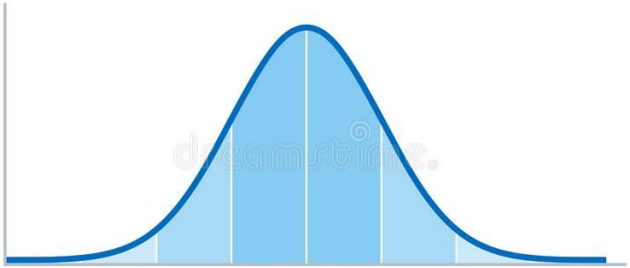
Módulo 12
Distribuciones de datos y muestreo

Cursado 2023
Instituto Superior Tecnológico Empresarial Argentino
Ing. Carla Pezzone

ISTEA

AGENDA DE LA CLASE

- Muestreo aleatorio y sesgo de la muestra
- Tamaño y calidad de los datos
- La distribución muestral
- Teorema del límite central
- Error estándar
- Bootstrap
- Intervalo de confianza
- Distribución normal
- Tamaño de la muestra



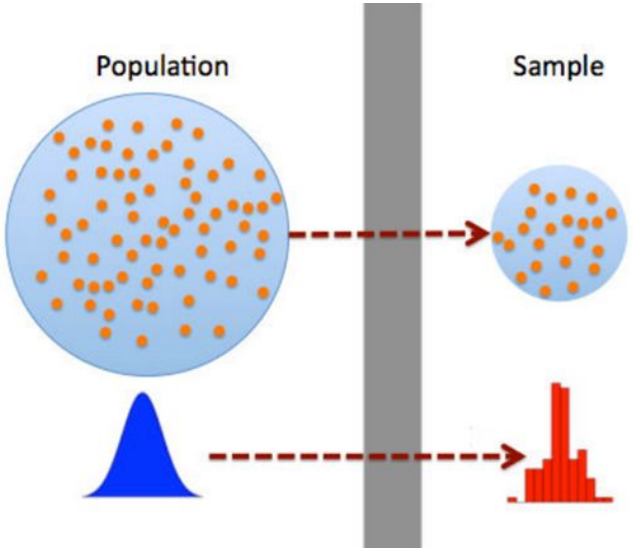
01 - Muestreo aleatorio y sesgo de la muestra

Una **muestra** es un **subconjunto de datos de un conjunto más grande**.

Los estadísticos llaman **población** a este conjunto de datos más grande.

¿Con la era del Big Data es el fin de los muestreos? NO

- Datos de diferentes calidades
- Necesidad de analizar menos cantidad de datos para aumentar eficiencia
- Modelos predictivos se ponen a prueba con muestras
- Con distintas muestras se hacen comparaciones



La **estadística tradicional** se ha centrado en teorías basadas en **supuestos sólidos sobre la población**

La **estadística moderna** se ha movido hacia el lado derecho, donde tales suposiciones no son necesarias

01 - Muestreo aleatorio y sesgo de la muestra

El **muestreo aleatorio** es un **proceso** en el que **cada miembro disponible de la población que se muestrea tiene la misma probabilidad de ser elegido** para la muestra en cada extracción.

La **muestra resultante** se denomina **muestra aleatoria simple**.

El muestreo puede ser:

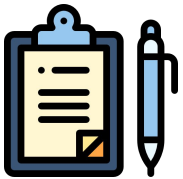
- **Con reposición**: en el que **las observaciones se vuelven a colocar en la población** después de cada extracción para una posible nueva selección
- **Sin reposición**: en donde las observaciones, **una vez seleccionadas, no están disponibles para futuras extracciones**

01 - Muestreo aleatorio y sesgo de la muestra

El **sesgo de la muestra** es aquella muestra que de manera **significativa** y **no aleatoria**, es **diferente de la población** más grande a la que se pretendía representar.

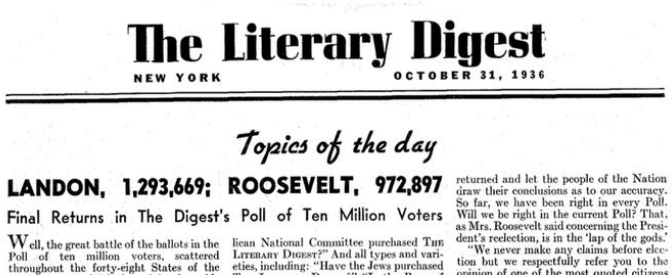
Casi ninguna muestra, incluídas las aleatorias, será exactamente representativa de la población

El **sesgo ocurre cuando la diferencia es significativa y se puede esperar que continúe** para otras muestras extraídas del mismo modo que la primera.



Veamos un ejemplo histórico

01 - Muestreo aleatorio y sesgo de la muestra



The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

Today — AMERICA SPEAKS — THE NATIONAL WEEKLY POLL of PUBLIC OPINION — Next Sunday

Institute Forecasts the Re-election of Franklin D. Roosevelt, Gives Him 54% of Popular Vote, Minimum of 315 Electors

Major Party Percent Is 55.7; New York in F.D.R. 'Sure' Column

Election Forecast

Election Will Test Clashing Poll Methods

- **Literary Digest** es uno de los principales diarios.
- Encuestó a sus suscriptores además de hacerlo a los individuos que figuraban en otras listas. Total: **10 millones de personas**.
- Predijo una **victoria** aplastante de **Landon**.

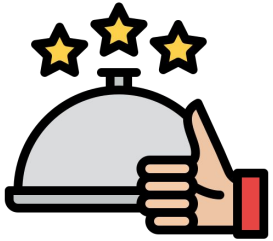
- George Gallup, fundador de **Gallup Poll**.
- Realizó encuestas quincenales a solo **2000 personas**.
- Predijo acertadamente una **victoria de Roosvelt**.

Literary Digest se relacionaba con personas con un estatus socioeconómico relativamente alto

La muestra estaba **sesgada**

01 - Muestreo aleatorio y sesgo de la muestra

Ejemplos similares es cuando vemos las reseñas en internet, en las redes sociales.



Esto conduce a un **sesgo de autoselección**: las personas motivadas para escribir reseñas pueden haber tenido **malas experiencias**, pueden estar **asociadas con el establecimiento** o simplemente pueden ser un tipo de persona **diferente de las que no escriben reseñas.**

El **sesgo estadístico** se refiere a **errores de medición o muestreo** que son **sistemáticos y se producen por el proceso de medición o muestreo**.

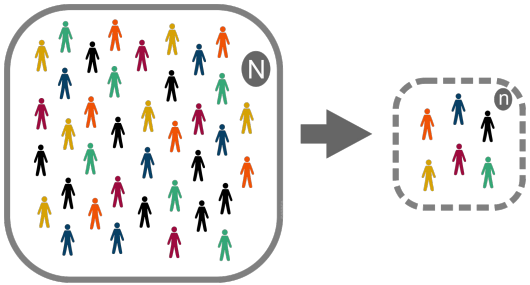
El **muestreo aleatorio no siempre es fácil.**
La **definición adecuada de una población** accesible es clave

En el **muestreo estratificado** la población se divide en **estratos** y se toman muestras aleatorias de cada estrato. De esta manera se generan tamaños de muestra equivalente para cada estrato.

02 - Tamaño y calidad de los datos

En la era del big data, a veces resulta sorprendente que **cuánto más pequeño, mejor**

El **tiempo** y el **esfuerzo** dedicados al muestreo aleatorio no solo **reducen el sesgo**, sino que también permite una **mayor atención a la exploración y calidad de los datos.**



¿Cuál debe ser el **tamaño de una muestra**?

Tenemos que entender los conceptos de:

- Margen de error
- Nivel de confianza
- Variabilidad de los datos

03 - La distribución muestral

La **estadística descriptiva** son los **métodos** que **analizan los datos de una muestra**, que es un subconjunto de una población

Las **estadísticas de una muestra** son la media, el desvío estándar, etc.

La **estadística inferencial** usa las **estadísticas de muestras** calculadas para sacar **inferencias** (conclusiones) sobre los parámetros de la población completa

Usan los **parámetros de la población**, que no son más que las **estadísticas** que vimos para una **muestra**, **pero calculadas para toda la población**

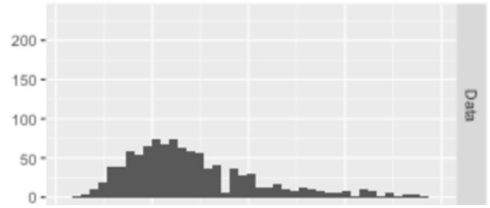
El término **distribución muestral** de un estadístico se refiere a la **distribución del estadístico de una muestra** **sobre muchas muestras extraídas de la misma población**

No confundir este término con la **distribución de datos** que se refiere a la **distribución de los puntos de datos individuales**

03 - La distribución muestral

Por ejemplo, tomamos 3 muestras de los ingresos anuales de solicitantes de préstamos para LendingClub:

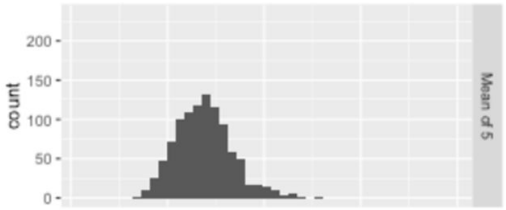
Una muestra de **1000 valores**



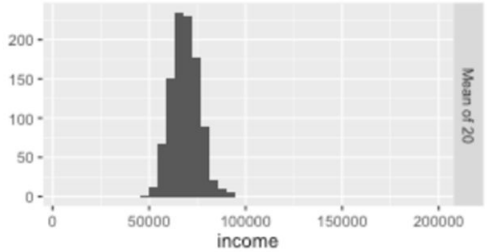
Está ampliamente distribuido y sesgado hacia los valores altos (sesgado hacia la derecha).

A mayor ingreso de los solicitantes, menos cantidad de ellos solicitan préstamos.

Una muestra de **1000 medias de 5 valores**



Una muestra de **1000 medias de 20 valores**



Los histogramas de medias de 5 y 20 son cada vez más compactos y tienen la forma de campana.

Distribución muestral de la media muestral, resulta de la distribución que se obtiene de extraer un gran número de muestras de su población y se calcula la media de todas las muestras.

04 - Teorema del límite central

El fenómeno que acabamos de describir se denomina **teorema del límite central**

Asegura que las **medias extraídas de varias muestras** se asemejan a la conocida curva normal en forma de campana (**Distribución Normal**), incluso si la población de origen no se distribuye normalmente, siempre que el **tamaño de la muestra sea lo bastante grande** y la **desviación de los datos de lo habitual no sea demasiado grande**.

05 - Error estándar

El **error estándar** es una **métrica** única que **resume la variabilidad de la distribución muestral de un estadístico.**

No es lo mismo que la **desviación estándar**, que mide la **variabilidad de los puntos de datos individuales**.

Se puede estimar utilizando un estadístico basado en la **desviación estándar s** de los valores de la muestra y el **tamaño de la muestra n**

$$Error\ estándar = SE = \frac{s}{\sqrt{n}}$$

A medida que aumenta el tamaño de la muestra, el error estándar disminuye.

La validez de la fórmula del error estándar **surge del teorema del límite central**:

1. Recolectamos varias muestras nuevas de la población
2. Para cada nueva muestra calculamos el estadístico (por ejemplo, la media)
3. Calculamos la desviación estándar de las medias calculadas en el Paso 2. Utilizamos este valor para la estimación del error estándar.

Imposible

06 - Bootstrap

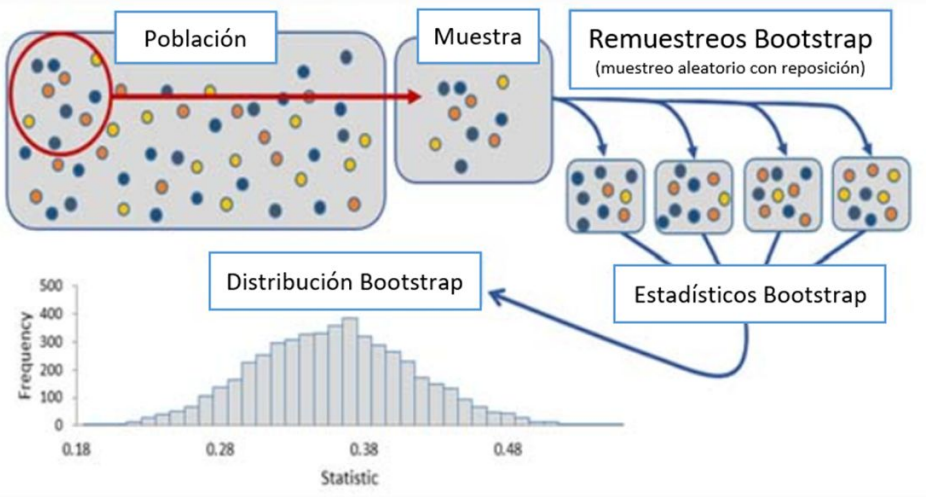
Método Bootstrap

Se ha convertido en la forma estándar de **estimar el error estándar**

Se puede utilizar para prácticamente **cualquier estadístico**

No se basa en el teorema del límite central ni en otros supuestos distributivos

Se hace **muestreo con reposición**



Conceptualmente, podemos imaginar que el método bootstrap **replica la muestra original** miles o millones de veces, de modo que tengamos una **hipotética población que incorpore todo el conocimiento de nuestra muestra original**.

06 - Bootstrap

El método Bootstrap se utiliza para determinar el tamaño de la muestra.

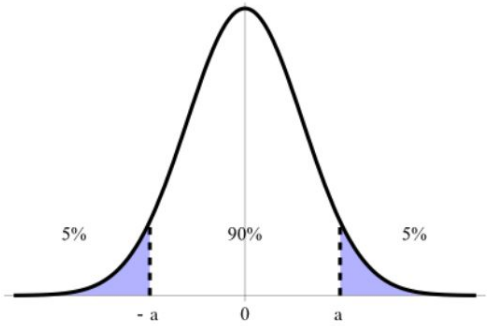
Podemos experimentar con diferentes valores de n para ver cómo se ve afectada la distribución muestral.

No compensa para muestras de pequeño tamaño, no crea ni rellena huecos en un conjunto de datos existentes

Ejemplo

07 - Intervalo de confianza

Un **intervalo de confianza** de x% **en torno a la estimación de la muestra** debería, en promedio, contener estimaciones de muestras similares el x% del tiempo



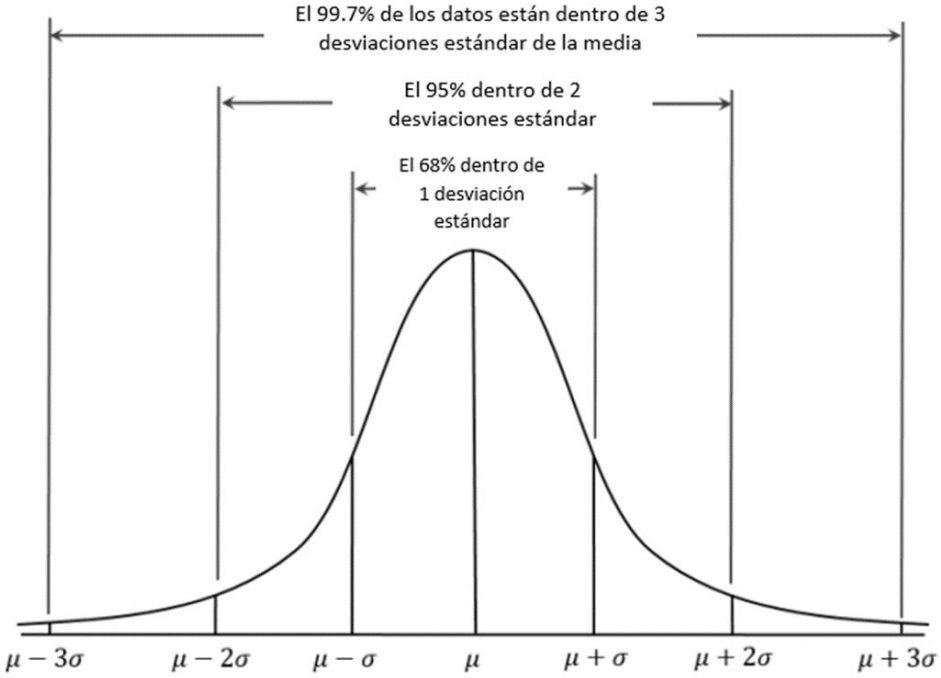
Una forma de pensar en un intervalo de confianza del 90% es la siguiente: es el intervalo que encierra el 90% de la parte central de la distribución de muestreo Bootstrap del estadístico de una muestra

El porcentaje asociado al intervalo de confianza se denomina **Nivel de Confianza** (**alfa**).

- Cuanto **mayor sea el nivel de confianza**, más amplio será el intervalo
- Cuanto **más pequeña es la muestra**, más amplio es el intervalo (es decir, mayor es la incertidumbre)

08 - Distribución normal

La **distribución normal** en forma de campana es icónica en la estadística tradicional.



Una **distribución normal estándar** es aquella en la que las unidades del eje x se expresan **en términos de desviaciones estándar de la media**.

08 - Distribución normal

Para comparar los datos con una distribución normal estándar, restamos la media y luego la dividimos por la desviación estándar.

A esto también se le denomina
normalización o **estandarización**

$$z = \frac{x - \mu}{\sigma}$$

El valor transformado se denomina **puntuación z**
A la distribución normal de estos valores transformados se le llama **distribución z**

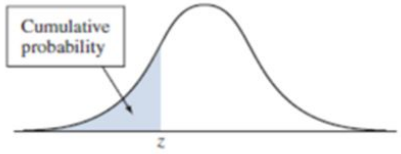
Es importante tener en cuenta que **convertir los datos a puntuación z** (es decir, estandarizar o normalizar los datos) **no hace que los datos se distribuyan normalmente**.

Coloca los datos en la **misma escala que la distribución normal estándar**, a menudo con fines de comparación

08 - Distribución normal

La **distribución z acumulada** suele presentarse en una **tabla que da la probabilidad** de que el resultado de la variable aleatoria distribuida normalmente sea menor o igual a mu más z veces sigma

z Table



$$z = \frac{x - \mu}{\sigma}$$

$$x_i = \mu + z_i\sigma$$

Standard normal cumulative probabilities

Valor de z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	Valor de p	006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	p-value	009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013		013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064

08 - Distribución normal

Ejemplo: Una población de gansos verdes migra cada otoño de la región del Báltico a la costa atlántica en Europa. La **duración de la migración se distribuye normalmente** con una **media de 4** y una **desviación estándar de 1,3 días**.
Ahora, ¿cuál sería la probabilidad de que los gansos completaran su migración dentro de 6 días hasta aquí?

1

$$z = \frac{x - \mu}{\sigma} = \frac{6\ días - 4\ días}{1.3\ días} = 1.54$$

=DISTR.NORM.ESTAND.N

2

Standard normal cumulative probabilities

z	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
...										
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

	A	B	C	D	E
1	x=	6			
2	μ =	4			
3	σ =	1.3			
4					
5	z=	1.54	=(B1-B2)/B3		
6					
7	p=	0.9380	=DISTR.NORM.ESTAND.N(B5;VERDADERO)		
8					

3

p = 0.9382. Esta es la probabilidad de que los gansos completen la migración dentro de los 6 días de un año dado.

08 - Distribución normal

Ahora, **si conocemos la probabilidad** y queremos **encontrar el valor crítico** correspondiente a la variable aleatoria, entonces

Por ejemplo, si queremos saber **cuántos días** les toma a los gansos migrar en el **10% de los casos** (es decir, queremos calcular el percentil 10 de duración de la migración)

1. La probabilidad que estamos buscando es 0.1 (10%)

Standard normal cumulative probabilities

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
...										
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170

Vemos en la tabla que el valor p es 0.1003 para un z de -1.28

=INV.NORM

29			
30	p=	0.1	
31	μ =	4	
32	σ =	1.3	
33			
34	x=	2.33	=INV.NORM(B30;B31;B32)
35			

2. $x = \mu + z\sigma = 4\ días + (-1.28) * 1.3\ días = 2.34\ días$

3. La respuesta es entonces que el 10% de las veces, a los gansos le lleva finalizar la migración 2,34 días.

09 - Tamaño de muestra

La elección del tamaño de la muestra depende de 3 factores:

1. **Qué tan preciso quieres ser**
   Recuerda que el intervalo de confianza se calcula ± de un estimador puntual de un cierto **margen de error**. Que tan preciso quieras ser depende de cuán grande permites ese margen. *A menor margen, mayor será el tamaño de la muestra.*

2. El tamaño depende del **nivel de confianza** que quieras usar.
   A mayor nivel de confianza mayor tamaño.

3. Depende de la **variabilidad de tus datos**.
   A mayor desviación estándar de tu variable, mayor tamaño.

09 - Tamaño de muestra

Fórmulas para el tamaño de la muestra:

Para una población infinita

$$n = \frac{z^2 * p * q}{e^2}$$

Para una población finita

$$n = \frac{N * z^2 * p * q}{e^2 * (N - 1) + z^2 * p * q}$$

Cuando desconocemos la cantidad total de la población a muestrear

q = proporción complementaria de p, es decir, **q = 1 - p**

El valor p, en general, no lo conoces, por lo que es necesario hacer una conjetura llamada **Enfoque seguro**: **p = 0.5**

09 - Tamaño de muestra

Por ejemplo: queremos calcular el tamaño de muestra para una **población desconocida**, con un **nivel de confianza de 95%** y un **margen de error del 3%**. Se desconoce la probabilidad **p** del evento estudiado, por lo que podemos tomar un enfoque seguro y **adoptar 0.5**.

$$n = \frac{z^2 * p * q}{e^2} = \frac{(-1.96)^2 * 0.5 * (1 - 0.5)}{0.03^2} = 1067$$

z = -1.95 es el z para la probabilidad de 0.025 y sale de tabla.
0.025 sale de 0.5/2 y 0.5 sale de considerar 1- 95% del IC.

Standard normal cumulative probabilities

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
...										
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294

09 - Tamaño de muestra

Por ejemplo: si queremos calcular el tamaño de muestra para una **población de 543098** consumidores de una marca de gaseosa determinada, con un **nivel de confianza de 95%** y un **margen de error del 3%**. Se desconoce la probabilidad **p** del evento estudiado, por lo que podemos tomar un enfoque seguro y adoptar **0.5**

$$n = \frac{N * z^2 * p * q}{e^2 * (N-1) + z^2 * p * q} = \frac{543098 * (-1.96)^2 * 0.5 * (1-0.5)}{(0.03)^2 * (1-543098) + (-1.96)^2 * 0.5 * (1-0.5)} = 1065$$

¿PREGUNTAS?

¡Muchas gracias!