Resumen prueba 1 Correlacional

- Hacer resumen de la materia pasada en clases (en base a mis apuntes y lo de la pau)
- Revisar qué textos tienen relación con lo visto en clases (Moore 1-54; 97-131; Field 206-244; Field 205-244 correlation
- Relacionar cosas de los textos con la materia
- Ver qué onda las fórmulas
- Pasarlo a papel

Conceptos principales:
Inferencia:
Asociación: entre variables, datos y significaciones estadísticas.
Reporte y reproducibilidad:

Sesión 1

- **Imaginación sociológica:** Relación del individuo con la sociedad y la historia, depende del contexto.
- **Imaginación estadística:** Apreciación de qué tan usual o inusual es un evento en relación a un conjunto de eventos similares mayores.
- **Estadística descriptiva:** Número de observaciones registradas y frecuencia de esas observaciones.
- **Estadística inferencial:** Contraste de hipótesis y teorías científicas en base a datos de investigación. Concentración teórica.

Sesión 2: Datos y variables
v
Datos entendidos como la expresión numérica de una característica (como esperanza de vida) de una unidad (años) en un punto del tiempo.

Medir vincula conceptos abstractos con valores empíricos, sus requisitos básicos son:

- **Exhaustividad:** mayor número de categorías significativas;
- **Exclusividad:** atributos mutuamente excluyentes.

Base de datos: fila = unidad; columna = variable; cada variable posee valores numéricos

- **Variables:** Una variable representa cualquier cosa o propiedad que varía y a la cual se le asigna un valor, pueden ser visibles o no visibles/latentes (ej: peso/inteligencia). Tipos de variables: **Discretas** (dicotómicas politómicas) y **Continuas** (Rango, [teóricamente] infinitos valores.

Escalas de medición de variables: NOIR- Nominal (uso de números en lugar de palabras / identidad / ejemplo: nacionalidad), Ordinal (números que se usan para ordenar series / ranking / nivel educacional), Intervalar (Intervalos iguales entre números / igualdad / Temperatura), Razón (Cer real / aditividad / distancia).

Tipos de datos en relación a escalas de medición.

- **Datos categóricos**:
  - pueden ser medidos sólo mediante escalas nominales, u ordinales en caso de orden de rango
- **Datos continuos**:
  - Medidos en escalas intervalares o de razón
  - Pueden ser transformados a datos categóricos

Tipos de análisis:

- **Variable dependiente (y):** lo que quiero explicar;
- **Variable independiente (x):** lo que me permite explicar la dependiente.

Tendencia central y dispersión

Tendencia central:

- **Moda:** valor que ocurre más frecuentemente;
- **Mediana:** valor medio de distribución ordenada. Si N es par , entonces es el promedio de los valores medios;
- **Media o promedio aritmético:** suma de los valores divididos por el total de casos.

Dispersión:

- **Varianza:** equivale al promedio de la suma de las diferencias del promedio al cuadrado, muestra qué tanto varían numéricamente los resultados entre los elementos de estudio.

Dispersión:

$$varianza = \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N - 1}$$

xi - x con la wea arriba es la distancia del elemento con el promedio, buscar google.

- **Desviación estándar:** es la raíz cuadrada de la varianza, mide la distribución de dispersión de datos.

Midiendo asociación, pendiente a apuntes.

Asociación: Cómo se relacionan las variables, la correlación de Pearson es una forma de hacerlo.

Sesión 3: Correlación de Pearson

Qué es la correlación

La correlación es una medida estadística que describe la asociación entre dos variables. Cuando existe correlación entre dos variables, el cambio en una de ellas tiende a estar asociado con un cambio en la otra variable. Lo que observamos es cómo se comportan los valores de dos (o más) variables para cada observación, y si podemos suponer que ese comportamiento conjunto tiene algún patrón.

Correlación producto-momento (Pearson)

$$Covarianza = cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Correlación = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y}$$

$$= \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Correlación de Pearson

Es una medida estadística que cuantifica la relación lineal entre dos variables continuas, va desde -1 a 1:

- r = 1: Correlación positiva perfecta. Una variable aumenta y la otra también en proporción constante.
- r = -1: Correlación negativa perfecta. Una variable aumenta la otra disminuye en proporción constante.
- r = 0: No hay correlación lineal entre las variables. No hay relación entre los cambios de las variables.

La varianza genera relaciones entre variables= **covarianza:** esta es una medida de asociación entre variables basada en la variabilidad de cada una de ellas, ¿la distancia que toma una variable del promedio tendrá relación con la distancia que tome otra variable del promedio? Expresa la distancia al promedio de las variables.

Covarianza

Varianza educación (x)

$$\sigma^2_{edu} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

$$\sigma^2_{edu} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

Varianza ingreso (y)

$$\sigma^2_{ing} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$$

$$\sigma^2_{ing} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})}{N-1}$$

$$Covarianza = cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

43

Lo de educación e ingreso es un ejemplo de variables asociadas.

$$Covarianza = cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Correlación = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y}$$

$$= \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Vamos por paso:

- $(x_i - \bar{x})$ es la diferencia de un valor de una variable respecto de su promedio. Por ejemplo, si x=educación, y un caso tiene educación 6 y el promedio de educación es 2, entonces el valor de $(x_i - \bar{x}) = 4$
- $(y_i - \bar{y})$: lo mismo pero para la otra varible, en el caso de nuestro ejemplo es ingreso.
- $(x - \bar{x})(y - \bar{y})$ es la multiplicación de los dos pasos anteriores para cada caso.
- $\sum(x - \bar{x})(y - \bar{y})$ es la suma de estos valores para el total de los casos

Limitaciones de Pearson:

- Medida de asociación lineal entre variables.
- No captura bien asociaciones no lineales.

- Posee supuestos distribucionales de x e y. las hipótesis o condiciones que se asumen acerca de la distribución de una variable aleatoria en un análisis estadístico o en la aplicación de ciertos métodos estadísticos
- Sensible a valores externos.
- Un mismo coeficiente puede reflejar distintas distribuciones bivariadas.

Sesión 4: Tamaños de efecto, otros coeficientes y matrices

Correlación sociológicamente: en los rangos de cohen son para las ciencias sociales,al hablar de dos variables se pueden relacionar dentro de sus tamaños del efecto, con la interpretación y los valores estadísticos. Cohen pensado para las ciencias sociales. No conviene Cohen como el estándar

Magnitud de correlación:
Si dos variables covarían entonces comparten varianza
¿Cómo se establece cuánta varianza comparten? La fórmula.

Varianza compartida y varianza no compartida
La varianza compartida puede pensarse como 1. Varianza no compartida se asocia a residuos, es decir, la cantidad de varianza que no está contenida en la correlación. Para poder obtener los residuos vamos a generar una recta que represente la asociación entre las variables. Esta es la recta de **regresión**. Esta recta nos permite un valor estimado de *y* para cada valor de *x*.

La varianza compartida (representa qué tan significativa es la relación entre variables) es la parte de la variabilidad en la variable dependiente que se puede atribuir a las variables independientes o covariables en un análisis estadístico, mientras que la varianza no compartida es la parte de la variabilidad que no se puede explicar y que se considera como ruido o error residual. En términos generales, en un análisis estadístico, se busca que la varianza compartida sea grande en comparación con la varianza no compartida, ya que esto sugiere que las variables independientes o covariables están teniendo un efecto significativo en la variable dependiente.

Coeficiente de determinación / R Cuadrado: Varianza compartida entre variables; Proporción de la varianza y que se asocia a x; se presenta entre 0 y 1 y se expresa en porcentaje.

Criterios de Cohen para tamaños de efecto
- Coeficiente de Pearson nos indica la dirección, y fuerza/intensidad de la relación.
- ¿Qué nos dice el tamaño del coeficiente? 0,5 es mediano?
- Cohen (88-92) sugiere criterios convencionales para clasificar efectos pequeños medianos o grandes-
- 0,10 efecto pequeño; 0,30 efecto mediano; 0,5 y más efecto grande.

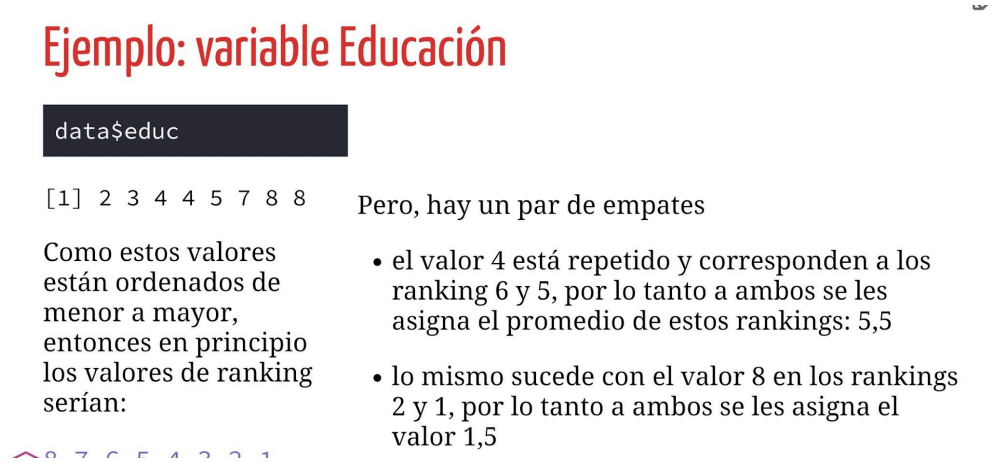
Tamaño de efecto: Valores convencionales para establecer si una magnitud es baja, mediana o alta

Otros coeficientes de correlación

Coeficiente de correlación de Spearman

- Se utiliza para variables ordinales y/o cuando se violan supuestos de distribución normal.
- Equivalente a la correlación de Pearson del ranking de las observaciones analizadas
- Es alta cuando las observaciones tienen un ranking similar

Cálculo de Spearman: Se le asigna un número de ranking a cada valor, el valor más bajo obtiene el mayor ranking y el más alto obtiene un ranking más bajo (el valor determina la lejanía o cercanía entre variables); los valores repetidos significan un empate y el ranking se promedia.



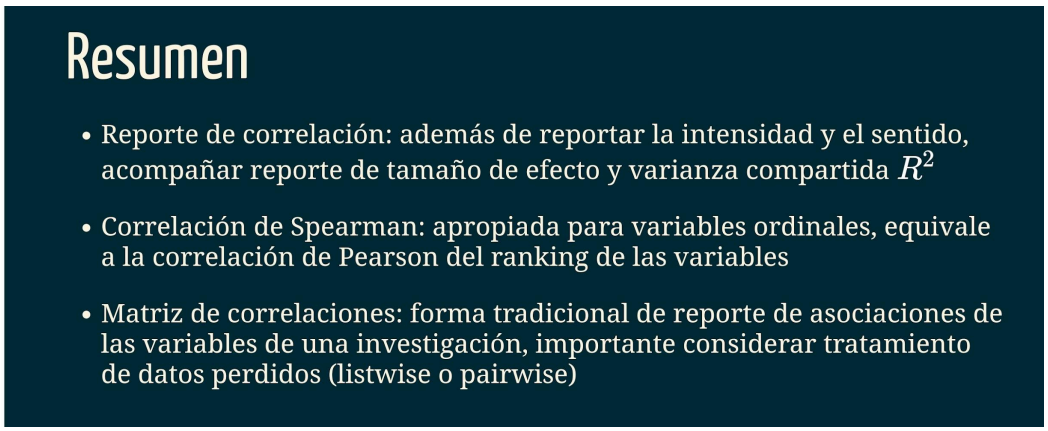
Coeficiente de correlación de Tau de Kendall (menos usado)
- Recomendado cuando hay un set de datos pequeños y/o hay mucha repetición de observaciones en el mismo ranking.
- Se basa en una comparación de pares de observaciones concordantes y discordantes.

Recomendaciones:
- Pearson es el coeficiente de correlación por defecto.
- En caso de datos de escala de medición ordinal se puede aplicar Spearman (aunque Pearson es también aceptado por este contexto).
- Kendall se usa en casos muy específicos.

Matrices de correlación

Matriz de correlación
- Se conforma cuando se representan simultáneamente más de un par de asociaciones bivariadas.
- Tabla de doble entrada donde las variables se representan en columnas.

Consideraciones sobre casos pérdidos
- Cuántos casos perdidos en las variables ¿cuál es el número de casos de matriz de correlaciones?
- La correlación bivariadas se calcula con información completa por lo tanto si hay un dato perdido en una de las variables se elimina el dato completo.
- Algunas variables lo hacen automáticamente, otras hay que especificarlas.



Respecto al proceso de investigación,

Reproducibilidad: Detallar el proceso de investigación para que cualquiera pueda seguirlo y obtener los mismos resultados, genera confianza en el trabajo y permite construir sobre los hallazgos que se obtuvieron y realizar un conocimiento colectivo. Garantiza resultados confiables y evita casualidades.

El coeficiente de determinación RCuadrado es una medida estadística que indica la proporción de la varianza total de una variable que es explicada por otra(s) variable(s). En pocas palabras,

Se utiliza para evaluar cuánta de la variabilidad de una variable se debe a otra variable.
sus valores van desde 0 a 1, en donde 0 indica que ambas variables comparten el 0% de su varianza, y 1 que comparten el 100% de su varianza. En el contexto de la correlación entre solo **dos variables**, el RCuadrado es igual a elevar al cuadrado el coeficiente de correlación = (r)^2. Esto nos permite conocer qué tanto la variabilidad de una variable X estaría asociado a la variabilidad de otra variable Y.

Textos

Texto I: La imaginación estadística (Ritchey)

El campo de la estadística: Conjunto de procedimientos para reunir, medir, clasificar, codificar, computar, analizar y resumir información numérica adquirida sistemáticamente (p. 2).

La imaginación estadística: Una apreciación de que tan usual o inusual es un evento, circunstancia o comportamiento, en relación con un conjunto mayor de eventos similares y una apreciación de las causas y consecuencias del mismo (p. 3).

En este sentido, **cualquier estadística está sujeta a la cultura**, es decir, es normativa: su interpretación depende del lugar, tiempo y cultura donde se observa (p. 4).

Una **norma estadística** es una tasa promedio de ocurrencia de un fenómeno (p. 5) Un **ideal estadístico** es una tasa de ocurrencia socialmente deseada de un fenómeno (p. 5-6).

La estadística trata de observar y organizar información numérica sistemáticamente adquirida. La información sistemáticamente adquirida que se organiza siguiendo los procedimientos de la ciencia y la estadística se llama dato o datos (p. 7).

Error estadístico: Grado conocido de imprecisión en los procedimientos utilizados para reunir y procesar información (p. 7).

La **estadística descriptiva** informa cuántas observaciones fueron registradas y qué tan frecuentemente ocurrió en los datos cada puntuación o categoría de observaciones (p. 8). La **estadística inferencial** se calcula para mostrar relaciones de causa-efecto, así como para probar hipótesis y teorías científicas (p. 8).

Variable: Fenómeno medible que varía (cambia) a través del tiempo, o difiere de un lugar a otro o de un individuo a otro (p. 10). **Variación** para referirse a cuánto difieren las mediciones de una variable entre los sujetos en estudio (p. 10).

Constante (cuando no varía y se mantiene 'constante'). **Variable dependiente**, la variable cuya variación queremos explicar (p. 11). **Variables independientes,** aquellas variables de predicción, que están relacionadas o predicen la variación en la variable dependiente (p. 11).

Hipótesis: Predicción sobre la relación entre dos variables; en ella se afirma que, los cambios en la medida de una variable independiente corresponderán a cambios en la medida de una variable dependiente (p. 11).

Los **siete pasos** del proceso de investigación son: especificar la pregunta de investigación, revisar la literatura científica, proponer una teoría y formular las hipótesis, seleccionar un diseño de investigación, recolectar datos, analizar los datos y sacar conclusiones, así como diseminar los resultados (p. 21).

Texto II: Análisis de relaciones (Moore)

Una **variable respuesta** mide el resultado de un estudio. Una **variable explicativa** influye o explica cambios en la variable respuesta (p. 98). A menudo, encontrarás que a las variables explicativas se les llama variables independientes y a las variables respuesta, variables dependientes (p. 98). **Ej:** El alcohol produce muchos efectos sobre el cuerpo humano. Uno de ellos es la bajada de la temperatura corporal. Para estudiar este efecto, unos investigadores suministraron distintas dosis de alcohol a unos ratones y al cabo de 15 minutos midieron la variación de temperatura de su cuerpo. La cantidad de alcohol es la variable explicativa y el cambio de temperatura corporal es la variable respuesta (p. 99).

Un **diagrama de dispersión** muestra la relación entre dos variables cuantitativas medidas en los mismos individuos. Los valores de una variable aparecen en el eje de las abscisas y los de la otra en el eje de las ordenadas. Cada individuo aparece como un punto del diagrama. Su posición depende de los valores que toman las dos variables en cada individuo (p. 104). Sitúa siempre a la variable explicativa, si una de ellas lo es, en el eje de las abscisas del diagrama de dispersión. En general, llamamos a la **variable explicativa x y a la variable respuesta y (p. 104).**

Examen de un **diagrama de dispersión**: En cualquier gráfico de datos, identifica el aspecto general y las desviaciones sorprendentes del mismo. Puedes describir el aspecto general de un diagrama de dispersión mediante la forma, la dirección y la fuerza de la relación. Un tipo importante de desviación son las observaciones atípicas, valores individuales que quedan fuera del aspecto general de la relación (p. 105).

Asociación positiva y asociación negativa: Dos variables están asociadas positivamente cuando valores superiores a la media de una de ellas tienden a ir acompañados de valores también situados por encima de la media de la otra variable, y cuando valores inferiores a la media también tienden a ocurrir conjuntamente. Dos variables están asociadas negativamente cuando valores superiores a la media de una de ellas tienden a ir acompañados de valores inferiores a la media de la otra variable, y viceversa (p. 106).

Si creemos que los cambios de una variable x explican o que incluso son la causa de los cambios de una segunda variable y, a la variable x la llamaremos variable explicativa y a la variable y variable respuesta (p. 112).

(**Correlación**) Un diagrama de dispersión muestra la forma, la dirección y la fuerza de la relación entre dos variables cuantitativas. Las relaciones lineales son especialmente importantes, ya que una recta es una figura sencilla bastante común. Decimos que una relación lineal es fuerte si los puntos del diagrama de dispersión se sitúan cerca de la recta (p. 120). La correlación mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas (p. 122).

La correlación se simboliza con la letra r. Supón que tenemos datos de dos variables x e y para n individuos. Los valores para el primer individuo son x1 e y1, para el segundo son x2 e y2, etc. Las medias y las desviaciones típicas de las dos variables son $\bar{x}$ y sx para los valores de x, e $\bar{y}$ y sy para los valores de y.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

La correlación r entre x e y es: (Características de la correlación) La fórmula de la correlación ayuda a ver que r es positivo cuando existe una asociación positiva entre las variables. De la misma manera, podemos ver que r es negativa cuando la asociación entre x e y es negativa. (p. 123-124).

1. La correlación no hace ninguna distinción entre variables explicativas y variables respuesta. Da lo mismo llamar x o y a una variable o a otra.
2. La correlación exige que las dos variables sean cuantitativas para que tenga sentido hacer los cálculos de la fórmula de r. No podemos calcular la correlación entre los ingresos de un grupo de personas y la ciudad en la que viven, ya que la ciudad es una variable categórica.
3. Como r utiliza los valores estandarizados de las observaciones, no varía cuando cambiamos las unidades de medida de x, de y o de ambas. Si en vez de medir la altura en centímetros lo hubiéramos hecho en pulgadas, o si en lugar de medir el peso en kilogramos lo hubiéramos hecho en libras, el valor de r sería el mismo. La correlación no tiene unidad de medida. Es sólo un número.

4. Una r positiva indica una asociación positiva entre las variables. Una r negativa indica una asociación negativa.
5. La correlación r siempre toma valores entre −1 y 1. Valores de r cercanos a 0 indican una relación lineal muy débil. La fuerza de la relación lineal aumenta a medida que r se aleja de 0 y se acerca a 1 o a −1. Los valores de r cercanos a −1 o a 1 indican que los puntos se hallan cercanos a una recta. Los valores extremos r = −1 o r = 1 sólo se dan cuando existe una relación lineal perfecta y los puntos del diagrama de dispersión están exactamente sobre una recta.
6. La correlación sólo mide la fuerza de una relación lineal entre dos variables. La correlación no describe las relaciones curvilíneas entre variables aunque sean muy fuertes.

(Apuntes prácticos) La correlación es una medida estadística que describe la asociación entre dos variables. Cuando existe correlación entre dos variables, el cambio en una de ellas tiende a estar asociado con un cambio en la otra variable. En términos concretos, lo que observamos es cómo se comportan los valores de dos (o más) variables para cada observación, y si podemos suponer que ese comportamiento conjunto tiene algún patrón. La correlación describe el sentido (dirección) y fuerza de la asociación. En otras palabras, nos permite conocer cómo y cuánto se relaciona la variación de una variable, con la variación de otra variable.

Interpretación
Recordemos nuestra matriz del comienzo:
Tenemos que la correlación entre la variable de estatus social subjetivo y años de educación es 0.3. ¿Cómo interpreto esto?

Una manera recomendable es la siguiente:
El coeficiente de correlación de Pearson entre estatus social subjetivo y años de educación es positivo y moderado (r = 0.3) según Cohen (1988).

Texto III: Correlation (Field)

To understand what covariance is, we first need to think back to the concept of **variance**. Remember that the variance of a single variable represents the average amount that the data vary from the mean (p. 206).

Calculating the covariance is a good way to assess whether two variables are related to each other. A **positive covariance** indicates that as one variable deviates from the mean, the other variable deviates in the same direction. On the other hand, **a negative covariance** indicates that as one variable deviates from the mean (e.g., increases), the other deviates from the mean in the opposite direction (e.g., decreases) (p. 208).

There are t**wo types of correlation: bivariate and partial.** A **bivariate** correlation is a correlation between two variables (as described at the beginning of this chapter) whereas a partial correlation (see section 6.6) looks at the relationship between two variables while 'controlling' the effect of one or more additional variables. Pearson's product-moment correlation coefficient (described earlier), Spearman's rho (see section 6.5.5) and Kendall's tau (see section 6.5.6) are examples of bivariate correlation coefficients (p. 213-214).

A **partial** correlatíon quantifies the relationship between two variables while controlling for the effects of a third variable on both variables in the original correlation.

A **semi-partíal** correlatíon quantifies the relationship between two variables while controlling for the effects of a third variable on only one of the variables in the original correlation (p. 238).

Ejercicio dispersión varianza:

Promedio/media aritmética: 5,12

2-P= -3,12
3-P= -2-12
(4-P)x2= -1,12
5-P= -0,12
7-P= 1,88
(8-P)x2= 2,88

Se elevan al cuadrado:
9,7344; 4,4944; 1,2544; 1,2544; 0,0144; 3,5344; 8,2944; 8,2944 = 36,8752 / 7 = **5,267**