Capítol 3

TEOREMA DEL LÍMIT CENTRAL

3.1 La mitjana mostral

Volem estudiar com es comporta la mitjana mostral. Considerem $X_1, X_2, \ldots, X_n$ una mostra aleatòria d'una distribució determinada, o sigui n variables aleatòries independents idènticament distribuïdes (v.a.i.i.d), aleshores podem definir la seva mitjana mostral.

Definició 18. *La* **mitjana mostral** $\overline{X}_n$ *d'una mostra aleatòria* $X_1, X_2, \ldots, X_n$ *d'una distribució determinada és una nova variable aleatòria que es defineix com*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Ens plantegem ara quins valors tindran l'esperança i la variància de la mitjana mostral. Com que $X_1, X_2, \ldots, X_n$ són v.a.i.i.d totes tenen la mateixa esperança i variància: $E(X_i) = \mu$ i $Var(X_i) = \sigma^2$ per a $i = 1, \ldots, n$. Utilitzant les propietats de l'esperança i de la variància podem calcular el que volíem:

$$\begin{aligned} E(\overline{X}_n) &= E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu, \\ Var(\overline{X}_n) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Per tant, observem que que l'esperança de la mitjana mostral coincideix amb l'esperança de X_i. Pel que fa a la desviació estàndard de la mitjana mostral (que rep el nom d'**error estàndard de la mostra**) serà $\frac{\sigma}{\sqrt{n}}$.

Observació 14. Si augmentem la mida de la mostra aleshores la variància de $\overline{X}_n$ es fa més petita, o sigui la dispersió de la mitjana mostral al voltant de μ disminueix.

El problema és que necessitem una mica més informació de la mitjana mostral, no en fem prou coneixent la seva esperança i la seva variància. Anem a estudiar en diferents casos com es comporta la mitjana mostral.

3.1.1 Mitjana mostral per a variables normals

Recordem en primer lloc el següent resultat que vam donar en el capítol anterior.

Proposició 8. *Si X i Y són dues variables aleatòries independents amb lleis $N(\mu_1, \sigma_1^2)$ i $N(\mu_2, \sigma_2^2)$ respectivament, aleshores la variable aleatòria*

$$Z = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Observem que aquest resultat es pot estendre al cas que tinguem n variables que segueixin una distribució normal. Aleshores ens quedarà:

Proposició 9. *Si $X_i \sim N(\mu_i, \sigma_i^2)$ per a $i = 1, 2, ..., n$, són n variables aleatòries independents aleshores la variable aleatòria*

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

.

Aleshores, a partir d'aquests resultats obtenim fàcilment el resultat de la següent proposició sobre la distribució de la mitjana mostral.

Proposició 10. *Si $X_1, X_2, ..., X_n$ és una mostra aleatòria simple d'una distribució normal $N(\mu, \sigma^2)$, (això vol dir que $X_i \sim N(\mu, \sigma^2)$) aleshores la mitjana mostral*

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

Aquest resultat és molt interessant però el problema amb què ens trobarem quan el vulguem utilitzar és que en general coneixerem el valor de μ però en canvi, no coneixerem el valor de σ. Veiem ara que com varia el resultat depenent de si el valor de σ és conegut o no ho és.

Cas de σ coneguda.
En el cas en què μ i σ són conegudes podem aplicar simplement la proposició anterior i per tant tenim que

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

Cas de σ desconeguda.
En el 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 mostral:

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

Un com solucionat aquest problema ens trobem que ara el que no coneixem és la distribució de $\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}}$, ja que ara no és cert que segueixi una llei normal. Per això en el cas que el valor de σ sigui desconegut i utilitzem en el seu lloc la desviació mostral farem servir el resultat següent:

Proposició 11. *Si* $X_1, X_2, ..., X_n$ *és una mostra aleatòria simple d'una distribució normal* $N(\mu, \sigma^2)$*, aleshores*

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

on t_{n-1} *és una distribució* t **de Student amb** $n-1$ **graus de llibertat**.

La distribució t_n, t **de Student amb** n **graus de llibertat**, és una distribució:

- Simètrica al voltant del 0 (igual que $N(0,1)$),
- la seva variància és més gran que la d'una $N(0,1)$, $Var(t_n) = \frac{n}{n-2}$,
- per a valors de n grans ($n > 30$) es pot aproximar per una $N(0,1)$.

Exemple 40. *El nivell de glucosa a la sang quan ens aixequem al matí segueix una distribució normal amb mitjana* $\mu = 87$*mg/dL i desviació estàndard* $\sigma = 10$*. Agafem* 150 *persones a l'atzar i els analitzem el nivell de glucosa quan s'aixequen al matí. Quina és la probabilitat que la mitjana observada sigui menor que* 83*?*

La distribució de la mitjana mostral del nivell de glucosa és una distribució normal amb esperança 87 i desviació estàndard $\frac{10}{\sqrt{150}} = 0.8165$, és a dir que

$$\frac{\overline{X}_n - 87}{0.8165} \sim N(0,1).$$

Així hem de calcular

$$P\left(\overline{X}_n < 83\right) = P\left(\frac{\overline{X}_n - 87}{0.8165} < \frac{-4}{0.8165}\right) = P(N(0,1) < -4.8989) = \texttt{pnorm(-4.8989)} = 0.00000048.$$

Exemple 41. *Volem mesurar també el nivell de bilirubina. Sabem que el nivell de bilirubina segueix una distribució normal amb desviació estàndard desconeguda. Agafem ara* 17 *persones a l'atzar i els analitzem el nivell de bilirubina. Si a partir de les dades obtingudes, ens surt una desviació típica mostral de* 0.2*, quina és la probabilitat que la distància entre la mitjana poblacional i la mitjana mostral sigui més gran que* 0.1*?*

Ara en aquest cas tenim que la desviació estàndard és desconeguda i que la desviació mostral és $s = 0.2$. Aleshores fixeu-vos que

$$\frac{\overline{X}_n - \mu}{\frac{0.2}{\sqrt{16}}} = \frac{\overline{X}_n - \mu}{0.05} \sim t_{16}.$$

Ens demanen que calculem

$$\begin{aligned} P\left(\left|\overline{X}_n - \mu\right| > 0.1\right) &= 1 - P\left(-0.1 \leq \overline{X}_n - \mu \leq 0.1\right) \\ &= 1 - P\left(\frac{-0.1}{0.05} \leq \frac{\overline{X}_n - \mu}{0.05} \leq \frac{0.1}{0.05}\right) \\ &= 1 - P\left(-2 \leq t_{16} \leq 2\right), \end{aligned}$$

on t_{16} és una t de Student amb 16 graus de llibertat. Utilitzant

$$P\left(-2 \leq t_{16} \leq 2\right) = P\left(t_{16} \leq 2\right) - P\left(t_{16} < -2\right) = 2 \cdot P\left(t_{16} \leq 2\right) - 1 = \texttt{2*pt(2,16)-1} = 0.9372,$$

obtenim que

$$P\left(\left|\overline{X}_n - \mu\right| > 0.1\right) = 1 - 0.9372 = 0.0628.$$

3.2 Teorema del límit central

El teorema del límit central ens servirà per estudiar el comportament de la suma de n variables aleatòries independents i idènticament distribuïdes per a valors de n grans, sigui quina sigui la distribució de les variables aleatòries. Fixeu-vos que els resultats de l'apartat anterior els podem utilitzar només en el cas en què les variables aleatòries segueixen una distribucíó normal.

Estudiem en primer lloc el cas en què les variables aleatòries segueixin una distribució binomial.

3.2.1 Cas binomial

Sigui X_1, X_2, ..., X_n una mostra aleatòria d'una distribució $\mathrm{Ber}(p)$, si considerem la suma d'aquestes variables aleatòries Bernouilli independents, sabem que $X_1 + \ldots + X_n \sim \mathrm{Bin}(n, p)$. Recordem que l'esperança i la variància d'una $\mathrm{Bin}(n, p)$ eren respectivament $n \cdot p$ i $n \cdot p \cdot (1-p)$.

En aquest cas, tenim el següent resultat

Teorema 3. *Sigui* $X_1, X_2, ..., X_n$ *una mostra aleatòria d'una distribució* $\mathrm{Ber}(p)$. *Aleshores per a tot* $a < b$ *es compleix*

$$\lim_{n\to\infty} P\left(a < \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \leq b\right) = P\left(a < N(0,1) \leq b\right).$$

El que ens està dient aquest teorema és que quan el valor de n ($n > 30$) és prou gran aleshores

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$$

es comporta com una distribució $N(0,1)$, o sigui que a efectes pràctics, si la n és gran podem substituir una llei per l'altra.

Observació 15. Podem reescriure el resultat del teorema anterior de diverses maneres. Sabem que $\sum_{i=1}^{n} X_i$ es distribueix com una $\mathrm{Bin}(n, p)$, però quan n és prou gran, el teorema ens diu que:

- $\dfrac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$ es comporta com una distribució $N(0,1)$.
- $\displaystyle\sum_{i=1}^{n} X_i - np$ es comporta com una distribució $N(0, np(1-p))$.
- $\displaystyle\sum_{i=1}^{n} X_i$ es comporta com una distribució $N(np, np(1-p))$.
- $\overline{X}_n$ es comporta com una distribució $N\left(p, \frac{p(1-p)}{n}\right)$

Exemple 42. *Si tirem un dau perfecte* 1.000 *vegades, quina és la probabilitat que traguem més de* 150 *sisos?*

Considerem la variable aleatòria $X =$ *nombre de sisos obtinguts.* És clar que $X \sim B(1.000, \frac{1}{6})$. Podem per tant calcular la probabilitat que ens demanen, utilitzant la distribució binomial

$$\begin{aligned} P(X > 150) &= 1 - P(X \leq 150) = 1 - \sum_{k=0}^{150} P(X = k) \\ &= 1 - \sum_{k=0}^{150} \binom{150}{k} \left(\frac{1}{6}\right)^k \left(1 - \frac{1}{6}\right)^{1000-k} = 0.916311, \end{aligned}$$

però aquesta quantitat sembla una mica llarga de calcular.

Aleshores, ho podem calcular també utilitzant l'aproximació per la distribució normal: observeu que una distribució $B(1.000, \frac{1}{6})$ es comporta com una distribució

$$N\left(1.000 \cdot \frac{1}{6}, 1.000 \cdot \frac{1}{6} \cdot \frac{5}{6}\right).$$

Així si considerem una variable aleatòria $Y \sim N(\frac{1.000}{6}, \frac{5.000}{36})$, tindrem

$$\begin{aligned} P(X > 150) &\simeq P(Y > 150) = P\left(\frac{Y - \frac{1.000}{6}}{\sqrt{\frac{5.000}{36}}} > \frac{150 - \frac{1.000}{6}}{\sqrt{\frac{5.000}{36}}}\right) \\ &= P(N(0,1) > -1.4142) = \texttt{1-pnorm(-1.4142)} = 0.921349. \end{aligned}$$

3.2.2 Cas general

Acabem de veure un cas particular del teorema del límit central, però la importància d'aquest teorema recau en el fet que es pot utilitzar sigui quina sigui la distribució de la nostra mostra.

Una versió més general del teorema és la següent:

Teorema 4. *Sigui $X_1, X_2, ..., X_n$ una mostra aleatòria d'una distribució amb esperança μ i variància σ^2. Aleshores quan n és prou gran,*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}},$$

es comporta com una distribució normal estàndard.

Observació 16. Igual que em fet en el cas en què hem estudiat la mitjana mostral de variables aleatòries normals, en el cas que no coneguéssim σ, podríem utilitzar la desviació típica mostral i aleshores tindríem que

$$\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}}$$

es comporta com una $N(0,1)$ (com que els valors de n són grans, la t de Student amb n graus de llibertat es comporta com una $N(0,1)$).

Exemple 43. *Un servei tècnic de reparació d'ordinadors tarda una mitjana de 3 hores a arreglar un ordinador, amb una desviació estàndard de 100 minuts. Durant la setmana vinent aquest servei ha d'arreglar 100 ordinadors. Per fer-ho disposa de 7 treballadors, és a dir, de 280 hores de treball. Quina és la probabilitat que els pugui arreglar tots?*

Considerem ara, per a cada ordinador, una variable X_i que ens doni el temps que tarda el servei a arreglar aquest ordinador. Per les dades que ens donen ja sabem que la seva esperança d'aquesta variable serà 180 –posem totes les unitats en minuts– i la seva desviació estàndard 100. No sabem, però, si aquesta variable segueix una distribució normal.

Com que 280 hores representen 16.800 minuts, el que nosaltres volem calcular és

$$P(\sum_{i=1}^{100} X_i \leq 16.800).$$

Per calcular aquesta probabilitat utilitzarem que, segons el teorema del límit central, la variable aleatòria

$$\frac{\sum_{i=1}^{100} X_i - 18.000}{100\sqrt{100}} = \frac{\sum_{i=1}^{100} X_i - 18.000}{1.000}$$

es comporta com una normal estàndard. Aleshores:

$$\begin{aligned} P(\sum_{i=1}^{100} X_i \leq 16.800) &= P\left(\frac{\sum_{i=1}^{100} X_i - 18.000}{1.000} \leq \frac{16.800 - 18.000}{1.000}\right) \\ &= P(N(0,1) \leq -1.2) = \texttt{pnorm(-1.2)} = 0.1150. \end{aligned}$$

És a dir, la probabilitat que es puguin arreglar tots els ordinadors és força petita.