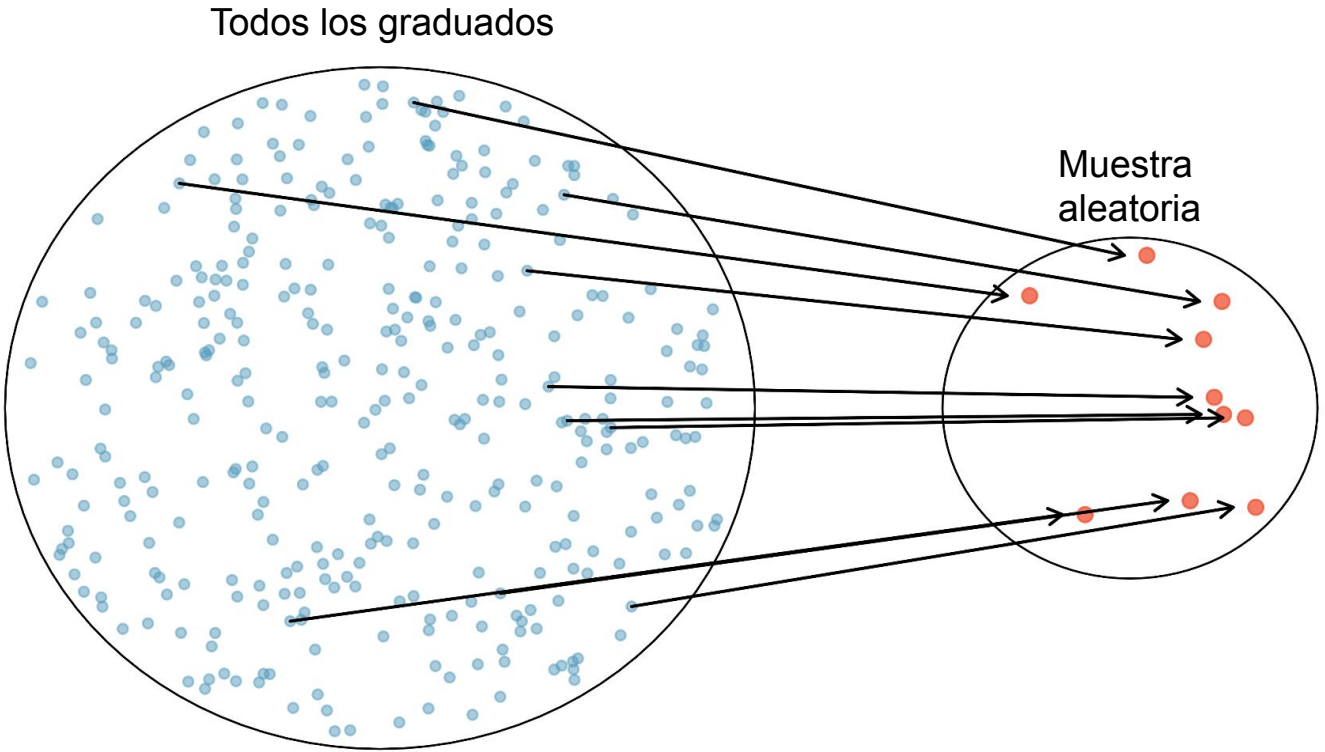
Procesamiento de datos en R y estadística para Ciencias Sociales

Clase 8. Introducción a la estadística inferencial

factor~data
IDAES_UNSAM

Población y muestra



Inferencia

- Al trabajar con muestras => error muestral
- La inferencia supone poder

  estimar un **parámetro** (característica de la población) a partir de una **estimación** (característica de la muestra) calculada a partir de un **estimador**

Inferencia

¿Cuál era el % de trabajadores en la población ocupada en las zonas urbanas de Argentina?

Parámetro => CENSO

Estimación => ENES / EPH / Etc.

Dos tipos de **estimación:**

Puntual

Por Intervalos de Confianza

Inferencia

En ambos casos, se usa un estimador:

- expresión matemática,
- se construye con los valores de una muestra
- sirven para obtener estimaciones del parámetro
- pueden existir una gran variedad de fórmulas diferentes

Parámetros		Estimadores	
Promedio Poblacional	$\mu = \frac{\Sigma x_i}{N}$	Promedio Muestral	$\bar{x} = \frac{\Sigma x_i}{n}$
Total Poblacional	$X = \Sigma x_i = N.\mu$	Total Muestral	$N.\bar{x}$
Proporción Poblacional	$P = \frac{\Sigma x_i}{N} = \frac{N^*}{N} \; ; x_i = 0;1$	Proporción Muestral	$\hat{P} = p = \frac{\Sigma x_i}{n} = \frac{n^*}{n} \; ; x_i = 0;1$
Cantidad de Casos Favorables Poblacional	$N^* = \Sigma x_i \quad ; x_i = 0;1$	Cantidad de Casos Favorable Muestral	$\hat{N}.p \; ; x_i = 0;1$
Variancia Poblacional	$\sigma_x^2 = \frac{\Sigma (x_i - \mu)^2}{N}$	Variancia Muestral	$S_x^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$

Inferencia

¿Cuál era el % de mujeres entre les trabajadores en las zonas urbanas de Argentina?

Parámetro => CENSO => En general, imposible
Estimación => ENES / EPH / Etc.
Estimador => fórmula para hacer la estimación

Un estimador puede ser considerado una variable “aleatoria”... ¿por qué?

Distribución muestral

Completar esta hoja de cálculo

https://docs.google.com/spreadsheets/d/14YHOQfrteEB0Qjg_8TwLuy8DLoL6J4SCD_B7ZID9oCQ/edit?usp=sharing

Distribución muestral

Sin entrar en detalles teóricos que no son necesarios a los fines de este curso, podemos decir que existen diferentes elementos que influyen en la forma de las diferentes distribuciones muestrales. Tres son, probablemente los más importantes:

- La forma de la distribución de la variable o población original.
- El conocimiento previo (o no) de la dispersión de la población.
- El tamaño de la muestra (n).

Distribución muestral

Entonces, la distribución muestral del estimador expresa todos los resultados posibles de un estimador (por ejemplo, la media) en todas las muestras posibles extraídas de una población.

La distribución muestral es un concepto teórico que nos permite formalizar esa noción de error muestral.

En la distribución muestral de un estimador se encuentra mucha información necesaria para aproximarnos a los errores del estimador.

Distribución muestral

Hay (como mínimo) dos problemas

Distribución muestral

Hay (como mínimo) dos problemas

- En general, no tenemos información de la población (si no, no haríamos la encuesta)
- Aún si la tuviéramos sería imposible obtener todas las muestras de una población. De hecho, casi siempre trabajamos con UNA SOLA MUESTRA

Distribución muestral

¿Entonces? Hay varias opciones para proceder.
Vamos a mencionar dos

- Hacer supuestos sobre las distribuciones muestrales (por ejemplo, asumir que bajo ciertas condiciones la distribución muestral de medias es aproximadamente normal)
- Tratar de estimar las distribuciones muestrales a partir de la muestra (usando técnicas de remuestreo)

¿Para qué sirve una distribución muestral?

Distribución muestral

¿Para qué sirve una distribución muestral? Bueno, para casi todo en estadística inferencial.

Si la conocemos (ya sea porque la asumimos o porque la estimamos), entonces podemos calcular errores, intervalos de confianza, hacer pruebas de hipótesis, etc.

Distribución muestral

Importante: hasta aquí solo vimos estadísticos simples (medias o proporciones) pero cualquier cosa que calculemos con una muestra tiene una distribución muestral

- una medida de dispersión
- un coeficiente de regresión
- una medida de asociación
- etc…

Distribución muestral

Intervalo de confianza

Metodología

- Estos son hallazgos de la encuesta de Satisfacción Política y Opinión Pública de la Universidad de San Andrés. En total fueron realizadas 1025 entrevistas entre el 15 y el 23 de Marzo de 2022 a adultos de 18 años en adelante, conectados a internet, en Argentina.
- La muestra es estratificada por región con cuotas de NSE y Edad. Se realiza en 8 regiones: NOA, NEA, Cuyo, Centro, Sur, y Buenos Aires dividida a su vez en CABA, GBA e interior de la Provincia de Buenos Aires. Los datos se ponderan de acuerdo a parámetros de NSE y Edad.
- La encuesta versa sobre satisfacción con la marcha general de las cosas, el desempeño de los poderes políticos y las políticas públicas, la aprobación del gobierno y sobre la opinión respecto de los principales líderes políticos nacionales. También indaga sobre tendencias sociales y temas de coyuntura política.
- En algunos casos los resultados de la suma de los porcentajes de respuesta no suman 100, ello puede deberse a redondeos computacionales, respuestas múltiples o la exclusión de los que no saben y no contestan. En algunos resultados se indica si fueron ponderados siguiendo algún criterio diferente al resto.
- Todas las encuestas están sujetas a otras fuentes de error adicionales a los errores muestrales, tales como errores de cobertura o medición. La precisión de esta encuesta online se estima mediante el cálculo de intervalos de credibilidad bayesianos. Usando una aproximación simple del posterior en una distribución normal, el intervalo de credibilidad del 95% está dado por, aproximadamente: $\hat{y} + \frac{1}{\sqrt{n}}$, de lo que se deduce que para una encuesta de 1000 casos es de aproximadamente +/− 3.15 puntos porcentuales.

FICHA TÉCNICA

celag.

- Universo de estudio: Población mayor de 16 años en todo el territorio nacional habilitada para votar en su lugar de residencia.
- Tipo de estudio: Cuantitativo, realizado a través de encuestas telefónicas mediante sistema CATI (computer aided telephone interviewing), realizadas por encuestadores profesionales. Llamadas realizadas a teléfonos fijos y celulares.
- Diseño muestral: Probabilístico aleatorio simple, representativo del universo de la población del país, basado en 2 muestras independientes de teléfonos fijos y celulares, con ajuste final en la selección del encuestado por cuotas de sexo y edad. El marco muestral de teléfonos fijos procede de la guía telefónica. Para detectar los números celulares se utilizó un algoritmo de rangos telefónicos existentes.
- Ponderación de resultados: Se diseñó una muestra no autoponderada, para asegurar cantidad de casos mínima por región en el desagregado de los datos. Para el análisis de los totales, los datos fueron ponderados para asignarle a cada distrito un peso proporcional al del parámetro poblacional.
- Tamaño de la muestra: 2.000 casos totales efectivos.
- Control de calidad: Todas las etapas de la investigación se desarrollan cumpliendo estándares internacionales de calidad para la investigación de opinión pública. Revisión del 100% de los cuestionarios, supervisión in situ del 20% y supervisión telefónica por terceros del 10% de cada encuestador.
- Margen de error: El margen de error para el total de la muestra oscila entre +/- 0,9% y +/- 2,2%, de acuerdo a la dispersión de la distribución, con un 95% de intervalo de confianza.
- Fecha de trabajo de campo: 3 de junio al 12 de junio de 2019
- Equipo CELAG: Alfredo Serrano Mancilla / Gisela Brito / Sergio Pascual

FICHA TÉCNICA

Fecha de realización
JUNIO 2015 (CABA, Córdoba, Santa Fe, Mendoza, Tucumán, Corrientes, Entre Ríos, Neuquén, Río Negro).
JULIO 2015 (Provincia de Buenos Aires).

Tipo de muestreo
Ajustado por cuotas de sexo, edad, sección electoral, y provincia.

Tamaño de la muestra
1500 CASOS.

Modalidad
Cuestionario estructurado con preguntas abiertas y cerradas.

Sistema de consulta
Domiciliaria bajo sobre cerrado.

Error muestral
+/- 2,58.

factor~data
IDAES_UNSAM

Intervalo de confianza

En términos muy generales, es un rango de estimaciones para un determinado parámetro desconocido.

Puede tomar la forma de afirmaciones tales como “la proporción de trabajadores entre la población ocupada en Argentina está entre un 0.6 y un 0.617 con un nivel de confianza del 95%”.

¿De dónde puede salir esto?

Intervalo de confianza

En términos muy generales, es un rango de estimaciones para un determinado parámetro desconocido.

Puede tomar la forma de afirmaciones tales como “la proporción de **trabajadores** entre la **población ocupada** en Argentina está entre un 0.6 y un 0.617 con un nivel de confianza del 95%”.

trabajadores => remite a una variable (clase social según EOW)
población ocupada => remite a una población

Intervalo de confianza

En términos muy generales, es un rango de estimaciones para un determinado parámetro desconocido.

Puede tomar la forma de afirmaciones tales como “la **proporción** de trabajadores entre la población ocupada en Argentina está entre un **0.6 y un 0.617** con un **nivel de confianza del 95%**”.

proporción => remite a un estimador
0.6 y un 0.617 => remite a un rango de estimaciones
nivel de confianza del 95% => remite a una región de la distribución muestral

Prueba de hipótesis

- Una prueba de hipótesis nos permite formular afirmaciones sobre la población y testearlas.
- La teoría que sustenta los tests/pruebas/ensayos de hipótesis incorpora elementos de la estadística descriptiva, de la teoría de probabilidades y de la teoría de la decisión.
- “Probar”, entonces, involucra ya sea a un parámetro o a alguna forma funcional no conocida de la distribución a partir de la cual se obtiene una muestra aleatoria.

Prueba de hipótesis

- La decisión acerca de si los datos muestrales apoyan estadísticamente la afirmación, se toma en base a probabilidades, y como se verá, si ésta es pequeña será rechazada la hipótesis.
- Comparación con un proceso judicial:
  - Supuesto: el acusado es inocente.
  - Se realiza el juicio => evidencias para probar la culpabilidad.
  - Si los testimonios y pruebas recogidos no permiten rebatir el supuesto original, el acusado permanecerá inocente, y en caso contrario (es decir si los testimonios y pruebas lo condenan), se lo declarará culpable.

Prueba de hipótesis

Al realizarse un Ensayo de Hipótesis sucede lo mismo:

- conocimiento previo de la población o los supuestos que se realicen sobre ella = Hipótesis básica (o nula) sobre algunos de sus parámetros (= supuesto de “inocencia” en el Juicio)
- A partir de una muestra se obtienen los “testimonios o pruebas”, y los resultados de la misma determinarán si los supuestos previos son rechazados o no, es decir “si el jurado declara al acusado inocente o culpable”.

Prueba de hipótesis

	Hombre	Mujer
Trabajadores	4116	3444
No trabajadores	3036	1818
Total	7152	5262

¿Cuál es la proporción de hombres que son trabajadores?
¿Y de mujeres?

Prueba de hipótesis

	Hombre	Mujer
Trabajadores	57.6%	65.5%

- Más precisamente, ¿cuál es la probabilidad de que las diferencias observadas entre esos valores provengan de dos subpoblaciones (hombres y mujeres) con parámetros diferentes? ¿Las diferencias observadas son producto del azar o de que las dos subpoblaciones son efectivamente diferentes?

Elementos de una prueba de hipótesis

- Lo primero es lo primero. Es necesario formular “hipótesis” sobre la población que deben ser testeadas.
- Sin embargo, en este tipo de técnicas, en realidad, lo que se hace es formular dos hipótesis:
  - Hipótesis nula (Ho): basada en los conocimientos previos de la población que se desean comprobar. En general, la Ho es la hipótesis que se desea “refutar”. Por ello, suele formularse de forma “contrafáctica”.
  - Hipótesis alternativa (Ha): es la que se tomará como (probablemente) cierta en caso de que a partir de los datos de la muestra se derive en el rechazo de la Ho
  -

Elementos de una prueba de hipótesis

- Las hipótesis se realizan sobre alguno de los parámetros de la población.
- Pueden involucrar una o más variables

- ¿Cuál sería la hipótesis en nuestro caso?

Elementos de una prueba de hipótesis

- Luego, se debe elegir el test más apropiado en función del tipo de variable, del tipo de hipótesis.
- En nuestro caso, se trata de una prueba de diferencia de proporciones porque estamos tratando de evaluar si la diferencia entre la proporción de hombres trabajadores es diferente a la de mujeres trabajadoras.
- Por último, hay que definir un nivel de confianza.

Errores en prueba de hipótesis

Decisión Adoptada	Hipótesis H_0	
	Cierta	Falsa
No rechazar H_0	***Decisión Acertada*** $(1 - \alpha)$	***Error de Tipo II*** (β)
Rechazar H_0	***Error de Tipo I*** (α)	***Decisión Acertada*** $(1 - \beta)$

Elementos de una prueba de hipótesis

- Una vez planteada la Ho, es necesario “suponer” cómo sería la distribución muestral del parámetro si Ho es cierta.
- Es decir, si Ho fuera cierta, todos los valores del parámetro de todas las muestras posibles se distribuirían de alguna manera.
- En algunos casos podremos utilizar la ley de los grandes números y el TCL y podremos suponer una distribución muestral del parámetro (normal, chi-cuadrado, etc.)
- En otros, la estimaremos mediante bootstrap

Elementos de una prueba de hipótesis

- La cuestión es que una vez que pudimos determinar qué forma tiene la distribución muestral del parámetro en aquellos casos en que Ho es cierta, definimos un subconjunto de muestras, cuyas estimaciones del parámetro son tan extremas que la probabilidad de la muestra que observamos se encuentre entre ellas es muy pequeña.
- Si el valor que estimamos de nuestro parámetro “cae” en alguna de estas muestras, rechazaremos Ho. La proporción que este subconjunto de muestras ocupa en la distribución muestra, es lo que se denomina: nivel de significación (α), habitualmente, 0,05 o 0,01.

Ponderación / factor de expansión

- Ponderar es asignarle a cada unidad de muestreo su peso muestral, es decir, el valor que indica el número de unidades de la población que representa cada individuo o caso de la base de datos.
- El peso muestral es calculado por el/la muestrista en base a las características de la población, al tipo de diseño de la muestra y a las limitaciones que puedan surgir en la etapa de recolección de datos (situaciones de no respuesta total o parcial) que puedan afectar el diseño muestral original.
- Por ello, cuando se trabaja con datos secundarios el peso muestral ya se encuentra definido y sólo hay que activar la ponderación para el procesamiento de los mismos.