

# **Analítica de Datos**

## **Pruebas de Hipótesis**

**Carlos Cardona Andrade**



# Tabla de contenido

1. Intervalos de Confianza
2. Pruebas de Hipótesis
3. Algunas consideraciones sobre Pruebas de Hipótesis
4. Ideas erróneas sobre Pruebas de Hipótesis

# **Un poco más sobre Intervalos de Confianza**

# Spotify - Data

```
1 # Recuerden SIEMPRE cargar los paquetes!!
2 library(tidyverse)
3
4 # Tomes los datos de:
5 # https://www.kaggle.com/code/lusfernandotorres/spotify-top-hits-2000-2019-eda/data
6 # Llamen los archivos desde un directorio relativo y no absoluto
7
8 # Este es absoluto
9 #data <- read.csv("/Users/ccard/Dropbox/analitica_datos/slides/lecture5/data/spotify_data.csv")
10
11 # Este es relativo
12 data <- read.csv("data/spotify_data.csv")
```

# Spotify - Data

```
1 # Top 8 más populares
2 data %>%
3   select(artist, popularity, danceability) %>%
4   arrange(desc(popularity)) %>%
5   head(8)
```

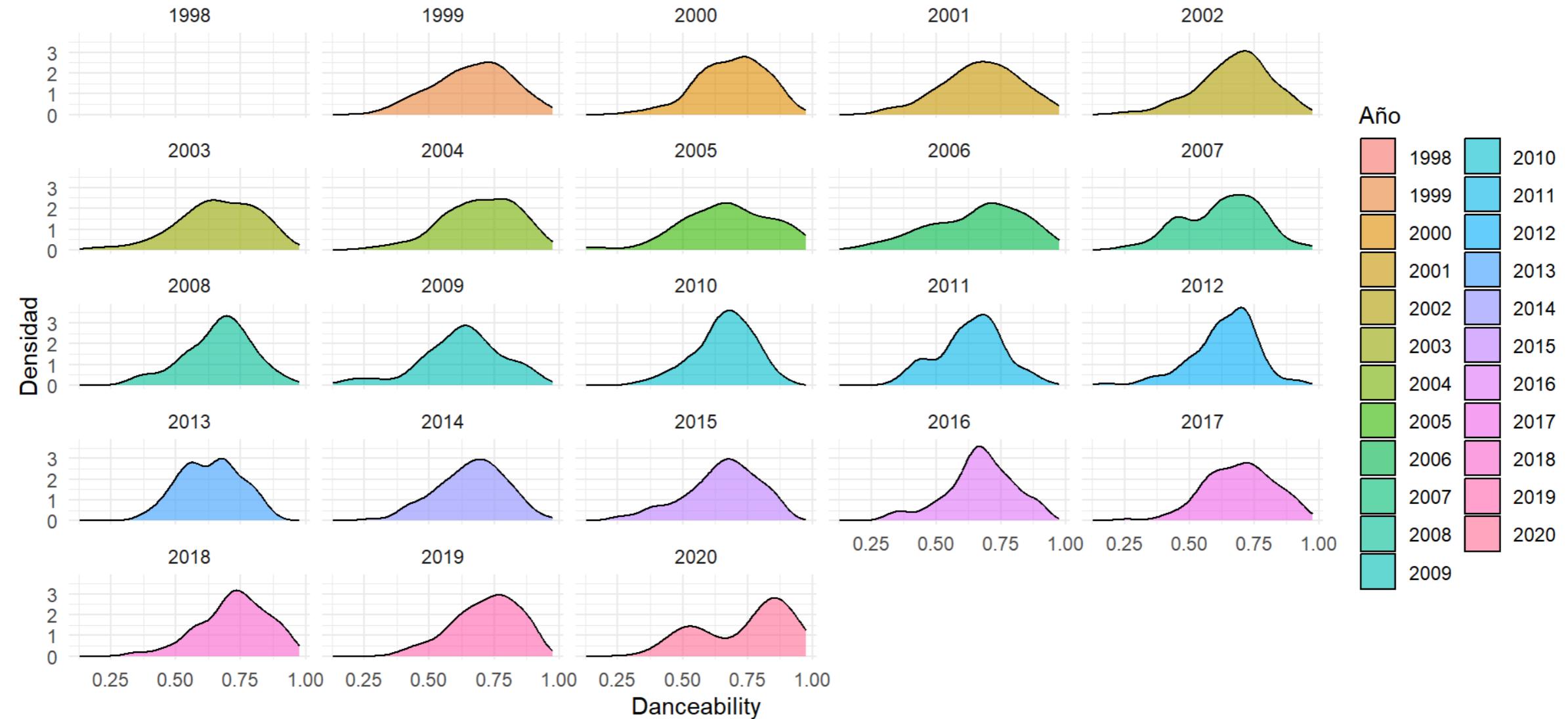
	artist	popularity	danceability
1	The Neighbourhood	89	0.612
2	Tom Odell	88	0.445
3	Eminem	87	0.908
4	Eminem	86	0.949
5	WILLOW	86	0.764
6	Billie Eilish	86	0.351
7	Billie Eilish	86	0.351
8	Eminem	85	0.548

```
1 # Cómo le va a Bad Bunny?
2 data %>%
3   select(artist, song, popularity, danceability) %>%
4   filter(artist=="Bad Bunny")
```

	artist	song	popularity	danceability
1	Bad Bunny MIA (feat. Drake)		77	0.817
2	Bad Bunny	Callaita	81	0.610

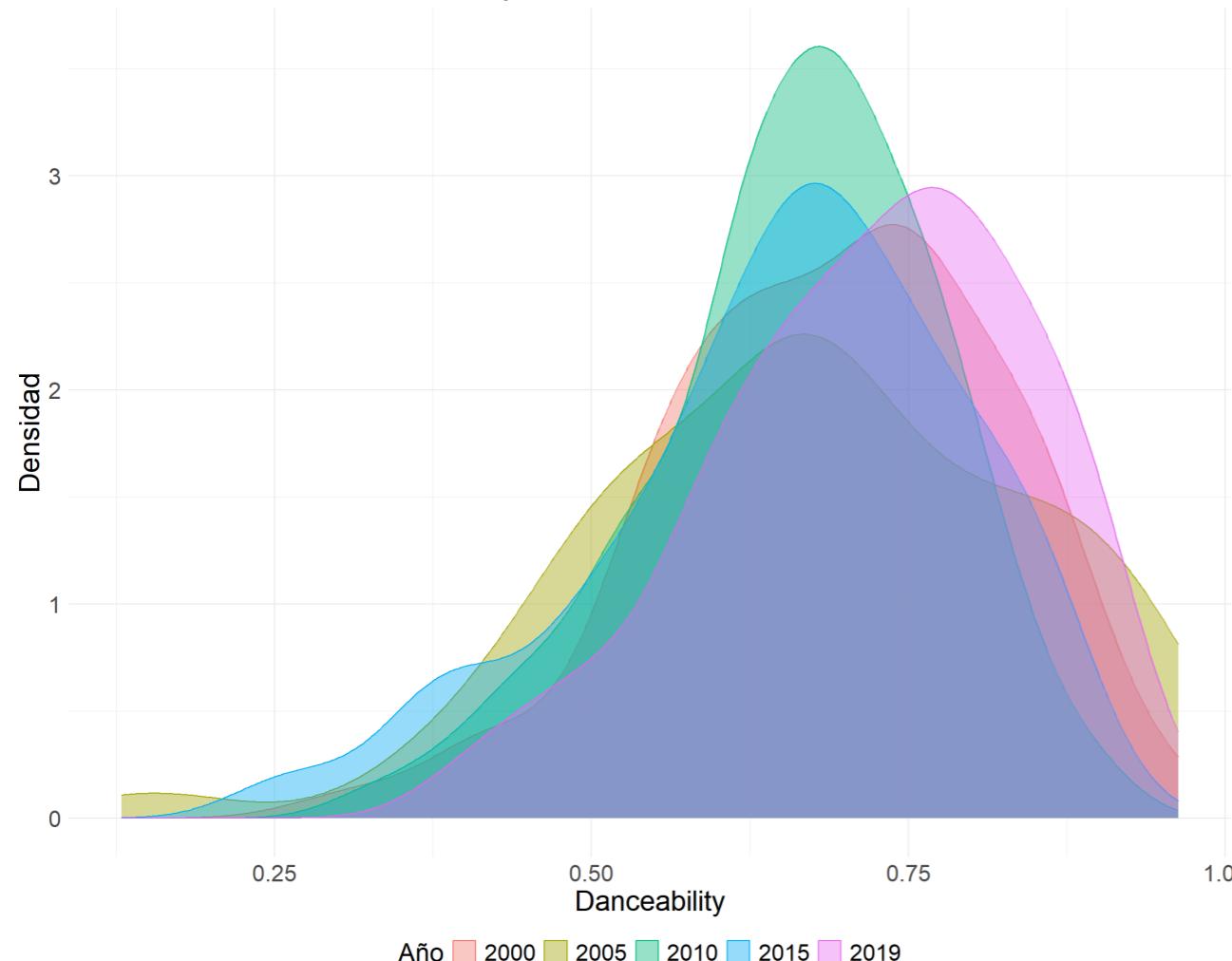
# Danceability a través del tiempo

Distribución de Danceability por Año



# Danceability a través del tiempo

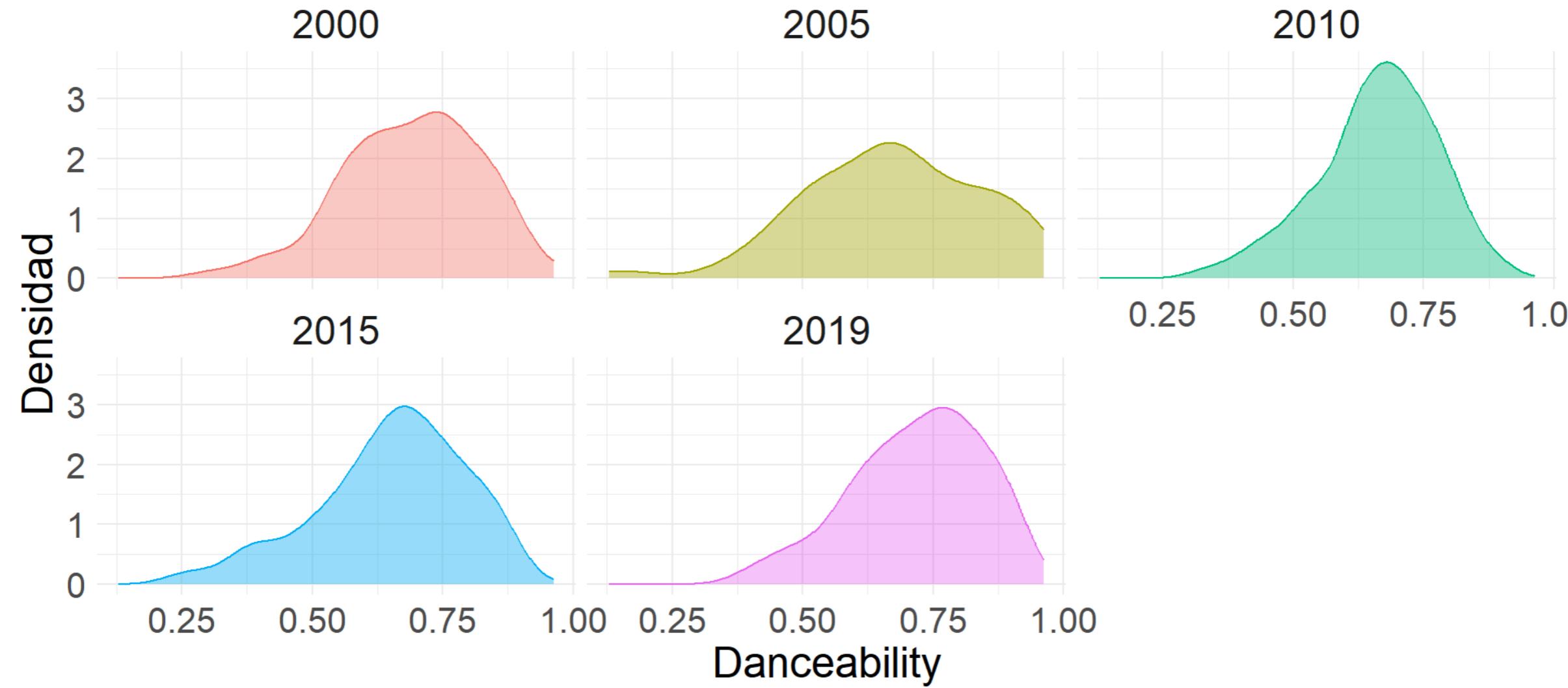
Distribución de Danceability (2000-2019)



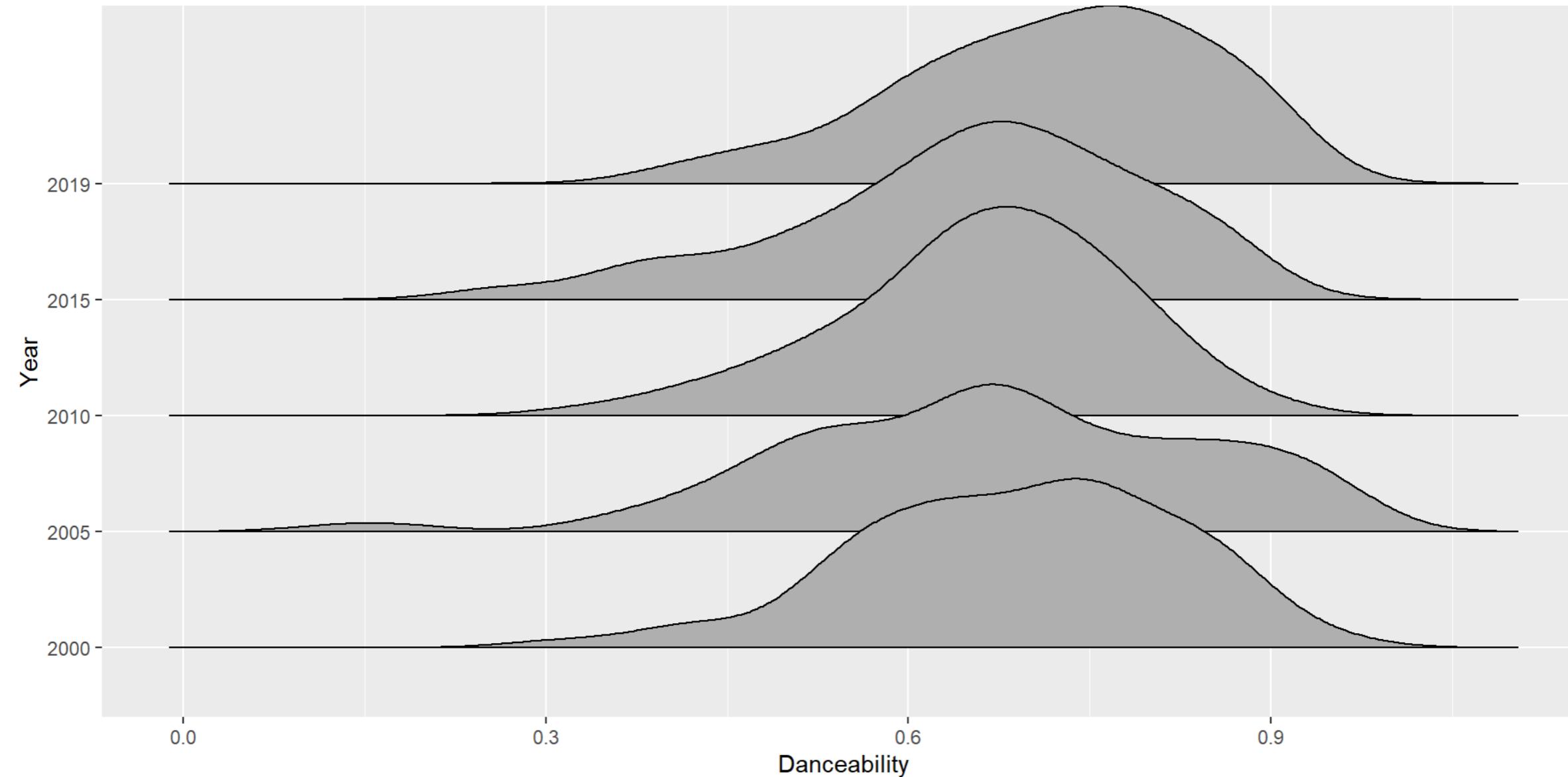
```
1 años_interes <- c(2000, 2005, 2010, 2015, 2019)
2 data_filtrada <- data %>% filter(year %in% años_interes)
3
4 ggplot() + # Density for the selected years
5   geom_density(data = data_filtrada, aes(x = danceability,
6     fill = año)) +
7   labs(title = "Distribución de Danceability (2000-2019)",
8       x = "Danceability",
9       y = "Densidad",
10      color = "Año",
11      fill = "Año") +
12   theme_minimal() +
13   theme(
14     legend.position = "bottom",
15     plot.title = element_text(size = 26),
16     axis.title.x = element_text(size = 20),
17     axis.title.y = element_text(size = 20),
18     axis.text = element_text(size = 16),
19     legend.text = element_text(size = 16),
20     legend.title = element_text(size = 18))
```

# Danceability a través del tiempo

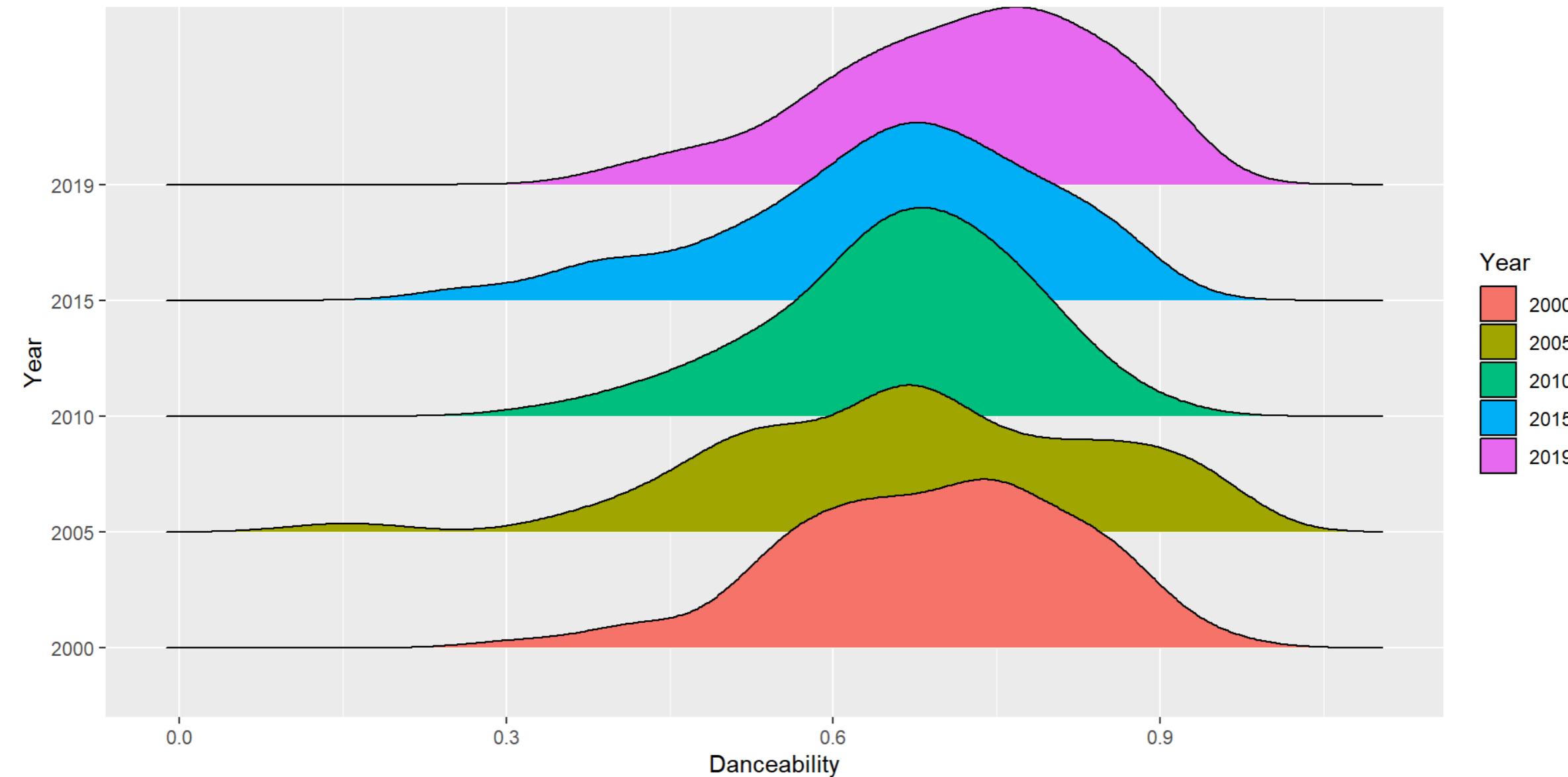
Distribución de Danceability (2000-2019)



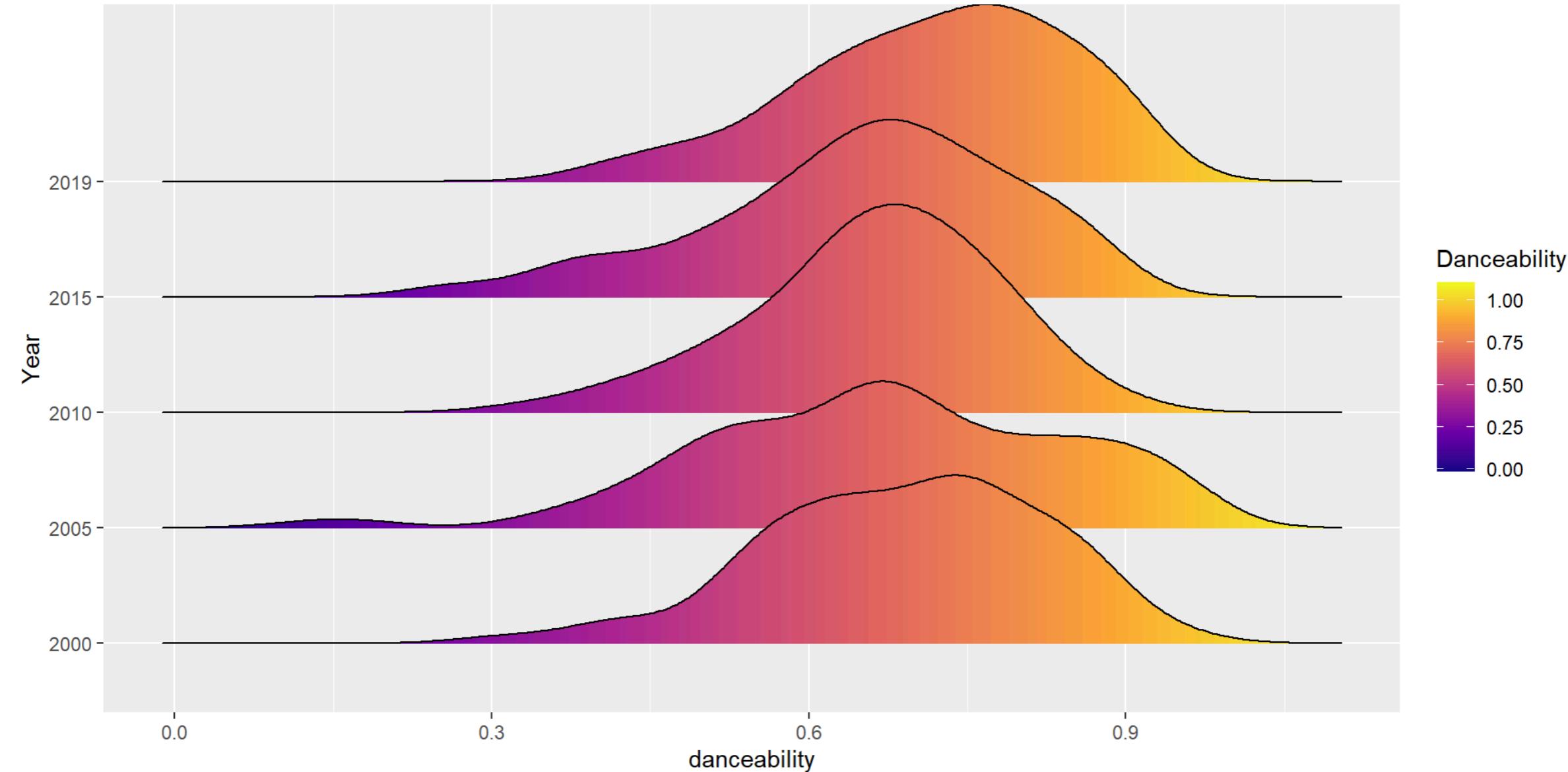
# Danceability a través del tiempo



# Danceability a través del tiempo



# Danceability a través del tiempo



# IC para Danceability - Cálculo Manual

```
1 # Calculamos la media poblacional  
2 media_poblacional <- mean(data$danceability, na.rm = TRUE)  
3 media_poblacional
```

```
[1] 0.6674375
```

```
1 # Set seed for reproducibility  
2 set.seed(123)  
3 # Tomamos una muestra de 100 canciones  
4 muestra <- data %>% sample_n(100)  
5  
6 # Calculamos la media muestral  
7 media_muestral <- mean(muestra$danceability, na.rm = TRUE)  
8 media_muestral
```

```
[1] 0.67731
```

```
1 # Calculamos el error estándar  
2 error_std <- sd(muestra$danceability, na.rm = TRUE) / sqrt(nrow(muestra))  
3 error_std
```

```
[1] 0.01394942
```

# IC para Danceability - Cálculo Manual

```
1 # Definimos el nivel de confianza (e.g., 95%)  
2 confidence_level <- 0.95  
3 z_score <- qnorm((1 + confidence_level) / 2)  
4 z_score
```

```
[1] 1.959964
```

```
1 # Calculamos los intervalos de confianza  
2 ic_inf <- media_muestral - z_score * error_std  
3 ic_sup <- media_muestral + z_score * error_std  
4  
5 # Definamoslo como intervalo  
6 ci <- c(ic_inf, ic_sup)  
7 ci
```

```
[1] 0.6499696 0.7046504
```

# IC para Danceability - Comando t.test

```
1 ## Otra manera es usando el comando t.test  
2 result <- t.test(muestra$danceability, conf.level = 0.95)  
3 result
```

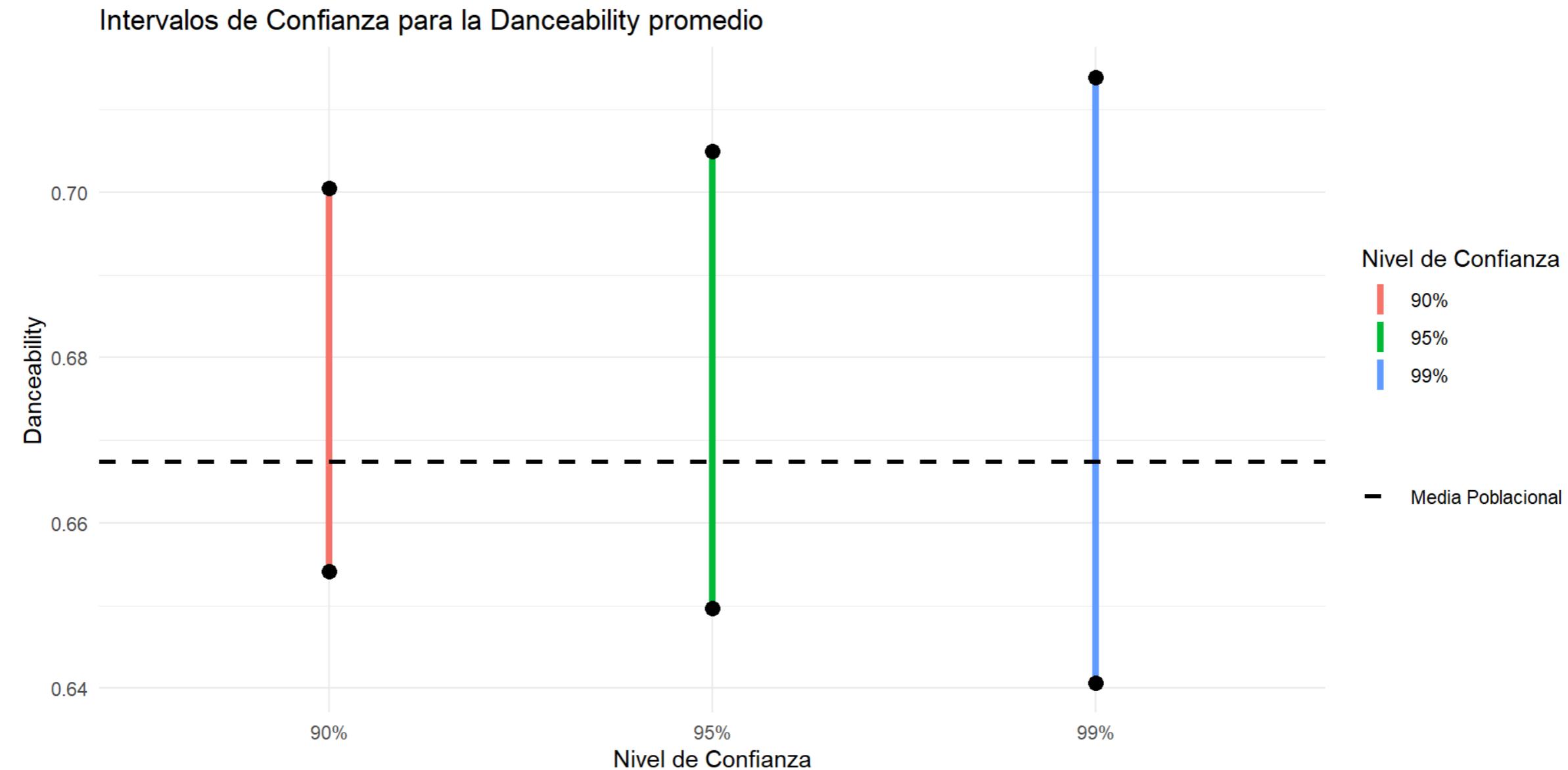
One Sample t-test

```
data: muestra$danceability  
t = 48.555, df = 99, p-value < 2.2e-16  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
 0.6496313 0.7049887  
sample estimates:  
mean of x  
 0.67731
```

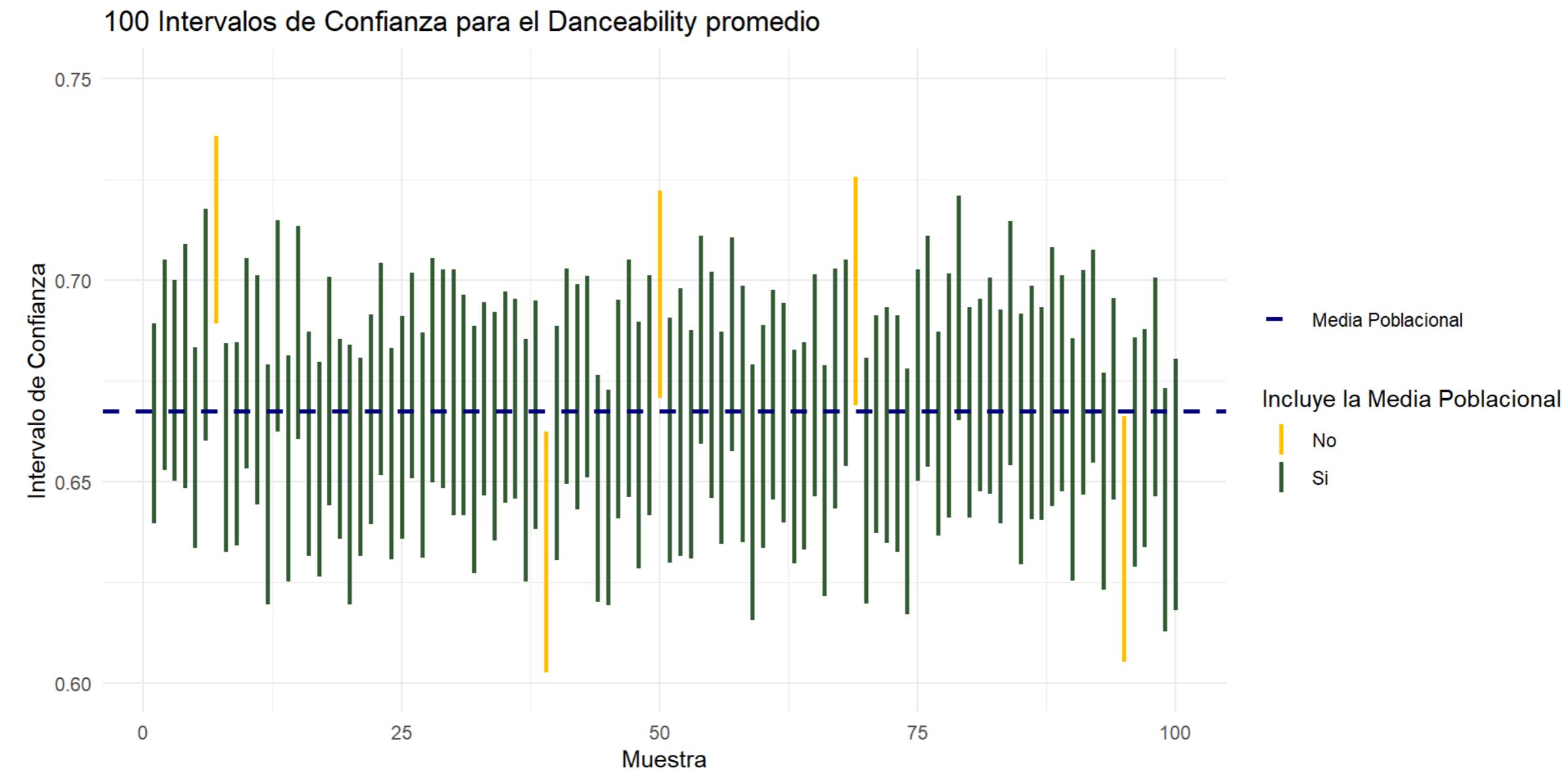
```
1 # Extraemos los intervalos  
2 ic_2 <- result$conf.int  
3 ic_2
```

```
[1] 0.6496313 0.7049887  
attr("conf.level")  
[1] 0.95
```

# IC con distintos niveles de confianza

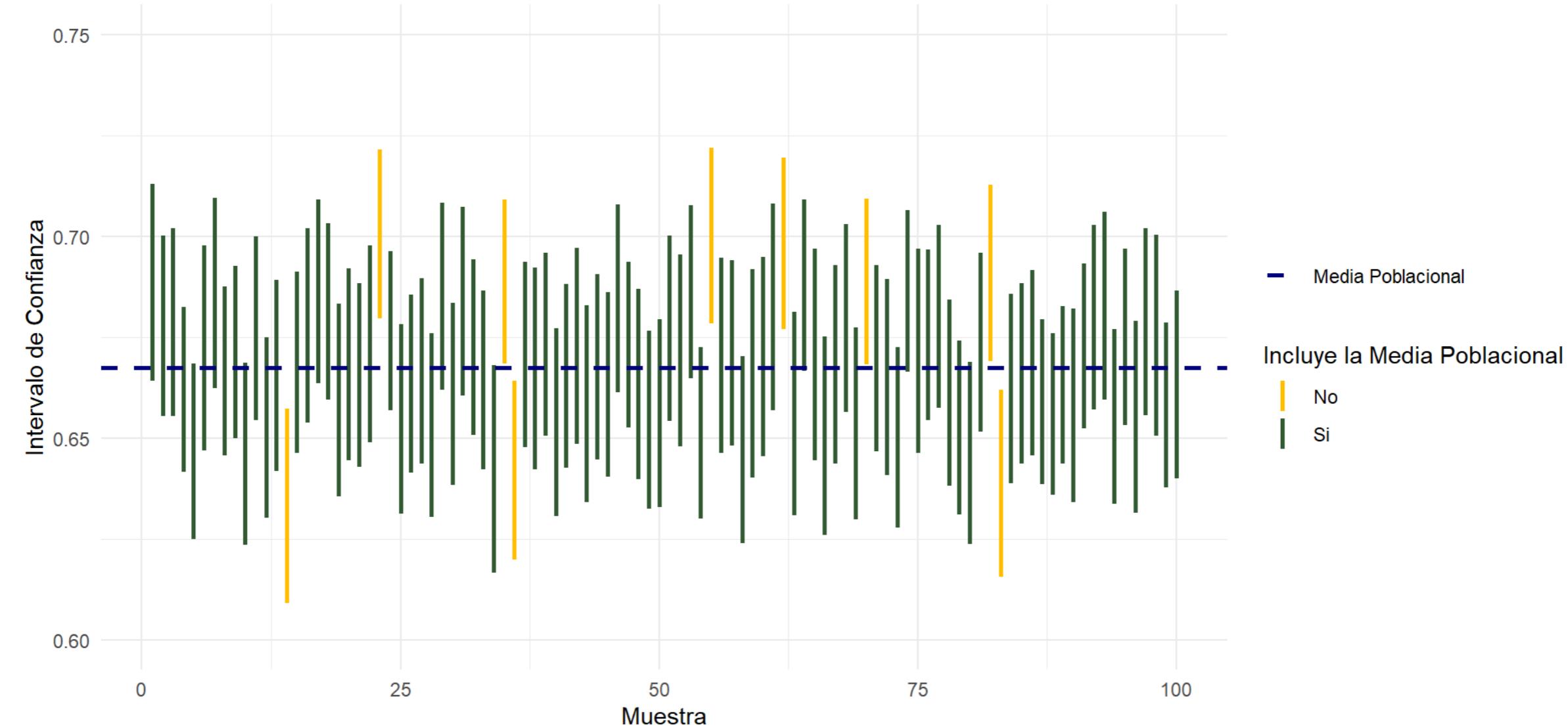


# 95% IC para 100 muestras



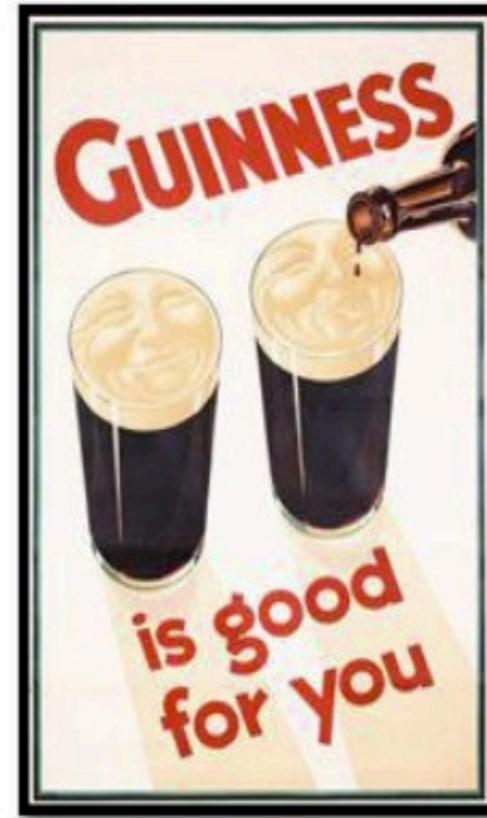
# 90% IC para 100 muestras

100 Intervalos de Confianza para el Danceability promedio



# Pruebas de Hipótesis

# Guinness - William Gosset



*"Guinness is the best beer available, it does not need advertising as its quality will sell it, and those who do not drink it are to be sympathized with rather than advertised to."* –W.S. Gosset (aka "Student")

# Guinness - William Gosset

“On the other hand, it is generally agreed that to leave the rejection of experiments entirely to the discretion of the experimenter is dangerous, as he is likely to be biased. Hence it has been proposed to adopt a criterion depending on the probability of such a wide error occurring in the given number of observations.”

# Ejemplo: Control de Calidad en Guinness

El equipo de control de calidad de la Cervecería Guinness quiere asegurarse de que su cerveza tenga un contenido de alcohol consistente, lo cual es crucial para mantener el sabor y la satisfacción del cliente. Para monitorear esto, miden regularmente el contenido de alcohol en sus lotes de cerveza.

El maestro cervecer cree que el contenido de alcohol óptimo para su stout estándar es del 4.5%. Para evaluar si el proceso de elaboración está manteniendo este objetivo, el equipo toma una muestra aleatoria de 106 lotes de cerveza de su última producción. Esta muestra arroja un contenido de alcohol promedio del 4.6% con una desviación estándar de 0.5%.

**¿Proporcionan estos datos evidencia convincente de que el contenido promedio de alcohol de todos los lotes de cerveza es mayor que el objetivo establecido por el maestro cervecer?**

# Ejemplo: Control de Calidad en Guinness

## Hipótesis

- Población: todos los lotes de cerveza
- El parámetro de interés  $\mu$  es el contenido promedio de alcohol de *todos* los lotes de cerveza
- Hay dos explicaciones de por qué la media muestral es mayor que el 4.5% recomendado por el cervecero:
  1. La media real de la población es diferente.
  2. La media real de la población es 4.5%, y la diferencia entre la media real de la población y la media de la muestra se debe simplemente a la variabilidad natural del muestreo.
- (El contenido de alcohol de los lotes es 4.5%)
- (El contenido de alcohol de los lotes es 4.5%)

# Maneras incorrectas de establecer $H_0$ y $H_A$

- y **SIEMPRE** se expresan en términos de parámetros de población, no de estadísticas de muestra.

- Ni:

- ni:

el contenido de alcohol promedio *en la muestra* es 4.5%

el contenido de alcohol promedio *en la muestra* es 4.6%

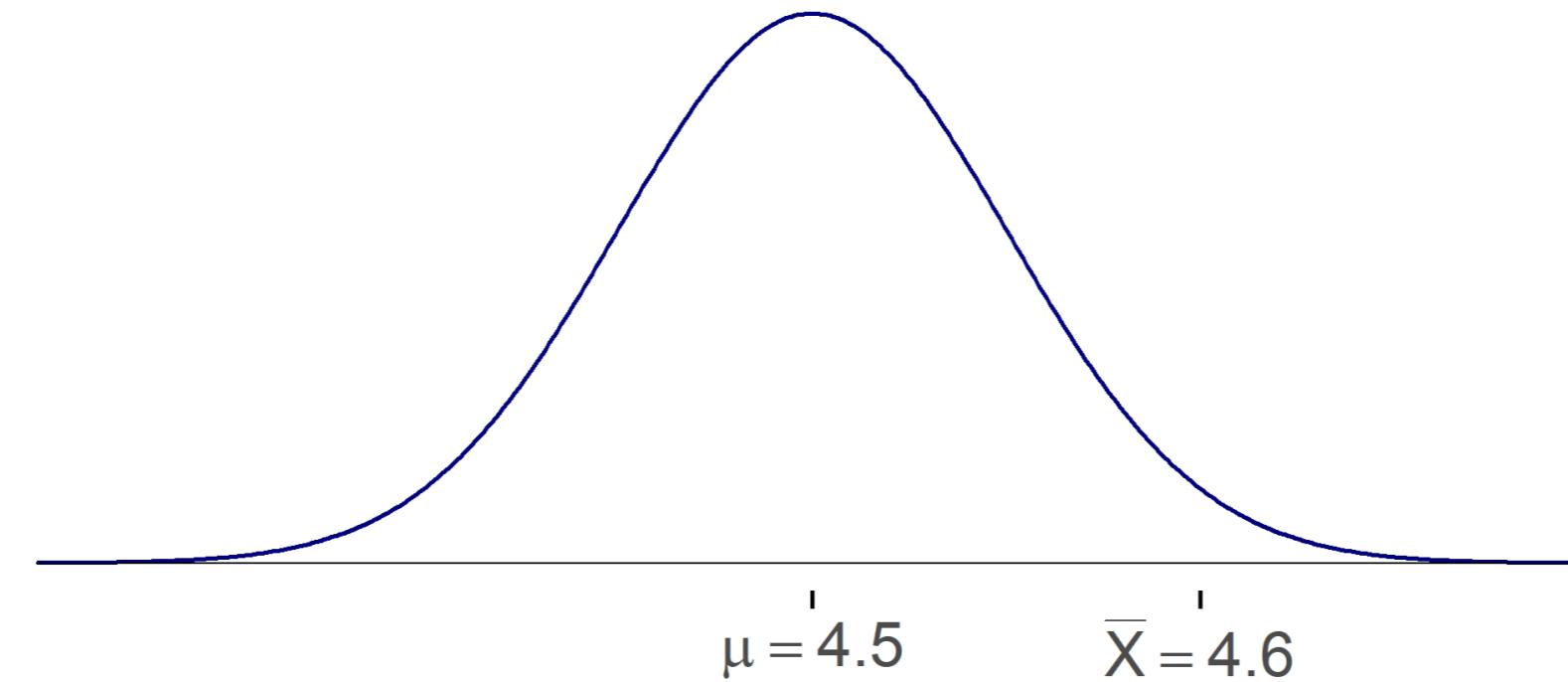
son correctas. Las hipótesis son:

También siempre **especifíquen claramente** qué es

e.g., es el contenido de alcohol promedio en los lotes de cerveza

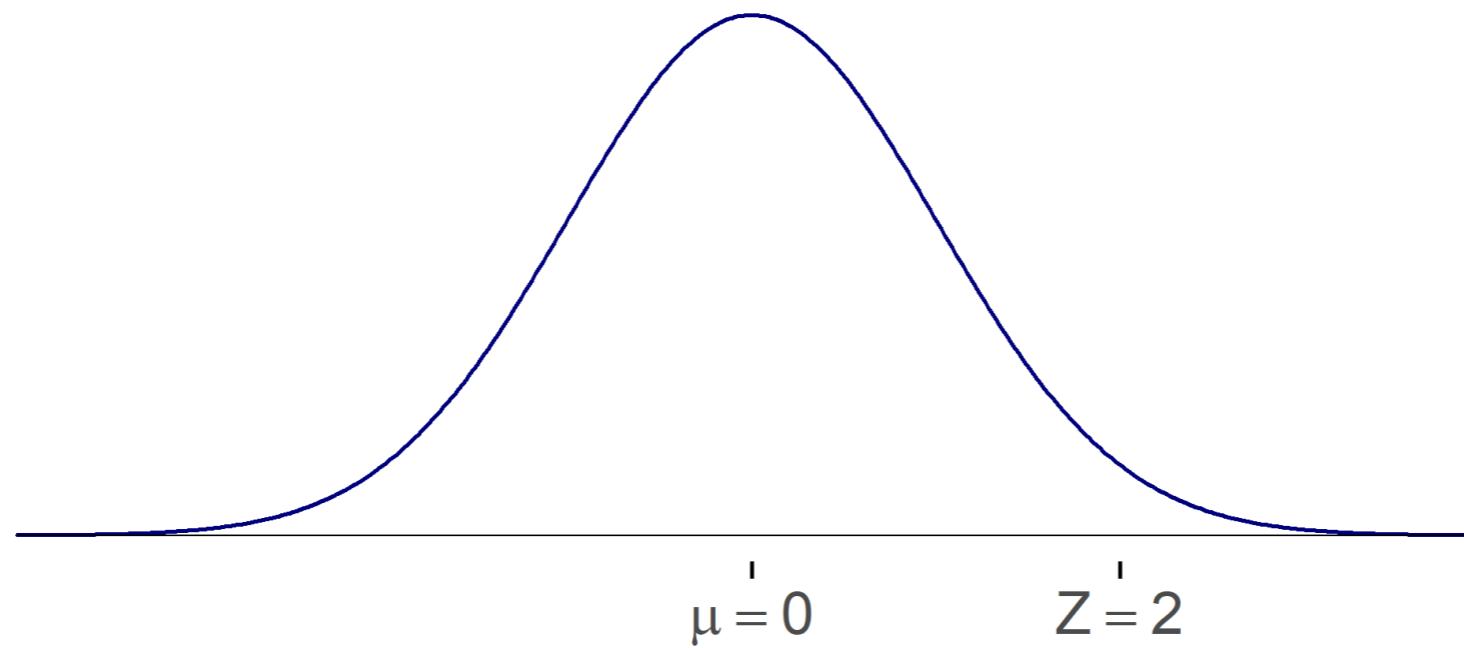
# Alcohol en la Guinness - Test

Por el TLC, bajo , la distribución muestral de la media muestral es:



Para medir qué tan *inusual* es la media muestral observada en relación con su distribución muestral, la estadística de prueba que usamos es el z-score.

# Alcohol en la Guinness - Test



# ¿Qué tan inusual es la media?



- Las medias muestrales que son probables de obtener si es cierta son las medias muestrales cercanas a la hipótesis nula.
- Las medias muestrales que son poco probables de obtener si es cierta son aquellas lejanas a la hipótesis nula.

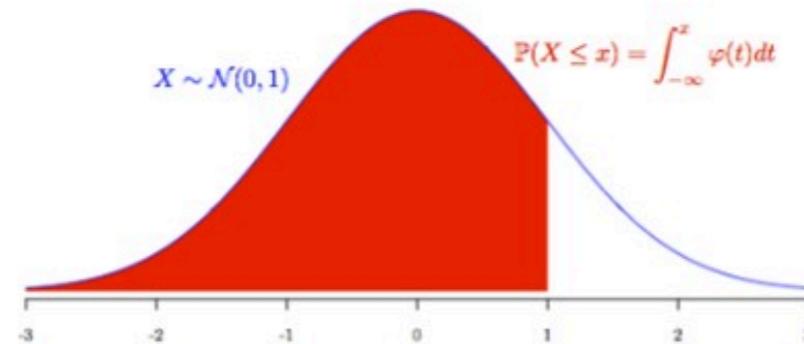
# ¿Qué significa “alta” y “baja” probabilidad?

- Esto se establece a partir de una probabilidad específica , la cual se conoce como *nivel de significancia* (se denota con ), para la prueba de hipótesis.
- El valor es una probabilidad pequeña que se utiliza para identificar muestras de poca probabilidad o inusuales.
- Por convención, los valores más comunes son  $\alpha = 0,05$  y  $\alpha = 0,01$ . Por ejemplo, si usamos  $\alpha = 0,05$ , separamos el 5% de las medias más improbables (valores extremos) del 95% de las medias muestrales más probables (valores centrales).

# Región y Z-Score críticos

- Los valores extremos que son poco probables, definidos por el nivel de significancia, constituyen lo que se conoce como **región crítica**.
- Estos valores son inconsistentes con la hipótesis nula. También se pueden interpretar como valores muestrales que proveen evidencia convincente de que el tratamiento/condición tienen algún efecto.
- Al igual que con los intervalos de confianza, para determinar la ubicación exacta de los límites se utilizan el  $\alpha$  y la tabla de la normal para encontrar el z-score crítico.

# ¿Cuál es nuestro $z$ o valor crítico?



- Si nuestro  $\alpha$ , el z-crítico será 1.64 de acuerdo a la tabla de la distribución normal

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

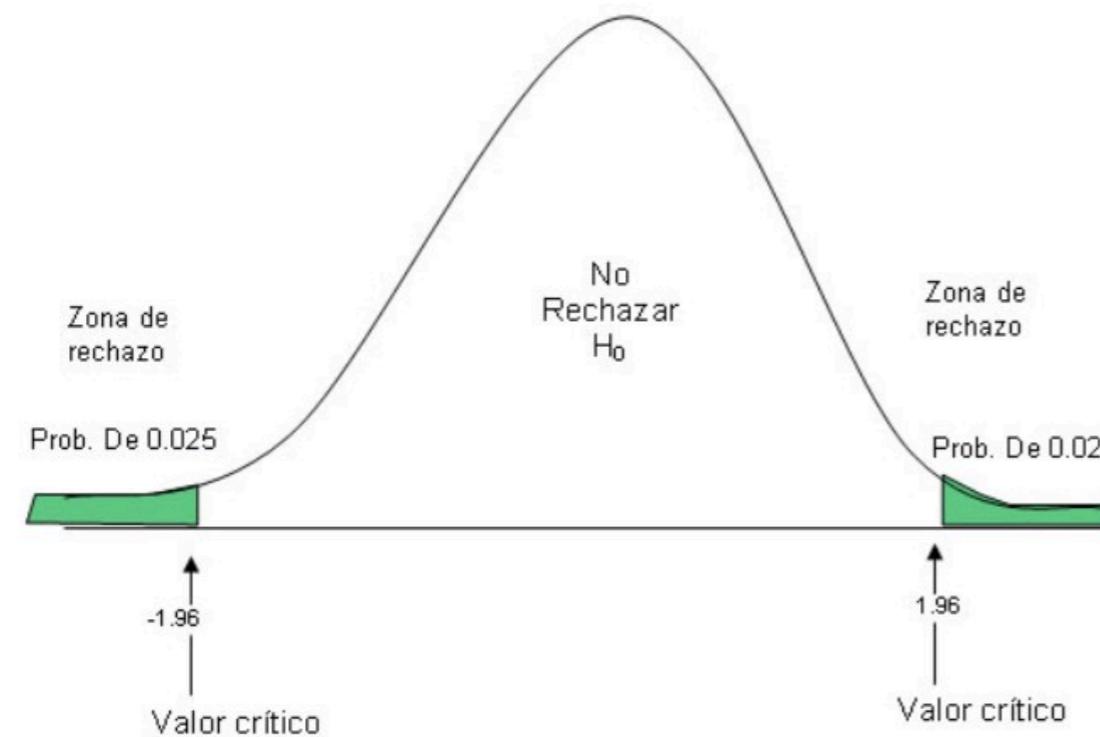
# Resultado de la prueba de hipótesis

- La media muestral se ubica en la región crítica.
- Un valor muestral en esta región es poco probable si es cierta.
- **Rechazamos**
- Los datos proveen evidencia convincente de que los lotes de cerveza tienen un contenido de alcohol promedio mayor a 4.5%.
- En otras palabras, la media muestral es **estadísticamente** diferente de 4.5%.

# Prueba de hipótesis a dos colas

Si el maestro cervecero quisiera saber si los datos proveen evidencia consistente que el contenido promedio de alcohol es **diferente** que el 4.5% recomendado, la hipótesis alternativa cambiaría:

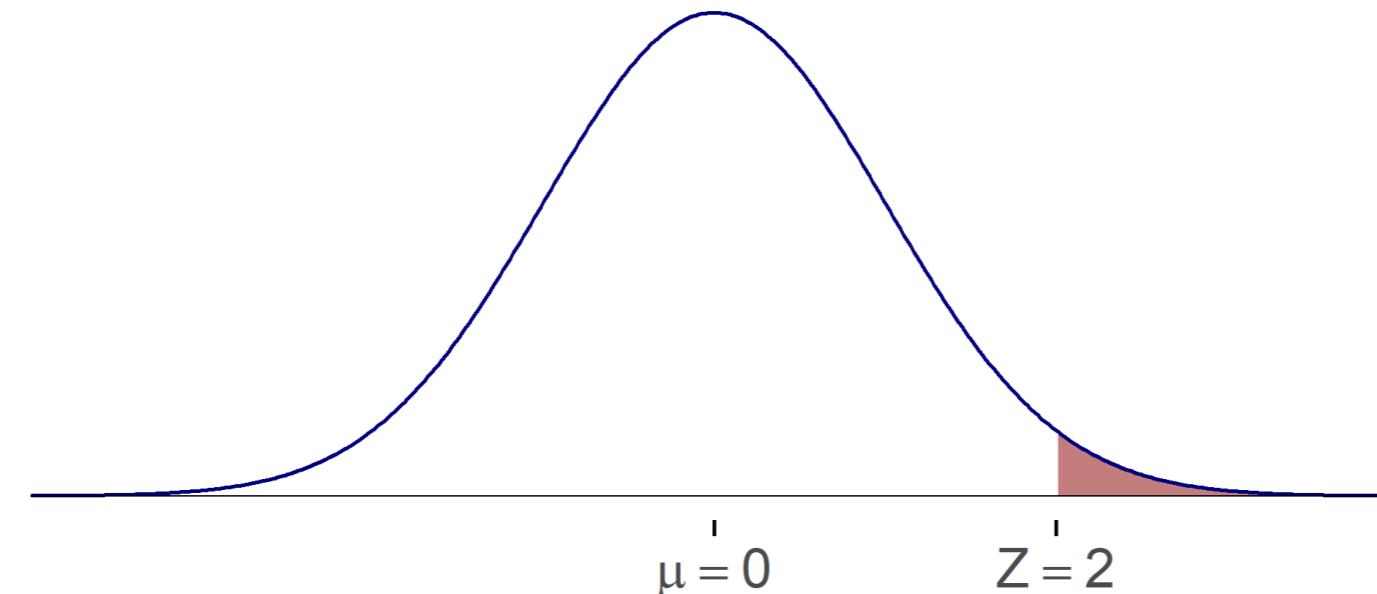
# Prueba de hipótesis a dos colas



- En este caso, una media muestral mucho menor a 4.5 también sería evidencia en favor de .
- Cuando no hay una dirección en la , se tienen dos regiones críticas.

# P-Value

- El z-score tiene una probabilidad asociada dada la forma de la distribución normal.
- Por ende, la decisión de una prueba de hipótesis puede basarse tanto en el z-score como en su probabilidad asociada (p-value).



# P-Value

- Para un nuestro , sabemos que . Si , se rechaza la hipótesis nula dado que .
- Esto sucede para un nivel de significancia del 5%, ¿qué sucede si este cambia a 1%?
- El p-value es la probabilidad de observar los resultados del estudio (), u otros más alejados de la hipótesis nula, si la hipótesis nula fuera cierta.

# Resumen - Prueba de hipótesis

1. Establezca las hipótesis

- 
- $< \text{ o } > \text{ o }$

2. Revise los supuestos y condiciones

- Independencia
- Normalidad:

3. Calcular el z-score y el p-value (dibujen la gráfica!!)

4. Tomen una decisión

- Si : Se rechaza
- Si : No se rechaza

# Prueba de hipótesis con dos muestras

- ¿Las empresas que cotizan en NYSE tienen mayores rendimientos promedio de acciones que las que cotizan en NASDAQ?
- ¿Las tasas de interés de las hipotecas ofrecidas por el banco A son más bajas que las ofrecidas por el banco B?
- ¿Los salarios promedio de los empleados en empresas tecnológicas son diferentes de los de las empresas manufactureras?
- El objetivo ahora es comparar las medias (o alguna cantidad) y de dos poblaciones

# Muestras independientes y relacionadas

- Las hipótesis en este caso son:
  - 
  -
- Las muestras pueden ser:
  1. Independientes: medición de unidades en distintos grupos.
  2. Relacionadas: medición de la misma unidad antes y después de alguna intervención/suceso.

# Un ejemplo

Una empresa de gestión de activos recientemente cambió al gerente del fondo que administra un bono de alto rendimiento. Se espera que este cambio tenga un impacto positivo en los retornos del bono. Para evaluar si el cambio de manager ha mejorado significativamente los retornos del bono, se recogieron los retornos mensuales del bono durante 10 meses antes y 10 meses después del cambio de manager.

Nuestra hipótesis nula es que no existe ningún cambio en el rendimiento promedio del bono con el cambio de gerente ()�.

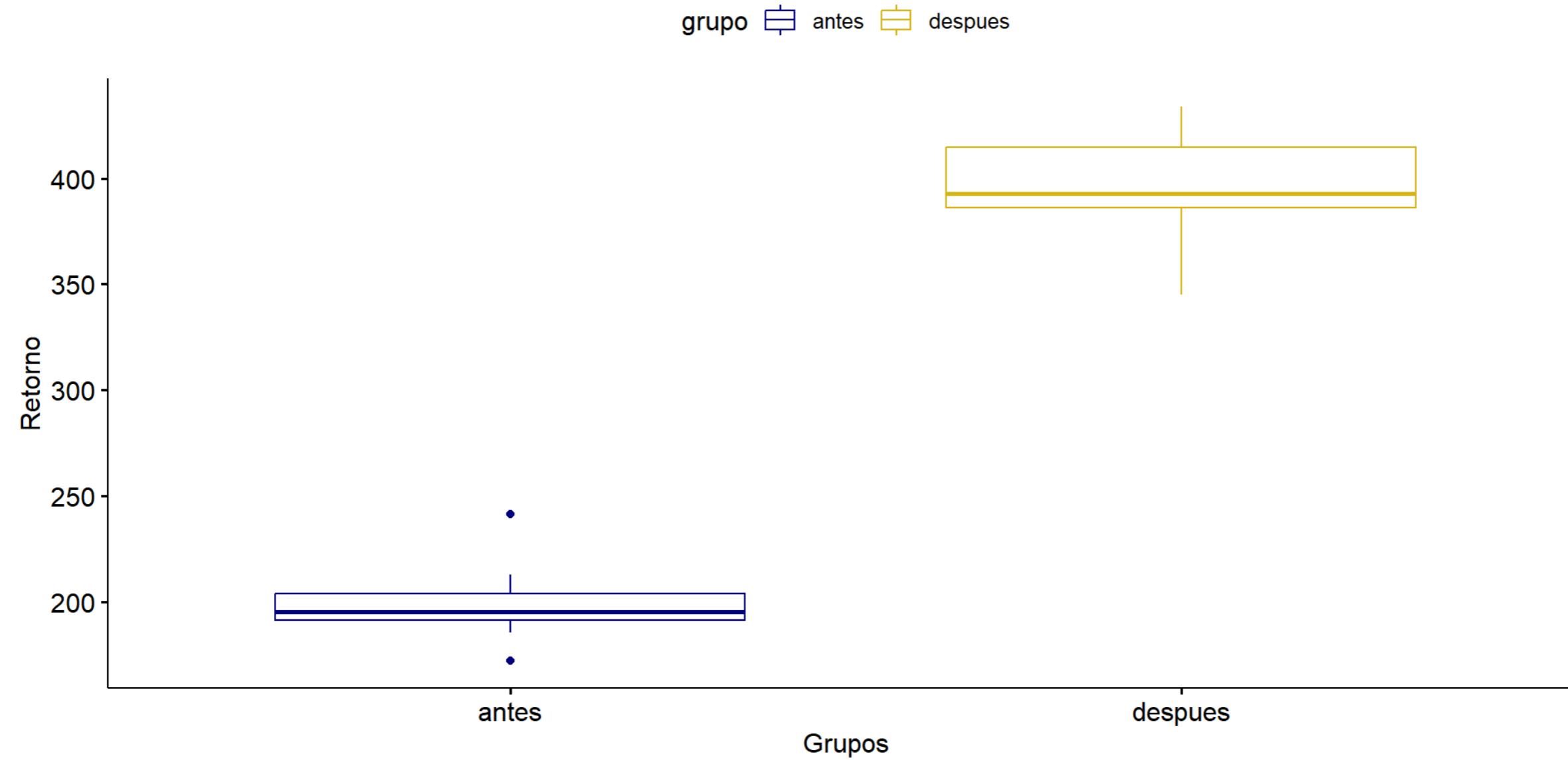
# Un ejemplo

```
1 # Retornos del bono antes del cambio de gerente
2 antes <-c(200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2, 185.5, 205.2, 193.7)
3 # Retornos del bono después del cambio de gerente
4 despues <-c(392.9, 393.2, 345.1, 393, 434, 427.9, 422, 383.9, 392.3, 352.2)
5 # Creamos el data frame
6 bono <- data.frame(
7   grupo = rep(c("antes", "despues"), each = 10),
8   retorno = c(antes, despues)
9 )
10 print(bono)
```

	grupo	retorno
1	antes	200.1
2	antes	190.9
3	antes	192.7
4	antes	213.0
5	antes	241.4
6	antes	196.9
7	antes	172.2
8	antes	185.5
9	antes	205.2
10	antes	193.7
11	despues	392.9

12	despues	393.2
13	despues	345.1
14	despues	202.0

# Ejemplo - Gráfica



# Ejemplo - t.test en R

```
1 # Calculemos el test  
2 test_resultado <- t.test(retorno ~ grupo, data = bono, paired = TRUE)  
3 test_resultado
```

Paired t-test

```
data: retorno by grupo  
t = -20.883, df = 9, p-value = 6.2e-09  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -215.5581 -173.4219  
sample estimates:  
mean of the differences  
 -194.49
```

```
1 # p-value  
2 test_resultado$p.value<0.05
```

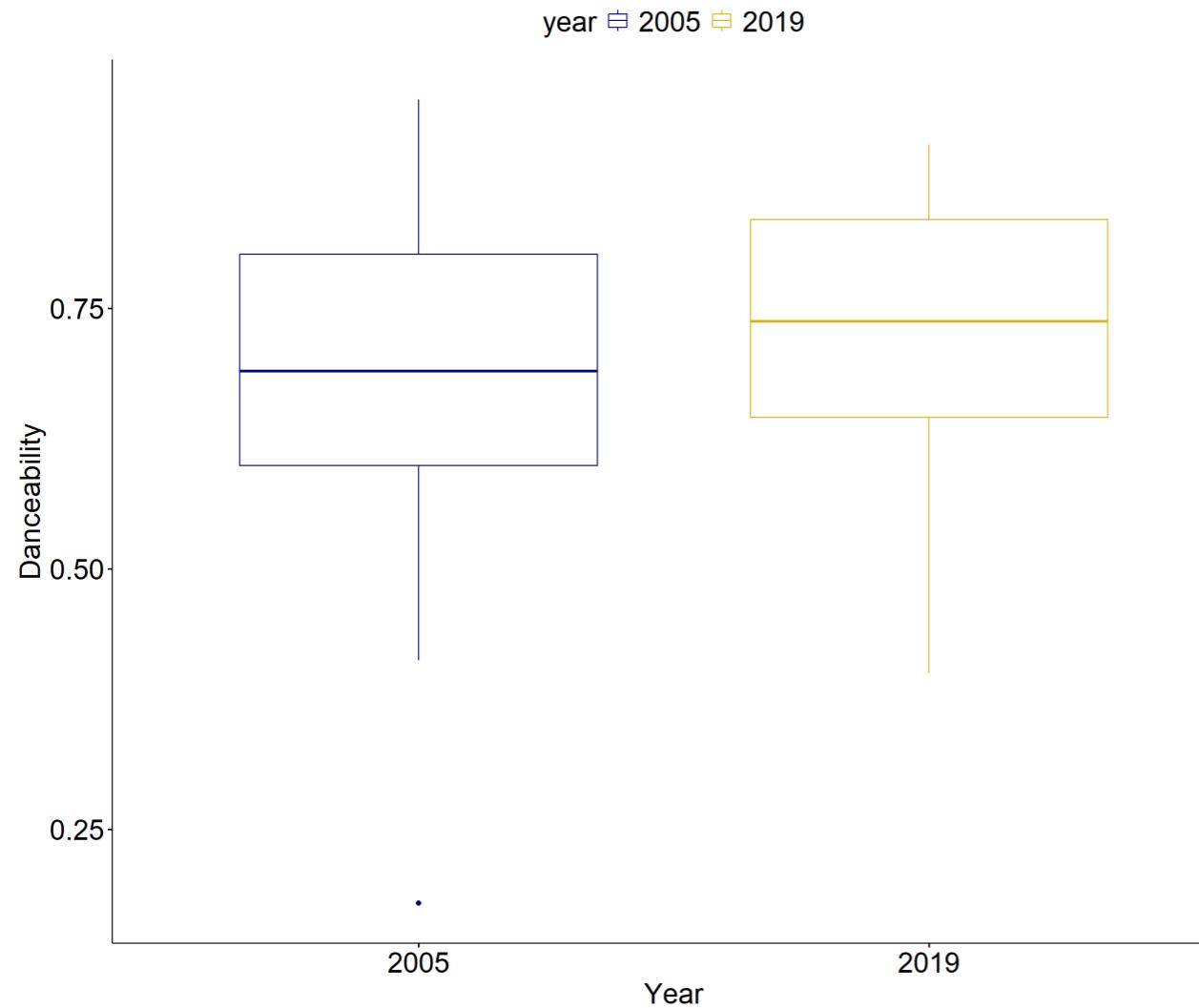
```
[1] TRUE
```

# Ejemplo Danceability

¿ Hubo un cambio en Danceability a través del tiempo?

Nuestra hipótesis nula es que no existe ningún cambio en el danceability promedio anual luego de 14 años () .

# Ejemplo Danceability - Gráfica



```
1 danceability <- data %>%
2   filter(year==c(2005,2019))%>%
3   select(year, danceability)
4
5 ggboxplot(danceability,
6   x = "year",
7   y = "danceability",
8   color = "year",
9   palette = c("navy", "#dbb818"),
10  order = c("2005", "2019"),
11  ylab = "Danceability",
12  xlab = "Year")
```

# Ejemplo Danceability - t.test en R

```
1 # Calculemos el test  
2 test_resultado2 <- t.test(danceability ~ year, data = danceability, paired = FALSE)  
3 test_resultado2
```

Welch Two Sample t-test

```
data: danceability by year  
t = -1.1142, df = 96.999, p-value = 0.2679  
alternative hypothesis: true difference in means between group 2005 and group 2019 is not equal to 0  
95 percent confidence interval:  
-0.08892705 0.02498037  
sample estimates:  
mean in group 2005 mean in group 2019  
0.6863962 0.7183696
```

```
1 # p-value  
2 test_resultado2$p.value<0.05
```

[1] FALSE

El resultado de la prueba es que el cambio en danceability no es **estadísticamente significativo**

Algunas consideraciones  
sobre

**Pruebas de Hipótesis**

# ¿De dónde vienen los valores poblaciones?

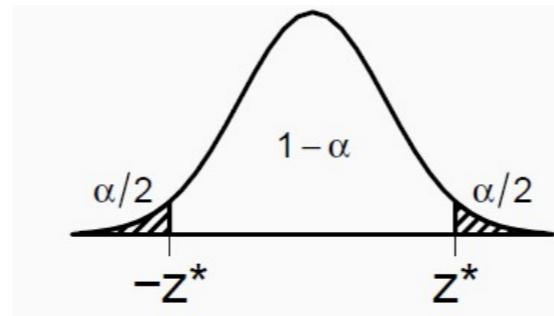
- Los valores poblaciones u objetivos de las pruebas de hipótesis pueden venir de:
- Un parámetro poblacional conocido a partir de un grupo de comparación o de un censo poblacional (Representatividad de la muestra).
- Parámetros conocidos a partir de un período anterior.
- Una idea objetivo (La producción deseada en una empresa).

# Intervalos de confianza y prueba de hipótesis a dos colas

En una prueba de hipótesis a dos colas:

Los siguientes son equivalentes:

- p-value (Por lo tanto no se rechaza al nivel de significancia )
- , donde es el valor tal que:



- está en el % intervalo de confianza para



# Ejemplo - IC

Supongamos que en un estudio:

- 90% IC para es (4.81, 11.39)
- 95% IC para es (4.18, 12.02)
- 99% IC para es (2.95, 13.25)

Entonces

- es rechazada al 5% pero no al 1% (*el p-value a dos colas está entre 0.01 y 0.05*) porque 4 está en el 99% IC pero no en el 95% IC
- es rechazada al 10% pero no al 5% porque 4.5 está en el 95% IC pero no en el 90% IC

# Errores Tipo I y II

	Decisión: Rechaza	Decisión: No Rechazar
<b>es verdadera</b>	Error Tipo I (Falso Positivo)	
<b>es falsa</b>		Error Tipo II (Falso Negativo)
• (Casi) Nunca sabremos si o son verdaderas, pero se necesita considerar todas las posibilidades		

# Errores Tipo I y II

Type I Error



Type II Error



# Consecuencias de los errores Tipo I y II

Los errores de tipo I y tipo II son diferentes tipos de equivocaciones y tienen consecuencias distintas:

- Generalmente, es el status quo, algo que generalmente creemos que es cierto.
- Si no se rechaza , usualmente significa que el status quo está bien. No se necesita tomar ninguna acción.
- Rechazar significa que algo en lo que solíamos creer ha sido refutado. Podría ser un avance científico (por ejemplo, la identificación de una nueva estrategia de inversión).
- Un error de tipo I introduce una conclusión falsa en la comunidad científica y puede llevar a un tremendo desperdicio de recursos antes

de que investigaciones posteriores invaliden el hallazgo original.

# Consecuencias de los errores Tipo I y II

Un error de tipo II (no reconocer un avance científico o una nueva estrategia financiera) representa una oportunidad perdida para el progreso científico o para la empresa.

- Los errores de tipo II también pueden ser costosos, pero generalmente pasan desapercibidos.
- Por eso, es más importante controlar la tasa de error de tipo I que la tasa de error de tipo II.

# Nivel de significancia = $p(\text{Error Tipo I})$

Cuando es verdadera, solo hay un 5% de probabilidad de obtener un valor < 5%.

- Esto significa que, en aquellos casos donde es realmente verdadera, no la rechazaremos incorrectamente en más del 5% de esas veces a largo plazo.
- En otras palabras, al usar un nivel de significancia del 5%, hay aproximadamente un 5% de probabilidad de cometer un error de tipo I si  $H_0$  es verdadera.
- Por eso preferimos valores pequeños de — aumentar incrementa la tasa de error de tipo I.

- Sin embargo, el nivel de significancia no controla la tasa de error de tipo II.

# Reportando el p-value

No se limiten a reportar la conclusión de si se rechaza . Muestren el p-value.

- Un p-value de 0.04 y un p-value de 0.000001 no son lo mismo. Aunque se rechace en ambos casos, la fuerza de la evidencia es muy diferente.
- Reportar simplemente si es rechazada sin el p-value es como reportar la temperatura como “fría” o “caliente”.
- Es mucho mejor reportar el p-value y permitir que la gente elija su propio nivel de significancia. Es similar a decirle a alguien la temperatura y dejar que decida cómo interpretarla.

# **Algunas ideas incorrectas sobre la prueba de hipótesis**

# El método científico: prueba y refutación

- Hay una verdad sutil pero fundamental en el método científico, y es que nunca se puede realmente probar una hipótesis con él, solo **refutar** la hipótesis.
- En palabras de Albert Einstein:

“No amount of experimentation can ever prove me right; a single experiment can prove me wrong.”

- Por lo tanto, nunca decimos que la hipótesis nula es verdadera.
- Cuando la evidencia no es lo suficientemente fuerte como para rechazar la nula, no decimos “aceptamos la hipótesis nula”, sino que decimos “no podemos rechazar la hipótesis nula”

# Fallar al rechazar $H_0$ no prueba que $H_0$ sea cierta

Un error común es concluir a partir de un p-value alto que la es probablemente verdadera.

- Un p-value bajo es evidencia de que no es verdadera.
- Si nuestro p-value es alto, ¿podemos concluir que es verdadera?
  - No, podríamos cometer un error de tipo II al no rechazar .
  - Además, la tasa de error de tipo II suele ser considerablemente más alta en comparación con la tasa de error de tipo I, la cual se mantiene controlada a un nivel bajo.
  - Es bastante común que no sea verdadera, pero los datos no la rechacen.
- Cuando no rechazamos , a menudo significa que los datos no son capaces de distinguir entre y (porque los datos son demasiado ruidosos, etc.).

# Ejemplo de la vida real

- Women's Health Initiative encontró que las dietas bajas en grasa reducen el riesgo de cáncer de seno con un p-value de **0.07**.
- El titular del New York Times: “[Estudio encuentra que las dietas bajas en grasa no detendrán el cáncer](#)”.
- El editorial principal afirmó que el estudio presentaba “evidencia sólida de que la guerra contra las grasas fue en vano” y añadió “este es el fin para la creencia de que reducir el porcentaje de grasa total en la dieta es importante para la salud”.
- No encontrar evidencia del efecto de las dietas bajas en grasa no significa que las dietas bajas en grasa no tengan ningún efecto.

# No tomen la significancia al 0.05 demasiado en serio

- Un p-value de 0.049 y un p-value de 0.051 ofrecen casi la misma evidencia contra .
- Por ejemplo, un estudio famoso de 2009 sobre una vacuna que podría proteger contra el VIH reportó un p-value a dos colas de 0.08, mientras que el p-value a una cola fue 0.04.
- Se desató mucho debate y controversia, en parte porque las dos formas de analizar los datos producen p-values a ambos lados de 0.05.
- Gran parte de este debate y controversia es bastante inútil; ambos p-values te dicen esencialmente lo mismo: que la vacuna tiene potencial,

# Las pruebas de hipótesis no pueden decirnos...

Las pruebas de hipótesis no pueden decirnos:

- si el diseño de un estudio está defectuoso
- si los datos se han recolectado adecuadamente

Por lo tanto, no podemos concluir a partir de un p-value pequeño si una variable tiene un efecto causal sobre otra variable o si la conclusión se puede generalizar a una población más grande.

# Significancia estadística no significa importancia práctica

Otro error es leer demasiado en el término "estadísticamente significativo".

- Decir que los resultados son estadísticamente significativos informa al lector que los hallazgos son poco probables de ser resultado del azar.
- Sin embargo, no dice nada sobre la importancia práctica del hallazgo.
- E.g., rechazar solo nos dice que , pero no qué tan grande o importante es . Puede ser que la diferencia no sea relevante por ser muy pequeña a pesar de ser significativa.
- Remedio: Reporten un intervalo de confianza del parámetro para que la gente pueda decidir si la diferencia es lo suficientemente grande como para ser relevante.

# Ejemplo

Un IC al 95% del contenido promedio de alcohol en este lote de cervezas será:

del cual uno puede decidir si la diferencia con respecto al 4.5 es suficientemente grande para ser relevante.

# En resumen...

- Rechazar no significa que estamos 100% seguros que es falsa.  
Podemos cometer errores Tipo I
- El p-value no es la probabilidad de que  $H_0$  sea verdadera.
- No rechazar no prueba que sea verdadera.
- No tomen el nivel de significancia de 0.05 demasiado en serio.
- Las pruebas de hipótesis no pueden decírnos si los datos se recolectaron adecuadamente o si el diseño de un estudio es malo.
- La significancia estadística no se traducen en importancia práctica.

