Práctica 8

Alejandro Cáceres
UPC - Statistics 2019/2020

Objetivo

- Contraste de hipótesis

Contraste de hipótesis

En la universidad de Maryland está el grupo de investigación CALCE en baterías de litio que se usan entre muchas otras cosas para teléfonos móviles y coches eléctricos.

Los datos de sus experimentos se pueden descargar de https://web.calce.umd.edu/batteries/data.htm

Contraste de hipótesis

Datos sobre experimentos en la capacidad de descarga de las baterías en función de las condiciones de su almacenamiento se encuentran en el fichero

```
baterias <- read.table("PLN_Number_SOC_Temp_StoragePeriod.csv",
header=TRUE, sep=",")
> head(baterias)
  PLN SOC TEMP Time Discharge.Capacity   X
1   1  NA   NA <NA>           1.421630 bad
2   2  NA   NA <NA>           1.439746 bad
3   3   0   50   3W           1.568073
4   4   0   50   3W           1.557777
5   5   0   50   3W           1.571983
6   6   0   50   3W           1.563704
```

Contraste de hipótesis

Quita los filas que tienen NA en alguna variable usando la función **complete.cases**

Contraste de hipótesis

```
> baterias <- baterias[complete.cases(baterias),]
> head(baterias)
  PLN SOC TEMP Time Discharge.Capacity X
3   3   0   50   3W           1.568073
4   4   0   50   3W           1.557777
5   5   0   50   3W           1.571983
6   6   0   50   3W           1.563704
7   7   0   50   6M           1.576870
8   8   0   50   6M           1.562722
```

Contraste de hipótesis

- La variable SOC (state of charge) es el el nivel de carga a la que se ha guardado la batería.
- La variable TEMP es la temperatura en Fareheit del almacén
- La variable TIME es el tiempo de guardado
- La variable Discharge.Capacity (en miliamperios-hora) es la capacidad de descarga medida depués de haber sido guarda.

Contraste de hipótesis

Una pregunta imporante es saber a qué nivel de carga se deben guardar las baterías para que no pierdan su capacidad de descarga.

Supongamos que el procedimiento estándard para almacenar las baterías es cargarlas al 100% y depositarlas en un almacén a 50 grados Farenheit (10 Celcius), lo que facilita su venta directa.

Estimemos la media y la desviación típica en la capacidad de descarga de estas baterías en la base de datos.

Contraste de hipótesis

Datos de descarga para TEMP=50 y SOC=100: práctica estándard

```
descarga <- baterias$Discharge.Capacity
 selectEstandard <- baterias$TEMP==50 & baterias$SOC==100
descargaEstandard <- descarga[selectEstandard]

mu0 <- mean(descargaEstandard)
mu0
[1] 1.561935

sigma0 <- sd(descargaEstandard)
sigma0
[1] 0.01294606
```

Tomaremos μ_0 y σ_0 como valores de referencia.

Contraste de hipótesis

Queremos saber ahora si al guardar las baterías a 50 grados Farenheit y totalmente descargadas (0% nivel de carga) incrementamos el rendimento en capacidad de descarga con respecto al procedimeinto estándard de guardarlas totalmente cargadas al 100%.

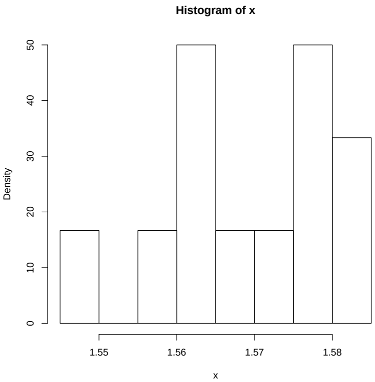
Consigamos los datos de descarga para TEMP=50 y SOC=0, y hagamos su histograma.

Contraste de hipótesis

Datos de descarga para TEMP=50 y SOC=0: nueva práctica

```
selectTest <- baterias$TEMP==50 & baterias$SOC==0
x <- descarga[selectTest]
barx <- mean(x)
barx
[1] 1.569232
n <- length(x)
n
[1] 12
hist(x, freq=FALSE)
```

Contraste de hipótesis



Contraste de hipótesis

Asumimos que la capacidad de descarga cuando la batería se guarda descargada (0%) es una variable aleatoria X que se distribuye normalmente $N(\mu_X, \sigma_X)$.

Inferimos el valor de un parámetro μ_X, que por simplicidad lo llamamos μ, usando el estadístico $\bar{X}$ que mide el promedio de la capacidad de descarga de baterías almacenadas con $TEMP = 50$ y $SOC = 0$ en una muestra de n mediciones.

Contraste de hipótesis

Contraste de hipótesis:

- Hipótesis nula H0: Guardar al 0% las baterías no es diferente a guardarlas al 100% en terminos de la capacidad de descarga, o sea $\mu = \mu_0 = 1.561935$.
- Hipótesis alternativa H1: Guardar a 0% tiene algún efecto (mejoró o empeoró) la capacidad de descarga, o sea $\mu \neq \mu_0$

Contraste de hipótesis

Almacenar las baterías descargadas altera su capacidad de descarga?
Recordemos los datos de nuestras mediciones

- $barx = 1.569232$
- $n = 12$

Recordemos que $\bar{x} = barx$ es una observación de la variable aleatoria $\bar{X}$.
Si suponemos que X es normal y que conocemos σ_X dado por el procedimiento estandard σ_0, entonces

$$\bar{X} \rightarrow N(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

Qué es $\sigma_{\bar{X}}$ en términos de σ_0?

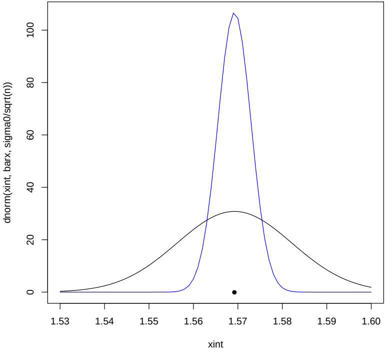
Contraste de hipótesis

- $barx = 1.569232$
- $\sigma_{\bar{X}} = sigma0/\sqrt{n} = 0.01294606/\sqrt{12} = 0.003737206$

En un intervalo `xint <- seq(1.53,1.6,0.001)` grafica las distribuciónes estimadas para X y $\bar{X}$.

Contraste de hipótesis

```
xint<-seq(1.53,1.6,0.001)
plot(xint,dnorm(xint,barx,sigma0/sqrt(n)), type="l",col="blue")

lines(xint,dnorm(xint,barx,sigma0))

points(barx,0,pch=16,col="black")
```



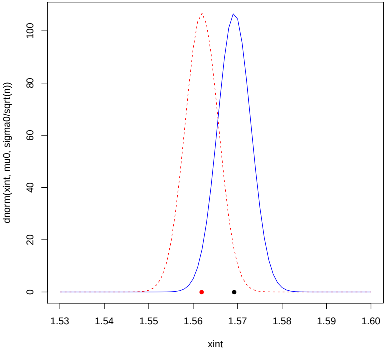
Contraste de hipótesis

Ahora sobre la distribución de $\bar{X}$ según los datos pintemos

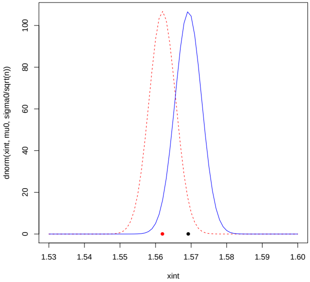
- la distribución de la hipótesis nula $H0 : \mu_0 = 1.561935$
- los valores medios de cada distribución

Contraste de hipótesis

```
plot(xint,dnorm(xint,barx,sigma0/sqrt(n)),type="l",col="blue")
points(barx,0,pch=16,col="black")

lines(xint,dnorm(xint,mu0,sigma0/sqrt(n)),col="red",lty=2)
points(mu0,0,pch=16,col="red")
```



Contraste de hipótesis



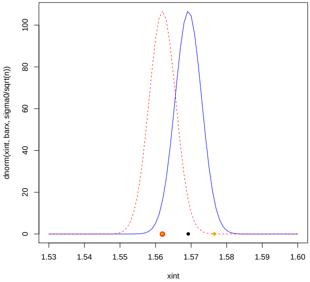
Calcula el intervalo de confianza al 95% para $\bar{x}$ (**z.test** sabemos σ_X) y añadelos al gráfico con points.

Contraste de hipótesis

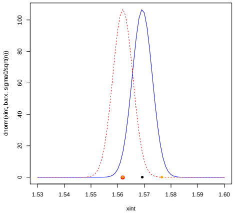
```
library(TeachingDemos)
> z.test(x, sd=sigma0, mu=mu0, conf.level=0.95)
One Sample z-test

data:  x
z = 1.9526, n = 12.0000000, Std. Dev. = 0.0129461, Std. Dev. of
the sample mean = 0.0037372, p-value = 0.05087
alternative hypothesis: true mean is not equal to 1.561935
95 percent confidence interval:
 1.561907 1.576557
sample estimates:
mean of x
 1.569232
```

Contraste de hipótesis

```
plot(xint,dnorm(xint,barx,sigma0/sqrt(n)),type="l",col="blue")
points(barx,0,pch=16,col="black")

lines(xint,dnorm(xint,mu0,sigma0/sqrt(n)),col="red",lty=2)
points(mu0,0,pch=16,col="red",cex=1.5)

points(c(1.561907, 1.576557),c(0,0),pch=16,col="orange")
```



Contraste de hipótesis



μ_0 cae dentro del intervalo de confianza al 95% de $\bar{x}$.

Conclusión: no podemos rechazar la hipótesis nula; no podemos afirmar con confianza del 95% que el nuevo procedimiento mejora la capacidad de descarga de la batería.

Contraste de hipótesis

Segundo argumento para contrastar hipótesis:

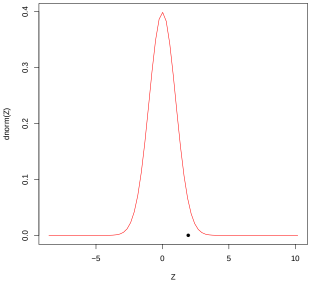
Tipifiquemos la hipótesis nula H_0; es decir, supongamos que las mediciones $\bar{X}$ provienen de esta hipótesis, restemosle μ_0 a $\bar{X}$ y dividamosla por su desviación típica

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{(n)}}$$

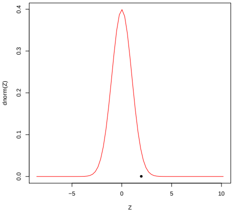
Z es una distribución estandard. Dibujemos la distribución y el resulado de nuestra observación z

Contraste de hipótesis

```
Z<-(xint-mu0)/(sigma0/sqrt(n))

plot(Z,dnorm(Z),col="red",type="l")
z<-(barx-mu0)/(sigma0/sqrt(n))
z
[1] 1.952611
points(z,0,pch=16)
```



Contraste de hipótesis



Bajo la hipótesis nula μ_0 toma el valor de 0 en la variable tipificada ($z_0 = 0$) mientras que nuestra obervación toma el valor $z = 1.952611$. Nuestro valor es diferente de 0, pero que tan raro es?

Si es raro podemos suponer que este valor no fue producido por la hipótesis nula y por lo tanto la rechazamos como posible explicación.

Contraste de hipótesis

Calculemos los cuantíles al 0.025% y 0.975% de una distribución estandard (usa la función **qnorm**)

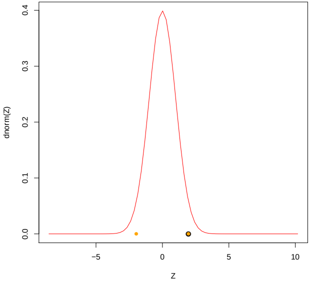
```
> qnorm(c(0.025,0.975))
[1] -1.959964  1.959964
```

El primer cuantíl es la cola izquierda de la distribución, el segundo es para la cola derecha. Ambos contienen el 95% de los datos bajo la hipótesis nula.

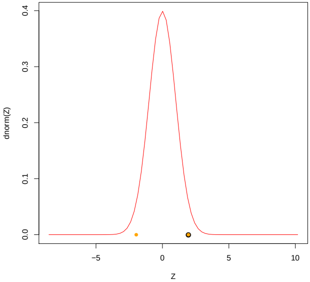
Añadamos los cuantíles al gráfico usando points

Contraste de hipótesis

```
plot(Z,dnorm(Z),col="red",type="l")
points(z,0,pch=16,cex=1.5)
points(c(-1.959964,1.959964),c(0,0),pch=16,col="orange")
```



Contraste de hipótesis



La región de la hipótesis nula: $z < -1.959964$ y $z > 1.959964$ es la **region crítica** para rechazar $H0$ con confianza del 95% usando un criterio de **dos colas**

Conclusión: Nuestro estadístico z tomó un valor de 1.952611 y no podemos descartar la hipótesis nula al 95%.

Contraste de hipótesis

Tercer criterio de evaluación de la hipótesis nula:

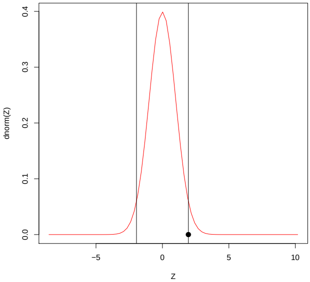
Cuál es la probabilidad de haber obtenido un valor a una distacia como mínimo de $z = 1.952611$ de la hipótesis nula?

La probabilidad acumulada hasta $z = 1.952611$ es

```
> pnorm(z)
[1] 0.9745671
```

Contraste de hipótesis

```
plot(Z,dnorm(Z),col="red",type="l")
points(z,0,pch=16,cex=1.5)
abline(v=z)
abline(v=-z)
```



Contraste de hipótesis

La probabilidad de haber obtenido un valor a una distacia de por lo menos $z = 1.952611$ de la hipótesis nula? es

```
> pval<-2*(1-pnorm(z))
> pval
[1] 0.05086572
```

Conclusión: No podemos rechazar la hipótesis nula al 95%, porque como *pval* no es menor a 0.05 nuestra observación no es lo suficientemente rara.

Contraste de hipótesis

Volvamos al **z.test**

```
> z.test(x, sd=sigma0, mu=mu0, conf.level=0.95)
One Sample z-test

data:  x
z = 1.9526, n = 12.0000000, Std. Dev. = 0.0129461, Std. Dev. of
the sample mean = 0.0037372, p-value = 0.05087
alternative hypothesis: true mean is not equal to 1.561935
95 percent confidence interval:
 1.561907 1.576557
sample estimates:
mean of x
 1.569232
```

Ya entendemos todo?

Contraste de hipótesis

Qué pasa si queremos probar únicamente si guardar la batería descargada es mejor que guardarla cargada?

Test de una cola:

- Hipótesis nula H0: Guardar al 0% no tiene diferencia que guardar al 100%, o sea $\mu = \mu_0$.
- Hipótesis alternativa H1: Guardar a 0% mejora la capacidad de descarga, o sea $\mu > \mu_0$.

Contraste de hipótesis

Reconsideremos qué tan rara es nuestra medición $\bar{x}$ si esperamos que sea mayor que μ_0

Contraste de hipótesis

Calculemos el cuantíl al 0.95% de una distribución estandard (usa la función **qnorm**).

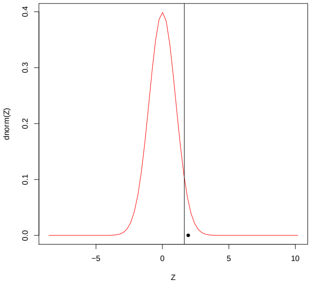
```
> qnorm(0.95)
[1] 1.644854
```

El cuantíl testa que tan rara es la observacion al 95% en la cola derecha de la distribución

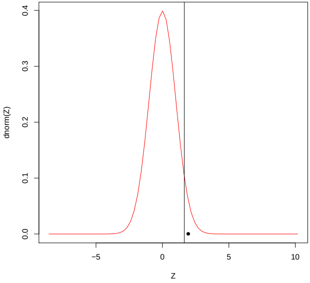
Añadamos este cuantíl a la distribución de la hipótesis nula estandarizada

Contraste de hipótesis

```
plot(Z,dnorm(Z),col="red",type="l")
points(z,0,pch=16)
abline(v=qnorm(0.95))
```



Contraste de hipótesis



La región de la hipótesis nula: $z > 1.644854$ es la **region crítica** para rechazar $H0$ con confianza del 95% usando un criterio de **una cola**.
Conclusión: Nuestro estadístico z tomó el valor de 1.952611 y entonces sí podemos descartar la hipótesis nula al 95%.

Contraste de hipótesis

La misma conclusión se obteiene del p-valor (el criterio de una cola no se multiplica por 2)

```
pval <- 1-pnorm(z)
pval
[1] 0.02543286
```

como *pval* < 0.05 entonces podemos rechazar la hipótesis nula con un contraste de una cola al 95% de confianza.

Contraste de hipótesis

Volvamos al **z.test**

```
z.test(x, sd=sigma0, mu=mu0, conf.level=0.95,
alternative="greater")
One Sample z-test



data:  x
z = 1.9526, n = 12.0000000, Std. Dev. = 0.0129461, Std. Dev. of
the sample mean = 0.0037372, p-value = 0.02543
alternative hypothesis: true mean is greater than 1.561935
95 percent confidence interval:
 1.563085      Inf
sample estimates:
mean of x
 1.569232
```

para el contraste de una cola **superior** usamos el parametro `alternative="greater"`

Contraste de hipótesis

- Qué conclusión obtenemos si incrementamos la confianza al 99%?
- Qué conclusión obtenemos si estimamos la varianza de los datos?

Contraste de hipótesis

Volvamos al **z.test**

```
z.test(x, sd=sigma0, mu=mu0, conf.level=0.99,
alternative="greater")
One Sample z-test

data:  x
z = 1.9526, n = 12.0000000, Std. Dev. = 0.0129461, Std. Dev. of
the sample mean = 0.0037372, p-value = 0.02543
alternative hypothesis: true mean is greater than 1.561935
99 percent confidence interval:
 1.560538      Inf
sample estimates:
mean of x
 1.569232
```

No rechazamos H0: $\mu_0 = 1.561935$ está dentro del intervalo de confianza, y el p-valor no es menor a 0.01.

Contraste de hipótesis

Cuando no sabemos la varianza muestral o la queremos estimar de los datos entonces usamos **t.test**

```
t.test(x, mu=mu0, conf.level=0.95)
One Sample t-test

data:  x
t = 2.475, df = 11, p-value = 0.03084
alternative hypothesis: true mean is not equal to 1.561935
95 percent confidence interval:
 1.562743 1.575721
sample estimates:
mean of x
 1.569232
```

Contraste de hipótesis

```
t.test(x, mu=mu0, conf.level=0.95)
One Sample t-test

data:  x
t = 2.475, df = 11, p-value = 0.03084
alternative hypothesis: true mean is not equal to 1.561935
95 percent confidence interval:
 1.562743 1.575721
sample estimates:
mean of x
 1.569232
```

Para el test de dos colas podemos rechazar la hipótesis nula, porque μ_0 cae fuera del intevalo de confianza y $pval < 0.05$.

Contraste de hipótesis

- nuestros análisis muestran que hay evidencias para considerar que guardar las baterías descargadas incrementa la capacidad de descarga para cuando se quieran usar.
- tenemos una confianza en la evidencia del 95% (no llegamos al 99%).
- es mas eficiente calcular la varianza de los datos. La varianza en la capacidad de descarga también varía dependiendo del nivel de carga que se utilice en el almacenamiento.
- Deberíamos hacer una prueba de hipótesis para la varianza.