Examen de estadística básica

1. **Estadística Descriptiva**: el objetivo es el dado los datos, ordenarlos, simplificarlos, resumirlos, clasificarlos, etcétera.
    - introducción a la estadística
        - población e individuo
            - el conjunto de todos los individuos recibe el nombre población.
            - Población: se debe tomar como un colectivo del que queremos sacar conclusiones
            - Individuo es la unidad o elemento que compone la población
        - muestras aleatorias
        - variable aleatoria y modelo probabilístico
            - la situación que se representa es la de una característica o valor poblacional, objeto de investigación = parámetro
        - diferentes estadísticas
    - **conceptos fundamentales de la estadística descriptiva**
        - caracteres
            - **cuantitativos**
                - cuando son tales que su observación en un individuo determinado proporciona un valor numérico asociada.
                - Edad, curso, altura…
            - **Cualitativos**
                - Cuando su observación en los individuos no suministra un numero, sino la pertenencia a una clase determinada.
                - Sexo, facultad matriculada….
        - **Modalidad de los caracteres**
            - las posibilidades, tipos o clases que pueden presentar los caracteres las denominaremos modalidades
            - sexo= femenino, masculino
        - **La matriz de datos**
            - la primera columna los individuos identificados y en la siguiente columna las diferentes características de estudio.
        - **Clases de datos**
            - I. Datos correspondientes a un carácter cualitativo
            - II. Datos correspondientes a un carácter cuantitativo
            - III. Datos agrupados en intervalos correspondientes a un carácter cuantitativo
        - **agrupamientos en intervalos**
            - extremos: la clase j-esima a c,-1 y a cj
            - amplitud: intervalo a la diferencia de sus extremos
            - formula de Sturges para calcular el número de intervalos.
                - K = 1 + 3,322 log_{10^n}
                - siendo n el número total de datos

    - **distribuciones unidimensionales de frecuencias**
        - tipos de frecuencia
            - frecuencia total
                - número total de datos
            - frecuencia absoluta n_i
            - frecuencia relativa M_i
            - frecuencia absoluta acumulada N_i
            - frecuencia relativa acumulada F_i
        - **representaciones graficas de las distribuciones unidimensionales de frecuencia**

        - **carácter cualitativo**
            - diagrama de sectores
            - diagrama de rectángulos
        - **carácter cuantitativo sin agrupar**
            - diagrama de barras
            - histograma
            - diagrama de hojas y ramas
            - diagrama de frecuencias acumuladas
            - función de distribución Empirica
        - **carácter cuantitativo agrupado en intervalos**
            - histograma
    - **medidas de tendencia central de caracteres cuantitativos**
        - media aritmética
        - mediana
            - datos agrupados
            - datos sin agrupar
        - moda
            - datos agrupados
            - datos sin agrupar
        - cuantiles
            - datos sin agrupar
            - datos agrupados
    - **medidas de dispersión**
        - recorrido
        - varianza
        - desviación típica
        - coeficiente de variación de pearson
    - **medidas de asimetría**
        - coeficiente de asimetría de Pearson
        - coeficiente de asimetría de fisher
    - medidas de posición y dispersión con R
- **distribuciones bidimensionales de frecuencias**
    - **tabla de doble entrada o tabla de contingencia**
        - distribuciones marginales
            - son distribuciones de frecuencia unidimensionales
        - distribuciones condicionadas
    - **representaciones graficas de las distribuciones bidimensionales de frecuencia**
        - datos agrupados en intervalos correspondiente a un carácter cuantitativo
            - histograma tridimensional
        - datos sin agrupar correspondientes a un carácter cuantitativo
            - diagrama de barras tridimensional
            - nube de puntos
    - **ajustes por mínimos cuadrados**
        - cuándos los datos no actúan de una forma independiente sino que es condicionado unos con otros, por ejemplo la altura de una persona y su peso, a mayor altura probablemente tendrá mayor peso.
    - precisión de ajustes por mínimos cuadrados

2. Probabilidad

- Introduccion
    - **Espacios probabilísticos**
        - Espacio muestral= Ω
        - probabilidad= P

        - espacio de sucesos= A
        - variable aleatoria= X
    - **Espacio muestral**
        - Ω= se les denomina sucesos elementales
            - Ejemplo: las posibles caras de un dado Ω= {1,2,3,4,5,6}
    - **Conceptos de probabilidad**
        - **Concepto frecuentista**
        - **Concepto clásico**
            - Principio de la razón insuficente
        - **Concepto subjetivo**
        - **Definición formal de probabilidad**
            - Axioma 1: para todo suceso A de A sea P(A)≥0
            - Axioma 2: sea P(Ω)= 1
            - Axioma 3: Paara toda colección de sucesos icompatibles, $\{A_i\}$ con $A_i \cap A_j = \emptyset$, i≠ j
    - **Propiedades elementales de la probabilidad**
        - P (∅)= 0
        - Se cumple la aditiva finita para sucesos incompatibles
        - La probabilidad del complementario de un A es P(A*)= 1 – P(A)
        - Si dos sucesos son tales que A ⊂ B, entonces es P(A) $\leq P(B)$
        - Si dos sucesos no son incompatibles, la probabilidad de su unión debe calcularse por la siguiente reglas: P(A∪B) = P(A) + P(B) – P(A∩B).
    - **Asignación de probabilidad en espacios muestrales discretos**
        - Ejemplo: si la probabilidad de obtener 1 es P1, la de un 3 P2 y la de 5 es P3 la probabilidad será: P1+P2+P3
    - **Modelo uniforme**
        - P(A) = $\frac{k}{n} = \frac{casos\ favorables\ a\ A}{casos\ posibles}$
    - **Probabilidad condicionada**
    - **Independencia de sucesos**
    - **Teorema de la probabilidad total**
    - **Teorema de Bayes**

3. Modelos probabilísticos

- Introducción
    - En este capitulo se estudia un catalago de modelos probabilísticos analizando sus principales propiedades y viendo cuales son los fenómenos aleatorios típicos a los que asociar estos modelos. Tratando bien de inferir un valor para dicho parametro poblacional, bien de construir un intervalo de confianza o mediante un contrate de hipótesis.
- Distribución de probabilidad
    - Funciones básicas de R en probabilidad
- **Variables aleatorias multivariantes**
    - Independencia de variables aleatorias
- **Modelos unidimensionales discretos**
    - Distribución binomial
        - Pruebas de Bernoulli
            - Consiste en la realización de ensayos repetidos e independientes
            - Solo dos posibles resultados
                - Éxito
                - Fracaso

- **Distribución de poisson**
  - Se utiliza por lo general para modelizar el numero de veces que ocurren cosas raras, como por ejemplo el numero de incendios por año, o el numero de suicidios por año
  - Aproximación de la distribución binominial por la de Poisson.
- **Distribución geométrica**
  - Parametro p, se asocia también al modelo de Bernoulli, es decir de tipo éxito/fracaso con probabilidades de éxito p, (numero de fallos antes del primer éxito)
- **Distribución hipergeométrica**
  - Se utiliza para situaciones que se adaptan al siguiente esquema: se supone una caja con N piezas de las cuales D son defectuosas y N – D no defectusas. Se extraen sin reemplazamiento n piezas de la caja y estamos interesados en modelizar el numero de piezas defectuosas extraídas en las n seleccionas.
- **Distribución binomial negativa**
  - De nuevo experimento de Bernoulli, del tipo éxito/fracaso con probabilidad de éxito p, pero analizando ahora la variable X = numero de fallos antes de éxito n-esimo.

- **Modelos unidimensionales continuos**
  - **Distribución normal**
    - La distribución normal se define como aquella distribución cuya función de densidad es…
  - **Distribución uniforme**
    - De parámetros (a, b) es un modelo que asigna, de forma continua, igual probabilidad a todas las partes del intervalo (a, b) en el que esta definida.
  - **Distribución Beta**
    - De parámetros (a, b) tiene por función de densidad
  - **Distribuciones Gamma y Exponencial**
    - De parámetros (a, b) tiene por función de densidad
  - **Distribución de Cauchy**
    - De parámetros (a, b) tiene por función de densidad
- **Modelos bidimensionales**
  - **Distribución normal bivariante**
    - Es de tipo continuo y que nos da la idea de un caso multivariante
- **Teorema central del limite**
  - Tiene ese nombre porque se creía que era el modelo habitual de la mayoría de los fenómenos de la naturaleza.

4. **Estimadores**. Distribución en el muestreo: poder llegar a dar un valor como estimación suya.
   - Introducción
     - Como ya se indica en el capitulo dos, el objetivo de la inferencia Estadística es el obtener conclusiones sobre la una población mediante la observación de una parte de la misma denominada muestra, el propósito de la inferencia estadística es el obtener conclusiones sobre θ en base a una muestra aleatoria simple.
     - Distribución muestral: (X1,…..,Xn)
     - La estimación, como dijimos en el capitulo 2, estaremos interesados bien en asignar- o mejor dicho inferir- un valor numérico al parámetro θ (**estimación por puntos**), o bien en inferir un conjunto de valores plausibles para θ (**estimación por intervalos de confianza y contraste de hipótesis**)
   - **Método de la máxima verosimilitud**
     - Consiste en dar como estimación del parámetro aquel valor, de entre los posibles, que haga máxima la probabilidad del suceso observado, es decir, de la muestra obtenida, como hemos dicho el método de la máxima verosimilitud propone como estimador del parámetro aquel que maximice la probabilidad del suceso observado.
   - **Distribuciones asociadas a poblaciones normales**
     - En esta sección se estudian tres distribuciones de probabilidad continuas.

- Distribución x^2 de Pearson
  - Sean X1,…,Xn, n variables aleatorias independientes, cada una de las cuales sigue una distribución N(0, 1).
  - Llamaremos a la distribución de la variable aleatoria suma de los cuadrados de las n variables N(0, 1)
  - Y = X_1^2+X_2^2+…+X_n^2
- Distribución t de student
  - En poblaciones normales N(μ, σ), la distribución en el muestro en la media muestral x es tam,bien normal, aunque dependiente de σ, siendo este parámetro habitualmente desconocido.
  - T = $\frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X^2}}$
- Distribución F de Snedecor
  - Una distribución relacionada con la estimación de l cociente de varianzas de dos poblaciones normales es la de nominada F de Snedecor.variables aleatorias independientes e idénticamente distribuidas N(0, σ)
  - F= $\frac{\frac{1}{n}\sum_{i=1}^{n} X^2}{\frac{1}{n}\sum_{i=1}^{n} Y^2}$

- **Estimación de la media de una población normal**
  - En esta sección estudiaremos cual debe ser el estimador a utilizar para estimar la media μ, cuando para la variable X se supone como modelo una N(μ, σ)
  - **Teorema de Fisher**
    - Sea X1,…,Xn una muestra aleatoria simple de una población N(μ, σ). Entonces, si X y S2 son, respectivamente, la media y la cursivarianza muestrales se tiene que:
      - .
      - .
      - .
    - σ Conocida: cuando la varianza poblacional es conocida, es razonable utilizar la media muestral x para estimar μ.
    - σ Desconocida: si σ es desconocida, el resultado anterior sigue siendo valido, pero de poco nos va a servir al depender la distribución de x de este parámetro desconocido.
- **Estimación de la media de una población no necesariamente normal (muestras grandes)**
  - Se estudiar la situación en la que el modelo que se supone para la variable en estudio no es normal, o al menos no estamos lo suficientemente seguros de que lo sea como para poder utilizar los resultados.
  - **Población no necesariamente normal**
    - Si no conocemos o no queremos suponer uun modelo determinado para la variable de estudio, siempre que esta tenga varianza finita σ^2, podemos utilizar el teorema central del limite obteniendo para muestras suficientemente grandes, digamos n > 30
  - **Población binomial**
    - Si estamos interesados en estimar una proporción poblacional p, como por ejemplo la de alérgicos al polen de las acacia en España, es razonable establecer un modelo binominal para la variable dicotómica en estudio con p probabilidades de éxito X ~ B (1, p)
  - **Población poisson**
    - Si se admite un modelo X ~ P(λ), el estimador de máxima verosimilitud para λ basado en una muestra aleatoria simple de tamaño n de X, erta la media muestral x, el cual tiene distribución en el muestro tal que n x ~P(n λ)
- **Estimación de la varianza de una población normal**
  - Habrá que distinguir la situación en la que la media es conocida de la que no lo es.
    - μ desconocida
    - μ conocida
- **Estimación del cociente de varianzas de dos poblaciones normales independientes**

- Estudiaremos la distribución del muestreo de los estimadores utilizados en inferencias sobre cociente de las varianzas de dos poblaciones normales independietes.
    - μ_1 y μ_2 desconocidas
    - μ_1 y μ_2 conocidas

- **estimación de la diferencia de medias de dos poblaciones normales independientes**
    - un resultado de cálculo de probabilidades es que si tenemos dos normales independiestes, su suma (o diferencia) es una normal con la media la suma (o diferencia) de las medias y deviación típica la raíz cuadrada de la suma de las varianzas.
        - σ_1 y σ_2 desconocidas y las muestras son pequeñas
        - σ_1 y σ_2 se suponen iguales
        - σ_1 y σ_2 no se suponen iguales
        - σ_1 y σ_2 conocidas
- **Estimación de la diferencia de medias de dos poblaciones independientes no necesariamente normales (muestras grandes)**
    - Si las muestras son lo suficientemente grandes como para aplicar el teorema central del limite, digamos n1 + n2 > 30 en el primer caso que y n1 + n2 > 100 en el segundo y tercero, además de ser los tres casos n1 aproximadamente igual a n2:
        - σ_1 y σ_2 conocidas
        - σ_1 y σ_2 desconocidas
- **Datos apareados**
    - En el caso en que tengamos pares de datos dependientes (X1,Y1),…,(Xn,Yn). Situación que se da por ejemplo cuando se quiere estudiar la eficacia de una dieta de adelgazamiento, se determina su peso antes de iniciar el tratamiento Xi y después de finalizarlo Yi
- **Tamaño muestral para una precisión dada**
    - En estadística es frecuente que se quiera determinar resultados con una determinada precisión, medida en términos de probabilidad. Cuando se puede estar interesados en determinar el tamaño muestral para que el error en la estimación sea menor que 2, de forma que esto ocurra en el 95% de las veces de muestreamos, es decir, que haya probabilidad 0,95 de que esa diferencia sea así de pequeña.

5. **Intervalos de confianza**: dar un intervalo numérico en el que verosímilmente se encuentra
    - **α** = si hay que calcular el intervalo de confianza de 99% = 0,99, α será 1 – 0,99 = 0,01 y en la atabla en la que sea el intervalo.
- introducción
    - calculo de intervalos de confianza de R
- **intervalos de confianza para la media de una población normal (istudent)**
    - tanto en eata sección como en las siguientes, determinaremos intervalos de confianza de colas iguales. Es decir, aquellos tales que, si el coeficiente de confianza es 1 – Alpha, dejan en cada uno de los extremos la mitad de la probabilidad, Alpha / 2
    - en esta sección suponemos que los n datos proceden de una población N(μ, σ), y lo que pretendemos determinar es el intervalo de confianza para la media μ.
        - σ conocida
        - σ desconocida
- **intervalo de confianza para la media de una población no necesariamente normal (muestras grandes) (istudent)**
    - aqui consideramos primero una situación general y luego supondremos -realmente como casos particulares- que la población es binomial y que es de poisson
        - **población no necesariamente normal**
            - si no suponemos modelo alguno para la variable aleatoria en estudio, excepto que tenga varianza σ^2 finita y que la muestra de tamaño n sea suficientemente grande, tenemos dos situaciones posibles dependiendo del conocimiento o no de la varianza poblacional
                - σ conocida

        - σ desconocida
    - **población binomial**
        - si suponemos que X ~ B(1,p) y que la muestra es sufientemente grande, el intervalo de confianza para p de coeficientes 1 – alpha es:
    - **población de poisson**
        - suponiendo que X ~ P(landa) y que la muestra es sufientemente grande, el intervalo de confianza para landa de coeficiente 1 – Alpha es:

- **intervalo de confianza para la varianza de una población normal**
    - dada una muestra aleatoria simple X1,…,Xn de una población N(μ, σ), vamos a determinar el intervalo de confianza para σ^2 , distinguiendo dos casos según sea desconocida o no la media de la población μ.
        - μ desconocida
        - μ conocida
- **intervalos de confianza para el cociente de varianzas de dos poblaciones normales independientes(F de Snedecor)**
    - supondremos que X1,…,Xn e Y1,…,Yn son dos muestras de tamaño n1 y n2 extraídas respectivamente de dos poblaciones independientes N(μ, σ) y N(μ, σ),
        - μ1 y μ2 conocidas
        - μ1 y μ2 desconocidas
- **intervalos de confianza para la diferencia de medias de dos poblaciones normales independientes (tstudent)**
    - al igual que en la sección anterior suponemos que X1,…,Xn e Y1,…,Yn son dos muestras de tamaño normales independientes N(μ, σ) y N(μ, σ)
        - σ_1 y σ_2 conocidas
        - σ_1 y σ_2 desconocidas, muestras pequeñas
- **intervalos de confianza para la diferencia de medias de dos poblaciones independientes no necesariamente normales (muestras grandes)(tstudent)**
    - si ahora X1,…,Xn e Y1,…,Y2 son dos muestras de tamaños n1 y n2 suficientemente grandes, extradas de dos poblaciones independientes de medias μ1 y μ2 respectivamente, de las que solo suponemos que tiene varianza σ_1 y σ_2 finitas, tendremos que:
        - σ_1 y σ_2 conocidas
        - σ_1 y σ_2 desconocidas
    - Poblaciones binominiales
- **intervalos de confianza para datos apareados**
    - si la muestra que tenemos es de datos emparejados (X1,Y1),….,(Xn,Yn), en el sentido de proceder de una población bidimensional, la forma de actuar consiste en definir la variable unidimensional diferencial D = X1-Yi y aplicar a sus parámetros los intervalos de confianza antes determinados.

6. **Contrates de hipótesis**: poder decidir si puede considerarse razonable un valor u otro para dicho parámetro.
    - introducción y conceptos fundamentales
        - Métodos estadísticos basados en contrastes de hipótesis
        - Una de las dos hipótesis, generalmente la quew corresponder con la situación estándar, recibe el nombrte de hipótesis nula h sub 0, mientras la otra recibe el nombre de hipótesis alternativa h sub 1, siendo el contraste de hipótesis el proceso de decisión basado en técnicas estadísticas en la cula decidimos -inferimos- cual de las dos hipótesis creemos correcta, acepptandola y rechazando en consecuencia la otra.
        - Errores tipo I y II
            - Rechaar la hipótesis nula h sub 0 cuando es cierta, error de tipo I
            - El aceptar H sub 0 cuando es falsa denominaremos error tipo II
        - P- valor

- Una critica que puede plantearser al lector respecto a la técnica de los test de hipótesis, es la dependencia de nuestros resultadsos e el nivel de significación Alpha elegido antes de efectuar el contraste
- El calculo de p-valor pertmite valorar la decisión ya tomadfa de rechazar o aceptar h sub 0, de forma que eun p-valor grande – digamos 0,2 o mas – confirma una decisión de aceptación de H sub 0. Tanto mas no lo congfirma cuanto mayor se a el p -valor
  - Por el contrario cuanto mas pequeño se el p-valor digamos 0,01 o menos – cofirma una decisión de rechazo de h sub 0.
  - P-vlalor del contrate unilateral será
    - P-valor = P{t9 > 0,372} = 0,35925
  - P-valor en el bilateral
    - P-valor = P{|t9| > 0,372} = 2 . P{t9 > 0,327} = 0,7185

- Contraste de hipótesis relativas a la media de una población normal
- Contraste de hipótesis relativas a la media de una población no necesariamente normal (muestras grandes)
- Contraste de hipótesis relativas a la varianza de una población normal
- Contraste de hipótesis relativas a las varianzas de dos poblaciones independientes
- Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones normales independientes
- Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones independientes no necesariamente normales (muestras grandes)
- Contraste de hipótesis relativas para datos apareados

7. Contrastes no paramétricos

- Introducción
  - Esto test que un por un lado no requieren espicificar un modelo para la variable de estudio y por otro contrstan hipótesis que no se refierten a los valores de la media ni de la varianza.
- Pruebas x^2
  - Pruebas x^2 con R
  - Contraste de bondad del ajuste
  - Contraste de homogeneidad de varias muestras
  - Contraste de independencia de caracteres
- Test relativos a una muestra y datos apareados
  - El contraste de los signos
  - El contraste de los rangos signados de wilcoxon
- Test relativos a dos muestras independientes
  - El contraste de wilcoxon-mann-whitney
  - El contraste de la mediana

8. Análisis de la varianza

- introducción
- Análisis de la varianza para un factor: diseño completamente aleatorizado
- Análisis de varianza con R
- Análisis de las condiciones
- Comparaciones múltiples
- Comparaciones múltiples con R

9. Regresión lineal y correlación

- introducción
- Modelo de la regresión lineal simple
  - Interpretación de los coeficientes de regresión
- Contraste de la regresión lineal simple
  - Análisis de la variación explicada frente a la no explicada por la recta de regresión
  - Contraste de hipótesis para β_1
- Regresión lineal con R
- correlación lineal
  - Estimación por punto de ρ

        - Contraste de hipótesis sobre ρ
    - Modelo de la regresión lineal múltiple
        - Contraste de la regresión lineal múltiple