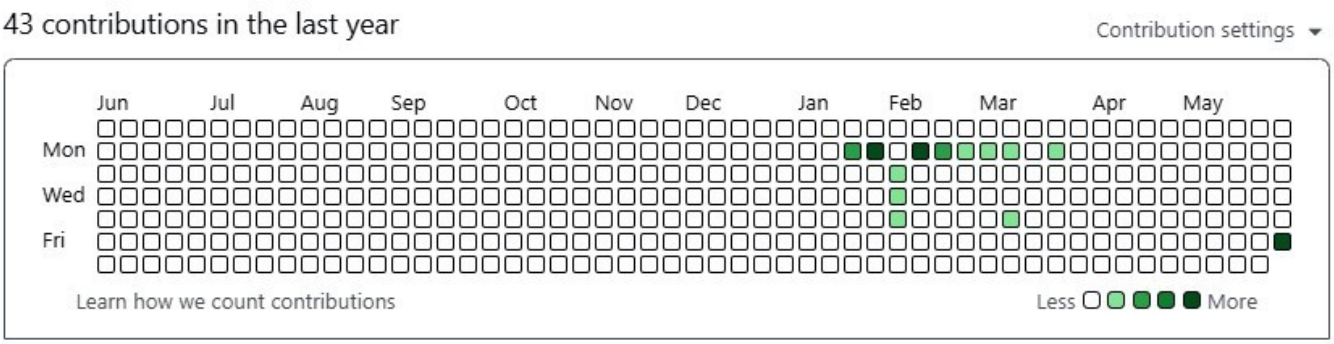
Tania Isela Sarmiento Muñoz

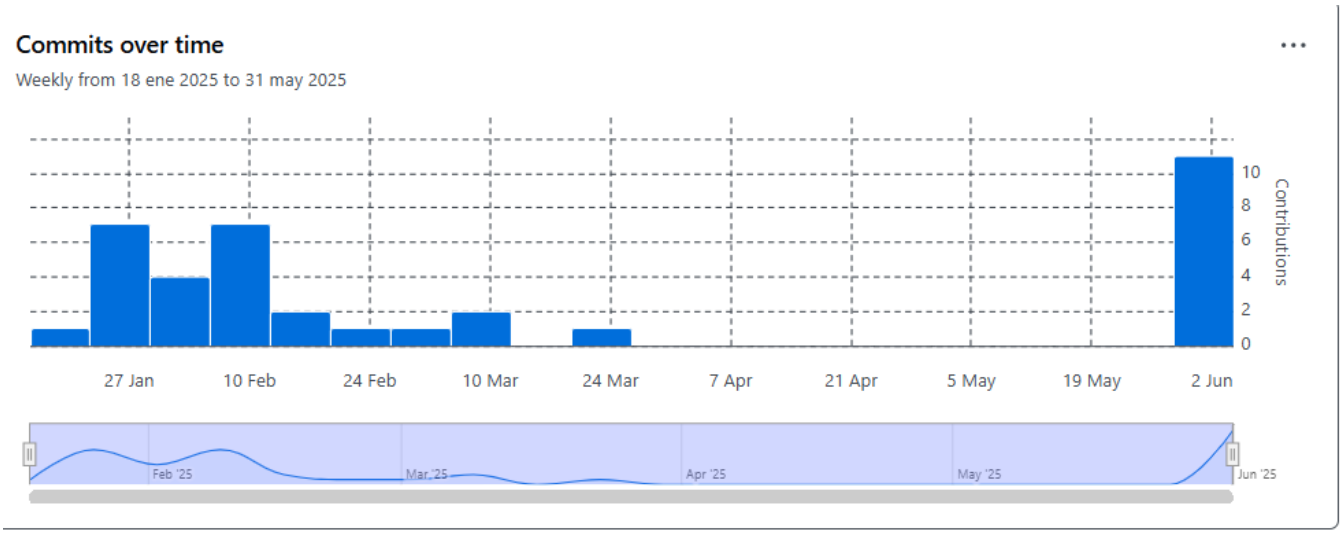
Matricula: 1503831

Estádistica posgrado 25

- Estadística en la Investigación Científica

Contenido del curso (8 semanas 20/01/2025 - 10/03/2025)





Carpeta y Scripts

https://github.com/tania-sarmz/Est_posgrado25

Clase 1: 20.01.2025 (Creación del repositorio *Github*)

✅ Lenguaje *Markdown*

- Descargamos Software R versión 4.2.3.
- Exploramos la ventana de Consola para comandos simples
- Mediante Nuevo Archivo, abrimos un Archivo de tipo *R Markdown*

🔲 Lenguaje *Markdown*: Lenguaje de programación más sencillo (WhatsApp, Word, etc.), se utiliza para marcar titulos, listados, tablas de contenido

✅ Creación de repositorio en *Github*

- Se accesa a la plataforma en línea (https://github.com/)
- Mediante la opción *Sign up* creamos un usuario y contraseña
- Posteriormente mediante la opción *Create repository* creamos un repositorio con las siguientes características:
    - Nombre del repositorio: Est_posgrado25
    - Descripción: Estadística en la Investigación Científica
    - Tipo: Pública
    - Marcar la opción de iniciar repositorio con un archivo *README*

🔲 Archivo *README*: Archivo de texto en el cual se redactarán las actividades que se realizan en el repositorio. Esta abierto al público en general.

-- FIN DE CLASE --

Clase 2: 27.01.2025 (Clonación del repositorio y creación del primer proyecto en Rstudio)

✅ Uso de la aplicación *Git Bash*

- Se descarga la aplicación Git Bash

🔲 Aplicación *Git Bash*: Sistema de control gratis y abierta para el manejo de proyectos desde una consola directa en nuestra computadora.

- Se abre el programa R versión 4.2.3.
- En la pestaña de Terminal se revisa la conexión con la nube de la computadora git mediante el comando: git--version
- Para sincronizar los comandos que se trabajan en la consola de R con la aplicación *Github* se siguen los siguientes pasos:
    - Configurar el nombre de usuario mediante el comando: $ git config –global user.name “taniaPC”
    - Se recomienda un usuario para PC y uno para LAPTOP, esto para identificar desde que equipo se esta trabajando o desde cual se hicieron correcciones o avances. Son conexiones locales, por ello se evita duplicidad de usuarios
    - Configurar correo mediante el comando: $ git config --global user.email “tania.sarmientomnz@uanl.edu.mx”
    - Con los avances se sincronizan desde cualquier dispositivo, guardando los cambios.
    - Verificar nombre usuario mediante el comando: $ git config --list

✅ Pasos para realizar la primera clonación

- Se abre el programa R versión 4.2.3. y nos colocamos en la pestaña de Consola
- Se accesa a repositorio Est_posgrado25 desde *Github*

- Desde la opción de *Code* en *Github* se abre la pestaña Local y se selecciona el apartado HTTPS para mostrar la dirección de acceso del repositorio: https://github.com/tania-sarmz/Est_posgrado25.git
- En R desde el menú Archivo se selecciona la opción de Crear proyecto
- En la ventana que se despliega se seleccionan las siguientes opciones: Version Control / Git
- En el espacio de la URL del respositorio se pega la dirección de acceso del repositorio: https://github.com/tania-sarmz/Est_posgrado25.git
- En el nombre del directorio del proyecto se coloca el nombre del repositorio: Est_posgrado25
- En la ruta de guardado se crea una carpeta directa en el Disco local C: Repositorios
- Como resultado de la clonación se van mostrado los archivos guardados en la ventana inferior izquierda

✅ Pasos para la sincronización al finalizar cada sesión de trabajo

- Desde la ventana superior izquierda se abre la pestaña nombrada Git
- Se seleccionan mediante el *checkmark* los archivos generados o actualizados durante la sesión de trabajo
- Se activa la "palomita" verde en la opción de Commit
- Se abre una ventana en la cual se resumen los comandos trabajados (todo lo eliminado en rojo y lo agregrado en verde)
- Se describe un mensaje de actualización en el apartado de *Commit message*
- Primero se selecciona la opción Pull (señalado con una flecha apuntando hacia abajo azul)
- Luego se selecciona la opción Push (señalado con una flecha apuntando hacia arriba verde)
- Al finalizar el proceso se debe arrojar el mensaje "Already up to date"
- Al volver a *Github* se deben mostrar los archivos generados y actualizados.

✅ Creación y lenguaje de un *Script*

- Desde el botón de Nuevo Archivo se selecciona la opción para desplegar la lista de archivos
- Se selecciona la primera que corresponde a R Script
- Al seleccionarla se abre la ventana para indicar la ruta de guardado y nombre del Script

🔲 *Script*: Espacio de trabajo para colocar el texto mediante el cual se describen los comandos de forma ordenada

- Todo los comentarios que no correspondan a un comando deben indicarse con un #
- Si se desea insertar una sección dentro del Script, se selecciona desde el menú *Code* la opción *Insert Section*
- Se abre una ventana para indicar el nombre de la sección a colocar
- Para hacer operaciones básicas se escribe el código mediante simbolos como 3 + 5
- Se selecciona la opción Run (resaltando el código) para correr la operación
- En la consola se muestra resultado de la operación
- Para asignar nombre a alguna operación se utilizan los simbolos <-
- Ya solamente se escribe el nombre asignado al código y se corre la formula asignada
- Se puede utilizar un nombre asignado en conjunto con otras funciones
- Se puede asignar datos a una variable mediante parentesis y comas, como se muestra a continuación: edad <- c (50, 25, 30, 18)
- Algunas recomendaciones de escritura de códigos:
    - Hay que utilizar de forma correcta los códigos asignados.

- Rstudio si distingue entre mayúsculas y minúsculas, según lo asignado.
- Librería para hacer graficas: GGPLOT2
- Graficas interactivas: shiny appt
- Mejores programas para hacer gráficas MathLab y Rstudio. Da mejor resolución de imagen y más personalizadas, a diferencia de gráficas Excel.

Clase 3: 10.02.2025 (Importación de base datos)

✅ Importación de datos

- Se bajo base de datos en linea de la página: **https://mgtagle.wordpress.com/category/laboratorios-r/base-de-datos/ https://mgtagle.wordpress.com/category/laboratorios-r/base-de-datos/** Específicamente trabajamos con los datos de *Ocampo Ocampo* refiere a una base de datos en tiempo real de un periodo determinado. Describe un número de grupos de variables.

- Pasos seguidos:

  - Se descargo el archivo en Excel.
  - Se convirtió a formato CSV. Nota: En McCintosh es en formato TF8.
  - Se guardo en la ruta del Respositorio, para una mejor accesibilidad.
  - Al guardarlo en el repositorio aparece de manera automática en la ruta de archivos y ya se puede abrir el archivo.
  - Ver Archivo, mediante la opción de View File.

- Abrimos Script 2 para ir guardando los comandos.

  - Para ver el contenido de los datos usar la Pestaña Environment
  - Al darle click se abre una pestaña mostrando los datos
  - También se puede usar la Función View

- Pasos para importar:

  - Darle objeto a mis datos mediante la función de asignar (<-) e indicar la importación de datos mediante la función de read.csv
  - Se coloca el nombre del archivo entre comillas “nombre del archivo”
  - Se sigue de una coma, para indicar la función header, la cual nos traerá el nombre de las columnas usando la indicación TRUE para que se las traiga igual
  - Para visualizar los tipos de datos importados, se utiliza la flecha azul, ubicada a la izquierda del nombre del archivo en la pestaña Environment.
  - Para visualizar los nombres de las variables o columnas usar la función names. Para promediar datos de una columna en la base de datos, se utiliza la función mean, dentro de la cual se asigna el nombre de la base de datos y se especifica mediante $ el nombre de la variable, o nombre de columna donde se encuentran los datos.

- Tipos de gráficas:

  - Histogramas: determinan frecuencias de categorías.
  - Barplot: datos acumulados o totales existentes (valor absoluto) de una variable
  - Pastel: solo describen la proporción

    - Boxplot: indica valor mínimo y máximo, se distribuye por cuartiles (25, 50, 75, 100%), solamente se indica en la cajar el 1º y 3er cuartil, la línea negra indica la mediana, la caja es el 50% de mis datos.
    - Tallo y hoja

✅ Insertar Boxplot y características:

- Por comas e instrucciones puedo ir indicando las características del gráfico
- Por ejemplo color = col; Titulo de gráfico = main; nombre de eje x = xlab; nombre de eje y = ylab.
- Para visualizar una gráfica de manera completa, o en una ventana, utilizar la función de zoom de la pestaña Plots.
- Valores máximos y mínimos pueden indicar errores o situaciones atípicas. Esos valores se pueden sacar o a tratar diferente, porque afecta al valor medio de mis datos.

✅ Características de histogramas:

- Usar la función hist, para insertar un histogramas de datos. Seguido del nombre de la base de datos y luego coma, $ para indicar variables deseadas o nombre de columna de los datos.
- Cada barra indica un intervalo de datos
- Para que en la gráfica se visualicen intervalos se usa la función de breaks, usando la amplitud y número de observaciones. Para indicar el número de barras que me conviene visualizar.

✅ Gráficas de tallo y hojas:

- Función de stem
- Se utilizaba anteriormente para ordenar los datos conforme a grupos de temperatura
- Desglosa el número de datos

✅ Estadística paramétrica:

- Todo tiene una contraparte en la estadística no paramétrica
- Detecta los cambios mínimos, es más robusta
- Se determina por la naturaleza de los datos.
- Primero se intenta la transformación de los datos, para no usar la estadística no paramétrica.
- Por ejemplo: Prueba de t, Pearson

✅ Estadística no paramétrica:

- Menos robusta, no detecta cambios sutiles.
- La diferencia entre los datos tiene que estar bien marcada.
- Por ejemplo: Prueba de rangos

✅ Comparaciones simples:

- Experimento balanceado, mismo número de representaciones de las condiciones / variables determinadas.
- Comparación de medias, requiere ajustes. Por ejemplo corrección Bonferroni.
- Se afina con la lectura de los datos con los que estoy trabajando.

✅ Datos de vivero (ejemplo de base de datos):

- En las mediciones se indica tamaño de tallo y raíz (en conjunto)
- Ctrl = sólo riegos
- Fert = se le aplico fertilizante
- Objetivo: Prueba de comparación de medias de muestras independientes
- Atajo de <- es Alt -

✅ Función de factores (Para que los tratamientos me los trate como factores)

- Declarar la variable de Tratamiento como factor, con la función as.factor
- Por ejemplo: vivero$Tratamiento <- as.factor (vivero$Tratamiento)

✅ Indicación de factores en las gráficas (Para que la gráfica indica por factores/ tratamientos)

- Se usa el símbolo ~ el cual se pone por comando de teclado AltGr *
- Por ejemplo: boxplot (vivero$IE ~ vivero$Tratamiento)
- Como es una sola variable para dos tratamientos se utilizar el mismo color para ambas barras.
- Se observan tendencia de los datos
- En este ejemplo son datos de muestras independientes

✅ Pruebas de t:

- Comparaciones de dos medias
- Solamente se utiliza para dos tratamientos
- Tres tipos :1 muestra, dos muestras independientes, dos muestras dependientes.
- *1 muestra* = existe una media teorética determinada. Por ejemplo tamaño de planta establecida.
- *2 muestras independientes* = comparación de dos grupos diferentes pero una variable en común. Por ejemplo altura entre hombres y mujeres al mismo tiempo.
- *dos muestras dependientes* = una variable de un grupo de individuos que se mide a través del tiempo. Por ejemplo producción de un año con respecto a otro.
- Tienen que tener una distribución normal, se forma una campana de Gauss = normalidad
  - Shapiro – Wilkins, menos robusta
  - Kolmogorov, más estricta
- Debe cumplir homogeneidad de varianzas = homogeneidad

✅ Hipotésis:

- Hipótesis nula = no hay diferencias
- Hipótesis alternativa = existen diferencias significativas entre las variables que quiero comparar.

✅ Características de los análisis estadísticos:

- Alfa (0.05) = límite de trabajo aceptable para interpretar a la hipótesis como nula o alternativa.
- En ciencias naturales o sociales el valor más aceptable es el 0.05.
- Valor de p-value = sirve de contraste con respecto al límite, o sea alfa. Oscila de 0 a 1. Valores menores a 0.05 indican aceptación de hipótesis nula. Valores mayores a 0.05 indican aceptación de hipótesis alternativa.

Clase 4: 17.02.2025 (Estadística de los datos)

✅ Medidas de tendencia

- Medidas de tendencia central
  - Media
  - Moda
  - Mediana
- Medidas de dispersión
  - Varianza
  - Desviación estandár
  - Coeficiente de variación

✅ Tipos de contraste

- Una cola: Solamente diferencias significativas entre dos grupos de datos dependientes
  - Debes saber si el software te lo esta distribuyendo en una cola o dos colas. Saber que se esta utilizando en los tipos de software.
- Dos colas: Solamente diferencias significativas entre dos grupos de datos independientes
  - Se le especifica en Rstudio, si se desea que sea de dos colas; aunque lo coloca por default

✅ Para hacer prueba de t

- Función que se utiliza: t.test
- Observación de los datos:
  - Aunque aparentemente la media es diferente, ¿como lo se estadísticamente?
  - ¿Es significativa o no es significativa?
  - Se la aplica la normalidad al índice de esbeltez
- Prueba de Shapiro (búsqueda de la normalidad)
  - Ho = son normales
  - H1 = no son normales
  - Nuestra orientación es conforme al p-value
  - Prueba de normalidad se realiza con una prueba de una cola.
- Restricciones R en las bases de datos
  - == igual
  - != diferente a
  - < mayor
  - > menor
  - <= mayor o igual a
  - >= menor o igual a ✅ Función subset, para hacer un subconjunto
  - Es para extraer una parte específica de la base de datos
  - Como un tipo filtro en Excel ✅ Comparación de varianzas (homogeneidad de varianzas
- Esta ligeramente “0.003” mayor, por lo que se acepta la hipótesis nula, cumple homogeneidad de varianzas.
- Si hubiese salido 0.049 es menor, por lo que se acepta la hipótesis alternativa, que no cumple homogeneidad de varianzas

- var. equal, me indica que mis varianzas son iguales (normales). Si no se coloca te “proyecta” los grados de libertad y el valor de p cambia ligeramente, lo cual puede afectar cuando las diferencias no son tan marcadas.
- t= valor estimado de la prueba = -2.9813
- df = grados de libertad = 40
- p-value = valor de p = 0.004868, es menor a 0.025, por lo que se acepta la hipótesis alternativa, hay diferencias significativas. En este caso, es MUY marcada la diferencia menor a 0.025, por lo que no es estrictamente necesario verificar distribución de 1 cola.

✅ Prueba de t (suponiendo que los datos son dependientes)

- Se utiliza la función paired, para indicar que son datos dependientes.
- Cambian los grados de libertad a 21 -1, ya que son 21 individuos tomados en dos momentos diferentes.
- El valor de p es menor a 0.025, por lo que si hay diferencias significativas. Lo cual en un ejemplo hipotético nos dice que la acción tomada en un momento diferente, si hace una diferencia en el “índice de esbeltez”.

✅ Prueba de t (suponiendo que los datos son de 1 muestra)

- Se establece una media teorética (mu =), por ejemplo “CONAFOR establece un índice de esbeltez de 0.85”
- De manera absoluta el promedio de IE es de 0.83, pero hay que probarlo ahora estadísticamente.
- El valor de p es mayor a 0.025, por lo que las medias son iguales; no hay diferencias significativas.

✅ Prueba de t (suponiendo que los datos son de 1 muestra)

- Si la media teorética fuese 0.90
- Entonces p es menor a 0.025, por lo que las medias ya NO son estadísticamente iguales.

Clase 5: 24.02.2025 (Prueba de varianzas)

✅ Pruebas de varianza (tres tratamientos o más diferentes)

- Análisis de varianzas para tres tratamientos o más
- Revisar distribución normal
- Revisar homogeneidad de varianzas
- Hay varianza entre los grupos (varianza entre niveles de factores, tratamientos)
- Hay varianza dentro de los grupos (varianza entre cada bloque, observaciones, grupos de individuos)
- Si se encuentra diferencia en por lo menos una varianza, se aplica una diferencia de medias (prueba de Tukey la más común, solo se aplica a experimentos balanceados).
- Diseño de experimentos balanceados, mismo número de observaciones por bloque.
- Cuando no hay experimentos balanceados, se pueden aplicar correcciones, como la de Bon Ferroni (prueba de medias diferente a Tukey).
- Observaciones, número de individuos en cada bloque.
- Repeticiones, número de veces que se repiten un número de observaciones por los bloques en total.

✅ Niveles de factor

- Niveles de factor, interacción entre variables, se refiere a las interacciones de factores que interfieren en las variables de estudios. Por ejemplo estoy midiendo altura con respecto a un tratamiento de riego, pero tomo en cuenta como le afectan factores como el clima o la temperatura.
    - Dos vías
    - Tres vías
- ¿Qué son los factores y los niveles de factor?
    - Utilice factores durante un experimento para determinar su efecto sobre la variable de respuesta. Los factores solo pueden asumir un número limitado de valores posibles, conocidos como niveles de los factores. Los factores pueden ser una variable categórica o estar basados en una variable continua, pero solo use unos pocos valores controlados en el experimento.
- Ejemplo de niveles de factor
    - Por ejemplo, usted estudia los factores que podrían afectar la resistencia del plástico durante el proceso de manufactura. Decide incluir Aditivo y Temperatura en su experimento. El aditivo es una variable categórica. Solo puede ser tipo A o tipo B. En cambio, la temperatura es una variable continua, pero aquí es un factor porque en el experimento solo se prueban tres valores de configuración de temperatura de 100C, 150C y 200C.

✅ Ejercicio con datos para ANOVA

- Instalar paquetería: blob:https://lite.evernote.com/f5b9a6f8-27fd-49cc-9ddd-0ec0098407c7
- Correr la librería con source data y la ruta de la librería
- Visualizar en un Boxplot: Se observan los factores de tratamiento, niveles de tratamiento. En este caso niveles de sitio son 4: Chinatu, Laguna, Trinidad y Tule
- Prueba de normalidad: Se observa un p-value menor a 0.05; entonces no son normales.
- Se comprueba en un histograma: Hay un sesgo a la derecha, si fuesen normales se concentrarían en los diámetros 30 y 40.

✅ Ejercicio de normalidad

- A mayor cantidad de datos, más nos acercamos a la normalidad
- Se observa un p-value mayor a 0.05; entonces si hay homogeneidad de varianzas
- Ante datos no normales y homogeneidad de varianzas
    - Es común que en recursos naturales no haya suficientes observaciones, por lo que no hay normalidad de datos.
    - Una solución puede ser la transformación de los datos (se reducen):
        - Log (x + 1)
        - √x
        - √x – 0.5 ✅ Transformación de datos desde R
- Se crea una nueva columna, desde R
- Se le asigna la función de log10 a los datos
- También se pueden redondear los datos para una mejor representación de normalidad
- Se correo nuevamente Shapiro Test, y ya se observa una normalidad de los datos.

✅ Para limpiar el sesgo en los datos

- Se instalo librería e1071, desde repmis
- La función skewness, me indica si se redujo el sesgo
    - Arriba de 1, totalmente sesgado

    - Entre 0.5 a 1 moderamente sesgado
    - Menos de 0.5, tendientes a 0, no tienen sesgo
- Tomar en cuenta la media ideal del grupo (media)
    - Hacia la derecha
    - Hacia la izquierda

✅ Otra manera de transformación de datos

- Mediante la aplicación de raíz cuadrada a los datos
- Con ello ya se alcanza la normalidad de los datos
- También se comprobó homogeneidad de varianzas
- Mediante la aplicación de raíz cuadrada a los datos
- Con ello ya se alcanza la normalidad de los datos
- También se comprobó homogeneidad de varianzas []]

✅ Para quitar “limpiar tu base” quitando columnas

- Se utilizan los corchetes y mediante coma y un menos se indica que columna de mi base quitar [,-#]

✅ Análisis de varianza – Tabla de ANOVA

- Cuando el CM tratamiento es mayor que el CM error, hay diferencias, de lo contrario no hay diferencias.
- Si el error es mayor, probablemente tengo una fuente de variación que esta alterando mis datos.
- Correr objeto dap.anova
- Mas resumido es con la función summary, que te da el resumen de cualquier objeto.
- Se corrobora que hay diferencias significativas mediante:
    - Valor de P
    - La probabilidad de F
- Como ya se comprobó diferencias significativas, ahora se debe comprobar la comparación entre variables
- Se grafican los datos de la prueba de Tukey

✅ Para ir colocando textos en la grafica boxplot

- Función de text, se pone primero la posición de la variable y luego la altura a la que queremos que aparezca el texto de acuerdo a la altura de la gráfica.
- También se pueden poner puntos decimales. Por ejemplo text (1,7.7, “a”)
- Nota: esta función se va poniendo sobre capas.

✅ Para ir colocando textos en los ejes de las gráficas

- Función de mtext, se pone primero el texto deseado entre comillas, y después en que posición de la gráfica).
- Siempre se puede volver a resetear, utilizando nuevamente la función de boxplot
- Buscar librería para adjudicar letras de diferencias significativas. Mult, four letters

Clase 6: 03.03.2025 (Pruebas de t de 1 y 2 colas)

✅ Ejemplo real

- Se ingresan los datos
- Con la función length se comprueba el número de datos total
- El promedio se comprueba con mean
- Se grafica con función plot ((density)), para ver los datos en una curva de distribución.
- Aparentemente se distribuye normal, sin embargo la media baja del peso de 80 kg

✅ Se coloca propiedades a la gráfica

- Se agrega línea de media mediante función abline, que es una mascara directa en la gráfica. V es vertical, lwd es ancho y lty para modificar como es la línea, en este caso 2 es descontinua.
- Se agrega línea de media teorética mediante función abline
- De manera visual se nota que la media observada es menor a la media teorética, sugiriendo la hipótesis alternativa. Que los costales tienen contenido menos peso de lo que viene en los productos. Sugiere un defecto en los procesos de calidad de llenado de los costales.

✅ Ahora en un histograma

- En un histograma nos interesaría más una línea horizontal de referencia.

✅ Prueba de t de 1 cola

- Ahora se corre la prueba de t, con mu teorética de 80 y referencia de menos (“less“ condiciona prueba de 1 cola).
- Con el valor de p = 0.01132, valor menor a 0.05, por lo que se acepta la hipótesis alternativa, que la media observada es menor a la media de 80 kg. Por lo que estadísticamente si hay diferencias significativas entre el peso de los costales y la declaración del peso de los costales en su empaque.
- Si la corremos sin condicionar “less“, sería una prueba de dos colas, por lo que el valor de referencia de p se toma como 0.025.
- Continua habiendo diferencias significativas, ya que p-value es de 0.02264, sin embargo no tomo de referencia si el valor que busca es menos o más a la media teorética de referencia, sólo que hay diferencias.

✅ Prueba de t de 2 colas

- El grupo 2 duerme en promedio 20 minutos más, a diferencia del primer grupo.
- No hay diferencia significativa entre ambos grupos, más de 0.025

✅ Ejemplo de calidad del aire

- Si el mes de mayo, es más caliente o más frío que la temperatura promedio (una muestra, con respecto a una media teorética).
- La condiciono si es una temperatura menor, con la función alternative = “l”.
- Se confirmo que la temperatura de mayo es menor, ya que p-value es menor a 0.05
- No hay varianza entre mayo y el resto de los meses de comparación
- Ahora veré si esa variación esta relacionada a otro factor (viento, ozono, etc.).
- Lo puedo hacer mediante una regresión lineal.
- Ozono no se observa variables
- Ver si el viento es variable de acuerdo a la varianza
- Aparentemente hay una diferencia significativa entre los meses, pero el grado de libertad es sólo uno

- La variable mes no me la esta reconociendo como factor
- Corrección como factor, se nota en los grados de libertad
- Se confirma que hay diferencias significativas.
- Como son diferentes hay que hacer prueba de Tukey.
- Sólo hay diferencias entre mayo con julio y mayo con agosto.

✅ Correlación de Pearson

- Correlación positiva, a medida que aumenta x aumenta y.
- Correlación negativa, a medida que disminuye x disminuye y.
- O bien, no hay correlación entre las dos variables.
- Va de -1 hasta 1
- En este caso hay una correlación negativa y es significativo, sin embargo es baja la influencia.
- Osea que el Viento influye negativamente en la temperatura, pero no es tan marcada
- Ahora sacare una grafica de puntos. Pch se refiere al tipo de carácter de los puntos.
- Se observa una tendencia, pero no tan clara.
- Se determina la lineal de regresión con r
- R2 explica el porcentaje que tanto una variable esta correlacionada con la otra.
- Se determina la lineal de regresión con r (coeficiente de relación), en este caso r = 0.45
- R2 (coeficiente de determinación) explica el porcentaje que tanto una variable esta correlacionada con la otra; en este caso r2 = 0.2025; o sea un 20%. Se tendría que acerca a 1 o -1 para una mayor tendencia de correlación en los datos.

Clase 7: 10.03.2025 (Regresión lineal)

✅ Regresión lineal

- Regresión lineal (r2): conocer el comportamiento de las variables, como puede variar una variable con respecto a la otra. Se usa coeficiente de r2 ajustada (en que porcentaje el índice de la variable x predice la variable y).
- Función de regresión lineal. lm (y x), siempre se pone la variable y (dependiente) primero, ya que es la que se ve descrita por la variable x (independiente).
- Nube de puntos: correlaciona dos variables, desde el fundamento de que para cada valor de x (independiente) existe un valor en y (dependiente), o sea pares de observaciones.
- Línea de tendencia central: Mide el ajuste en la nube de puntos. Parte de la relación y´= alfa + beta (x). No sirve para predecir, sólo describir.
- Alfa y beta son dos condicionantes de similitud, son dos valores constantes para obtener la y´.
- Alfa: Valor que tiene y cuando x vale 0
- Beta: Pendiente que indica cual es el valor que incrementa y cuando x incrementa una unidad.

✅ Ejercicio de correlación

- Importar datos: La variable eruptions es la duración de chorro de agua en cada erupción, y la variable waiting es el tiempo de espera entre cada erupción.
- Se gráfican los datos en donde visualmente si se observa un correlación. Ahora se comprueba numéricamente.
- Cor solo valor de correlación
- Cor.test, para ver no solamente el valor de cor, sino para conocer si es significativo o no.

- Se observa que existe una fuerte correlación entre ambas variables y de manera positiva, pues el valor se aproxima bastante al valor de 1.

✅ Modelo de regresión lineal, para determinar la línea de tendencia central.

- Se guarda con el objeto g.lm
- Al correr la regresión de g.lm se observa que cuando el valor de x es 0, y intercepta en el valor de 33.47, sin embargo no se observa algún valor similar a esta situación
- Por ello, se comprueba por medio de una matriz con la función de summary, en donde se observa la significancia de los valores de alfa (intercept, en este caso 33.47) y beta (geysereruptions, en este caso).
- Los datos deben pasar por el centro, se debe tener en cuenta cuando hay huecos entre datos.
- Las líneas de tendencia central tienen una sumatoria de las diferencias, las cuales deben acercarse a 0. Cuando no sucede esto no esta en el ajuste correcto.
- Para asegurarse que esa línea pase por el centro, usar la función abline
- Se corrige la relación de la variable independiente y dependiente
- Ahora si se observa coherencia en el valor de alfa (-1.87), acorde a los datos
- La r2 esta indicando que el ajuste es del 81.08%, con esa probabilidad y esta descrito por x.

✅ Análisis de varianza de la regresión, mediante un modelo de anova

- Análisis de varianza de la regresión, mediante un modelo de anova (no para buscar diferencias en comparación de medias), para que describa la varianza. Para conocer como es la variación, rectificar que el modelo que estoy utilizando funciona. Te resume los valores que fuimos sacando por independientes. Nos interesa observar la varianza (0.247) y la significancia.
- Aquí se observan los grados de libertad en 270, la sumatoria de la diferencias al cuadrado es 66.56