Introducción al razonamiento bayesiano con modelos generativos

Dr. Héctor Nájera
Dr. Curtis Huffman

19/02/2026

La semana pasada

- La inferencia estadística usa **modelos matemáticos** para hacer afirmaciones generales a partir de un **conjunto particular de datos**
- La investigación empírica recurre a la inferencia estadística traduciendo hipótesis de investigación científica/económica (conocimiento científico cualitativo) en clave de **parámetros de modelos** probabilísticos
- Los modelos aprenden de los datos restringiendo la región del espacio de parámetros, que esos modelos definen, donde viven los mejores predictores

Modelo generativo

Datos

Predicciones

Estimador

Modelo(s) estadístico(s)

Estimando: eso que
quieres estimar

Modelo generativo

¿Cómo describir a los datos?

Modelo(s) estadístico(s)

Estimando: eso que quieres estimar

Preguntas científicas que implican algún tipo de generalización

Razonamiento bayesiano

Prior / A priori

Creibilidad/nivel de conocimiento respecto a algún elemento de un sistema (estimando) p(H), p(θ)

Verosimilitud

Evidencia / Datos y descripción matemática de la plausibilidad de los datos

Distribución Posterior

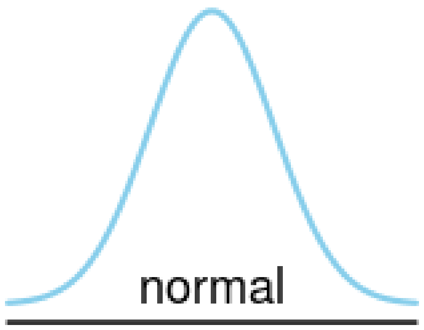
Actualización de p(θ) toda vez que observamos la evidencia

¿Dónde entra la probabilidad?

No hay desacuerdos sobre su definición (Kolgomorov, 1933)

Es en su interpretación donde hay diferencias. En este curso (Bayes-Laplace) se entiende como:

- El nivel de conocimiento/ignorancia que tenemos respecto a la magnitud de algún(os) elemento(s) de un sistema
- El grado de creencia en un enunciado o hipótesis de un sistema (*No es una propiedad de un sistema*)



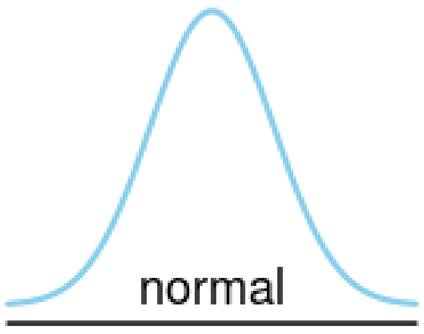
La probabilidad permite codificar matemáticamente nuestros niveles de:

1. creencia,
2. la evidencia,
3. y los ajustes a nuestras creencias

¿Dónde entra la probabilidad?

Nos preguntamos sobre la probabilidad de los valores de parámetros $p(\theta)$:

1. Muestra a población: % de hogares sin agua mayores condicional en la muestra
2. Medición: % de personas en pobreza condicional en los observables, modelo de medición y en la muestra
3. Tratamiento: El efecto del transito formal-informal sobre el riesgo de caer en pobreza condicional en un modelo explicativo (+1 y 2)



Solo podemos hablar de que la probabilidad es 0, 50, 90 o 100% con distribuciones de probabilidad: Densidad

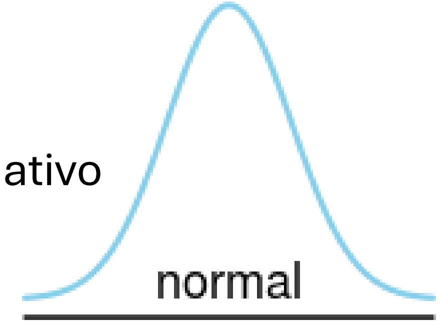
Razonamiento bayesiano

Prior / A priori

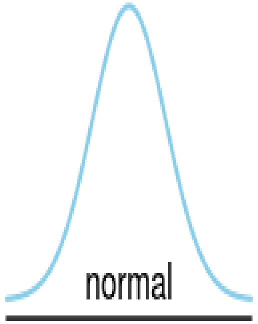
p(θ)

No sé nada

Más informativo

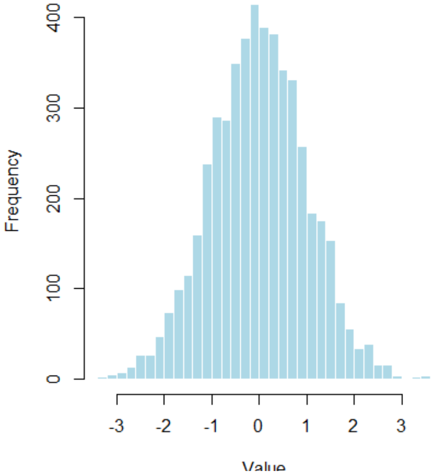


Mucho más informativo



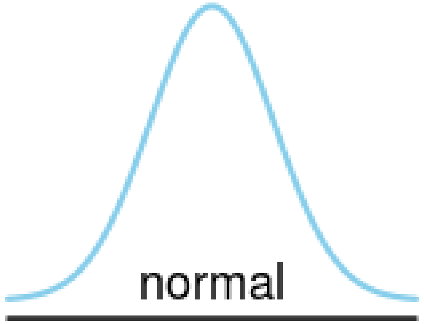
Verosimilitud

P(D|θ)

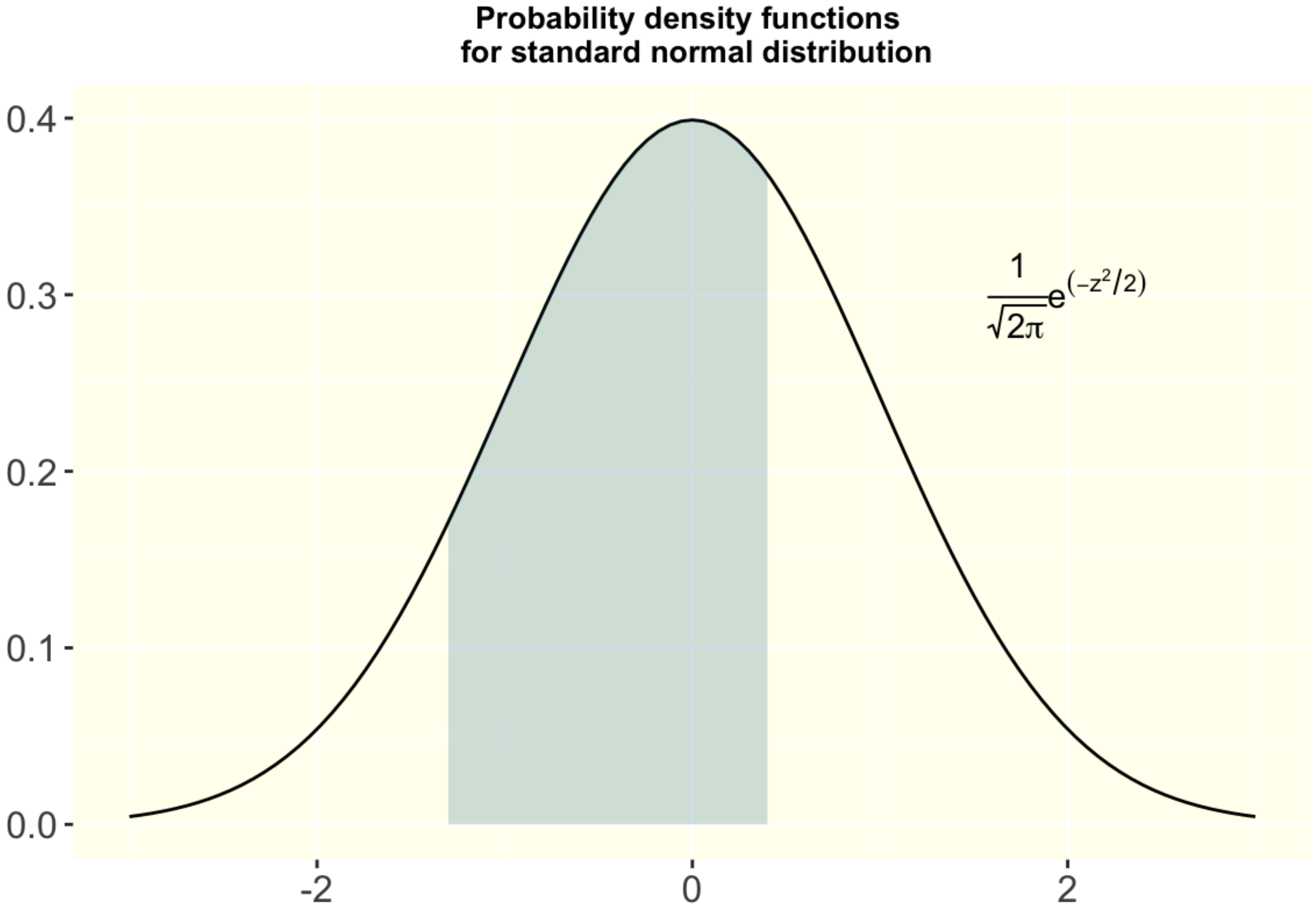


Posterior

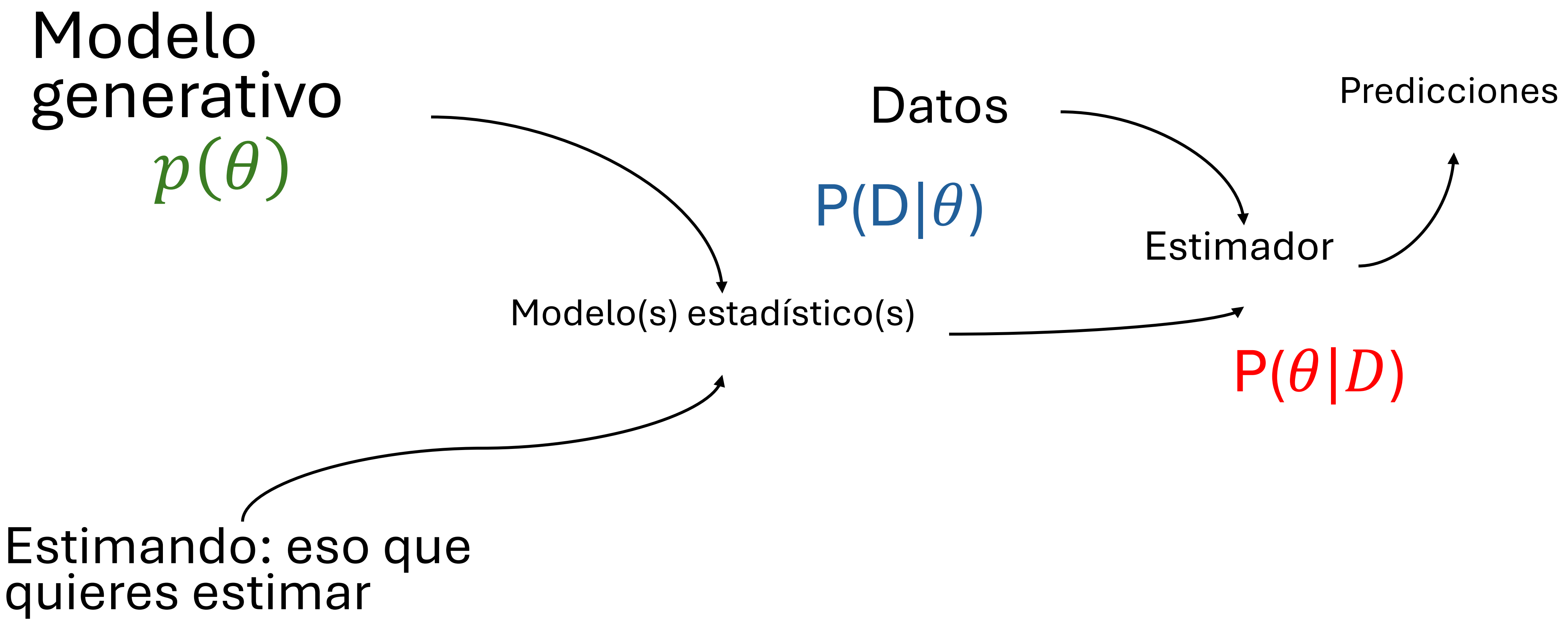
P($\theta|D$)



Para hacer inferencia sobre la posterior necesitamos el área



Para calcular el 95% (ignorar el 2.5% extremo), necesitamos integrar sobre todas las posibilidades de los valores posible del parámetro


Modelo generativo
$p(\theta)$
Datos
Predicciones
$\mathrm{P}(\mathrm{D}|\theta)$
Estimador
Modelo(s) estadístico(s)
$\mathrm{P}(\theta|D)$
Estimando: eso que quieres estimar

Modelo
generativo
$p(\theta)$

P(D|θ)

Modelo(s)
estadístico(s)

Estimando: eso
que quieres
estimar

Ejemplo de un objetivo típico de la generalización

- ¿Cuál es la proporción de agua/tierra en el planeta? $p(\theta)$

- Obtengo 29 observaciones aleatoriamente

- ¿Qué tipo de inferencia me gustaría hacer?

$$\mathrm{P}(\theta|D)$$

“La distribución posterior de la proporción estimada de agua conforme a la evidencia que acumulamos y nuestro conocimiento previo”

Observaciones

Tierra=0

Agua=1

D = 0 1 0 0 0 1 1 1 1 0 1 1 0 1 1 0 1 0 1 1 0 1 1 1

¿Por qué no puedo usar la media aritmética?

¿Por qué la media de D es insuficiente para hacer inferencias?

- Si tuviéramos los datos que queremos no necesitaríamos incorporar la incertidumbre en nuestros cálculos

- Si tuviéramos una hipótesis sobre p(θ), ¿Cómo reubicaríamos nuestra credibilidad de p(θ) toda vez que vimos D?

- Las fuentes de incertidumbre nos obligan a trabajar con parámetros que a su vez sea expresiones de la propia incertidumbre de p(θ)

Modelos generativos

- Si los datos fueran perfectos y cómo los quisiéramos, bien podríamos afirmar que no tenemos que buscar hacer inferencias
- Modelamos porque buscamos describir y descontaminar los patrones
- Son los parámetros de los modelos los que nos permiten hacer tales descripciones y conclusiones

¿Qué haría un Bayesiano?

- ¿Qué sabemos sobre la proporción de agua en el planeta?
- ¿Cuál sería un modelo razonable generativo de los datos?
- Codificaría el modelo generativo:
  - Parámetros de distribuciones
  - Distribuciones de parámetros
- Calcularía usando el teorema de bayes (Próximas clases)

Modelo generativo

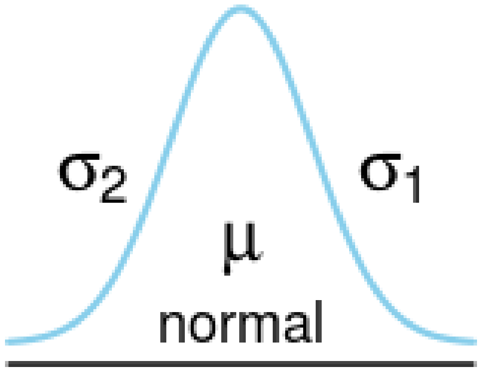
¿Cómo podemos describir (matemáticamente) el proceso que *describe* a los datos?

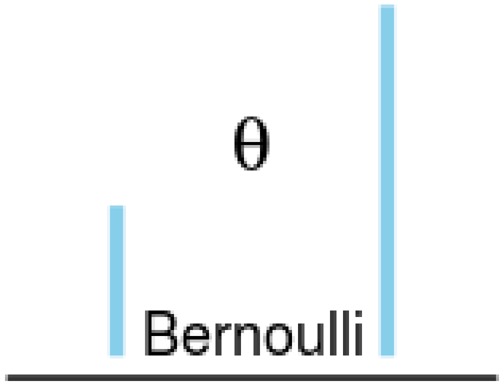
Necesito un modelo razonable (*no el verdadero*) sobre el conjunto posible de resultados… veamos algunos candidatos

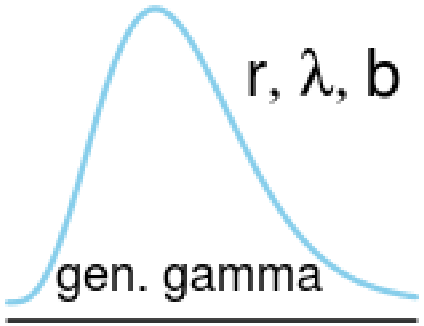
Componentes del modelo generativo

¿De qué están hechos estos modelos? I

- **Parámetros de distribuciones:** Eso de lo que están hechas las distribuciones





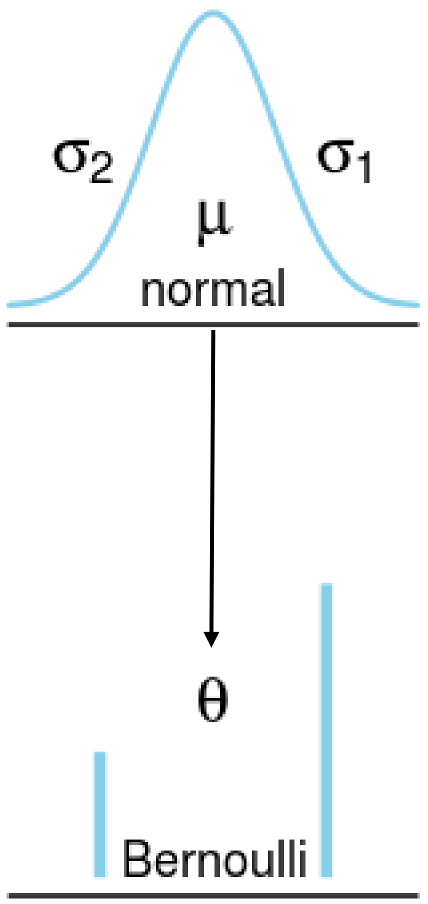


¿De qué están hechos estos modelos? II

Distribuciones de parámetros:

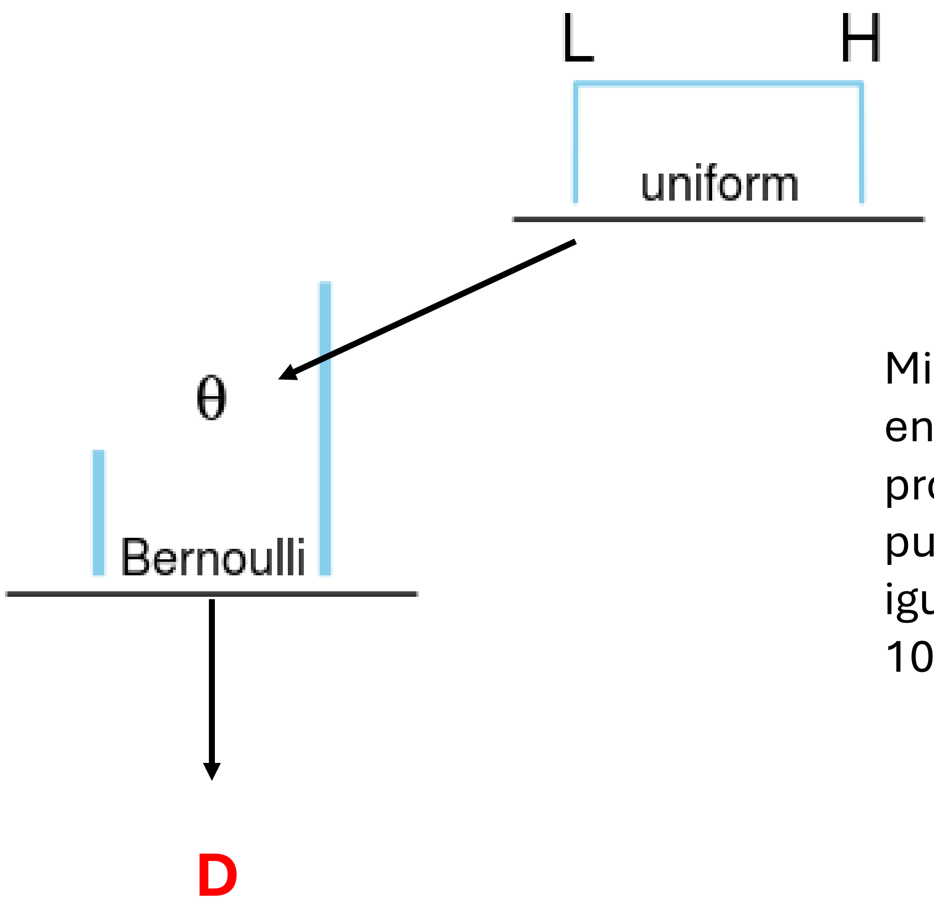
- Los posibles valores de los parámetros que general al parámetro de interés

- Hiperparámetro: Distribuciones de parámetros que generan otros parámetros



Scaffolding

"Understanding your model by comparing it to related models."
Gelman et al.



Mi modelo “de entrada/a priori” propone que θ (**p**) puede ser igualmente 0 o 100%

```
y1 <- rbinom(n = 100, size = 1, prob = 0)
y2 <- rbinom(n = 100, size = 1, prob = 0.5)
y3 <- rbinom(n = 100, size = 1, prob = 1)
mean(y1)
```

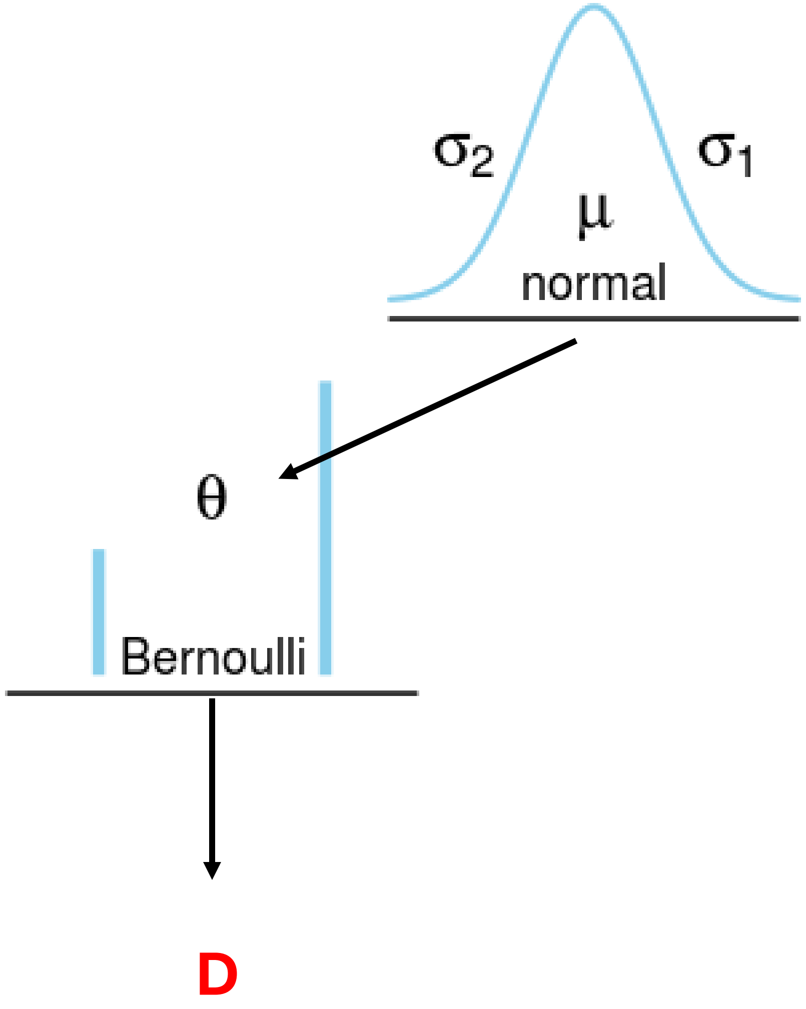
≈ 0

```
mean(y2)
```

$\approx .5$

```
mean(y3)
```

≈ 1

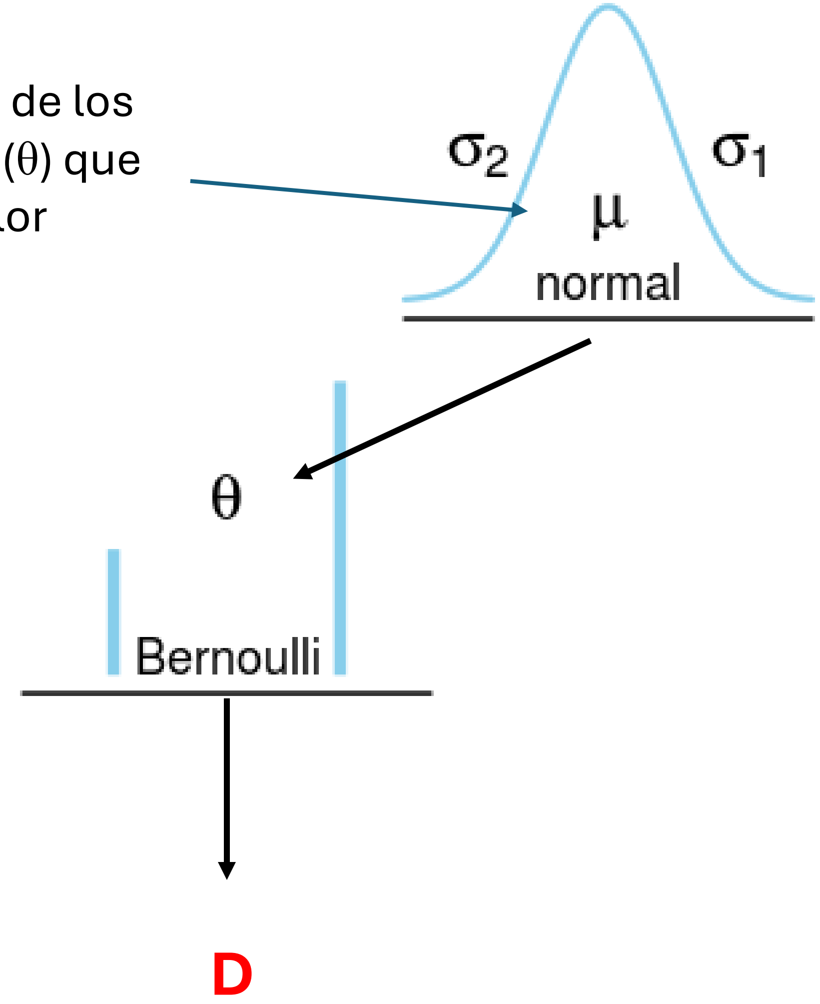


Mi modelo "de entrada/a priori" propone que θ (**p**) tiene como valores más probables μ=.5 pero permite cierta dispersión σ.

```
y1 <- rbinom(n = 100, size = 1, prob = 0.5)
y2 <- rbinom(n = 100, size = 1, prob = 0.55)
y3 <- rbinom(n = 100, size = 1, prob = 0.45)

mean(y)
≈ 0.5
mean(y)
≈ 0.55
mean(y)
≈ 0.45
```

No es el valor central de los posibles valores de p(θ) que van de 0-1, sino el valor promedio esperado.

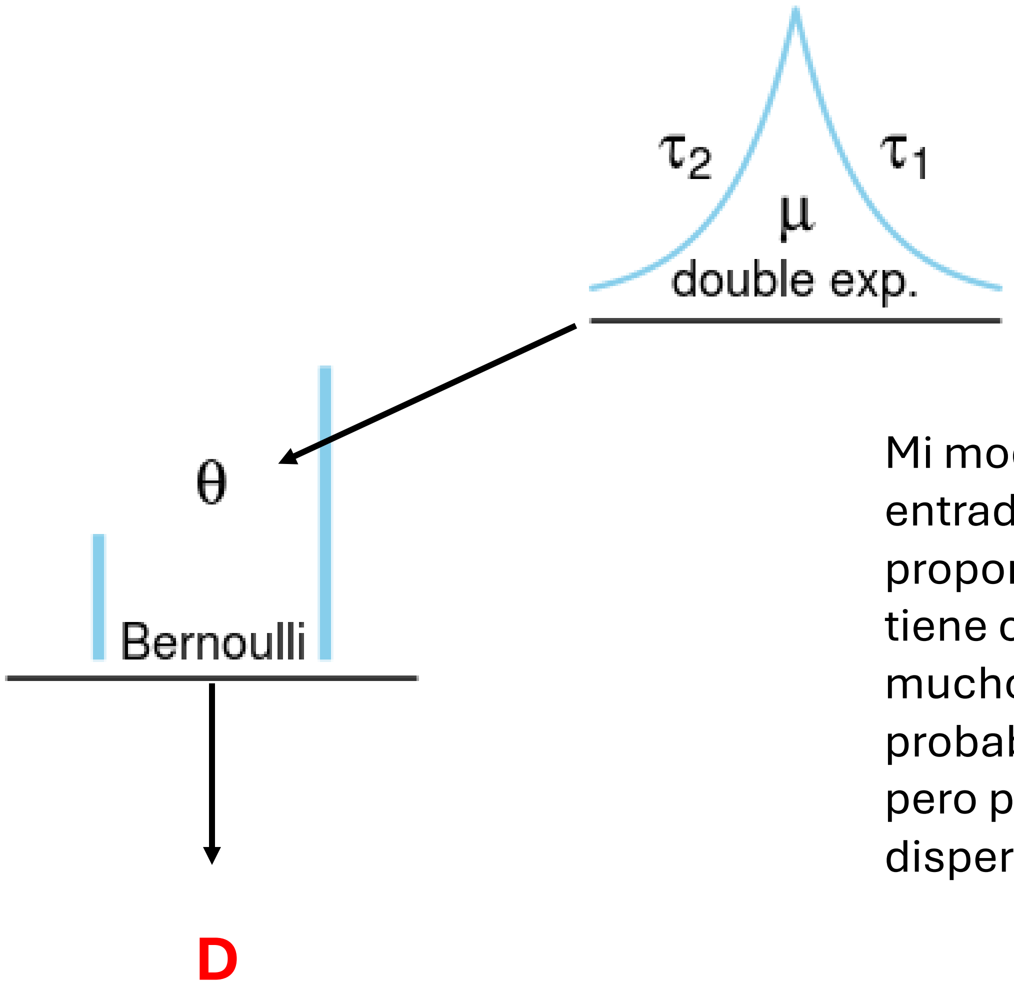


Mi modelo "de entrada/a priori" propone que θ (**p**) tiene como valores más probables μ=.7 pero permite cierta dispersión σ.

```
y1 <- rbinom(n = 100, size = 1, prob = 0.7)
y2 <- rbinom(n = 100, size = 1, prob = 0.65)
```

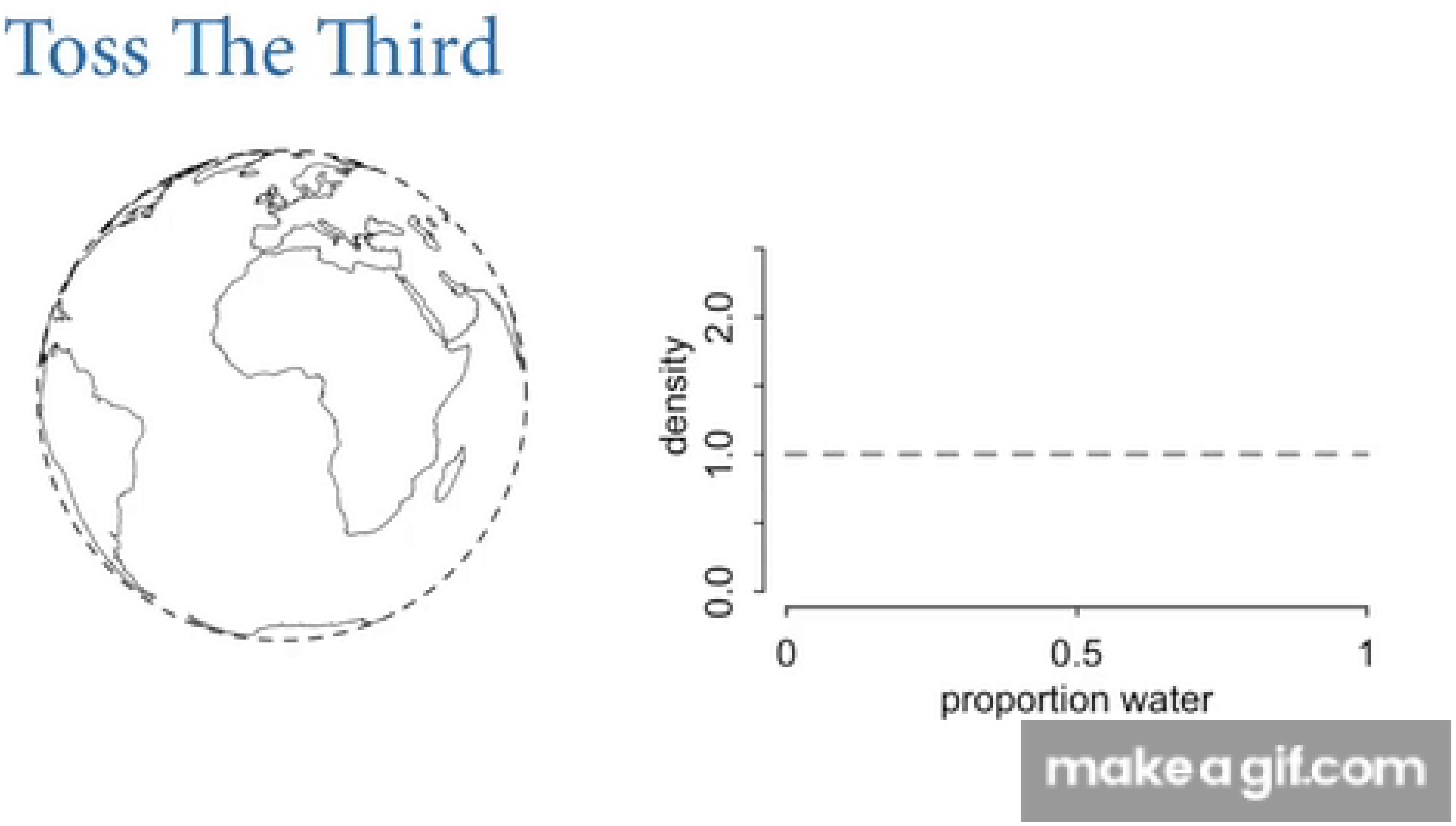
mean(y1)
$\approx$ 0.7
mean(y2)
$\approx$ 0.65

```
y <- rbinom(n = 100, size = 1, prob = 0.5)
mean(y)
```
$\approx$ 0.5

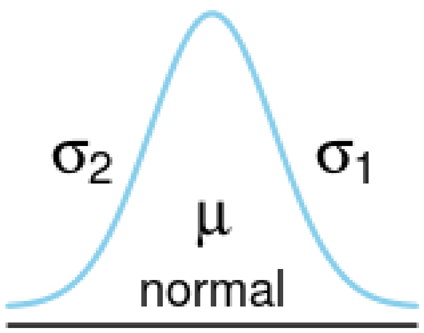


Mi modelo “de entrada/a priori” propone que θ (**p**) tiene como valores mucho más probables μ=.5 pero permite cierta dispersión τ.

Reubicación de credibilidad



$P(\theta|D)$



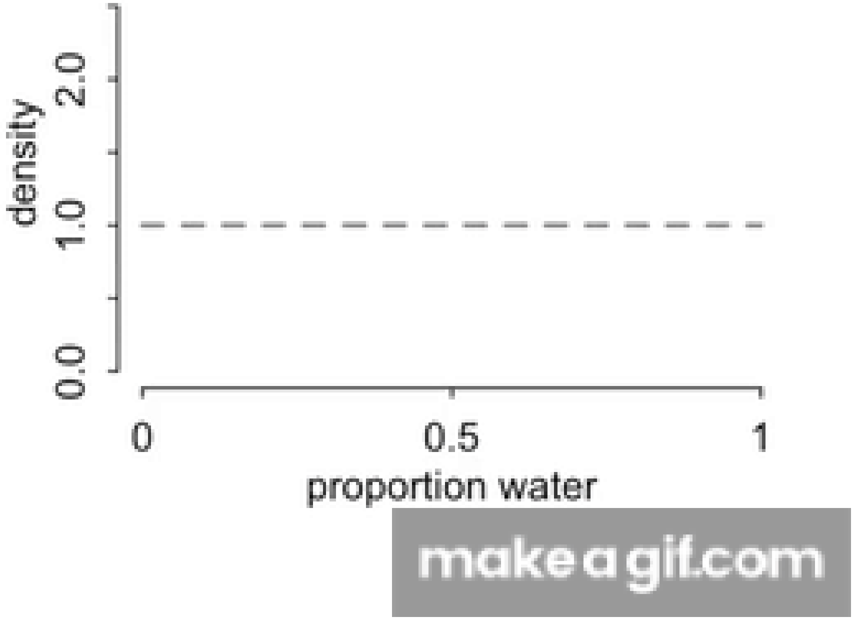
$\mu = .65$

$\sigma = .07$

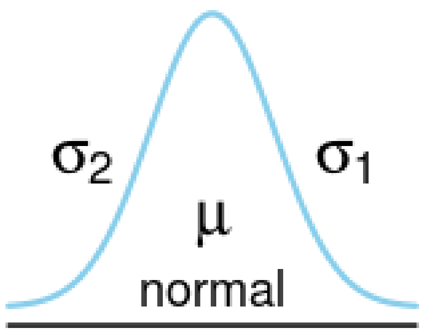
Cr 95% = .53 - .76

Reubicación de credibilidad II (n=29+100)

Toss The Third



P($\theta|D$)



$\mu = .71$

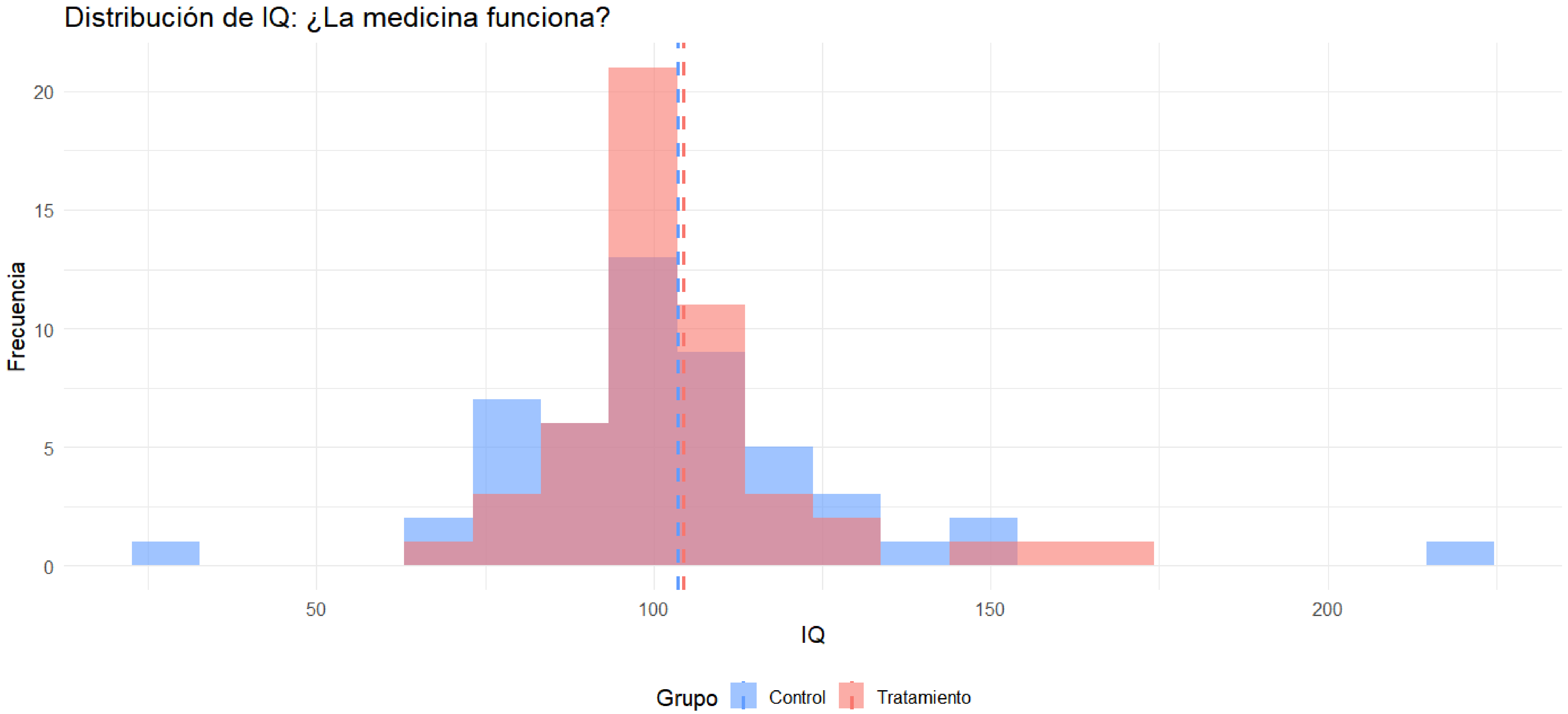
$\sigma = .02$

Cr 95% = .65 - .75

Modelos generativos

Segundo ejemplo

La facultad de medicina nos da el siguiente problema

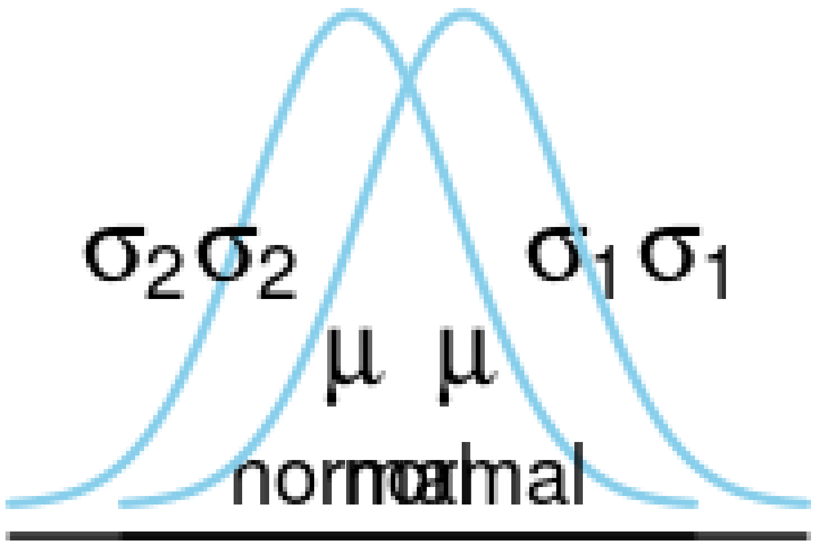


Generalización del control al tratamiento

- ¿Qué tan diferente es el resultado de un grupo respecto a otro?
- ¿Qué tanta certeza temenos sobre la magnitudes de la diferencia?
- ¿Podemos decir con cierta credibilidad que la diferencia supera cierto tamaño?

¿Qué necesitamos para contestar tal pregunta?

- La distribución de los posibles valores de las medias estimadas, condicionadas en el modelo y los datos, de cada uno de lo grupos

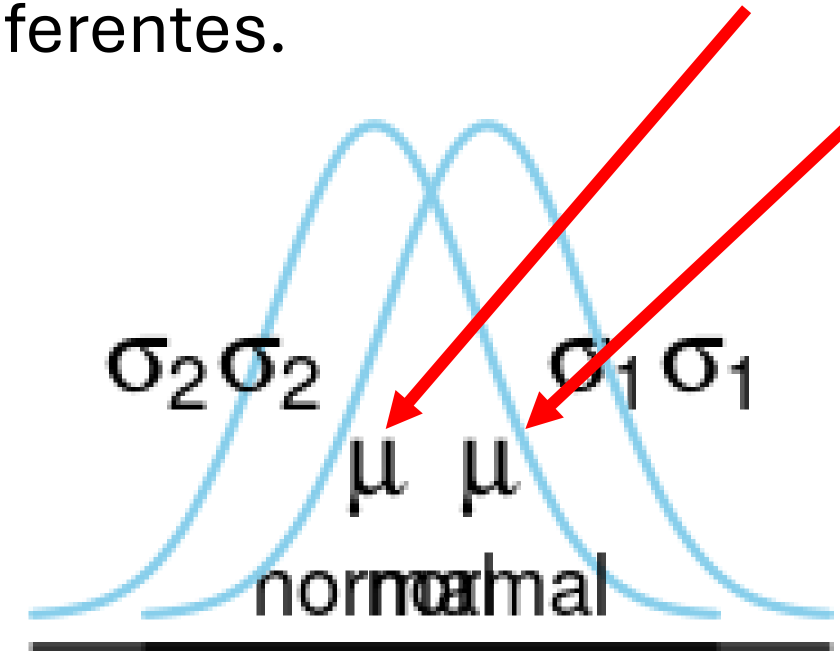


¿Qué haría un Bayesiano?

- ¿Qué sabemos sobre el tratamiento y sobre el experimento?
- ¿Cuál sería un modelo razonable generativo de los datos?
- Codificaría el modelo generativo:
    - Parámetros de distribuciones
    - Distribuciones de parámetros
- Calcularía usando el teorema de bayes

Modelos generativos

- Cuando me pregunto si el efecto es distinto, en el fondo me estoy preguntando si los parámetros que hacen a tales distribuciones observadas son diferentes.



Componentes del modelo generativo

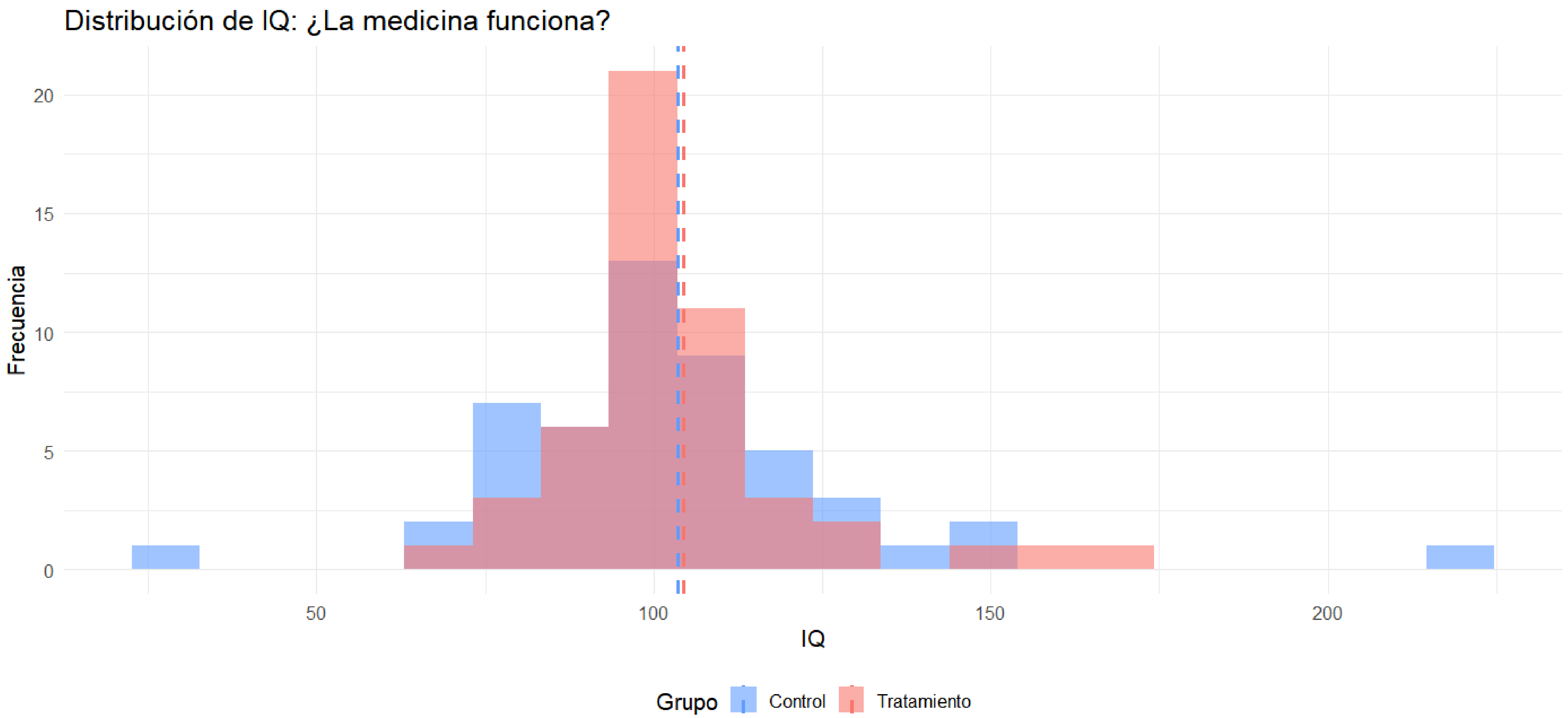
Componentes del modelo generativo

Valores posibles

- Las medias de ambos grupos μ_1 & μ_2
- La dispersión respecto a la media σ_1 & σ_2
- Los parámetros Θ tienen su propia distribución (distribuciones de parámetros)

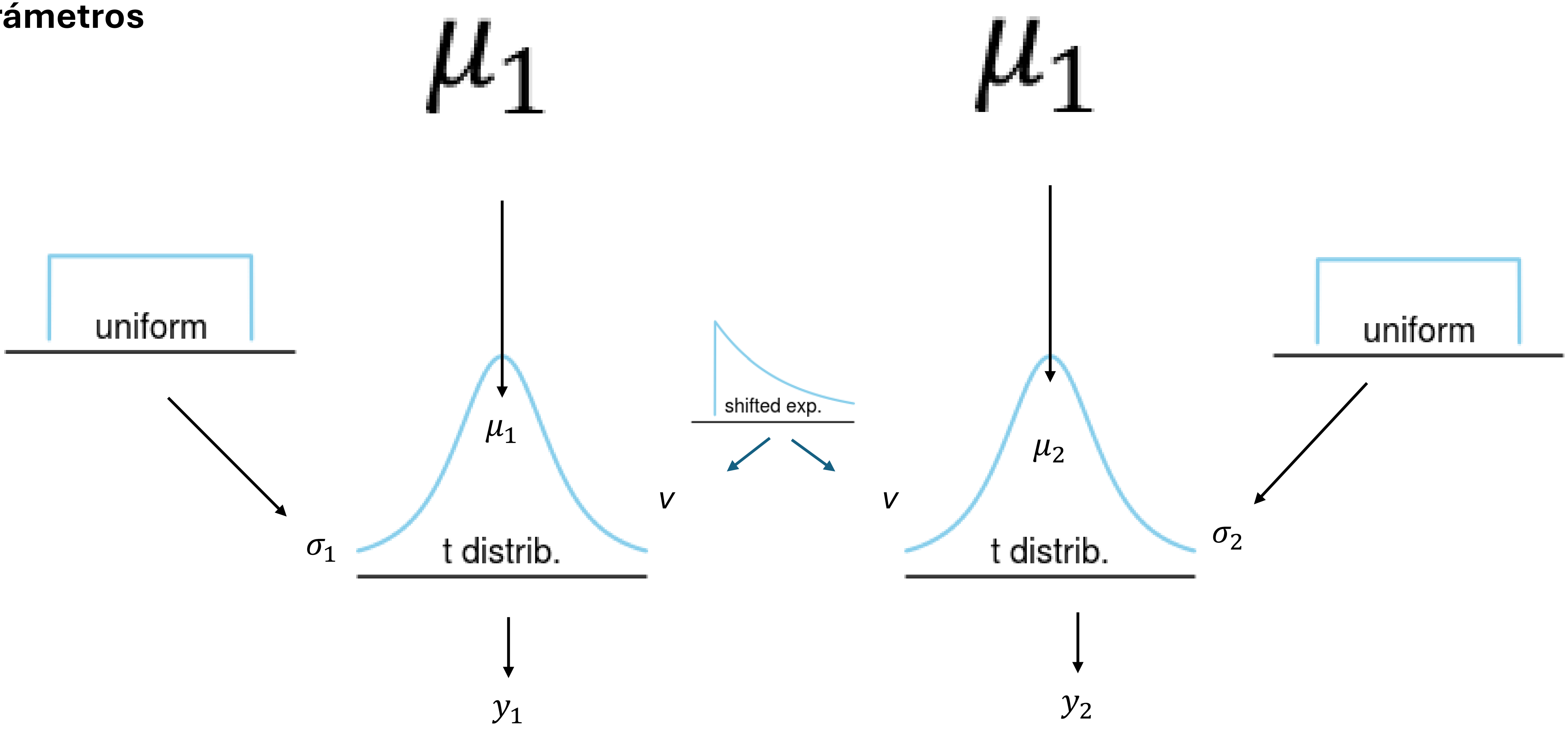
Si estimamos los valores creíbles de tales parámetros (μ_1 & μ_2) podemos proceder a hacer las comparaciones buscadas: $\mu_1 - \mu_2$

¿Cómo describir estos datos?

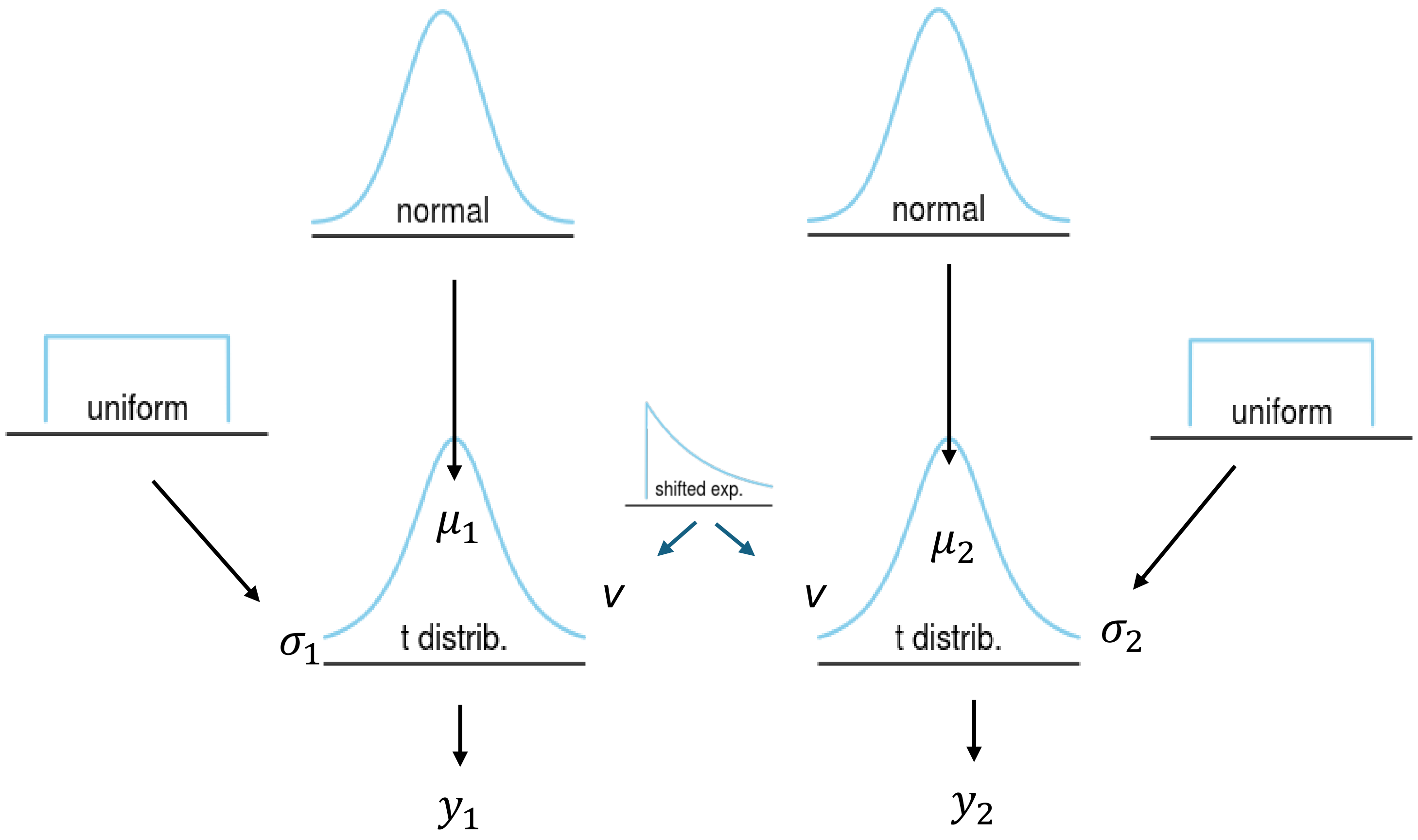


Modelo estadístico generativo de cinco parámetros

Distribuciones a priori de los parámetros



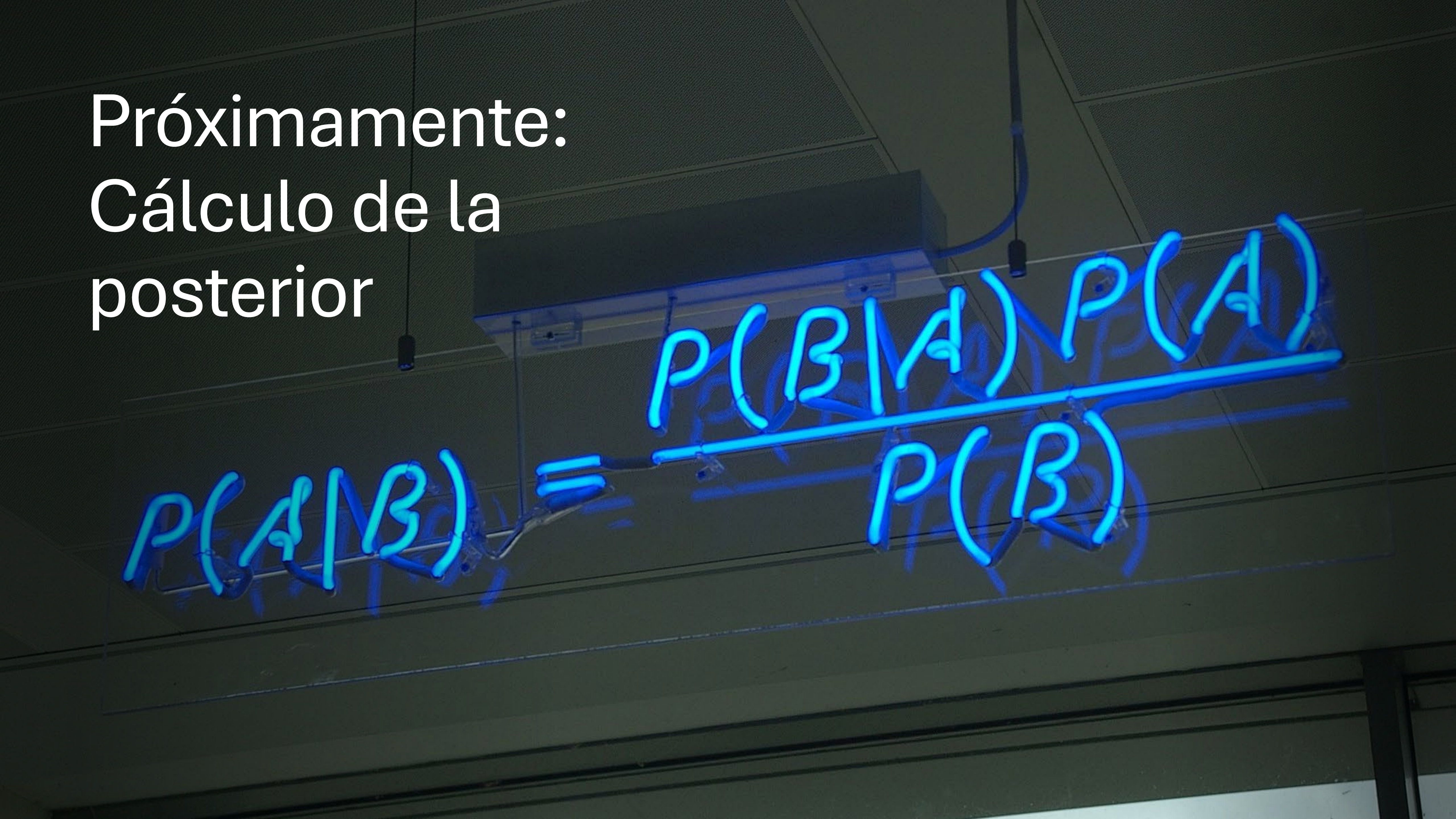
Modelo estadístico generativo



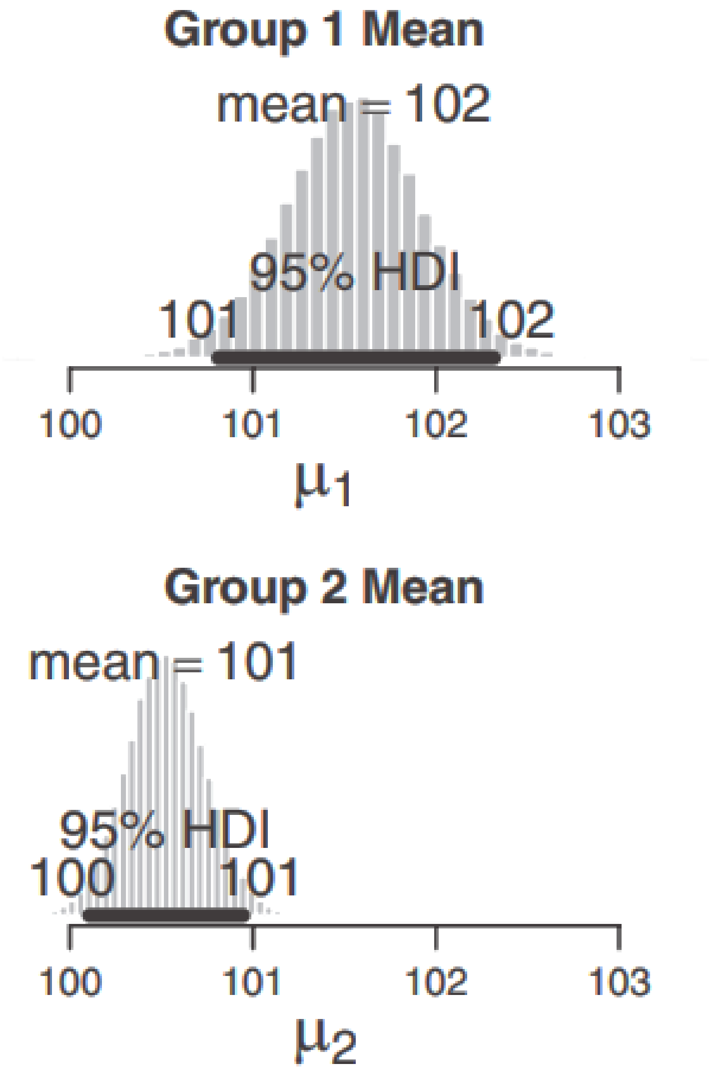
$$P(\Theta \mid D)$$

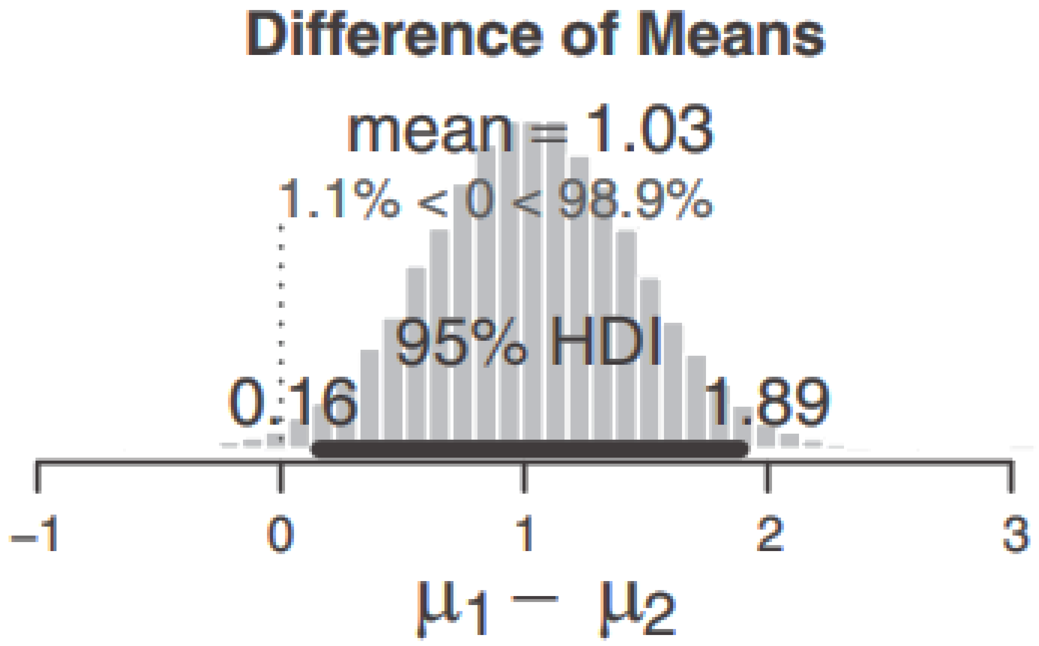
Distribuciones posteriores de los parámetros

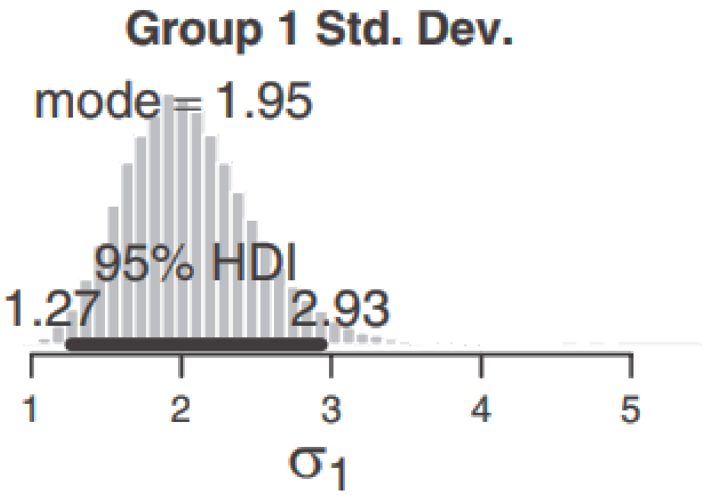
Próximamente: Cálculo de la posterior

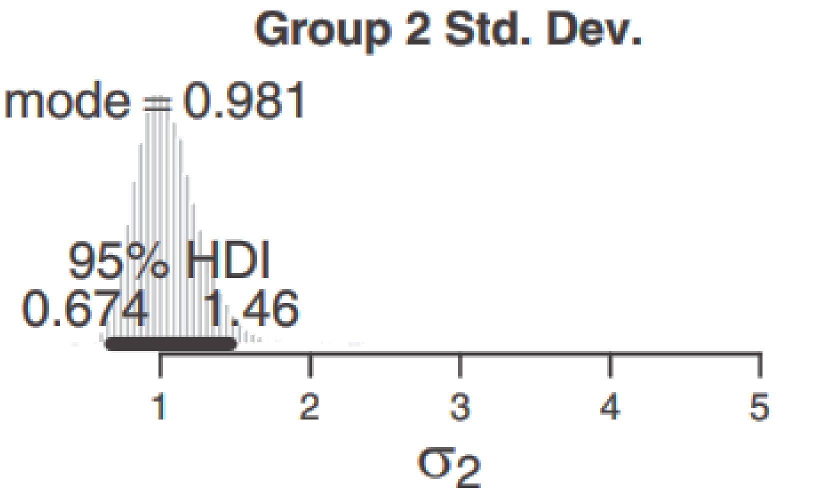


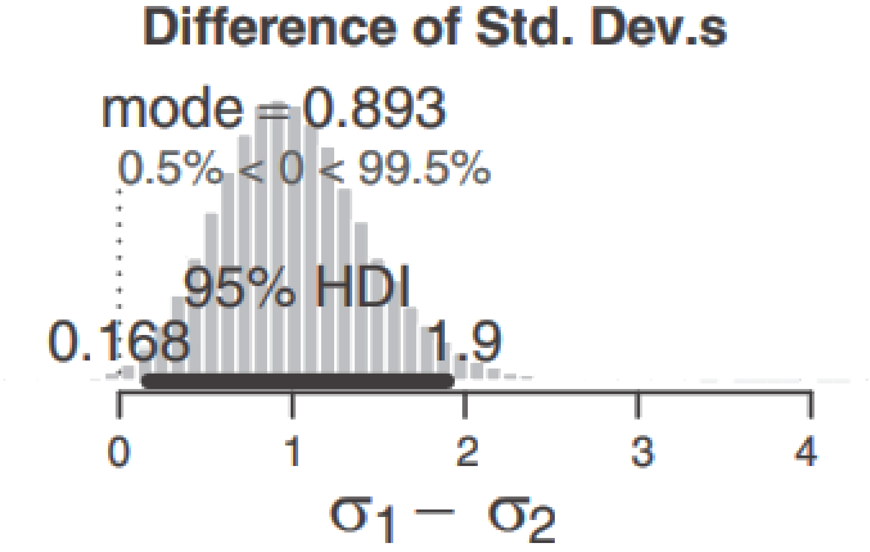
Distribuciones posteriores de los parámetros











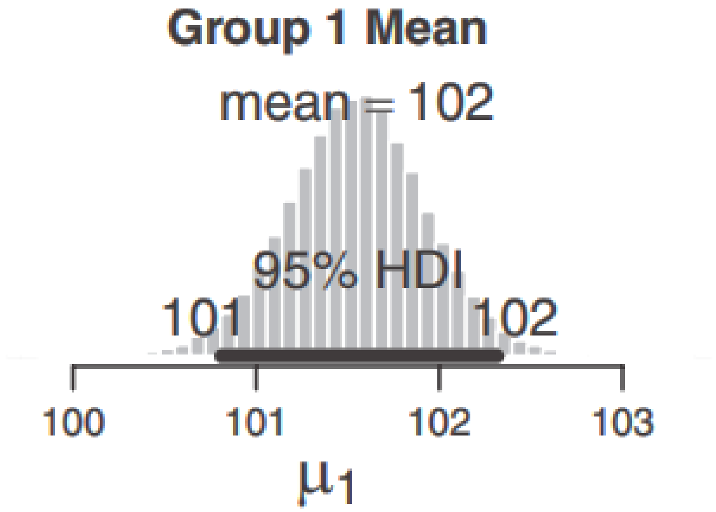
Distribuciones posteriores de los parámetros

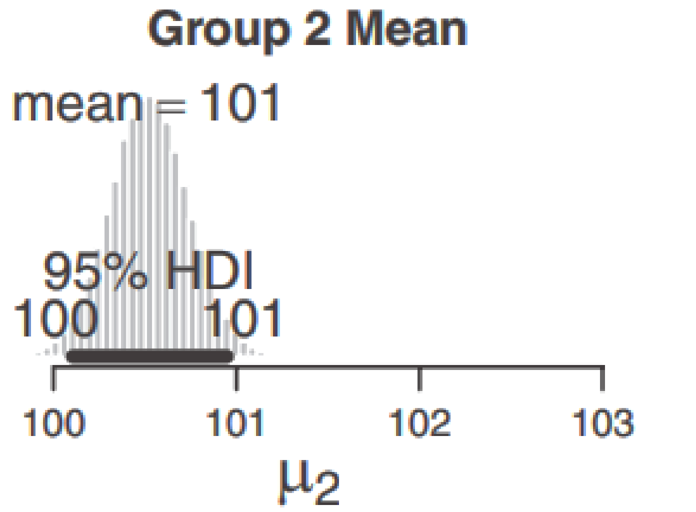

HULK, IT'S GREAT HAVING YOU IN THE AVENGERS! CAN YOU OPEN THIS JAR OF PICKLES?
NICKEL

¿El t-test clásico/frecuentista hace lo mismo cierto?

- Para poder hacer cualquiera de las siguientes afirmaciones:
  - Es 95% probable que las medias del grupo 1 y 2 sean diferentes
  - La magnitud de la diferencia es de 1.03 con un rango creíble de .4-1.8

- Se necesita la distribución de los valores probables de μ_1& μ_2





¿Qué tenemos entonces?

```
> print(t_test_result)

        Two Sample t-test

data:  value by group
t = 1.5587, df = 87, p-value = 0.1227
alternative hypothesis: true difference in means between group Drug and group Placebo is not equal to 0
95 percent confidence interval:
 -0.428653  3.544155
sample estimates:
   mean in group Drug mean in group Placebo
             101.9149              100.3571

>
```

El t-test frecuentista no nos dice nada de la probabilidad de que las medias sean exactamente iguales o diferentes.

Lo que sí nos dice es que es probable que la diferencia observada sea producto de la suerte, dado que son exactamente iguales.

El resultado no sólo es contrario al que obtuvimos con la aproximación bayesiana sino que nos da poca evidencia para hacer la inferencia que queríamos hacer.

Sin la distribución del parámetro de interés, no hay inferencias ricas y razonables.

“Statistics is said to be the science of defaults. One of our challenges is to defaultize things.” Gelman et al (2003)

Pero el t-test es un modelo particular

- Todos los modelos estadísticos tienen parámetros
- Estos parámetros tienen distribuciones
- Es sobre dichas distribuciones que busco hacer inferencias sobre la credibilidad de sus valores

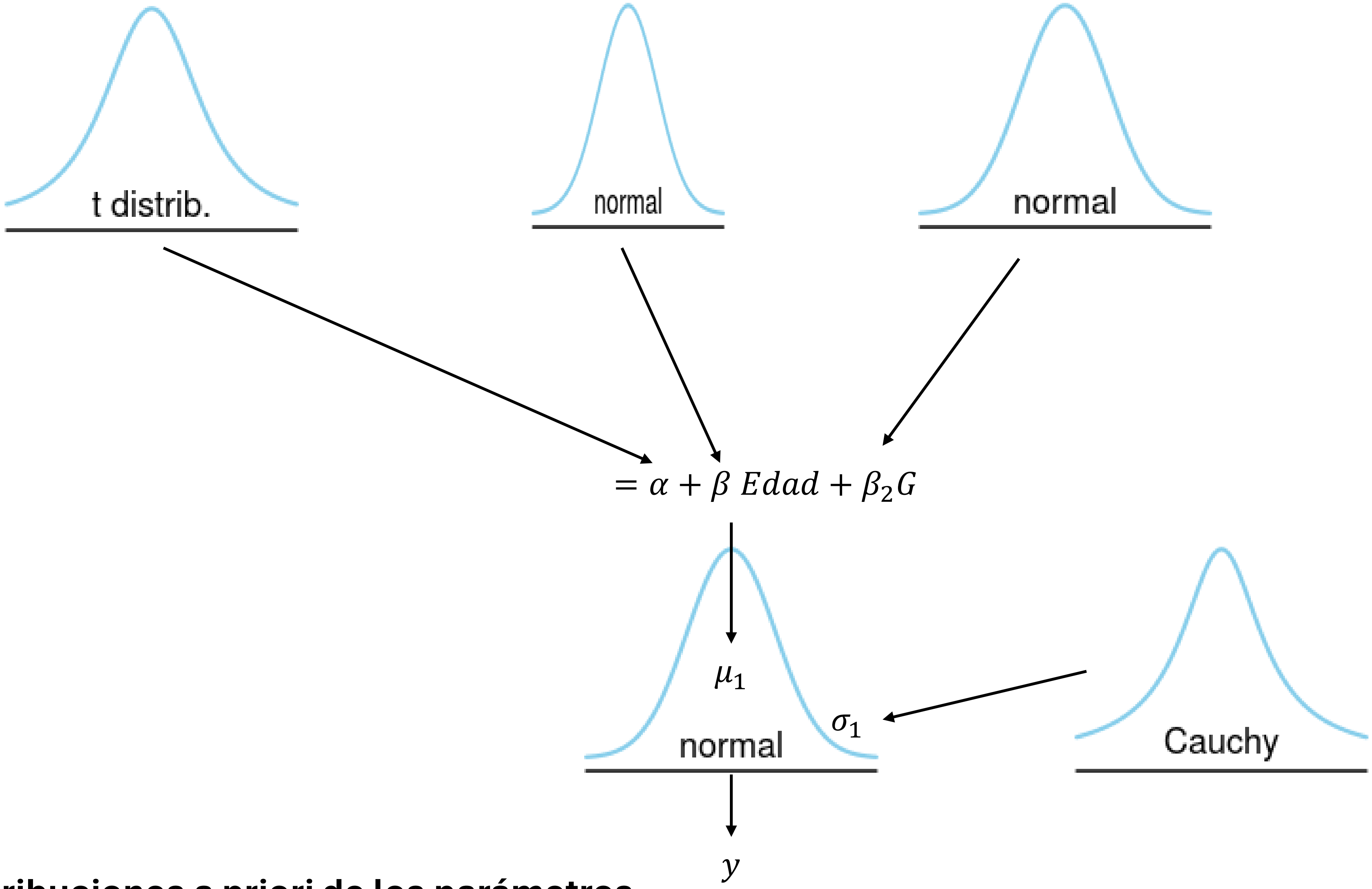
Por ejemplo

¿Qué tan distinto es un grupo respecto a otro?

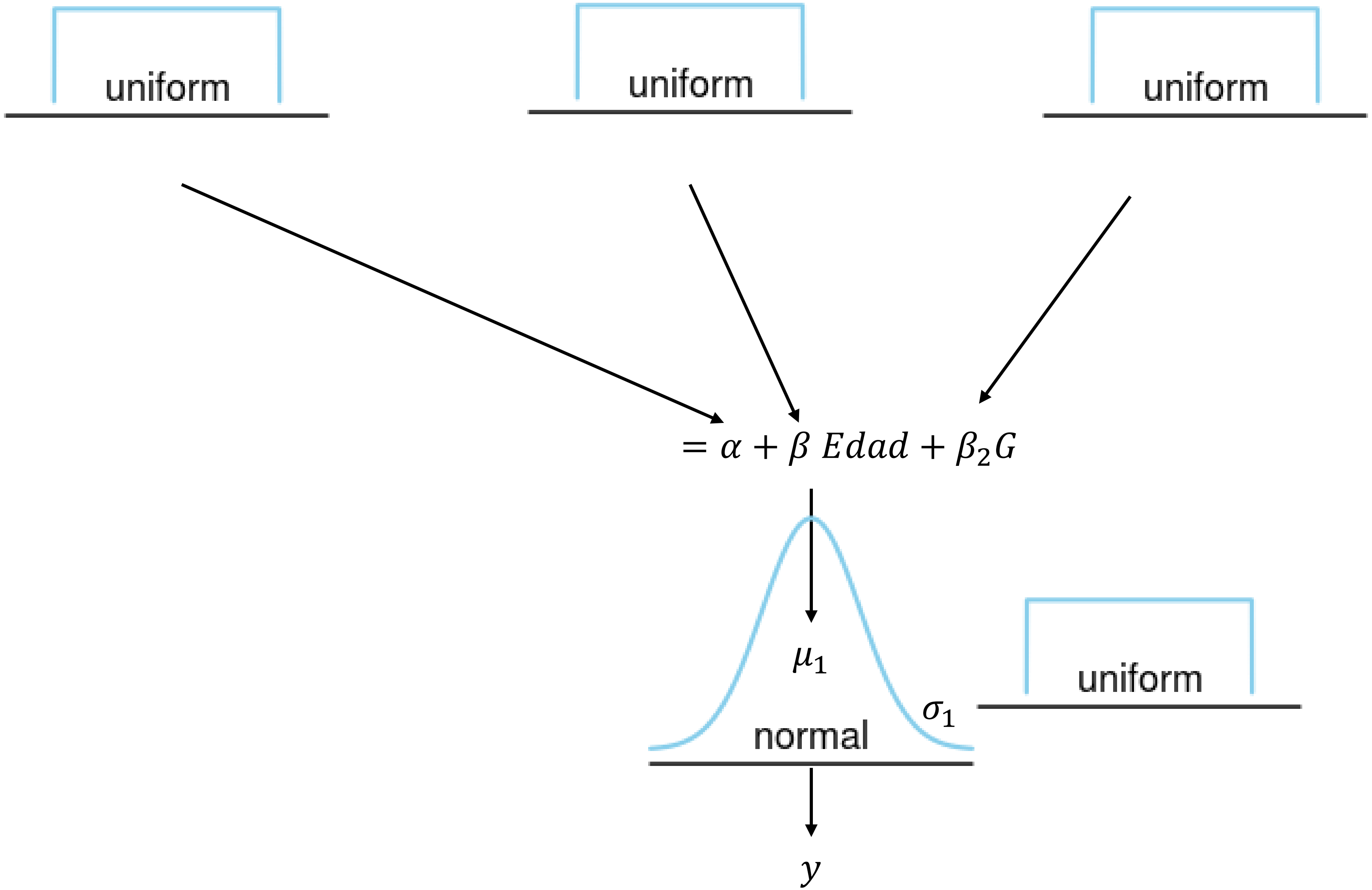
H: El salario promedio del grupo 1 es mayor al del grupo 2 condicionando por edad

Si tomo una muestra de ambos grupos:

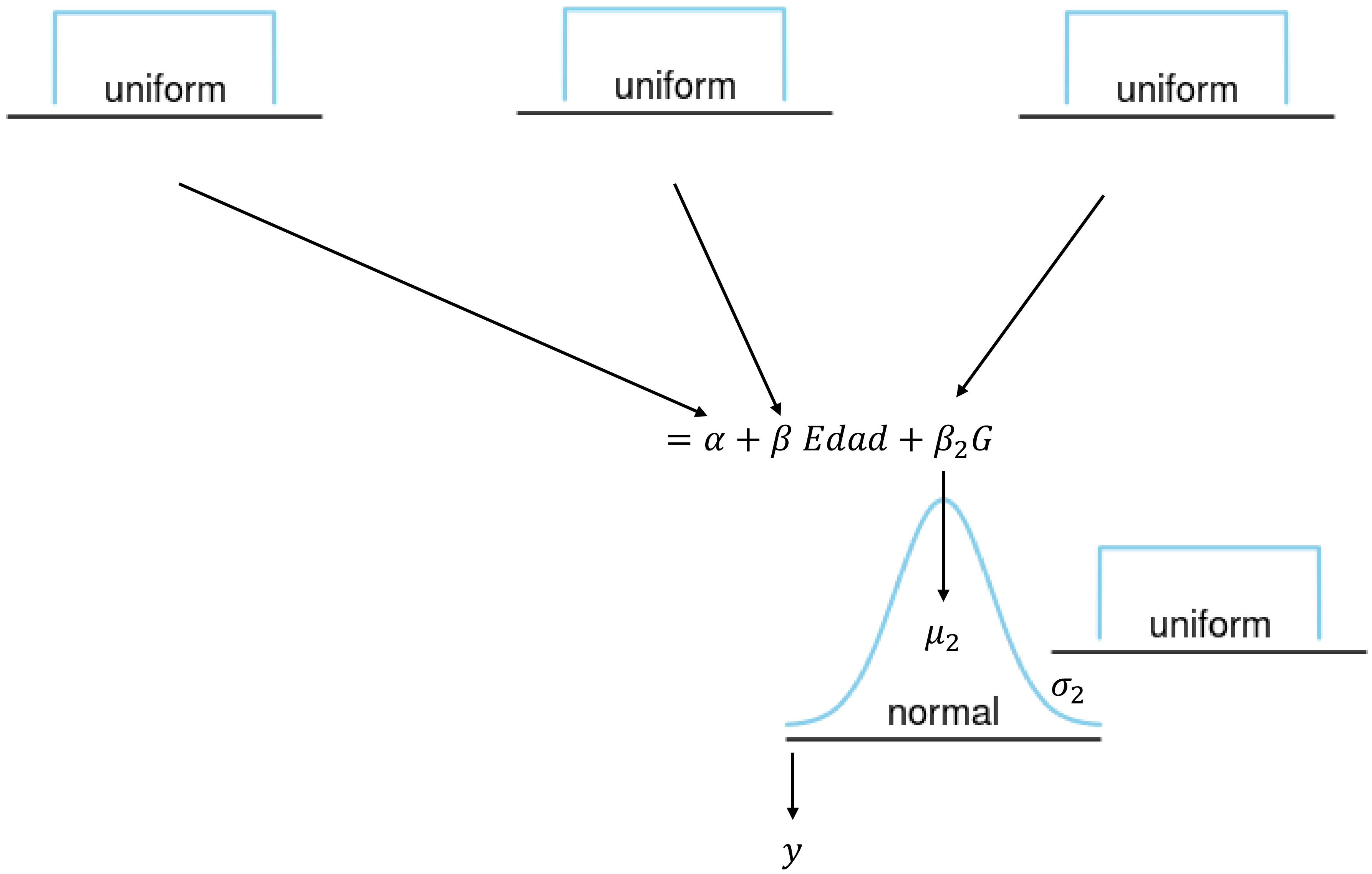
¿Basta con comparar aritméticamente las distribuciones de los datos observados para poder **generalizar** mis hallazgos respecto a H?



Distribuciones a priori de los parámetros



Distribuciones a priori de los parámetros: El modelo "determinístico" es el mismo pero el modelo generativo no



Distribuciones a priori de los parámetros: El modelo “determinístico” es el mismo pero el modelo generativo no

Conclusiones

- Los modelos generativos establecen vínculos explícitos entre la pregunta de investigación, la evidencia (datos) y las hipótesis sustantivas
- El razonamiento bayesiano es fácil de seguir, hace sentido y nos permite contestar las preguntas que nos planteamos
- La distribución posterior permite hacer inferencias directas sobre la magnitud e incertidumbre de las hipótesis sustantivas de investigación
- Los datos no se acomodan al modelo, al revés: Lo que importa es que información utiliza no lo que le hace a los datos

Próxima clase

Capítulo 3: Sampling the imaginary. Statistical Rethinking. McElreath

Capítulo 5: Bayes rule. Doing Bayesian data análisis. Krushcke

Referencias

- ¿Qué es la probabilidad?

  - Lennox, K. [Lawrence Livermore National Laboratory]. (2016, September 27). *All About that Bayes: Probability, Statistics, and the Quest to Quantify Uncertanty* [Video]. YouTube. https://youtu.be/eDMGDhyDxuY
  - Kozyrkov C. [Cassie Kozyrkov]. (2020, September 2). *Are you Bayesian or Frequentist?* [Video]. YouTube. https://youtu.be/GEFxFVESQXc?si=VA2vqYQEhqSsgN55