nombre d'individus de la descendència.

8.2.1 La primera generació dels híbrids estudiant un caràcter

Donem els resultats obtinguts per cada caràcter estudiat per Mendel en la F_2.

Fenotip parental	Quantitat total	F_2	Proporció
Llavor rodona-arrugada	7324	5474 rodones i 1850 arrugades	2.96:1
Llavor groga-verda	8023	6022 grogues i 2001 verdes	3.01:1
Coberta llavor lila-blanca	929	705 liles i 224 blanques	3.15:1
Beina llisa-arrugada	1181	882 llises i 299 arrugades	2.95:1
Beina verda-groga	580	428 verdes i 152 grogues	2.82:1
Flor axial-terminal	858	651 axials i 207 terminals	3.14:1
Tija llarga-petita	1064	787 llargues i 277 petites	2.84:1

Observació 8.2.1.1.

- La proporció està arrodonida a les centèssimes.
- La mitjana de les proporcions entre fenotips dominants i recessius és, truncant a les centèssimes, $2.98:1$, si arrodonim a les dècimes obtenim la relació $3:1$.

8.2.2 La segona generació dels híbrids estudiant un caràcter

Mendel estudia els descendents amb fenotip dominant que ha obtingut en la F_2, és a dir, quins del quals tenen genotip AA o bé Aa. Per tant, autofecunda un altra vegada per tal d'esbrinar-ho. Els que li han sortit amb genotip aa, no els té en compte ja que sap que són constants. La taula que tenim a continuació dóna els resultats.

Fenotip parental	F_2 dominant	Distinció de F_2 segons F_3	Proporció
Llavor rodona-arrugada	565 rodones	372 rodones-arrugades i 193 arrugades	1.93:1
Llavor groga-verda	519 grogues	353 grogues-verdes i 166 verdes	2.13:1
Coberta llavor lila-blanca	100 liles	64 liles-blanques i 36 blanques	1.78:1
Beina llisa-arrugada	100 llises	71 llises-arrugades i 29 arrugades	2.45:1
Beina verda-groga	100 verdes	60 verdes-grogues i 40 grogues	1.5:1
Flor axial-terminal	100 axials	67 axials-terminals i 33 terminals	2.03:1
Tija llarga-petita	100 llargues	72 llargues-petites i 28 petites	2.57:1

Observació 8.2.2.

- La proporció està arrodonida a les centèssimes.
- La mitjana de les proporcions entre fenotips dominants i recessius és, truncant a les centèssimes, $2.05:1$, si trunquem a les dècimes obtenim la relació $2:1$.
- Així F_2 segueix la relació $2:1:1$.

8.2.3 La primera generació dels híbrids estudiant dos caràcters

Observem la descendència de F_1 creuant dos caràcters.

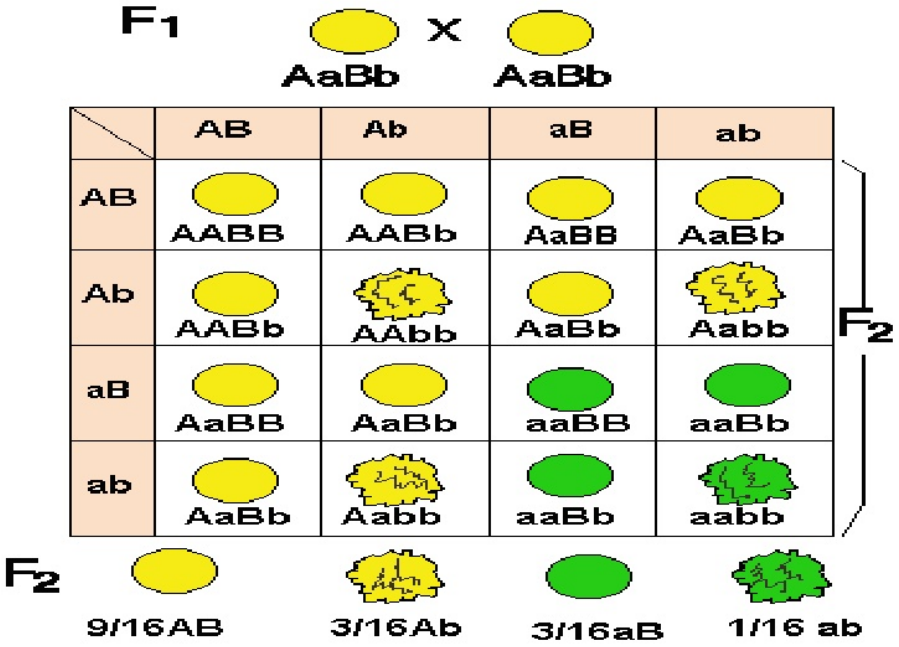

Figura 5: Creuaments.

Mendel explica que fa el creuament següent:

AB llavor parental	ab pol·len parental
A rodona	a arrugada
B groga	b verda

Pel genotip de F_2 obté els resultats següents, arrodonint al quart decimal:

Genotip	Teòric	Pràctic
AABB	$1/16 = 0.0625$	$38/529 \approx 0.0718$
aaBB	$1/16 = 0.0625$	$28/529 \approx 0.0529$
AAbb	$1/16 = 0.0625$	$35/529 \approx 0.0662$
aabb	$1/16 = 0.0625$	$30/529 \approx 0.0567$
AaBB	$2/16 = 0.125$	$60/529 \approx 0.1134$
AABb	$2/16 = 0.125$	$65/529 \approx 0.1229$
aaBb	$2/16 = 0.125$	$68/529 \approx 0.1285$
Aabb	$2/16 = 0.125$	$67/529 \approx 0.1267$
AaBb	$4/16 = 0.25$	$138/529 \approx 0.2609$

Formem 4 grups segons si per cada caràcter tenim al.lels iguals o no. El primer grup consta d'aquells que pels dos caràcters els seus al·lels són idèntics, el segon grup són els que per un dels caràcters tenen al·lels diferents i l'últim grup és el que té pels dos caràcters al·lels diferents.
Calculem la mitjana per cada grup: $\frac{38+28+35+30}{4}$, $\frac{60+65+68+67}{4}$ i $\frac{138}{1}$.

- La mitjana de les proporcions per cada grup és: $32.75 : 65 : 138$ que presenta una aproximació $33 : 66 : 132$, així obtenim la relació $1 : 2 : 4$.

La següent taula dóna els resultats corresponents al fenotip de F_2, arrodonint a quatre decimals.

Fenotip	Tèoric	Pràctic
AB	$9/16 = 0.5625$	$301/529 \approx 0.5690$
Ab	$3/16 = 0.1875$	$102/529 \approx 0.1928$
aB	$3/16 = 0.1875$	$96/529 \approx 0.1815$
ab	$1/16 = 0.0625$	$30/529 \approx 0.0567$

- La proporció obtinguda és: $301 : 102 : 96 : 30$ que presenta una aproximació $10.03 : 3.4 : 3.2 : 1$ que s'apropa a la relació $9 : 3 : 3 : 1$.

8.2.4 La primera generació dels híbrids estudiant tres caràcters

Mendel explica que fa el creuament següent:

ABC llavor parental	abc pol·len parental
A rodona	a arrugada
B groga	b verda
C lila	c blanca

En la F_2 obté pel genotip els resultats següents, arrodonint a 4 decimals:

Genotip	Teòric	Pràctic
AABBCC	$1/64 \approx 0.0156$	$8/639 \approx 0.0125$
AABBcc	$1/64 \approx 0.0156$	$14/639 \approx 0.0219$
AAbbCC	$1/64 \approx 0.0156$	$9/639 \approx 0.0141$
AAbbcc	$1/64 \approx 0.0156$	$11/639 \approx 0.0172$
aaBBCC	$1/64 \approx 0.0156$	$8/639 \approx 0.0125$
aaBBcc	$1/64 \approx 0.0156$	$10/639 \approx 0.0156$
aabbCC	$1/64 \approx 0.0156$	$10/639 \approx 0.0156$
aabbcc	$1/64 \approx 0.0156$	$7/639 \approx 0.0110$
AABBCc	$2/64 \approx 0.0313$	$22/639 \approx 0.0344$
AABbCC	$2/64 \approx 0.0313$	$15/639 \approx 0.0235$
AABbcc	$2/64 \approx 0.0313$	$18/639 \approx 0.0282$
AAbbCc	$2/64 \approx 0.0313$	$17/639 \approx 0.0266$
AaBBCC	$2/64 \approx 0.0313$	$14/639 \approx 0.0219$
AaBBcc	$2/64 \approx 0.0313$	$18/639 \approx 0.0282$
AabbCC	$2/64 \approx 0.0313$	$20/639 \approx 0.0313$
Aabbcc	$2/64 \approx 0.0313$	$16/639 \approx 0.0250$
aaBBCc	$2/64 \approx 0.0313$	$25/639 \approx 0.0391$
aaBbCC	$2/64 \approx 0.0313$	$19/639 \approx 0.0297$
aaBbcc	$2/64 \approx 0.0313$	$24/639 \approx 0.0376$
aabbCc	$2/64 \approx 0.0313$	$20/639 \approx 0.0313$
AABbCc	$4/64 = 0.0625$	$45/639 \approx 0.0704$
AaBBCc	$4/64 = 0.0625$	$38/639 \approx 0.0595$
AaBbCC	$4/64 = 0.0625$	$49/639 \approx 0.0767$
AaBbcc	$4/64 = 0.0625$	$48/639 \approx 0.0751$
AabbCc	$4/64 = 0.0625$	$40/639 \approx 0.0626$
aaBbCc	$4/64 = 0.0625$	$36/639 \approx 0.0563$
AaBbCc	$8/64 = 0.125$	$78/639 \approx 0.1221$

Formem 5 grups segons si per cada caràcter tenim al.lels iguals o no. El primer consta d'aquells que pels tres caràcters els seus al·lels són idèntics, el segon grup són els que per un dels caràcters tenen al·lels diferents, el tercer són aquells que tenen dos caràcters amb al·lels diferents i l'últim grup és el que té per tots els caràcters al·lels diferents.
Calculem la mitjana per cada grup: $\frac{8+14+9+11+8+10+10+7}{8}$,
$\frac{22+15+18+17+14+18+20+16+25+19+24+20}{12}$,
$\frac{45+38+49+48+40+36}{6}$ i $\frac{78}{1}$.

- La mitjana de les proporcions per a cada grup és: $9.625 : 19 : 42.666 : 78$ que podem aporximar a la relació $10 : 20 : 40 : 80$ que és el mateix que $1 : 2 : 4 : 8$.

La següent taula dóna els resultats corresponents al fenotip de F_2, arrodonint a quatre decimals.

Fenotip	Teòric	Pràctic
ABC	$27/64 \approx 0.4219$	$269/639 \approx 0.4210$
ABc	$9/64 \approx 0.1406$	$98/639 \approx 0.1534$
AbC	$9/64 \approx 0.1406$	$86/639 \approx 0.1346$
aBC	$9/64 \approx 0.1406$	$88/639 \approx 0.1377$
Abc	$3/64 \approx 0.0469$	$27/639 \approx 0.0423$
aBc	$3/64 \approx 0.0469$	$34/639 \approx 0.0532$
abC	$3/64 \approx 0.0469$	$30/639 \approx 0.0469$
abc	$1/64 \approx 0.0156$	$7/639 \approx 0.0110$

- La proporció obtinguda és: $269 : 98 : 86 : 88 : 27 : 34 : 30 : 7$ que presenta una aproximació $38.42 : 14 : 12.28 : 12.57 : 3.85 : 4.85 : 4.28 : 1$ que s'apropa a la proporció $27 : 9 : 9 : 9 : 3 : 3 : 3 : 1$.

8.3 Lleis de Mendel

La Genètica Mendeliana o les Lleis de Mendel són el conjunt de regles bàsiques sobre la transmissió per herència genètica de les característiques dels organismes parentals als seus fills. Aquestes regles bàsiques d'herència constitueixen el fonament de la genètica. Aquestes lleis deriven dels resultats dels experiments.

8.3.1 Primera Llei de Mendel o Principi de la Uniformitat

Les plantes híbrides Aa de F_1 obtingudes pel creuament de dues línies pures que es diferencien en un únic caràcter tenen totes la mateixa aparença externa, fenotip, sent idèntiques entre si (uniformes) i s'assemblen a un dels dos parentals. Al caràcter que es manifesta en les plantes de F_1, híbrids Aa, s'anomena *dominant* i al caràcter que no es manifesta s'anomena *recessiu*. Aquest resultat és independent de la direcció en la que s'han dut a terme els creuaments.

8.3.2 Segona Llei de Mendel o Principi de Segregació

L'autofecundació de les plantes híbrides Aa procedents del creuament entre dues línies pures que es diferencien en un caràcter originen una segona generació filial F_2 en la que apareixen $\frac{3}{4}$ parts de les plantes d'aparença externa *dominant* i $\frac{1}{4}$ de plantes amb aparença externa *reccessiva*. De manera que el caràcter *recessiu* reapareix en la F_2 i de cada quatre plantes una té fenotip *recessiu*. Aquest resultat és conseqüència dels híbrids de la F_1 que formen els seus gamets, els al·lels del mateix locus es separen donant lloc a dues classes de gamets en igual proporció, meitat dels gamets amb al·lel *dominant* A i meitat dels gamets amb al·lel *recessiu* a. Això succeeix tant per la banda femenina com per la masculina.

8.3.3 Tercera Llei de Mendel o Principi de la Combinació Independent

Els membres de parelles d'al·lels diferents es distribueixen o combinen de forma independent quan formen els gamets d'un heterozigot per als caràcters corresponents. En el cas d'un doble heterozigot $AaBb$ els al·lels dels locus A, a i els del locus B, b es combinen de forma independent per formar quatre classes de gamets en igual proporció.

9 Anàlisi estadístic de les dades de Mendel

En aquesta secció analitzem els resultats pràctics i teòrics obtinguts per Mendel. Realitzem aquest anàlisi amb el test d'ajustament de la χ^2. Apliquem el test pels experiments analitzant un, dos i tres caràcters.

L'objectiu que estem buscant és acceptar la hipòtesi H_0, per tant ens interessa un p-valor gran i l'estadístic de contrast $D_n(x)$ sigui petit.

Utilitzem els valors crítics de la distribució χ^2 fixant $\alpha = 0.05$. Els que necessitem són els següents:

g.d.l	$\alpha = 0.05$
1	3.841
2	5.991
3	7.815
7	14.067

Comencem l'estudi. Per a cada cas considerem el test d'hipòtesis adequat i fixem el nivell de significació, optem per $\alpha = 0.05$. Calculem l'estadístic de contrast de Pearson $D_n(x)$ gràcies al programa R i obtenim els graus de llibertat i el p-valor corresponent a cada test. Definim la regió d'acceptació A_0 i realitzem les gràfiques de la funció de densitat per a cada test amb els resultats obtinguts, marquem en el eix OX els resultats de $D_n(x)$ i tracem per cada resultat un segment vertical fins a arribar a la imatge de la funció de densitat de la χ^2_n corresponent. Cada segment de color mostra el resultat per cada experiment, incloem el resultat del test per a totes les dades juntes en alguns casos i marquem l'àrea obtinguda en gris que equival al p-valor.

Si calculem $P(\chi^2_n > D_n(x)) = p$-valor, ens dóna una idea més clara de la bondat d'ajustament ja que és la probabilitat que una variable amb llei χ^2_n prengui un valor més gran o igual que $D_n(x)$. El p-valor indica quin és el límit de α per seguir acceptant H_0.

9.1 La primera generació dels híbrids per un únic caràcter

Analitzem F_2 i el model utilitzat en aquest cas té distribució χ^2_1. Partint de la proporció esperada $3:1$, considerem el test d'hipòtesis següent:

$$H_0 : p = (\frac{3}{4}, \frac{1}{4}) \qquad H_1 : p \neq (\frac{3}{4}, \frac{1}{4})$$

Recordem en aquesta taula les dades pràctiques obtingudes. Cada experiment es va fer per un caràcter determinat.

Experiment: Caràcter	Fenotip dominant	Fenotip recessiu
Exp 1: Forma de les llavors madures	5474	1850
Exp 2: Color de la llavor	6022	2001
Exp 3: Color de les flors	705	224
Exp 4: Forma de les beines madures	882	299
Exp 5: Color de les beines inmadures	428	152
Exp 6: Posició de les flors	651	207
Exp 7: Longitud de la tija	787	277
Total	14949	5010

Apliquem el test per cada experiment i per la suma de tots i obtenim els resultats:

Experiment	$D_n(x)$	df	p-valor
1	0.26288	1	0.6081
2	0.014999	1	0.9025
3	0.39074	1	0.5319
4	0.063506	1	0.801
5	0.45057	1	0.5021
6	0.34965	1	0.5543
7	0.60652	1	0.4361
Total	0.10957	1	0.7406

Sabent que la distribució asimptòtica de $D_n(x)$ és una χ^2_1, definim la regió d'acceptació del test a nivell $\alpha = 0.05$

$$A_0 = \{x; D_n(x) \leq 3.841\}$$

per tant, acceptem H_0.

Observació 9.1. La probabilitat més alta ens la dóna l'experiment 2, la més baixa l'experiment 7 i finalment podem dir que la probabilitat obtinguda per a totes les dades juntes és molt bona.

Observem la gràfica corresponent als resultats obtinguts:

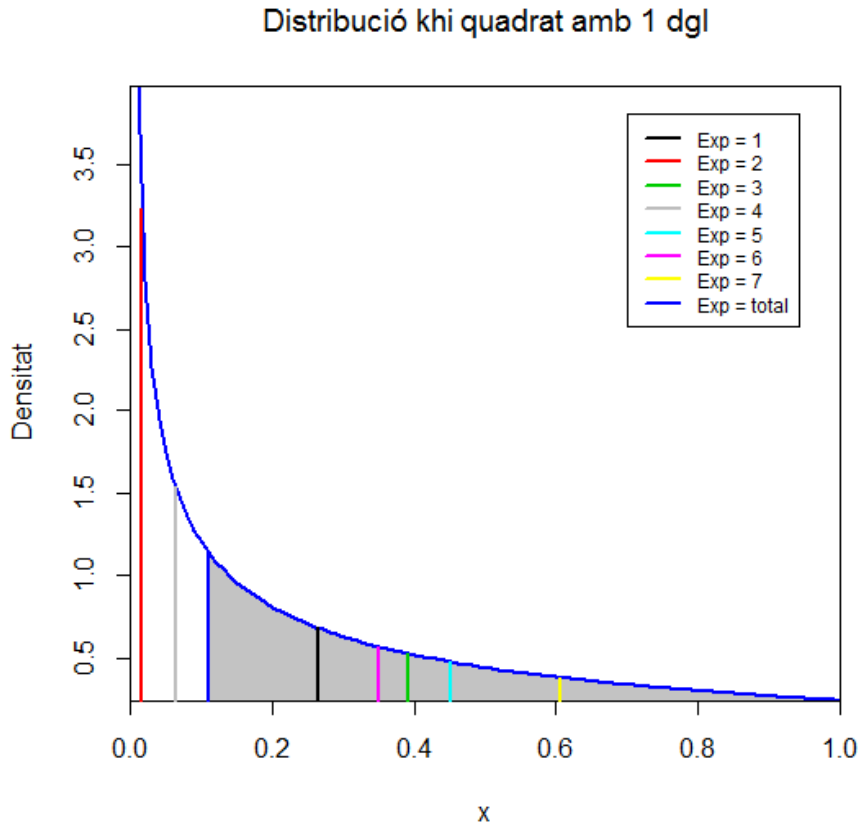

Figura 6: F_2 per un únic caràcter.

9.2 La segona generació dels híbrids per un únic caràcter

Volem analitzar F_2, per-ho ens quedem amb els que mostren fenotip *dominant*, aquests els autofecunden i el que volem és estudiar quins de F_3 provenen de AA i quins de Aa. El model utilitzat en aquest cas té distribució χ^2_1.
Considerant la proporció esperada $2:1$, considerem el test d'hipòtesis següent:

$$H_0 : p = (\frac{2}{3}, \frac{1}{3}) \qquad H_1 : p \neq (\frac{2}{3}, \frac{1}{3})$$

Recordem en aquesta taula les dades pràctiques obtingudes. Cada experiment es va fer per un caràcter determinat i per la suma de tots.

Experiment: Caràcter	Genotip AA	Genotip AA, Aa i aa
Exp 1: Forma de les llavors madures	193	372
Exp 2: Color de la llavor	166	353
Exp 3: Color de les flors	36	64
Exp 4: Forma de les beines madures	29	71
Exp 5: Color de les beines	40	60
Exp 6: Posició de les flors	33	67
Exp 7: Longitud de la tija	28	72
Total	525	1059

Apliquem el test per cada experiment i per la suma de totes les dades i obtenim els resultats:

Experiment	$D_n(x)$	df	p-valor
1	0.17345	1	0.6771
2	0.42486	1	0.5145
3	0.32	1	0.5716
4	0.845	1	0.358
5	2	1	0.1573
6	0.005	1	0.9436
7	1.28	1	0.2579
Total	0.025568	1	0.873

Utilitzant que la distribució asimptòtica de $D_n(x)$ és una χ^2_1, llavors la regió d'acceptació del test a nivell $\alpha = 0.05$ és

$$A_0 = \{x; D_n(x) \leq 3.841\}$$

per tant, acceptem H_0.

Observació 9.2. La probabilitat més alta ens la dóna l'experiment 6, la més baixa l'experiment 5 i la probabilitat obtinguda per a totes les dades és molt bona.

La gràfica següent ens mostra els resultats:

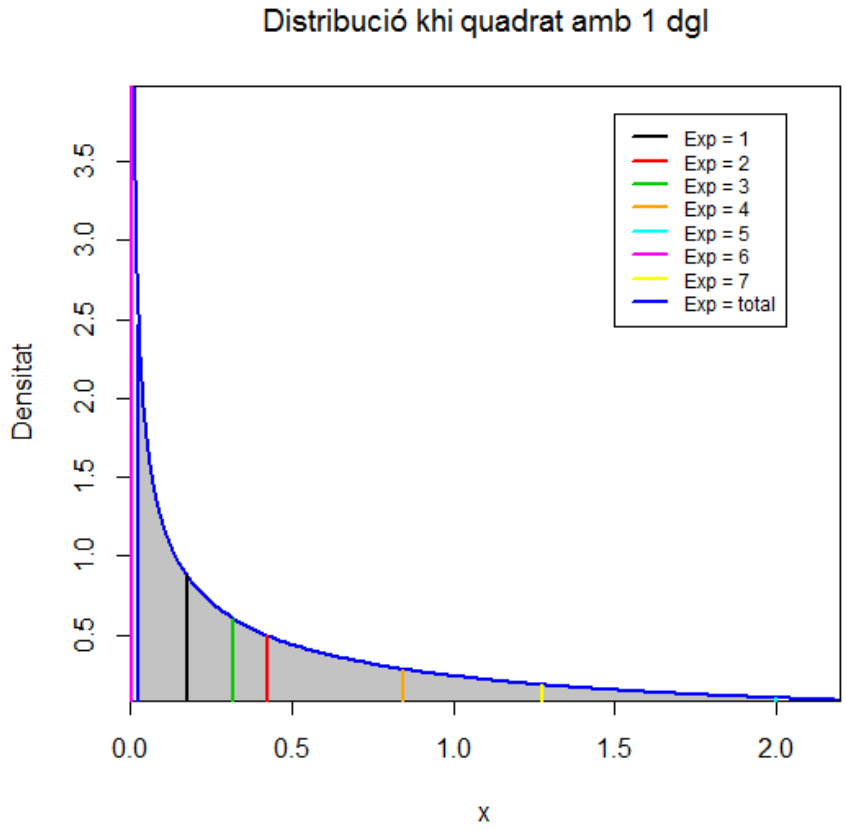

Figura 7: Els dominants de F_2 per un únic caràcter.

Podem dir que acceptem H_0 per tots els tests realitzats, que quantifiquem com a 14 tests independents, excloent als que engloben totes les dades. Podem dir que la proporció obtinguda per les dades pràctiques és molt semblant a l'esperada.

9.3 La primera generació dels híbrids per dos caràcters

9.3.1 Genotip

Analitzem el genotip de F_2 per dos caràcters i el model utilitzat en aquest cas té distribució χ^2_2.

Partint de la proporció esperada $1:2:4$, considerem el test d'hipòtesi següent:

$$H_0 : p = (\frac{1}{7}, \frac{2}{7}, \frac{4}{7}) \qquad H_1 : p \neq (\frac{1}{7}, \frac{2}{7}, \frac{4}{7})$$

Recordem en aquesta taula les dades pràctiques obtingudes pel genotip.

Genotip	Resultats pràctics
AABB	38
aaBB	28
AAbb	35
aabb	30
AaBb	138
AaBB	60
AABb	65
aaBb	68
Aabb	67
Total	529

Amb aquestes dades pràctiques, considerem la mitjana per al tres grups: 32.75,138 i 65, apliquem el test i obtenim el resultat:

Experiment	$D_n(x)$	df	p-valor
Genotip	0.18823	2	0.9102

La distribució asimptòtica de $D_n(x)$ és una χ^2_2, definim la regió d'acceptació del test a nivell $\alpha = 0.05$ com

$$A_0 = \{x; D_n(x) \leq 5.991\}$$

així acceptem H_0.

Realitzem la gràfica corresponent amb els resultats obtinguts:

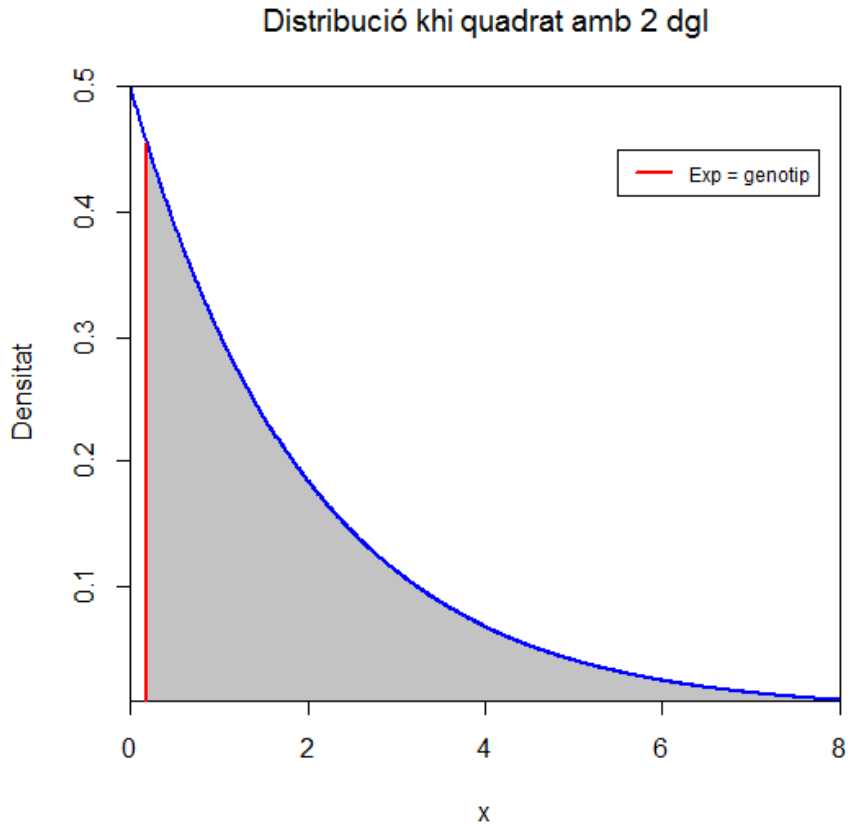

Figura 8: Genotip de F_2 per dos caràcters.

9.3.2 Fenotip

Analitzem el fenotip de F_2 per dos caràcters i el model utilitzat en aquest cas té distribució χ^2_3, partint de la proporció esperada $9:3:3:1$ considerem el test d'hipòtesi següent:

$$H_0 : p = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}) \qquad H_1 : p \neq (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16})$$

Recordem en aquesta taula les dades pràctiques obtingudes pel fenotip.

Fenotip	Resultats pràctics
AB	301
Ab	102
aB	96
ab	30

Apliquem el test i obtenim el resultat:

Experiment	$D_n(x)$	df	p-valor
Fenotip	0.50557	3	0.9177

Utilitzant que la distribució asimptòtica de $D_n(x)$ és una χ^2_3, definim la regió d'acceptació del test a nivell $\alpha = 0.05$ com

$$A_0 = \{x; D_n(x) \leq 7.815\}$$

per tant, acceptem H_0.

La gràfica que mostra els resultats obtinguts és:

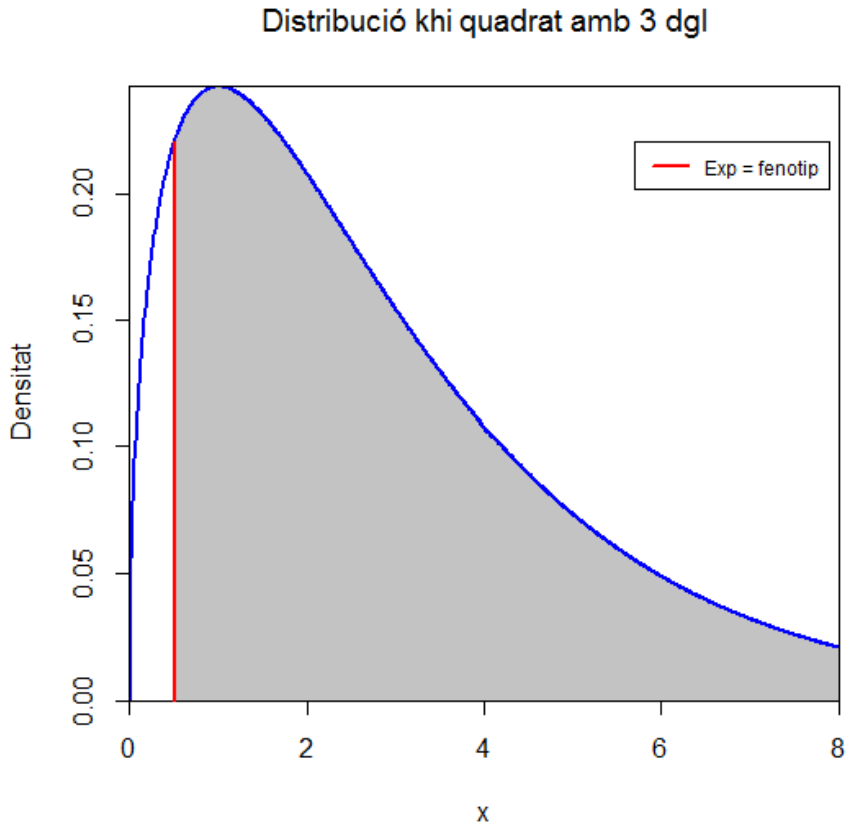

Figura 9: Fenotip de F_2 per dos caràcters.

Finalment podem dir que acceptem H_0 per a les proporcions esperades tant del fenotip com del genotip de F_2.

9.4 La primera generació dels híbrids per tres caràcters

9.4.1 Genotip

Analitzem el genotip de F_2 creuant tres caràcters i el model utilitzat en aquest cas té distribució χ^2_3. Partint de la proporció esperada $1:2:4:8$, considerem el test d'hipòtesi següent:

$$H_0 : p = (\frac{1}{15}, \frac{2}{15}, \frac{4}{15}, \frac{8}{15}) \qquad H_1 : p \neq (\frac{1}{15}, \frac{2}{15}, \frac{4}{15}, \frac{8}{15})$$

Recordem en aquesta taula les dades pràctiques obtingudes pel genotip.

Genotip	Resultats pràctics
AABBCC	8
AABBcc	14
AAbbCC	9
AAbbcc	11
aaBBCC	8
aaBBcc	10
aabbCC	10
aabbcc	7
Total	77
Mitjana	77/8

Genotip	Resultats pràctics
AABBCc	22
AAbbCc	17
aaBBCc	25
aabbCc	20
AABbCC	15
AABbcc	18
aaBbCC	19
aaBbcc	24
AaBBCC	14
AaBBcc	18
AabbCC	20
Aabbcc	16
Total	228
Mitjana	228/12

Genotip	Resultats pràctics
AABbCc	45
aaBbCc	36
AaBBCc	38
AabbCc	40
AaBbCC	49
AaBbcc	48
Total	256
Mitjana	256/6

Genotip	Resultats pràctics
AaBbCc	78
Total	78
Mitjana	78/1

Amb aquestes dades pràctiques, considerem la mitjana per als quatre grups: 9.625, 19, 42.67 i 78, apliquem el test i obtenim el resultat:

Experiment	$D_n(x)$	df	p-valor
Genotip	0.29034	3	0.9618

Utilitzant que la distribució asimptòtica de $D_n(x)$ és una χ^2_3 considerem la regió d'acceptació del test a nivell $\alpha = 0.05$ com

$$A_0 = \{x; D_n(x) \leq 7.815\}$$

així acceptem H_0.

La gràfica corresponent als resultats és:

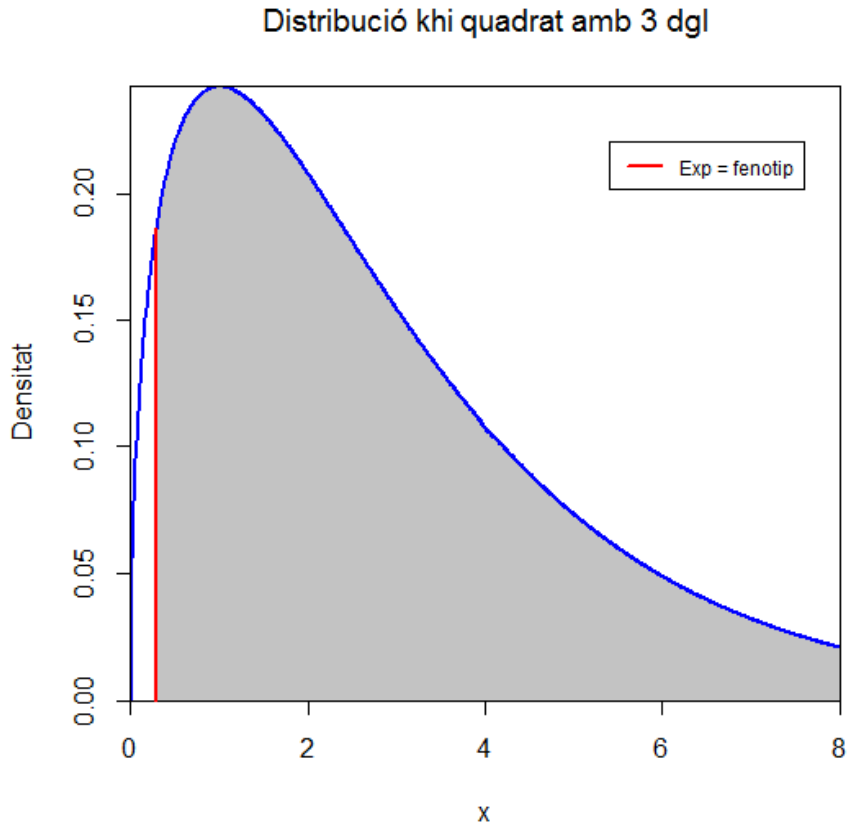

Figura 10: Genotip de F_2 per tres caràcters.

9.4.2 Fenotip

Analitzem el fenotip de F_2 creuant tres caràcters i el model utilitzat en aquest cas té distribució χ^2_7. Partint de la proporció $27:9:9:9:3:3:3:1$, considerem el test d'hipòtesi següent:

$$H_0 : p = (\frac{27}{64}, \frac{9}{64}, \frac{9}{64}, \frac{9}{64}, \frac{3}{64}, \frac{3}{64}, \frac{3}{64}, \frac{1}{64}) \qquad H_1 : p \neq (\frac{27}{64}, \frac{9}{64}, \frac{9}{64}, \frac{9}{64}, \frac{3}{64}, \frac{3}{64}, \frac{3}{64}, \frac{1}{64})$$

Recordem en aquesta taula les dades pràctiques obtingudes pel fenotip.

Fenotip	Resultats pràctics
Abc	27
ABC	269
ABc	98
AbC	86
aBC	88
aBc	34
abC	30
abc	7

Apliquem el test i obtenim el resultat:

Experiment	$D_n(x)$	df	p-valor
Fenotip	2.673	7	0.9135

Utilitzant que la distribució asimptòtica de $D_n(x)$ és una χ^2_7, definim la regió d'acceptació del test a nivell $\alpha = 0.05$ com:

$$A_0 = \{x; D_n(x) \leq 14.067\}$$

per tant, acceptem H_0.

La gràfica que mostra els resultats obtinguts és:

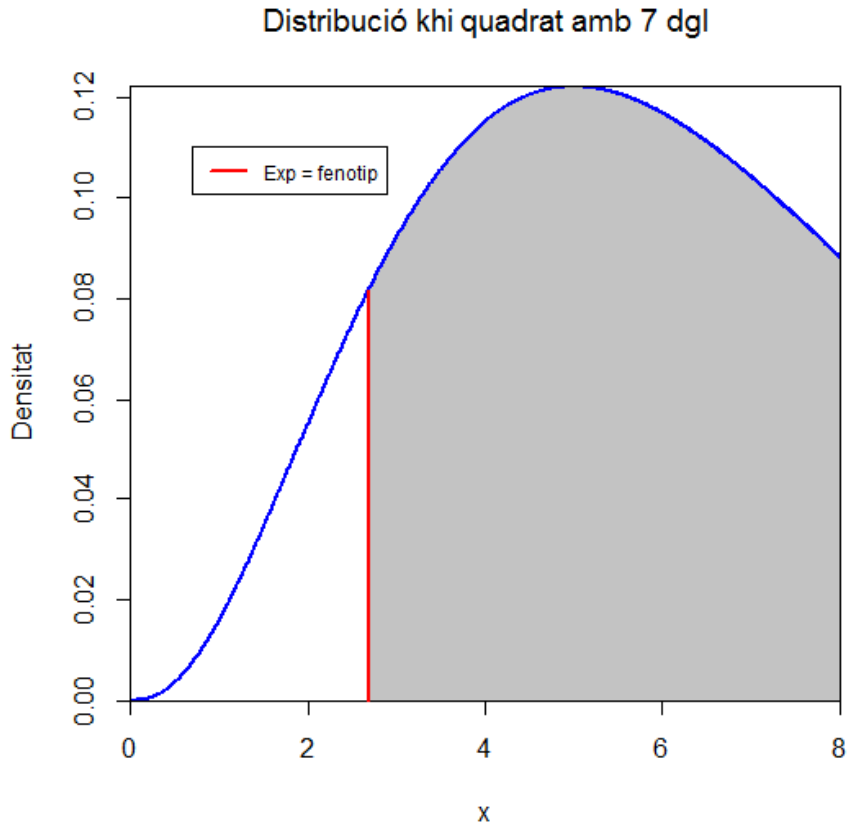

Figura 11: Fenotip de F_2 per tres caràcters.

Podem dir que acceptem H_0 creuant tres caràcters tant pel genotip com pel fenotip.

9.5 Massa bo per ser veritat!

Hem fet 18 tests χ^2. Sabem ara, amb el coneixement actual, que ens trobem sempre sota H_0. Utilitzant $\alpha = 0.05$, ens trobem que en un 5% dels tests ens hauriem d'haver equivocat, o sigui $(18 \cdot 5)/100 = 0.9$ tests, però no ho hem fet en cap ocasió. Aquesta consideració no és incompatible amb els resultats ja que $0.9 < 1$.

Veiem ara però un raonament a partir del p-valor. El p-valor més petit que hem trobat és 0.1573; prenent $\alpha = 0.15$ no hauriem rebutjat mai H_0, mentre que teòricament $(18 \cdot 15)/100 = 2.7$, seria el nombre esperat de tests a rebutjar; això contradiuria clarament els resultats de l'experiment.

Aquesta consideració ens porta a pensar que Mendel o els seus ajudants van maquillar les dades. Fisher comenta textualment durant la difusió de l'article de Mendel: *The general level of agreement between Mendel's expectations and his reported results shows that it is closer than would be expected in the best of several thousand repetitions... I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made.*

10 Conclusions

Aquest treball de fi de carrera l'he portat a terme utilitzant els coneixements adquirits en les assignatures troncals de Probabilitats i Estadística del Grau de Matemàtiques, altres coneixements adquirits durant el Grau i també durant el desenvolupament del propi treball. Realitzem una tasca de comprensió de l'article de Mendel per obtenir les dades pràctiques i conèixer la metodologia desenvolupada. Els recursos consultats han estat tan en format paper com en digital.

Aquesta memòria constitueix un anàlisi dels resultats obtinguts.

A partir del plantejament fet hem realitzat una sèrie de tests d'ajustament de la χ^2 amb les dades mendelianes fent servir el programa R. Gràcies als resultats obtinguts es pot afirmar l'acceptació de H_0 en tots els casos i per tant es corroboren les **Lleis de Mendel**. Les dades pràctiques i teòriques presenten una bona concordança si bé observem particularment que alguns casos s'allunyen d'un bon ajust entre les dades que Mendel mateix justifica com: la falta de germinació d'algunes llavors, la idoneïtat dels experiments i l'especificitat individual d'alguns trets. Finalment s'observa que les dades presenten un comportament massa bo tenint en compte la variabilitat pròpia de l'atzar.

11 Annex

Comandes del R

- Figura 1: Gràfiques $N(\mu, \sigma^2)$

```
x <- seq ( -6, 6, len=100 )
y <- cbind ( dnorm ( x, -2, 1 ),
dnorm ( x, 2, .3 ),
dnorm ( x, .5, 2 ),
dnorm(x, 0, 1))
matplot ( x,y,xlab="x", ylab="Densitat", xaxs="i", yaxs = "i",
main=expression(paste("Distribucions Normals")),type="l", col=
cbind("blue","green","red","yellow" ),lwd=2:2)
legend ( -5, 1.2,paste( "mu =", c(-2,2,.5,0),";
sigma =",c(1,.3,2,1) ),lty=1:4,lwd=2:2, col=cbind("blue","green",
"red","yellow"), cex=.75 )
```

- Figura 2: Gràfiques χ^2_n

```
x=seq(0,40,0.01)
y<-cbind(dchisq(x,6),dchisq(x,11),dchisq(x,30))
matplot(x,y,xlab="x",ylab="Densitat",xaxs="i",yaxs="i",lwd=2:2,
 main=expression(paste("Distribució khi quadrat amb n graus de
llibertat")),type="l",col=cbind("blue","green","red"))
legend(30,.125,paste("n =", c(6,11,30)),lty=1:3,lwd = 2:2,col=
cbind("blue","green","red"),cex=.75)
```

- La primera generació dels híbrids per un únic caràcter
  - Càlcul de $D_n(x)$ per a cada experiment

```
x<-c(5474,1850)
p<-c(3/4,1/4)
chisq.test(x,p=p)
X-squared = 0.26288, df = 1, p-value = 0.6081

x<-c(6022,2001)
p<-c(3/4,1/4)
chisq.test(x,p=p)
X-squared = 0.014999, df = 1, p-value = 0.9025

x<-c(705,224)
p<-c(3/4,1/4)
chisq.test(x,p=p)
X-squared = 0.39074, df = 1, p-value = 0.5319
```

```
x<-c(882,299)
p<-c(3/4,1/4)
chisq.test(x,p=p)
X-squared = 0.063506, df = 1, p-value = 0.801

x<-c(428,152)
p<-c(3/4,1/4)
chisq.test(x,p=p)
X-squared = 0.45057, df = 1, p-value = 0.5021

x<-c(651,207)
p<-c(3/4,1/4)
chisq.test(x,p=p)
X-squared = 0.34965, df = 1, p-value = 0.5543

x<-c(787,277)
p<-c(3/4,1/4)
chisq.test(x,p=p)
X-squared = 0.60652, df = 1, p-value = 0.4361

x<-c(14949,5010)
p<-c(3/4,1/4)
chisq.test(x,p=p)
X-squared = 0.10957, df = 1, p-value = 0.7406\\
```

  - Comandes per la gràfica de la Figura 6

```
x<-seq(0,1,by=0.01)
y<-dchisq(x,1)
plot(x,y,xaxs="i",yaxs="i",xlab="x",ylab="Densitat",main=
expression(paste("Distribució khi quadrat amb 1 dgl")),
type="l",col=4,lwd=2:2)
legend(0.7,3.8,paste("Exp =", c(1,2,3,4,5,6,7,"total")),
lwd=2:2,col=cbind(1,2,3,8,5,6,7,4),cex=.75)

lines(0.26288,dchisq(0.26288,1),type="h",col=1,lwd=2:2)
lines(0.014999,dchisq(0.014999,1),type="h",col=2,lwd=2:2)
lines(0.39074,dchisq(0.39074,1),type="h",col=3,lwd=2:2)
lines(0.063506,dchisq(0.063506,1),type="h",col=8,lwd=2:2)
lines(0.45057,dchisq(0.45057,1),type="h",col=5,lwd=2:2)
lines(0.34965,dchisq(0.34965,1),type="h",col=6,lwd=2:2)
lines(0.60652,dchisq(0.60652,1),type="h",col=7,lwd=2:2)

lines(0.10957,dchisq(0.10957,1),type="h",col=4,lwd=2:2)
```

- La segona generació dels híbrids per un únic caràcter

  - Càlcul de $D_n(x)$ per a cada experiment

```
x<-c(372,193)
p<-c(2/3,1/3)
chisq.test(x,p=p)
X-squared = 0.17345, df = 1, p-value = 0.6771

x<-c(353,166)
p<-c(2/3,1/3)
chisq.test(x,p=p)
X-squared = 0.42486, df = 1, p-value = 0.5145

x<-c(64,36)
p<-c(2/3,1/3)
chisq.test(x,p=p)
X-squared = 0.32, df = 1, p-value = 0.5716

x<-c(71,29)
p<-c(2/3,1/3)
chisq.test(x,p=p)
X-squared = 0.845, df = 1, p-value = 0.358

x<-c(60,40)
p<-c(2/3,1/3)
chisq.test(x,p=p)
X-squared = 2, df = 1, p-value = 0.1573

x<-c(67,33)
p<-c(2/3,1/3)
chisq.test(x,p=p)
X-squared = 0.005, df = 1, p-value = 0.9436


x<-c(72,28)
p<-c(2/3,1/3)
chisq.test(x,p=p)
X-squared = 1.28, df = 1, p-value = 0.2579

x<-c(1059,525)
p<-c(2/3,1/3)
chisq.test(x,p=p)
X-squared = 0.025568, df = 1, p-value = 0.873
```

– Comandes per la gràfica de la Figura 7

```
x<-seq(0,2.2,by=0.01)
y<-dchisq(x,1)
plot(x,y,xaxs="i",yaxs="i",xlab="x",ylab="Densitat",main=
```

```
expression(paste("Distribució khi quadrat amb 1 dgl")),
type="l",col=4,lwd=2:2)
legend(1.5,3.8,paste("Exp =", c(1,2,3,4,5,6,7,"total")),
lwd=2:2,col=cbind(1,2,3,"orange",5,6,7,4),cex=.75)

lines(0.17345,dchisq(0.17345,1),type="h",col=1,lwd=2:2)
lines(0.42486,dchisq(0.42486,1),type="h",col=2,lwd=2:2)
lines(0.32,dchisq(0.32,1),type="h",col=3,lwd=2:2)
lines(0.845,dchisq(0.845,1),type="h",col="orange",lwd=2:2)
lines(2,dchisq(2,1),type="h",col=5,lwd=2:2)
lines(0.005,dchisq(0.005,1),type="h",col=6,lwd=2:2)
lines(1.28,dchisq(1.28,1),type="h",col=7,lwd=2:2)

lines(0.025568,dchisq(0.025568,1),type="h",col=4,lwd=2:2)
```

- La pirmera generació dels híbrids per dos caràcters
  - Genotip
    - Càlcul de $D_n(x)$

```
x<-c(32.75,65,138)
p<-c(1/7,2/7,4/7)
chisq.test(x,p=p)
X-squared = 0.18823, df = 2, p-value = 0.9102
```

    - Comandes per la gràfica de la Figura 8

```
x<-seq(0,8,by=0.01)
y<-dchisq(x,2)
matplot(x,y,xaxs="i",yaxs="i",xlab="x",ylab="Densitat",main=
expression(paste("Distribució khi quadrat amb 2 dgl")),
type="l",col=4,lwd=2:2)
legend(5.5,0.45,paste("Exp =", "genotip"),
lwd=2:2,col=2,cex=.75)

lines(0.18823,dchisq(0.18823,2),type="h",col=2,lwd=2:2)
```

  - Fenotip
    - Càlcul de $D_n(x)$

```
x<-c(301,102,96,30)
p<-c(9/16,3/16,3/16,1/16)
chisq.test(x,p=p)
```

```
X-squared = 0.50557, df = 3, p-value = 0.9177
```

* Comandes per la gràfica de la Figura 9

```
x<-seq(0,8,by=0.º01)
y<-dchisq(x,3)
matplot(x,y,xaxs="i",yaxs="i",xlab="x",ylab="Densitat",main=
expression(paste("Distribució khi quadrat amb 3 dgl")),
type="l",col=4,lwd=2:2)
legend(5.7,0.22,paste("Exp =", "fenotip"),
lwd=2:2,col=2,cex=.75)

lines(0.50557,dchisq(0.50557,3),type="h",col=2,lwd=2:2)
```

- La primera generació dels híbrids per tres caràcters
  - Fenotip
    * Càlcul de $D_n(x)$

```
x<-c(9.625,19,42.67,78)
p<-c(1/15,2/15,4/15,8/15)
chisq.test(x,p=p)
X-squared = 0.29034, df = 3, p-value = 0.9618
```

    * Comandes per la gràfica de la Figura 10

```
x<-seq(0,8,by=0.01)
y<-dchisq(x,3)
matplot(x,y,xaxs="i",yaxs="i",xlab="x",ylab="Densitat",main=
expression(paste("Distribució khi quadrat amb 3 dgl")),
type="l",col=4,lwd=2:2)
legend(5.4,0.22,paste("Exp =", "fenotip"),
lwd=2:2,col=2,cex=.75)

lines(0.29034,dchisq(0.29034,3),type="h",col=2,lwd=2:2)
```

  - Fenotip
    * Càlcul de $D_n(x)$

```
x<-c(269,98,86,88,27,34,30,7)
p<-c(27/64,9/64,9/64,9/64,3/64,3/64,3/64,1/64)
chisq.test(x,p=p)
X-squared = 2.673, df = 7, p-value = 0.9135
```

∗ Comandes per la gràfica de la Figura 11

```
x<-seq(0,8,by=0.01)
y<-dchisq(x,7)
matplot(x,y,xaxs="i",yaxs="i",xlab="x",ylab="Densitat",main=
expression(paste("Distribució khi quadrat amb 7 dgl")),
type="l",col=4,lwd=2:2)
legend(0.7,0.11,paste("Exp =", "fenotip"),
lwd=2:2,col=2,cex=.75)

lines(2.673,dchisq(2.673,7),type="h",col=2,lwd=2:2)
```

Referències

[1] Besalú M.; Rovira C.: *Probabilitats i Estadística*, Publicacions i Edicions de la Universitat de Barcelona, 2013.

[2] Márquez D.; Julià O.: *Un primer curs d'estadística*, Edicions Universitat de Barcelona, 2011.

[3] Peña D.: *Estadística: Modelos y métodos; 1.Fundamentos*, Alianza Editorial, 1989.

[4] Peña D.: *Fundamentos de Estadística*, Alianza Editorial, 2004.

[5] Corcuera J.M.: *Apunts Estadística: Test khi quadrat*, 2011.

[6] C.R.Reeve E: *Encyclopedia of Genetics*, FIZROY DEARBORN PUBLISHERS, 2001.

[7] Mendel G: *Experiments in Plant Hybridization*, **http://www.mendelweb.org/Mendel.html**, 1865.