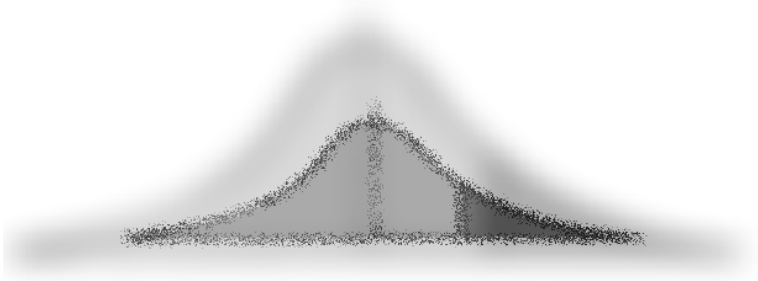
Anexo

III

Métodos estadísticos de análisis

- Estadística descriptiva
- Análisis de muestras y contraste de hipótesis
- Análisis de varianza factorial
- Prueba binomial
- Prueba Chi-cuadrado de Pearson
- Modelos de regresión lineal



Estadística descriptiva

La estadística descriptiva utiliza procedimientos, que resumen la información contenida en los extensos datos, a la vez que nos ayudan a depurar la muestra, ya que se estima empíricamente que entre un 2 y un 5% de las observaciones, pueden tener errores de medición o transcripción (Peña, 1991). Los procedimientos utilizados, pueden variar en función de los objetivos que nos planteamos en el estudio de la variable en cuestión. Si nos interesa un gran conocimiento de su comportamiento, podemos llegar a aplicar toda la metodología, mientras que en otros casos el lector observará que en la descripción de la variable, sólo se han utilizado algunos de los procedimientos aquí explicados. En cualquier caso, es necesario señalar que los métodos empleados son diferentes según la naturaleza de la variable estudiada (cualitativa o cuantitativa).

- **Variables cuantitativas.**

Una primera forma de sintetizar la información es la representación de la frecuencia de aparición de cada intervalo en tablas, o gráficamente en histogramas. Esto nos permite conocer la pauta que sigue la variable, así como la detección de aquellas observaciones que más se apartan de este comportamiento, permitiendo su revisión para evitar errores.

Las medidas que tienen una más fácil interpretación intuitiva sobre el valor de la variable son las medidas de centralización: Media aritmética, Mediana y Moda. La primera es preferible en datos homogéneos, mientras que las otras dos, son más robustas frente a errores, observaciones atípicas o datos muy heterogéneos. En cualquier caso, cuanto más simétrica es la distribución más similares son las tres medidas. Las medidas de dispersión son indicativas de lo alejados que están los valores de la media, utilizando en nuestro caso la desviación típica que puede interpretarse como el error de medición. Estas medidas pueden resumirse o fundirse en una única información mediante el cálculo del intervalo de confianza estimado como:

$$IC_{95\%} \approx \bar{x} \pm 2 \cdot \frac{\hat{s}}{\sqrt{n}} = \hat{p} \pm 2 \cdot \sqrt{\frac{p \cdot q}{n}}$$ (Peña, 1991)

La primera expresión se usa para obtener el intervalo de confianza de una media, estimando la varianza poblacional a través de la desviación típica muestral y redondeando el valor de Z-normal a dos por comodidad en los cálculos. La segunda expresión permite estimar el intervalo de confianza de una proporción (*p)*, siendo *q* la probabilidad del suceso contrario (1-*p*). Nótese en ambos casos, la gran importancia que tiene el tamaño muestral (*n*) para reducir el tamaño del intervalo de confianza.

Las medidas descritas serían las más habituales, pero existen otras que completan esta información o aportan otra adicional como el coeficiente de variación, coeficiente de asimetría, curtosis, rango...etc. O incluso otras que permiten la descripción conjunta de variables como el coeficiente de correlación o la matriz de varianzas-covarianzas. En cualquier caso no se ha considerado necesario el uso de estos procedimientos para el estudio, aunque si que se ha utilizado y ha resultado de gran interés el diagrama de caja como resumen de la mayoría de las anteriores variables en un simple gráfico, permitiendo observar de forma aproximada la distribución de frecuencias, la simetría, la media y la mediana, así como los valores extremos y atípicos que pudieran contener errores.

- **Variables cualitativas.**

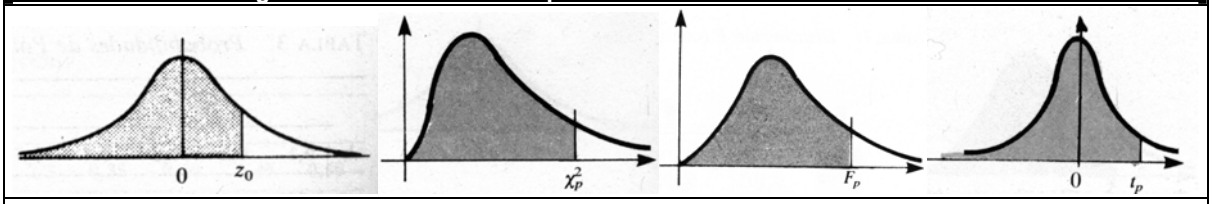
Al igual que en las variables cuantitativas, una buena forma de comenzar es la representación de la frecuencia de aparición de cada clase utilizando tablas, diagramas de barras (no confundir con los histogramas que representan variables continuas) o diagramas de sectores según el objetivo. Estas tablas mostrarán el tamaño muestral, así como el porcentaje de cada clase, pudiendo incorporar el intervalo de confianza de este porcentaje según el procedimiento anteriormente explicado.

Análisis de muestras y contraste de hipótesis

El **análisis de muestras** es una herramienta utilizada en el caso de que no se pueda estudiar toda la población (como ocurre en nuestro proyecto), por lo que se intenta explicar su comportamiento a partir del estudio de una muestra de la población, y el establecimiento del correspondiente modelo de distribución.

La distribución más habitual en variables continuas para muchos procesos de medición sin errores sistemáticos, es la distribución normal, como justifica el teorema central del límite, que establece que cuando los resultados de un experimento son debidos a un conjunto muy grande de causas independientes, que actúan sumando sus efectos, siendo cada efecto individual de poca importancia respecto al conjunto, es esperable que los resultados sigan una distribución normal N. Otras distribuciones importantes relacionadas con la normal son: la $\chi 2$ de Pearson, la F de Fisher (útil para comparar varianzas de poblaciones normales) y la t de Student.

Figura 19: Distribuciones de probabilidad más usadas en estadística



(Fuente: Peña, 1991)

En la práctica para muestras mayores de 30 individuos, la media muestral seguirá una distribución normal, mientras que esto sólo se produce para la varianza muestral en tamaños muestrales muy grandes, y con una aproximación muy lenta, por lo que suelen usarse distribuciones asimétricas. Los estimadores deben ser centrados (próximos al parámetro poblacional), precisos (con baja variabilidad), con bajo error cuadrático medio, y en el peor de los casos, consistentes (deben acercarse al parámetro poblacional al aumentar el tamaño muestral).

Además, no solo nos interesa conocer el valor del estimador, si no también su intervalo de confianza con un determinado nivel de confianza (1-α), por ejemplo, un intervalo para la media de una población de nivel 0,95 (α=0,05) indica que el 95% de los intervalos construidos contienen el verdadero valor de la media.

El **contraste de hipótesis** sirve para comprobar si cierta propiedad supuesta para una población, es compatible con lo observado en una muestra de ella. Para ello se parte de una hipótesis neutra, llamada hipótesis nula (H_0), que se contrasta para intentar probar su falsedad y poder rechazarla si es así, pero en el caso de no quedar probada su falsedad, tampoco podemos considerarla como probada, ya que para ello debería estudiarse toda la población.

La región de rechazo se establece fijando un nivel de significación (α) o nivel de probabilidad tal que sucesos con probabilidad de ocurrir menor que α los consideramos despreciables según la importancia de la hipótesis (en nuestro trabajo consideraremos generalmente α=0,05 y más raramente 0,01). Sin embargo, este rechazo puede ser muy subjetivo al ser nosotros los que fijamos α y no evaluar la distancia a la región de aceptación, por ello se define el nivel crítico p, como la probabilidad de obtener una discrepancia mayor o igual que la observada en la muestra, cuando la H_0 es cierta. De este modo, si p es menor de 0,01 se rechaza H_0, si es mayor de 0,25 no podemos rechazarla, mientras que si estamos entre 0,2 y 0,01 nosotros podemos decidir en función de la importancia del rechazo, pero siendo conscientes de que estamos en la región crítica. En cualquier caso, el contraste de hipótesis también puede realizarse comparando los intervalos de confianza, que son mucho más informativos que el test.

Análisis de varianza factorial

Se trata de un procedimiento que permite dividir la variabilidad de un experimento en distintos componentes según sus causas (variabilidad explicada y no-explicada) esto nos permite comparar medias entre grupos (niveles) y explicar las posibles diferencias. Este método denominado ANOVA según sus siglas en inglés, o más correctamente ADEVA en castellano, realiza un contraste de hipótesis en el que se considera como hipótesis nula, el que las medias de todos los niveles son iguales, por lo que no se ven influidas por el nivel al que pertenecen. El rechazo de esta hipótesis implicará que alguna de las medias no es igual y, por tanto, alguno de los niveles tiene una media diferente y podría ser influyente, por lo que entonces se procede a evaluar la diferencia o comparación entre medias.

Si el análisis se aplica a más de un factor, la tabla ADEVA incluye un nuevo factor que es la interacción de los factores simples. En este caso, para la correcta interpretación del modelo, si la interacción resulta significativa hay que conocer de qué tipo es mediante los gráficos de interacción. Si las líneas del gráfico son paralelas no hay interacción; si son divergentes o convergentes indica que hay interacción entre factores. En este último supuesto, si las líneas no llegan a cruzarse, indica que la interacción es cuantitativa por lo que podríamos seguir interpretando los factores por separado. Pero si las líneas se cruzaran, sería una interacción cualitativa, produciéndose un cambio de rango entre los distintos niveles de los factores, por lo que los factores sólo se podrían interpretar conjuntamente.

La tabla ADEVA presenta los factores entre los que se reparte la variabilidad existente (uno de ellos es el error o variabilidad no explicada por los factores del modelo), la suma de cuadrados como medida de la variabilidad, los grados de libertad o términos estadísticamente independientes y la media cuadrática qué es la relación entre la suma de cuadrados y los grados de libertad. Como estadístico de contraste para conocer la significación de un factor, se utiliza la F que se distribuye según una F de Fisher y se calcula como la relación entre la media cuadrática del factor y la media cuadrática del error.

Debe advertirse que la suma de cuadrados empleada es de tipo III, pues de este modo no es influyente que el diseño del experimento esté desequilibrado como en nuestro estudio. En este caso, la media cuadrática del error es el mismo para todos los factores y se estima cada factor controlando por todos los restantes incluidos en el modelo saturado (en Ponz-2000 citado de Díaz Esteban-2000)

Tras el estudio de la interacción se procede a determinar cuales son las medias diferentes (en el caso de que algún factor resulte significativo), estudiando la diferencia entre medias si sólo tenemos dos niveles en el factor, o métodos de separación de medias cuando existen más niveles. De los múltiples métodos existentes hemos aplicado dos: el método Student-Newman-Keuls como menos conservativo y el método DHS de Tukey como más conservativo.

Finalmente debe realizarse la diagnosis y crítica del modelo, para determinar si los datos disponibles son consistentes con las hipótesis básicas que establece el modelo. Para este tipo de análisis deben estudiarse tres condiciones: normalidad, homocedasticidad e independencia.

El incumplimiento de la normalidad supone que el procedimiento deja de ser óptimo, pero afortunadamente el ADEVA es bastante robusto frente a este supuesto, por lo que su incumplimiento no es de gran importancia. Para la determinación de esta condición se ha aplicado el Test de Kolmogorov-Smirnov con la corrección de significación de Lilliefors.

La homocedasticidad se ha estudiado mediante el Contraste de homogeneidad de la varianza de Levene, por su independencia de la condición de normalidad (Lizasoain y Joaristi, 1999). Su incumplimiento es más problemático sobretodo en diseños desequilibrados. No obstante, el nivel de significación es mayor que si se cumpliese la condición (Lizasoain y Joaristi, 1999), por lo que las diferencias detectadas deben tenerse en consideración del mismo modo, aunque no tendremos en cuenta los intervalos de confianza calculados. Además, según Peña (1991), la distinta variabilidad existente, no tiene por qué deberse a errores del muestreo, sino que puede reflejar la propia variabilidad que tienen cada uno de los grupos en que se divide la población.

En el caso de que las observaciones no sean independientes, las expresiones usadas para las varianzas de los estimadores son erróneas, por tanto, los intervalos de confianza y contrastes, tendrán un nivel de

confianza distinto al supuesto, por lo que es imprescindible el cumplimiento de esta hipótesis. Para comprobar este supuesto se han aplicado dos pruebas: la Prueba de Rachas y la "d" de Durbin-Watson que es una medida de la autocorrelación.

Prueba binomial

El procedimiento Prueba binomial compara las frecuencias observadas de las dos categorías de una variable dicotómica con las frecuencias esperadas en una distribución binomial con un parámetro de probabilidad especificado. Por defecto, el parámetro de probabilidad para ambos grupos es 0,5. Para cambiar las probabilidades, puede introducirse una proporción de prueba para el primer grupo. La probabilidad del segundo grupo será 1 menos la probabilidad especificada para el primer grupo. (SPSS, 2002). La hipótesis nula planteada en este test es que la frecuencia de aparición de las dos categorías es coincidente, mientras que en la hipótesis alternativa supondría que presentan distinta frecuencia.

Prueba Chi-cuadrado de Pearson

El procedimiento Prueba de chi-cuadrado tabula una variable en categorías y calcula un estadístico chi-cuadrado. Esta prueba de bondad de ajuste compara las frecuencias observadas y esperadas en cada categoría para contrastar si todas las categorías contienen la misma proporción de valores o si cada categoría contiene una proporción de valores especificada por el usuario. Las pruebas no paramétricas no requieren supuestos sobre la forma de la distribución subyacente. Se asume que los datos son una muestra aleatoria. Las frecuencias esperadas para cada categoría deberán ser 1 como mínimo. No más de un 20% de las categorías deberán tener frecuencias esperadas menores del 5%. (SPSS, 2002)

Según Peña (1991) la muestra debe contener al menos 25 individuos, agrupados en un mínimo de 5 clases, siendo conveniente que los intervalos cubran todo el rango de datos de la variable y que aproximadamente, todas las clases tengan el mismo número de datos y no menos de 3.

La condición del tamaño muestral indicada se da como mínimo para tener una muestra aleatoria, mientras el resto de condiciones señaladas por los autores en cuanto al número de clases y frecuencia de cada clase, se debe al hecho de que cuantas menos clases tenemos o menos frecuentes son, baja la probabilidad de detectar diferencias, es decir, sólo podremos encontrar diferencias cuando ambas distribuciones sean muy distintas. Con el fin de unificar condiciones nosotros nos hemos fijado como límites de la prueba el tener un tamaño muestral mayor de 25 con un mínimo de 5 clases y frecuencia mayor del 5% en todos los grupos.

Modelos de regresión lineal

La regresión lineal permite establecer modelos predictivos mediante la relación de una variable dependiente, con una o más variables independientes explicativas, aplicando el método de los mínimos cuadrados ordinarios. Los modelos de regresión simple deben cumplir cuatro hipótesis básicas: linealidad, normalidad, homocedasticidad e independencia. Si el modelo es de regresión múltiple se añade una quinta hipótesis: multicolinealidad.

Es de gran importancia para establecer un buen modelo predictivo el comprobar las hipótesis y en caso de incumplimiento acometer los procedimientos oportunos para salvar los inconvenientes que se deriven. En el estudio que nos ocupa los modelos de regresión no han sido utilizados con el objeto de crear modelos predictivos, sino con la finalidad de conocer la relación entre algunas variables. Al tener esta finalidad no hemos sido tan estrictos en la corrección de los modelos según su diagnosis, pues el incumplimiento de las hipótesis no tiene por qué impedir el estudio de la relación entre las variables. En cualquier caso, si que se han aplicado métodos de diagnosis de las hipótesis con el fin de conocer las limitaciones de los modelos o incluso, que puedan servir como inicio de un estudio para obtener modelos predictivos, que naturalmente, se saldrían del objetivo tan general que tiene este trabajo.

Para el estudio de la linealidad se han utilizado los diagramas de dispersión de las observaciones y de los residuos, así como el coeficiente de correlación (R). La falta de linealidad cuestiona al modelo en sí, pues sino es lineal ¿cómo podemos pretender predecir resultados con un modelo lineal?. En este caso obtendríamos un mal ajuste de la recta (R^2), y habría que recurrir al modelo curvilíneo que mejor se ajustase (procedimiento mucho más complejo que el lineal) o se puede optar por transformar las variables originales para intentar obtener linealidad.

La homocedasticidad puede intuirse con los diagramas de dispersión de los residuos. Su incumplimiento no es razón para descartar el modelo, pero afectará a los intervalos de confianza calculados y las predicciones pueden ser peligrosas según el rango de la x en que trabajemos. La solución a este problema es el uso del método de los mínimos cuadrados ponderados o la transformación de las variables.

Para detectar la normalidad se ha utilizado el histograma de los residuos superpuesto a una curva normal. La falta de normalidad provoca que no sean válidos los intervalos de confianza, y por tanto, tampoco serán válidos los contrastes realizados. Para salvar este inconveniente podemos renunciar a los contrastes, eliminar aquellas observaciones atípicas erróneas, transformar la variable respuesta o introducir en el modelo otras variables relevantes.

La independencia se ha comprobado con el estadístico “d” de Durbin-Watson que mide la autocorrelación entre las observaciones. La falta de independencia implica que la muestra no es aleatoria, por lo que el modelo no podrá ser válido, pues se calcula a partir de una muestra incorrecta. Para que nunca tengamos este problema es imprescindible que la muestra se tome aleatoriamente.

La multicolinealidad o relación entre variables independientes es totalmente indeseable en el modelo, pues aunque no afecta a las predicciones, puede impedir la correcta interpretación de los coeficientes del modelo o aumentar el tamaño de los intervalos de confianza hasta hacer que el coeficiente no sea significativo. Se considera que existe multicolinealidad cuando el factor de incremento de la varianza (VIF) es superior a 10. Se puede solucionar eliminando las variables con mayor VIF o uniendo las variables en una nueva variable.