Estadística con R (II)

Descripción general

Este material reúne prácticas intermedias de análisis estadístico realizadas en el entorno **RStudio**, como continuación del *Manual de Prácticas de Estadística con R (I)* y en coordinación con el *Manual de Prácticas de Estadística con SPSS*. Está orientado al aprendizaje práctico de los fundamentos de la inferencia estadística, desde el muestreo hasta los intervalos de confianza y los contrastes de hipótesis, tanto paramétricos como no paramétricos (2 poblaciones, Chi-Cuadrado). A través de ejemplos guiados y conjuntos de datos reales, el alumnado desarrolla competencias en formulación de hipótesis, estimación por intervalos, comparación de medias y proporciones, y selección adecuada de pruebas estadísticas según la naturaleza de los datos. El contenido incluye enunciados, prácticas resueltas y archivos de datos que facilitan el aprendizaje autónomo, la interpretación estadística y la aplicación de métodos de inferencia en R..

Objetivo general	Descripción y competencias asociadas
1. Aplicar métodos de muestreo y simulación de datos.	Generar muestras aleatorias con distintos métodos (aleatorio simple, sistemático y estratificado) y explorar su comportamiento mediante herramientas de R.
2. Estimar parámetros poblacionales mediante intervalos de confianza.	Construir e interpretar intervalos de confianza para medias, varianzas y proporciones en una y dos poblaciones.
3. Comparar poblaciones y evaluar diferencias significativas.	Analizar la homogeneidad entre grupos o periodos mediante intervalos de diferencia de medias, proporciones y varianzas.
4. Formular y contrastar hipótesis estadísticas.	Realizar pruebas t y F para una o dos muestras, contrastes pareados y de homogeneidad, interpretando el valor p y la significación estadística.
5. Aplicar contrastes no paramétricos y asociación	Utilizar pruebas como Kolmogorov–Smirnov, Wilcoxon, Mann–Whitney y Ji-cuadrado para el análisis de independencia o asociación.
6. Fomentar la interpretación crítica y la comunicación de resultados.	Presentar conclusiones con rigor, diferenciando significación estadística de relevancia práctica.

Autoras: Nerea González García · Ana B. Nieto Librero
Departamento de Estadística, Universidad de Salamanca
Curso académico: 2023–2024
Tipo de recurso: Material docente / prácticas resueltas
Repositorio institucional: GREDOS – Universidad de Salamanca

ESTRUCTURA DEL MANUAL

Práctica 1. Métodos de muestreo

- Muestreo aleatorio simple, sistemático y estratificado.
- Afijación proporcional, de Neyman y óptima.
- Aplicación sobre la base *cancer* del paquete survival.

Práctica 2. Intervalos de confianza

- Intervalos para medias, varianzas y proporciones.
- Diferencias de medias y proporciones entre poblaciones.
- Comparaciones entre regiones y periodos (caso: indicadores de sostenibilidad).
- Representación gráfica de intervalos mediante plotmeans().

Práctica 3. Contrastes de hipótesis paramétricos

- Contrastes t de Student (una y dos muestras).
- Pruebas F de comparación de varianzas.
- Prueba T para la igualdad de medias (independientes, apareados).
- Interpretación p-valor y decisión estadística.

Práctica 4. Contrastes no paramétricos y asociación

- Test de Kolmogorov–Smirnov, Wilcoxon y Mann–Whitney.
- Análisis de asociación entre variables categóricas.

Material complementario

- Conjuntos de datos utilizados.

PRÁCTICA 1: MUESTREO

Trabajaremos con la base de datos *cancer* (Loprinzi et al, 1994) que se encuentra disponible en la librería *survival* de R. En ella se recoge información sobre la supervivencia en pacientes con cáncer avanzado de pulmón del North Central Cancer Treatment Group. Las variables que se incluyen en la base de datos son las siguientes:

- **inst**: código de la institución a la que pertenece el paciente.
- **time**: tiempo de supervivencia en días.
- **status**: estado del paciente (1=censurado, 2=fallecido).
- **edad**: edad del paciente en años.
- **sex**: género del paciente (1=Hombre, 2=Mujer).
- **ph.ecog**: puntuación ECOG (0=bueno, 5=fallecido).
- **ph.karno**: puntuación Karnofsky física (0=malo, 100=bueno).
- **pat.karn**: tasa Karnofsky por paciente. (Estas dos variables miden la capacidad del paciente para realizar actividades cotidianas).
- **meal.cal**: calorías consumidas en las comidas principales.
- **wt.loss**: pérdida de peso en los últimos seis meses.

Suponiendo que dicha base de datos representa la población total de enfermos de cáncer de pulmón avanzado, realice las siguientes tareas:

1. **Muestreo aleatorio simple**. Seleccione una muestra de tamaño 50 sin reemplazo.

```
library(survival)
data(cancer)
dim(cancer)
 Base:
        n<-50
        indice_50<-sample(1:nrow(cancer), size=n,
        replace=FALSE)
        indice_50
        muestra_50<-cancer[indice_50,]
        head(muestra_50)
```

Dplyr:

```
library(dplyr)
muestra_50b<-cancer%>%
        sample_n(size=n, replace=F)
head(muestra_50b)
```

Seleccione una muestra que suponga un 75% del tamaño de la población.

```
muestra_prop<- cancer %>%
        sample_frac(0.75)
head(muestra_prop)
```

2. **Muestreo aleatorio sistemático.** Seleccione una muestra de tamaño 50 sin reemplazo.

```
install.packages("devtools")
library(devtools)
install_github("DFJL/SamplingUtil")
library(SamplingUtil)
indice_sis<- sys.sample(N=nrow(cancer), n=50)
indice_sis
muestra_sis<- cancer[indice_sis, ]
head(muestra_sis)
```

3. **Muestreo aleatorio estratificado.** Seleccione una muestra de tamaño 50 sin reemplazo.

Afijación proporcional:

```
estratos<- cancer %>%
select(sex,time) %>%
group_by(sex) %>%
summarise(n=n(),
        s=sd(time)) %>%
mutate(p=n/sum(n))

estratos

nsizeProp<-nstrata(n=n,wh=estratos[,4],method="proportional")
nsizeProp
```

```
library(sampling)
indices_estratos <- strata( cancer, stratanames = c("sex"),
size= unlist(nsizeProp), method = "srswor" )
cancer.estrata <- getdata( cancer, indices_estratos )
cancer.estrata
```

Afijación de Neyman:

```
#Asignación de la muestra óptima a los estratos (asume costes iguales)

nsizeNeyman<-
nstrata(n=n,wh=estratos[,4],sh=estratos[,3],method="neyman")
nsizeNeyman


indices_estratosney <- strata(cancer, stratanames = c("sex"), size =
unlist(nsizeNeyman), method = "srswor" )
cancer.estrataney <- getdata( cancer, indices_estratosney )
cancer.estrataney
```

Afijación óptima:

```
#Asignación de la muestra óptima a los estratos
#Coste de muestreo por estrato
ch<-c(5,10)
nsizeOpt<-
nstrata(n=n,wh=estratos[,4],sh=estratos[,3],ch,method="optimal")
nsizeOpt


indices_estratosop <- strata( cancer, stratanames = c("sex"),
size = unlist(nsizeOpt), method = "srswor" )
cancer.estrataop <- getdata( cancer, indices_estratosop )
cancer.estrataop
```

PRÁCTICA 2 – INTERVALOS DE CONFIANZA PARAMÉTRICOS

- 1 POBLACIÓN Y 2 POBLACIONES –

Caso práctico

Desde que en el año 1987 se definió el concepto de sostenibilidad en el Informe Brundtland, se han desarrollado diferentes índices e indicadores en este ámbito para medir gradualmente el desarrollo sostenible de una Sociedad. El Índice de Sociedad Sostenible (SSI), es una idea de los fundadores de la *Sustainable Society Foundation* (SSF) que se publica bianualmente desde hace casi dos décadas por la SSF, se ha convertido en una fuente de información bien establecida en todo el mundo para los organismos gubernamentales, las organizaciones no gubernamentales, las empresas privadas y el mundo académico para valorar el nivel de sostenibilidad de los países en torno a tres componentes: bienestar humano, bienestar medioambiental y bienestar económico. Este índice valora el desarrollo sostenible de una sociedad a través de un conjunto de 21 indicadores clasificados en siete ejes diferentes:

-

Bienestar humano			**Bienestar medioambiental**		**Bienestar económico**	
Necesidades básicas	*Desarrollo personal y salud*	*Sociedad equilibrada*	*Recursos naturales*	*Clima & Energía*	*Transición*	*Economía*
Alimentación suficiente Agua potable suficiente Saneamiento seguro	Educación Vida sana Igualdad de género	Distribución de la renta Crecimiento de la población Buena gobernanza	Biodiversidad Recursos hídricos renovables Consumo	Uso de energía Ahorro de energía Gases de efecto invernadero Energía renovable	Agricultura ecológica Ahorro	PIB Empleo Deuda pública

Trabaje sobre el archivo DatosIC2.xlsx.

Parte 1: Intervalos de confianza paramétricos – 1 población

En esta parte se van a trabajar algunas de las funciones disponibles para la construcción de intervalos de confianza de los siguientes parámetros:

- Media
- Proporción
- Varianza
- Diferencia de medias
- Diferencia de proporciones
- Razón de varianzas

IC para la media de una población normal con varianza desconocida

IC al nivel $1-\alpha$ para la media μ cuando σ^2 es desconocida

$$\left(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n-1}}, \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n-1}}\right)$$

1.- Halle el intervalo de confianza para el valor medio de “agua potable suficiente" para toda la población, y por “IDH_BAJO”, y conteste a las siguientes preguntas (suponga que la variable se distribuye según una distribución normal N(μ,σ)).

```
library(boot)

norm.ci(conf=0.95, t0 = mean(datos2$Agua_potable_suficiente), var.t0 =
var(datos2$Agua_potable_suficiente)/length(datos2$Agua_potable_suficiente))
```

```
t.test(x=datos2$Agua_potable_suficiente, conf.level=0.95)$conf.int
```

1.1.- ¿Qué puede concluirse en relación a la puntuación media de toda la población?

$$IC_{\mu}^{0.95} = [8.46, 8.75]$$

1.2.- Si queremos ver qué ocurre en relación a los dos grupos: ¿qué puede afirmarse a la vista de los resultados obtenidos hallando los intervalos al 95%?

```
library(dplyr)

si<-filter(datos2, datos2$IDH_BAJO=="SÍ")

no<-filter(datos2, datos2$IDH_BAJO=="NO")

t.test(x=si$Agua_potable_suficiente, conf.level=0.95)$conf.int
```

t.test(x=no$Agua_potable_suficiente, conf.level=0.95)$conf.int

$$IC_{\mu_{si}}^{0.95} = [6.12, 6.65]$$

$$IC_{\mu_{no}}^{0.95} = [9.01, 9.25]$$

1.3.- Sin calcular los intervalos de confianza, ¿la conclusión sería la misma al 99% de confianza? Razone su respuesta.

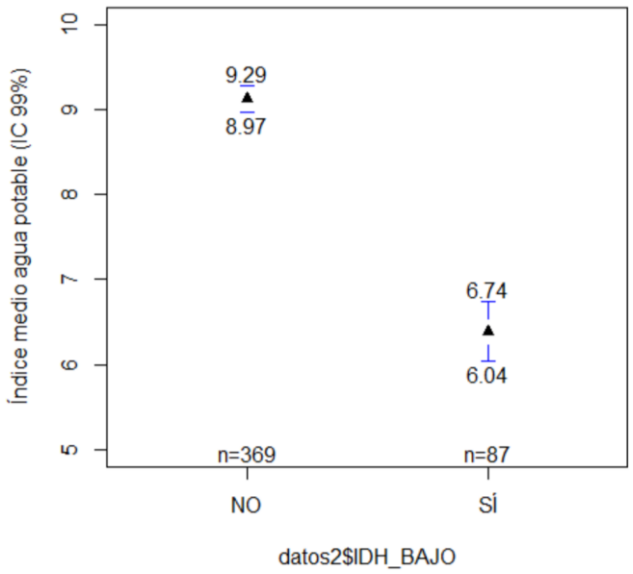
No tiene por qué

¿Y al 90% de confianza?

Sí

1.4.- Represente estos intervalos de confianza al 99%.

```
library(gplots)
plotmeans(datos2$Agua_potable_suficiente~datos2$IDH_BAJO, data=datos2,
p=0.99, ci.label=T, connect=F, mean.labels = F, pch=17, digits=2,ylim=c(5,10), col="black",
ylab="Índice medio agua potable (IC 99%)")
```



IC para la varianza de una población normal con media desconocida

IC al nivel $1-\alpha$ para la varianza σ^2 cuando μ es desconocida

$$\left(\frac{(n-1)S_c^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S_c^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

2.- Calcule un intervalo de confianza para la varianza del índice de gases de efecto invernadero a un nivel de confianza del 99%.

```
library(EnvStats); varTest(datos2$Gases_de_efecto_invernadero,  conf.level=0.99
```

$$IC_{\sigma^2}^{0.99} = [9.26, 13.04]$$

$$IC_{\sigma}^{0.99} = [3.04, 3.61]$$

IC para una proporción

IC al nivel $1-\alpha$ para la proporción p

$$\left(\hat{p} - z\alpha/_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z\alpha/_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

3.- Supongamos ahora que queremos calcular un intervalo de confianza al 99% para la proporción de países que presentan un IDH bajo.

```
length(datos2$IDH_BAJO)
```

```
table(datos2$IDH_BAJO)
```

```
prop.test(x=87, n=456, conf.level=0.99)$conf.int
```

$$IC_{p}^{0.99} = [0.14, 0.24]$$

Ejercicio Propuesto

Calcular intervalos de confianza al 95% para la media de tres índices recogidos en la base de datos suponiendo que los tres provienen de poblaciones con distribuciones normales y no se dispone de información acerca de sus varianzas. Proporcione también intervalos de confianza al 95% para las desviaciones típicas de cada una de ellas.

Parte 2: Intervalos de confianza paramétricos – 2 poblaciones

En esta parte se van a trabajar algunas de las funciones disponibles para la construcción de intervalos de confianza de los siguientes parámetros:

- Media
- Proporción
- Varianza
- Diferencia de medias
- Razón de varianzas
- Diferencia de proporciones

Intervalo de confianza para la diferencia de medias de poblaciones normales independientes con varianzas desconocidas e iguales

IC al nivel $1-\alpha$ para la diferencia de medias $\mu_1-\mu_2$

$$\left((\bar{X}_1 - \bar{X}_2) - t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}\right)$$

4.- Calcule el intervalo de confianza al 90% para la diferencia de medias entre el indicador de Alimentación Suficiente en el año 2006 entre África y Asia y deduzca si existen diferencias en el indicador medio de los países de ambos continentes. Asuma que la distribución es normal y suponga homocedasticidad entre varianzas. ¿Considera que los continentes tienen la misma media poblacional?

```
datos<-filter(datos2, Continente=="África"|Continente=="Asia", Year=="a2006")
t.test(datos$Alimentacion_suficiente~datos$Continente, var.equal=T, conf.level=0.90)$conf.int
```

$$IC^{0.95}_{\mu_{África}-\mu_{Asia}} = [-1.43, -0.37]$$

- o -

```
africa06<-filter(datos2, Continente=="África", Year=="a2006")
asia06<-filter(datos2, Continente=="Asia", Year=="a2006")

t.test(africa06$Alimentacion_suficiente, asia06$Alimentacion_suficiente, var.equal=TRUE,
conf.level=0.90)$conf.int
```

$$IC^{0.90}_{\mu_{África}-\mu_{Asia}} = [-1.43, -0.37]$$

Intervalo de confianza para la diferencia de medias de poblaciones normales independientes con varianzas desconocidas y diferentes

IC al nivel $1-\alpha$ para la diferencia de medias $\mu_1-\mu_2$

$$\left((\bar{X}_1 - \bar{X}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}}, (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}}\right)$$

5.- Realice el mismo análisis anterior, asumiendo que las varianzas poblacionales, desconocidas, son diferentes (suponga heterocedasticidad).

```
t.test(africa06$Alimentacion_suficiente, asia06$Alimentacion_suficiente,
var.equal=FALSE, conf.level=0.90)$conf.int
```

$$IC^{0.90}_{\mu_{África}-\mu_{Asia}} = [-1.42, -0.38]$$

Intervalo de confianza para el cociente de varianzas de poblaciones normales independientes

IC al nivel $1-\alpha$ para el cociente de varianzas

$$\left(\frac{S_{c1}^2/S_{c2}^2}{F_{\alpha/2,n_1-1,n_2-1}}, \frac{S_{c1}^2/S_{c2}^2}{F_{1-\left(\frac{\alpha}{2}\right),n_1-1,n_2-1}}\right)$$

6.- Calcule el intervalo de confianza a un nivel de confianza del 90% para el cociente de varianzas entre el indicador de Alimentación Suficiente en el año 2006 entre África y Asia. Puede asumirse que ambas varianzas son iguales al 90% de confianza?

```
library(stats)
res<-var.test(africa06$ Alimentacion_suficiente, asia06$ Alimentacion_suficiente, conf.level = 0.90)
res$conf.int
```

$$IC^{0.90}_{\sigma^2_{África}/\sigma^2_{Asia}} = [0.87, 2.51]$$

¿Y al 99%?

$$IC^{0.99}_{\sigma^2_{África}/\sigma^2_{Asia}} = [0.64, 3.38]$$

¿Y en el año 2016 a un nivel del confianza del 95?

$$IC^{0.95}_{\sigma^2_{África}/\sigma^2_{Asia}} = [1.22, 4.30]$$

¿Concluiríamos lo mismo al 90% de confianza? Sí ¿Y al 99%? No tiene por qué

Intervalo de confianza para la diferencia de medias de poblaciones normales no independientes

IC al nivel $1-\alpha$ para la diferencia de medias $\mu_1-\mu_2$ (poblaciones no independientes)

$$\left(\overline{D}-t\alpha_{/2}\frac{s_D}{\sqrt{n-1}},\overline{D}-t\alpha_{/2}\frac{s_D}{\sqrt{n-1}}\right)$$

7.- Construya un IC a un nivel de confianza del 95% para la diferencia de indicadores de empleo medios en los países en Centro América y Caribe entre los años 2006 y 2016.

```
centroamerica<-filter(datos2, Region_Continente=="Centro América y Caribe")
ca06<-filter(centroamerica, Year =="a2006")
ca16<-filter(centroamerica, Year =="a2016")
t.test(ca06$Empleo,ca16$Empleo, paired = TRUE, conf.level = 0.95)$conf.int
```

$$IC^{0.95}_{\mu_{D_{06-16}}} = [-1.53, 0.26]$$

¿Y entre los años 2006 y 2010?

$$IC^{0.95}_{\mu_{D_{06-10}}} = [-1.44, -0.32]$$

Intervalo de confianza para la diferencia de proporciones de poblaciones independientes

IC al nivel $1-\alpha$ para la diferencia de proporciones $p_1 - p_2$

$$\left((\hat{p}_1-\hat{p}_2)-z\alpha_{/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}},(\hat{p}_1-\hat{p}_2)+z\alpha_{/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\right)$$

6.- Calcular un IC al 97% para la diferencia entre la proporción de países de África y Asia que tienen un alto nivel de gases de efecto invernadero en el año 2016. Las Naciones Unidas declaran que se considera que los niveles de gases de efecto invernadero son perjudiciales para la salud cuando el indicador supera los 9 puntos. ¿Puede considerarse a ese nivel de confianza que las proporciones de países en riesgo son iguales en ambos continentes?

```
Africa<-filter(datos2,Continente=="África",Year=="a2016")
Asia<-filter(datos2,Continente=="Asia",Year=="a2016")
nAfr<-nrow(filter(Africa, Gases_de_efecto_invernadero>9)) ; nAsia<-nrow(filter(Asia,
Gases_de_efecto_invernadero>9))
Recuento<- c(nAfr, nAsia) ; Total<- c(nrow(Africa), nrow(Asia))
prop.test(Recuento, Total, conf.level = 0.97)$conf.int
```

$$IC^{0.97}_{p_{\acute{A}frica}-p_{Asia}} = [0.27, 0.74]$$


Ejercicio Propuesto

1. **Calcule el intervalo de confianza al nivel de 95% para la diferencia de medias entre el indicador de Alimentación Suficiente en el año 2006 entre Centro América y Caribe y Sudamérica y deduzca si existen diferencias en el indicador medio de los países de ambas regiones. Analice previamente la homocedasticidad entre varianzas. Suponga normalidad.**
2. **Construya un intervalo de confianza a un nivel de confianza del 95% para la diferencia de gases de efecto invernadero medios de los países en Europa en 2006 y 2016. Razone si la conclusión sería la misma a un 90% y a un 99% de confianza.**
3. **Construya un intervalo de confianza a un nivel de confianza del 95% para la diferencia de indicadores de igualdad de género medios de los países en Centroamérica/Caribe en 2006 y 2016. Razone si la conclusión sería la misma a un 90% y a un 99% de confianza.**
4. **Calcular un intervalo de confianza al 97% para la diferencia entre la proporción de países de América y Europa que tienen un desarrollo humano muy alto en 2016. ¿Puede considerarse a ese nivel de confianza que ambas proporciones son iguales?**

PRÁCTICA 3

Contrastes de Hipótesis paramétricos en R

Entre los tratamientos posibles para el cáncer de colon se incluyen la cirugía, radioterapia y tratamientos farmacológicos, como la quimioterapia o la inmunoterapia. La cirugía es usualmente el primer tratamiento o el tratamiento principal para las personas con cáncer de colon que no se ha propagado a partes distantes, aunque también es común administrar como tratamiento adyuvante quimioterapia tras la cirugía (que se administran de manera complementaria después del tratamiento principal). Esto ayuda a disminuir la probabilidad de que se produzca una recurrencia (reproducción).

La base de datos "colon.sav" contiene los datos de uno de los primeros ensayos con éxito de quimioterapia adyuvante para el cáncer de colon en 888 pacientes, de los que se recogió la siguiente información:

- Tratamiento: tratamiento del paciente - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU. El levamisol es un compuesto de baja toxicidad que se utilizaba anteriormente para tratar las infestaciones de gusanos en los animales; el 5-FU es un agente quimioterapéutico moderadamente tóxico.
- Edad: edad del paciente en años
- Sexo: sexo del paciente (0=Mujer, 1=Hombre)
- Obstrucción: obstrucción del colon por el tumor
- Perforación: perforación del colon
- Adherencia: adherencia a órganos cercanos
- Ganglios: número de ganglios linfáticos con cáncer detectable
- Tiempo_muerte: tiempo hasta el evento (muerte) (días)
- Diferenciación: diferenciación del tumor (1=bien, 2=moderado, 3=mal)
- Extensión: Extensión de la diseminación local (1=submucosa, 2=músculo, 3=serosa, 4=estructuras contiguas)
- Tiempo_cirugia: tiempo desde la cirugía hasta el registro (0=corto, 1=largo)
- GangliosPos4: más de 4 ganglios linfáticos positivos

Con fines didácticos se han agregado dos nuevas variables a la base de datos: peso de los pacientes antes y después del tratamiento.

Los datos proceden del conjunto colon, incluido en el paquete *survival* de R (Therneau, 2024), que contiene información de un ensayo clínico del *North Central Cancer Treatment Group* sobre cáncer de colon.

Referencia:

Therneau, T. M. (2024). *survival: A Package for Survival Analysis in R* (version 3.5-0) [Dataset colon]. R Foundation for Statistical Computing. https://CRAN.R-project.org/package=survival

Trabaje sobre la base de datos "colon.sav" y resuelva los siguientes ejercicios en R.

UNA POBLACIÓN

Contraste para la media de una población normal

1.- Los especialistas creen que la edad media a la que los pacientes desarrollan cáncer de colon es de 55 años y por eso es esa la edad a la que comienzan la realización de pruebas periódicas de revisión. Contraste la hipótesis de que la edad media de los pacientes con cáncer de colon se sitúa en 55 años. *Asuma que la distribución de la edad es normal y un nivel de significación del 5% (α=0,05).*

H_0: $\mu = 55$
H_1: $\mu \neq 55$

Nivel de significación: $\alpha = 0.05$
Estadígrafo de contraste:

$$t = \frac{\bar{x} - 55}{s_c/\sqrt{n}} = \frac{\bar{x} - 55}{s/\sqrt{n-1}} \equiv t_{n-1}$$

Valor crítico:

$$t_{n-1,\frac{\alpha}{2}} = t_{888,0.025} = 1.96$$

Región crítica: $\{t \,/\, |t| > 1.96\}$ - Región de aceptación: $\{t/\ -1.96 \leq t \leq 1.96\}$

Valor experimental: $t_{exp} = \frac{59.81-55}{11.914/\sqrt{888}} = 12{,}032$

```
t.test(colon$Edad, mu = 55)
```

```
t.test(colon$Edad, mu = 55, alternative = "two.sided",conf.level=0.95)
```

```
        One Sample t-test

data:  colon$Edad
t = 12.032, df = 887, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 55
95 percent confidence interval:
 59.02610 60.59552
sample estimates:
mean of x
 59.81081
```

P-Valor: <0.001
Decisión estadística: Se rechaza la hipótesis nula
Decisión de investigación: Se han encontrado evidencias estadísticas (altamente significativas) que indican que la edad media es distinta de 55 años.

2.- Contraste la hipótesis de que el tiempo de supervivencia media de los pacientes con cáncer de colon que **no reciben tratamiento** es inferior a 4 años (1.440 días). *Asuma que la distribución del tiempo de supervivencia es normal y un nivel de significación del 5% (α =0,05).*

En primer lugar, se seleccionan los datos a analizar:

```
library(dplyr)
data.sub<-filter(colon, Tratamiento==1)
```

A continuación, se plantea el contraste de hipótesis a examinar:

H_0: $\mu = 1440$
H_1: $\mu < 1440$

Nivel de significación: $\alpha = 0.05$
Estadígrafo de contraste:

$$t = \frac{\bar{x} - \mu_0}{s_c/\sqrt{n}} = \frac{\bar{x} - 1440}{s_c/\sqrt{n}} \equiv t_{n-1}$$

Valor crítico: $t_{n-1,\alpha} = t_{304,0.05} = 1.645$
Región crítica: $\{t \,/\, t < -1.645\}$ - Región de aceptación: $\{t \,/\, t \geq -1.645\}$

Valor experimental: $t_{exp} = \frac{1599.68-1440}{855.81/\sqrt{305}} \approx 3.259$

```
t.test(data.sub$tiempo_muerte, alternative = "less", mu = 1440)
```

```
        One Sample t-test

data:  data.sub$tiempo_muerte
t = 3.2584, df = 304, p-value = 0.9994
alternative hypothesis: true mean is less than 1440
95 percent confidence interval:
     -Inf 1680.526
sample estimates:
mean of x
 1599.675
```

P-Valor: $p > \alpha$
Decisión estadística: No se rechaza la hipótesis nula
Decisión de investigación: No se han encontrado evidencias estadísticas que indiquen que la supervivencia media sea menor de 4 años.

DOS POBLACIONES

Comparación de dos medias (muestras independientes, poblaciones normales)

1.- Analice si el Peso de los pacientes con tratamiento Lev es igual que el de los pacientes tratados con Lev+5-FU. *Asuma que los datos siguen una distribución normal y un nivel de significación del 5% (α =0,05).*

Previamente, seleccionaremos los datos:

```
trt<-filter(colon, Tratamiento>1)
```

Previo al contraste de igualdad de medias, se lleva a cabo el respectivo contraste de homocedasticidad.

$H_0: \sigma^2_{Lev} = \sigma^2_{Lev+5FU}$
$H_1: \sigma^2_{Lev} \neq \sigma^2_{Lev+5FU}$

```
library(car)
leveneTest(trt$pesoPOST,trt$Tratamiento) # SPSS

lev<-filter(colon, Tratamiento==2)
lev_fu <-filter(colon, Tratamiento==3)
var.test(lev$pesoPOST, lev_fu$pesoPOST)  #F-Fisher
```

```
        F test to compare two variances

data:  filter(colon, Tratamiento == 2)$pesoPOST and filter(colon, Tratamiento == 3)$pesoPOST
F = 1.0667, num df = 293, denom df = 288, p-value = 0.583
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8470244 1.3429533
sample estimates:
ratio of variances
          1.066667
```

Resulta un p-valor (0.583) mayor que el nivel de significación (0.05), por lo que no puede rechazarse H_0, es decir, se asume la hipótesis de igualdad de varianzas. Ahora:

"¿Existen diferencias significativas en el peso medio tras el tratamiento de los pacientes tratados con Levamisole y los tratados con el fármaco combinado?"

$H_0: \mu_{Lev} = \mu_{Lev+5FU}$
$H_1: \mu_{Lev} \neq \mu_{Lev+5FU}$

```
t.test(lev$pesoPOST, lev_fu$pesoPOST, var.equal = TRUE)
```

```
        Two Sample t-test

data:  lev$pesoPOST and lev_fu$pesoPOST
t = 1.2414, df = 581, p-value = 0.215
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7901784  3.5050761
sample estimates:
mean of x mean of y
 59.13946  57.78201
```

Estadígrafo de contraste:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \equiv t_{n_1+n_2-2}$$

Valor crítico:

$$t_{581,\alpha/2} = t_{581,0.025} = 1.96$$

Región crítica: $\{t \ / \ |t| > 1.96\}$

Región de aceptación de H_0: $\{t \ / -1.96 \leq t \leq 1.96\}$

Valor experimental:

$$t_{exp} = \frac{(59.1395 - 57.782)}{\sqrt{\frac{293 \cdot (13.41^2) + 288 \cdot (12.98^2)}{294 + 289 - 2}}\sqrt{\frac{1}{294} + \frac{1}{289}}} = 1.241$$

P-Valor: $p = 0.215 > 0.05 = \alpha$

Decisión estadística: No se han encontrado evidencias significativas para rechazar la hipótesis nula.

Decisión de investigación: No encontramos evidencias estadísticas que nos indiquen que el peso medio posterior al tratamiento sea diferente entre los grupos de pacientes que reciben cada tratamiento.

Comparación de dos medias (datos emparejados, poblaciones normales)

2.- A los pacientes se les hizo un control regular de su peso, tras ponerles el tratamiento contra el cáncer de colon correspondiente. La hipótesis de los médicos es que la administración del tratamiento combinado ayuda a estabilizar el peso de los pacientes. Para pacientes con tratamiento combinado, contrastar la hipótesis de que ha habido cambio significativo en los valores de PESO con dicho tratamiento. *Asuma que la distribución del peso es normal en ambos grupos de pacientes y un nivel de significación del 5% (α =0,05).*

H0: $\mu_d = 0$
H1: $\mu_d \neq 0$

Nivel de significación: 0.05
Estadígrafo de contraste:

$$t = \frac{\bar{x}_d - \mu_{d_0}}{s_{c_d}/\sqrt{n}} \equiv t_{n-1}$$

Valor crítico: $t_{288,\alpha/2} = t_{288,0.025} \approx z_{0.025} = 1.96$

Valor experimental: $t_{exp} = -10.069$

t.test(lev_fu$pesoPRE, lev_fu$pesoPOST, paired = TRUE)

```
        Paired t-test

data:  lev_fu$pesoPRE and lev_fu$pesoPOST
t = -10.069, df = 288, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.632971 -3.117894
sample estimates:
mean of the differences
              -3.875433
```

P-Valor: p<0.001 => $p < \alpha$
Decisión estadística: Se han encontrado evidencias altamente significativas para rechazar H0.
Decisión de investigación: Se han encontrado evidencias altamente significativas para pensar que el peso de los pacientes con el tratamiento combinado cambia.

¿Qué contraste plantearíamos en el caso de que la afirmación fuese que el peso medio para pacientes con tratamiento combinado aumenta con el tiempo? Comente los resultados obtenidos.

H0: $\mu_d = 0$
H1: $\mu_d < 0$

t.test(lev_fu$pesoPRE, lev_fu$pesoPOST, paired = TRUE, alternative = "less")

P-Valor: p<0.001 => $p < \alpha$

Decisión estadística: Se han encontrado evidencias altamente significativas para rechazar la hipótesis nula.

```
        Paired t-test

data:  lev_fu$pesoPRE and lev_fu$pesoPOST
t = -10.069, df = 288, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.240315
sample estimates:
mean of the differences
              -3.875433
```

PRÁCTICA 4: Contrastes de hipótesis no paramétricos

- 2 poblaciones -

Trabaje sobre la base de datos “colon.sav” y resuelva los siguientes ejercicios en R.

1. Queremos comprobar si el número elevado de ganglios linfáticos positivos es una variable a tener en cuenta en la supervivencia de los pacientes con cáncer de colon. ¿Existen diferencias significativas en el tiempo de supervivencia de los pacientes con más de 4 ganglios linfáticos positivos y los pacientes con menos?
   Establezca cuáles son la hipótesis nula y alternativa; el nivel de significación, el valor del estadígrafo de contraste, la región crítica, región de aceptación de la hipótesis nula, y las conclusiones estadística y clínica. Analizar primero si los datos se ajustan a una ley Normal o no y si hay homocedasticidad o heterocedasticidad y realizar el contraste sobre la edad acorde a los resultados obtenidos.

La primera hipótesis a contrastar es:

H0 ≡ La variable en estudio sigue una distribución normal

Ha ≡ La variable en estudio no sigue una distribución normal

```
ks.test(gneg$tiempo_muerte, pnorm,
mean(gneg$tiempo_muerte), sd(gneg$tiempo_muerte))

        One-sample Kolmogorov-Smirnov test

data:  gneg$tiempo_muerte
D = 0.16053, p-value = 4.885e-15
alternative hypothesis: two-sided
```

```
ks.test(gpos$tiempo_muerte, pnorm,
mean(gpos$tiempo_muerte), sd(gpos$tiempo_muerte))

        One-sample Kolmogorov-Smirnov test

data:  gpos$tiempo_muerte
D = 0.13992, p-value = 0.0002017
alternative hypothesis: two-sided
```

¿Existen diferencias significativas en el tiempo de supervivencia entre pacientes con más y menos de 4 ganglios afectados?

Nivel de significación: α=0,05

```
wilcox.test(gneg$tiempo_muerte,gpos$tiempo_muerte)
        Wilcoxon rank sum test with continuity correction

data:  gneg$tiempo_muerte and gpos$tiempo_muerte
W = 107307, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

¡cuidado con el orden de las variables!

```
> wilcox.test(gneg$tiempo_muerte,gpos$tiempo_muerte, conf.level=0.95,
+             paired=F, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  gneg$tiempo_muerte and gpos$tiempo_muerte
W = 107307, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

```
> wilcox.test(gpos$tiempo_muerte,gneg$tiempo_muerte, conf.level=0.95,
+             paired=F, alternative = "two.sided")

        Wilcoxon rank sum test with continuity correction

data:  gpos$tiempo_muerte and gneg$tiempo_muerte
W = 46149, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Se rechaza H0: Podemos afirmar que existe diferencia en el tiempo de supervivencia entre los pacientes con más de 4 ganglios positivos y menos.

2.- A los pacientes se les hizo un control regular de su peso, tras ponerles el tratamiento contra el cáncer de colon correspondiente. Analice si ha habido cambio significativo en los valores de peso de los pacientes con tratamiento combinado, antes y después del tratamiento a un nivel de confianza del 95%.

La primera hipótesis a contrastar es:

H0 ≡ La variable pesoPRE sigue una distribución normal

Ha ≡ La variable pesoPOST en estudio no sigue una distribución normal

```
t3=filter(colon, Tratamiento==3)
ks.test(t3$pesoPRE, pnorm, mean(t3$pesoPRE), sd(t3$pesoPRE))
ks.test(t3$pesoPOST, pnorm, mean(t3$ pesoPOST), sd(t3$ pesoPOST))
```

Ahora:

```
wilcox.test(t3$pesoPRE, t3$pesoPOST, paired=T,
            alternative ="two.sided")
```

```
        Wilcoxon signed rank test with continuity correction

data:  t3$pesoPRE and t3$pesoPOST
V = 8053, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

⇨ P-Valor<0.001 => Se rechaza H0.

3. Analizar si hay relación entre el tratamiento y el tiempo de supervivencia de los pacientes. Complete la siguiente tabla, analice la situación y contraste, a un nivel de confianza del 95%, si los años de vida y el tratamiento administrado están asociados.

```
tabla<-table(colon$Tratamiento, colon$tiempoREC)
colnames(tabla)<-c("<1año", "1-3 años", "3-6 años", ">6 años")
row.names(tabla)<-c("Obs", "Lev", "Combinado")
```

Porcentajes fila: `prop.table(tabla, margin=1)`

```
                <1año   1-3 años   3-6 años    >6 años
Obs        0.07540984 0.27540984 0.32459016 0.32459016
Lev        0.09863946 0.26870748 0.27551020 0.35714286
Combinado  0.08304498 0.16955017 0.32525952 0.42214533
```

Porcentajes columna: `prop.table(tabla, margin =2)`

```
              <1año  1-3 años  3-6 años   >6 años
Obs       0.3026316 0.3962264 0.3613139 0.3036810
Lev       0.3815789 0.3726415 0.2956204 0.3220859
Combinado 0.3157895 0.2311321 0.3430657 0.3742331
```

`res<-chisq.test(tabla)`

```
        Pearson's Chi-squared test

data:  tabla
X-squared = 15.117, df = 6, p-value = 0.01937
```

Valores esperados: `res$expected`

```
              <1año 1-3 años 3-6 años  >6 años
Obs       26.10360 72.81532 94.11036 111.9707
Lev       25.16216 70.18919 90.71622 107.9324
Combinado 24.73423 68.99550 89.17342 106.0968
```

Residuos: `res$residuals`

```
              <1año   1-3 años   3-6 años    >6 años
Obs       -0.6074577  1.3107274  0.5040320 -1.2257781
Lev        0.7650902  1.0516731 -1.0201281 -0.2822618
Combinado -0.1476337 -2.4072534  0.5111184  1.5439456
```

R Studio®


Nota:
Este material se ha utilizado como apoyo en la asignatura *Estadística* del Grado en Matemáticas y la Doble Titulación de Grado en Física y en Matemáticas y de la Universidad de Salamanca.