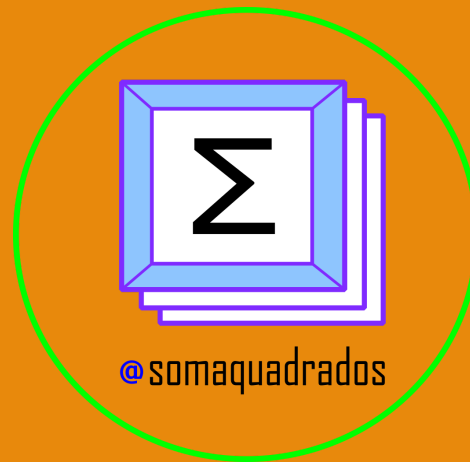


Introducción al análisis de datos biológicos con R

Clase 3

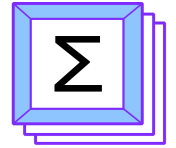


Eliana F. Burgos

Clase 3

Contenido

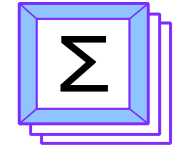
1. Variables Estadísticas
2. Distribución de probabilidad
3. Estadística descriptiva: medidas de posición
4. Estadística descriptiva: medidas de dispersión



@somaquadrados



IMPORTANTE: LEER!!

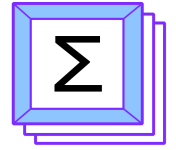


@somaquadrados

Consejos útiles

1. Recuerde leer bien las consignas y escribir todo lo que necesite en su script.
2. NO hay calificación en estos ejercicios, son solo para aplicar lo que vimos en la clase teórica y que podamos cerrar la idea.
3. Copie también en su script los errores que se vaya encontrando, para poder debatir su solución y en caso de que haya podido resolverlos, anote la solución y si usted llegó por análisis propio o mediante búsqueda en la web.
4. Trabaje con el tipo de dato correctamente! Setee las variables numéricas como numéricas, factores como factores, etc...
5. Al finalizar todos los ejercicios, complete el formulario que se encuentra al final de esta guía.

Además **PACIENCIA Y COMIDITA RICA** a la hora de hacer los ejercicios
Cualquier duda nos escriben al mail o al grupo de telegram.



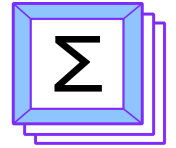
@somaquadrados

efburgos@conicet.gov.ar ; mariliabioufpr@gmail.com

Vamos a programar



Ejercicio 1

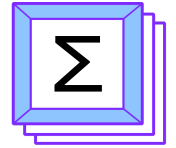


@somaquadrados

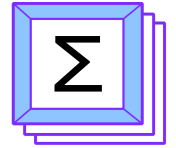
Estudiamos la abundancia de escarabajos en cuatro sitios (a,b,c,d) con diferentes grados de conservación. Medimos la humedad del suelo y el porcentaje de cobertura herbácea a un metro del suelo.

OBJETIVO: Queremos evaluar si la humedad y la cobertura vegetal tienen alguna influencia sobre la abundancia de los escarabajos.





1. Identifique la **variable respuesta**: ¿Qué tipo de variable es? ¿Qué valores **mínimos** y **máximos** toma?
2. Identifique las **variables explicativas**: ¿Qué tipo de variable son? ¿Qué valores mínimos y máximos toman?
3. Calcule la **media** para la variable respuesta y las variables explicativas para todo el estudio
4. Calcule la **moda** para la variable respuesta y las explicativas para todo el estudio.
5. Calcule los **cuartiles** para la variable respuesta y explicativas para todo el estudio.
6. ¿Hay diferencias entre los sitios? Haga un **boxplot** donde se visualicen las medidas de posición para la abundancia en cada sitio
7. ¿Cuál es la **variabilidad** de la abundancia en cada sitio? ¿Es **simétrica** mi variable respuesta?

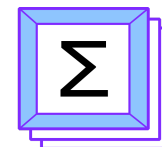


@somaquadrados

Ejercicio 1

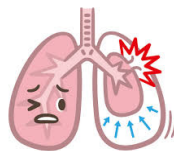
8. Piense en el tipo de datos que componen la variable respuesta. ¿Qué **distribución de probabilidad** cree que presentan?
9. Modele las distribuciones de probabilidad posibles para su variable respuesta y compare el resultado con su identificación previa.
10. Grafique la abundancia de los escarabajos en función de la **humedad**.
11. Grafique la abundancia de los escarabajos en función de la **cobertura vegetal**.
12. La humedad del suelo depende en cierta manera de la cobertura vegetal presente en un determinado sitio. Entonces, ¿Usted cree que son **variables independientes**? ¿Cómo corroboraría esto?
13. Haga una tabla con los estadísticos descriptivos calculados hasta el momento. Puede hacerlo en Excel o con la función **flextable** del paquete “*flextable*”(Gohel, 2021).

Recuerde completar el formulario copiando el script correspondiente a cada pregunta



@somaquadrados

Ejercicio 2



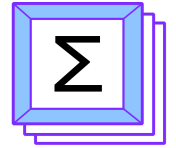
Un amigo médico nos escribe un mail pidiéndonos ayuda con los análisis de datos de pacientes jóvenes que sufren fibrosis quística y nos envía la base de datos. Él quiere saber si la capacidad pulmonar en estos jóvenes varía con la edad o el sexo. Entonces, nuestra variable respuesta será “capacidad pulmonar” y las variables explicatorias “edad” y “sexo”. Note que la edad esta expresada en rangos y que el sexo esta expresado en 0,1, 2 siendo 0=hombre, 1= mujer, 2=intersex.

OBJETIVO: Evaluar si existen cambios en la capacidad de pulmonar del paciente con fibrosis quística de acuerdo a su edad y sexo.

Fibrosis Pulmonar Idiopática →	← Fibrosis Quística Pulmonar
Tos, falta de aliento, restricción respiratoria progresiva, cansancio, deformación de los dedos de las manos (acropaquias), pérdida de apetito y de peso.	Tos persistente con mucosidad, dificultad respiratoria, infecciones pulmonares recurrentes, y síntomas digestivos, como heces malolientes y grasas, dificultad para aumentar de peso, obstrucción intestinal, estreñimiento o pancreatitis en casos atípicos.

SÍNTOMAS

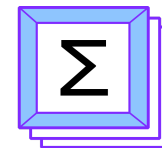
Resuelva



@somaquadrados

1. Identifique qué **tipo de variable** es “capacidad pulmonar” y evalúe la **distribución de probabilidad** que presenta mediante un histograma.
2. ¿Qué medida estadística puede identificar en el histograma? .
3. Calcule la **media, mediana y moda** de la capacidad pulmonar para cada **franja etaria**.
4. Calcule la **media, mediana y moda** de la capacidad pulmonar para cada **sexo**.
5. ¿Qué tan variable es la “capacidad pulmonar” en nuestra población muestra? Mida el **sesgo, *curtosis** y el **coeficiente de variación**.
6. Grafique la capacidad pulmonar en función de la edad y en función del sexo. Puede elegir el gráfico que le resulte más visual o explicativo
7. Concluya brevemente en función de los resultados.

Recuerde completar el formulario copiando el script correspondiente a cada pregunta



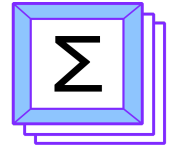
@somaquadrados

Ejercicio 3



En el marco de la evaluación del estado de conservación de mamíferos de la provincia, nos designaron la especie *Chironectes minimus* mejor conocida como **cuica de agua**, una zarigüeya propia de la Selva Paranaense y que, hasta el momento, en nuestro país solo se encuentra en Misiones. Sabemos que en otras regiones su peso es menor en áreas degradadas respecto de las áreas conservadas. Realizamos la captura de las zarigüeyas (con ayuda de un veterinario, las medidas de bioseguridad adecuadas y los permisos correspondientes), donde registramos el peso y el sexo de los individuos en dos sitios completamente distintos: un área natural protegida y un cultivo de yerba con arroyos y tajamares.



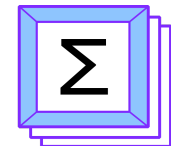


@somaquadrados

OBJETIVO: Evaluar de manera preliminar si las cuicas de agua sufren cambios en sus tamaños en función del ambiente en el que se encuentran.

Resuelva

1. Si la variable respuesta es el peso de las zarigüeyas ¿Qué **tipo de variable** es y cual sería la **distribución de probabilidades**?
2. Grafica un **histograma** del peso de las zarigüeyas ¿Qué se observa?
3. Identifica la **media** y la **moda** del peso.
4. Grafica el **peso** de las zarigüeyas en función del **uso del suelo**.
5. ¿Qué puedes concluir sobre el grafico anterior?
6. Dado que solo muestreamos una sola vez por estación y en un solo año, ¿Crees que estos valores son concluyentes? Calcula el **coeficiente de variación** del peso para uso del suelo y concluye sobre esto.

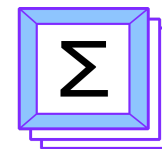


@somaquadrados

7. Suponiendo que nuestros datos son robustos y que pueden servir para hacer inferencias, ¿Cuál crees que es la implicancia de estos resultados para la conservación de esta especie?

8. ¿Qué otro **análisis descriptivo** podrías realizar? Pueden ser gráficos o el cálculo de otras medidas de posición/dispersión, o ambos.

Recuerde completar el formulario copiando el script correspondiente a cada pregunta



@somaquadrados

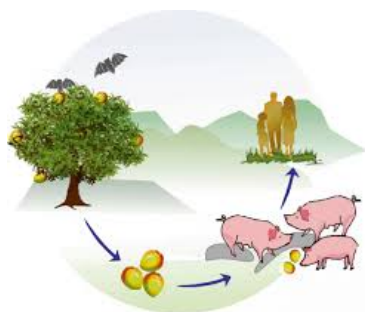
Ejercicio 4

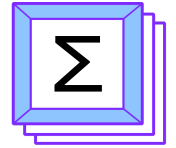


Nos salió una beca para hacer una pasantía en Malasia estudiando reservorios del virus Nipah, un patógeno transmitido por murciélagos frugívoros que viven cercanos a zonas de cria de cerdos para consumo humano. Queremos identificar las especies que intervienen en el ciclo silvestre de este virus y conocer como es la seroprevalencia de dicho patógeno en dos áreas naturales.

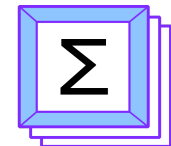
Fuimos a campo, hicimos la captura de murciélagos, buscamos anticuerpos, y armamos la tabla. Ahora nos toca comenzar a analizar estos datos.

Objetivo: Evaluar la seroprevalencia del virus Nipah en poblaciones de murciélagos en función del sitio, estación del año y de la riqueza y diversidad del ensamble.





1. Si la **variable respuesta** es la seroprevalencia del virus en el área natural protegida, ¿qué tipo de variable es? ¿Cómo mediría usted la simetría de esta variable? ¿A qué distribución corresponde?
2. Charlando con un colega le pregunta si la seroprevalencia del virus está relacionada con la riqueza y diversidad del ensamble, ¿Cómo podría usted, de manera exploratoria, **visualizar** si existe esta relación?
3. Dado que la diversidad está relacionada a la riqueza, ¿qué análisis podría hacer para evaluar estadísticamente la **relación** entre estas variables?



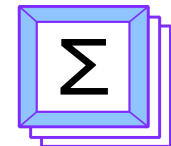
@somaquadrados

4.¿Cuál es la seroprevalencia **media** del virus en murciélagos? ¿Qué tan **variable** es esa seroprevalencia?

5.¿Qué sucede con la seroprevalencia **media** del virus entre estaciones del año? ¿Y entre áreas naturales protegidas?

6.En base a todos estos análisis **descriptivos**, ¿qué podría usted concluir?

Recuerde completar el formulario copiando el script correspondiente a cada pregunta



@somaquadrados



COMPLETE EL FORMULARIO -> [<https://forms.gle/tmLsFUB5VfVqEcTKA>]

FIN