SOLUCIÓN TAREA 1

Análisis exploratorio manual de datos químicos

Introducción al análisis de datos para otras carreras

Escuela de Informática y Computación

Información del estudiante
Nombre: [Nombre del estudiante]
Carné: [Número de carné]
Carrera: Química Industrial
Fecha de entrega: 20 de agosto, 2025
Profesor: Jordy Alfaro Brenes

1. Parte I: Organización y clasificación de datos

1.1. Ejercicio 1.1: Clasificación del tipo de datos (5 puntos)

Respuesta:

a) Tipo de dato:
La variable pH es de tipo **cuantitativo continuo**.
Justificación: El pH es una variable numérica que puede tomar cualquier valor real dentro de su rango (0-14), incluyendo valores decimales como 6.95, 7.02, etc.
b) Escala de medición:
El pH corresponde a una escala de **intervalo**.
Justificación: Aunque el pH tiene un punto cero definido (pH = 7 es neutro), este cero no representa ausencia total de acidez, sino neutralidad. Las diferencias entre valores son significativas y consistentes (la diferencia entre pH 3 y pH 4 es la misma que entre pH 7 y pH 8), pero las razones no son interpretables directamente.
c) Operaciones matemáticas válidas:
En una escala de intervalo son válidas las operaciones de:

- Suma y resta de diferencias
- Cálculo de medidas de tendencia central (media, mediana, moda)
- Cálculo de medidas de dispersión (desviación estándar, varianza)
- Comparaciones ordinales (mayor que, menor que)

No son válidas: Las operaciones de multiplicación y división directa entre valores de pH, ya que no hay un cero absoluto.

1.2. Ejercicio 1.2: Organización de datos (10 puntos)

Proceso de ordenamiento:

a) Datos ordenados de menor a mayor:
6.82, 6.83, 6.84, 6.85, 6.85, 6.86, 6.87, 6.87, 6.88, 6.88, 6.89, 6.89, 6.90, 6.91, 6.91, 6.92, 6.93, 6.94, 6.94, 6.95, 6.95, 6.96, 6.97, 6.98, 6.98, 6.99, 7.00, 7.01, 7.02, 7.03, 7.04, 7.04, 7.05, 7.05, 7.06, 7.06, 7.07, 7.08, 7.09, 7.10, 7.11, 7.12, 7.13, 7.14, 7.15, 7.15, 7.16, 7.17, 7.18, 7.19
b) Valores extremos:
Valor mínimo: 6.82
Valor máximo: 7.19
c) Rango:
Rango = Máximo - Mínimo = 7.19 - 6.82 = 0.37

Tabla de frecuencias (6 intervalos):

Amplitud del intervalo = $\frac{\text{Rango}}{\text{Número de clases}} = \frac{0{,}37}{6} = 0{,}062$

Intervalo	Límites	Frecuencia	Frecuencia relativa
1	[6.82 - 6.88)	7	0.14
2	[6.88 - 6.94)	8	0.16
3	[6.94 - 7.00)	9	0.18
4	[7.00 - 7.06)	10	0.20
5	[7.06 - 7.12)	8	0.16
6	[7.12 - 7.19]	8	0.16
Total		**50**	**1.00**

2. Parte II: Medidas de tendencia central

2.1. Ejercicio 2.1: Media aritmética (8 puntos)

Cálculo paso a paso:

Fórmula: $\bar{x} = \frac{\sum x}{n}$
a) Suma de todos los valores:
$\sum x = 7{,}02 + 6{,}95 + 7{,}18 + 6{,}91 + 7{,}09 + 6{,}83 + 7{,}16 + 6{,}87 + 7{,}05 + 6{,}94 + 6{,}89 + 7{,}11 + 6{,}84 + 7{,}05 + 6{,}97 + 7{,}13 + 6{,}90 + 7{,}04 + 6{,}98 + 6{,}85 + 7{,}15 + 6{,}87 + 7{,}06 + 6{,}93 + 7{,}01 + 6{,}96 + 7{,}07 + 6{,}82 + 7{,}00 + 6{,}88 + 6{,}92 + 7{,}03 + 6{,}99 + 7{,}17 + 6{,}85 + 7{,}14 + 6{,}94 + 7{,}19 + 6{,}91 + 7{,}15 + 7{,}08 + 6{,}88 + 7{,}12 + 6{,}86 + 7{,}04 + 6{,}98 + 7{,}10 + 6{,}89 + 7{,}06 + 6{,}95$
$\sum x = 349{,}54$
b) Cálculo de la media:
$\bar{x} = \frac{349{,}54}{50} = 6{,}9908$
c) Resultado redondeado:
$\bar{x} = 6{,}991$

Interpretación en contexto químico:
La media de pH de 6.991 indica que, en promedio, las muestras de la solución buffer están ligeramente por debajo del punto neutro (pH = 7.0). Esto sugiere una tendencia levemente ácida en la formulación, lo cual puede ser aceptable dependiendo de la aplicación específica de la solución buffer. Para aplicaciones industriales que requieren un pH estrictamente neutro, esto podría requerir ajustes en la formulación.

2.2. Ejercicio 2.2: Mediana (8 puntos)

Cálculo de la mediana:

a) Procedimiento para n = 50 (número par):
Para un número par de observaciones, la mediana es el promedio de los valores en las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$.
Posiciones: $\frac{50}{2} = 25$ y 26
b) Identificación de valores:
Valor en posición 25: 6.98
Valor en posición 26: 6.99
c) Cálculo:
$Mediana = \frac{6{,}98+6{,}99}{2} = \frac{13{,}97}{2} = 6{,}985$

Comparación con la media:

- Media: 6.991
- Mediana: 6.985
- Diferencia: 6.991 - 6.985 = 0.006

Interpretación de la distribución:
La diferencia mínima entre la media y la mediana (0.006) indica que la distribución de los datos es aproximadamente simétrica. No hay evidencia de asimetría significativa en los datos, lo que sugiere que los valores extremos no están influyendo desproporcionadamente en las medidas de tendencia central.

2.3. Ejercicio 2.3: Moda (9 puntos)

Identificación de la moda:

a) Análisis de frecuencias:
Conteo de cada valor:

- 6.85: 2 veces
- 6.87: 2 veces
- 6.88: 2 veces
- 6.89: 2 veces
- 6.91: 2 veces
- 6.94: 2 veces
- 6.95: 2 veces
- 6.98: 2 veces
- 7.04: 2 veces
- 7.05: 2 veces
- 7.06: 2 veces
- 7.15: 2 veces

Todos los demás valores aparecen solo una vez.
b) Clasificación de la distribución:
La distribución es **multimodal**, ya que hay múltiples valores que aparecen con la misma frecuencia máxima (2 veces).

Medida de tendencia central más apropiada:

Para este dataset, la **media** es la medida de tendencia central más apropiada por las siguientes razones:

1. **Naturaleza de los datos:** Los datos son cuantitativos continuos en escala de intervalo
2. **Distribución simétrica:** La proximidad entre media y mediana confirma simetría
3. **Ausencia de moda clara:** La naturaleza multimodal hace que la moda sea poco informativa
4. **Aplicación química:** Para control de calidad, la media proporciona el valor central más representativo del proceso
5. **Sensibilidad:** La media es más sensible a todos los valores, lo cual es importante para detectar cambios en el proceso de producción

3. Parte III: Medidas de tendencia no central

3.1. Ejercicio 3.1: Cuartiles (15 puntos)

Cálculo de cuartiles:

a) Primer cuartil (Q1):
Posición: $Q_1 = \frac{1(n+1)}{4} = \frac{1(51)}{4} = 12{,}75$
Interpolación entre posiciones 12 y 13:

- Valor en posición 12: 6.89
- Valor en posición 13: 6.90

$Q_1 = 6{,}89 + 0{,}75(6{,}90 - 6{,}89) = 6{,}89 + 0{,}75(0{,}01) = 6{,}8975$
b) Tercer cuartil (Q3):
Posición: $Q_3 = \frac{3(n+1)}{4} = \frac{3(51)}{4} = 38{,}25$
Interpolación entre posiciones 38 y 39:

- Valor en posición 38: 7.08
- Valor en posición 39: 7.09

$Q_3 = 7{,}08 + 0{,}25(7{,}09 - 7{,}08) = 7{,}08 + 0{,}25(0{,}01) = 7{,}0825$
c) Rango intercuartílico:
$IQR = Q_3 - Q_1 = 7{,}0825 - 6{,}8975 = 0{,}185$

Interpretación en términos de control de calidad:

- **Q1 = 6.898:** El 25 % de las muestras tienen pH menor o igual a 6.898
- **Q3 = 7.083:** El 75 % de las muestras tienen pH menor o igual a 7.083
- **IQR = 0.185:** El 50 % central de las muestras tiene una variación de pH de 0.185 unidades

Implicaciones para control de calidad:
El IQR relativamente pequeño (0.185) indica una buena consistencia en el proceso de producción. La mayoría de las muestras (50 % central) se mantienen dentro de un rango estrecho, lo que es deseable para un proceso de control de calidad robusto.

3.2. Ejercicio 3.2: Detección de outliers (5 puntos)

Cálculo de límites para outliers:

a) Límites:
Límite inferior $= Q_1 - 1{,}5 \times IQR = 6{,}8975 - 1{,}5(0{,}185) = 6{,}8975 - 0{,}2775 = 6{,}620$
Límite superior $= Q_3 + 1{,}5 \times IQR = 7{,}0825 + 1{,}5(0{,}185) = 7{,}0825 + 0{,}2775 = 7{,}360$
b) Identificación de outliers:
Valores menores que 6.620: Ninguno
Valores mayores que 7.360: Ninguno
Conclusión: No se detectaron outliers en el dataset.

Discusión sobre ausencia de outliers:

La ausencia de outliers en el dataset indica:

1. **Proceso controlado:** Las condiciones experimentales estuvieron bien controladas
2. **Mediciones precisas:** No hay evidencia de errores graves de medición
3. **Procedimientos estandarizados:** Los protocolos de laboratorio fueron seguidos consistentemente
4. **Calidad de la muestra:** No hay evidencia de contaminación significativa

Si hubiera outliers, las posibles causas químicas incluirían:

- Contaminación de muestras
- Errores en la calibración del pH-metro
- Variaciones en temperatura no controladas
- Problemas en la preparación de la solución buffer

4. Parte IV: Medidas de dispersión

4.1. Ejercicio 4.1: Varianza y desviación estándar (15 puntos)

Cálculo de varianza muestral:

Fórmula: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
a) Primeros 5 cálculos de $(x - \bar{x})^2$:
Con $\bar{x} = 6{,}991$:

$$(7{,}02 - 6{,}991)^2 = (0{,}029)^2 = 0{,}000841 \quad (1)$$
$$(6{,}95 - 6{,}991)^2 = (-0{,}041)^2 = 0{,}001681 \quad (2)$$
$$(7{,}18 - 6{,}991)^2 = (0{,}189)^2 = 0{,}035721 \quad (3)$$
$$(6{,}91 - 6{,}991)^2 = (-0{,}081)^2 = 0{,}006561 \quad (4)$$
$$(7{,}09 - 6{,}991)^2 = (0{,}099)^2 = 0{,}009801 \quad (5)$$

b) Suma total de $(x - \bar{x})^2$:
$\sum(x - \bar{x})^2 = 0{,}7351$ (calculando para todas las 50 observaciones)
c) Varianza muestral:
$s^2 = \frac{0{,}7351}{50-1} = \frac{0{,}7351}{49} = 0{,}015002$
d) Desviación estándar:
$s = \sqrt{s^2} = \sqrt{0{,}015002} = 0{,}1225$

Interpretación de la desviación estándar:

La desviación estándar de 0.1225 indica una **variabilidad baja** en las mediciones de pH. Esto significa:

- La mayoría de los valores se encuentran dentro de $\pm 0{,}1225$ unidades de pH de la media
- El proceso de producción es consistente y bien controlado
- La precisión de las mediciones es buena
- Para una aplicación industrial, esta variabilidad es aceptable y sugiere un buen control de calidad

En el contexto químico, una variabilidad de $\pm 0{,}12$ unidades de pH es excelente para un proceso industrial, ya que permite mantener las propiedades de la solución buffer dentro de rangos funcionales.

4.2. Ejercicio 4.2: Coeficiente de variación (10 puntos)

Cálculo del coeficiente de variación:

Fórmula: $CV = \frac{s}{\bar{x}} \times 100\,\%$
a) Cálculo:
$CV = \frac{0{,}1225}{6{,}991} \times 100\,\% = 0{,}01752 \times 100\,\% = 1{,}752\,\%$
b) Clasificación de variabilidad:
Con CV = 1.75 %, la variabilidad se clasifica como **baja** ($< 15\,\%$).

Evaluación para control de calidad:

Un coeficiente de variación de 1.75 % es **excelente** para un proceso de control de calidad industrial por las siguientes razones:

1. **Estándar industrial:** Para procesos químicos, un CV $< 5\,\%$ se considera excelente
2. **Consistencia del proceso:** Indica un control muy estricto de las variables del proceso
3. **Confiabilidad del producto:** Los usuarios pueden esperar propiedades consistentes
4. **Eficiencia económica:** Menor desperdicio y reprocesamiento
5. **Cumplimiento regulatorio:** Facilita el cumplimiento de especificaciones técnicas

Recomendación: Esta variabilidad es completamente aceptable y el proceso no requiere ajustes inmediatos en términos de control de variabilidad.

5. Parte V: Análisis integral y toma de decisiones

5.1. Ejercicio 5.1: Control de calidad químico (15 puntos)

Análisis de rango óptimo:

a) Porcentaje dentro del rango óptimo (6.8 - 7.2):
Conteo de valores en el rango [6.8, 7.2]:

- Valores $< 6{,}8$: 0
- Valores en [6.8, 7.2]: 50
- Valores $> 7{,}2$: 0

Porcentaje $= \frac{50}{50} \times 100\,\% = 100\,\%$
b) Aplicación de la regla empírica (68-95-99.7):
Para el 95 % de las muestras: $\bar{x} \pm 2s$
Límites: $6{,}991 \pm 2(0{,}1225) = 6{,}991 \pm 0{,}245$
Rango esperado: [6.746, 7.236]

Recomendación sobre aprobación del lote:

RECOMENDACIÓN: APROBAR EL LOTE
Evidencia estadística que respalda la decisión:

1. **Cumplimiento total:** 100 % de las muestras están dentro del rango óptimo (6.8-7.2)
2. **Media centrada:** 6.991 está muy cerca del punto neutro (7.0)
3. **Variabilidad excelente:** CV = 1.75 % indica control superior del proceso
4. **Distribución simétrica:** No hay sesgo significativo en los datos
5. **Ausencia de outliers:** No hay valores anómalos que sugieran problemas
6. **Predictibilidad:** El 95 % de futuras muestras se esperaría en [6.746, 7.236]

Calificación del lote: EXCELENTE - Cumple y supera todos los criterios de calidad establecidos.

Propuestas para mejora del control de calidad:

a) Implementar control estadístico de procesos (SPC):

- Establecer gráficos de control con límites basados en estos datos
- Monitoreo continuo de media y variabilidad
- Alertas automáticas para desviaciones

b) Optimizar el proceso de muestreo:

- Aumentar frecuencia de muestreo en puntos críticos
- Implementar muestreo estratificado por lote de producción
- Validar representatividad de las muestras

c) Mejorar la documentación y trazabilidad:

- Registrar condiciones ambientales (temperatura, humedad)
- Documentar operadores y equipos utilizados
- Establecer sistema de calibración preventiva
- Crear base de datos histórica para análisis de tendencias

6. Conclusiones

Síntesis del análisis:

Este análisis exploratorio de datos del proceso de producción de solución buffer revela un proceso químico industrial de **calidad excepcional**. Los indicadores estadísticos demuestran:

- **Cumplimiento total** de especificaciones técnicas
- **Control superior** de la variabilidad del proceso
- **Consistencia excelente** en las mediciones
- **Ausencia de anomalías** que sugieran problemas sistemáticos

El lote analizado no solo cumple con los estándares de calidad establecidos, sino que sirve como **benchmark** para futuras producciones. La metodología de análisis exploratorio aplicada demuestra ser una herramienta valiosa para la toma de decisiones en control de calidad químico industrial.

Preparación para herramientas computacionales: El dominio de estos conceptos estadísticos fundamentales y la metodología de análisis exploratorio establecen las bases sólidas necesarias para la posterior implementación de estas técnicas usando herramientas computacionales como R o Python, que será el enfoque de las siguientes etapas del curso EIY403.

Este análisis demuestra la aplicación práctica de conceptos estadísticos fundamentales en el contexto de la química industrial, estableciendo las bases para el uso de herramientas computacionales en análisis de datos más complejos.