Análisis de entrada

Carlos Javier Uribe Martes

Ingeniería Industrial
Universidad de la Costa

Marzo 24, 2020

Contenido

Análisis de entrada

- Para llevar a cabo una simulación utilizando entradas aleatorias debemos especificar sus distribuciones de probabilidad [2].
- Seleccionar distribuciones de probabilidad adecuadas es una tarea importante y que requiere tiempo y buenos análisis estadísticos. [1].

Análisis de entrada

La metodología para conducir un Análisis de entrada incluye los siguientes pasos:

1. Recolección de datos.
2. Análisis de datos.
3. Modelado de datos.
4. Pruebas de bondad de ajuste.

Recolección de datos

- Aún teniendo un modelo válido, si los datos se recogen de una manera inadecuada, o son analizados incorrectamente, los resultados del modelo serán erróneos y pueden conducir a malas decisiones.
- Hay riesgos derivados de la escasez de datos disponibles, o de datos irrelevantes, desactualizados o simplemente erróneos.

Recolección de datos

- Algunas sugerencias para la recolección de datos son:
  - Planeación: observación del sistema actual y situaciones atípicas.
  - Análisis de los datos a medida que son recolectados.
  - Verificar homogeneidad en los diferentes grupos de datos.
  - Revisar la relación entre variables.
  - Revisar autocorrelación.
  - Diferenciar claramente entre datos de entrada y de salida.

Modelos de entrada sin datos

- Si el sistema a modelar No existe, el analista debe confiar en datos más vagos, que pueden incluir:
  1. Experiencias previas.
  2. Opinión de expertos.
  3. Intuición.
  4. Conjeturas a partir de las limitaciones físicas, datos de ingeniería, estándares o la naturaleza del proceso.

Modelos de entrada sin datos

- En muchas aplicaciones de la vida real, se utilizan heurísticas como:
  - Variables aleatorias con poca variabilidad se simplifican y modelan como deterministas.
  - Para distribuciones desconocidas se postulan forma funcionales particulares que incorporen cualquier información disponible (Triangular, Uniforme).
  - La experiencia a veces puede proporcionar información sobre la forma funcional de las distribuciones.

Modelos de entrada con datos

- Si el sistema a modelar ya existe, entonces puede proporcionar los datos empíricos necesarios a partir de mediciones en campo.

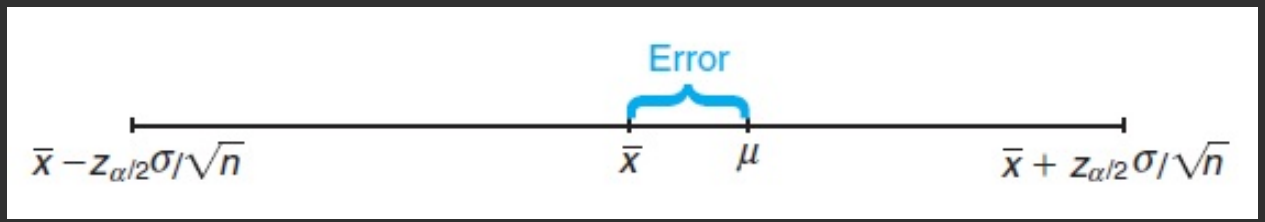
¿Qué datos se requieren tomar?

- Dentro de los datos de entrada a recolectar se requieren aquellos relacionados con tiempos entre llegadas, tiempos de servicio, tiempo a la falla de máquinas/recursos, duración de las fallas, tiempos de desplazamiento, probabilidades de tener ciertos valores de atributos (tipos de cliente, cantidad de demanda).
- La recolección de datos de las medidas de desempeño del sistema en estudio también es esencial para la validación del modelo.

¿Cuántos datos se requiere tomar?

- Si $\bar{x}$ es un estimador de μ, se puede tener un $100\,(1-\alpha)\,\%$ de confianza de que el error no excederá una cantidad específica e cuando el tamaño de muestra sea:

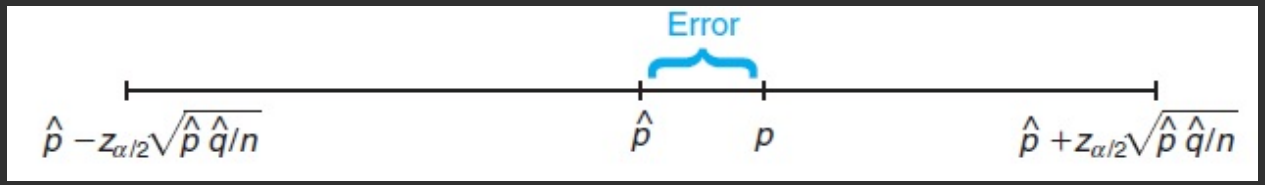
$$n = \left(\frac{z_{\alpha/2} \times \sigma}{e} \right)^2$$



¿Cuántos datos se requiere tomar?

- Si $\hat{p}$ es un estimador de p, se puede tener un $100\,(1-\alpha)\,\%$ de confianza de que el error no excederá una cantidad específica e cuando el tamaño de muestra sea:

$$n = \left(\frac{z^2_{\alpha/2} \times \hat{p} \times \hat{q}}{e^2} \right)$$



Análisis de datos

- El análisis de datos implica el cálculo de varias estadísticas a partir de los datos recopilados:
  - Estadísticas relacionadas con los momentos (media, desviación estándar, coeficiente de variación, etc.).
  - Estadísticas relacionadas con distribuciones (histogramas, q-q plot, p-p plot).
  - Estadísticas relacionadas con la dependencia temporal (autocorrelaciones dentro de una serie temporal empírica o correlaciones cruzadas entre dos o más series temporales distintas).

Estimación de media y varianza muestral

- Si las observaciones en una muestra de tamaño n son $x_1, x_2, \ldots, x_n$, la media muestral y la varianza muestral están dadas por:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}{n-1}$$

Estimadores de estadísticas para distribuciones comunes

Distribución	Parámetros	Estimadores
Poisson	α	$\hat{\alpha} = \bar{X}$
Exponencial	λ	$\hat{\lambda} = \frac{1}{\bar{X}}$
Gamma	β, θ	$\hat{\beta}$ ver Tabla estimadores de máxima verosimilitud $\hat{\theta} = \frac{1}{\bar{X}}$
Normal	μ, σ^2	$\hat{\mu} = \bar{X}$ $\hat{\sigma}^2 = S^2$
Lognormal	μ, σ^2	$\hat{\mu} = \bar{X}$ luego de sacar $\ln$ a los datos $\hat{\sigma}^2 = S^2$ luego de sacar $\ln$ a los datos

Histogramas

- Son útiles para la identificación de la forma de una distribución.
  - El número de clases, k depende del número de observaciones n y de la dispersión de los datos. Puede usarse:

$$k = \sqrt{n}$$

o

$$k = 1 + 3{,}322 \log_{10} n$$

  - Si los intervalos son muy anchos el histograma no mostrará un comportamiento claramente.

Q-Q plot

- Sea X una variable aleatoria con función acumulada de probabilidad $F_x(X)$, entonces el q-cuantil de X es aquel valor γ tal que $F_x(X) = P(X \leq \gamma) = q$. Si F tiene inversa entonces $\gamma = F^{-1}(q)$.
- Para realizar una Q-Q plot se emplea el siguiente algoritmo:
  1. Tomar una muestra de los datos $x_i,\ i = 1, 2, \ldots, n$ y ordenarlos para obtener $y_j,\ j = 1, 2, \ldots, n$.
  2. y_j es una estimación del $\left[\frac{j-\left(\frac{1}{2}\right)}{n}\right]$ cuantil de X. Esto es $y_j \sim F^{-1}\left[\frac{j-\left(\frac{1}{2}\right)}{n}\right]$.
  3. Graficar y_j vs $F^{-1}\left[\frac{j-\left(\frac{1}{2}\right)}{n}\right]$
  4. Si los datos corresponden a la distribución que se está probando, la gráfica debe ser aproximadamente una línea recta.

Q-Q plot

Algunas consideraciones al realizar una Q-Q plot:

- Nunca es realmente una línea recta.
- Un punto encima de la línea será probablemente seguido por otro.
- La variación en los extremos es más grande. La linealidad en el centro es más importante que la linealidad en los extremos.

P-P plot

- Utiliza la probabilidad acumulada para verificar si una muestra de datos sigue una distribución de probabilidad en particular, mediante el siguiente algoritmo:
  1. Tomar una muestra de los datos $x_i,\ i = 1, 2, \ldots, n$ y ordenarlos para obtener $y_j,\ j = 1, 2, \ldots, n$.
  2. Para cada valor de la muestra calcular $q_j = \left[\frac{j-\left(\frac{1}{2}\right)}{n}\right]$.
  3. Graficar q_j vs $F_x(y_j)$.
  4. Si los datos corrsponden a la distribución que se está probando, la gráfica debe ser aproximadamente un línea recta.

Diferencias entre Q-Q plot y P-P plot

- Un P-P plot compara la función de probabilidad acumulada de una muestra de datos con una función de probabilidad específica $F(\cdot)$, mientras que un Q-Q plot compara los cuantiles estimados dada una función de probabilidad con una muestra de datos.
- El rango de un P-P plot siempre es entre 0 y 1, el rango del Q-Q plot depende del rango de la función de probabilidad y de los datos observados.
- Un Q-Q plot amplifica las diferencias existentes en las colas del gráfico, mientras que un P-P plot amplifica las diferencias en el centro.

Modelado de datos

- En esta etapa, un modelo probabilístico es ajustado a las series de tiempo empíricas recolectadas.
- Dependiendo del tipo de datos de series de tiempo que se van a modelar, esta etapa se puede clasificar en dos categorías:
  1. Las observaciones independientes se modelan como una secuencia de variables aleatorias iid. En este caso, se busca identificar (ajustar) una distribución y sus parámetros a los datos empíricos.
  2. Las observaciones dependientes se modelan como procesos aleatorios con dependencia temporal. En este caso, se requiere identificar (ajustar) una ley de probabilidad a los datos empíricos.

Identificación de distribuciones de probabilidad

- Existen literalmente cientos de distribuciones de probabilidad.
- Algunas distribuciones aparecen muy a menudo en estudios de simulación:
  - Binomial, Poisson, Normal, Lognormal, Exponencial, Gamma, Beta, Erlang, Weibull, Uniforme, Triangular, ...

Pruebas de bondad de ajuste

- Las pruebas de bondad de ajuste son pruebas de hipótesis para verificar si los datos observados en una muestra aleatoria se ajustan con algún nivel de significancia a determinada distribución de probabilidad.
- Las hipótesis son:
  - H_0: la variable aleatoria X sigue una distribución asumida con los parámetros estimados.
  - H_1: la variable aleatoria X no sigue la distribución asumida.
- Para realizar la prueba, se clasifican los datos observados en k clases y se contabiliza el número de observaciones en cada clase, posteriormente se compara la frecuencia observada en cada clase con la frecuencia esperada en esa clase si la hipótesis nula es correcta.

Prueba chi-cuadrado

- Considere $k > 2$ el núméro de clases, O_i la frecuencia observada en la clase i, E_i la frecuencia esperada en la clase i si H_0 es correcta.
- La prueba se basa en el estadístico de prueba chi-cuadrado:

$$Y = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

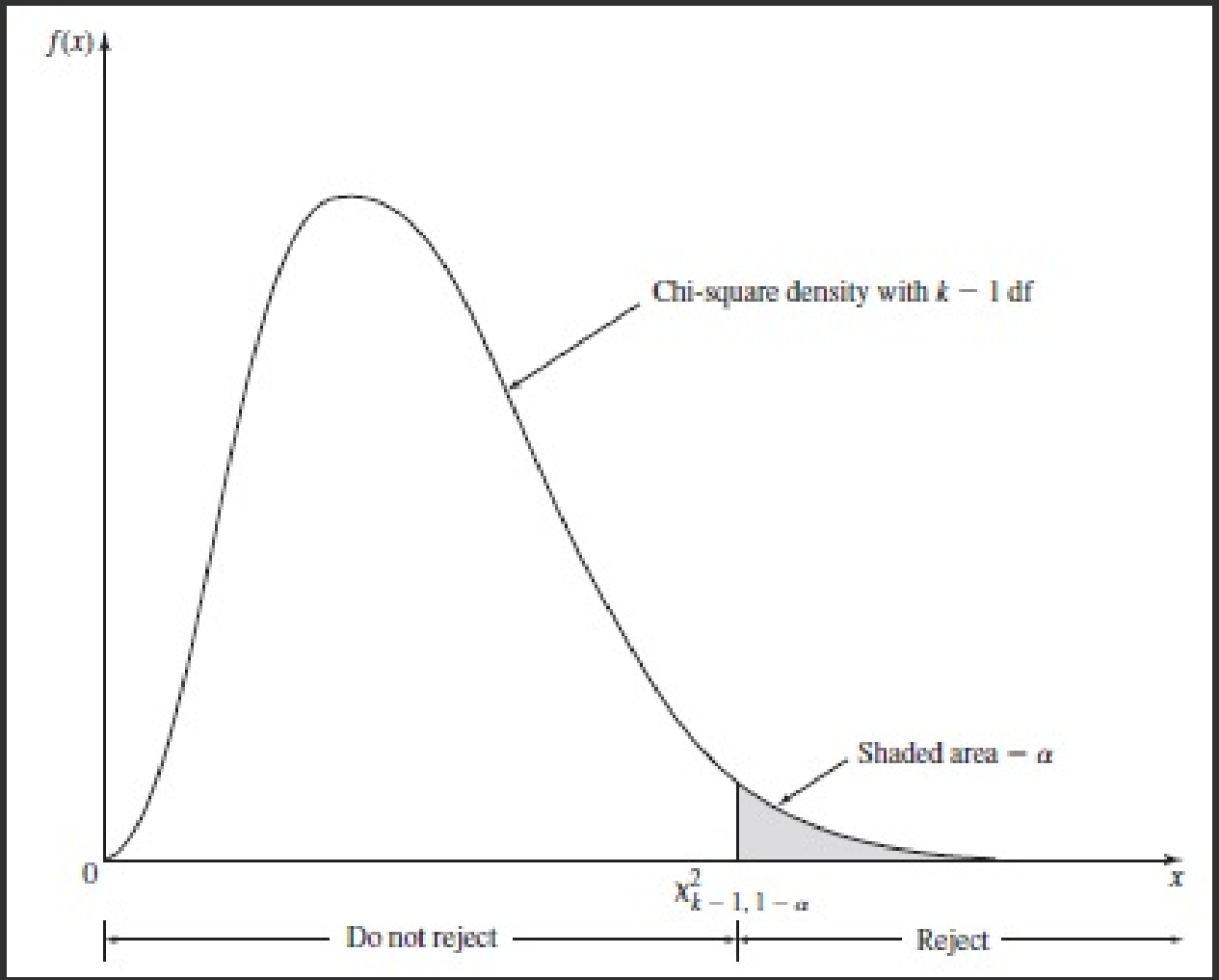
- El estadístico sigue una distribución chi-cuadrado con $k - r - 1$ grados de libertad, donde r es el número de parámetros estimados en $f_0(x)$ para encontrar E_i.

Prueba chi-cuadrado

Consideraciones

- Si las diferencias $O_i - E_i$ son pequeñas, el valor del estadístico es pequeño, por el contrario si esas diferencias son grandes el valor del estadístico es grande.
- El tamaño de la muestra deberá ser moderadamente grande, pues si la muestra es muy pequeña no se podrá formar un número suficiente de clases y si la muestra es muy grande la prueba conducirá a rechazo casi con seguridad.
- Se sugiere evitar tener clases con E_i menores que 5, esto puede conseguirse combinando clases vecinas. Tenga en cuenta que para calcular los grados de libertad, k es el número de clases efectivas.

Prueba chi-cuadrado



Prueba de Kolmogorov-Smirnov

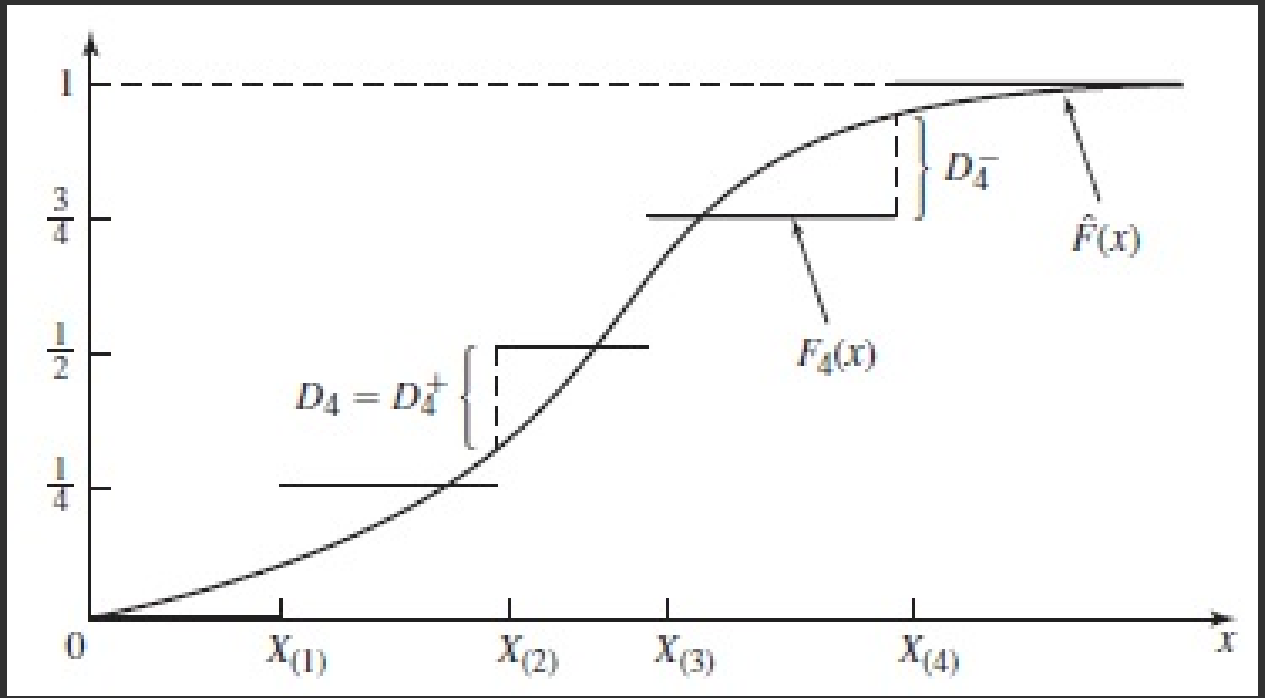
- Esta prueba formaliza la idea de una gráfica cuantil-cuantil y compara una función empírica de probabilidad con la función de la distribución hipotética.
- No requiere de especificación de intervalos y es válida para cualquier tamaño de muestra.

Prueba de Kolmogorov-Smirnov

Algoritmo

1. Tomar una muestra de los datos $x_i,\ i = 1, 2, \ldots, n$ y ordenarlos para obtener $y_j,\ j = 1, 2, \ldots, n$.
2. Estimar las diferencias por arriba y por abajo: $D^+ = \text{máx}\left[\frac{j}{n} - F(y_j)\right]$ y $D^- = \text{máx}\left[F(y_j) - \frac{j-1}{n}\right]$. El estadístico de prueba está dado por: $D : \text{máx}\left[D^+, D^-\right]$
3. Determine el valor crítico D_α para un nivel de significancia α y un tamaño de muestra N.
4. Si el estadístico calculado es mayor que el valor crítico, entonces se rechaza la hipótesis nula. De lo contrario, se concluye que no hay evidencia estadística para recharzarla.

Prueba de Kolmogorov-Smirnov



p-value

- El p-value es el nivel de significancia en el que se rechazaría H_0 para el valor dado del estadístico de prueba. Por lo tanto, un p-value alto tiende a indicar un buen ajuste (tendríamos que aceptar una gran posibilidad de error para rechazar), mientras que un p-value pequeño sugiere un mal ajuste.
- El p-value se puede ver como una medida de ajuste, siendo mejores los valores más grandes. Regla de rechazo: si el p-value es menor que el nivel de significancia entonces se debe rechazar la hipótesis nula.

Referencias

Banks, J., Carson II, J. S., Nelson, B. L. y Nicol, D. M. *Discrete-Event System Simulation*. Fifth (Pearson, 2014).

Law, A. M. *Simulation modeling and analysis*. Fifth (McGraw-Hill, 2015).