UNIR México – Universidad Rioja de España

Dra. Mariana Edith Miranda Varela

4-8-2025

Estadística Inferencial

Análisis muestral a la predicción poblacional

Alumno: Leonard José Cuenca Roa

Tabla de contenido

1. Descripción Contrastes de Hipótesis

El contraste de hipótesis, tras exhaustivas revisiones y validaciones de la información en medios electrónicos y consultas bibliográficas, se puede considerar una herramienta fundamental en la inferencia estadística y un pilar esencial en el análisis de datos. Su origen se deriva de las pruebas por contradicción, ya que partimos de una suposición y calculamos la probabilidad de observar dichos datos bajo la premisa de que la suposición es cierta. Si esta probabilidad es muy reducida, concluimos que es improbable que nuestra hipótesis sea incorrecta. Esto nos permite tomar decisiones informadas sobre las características de una población basándonos en datos de una muestra. Su importancia reside en la capacidad de cuantificar la evidencia para respaldar o refutar una afirmación, en lugar de simplemente aceptar o rechazar ideas de forma intuitiva. Para su aplicación, es crucial dominar los siguientes puntos:

- **Hipótesis Nula (H0):** Es la premisa que se considera correcta por intuición, respaldada por cierta teoría o consolidada empíricamente.
- **Hipótesis Alternativa (H1 o Ha):** Es la premisa desafiante que rompe lo establecido; es la negativa de H0, y lo que el investigador intenta probar con estudios persistentes y cuidadosamente validados.
- **Nivel de Significancia (α):** Este concepto es fundamental en los contrastes, ya que se basa en los intervalos de confianza. Decidimos los márgenes que marcan los límites inferior y superior, y gracias a estos valores, determinamos si rechazamos o no la hipótesis nula.
- **Estadístico de Prueba:** Es un valor numérico calculado a partir de los datos de la muestra. Su distribución muestral debe ser conocida bajo la suposición de que la hipótesis nula es verdadera.
- **Región de Rechazo:** Es el conjunto de valores del estadístico de prueba que son tan extremos o improbables bajo la hipótesis nula que nos llevarían a rechazarla. Los límites de esta región se definen en función del nivel de significancia (α).
- **Valor p (p-value):** Es una medida de la fuerza de la evidencia en contra de la hipótesis nula; un valor p pequeño indica una fuerte evidencia para rechazar H0.
- **Error Tipo I (α):** Este error se produce cuando se rechaza la hipótesis nula (H0) siendo esta verdadera. Metafóricamente, es como condenar a un inocente.
- **Error Tipo II (β):** Este error se produce cuando no se rechaza la hipótesis nula (H0) siendo esta falsa. Metafóricamente, es como dejar libre a un culpable.
- **Potencia del Contraste (1−β):** Es la probabilidad de rechazar correctamente una hipótesis nula falsa. Representa la capacidad de la prueba para detectar un efecto real cuando este existe; un buen diseño experimental busca maximizar esta potencia.

Es de mucha importancia mencionar los tipos de contrastes, tales como:

- **Contrastes de hipótesis para una media:** Permite evaluar una afirmación sobre el valor promedio (media, μ) de una sola población. Partimos de una hipótesis sobre cuál creemos que es el valor de la media poblacional y, a través de una muestra, determinamos si hay suficiente evidencia para rechazar esa hipótesis.
- **Contrastes de hipótesis para la proporción:** Se utiliza para evaluar una afirmación sobre el porcentaje de individuos en una población que poseen una característica particular. Es útil cuando la variable de interés es categórica binaria.
- **Contrastes de hipótesis sobre la varianza:** Se emplea para evaluar una afirmación sobre la dispersión de los datos en una población.

- **Contrastes paramétricos para dos muestras:** Se utilizan para comparar las características de dos poblaciones diferentes o el efecto de dos tratamientos distintos, basándose en la información de dos muestras.
- **Contrastes de hipótesis robustos:** Son métodos diseñados para ser menos sensibles a las violaciones de las suposiciones subyacentes de los contrastes paramétricos tradicionales, usándose cuando los datos no cumplen los supuestos de normalidad o cuando hay valores atípicos significativos.

En este mundo tan globalizado, los datos son la nueva fiebre del oro. Gracias a la abundancia de estos, facilitada por la economización del almacenamiento de datos, las grandes y pequeñas industrias pueden almacenar mucha data en bruto. Posteriormente, aplican análisis para lograr una mejora en los procesos o una temprana y crucial toma de decisiones que ayude a guiar a los conglomerados en un buen camino y evitar o prevenir un rumbo perjudicial.

El contraste de hipótesis, como ya se mencionó, es una herramienta clave del presente. Podemos mencionar algunas áreas cruciales donde se está utilizando:

- **Medicina y Salud Pública:** Crucial para validar las nuevas o viejas técnicas de la medicina, también en el área de la farmacología.
- **Negocios y Marketing:** El área más explotada sin duda alguna; grandes comercios usan el comportamiento de compra y de preferencias para validar hipótesis de gustos, modas y tendencias.
- **Manufactura y Control de Calidad:** Para validar procesos operacionales, medir la calidad y la efectividad de los procesos.
- **Ciencias Sociales:** Otra área que explota el análisis de hipótesis, ya que ayuda en el desarrollo, creación y validación de proyectos sociales.

Para concluir, el contraste de hipótesis es, sin duda alguna, una de las tantas herramientas que, al ser dominada, desbloquea una gran habilidad para resolver situaciones que ayuden a aportar un granito de arena en cualquier área

2. Algoritmo Contrastes de Hipótesis

De acuerdo con la recopilación, validación y analisis procedo a crear un modelo algoritmico y de pasos a seguir para generar un tipo de contrastación de hipotesis, puedo describir que un contraste de hipótesis sigue una serie de pasos lógicos que permiten evaluar una afirmación sobre una población o comparar dos poblaciones basándose en datos muestrales.

Paso 1: Formulación de las Hipótesis

Iniciamos con el Objetivo:

Objetivo: Determinar si la evidencia muestral es suficientemente fuerte para rechazar una afirmación inicial sobre una o más poblaciones.

Este es el punto de partida de cualquier contraste. Se establecen dos afirmaciones opuestas sobre el parámetro de interés (media, proporción, varianza, etc.).

- **1.1. Hipótesis Nula (H0):**

    - **Definición:** Representa la afirmación del "status quo", la no diferencia o no efecto. Es lo que se asume como cierto hasta que la evidencia demuestre lo contrario. Siempre incluye un signo de igualdad (=,≤,≥).
    - **Ejemplo para una población:** "H0: El peso promedio de las bolsas de café es de 250 gramos." (μ=250).
    - **Ejemplo para dos poblaciones:** "H0: No hay diferencia en la efectividad promedio de dos tratamientos médicos." (μ1=μ2).
- **1.2. Hipótesis Alternativa (H1 o Ha):**
    - **Definición:** Es la negación de la hipótesis nula y la afirmación que el investigador busca probar o encontrar evidencia a favor. Puede ser unilateral (mayor que, menor que) o bilateral (diferente de).
    - **Ejemplo para una población:** "H1: El peso promedio de las bolsas de café no es de 250 gramos." (μ□=250) o "El peso promedio de las bolsas de café es menor de 250 gramos." (μ<250).
    - **Ejemplo para dos poblaciones:** "H1: Existe una diferencia en la efectividad promedio de dos tratamientos médicos." (μ1□=μ2) o "El tratamiento A es más efectivo que el tratamiento B." (μA >μB).

Paso 2: Selección del Nivel de Significancia (α)

Se define la probabilidad máxima de cometer un **Error Tipo I** (rechazar H0 cuando es verdadera). Este valor se establece *antes* de recolectar y analizar los datos. Los valores comunes son 0.05 (5%) o 0.01 (1%).

- **Importancia:** Un α más pequeño hace que la prueba sea más estricta, requiriendo evidencia más fuerte para rechazar H0.

Paso 3: Selección del Estadístico de Prueba

Se elige la fórmula estadística adecuada que se utilizará para calcular un valor a partir de los datos de la muestra. La elección depende del tipo de datos (cuantitativos o cualitativos), el número de poblaciones, el conocimiento de los parámetros poblacionales (e.g., varianza), y el cumplimiento de ciertos supuestos.

- **Para una población:**
    - Para medias: Z-test (si σ conocida o n grande) o T-test (si σ desconocida o n pequeña).
    - Para proporciones: Z-test para proporciones.
    - Para varianzas: Chi-cuadrado (χ2).
- **Para dos poblaciones:**
    - Para medias: T-test para dos muestras independientes (si son grupos distintos) o T-test para muestras pareadas (si son las mismas unidades medidas dos veces).
    - Para proporciones: Z-test para dos proporciones.
    - Para varianzas: F-test (para comparar la igualdad de varianzas).

Paso 4: Recolección de Datos y Cálculo del Estadístico de Prueba

Se toma una muestra representativa de la población (o poblaciones) y se recopilan los datos necesarios. Con estos datos, se calcula el valor del estadístico de prueba seleccionado en el Paso 3.

- **Importancia en Big Data:** Aquí es donde la capacidad de manejar y procesar grandes volúmenes de datos de manera eficiente (limpieza, transformación y muestreo si es necesario) es crucial para obtener una muestra representativa y precisa.

Paso 5: Toma de Decisión (Comparación con Valor P o Región Crítica)

Este paso crucial implica comparar el resultado de nuestro cálculo con el criterio de decisión establecido por el nivel de significancia.

- **Opción A: Uso del Valor p (p-value):**
  - **Definición:** El valor p es la probabilidad de observar un resultado muestral tan extremo o más extremo que el obtenido, *si la hipótesis nula (H0) fuera verdadera*.
  - **Regla de Decisión:**
    - Si **Valor p ≤α**: Se **rechaza la Hipótesis Nula (H0)**. Hay suficiente evidencia estadística para apoyar la hipótesis alternativa.
    - Si **Valor p >α**: **No se rechaza la Hipótesis Nula (H0)**. No hay suficiente evidencia estadística para rechazarla. (Esto no significa que H0 sea verdadera, solo que los datos no la contradicen fuertemente).
  - **Ventaja:** El p-valor proporciona una medida continua de la evidencia en contra de H0, lo que facilita la interpretación.
- **Opción B: Uso de la Región Crítica (o de Rechazo):**
  - **Definición:** La región crítica es el rango de valores del estadístico de prueba que llevarían al rechazo de H0. Sus límites (valores críticos) se determinan a partir de la distribución del estadístico de prueba y el nivel α.
  - **Regla de Decisión:**
    - Si el valor del Estadístico de Prueba **cae dentro de la Región Crítica**: Se **rechaza la Hipótesis Nula (H0)**.
    - Si el valor del Estadístico de Prueba **cae fuera de la Región Crítica**: **No se rechaza la Hipótesis Nula (H0)**.

Paso 6: Conclusión e Interpretación

Se formula una conclusión clara en el contexto del problema original, evitando la jerga estadística excesiva. Esta conclusión debe responder a la pregunta de investigación inicial.

3. Elección de escenario modelo, planteamiento del estudio y definición del tipo de información

Escenario A

Problema en la Vida Diaria (Planteamiento del Estudio):

Soy un analista de ventas en la plataforma de Mercado Libre para un Cliente Marca NUBE y administro los productos de bañeras para bebés. He notado que en el mes de julio, específicamente a partir del día 19, las ventas diarias de la bañera color Rosa han disminuido notablemente en comparación con los días anteriores de julio. Sospecho que esto podría estar relacionado con ajustes de precios de la competencia.

Pregunta de Investigación:

¿Existe una disminución estadísticamente significativa en el promedio de ventas diarias de la bañera color Rosa a partir del 19 de julio, en comparación con el promedio de ventas diarias de la misma bañera antes del 19 de julio?

Tipo de Contraste:

Contraste de hipótesis paramétrico para dos medias (muestras independientes). Esto porque estás comparando el promedio de ventas de dos períodos distintos para el mismo producto (rosa), antes y después de una fecha clave.

Unidad de Observación:

La unidad de observación **ventas diarias por producto**. Ya que mi reporte representa las ventas de un color específico en un día.

Variable de Interés:

- total_ventas_diarias_por_color (cuantitativa, medida en unidades vendidas por día). Esto es lo que actualmente tienes en tu tabla y lo que se puede promediar.

Variables Adicionales:

- titulo_publicacion: (categórica) "Bañera TINA De Baño Para Bebe Plegable Portatil Casa Y Viaje", seguida del color (Rosa, Azul, etc.).
- fecha_venta: (fecha) La fecha del día de la venta (ej. 01/07/2025).
- ventas: El número de unidades vendidas para ese titulo_publicacion en esa fecha_venta.
- color_producto: Puedes derivar esta variable de titulo_publicacion (ej. "Rosa", "Azul", "Celeste").

Método de recolección:

La información se recopilará directamente de los reportes de ventas diarios exportados de la plataforma de Mercado Libre.

Escenario B

Problema en la Vida Diaria (Planteamiento del Estudio):

Soy un analista de ventas en la plataforma de Mercado Libre para un Cliente Marca NUBE y administro los productos de bañeras para bebés mi equipo tiene como objetivo principal mantener un alto nivel de satisfacción del cliente, medido a través de una encuesta Post Venta posterior al servicio. Históricamente, se ha asumido que al menos el 85% de los clientes están "Satisfechos" o "Muy Satisfechos" con la atención recibida.

Pregunta de Investigación:

¿La proporción actual de clientes "Satisfechos" o "Muy Satisfechos" en el servicio de atención al cliente de Mercado Libre es inferior al 85%?

Tipo de Contraste:

Contraste de hipótesis para una proporción.

Unidad de Observación:

Cada cliente que ha completado la encuesta de satisfacción después de recibir atención

Variable de Interés:

- satisfaccion_cliente (categórica binaria, con dos posibles resultados: "Satisfecho/Muy Satisfecho" o "No Satisfecho"). Esta se derivaría de las respuestas a la encuesta.

Variables Adicionales:

- id_cliente: Identificador único del cliente.
- fecha_interaccion: Fecha y hora de la interacción con el servicio.
- canal_atencion: Canal por el cual el cliente fue atendido (ej. chat, teléfono, email).

- tipo_problema: Categoría del problema que el cliente consultó (ej. envío, devolución, producto, cuenta).
- puntuacion_encuesta: La puntuación original de la encuesta (ej. escala del 1 al 5, donde 5 es muy satisfecho).

Método de recolección:

Los datos se recopilarían de los registros de las encuestas de satisfacción post-servicio. Cada vez que un cliente completa una encuesta, su respuesta se almacenaría. Para el estudio, se tomaría una muestra aleatoria de las encuestas completadas en un período reciente.

4.- Desarrollar modelos numéricos

Planteamiento del Problema:

Como gestor de operaciones en Mercado Libre, necesitamos verificar si la proporción de ventas entregadas cumple con el estándar del 98%.

NOTA: Para este ejercicio, he logrado obtener y compartir, con la validación de mis superiores y con fines educativos, un conjunto de datos que representa valores de un universo real. Utilizaré esta información para aplicar y formular mi hipótesis. A diferencia de las secciones anteriores, donde los planteamientos eran reales, pero no fue posible emplear datos auténticos para desarrollar las hipótesis, en este caso sí dispondré de información verídica.

Pregunta de Investigación:

¿La proporción actual de ventas con estado 'Entregado' es inferior al 98%?

Hipótesis Estadísticas:

- H_0: $p \geq 0.98$ (La proporción de ventas entregadas es al menos 98%)
- H_1: $p < 0.98$ (La proporción de ventas entregadas es menor al 98%)

Tipo de Contraste:

- Unilateral izquierdo: Justifico usar este tipo de contraste ya que mi hipótesis alternativa (H1) especifica que el parámetro poblacional de interés es **menor que** un valor hipotético.

Nivel de Significancia:

- $\alpha = 0.05$ (95% de confianza)

Datos Resaltantes del ejercicio

Características de los Datos:	**Pruebas Z para Una Proporción**	**Justificación prueba Z**
• Total, de ventas: 9198 • Ventas entregadas: 8586 • Ventas no entregadas: 612 • Proporción observada: 0.9335 (93.35%)	• Estamos probando una proporción poblacional • Tenemos una muestra suficientemente grande • Se debe cumplir: $n \cdot p_0 \geq 5$ y $n \cdot (1-p_0) \geq 5$	• Estamos probando una proporción poblacional • Tenemos una muestra suficientemente grande • Se debe cumplir: $n \cdot p_0 \geq 5$ y $n \cdot (1-p_0) \geq 5$

Verificación de Condiciones	Fórmula empleada	Cálculos Estadísticos
$n \cdot p_0 = 9198 \times 0.98 = 9014.0$ ✅ ≥ 5 $n \cdot (1-p_0) = 9198 \times 0.020000000000000018 = 184.0$ ✅ ≥ 5	• $Z = (\hat{p} - p_0) / \sqrt{[p_0(1-p_0)/n]}$ Donde: • $\hat{p} = 0.9335$ (proporción observada) • $p_0 = 0.98$ (proporción hipotética) • $n = 9198$ (tamaño de muestra)	Error estándar: $\sqrt{[0.98 \times 0.020000000000000018/9198]} = 0.001460$ Estadístico Z: $(0.9335 - 0.98) / 0.001460 = -31.8794$ Valor crítico $Z_{0.05} = -1.6449$ P-value = $P(Z < -31.8794) = 0.000000$

Resultado P-value	Comparación con Nivel Significación	Decisión Estadística
P-VALUE EXACTO: 0.000000	• Nivel de significancia empleado: $\alpha = 0.05$ • P-value = 0.000000 < $\alpha = 0.05$ • Criterio adicional: $Z = -31.8794 < Z_{0.05} = -1.6449$	- SE RECHAZA la hipótesis nula (H_0)
CRITERIOS DE DECISIÓN		
• Como p-value = 0.000000 < $\alpha = 0.05$ • También: $Z = -31.8794 < Z_{0.05} = -1.6449$ • Conclusión: Se rechaza H_0 al nivel de significancia del 5%		

Conclusiones

Puedo llegar a la siguiente conclusión con base en la evidencia estadística analizada de 9198 ventas, se encontró evidencia SUFICIENTE para concluir que la proporción actual de ventas entregadas (93.35%) es significativamente menor al estándar requerido del 98%.

El impacto que puede generar a las operaciones son las siguientes:

- El proceso de entregas NO está cumpliendo con el KPI establecido
- Se requiere intervención INMEDIATA en la operación logística
- Existe riesgo de impacto negativo en la satisfacción del cliente
- Es necesario implementar acciones correctivas URGENTES

Puedo sugerir las siguientes recomendaciones para el equipo luego del análisis se debe realizar una auditoría completa del proceso de entregas para identificar los puntos de falla en la cadena logística e implementar mejoras operativas para alcanzar el estándar del 98%. Es importante resaltar que este estudio es con razones académicas y se basó en una muestra de 9198 ventas, utilizando un contraste de hipótesis unilateral izquierdo con un nivel de confianza del 95% ($\alpha = 0.05$).

Evidencias

A continuación, se dejan los enlaces correspondientes para que se pueda validar los análisis y fuente de datos.

✓ Algoritmo Del Análisis Jupyter Clic Enlace Data Set de Estudio Clic Enlace