Introducción al análisis de varianza

Comparando varias poblaciones

- Vimos extensiones de los métodos de inferencia
  - Métodos para **una población**, los extendimos para comparar los **parámetros de dos poblaciones**
  - Tanto con medias como con proporciones
- Trataremos de ir más allá
  - Es natural pensar en extensiones para **más de dos grupos**
  - De hecho, la familia de pruebas χ^2 soporta inferencias con **más de dos proporciones**
  - Ahora trataremos de contrastar la **igualdad de un conjunto de medias poblacionales**

Comparando varias poblaciones

- ¿Por qué comparar varias medias?
  - Es fácil pensar en varios ejemplos:
    - varias **áreas geográficas**
      - nivel de felicidad en ciudades, pueblos y áreas rurales
    - varias **dimensiones sociales**:
      - rendimiento de estudiantes provenientes de liceos municipales, particulares-subvencionados y particulares pagados
    - varios **grupos etarios**:
      - niveles de glucosa en la sangre de menores, adolescentes, adultos jóvenes, adultos y adultos mayores
    - varios individuos u **objetos de estudio diferentes**:
      - tiempo para 100 m planos de 5 atletas
      - error medio cometido por cuatro algoritmos para predecir demanda eléctrica

Comparando varias poblaciones

- Seguimos con **un factor**
  - Pero ahora vamos a comparar **más de dos niveles**
  - Aquí podemos seguir **dos tipos de modelos**
  - Un modelo de **efectos fijos**[1, cap. 6]
    - en donde se trabaja como si **no hubiera más tratamientos** que los niveles que se están estudiando
    - es decir, los efectos que se observen están fijos y no son resultado de alguna decisión aleatoria
  - Un modelo de **efectos aleatorios**[1, cap. 6]
    - en donde se considera que los tratamientos son una muestra de una **población de niveles** posibles
    - es decir, el factor es una variable aleatoria y los niveles usados en el modelo son el resultado de una **muestra aleatoria** de esta variable

Comparando varias poblaciones

- Veamos un ejemplos clarificador:
  - Un par de colegios quiere determinar si las notas de sus alumnos de 2do medio en matemáticas **dependen del profesor** con quien tienen la asignatura (o solo del esfuerzo que de cada alumno pone)
    - uno de los colegios tiene tres 2º medios y comparará los tres cursos
      - este caso, **observaremos efectos fijos**
    - pero otro colegio tiene ¡quince 2º medios! por lo que ha decidido comparar solo cuatro de ellos (elegidos al azar)
      - este caso, los **efectos son aleatorios**
- Los procedimientos son un tanto distintos
- También existen los modelos mixtos
  - Cuando hay más de un factor, claro

Ejemplo

- Usemos un ejemplo para ver las complicaciones:
  - El dueño de una empresa de desarrollo de software quiere invertir más eficientemente en su capital humano y para eso realiza el siguiente experimento:
    - los desarrolladores se dividen aleatoriamente en **4 grupos**
    - un **grupo de control** (sin intervención)
    - los otros grupos se envían a un **curso de capacitación** en desarrollo ágil de aplicaciones con distinta duración: 2, 4 o 6 días
    - se ha medido el **número de pruebas unitarias falladas por *sprint***, para varios *sprints* elegidos aleatoriamente
    - la idea es determinar **si la capacitación tuvo un efecto** en el número promedio de fallas que cometen los desarrolladores

Ejemplo

- Preguntas
  - ¿Cuál es el **factor** en este caso?
  - ¿Cuáles son sus **niveles**?
  - ¿Efectos fijos o efectos aleatorios?

Ejemplo

- El experimento tuvo los siguientes resultados:

0 día	2 días	4 días	6 días
26	22	19	19
27	23	20	20
28	24	21	23
28	27	23	24
33	27	27	24

  - ¿Tuvo **un efecto** la capacitación en el número promedio de fallas que cometen los desarrolladores?
  - ¿Cuántos días de capacitación son ideales?

Extendiendo las ideas de Student

- Pregunta
  - ¿Podemos responder con un gráfico tal vez?
  - ¿Podemos extender la prueba t de Student a este caso?
    - ¿cómo calcular la **varianza combinada** en este ejemplo? o
    - ¿tal vez usar **más una prueba t**?

0 día	2 días	4 días	6 días
26	22	19	19
27	23	20	20
28	24	21	23
28	27	23	24
33	27	27	24

Extendiendo las ideas de Student

- Extendiendo las ideas:
  - El procedimiento que usamos en la prueba t de Student **no es fácil** de extender
  - ¿Por qué no usar pares de t-tests?
    - se **complica el cálculo**: hay que hacer $k \cdot (k-1)/2$ pruebas para k niveles
    - peor, el **nivel de significación α se distorsiona**[1, cap. 6]
      - e.g. si se usa α = 0.05 en **tres pares de pruebas** (3 niveles: A, B, C)
      - **H_0: todas medias son iguales**, puede ser rechazada por la primera prueba (A-B), **o** la segunda prueba (A-C) **o** por la tercera (B-C)
      - si las pruebas son independientes, la **probabilidad de no rechazar H_0** es $(0.95)^3 \approx 0.857$
      - **¡$\alpha \approx 0.143$, no 0.05!**

Análisis de varianza

- Debemos buscar un método alternativo
- Uno de los más populares y estudiados[1, cap. 6]:
  - El **análisis de varianza** (ANOVA o AoV)

- En resumen el método[1, cap. 6]:
  - Compara la varianza **entre** las medias de las poblaciones, con la varianza **dentro** de cada población

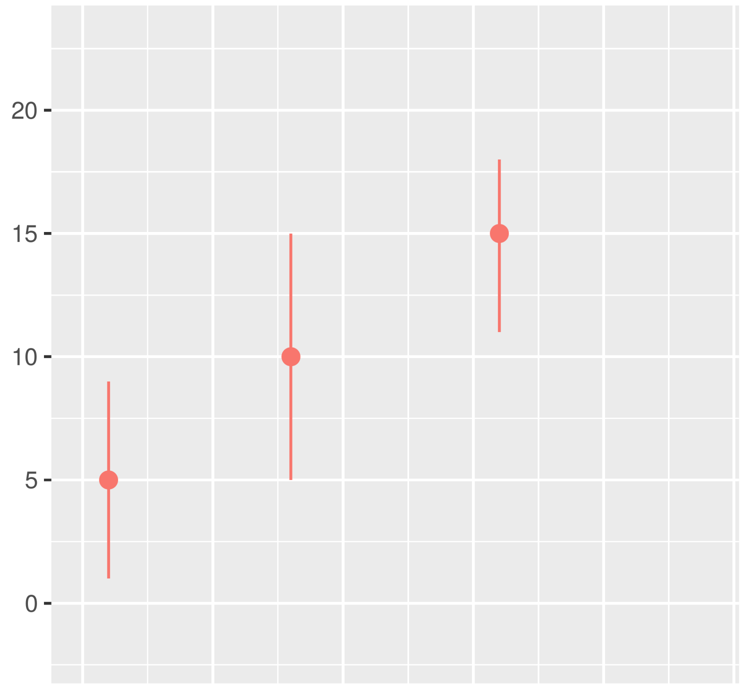
Análisis de varianza

- Pregunta
- ¿Por qué analizar varianza si queremos **comparar medias**?

Análisis de varianza

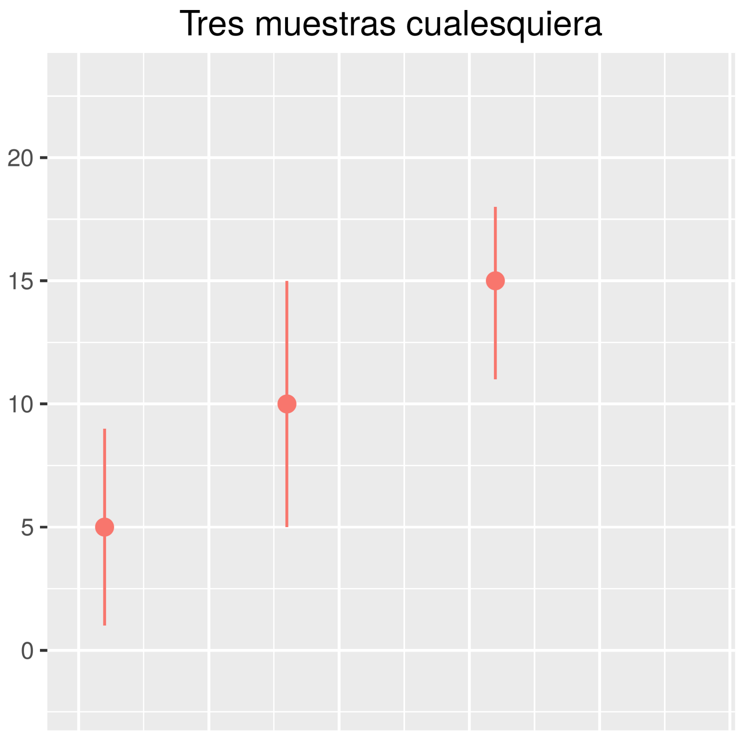
- Veamos un ejemplo genérico para aclarar
  - Consideremos tres muestras con medias 5, 10 y 15
  - ¿Vienen de la **misma población**?

Tres muestras cualesquiera



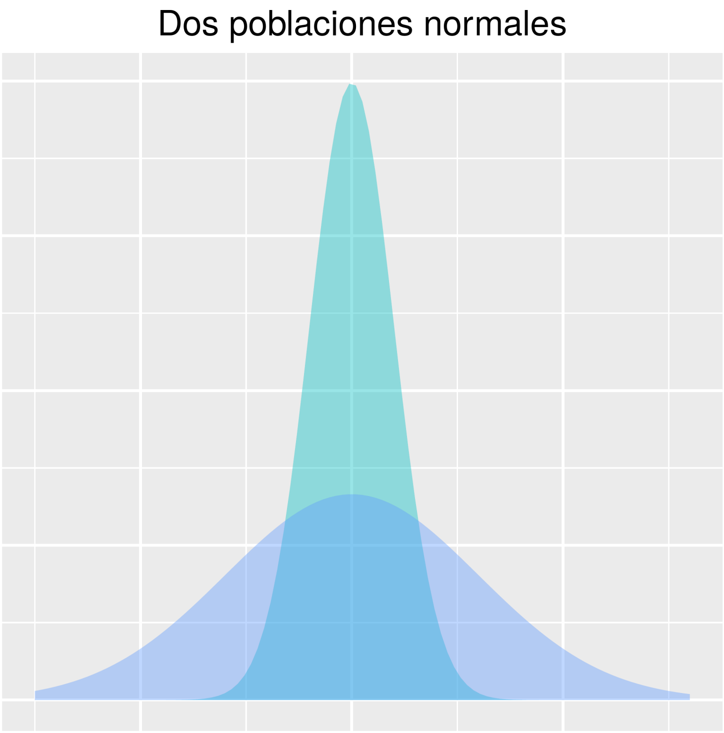
Análisis de varianza

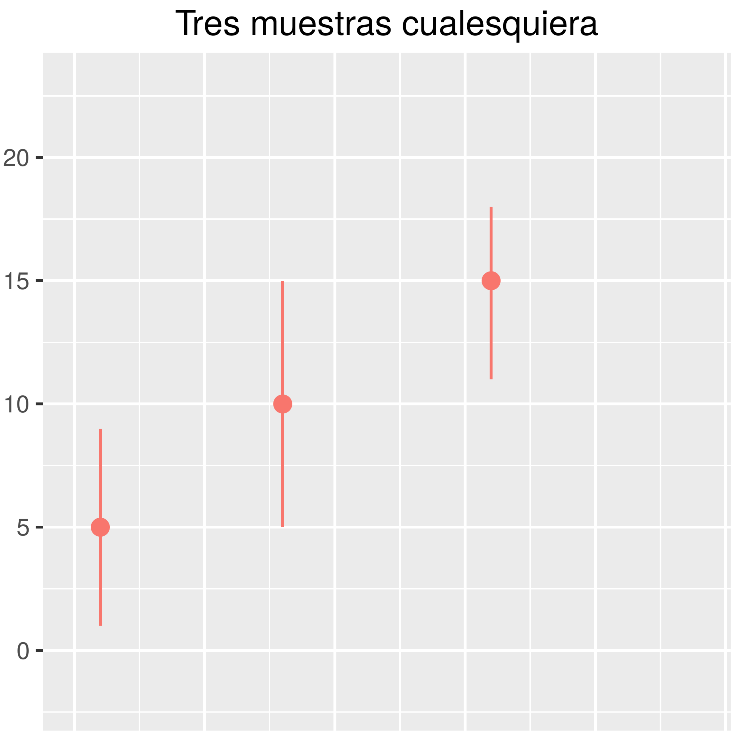
- Veamos un ejemplo genérico para aclarar
  - Consideremos tres muestras con medias 5, 10 y 15
  - ¿Vienen de la **misma población**?
  - Nuestra respuesta ha de ser: **¡depende de la varianza!**



Análisis de varianza

- ¿Depende de la varianza?
  - Para ilustrar, consideremos dos poblaciones con **niveles de variabilidad distintos**

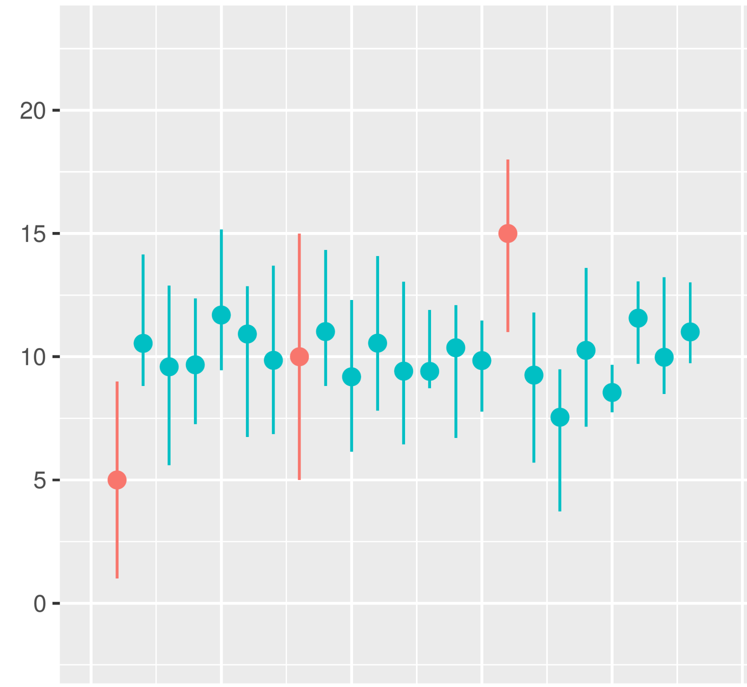




Análisis de varianza

- Tomemos muestras aleatorias de la 1ª población
  - Hay muestras que tienen **valores 5 o 15**
  - Pero **ninguna** muestra nos dio con **media 5 o 15**
    - es decir, la **variación** de las medias **entre** las muestras es **más grande** que lo esperado por el **muestreo aleatorio**
    - esto sugiere que hay una **diferencia real** entre las muestras
    - tal vez vienen de poblaciones distintas

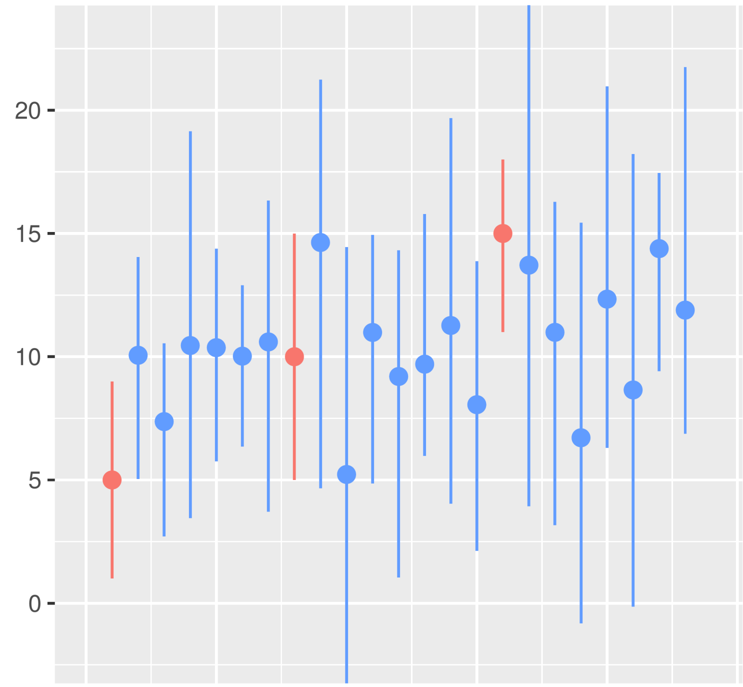
Las tres muestras y muestras de la población 1



Análisis de varianza

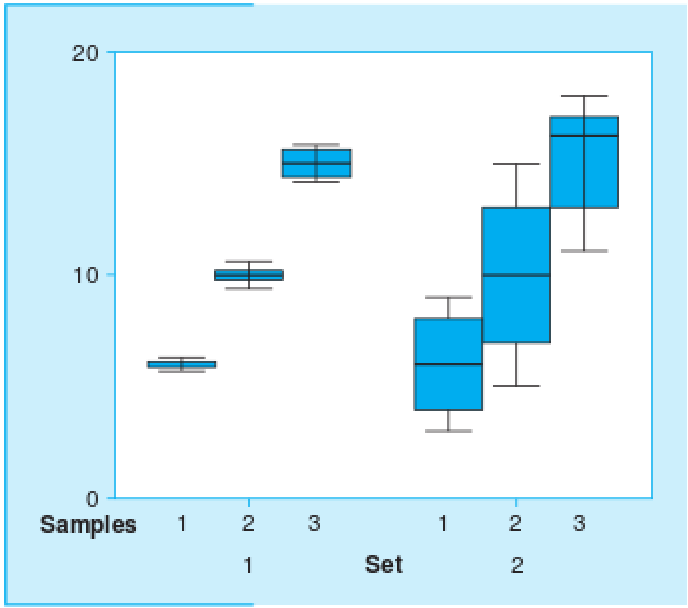
- Veamos ahora qué pasa con la 2ª población
  - Las muestras son claramente **más dispersas**
  - Hay muestras con **medias muy cercanas a 5 y a 15**
    - luego la **variación** de las medias **entre** las muestras **no es mucho más grande** que la variación **dentro de las muestras**
    - estas diferencias **podrían deberse al azar** introducido por el muestreo
    - esto nos hace **dudar** poder declarar fehacientemente que existe una **diferencia real** entre las muestras

Las tres muestras y muestras de la población 2



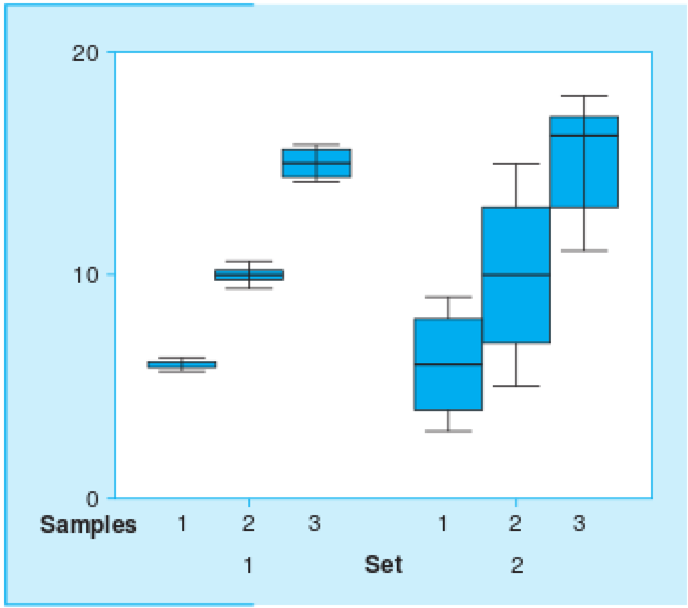
Procedimiento Anova

- ¿Cómo llevamos esto a un procedimiento?
  - Lo usual es que **no conocemos** las poblaciones
    - solo nos queda *confiar* en las muestras
  - Veamos estos dos casos[1, cap. 6]:
    - las **mismas medias** por grupo
    - es decir, las mismas diferencias entre las medias, **misma variabilidad entre los grupos**
    - ¿qué hay de la **evidencia** que soporta diferencias reales?
    - ¿diríamos que tenemos **el mismo nivel de evidencia** en ambos casos?



Procedimiento Anova

- ¿Cómo llevamos esto a un procedimiento?
  - La evidencia de diferencias entre grupos es **más fuerte** en el primer conjunto que en el segundo[1, cap. 6]
    - esto porque las observaciones en cada grupo **están más aglutinados** en el 1er caso
    - lo que sugiere que **la población** desde donde vienen estas muestras tiene **menor varianza**
    - luego, aunque la **varianza entre las medias** es la misma, la **varianza entre observaciones** de cada muestra es menor en el 1º

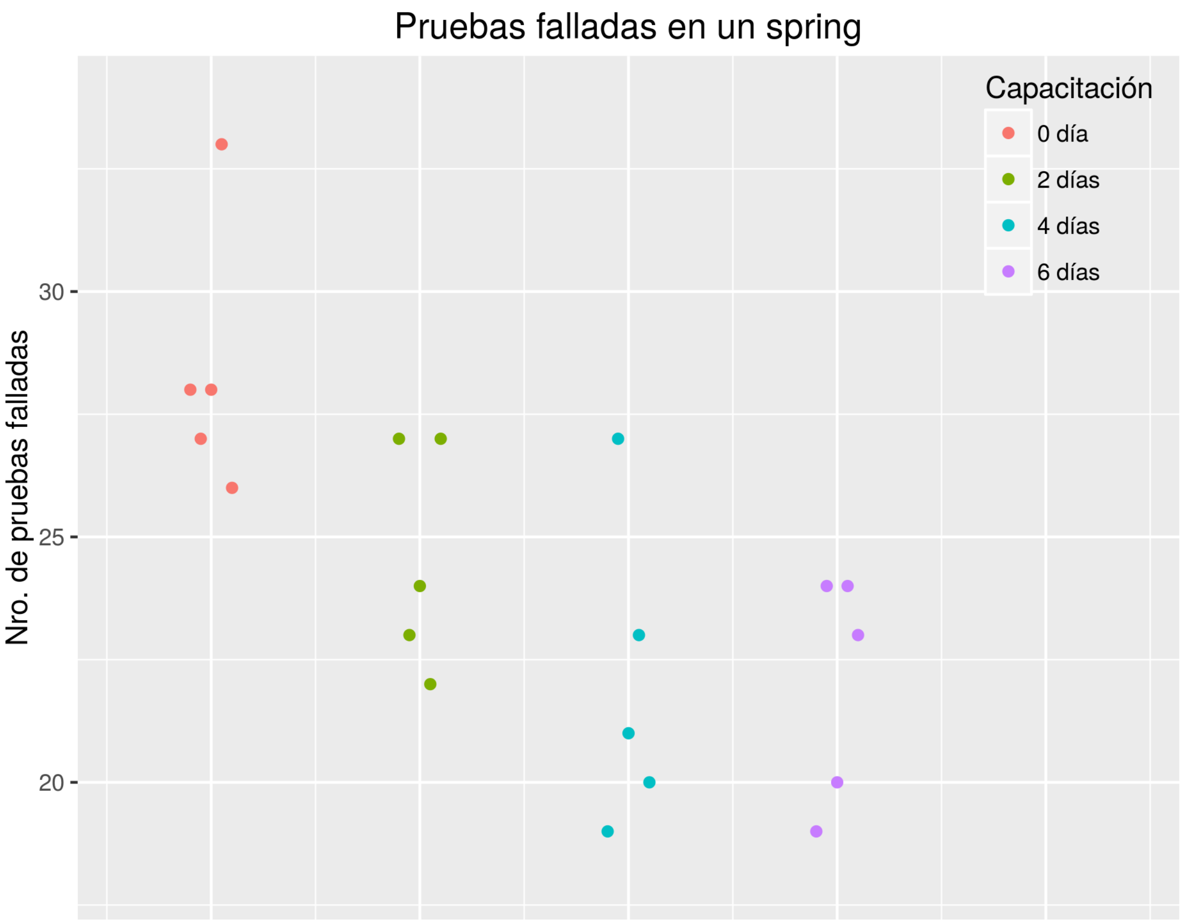


Procedimiento Anova

- ¿Cómo llevamos esto a un procedimiento?
  - Esta es la base del análisis de varianza[1, cap. 6]
    - se compara la varianza **entre las medias de las poblaciones** con la varianza **entre las observaciones al interior de cada población**
  - Veamos esta idea en nuestro ejemplo

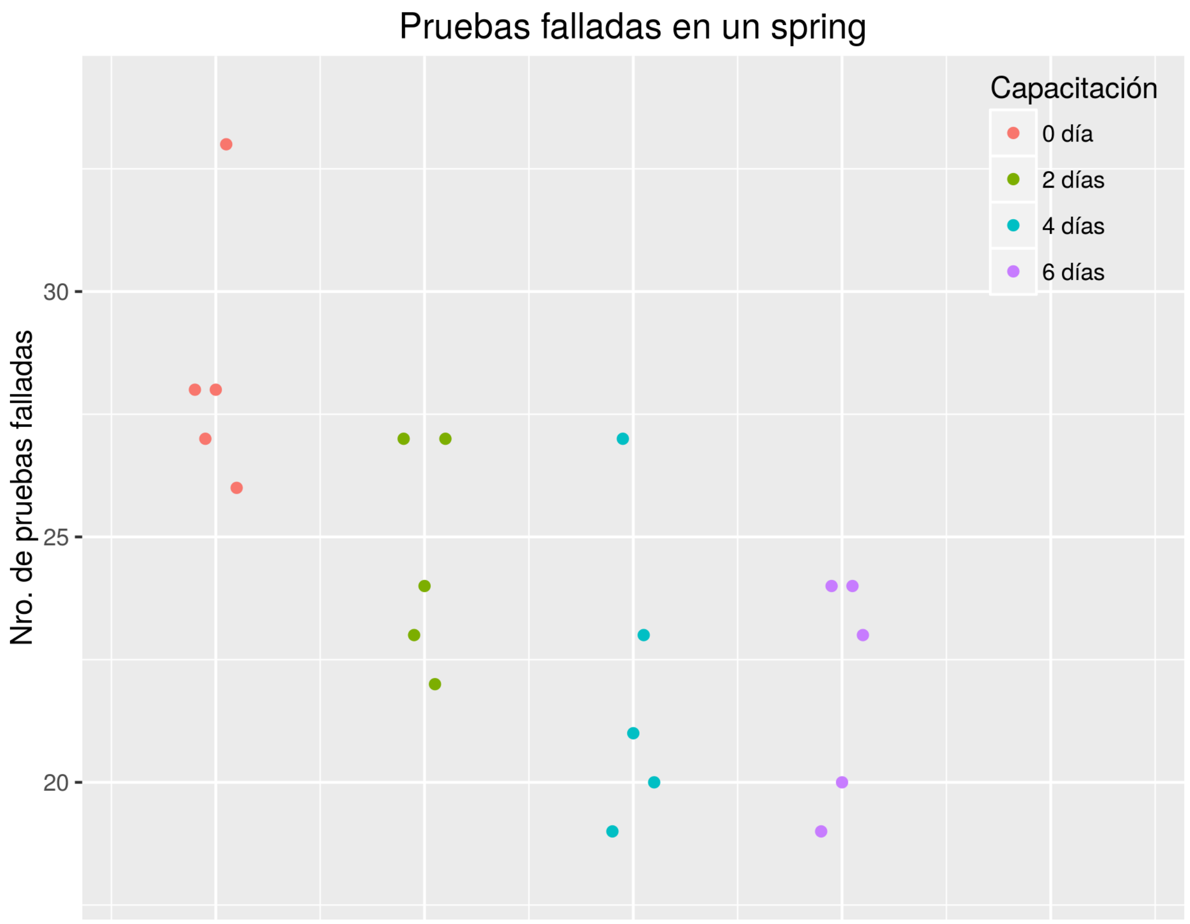
Ejemplo

- Visualmente:
  - ¿Dónde hay más varianza?
  - ¿Entre las medias de cada grupo? o
  - ¿al interior de cada grupo?



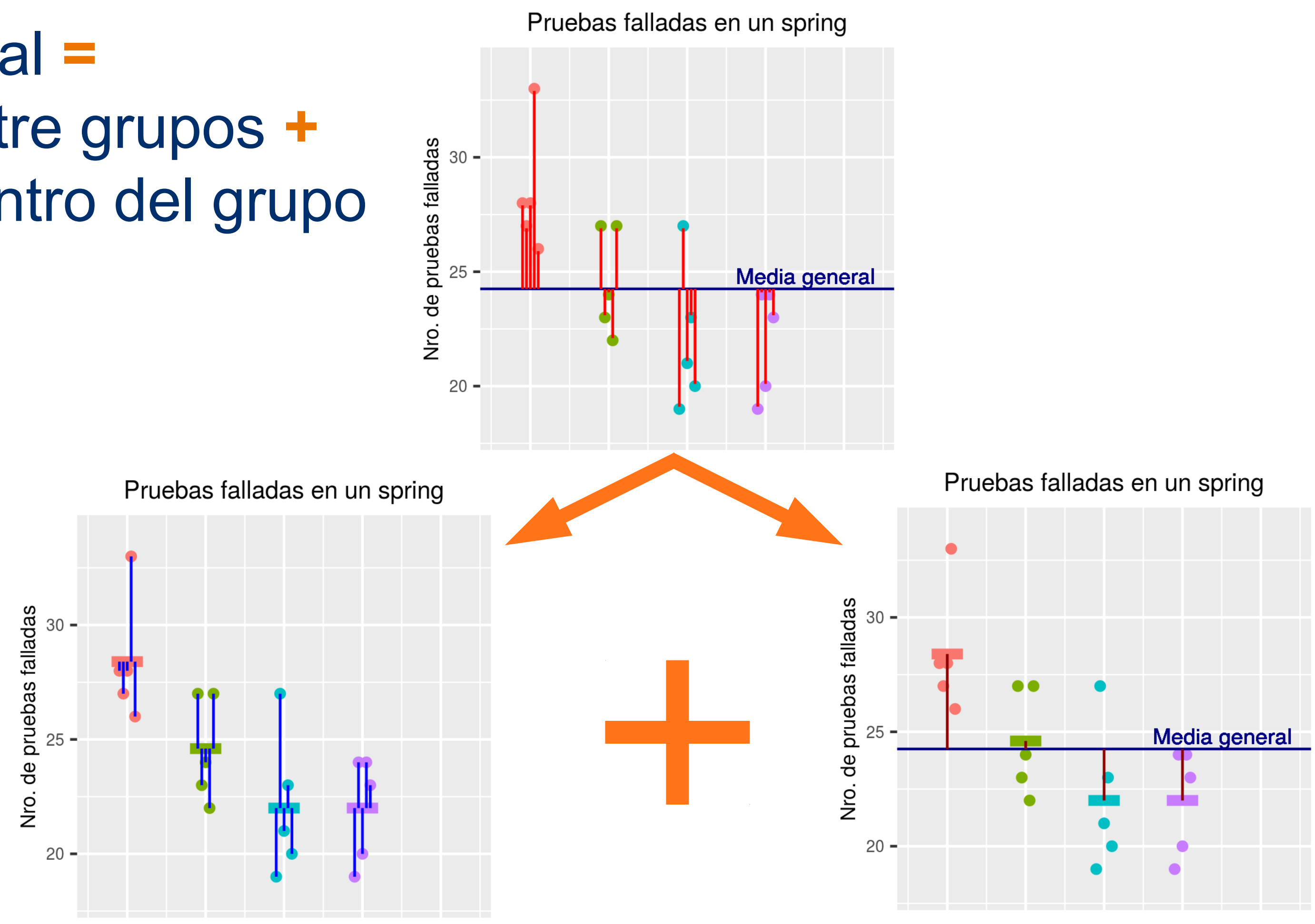
Ejemplo

- Visualmente:
  - ¿Dónde hay más varianza?
  - ¿Entre las medias de cada grupo? o
  - ¿al interior de cada grupo?
  - ¡No es tan fácil!
  - Mejor tener un **procedimiento matemático**



Ejemplo

- La genialidad[2, cap. 13]:
  - Variación total =
    variación entre grupos +
    variación dentro del grupo



Procedimiento Anova

- Algunas ideas y notaciones
  - "Entre" (grupos, muestras, poblaciones) ≈ *between*
  - "Al interior de" (grupos, muestras, poblaciones) ≈ *within*
  - Estimación de la **varianza total**:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{SS_T}{df_T}$$

    - esto es, la suma del total de las desviaciones cuadradas (***total sum of squared***, $\mathbf{SS_T}$) dividida por los grados de libertad totales (***total degrees of freedom***, $\mathbf{df_T}$)
  - Al separar la varianza total:

$$SS_T = SS_B + SS_W$$

$$df_T = df_B + df_W$$

Procedimiento Anova

- Pero usamos medidas normalizadas:
  - Promedio de desviaciones cuadradas (***mean squared*, MS**)
    - así se tiene el **MS entre grupos (MS_B)**, y el **MS dentro de los grupos (MS_W)**
    - si ambas medidas de variación **vienen de la misma población (H0)**, entonces su razón debe ser cercana a uno: MS_B / $MS_W \approx 1$
    - pero MS es, en casos simples, lo que llamamos **varianza**
    - y sabemos que una **división de varianzas** sigue **distribución F**
      - eso permite obtener un p-valor para diferentes valores de esta razón de varianzas

Referencias

[1] Rudolf J. Freund, Donna Mohr, William J. Wilson (2010). Statistical Methods, 3rd Edition. Academic Press.

[2] Richard Lowry (2016). VassarStats. Vassar College, http://www.vassarstats.net/.

Ejercicio

- Existen muchos métodos para hacer Anova en R

0 día	2 días	4 días	6 días
26	22	19	19
27	23	20	20
28	24	21	23
28	27	23	24
33	27	27	24

  - Averigüe cómo hacerlo con la función `aov()`
  - Obtenga el mismo resultado con la función `ezANOVA()` del paquete `ez`