PROBABILIDAD Y ESTADÍSTICA FUNDAMENTAL - TALLER 1

Profesor: Andrés Nicolás López. Departamento de Estadística

Preguntas 1

1.1 Según lo discutido en clase responda:

a. ¿Cuál es la principal diferencia entre la estadística y la matemática?

b. Explique 3 razones por las cuales se cuestiona la posibilidad inferencial en los sondeos electorales de las firmas encuestadoras del plebiscito por la paz en Colombia

c. ¿Es necesaria la inferencia estadística cuando se observa toda la población?

d. ¿Por qué muestrear de una población?.

e. ¿Cuál es la importancia de la estadística para la toma de decisiones?

f. ¿Cuál es la principal diferencia entre la estadística descriptiva y la estadística inferencial?

1.2 Responda Verdadero (V) o Falso (F) según corresponda **y explique**:

a. En el estudio de la estadística se presenta un error inherente que en la práctica es despreciable

b. La característica común entre las 4 escalas de medida estudiadas (nominal, ordinal, intervalo y razón) es que entre las modalidades de respuesta se cumple la relación de igualdad o desigualdad

c. La escala de intervalo y de razón proveen el mismo nivel de sofisticación.

d. La representación gráfica de una variable cuantitativa discreta es siempre igual a la representación de una variable cualitativa nominal u ordinal.

e. La variable sexo al nacer codificada como 0 (igual a hombre) y 1 (igual a mujer) es cuantitativa, pues sus atributos son ahora representados por números.

f. La detección de datos atípicos a partir del diagrama de caja, la regla empírica y la desigualdad de Chebyshev aplica únicamente para distribuciones acampanadas.

Preguntas 2

2.1 Identifique las unidades experimentales en las que se miden las siguientes variables. Adicionalmente, determine el tipo de variable involucrada y la escala de medición de la misma:

a. Tamaño del tumor cancerígeno de un paciente.

b. Intención de voto para las elecciones presidenciales.

c. Estadío del cáncer en un paciente.

d. Cociente de inteligencia (IQ) de los candidatos a la alcaldía.

e. Grado de escolaridad de un votante.

2.2 Identifique cada una de las variables cuantitativas como discretas o continuas:

a. Número de accidentes en botes en un tramo de 50 millas del río.

b. Tiempo para completar un cuestionario.

c. Rendimiento en kilogramos de una cosecha de papas.

d. Población en una región particular de un país.

e. Número de pensamientos intrusivos después del diagnóstico de cáncer.

2.3 Un investigador médico desea estimar el tiempo de supervivencia de un paciente con cáncer después de un régimen particular de radioterapia:

a. ¿Cuál es la variable de interés para el investigador médico?

b. ¿La variable del inciso anterior es cualitativa, cuantitativa discreta o cuantitativa continua?

c. Identifique la población de interés para el investigador médico.

d. De manera simple, describa la forma en que el investigador podría seleccionar una muestra de entre la población.

e. ¿Qué problemas podrían surgir al muestrear desde esta población?

Preguntas 3

3.1 Responda Verdadero (V) o Falso (F) según corresponda **y explique**:

a. A diferencia del diagrama de tallo y hojas, el histograma permite recuperar los valores individuales de la variable de interés.

b. El número de intervalos seleccionados para la construcción de un histograma debe seleccionarse cuidadosamente y generalmente de manera experta (es decir, por parte del investigador).

c. La descripción de datos con medidas numéricas es importante en el estudio de la variable, sin embargo, un muy buen resumen de la distribución de frecuencias es suficiente para caracterizar de manera completa la variable de interés.

d. La mediana es la única medida de tendencia central que puede calcularse para todas las escalas de medida estudiadas.

e. Para la media muestral, todos los individuos tienen el mismo peso, por lo cual, esta es una estadística bastante robusta ante valores atípicos

f. El rango muestral generalmente sobreestima el poblacional.

g. A diferencia de la varianza, el coeficiente de variación no tiene unidades.

3.2 Para una muestra de 20 días del mes de Enero de 2022 se obtuvo la distribución de frecuencias absolutas de la variable *Número de personas que ingresan diariamente a la unidad de urgencias del hospital de Usaquén.*

```
The decimal point is 1 digit(s) to the right of the |

0 | 2
0 | 778889
1 | 0001111222244
```

Figure 1: Diagrama de tallo y hojas. Ejercicio 4.2.

La Figura 1 presenta la distribución de frecuencias muestral mediante un diagrama de tallo y hojas. A partir de esta:

a. Recupere la información original del gráfico de tallo y hojas y represéntela de manera adecuada en una pequeña base de datos.

b. Describa el tipo de variable medida y la escala de medición de la misma.

c. Caracterice de manera **cualitativa** la distribución de la variable *Número de personas que ingresan diariamente a la unidad de urgencias del hospital de Usaquén*: tendencia, dispersión y forma.

c. Caracterice de manera **cuantitativa** la distribución de la variable *Número de personas que ingresan diariamente a la unidad de urgencias del hospital de Usaquén*: tendencia (media, mediana, moda), dispersión (varianza, desviación estándar, rango e IQR) y forma.

Preguntas 4

4.1 a. Muestre que la suma de desvíos respecto a $\bar{x}$ para un conjunto de observaciones x_1, ... ,x_n es igual a cero, es decir, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

b. Si un conjunto de observaciones x_1, ... ,x_n mayores a cero es transformado conforme a $y_i = \ln x_i$ para $i = 1, ..., n$ ¿A qué es igual exp $\bar{y}$ en términos de los datos originales?

c. Suponga que tiene un conjunto de observaciones $x_1, ..., x_n$ con media $\bar{x}$ y varianza s_x^2 el cual es transformado conforme a

$$z_i = \frac{x_i - x}{s_x} \text{ para } i = 1, ..., n$$

¿A qué es igual $\bar{z}$ y s_z^2?

Preguntas 5

5.1 En caso de ser posible, construya el gráfico boxplot para la variable *Estatura en cms de los estudiantes del curso de Estadística* con la siguiente información:

- La persona más baja mide 100 cm.
- ¡La distribución es completamente simétrica!.
- La segunda persona más alta mide 195 cm, la más alta 220 cm.
- El tercer cuartíl es 170 cm y la mediana 155 cm.

5.2 Según lo aprendido en clase respecto al diagrama de caja y bigotes o boxplot:

a. Describa detalladamente el proceso para su construcción.

b. Represente gráficamente mediante un bosquejo la caracterización de las siguientes distribuciones:
   - Asimétrica, alto apuntamiento y sesgo a la derecha. Un outlier superior.
   - Platicúrtica y simétrica. Sin outliers.
   - Asimetría negativa, bajo apuntamiento y dos outliers inferiores.

5.3 Basado en la lectura del primer capítulo de la monografía Gráficos Estadísticos con R

https://cran.r-project.org/doc/contrib/grafi3.pdf

Responda:

a. ¿Cuál es la motivación de la distorsión de los gráficos estadísticos por parte de los medios de comunicación?.

b. Escriba y explique 3 de los principios de William Playfair en la elaboración de gráficos.

c. ¿Cuál es la principal desventaja de los gráficos que el autor denomina *de paquete*?.

Pregunta 6

El conjunto de datos Wage de la librería ISLR brinda información de un grupo de trabajadores. Los datos presentan la información del salario y otras variables de interés de una muestra de tamaño $n = 3000$.

a. Extraiga las mediciones de la variable *age*, que corresponde a la edad en años de los trabajadores, del conjunto de datos e identifique el tipo de variable y su escala de medición.

b. Realice el diagrama de barras para la variable *age* y analice gráficamente la distribución de frecuencias relativas de los datos. Tenga en cuenta características tales como tendencia, dispersión y forma. ¿Requiere categorizar (construir intervalos) para esta variable?, en dado caso, utilice la regla de Sturges y usando el método de inclusión a izquierda.

c. Realice un análisis completo de la distribución de los datos a través de la descripción numérica de la variable.

Ayuda en R

- `summary()`. Centro, localización y dispersión.
- `install.packages(e1071)`. Forma asimetría.

Pregunta 7

Del mismo conjunto de datos del punto anterior (Wage), considere ahora la variable wage, que es propiamente aquella que proporciona el valor del salario semanal de los 3000 trabajadores de la muestra.

a. Trace un histograma que considere adecuado y que permita analizar preliminarmente la tendencia, dispersión y forma de los datos.

b. Construya una gráfica de caja. En comparación con el primer literal de este punto, comente si sus observaciones preliminares fueron adecuadas.

c. ¿Qué puede decir de los datos atípicos de esta variable? Tenga en cuenta los tres criterios vistos en clase que apliquen para este conjunto de datos.

Ayuda en R

- `hist()`. Histograma.
- `boxplot()`. Diagrama de caja.

Pregunta 8

Se quiere estudiar el número de horas que emplean los estudiantes de Probabilidad y Estadística Fundamental en transportarse diariamente. Por el gran volumen de estudiantes inscritos en la asignatura, se decide encuestar únicamente a 50 estudiantes. A continuación se muestra el conjunto de datos recolectados:

1, 2, 2, 3, 1, 3, 4, 2, 2, 1
1, 1, 2, 2, 3, 2, 2, 5, 2, 3
4, 1, 2, 1, 1, 2, 1, 2, 3, 1
2, 2, 2, 1, 3, 2, 3, 1, 2, 2
3, 2, 3, 2, 2, 3, 1, 2, 2, 2

a. ¿Es este conjunto de mediciones una población o una muestra?.

b. Identifique el tipo de variable (cualitativa, cuantitativa discreta ó cuantitativa continua) y la escala de medición correspondiente (nominal, ordinal, intervalo o razón).

c. Realice la representación tabular de la variable empleando la función adecuada en R. Tenga en cuenta el tipo de variable y su escala de medida.

d. Represente gráficamente el comportamiento de los datos en R. Tenga en cuenta la misma recomendación del numeral anterior.

Ayuda en R

- `table()`. Frecuencias absolutas.
- `prop.table()`. Frecuencias relativas.
- `cut()`. Categorización de una variable.