Metodología Six Sigma

Herramientas para la fase de análisis: Estadística inferencial con R

Profesor: Gustavo Ahumada

Introducción

La inferencia estadística es la rama de la estadística por la cual llegamos a conclusiones sobre una población a través de una muestra de la población. Podemos hacer inferencias sobre varios temas relacionados con los datos, por ejemplo, los parámetros de la distribución de probabilidad, los parámetros de un modelo dado que explica la relación entre variables, la bondad de ajuste a una distribución de probabilidad y las diferencias entre grupos (por ejemplo, con respecto a la media o la varianza).

En los proyectos Six Sigma, la mejora está estrechamente relacionada con el efecto que algunos parámetros del proceso (entrada) tienen en las características del proceso (salida). La inferencia estadística proporciona la base científica necesaria para lograr los objetivos del proyecto y validar sus resultados.

En esta sección revisamos algunas herramientas de inferencia estadística adecuadas a los proyectos Six Sigma como son: **Análisis de Regresión y Análisis de Varianza (ANOVA)**. Estas técnicas estudian la relación entre variables.

Nota: herramientas estadísticas como intervalos de confianza y estimación puntual, las cuales fueron evaluadas en unidades anteriores de este curso, deben ser estudiadas por parte del estudiante. De igual manera, el concepto de prueba de hipótesis (revisar sección 10.2 a 10.3).

Regresión

Identificación del modelo

Una de las inferencias más importantes que se deben hacer en un proyecto Six Sigma se refiere a la relación entre las características críticas para la calidad (CTQ) de nuestro proceso (Ys) y los parámetros del proceso (Xs), es decir, las variables que afectan el proceso. Esta relación se representa como una función donde se puede inferir el valor de las Y desde el conocimiento de las Xs, además de un componente de error (ε)

$$Y = f(X) + \varepsilon$$

Esta función es un modelo estadístico, con algunos parámetros que deberían ser estimados. Además, el término de error del modelo debería ser evaluado.

Le regresión es la técnica estadística empleada para estimar la función f en la ecuación anterior. El caso más sencillo es la regresión simple, donde f corresponde a una linea recta. Si tenemos solamente una variable dependiente (y) y una variable independiente (x), el modelo de regresión es:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

y los parámetros estimados son β_0 (el intercepto de la linea recta) y β_1 (la pendiente de la linea recta). Entonces el modelo estimado será:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

La diferencia entre los valores estimados y el valor actual en la muestra, $\hat{y} - y$, son los *residuos*. Los residuos son empleados para verificar la validez del modelo.

Antes de estimar el modelo de regresión, es necesario encontrar alguna encontrar alguna evidencia concerniente a la relación lineal entre las variables. El estadístico apropiado para tener una primera aproximación es el coeficiente de correlación, definido como:

$$r = \frac{s_{xy}}{s_x y_y}$$

donde $s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ es la covarianza entre dos muestras y s_x y s_y son las desviaciones estándar muestrales de x y y, respectivamente. Este coeficiente va de -1 a 1.

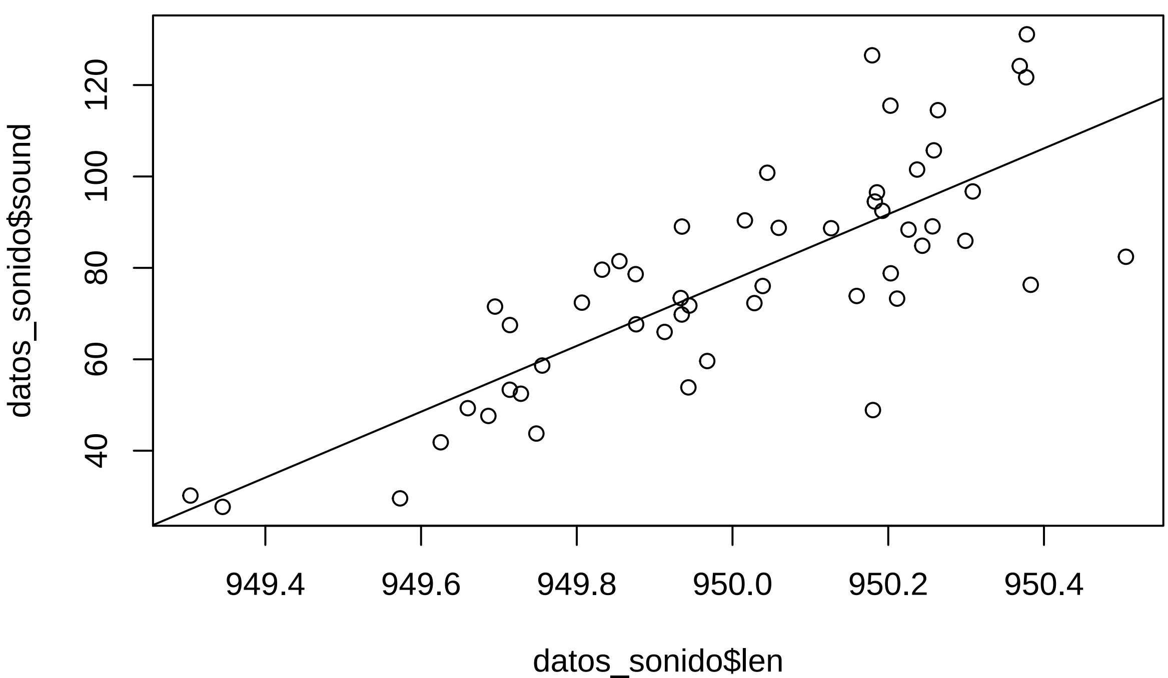
Ejemplo. Cuerdas de guitarra. Durante una prueba de resistencia de cuerdas de guitarra producidas, una nueva media fue tomada para aquellas cuerdas con un nivel de tensión superior a 8. Esta medida esta relacionada al volumen del sonido producido por la cuerda, la cual es considerada una característica CTQ.

El Master Black belt pregunta si la longitud de las cuerdas está relacionda al volumen del sonido. Primero

```
library(SixSigma)
datos_sonido <- na.omit(ss.data.strings)
cor(datos_sonido$sound, datos_sonido$len)
```

```
## [1] 0.7904747
```

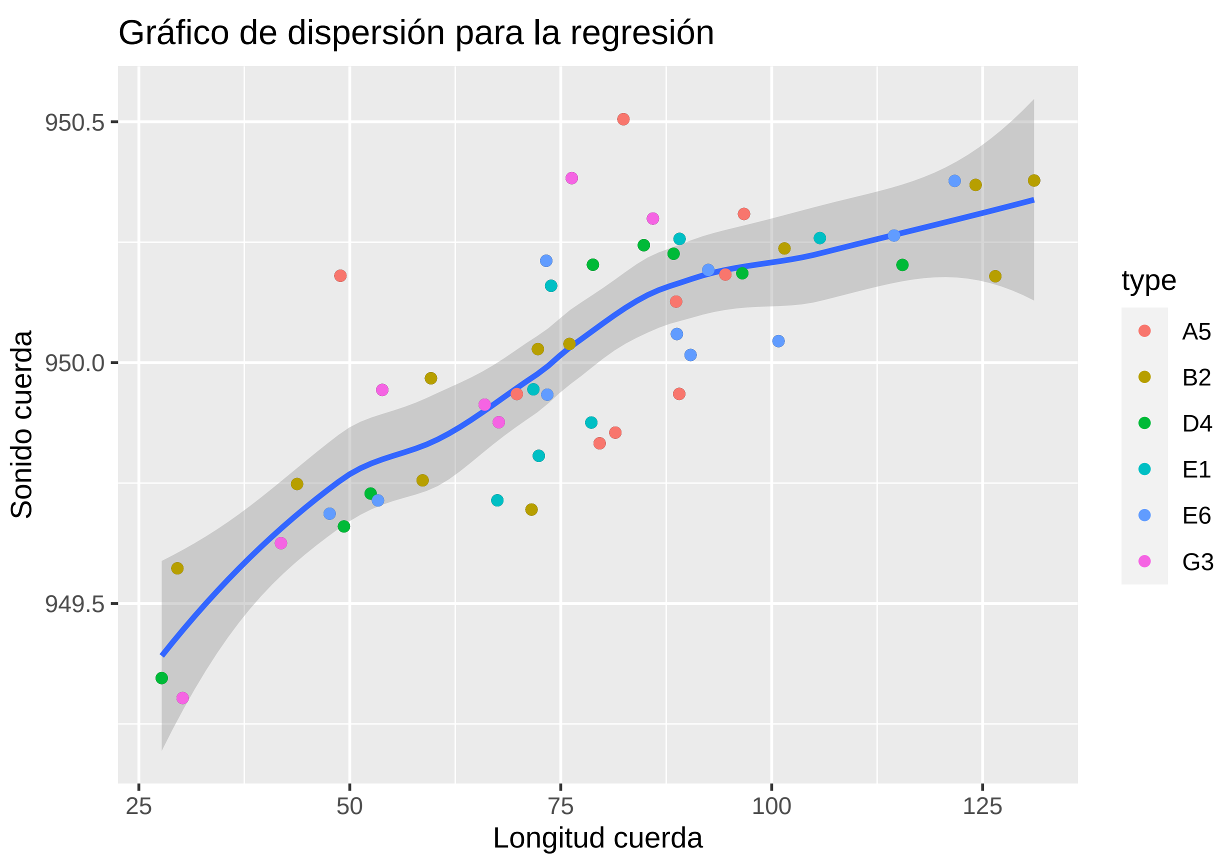
```
plot(datos_sonido$sound ~ datos_sonido$len, data = datos_sonido) +
abline(lm(sound ~ len, data = datos_sonido))
```



```
## integer(0)
```

```
library(ggplot2)
qplot(datos_sonido$sound, datos_sonido$len, data = datos_sonido,
      geom = c("point", "smooth"),
      xlab = "Longitud cuerda",
      ylab = "Sonido cuerda") +
        geom_point(aes(col = type)) +
  labs(title = "Gráfico de dispersión para la regresión")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Ajustando el modelo