CLAVES DICOTOMICAS PARA ESCOGER PRUEBAS ESTADISTICAS

By Pavel J. Atauchi Rojas
Museo de Historia Natural Cusco

1a. Preguntas acerca de asociación entre variables …….…… (2)
1b. Preguntas acerca de diferencia entre grupos …… (14)

2a. Las dos variables varían, pero uno no depende del otro ……... (3)
2b. Las dos variables varian, y el valor de uno depende del valor del otro [incluye variables predictivas u observación de causa-efecto]…….. (5)

3a. Comprobar supuestos paramétricos para el análisis correlacional **(Correlación)**

- Normalidad bivariante de los variables de respuesta/predictor – Marginal Scatterplot boxplot

```
> library(car)
> scatterplot(v1~v2, dataset)
# donde v1 y v2 son variables continuas dentro del data.frame dataset
```

- Linearidad de puntos de los datos en un scatterplot, treendline and lowess smoother.

```
>library(car)
>scatterplot(v1~v2, dataset, reg.line = F)
# donde v1 y v2 son variables continuas dentro del data.frame dataset y reg.line=F excluye la línea de regresión desde el grafico.
```

- Reune los supuestos paramétricos **(Correlación de Pearson)**

```
>corr.test(~v1+v2, data=dataset)
# Donde v1 y v2 son variables continuas

# Producir grafico de resumen ……... (13)
```

3b. Los supuestos paramétricos no se justifican y la escala de transformación no son exitosos o son inadecuados ……. (4)

4a. El tamaño de muestra se encuentra entre 7 y 30 **(Rango de Correlación de Spearman)**

```
>cor.test(~v1+v2, data=dataset, method="sperman")
# Donde v1 y v2 son variables continuas
# Producir grafico de resumen ……... (13)
```

4b. El tamaño de muestra es mayor a 30 **(Correlación de Tao Kendall's)**

```
>cor.test(~v1+v2, data=dataset, method="kendall")
# Donde v1 y v2 son variables continuas
# Producir grafico de resumen ……... (13)
```

5a. Comprobar los supuestos paramétricos para el análisis de regresión **(Regresión)**

- Normalidad en las variables de respuesta (y variable predictor si fue medido) - marginal scatterplot boxplots
- Homogenidad de varianza – Spread of data around scatterplot trendline
- Linearidad de puntos de los datos en un scatterplot, trendline and lowess soother useful

Reúne los supuestos paramétricos ……..…….. (6)

5b. No reúne los supuestos paramétricos y las escalas de transformación no son éxitos e inapropiados …….…….. (8)

6a. Los niveles de variables de predicción están agrupados (no fueron medidos), no hay incertidumbre en la variable de predictor o el objetivo principal es:

- Prueba de Hipótesis (Ho : $\beta = 0$)
- Generando un modelo predictivo ($y = \beta o + \beta_1 x$)

Mínimos Cuadrados Ordinarios (OLS) de regresión …….. (7)

6b. Los niveles de variables de predicción no están agrupados (no fueron medidos), y el objetivo principal del análisis es estimar la pendiente de la relación poblacional (Modelo II de regresión)

```
>library(biology)
>data.lm <-lm.II(DV~IV, christ, type="RMA")
>summary(data.lm)
```

Donde DV e IV son variables de repuesta y predictor respectivamente en el grupo de datos data.frame. type puede ser uno de "MA", "RMA", "rMA" o "OLS". Para type="rMA", también es posible forzar una respuesta mínima de cero (zero=T)

Producir grafico de resumen …….……... (13)

7a. Valor de respuesta individual para cada nivel de la variable predictora

```
>dataset.lm<-lm(IV~DV, dataset)
>plot(dataset.lm)
>influence.measures(dataset.lm)
>summary(dataset.lm)
```

Donde DV e IV son variables de respuesta y predictor respectivamente dentro de dataset data.frame

#Opcional:

Para obtener intervalos de confianza del parámetros (f) …… (11)

nota: Si hay incertidumbre en la variable predictora, los intervalores de confianza de los parámetros pueden ser inapropiados

Predecir nuevos valores de la variable de respuesta ……..……... (12)

Producir grafico de resumen ……...……... (13)

7b. Valor de respuesta múltiple para cada nivel de la variable predictora

```
>anova(lm(DV~IV+as.factor(IV), dataset))
```

- Acumulación de residuales temporales

```
>dataset.lm<-lm(DV~IV, dataset)
```

```
>summary(dataset.lm)
```

- No acumulación de residuales temporales

```
>dataset.lm<-aov(DV~IV+error(as.factor(IV)), dataset)
>summary(dataset.lm)
>lm(DV~IV, dataset)
# Donde DV e IV son variables de respuesta y predictor respectivamente.
```

8a. El muestreo han sido colectado al azar/aleatoriamente/sin orden ni concierto, no hay razón para sospechar falta de independencia……..……. (9)

8b. El muestreo no ha sido posible aleatorizar, los datos no son necesariamente independientes **(Prueba de Aleatorización)**

```
>stat<-function(data,index){
+ summary (lm(DV~IV, data))$coef[2, |3]
+ }

>rand.gen<-function(data, mle){
+out <-data
+out$IV<-sample(out$IV, replace=F)
+out
+}

>library(boot)
>dataset.bot<-boot(dataset, stat, R=5000, sim=”parametric”, ran.gen=rand.gen)
>plot(dataset.boot)
>dataset.boot
# Donde DV e IV son variables de repuesta y predictor respectivamente

#Opcional:
# Para obtener intervalos de confianza del parámetros (f) ……...… (11)
nota: Si hay incertidumbre en la variable predictora, los intervalores de confianza de los parámetros pueden ser inapropiados
# Predecir nuevos valores de la variable de respuesta ……..…... (12)
# Producir grafico de resumen ……..……... (13)
```

9a. Ligera no normalidad debido principalmente a los valores atípicos (observaciones influyentes), los datos lineales (M-regresion)

```
>library(MASS)
>data.rlm<-rlm(DV~IV, dataset)
# Donde DV e IV son variables de respuesta y predictora respectivamente
#Opcional:
# Para obtener intervalos de confianza del parámetros (f) ……..…… (11)
nota: Si hay incertidumbre en la variable predictora, los intervalores de confianza de los parámetros pueden ser inapropiados
# Predecir nuevos valores de la variable de respuesta ……..……... (12)
# Producir grafico de resumen ……..…... (13)
```

9b. Los datos no son normales y/o no son lineales ……..… (10)

10a. Repuesta binario (e.g. vivo/muerto, presente/ausente)……. **(Regresión Logistica)**

10b. Distribución subyacente de la variable de repuesta y los residuales son conocidos …… **(Modelo General Linealizado)**

10c. Datos curvilíneos ……..…… **(Regresión no Lineal)**
10d. Datos monotonicos no lineales ……..…. **(Regresión no paramétrica)**

- Theil-Sen, mediana únicas de regresión robusta **(Kendall's)**

```
>library(mblm)
>data.mblm<-mblm(DV~IV, dataset, repeated=F)
>summary(data.mblm)
```

- Siegel, repite las medianas de regresión

```
>library(mblm)
>data.mblm<-mblm(DV~IV, dataset, repeated=T)
>summary(data.mblm)
```

Donde DV e IV son variables de repuesta y predictor respectivamente
#Opcional:
Para obtener intervalos de confianza del parámetros (f) ……..…… (11)
nota: Si hay incertidumbre en la variable predictora, los intervalores de confianza de los parámetros pueden ser inapropiados
Predecir nuevos valores de la variable de respuesta ……..….... (12)
Producir grafico de resumen ……..….... (13)

11. Generando Parametros de Intervalo de Confianza

```
>confint(model, level=0.95)
```

#donde model es un modelo ajustado

Obtener estimaciones de parámetros aleatorizados y sus intervalos de confianza

```
>par.boot<-function(dataset, index){
+     x<-dataset$ALT[index]
+     y<-dataset$HK[index]
+     model<-lm(y~x)
+     coef(mol)
+}

>data.boot<-boot(dataset, par.boot, R=5000)
>boot.ci(dataset.boot, index=2)
```

Donde dataset es un data.frame. El argumento opcional (R=5000) indica 5000 randomizaciones y el argumento (index=2) indica que parámetro genera el intervalo de confianza para (y-intercept=1, slope=2). Note el uso de la función lm() para estimar parámetros y podría ser remplazado por alternativas robustas como rlm() o mblm().

12. Generando nuevos valores de repuesta (e Intervalos de predicción correspondientes)

```
>predict(model, data.frame(IV=c()), interval="p")
```

donde model es un modelo ajustado y IV is la variable predictor y c() es un vector de nuevos valores (e.g. c(10,13.4))

Para obtener los intervalos de predicción de la aleatorización

```
>pred.boot<-function(dataset, index){
+     dataset.rs<-dataset[index, ]
+     dataset.lm<-lm(HK~ALT, dataset.rs)
+     predict(dataset.lm, data.frame(ALT=1))
+}

>dataset.boot<-boot(dataset, pred.boot, R=5000)
```

```
>boot.ci(dataset.boot)
```
Donde dataset es el nombre del dataframe. Notar el uso de la función lm() para estimar los parametros. Esto podría ser remplazado por alternativas mcomo rlm() o mblm().

13. Resumen del diagrama base para la correlación y regresión

```
>plot(v1~v2, data, pch=16, axes=F, xlab="", y lab="")
>axis(1, cex.axis=0.8)
>mtext(text="x-axis tittle", side=1, line=3)
>axis(2, las=1)
>mtext(text="y-axis title", side = 2, line=3)
>box(bty="1")
```
donde v1 y v2 son variables continuas en el dataset data.frame. Para regresión, v1 representa la variable de respuesta y v2 representa la variable predictor.

Agregando elipse de confidencia
```
>data.ellipse(v2, v1, levels=0.95, add=T)
```

Agregando línea de regresión
```
>abline(model)
```
#donde model representa el modelo de regresión ajustado.

Agregando intervalo de confianza de la regresión
```
>x<-seq(min(IV), max (IV), 1=1000)
>y<-predict(object, data.frame (IV=x), intervalo="c")
>matlines(x, y, lty=1, col1)
```
donde IV es el nombre de la variable predictora (incluyendo el data.frame) model representa el modelo de regresión ajustado.

16a. Las frecuencias calculadas esperadas a partir de datos de la muestra de acuerdo con una relación teórica - frecuencias homogéneas **(Prueba de chi-cuadrado)**

```
>chisq.test(c(c1, c2, …))
```
OR
```
>chisq.test(data.xtab)
```
Donde c1, c2, … son frecuencias tabuladas de cada clasificación y data.xtab es una tabla de valores observados.

Comprobar los supuestos que no mas del 20% de frecuencias esperadas son menos de 5, añadir la función anterior con $res, e.g. chisq.test(data.xtab)$res

Especificar una tasa de alternativa de valores esperados, usando el argumento p=c()

La performance de la prueba G, usar la función g.test() en el paquete biology

16b. Las frecuencias calculadas esperadas a partir de un modelo matemático representando una distribución **(Goodness of fist test- Kolmogorov-Smiirnov test)**

```
>ks.test(DV, DIST, …)
# OR
>ks.test(DV, “dist”, …)
# e.g.
>ks.test(DV, “pnorm”, mean(DV),  sd(DV))
```

Donde DV is el nombre de la variable dependiente. El segundo argumento es también un vector numero (DIST) representando la distribución para comparar la variable dependiente o mas de una cadena de caracteres(“dist”) representando la función de distribución acumulativa.

17a. Tabla de contingencia de dos vías …… (18)
17b. Tabla de contingencia de tres o mas vías **(Considerar GLM)…… (32)**

18a. Comprobando el supuesto que no mas del 20% de frecuencias esperadas son mas de 5.

```
>chisq.test(data.xtab, corr=F)$exp
```

Reune los supuestos ……... (19a)
18b. No reúne los supuestos ……..…….. (19b)

19a. Analizar la tabla de contingencia **(Chi-cuadrado – Todos los valores esperados son mas de 5)**

```
>chisq.test(data.xtab, corr=F)
```

Performance prueba de G, usando la función g.test() in el paquete biology

Si se rechaza la hipótesis nula

Examinar los residuales
Anexar la función anterior con $res, e.g. chisq.test(data.xtab, corr=F)$res

Examinar – odds ratios ……..……...…… (20)
Construir la figura de resumen …… (21)

19b. Analizar la tabla de contingencia **(Prueba exacta de Fisher)**

```
>Fisher.test(data.xtab)
```

Si la hipótesis nula es rechazada

Examinar – odds ratios ……..…...…… (20)
Construir la figura de resumen ……..…… (21)

20. Calcular –**Odds ratios**

```
>library()
>oddsratios(data.xtab)
```

21. Estructurar la Figura de resumen

```
>library(vcd)
>strucplot(data.xtab, shade=T)
```

23a. Comprobar los supuestos paramétricos

- Normalidad de las variables de respuesta de cada nivel de las variables categóricas – boxplots

  ```
  >boxplot(DV~Factor, dataset)
  ```

  # Donde DV y Factor son variables de respuesta y factor respectivamente en el grupo de dato data.frame

- Homogeneidad de varianza – boxplots y scatterplot de medias versus varianza

  ```
  >plot(tapply(dataset$DV, dataset$Factor, var), tapply(dataset$DV, dataset$Factor, mean))
  ```

  # Donde DV and Factor son variables de respuesta y factor respectivamente

Reúne los supuestos paramétricos …… (24)
23b. No reúne los supuestos paramétricos ……... (27)

25a. Modelo I – Un solo factor fijo

```
>data.aov<-aov(DV~Factor, dataset)
>plot(data.aov)
>anova(data.aov)
```

Si, la hipótesis nula se rechaza – Diferencias significativa entre grupos de medias detectadas …… (34)

25b. Modelo II – Un solo factor al azar

```
>anova(aov(DV~Factor, dataset))
```

Si, la hipótesis nula se rechaza – Diferencias significativa entre grupos de medias detectadas – calcular los componente de la varianza

```
>library(nlme)
>data.lme<-lme(DV~1, radom = ~1|Factor, data = dataset, method="ML")
>varCorr(data.lme)
>data.lme<-lme(DV~1, random=~1 | Factor, data=dataset, method="REML")
>VarCorr(data.lme)
```

26a. Con las comparaciones de medias previstas

```
>contrasts(dataset$Factor)<-cbind(c(contrasts), c(contrasts), …)
>round(crossprod(constrasts(dataset$Factor)), 2)
>data.list<-list(Factor=list(lab=1, …), …)
>data.aov<-aov(DV~Factor, data=dataset)
>plot(data.aov)
>summary(data.aov, split=data.list)
```

26b. Con las tendencias polinómicas previstas

```
>contrasts(dataset$Factor)<-"contr.poly"
>data.list<-list(Factor=list(linear=1))
```

```
>data.aov<-aov(DV~Factor, data=dataset)
>plot(data.aov)
>summary(data.aov, split=data.list)
```

27a. Intentar una transformación de escala de datos ……..…… (23)
27b. La transformación fue inapropiada ……...….. (28)

28a. La distribución subyacente de la variable de respuesta es normal pero las varianzas no son iguales **(Prueba de Welch's)**

```
>oneway.test(DV~Factor, var.equal=F)
```

si se rechaza la hipótesis nula – Las diferencias significativas entre medias de grupos detectado ……..…… (31)

OR considerar GLM

28b. La distribución subyacente de la variable de respuesta no es normasl ……..…. (29)

29a. La distribución subyacente de la variable de respuesta y los residuales son conocido **(GLM)**
29b. La distribución subyacente de la variable de respuesta y los residuales no son conocidos …… (30)

30a. Varianzas no son tremendamente desiguales, pero presenta valores atípicos **(Prueba no paramétrica de Kruskal-Wallis)**

```
>kruskal.test(DV~Factor, var.equal=F)
```

Si la hipótesis nula es rechazada – Las diferencias significativa entre grupos de medias detectadas …… (31 c,b/c)

30b. Varianzas no son tremendamente desiguales, no es posible el muestreo aleatorio **(Prueba de Randomización)**

```
>library(boot)
>data.boot<-boot(dataset, stat, R=999, sim="parametric",
rand.gen=rand.gen)
>plot(data.boot)
>print(data.boot)
```

Donde stat es la estadística para calcular repetidamente y rand.gen define como los datos fueron aleatorizados.

31a. Comparaciones paramétricas simultaneas múltiples **(Prueba de Tukey)**

```
>library(multcomp)
>summary(glht(model, linfct=mcp(Factor="Tukey")))
```

31b. Comparaciones no paramétricas simultaneas múltiples **(Prueba de Steel)**

```
>library(npmc)
>data<-data.frame(var=dataset$|DV, class=dataset$Factor)
>summary(npmc(data), type ="steel")
```

31c. Comparaciones multiples basados en ajustes de p-valor

```
>library(multtest)
>mt.rawp2adjp(pvalues, proc="sidakSD")
>p.adjust(pvalues, method="holm")
```

Donde pvalor es una lista de p valor desde cada comparación de pares y 'holm' y 'sidakSD' son los nombres de los procedimientos ajustados de los p-valores. Por procedimientos alternativos.

La función p.adjust de encima también puede ser llamado a partir de rutinas sin otras pares

(Prueba de pares paramétricos)

```
>pairwise.t.test(DV~Factor, pool.sd=F, p.adjust="holm")
```

(Prueba de pares no paramétricos)

```
>pairwise.wilcox.test(DV~Factor, p.adjust="holm")
```

Transformación de datos

Naturaleza de datos	**Transformación**	**R syntax**
Mediciones (tamaño, altura, etc)	Log_e	Log(x)
	Log_{10}	Log(x, 10)
	Log_{10}	Log10(x)
	Log x + 1	Log(x+1)
Conteos (numero de individuos, etc)	√	Sqrt(x)
Porcentaje (puede ser proporciones)	arcsin	Asin (sqrt(x))*180/pi

Ajuste de alternativas de p-valor para uso de pairwise.wilcoxon.test y pariwise.t.test

Syntax	Correction	Descripcion
Bonferroni	Correcion de paso simple de bonferroni	p-valor multiplicado por numero de comparaciones para control de la tasa de error de familia de pares.
Holm		

33a. Regresion Logistico – Variable predictor simple

> data.glm<-glm (DV~IV, dataset, family="Poisson")

Obtener los parámetros del modelo
> summary(data.glm)

Obtener la tabla de desviación
>anova(data.glm, test="chisq")

33b. Variable predictor multiple – **Regresion Logistica Multiple**

>data.glm<-glm(DV~IV1+IV2+IV3+…+IVn, dataset, family="Poisson")

Obtener las estimaciones de los parámetros del modelo
> summary (data.glm)

OR

> anova(data.glm, dataglmR, test="chisq")

Where data.glmR is un modelo reducido con adicion, omitiendo la variable de interés.

34a. Comprobar los supuestos

En lo que prosigue data.glm es el modelo lineal generalizado equipado

Falta de ajuste

a. Estadistico de prueba de normalidad Le Cessie-van Houwelingen

```
>librry(Design)
>data.lrm<-lrm(formula, dataset, y=T, x+T)
>resid(data.lrm)
```

Where Formula *es una formula relacionada a la variable de respuesta de la combinación lineal de variables predictoas*

b. Pearson X2

```
>pp<-sum(resid(data.lrm, type="Pearson")^2)
>1-pchisq(pp, data.glm$df.resid)
```

c. Desviacion

```
>1-pchisq(data.glm, data.glm$df.resid)
```

Relacion lineal entre los predictores y la función de enlace (grafico de componentes + residuales)

```
>library(car)
>cr.plots(data.glm, ask=F)
```

Influencia

```
>influence.measures(data.glm)
```

Reune los supuestos ……. (Go Back)

34b. No reúne los suspuestos, la transformación de escala de las variables predictoras pueden ser usados para mejorar la linearidad, caso contrario considerar **(GAM)** …. ()

35a. Examinar la sobredispersion

Residuales de Pearson

>sum(resid(data.glm, type="pearson")^2)/data.glm$df.resid

Desviacion

>data.glm$deviance/data.glm$df.resid

Dispersion no se aparta sustancialmente de 1 ……. (Go back)

35b. Modelo de sobre dispersión

Volver a correr el modelo con distribución "quasi"

>data.glm<-glm(DV~IV, dataset, family="quasibinomial")

>anova(data.glm, test="F")

Considerar un binomial negativo

>data.glm<-glm.nb(DV~IV, dataset)

>anova(data.glm, test="F")

36a. Variable predictor continuo **(Regresion Poisson)**

>data.glm<-glm(DV~IV1+…, dataset, family="poisson")

comprobar que el modelo reuna los suspuestos

Examinar la sobredispersion

Obtener los estimadores de los parámetros del modelo

>summary(data.glm)

OR

>anova(data.glm, data.glmR, test="chisq")

Donde data.glmR, es un modelo reducido construirdo por omisión del termino de interés

Calcular valores de ratios ……..…. (37)

Performance de selección de modelo y modelo promedio ……. (39)

36b. Solo variables categóricas **(Modelamiento logarítmico linear)**

>data.glm<-glm(DV~CAT1*CAT2*…., dataset, family="poisson")

Examinar independiente condicional

>data.glm1<-update(data.glm, ~. –CAT1:CAT2, dataset)

>anova(data.glm, data.glm1, test="chisq")

Ver tabla para modelos log-linear reducidos y completamente apropiados para examinar completo y dependencia e independiencia condicional

Calcular los ratios …….. (37)

Performance de selección de modelo ……. (39)

37a. Calcular los ratios

>library(biology)

>odds.ratio(data.glm)

38a. Modelos Aditivos Generalizados **(GAM)**

>library(gam)
>data.gam<-gam(DV~lo(CAT1)+lo(CAT2)+…, family=”gaussian”, dataset)

La family=parameter, puede ser usado para especificar el error de distribución apropiado

Comprobar que los modelos están adheridos a los supuestos
Examinar la estimación de los parámetros del modelo
>summary(data.gam)

Performance selección de modelo ……. (39)

39a. Performance de Selección de Modelo

>library(MuMIn)
>dredge(data.glm)
>model.avg(get.models(dredge(model)))

OR

>library(biology)
>model.selection.glm(model)