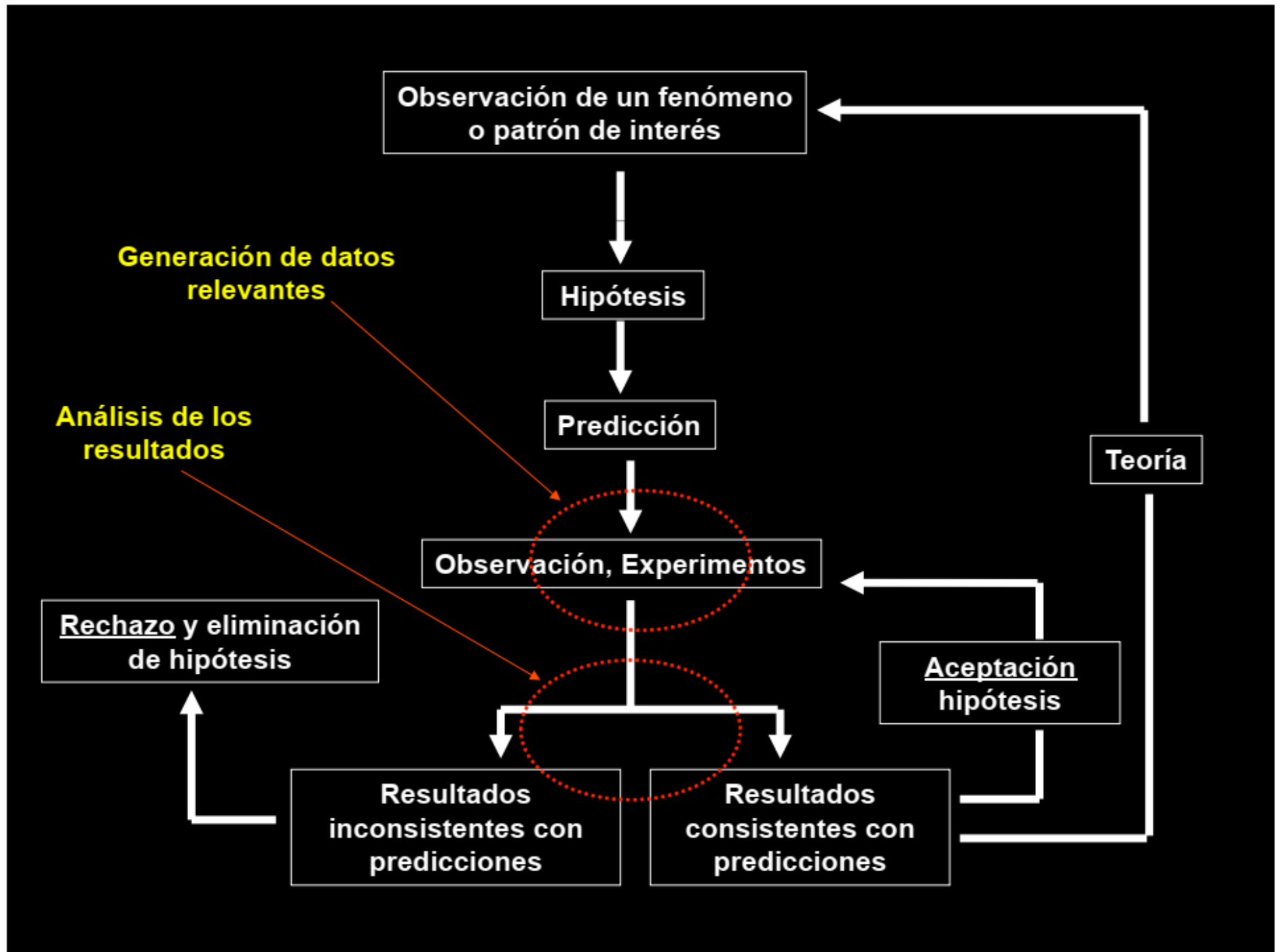


Sesión 2: Estadística para Bioinformática

Conceptos básicos de estadística



Definiciones importantes

Población:

conjunto de individuos o elementos que cumplen ciertas propiedades y entre los cuales se desea estudiar un determinado fenómeno.



Muestra:

Subconjunto representativo de una población.



Unidad estadística

un miembro del conjunto de entidades que están siendo estudiadas

muestreo



Parámetros poblacionales

μ = Media poblacional
 σ^2 = Varianza poblacional
 σ = Desviación estándar

inferencia



Estadígrafos

\bar{X} = media muestral
 S^2 = Varianza muestral
 S = Desviación estándar

Estadígrafos

- **Estadígrafos de Tendencia Central**

Propuestos para describir la típica concentración de puntos u observaciones en torno a algún valor dentro del conjunto de datos

Media, Mediana, Moda

- **Estadígrafos de Dispersión**

Describen la variabilidad de los datos alrededor de la tendencia central

Rango, Varianza, Desviación estándar

Estadígrafos de Tendencia Central

Media:
$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Mediana: valor central tal que una mitad de los registros de la muestra quedan ubicados a su izquierda mientras que la otra mitad estarán ubicados a su derecha

Moda: valor más frecuente de una muestra

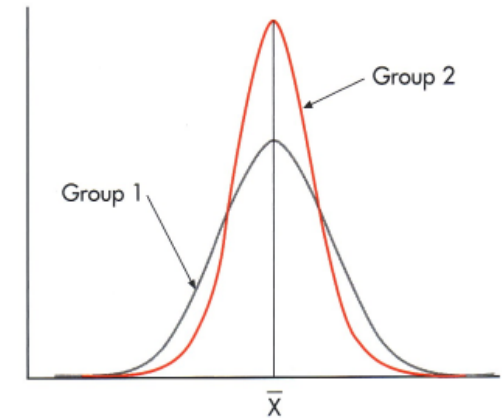
Estadígrafos de Dispersión

Rango

$$X_{(\max)} - X_{(\min)}$$

Varianza

$$S^2 = \frac{\sum_{i=1}^n (X_j - \bar{X})^2}{n - 1}$$



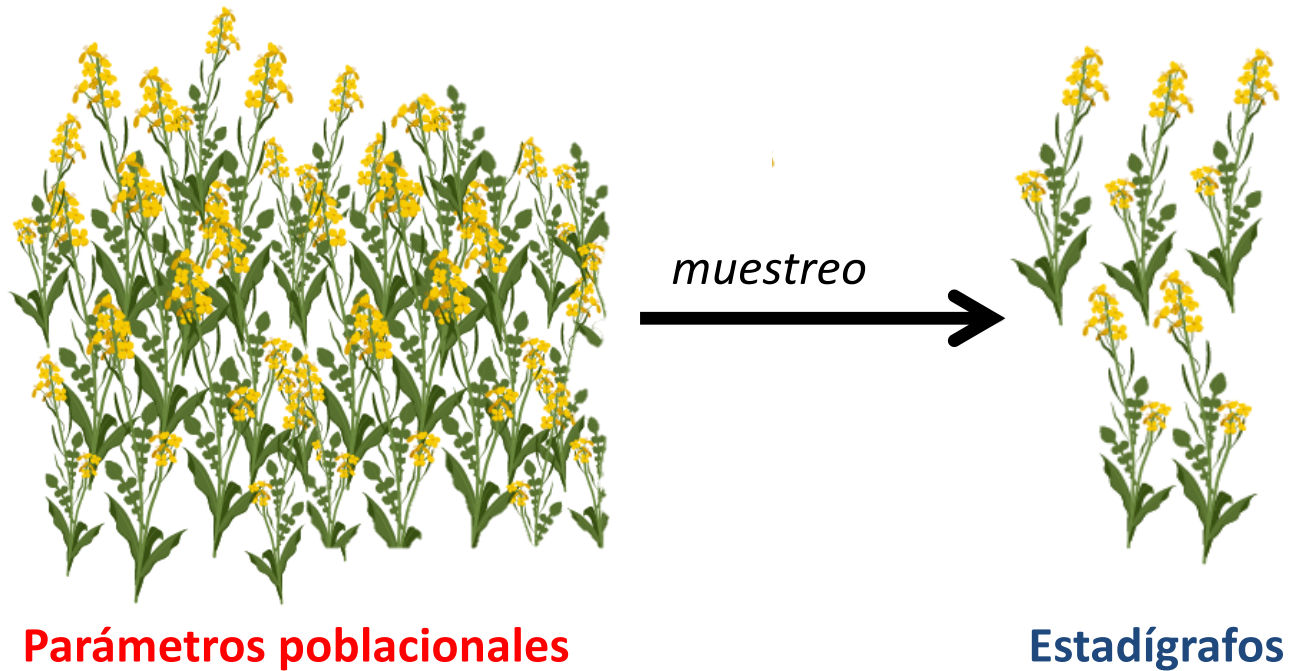
Desviación estándar

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

Coefficiente de variación

$$C.V. = \frac{S}{\bar{X}} * 100$$

Función principal de la Estadística



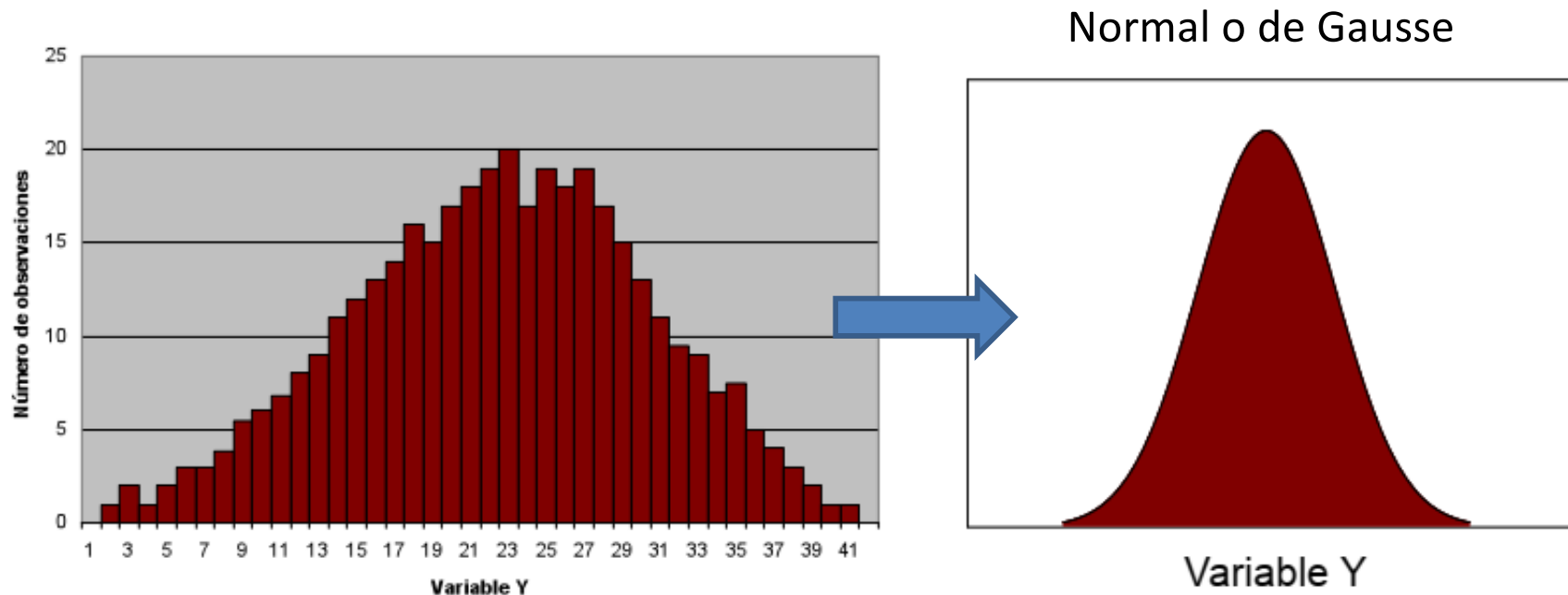
Parámetros poblacionales

Estadígrafos

Inferencia

Distribución normal

En muchas variables y conjuntos de datos en biología, la distribución de éstos se parece a una distribución con forma de “campana”

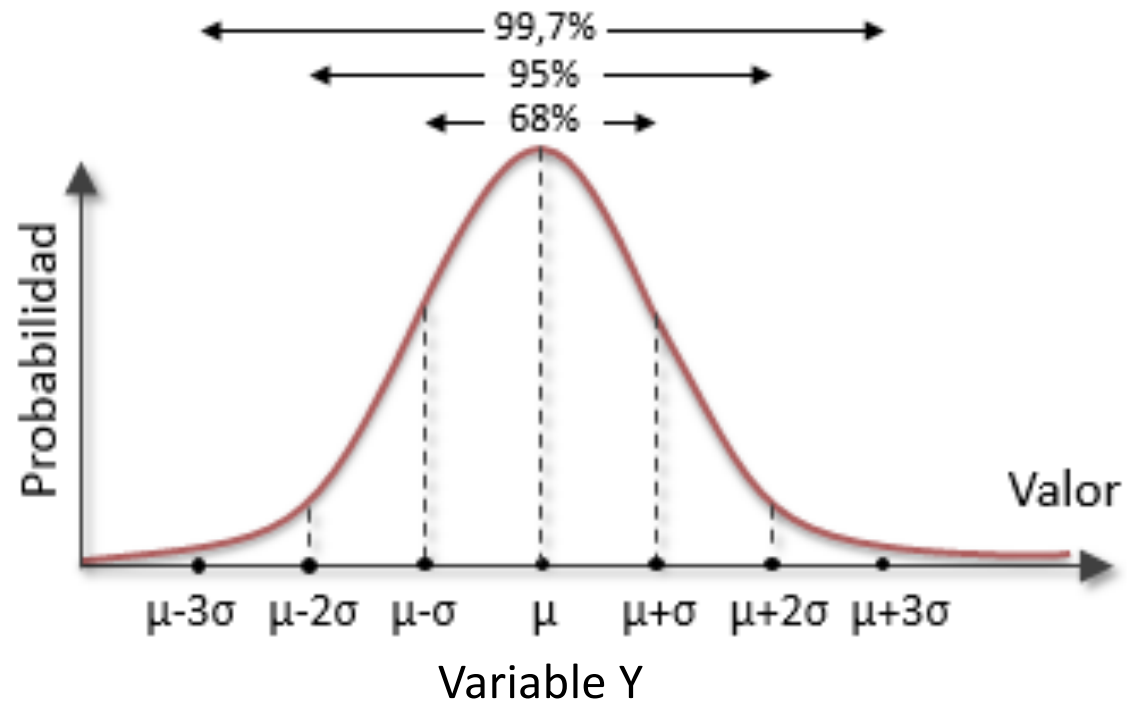


***Una gran familia de pruebas estadísticas (paramétricas) tienen como supuesto que los datos tienen una distribución con forma de campana

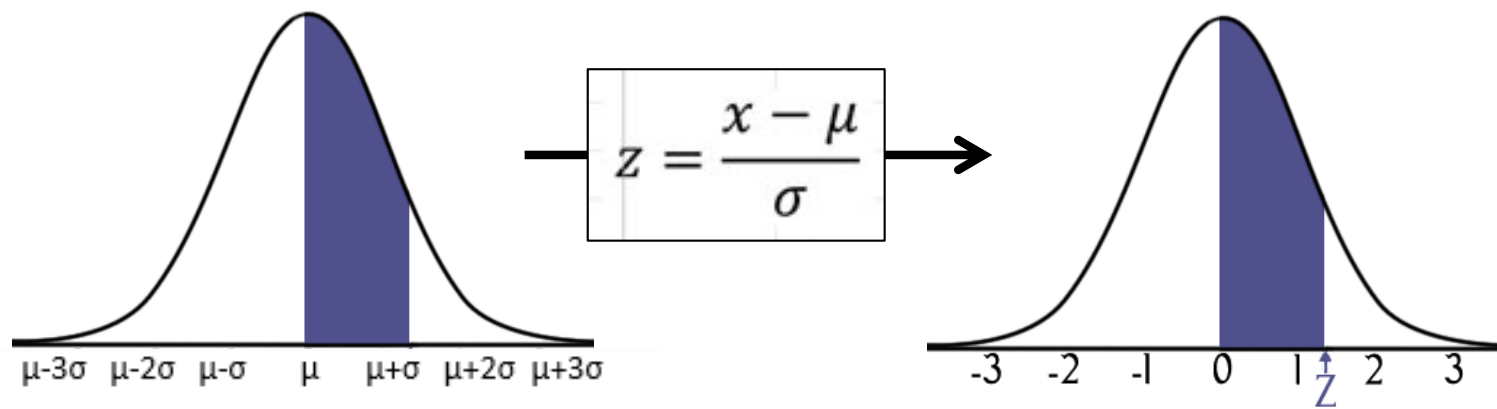
Propiedades de la distribución Normal

- Toda distribución Normal se define por dos parámetros: μ y σ
- Simétrica por ambos lados
- Media = mediana = moda
- El área bajo la curva entre dos valores de Y corresponde a la *probabilidad* de observar un valor de Y dentro de dicho rango

μ y σ pueden tener infinitos valores



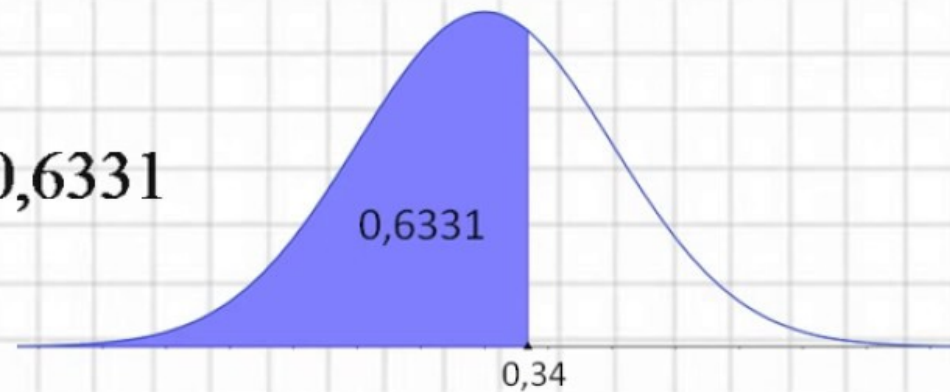
Curva normal estandarizada (z)



Curva normal estandarizada (z)

¿Qué área queda por debajo del valor 0,34?

$$P(Z \leq 0,34) = 0,6331$$



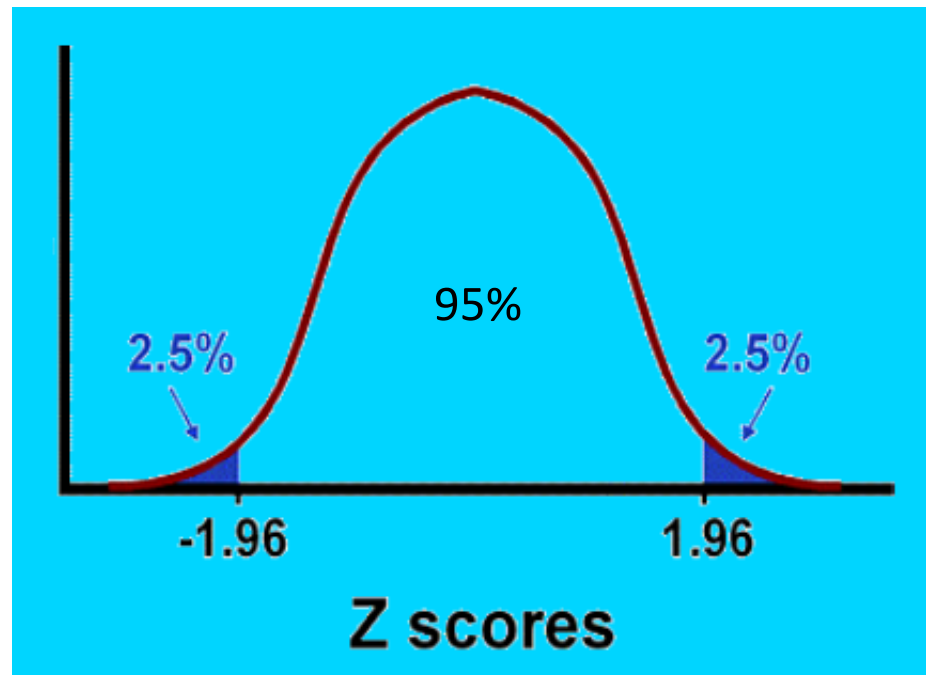
z	,00	,01	,02	,03	,04	,05	,06
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7421	0,7454



Para inferencias

Intervalo de confianza para la media

Estimamos el rango de valores de z donde la probabilidad de encontrar al verdadero valor de la media de la población (estandarizado) es 0,95



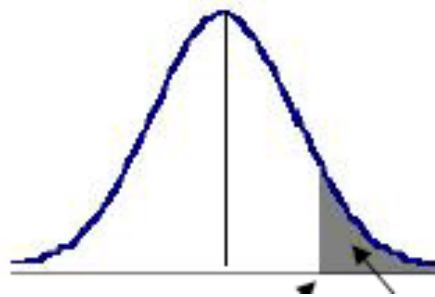
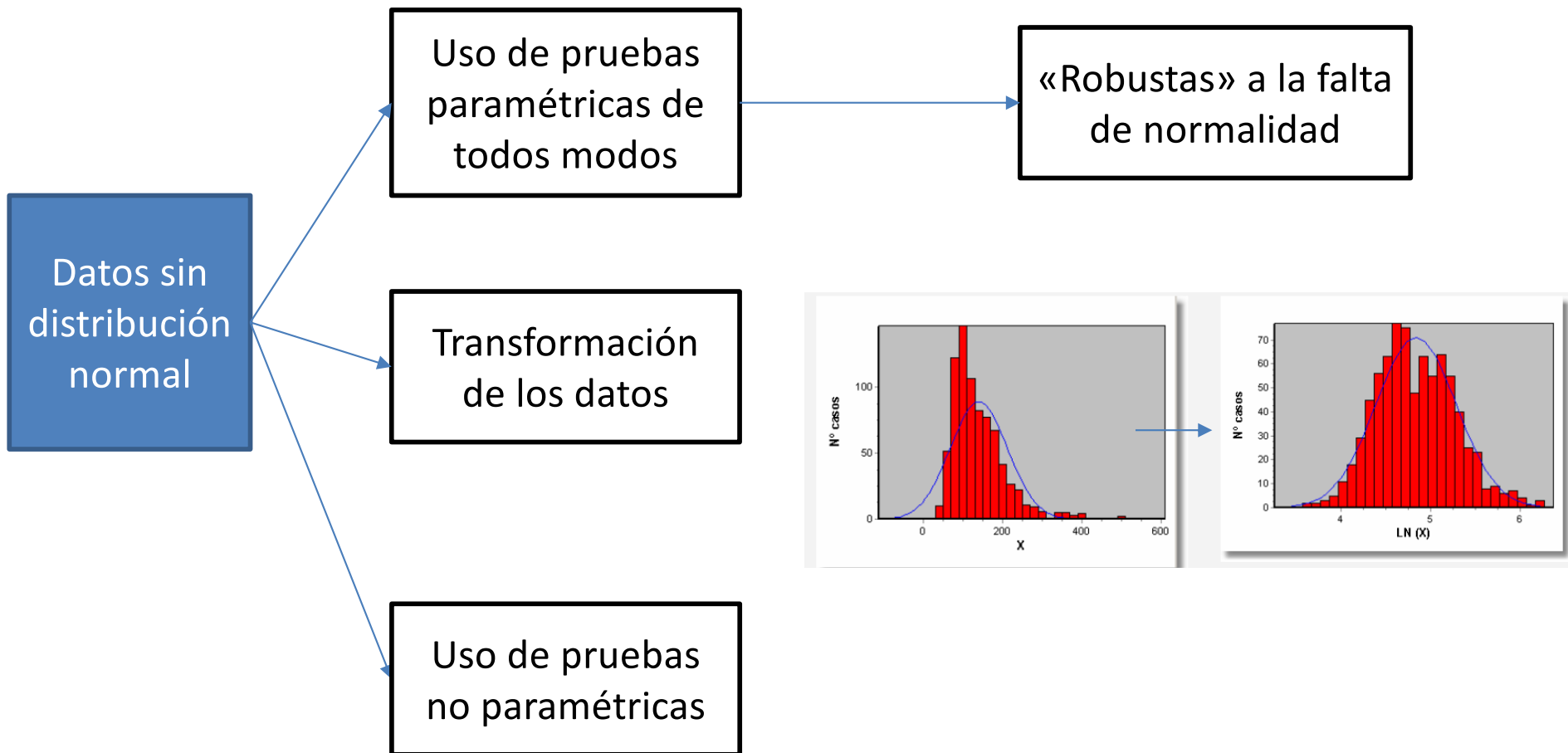


TABLE B.2 Proportions of the Normal Curve (One-Tailed)

This table gives the proportion of the normal curve that lies beyond (i.e., is more extreme than) a given normal deviate; e.g., $Z = (X_i - \mu)/\sigma$ or $Z = (\bar{X} - \mu)/\sigma \bar{X}$. For example, the proportion of a normal distribution for which $Z \geq 1.51$ is 0.0655.

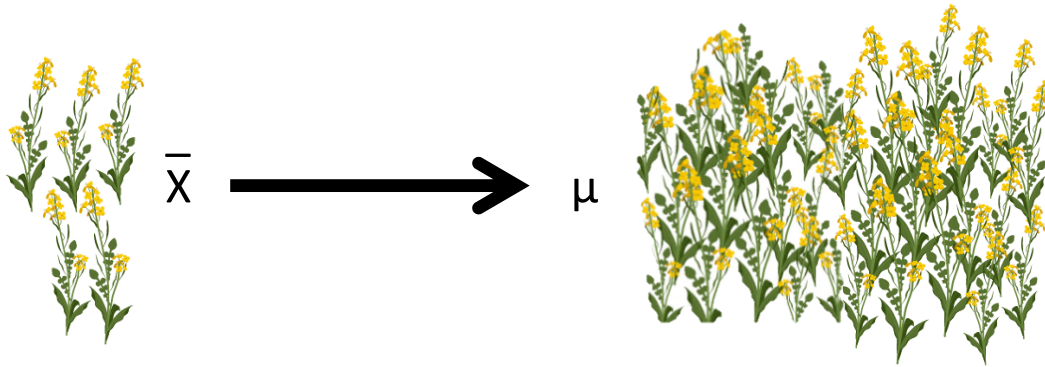
Z	0	1	2	3	4	5	6	7	8	9	Z
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641	0.0
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247	0.1
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859	0.2
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483	0.3
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121	0.4
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776	0.5
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451	0.6
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148	0.7
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867	0.8
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611	0.9
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379	1.0
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170	1.1
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985	1.2
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823	1.3
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681	1.4
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559	1.5
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455	1.6
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367	1.7
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294	1.8
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233	1.9
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183	2.0
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143	2.1
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110	2.2
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084	2.3
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064	2.4

¿ Y si el supuesto de normalidad no se cumple?



Usos de la estadística

- Inferencias



- Prueba de hipótesis



Grupo experimental 1



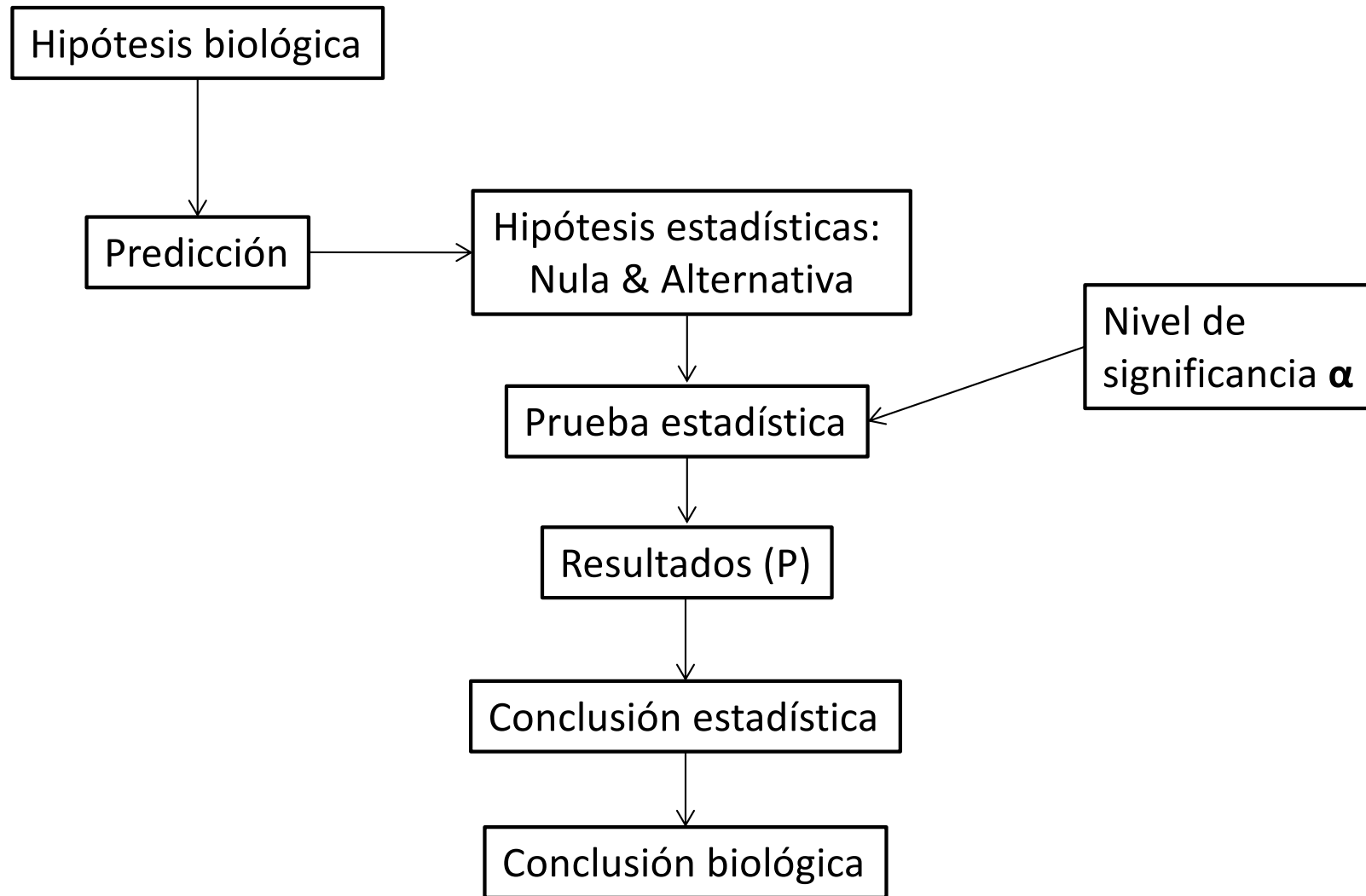
Grupo experimental 2

Prueba de hipótesis

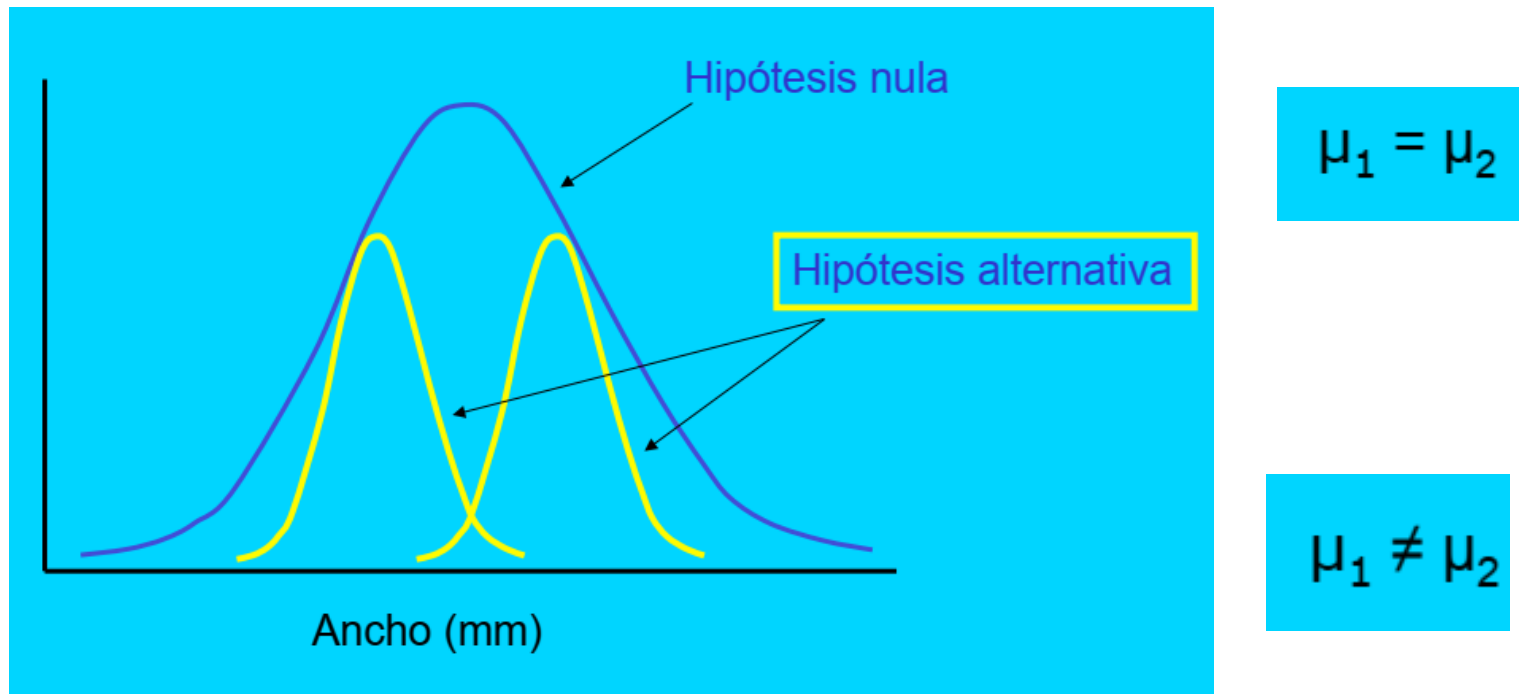
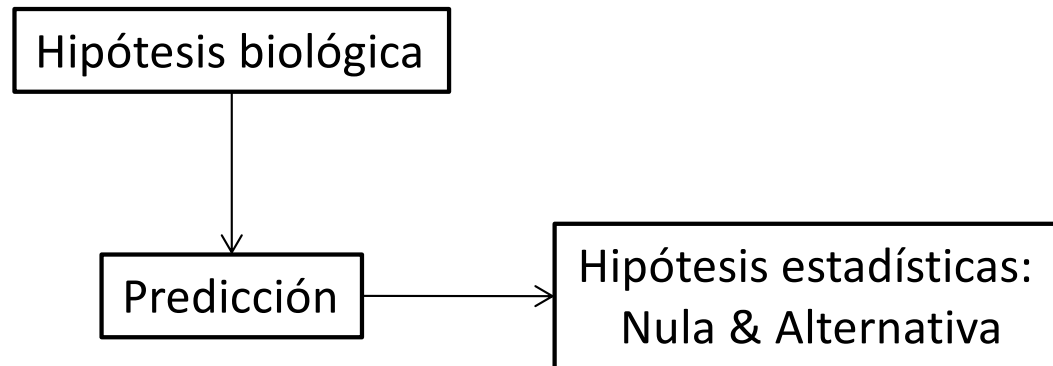
Hipótesis= Una afirmación transitoria que debe ser sometida a prueba

El método de las pruebas de hipótesis consiste fundamentalmente en **establecer la probabilidad de que la diferencia observada entre dos grupos sea consecuencia del azar.**

Procedimiento para la prueba de hipótesis

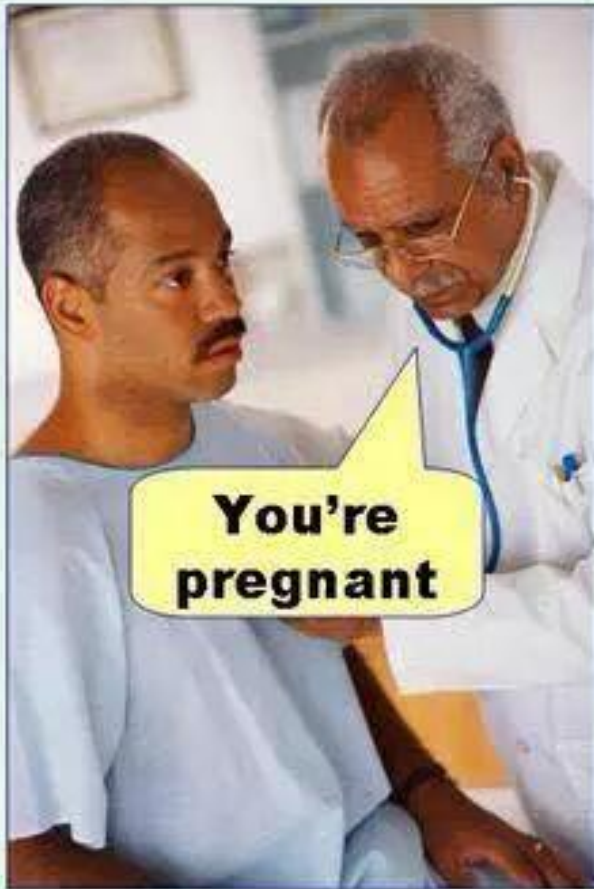


Procedimiento para la prueba de hipótesis

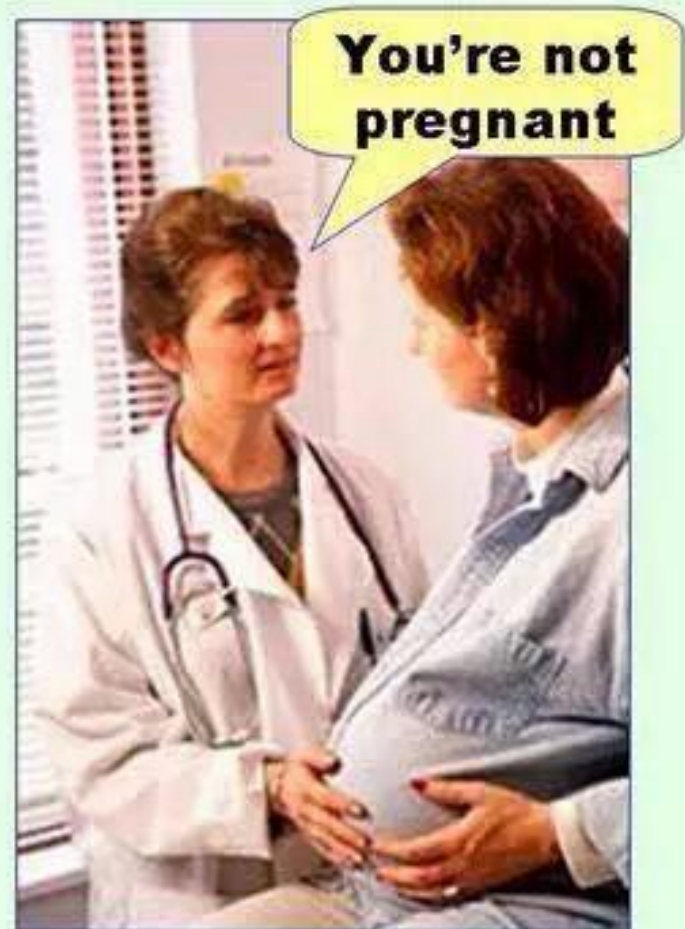


Errores tipo I y II



Type I error
(false positive)



Type II error
(false negative)



Errores tipo I y II

	Aceptar H_0	Rechazo H_0
H_0 verdadera		Error tipo I (α) (falso positivo)
H_0 falsa	Error tipo II (β) (falso negativo)	

α es la probabilidad de rechazar la hipótesis nula cuando esta es verdadera,

Dado que β es la probabilidad de aceptar H_0 cuando esta es falsa,
 $1-\beta$ es la probabilidad de realizar una decisión correcta y rechazar H_0 cuando esta es falsa = **potencia de la prueba**

Errores tipo I y II

Factores que afectan a β (y a la potencia de la prueba)

- **Controlables:**
 - Tamaño de la muestra
 - Valor de significancia (α)
- **Incontrolables**
 - Tamaño de efecto
 - Varianza de la población

Significancia estadística

- Probabilidad de cometer Error tipo I (α)
- Por convención: $\alpha=0.05$
- OJO= La significancia no dice nada respecto al tamaño del efecto

BEWARE FALSE CONCLUSIONS

Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.

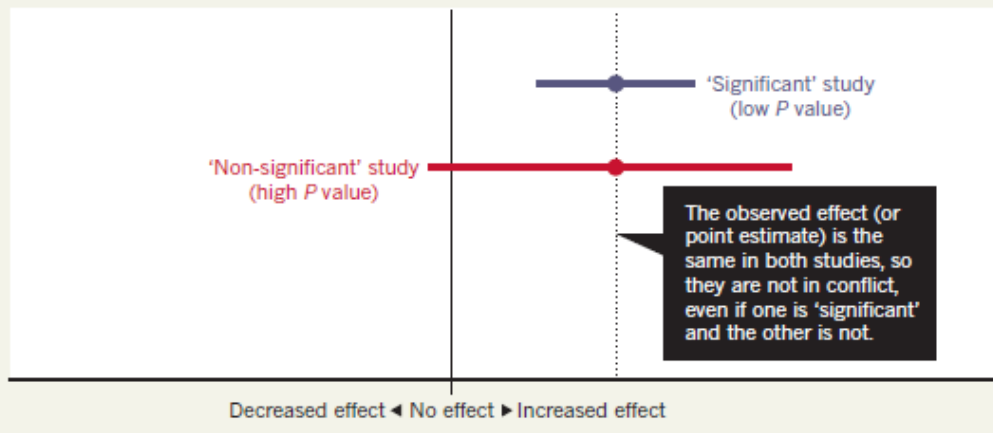


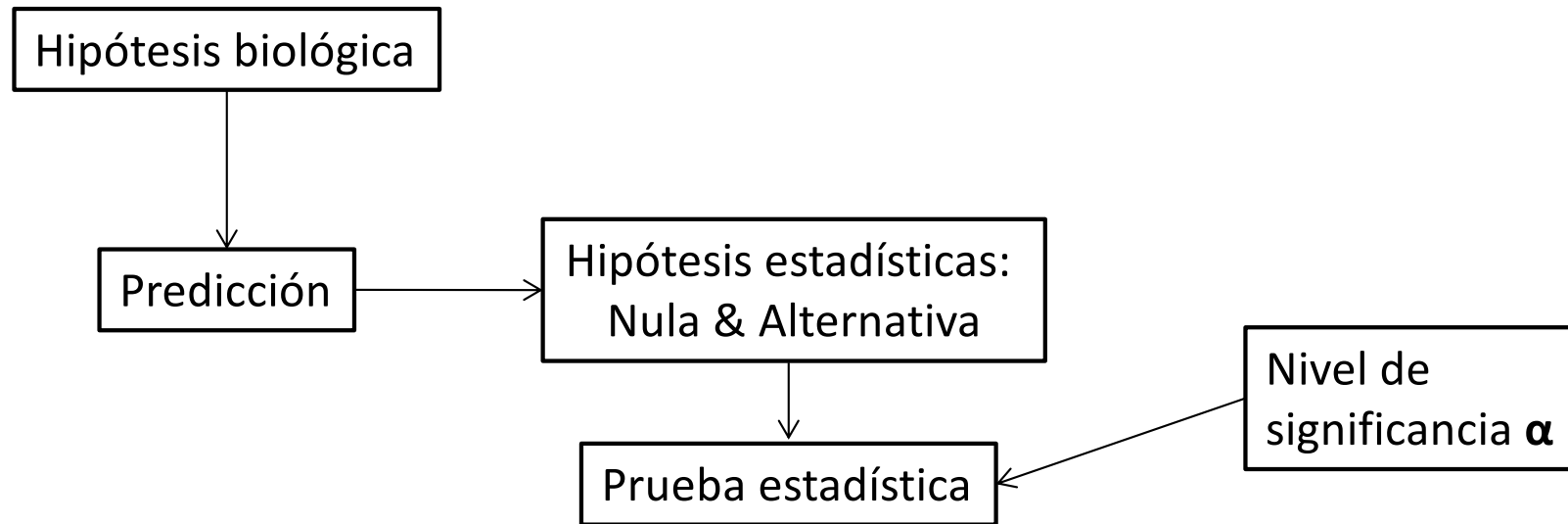
Figure 1. "Significant-itis" — a maniac-depressive disorder characterized by an obsession with the P-value.

EDITORIAL

Moving to a World Beyond “ $p < 0.05$ ”

- Don't base your conclusions solely on whether an association or effect was found to be “statistically significant” (i.e., the p -value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p -value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Procedimiento para la prueba de hipótesis

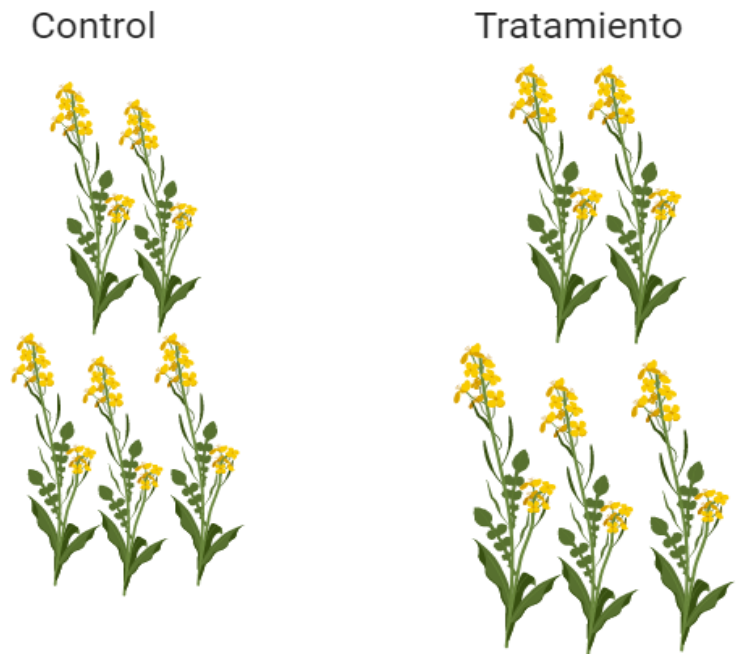
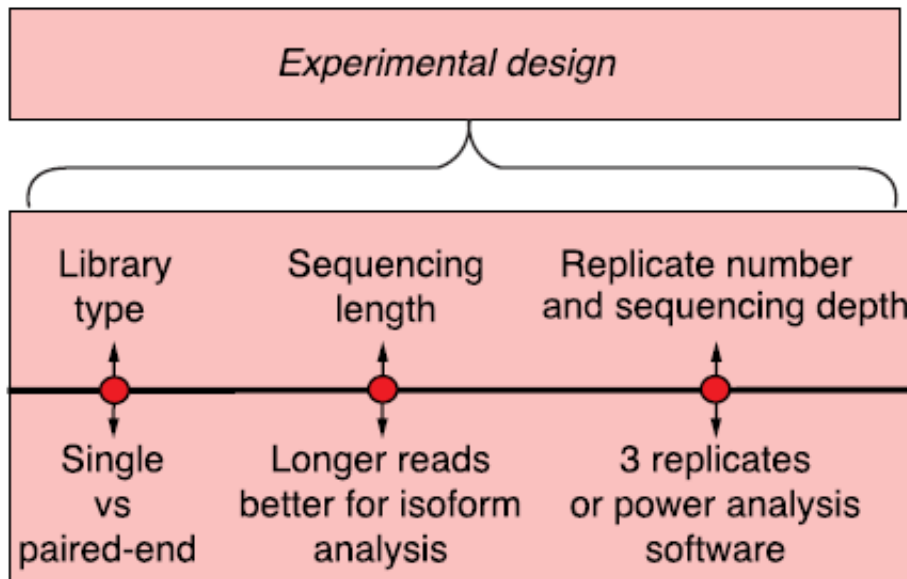


Depende del diseño experimental
y del tipo de variable

Diseño Experimental

En un diseño experimental se manipulan deliberadamente una o más variables, vinculadas a las causas, para medir el efecto que tienen en otra variable de interés.

Que los datos generados puedan responder la pregunta biológica de interés!!



Prueba estadística

		Tipo de datos			
		Numéricos (gaussiana)	Ordinal o numérica (no gaussiana)	Numéricos (outliers)	Nominal binaria (2 resultados posibles)
Objetivo	Comparar 2 grupos independientes	Prueba t para 2 muestras independientes	Prueba de Mann-Whitney	Prueba de Yuen para muestras independientes	Prueba de Fisher o Chi-cuadrado (para muestras grandes)
	Comparar 2 grupos relacionados	Prueba t para 2 muestras relacionadas	Prueba de Wilcoxon para muestras relacionadas	Prueba de Yuen para muestras relacionadas	Prueba de McNemar
	Comparar 3 o más grupos independientes	ANOVA de 1-vía para muestras independientes	Prueba de Kruskal-Wallis	ANOVA robusto de 1-vía para muestras independientes	Prueba Chi-cuadrado
	Comparar 3 o más grupos relacionados	ANOVA de 1-vía para muestras relacionadas	Prueba de Friedman	ANOVA robusto de 1-vía para muestras relacionadas	Prueba Q de Cochran
	Asociar 2 variables	Correlación de Pearson	Correlación de Spearman o Kendall	Correlación robusta	Coefficiente V de Cramer