Curso avanzado de estadística

Sesión 2: Contrastes bivariados y algunas utilidades

Carlos Vergara Hernández

7 de julio de 2022

Servicio de estudios estadísticos, FISABIO

Contrastes bivariados

Ideas iniciales

- Este tipo de contrastes se dirigen a evaluar la asociación entre dos variables.
  - Siempre consideraremos una variable respuesta (dependiente) y una predictora (independiente).

Ideas iniciales

- Este tipo de contrastes se dirigen a evaluar la asociación entre dos variables.
  - Siempre consideraremos una variable respuesta (dependiente) y una predictora (independiente).

- La palicación directa de este tipo de pruebas se reserva a contextos con diseños experimentales clásicos y bien definidos (por ejemplo, al rey de los diseños: experimento controlado aleatorizado por bloques).
  - La aplicación indiscriminada de estas pruebas debe eliminarse como elemento de soporte en tablas descriptivas (o debe controlarse el error tipo I frente a comparaciones múltiples).

Ideas iniciales

- Este tipo de contrastes se dirigen a evaluar la asociación entre dos variables.
  - Siempre consideraremos una variable respuesta (dependiente) y una predictora (independiente).

- La palicación directa de este tipo de pruebas se reserva a contextos con diseños experimentales clásicos y bien definidos (por ejemplo, al rey de los diseños: experimento controlado aleatorizado por bloques).
  - La aplicación indiscriminada de estas pruebas debe eliminarse como elemento de soporte en tablas descriptivas (o debe controlarse el error tipo I frente a comparaciones múltiples).

- Como ya se comentó, hay pruebas paramétricas y no paramétricas.
  - Desde mi perspectiva (no solo la mía), y dado que los requisitos de las paramétricas no se suelen cumplir en entornos reales, resulta apropiado realizar de base una prueba no paramétrica.

Esquema de pruebas habituales

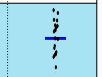
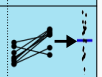
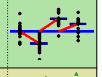
Prueba	Tipo respuesta	Variable independiente	Tipo de test
Test t	Númerica independiente	2 categorías	Paramétrico
Test t pareado	Númerica pareada	2 categorías	Paramétrico
Mann-Whitney	Ordinal independiente	2 categorías	No paramétrico
Mann-Whitney pareado	Ordinal pareada	2 categorías	No paramétrico
ANOVA una vía	Númerica independiente	≥ 2 categorías	Paramétrico
ANOVA medidas repetidas	Númerica pareada	≥ 2 categorías	Paramétrico
Kruskal-Wallis	Ordinal independiente	≥ 2 categorías	No paramétrico
Friedman	Ordinal pareada	≥ 2 categorías	No paramétrico
Chi-cuadrado	Categórica independiente	≥ 2 categorías	Nominal
Test de Fisher	Categórica independiente	≥ 2 categorías	Nominal

Extensión: pruebas bivariadas como LM y GLM

Common statistical tests are linear models

Last updated: 28 June, 2019.Also check out the Python version!

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=…[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +…)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +…+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + S_2 + S_3 +…+ S_K + G_2*S_2 + G_3*S_3 + … + G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: $\mathbf{G_{2\,to\,N}}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $\mathbf{S_{2\,to\,K}}$ for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2 + G_3*S_3 +…+ G_N*S_K, family=…)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $log(y_i) = log(N) + log(\alpha_i) + log(\beta_j) + log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +…+ G_N, family=…)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.

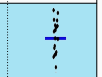
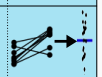
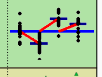
Jonas Kristoffer Lindeløv
https://lindeloev.net

Extensión: pruebas bivariadas como LM y GLM

Common statistical tests are linear models

Last updated: 28 June, 2019.Also check out the Python version!

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=…[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +…)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +…+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + S_2 + S_3 +…+ S_K + G_2*S_2 + G_3*S_3 + … + G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: $\mathbf{G_{2\,to\,N}}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $\mathbf{S_{2\,to\,K}}$ for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2 + G_3*S_3 +…+ G_N*S_K, family=…)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments: `glm(model, family=poisson())`* *As linear-model, the Chi-square test is $log(y_i) = log(N) + log(\alpha_i) + log(\beta_j) + log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +…+ G_N, family=…)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

Pero...

En este apartado nos limitaremos a ver las pruebas bivariadas y, llegado el momento, comprobaremos la equivalencia con LM y GLM.

Datos que usaremos

La base de datos viene como elemento de soporte al paquete de R `compareGroups`, y hace referencia al estudio **PREDIMED** (Estruch et al., 2018).

- Ensayo clínico aleatorizado, evaluando el efecto de suplementos de aceite de oliva o frutos secos en la salud cardiovascular.
- Los datos que acompañan al paquete no tienen pinta de ser reales, pero quiero pensar que puede tratarse de una simulación en base a los reales.

Datos que usaremos

La base de datos viene como elemento de soporte al paquete de R `compareGroups`, y hace referencia al estudio **PREDIMED** (Estruch et al., 2018).

- Ensayo clínico aleatorizado, evaluando el efecto de suplementos de aceite de oliva o frutos secos en la salud cardiovascular.
- Los datos que acompañan al paquete no tienen pinta de ser reales, pero quiero pensar que puede tratarse de una simulación en base a los reales.

Para cargar los datos y echar un vistazo a las primeras observaciones de las primeras columnas:

```
library(compareGroups)
data(predimed)
predimed[1:6, 1:6]
```

```
#            group    sex age   smoke   bmi waist
# 1        Control   Male  58  Former 33.53   122
# 2        Control   Male  77 Current 31.05   119
# 4  MedDiet + VOO Female  72  Former 30.86   106
# 5 MedDiet + Nuts   Male  71  Former 27.68   118
# 6  MedDiet + VOO Female  79   Never 35.94   129
# 8        Control   Male  63  Former 41.66   143
```

Dinámica de trabajo

1. Importar datos (en bruto: **NUNCA SE MODIFICAN**).

Dinámica de trabajo

1. Importar datos (en bruto: **NUNCA SE MODIFICAN**).

2. Decidir si se reservarán datos para validación posterior y, si así fuera, reservarlos ya mismo.

Dinámica de trabajo

1. Importar datos (en bruto: **NUNCA SE MODIFICAN**).

2. Decidir si se reservarán datos para validación posterior y, si así fuera, reservarlos ya mismo.

3. Limpiar base de datos y guardar los procesados.

Dinámica de trabajo

1. Importar datos (en bruto: **NUNCA SE MODIFICAN**).

2. Decidir si se reservarán datos para validación posterior y, si así fuera, reservarlos ya mismo.

3. Limpiar base de datos y guardar los procesados.

4. Descriptiva de los datos (tabla 1), incluyendo visualizaciones estándar (¿pruebas bivariadas?).

Dinámica de trabajo

1. Importar datos (en bruto: **NUNCA SE MODIFICAN**).

2. Decidir si se reservarán datos para validación posterior y, si así fuera, reservarlos ya mismo.

3. Limpiar base de datos y guardar los procesados.

4. Descriptiva de los datos (tabla 1), incluyendo visualizaciones estándar (¿pruebas bivariadas?).

5. Confección del análisis: modelo apropiado.

Dinámica de trabajo

1. Importar datos (en bruto: **NUNCA SE MODIFICAN**).

2. Decidir si se reservarán datos para validación posterior y, si así fuera, reservarlos ya mismo.

3. Limpiar base de datos y guardar los procesados.

4. Descriptiva de los datos (tabla 1), incluyendo visualizaciones estándar (¿pruebas bivariadas?).

5. Confección del análisis: modelo apropiado.

6. Ejecución del modelo, extrayendo medidas resumen del modelo (tabla 2).

Dinámica de trabajo

1. Importar datos (en bruto: **NUNCA SE MODIFICAN**).

2. Decidir si se reservarán datos para validación posterior y, si así fuera, reservarlos ya mismo.

3. Limpiar base de datos y guardar los procesados.

4. Descriptiva de los datos (tabla 1), incluyendo visualizaciones estándar (¿pruebas bivariadas?).

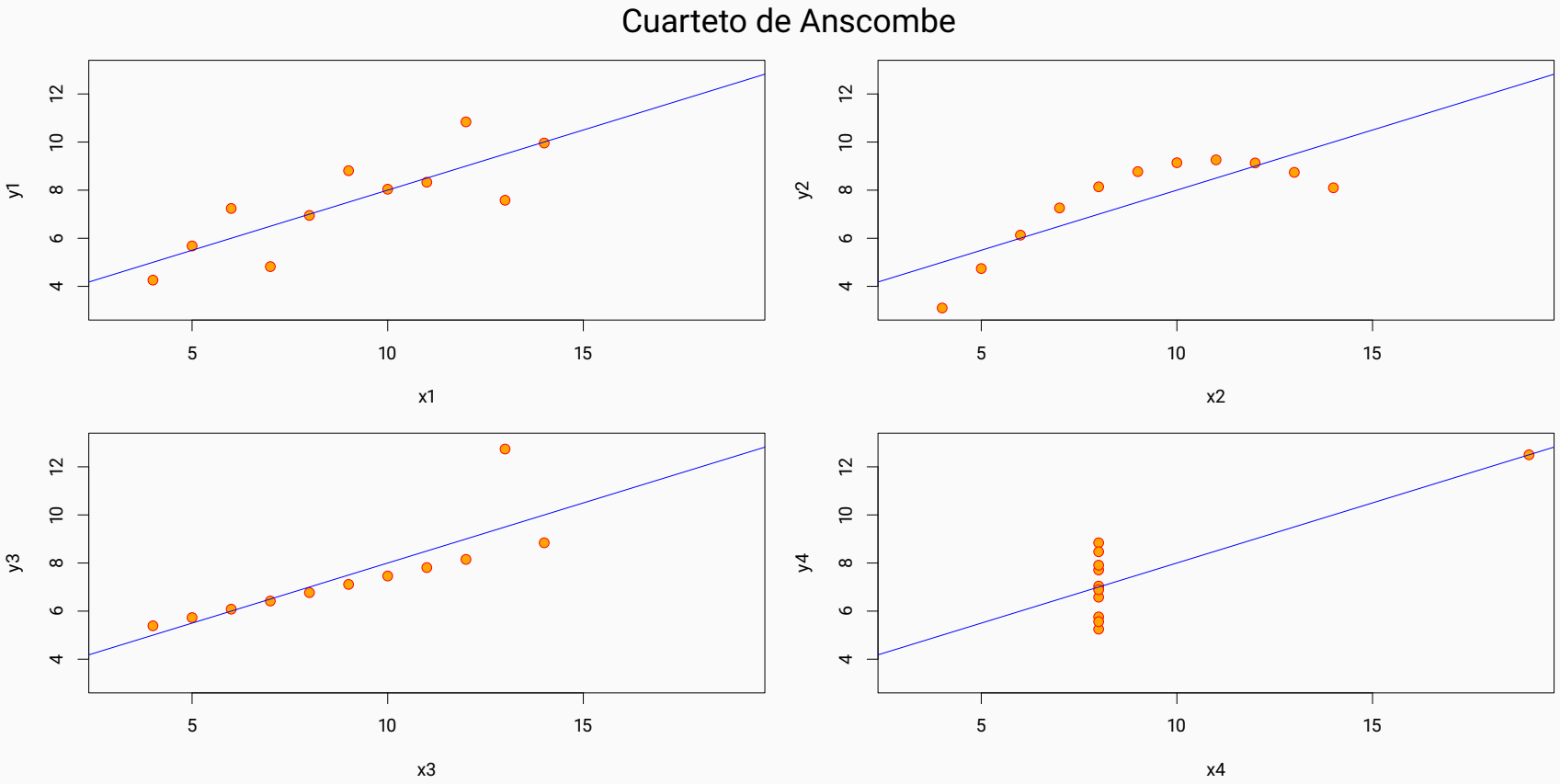
5. Confección del análisis: modelo apropiado.

6. Ejecución del modelo, extrayendo medidas resumen del modelo (tabla 2).

7. Presentación de resultados de manera dinámica (paquetes `rmarkdown` (Xie et al., 2020) o `shiny` (Wickham, 2021)).

¿Por qué incluir visualizaciones?

El mejor ejemplo que justifica esta práctica lo supone el cuarteto de Anscombe.



A continuación…

Lo que resta del día trabajaremos directamente con R.

- Concretamente, con el *script* `01_bivariado_utilidades.R` que está en el repositorio de **Github**.

Gracias por la atención

vergara_car@gva.es @carlos_verher @carlosvergara

Referencias bibliográficas i

Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.-I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J., Lamuela-Raventos, R. M., Serra-Majem, L., Pintó, X., Basora, J., Muñoz, M. A., Sorlí, J. V., Martínez, J. A., Fitó, M., Gea, A., … Martínez-González, M. A. (2018). Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts. *New England Journal of Medicine*, *378*, e34. https://doi.org/10.1056/NEJMOA1800389

Wickham, H. (2021). *Mastering Shiny : build interactive apps, reports, and dashboards powered by R*. O'Reilly. https://mastering-shiny.org/

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R Markdown Cookbook*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook