De pruebas t a regresión a ANOVA

Christian DiCanio
cdicanio@buffalo.edu

University at Buffalo - Department of Linguistics

25/6/18

Pruebas t

Una prueba t es un método de evaluar la probabilidad que un promedio de una muestra viene de la misma distribución de un promedio predecido usando la distribución t.

El nivel de la prueba es, por defecto, 0.05 en R.

Para un análisis dado, podemos crear una muestra y usar t.test() para determinar donde se ubican los intervalos de confianza de 95 %.

muestra <- rnorm(10, mean = 50, sd = 5)
t.test(muestra)$conf.int
Véase código #1.

Intervalos de confianza

El número "95 %" refleja los promedios hipotéticos; no es el caso que una muestra normal contendrá el promedio de la población; 5 % no lo contendrá.

Puede ser que tenemos mala suerte y descubrimos que nuestra muestra no es representativa de la población.

Un CI (intervalo de confianza) en 95 % de las muestras repetidas contentrá el promedio de la población.

Tomando muestras

Ahora qué sabemos?

- Es más probable que el promedio de una muestra aleatoria está cerca del promedio de la población que no.
- El límite de la distribución de la muestra es igual al parametro de la población. Cuando n se acerca a la infinidad, la distribución de la muestra se acerca a la distribución de la población.
- No sabemos σ, pero lo podemos estimar usando SE y podemos inferirlo usando una distribución parecida a la distribución normal (t).

Qué es p?

Refleja una probabilidad condicional; es la probabilidad de observar un promedio de una muestra dada asumiendo que la hipótesis nula es la verdad.

Es una medida de verosimilitud.

Cómo comparamos dos muestras?

Hasta ahora, hemos examinado solamente una sola muestra en una prueba de t. Se llama una prueba t de una muestra.

Normalmente nos interesa si una muestra es diferente de otra muestra.

Cuál es la probabilidad que una muestra viene de otra? Se llama una prueba t de dos muestras.

Comparamos dos muestras:

Cuadro: Tiempo de leer con adultos y niños

Grupo	Tamaño de muestra	$\overline{x}$ (s)	s
niños	$n_1 = 10$	$\overline{x}_1 = 30$	$s_1 = 43$
adultos	$n_2 = 20$	$\overline{x}_2 = 7$	$s_2 = 25$

Los tamaños de las muestras, sus promedios para leer un grupo de cláusulas y sus sd son diferentes.

Propiedades de pruebas t de dos muestras

- La meta es comparar las respuestas de dos grupos.
- Cada grupo es una muestra de una población distinta (diseño entre sujetos)
- Las respuestas de cada grupo son independientes de los que observamos del otro grupo.
- Los tamaños de las muestras pueden ser diferentes.

Cuál es la hipótesis nula cuando comparamos dos grupos?

Asumimos que los promedios de cada grupo son iguales.

H_0: $\mu_1 = \mu_2$
μ_1 - $\mu_2 = 0$
$\delta = 0$

No sabemos los promedios de la poblaciones pero podemos aproximarlos con una estadística, d.

d = $\overline{x}_1$ - $\overline{x}_2$

Debemos incluir n con pruebas de t.

$t = (\overline{x} - \mu)/s_{\overline{x}}$ where $s_x = s/\sqrt{n}$.

Cómo se calcula N si son diferentes a través de las muestras? Es la suma.

$$t = (\overline{x} - \overline{y})/(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$$

$$t = (30 - 7)/(\sqrt{\frac{43^2}{10} + \frac{25^2}{20}})$$

(Véase código #2.)

Cuál es el output? Qué significa?

Está en términos del número de desviaciones estandáres del promedio.

Podemos calcular el valor de p para nuestro output usando la función pt().

Cuántos grados de libertad tenemos? Cómo lo calculamos?
El df con esta prueba es la suma de los df de cada muestra: $df_1 = 9$; $df_2 = 19$; 28 entonces.

Nuestros resultados significan que tenemos una probabilidad de 6.4 % que las muestras vienen de la misma distribución.

Es más alto de nuestro nivel de α (0.05).

Debemos concluir que las muestras no son diferentes estadísticamente (no significativo).

Otro ejemplo con datos reales

Investigamos nuestros datos de VOT otra vez.

Nuestra hipótesis es que el promedio de VOT de /k/ es igual al promedio de VOT de /k^{w}/.

Véase código #3. Cómo varían los parametros para cada muestra?

Normalmente no necesitamos extraer todos los parametros para hacer una prueba t. R lo hace.

t.test(data.k$VOT, data.kw$VOT); véase código #4.

Cuál es la probabilidad que nuestros promedios de las muestras vienen de la misma distribución?

Qué podemos concluir?

Corelación y Covarianza

Hemos platicado de promedios y desviaciones estandares.

También nos interesan saber como una distribución de una muestra varía al respecto de otra distribución de una muestra.

Una manera de explorar la relación entre dos factores es por mirar totales en una tabla de contingencias.

Una muestra - valores promedios de F1 de hombres y mujeres a través de idiomas diferentes. (Abra los datos "F1_data.txt".)

Véase código #5 (la parte arriba).

Qué pasa cuando comparamos dos grupos de valores de F1?

Por dibujar los valores en esta manera, podemos examinar si las tendencias en una distribución dada corresponden a valores en otra distribución.

Nos interesa si las desviaciones de una distribución tienen la misma magnitud de las desviaciones de otra distribución.

Asumimos que podemos comparar estas muestras y que cada punto está emparejado. Para evaluar esto, necesitamos tener el mismo N de cada muestra.

Los productos de las desviaciones de los promedios para dos muestras de valores de F1:

$$\sum_{i=0}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Este producto debe ser sensible al tamaño del grupo de datos. Entonces tomamos un promedio dividido por n.

$$\frac{\sum_{i=0}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

El producto promedio de las desviaciones se llama **la covarianza** de X y Y.

Nótese que la covarianza es sensible al tipo de medida que analizamos (si no está normalizado). Si queremos comparar diferentes tipos de medidas, debemos normalizar los datos (pero R lo hace automáticamente).

La suma de los productos de las desviaciones normalizadas (z), dividido por el tamaño de la muestra nos da el valor r_{xy}. Se llama **el coeficiente de corelación**.

La covarianza es igual a la covariación. El último usa una escala normalizada.

El rango de corelación: -1 (negativo) a 0 (ningún) a 1 (positivo).

La línea de regresión

La línea de regresión es la mejor estimación de mínimos cuadrados del ajuste linear de dos muestras de datos.

$$\hat{y}_i = BX + A$$

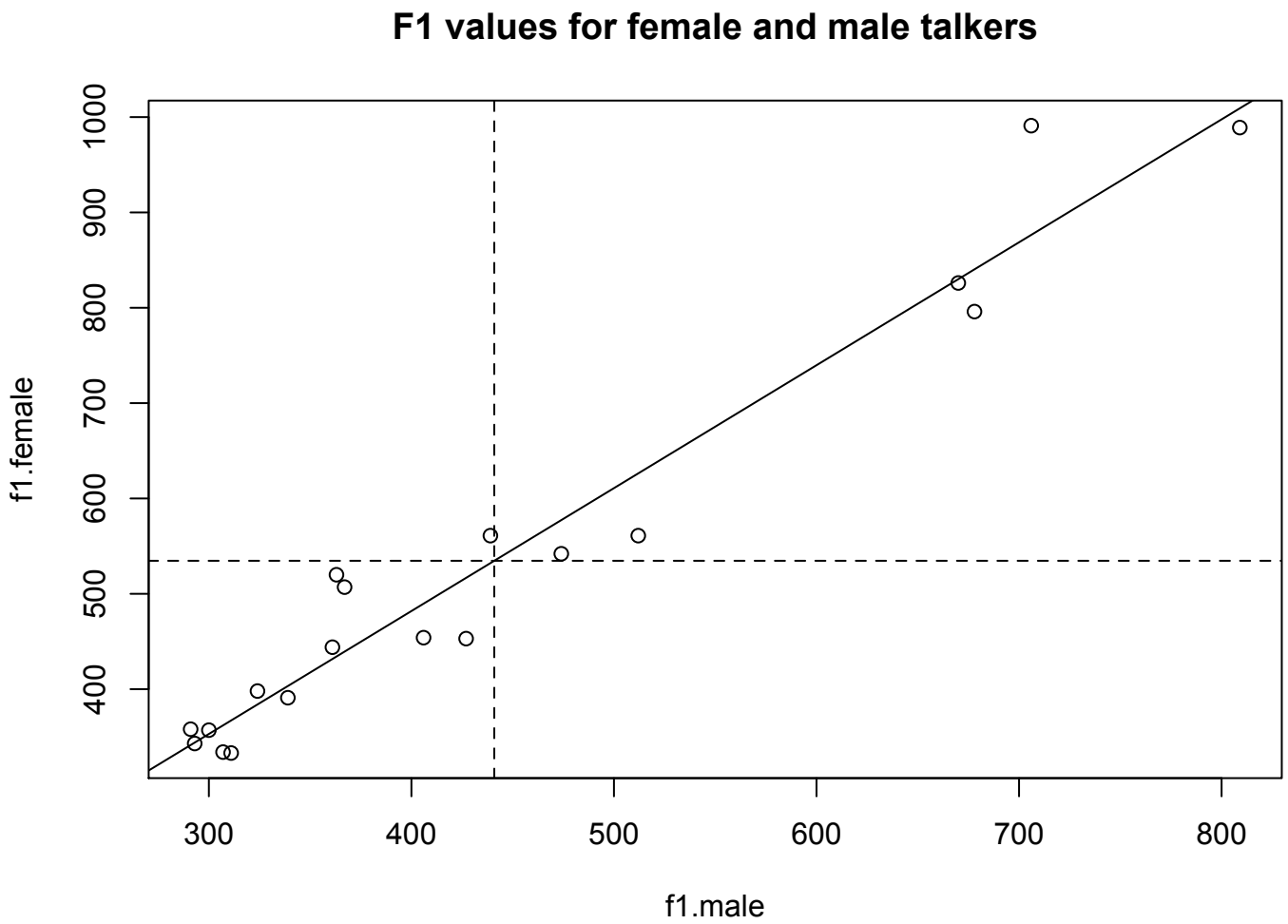
Intentamos minimizar las desviaciones cuadradas entre los valores predecidos y observados en trazar una línea.

El ajuste de una función linear está basado en el matriz de corelación (que tanto están corelados los valores).

Nuestra muestra de F1:

1. B = cor(male, female)*(sd(male)/sd(female))
2. A = mean(male) - B*mean(female)

A es nuestra intersección y B refleja nuestra pendiente.


F1 values for female and male talkers
f1.female
400
500
600
700
800
900
1000
f1.male
300
400
500
600
700
800

La línea de regresión refleja que tanto podemos predecir X en términos de Y.

También es un modelo de la relación entre dos muestras. Podemos formar predicciones basadas en el ajuste a nuestros datos. Por ejemplo, qué sería un valor de F1 de hombres si tenemos un valor de mujeres?

Véase código #6.

La varianza en modelos lineares

El ajuste de un model linear se capta por el valor de corelacción, r, que se llama *Pearson's product moment correlation*.

La varianza explicada en el modelo es r^2. Porque r_{xy} está normalizado esta valor refleja una proporción.

Corelaciones mejores (o ajustes mejores) tienen mejores coeficientes y por cuadrarlas, eliminamos números negativos.

Raras veces observamos una corelación perfecta. Entonces siempre hay una proporción de la varianza de la muestra que no está predecida de nuestro modelo.

Nos interesa tener un ajuste bueno porque queremos observar las relaciones entre los factores.

Otra manera de evaluar nuestro ajuste es examinar el **residuo**. Es la proporción de varianza en nuestro modelo que no está explicada.

El residuo

La varianza en regresión es una combinación de la varianza explicada del modelo y el residuo.

La varianza no predecida es la varianza de la desviación entre los valores observados y predecidos.

Es una proporción y se puede observar los valores cuando se examina un modelo linear.

Funciones en R para modelos lineares

La función *lm()* crea un modelo linear entre un factor y uno o más factores.

El sintáxis: lm(x∼y, data=mis.datos), donde *x* es la variable dependiente y *y* es la variable independiente.

summary() de un modelo linear nos dice los parametros del modelo y su ajuste.

Ejemplo usando lm() con los datos de F1:

Cuales son las hipótesis que investigamos con esta estadística?

Qué significan los números?

Parametros del modelo linear

- Residuos = las desviaciones entre los valores predecidos del modelo y lo que observamos.
- Coeficientes (Estimaciones) = el pendiente y la intersección de valores en el model linear
- r^2 = la proporción de varianza que explica el modelo (el ajuste)
- Las pruebas de t evaluan si la intersección y el pendiente son diferentes de "0", nuestra hipótesis nula.

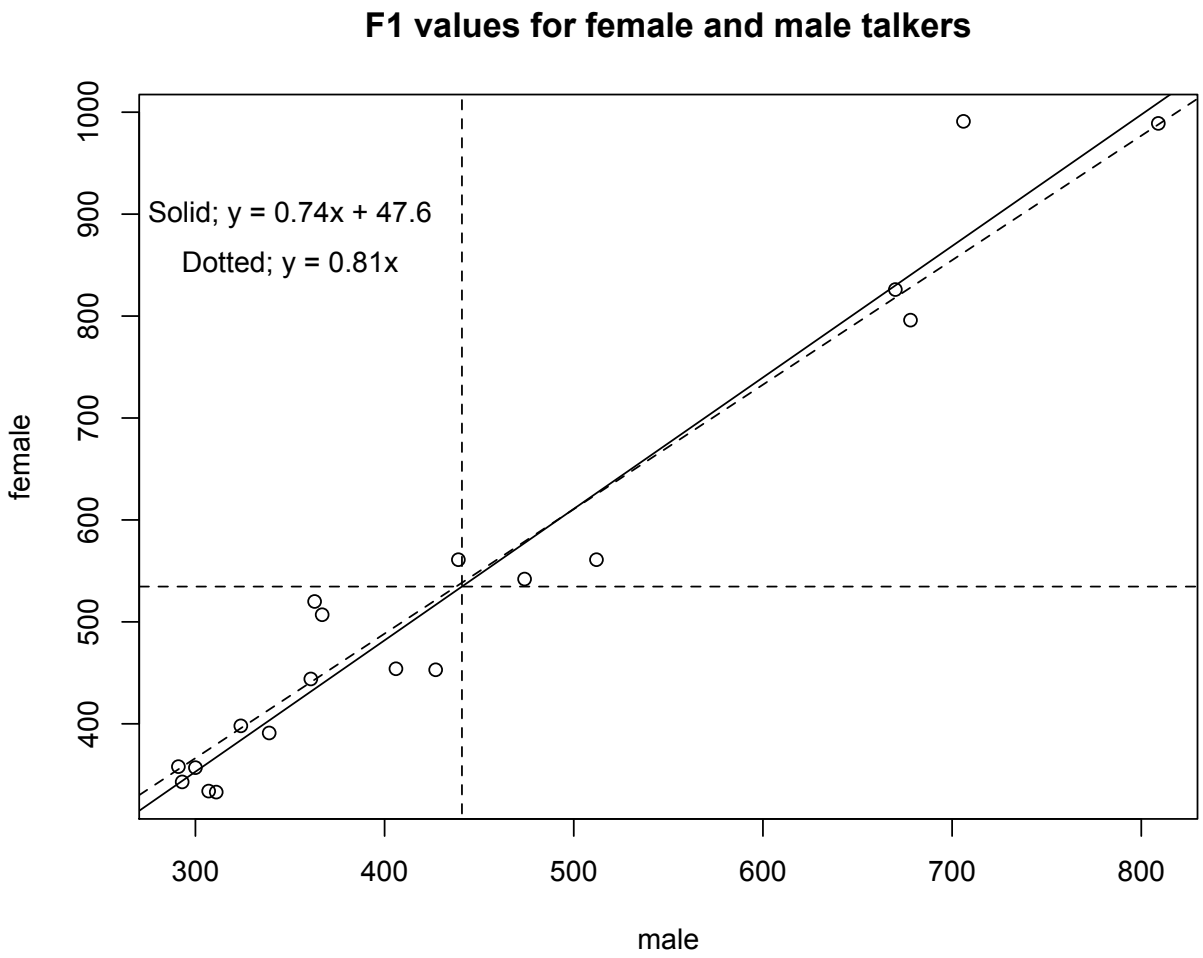
A ver los valores...

La intersección

Parece que el valor de t está encima de nuestro nivel de α. Sugiere que no debemos rechazar la hipótesis nula que la intersección es diferente de "0."

Podemos evaluar models multiples y determinar si uno sin una intersección es diferente de un modelo con una intersección.

Véase código #8 (arriba).



Los ajustes lineares no nos aperecen muy diferentes. Hay maneras de evaluar si los ajustes son diferentes (pero dejamos esta discusión para mañana).

Vamos a evaluar las relaciones entre componentes diferrentes de oclusivas en Triqui de Itunyoso.

El idioma posee oclusavas fortis (geminadas) y lenis (sencillas). Para cada oclusiva, examiné la duración del cierre, de la soltura y de la aspiración.

Ejercicio

Sigue el código #9 para importar los datos en RStudio. Dividimos los datos en dos partes: oclusivas lenis y fortis.

1. Para cada subconjunto, evalúe un modelo linear donde la duración de aspiración es la variable dependiente y la duración del cierre es la variable independiente.
2. Examine la varianza del modelo, el valor de t y el valor de p.
3. Ahora, la duración de la soltura será la variable independiente. Evalúe un modelo de este tipo.
4. Examine la varianza del modelo, el valor de t y el valor de p.

ANOVA

Hasta ahora examinamos dos factores continuos, pero normalmente queremos entender como funciona un factor continuo al respecto de factores discretos.

Si tenemos un factor de 3 niveles: A, B, C y un factor dependiente, p.ej. la duración. Nos interesa saber si nuestro factor tiene un efecto en la duración.

Si hacemos una prueba de t, contestaríamos nuestra pregunta? Cuales son las comparaciones?

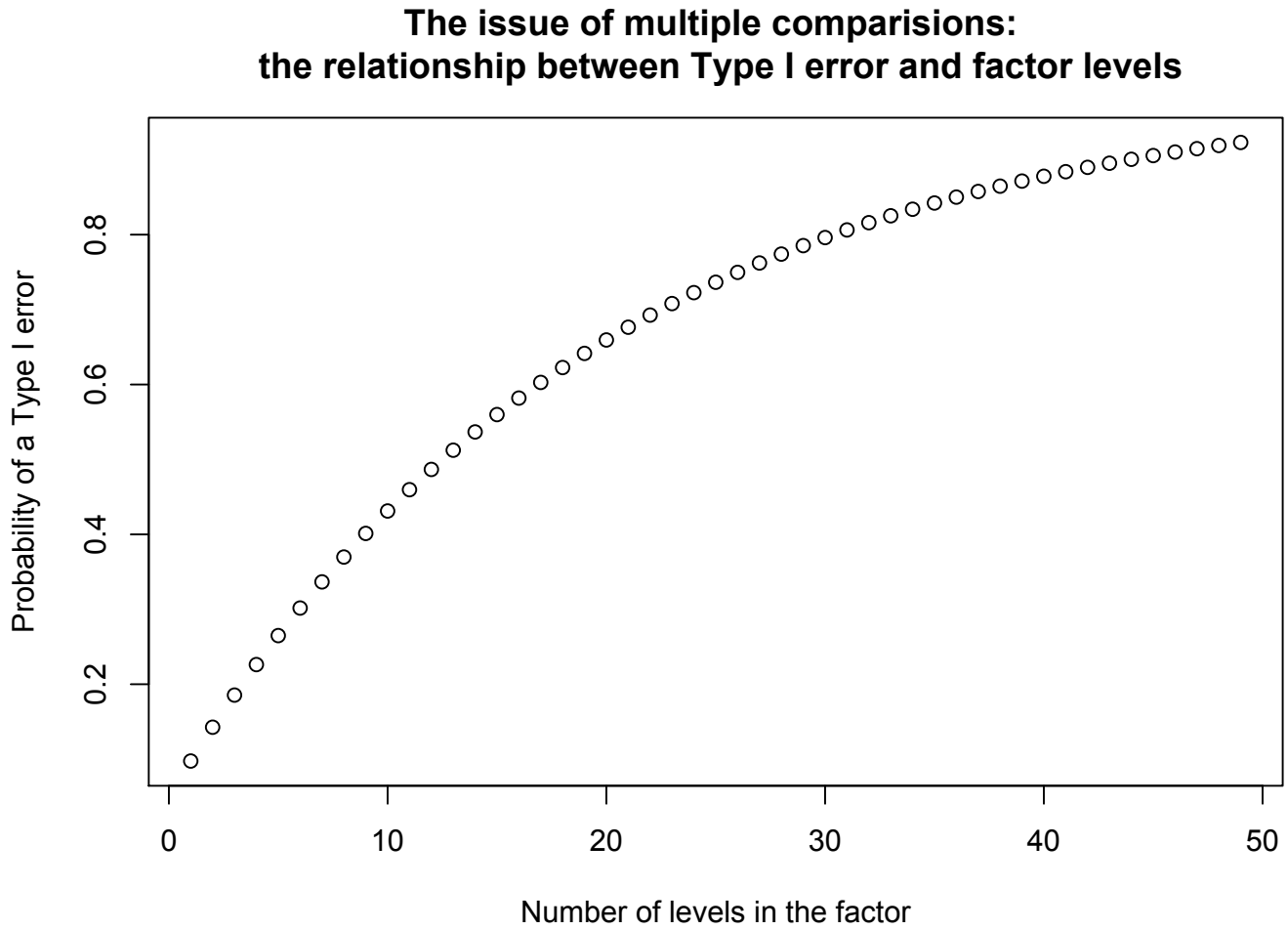
A vs. B; A vs. C; y B vs. C

Comparaciones multiples

Puede ser que encontramos una diferencia entre cada uno de estas comparaciones, pero cuando aumentamos el número de niveles de nuestro factor, aumentamos la probabilidad de producir un error Tipo I.

La probabilidad de rechazar al menos una hipótesis nula es la suma de probabilidades de cada evento mutuamente exclusivo.

Nuestro nivel de α no es 0.05 para este factor, sino 0.143. Por qué?



La probabilidad de al menos un éxito (de los tres) = 1 - la probabilidad de ningún éxitos.

1 - pbinom(0, 3, 0.05)

La probabilidad del éxito de *al menos un* experimento es 0.14, no 0.05. Se pone más probable que nuestros resultados de pruebas t resultarán por chanza.

Podemos bajar α, pero eso reduce nuestro poder. Qué hacemos? ANOVA.

ANOVA de un solo factor

La prueba bien aceptada para el análisis de variables continuas con variables discretas.

Ejemplo de Pitt & Shoaf (2002), de Johnson (2008): Los sujetos oyen un prime y un blanco. El sujeto repite el blanco. El prime varía en su grado de traslapar fonológicamente con el blanco.

Tomamos 96 ejemplos (1 cada hablante) para examinar que tanto demora la respuesta del sujeto por con el grado de traslapar (0, 3 fonemas) y por cuando contestaron las preguntas en el experimento (al inicio, en el medio, al final).

Datos de una muestra - Johnson (2008)

Sujeto	Demora	Tratamiento
x1	1179	early
x2	729	early
x3	845	mid
x4	804	mid
x5	744	late
...	...	...

Nuestra hipótesis es que los sujetos contestaron más rápidamente al experimento más tarde. Esto es el "Tratamiento."

Independencia

Cada observación (sujeto) es independiente de las otras (no hay medidas repetidas).

Tomamos solamente *una* observación de cada sujeto.

Típicamente no se lo hace (usamos análisis de medidas repetidas), pero asumimos que cada observación es independiente acá.

Examinar promedios de las muestras

Un promedio de una muestra es la suma del promedio de la población y alguna varianza de error.

$$\overline{x}_1 = \mu - \epsilon_1$$

ϵ_1 refleja que tanto nuestra muestra desvía del promedio de la población. Tiene distribución normal.

Podemos fijar el promedio de la población y ajustar ϵ para examinar como nuestro model corresponde a los datos.

Con ANOVA, nos interesa cuanta varianza observamos dentro de un grupo y cuánta varianza observamos entre cada grupo.

$y = \mu + \alpha + \epsilon$
$y =$ *promedio de muestra* $+$ *desviación de nivel de* μ $+$ *resto*

Parece mucho al modelo de regresión!

Resulta que podemos modelar la contribución de cada nivel de nuestro tratamiento como sigue. Para los 3 niveles, la medida de RT al inicio es:

	Beginning	Middle	End	Overall
$\overline{x}_i$	888	805	741	$\overline{x}.. = 810$
$\overline{x}_i - \overline{x}..$	78	-4.6	-69	
$(\overline{x}_i - \overline{x}..)^2$	6,111	21.1	4,759	sum = 10,891

$$RT_{beg} = \mu + \tau_{beg} + \epsilon_{beg,j}$$
$$RT_{beg} = 810 + 78 + \epsilon_{beg,j}$$

H_0 = las diferencias entre los promedios de niveles = 0.

$\tau_{beg} = \tau_{mid} = \tau_{end}$

ANOVA evalúa H_0 por comparar un model con el tratamiento con otro modelo sin el tratamiento.

Se anticipa que la magnitud de la desviación de variación aleatoria es comparable a la magnitud de desviación de los efectos del tratamiento. Si es la verdad, no hay una diferencia significativa.

Evaluar la varianza

Comparamos la varianza aleatoria (varianza dentro del factor) a la varianza del tratamiento (varianze entre-factor). Crea un ratio.

Calculamos la suma de cuadrados (SS) y el promedio de cuadrados (MS) de los dos tipos de varianza.

$$SS_{entre} = r\Sigma(\overline{x}_i - \overline{x}..)^2$$

r es el grado de libertad para **cada** nivel. ANOVA **asume que los datos son balanceados** dentro de los niveles.

Cuadrados del promedio

De la tabla arriba ya sabemos que $\Sigma(\overline{x}_i - \overline{x}..)^2 = 10,891$. Entonces, $SS_{entre} = 348,512$

$$MS_{entre} = \frac{SS_{entre}}{df}$$

$$MS_{entre} = \frac{348512}{2} = 174,256$$

Si MS = 174,256, entonces nuestra H_0 es que la desviación de error será aproximadamente igual.

SS_{error}

Cómo separamos la varianza del tratamiento y la varianza aleatoria?

Restamos $SS_{tratamiento}$ de SS total. Se encuentra SS total por agregar las desviaciones cuadradas entre cada dato y el promedio sin tomar en cuenta su tratamiento.

$SS_{total} = \Sigma(x_{ij} - \overline{x}..)^2 = 3{,}483{,}523$
$SS_{adentro} = SS_{total} - SS_{adentro} = 3,136,368$

Si tenemos $SS_{adentro}$, podemos calcular $MS_{adentro}$

$$MS_{adentro} = \frac{SS_{adentro}}{df}$$
$$MS_{adentro} = \frac{3{,}136{,}368}{93} = 33{,}724$$

$$DF_{adentro} = DF_{total} - N_{entre}$$
$$DF_{adentro} = 96 - 3$$

La estadística F

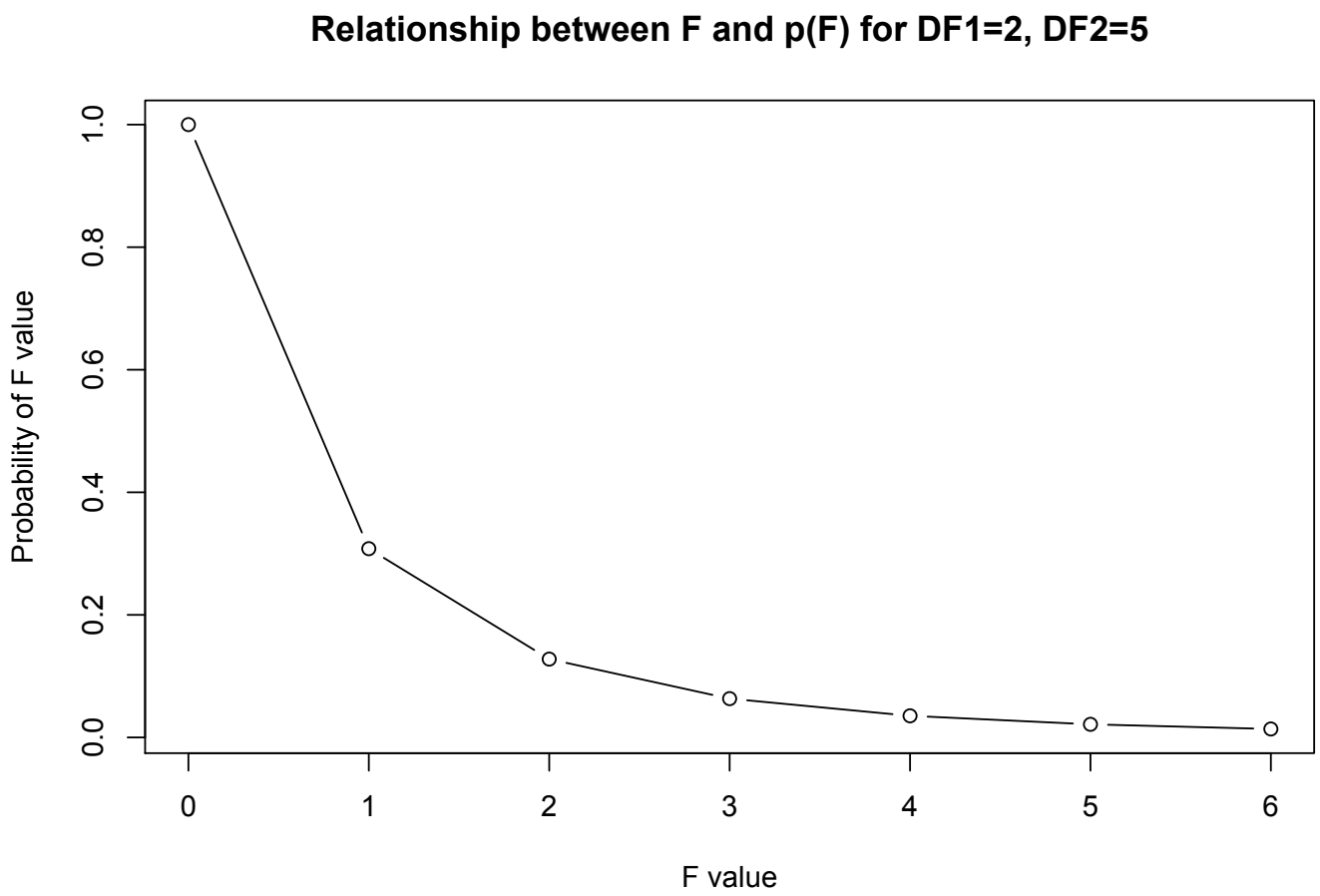
Todas las observaciones se consiste de una porción que es el efecto del tratamiento y una porción que es el efecto aleatorio. Queremos determinar si la agrupación que tenemos con niveles del tratamiento explica más varianza de lo que es aleatoria.

Es decir, hay una ventaja de incluir el tratamiento?

Nuestro modelo es: $x_{ij} = \mu + \tau_i + \epsilon_{ij}$,
pero H_0 es $x_{ij} = \mu + \epsilon_{ij}$

El ratio de MS = F. $F = \frac{MS_{entre}}{MS_{adentro}}$

La distribución F



Nuestra H_0 es que los efectos de tratamiento son 0. El valor de F debe ser cerca de 1.

$$F = \frac{MS_{entre}}{MS_{adentro}} \qquad F = \frac{174{,}256}{33{,}724} = 5{,}17$$

F es más grande de 1. Qué nos dice?

Cómo calculamos p?

Qué es la probabilidad que podríamos echar a suertes un valor F 5.17 de una población donde las diferencias entre los tratamientos fuera 0, $df_1 = 3$, y $df_2 = 93$?

pf(5.17, 3, 93, lower.tail=FALSE) = 0.0024

Un valor significativo de F indica que el modelo con el mejor ajuste estadístico sería uno que incluye los efectos del tratamiento. El ratio F acá es demasiado grande que es improbable que se hubiera sido creado por un modelo sin el tratamiento.

Suposiciones

ANOVA asume que las varianzas del error tienen una distribución normal y son independientes.

ANOVA es fuerte si violamos unas suposiciones, pero la suposición de independencia es importante.

Como una regla de oro, si la desviación la más pequeña del nivel de tratamiento es menos de doble la desviación estandár la más grande del tratamiento, es estable ANOVA.

El método más fácil

Cargamos los datos de Pitt y Shoaf.
Corremos la función: anova(lm(rt∼position, data=ps))

Examinamos que significa cada celda.

Otro ejemplo

Ahora examinamos los datos de VOT de mixteco.

El VOT varía con el consonante?

Cuántas pruebas de t necesitaríamos hacer para examinar este efecto? Hacemos ANOVA. Véase código #10 - 12.

ANOVA de dos factores

Experimentos factoriales son los que tienen dos o más predictores.

Un diseño clásico es 2 x 2, con dos niveles de cada factor. Podemos también hacer experimentos de 3 x 2 x 4, etc.

Con diseños factoriales, nos interesa un efecto de interacción - la posibilidad que nuestro efecto de un factor no es el mismo a través de los niveles del otro.

	Factor Y	
Factor X	P	Q
A	AP	AQ
B	BP	BQ

Hay tres relaciones: la desviación que encontramos con tratamiento X, con tratamiento Y y la interacción entre ellos.

Como antes, calculamos una varianza de error total. Se lo usa como base para compararlo a las otras varianzas.

No solamente nos interesa cada factor, pero si una desviación de un nivel dado variaría en la misma manera cuando lo comparamos con otro nivel.

Para examinar las interacciones, determinamos la desviación del promedio de cada nivel del factor X en comparación con cada nivel P y Q de factor Y. Las desviaciones observadas de cada nivel en X están comparadas a la desviación promedio entre los niveles.

Calculamos SS_τ por cada nivel del tratamiento como ANOVA de un solo factor.

$$\mu + \alpha_i + \beta_j + \alpha_i\beta_j + \epsilon_{ijk}$$

Un ejemplo

Pitt & Shoaf data de Johnson (2008).

	Temprano	Medio	Tarde	$\overline{x}$
0-phone	745	905	701	784
3-phone	910	820	691	807
$\overline{x}$	828	863	696	795

Sabemos algo del un efecto de tratamiento (el tiempo de la prueba experimental), pero los oyentes escucharon pruebas experimentales que traslaparon con el prime y otros que no. Este tratamiento tiene solamente dos niveles.

	Temprano	Medio	Tarde	$\overline{x}$
0-phone	745	905	701	784
3-phone	910	820	691	807
$\overline{x}$	828	863	696	795

Qué es α_{3P}? $\overline{x}_\tau - \overline{x}$..
Qué es α_{0P}? "
Qué es β_{early}? "

Calculamos valores de α y β como antes.

Valores de interacción

	Early	Mid	Late	$\overline{x}$
0-phone	745	905	701	784
3-phone	910	820	691	807
$\overline{x}$	828	863	696	795

$$RT_{3P.beg} = \mu + \alpha_{3P} + \beta_{beg} + +\alpha_{3P}\beta_{beg} + \epsilon_{ijk}$$
$$RT_{3P.beg} = 795 + 11 + 33 + 71 + \epsilon_{ijk}$$

Cómo calculamos $SS_{\tau_1\tau_2}$?

Recuérdense que $SS_\tau = r\Sigma(\overline{x}_i - \overline{x}..)^2$

Nuestro modelo tiene cuatro SS diferentes y no dos; necesitamos SS_{τ_1}, SS_{τ_2}, $SS_{\tau_1\tau_2}$, and SS_{total}.

Las etapas para calcularlas:
$X = (N_1-1)(N_2-1)*\Sigma(\Sigma(\overline{x}_{L1\tau_1*L1\tau_2}-\overline{x}..)^2+\Sigma(\overline{x}_{L2\tau_1*L1\tau_2}-\overline{x}..)^2...)$

L = nivel del tratamiento. Nótese que significa "N". Qué asumimos?

Lo que hacemos acá es calcular las desviaciones cuadradas de cada interacción posible, agregarlas y multiplicarlas con un grado de libertad grande.

Queremos excluir SS de cada tratamiento de la varianza total que observamos con la suma de todas las interacciones. La próxima etapa es restar el SS del tratamiento.

$$SS_{\tau_1\tau_2} = X - SS_{\tau_1} - SS_{\tau_2}$$

Si calculamos SS, MS es fácil. Lo dividmos por df. Después aplicamos la prueba de F a los cuatro MS que hemos calculado.

Output

Véase código #13.

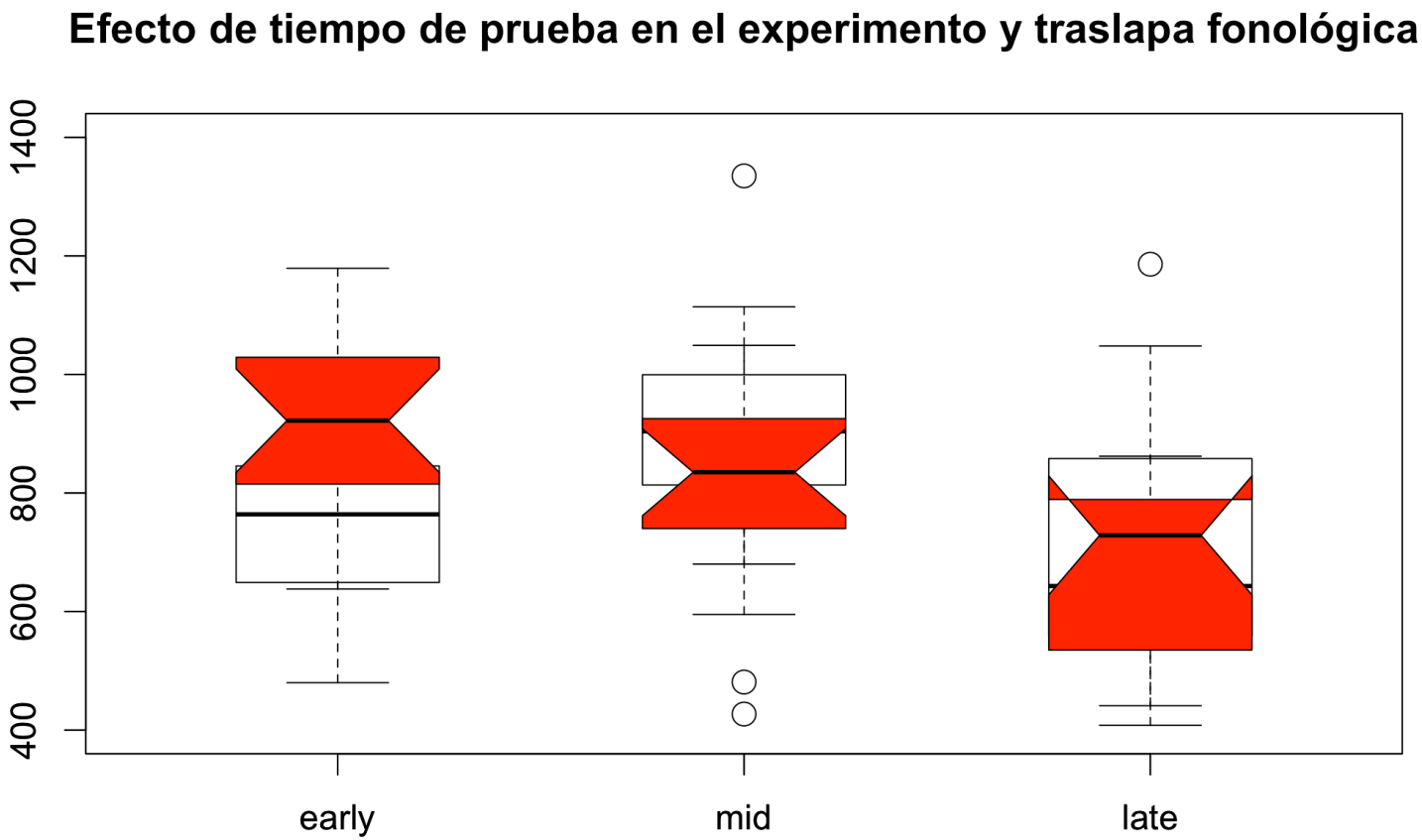
1: Qué encontramos con nuestro efecto principal de "Overlap"? Qué significa?

2: Qué tipo de interacción encontramos?

3: Qué significa?

(Ahora véase código #14.)

Las interacciones en nuestros datos

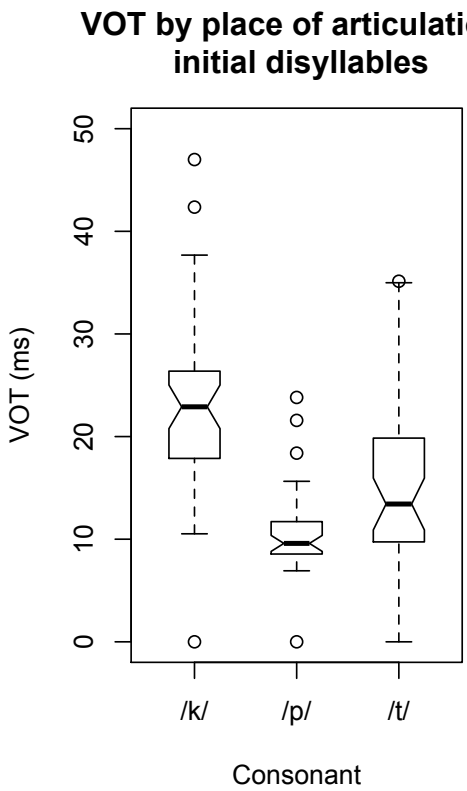


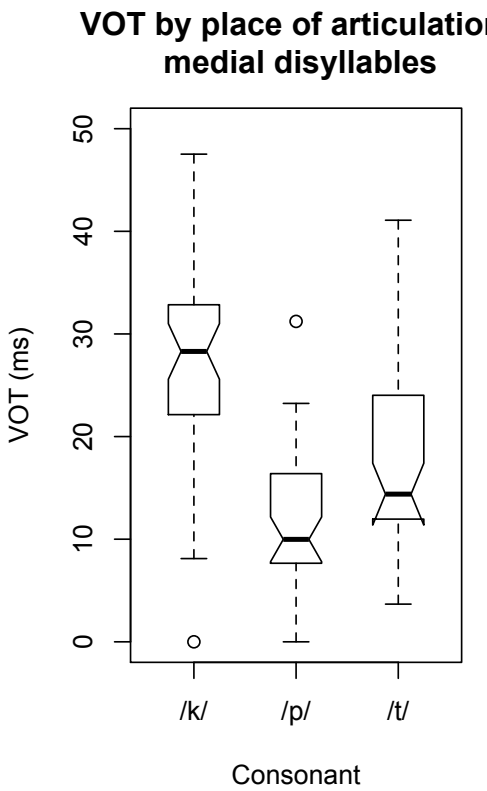
Puede ser que tenemos alguna idea de las diferencias con nuestros efectos principales, pero normalmente no podemos predecir lo que pasa con nuestras interacciones hasta que visualizamos los datos.

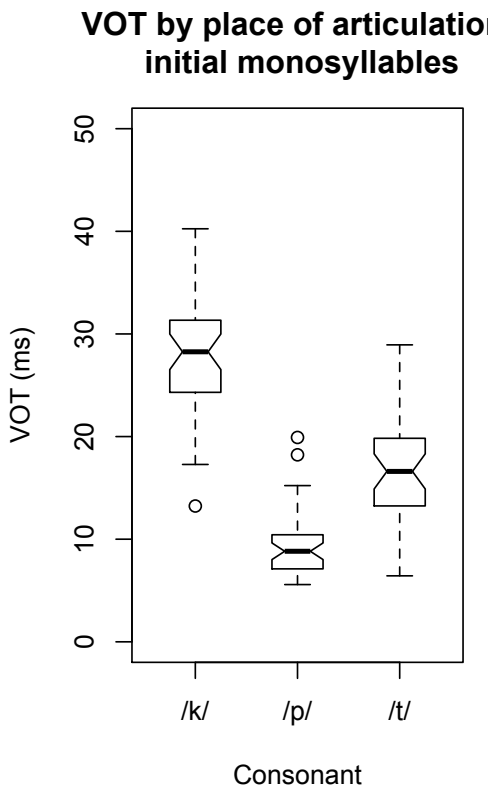
Cuando reportamos las estadísticas, incluya una discusión de las magnitudes de los efectos con las estadísticas y dirija el lector a sus observaciones.

Véase código #15.

Datos de mixteco







Pruebas post-hoc

A veces nos interesa saber si las diferencias en la varianza son generales o se los atribuye solamente a un nivel específico de un factor.

Después de aplicar ANOVA, puede ser útil investigar unas "pruebas post-hoc" que nos permite examinar todas las interacciones individuas de nuestros datos.

Si encontramos que solamente una o dos interacciones son significativas, las incluimos por decir "Una prueba post-hoc nos dice que solamente dos comparaciones eran significativas: A con B y A con C (pero no A con D)."

Comparaciones emparejadas

Un tipo de prueba post-hoc es aplicar una prueba t de dos muestras a través de los niveles de tratamiento.

Es necesario crear subconjuntos para hacer esto.

Véase código #16.

Hay que fijarse bien - necesitamos evitar comparaciones multiples.

Las correcciones de **Bonferroni** - por cada comparación N, lo dividmos α por N.

Si p $< \alpha$/N, el efecto dado es significativo.

Entonces si tenemos 3 niveles y hay 3 comparaciones posibles, $\alpha = 0.05/3 = 0.0167777$.