CI0131 Diseño de experimentos I Ciclo 2024
Laboratorio

T-test: Comparación de medias independientes

Basado en el ejercicio de Joaquín Amat Rodrigo

https://www.cienciadedatos.net/documentos/12_t-test

Introducción

Una de las opciones cuando se quiere comparar una variable continua entre dos grupos consiste en comparan los resultados promedio obtenidos para cada uno. El hecho de que los valores promedio de cada grupo no sean iguales no implica que haya evidencias de una diferencia significativa. Dado que cada grupo tiene su propia variabilidad, aunque el tratamiento no sea eficaz, las medias muestrales no tienen por qué ser exactas.

Con el fin de estudiar si la diferencia observada entre las medias de dos grupos es significativa, se puede recurrir a métodos paramétricos como el basado en Z-scores o en la distribución T-student. En ambos casos se pueden calcular tanto intervalos de confianza para saber entre que valores se encuentra la diferencia real de las medias poblacionales o test de hipótesis para determinar si la diferencia es significativa.

La distribución T-student se asemeja en gran medida a la distribución normal. Tiene como parámetros la media, la varianza y además incorpora a través de los grados de libertad una modificación que permite flexibilizar las colas en función del tamaño que tenga la muestra. A medida que se reduce el tamaño muestral, la probabilidad acumulada en las colas aumenta, siendo así menos estricta de lo cabría esperar en una distribución normal. Una distribución Tstudent con 30 o más grados de libertad es prácticamente igual a una distribución normal.

El número de grados de libertad de la distribución T-student se calcula:

• Para estudiar una sola muestra: $df = tamaño\ muestra - 1$
• Cuando se comparan dos muestras: existen varios métodos, uno de los utilizados es emplear los grados de libertad de la muestra de menor tamaño $minimo((n1 - 1), (n2 - 1))$.

En los programas informáticos se emplean métodos para ajustar de forma más precisa los grados de libertad.

Condiciones de un t-test para muestras independientes

Las condiciones para calcular intervalos de confianza o aplicar un test de hipótesis basados en la distribución T-student son las mismas que para el teorema del límite central.

• Independencia: Las observaciones tienen que ser independientes unas de las otras. Para ello el muestreo debe ser aleatorio y el tamaño de la muestra inferior al 10% de la población.

• Normalidad: Las poblaciones que se comparan tienen que distribuirse de forma normal. A pesar de que la condición de normalidad recae sobre las poblaciones, normalmente no se dispone de información sobre ellas por lo que las muestras (dado que son reflejo de la población) tiene que distribuirse de forma aproximadamente normal. En caso de cierta asimetría los t-test son considerablemente robustos cuando el tamaño de las muestras es mayor o igual a 30.

• Igualdad de varianza (homocedasticidad): la varianza de ambas poblaciones comparadas debe de ser igual. Tal como ocurre con la condición de normalidad, si no se dispone de información de las poblaciones, esta condición se ha de asumir a partir de las muestras. En caso de no cumplirse esta condición se puede emplear un Welch Two Sample t-test. Está corrección se incorpora a través de los grados de libertad permitiendo compensar la diferencia de varianzas pero con el inconveniente de que pierde precisión.

Test de hipótesis

Los pasos a seguir para realizar un t-test de medias independientes son:

1. Establecer las hipótesis.
2. Calcular el estadístico (parámetro estimado) que se va a emplear.
3. Validar los supuestos para aplicar un t-test
4. Determinar el tipo de test, una o dos colas.
5. Determinar el nivel de significancia α.
6. Cálculo de p-value y comparación con el nivel de significancia establecido.
7. Cálculo del tamaño del efecto (opcional pero recomendado).
8. Conclusiones. Analizar los resultados

1. Establecer las hipótesis

Hipótesis nula ($H0$): por lo general es la hipótesis escéptica, la que considera que no hay diferencia o cambio. Suele contener en su definición el símbolo =. En el caso de comparar dos medias independientes la hipótesis nula considera que $\mu 1 = \mu 2$.

Hipótesis alternativa (HA): considera que el valor real de la media poblacional es mayor, menor o distinto del valor que establece la Ho. Suele contener los símbolos >, <, ≠. En el caso de comparar dos medias independientes la hipótesis alternativa considera que $\mu 1 \neq \mu 2$.

2. Calcular el estadístico (parámetro estimado)

El estadístico es el valor que se calcula a partir de la muestra y que se quiere extrapolar a la población de origen. En este caso es la diferencia de las medias muestrales ($X1 - X2$).

3. Validar los supuestos para aplicar un t-test

Validar las condiciones de independencia, normalidad y homogeneidad de varianzas que permitan aplicar el t-test.

4. Determinar el tipo de test, una o dos colas

Los test de hipótesis pueden ser de una cola o de dos colas. Si la hipótesis alternativa emplea ">" o "<" se trata de un test de una cola, en el que solo se analizan desviaciones en un sentido. Si la hipótesis alternativa es del tipo "diferente de" se trata de un test de dos colas, en el que se analizan posibles desviaciones en las dos direcciones. Solo se emplean test de una cola cuando se sabe con seguridad que las desviaciones de interés son en un sentido y únicamente si se ha determinado antes de observar la muestra, no a posteriori.

5. Determinar el nivel de significancia α

El nivel de significancia α determina la probabilidad de error que se quiere asumir a la hora de rechazar la hipótesis nula. Se emplea como punto de referencia para determinar si el valor de p-value obtenido en el test de hipótesis es suficientemente bajo como para considerar significativas las diferencias observadas y por lo tanto rechazar $H0$.

A menor valor de alpha, menor probabilidad de rechazar la hipótesis nula. Por ejemplo, si se considera α = 0.05, se rechazará la hipótesis nula en favor de la hipótesis alternativa si el p-value obtenido es menor que 0.05, y se tendrá una probabilidad del 5% de haber rechazado $H0$ cuando realmente es cierta.

En nivel de significancia debe establecerse en función de que error sea más costoso:
• Error tipo I: Error de rechazar la $H0$ cuando realmente es cierta
• Error tipo II: Error de considerar como cierta $H0$ cuando realmente es falsa.

6. Cálcular p-value y comparar con el nivel de significancia

Si las condiciones mencionadas previamente se cumplen, se puede considerar que:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{SE} \sim t_{(df)}$$

$$siendo\ \ SE = \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

$S\hat{}$ es la cuasidesviación típica muestral o desviación típica muestral corregida.

$$pvalue = P(|Tcalculada| >= tdf, 1-\alpha/2)$$

7. Identificar el tamaño del efecto

El tamaño del efecto o también llamado effect size es la diferencia neta observada entre los grupos de un estudio. No se trata de una medida de inferencia estadística ya que no se pretende identificar si las poblaciones son significativamente diferentes, sino que simplemente indica la diferencia observada entre muestras, independientemente de la varianza que tengan.

Se trata de un parámetro que siempre debe acompañar a los p-values, ya que un p-value solo indica si hay evidencias significativas para rechazar la hipótesis nula pero no dice nada de si la diferencia es importante o práctica. Esto último se averigua mediante el tamaño del efecto.

En el caso de los t-test de medias independientes, existen dos medidas posibles del tamaño del efecto: la **d** de Cohen y la r de Pearson. Ambas son equivalentes y pueden transformarse de una a otra. Cada una de estas medidas tiene unas magnitudes recomendadas para considerar el tamaño del efecto como pequeño, mediano o grande. La función cohen.d() del paquete *effsize* permite calcular el tamaño del efecto de la diferencia de medias independientes.

D de Cohen para muestras independientes

$$d = \frac{|diferencia\ de\ medias\ entre\ los\ grupos|}{sd}$$

Existen dos formas distintas se utilizan para calcular la sd conjunta de ambas muestras:

$$sd = \sqrt{\frac{(n-1)sd_1^2 + (n-2)sd_2^2)}{n1 + n2 - 2}}$$

$$sd = \sqrt{\frac{sd_1^2 + sd_2^2}{n1 + n2}}$$

Los límites más utilizados para clasificar el tamaño del efecto con d-Cohen son:

d ≤ 0.2 pequeño
d ≥ 0.5 mediano
d = 0.8 grande

R de Pearson para t-test independiente:

$$r = \sqrt{t^2}(t^2 + gl)$$

t = estadístico t obtenido en el test
gl = grados de libertad del test

Los límites más utilizados para clasificar el tamaño del efecto con r son:

d ≤ 0.1 pequeño
d ≥ 0.3 mediano
d = 0.5 grande

8. Interpretar los resultados

Si el p-value es menor que el valor de α seleccionado, existen evidencias suficientes para rechazar $H0$ en favor de HA .

Ejemplo utilizando R

El datset "nacim.csv" contiene información sobre 150 nacimientos junto con información de las madres.

Se quiere determinar si existen evidencias significativas de que el peso (weight) de los recién nacidos cuyas madres fuman (smoker) difiere de aquellos cuyas madres no fuman (nonsmoker).

Cargue nacim.csv y la biblioteca tidyverse

```
library(tidyverse)
nacim <- read.csv("nacim.csv", header=TRUE, stringsAsFactors=TRUE)
head(nacim, 10)
```

Nota: El read.csv también lo puede hacer con file.choose()

	Column1	f_age	m_age	weeks	premature	visits	gained	weight	sex_baby	smoke
1	1	31	30	36	premie	13	1	6.88	male	smoker
2	2	34	36	39	full term	5	35	7.69	male	nonsmoker
3	3	36	35	40	full term	12	29	8.88	male	nonsmoker
4	4	41	40	40	full term	13	30	9.00	female	nonsmoker
5	5	42	37	40	full term	NA	10	7.94	male	nonsmoker
6	6	37	28	40	full term	12	35	8.25	male	smoker
7	7	35	35	28	premie	6	29	1.63	female	nonsmoker
8	8	28	21	35	premie	9	15	5.50	female	smoker
9	9	22	20	32	premie	5	40	2.69	male	smoker
10	10	36	25	40	full term	13	34	8.75	female	nonsmoker

1. Hipótesis

H0: no hay diferencia entre las medias poblacionales: μ(nonsmoker)−μ(smoker)=0

Ha: si hay diferencia entre las medias poblacionales: μ(nonsmoker)−μ(smoker)≠0

2. Parámetro estimado (estadístico)

Diferencia entre las medias muestrales:

```r
smoker    <- nacim %>% filter(smoke == "smoker") %>% pull(weight)
nonsmoker <- nacim %>% filter(smoke == "nonsmoker") %>% pull(weight)

mean(nonsmoker) - mean(smoker)
```

[1] 0.4005

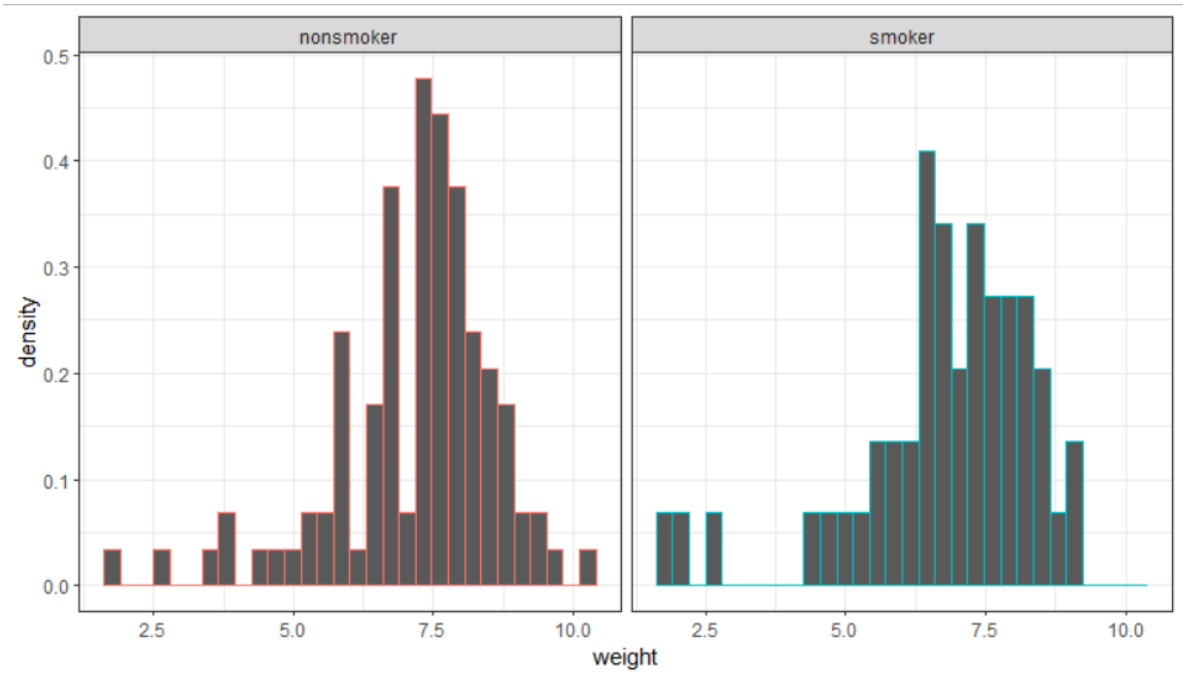
3. Validar las condiciones para aplicar un t-test

3.1. Independencia:

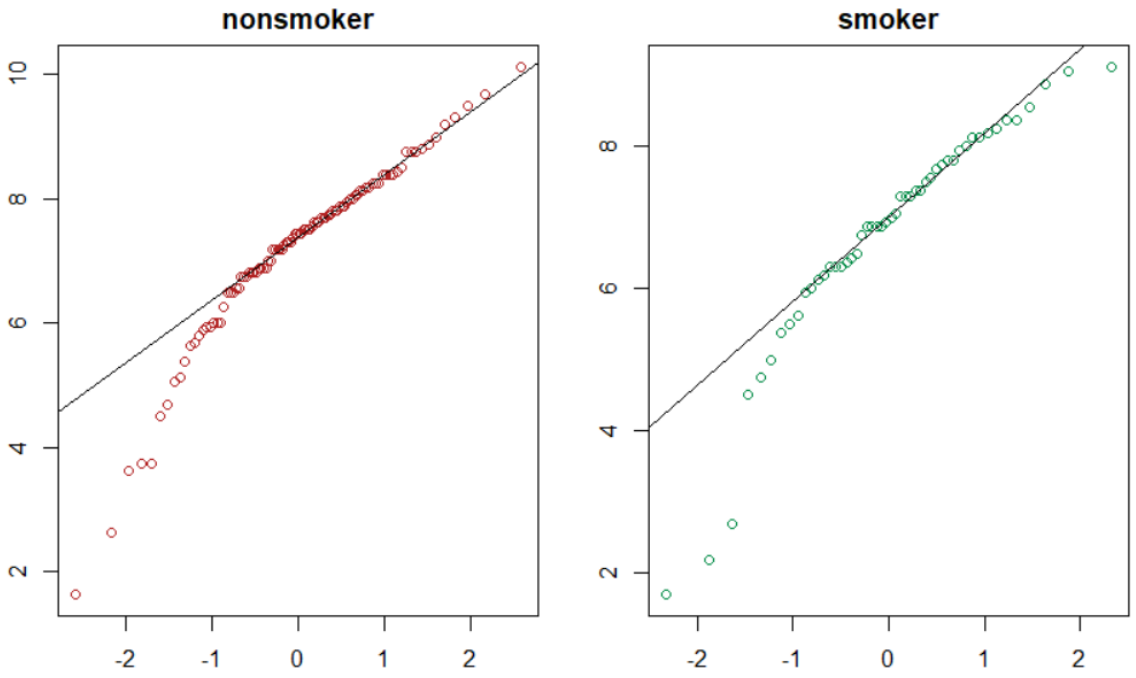
Se trata de un muestreo aleatorio. Se puede afirmar que los eventos son independientes.

3.2. Normalidad:

```r
ggplot(nacim,aes(x = weight)) +
  geom_histogram(aes(y = ..density.., colour = smoke)) +
  facet_grid(.~ smoke) +
  theme_bw() + theme(legend.position = "none")
```



```r
par(mar = c(2, 2, 2, 2))
par(mfrow = c(1, 2))
qqnorm(nonsmoker, xlab = "", ylab = "",
       main = "nonsmoker", col = "firebrick")
qqline(nonsmoker)
qqnorm(smoker, xlab = "", ylab = "",
       main = "smoker", col = "springgreen4")
qqline(smoker)
```




```r
shapiro.test(smoker)
```

Shapiro-Wilk normality test

data: smoker

W = 0.89491, p-value = 0.0003276

Esta prueba utiliza la estrategia de hipótesis nula (los datos provienen de una población normal) y alternativa (los datos no provienen de una población normal). Si el valor p obtenido es menor que el alfa escogido con antelación, se rechaza la hipótesis nula.

```
shapiro.test(nonsmoker)
```

```
        Shapiro-Wilk normality test

data:  nonsmoker

W = 0.92374, p-value = 2.234e-05
```

Los gráficos qqnorm muestran asimetría hacia la izquierda y los test encuentran evidencias significativas de que los datos **no** proceden de poblaciones con distribución normal.

Una prueba no paramétrico basado en la mediana (Mann-Withney-Wilcoxon test) o un test de Bootstraping serían más adecuados que la prueba t. Otra opción sería estudiar si los datos anómalos son excepciones que se pueden excluir del análisis.

Sin embargo, vamos a continuar con la prueba t-test para efecto del ejercicio de utilizar la prueba t clásica de contraste de hipótesis.

Se podría argumentar que dado que el tamaño de cada grupo es mayor que 30 la prueba t-test sigue siendo robusta. En general el hecho de contar con una muestra grande se podría usar si estuviéramos realizando una prueba Z_0 para la que necesitaríamos conocer también las varianzas poblacionales (que no es este caso). Aun en el caso de una prueba Z_0 la no normalidad debe considerarse con mucha precaución, especialmente si la distribución de los datos subyacentes es muy asimétrica o presenta colas pesadas. Además, este hecho es necesario mencionarlo en las conclusiones.

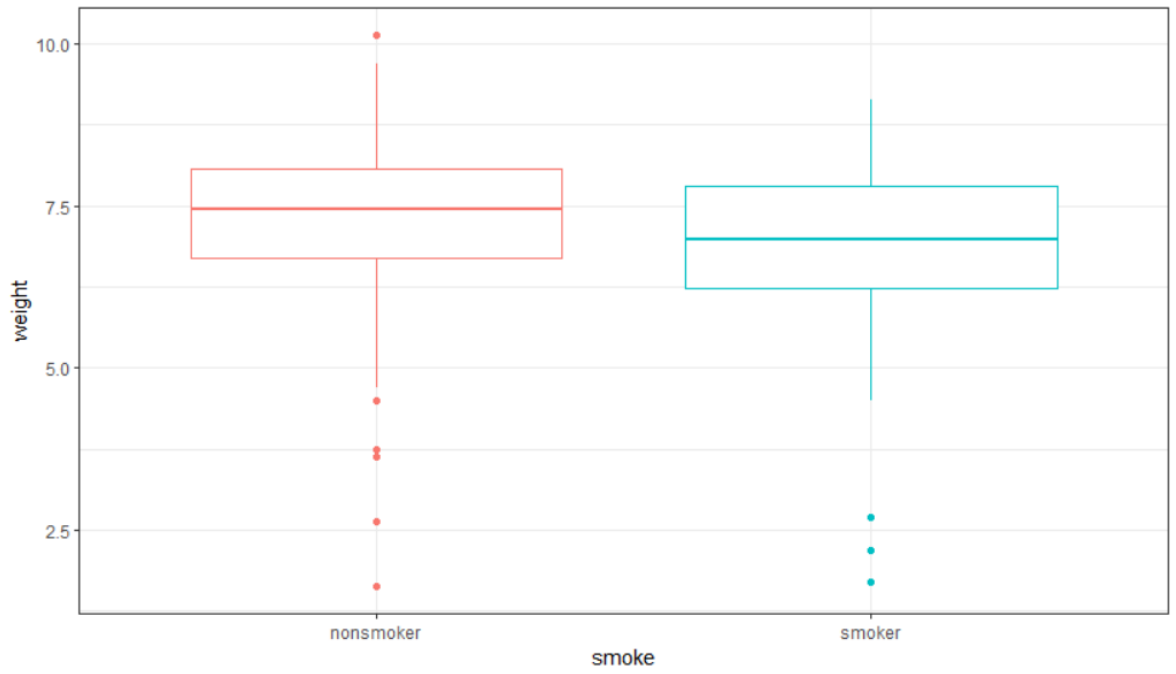
3.3. Igualdad de varianza:

Existen varias pruebas que permiten comparar varianzas. Dado que no se cumple el criterio de normalidad, uno de los recomendados es el test Levene o el test no paramétrico de Fligner-Killeen (ambos basados en la mediana).

Nuevamente estas pruebas utilizan la estrategia de hipótesis nula (los datos poseen varianzas iguales) y alternativa (los datos no poseen varianzas iguales). Si el valor p obtenido es menor que el alfa escogido con antelación, se rechaza la hipótesis nula.

Inicialmente podemos graficar los datos para observar visualmente las distribuciones:

```
ggplot(data = nacim) +
  geom_boxplot(aes(x = smoke, y = weight, colour = smoke)) +
  theme_bw() + theme(legend.position = "none")
```



Ejecutamos las pruebas estadísticas:

```
require(car)
fligner.test(weight ~ smoke, data = nacim)
```

Fligner-Killeen test of homogeneity of variances

data: weight by smoke

Fligner-Killeen:med chi-squared = 0.56858, df = 1, p-value = 0.4508

```
leveneTest(weight ~ smoke, data = nacim, center = "median")
```

Levene's Test for Homogeneity of Variance (center = "median")

	Df	F value	Pr(>F)
group	1	0.4442	0.5062
	148		

Ninguno de los dos test encuentra evidencias significativas (para alpha = 0.05) de que las variancias sean distintas entre ambas poblaciones. Si las varianzas no fuesen iguales se tendría que realizar el t-test con la corrección de Welch.

4. Determinar el tipo de test

Se trata de una prueba de dos colas.

5. Determinar el nivel de significancia

$\alpha = 0.05$

6. Cálculo de p-value

R tiene una función integrada que permite realizar t-test para una o dos muestras, tanto con corrección (en caso de que las varianzas no sean iguales) como sin ella. Esta función devuelve tanto el p-value del test como el intervalo de confianza para la verdadera diferencia de medias.

```
t.test(
  x           = smoker,
  y           = nonsmoker,
  alternative = "two.sided",
  mu          = 0,
  var.equal   = TRUE,
  conf.level  = 0.95
)
```

Two Sample t-test

data: smoker and nonsmoker

t = -1.5517, df = 148, p-value = 0.1229

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.9105531 0.1095531

sample estimates:

mean of x mean of y

6.7790 7.1795

7. Cálculo tamaño de efecto

```
library(effsize)
cohen.d(formula = weight ~ smoke, data = nacim, paired = FALSE)
```

Cohen's d

d estimate: 0.2687581 (small)

95 percent confidence interval:

lower upper

-0.07488708 0.61240332

8. Conclusiones. Análisis de resultados

Dado que p-value (0.1229) es mayor que alpha (0.05) , no se dispone de evidencia suficiente para considerar que existe una diferencia entre el peso promedio de niños nacidos de madres fumadores y el de madres no fumadoras. El tamaño de efecto medido por d-Cohen es pequeño (0.27).

Trabajo a entregar

A. Realice un análisis similar sobre el peso de los bebés pero en lugar de considerar madres fumadoras o no fumadoras (variable smoke), utilice la variable **sex_baby**. Es decir, si el sexo del bebé está relacionado con el peso al nacer.
B. Haga lo mismo para la variable **premature**, es decir, determinar si que el bebé sea prematuro está relacionado con el peso al nacer.

Indicaciones

- El trabajo debe realizarse en grupos de dos personas.
- En ambos casos (A y B) debe realizar los pasos 1 al 8, incluyendo en el reporte el código R utilizado y los resultados de la ejecución del código en R (que pueden ser valores o gráficos).
- Puede usar la prueba t-test a pesar de que no se cuente con normalidad. Sin embargo, notifique en su reporte si esta situación se presenta. Adicionalmente, ponga atención a los resultados de los estadísticos que usa para verificar supuestos, por si eventualmente debe aplicar ciertas correcciones a la hora de aplicar otros estadísticos o la misma prueba t.

- El reporte es un documento con un formato uniforme y coherente, en el que se incluye una pequeña introducción (al menos una línea de texto) para cada gráfico o código R que se incluye.
- Los gráficos deben ser apropiadamente presentados, con etiquetas correctas y debidamente rotulado con la información que presenta.
- El código R debe ser también incluido apropiadamente en el reporte, en forma de imágenes o de texto debidamente formateado e indentado para que se distinga del resto del texto.
- Las salidas o los gráficos de código R que deba incluir deben aparecer junto al código que los genera. No es válido incluir varias instrucciones de código y luego una serie de respuesas y gráficos. Cada instrucción que produzca una salida o un gráfico debe estar asociada directamente con el resultado (el resultado inmediatamente después de la instrucción que lo generó).
- Las entregas se realizarán en formato PDF. La letra debe ser Times New Roman, Arial o Cambria. Los tamaños permitidos son 10-12.
- Debe incluir una portada de página completa en la que incluya el nombre del curso, del laboratorio, así como los nombres completos de los estudiantes. Esta página será utilizada por el profesor para apuntar la nota y también para incluir comentarios en caso de tener que incluir generales del informe.
- El reporte se debe entregar en Mediación Virtual, en un archivo .pdf que pueda ser leído en programas comerciales de uso habitual. Debe verificar que el .pdf que subió a Mediación Virtual contiene los ejercicios resueltos y que el archivo puede abrirse correctamente. En caso de problemas con el archivo .pdf (no abre correctamente, está corrupto, etc.) se considerará que no entregó la tarea.
- Las entregas tardías se penalizarán con un 10% de la nota luego de vencida la fecha y hora de entrega, más un 10% adicional por cada hora de retraso.