Máster Sector Farmacéutico

Estadística aplicada

Teresa Villagarcía

Indice

Probabilidad

- ¿Por qué estudiamos probabilidad?
  - Ejemplo: Proporción de piezas/pastillas defectuosas producidas en un proceso industrial. Tomamos una muestra y obtenemos que el 3% es defectuoso
  - A continuación tomamos otra y resultan defectuosas el 2.5%. ¿Ha mejorado el proceso?
  - Si a continuación obtenemos un 3,2% ….¿Debemos parar el proceso productivo y hacer un mantenimiento?

Probabilidad

- Sabemos que cuando el proceso funciona bien produce pastillas de peso medio 2 y desviación 0.1. ¿Cuántas salen defectuosas?

- Tomamos una muestra y el peso medio de las pastillas es 2.1 ¿Ahora el proceso funciona bien? ¿Lo arreglamops?

Probabilidad

- Sabemos que si nuestro servicio funciona bien recibimos en promedio 156 reclamaciones al mes.
  - Este mes hemos recibido 150. ¿Hemos mejorado?
  - El mes pasado recibimos 160. ¿Hemos empeorado?
  - Si recibimos 180 ¿qué debemos hacer?

- Cuando todo funciona bien nuestra máquina produce un 2% de piezas defectuosas.
  - Hoy hemos producido 1000 piezas y ha habido 23 piezas defectuosas. ¿Está mal la máquina?

La probabilidad es el arma que vamos a utilizar para poder generalizar nuestras conclusiones a toda la población de referencia y contestar a ese tipo de preguntas

Propiedades de la probabilidad

- Probabilidad está comprendida entre 0 y 1.
- O entre 0 y 100

Todos tenemos una idea intuitiva de lo que es la probabilidad:

Sirve para medir incertidumbres.

¿Es probable que mañana haga 40ºC?

¿Y que yo termine la clase afónica?

Variables Aleatorias

- Son las variables numéricas que pueden tomar valores al azar:
- Ejemplos:
  - Nº de reclamaciones al mes en un servicio
  - Nº de servicios prestados
  - Nº personas que hay en un hotel.
  - Retrasos promedio en los trenes
  - Peso pastilla
  - % de componente

Variables Aleatorias

- Pueden ser
  - Proporciones (Muchas vienen de variables cualitativas)
    - % de componente activo en una pastilla
    - % Pastillas defectuosas en un lote (viene de cualitativa)
  - Continuas
    - Pueden tomar cualquier valor
      - Retrasos del tren: puede ser 5.34 minutos

¿Cómo se estudian las Variables Aleatorias?

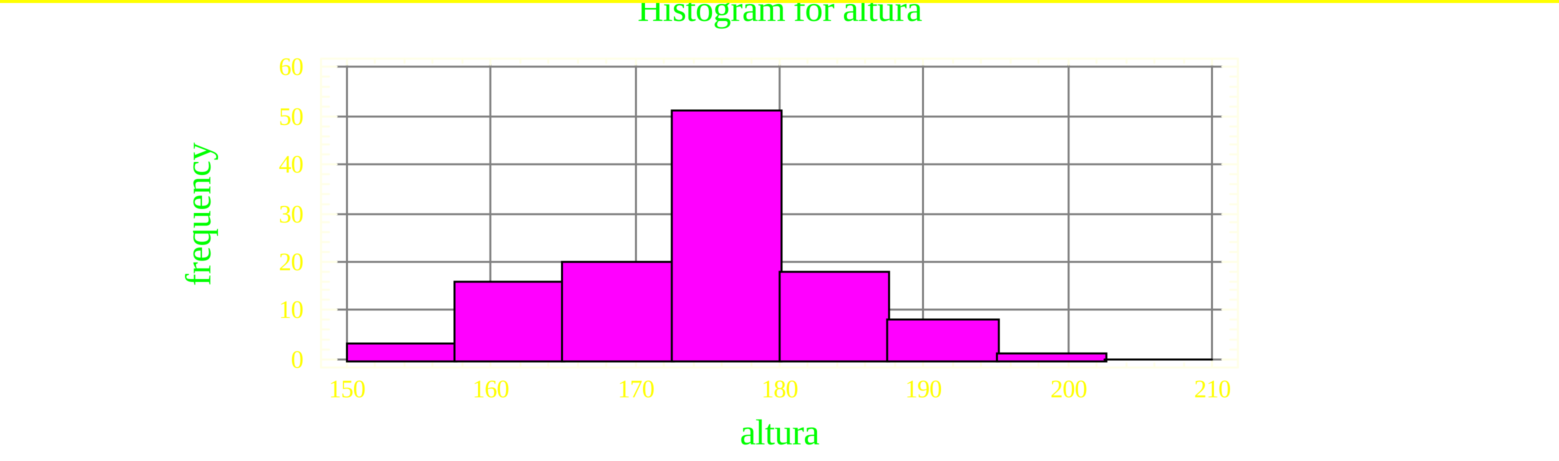
- Como pueden tomar valores al azar sólo se pueden estudiar mediante probabildades

- Definir en cada problema los valores y las probabilidades puede ser muy complejo.
  - Ejemplo: Niños y niñas en familias de cuatro hijos

- Lo simplificamos mediante distribuciones de probabilidad
  - Continuas (y proporciones): Distribución normal

Distribución normal

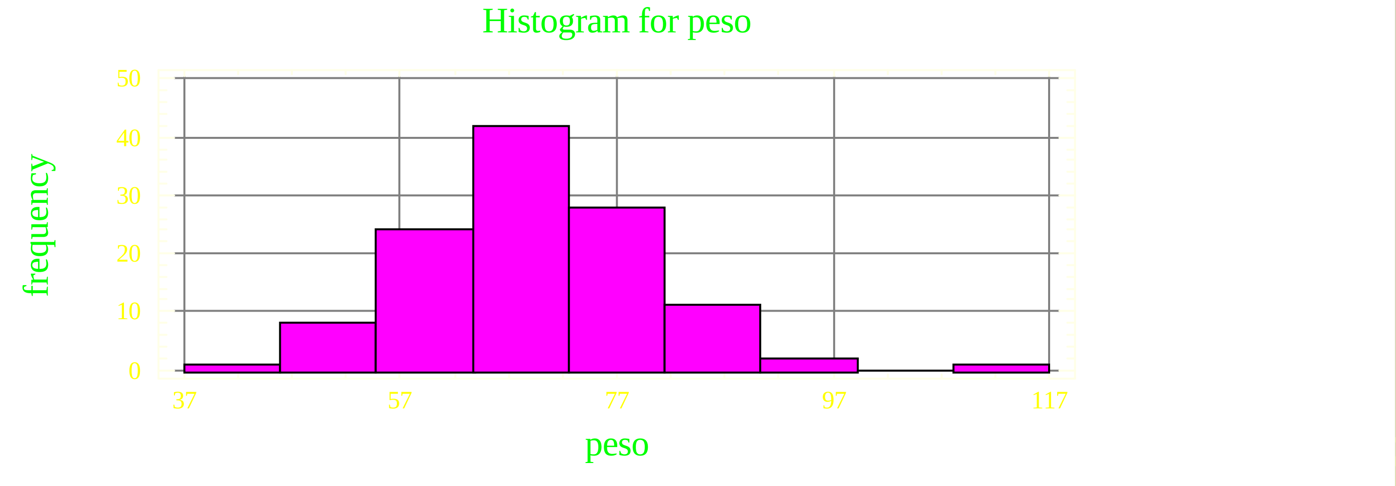
¿En qué rango es más probable encontrar gente?



¿Cómo será la altura en general?

Distribución normal

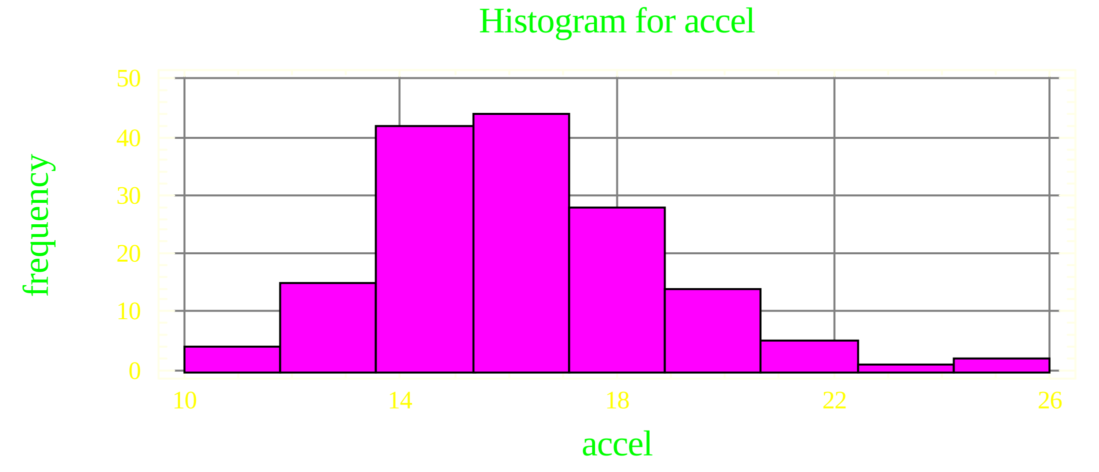
¿En qué rango es más probable encontrar gente?



¿Cómo será el peso en general?

Distribución normal

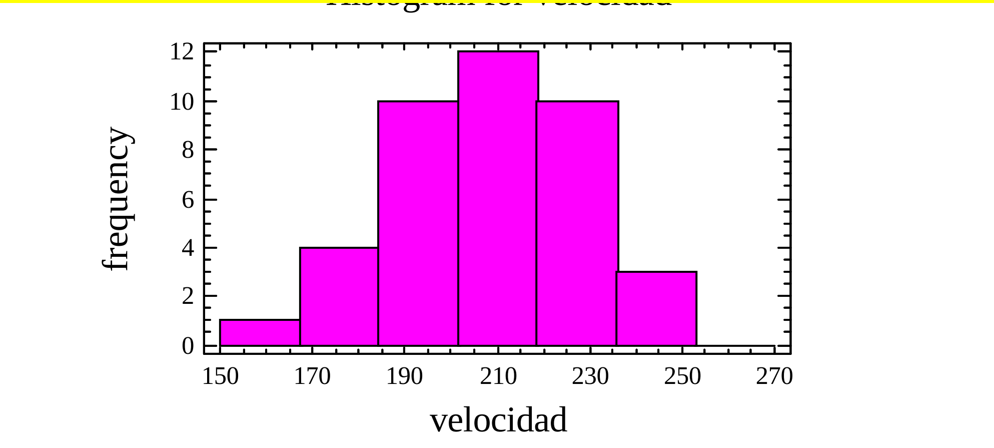
¿En qué rango es más probable encontrar coches?



¿Cómo será en general la aceleración?

Distribución normal

¿En qué rango es más probable encontrar coches?

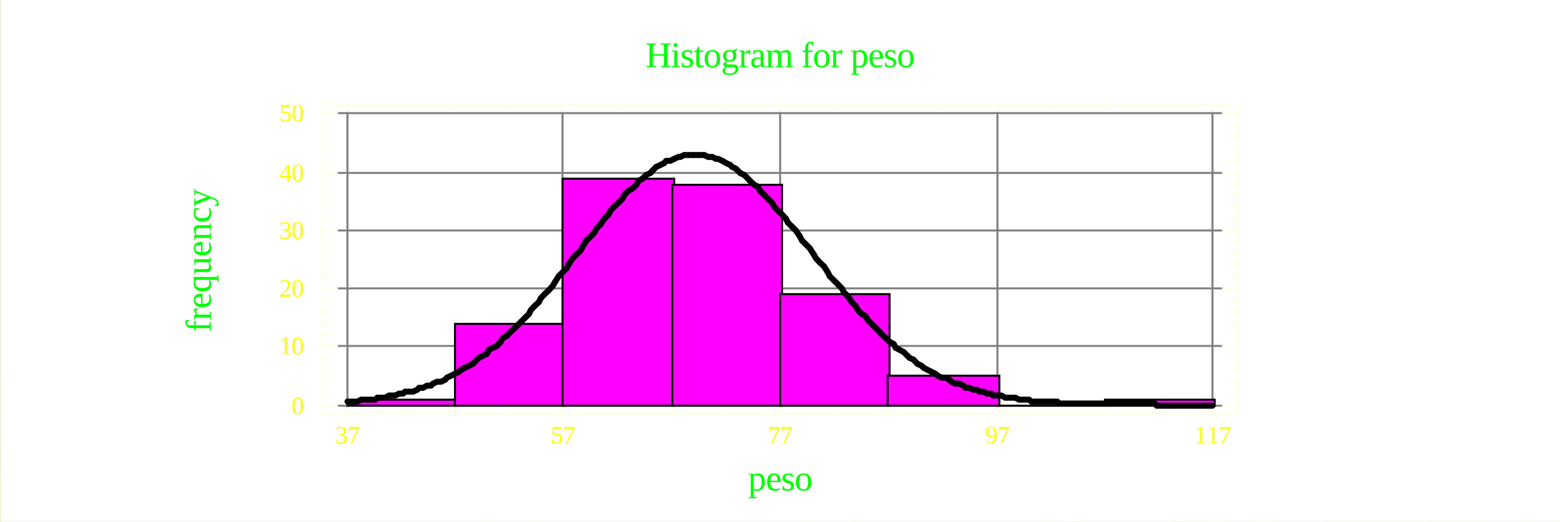


¿Cómo será en general al velocidad?

Distribución normal

Queremos encontrar una curva que represente los datos y

Usarla para generalizar nuestras conclusiones.



La probabilidad va ligada al área de la curva.

Distribución Normal

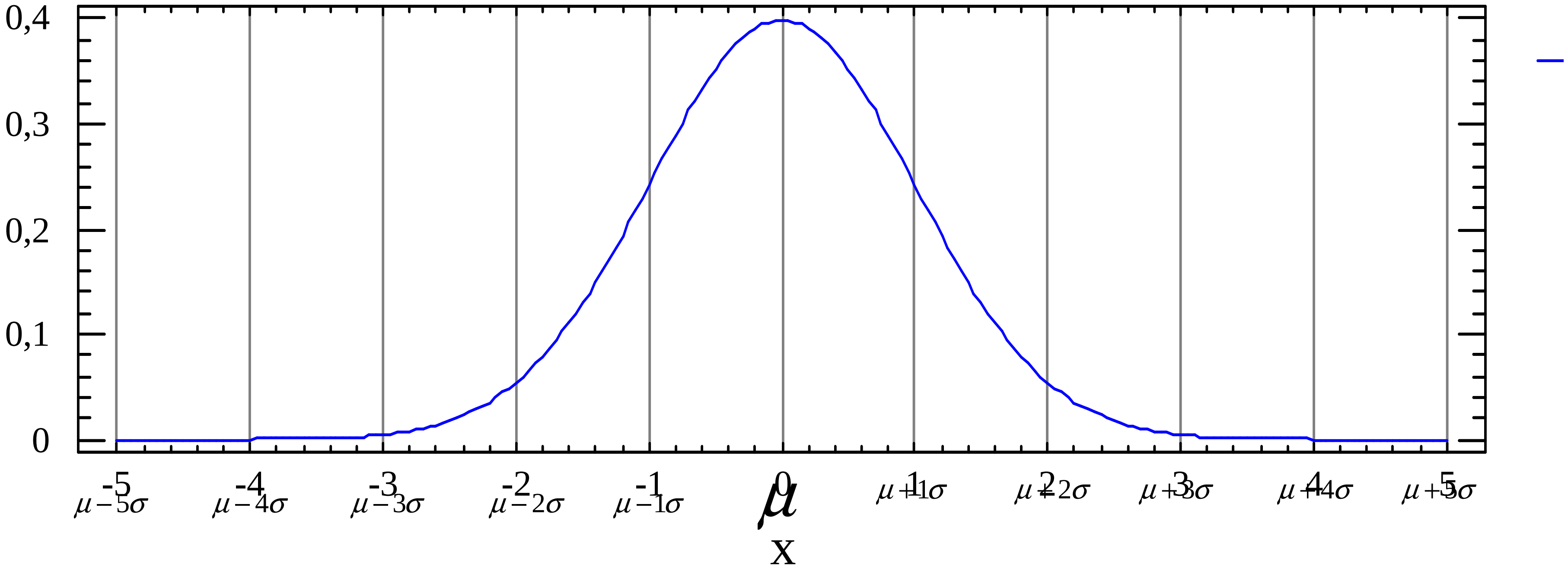
- La distribución normal se llama también CAMPANA DE GAUSS.
- El área que abarca bajo ella es 1
- Se caracteriza por dos valores: Media μ y la desviación típica σ
- La normal se escribe:

$$N(\mu,\sigma)$$

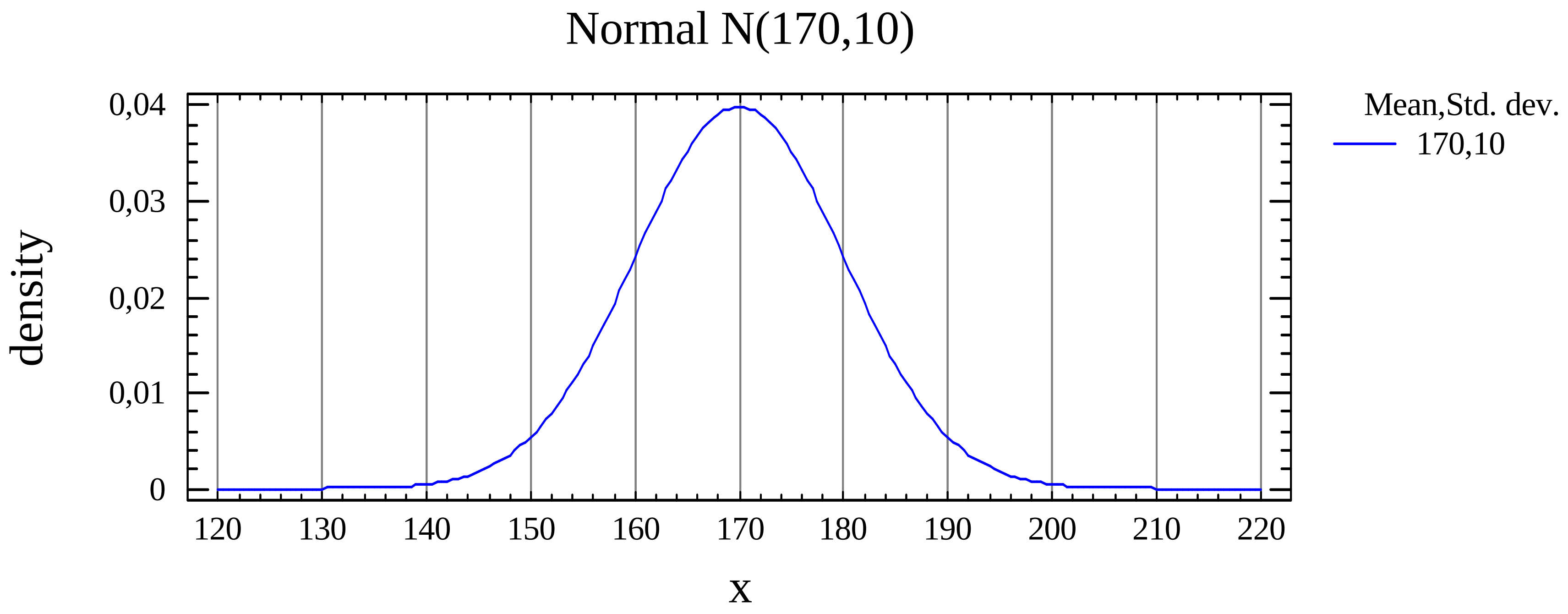
Dibujar Normales

- Entre la media y 2 desviaciones a cada lado abarca el 95% de probabilidad
- Entre la media y 3 desviaciones abarca el 99,7%

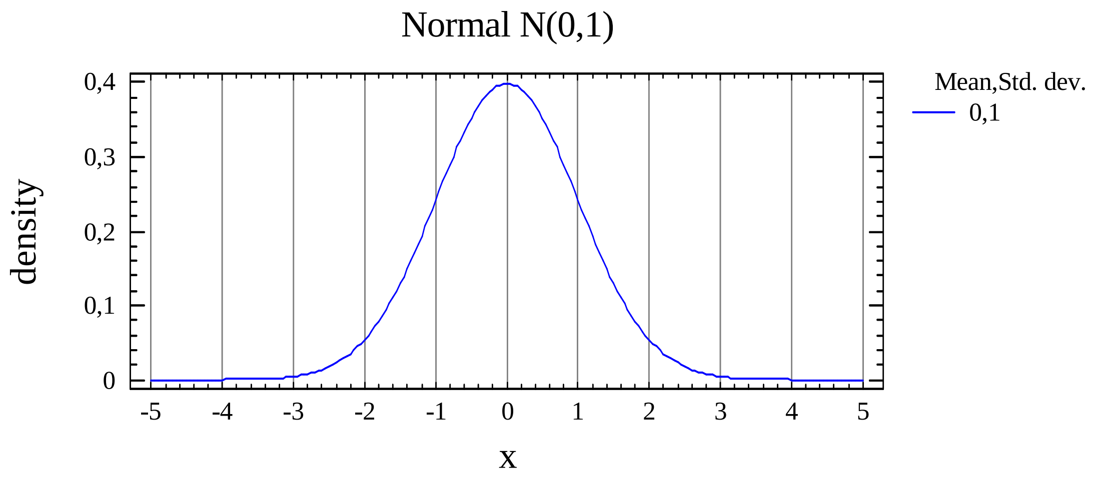
Distribución Normal



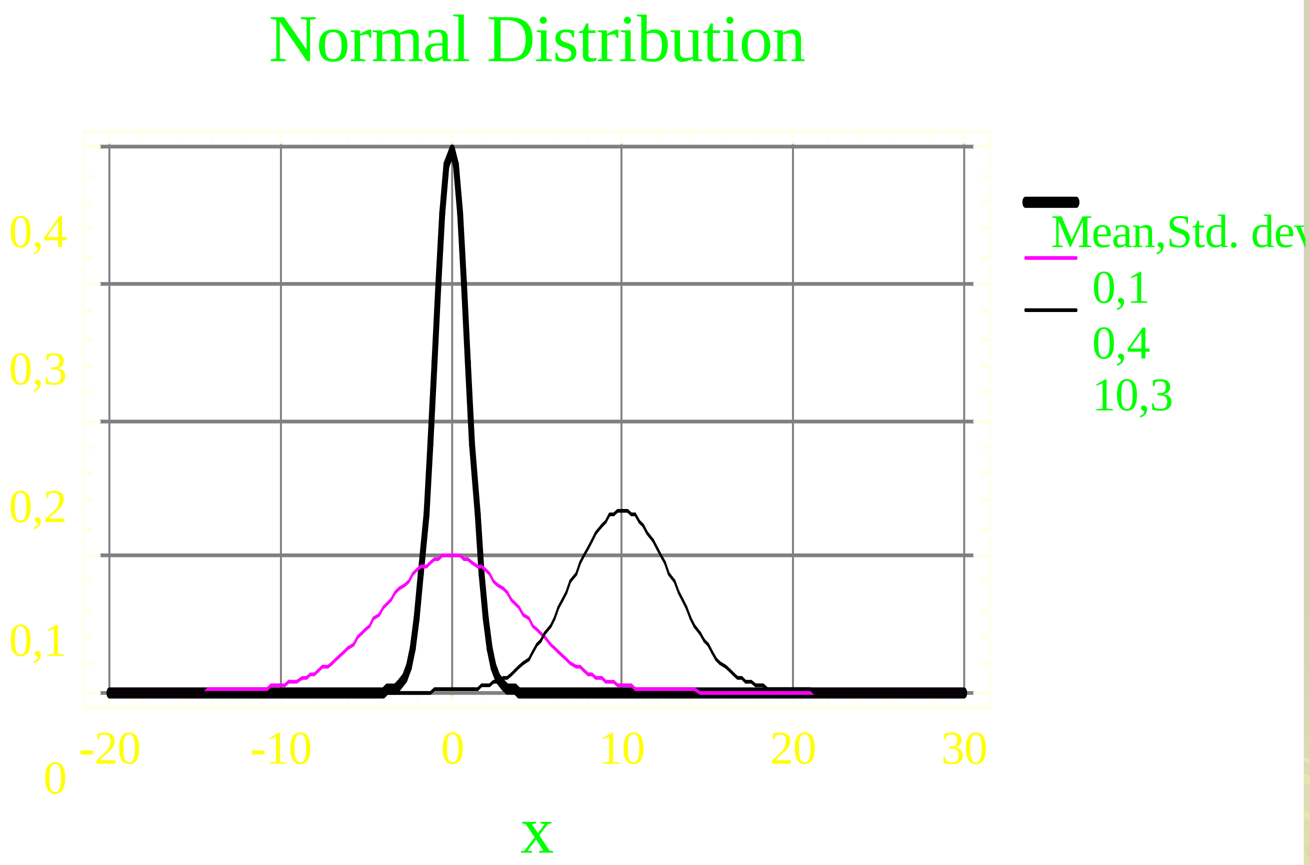
Distribución Normal



Distribución Normal

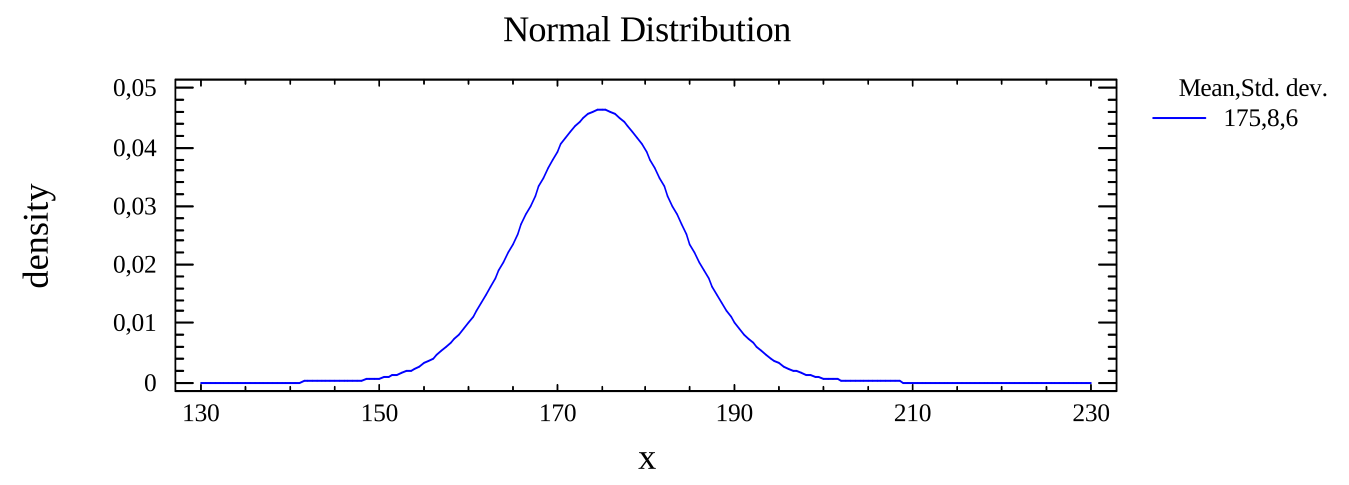


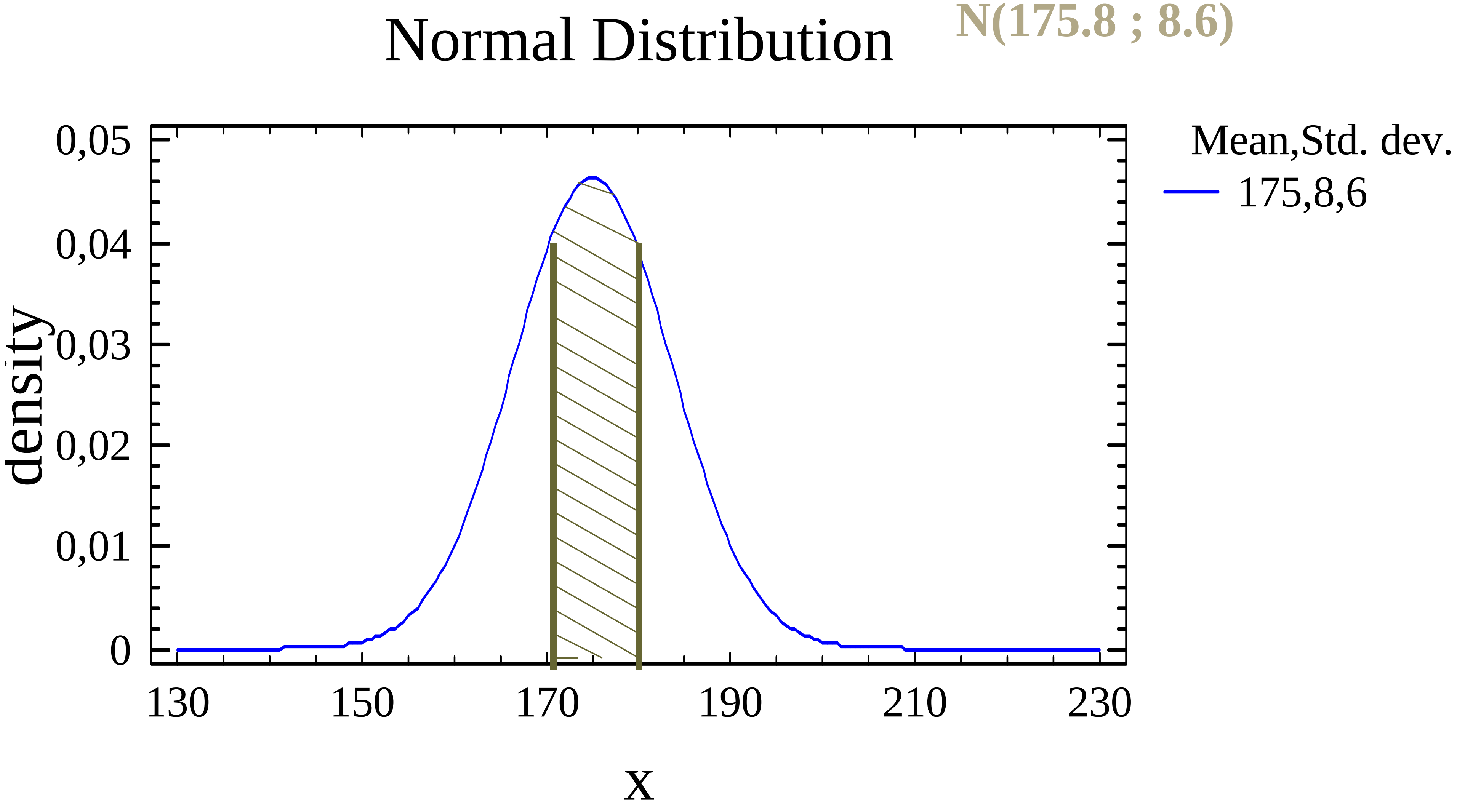
Distribución Normal



Distribución Normal

El área comprendida bajo la curva normal es la probabilidad de encontrar observaciones en ese intervalo.

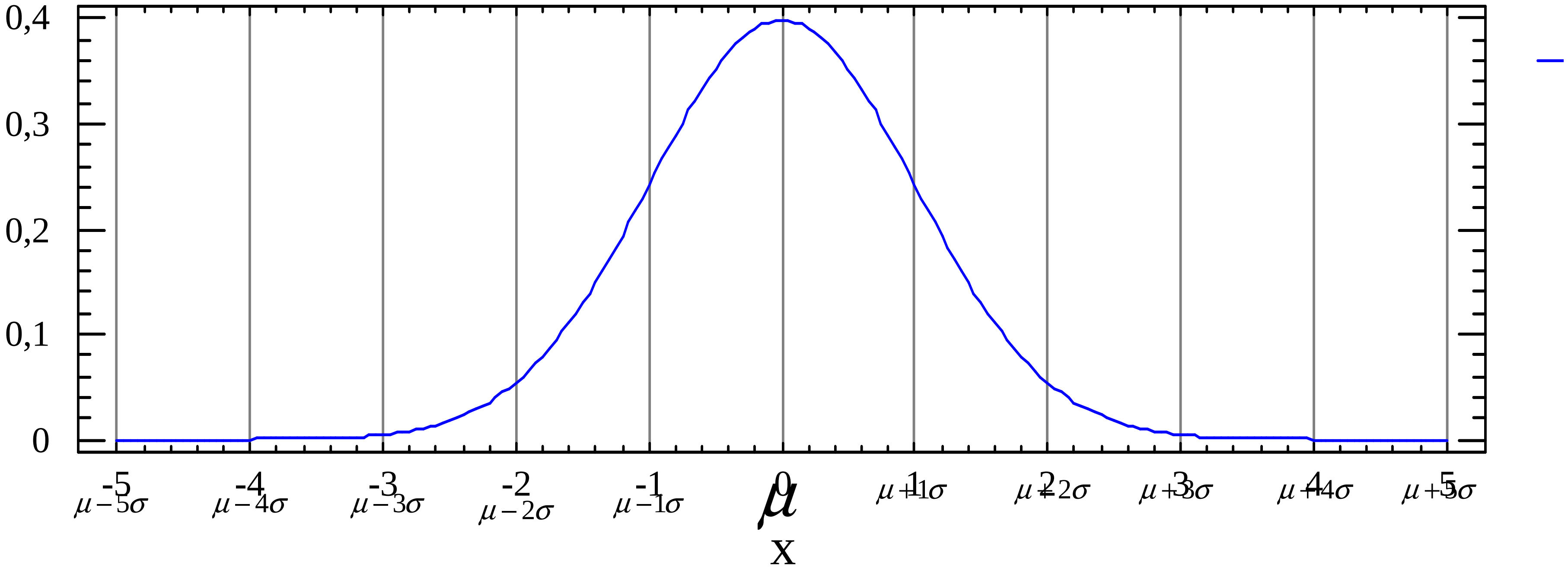




N(175.8 ; 8.6)

P(170<X<180)=Area bajo la curva.   Lo calcula la máquina

Distribución Normal: Algunas probabilidades

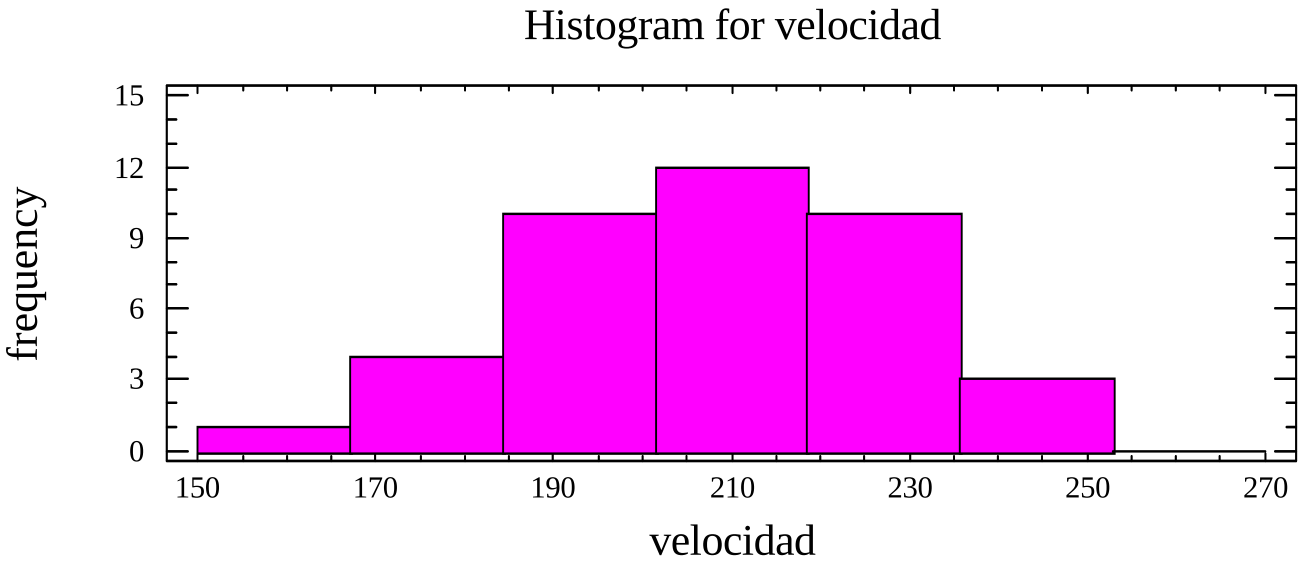


Entre $\mu-2\sigma$ y $\mu+2\sigma$ hay un 95% de probabilidad

Entre $\mu-3\sigma$ y $\mu+3\sigma$ hay un 99.7% de probabilidad

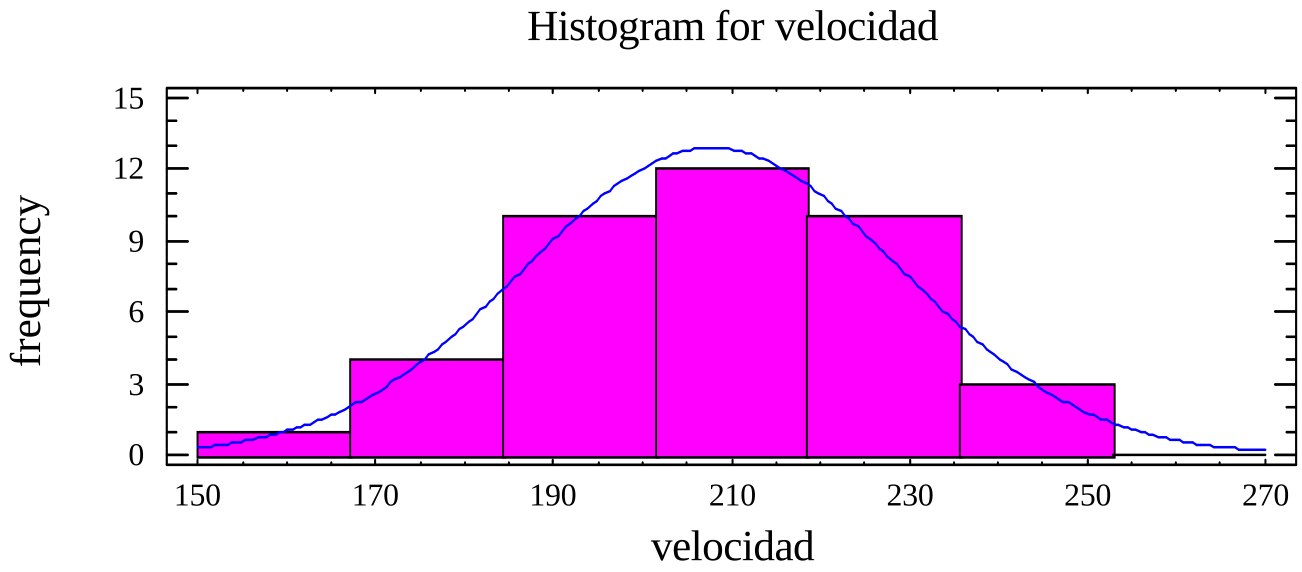
Cálculo de probabilidades con la normal

- El objetivo de las distribuciones es AJUSTARLAS a los datos.
- Este proceso se llama estimación. Y consiste en encontrar la normal que mejor ajusta a los datos:



Cálculo de probabilidades con la normal

- El objetivo de las distribuciones es AJUSTARLAS a los datos.
- Este proceso se llama estimación. Y consiste en encontrar la normal que mejor ajusta a los datos:



Se puede demostrar

- La normal que mejor ajusta a los datos es

$$N(\mu,\sigma) = N(Media, desviación)$$

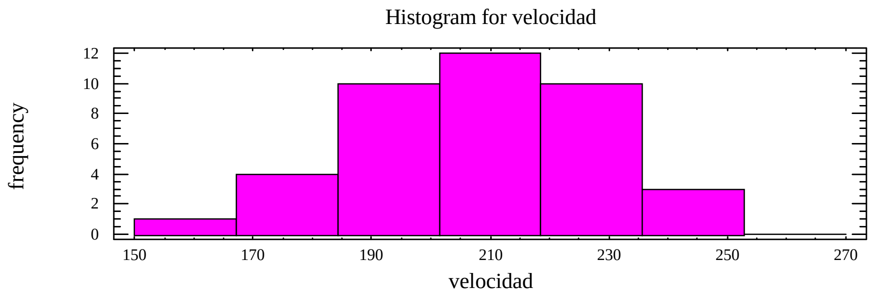
Ejemplo velocidad

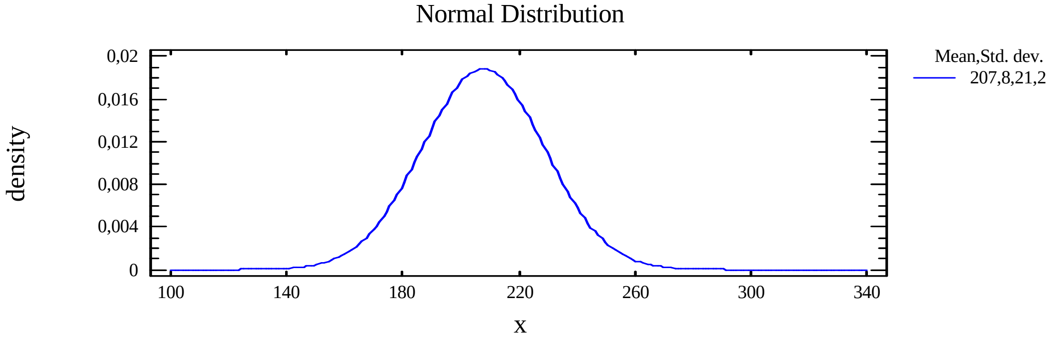
- La velocidad de una muestra de coches bastante deportivos, tiene un valor medio
- Media=207,82
- Desviación típica=21,219

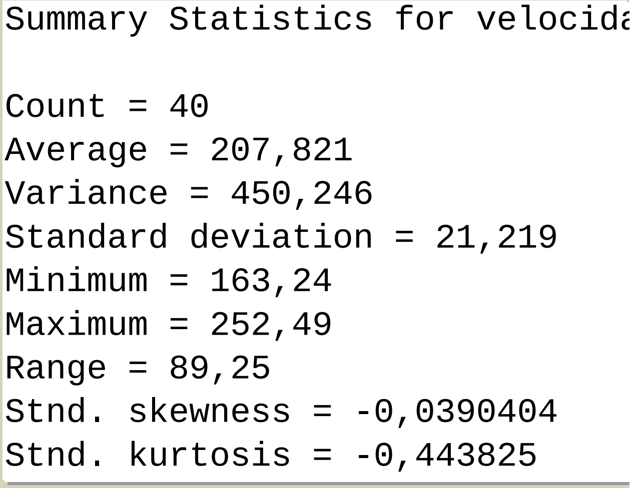
- POR TANTO LA NORMAL QUE AJUSTA ESTOS DATOS ES:

$$N(207.8\ ;\ 21.2)$$

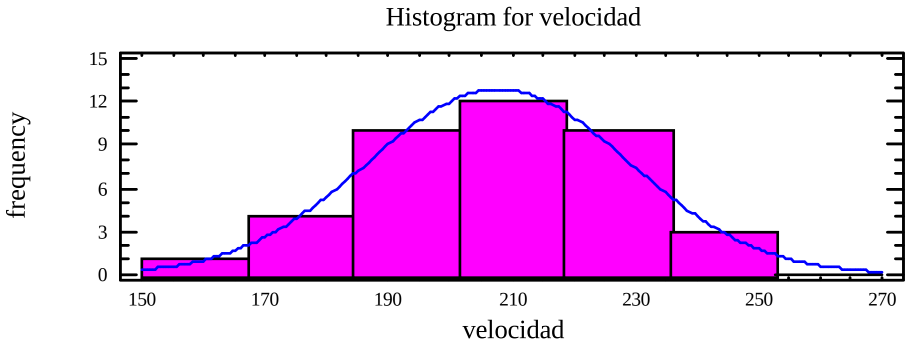
- Veámoslo gráficamente.







```
Summary Statistics for velocida

Count = 40
Average = 207,821
Variance = 450,246
Standard deviation = 21,219
Minimum = 163,24
Maximum = 252,49
Range = 89,25
Stnd. skewness = -0,0390404
Stnd. kurtosis = -0,443825
```



Una vez estimada la NORMAL Podemos calcular probabilidades N(207.8, 21.2)

- P(Velocidad < 200)=0.356
- P(Velocidad < 250)=0.977
- P(200<Velocidad<250)= 0.977 -0.356=0.621
- En STATGRAPHICS vamos a:
  - describe

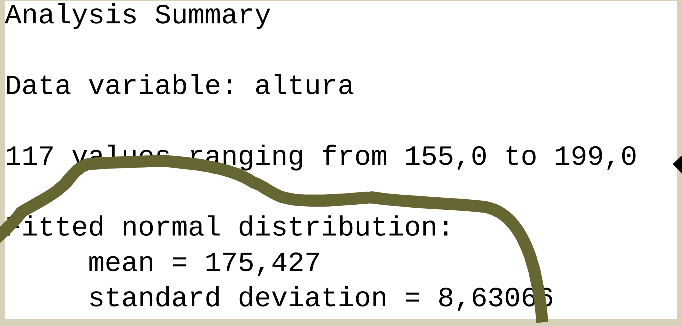
    Distributions

    Distribution fitting (uncensored)

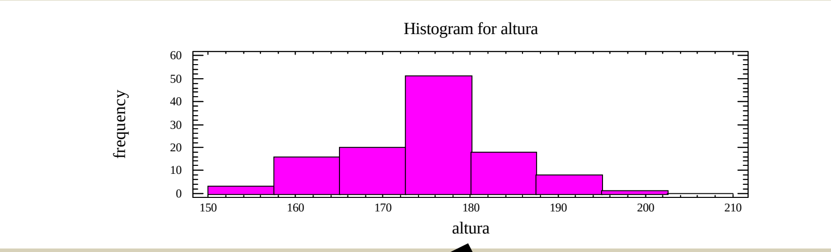
    Ajustamos la normal

    Tail Areas

Otro ejemplo: Alturas de los estudiantes

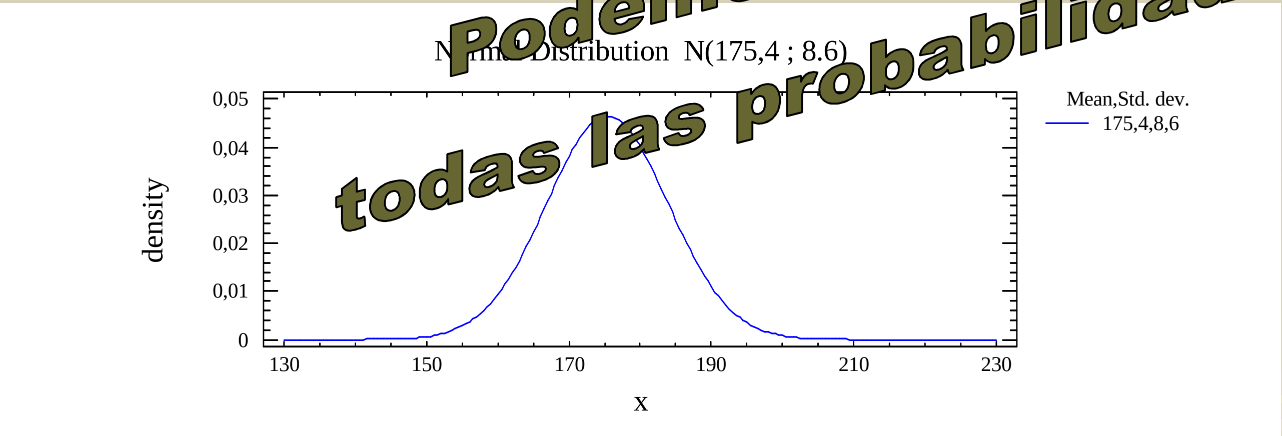


```
Analysis Summary

Data variable: altura

117 values ranging from 155,0 to 199,0

Fitted normal distribution:
     mean = 175,427
     standard deviation = 8,63066
```



N(175.4 ; 8.6)

Podemos calcular todas las probabilidades



En Tail areas obtenemos:

P(Altura<170) = 0,265031

P(Altura<180) = 0,703636 P(170<Altura<180)=0,70-027=0,43

Proceso de estimación (resumen)

- Miramos los datos
- Vemos si parecen normales
- Les ajustamos la distribución normal N(media, desviación)
- Ya podemos calcular probabilidades y...

- TOMAR DECISIONES

Estimación proporciones (distribución binomial):
Proporción de defectuosos

- Para estimar la proporción de defectuosos, se toma una muestra grande cuando el proceso se sabe que está bien y se estima

P(D)=número de defectuosos/número total

Ejemplo

- Hemos tomado 200 piezas y obtenido 4 defectuosas:

- P(D)=4/200=0.02 (El 2%)
- Con este dato ya podemos calcular probabilidades:
- Tomamos normalmente muestras de tamaño 20. Como P(D)=0.02

- X=Número de defectuosos en 20 con P(D)=0.02 es una Binomial.
  - B(20, 0.02)

INTERVALOS de confianza

- Los intervalos de confianza proporcionan una zona en la que previsiblemente estará el valor auténtico del parámetro.
- Se calculan con una confianza determinada
- Normalmente 95%

INTERVALOS de confianza

- Se calculan para:
  - Media
  - Varianza o desviación típica
  - Proporciones

INTERVALOS DE CONFIANZA MEDIA

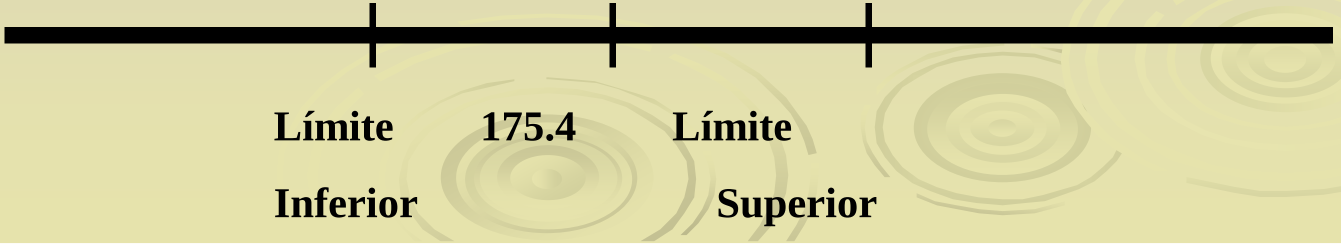
- Hemos estimado la altura media de los estudiantes de ingeniería en

$$\boxed{\bar{x} = 175.4}$$

Un intervalo de confianza proporciona una zona en la que

con una confianza predeterminada estará la altura

media de verdad de esa población



Límite Inferior

175.4

Límite Superior

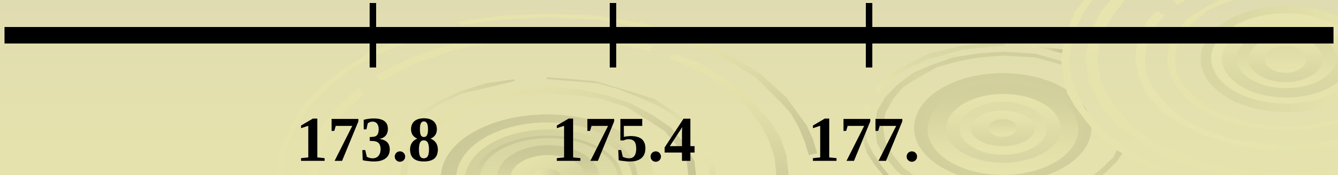
INTERVALOS DE CONFIANZA MEDIA

- Statgraphics:
- Describe-One variable-Confidence intervals

Confidence Intervals for altura

95,0% confidence interval for mean: 175,427 +/- 1,58035 [173,847;177,008]

95,0% confidence interval for standard deviation: [7,64853;9,90463]



INTERVALOS DE CONFIANZA
MEDIA

- La fórmula que sirve para calcularlo (LO PONGO PARA ASUSTAR)y es:

$$\bar{x} - t_{\alpha/2}\hat{s}/\sqrt{n} \qquad \bar{x} + t_{\alpha/2}\hat{s}/\sqrt{n}$$

n=número de datos

t es una distribución que viene en las tablas

s es la desviación típica

Si n>30 y la confianza es del 95% entonces t=2 y el intervalo es:

$$\bar{x} - 2\hat{s}/\sqrt{n} \qquad \bar{x} + 2\hat{s}/\sqrt{n}$$

Intervalo de confianza
Varianza o desviación típica

- Hemos estimado que la media de altura era 175.4
- La desviación es 8.63
- En describe-one variable-confidence intervals

```
Confidence Intervals for altura

95,0% confidence interval for mean: 175,427 +/- 1,58035   [173,847;177,008]

95,0% confidence interval for standard deviation: [7,64853;9,90463]
```

Contrastes de hipótesis

- Contraste de hipótesis en estadística es contestar a preguntas
- Siempre se contesta en términos de probabilidad.
- Es semejante a intervalos de confianza
- Un intervalo también puede resolver estos problemas.

Contrastes de hipótesis

Problema:

Tenemos un muestra de alturas de 2000 españoles.

Media=173 y desviación=10cm.

Hace 10 años la media era de 168.

¿Ha aumentado la altura de los españoles o es una casualidad de la muestra?

Contrastes de hipótesis

Problema:

Las bolsas de un fabricante soportan 16 Kg tomamos una muestra y, a la vista de su media y desviación ¿Debemos creer al fabricante?

Problema:

La duración de una enfermedad es de 15 días. Se estudia un nueva droga en 20 pacientes y se observa que, en promedio la duración son 13 días. ¿Es mejor la nueva droga? ¿Es una casualidad?

Problema:

Tenemos un muestra de alturas de 2000 españoles. Media=173 y desviación=10cm.

Hace 10 años la media era de 168.

¿Ha aumentado la altura de los españoles o es una casualidad de la muestra?

Metodología:

Definimos un Hipótesis nula H_0 que es lo que se quiere comprobar

y una hipótesis alternativa H_1 que es lo que observamos

H_0: Media=168

H_1: Media>168

H_0: Media=168

H_1: Media distinta 168

Problema:

Tenemos un muestra de alturas de 2000 españoles. Media=173 y desviación=10cm.

Hace 10 años la media era de 168.

¿Ha aumentado la altura de los españoles o es una casualidad de

la muestra?

La decisión se toma en términos de probabilidad de que ocurra H_0 con los datos que obtenemos.

La herramienta para decidir se llama p-valor

Describe

Hipotesis Test

H_0: Media=168

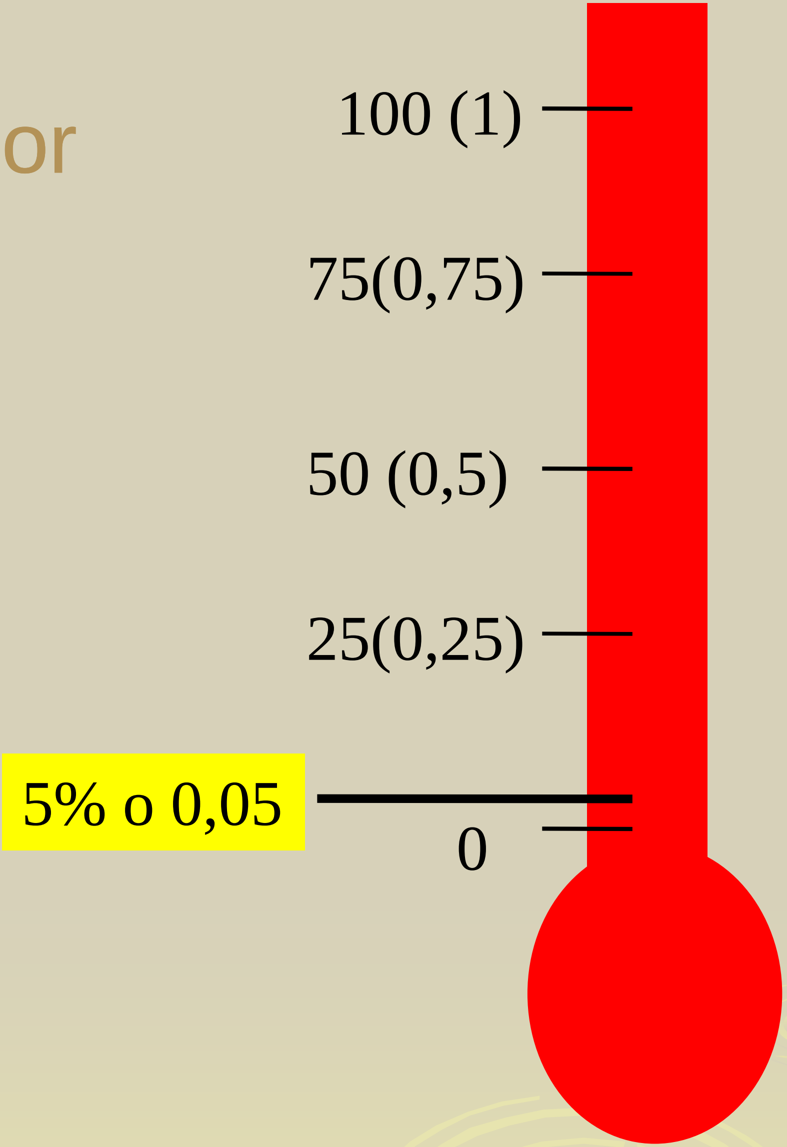
H_1: Media>168

El p-valor

- El p-valor da una idea de lo verosímil que es la hipótesis nula con los datos que tenemos
- P-valor es bajo….Poco razonable que Ho sea verdad
- P-valor alto……Es bastante posibe que Ho sea verdad

P-valor

- El p-valor da la verosimilitud de Ho con los datos que tenemos. Si es bajo (menor que 0,05) rechazamos Ho. Si es alto, no rechazamos Ho



100 (1)

75(0,75)

50 (0,5)

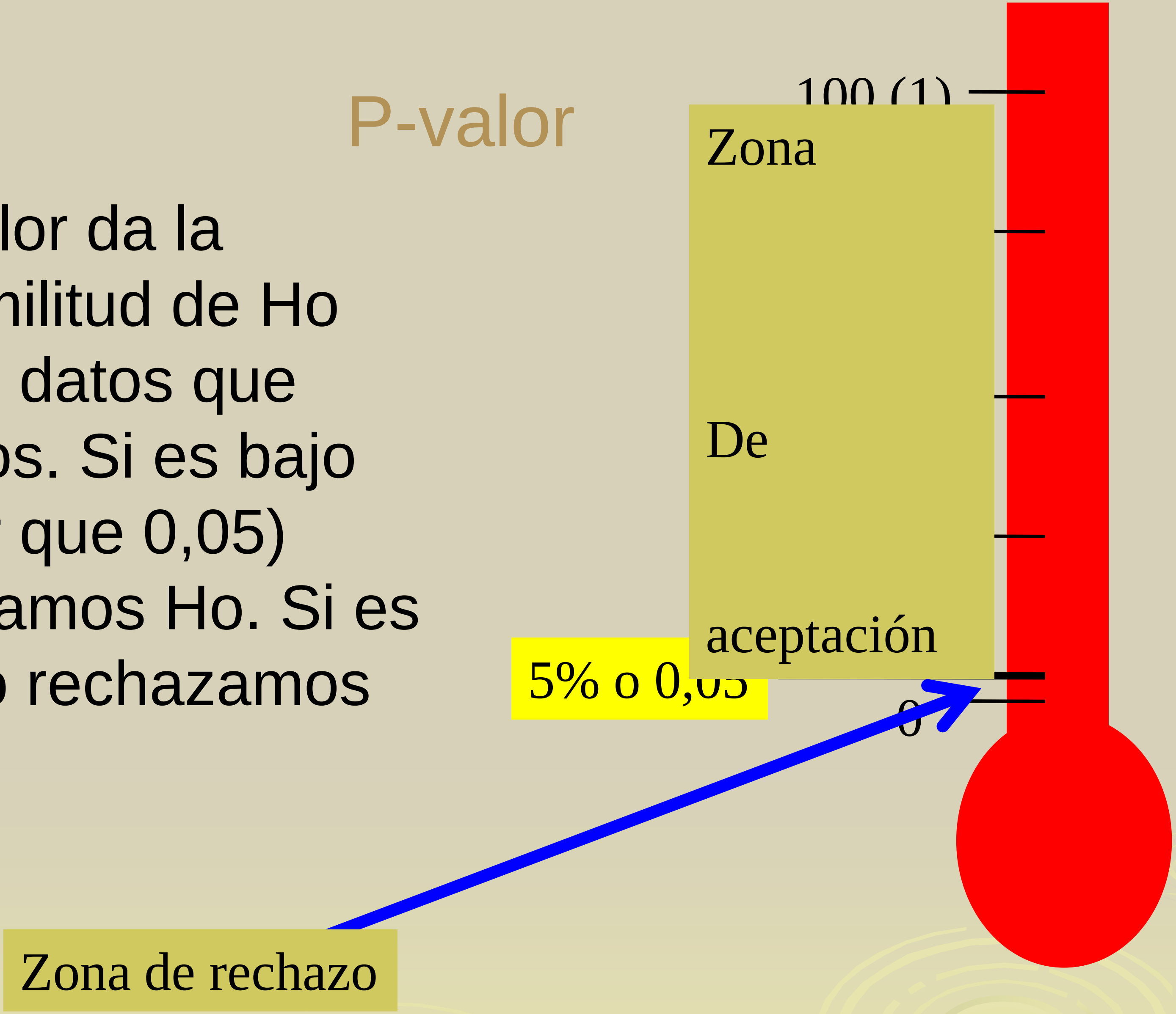
25(0,25)

5% o 0,05

0

P-valor

- El p-valor da la verosimilitud de Ho con los datos que tenemos. Si es bajo (menor que 0,05) rechazamos Ho. Si es alto, no rechazamos Ho



100 (1)

Zona

De

aceptación

5% o 0,05

0

Zona de rechazo

Problema:

Tenemos un muestra de alturas de 2000 españoles. Media=173 y desviación=10cm.

Hace 10 años la media era de 168.

¿Ha aumentado la altura de los españoles o es una casualidad de la muestra?

La decisión se toma en términos de probabilidad de que ocurra H_0 con los datos que obtenemos.

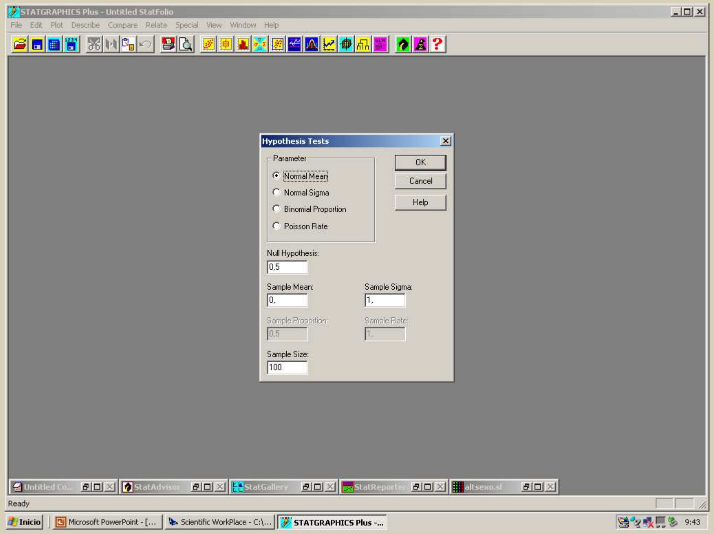
La herramienta para decidir se llama p-valor

Describe

Hipotesis Test

H_0: Media=168

H_1: Media>168



Parameter
● Normal Mean
○ Normal Sigma
○ Binomial Proportion
○ Poisson Rate

Null Hypothesis: 0.5

Sample Mean: 0.

Sample Sigma: 1.

Sample Proportion: 0.5

Sample Rate: 1.

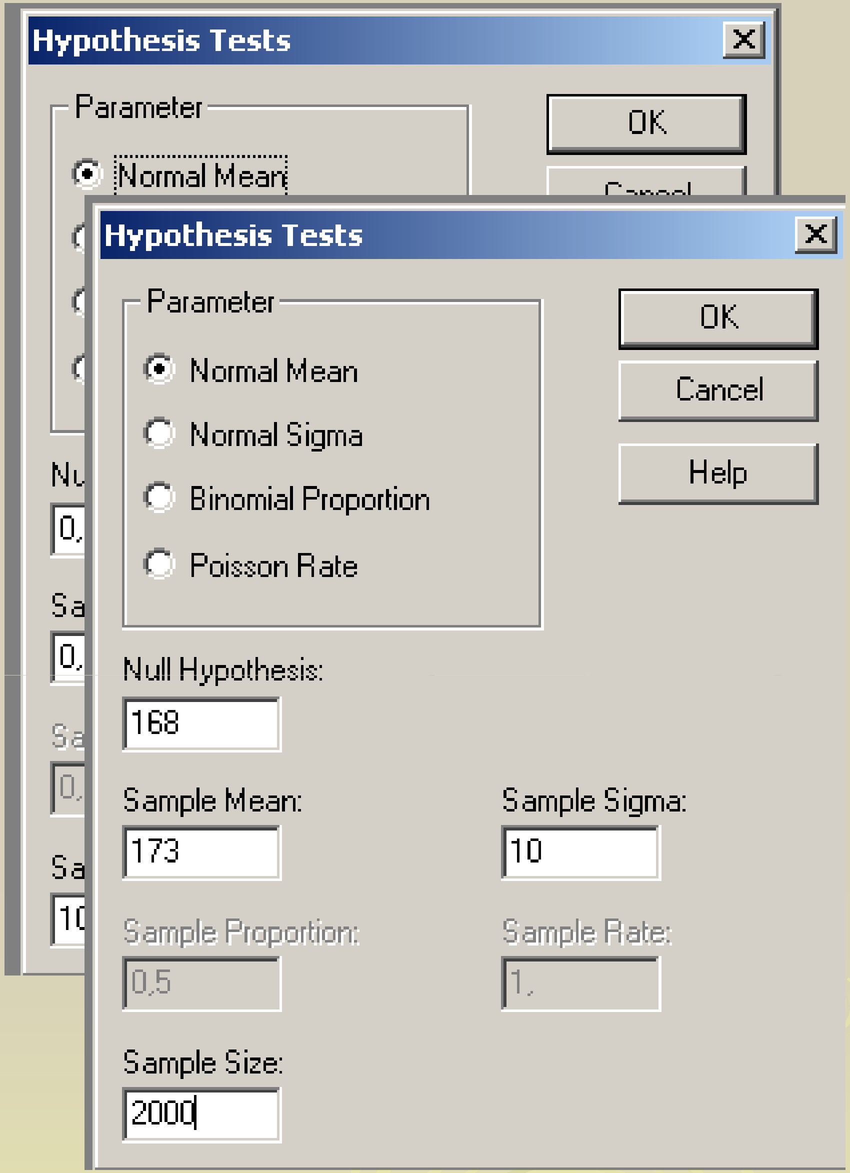
Sample Size: 100

Pide: Null hypothesis. H_0 Media=168

Media de la muestra: 173

Desviación de la muestra: 10

Tamaño de la muestra: 2000



Parameter
● Normal Mean
○ Normal Sigma
○ Binomial Proportion
○ Poisson Rate

Null Hypothesis: 168

Sample Mean: 173

Sample Sigma: 10

Sample Proportion: 0,5

Sample Rate: 1,

Sample Size: 2000

RESULTADOS

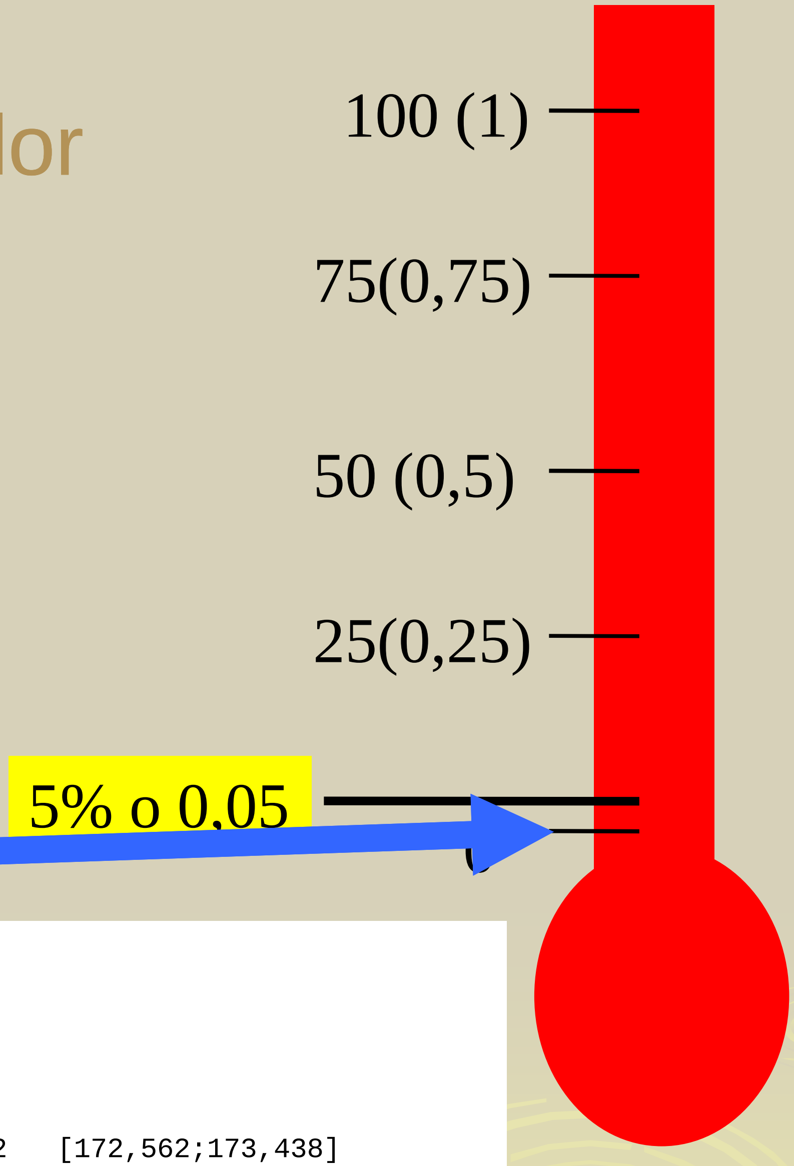
```
Hypothesis Tests
----------------
Sample mean = 173,0
Sample standard deviation = 10,0
Sample size = [illegible]000

95,0% confidence interval for mean: 173,0 +/- 0,438262   [172,562;173,438]

Null Hypothesis: mean = 168,0
Alternative: not equal
Computed t statistic = 22,3607
P-Value = 0,0
Reject the null hypothesis for alpha = 0,05.
```

Más altos?????

P-valor

H_0: Media=168

H_1: Media>168



Zona de rechazo

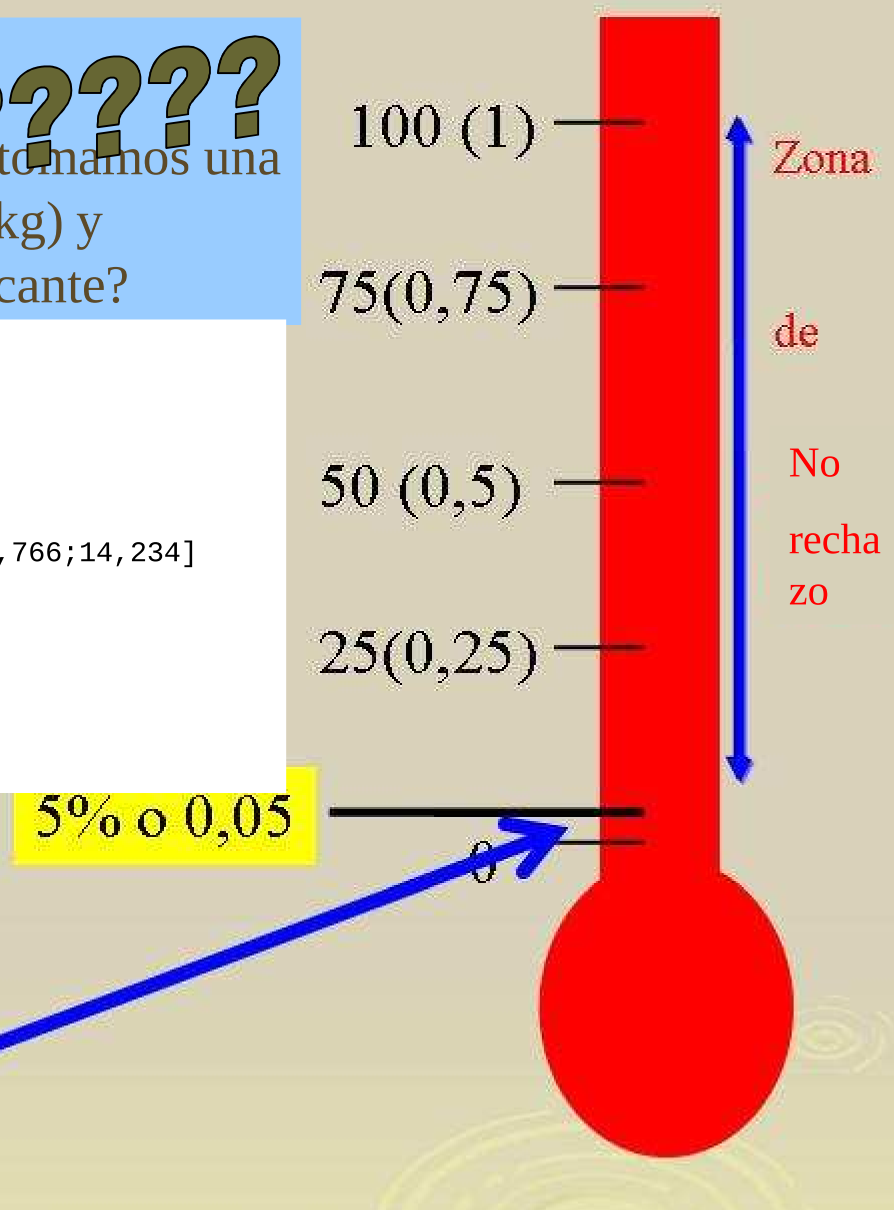
```
Hypothesis Tests
----------------
Sample mean = 173,0
Sample standard deviation = 10,0
Sample size = 2000

95,0% confidence interval for mean: 173,0 +/- 0,438262   [172,562;173,438]

Null Hypothesis: mean = 168,0
Alternative: not equal
Computed t statistic = 22,3607
P-Value = 0,0
Reject the null hypothesis for alpha = 0,05.
```

Problema:

Las bolsas de un fabricante soportan 16Kg tomamos una muestra n=20 y a la vista de su media (14kg) y desviación (0.5Kg) ¿Debemos creer al fabricante?

Mejor o peor?????

```
Hypothesis Tests
----------------
Sample mean = 14,0
Sample standard deviation = 0,5
Sample size = 20

95,0% confidence interval for mean: 14,0 +/- 0,234008   [13,766;14,234]

Null Hypothesis: mean = 16,0
Alternative: not equal
Computed t statistic = -17,8885
P-Value = 2,39586E-13
Reject the null hypothesis for alpha = 0,05.
```



Zona de rechazo

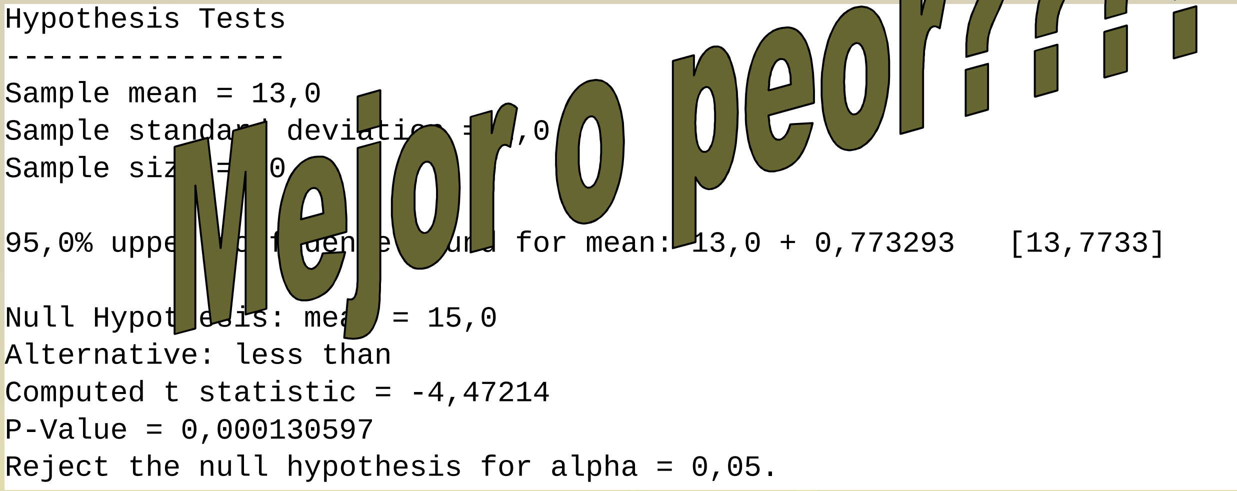
H_0: Media=16

H_1: Media<16

Problema:

La duración de una enfermedad es de 15 días. Se estudia un nueva droga en 20 pacientes y se observa que, en promedio la duración son 13 días con desviación típica de 2 días.

¿Es mejor la nueva droga? ¿Es una casualidad?



```
Hypothesis Tests
----------------
Sample mean = 13,0
Sample standard deviation = [illegible],0
Sample size = [illegible]0

95,0% upper confidence bound for mean: 13,0 + 0,773293   [13,7733]

Null Hypothesis: mean = 15,0
Alternative: less than
Computed t statistic = -4,47214
P-Value = 0,000130597
Reject the null hypothesis for alpha = 0,05.
```

Mejor o peor?????

H_0: Media=15

H_1: Media<15

Podemos contrastar

- Una media
- Dos medias
- Muchas medias (ANOVA)
- Una proporción
- Dos proporciones
- Independencia entre variables cualitativas (Tablas de Contingencia)

- **Siempre con la misma idea de p-valor**

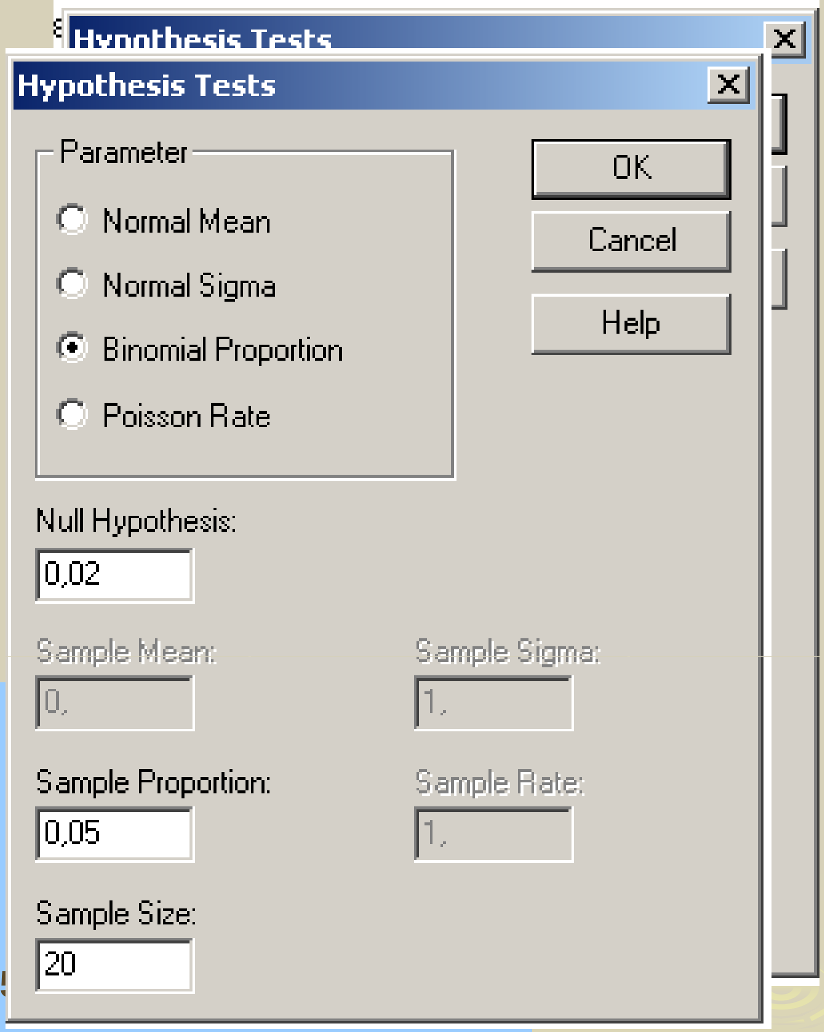
Proporción

- El proceso debe producir en buenas condiciones un 2% de defectuosos.
- ¿Cómo saber si funciona bien?
- Tomando muestras constantemente y contrastando si la proporción de verdad puede ser 2%.
- Tomamos una muestra:
  - n=20 obtenemos 1 defectuoso.

Proporc

H_0: prop= 0,02

H_1: prop distinta 0,02



Parameter

○ Normal Mean

○ Normal Sigma

● Binomial Proportion

○ Poisson Rate

Null Hypothesis: 0,02

Sample Mean: 0,

Sample Sigma: 1,

Sample Proportion: 0,05

Sample Rate: 1,

Sample Size: 20

```
Hypothesis Tests

----------------

Sample proportion = 0,05

Sample size = 20

Approximate 95,0% confidence interval for p: [0,00126

Null Hypothesis: proportion = 0,02

Alternative: not equal

P-Value = 0,664784

Do not reject the null hypothesis for alpha = 0,05.
```

Contraste para la diferencia de medias o proporciones

Problema:

Probamos dos marcas de neumáticos. El resultado es:

TIPO 1: Tamaño de la muestra: n_1

Media 1

Desviación típica 1

¿Son iguales o distintos?

TIPO 2: Tamaño de la muestra: n_2

Media 2

Desviación típica 2

H_0: media 1= media2 (Diferencia=0)

H_1: media 1 distinta de media 2

Contrasto para la di
propo

Compare-Two Samples-Hypoth

Problema:

Probamos dos marcas de neumáticos. El

TIPO 1: Tamaño de la muestra: n_1=10

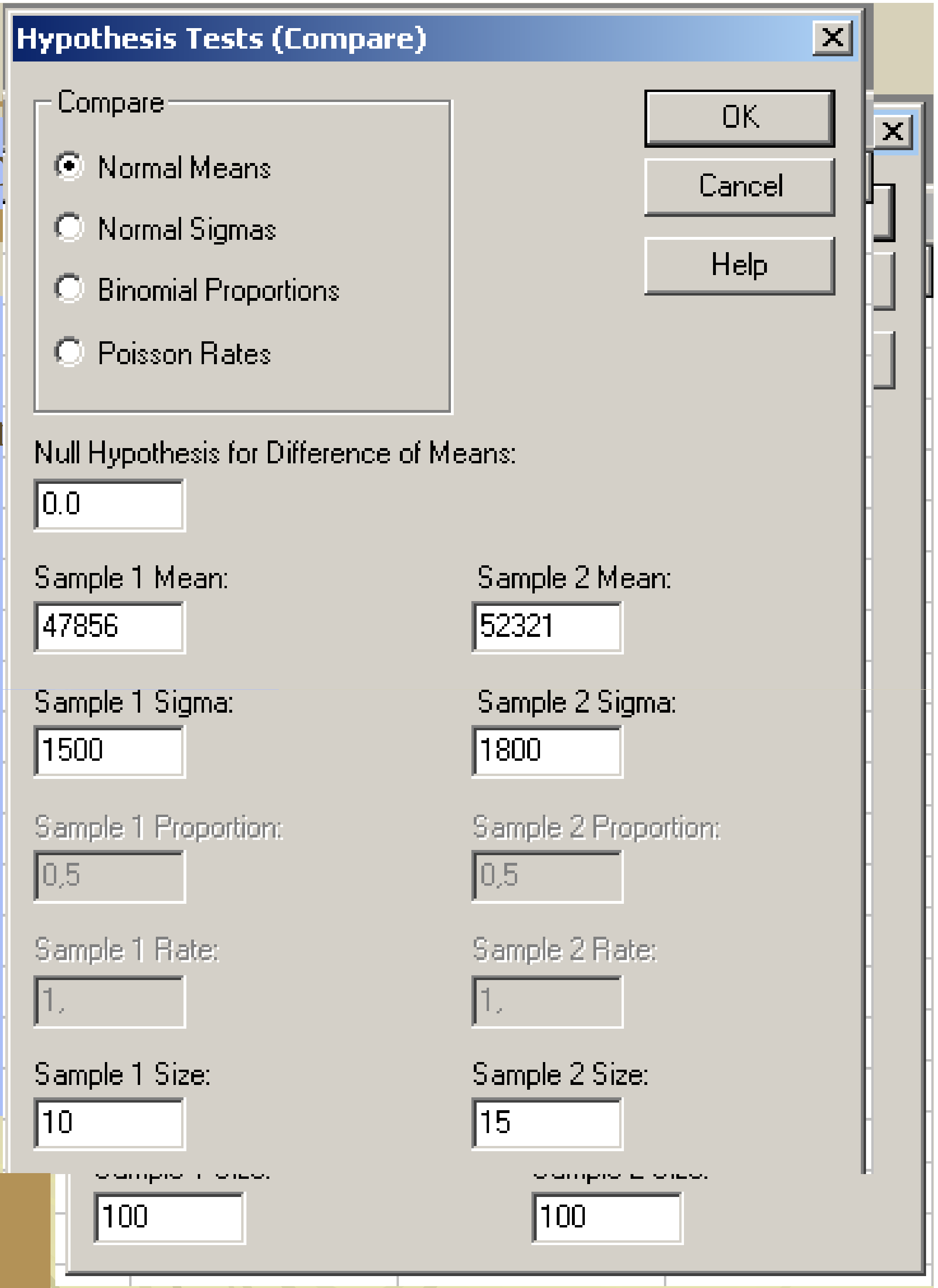
Media 1 =47.856Km

Desviación típica 1: 1500 Km

TIPO 2: Tamaño de la muestra: n_2 = 15

Media 2= 52.321

Desviación típica 2: 1800Km



Compare
- ● Normal Means
- ○ Normal Sigmas
- ○ Binomial Proportions
- ○ Poisson Rates

Null Hypothesis for Difference of Means: 0.0

Sample 1 Mean: 47856	Sample 2 Mean: 52321
Sample 1 Sigma: 1500	Sample 2 Sigma: 1800
Sample 1 Proportion: 0,5	Sample 2 Proportion: 0,5
Sample 1 Rate: 1,	Sample 2 Rate: 1,
Sample 1 Size: 10	Sample 2 Size: 15
100	100

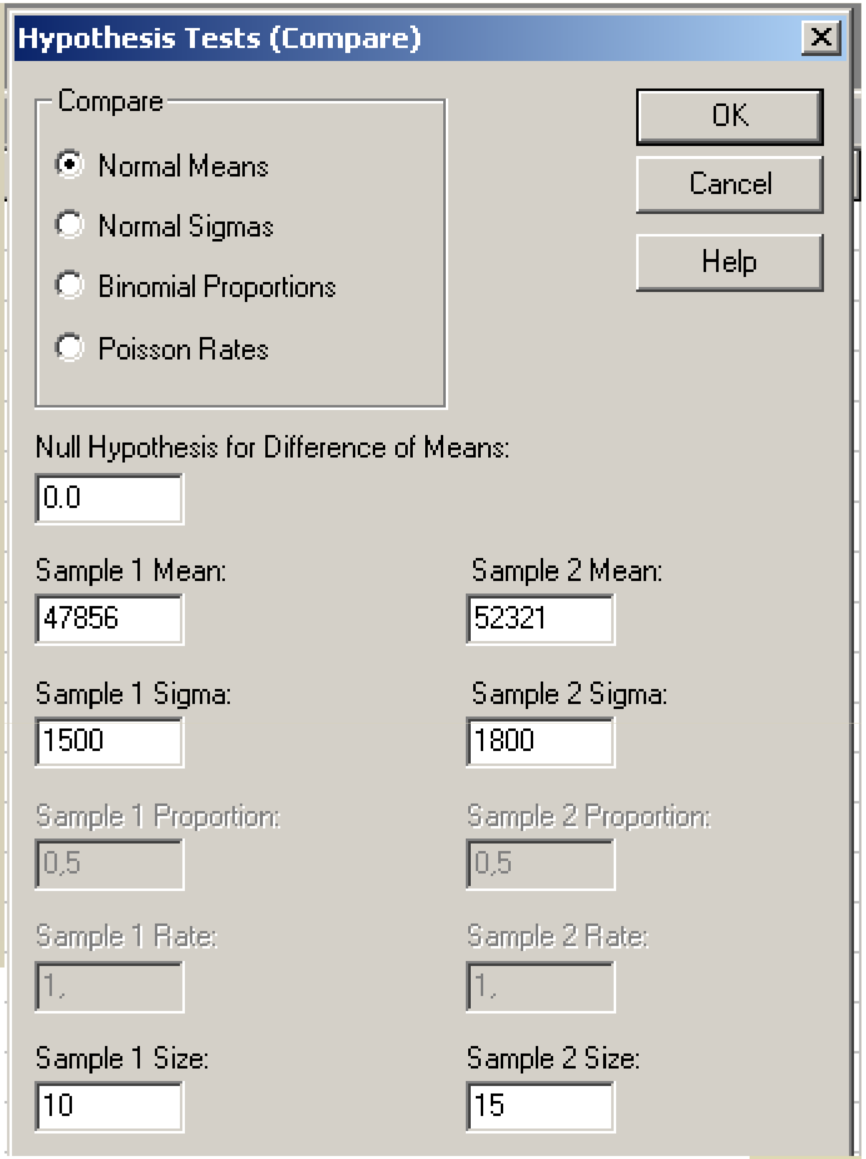
H_0: media 1= media2

H_1: media 1 distinta de media 2

RESULTADOS

H_0: media 1= media2

H_1: media 1 distinta de media 2



Compare

● Normal Means

○ Normal Sigmas

○ Binomial Proportions

○ Poisson Rates

Null Hypothesis for Difference of Means: 0.0

Sample 1 Mean: 47856

Sample 2 Mean: 52321

Sample 1 Sigma: 1500

Sample 2 Sigma: 1800

Sample 1 Proportion: 0,5

Sample 2 Proportion: 0,5

Sample 1 Rate: 1,

Sample 2 Rate: 1,

Sample 1 Size: 10

Sample 2 Size: 15

```
Hypothesis Tests
----------------
Sample means = 47856,0 and 52321,0
Sample standard deviations = 1500,0 and 1800,0
Sample sizes = 10 and 15

95,0% confidence interval for difference between means: -4465,0 +/- 1426,38   [-5891,38;-3038,62]

Null Hypothesis: difference between means = 0,0
Alternative: not equal
Computed t statistic = -6,47554
P-Value = 0,00000131452
Reject the null hypothesis for alpha = 0,05.
```

¿Son iguales o distintas?

Comparación de proporciones.

Problema:

Probamos dos máquinas y estudiamos el número de defectuosos. El resultado es:

MAQUINA 1: Tamaño de la muestra: n_1

Defectuosos: D_1

Proporción de defectuosos: $p_1=D_1/n_1$

MAQUINA 2: Tamaño de la muestra: n_2

Defectuosos: D_2

Proporción de defectuosos: $p_2=D_2/n_2$

H_0: $p_1=p_2$

H_1: p_1 distinta p_2

¿Son iguales o distintas?

Comparación de proporciones.

Problema:

Probamos dos máquinas y estudiamos el número de d

MAQUINA 1: Tamaño de la muestra: n_1

Defectuosos: D_1

Proporción de defectuosos: $p_1=D_1/n_1$

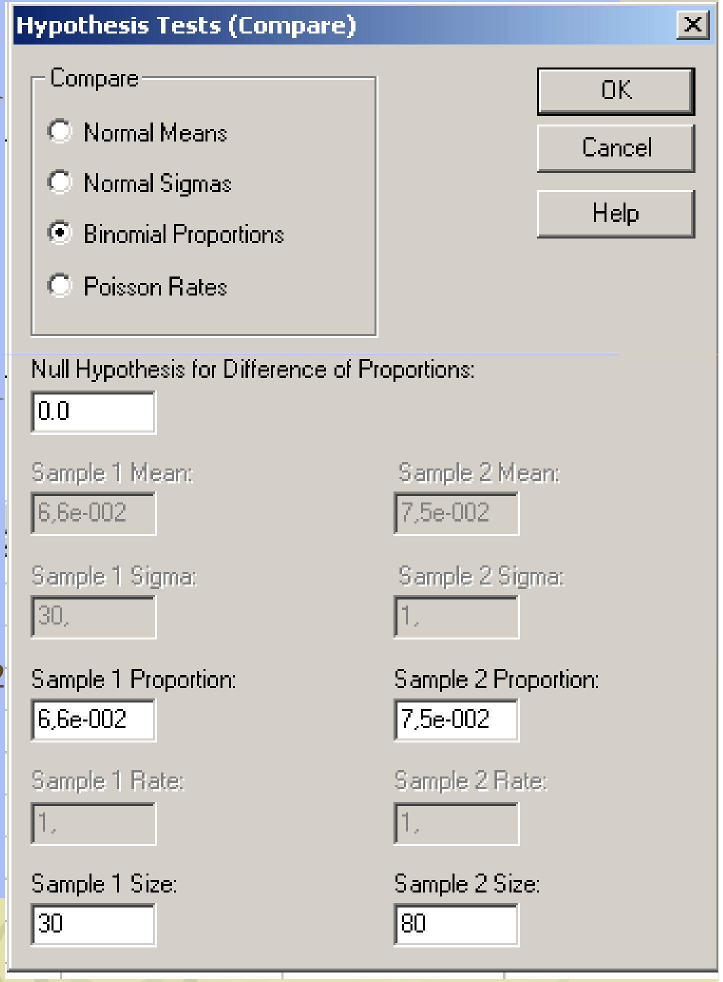
MAQUINA 2: Tamaño de la muestra: n_2

Defectuosos: D_2

Proporción de defectuosos: $p_2=D_2/n_2$

H_0: $p_1=p_2$

H_1: p_1 distinta p_2



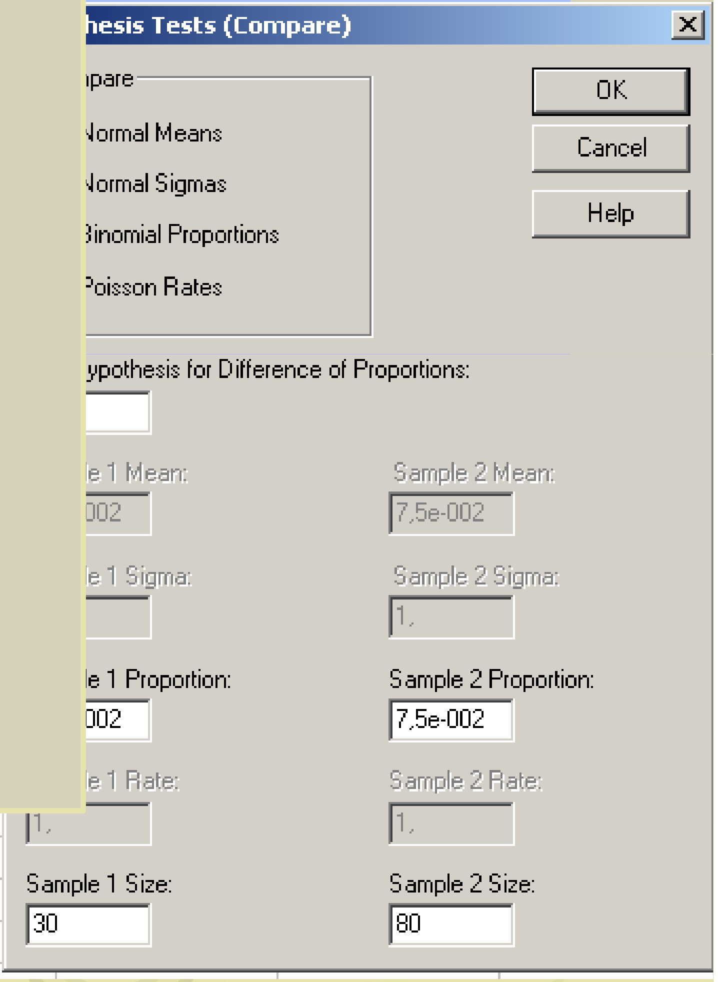
Compare
- ○ Normal Means
- ○ Normal Sigmas
- ● Binomial Proportions
- ○ Poisson Rates

Null Hypothesis for Difference of Proportions: 0.0

Sample 1 Mean: 6,6e-002	Sample 2 Mean: 7,5e-002
Sample 1 Sigma: 30,	Sample 2 Sigma: 1,
Sample 1 Proportion: 6,6e-002	Sample 2 Proportion: 7,5e-002
Sample 1 Rate: 1,	Sample 2 Rate: 1,
Sample 1 Size: 30	Sample 2 Size: 80

ciones.

```
Hypothesis Tests
----------------
Sample proportions = 0,066 and 0,075
Sample sizes = 30 and 80


Approximate 95,0% confidence interval for difference
between proportions: [-0,114947;0,0969469]


Null Hypothesis: difference between proportions = 0,0
Alternative: not equal
Computed z statistic = -0,162069
P-Value = 0,871247
Do not reject the null hypothesis for alpha = 0,05.
```



le 1 Proportion: 002

Sample 2 Proportion: 7,5e-002

Sample 1 Size: 30

Sample 2 Size: 80

H_0: $p_1=p_2$

H_1: p_1 distinta p_2

¿Qué pasa si tenemos varios grupos que comparar?

Problema:

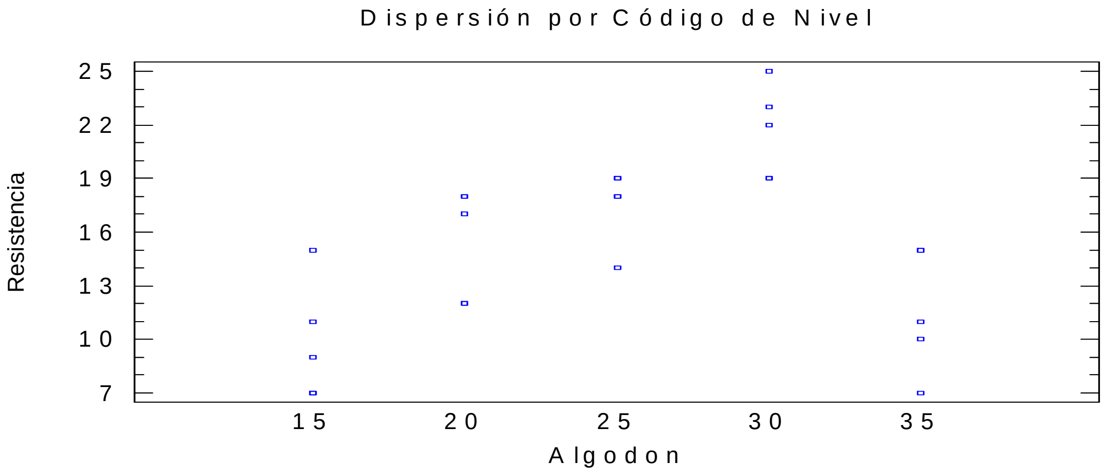
Queremos analizar las propiedades (resistencia) que confiere a una fibra textil el contenido (%) de algodón:

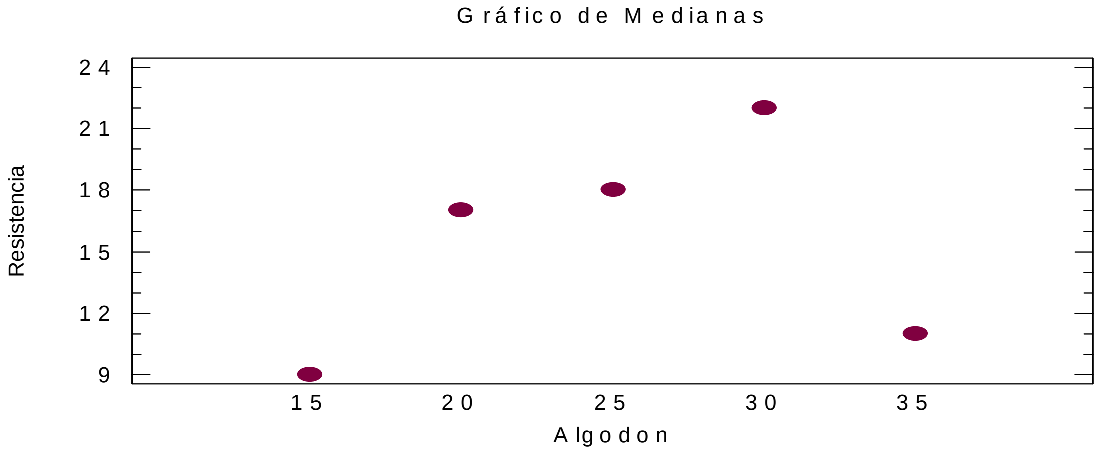
Porcentaje de algodón				
15%	20%	25%	30%	35%
7	12	14	19	7
7	17	18	25	10
15	12	18	22	11
11	18	19	19	15
9	18	19	23	11

¿Influye el % de algodón?

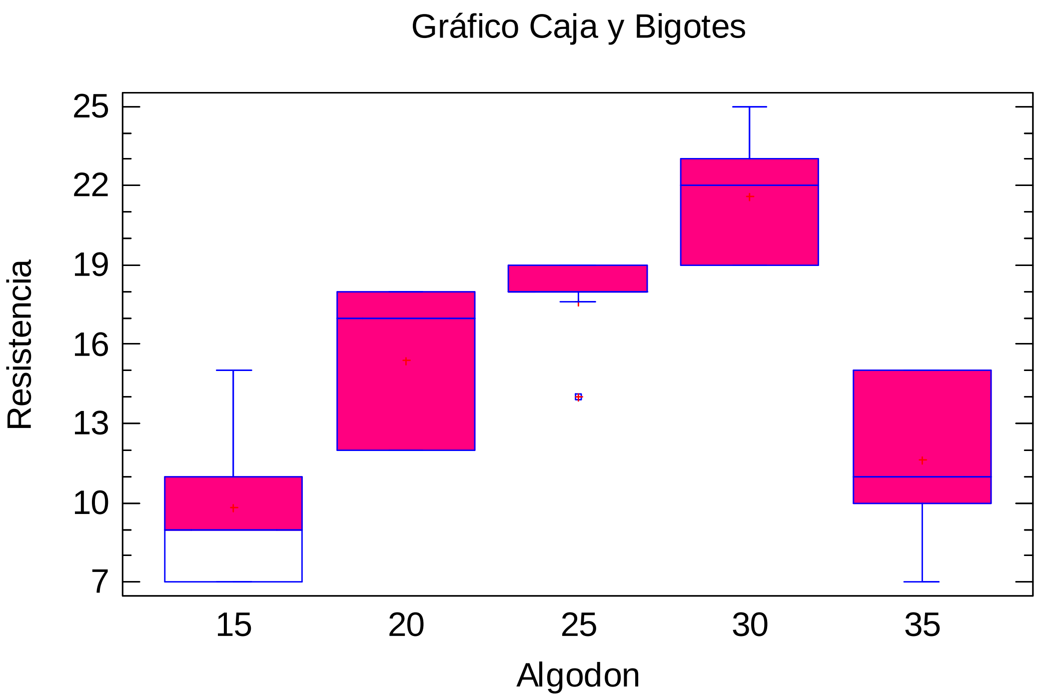
Porcentaje de algodón				
15%	20%	25%	30%	35%
7	12	14	19	7
7	17	18	25	10
15	12	18	22	11
11	18	19	19	15
9	18	19	23	15
Medias				
9.8	15.4	17.6	21.6	10.8

¿Influye el % de algodón?





¿Influye el % de algodón?



Objetivo del análisis

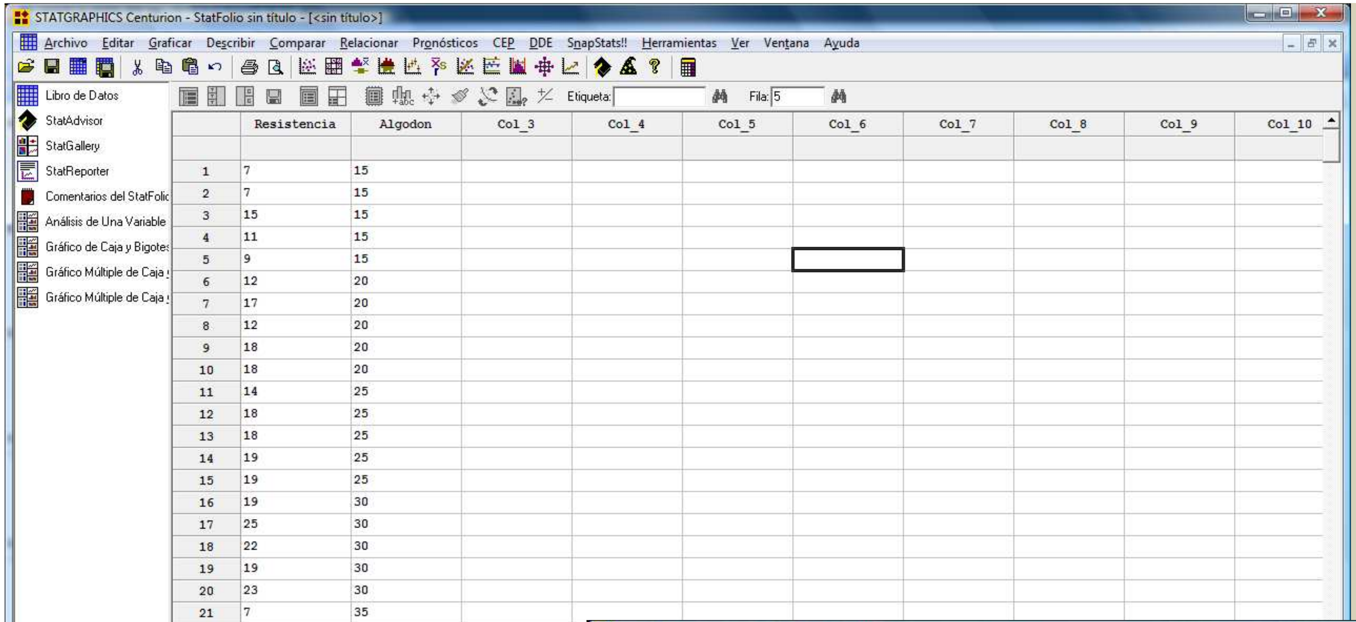
- Saber si el grupo (Factor) influye
- En el modelo equivale a:
  - ¿Son iguales $\mu_1, \mu_2, \mu_3.$ …..?
  - ¿Alguna media es diferente?

- Estadísticamente:
  - ¿$H_o: \mu_1 = \mu_2 = \mu_3 = \ldots. = \mu_I$?
  - ¿H_1: Alguna es diferente?

La técnica estadística empleada se denomina ANOVA, pero equivale a un contraste de hipótesis múltiple !!!

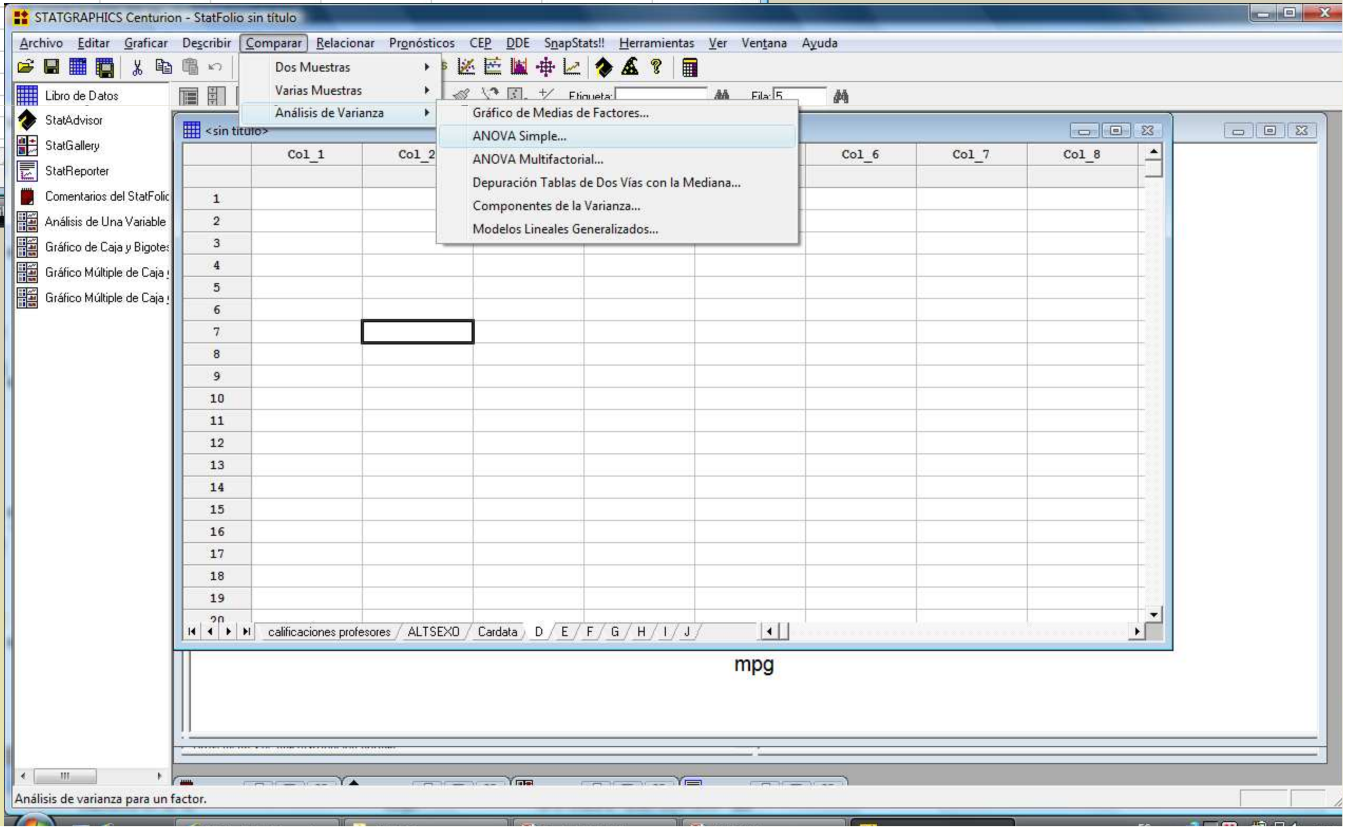
Porcentaje de algodón				
15%	20%	25%	30%	35%
7	12	14	19	7
7	17	18	25	10
15	12	18	22	11
11	18	19	19	15
9	18	19	23	15
Medias				
9.8	15.4	17.6	21.6	10.8

Introducimos los datos

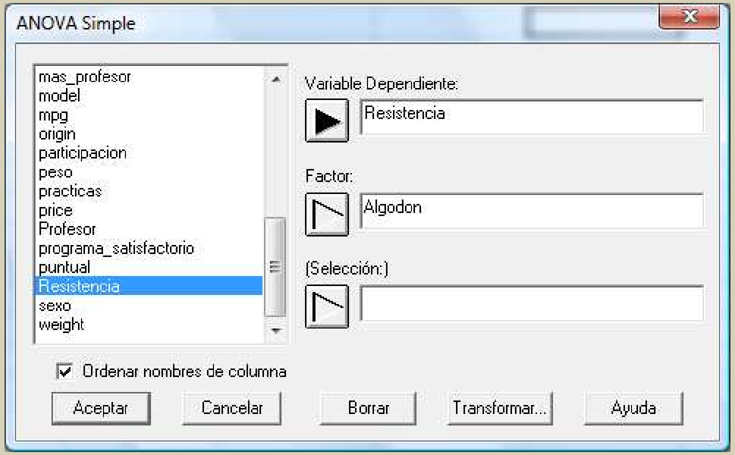


	Resistencia	Algodon
1	7	15
2	7	15
3	15	15
4	11	15
5	9	15
6	12	20
7	17	20
8	12	20
9	18	20
10	18	20
11	14	25
12	18	25
13	18	25
14	19	25
15	19	25
16	19	30
17	25	30
18	22	30
19	19	30
20	23	30
21	7	35
22	10	35
23	11	35
24	15	35
25	15	35
26		
27		

Hacemos el analisis….



Análisis de varianza para un factor.



Variable Dependiente: Resistencia

Factor: Algodon

(Selección:)

☑ Ordenar nombres de columna

Hacemos el analisis….

Este es el resultado

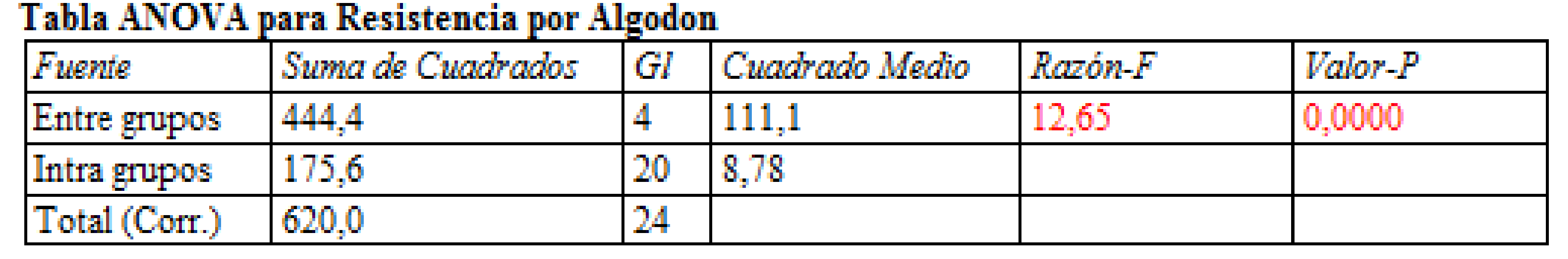


Tabla ANOVA para Resistencia por Algodon

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Entre grupos	444,4	4	111,1	12,65	0,0000
Intra grupos	175,6	20	8,78		
Total (Corr.)	620,0	24			

P-valor

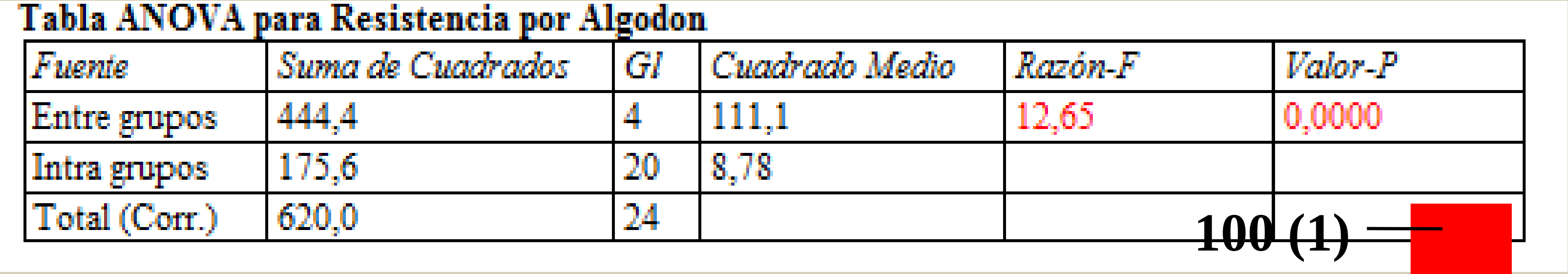
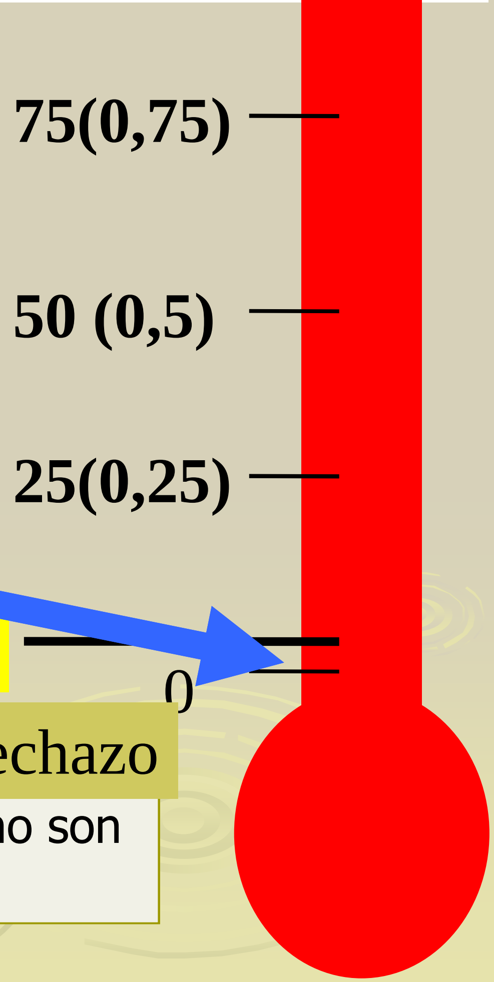


Tabla ANOVA para Resistencia por Algodon

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Entre grupos	444,4	4	111,1	12,65	0,0000
Intra grupos	175,6	20	8,78		
Total (Corr.)	620,0	24			



- Estadísticamente:
  - ¿H_o: $\mu_1 = \mu_2 = \mu_3 = \ldots . = \mu_l$?
  - ¿H_1: Alguna es diferente?

Valor - P

5% o 0,05

Zona de rechazo

¿Influye el algodón?. NO se cumple Ho → todas las medias no son Iguales para cualquier % de algodón