“AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN DE LA ECONOMÍA PERUANA”

UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA

FACULTAD DE CIENCIAS DE LA SALUD

ESCUELA PROFESIONAL DE MEDICINA HUMANA-EPMH

ASIGNATURA:

“SISTEMATIZACIÓN Y MÉTODOS ESTADÍSTICOS”

INTEGRANTES:

- MARÍA LUCIA JACOBO ATUNCAR
- GAMBOA CANALES MARIPAZ
- ARIANA ABIGIAL VIDAL ROMUCHO
- FERNANDA GIANELLA CASTILLA SALVADOR
- SEBASTIAN PALOMINO ROJAS
- KRISTY STEFANY ALVAREZ PEVES

DOCENTE:

DR. SEGUNDO VICENTE

SAN BORJA

2025-2

Cargar e instalar paquetes

```{r}
install.packages("car") # Para la prueba de Levene
```

```{r}
library(tidyverse)
library(here)
library(rio)
library(gtsummary)
library(car)
library(ggplot2)
```

Cargando los datos

```{r}
circun_glucosa <- import(here("cirrosis.csv"))
```

Sobre los datos para esta práctica
El dataset circun_glucosa, de 1000 personas adultas (>=20 años de edad), contiene datos glucosa medida en ayunas (en mg/dL), circunferencia de cintura (en centimetros), tabaquismo y otros datos demográficos.

```{r}
names(cirrosis)
```

```
 [1] "ID"                "Dias_Seguimiento"
 [3] "Estado"            "Medicamento"
 [5] "Edad"              "Sexo"
 [7] "Ascitis"           "Hepatomegalia"
 [9] "Aracnoides"        "Edema"
[11] "Bilirrubina"       "Colesterol"
[13] "Albumina"          "Cobre"
[15] "Fosfatasa_Alcalina" "SGOT"
[17] "Trigliceridos"     "Plaquetas"
[19] "Tiempo_Protrombina" "Etapa"
```

1.1 El problema en este ejercicio

El desenlace Y de interés para este ejercicio es la variable glucosa medida en ayunas. Veamos la distribución de la variable y el promedio en en un histograma.

```{r}
cirrosis |>
  ggplot(aes(x = Colesterol)) +
  geom_histogram(
    color = "white",
    ) +
  labs(y = "Frecuencia",
       x = "resultados") +
  geom_vline(xintercept = mean(cirrosis$Colesterol, na.rm = TRUE),
             color = "darkorange", size = 1.5)
```

En estos datos, el promedio de la glucosa es:

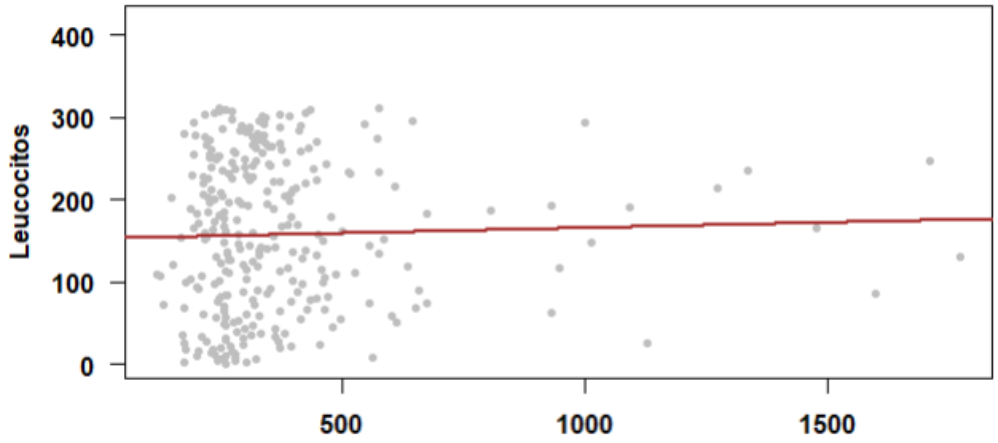
```{r}
mean(cirrosis$Colesterol, na.rm = TRUE)
```

Una observación importante a partir del histograma y el promedio (el valor esperado) es que existe una gran variación entre los valores de glucosa de los individuos de quienes provienen los datos. Podemos hipotetizar de que otras variables (predictores) podrían influir en esta variación, por ejemplo, la circunferencia de cintura.

1.2 Notación en el método de regresión lineal simple

```{r}
plot(ID ~ Colesterol , data = cirrosis,
     col = "gray",
     ylab = "Leucocitos",
     xlab = "edad",
     las = 1,
     pch = 20,
     font.lab = 2, font.axis = 2)

# La función lm() ajusta el modelo de regresión lineal
abline(lm(ID ~ Colesterol , data = cirrosis), lwd = 2, col = "brown")
```



La ecuación siguiente ecuación describe un modelo de regresión lineal simple para Y usando un predictor continuo X.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Cuando ajustamos un modelo de regresión lineal simple a nuestros datos, estimamos (hallamos) los parámetros del modelo que mejor explican la relación entre las dos variables (desenlace y predictor), incluyendo los coeficientes (β_0, β_1) y el error (ε), que representa la variabilidad no explicada por el modelo.

Para un predictor continuo, el intercepto (β_0) es el valor esperado de Y cuando X = 0 (es decir, el promedio del resultado cuando el predictor es cero). La pendiente (β_1) es el cambio promedio en Y por cada unidad de cambio en X. El término de error (ε) representa la diferencia entre los valores observados y los valores predichos por el modelo.

Aplicado a nuestro ejemplo, el intercepto (β_0) representa la circunferencia de cintura promedio cuando la glucosa en ayunas es cero (aunque este valor puede no tener sentido práctico, es necesario matemáticamente). La pendiente (β_1) indica cuánto aumenta (o disminuye) en promedio la circunferencia de la cintura por cada unidad

adicional de glucosa en ayunas (medida en mg/dL). El error (ε) recoge la variación individual que no es explicada solo por la glucosa.
Asi que, como el objetivo es hallar los valores de los parámetros (β_0,β_1,ε), es apropiado decir que estamos 'ajustando el modelo de regresión lineal simple' para el problema planteado (a.k.a la asociación entre glucosa y la circunferencia de cintura)

1.3 Ajustando el modelo de regresión lineal simple para nuestro problema

En R, usamos la función lm() para ajustar un modelo de regresión lineal. "lm" es la abreviatura para "linear model". Dentro de la función debemos indicarle como argumentos el desenlace X, el predictor Y y la data donde se encuentran las variables. Esta es la estructura para ajustar el modelo con la función lm: lm(y ~ x, data = mis_datos).

Ajustando el modelo para nuestros datos

```{r}
modelo_ejemplo = lm(ID ~ Colesterol, data = cirrosis)
```

```{r}
summary(modelo_ejemplo)
```

```
Call:
lm(formula = ID ~ Colesterol, data = cirrosis)

Residuals:
     Min       1Q   Median       3Q      Max
-155.935  -78.905   -1.394   81.004  154.247

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 153.55061   10.22862  15.012   <2e-16 ***
Colesterol    0.01297    0.02346   0.553    0.581
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.53 on 282 degrees of freedom
  (134 observations deleted due to missingness)
Multiple R-squared:  0.001082,  Adjusted R-squared:  -0.00246
F-statistic: 0.3055 on 1 and 282 DF,  p-value: 0.5809
```

Para ver los resultados, usamos la función summary() y dentro, el objeto modelo_ejemplo.

1.4 Interpretando los resultados

La sección Coefficients del resultado:

```{r}
summary(modelo_ejemplo)$coef
```

```
                Estimate Std. Error    t value
(Intercept) 153.55061416  10.228620 15.0118599
Colesterol    0.01296562   0.023457  0.5527401
                Pr(>|t|)
(Intercept) 7.705241e-38
Colesterol  5.808792e-01
```

1.5 ¿Cómo reportar los resultados del ajuste del modelo de regresión lineal simple?

Tanto si se trata de una tesis o un artículo, abajo un ejemplo de cómo reportar los resultados del presente problema:

Adicionalmente, es buena idea presentar los resultados en un tabla.

```{r}
theme_gtsummary_language("es")

tabla_reporte <- modelo_ejemplo |>
  tbl_regression(intercept = T, estimate_fun = function(x) style_sigfig(x, digits = 4),
                 pvalue_fun    = function(x) style_pvalue(x, digits = 3),
                 label = list(Colesterol ~ "volumen Colesterol")) |>
  modify_caption("resultado de colesterol")

tabla_reporte
```

resultado de colesterol

Característica	Beta	95% CI	p-valor
(Intercept)	153.6	133.4, 173.7	<0.001
volumen Colesterol	0.0130	-0.0332, 0.0591	0.581

Abreviacion: CI = Intervalo de confianza

Exportamos la tabla

```{r}
tabla_reporte |>
  as_flex_table()  |>
  flextable::save_as_docx(path = "tabla_reporte.docx")
```

2 Prueba t de Student para muestras independientes

Imagina que, ahora, luego de haber tomado las mediciones de medidas de glucosa en ayunas (mg/dL) queremos saber si el promedio de glucosa en varones es significativamente diferente del promedio de glucosa en mujeres. Es esta situación, hay dos grupos (varones y mujeres) de muestras independientes.

2.1 ¿Cuándo usar la prueba t de Student para muestras independientes?

Cuando los dos grupos de muestras a comparar han sido muestreadas de una distribución normal. Aquí podemos usar la prueba de Shapiro-Wilk.

Cuando las varianzas de los dos grupos son iguales. Esto puede ser evaluado con la prueba F.

Usualmente, la hipótesis de la prueba t de Student son:

- Hipótesis nula (H_0): No hay diferencia entre las medias de los dos grupos.

$$H_0 : \mu_1 = \mu_2$$

- Hipótesis alternativa (H_1): Hay una diferencia entre las medias de los dos grupos.

$$H_1 : \mu_1 \neq \mu_2$$

2.2 Sobre los datos para esta práctica

El dataset circun_glucosa, de 1000 personas adultas (>=20 años de edad), contiene datos circunferencia de cintura (en centímetros), la variable sexo y otros datos demográficos.

2.3 Resumen y visualización

Resumen

Antes de realizar la prueba t de Student es importante conocer la distribución de los datos e identificar si hay valores perdidos o atípicos. Empecemos por el resumen:

```r
group_by(cirrosis, Edema) |>
  summarise(
    count = n(),
    mean = mean(Ascitis, na.rm = TRUE),
    sd = sd(Colesterol, na.rm = TRUE)
  )
```

tbl_df

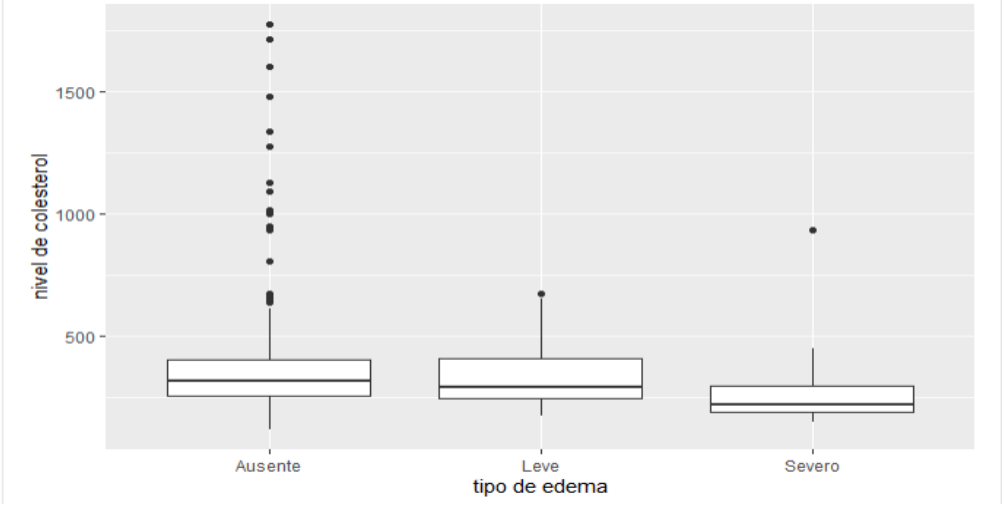
A tibble: 3 × 4

Edema <chr>	count <int>	mean <dbl>	sd <dbl>
Ausente	354	NA	241.8292
Leve	44	NA	135.0830
Severo	20	NA	186.2010

3 rows

Visualización

```r
cirrosis |>
  filter(!is.na(Colesterol)& !is.na(Edema))  |>
  ggplot(aes(y = Colesterol, x = Edema)) +
  geom_boxplot() +
  labs(y = "nivel de colesterol", x = "tipo de edema")
```



2.4 Pruebas preliminares para evaluar los supuestos de la prueba t de Student

Supuesto 1: los datos deben haber sido muestreados de una distribución normal. Para esto, usamos la prueba de Shapiro-wilk.

```{r}
cirrosis |>
  filter(Edema == "Severo") |>
  summarise(shapiro = list(shapiro.test(Plaquetas))) |>
  pull(shapiro)
```

```
[[1]]

        Shapiro-Wilk normality test

data:  Plaquetas
W = 0.93589, p-value = 0.2003
```

```{r}
cirrosis |>
  filter(Edema == "Ausente") |>
  summarise(shapiro = list(shapiro.test(Plaquetas))) |>
  pull(shapiro)
```

```
[[1]]

        Shapiro-Wilk normality test

data:  Plaquetas
W = 0.97477, p-value = 1.02e-05
```

Supuesto 2: Las varianzas de los dos grupos son iguales Para esto podemos usar la prueba F para evaluar la homogeneidad de varianzas. Esto esta implementado en la función var.test()

```{r}
var.test(Plaquetas ~ Aracnoides, data = cirrosis)
```

```
        F test to compare two variances

data:  Plaquetas by Aracnoides
F = 0.87656, num df = 217, denom df = 89, p-value
= 0.4411
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6093036 1.2286573
sample estimates:
ratio of variances
         0.8765571
```

El valor p de la prueba F es p = 0.8765571. Es mayor que el nivel de significancia α = 0.05. En conclusión, no hay una diferencia significativa entre las varianzas de los dos conjuntos (femenino y masculino) de datos. Por lo tanto, podemos usar la prueba t clásica que asume igualdad de varianzas

2.5 Realizamos la prueba t para nuestros datos.

```r
{r}
t.test(Plaquetas ~ Aracnoides, data = cirrosis, var.equal = TRUE)
```

```
        Two Sample t-test

data:  Plaquetas by Aracnoides
t = 2.8611, df = 306, p-value = 0.004512
alternative hypothesis: true difference in means between group No and group Sí is
not equal to 0
95 percent confidence interval:
 10.57863 57.17998
sample estimates:
mean in group No mean in group Sí
        271.8349         237.9556
```

3.1 ¿Cuándo usar el ANOVA de una vía?

Las observaciones se obtienen de forma independiente y aleatoria de la población definida por los niveles del factor.

Los datos de cada nivel del factor se distribuyen normalmente.

Hipótesis nula (H_0): No hay diferencia entre las medias de los dos grupos.
Estas poblaciones normales tienen una varianza común. (Se puede usar la prueba de Levene para verificar esto.)

3.2 Sobre los datos para esta práctica

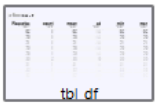
El dataset circun_glucosa, de 1000 personas adultas (>=20 años de edad), contiene datos de peso corpotal (kg), la variable tabaquismo y otros datos demográficos.

3.3 Resumen y visualización

Resumen

Antes de realizar la prueba de ANOVA es importante conocer la distribución de los datos e identificar si hay atípicos. Empecemos por el resumen:

```r
{r}
group_by(cirrosis, Plaquetas) |>
  summarise(
    count = n(),
    mean = mean(Plaquetas, na.rm = TRUE),
    sd = sd(Plaquetas, na.rm = TRUE),
    min = min(Plaquetas, na.rm = TRUE),
    max = max(Plaquetas, na.rm = TRUE) )
```

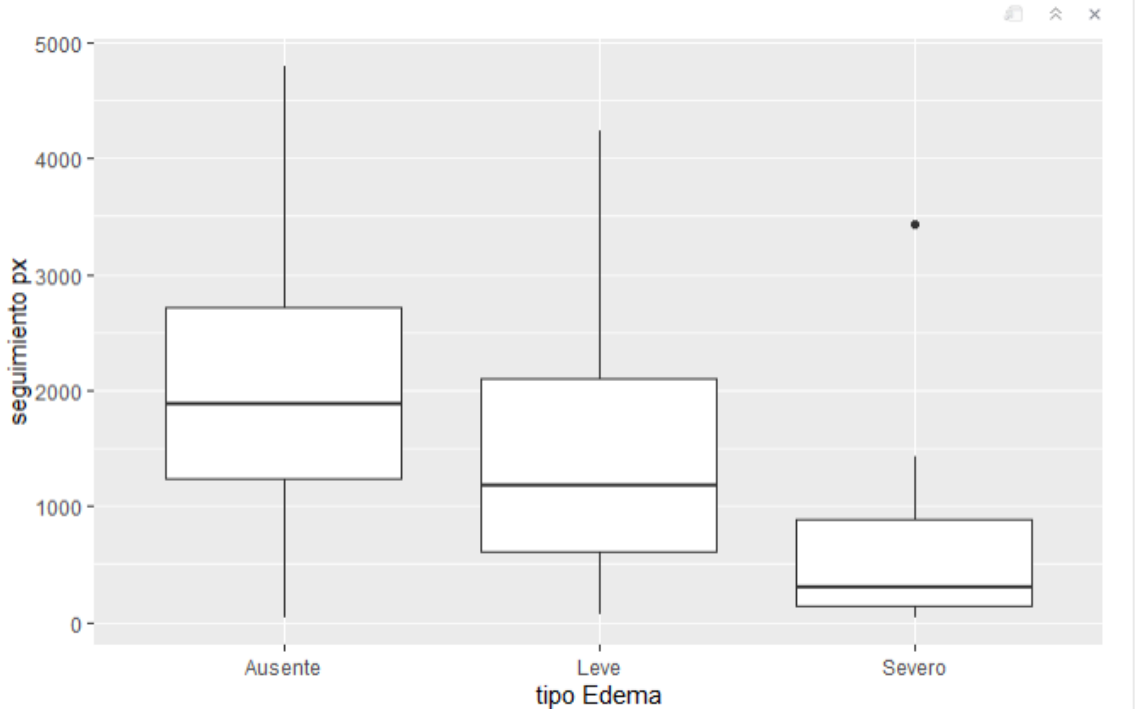
A tibble: 244 × 6

Plaquetas <dbl>	count <int>	mean <dbl>	sd <dbl>	min <dbl>	max <dbl>
62	1	62	NA	62	62
70	1	70	NA	70	70
71	1	71	NA	71	71
76	1	76	NA	76	76
79	1	79	NA	79	79
80	2	80	0	80	80
81	1	81	NA	81	81
88	1	88	NA	88	88
92	1	92	NA	92	92
95	2	95	0	95	95

1-10 of 244 rows Previous 1 2 3 4 5 6 ... 25 Next

Visualización

```{r}
cirrosis |>
  filter(!is.na(Dias_Seguimiento)& !is.na(Edema))  |>
  ggplot(aes(y = Dias_Seguimiento, x = Edema)) +
  geom_boxplot() +
  labs(y = "seguimiento px", x = "tipo Edema")
```



3.4 Pruebas preliminares para evaluar los supuestos del ANOVA

```{r}
cirrosis <- cirrosis |>
    mutate(Edema = as.factor(Edema))
```

Supuesto 1v: los datos deben haber sido muestreados de una distribución normal.

Para esto, usamos la prueba de Shapiro-wilk.

```{r}
cirrosis |>
  filter(Edema == "Severo") |>
  summarise(shapiro = list(shapiro.test(Albumina))) |>
  pull(shapiro)
```

```
[[1]]

        Shapiro-Wilk normality test

data:  Albumina
W = 0.98245, p-value = 0.9617
```

```{r}
cirrosis |>
  filter(Edema == "Leve") |>
  summarise(shapiro = list(shapiro.test(Albumina))) |>
  pull(shapiro)
```

```
[[1]]

        Shapiro-Wilk normality test

data:  Albumina
W = 0.97632, p-value = 0.4936
```

```{r}
cirrosis |>
  filter(Edema == "Ausente") |>
  summarise(shapiro = list(shapiro.test(Albumina))) |>
  pull(shapiro)
```

```
[[1]]

        Shapiro-Wilk normality test

data:  Albumina
W = 0.98862, p-value = 0.007275
```

```{r}
leveneTest(Albumina ~ Edema, data = cirrosis)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.4579 0.6329
      415
```

3.5 Realizamos la prueba de ANOVA de una vía para nuestros datos.

```{r}
res_anova = aov(Albumina ~ Edema, data = cirrosis)
```

```{r}
summary(res_anova)
```

Interpretando los resultados

Dado que el valor p es mayor que el nivel de significancia 0.05, podemos concluir que no existen diferencias significativas entre los grupos.

Aunque para este ejecicio no hemos encontrado una diferencia estadísticamente significativa, cuando sí lo hay, es importante realizar una prueba de comparación por pares para saber dónde se encuentra la diferencia. Para esto, se puede utilizar la prueba Tukey HSD (Tukey Honest Significant Differences)

```
{r}
TukeyHSD(res_anova)
```

MÉTODO DE REGRESIÓN EN R STUDIO

Paso 1: Preparar los datos: Primero, necesitas tener tus datos en un data.frame. Por ejemplo:

```
{r}

# Crear datos de ejemplo
datos <- data.frame(
  x = c(1, 2, 3, 4, 5),
  y = c(2, 4, 5, 4, 5))
```

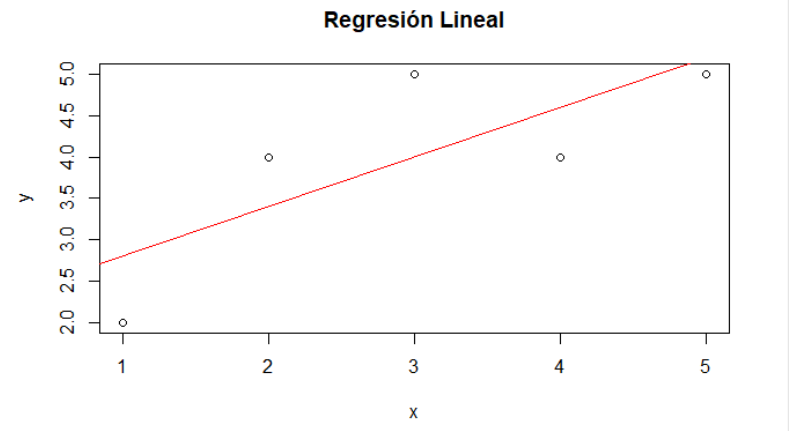
Paso 2: Ajustar el modelo de regresión: Usa la función lm() para ajustar el modelo:

```
{r}
modelo <- lm(y ~ x, data = datos)
```

```
{r}
summary(modelo)
```

Paso 3: Graficar la regresión: Para visualizar la línea de regresión:

```
{r}
plot(datos$x, datos$y, main = "Regresión Lineal", xlab = "x", ylab = "y")
abline(modelo, col = "red")
```



Paso 4: Diagnóstico del modelo

Puedes verificar los residuos y otros diagnósticos con:

```r
par(mfrow = c(2, 2))
plot(modelo)
```