INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5625 - INVESTIGACIÓN DE MERCADOS

Repaso: Estadística para la investigación de mercados

André Carboni

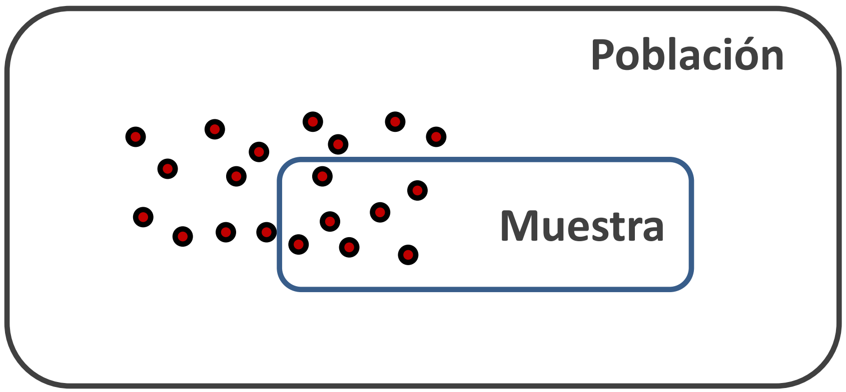
Semestre primavera 2012

MUESTREO:
DISEÑO Y PROCEDIMIENTOS

Muestra

¿Qué es una muestra?

- Grupo de elementos que es un subconjunto del universo y que serán contactados para obtener información.



Muestra v/s censo

	Características que favorecen el uso de:	
	Muestra	Censo
Presupuesto	Pequeño	Grande
Tiempo disponible	Poco	Mucho
Tamaño de la población	Grande	Pequeña
Varianza de la característica	Pequeña	Grande
Costo de los errores de muestreo	Bajo	Alto
Naturaleza de la medición	Destructiva	No destructiva
Atención a casos individuales	Sí	No

Proceso de muestreo

Definir la población meta

↓

Determinar el marco de muestreo

↓

Seleccionar la técnica de muestreo

↓

Determinar el tamaño de la muestra

↓

Llevar a cabo el proceso del muestreo

Proceso de muestreo

Definir la población meta

↓

Determinar el marco de muestreo

↓

Seleccionar la técnica de muestreo

↓

Determinar el tamaño de la muestra

↓

Llevar a cabo el proceso del muestreo

Definición de la población meta

- **Población meta:**
  - Conjunto de elementos y objetos que poseen la información buscada y acerca del cual se harán inferencias
  - Debe definirse con precisión, sino es ineficaz o engañosa

- La definición de la población meta es en base a:
  - Elementos: dueñas de casa, mujeres, entre 18 y 55 años, que han comprado leche en los últimos 30 días en…
  - Unidades de muestreo: Supermercados, hogares, …
  - Extensión: Límites geográficos
  - Tiempo

Proceso de muestreo

1. Definir la población meta
2. Determinar el marco de muestreo
3. Seleccionar la técnica de muestreo
4. Determinar el tamaño de la muestra
5. Llevar a cabo el proceso del muestreo

Determinación del marco de muestreo

- Consiste en listar un conjunto de instrucciones para identificar a la muestra:
  - Base de datos de e-mails
  - Generación aleatoria de números de teléfono
  - Ubicación geográfica con un mapa
  - …

- Formas de evitar el error de determinación del marco de muestreo:
  - Redefinir la población en términos del marco
  - Seleccionar a los encuestados ad hoc en la etapa de recolección de datos
  - Ajustar los datos recabados con un esquema de ponderaciones para equilibrar el error

Proceso de muestreo

Definir la población meta

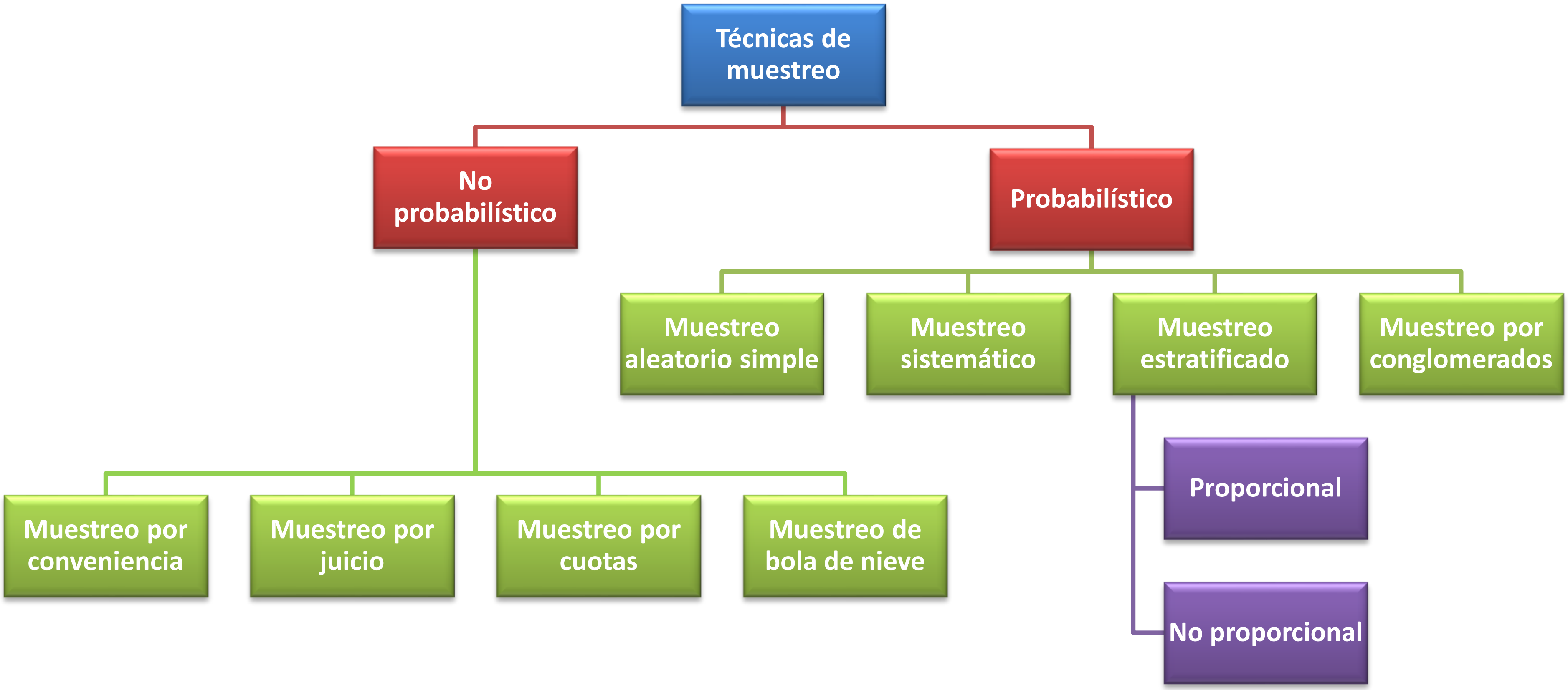
Determinar el marco de muestreo

Seleccionar la técnica de muestreo

Determinar el tamaño de la muestra

Llevar a cabo el proceso del muestreo

Elección de la técnica de muestreo



Muestreo no probabilístico

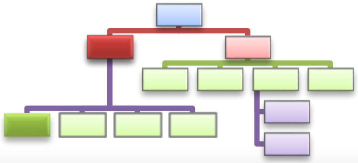
- No hay forma de establecer la probabilidad de seleccionar un determinado elemento → resultados no se pueden proyectar estadísticamente a la población.
- No necesariamente son inexactas o peores que las probabilísticas. Pueden ser o no representativas dependiendo del método y controles de selección.
- Más económico que el muestreo probabilístico.
- Aumentar el tamaño muestral no resuelve los problemas de sesgo que se presenten (por ej. una parte de la población objetivo puede quedar fuera del muestreo por definición).
- La decisión acerca del tamaño de la muestra está relacionada directamente con el costo de la investigación.
- El tamaño de la muestra no depende del tamaño de la población en estudio.

Muestreo por conveniencia

- La muestra la selecciona el entrevistador, sólo restringido a considerar elementos dispuestos a participar.
- La más económica y rápida
- No hay forma de determinar la representatividad de la muestra
  ➔ No se puede generalizar a la población.
  ➔ OK para Investigación exploratoria.

Ejemplos:

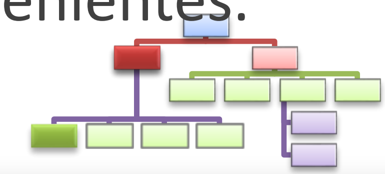
- Entrevistar mujeres en un centro comercial sin normas sobre cuotas.
- Encuestas por correo, e-mail o Internet.
- Muestra para un focus group (en general).



Ejemplo

A	B	C	D	E
1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

- El grupo D se reunió en un momento y lugar convenientes.

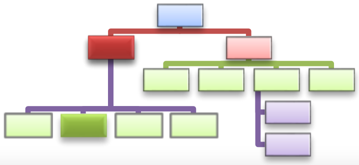


Muestreo por juicio

- Caso particular de muestreo por conveniencia.
- Se selecciona una muestra de elementos que parece representar a la población a analizar.
- Es económico, práctico y rápido.
- Es subjetivo.

Ejemplos:

- Mercados de prueba: selección de una región geográfica que se cree es representativa del todo el mercado.
- Tiendas elegidas para probar nuevo sistema de exhibición de mercancías.
- Elegir región metropolitana, descartar comunas muy pobres, descartar comunas poco seguras, seleccionar manzanas representativas, seleccionar 1 casa cada 10 en la manzana.



Ejemplo

A	B	C	D	E
1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

- Investigador considera que B, C y E son típicos y convenientes.

Muestreo por cuotas

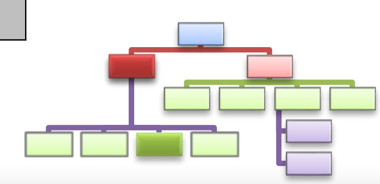
INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

- Es un muestreo por juicio restringido de dos etapas.
  - Desarrollar categorías de control, o cuotas, de los elementos de la población (ej: edad, sexo, GSE, etc.). Definir proporciones de cada cuota según la proporción en la población.
  - Seleccionar por conveniencia o juicio a los elementos de la muestra.
- Se agregan elementos a la muestra hasta que se completa la cuota.

Ejemplo:

- De alguna población particular, describo las siguientes cuotas.

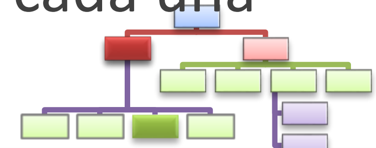
GSE	Segmento				Total
	Hombres		Mujeres		
	< 30 años	>= 30 años	< 30 años	>= 30 años	
ABC1	100	60	95	45	**300**
C2	85	40	75	50	**250**
C3	75	50	60	65	**250**
D	50	50	55	45	**200**
Total	310	200	285	205	**1000**



Ejemplo

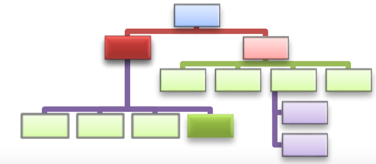
A	B	C	D	E
1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

- 5 cuotas (A, B, C, D, E) e impongo un elemento de cada una



Muestreo de bola de nieve

- Se selecciona un grupo inicial de encuestados al azar, a quienes después de entrevistar se les pide que identifiquen a otras personas que pertenezcan a la población meta de interés.
- La muestra se selecciona en base a referencias.
- Los seleccionados tendrán características demográficas y psicográficas más similares que si se eligen todos al azar.
- Útil para estimar características raras (ej: Hombres viudos menores a 35).



Ejemplo

A	B	C	D	E
1	6	11	16	21
$\boxed{2}$	7	$\boxed{12}$	17	22
3	8	$\boxed{13}$	$\boxed{18}$	23
4	$\boxed{9}$	14	19	24
5	10	15	20	25

- Elijo a miembros 2 y 9. El 2 refiere a 12 y 13. El 9 refiere al 18.

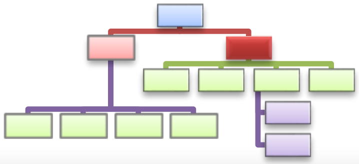
Muestreo probabilístico

- Todas las unidades de la población tienen una probabilidad conocida y mayor que cero de ser seleccionadas en la muestra.

- Permite analizar la representatividad de la muestra.

- A partir de la muestra es posible calcular un intervalo de confianza de la variable de interés en la población.

- Se utilizan cuando se requiere una estimación muy precisa de la variable a estudiar en la población.

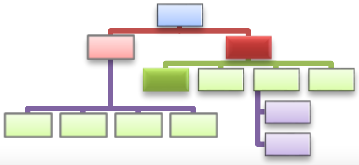


Muestreo aleatorio simple (MAS)

- Cada unidad o elemento tienen igual probabilidad de ser seleccionadas. Cada muestra posible de tamaño n tiene la misma probabilidad de ser elegida.
- Se requiere una clasificación organizada que enumere a todos las unidades de la población.
- Se selecciona aleatoriamente.

- **Ventajas:**
  - Fácil de entender y aplicar.

- **Desventajas:**
  - No siempre son factibles (ej.: enumeración).
  - Pueden ser costosas (ej.: área geográfica).
  - Distorsiones en muestras pequeñas.



Ejemplo

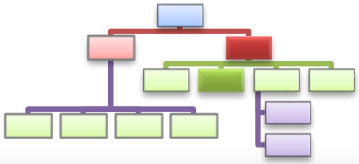
A	B	C	D	E
1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

- Seleccionar 5 números aleatorios del 1 al 25.

Muestreo sistemático

- La muestra se elabora partiendo de un elemento elegido arbitrariamente, seleccionando los elementos siguientes de la lista con un salto constante.
- Para calcular intervalo del muestre, se divide el tamaño de la población (N) por el tamaño de la muestra (n) y se redondea.
- Si la población no está ordenada ➔ Se parece al MAS.
- Si la población está ordenada según característica de interés ➔ Más representativo de la población

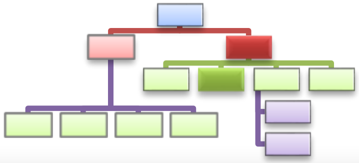
- **Ventajas:**
  - Fácil de utilizar
  - No hay que generar números aleatorios ➔ Más económica
  - Menor error de muestreo que el MAS si está ordenado
  - Se puede aplicar sin conocer marco de muestreo (ej: ver i-ésima persona que sale de una tienda).



Ejemplo

A	B	C	D	E
1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

- Intervalo = 25/5 = 5. Se elige número aleatorio entre 1 y 5 como punto de partida (2) y el intervalo es de 5 en 5.

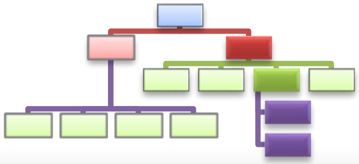


Muestreo estratificado

- Dividir a los elementos en subpoblaciones (estratos). Los **estratos** deben ser mutuamente excluyentes.
- Los elementos dentro de cada estrato deben ser homogéneos, mientras que entre estratos deben ser heterogéneos.
- Dentro de cada estrato, seleccionar una muestra independiente (ej.: MAS, muestreo sistemático,…)
- Se miden las variables de interés en cada muestra, se pondera y se estima un total para la población

- **Ventajas:**
  – Reduce significativamente el intervalo de confianza

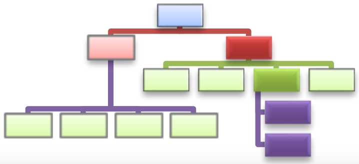
- **Desventajas:**
  – Más complejo que MAS
  – Más costoso que MAS



Muestreo estratificado

- **Asignación Proporcional:**
  - El número de elementos seleccionados es proporcional al tamaño relativo del estrato con respecto a la población. En el peor de los casos será tan eficaz como MAS.

- **Asignación No Proporcional u Óptima:**
  - Hay una doble ponderación de los elementos: tamaño del estrato al que pertenece y varianza de la variable a calcular.
    - Mayor tamaño ➔ Influye más en la media de la población.
    - Mayor varianza ➔ Menor precisión ➔ Tomar más elementos.



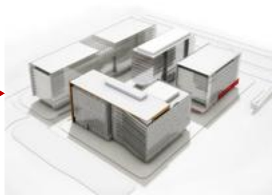
Ejemplo

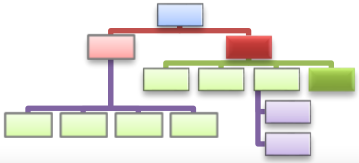
A	B	C	D	E
1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

- Se toma aleatoriamente un elemento de cada estrato.

Muestreo por conglomerados

- Se realiza una partición (excluyente y exhaustiva) de la población de unidades de muestreo en grupos cerrados (conglomerados) a partir de una variable de clasificación.
- Se selecciona una muestra dentro del conjunto de conglomerados (ej.: MAS, salto sistemático, …).
- Se selecciona una muestra dentro de cada conjunto de grupos cerrados elegidos en la etapa anterior (ej.: MAS).
- Puede haber varias fases.
- **Ejemplo:**
  - Región
  - Ciudad
  - Manzana
  - Casa
  - Familia
  - Población objetivo

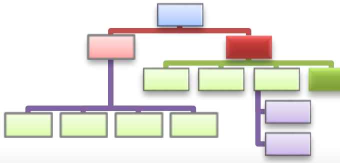




Ejemplo

A	B	C	D	E
1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

- Seleccionar al azar 3 conglomerados (B, D, E). Se elige al azar 1 o 2 elementos de cada uno.



Proceso de muestreo

- Definir la población meta
- Determinar el marco de muestreo
- Seleccionar la técnica de muestreo
- **Determinar el tamaño de la muestra**
- Llevar a cabo el proceso del muestreo

Determinar tamaño de la muestra

- Se deben tomar en cuenta los siguientes enfoques prácticos:
  - Reglas empíricas.
  - Restricción presupuestaria.
  - Valor de la información.
  - Utilización de estudios comparables.
  - Analizar los factores que determinan el tamaño:
    - Número de grupos y subgrupos dentro de la muestra (por ej. análisis detallado por segmento).
    - Variabilidad de la población (por ej. si todos los individuos de la población objetivo se comportaran igual nos bastaría considerar una muestra de 1 caso).

Proceso de muestreo

1. Definir la población meta
2. Determinar el marco de muestreo
3. Seleccionar la técnica de muestreo
4. Determinar el tamaño de la muestra
5. Llevar a cabo el proceso del muestreo

Llevar a cabo el proceso

- La realización del proceso de muestreo requiere una especificación detallada de cómo se llevarán a cabo las decisiones del diseño de muestreo relacionadas con la población, el marco de muestreo, las técnicas de muestreo y el tamaño de la muestra.

- Es necesario especificar los procedimientos a seguir en el caso de no respuesta (por ejemplo vivienda desocupada).

Tratamiento de la no respuesta

Motivos asociados a este fenómeno:

- Simplemente no quiere responder.
- No tiene la capacidad o conocimiento suficiente para contestar.
- No se encuentra.

Soluciones:

- Mejorar el diseño de la encuesta (por ej. preguntas de perfilamiento al final del cuestionario).
- Reintento de generar la respuesta (por ej. visitas repetidas al mismo hogar para maximizar su probabilidad de respuesta).
- Estimar el efecto asociado a la no respuesta (por ej. considerar que este grupo puede tener una opinión negativa de mi servicio).

TAMAÑO DE LA MUESTRA

Muestreo aleatorio simple (MAS)

- Todos los elementos de la población tienen una probabilidad conocida, no nula, de ser elegidos. Pero al no ser devueltos a la población, la probabilidad de que salga un elemento determinado depende de las extracciones anteriores en el caso de población finita.

- Consiste en elegir n elementos de una población de N elementos.

- Las muestras posibles son:

$$\binom{N}{n} = \frac{N!}{(N-n)!n!}$$

Ejemplo:

- ¿Cuántas muestras de 100 elementos pueden ser obtenidas de una población de 120?

$$\binom{120}{100} = \frac{120!}{20!100!} = 29.462.227.291.176.600.000.000$$

Muestreo aleatorio simple (MAS)

- La probabilidad de que una muestra sea elegida es:

$$\text{Prob(muestra específica)} = \frac{1}{\binom{N}{n}}$$

- La probabilidad de que un elemento forme parte de una muestra dada será:

$$\frac{\text{Número de muestras probables}}{\text{Número de muestras posibles}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

- Por lo tanto, la media y varianza muestral de una variable podrá variar **de muestra en muestra**.

- En la medida que se aumenta el tamaño de la muestra (n), entonces el error cometido por elegir una muestra determinada disminuye (error muestral). Esto debido a que la cantidad de muestras posibles de escoger disminuye y en consecuencia, también la probabilidad de variación del indicador de muestra en muestra.

Muestreo aleatorio simple (MAS)

¿Cómo hacer un muestreo aleatorio?

- Con un listado de elementos de la población y usando números aleatorios.

Ejemplo:

- ¿Cómo elegir una muestra de 10 elementos de una población de 1000?

N° Muestra	N° Aleatorio	N° Población
1	0.73812	738
2	0.29499	295
3	0.34016	340
4	0.93316	933
5	0.79095	791
6	0.17945	180
7	0.59108	591
8	0.20253	203
9	0.22276	223
10	0.95729	957

Muestreo aleatorio simple (MAS)

¿Cómo hacer un muestro aleatorio?

- Con un listado de elementos de la población y muestreo sistemático con arranque aleatorio.

Ejemplo:

- Elegir una muestra de 6 a partir de una población de 20.

$$K = \left\lfloor \frac{N}{n} \right\rfloor = \left\lfloor \frac{20}{6} \right\rfloor = \lfloor 3{,}33333\ldots \rfloor = 3$$

Elegido aleatoriamente entre los K primeros elementos

N° Población
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Muestreo aleatorio simple (MAS)

- **¿Cómo estimar parámetros de la población a partir de la muestra?**

	Población	Muestra
Total	$Y = \sum_{i=1}^{N} Y_i$	$\hat{Y} = N\bar{y} = y$
Media	$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$	$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$
Proporción	$P = \frac{\sum_{i=1}^{N} A_i}{N}$	$\hat{P} = \frac{\sum_{i=1}^{n} a_i}{n} = p$

- Las Yi representan los valores de la variable medida en el i-ésimo elemento de la población (por ej. edad, frecuencia de consumo, opinión sobre un servicio, etc.).
- Las Ai son variables dicotómicas (0 ó 1), tomando el valor 1 en el caso de poseer la característica (por ej. género).

Muestreo aleatorio simple (MAS)

- Las estimaciones obtenidas a partir de la muestra están afectadas por el error estadístico o error muestral. Este error, en el caso del muestreo probabilístico es medible.

	Población	Muestra
Total	$S^2 = \frac{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}{N}$	$\hat{S}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}{n-1} = s^2$
Media	$S_{\overline{Y}}^2 = \frac{S^2}{N}$	$\hat{S}_{\overline{Y}}^2 = \frac{N-n}{N} * \frac{s^2}{n} = s_{\overline{y}}^2$
Proporción	$S_{\overline{P}}^2 = \frac{P*(1-P)}{N}$	$\hat{S}_{\overline{P}}^2 = \frac{N-n}{N} * \frac{p*(1-p)}{n-1} = s_{\overline{p}}^2$

Factor de corrección

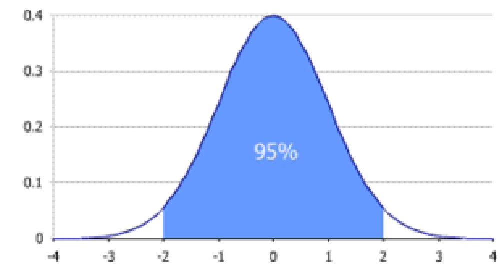
$$P = \frac{\sum_{i=1}^{N} A_i}{N}, \quad Q = 1 - P$$

Si n/N<5% se aproxima a población infinita y se eliminan los factores de corrección por tamaño de muestra.

Muestreo aleatorio simple (MAS)

- El grado de confianza de una estimación es la probabilidad de que el verdadero valor en la población se encuentre entre dos valores determinados, llamados límites del intervalo de confianza.

$$\text{Estimación} \pm \text{Error} = \text{E} \pm K * \hat{S}_E$$



- El valor de la población que deseamos estimar estará comprendido con una probabilidad o grado de confianza definido, entre los valores que resultan de sumar o restar el error estadístico al valor obtenido del estimador muestral.

- El coeficiente K depende del grado de confianza elegido.

Muestreo aleatorio simple (MAS)

INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

- En estricto rigor K sigue una distribución t-Student de n-1 grados de libertad en vez de una normal, pues la varianza poblacional es desconocida. Sin embargo, la distribución normal es una buena aproximación para muestras grandes (n>50).

Confianza	t(n=10)	t(n=25)	t(n=50)	t(n=75)	t(n=100)	t(n=1000)	N(0,1)
99.73%	4.09	3.34	3.16	3.10	3.08	3.01	3.00
99.00%	3.25	2.80	2.68	2.64	2.63	2.58	2.58
98.00%	2.82	2.49	2.40	2.38	2.36	2.33	2.33
96.00%	2.40	2.17	2.11	2.09	2.08	2.06	2.05
95.45%	2.32	2.11	2.05	2.03	2.03	2.00	2.00
95.00%	2.26	2.06	2.01	1.99	1.98	1.96	1.96
90.00%	1.83	1.71	1.68	1.67	1.66	1.65	1.64
80.00%	1.38	1.32	1.30	1.29	1.29	1.28	1.28
68.27%	1.06	1.02	1.01	1.01	1.01	1.00	1.00

Muestreo aleatorio simple (MAS)

- Para el caso de la **media**, tenemos:

$$\text{Error} = K * \hat{S}_{\bar{Y}} = K * \sqrt{\frac{N-n}{N} * \frac{s^2}{n}}$$

- Despejando n:

$$\boxed{n = \frac{NK^2s^2}{Ne^2 + K^2s^2}}$$

- Es decir, el tamaño de la muestra depende del K (grado de confianza), el error estadístico que estamos dispuestos a tolerar y de la variabilidad de los datos.

- Normalmente “s” es desconocida y lo que se hace es hacer una sobreestimación de él para asegurar que la muestra no tenga un error estadístico mayor al deseado (por ej. ponerse en el peor de los casos). Otra alternativa es estimarla a través de una muestra piloto.

- Si la población es infinita:

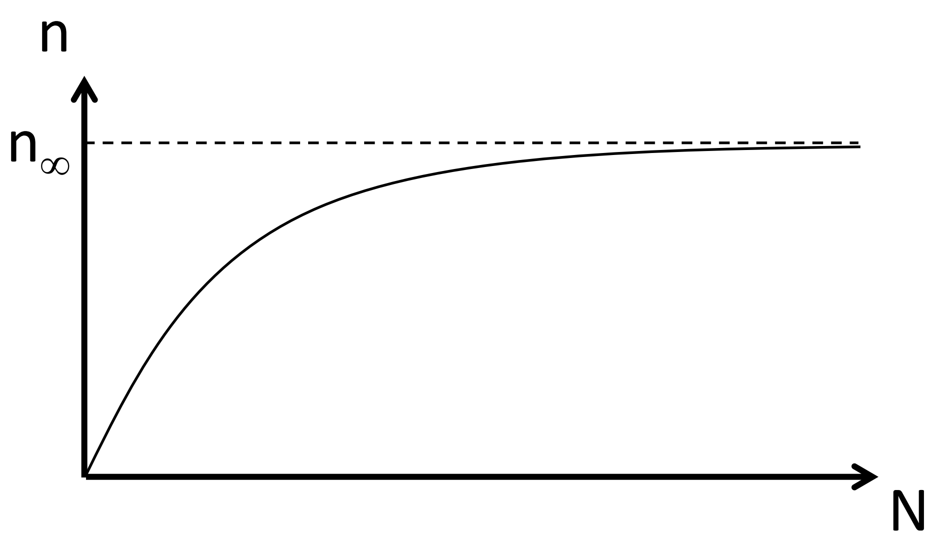
$$\boxed{n_\infty = \frac{K^2s^2}{e^2}}$$

Muestreo aleatorio simple (MAS)

- La muestra con población finita se relaciona de la siguiente forma con la de población infinita:

$$n = \frac{N * n_{\infty}}{N + n_{\infty}}$$

- Gráficamente:



n_{∞}: Es el tamaño de muestra para una población infinita

- Es decir, por mucho que aumente el tamaño de la población, el tamaño de muestra necesario para niveles dados de confianza, error estadístico y dispersión de los datos es prácticamente el mismo.

Muestreo aleatorio simple (MAS)

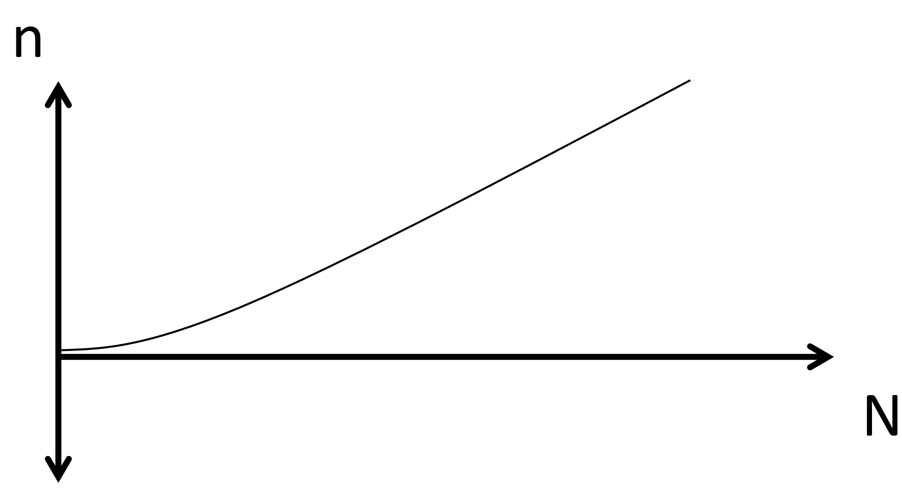
- Para el caso de estimaciones de **totales** tenemos:

$$\text{Error} = K * \hat{S}_Y = K * \sqrt{N(N-n)\frac{s^2}{n}}$$

- Despejando n:

$$\boxed{n = \frac{N^2 K^2 s^2}{e^2 + N K^2 s^2}}$$

- En este caso, el tamaño de la muestra aumenta al aumentar el N. Gráficamente:



Muestreo aleatorio simple (MAS)

- Para el caso de una **proporción** tendremos:

$$\text{Error} = K * \hat{S}_P = K * \sqrt{\frac{N-n}{N-1}\frac{pq}{n}}$$

- Despejando n:

$$\boxed{n = \frac{K^2 pqN}{e^2(N-1) + K^2 pq}}$$

- Si la población es infinita:

$$\boxed{n_\infty = \frac{K^2 pq}{e^2}}$$

- Cuando los valores de p y q no se conocen se recomienda tomar p = q = 0.5, pues eso considera la máxima varianza posible de los datos (el peor de los casos).

$$n = \frac{\frac{N}{N-1} n_\infty}{1 + \frac{1}{N-1} n_\infty} \cong \frac{N n_\infty}{N + n_\infty}$$

Es decir, aproximadamente igual a la ecuación vista para la media.

Muestreo aleatorio simple (MAS)

Ejemplo:

- Se tienen 200.000 facturas y se desea estimar la proporción de ellas que tienen errores.
- Se seleccionaron 56 facturas a través de un muestreo aleatorio simple y se encontraron 2 con errores (3.57%)
- ¿Cuál es el error de la estimación con un 95% de confianza?

$$\text{Error} = K * \hat{S}_P = 2 * \sqrt{\frac{200.000 - 56}{200.000 - 1} * \frac{0.0357 * 0.9643}{56}} = 0.0496$$

- Es decir, el verdadero valor en la población está con un 95% de probabilidad entre:

$$\mathrm{L}_{\text{inferior}} = \overline{p} - Error = 0.0357 - 0.0496 = -0.0139 = -1.4\%$$
$$\mathrm{L}_{\text{superior}} = \overline{p} + Error = 0.0357 + 0.0496 = 0.0853 = 8.5\%$$

- Por lo tanto, dado los resultados de la muestra y su error no se puede concluir si existen o no errores significativos en las facturas (por ej. si el límite fijado fuera de 5%).

Muestreo aleatorio simple (MAS)

- ¿Cuál debería ser el tamaño de la muestra para que el error no sea más de un punto porcentual?

$$n = \frac{K^2 pqN}{e^2(N-1)+K^2 pq} = \frac{2^2 * 0.0357 * 0.9643 * 200000}{0.01^2 * (200000-1) + 2^2 * 0.0357 * 0.9643} = 1367{,}6 \cong 1368$$

Muestreo Aleatorio Estratificado

- Se usa si el conocimiento de la población permite agrupar previamente sus elementos en subconjuntos o estratos de forma que sean más homogéneos.
- La obtención de una muestra con elementos de cada uno de ellos conlleva una mayor precisión de las estimaciones que el muestreo aleatorio simple.
- Si la selección de la muestra en cada estrato se hace mediante un muestreo aleatorio simple, se llama muestreo aleatorio estratificado.
- En un caso extremo, si cada estrato fuera de forma tal que cada uno tuviera elementos iguales dentro de él para la variable a estudiar, bastaría un elemento de cada estrato para tener una muestra representativa.

Muestreo Aleatorio Estratificado

- No hay reglas definidas sobre el número de estratos a utilizar. En general, se puede decir que al aumentar el número de estratos aumenta la precisión.

- Por otro lado, un número elevado de estratos complica los cálculos y puede que su aporte a reducir la muestra sea mínimo.

- Para definir bien los estratos es necesario utilizar alguna variable conocida de la población que esté correlacionada con aquella que queremos investigar.

Muestreo Aleatorio Estratificado

- Si representamos las poblaciones de L estratos por $N_1, N_2, \ldots, N_L$:

$$\sum_{h=1}^{L} N_h = N \qquad W_h = \frac{N_h}{N} \qquad \sum_{h=1}^{L} W_h = 1$$

  – W_h es el peso relativo del estrato h en la población.

- Se llama **afijación** al reparto que se hace del tamaño de muestra *n* entre los diferentes estratos.
- Existen tres tipos de afijación:
  – **Afijación igual**: se le asigna a cada estrato el mismo número de muestra.

$$n_h = \frac{n}{L}$$

  – **Afijación proporcional**: a cada estrato se le asigna una fracción de la muestra que es igual a la proporción que tiene el estrato en la población.

$$n_h = n * \frac{N_h}{N}$$

Muestreo Aleatorio Estratificado

- **Afijación óptima:** se asigna a cada estrato un número de muestra proporcional al producto de su población por la desviación estándar del estrato en la muestra.

$$n_h = n * \frac{N_h * s_h}{\sum_{l=1}^{L} N_l * s_l}$$

- Este tipo de afijación es la que entrega una mejor precisión para un mismo tamaño de muestra, pues toma en cuenta el tamaño del estrato y la variabilidad de los datos a su interior.

Muestreo Aleatorio Estratificado

- ¿Cómo estimar parámetros de la población a partir de una muestra estratificada?

	Población	Muestra
Total	$Y = \sum_{i=1}^{N} Y_i$	$\hat{Y}_e = N\overline{y_e} = \sum_{l=1}^{L} N_l \overline{y_l}$
Media	$\overline{Y} = \dfrac{\sum_{i=1}^{N} Y_i}{N}$	$\hat{\overline{Y}}_e = \dfrac{\sum_{l=1}^{L} N_l * \overline{y}_l}{N} = \sum_{l=1}^{L} W_l * \overline{y}_l = \overline{y}_e$
Proporción	$P = \dfrac{\sum_{i=1}^{N} A_i}{N}$	$\hat{P}_e = \dfrac{\sum_{l=1}^{n} N_l * p_l}{N} = \sum_{l=1}^{n} W_l * p_l = p_e$

- ,

Muestreo Aleatorio Estratificado

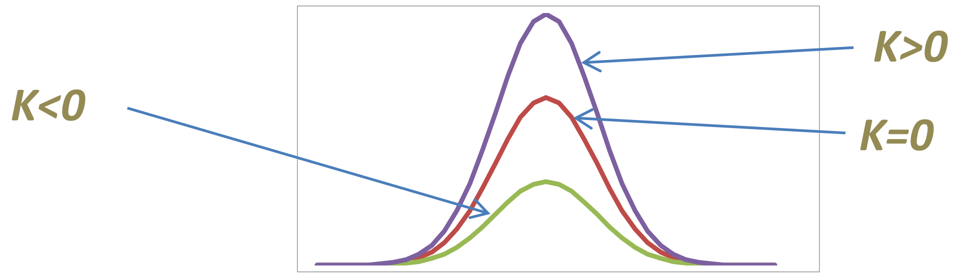
- ¿Cómo estimar la variabilidad de los datos a partir de una muestra estratificada?

	Población	Muestra
Total	$S^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N}$	$\hat{S}^2_{Y_e} = \sum_{l=1}^{L} N_l (N_l - n_l) \frac{s_l^2}{n_l} = s^2_{y_e}$
Media	$S^2_{\bar{Y}} = \frac{S^2}{N}$	$\hat{S}^2_{\bar{Y}_e} = \frac{1}{N^2} \sum_{l=1}^{L} N_l (N_l - n_l) \frac{s_l^2}{n_l} = s^2_{\bar{y}_e}$
Proporción	$S^2_{\bar{P}} = \frac{P*(1-P)}{N}$	$\hat{S}^2_{\bar{P}_e} = \frac{1}{N^2} \sum_{l=1}^{L} N_l^2 \frac{N_l - n_l}{N_l - 1} * \frac{p_l*(1-p_l)}{n_l} = s^2_{\bar{p}_e}$

TEST DE HIPÓTESIS

Pero primero… Kurtosis

- Es una medición de qué tan concentrado o aplanado está un conjunto de datos en comparación con la distribución normal.



- Un índice positivo indica una distribución más concentrada. En tanto que uno negativo indica una distribución relativamente plana.

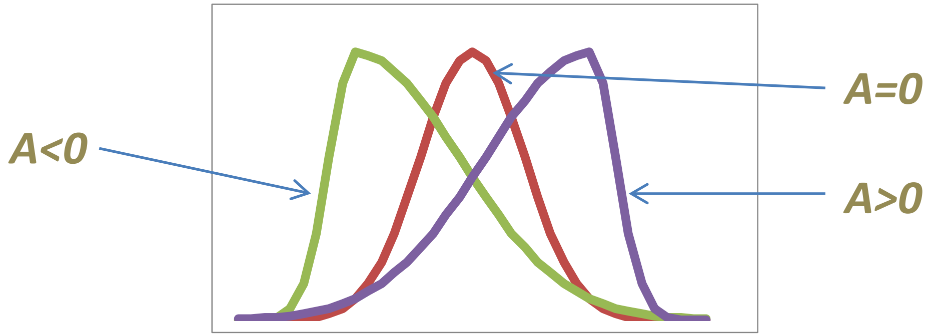
$$K = \frac{\sum_{i=1}^{N}\left(\frac{Y_i - \bar{Y}}{S}\right)^4}{N} - 3$$

- Donde S es la desviación estándar de la población. En el caso de estar trabajando con una muestra y desconocer S, entonces el estimador insesgado de K es:

$$\hat{K} = k = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s}\right)^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Pero primero… Skewness (asimetría)

- Es una medición del grado de asimetría de una distribución alrededor de su media.



- Un índice positivo indica una distribución asimétrica hacia la derecha. En tanto que uno negativo indica asimetría hacia la izquierda.

$$A = \frac{\sum_{i=1}^{N} \left(\frac{Y_i - \bar{Y}}{S} \right)^3}{N}$$

- Donde S es la desviación estándar de la población. En el caso de estar trabajando con una muestra y desconocer S, entonces el estimador insesgado de A es:

$$\widehat{A} = a = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{s} \right)^3$$

Test de hipótesis

- Se usan cuando se requiere probar o rechazar una idea prestablecida.
- **Hipótesis:** presunción de sobre una característica de la población.

- Ej. Un gerente de una empresa de telefonía estima que en promedio transcurren 7 minutos desde que una persona saca un número hasta que la atienden en caja. Se ha tomado una muestra de 36 personas y se obtuvo que el tiempo promedio transcurrido fue de 8 minutos, con una desviación estándar de la muestra de 2 minutos. ¿Debemos aceptar o rechazar la hipótesis del gerente?

- Es decir, ¿es 8 lo suficientemente lejano de 7 como para descartar la estimación del gerente? O al contrario, ¿es lo suficientemente cercana como para validarla?

Test de hipótesis

- Existen dos explicaciones para la diferencia entre el valor hipotético y el de la muestra:

  a) La hipótesis es cierta y la diferencia observada se debe al error estadístico.
  b) La hipótesis es falsa y el valor real será algún otro valor.

- Un test de hipótesis ayuda a determinar cuál es la explicación más probable.

Pasos de un test de hipótesis

Ejemplo:

- **Hipótesis nula:** Lo que se quiere probar

$$H_0 : \bar{Y} = 7\ días$$

- **Hipótesis alternativa:** Compite con la H_0, puede ser unidireccional o bidireccional.

$$H_1 : \bar{Y} \neq 7\ días$$

$$t = \frac{\bar{y} - \bar{Y}}{s / \sqrt{n}} = \frac{8-7}{2/\sqrt{36}} = 3$$

Pasos de un test de hipótesis

- La regla de decisión puede ser, por ejemplo:
  - α=5%
  - n=36
  - $t_{1-\alpha/2,n-1}$=2,03

- Como $t>t_c$, entonces se rechaza con un 95% de confianza que una muestra de 36 elementos con media de 8 días y desviación estándar de 2 provenga de una población con media 7 días.

- La verificación de hipótesis está sujeta a dos tipos de error (I y II).
  - **Tipo I:** cuando la hipótesis nula es verdadera e incorrectamente la rechazamos
  - **Tipo II:** cuando la hipótesis nula es falsa e incorrectamente no la rechazamos

Test de hipótesis para proporciones de una muestra

- **Definición de Hipótesis:**
  - Hipótesis Nula o Indiferente: Lo que se quiere probar

$$H_0 : P = P^*$$

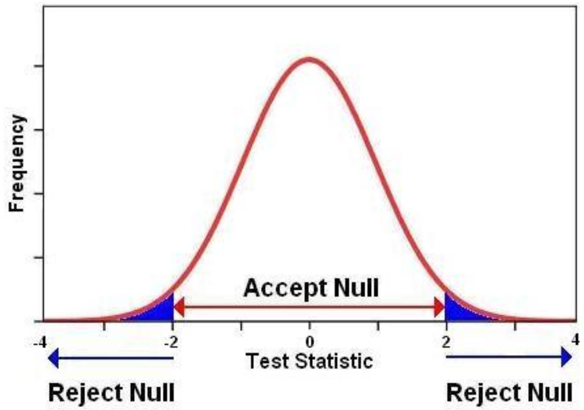
  - Hipótesis Alternativa: Compite con la H0, puede ser bidireccional o unidireccional.

$$H_1 : P \neq P^*$$

- **Método estadístico:**

$$Z = \frac{p - P^*}{\sqrt{\frac{P^*(1-P^*)}{n}}} \qquad Z_c = Z_{1-\alpha/2}$$

- **Regla de decisión:**
  - Rechazar H_0 si $Z>Z_c$ ó $Z<-Z_c$



Ejemplo:

- Se realizaron 300 entrevistas para evaluar un nuevo producto y 74 de ellos declararon una intención futura de compra. La empresa ha establecido como su criterio de decisión que lanzará el producto si la proporción que está dispuesto a comprarlo es igual o superior al 19.5%. ¿Debería lanzar el producto?

$$H_0: P \leq 0.195 \qquad H_1: P > 0.195 \qquad p = \frac{74}{300} = 0.247$$

$$Z = \frac{p - P^*}{\sqrt{P^*(1-P^*)/n}} = \frac{0.247 - 0.195}{\sqrt{0.195 * 0.805 / 300}} = 2.27 \qquad Z_c = Z_{1-\alpha} = Z_{0.95} = 1.64$$

Test unidireccional

- Como Z>Zc, entonces se rechaza la hipótesis nula y se debería lanzar el producto.

Test de hipótesis para proporciones de dos muestras independientes

- **Definición de Hipótesis:**
  - Hipótesis Nula o Indiferente: Lo que se quiere probar

$$H_0 : P_1 = P_2$$

  - Hipótesis Alternativa: Compite con la H0, puede ser direccional o unidireccional.

$$H_1 : P_1 \neq P_2$$

- **Método estadístico:**

$$Z = \frac{p_1 - p_2}{s_{\bar{p}_1 - \bar{p}_2}} \qquad s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\left[p^*(1 - p^*)\right]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \qquad p^* = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \qquad Z_c = Z_{1-\alpha/2}$$

- **Regla de decisión:**
  - Rechazar H_0 si $Z>Z_c$ ó $Z<-Z_c$

Test de hipótesis

Ejemplo:

- Suponga que las proporciones de disposición positiva a la compra del nuevo producto en hombres y mujeres son de 28% y 21% respectivamente. En la muestra hay 150 hombres y 150 mujeres. ¿Hay diferencia entre hombres y mujeres en la intención de compra del producto?

$$H_0 : P_1 = P_2 \qquad H_1 : P_1 \neq P_2$$

$$p^* = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{150 * 0.28 + 150 * 0.21}{300} = 0.247$$

$$Z = \frac{p_1 - p_2}{s_{\bar{p}_1 - \bar{p}_2}} = \frac{0.28 - 0.21}{\sqrt{0.247 * 0.753 * \left[\frac{1}{150} + \frac{1}{150}\right]}} = 1.41$$

$$Z_c = Z_{1-\alpha/2} = Z_{0.975} = 1.96$$

- Como $Z<Z_c$, entonces no se puede rechazar H_0.

Test de hipótesis para proporciones para más de dos muestras independientes

- Definición de Hipótesis
  - Hipótesis Nula o Indiferente: Lo que se quiere probar
  - $$H_0: P_1 = P_T, P_2 = P_T, \ldots, P_k = P_T$$
  - Hipótesis Alternativa: Compite con la H0, puede ser bidireccional o unidireccional.

$$H_1: \exists i / P_i \neq P_T$$

- Método estadístico
  - Consideremos k muestras de tamaño n_k, cada una de las cuales tiene x_k éxitos y $(n_k - x_k)$ fracasos. Sea:

$$x_T = x_1 + \cdots + x_k \qquad n_T = n_1 + \cdots + n_k$$

$$\chi_c^2 = \chi^2_{1-\alpha/2, k-1}$$

- Regla de decisión
  - Rechazar H_0 si $\chi^2 > \chi_c^2$ ó $\chi^2 < -\chi_c^2$

Test de hipótesis

- Entonces:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(o_{ij}-e_{ij}\right)^2}{e_{ij}} \qquad \chi_c^2 = \chi^2_{1-\alpha/2,\,k-1}$$

- Donde:
  - o_{ij}=frecuencia observada en la muestra j (i=1 → éxito, i=2 → fracaso).
  - e_{ij}=frecuencia esperada en la muestra j (i=1 → éxito, i=2 → fracaso).

$$e_{1j} = n_j \cdot P_T = n_j \cdot \frac{x_T}{n_T} \qquad e_{2j} = n_j \cdot \left(1-P_T\right) = n_j \cdot \frac{\left(n_T - x_T\right)}{n_T}$$

- Regla de decisión
  - Rechazar H_0 si $\chi^2 > \chi_c^2$ ó $\chi^2 < -\chi_c^2$

Test de hipótesis

INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

Ejemplo:

- Si 100 entrevistados eran entre 18-34 años, otros 100 entre 35-54 años y otros 100 de más de 55 años, y que los porcentajes de intención de compra de cada grupo fueron 5%, 21% y 48% respectivamente. ¿Podemos concluir que hay diferencias por edad?

$$H_0 : P_1 = P_2 = P_3 = 0.247$$

$$H_1 : \exists\, i \,/\, P_i \neq 0.247$$

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{(5-24.7)^2}{24.7} + \frac{(21-24.7)^2}{24.7} + \frac{(48-24.7)^2}{24.7} + \frac{(95-75.3)^2}{75.3} + \frac{(79-75.3)^2}{75.3} + \frac{(52-75.3)^2}{75.3} = 50.79$$

$$\chi_c^2 = \chi^2_{1-\alpha/2,\,k-1} = \chi^2_{0.975,2} = 5.99$$

- Como $\chi^2 > \chi_c^2$ se rechaza H0 → Sí, hay diferencias significativas por edad.

Test de hipótesis para la media de una muestra

- **Definición de Hipótesis:**
  - Hipótesis Nula o Indiferente: Lo que se quiere probar.

$$H_0 : \bar{Y} = Y^*$$

  - Hipótesis Alternativa: Compite con la H0, puede ser bidireccional o unidireccional.

$$H_1 : \bar{Y} \neq Y^*$$

- **Método estadístico:**
  - Si n>30 se puede aproximar por la distribución normal.

$$t = \frac{\bar{y} - Y^*}{s / \sqrt{n}} \qquad t_c = t_{1-\alpha/2,\, n-1}$$

- **Regla de decisión:**
  - Rechazar H_0 si $t > t_c$ ó $t < -t_c$

Test de hipótesis

Ejemplo:

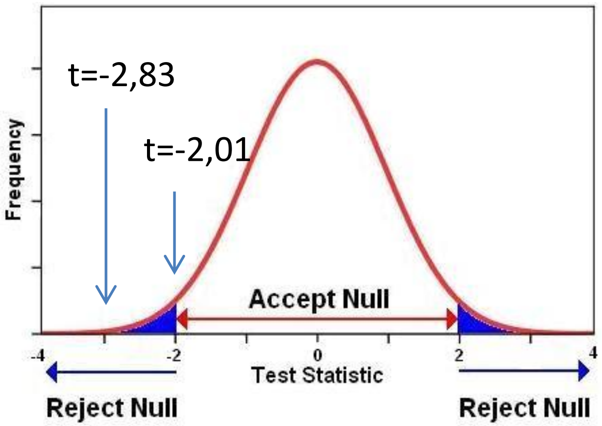
- A partir de una muestra aleatoria de 50 automóviles de un mismo modelo se midió el rendimiento en kilómetros por litro. El rendimiento promedio es de 11.4 km/lt. y la desviación estándar de la muestra es de 1.5. Si las especificaciones técnicas del rendimiento son 12 km/lt., ¿se rechaza o no la hipótesis de que las especificaciones son verdaderas?

$$H_0 : \bar{Y} = 12$$

$$H_1 : \bar{\mathrm{Y}} \neq 12$$

$$t = \frac{\bar{y} - Y^*}{s/\sqrt{n}} = \frac{11.4 - 12}{1.5/\sqrt{50}} = -2.83$$

$$t_c = t_{97.5\%,49} = 2.01$$



- Se rechaza H_0 pues $t < -t_c$

Test de hipótesis para las medias de dos muestras

- **Definición de Hipótesis**
  - Hipótesis Nula o Indiferente: Lo que se quiere probar.

$$H_0 : \bar{Y}_1 = \bar{Y}_2$$

  - Hipótesis Alternativa: Compite con la H0, puede ser bidireccional o unidireccional.

$$H_1 : \bar{Y}_1 \neq \bar{Y}_2$$

- **Método estadístico**

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_{\bar{y}_1 - \bar{y}_2}} \qquad s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad t_c = t_{1-\alpha/2, n_1+n_2-2}$$

  - Si n>30 se puede aproximar por la distribución normal.

- **Regla de decisión**
  - Rechazar H_0 si $t > t_c$ ó $t < -t_c$

Test de hipótesis

Ejemplo:

- Dos muestras aleatorias de 50 y 75 automóviles, cada una de un modelo diferente. Se midió el rendimiento en kilómetros por litro en cada grupo. El rendimiento promedio para el primer grupo fue de 11.4 km/lt y de 10.9 km/lt para el segundo. Las desviaciones estándar de cada muestra fueron de 1.5 y 2. ¿Se rechaza o no la hipótesis que ambos modelos tienen el mismo rendimiento?

$$H_0 : \overline{Y}_1 = \overline{Y}_2 \qquad H_1 : \overline{Y}_1 \neq \overline{Y}_2$$

$$s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{1.5^2}{50} + \frac{2^2}{75}} = 0.314 \qquad t = \frac{11.4 - 10.9}{0.314} = 1.59$$

$$t_c = t_{97.5\%,123} = 1.98$$

- Se acepta H_0 pues $t < t_c$

Test de hipótesis para la varianza de dos muestras

- Definición de Hipótesis
  - Hipótesis Nula o Indiferente: Lo que se quiere probar.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

  - Hipótesis Alternativa: Compite con la H0, puede ser bidireccional o unidireccional.

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

- Método estadístico

$$F = \frac{\sigma_i^2}{\sigma_j^2}, \text{con } \sigma_i^2 > \sigma_j^2 \qquad F_c = F_{1-\alpha, n_1-1, n_2-2}$$

- Regla de decisión
  - Rechazar H_0 si $F > F_c$

Test de hipótesis

Ejemplo:

- Se desea saber si hay diferencia entre hombres y mujeres fumadores en la varianza del consumo de cigarrillos. Para ello se tomó una muestra de 149 hombres y 139 mujeres, con consumos promedio de 24.21 y 24.92 cigarrillos por semana y desviaciones estándar de 4.9 y 4.79 respectivamente.

$$H_0 : \sigma_m^2 = \sigma_f^2 \qquad H_1 : \sigma_m^2 \neq \sigma_f^2$$

$$F = \frac{4.9^2}{4.79^2} = 1.05 \qquad F_c = F_{95\%,148,138} = 1.32$$

- Se acepta H_0 pues $F < F_c$

Muestras relacionadas

- Cuando medimos en una misma muestra en distintos momentos del tiempo tendremos dos muestras que son dependientes.
- Normalmente entre ambas mediciones ha ocurrido un evento que afecta la medición (ej. publicidad).

Test de hipótesis para la media entre dos muestras relacionadas

- **Definición de Hipótesis**

$$H_0 : \bar{Y}_1 = \bar{Y}_2 \qquad H_1 : \bar{Y}_1 \neq \bar{Y}_2$$

- **Método estadístico**

$$t = \frac{\sum_{i=1}^{n} d_i \Big/ n}{s_d \Big/ \sqrt{n}} \qquad d_i = y_{i2} - y_{i1} \qquad t_c = t_{1-\alpha/2, n-1}$$

- **Regla de decisión**
  - Rechazar H_0 si $t > t_c$ ó $t < -t_c$

Test de hipótesis

Ejemplo:

- Se midió la actitud positiva de compra antes y después de la publicidad en una muestra de 10 personas. ¿Qué se puede concluir de los resultados del experimento?

n	Actitud Previa (yi1)	Actitud Posterior (yi2)	di
1	50	53	3
2	25	27	2
3	30	38	8
4	50	55	5
5	60	61	1
6	80	85	5
7	45	45	0
8	30	31	1
9	65	72	7
10	70	78	8
Promedio	50.5	54.5	4
desv. Est.	18.48	19.73	3.02

$$t = \frac{4}{3.02/\sqrt{10}} = 4.19 \qquad t_c = t_{97.5\%,9} = 2.26$$

- Como t>tc entonces se rechaza H0. Es decir, la publicidad tiene un efecto estadísticamente significativo en la actitud de compra.

Test de hipótesis

Ejemplo:

- ¿Y si hubiésemos elegido el test para medias con muestras independientes?

$$s_{\bar{y}_1-\bar{y}_2} = \sqrt{\frac{18.48^2}{10} + \frac{19.73^2}{10}} = 8.55 \qquad t = \frac{\bar{y}_1 - \bar{y}_2}{s_{\bar{y}_1-\bar{y}_2}} = \frac{50.5 - 54.5}{8.55} = -0.468 \qquad t_c = t_{97.5\%,18} = 2.10$$

- En este caso hubiéramos aceptado H_0 y concluido que no hay diferencia en la conducta de presentar una actitud positiva hacia la compra antes y después de la publicidad

Test de hipótesis para las proporciones entre dos muestras relacionadas

- **Definición de Hipótesis**

$$H_0 : P_1 \geq P_2 \qquad H_1 : P_1 < P_2$$

- **Método estadístico**

$$z = \frac{p_2 - p_1}{\left[\dfrac{b + c - \left\{(b-c)^2 / n\right\}}{n(n-1)}\right]^{1/2}}$$

		Antes		
		Sí	**No**	**Total**
Después	**Sí**	a	b	a+b
	No	c	d	c+d
	Total	a+c	b+d	a+b+c+d

$$Z_c = Z_{1-\alpha}$$

- **Regla de decisión**
  - Rechazar H_0 si $Z > Z_c$

Ejemplo:

- Se midió la disposición a la compra antes y después de la publicidad en una muestra de 100 personas. ¿Qué se puede concluir de los resultados del experimento?

		ANTES		
		SÍ	**NO**	**TOTAL**
DESPUÉS	**SÍ**	23	7	30
	NO	2	68	70
	TOTAL	25	75	100

$$z = \frac{0.30 - 0.25}{\left[\dfrac{7 + 2 - \left\{(7-2)^2 / 100\right\}}{100 * 99}\right]^{1/2}} = 1.68$$

$$z_c = z_{95\%} = 1.64$$

- Como Z>Zc entonces se rechaza H0.