

# Tutorial de herramientas estadísticas avanzadas en R

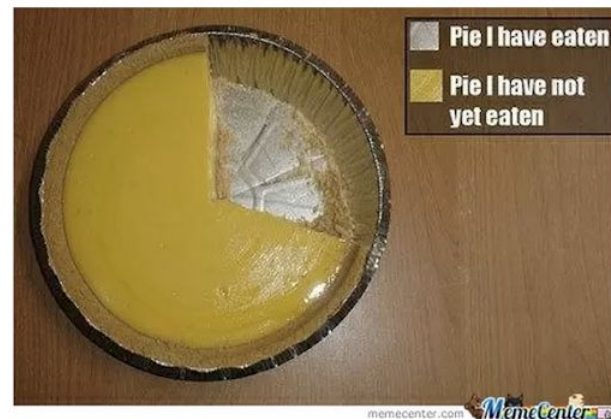


Daniel Herrera



Laboratorio de Neurociencias, Facultad de Ciencias, Udelar  
Centro Interdisciplinario de Ciencia de Datos y Aprendizaje Automático, Udelar

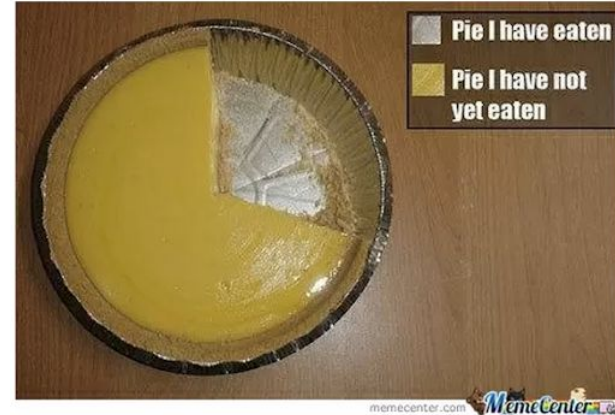
# ¿Para qué sirve la estadística?



# ¿Para qué sirve la estadística?

Respuestas incorrectas:

- Para que me publiquen el artículo. Necesito obtener un  $p < 0.05$
- Para poder usar métodos sofisticados, y que el trabajo parezca mejor
- Para usar métodos super complejos que compensen por datos que son malos o pocos



# ¿Para qué sirve la estadística?

Qué beneficios tienen para hacer ciencia:

- **Entender mejor los diseños experimentales.** Puede permitir diseñar mejores experimentos
- **Entender mejor las preguntas.** Me permite clarificar y formular mejor lo que quiero aprender de los datos.
- **Entender la estructura y las limitaciones de los datos.** Permite saber qué preguntas puedo hacerles (y cuáles no), y qué problemas pueden tener
- **Aplicar todo lo anterior a interpretar la literatura.**

# ¿Para qué sirve la estadística?

Qué beneficios tienen para hacer ciencia:

- Otro gran beneficio es aprender cómo/cuándo evitar la estadística



**Dr Will Harrison**  
@willjharrison

...

The more I master various statistical techniques and when to use them (or not), the more determined I become to design better experiments that don't need inferential statistics to draw conclusions.



**Sanjay Srivastava** @hardsci · Jul 21

There are a lot of statistical methods that are complicated and take a long time to learn. And that's a problem because they often also have a pretty high bar for when they're defensible to use, and no one who's invested the time wants to only ever say "nope not this time either"

# Objetivos de la clase

- Discutir una de las principales y más versátiles herramientas de estadística
- Ver ejemplos de implementación en R
- Aplicarlas a datos reales y simulados

La clase no va a ser hands-on porque insumiría mucho tiempo de clase

# Modelos lineales

Caja de herramientas de tests estadísticos



Modelos lineales y variaciones



# Modelos lineales

Los modelos lineales son modelos de la siguiente forma:

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \dots + x_n\beta_n + \epsilon$$

donde una variable dependiente  $y$  es modelada como una combinación lineal de variables independientes  $x_1, x_2, \dots, x_n$ , cada una con un coeficiente  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ . También se incluye en el modelo variabilidad (o ruido) representada por  $\epsilon$ .



# Modelos lineales

Los modelos lineales son modelos de la siguiente forma:

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \dots + x_n\beta_n + \epsilon$$

PD. Aunque la palabra “modelo” suene rimbombante, (casi) siempre que hacemos estadística estamos asumiendo un modelo de los datos. Es mucho mejor tener claro cuál es este modelo para entender lo que estamos haciendo.

# Modelos lineales

Un tipo de experimento muy común es medir una determinada **variable de respuesta**, y querer estudiar cómo esta cambia en **diferentes condiciones**, manipuladas experimentalmente u observadas.

En estos casos, una de las formas más simples de modelar nuestros datos es un modelo lineal:

$$\text{variable de respuesta} = \beta_0 + \text{condición}_1\beta_1 + \text{condición}_2\beta_2 + \dots + \text{condición}_n\beta_n + \epsilon$$

# Modelos lineales

## Ejemplos:

Mido tiempo de reacción (TR) en una tarea psicofísica, y voy variando el contraste del estímulo (Cr), así como su tamaño (Tm).

Posible modelo lineal:

$$TR = \beta_0 + \beta_1 Cr + \beta_2 Tm + \epsilon$$

# Modelos lineales

## Ejemplos:

Mido la cantidad de células fluorescentes (*Cel*) en hipocampo de ratón, con 2 tratamientos distintos (*Tr*), y extrayendo las muestras en diferentes momentos del desarrollo (*De*)

Posible modelo lineal:

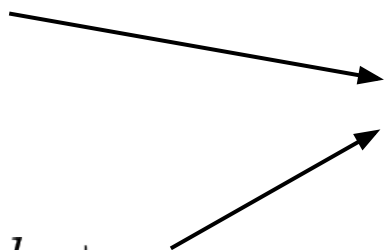
$$Cel = \beta_0 + \beta_1 Tr + \beta_2 De + \beta_3 (Tr \times De) + \epsilon$$

# Modelos lineales

Una vez que tengo mis datos, y que elegí el modelo, ajusto el modelo a los datos con algún software estadístico.

Lo que se hace en el ajuste, es que el software encuentra los valores de los  $\beta$  que mejor ajustan la relación entre las  $y$  y las  $x$ , dada la estructura del modelo. También calcula varias características del ajuste (que tan bueno es, cuanta varianza quedó sin explicar, etc).

Variable respuesta	Cond 1	Cond 2
y1	x11	x21
y2	x12	x22
y3	x13	x23
y4	x14	x24



Estimación de los  
coeficientes del modelo ( $\beta$ )

$$y = \beta_0 + \beta_1 Cond_1 + \beta_2 Cond_2 + \epsilon$$

# Modelos lineales

Vimos qué son, y cómo aplicar un modelo lineal. Pero, ¿para qué sirven? ¿qué hacemos ahora?

- El  $\beta$  asociado a cada condición nos da una medida de su efecto sobre  $y$
- Podemos comparar diferentes modelos estadísticamente. Esto nos permite testear la significancia estadística de diferentes condiciones (podemos obtener un p-valor).

# Modelos lineales

De hecho, muchos de los tests estadísticos clásicos son equivalentes a los modelos lineales. En vez de aprender un test para cada ocasión, con aprender modelos lineales alcanza para casi todo.

[https://lindeloev.github.io/tests-as-linear/#1\\_the\\_simplicity\\_underlying\\_common\\_tests](https://lindeloev.github.io/tests-as-linear/#1_the_simplicity_underlying_common_tests)

## Common statistical tests are linear models

Last updated: 02 April, 2019

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

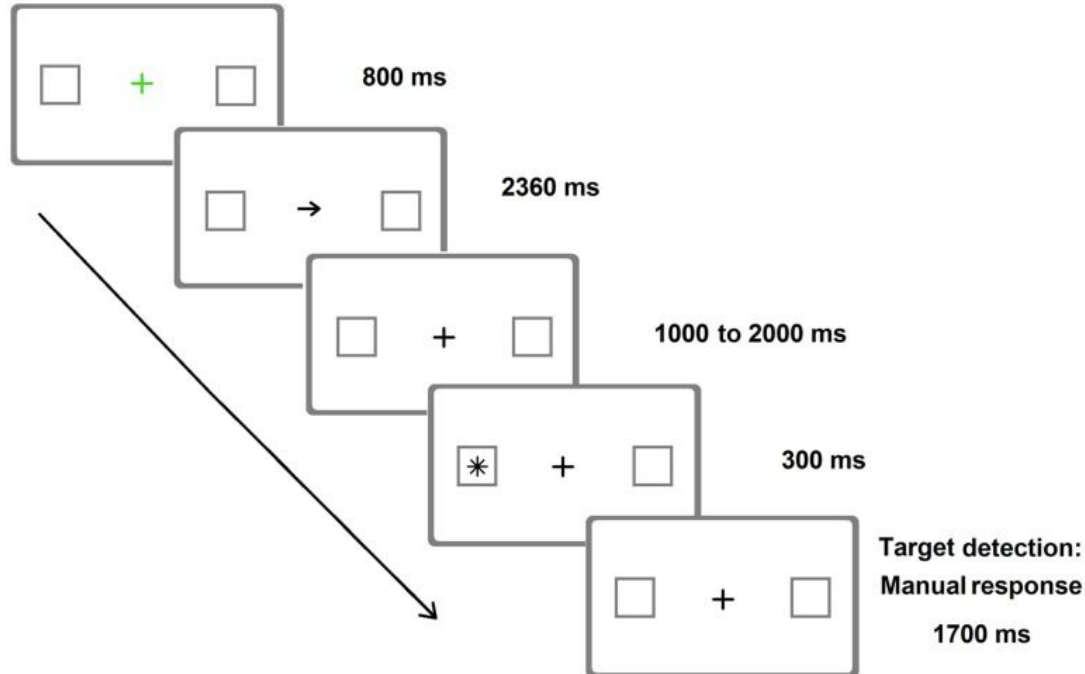
	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: $\text{lm}(y \sim 1 + x)$	<b>y is independent of x</b> P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	$\text{lm}(y \sim 1)$ $\text{lm}(\text{signed\_rank}(y) \sim 1)$	✓ for $N > 14$	One number (intercept, i.e., the mean) predicts y. - (Same, but it predicts the <i>signed rank</i> of y.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y1, y2, paired=TRUE) wilcox.test(y1, y2, paired=TRUE)	$\text{lm}(y_2 - y_1 \sim 1)$ $\text{lm}(\text{signed\_rank}(y_2 - y_1) \sim 1)$	✓ for $N > 14$	One intercept predicts the pairwise $y_2 - y_1$ differences. - (Same, but it predicts the <i>signed rank</i> of $y_2 - y_1$ .)	
	<b>y ~ continuous x</b> P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	$\text{lm}(y \sim 1 + x)$ $\text{lm}(\text{rank}(y) \sim 1 + \text{rank}(x))$	✓ for $N > 10$	One intercept plus x multiplied by a number (slope) predicts y. - (Same, but with <i>ranked x</i> and y)	
	<b>y ~ discrete x</b> P: Two-sample t-test N: Welch's t-test N: Mann-Whitney U	t.test(y1, y2, var.equal=TRUE) t.test(y1, y2, var.equal=FALSE) wilcox.test(y1, y2)	$\text{lm}(y \sim 1 + G_2)^4$ $\text{glm}(y \sim 1 + G_2, \text{weights} = \dots^8 y)$ $\text{lm}(\text{signed\_rank}(y) \sim 1 + G_2)^4$	✓ for $N > 11$	An intercept for <b>group 1</b> (plus a difference if <b>group 2</b> ) predicts y. - (Same, but with one variance <i>per group</i> instead of one common.) - (Same, but it predicts the <i>signed rank</i> of y.)	
	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_n)^4$ $\text{lm}(\text{rank}(y) \sim 1 + G_2 + G_3 + \dots + G_n)^4$	✓ for $N > 11$	An intercept for <b>group 1</b> (plus a difference if $\text{group} \neq 1$ ) predicts y. - (Same, but it predicts the <i>rank</i> of y.)	
Multiple regression: $\text{lm}(y \sim 1 + x_1 + x_2 + \dots)$	P: One-way ANCOVA	aov(y ~ group + x)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_n + x)^4$	✓	- (Same, but plus a slope on x.) <i>Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.</i>	
	P: Two-way ANOVA	aov(y ~ group * sex)	$\text{lm}(y \sim 1 + G_2 + G_3 + \dots + G_n + S_2 + S_3 + \dots + S_k + G_2 * S_2 + G_3 * S_3 + \dots + G_n * S_k)$	✓	Interaction term: changing <b>sex</b> changes the <b>y ~ group</b> parameters. <i>Note: <math>G_{2:n,k}</math> is an indicator (0 or 1) for each non-intercept levels of the group variable. Similarly for <math>S_{2:n,k}</math> for sex. The first line (with <math>G_2</math>) is main effect of group, the second (with <math>S_2</math>) for sex and the third is the <b>group * sex</b> interaction. For two levels (e.g. male/female), line 2 would just be '<math>S_2</math>' and line 3 would be '<math>S_2</math> multiplied with each <math>G_2</math>'.</i>	[Coming]
	<b>Counts ~ discrete x</b> N: Chi-square test	chisq.test(groupXsex_table)	<b>Equivalent log-linear model</b> $\text{glm}(y \sim 1 + G_2 + G_3 + \dots + G_n + S_2 + S_3 + \dots + S_k + G_2 * S_2 + G_3 * S_3 + \dots + G_n * S_k, \text{family} = \dots)^4$	✓	Interaction term: (Same as Two-way ANOVA.) <i>Note: Run glm using the following arguments: <code>glm(model, family=poisson())</code>. As linear-model, the Chi-square test is <math>\log(y) = \log(N) + \log(\alpha) + \log(\beta) + \log(\alpha\beta)</math> where <math>\alpha</math> and <math>\beta</math> are proportions. See more info in the <a href="#">accompanying notebook</a>.</i>	Same as Two-way ANOVA
	N: Goodness of fit	chisq.test(y)	$\text{glm}(y \sim 1 + G_2 + G_3 + \dots + G_n, \text{family} = \dots)^4$	✓	(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  $y \sim 1 + x$  is R shorthand for  $y = 1 + b + a \cdot x$  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  $\text{signed\_rank} = \text{function}(x) \text{ sign}(x) * \text{rank}(\text{abs}(x))$ . The variables G and S 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>.

<sup>4</sup> See the note to the two-way ANOVA for explanation of the notation.

<sup>8</sup> Same model, but with one variance per group: `glm(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.

# Ejemplo 1: Modelos lineales para analizar tiempos de reacción en test de Posner





# Modelos lineales generalizados (GLM)

- Los modelos lineales generalizados son una extensión de los modelos lineales que nos permiten analizar otro tipo de variables de respuesta y
- Ejemplo, datos de proporciones: Hago un experimento psicofísico con  $N$  ensayos por persona, y mido la cantidad de respuestas correctas. Si uso como  $y$  la cantidad de respuestas correctas se me pueden dar varios problemas. Por ejemplo, si cambio el  $N$  en un nuevo experimento, los datos ya no serían comparables.
- Ejemplo, datos de conteos: Tengo 2 grupos de ratones, unos con un tratamiento y otros sin. En cada ratón cuento cuantas neuronas “bebes” aparecen ( $B$ ), para medir neurogénesis. Si uso  $B$  como variable de respuesta en un modelo lineal me puede dar problemas. Ej. No es la misma variabilidad  $\epsilon$  en ratones con  $B \sim 3$  que en ratones con  $B \sim 20$ .

# Modelos lineales generalizados (GLM)

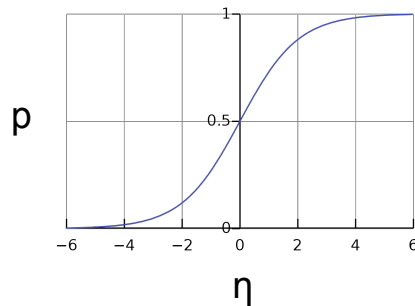
Los modelos lineales generalizados comienzan con un componente lineal (por eso se llama *modelos lineales*), pero le agregan modificaciones para adaptar la matemática al tipo de variable que tenemos.

Por ejemplo, para modelar las proporciones de la diapositiva anterior podemos usar una regresión logística (un tipo de GLM), que comienza con un componente lineal:

$$\eta = \beta_1 + \beta_2 Cond_1 + \beta_3 Cond_2$$

y luego le aplica una transformación logit-inversa, llevando el resultado al intervalo [0-1] para modelar la probabilidad del evento (ej. respuesta correcta) en las diferentes condiciones:

$$p = \text{logit}^{-1}(\eta) = \frac{e^{-\eta}}{1 - e^{-\eta}}$$

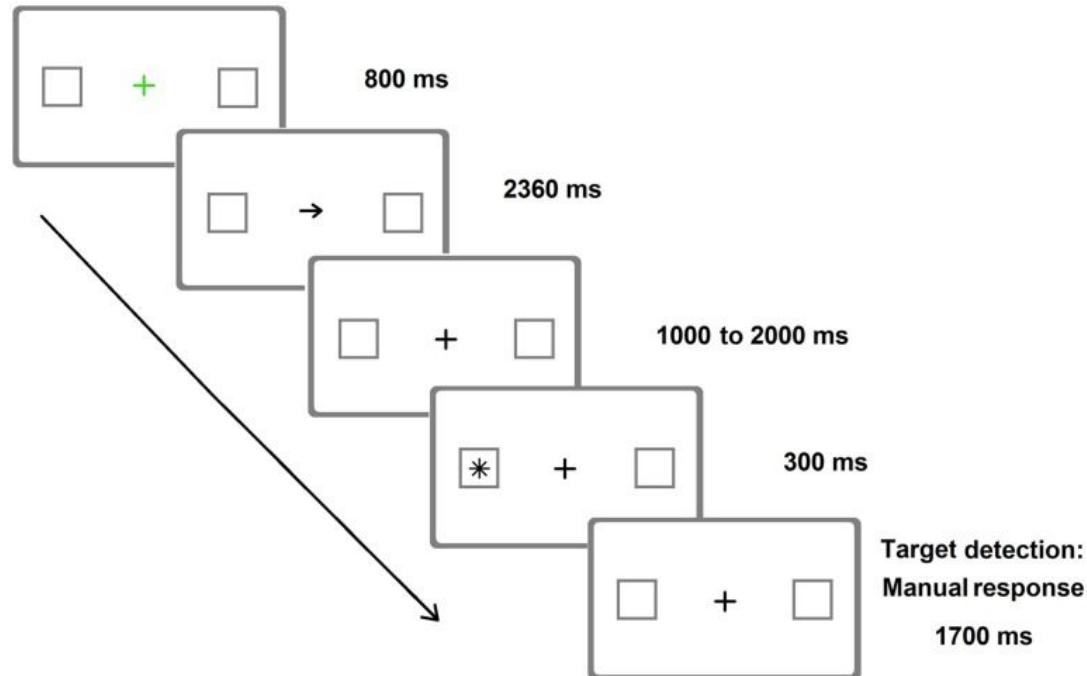


# Modelos lineales generalizados (GLM)

Ventaja sobre modelo lineal solo:

- Es lo correcto matemáticamente. Tenemos  $N$  ensayos y se dan  $E$  eventos, lo que queremos modelar es la probabilidad de  $E$ , no la cantidad de  $E$  (que puede variar si cambiamos  $N$ )
- Toma en cuenta correctamente la variabilidad de los datos. Con 100 ensayos y un  $p=0.5$ , el valor esperado de  $E$  es 50, pero es esperable que vea valores de 40 o 60. Sin embargo, con 100 ensayos y  $p=0.01$  el valor esperado de  $E$  es 1, y es muy poco probable que vea a  $E$  un valor de 10. En ambos casos los números absolutos de la variación son iguales, y un modelo lineal simple les daría probabilidad similar.
- Me permite comparar entre condiciones con diferentes  $N$
- Muchas ventajas más

## Ejemplo 2: Modelos lineales generalizados para analizar errores en test de Posner



# Experimentos de medidas repetidas: Variabilidad entre individuos (o células, o grupos, etc)

Hasta ahora no consideramos en nuestros análisis una característica importante del set de datos: Estamos midiendo diferentes personas, que pueden tener variaciones entre sí.

Esta es una característica muy común, que en general es bueno considerar. Por ejemplo, puede haber variabilidad entre los animales de los que tomo tejidos, las escuelas que incluyo en un análisis educacional, las personas que toman una encuesta psicométrica, o entre los peces que uso para medir comportamiento.

Generalmente, esta es variabilidad que no es el efecto que me interesa en sí, pero que quiero tener en cuenta (ej. mi resultado  $p < 0.05$  se debe a un solo ratón con valores atípicos?)

# Experimentos de medidas repetidas: Variabilidad entre individuos (o células, o grupos, etc)

Intuitivamente podemos pensar en varias formas de solucionar esto, por ejemplo, restarle la media a cada sujeto (o clase, o animal, etc).

Los modelos lineales (y GLMs) nos permiten incluir los efectos de las diferentes personas de forma natural. Una forma de hacerlo, con las herramientas ya vistas, es usar un intercept ( $\beta_0$ ) diferente para cada persona. Esto es equivalente a restarle la media a cada sujeto.

# Modelos mixtos

Si tenemos medidas repetidas, una mejor forma de modelar la variabilidad entre medidas es usar modelos mixtos.

En estos modelos, en lugar de ajustar un intercepto independiente a cada unidad de medida (sujeto, ratón, escuela, etc), se asume que los interceptos pueden variar, pero que salen de una distribución (en general Gausiana) que describe a la población de la que se tomó la muestra (la población de estudiantes de Fcien de donde se reclutó sujetos, la población de escuelas en Montevideo de las que se tomaron las muestras, etc).

Considera la variabilidad, y permite ver diferencias entre grupos (a diferencia de restar las medias a cada individuo). Permite agregar variabilidad en todos los coeficientes del modelo que querramos y que tenga sentido matemático.

# Modelos mixtos

Se llaman modelos mixtos porque tienen lo que se llaman efectos fijos y efectos aleatorios.

Matemáticamente:

$$y^i = \beta_0^i + \beta_1^i Cond_1 + \epsilon$$

donde  $\beta_0^i \sim N(\beta_0, \sigma_0)$      $\beta_1^i \sim N(\beta_1, \sigma_1)$

El modelo estima las medias de los efectos en la población, o efectos fijos,  $(\beta_0, \beta_1)$ , y también estima cómo estos varían entre las personas (o ratones, o escuelas), dado por los parámetros  $\sigma_0$  y  $\sigma_1$



# Modelos mixtos

Se están convirtiendo en una herramienta de análisis aceptada y común.

## Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models

[TF Jaeger](#) - Journal of memory and language, 2008 - Elsevier

This paper identifies several serious problems with the widespread use of ANOVAs for the analysis of categorical outcome variables such as forced-choice variables, question-answer accuracy, choice in production (eg in syntactic priming research), et cetera. I show that even after applying the arcsine-square-root transformation to proportional data, ANOVA can yield spurious results. I discuss conceptual issues underlying these problems and alternatives provided by modern statistics. Specifically, I introduce ordinary logit models (ie logistic ...

☆ 99 Cited by 3215 Related articles All 16 versions

## [HTML] Random effects structure for testing interactions in linear mixed-effects models

[DJ Barr](#) - Frontiers in psychology, 2013 - frontiersin.org

In a recent paper on mixed-effects models for confirmatory analysis, Barr et al.(2013) offered the following guideline for testing interactions:“one should have byunit [subject or item] random slopes for any interactions where all factors comprising the interaction are within-unit; if any one factor involved in the interaction is between-unit, then the random slope associated with that interaction cannot be estimated, and is not needed”(p. 275). Although this guideline is technically correct, it is inadequate for many situations, including mixed ...

☆ 99 Cited by 402 Related articles All 14 versions 99

## The anova to mixed model transition

[MP Boisgontier](#), [B Cheval](#) - Neuroscience & Biobehavioral Reviews, 2016 - Elsevier

A transition towards mixed models is underway in science. This transition started up because the requirements for using analyses of variances are often not met and mixed models clearly provide a better framework. Neuroscientists have been slower than others in changing their statistical habits and are now urged to act.

☆ 99 Cited by 197 Related articles All 12 versions

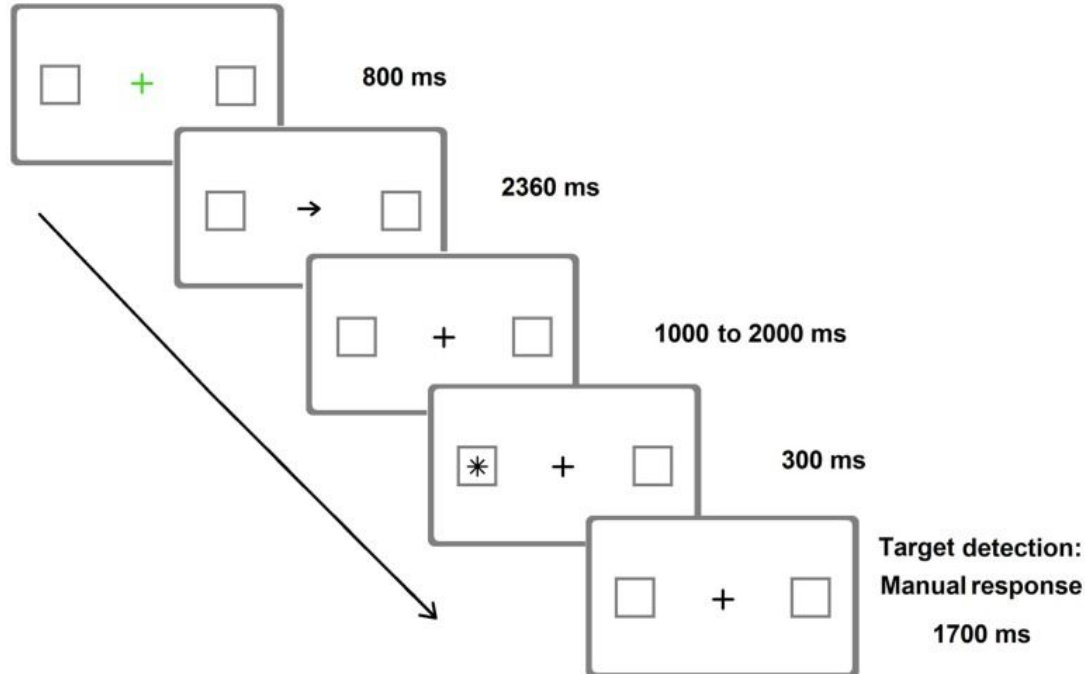
## Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem.

[CM Judd](#), [J Westfall](#), [DA Kenny](#) - Journal of personality and social ..., 2012 - psycnet.apa.org

Throughout social and cognitive psychology, participants are routinely asked to respond in some way to experimental stimuli that are thought to represent categories of theoretical interest. For instance, in measures of implicit attitudes, participants are primed with pictures ...

☆ 99 Cited by 945 Related articles All 12 versions

# Ejemplo 3: Modelos lineales generalizados para analizar errores en test de Posner



# Modelos mixtos

Los modelos mixtos son más difíciles de ajustar que los modelos lineales clásicos. Requieren más datos (ej. más personas) y pueden dar error si nuestros datos tienen particularidades raras (ej. grandes outliers, u otras)

¿Cuándo usar modelos mixtos? No siempre es viable usar modelos mixtos, quizás es muy caro recolectar datos de nuevas personas, nuevas escuelas, o probar con diferentes camadas de ratones, etc, y los datos disponibles no me permiten ajustar un modelo mixto. Además, incluir más efectos aleatorios (ej. 3 o 2 en vez de 1) exige bastantes más datos.

Pero eso me da información valiosa: si el modelo no ajusta bien, significa que mis datos por sí solos no son suficientes para estimar la variabilidad y extrapolar por fuera de mi muestra. Es válido aceptar esto, puedo reportar mis estimados para mi muestra (sin efectos aleatorios, o con menos efectos aleatorios) y discutir porqué o porqué no está bien extrapolar (ej. hay que extrapolar porque es muy caro, no queda otra)

# Modelos mixtos

Los modelos mixtos son conceptualmente más complicados que los modelos sin efectos aleatorios, pero eso es porque tienen más flexibilidad para ajustar a las complicaciones del mundo y de los diseños experimentales. No siempre es fácil identificar cómo modelar nuestros datos, pero el ejercicio de pensar en esto nos hace entender mejor lo que estamos haciendo experimentalmente.

Lo que vimos hasta ahora también se sigue extendiendo (ej. modelos jerárquicos). Estas extensiones suelen darse porque responden a problemas de análisis. Es probable que si tenés datos complicados que no sabés cómo analizar, alguna versión de estos modelos te pueda ayudar a entender los posibles caminos y sus limitaciones.

Los científicos tienen que entender nociones básicas de estadística (cuanto más mejor), pero no tienen que ser expertos, para eso hay expertos en estadística. Asesorate cuando estás con un estadístico cuando estás empesando tu proyecto (y tenerlo de co-autor) te puede ahorrar mucho tiempo, recursos y dolores de cabeza.

# Material extra

Si te interesan estos temas, Google está lleno de tutoriales de linear models, generalized linear models, y mixed effects models.

<https://www.rensvandeschoot.com/tutorials/lme4/>

<https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf>

[https://jontalle.web.engr.illinois.edu/MISC/lme4/bw\\_LME\\_tutorial.pdf](https://jontalle.web.engr.illinois.edu/MISC/lme4/bw_LME_tutorial.pdf)

# FIN

Quizás si consultan con un estadístico bajan estos números

