**¿Qué es un data scientist? Diferencias y similitudes entre data science y data analysis**

La ciencia de datos (data science) es una ciencia interdisciplinar que se nutre y utiliza múltiples áreas de conocimiento como la estadística, la programación e incluso otras ramas de la informática en combinación con la información que recibe del entorno de interés (una consultoría, un laboratorio que trabaja con datos biológicos, una industria cárnica, etc.) para entender, procesar, analizar y extraer conclusiones sobre los datos recogidos de dicho entorno. El científico de datos (data scientist), básicamente, se encarga de manejar y crear código utilizando conocimientos de informática y matemáticas para no solo analizar esos datos, sino llegar a sus propias conclusiones y sugerir y contestar a nuevas preguntas sobre el objeto de interés.

Un analista de datos (data analyst) también se encarga de procesar y analizar información para responder a preguntas a partir de un proceso de selección y análisis de datos. Sin embargo, las diferencias son claras entre el trabajo que realiza un data analyst y el trabajo que realiza un data scientist. Es más, en muchas empresas suelen convivir ambos tipos de especialistas que tienen focalizaciones distintas. Vemos las principales diferencias:

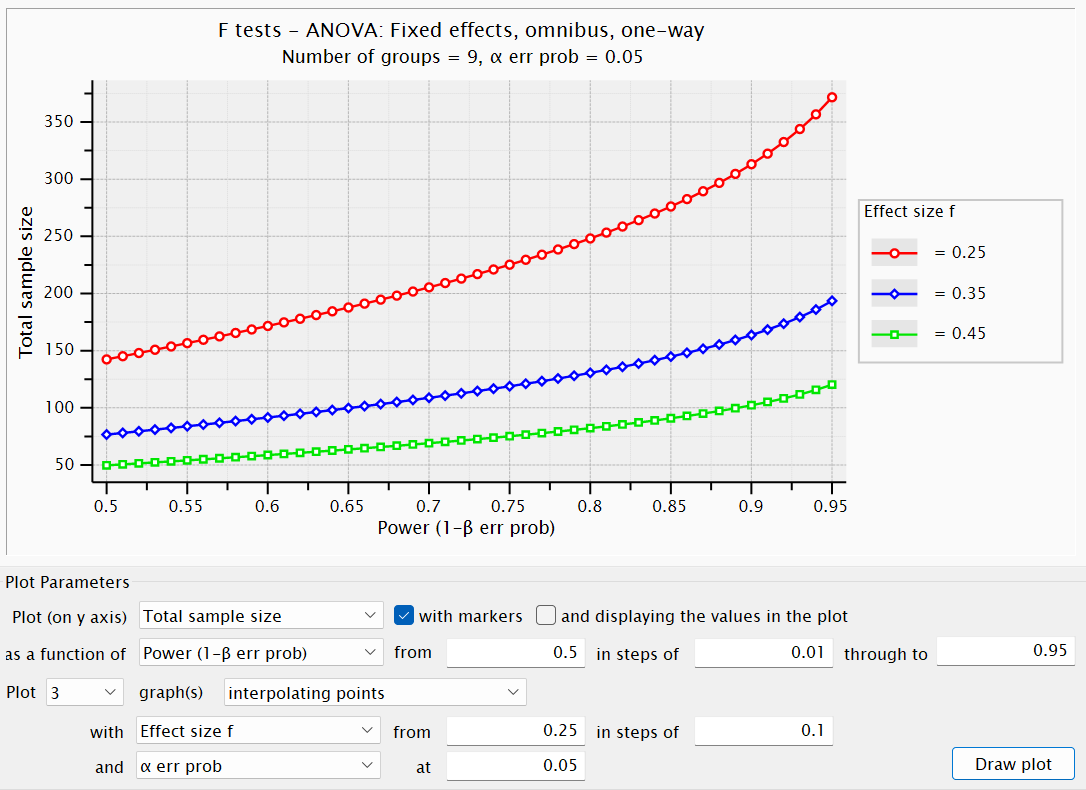
* En primer lugar, el data analyst normalmente solo responde a las preguntas que se le ofrecen en ese entorno de trabajo. Es decir, procesa y analiza los datos para obtener respuestas a las preguntas concretas que la empresa le encarga. Sin embargo, el data scientist va más allá y es capaz de entender la información de manera no tanto descriptiva sino para ser capaz de derivar nuevas preguntas a futuro.
* Por otro lado, en relación a lo anterior, el data analyst se centra más en lo que respecta a análisis descriptivo, visualización de datos y la identificación de patrones o tendencias para responder a esas preguntas (por ejemplo, hacer regresiones lineales o logísticas o un ANOVA/MANOVA de forma detallada para ver patrones, tendencias o interacciones entre datos). En cambio, el data scientist, debido a su enfoque más amplio, se centra más en análisis predictivo, modelización de datos, machine learning y sobre todo, en la generación de paquetes y software para ser usado a posteriori de ser necesario, reutilizando la información y los modelos para ser usados a futuro.
* Así, mientras que el data analyst no se sale tanto de su enfoque estadístico-matemático para ofrecer insights sobre los datos, el data scientist enfoca el problema desde una perspectiva más a futuro creando software que le pueda servir a la persona/empresa interesada, transformando los datos en modelos complejos para plantear otras preguntas, incluso por el propio data scientist.

**¿Qué es el poder estadístico? ¿Cómo se relaciona con el efecto estadístico?**

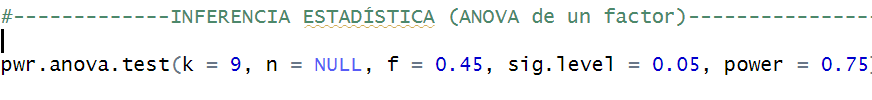
El poder estadístico o la potencia estadística de un test (suponemos no aleatorizado) mide la capacidad de un test de detectar errores de tipo II, es decir, la probabilidad de aceptar la hipótesis nula H0 (que conocemos) dado que H0 es falsa.

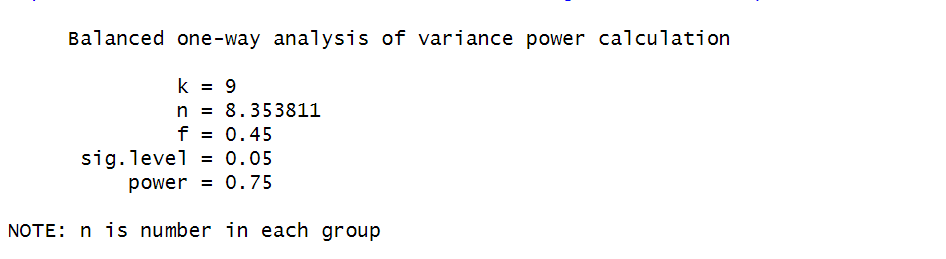
En otras palabras, el poder estadístico mide la capacidad de un estudio para identificar un efecto de estar este presente. Un poder estadístico bajo quiere decir que existe un mayor riesgo de no detectar efectos reales por lo anteriormente dicho, lo que puede llevar a una conclusión errónea. Por eso, en investigación y análisis de laboratorio, se suele buscar un poder estadístico mínimo de 0,80 (80%), lo que implica que hay un 80% de probabilidad de detectar un efecto si realmente existe, aunque dependiendo de la viabilidad de hacer más o menos réplicas independientes del experimento para cada grupo, podremos aceptar poderes estadísticos de 0,60-0,70 (menos de 0,50 en general no, aunque en algunos estudios no se suele hacer un análisis de poder previo al estudio y se aceptan poderes bajos), también dependerá de cuán formal queramos que sea el estudio. Aquí, por cómo es nuestro estudio (una probatina de laboratorio sencilla), hemos considerado un poder de 0,75 (cerca a 0,80) sin problemas, si bien como veremos es recomendable utilizar n > 30 ya que en tal caso en general la distribución de los datos intergrupo (las réplicas) aproxima bien a una distribución normal (podremos suponer normalidad). El efecto a priori queremos que sea alto normalmente porque no solo queremos que el fármaco tenga algún efecto en la germinación, sino que queremos que el fármaco en cuestión tenga mucho efecto en esta.

Así, lo que hemos estudiado primeramente es el tamaño mínimo de muestra necesario para detectar un efecto de interés con una probabilidad adecuada. Aquí se muestra la gráfica para varios f desde 0.25 a 0.45 y poder de 0,75 y un en un one-way ANOVA, que es lo que queremos estudiar.



Nuestra N (número total de réplicas independientes) es aproximadamente N = 81 según la gráfica en G\*Power para f = 0,45 y 1-β = 0,75 (el poder). Si lo hacemos en R, nos devuelve el número de réplicas por grupo que tenemos que tomar, que da n = 8,35, que redondeado hacia arriba es n = 9, que coincide con el análisis que hace G\*Power, con un efecto de f = 0,45 (muy alto, alto se considera en general más de 0,4).





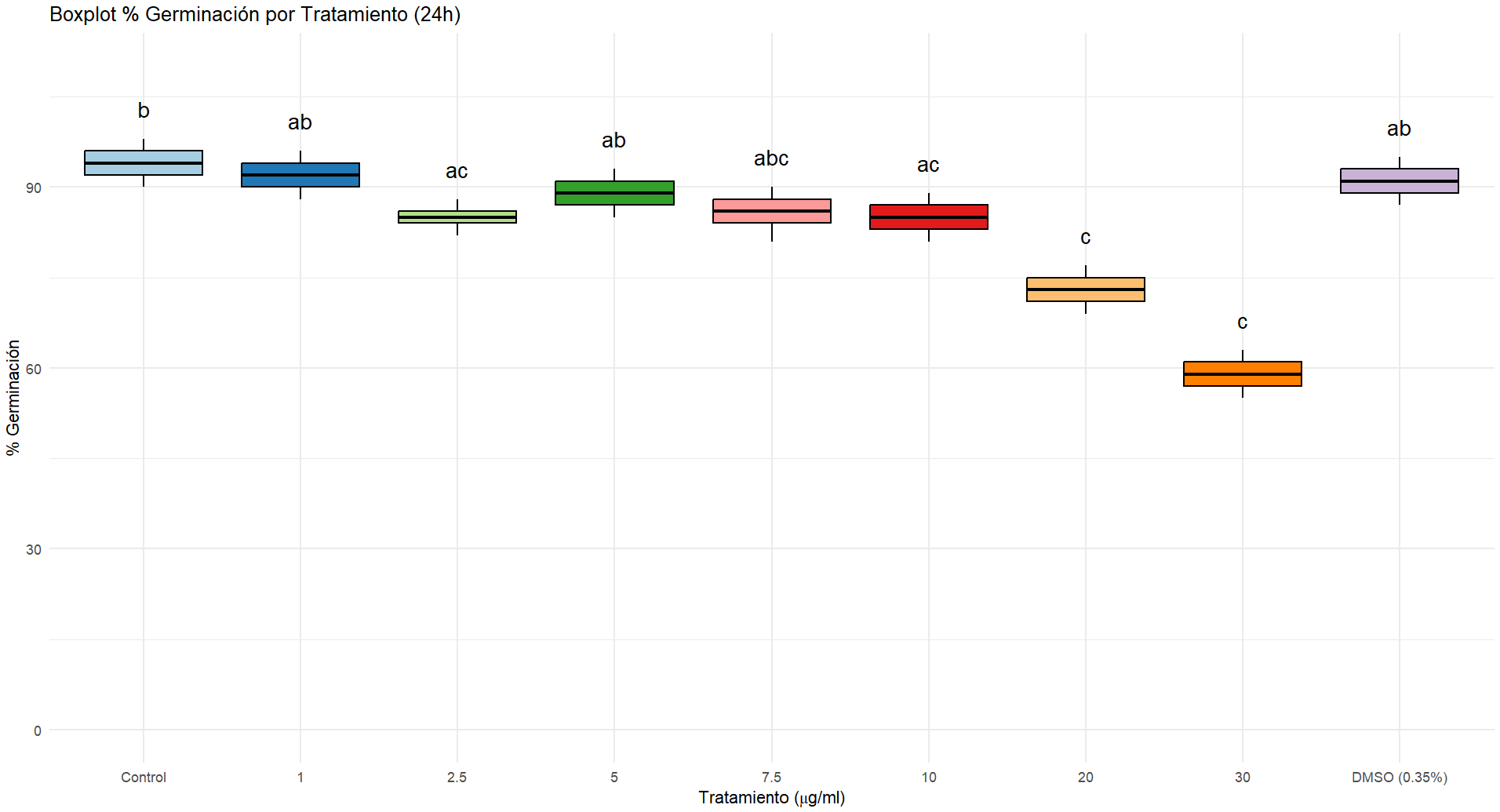
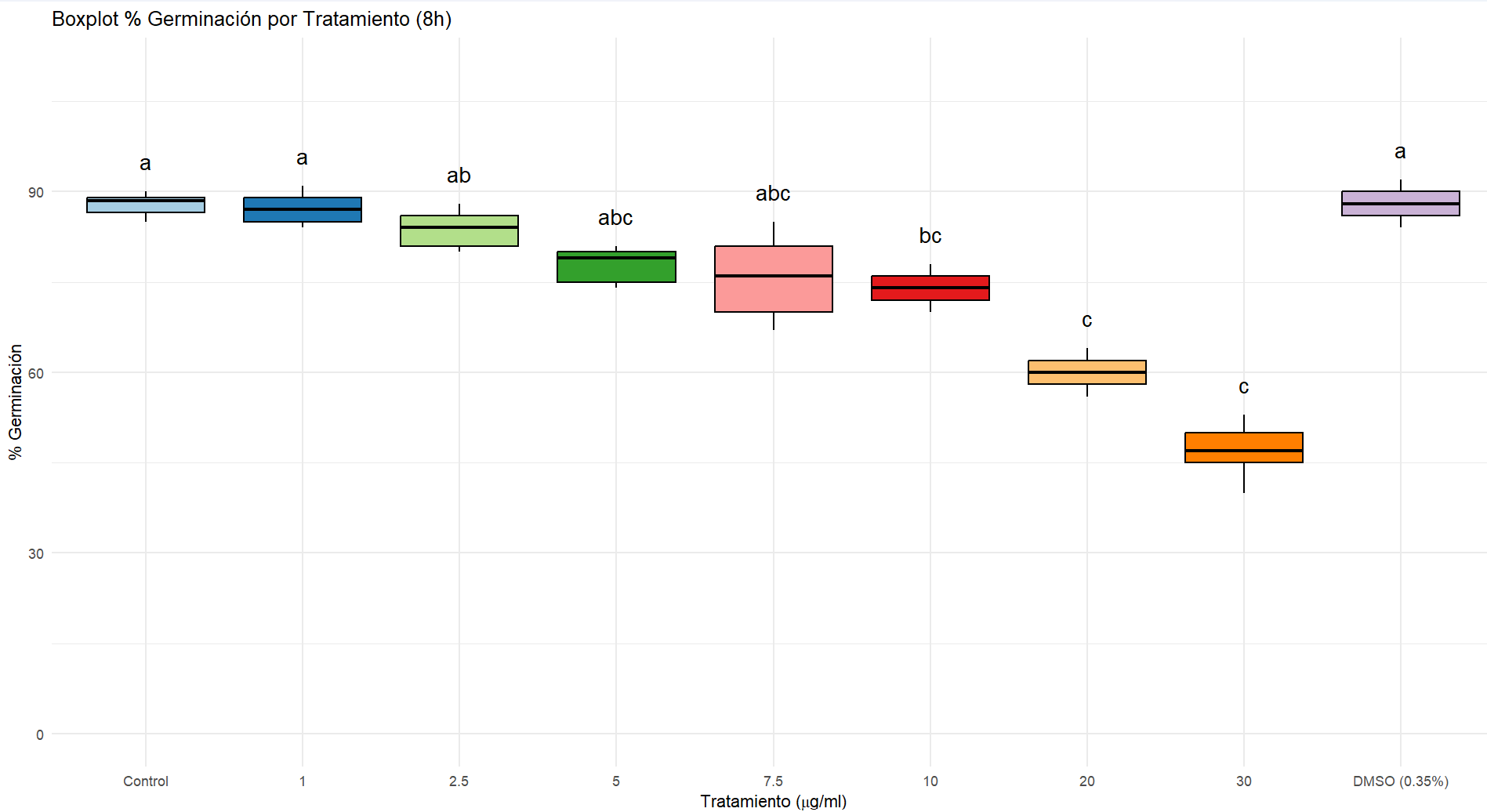
**Pasos para la actividad 1**

(rellenar todo esto con sus pasos y todo)

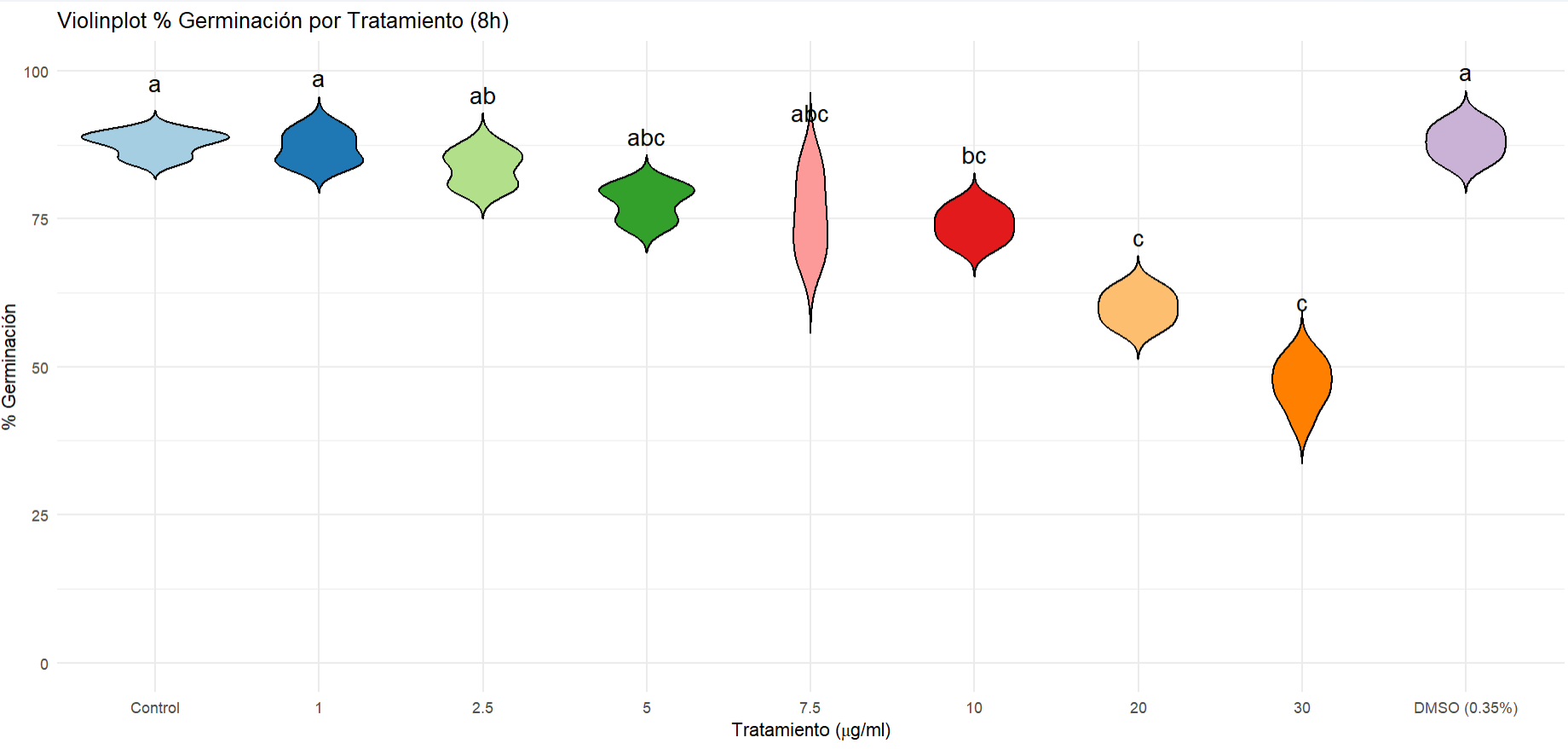
**Conclusiones de la actividad 1**

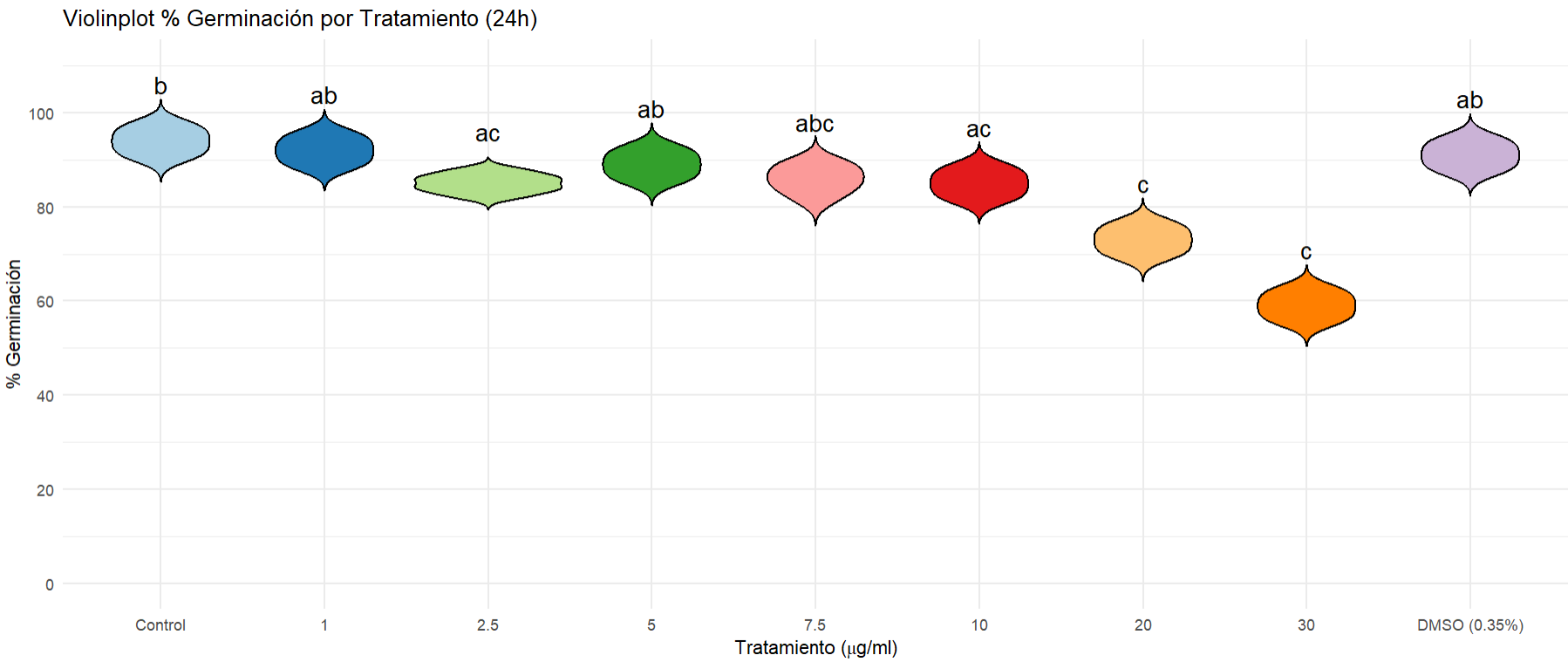
1. Para 8 h, observamos que con respecto al control, vemos una diferencia de medias (la media de las réplicas por grupo) en el % de germinación estadísticamente significativa a partir de 10 µg/ml de fármaco+DMSO. Sin embargo, el mayor efecto se alcanza con concentraciones de 20 y 30 µg/ml, pero puesto que no hay diferencia estadísticamente significativa de medias entre 20 y 30 µg/ml, recomendaría utilizar el **fármaco+DMSO** con una dosis de **20 µg/ml**. Para 24 h, la conclusión es idéntica, salvo que para 2.5 µg/ml la diferencia sí es estadísticamente significativa con el control, aunque no lo es con respecto a 1 µg/ml y 5 µg/ml, de manera que tampoco nos interesa porque el efecto es relativamente pequeño. Notamos también que en 24 h el % de germinación es en general mayor (lo cual tiene bastante sentido ya que dejamos más tiempo para que el hongo crezca).
2. Asimismo, notamos que la diferencia de medias en el % de germinación **se debe fundamentalmente a la concentración de fármaco en µg/ml y no al DMSO**. Esto es porque la diferencia de medias entre el control y la mezcla con DMSO puro 0.35% (sin fármaco) no es estadísticamente significativa, como se observa en el boxplot.

**BOXPLOTS**



**VIOLINPLOTS**

****

****

**Pasos de la actividad 2**

**Conclusiones de la actividad 2**

Comparando modelos glm, vemos que los modelos beta