

ESTADÍSTICA BÀSICA

Carles Rovira

Departament d'Estadística



UNIVERSITAT DE BARCELONA



TEXTOS DOCENTS

278

ESTADÍSTICA BÀSICA

Carles Rovira

Departament d'Estadística

Publicacions i Edicions



UNIVERSITAT DE BARCELONA



INTRODUCCIÓ

Un estudi estadístic té bàsicament tres etapes: recollir les dades, descriure-les i analitzar-les (per deduir-ne tanta informació com puguem sobre la població). En aquestes lliçons no parlarem gaire de com recollir-les; parlarem una mica de com descriure-les –**estadística descriptiva**– i ens centrarem en com analitzar-les –**inferència estadística i model de regressió**. Per fer aquesta etapa necessitarem introduir alguns conceptes bàsics de la teoria de la **probabilitat**.

Durant aquestes lliçons s'utilitzen diversos exemples per ajudar a comprendre els conceptes que s'exposen. Com que l'objectiu no és fer un tractat *comme il faut* sinó transcriure un material docent, els exemples són inventats i en la majoria de casos d'interès nul per ells mateixos. Només pretenen ajudar a aclarir els conceptes.

Carles Rovira i Escofet

Departament d'Estadística de la Universitat de Barcelona

ÍNDEX

1	MODELS ESTADÍSTICS	1
2	ESTADÍSTICA DESCRIPTIVA	3
2.1	Dades univariants	3
2.1.1	<i>Representació de dades: histogrames</i>	3
2.1.2	<i>Representació numèrica: estadístics</i>	4
2.1.3	<i>Altres mètodes de representació gràfica: el diagrama de caixa</i>	7
2.1.4	<i>Treball amb dades agrupades</i>	8
2.2	Dades bivariants	9
2.2.1	<i>Taules creuades</i>	9
2.2.2	<i>Representacions gràfiques</i>	9
2.2.3	<i>Mesures de dependència</i>	10
3	PROBABILITATS	11
3.1	Espais de probabilitat	11
3.1.1	<i>Esdeveniments</i>	12
3.1.2	<i>Definició de probabilitat</i>	13
3.1.3	<i>Espais de probabilitats finits</i>	14
3.1.4	<i>Propietats de la probabilitat</i>	15
3.2	Probabilitat condicionada	17
3.3	Esdeveniments independents	18
3.4	La fórmula de Bayes	20
3.5	Combinatòria	23
4	VARIABLES ALEATÒRIES	25
4.1	Variables aleatòries discretes	27
4.1.1	<i>Independència de variables aleatòries discretes</i>	30
4.1.2	<i>Exemples de variables aleatòries discretes</i>	30
4.2	Variables aleatòries absolutament contínues	33
4.2.1	<i>Independència de variables aleatòries absolutament contínues</i>	35
4.2.2	<i>Exemples de variables aleatòries absolutament contínues</i>	35
4.3	Esperança i variància	38
4.3.1	<i>L'esperança</i>	38
4.3.2	<i>Propietats de l'esperança</i>	39
4.3.3	<i>La variància</i>	40
4.3.4	<i>Propietats de la variància</i>	42

4.3.5	<i>Càlcul d'esperances i variàncies</i>	43
4.3.6	<i>La desigualtat de Tsebeixev</i>	44
4.4	La distribució normal	45
4.4.1	<i>Estandarditzar</i>	47
4.5	Simulació de variables aleatòries	49
4.5.1	<i>Variables aleatòries amb funció de distribució contínua</i>	49
4.5.2	<i>Variables aleatòries discretes</i>	50
5	MOSTREIG	51
5.1	Mostreig aleatori simple	51
5.2	Altres mètodes	52
5.2.1	<i>Mostreig estratificat</i>	52
5.2.2	<i>Mostreig per conglomerats</i>	52
5.2.3	<i>Mostreig polietàpic</i>	53
6	TEOREMA DEL LÍMIT CENTRAL	55
6.1	La mitjana mostral	55
6.1.1	<i>Mitjana mostral per a variables normals</i>	56
6.2	Teorema del límit central	58
6.2.1	<i>Cas binomial</i>	58
6.2.2	<i>Cas general</i>	61
6.3	Altres distribucions associades a la distribució normal	62
7	INTRODUCCIÓ A LA INFERÈNCIA ESTADÍSTICA	65
8	ESTIMACIÓ PUNTUAL	67
9	INTERVALS DE CONFIANÇA	71
9.1	Construcció d'interval	71
9.2	Interval de confiança per a una població normal	73
9.3	Interval de confiança per a una proporció	77
9.4	Interval de confiança per a mostres grans	78
9.5	Interval de confiança per a dues mostres	79
10	CONTRAST D'HIPÒTESIS	85
10.1	Teoria general de contrast d'hipòtesis	85
10.2	Contrastos per a una població normal	89
10.3	Contrast sobre una proporció	97
10.4	Contrast sobre la mitjana per a mostres grans	99
10.5	Contrastos per a dues mostres normals	100
11	EL MODEL DE REGRESSIÓ	107
11.1	Càlcul de la recta de regressió	108
11.1.1	<i>Mètode dels mínims quadrats</i>	109
11.2	Qualitat de l'ajustament	112
11.2.1	<i>Relació amb el coeficient de correlació lineal</i>	114

11.2.2 <i>Anàlisi dels residus</i>	114
11.3 Inferència en el model de regressió	115
11.4 Altres models	116

BIBLIOGRAFIA	119
---------------------	------------

CAPÍTOL I

1. MODELS ESTADÍSTICS

En tot estudi estadístic tractarem unes certes dades que volem estudiar i de les quals volem treure tota la informació que puguem. Els primers conceptes que hem de tenir clars són:

1. La **població** objecte del nostre estudi: el conjunt d'individus que volem estudiar. Penseu que segons l'estudi que fem un **individu** pot ser una persona, un iogurt, un avió, una flor, etc.
2. La **mostra** que tenim: és el subconjunt de la població que hem observat i del qual coneixem les dades.
3. Les **variables** que volem estudiar: les característiques –numèriques o no– observades sobre els individus.

En moltes situacions no és possible estudiar tota la població –imagineu que volem estudiar el gust d'un cert tipus de iogurts, no els podem pas tastar tots!– i ens hem de restringir a estudiar-ne una petita part, que és el que hem anomenat *mostra*. La mida de la mostra que utilitzem depèn de molts factors, fonamentalment econòmics. És, però, molt important saber com s'ha escollit, ja que ens cal que la mostra escollida sigui representativa de tota la població (més endavant parlarem de què significa i de com ho podem fer).

Entre les variables que volem estudiar n'hi ha de diversos tipus. De manera una mica simple les podem dividir en dues classes:

1. **Variables qualitatives.** Es refereixen a una característica no numèrica de l'individu. En principi no s'expressen numèricament, però a la pràctica moltes vegades es codifiquen numèricament per facilitar-ne el tractament. Per exemple, el sexe d'un individu és una variable qualitativa que pot prendre dos valors: M (masculí) i F (femení), però moltes vegades es codifica amb un 1 (M) o un 0 (F).

2. **Variables quantitatives.** Són les que es refereixen a característiques dels individus que s'expressen numèricament. Dintre d'aquestes en podem trobar de **discretes** –quan només poden prendre un nombre discret de valors– o de **contínues** –quan poden prendre qualsevol valor dins d'un interval.

Exemple. Considerem com a població d'un estudi tots els alumnes de la universitat. Una manera d'escollir una mostra podria ser la següent: agafem els 50 primers alumnes que passen per davant de secretaria un dilluns al matí. És, però, aquesta mostra representativa de la població, és a dir, de tots els alumnes? Algunes variables que podem estudiar són: sexe (qualitativa), nombre de fills (quantitativa discreta) o alçada (quantitativa contínua).

Donem un altre exemple. Considerem com a població els avions de passatgers del món. Una manera d'agafar una mostra podria ser la següent: agafem un avió a l'atzar de cada companyia. Algunes variables: antiguitat (quantitativa discreta), model (qualitativa) o quilòmetres recorreguts el darrer any (quantitativa contínua).

En general tindrem les observacions de les variables que volem estudiar per als individus de la nostra mostra. Aquestes observacions seran un conjunt de dades que serà amb el que treballarem.

CAPÍTOL X

10. CONTRAST D'HIPÒTESIS

Introduïrem ara una nova tècnica estadística: el contrast d'hipòtesis. El concepte no és gaire difícil: suposarem una certa hipòtesi i a partir de les nostres observacions haurem de decidir si aquesta hipòtesi és acceptable o si hi ha prou evidències per rebutjar-la.

Tornem a recordar el nostre marc general. Tenim una mostra aleatòria simple $X_1, \dots, X_n$ amb una distribució desconeguda que sabem que pertany a una família de distribucions $\{P_\theta, \theta \in \Theta\}$. Nosaltres tenim de nou només una observació d'aquesta mostra $x_1, \dots, x_n$.

En aquest marc, plantejar una hipòtesi vol dir només fer una conjectura sobre el valor del paràmetre θ . L'objectiu és, a partir de les nostres observacions $x_1, \dots, x_n$, dir si hi ha o no hi ha motius per rebutjar-la.

Exemple. Llançem 100 vegades una moneda i ens surten 45 cares i 55 creus. Volem saber si la moneda és perfecta. Cada llançament de la moneda el podem considerar como una realització d'una variable $B(p)$, on p és la probabilitat de treure una cara. Tenim per tant $X_1, \dots, X_{100}$ una mostra aleatòria simple d'una distribució $B(p)$ i volem saber si $p = \frac{1}{2}$. És evident que si ens haguessin sortit 50 cares i 50 creus diríem que sí, però amb la relació 45 – 55 obtinguda, què hem de fer?

Exemple. Imagineu que tenim una mostra aleatòria simple $X_1, \dots, X_n$ d'una distribució $N(\mu, 16)$. Una possible hipòtesi és plantejar-nos si a partir de les observacions $x_1, \dots, x_n$ podem acceptar que $\mu = 23$. Si per exemple les nostres observacions són 40, 42, 37 i 45 ja es veu clarament que la μ no serà 23. Malauradament, a la vida, les coses no són tan clares i ens cal una tècnica que ens permeti donar una resposta amb un cert rigor.

Una qüestió semàntica: moltes vegades s'utilitzen els mots contrast i test indistintament.

10.1 Teoria general de contrast d'hipòtesis

Veurem ara els conceptes bàsics dels contrastos d'hipòtesis. Anirem desenvolupant alhora l'exemple següent per tal d'anar aclarint els conceptes nous.

Exemple de l'alçada. Tornem al nostre exemple típic de l'alçada. Suposem ara que l'alçada segueix una distribució normal i que sabem segur que la desviació estàndard val 4. Tenim, per tant, una mostra aleatòria simple $X_1, \dots, X_n$ d'una distribució $N(\mu, 16)$. Creiem a més que la $\mu = 175$. Les nostres observacions són:

180 165 168 192 195 187 181 177 175 186,

i a partir d'aquestes observacions hem de dir si podem acceptar que la $\mu = 175$ o si hi ha prou evidències per dir que no és cert.

El primer pas és plantejar les hipòtesis.

Definició. La **hipòtesi nul·la**, H_0 , és la hipòtesi de sortida. Ens indica que les coses són com s'esperava que fossin –és a dir, la situació usual– i que la variabilitat observada és només una qüestió de casualitat.

La **hipòtesi alternativa**, H_1 , és la que ens indica que la hipòtesi esperada no era certa. Indica, per tant, que hi ha un canvi i que les observacions reflecteixen aquest canvi.

Plantejades les hipòtesis, ara, a partir de les nostres observacions, haurem de prendre una decisió. N'hi ha dues de possibles: o rebutgem la hipòtesi nul·la o no la rebutgem. També hi ha, però, dues possibles situacions: la hipòtesi nul·la pot ser certa o falsa. Tot això ho podem presentar a la taula següent:

	No rebutjar H_0	Rebutjar H_0
H_0 certa	Decisió encertada	Error de tipus I
H_0 falsa	Error de tipus II	Decisió encertada

Fixeu-vos que tenim, per tant, dos tipus d'errors: l'error que fem en rebutjar la hipòtesi nul·la quan aquesta és certa –error de tipus I– i l'error que fem en no rebutjar-la quan és falsa –error de tipus II.

En el nostre plantejament la hipòtesi nul·la representa l'*statu quo*, el que es creu que és cert. Per tant, si ens equivoquem quan no rebutgem la hipòtesi nul·la quan aquesta és falsa, l'error que fem no és greu. L'error greu, des del nostre punt de vista, és rebutjar la hipòtesi nul·la quan aquesta és certa, ja que, en aquest cas, estem detectant uns canvis en la situació de sortida que en realitat no existeixen. Aquest raonament traduït en termes d'errors ens diu que ens interessarà que l'error de tipus I estigui ben controlat.

Fixarem també l'anomenat *nivell de significació*.

Definició. El **nivell de significació**, α , és l'error màxim de tipus I que estem disposats a acceptar.

Determinar el nivell d'error no és un problema matemàtic, ja que depèn normalment de consideracions econòmiques. Normalment, però, s'utilitzen nivells de significació de 0,1; 0,05 o 0,01.

Exemple de l'alçada. Segons el nostre plantejament hem de considerar les hipòtesis

$$H_0 : \mu = 175 \quad \text{contra} \quad H_1 : \mu \neq 175.$$

Nosaltres creiem que la $\mu = 175$, així que ho posem com a hipòtesi nul·la. Només rebutjarem la hipòtesi nul·la i per tant ens decantarem per l'alternativa $H_1 : \mu \neq 175$, quan hi hagi evidències prou clares que ens indiquin que la hipòtesi nul·la és falsa!

Fixarem un nivell de significació $\alpha = 0,05$. Amb això indicarem que, si decidim rebutjar la hipòtesi nul·la, la probabilitat que ens estiguem equivocant volem que sigui més petita que α . Dit d'una altra manera, que de cada 100 vegades que decidim rebutjar la hipòtesi nul·la només ens equivocarem en 5. No sabem però la probabilitat d'equivocar-nos quan decidim acceptar la hipòtesi nul·la.

El pas següent consisteix a determinar un estadístic de contrast.

Definició. Un **estadístic de contrast** és un estadístic del qual coneixem la distribució si la hipòtesi nul·la H_0 és certa.

A partir de les nostres observacions tindrem un valor concret de l'estadístic de contrast. Aleshores, si aquest valor és molt poc probable –lligat al valor del nivell de significació– amb la distribució de l'estadístic sota la hipòtesi nul·la H_0 , deduirem que l'error està en el fet de suposar que la hipòtesi nul·la H_0 és certa i, per tant, la rebutjarem. D'altra banda, si el valor observat era bastant probable amb la distribució de l'estadístic sota la hipòtesi nul·la H_0 , podem pensar que ens ha sortit el que esperàvem i que per tant no hi ha prou evidències per rebutjar la hipòtesi nul·la.

Exemple de l'alçada. En el nostre cas podem agafar com a estadístic de contrast

$$\frac{\bar{X}_n - 175}{\frac{\sigma}{\sqrt{n}}}$$

(recordem que $\sigma = 4$ i $n = 10$), que sabem que sota la hipòtesi nul·la H_0 té una distribució normal estàndard. Fixeu-vos que a partir de les nostres dades obtenim un observació d'aquest estadístic

$$z = 4,4271.$$

Ara, creiem que aquest valor és un valor “prou probable” per a una normal estàndard?

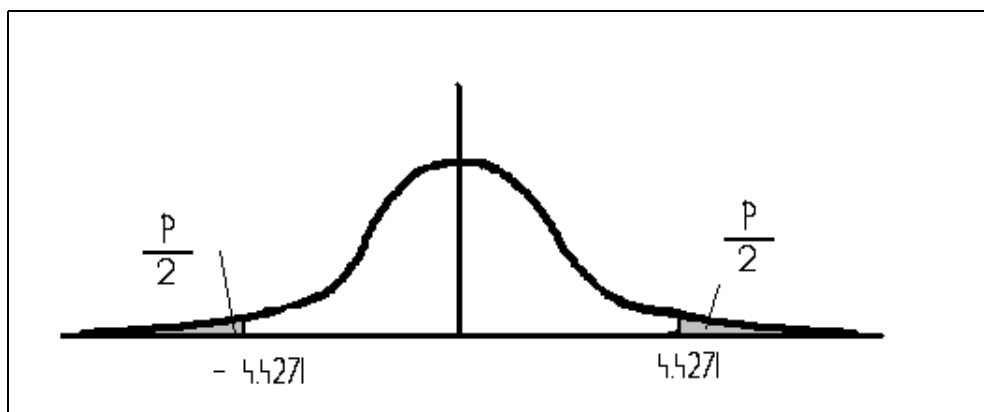
Hem d'aclarir, però, què entendrem per *prou probable*. Això ho farem a partir del concepte de p -valor.

Definició. El **p -valor** és la probabilitat del resultat de l'estadístic de contrast observat o d'un de més allunyat quan la hipòtesi nul·la és certa.

Fixeu-vos que el fet que el p -valor sigui petit ens indica que ens ha sortit un valor poc probable. Això ens fa pensar que la hipòtesi nul·la és falsa i per tant decidim rebutjar-la. Evidentment ens podem equivocar, ja que el valor encara que sigui improbable pot sortir. Però observeu que la probabilitat que ens equivoquem és el p -valor. I recordeu que volíem que aquesta probabilitat fos menor que el nivell de significació.

Quan el p -valor sigui més petit que el nivell de significació α rebutjarem la hipòtesi nul·la. Altrament, no la rebutjarem.

Exemple de l'alçada. Fixeu-vos en el gràfic que la probabilitat de l'estadístic de contrast observat o d'un de més allunyat quan la hipòtesi nul·la és certa és, si Z és una normal estàndard,



$$p = P(Z < -4.4271) + P(Z > 4.4271) = 0.0000094.$$

Com que és més petit que α (recordeu que $\alpha = 0.05$) haurem de rebutjar la hipòtesi nul·la.

En resum, realitzar un contrast d'hipòtesis té diverses etapes:

1. Plantejament:

fixar hipòtesis nul·la i alternativa,
fixar nivell de significació.

2. Resolució:

determinar l'estadístic de control i calcular el seu valor observat,
determinar el p -valor,
prendre una decisió, comparant el p -valor amb el nivell de significació.

Per acabar la teoria general de contrast d'hipòtesis cal fer un darrer comentari. Hem vist un exemple que amb una distribució normal fèiem un contrast del tipus

$$H_0 : \mu = 175 \quad \text{contra} \quad H_1 : \mu \neq 175.$$

Aquest contrast s'anomena **bilateral**. Podem considerar però també altres hipòtesis alternatives:

1. Si creiem que la mitjana és igual o més petita que 175 i en fer el contrast només volem rebutjar la hipòtesi nul·la si hi ha evidències que la mitjana és realment superior a 175, hem de plantejar el contrast

$$H_0 : \mu = 175 \quad \text{contra} \quad H_1 : \mu > 175.$$

2. Si creiem que la mitjana és igual o més gran que 175 i en fer el contrast només volem rebutjar la hipòtesi nul·la si hi ha evidències que la mitjana és realment inferior a 175, hem de plantejar el contrast

$$H_0 : \mu = 175 \quad \text{contra} \quad H_1 : \mu < 175.$$

Aquests dos contrastos s'anomenen **unilaterals**.

10.2 Contrastos per a una població normal

Considerarem en aquesta secció que tenim una mostra aleatòria simple $X_1, \dots, X_n$ que prové d'una distribució $N(\mu, \sigma^2)$. Els estadístics de contrast que apareixeran seran els mateixos que sortien com a funcions pivotants en el càlcul d'interval de confiança.

Contrast sobre la mitjana –cas de variància σ^2 coneguda. El test més senzill és quan la nostra hipòtesi nul·la és que la mitjana val un valor μ_0 determinat contra que no ho sigui quan la variància és coneguda –és el cas de l'exemple de l'alçada. És a dir,

$$H_0 : \mu = \mu_0 \quad \text{contra} \quad H_1 : \mu \neq \mu_0.$$

L'estadístic de contrast que utilitzem en aquest test és

$$\frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}},$$

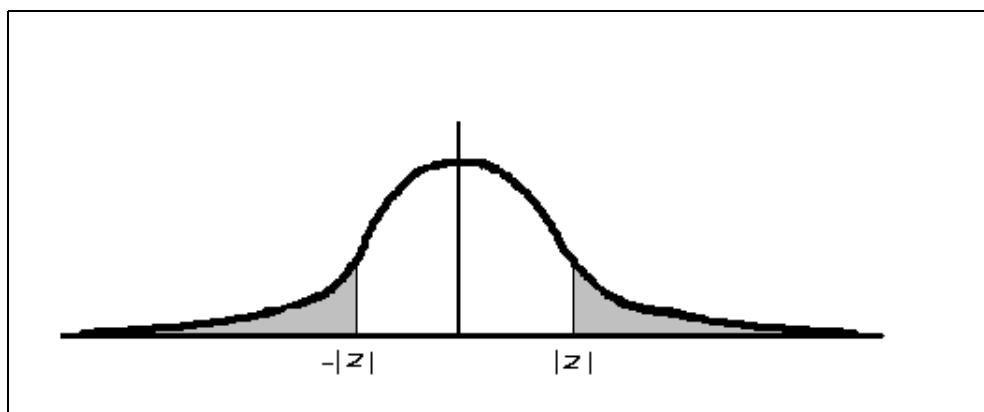
que sabem que sota la hipòtesi nul·la H_0 segueix una distribució normal estàndard. Tenim, per tant, l'observació de l'estadístic de contrast

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

Aleshores el p -valor el podem calcular com

$$p = P(Z > |z|) + P(Z < -|z|) = 2P(Z > |z|),$$

on ara Z segueix una distribució normal estàndard.



També es poden fer les versions unilaterals d'aquest test, en les quals fem variar la hipòtesi alternativa. És el cas en què considerem les hipòtesis

$$H_0 : \mu = \mu_0 \quad \text{contra} \quad H_1 : \mu < \mu_0,$$

o bé

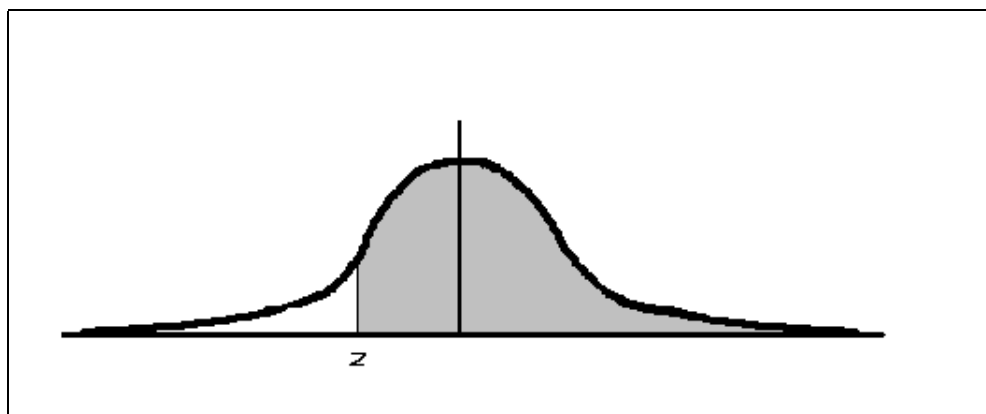
$$H_0 : \mu = \mu_0 \quad \text{contra} \quad H_1 : \mu > \mu_0.$$

L'estadístic de contrast és el mateix que en el cas bilateral però el p -valor es calcula d'una manera diferent en cada cas.

- Quan $H_1 : \mu > \mu_0$ aleshores

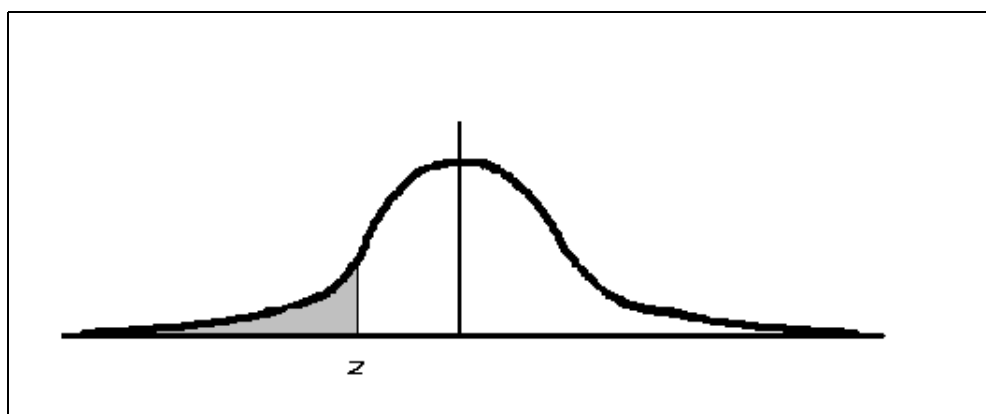
$$p = P(Z > z)$$

(observeu que si la $\mu < \mu_0$ aleshores esperem obtenir valors z negatius).



- Quan $H_1 : \mu < \mu_0$ aleshores

$$p = P(Z < z).$$



Exemple de l'alçada. Recordem que l'estadístic de contrast era

$$\frac{\bar{X}_n - 175}{\frac{\sigma}{\sqrt{n}}}$$