Formulario de Análisis e Interpretación de Datos

Tabulación de Datos

i	Modalidad. Valor específico o grupo de datos
k	Cantidad de modalidades
n_i	Frecuencia absoluta. Conteo de valores de la modalidad
$N = \sum_{i=1}^{k} n_i$	Total de datos
$N_i = \sum_{j=1}^{i} n_j$	Frecuencia acumulada
$f_i = \frac{n_i}{N}$	Frecuencia relativa
$F_i = \frac{N_i}{N}$	Frecuencia acumulada relativa

Medidas de Resumen (Posición y Forma)

$E[x] = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$	Media datos no agrupados o valor esperado
$E[x] = \bar{x} = \frac{\sum_{i=1}^{N} f_i x_i}{\sum_{i=1}^{N} f_i}$	Media datos agrupados o valor esperado
$s^2 = E[x^2] - (E[x])^2 = \frac{1}{n}\left(\sum {x_i}^2 - \frac{(\sum x_i)^2}{n}\right) = \frac{\sum(x_i - \bar{x})^2}{n}$	Varianza datos no agrupados
$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$	Cuasivarianza muestral o estimador insesgado
$s^2 = E[x^2] - (E[x])^2 = \frac{1}{n}\sum f_i {x_i}^2 - \left(\frac{1}{n}\sum f_i x_i\right)^2$	Varianza datos agrupados
$s = \sqrt{s^2}$	Desviación estándar

Cuartiles

$Pos_2 = \frac{(N+1)}{2}$	Posición del cuartil 2 o mediana	$P[x < Q_2] = 0.5$
$Q_i = L_i + \left[\frac{\frac{N}{4}i - N_{i-1}}{n_i}\right] a_i$	Valor del cuartil i: L_i = Valor del límite inferior de este intervalos N_{i-1} = Frecuencia acumulada intervalo anterior a_i = Amplitud del intervalor i n_i = Frecuencia absoluta del intervalo i	

Correlación y Regresión $\hat{y}_i = a + bx_i$

$s_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$	
$s_{xx} = SSE_x = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$	Suma de cuadrados de x
$s_x = \sqrt{s_{xx}}$	Raíz cuadrada de suma de cuadrados de x
$s_{yy} = SSE_y = \sum(y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$	Suma de cuadrados de y
$s_y = \sqrt{s_{yy}}$	Raíz cuadrada de s
$r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$	Correlación
$b = \frac{s_{xy}}{s_{xx}} = r\frac{s_x}{s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$	Coeficiente de x para el modelo lineal $\hat{y} = a + bx$
$a = \bar{y} - b\,\bar{x} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$	Intercepto para el modelo lineal $\hat{y} = a + bx$
$e_i = y_i - \hat{y}_i$	Error del modelo
$R^2 = \frac{Varianza\ de\ las\ predicciones}{Varianza\ de\ las\ observaciones} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$	Coeficiente de determinación
$\sum(y_i - \bar{y})^2 = \sum[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2]$ $\sum(\hat{y}_i - \bar{y})^2 = \sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2$	
$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum e^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$	

Teorema de Bayes

$P[A\|B] = \frac{P[A \cap B]}{P[B]}$	Teorema de Bayes
$P[A \cap B] = P[B \cap A]$ $P[A\|B]P[B] = P[B\|A]P[A]$	Probabilidad conjunta
$P[A\|B] = \frac{P[B\|A]P[A]}{P[B]} = \left(\frac{P[B\|A]}{P[B]}\right)P[A]$ $P[A\|B] = \frac{P[B\|A]P[A]}{P[B\|A]P[A] + P[B\|not\ A]P[not\ A]}$	Teorema de Bayes replanteado
$\left(\frac{P[B\|A]}{P[B]}\right)$	Tasa de verisimilitud o de posibilidades
$P[A\|B] = \frac{P[B\|A]P[A]}{P[B\|A]P[A] + P[B\|not\ A]P[not\ A]}$ $P[A\|B] = \frac{P[A]P[B\|A]}{P[A]P[B\|A] + P[B]P[B\|B] + P[C]P[B\|C]}$	Teorema general de Bayes

Distribuciones de Probabilidad Discretas

$P[x=k]=\begin{cases} p & k=1 \\ 1-p & k=0 \end{cases}$	Distribucion de Probabilidad Bernoulli
$P[x=k]=\binom{1}{k}p^k(1-p)^{1-k}, k=0,1$	
$E[x]=p$	
$VAR[x]=pq=p(1-p)$	
$P[x=k]=\binom{N}{k}p^k(1-p)^{N-k}, k=1..N$	Distribucion de Probabilidad Binomial
$E[x]=Np$	
$VAR[x]=Np(1-p)$	
$P[x=k]=(1-p)^{k-1}p$	Distribucion de Probabilidad Geometrica
$E[x]=\frac{1}{p}$	
$VAR[x]=\frac{1-p}{p^2}$	

Distribución Normal y Teorema de Limite Central

$z \to N(0,1)$	Distribución Normal Estándar
$x \to N(\mu,\sigma) \to z=\frac{x-\mu}{\sigma} \to N(0,1)$	Estandarización de una Distribución Normal
$\sum x \to N(n\mu,\sqrt{n}\sigma) \to z=\frac{\sum x-n\mu}{\sqrt{n}\sigma} \to N(0,1)$	Teorema del Límite Central
$E[x] \to N\left(\mu,\frac{\sigma}{\sqrt{n}}\right) \to z=\frac{E[x]-\mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$	

Tamaño de la Muestra e Intervalo de Confianza

$e = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \rightarrow \sqrt{n} = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{e}\right) \rightarrow n = \left(z_{\frac{\alpha}{2}}\left(\frac{\sigma}{e}\right)\right)^2$	Tamaño de Muestra
$n = p(1-p)\left(\frac{z_{\frac{\alpha}{2}}}{e}\right)^2$	Tamaño de Muestra por Proporcion
$x \pm e$	Intervalo de Confianza
$x \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = x \pm z_{\frac{\alpha}{2}} \cdot SE = x \pm SB$	Intervalo de Confianza

Intervalos de Confianza

Prueba	**Estimador**	**Tam**	**Varianza**
Media	$\mu = \bar{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$	n>30	Conocidas
Media	$\mu = \bar{x} \pm t_{\frac{\alpha}{2},df}\frac{\sigma}{\sqrt{n}},\ df = n-1$	n>30	Desconocidas
Diferencia en media de dos muestras	$(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$		Conocidas
Diferencia en media de dos muestras	$(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2},df}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ $df = n_1 + n_2 - 2$		Desconocidas pero iguales
Diferencia en media de dos muestras	$(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2},df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$		Desconocidas y diferentes
Proporción	$p = \bar{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$		
Diferencia en proporción de dos muestras	$(p_1 - p_2) = (\bar{p}_1 - \bar{p}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\left(\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}\right) + \left(\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}\right)}$	$n_1p_1 > 5$ $n_1(1-p_1) > 5$ $n_2p_2 > 5$ $n_2(1-p_2) > 5$	
Varianza	$\sigma^2 \in \left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right]$		

Estimadores Estadísticos Pruebas de Hipótesis

Prueba	Estimador	Tamaño	Varianza	Distribución
Media	$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	n>30	Conocidas	$N(0,1)$
Media	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}},\ df = n - 1$	n≤30	Desconocida	t(n-1)
Diferencia en media de dos muestras	$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$		Conocidas	$N(0,1)$
Diferencia en media de dos muestras	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ $df = n_1 + n_2 - 2$		Desconocidas pero iguales	$t(n_1 + n_2 - 2)$
Diferencia en media de dos muestras	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$		Desconocidas y diferentes	t(df)
Proporción	$z = \frac{\bar{p} - p}{\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}}$			$N(0,1)$
Diferencia en proporción de dos muestras	$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\left(\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1}\right) + \left(\frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}\right)}}$	$n_1 p_1 > 5$ $n_1(1 - p_1) > 5$ $n_2 p_2 > 5$ $n_2(1 - p_2) > 5$		$N(0,1)$
Varianza	$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$			$\chi^2(n - 1)$
Cociente de variazas	s_1^2 / s_2^2			$F(n_1 - 1, n_2 - 1)$

Pruebas de Hipótesis

$$\boldsymbol{Nivel\ de\ confianza\ NC = 1 - \alpha}$$

$$\boldsymbol{valor.p_z = P\left[x \geq \frac{x-\theta}{\frac{\sigma}{\sqrt{b}}}\right]}$$

k=grados de libertad

Tipo	Hipótesis	Se rechaza Ho si	Desviación estándard
Cola izquierda	$Ho: x = \theta$ $H1: x < \theta$	$\frac{x-\theta}{\frac{\sigma}{\sqrt{n}}} < -z_\alpha$ $valor.p_z < \alpha$	Conocida
Cola derecha	$Ho: x = \theta$ $H1: x > \theta$	$\frac{x-\theta}{\frac{\sigma}{\sqrt{n}}} > z_\alpha$ $valor.p_{-z} < \alpha$	Conocida
Doble cola	$Ho: x = \theta$ $H1: x \neq \theta$	$\frac{x-\theta}{\frac{\sigma}{\sqrt{n}}} < -z_{\frac{\alpha}{2}}$ ó $\frac{x-\theta}{\frac{\sigma}{\sqrt{n}}} > z_{\frac{\alpha}{2}}$ $valor.p < \frac{\alpha}{2}$	Conocida
Cola izquierda	$Ho: x = \theta$ $H1: x < \theta$ s cuasivarianza	$\frac{x-\theta}{\frac{s}{\sqrt{n}}} < -t_{\alpha,k-1}$ $valor.p_t < \alpha$	Desconocida
Cola izquierda	$Ho: x = \theta$ $H1: x > \theta$ s cuasivarianza	$\frac{x-\theta}{\frac{s}{\sqrt{n}}} > t_{\alpha,k-1}$ $valor.p_{-t} < \alpha$	Desconocida
Doble cola	$Ho: x = \theta$ $H1: x \neq \theta$ s cuasivarianza	$\frac{x-\theta}{\frac{s}{\sqrt{n}}} < -t_{\frac{\alpha}{2},k-1}$ ó $\frac{x-\theta}{\frac{s}{\sqrt{n}}} > t_{\frac{\alpha}{2},k-1}$ $valor.p < \frac{\alpha}{2}$	Desconocida

Nivel de Significancia	Desviaciones de la media	Prob
$\alpha = 0.1$	Z está a 1.28 desv std	$P[x \leq z_\alpha] = 0.9$
$\alpha = 0.05$	Z está a 1.645 desv std	$P[x \leq z_\alpha] = 0.95$
$\alpha = 0.025$	Z está a 1.96 desv std	$P[x \leq z_\alpha] = 0.975$
$\alpha = 0.01$	Z está 2.326 desv std	$P[x \leq z_\alpha] = 0.99$

Intervalos de Confianza Regresión $\widehat{y_i} = a + bx_i$

$se = \sqrt{VAR[r]} = \sqrt{\frac{\sum(y_i - \widehat{y_i})^2}{n-2}}$	Error estándar de la regresión
$r_i = y_i - \widehat{y_i}$	Residuos
$\widetilde{r_i} = \frac{r_i}{se}$	Residuos estandarizado
$se[b] = \left[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right] = \sqrt{\frac{se^2}{\sum(x_i - \bar{x})^2}} = se\sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}$	Varianza de la pendiente
$b = \bar{b} \pm t_{n-2} se[b] = \bar{b} \pm t_{n-2} se\sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}$	Intervalo confianza pendiente El valor 0 no debe estar dentro del intervalo
$t_{\alpha,n-2} = \frac{\bar{b}}{se[b]} = \frac{\bar{b}}{\sqrt{\frac{\sum(y_i - \widehat{y_i})^2}{(n-2)\sum(x_i - \bar{x})^2}}}$	Estimador prueba de hipótesis
$y_i = \widehat{y_i} \pm t_{\frac{\alpha}{2},n-2} se\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)\sigma_x^2}}$	Intervalo de confianza de la predicción
$y_i = \widehat{y_i} \pm t_{\frac{\alpha}{2},n-2} s\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)\sigma_x^2}}$	Intervalo de confianza de una predicción
$VAR[a] = VAR[y - bx] = VAR[y] + b^2 VAR[x]$ $se[a] = \sqrt{VAR[a]} = se\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)}$	Varianza de la intercepto No es importante