TP 4: Test

Vamos a resolver una versión levemente simplificada del proceso de test que usa Netflix descrito en este blog (Febrero 2024): link al blog.

Contexto

Netflix realiza actualizaciones constantes en su plataforma y necesita asegurar que éstas no perjudiquen la experiencia del usuario. Una métrica clave es el "play-delay", el tiempo entre que un usuario presiona "play" y el inicio de la reproducción.

Para prevenir que nuevas actualizaciones aumenten el "play-delay", Netflix usa un criterio de control de calidad: la actualización inicialmente se despliega a un grupo reducido de usuarios ($n = 200$) y, si se sospecha que la actualización aumenta el "play-delay", se la rechaza y se la devuelve al equipo de desarrolladores para que la auditen o la vuelvan a construir.

La variabilidad en el "play-delay" entre distintos usuarios depende de muchos factores: el dispositivo, la conectividad, el horario del día, etc. Disponemos de datos históricos del "play-delay" con la versión anterior (decenas de miles de observaciones). Esto nos permite afirmar que la distribución de delays (de la versión anterior) es una normal.

Requisitos del Test (citados del blog de Netflix)

1. Para minimizar el daño a los usuarios, si hay algún problema con la calidad de transmisión experimentada por los usuarios en el grupo de tratamiento (el grupo de usuarios a los que se les muestra la nueva versión), necesitamos abortar la prueba y revertir el cambio de software lo antes posible.

2. Este sistema es parte de un proceso semi-automatizado. Una prueba con un falso positivo (es decir, una prueba en la que se decide erróneamente que incrementó el "play-delay") interrumpe innecesariamente el proceso de lanzamiento de software, reduciendo la velocidad de entrega y haciendo que los desarrolladores busquen errores que no existen.

Ejercicios

1 Datos históricos

1.1

Grafique en un histograma los datos históricos de "play-delay". ¿Parecen distribuirse de forma Normal? Reporte la media y la varianza de los datos.

2 Grupo de prueba (nueva versión)

Consideramos que tenemos suficientes datos de la versión anterior (datos históricos) como para pensar que tenemos a toda la población. Con lo cual podemos asumir que la distribución de "play-delay" histórica es una $\mathcal{N}(\mu_0, \sigma_0^2)$ con μ_0 y σ_0^2 la media y la varianza calculadas a partir de los datos en el ítem anterior.
Llamaremos $X_1, \ldots, X_{200}$ al "play-delay" de los nuevos 200 usuarios evaluados (a los que se les mostró la nueva versión).
Asumimos que $X_1, \ldots, X_{200} \sim \mathcal{N}(\mu, \sigma_0^2)$, donde no necesariamente $\mu = \mu_0$, pero sí se mantiene σ_0^2.

2.1

Usando los datos de los nuevos 200 usuarios, estime μ, la esperanza del "play-delay" de la actualización.

3 Construir el test

3.1

Escriba la hipótesis nula H_0 y la alternativa H_1, entendiendo el contexto del problema. Recuerde que queremos mantener acotada la probabilidad de decidir que la actualización aumenta el "play-delay" cuando en realidad esto no es así. Aclare cuál es el estadístico utilizado en este test.

3.2

¿Cuál es la región de rechazo para este test, para $\alpha = 0.05$?

4 Usar el test

4.1

Explique cómo utilizar el test que ha construido. ¿Qué decisión se toma basándose en las nuevas 200 observaciones? ¿Se debe enviar el código para revisión?

4.2

¿Cómo es la región de rechazo con $\alpha = 0.01$? Compárela con la región calculada en 3.2. Haga lo mismo para $\alpha = 0.1$. ¿Cómo afecta α a la región de rechazo? ¿A medida que modifico α, mando más o menos seguido actualizaciones a auditar?

4.3

Evalúe qué decisión tomaría con el test diseñado, con cada uno de los siguientes alfas: $\alpha = \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10\}$.

4.4

Encuentre el valor mínimo de α (utilizando más decimales) para el cual se rechaza H_0 con los datos observados. Explique qué representa este valor en el contexto del problema.

5 Simulaciones del error Tipo 1

5.1

Simule una nueva muestra de tamaño $n = 200$ siguiendo la distribución $\mathcal{N}(\mu_0, \sigma_0^2)$, asumiendo que la hipótesis nula H_0 es verdadera.

Aplique el test estadístico con un nivel de significación $\alpha = 0.05$ a esta muestra simulada. ¿Qué decisión se toma para esta muestra? ¿Es la decisión correcta, considerando que H_0 es verdadera?

5.2

Repita el procedimiento anterior $R = 10000$ veces. ¿Qué porcentaje de las veces tomó la decisión correcta? ¿Qué porcentaje de las veces esperaba tomar la decisión correcta?

5.3

¿Cuál es la interpretación teórica de α?

6 Dos perspectivas sobre el p-valor

- El p-valor es el menor nivel de significación α para el que rechazamos H_0 para los datos observados.

- El p-valor es la probabilidad de obtener, asumiendo H_0, el valor observado o uno más extremo aún. En nuestro caso, $P(\bar{X} \geq \bar{x}_{obs}|H_0$ es Verdadera).

6.1

¿Cuál es la probabilidad de observar un resultado como el observado o más extremo aún, asumiendo H_0 verdadera?

6.2

Compare ese valor con el obtenido en el inciso 4.4) ¿Qué observa?