UNIVERSIDAD INTERNACIONAL DE LA RIOJA

MAESTRÍA EN ANÁLISIS DE DATOS Y BIG DATA


# Actividad 3: Uso avanzado de bases de datos NoSQL

**Autor:**
Julia Silvana Huayta Gómez

**Profesor:**
Yeray Mezquita
16 de febrero 2024

1. **Una breve descripción sobre qué son y para qué se utilizan los "contrastes de hipótesis".**

   Los contrastes de hipótesis son procedimientos estadísticos diseñados para evaluar afirmaciones o suposiciones sobre los parámetros de una población. Se utilizan para tomar decisiones informadas sobre afirmaciones basadas en evidencia muestral.

2. **Proponer un algoritmo general que describa cómo se hacen los contrastes de hipótesis a nivel de una población y para comparar dos poblaciones.**

   **Contraste de Hipótesis para una Población:**

   **Formulación de Hipótesis:**

   **Hipótesis Nula:** Define la afirmación de que se va a poner a prueba (por ejemplo, igualdad de medias, proporciones, etc.).

   **Hipótesis Alternativa:** Especifica la afirmación contraria a la hipótesis nula.

   Elección del Nivel de Significancia Decide el nivel de significancia, generalmente 0.05 o 0.01.

   **Recopilación de Datos:** Recoge una muestra representativa de la población.

   **Análisis de Datos:** Calcula el estadístico de contraste relevante (t, z, chi-cuadrado, etc.) basado en la muestra y la hipótesis nula.

   **Toma de Decisiones:** Compara el valor calculado del estadístico con el valor crítico o calcula el p-valor. Si el p-valor es menor que α, se rechaza la hipótesis nula.

   **Contraste de Hipótesis para Comparar Dos Poblaciones:**

   **Formulación de Hipótesis:**

   Define las hipótesis nula y alternativa para la comparación de dos poblaciones (por ejemplo, igualdad de medias, diferencia de proporciones, etc.).

   **Elección del Nivel de Significancia:**

   Decide el nivel de significancia.

   **Recopilación de Datos:**

   Recoge muestras representativas de ambas poblaciones.

   **Análisis de Datos:**

   Calcula el estadístico de contraste apropiado para la comparación de dos poblaciones (t, z, chi-cuadrado, etc.).

   **Toma de Decisiones:**

   Compara el valor calculado del estadístico con el valor crítico o calcula el p-valor.

   Si el p-valor es menor que α, se rechaza la hipótesis nula, lo que sugiere diferencias significativas entre las poblaciones.

3. **A continuación, elegir dos escenarios modelo para mostrar cómo puede usarse un contraste de hipótesis para buscar "respuesta" a un "problema" en el estudio de una población y en el estudio comparativo de dos poblaciones. Esos problemas deben estar relacionados con la vida diaria.**

   **Escenario Modelo 1: Estudio de una Población**

   Problema: Se quiere determinar si hay evidencia estadística para afirmar que la media de horas de sueño de adultos en una ciudad específica es diferente de la media nacional, que se establece en 7 horas.

   **Escenario Modelo 2: Estudio Comparativo de Dos Poblaciones**

   **Problema:** Se quiere determinar si hay diferencias significativas en las tasas de éxito de dos métodos de enseñanza de matemáticas en dos escuelas diferentes.

4. **Los modelos deben contener información como: (1) planteamiento del estudio y (2) definición del tipo de información que recoger para tener la base de datos apropiada en cada caso.**

   **Modelo 1: Estudio de una Población - Horas de Sueño en una Ciudad**

   **Planteamiento del Estudio:**

El objetivo es determinar si la media de horas de sueño de adultos en una ciudad específica difiere de la media nacional, que se establece en 7 horas. Esto tiene implicaciones para la salud y el bienestar de la población local.

**Definición del Tipo de Información a Recoger:**

**Variable a Medir:** Horas de sueño por noche.

**Muestra Representativa**: Seleccionar una muestra aleatoria de adultos en la ciudad.

**Recogida de Datos:** Encuestar a los participantes sobre la cantidad de horas de sueño que obtienen habitualmente.

**Otros Datos Demográficos:** Registrar información demográfica relevante (edad, género, ocupación) para controlar posibles variables confusas.

**Modelo 2: Estudio Comparativo de Dos Poblaciones - Métodos de Enseñanza de Matemáticas en Escuelas**

**Planteamiento del Estudio:**

Se busca determinar si hay diferencias significativas en las tasas de éxito de dos métodos de enseñanza de matemáticas en dos escuelas diferentes. Esto puede influir en la elección de métodos educativos más efectivos.

**Definición del Tipo de Información a Recoger:**

**Variable a Medir:** Tasa de éxito en matemáticas.

**Muestras Representativas:** Seleccionar muestras aleatorias de estudiantes de cada escuela.

**Recogida de Datos:** Aplicar un método específico de enseñanza de matemáticas a cada grupo y registrar las tasas de éxito.

**Datos Adicionales:** Registrar información sobre el rendimiento académico anterior de los estudiantes para controlar posibles variables confusas.

**5. Desarrollar modelos numéricos para estudiar las situaciones anteriores, atendiendo a las siguientes restricciones:**

- **Indica claramente los estudios que vas a realizar, es decir, los problemas que vas a investigar, y explica cómo elegirás la base de datos apropiada.**

**Estudio de Horas de Sueño en una Ciudad Específica**

**Estudio 1:**

**Problema:** Determinar si la media de horas de sueño de adultos en una ciudad específica es diferente de la media nacional (7 horas).

**Base de Datos Apropiada:**

Seleccionar una muestra aleatoria representativa de adultos en la ciudad.

Recoger información sobre las horas de sueño habituales por noche y datos demográficos relevantes (edad, género, ocupación).

**Estudio Comparativo de Métodos de Enseñanza de Matemáticas en Escuelas:**

**Estudio 2:**

Problema: Determinar si hay diferencias significativas en las tasas de éxito de dos métodos de enseñanza de matemáticas en dos escuelas diferentes.

**Base de Datos Apropiada:**

Seleccionar muestras aleatorias representativas de estudiantes de cada escuela.

Recoger información sobre la tasa de éxito en matemáticas y datos académicos previos de los estudiantes.

**En cada caso, propón una base de datos de al menos 1000 individuos (puedes obtenerla de fuentes reales o simular la base de datos con el método que elijas).**

**Dataset modelo 1:**

ID de persona: un identificador para cada individuo.
Género: El género de la persona (Hombre/Mujer).
Edad: La edad de la persona en años.
Ocupación: La ocupación o profesión de la persona.

Duración del sueño (horas): La cantidad de horas que la persona duerme por día.
Calidad del sueño (escala: 1-10): una calificación subjetiva de la calidad del sueño, que va del 1 al 10.
Nivel de actividad física (minutos/día): la cantidad de minutos que la persona realiza actividad física diariamente.
Nivel de estrés (escala: 1-10): una calificación subjetiva del nivel de estrés experimentado por la persona, que varía de 1 a 10.
Categoría de IMC: la categoría de IMC de la persona (por ejemplo, bajo peso, normal, sobrepeso).
Presión arterial (sistólica/diastólica): La medida de la presión arterial de la persona, indicada como presión sistólica sobre presión diastólica.
Frecuencia cardíaca (lpm): La frecuencia cardíaca en reposo de la persona en latidos por minuto.
Pasos diarios: el número de pasos que da la persona por día.
Trastorno del sueño: Presencia o ausencia de un trastorno del sueño en la persona (Ninguno, Insomnio, Apnea del sueño).

**Dataset modelo 2:**

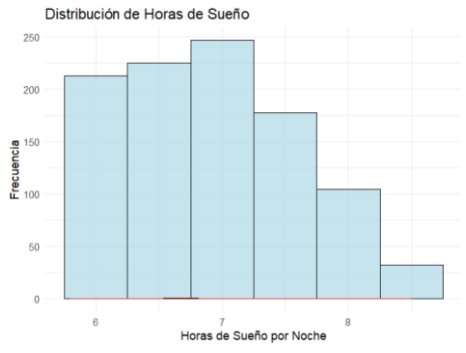
Estudiante: id de estudiante
Escuela: Dos instituciones educativas. Mi pequeño Gigante y Santa Rita
Metodo: Metodos de enseñanza de las Instituciones educativas. ABP (Aprendizaje basado en proyectos) y Competencias (Enseñanza basada en competencias)
Éxito previo: porcentaje de éxito en un examen anterior del alumno, las calificaciones son en un rango del 1 al 100
Éxito actual: porcentaje de éxito en el ultimo examen del alumno, las calificaciones son en un rango del 1 al 100

**Calcula y visualiza gráficamente los principales elementos de estadística descriptiva para cada uno de estos conjuntos de datos. Esto debes hacerlo con el software R.**

**Dataset modelo 1:**

```
summary(datos$Sleep.Duration)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
5.800   6.500   6.800   6.936   7.400   8.500
```

Se puede visualizar que el mínimo de horas que una persona duerme es 5.8, la media 6,94 y el máximo 8,5 horas.



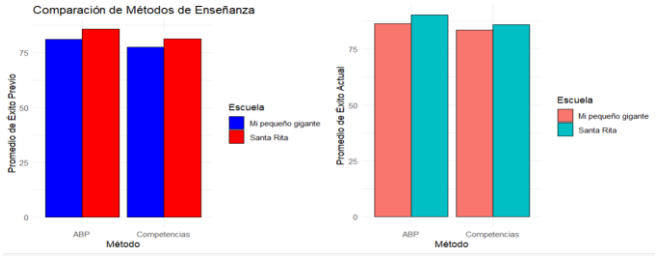

En el gráfico podemos interpretar que generalmente la mayoría de la población que se encuentra en nuestra muestra duerme en un rango de 7 horas aproximadamente y que son pocas las personas que duermen más de 8 horas.

**Dataset modelo 2:**

```
[1] "Estadísticas Descriptivas para ExitoPrevio:"
> print(stats_previo)
              Escuela       Metodo ExitoPrevio.Min. ExitoPrevio.1st Qu. ExitoPrevio.Median ExitoPrevio.Mean
1 Mi pequeño gigante          ABP         68.00000           78.00000          82.00000         81.12766
2         Santa Rita          ABP         68.00000           82.00000          85.50000         85.77083
3 Mi pequeño gigante Competencias         65.00000           72.00000          76.00000         77.62389
4         Santa Rita Competencias         65.00000           72.00000          80.00000         81.25794
  ExitoPrevio.3rd Qu. ExitoPrevio.Max.
1            85.00000          94.00000
2            90.00000          98.00000
3            81.00000          95.00000
4            88.00000          96.00000
`
[1] "Estadísticas Descriptivas para ExitoActual:"
> print(stats_actual)
              Escuela       Metodo ExitoActual.Min. ExitoActual.1st Qu. ExitoActual.Median ExitoActual.Mean
1 Mi pequeño gigante          ABP         76.00000           82.00000          87.00000         86.46454
2         Santa Rita          ABP         74.00000           86.00000          91.50000         90.34583
3 Mi pequeño gigante Competencias         73.00000           79.00000          82.50000         83.52212
4         Santa Rita Competencias         74.00000           76.00000          85.50000         86.02381
  ExitoActual.3rd Qu. ExitoActual.Max.
1            90.00000          98.00000
2            95.00000         100.00000
3            88.00000          98.00000
4            93.00000         100.00000
>
```

Podemos observar que el éxito de una calificación de un estudiante en la Institución Mi Pequeño Gigante antes de aplicar el método ABP en promedio es de 81.13 y después es 86,46 lo cual indica que hubo aumento.

Santa Rita obtuvo un resultado mayor aplicando el método competencias




Podemos visualizar que aplicar los métodos hace que el alumno mejore en su rendimiento

**Calcula, usando R, las cantidades que necesites para poder desarrollar los análisis orientados en 1.**

**Dataset modelo 1:**

Hipótesis Nula (H0): La media de las horas de sueño en la ciudad es igual a 7 horas.

Hipótesis Alternativa (H1): La media de las horas de sueño en la ciudad es diferente de 7 horas.

```
> print(resultado_prueba)

        One Sample t-test

data:  datos$Sleep.Duration
t = -3.0809, df = 999, p-value = 0.002121
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.895726 6.976874
sample estimates:
mean of x
   6.9363
```

Estadístico t (t-statistic): El valor es -3.0809. Indica cuántas desviaciones estándar está la

media observada de tus datos (6.9363) de la media hipotética que estás comparando (7 horas). Grados de libertad (df): Son 999, que es el tamaño de tu muestra menos 1 (n-1).

Valor p (p-value): Es 0.002121. Indica la probabilidad de obtener un estadístico t tan extremo o más extremo bajo la hipótesis nula de que la verdadera media es igual a 7. Un valor de p pequeño sugiere evidencia en contra de la hipótesis nula.

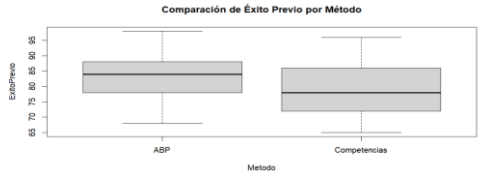
Intervalo de Confianza del 95%: Es [6.895726, 6.976874]. Esto indica el rango en el cual es probable que se encuentre la verdadera media de las horas de sueño en la población.

La interpretación final se basa en el valor p. En este caso, el valor p (0.002121) es menor que el nivel de significancia común de 0.05. Por lo tanto, hay evidencia estadística para rechazar la hipótesis nula de que la media de horas de sueño es igual a 7 horas. En otras palabras, la diferencia en las horas de sueño es estadísticamente significativa, y la media observada es significativamente diferente de la media nacional de 7 horas.

**Dataset Modelo 2:**

Hipótesis nula (H0): No hay diferencia significativa en el éxito previo entre los métodos de enseñanza ABP y Competencias en las dos instituciones.

Hipótesis alternativa (H1): Hay una diferencia significativa en el éxito previo entre los métodos de enseñanza ABP y Competencias en las dos instituciones.




```
        Welch Two Sample t-test

data:  Mejora by Metodo
t = -2.7908, df = 878.81, p-value = 0.005372
alternative hypothesis: true difference in means between group ABP a
nd group Competencias is not equal to 0
95 percent confidence interval:
 -0.5359582 -0.0933722
sample estimates:
        mean in group ABP mean in group Competencias
                 4.986590                  5.301255
```

El resultado del Welch Two Sample t-test proporciona información sobre si hay una diferencia significativa en la mejora entre los métodos de enseñanza ABP y Competencias. Aquí está la interpretación clave:

Estadístico t y grados de libertad (df):

t = -2.7908: Este es el valor del estadístico t para la prueba. En términos simples, cuanto más lejos esté este valor de cero, más evidencia hay en contra de la hipótesis nula.

df = 878.81: Estos son los grados de libertad ajustados según el método de Welch para acomodar posibles diferencias en las varianzas entre los grupos.

p-value:

p-value = 0.005372: Es el valor p. Si es menor que tu nivel de significancia (como comúnmente 0.05), puedes rechazar la hipótesis nula. En este caso, el p-valor es menor que 0.05, lo que sugiere que hay evidencia estadística para rechazar la hipótesis nula.

Intervalo de confianza (95%):

Intervalo de confianza: -0.536 a -0.093: Esto te da un rango plausible para la verdadera diferencia entre las medias de las mejoras en ABP y Competencias. Dado que no incluye cero, es consistente con la evidencia encontrada en el p-valor de que hay una diferencia significativa.

Estimaciones de la muestra:

Media en grupo ABP: 4.986590

Media en grupo Competencias: 5.301255

Estas son las medias estimadas de la mejora para cada grupo.

**Incluye una discusión relativa al tipo de conclusiones que obtienes e indica qué análisis has llevado a cabo para conseguirlo.**

**Dataset modelo 1:**

Análisis Realizado:

Se llevó a cabo un análisis de las horas de sueño en una ciudad específica, comparando la media observada con la media nacional establecida en 7 horas. El análisis incluyó una prueba de hipótesis utilizando la prueba t, donde la hipótesis nula planteaba que la media de sueño es igual a 7 horas,

y la alternativa sugería que la media es diferente de 7 horas.

Resultado del Análisis:

La prueba t arrojó un valor p de 0.002121, lo cual es menor que el nivel de significancia común de 0.05. Por lo tanto, se rechazó la hipótesis nula, indicando que la diferencia en las horas de sueño es estadísticamente significativa. La media observada en la muestra (6.9363 horas) es significativamente diferente de la media nacional.

Implicaciones:

Estos resultados tienen implicaciones para la salud y el bienestar de la población local. La diferencia estadísticamente significativa en las horas de sueño podría sugerir patrones de sueño distintivos en esta ciudad en comparación con la media nacional. Esto podría estar relacionado con factores ambientales, culturales o de estilo de vida que influencian las rutinas de sueño.

Conclusiones:

Diferencia Significativa: La prueba de hipótesis ha demostrado una diferencia estadísticamente significativa entre la media de horas de sueño en la ciudad específica y la media nacional de 7 horas. Esta diferencia, respaldada por el valor de p (0.002121), indica que las horas de sueño en la ciudad son distintas de la media nacional.

Media Observada: La media observada en la muestra es de aproximadamente 6.94 horas de sueño por noche, lo que sugiere que, en promedio, los adultos en la ciudad duermen menos de las 7 horas recomendadas a nivel nacional.

Discusión:

Implicaciones para la Salud: La diferencia significativa en las horas de sueño puede tener implicaciones para la salud de la población local. La falta de sueño crónica se ha asociado con diversos problemas de salud, desde dificultades cognitivas hasta riesgos cardiovasculares.

Factores Determinantes: La variabilidad en las horas de sueño podría estar influenciada por diversos factores, como el estilo de vida, las demandas laborales o culturales específicas de la ciudad. Sería beneficioso profundizar en estos aspectos para comprender mejor las razones detrás de esta diferencia.

Importancia de Investigaciones Adicionales: Este estudio proporciona una visión inicial, pero se necesita más investigación. Sería valioso explorar factores específicos que podrían contribuir a las diferencias en las horas de sueño, como el estrés laboral, el uso de tecnología antes de dormir u otros aspectos del estilo de vida.

Enfoque en Factores Demográficos: Futuros estudios podrían considerar el análisis de subgrupos demográficos para comprender si hay variaciones significativas en las horas de sueño entre diferentes grupos de edad, género u ocupación.

Recomendaciones de Intervención: Si la falta de sueño se asocia con riesgos para la salud, se podrían considerar intervenciones a nivel comunitario, como campañas de concientización sobre la importancia del sueño o programas para mejorar la calidad del sueño.

**Dataset modelo 2:**

**Conclusiones:**

Los estudiantes de la institución Santa Rita, en general, muestran un mayor éxito tanto previo como actual en comparación con los estudiantes de Mi pequeño gigante.

La enseñanza basada en proyectos (ABP) parece tener un impacto positivo en el éxito de los estudiantes, ya que, en promedio, los estudiantes que han experimentado esta metodología han tenido más éxito que aquellos que han recibido enseñanza basada en competencias.

La prueba de hipótesis sugiere que hay una diferencia significativa en el éxito entre los dos métodos de enseñanza (ABP y Competencias). El intervalo de confianza del 95% para la diferencia en medias no incluye cero, respaldando la evidencia de una diferencia real.

**Discusión:**

Aunque la diferencia es estadísticamente significativa, es crucial considerar el tamaño del efecto. El valor promedio de éxito actual para Competencias es ligeramente superior, pero la magnitud de esta diferencia puede no ser lo suficientemente grande como para tener implicaciones prácticas sustanciales.

La prueba t de Welch indicó que hay una diferencia estadísticamente significativa en el éxito entre estos dos métodos, respaldando la noción de que la elección del método de enseñanza puede afectar el rendimiento de los estudiantes.