Tópicos de Estadística: Enseñanza basada en simulaciones.

REM - La Plata, 2018

Dra. Aldana González Montoro
FAMAF - Universidad Nacional de Córdoba
agonzalez@famaf.unc.edu.ar

Introducción

Cada mañana los medios de comunicación nos enfrentan con datos estadísticos en tópicos que van desde la economía a la educación, de películas a deportes, de comida a medicina, de opinión pública hasta comportamiento social. La comprensión de esa información guía decisiones en nuestra vida personal y nos habilita a desarrollar nuestras responsabilidades como ciudadanos y en cualquier área de trabajo en la que nos desarrollemos. En este sentido, cada estudiante graduado del secundario debería ser capaz de producir razonamientos estadísticos sólidos y de utilizarlos. Generalmente, los estudiantes terminan los cursos introductorios de estadística con la impresión de que la estadística es una colección de fórmulas y métodos no relacionados entre sí. El uso de metodologías de enseñanza de la estadística basadas en simulaciones nos permite mostrar a los estudiantes el proceso estadístico completo desde el primer día, sin atascarnos en formalidades matemáticas y condiciones que son requeridas en los métodos teóricos. De esta manera, una vez comprendida la lógica estadística, se hace más fácil y natural la incorporación de los métodos formales.

En este curso abordaremos tópicos centrales de la inferencia estadística usando ejemplos reales y mostrando cómo las simulaciones permiten comprender conceptualmente los procedimientos estadísticos. Los dos temas centrales que se tratan en los cursos de inferencia estadística son "intervalos de confianza" y "contrastes de hipótesis". Como el tiempo para este curso es acotado, abarcaremos solo uno de ellos: los contrastes de hipótesis. En los cursos clásicos, se deben abordar una gran cantidad de temas teóricos antes de llegar a este capítulo. Aquí, nos centraremos en la comprensión conceptual de los contrastes, utilizando métodos de aleatorización de datos. Este enfoque tiene, por lo menos, estas ventajas:

- Permite un abordaje más cercano a la intuición de los/las estudiantes.
- Es fácilmente generalizable a otras situaciones.
- Hace uso de tecnología, lo que lo vuelve más llamativo para el estudiantado.

De esta manera, en mi opinión, deberían empezar los cursos de estadística. Este enfoque más intuitivo, en conjunto con el uso de datos reales, datos que el estudiantado recolecte o traiga de casa y el desarrollo de proyectos, tiene el potencial de hacer que los jóvenes entiendan el verdadero poder de la estadística y se interesen por seguir aprendiendo. Existen libros de clase con este enfoque. Yo me basé, principalmente, en estos dos:

- Statistics: Unlocking the power of data. Lock, R.H. y Lock, P.F. y Morgan, K. y Lock, E.F. y Lock, D.F. (2013).
- Introduction to Statistical Investigations: High School Binding. Tintle, N. y Chance, B.L. y Cobb, G.W. y Rossman, A.J. y Roy, S. y Swanson, T. y VanderStoep, J. (2015).

Muchos de los ejemplos, provienen de estos libros.

Debido al poco tiempo con el que contamos, no mostraremos cómo realizar las simulaciones, aunque muchas de ellas ni siquiera requieren de una computadora. Discutiremos conceptos como: “parámetros poblacionales”, “estadísticos muestrales”, “variabilidad”, “distribución muestral”, “estimaciones puntuales”, “contrastes de hipótesis”, “evidencia muestral”, “p valor” y “significación estadística”.

Para empezar hagamos un recorrido por el proceso estadístico general.

El proceso Estadístico

“La ESTADÍSTICA es la ciencia de recolectar, describir y analizar datos”.

La estadística es la ciencia de la producción de datos útiles para abarcar una pregunta, analizar los datos resultantes y sacar conclusiones a partir de ellos. La Figura 1. muestra el proceso de la investigación estadística en 6 pasos, como descripto en Tintle2015. El primer paso, es el de realizar una pregunta que pueda ser respondida mediante la recolección de datos. Estas preguntas pueden incluir, por ejemplo, la comparación de grupos, el efecto de alguna cosa sobre otra o la valoración de la opinión de las personas. Luego, en el paso 2, tenemos el diseño del estudio o experimento y la recolección de los datos. Este paso incluye la selección de las personas u objetos que participarán en el estudio, decidir cómo juntar los datos y llevar a cabo la recolección con mucho cuidado y de manera sistemática. En el tercer paso, se exploran los datos recolectados, buscando patrones relacionados con nuestra pregunta e intentando detectar resultados inesperados que puedan apuntar a nuevas preguntas de interés. En el paso 4, se hacen inferencias más allá de los datos. Esto es, se determina si los resultados encontrados a partir de los datos reflejan una tendencia y se estima, además, el tamaño o fuerza de esa tendencia. En el paso 5, se formulan las conclusiones y se considera el alcance de las inferencias realizadas en el paso anterior. Es importante saber a qué proceso o a qué grupo de individuos u objetos pueden ser generalizadas estas conclusiones. El último paso involucra el observar el estudio completo para notar limitaciones y sugerir nuevos estudios para avanzar en los descubrimientos.

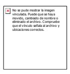


Figura 1. El proceso estadístico.

Un paseo rápido por la recolección y descripción de los datos

Nuestro objetivo es comprender el concepto de contrastes de hipótesis, una de las metodologías de la inferencia estadística más importantes. Sin embargo, para llegar a realizar un contraste de hipótesis, primero hay que diseñar un estudio o experimento, recolectar los datos y hacer un análisis exploratorio de los mismos. No nos vamos a detener en esa primera parte del proceso estadístico (la cual es sumamente importante) pero haremos un recorrido rápido por algunos conceptos que necesitaremos más adelante.

Un conjunto de datos consiste en una o más variables que recolectan información para cada uno de los casos en una muestra o población. Definamos los conceptos de esta frase.

Definiciones
Los sujetos u objetos de los cuales obtenemos información son los *casos* en un conjunto de datos y, generalmente, guardamos la información de cada caso en una fila de una tabla.
Una *variable* es una característica que se guarda de cada caso. Las variables generalmente corresponden a las columnas de la tabla.

Las variables son generalmente clasificadas como *categóricas* o *cuantitativas*.

Definiciones
Una variable *categórica* divide los casos en grupos, ubicando cada caso en, exactamente, una categoría.
Una variable *cuantitativa* mide una cantidad numérica para cada caso.

Generalmente no es posible conseguir datos sobre toda una población, la mayoría de las veces, debemos trabajar con una muestra.

Definiciones
Una *población* incluye a todos los individuos u objetos de interés.
Los datos se colectan, generalmente, sobre una *muestra*, que es un subconjunto de la población.

Lo que podamos inferir acerca de una población basándonos en los datos de una muestra depende del método de recolección de estos datos. Las muestras deben ser representativas de la población y evitar sesgos. Los métodos más eficaces para evitar sesgo son los muestreos aleatorios. También intentamos evitar otras fuentes de sesgo, como el tipo de preguntas en una encuesta. Pero tampoco nos vamos a detener en como diseñar experimentos.

En este curso trabajaremos con conjuntos de datos de una sola variable. Veamos algunas notaciones y definiciones.

Cuando tenemos una variable categórica, podemos hablar de la *proporción* de datos que partencen a una categoría en particular.

Definición
La *proporción* en una categoría se encuentra como
$\textrm{Proporción en una categoría} = \frac{\textrm{Número de datos en la categoría}}{\textrm{Número total de datos}}$
La proporción en una muestra la denotamos por $\hat{p}$ y leemos “p sombrero”. La proporción en una población la denotamos por p.

Cuando trabajamos con variables cuantitativas tenemos otras medidas como por ejemplo, la media de los datos.

Definición
La media (o promedio) de los datos, $x_1, x_2, \ldots, x_n$, de una variable cuantitativa es
Media $= \frac{x_1 + x_2 +\cdots+ x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$
La media de una muestra la denotamos por $\bar{x}$ y leemos ''x-barra".
La media de una población la denotamos por μ, la letra griega"mu".

También tenemos el desvío estándar, que nos da una medida de la distancia típica entre un valor cualquiera de los datos y la media. Mientras más grande es el desvío estándar, más dispersos son los datos.

Definición
El desvío estándar para una variable cuantitativa mide la dispersión de los datos en una muestra.
Desvío estándar $= \sqrt{\frac{\sum_{i =1}^{n} (x_i - \bar{x})^2}{n - 1}}$
El desvío estándar de una muestra se denota por s, y mide cuan dispersos están los datos de la muestra respecto de la media muestral $\bar{x}$.
El desvío estándar de una población se denota por σ, la letra griega "sigma" y mide cuan dispersos están los datos respecto de la media poblacional μ.

Cuando describimos una variable cuantitativa, consideramos varias cuestiones, dónde están centrados los datos, cuánto varían y qué forma tienen. Estos son aspectos de lo que llamamos la *distribución* de los datos. Dónde están centrados y cuánto varían lo podemos contestar con la media y el desvío estándar, entre otras. Para estudiar la forma de los datos, generalmente realizamos gráficos. Veamos un ejemplo.

Ejemplo

Hay una famosa base de datos *iris* que contiene medidas de variables para 50 flores de tres especies de iris. Consideraremos en este ejemplo la variable "ancho de sépalo" medida en centímetros. En la Figura 2 podemos observar los valores de esta variable para los casos de la muestra (en el gráfico de puntos de la izquierda) y un histograma (a la derecha). En el histograma, los datos se agrupan en intervalos y lo que muestra cada barra es la cantidad de datos que caen en cada intervalo. Podemos describir esta distribución como simétrica y con forma de campana. Este tipo de distribución lleva una atención especial en estadística y se la llama "normal".

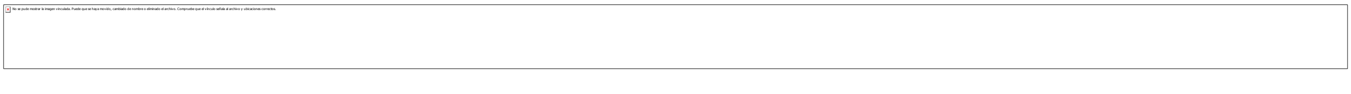


Figura 2. Ancho de sépalos para las flores iris. Conteos a la izquierda, histograma a la derecha.

Regla del 95%
Si una distribución es aproximadamente simétrica y con forma de

campana, cerca del 95% de los datos se encuentran a una distancia de la media menor a dos desvíos estándar.

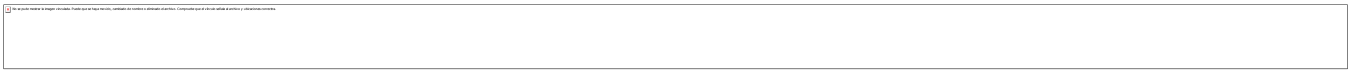


Figura 3. Regla del 95%.

Dentro del conjunto de datos *iris*, hay datos para tres especies distintas de la flor, setosa, versicolor y virgínca. La Figura 4 muestra los histogramas para cada una de estas especies. Las flechas indican el intervalo $\bar{x} \pm 2s$ en cada distribución y el porcentaje de datos que caen en él, mostrando que en este caso se cumple la regla del 95%.

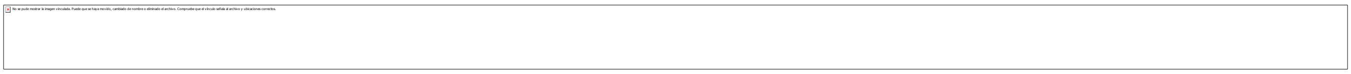


Figura 4. Ancho de sépalo para distintas variedades de la flor iris.

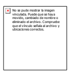
Podemos ver algunas diferencias entre las especies de iris. Por ejemplo, podemos ver que la media de la especie setosa es mayor a las otras dos y que hay muy pocos datos de la especie versicolor por encima de esa media. También podemos advertir que la especie setosa tiene valores más dispersos que las otras dos. Si nos dan un dato sin decirnos de qué especie es, y el valor de ese dato es 3.5, ¿diríamos que ese dato viene de una flor de la especie versicolor? ¿y de la especie virgínica? ¿Cuán atípico es ese dato si proviniera de esta última especie? ¿Cuán atípico es si proviniera de la especie setosa?

Las técnicas descriptivas nos permiten buscar patrones, encontrar anomalías y sugerir relaciones dentro de un conjunto de datos, sin embargo, las conclusiones que se pueden extraer son informales. Dada una muestra, podemos ver, por ejemplo, que las medias de dos grupos son distintas y que una es más grande que la otra, pero no podemos determinar si esa diferencia se puede extrapolar a la población o si es una característica de la muestra particular, obtenida por azar. Para poder sacar conclusiones sobre la población, necesitamos de la inferencia estadística.

Inferencia Estadística

Distribución muestral

Una vez que los datos fueron recolectados comenzamos con el proceso de usar la información de la muestra para entender qué puede ser verdad acerca de la población total. Primero se hace una inspección mediante medidas y gráficos descriptivos y luego, hacemos inferencia. Si todo lo que vemos son los datos de la muestra, ¿qué conclusiones podemos sacar acerca de la población? ¿Cuán seguros estamos acerca de la precisión de esas conclusiones? El proceso que busca responder estas preguntas se llama *inferencia estadística*.



La investigación estadística usa los datos de la muestra para entender una población.

> **Definición.**
> La *inferencia estadística* es el proceso de sacar conclusiones acerca de una población basadas en la información obtenida de una muestra.

Generalmente, las preguntas que nos hacemos tienen que ver con una característica de la distribución de los datos. Para identificar si estamos trabajando con la población entera o solo una muestra usamos el término *parámetro* si estamos hablando de una cantidad medida en la población y el término *estadístico* para identificar una cantidad medida en la muestra.

> **Definición.**
> Un *parámetro* es un número que describe una característica de una población.
> Un *estadístico* es un número que se calcula con los datos de una muestra.

Ejemplo

El censo de 2010 de la República Argentina dice que el $56,05\%$ de la población de tres años y más de la provincia de Córdoba utiliza computadora (GEOCENSO). Supongamos que tomamos una muestra aleatoria tamaño $n = 200$ de personas de tres años dos más de Córdoba, de las cuales 123 usan computadora. ¿Cuál es el parámetro poblacional y cuál el estadístico en este caso?

- El parámetro poblacional en este caso es la proporción personas, p, que usan computadora entre todas las personas de Córdoba, de tres o más años. Al tratarse de un censo, podemos decir que $p = 0,5605$. El estadístico muestral en este caso es la proporción de personas que usan computadora en la muestra. Lo denotamos con $\hat{p} = 123/200 = 0,615$.

Por lo general nos referimos con el mismo nombre (por ejemplo “media” o “proporción”) para referirnos tanto al parámetro como al estadístico muestral, pero generalmente usamos notaciones distintas:

	parámetro	**estadístico muestral**
media	μ	$\bar{x}$
desvío estándar	σ	s
proporción	p	$\hat{p}$

Ejemplo

La Encuesta Nacional de Factores de Riesgo (ENFR) se realiza en Argentina como estrategia nacional de prevención y control de enfermedades no transmisibles. La encuesta se realiza a una muestra de personas de 18 años o más que viven en conglomerados de 5000 o más habitantes. Según los datos obtenidos en la ENFR de 2005 la edad promedio a la que los hombres comienzan a fumar es de 16.7 años[1]. Para calcular este número se tuvieron en cuenta 10524 hombres fumadores o ex-fumadores participantes de la ENFR. ¿Cuál es el parámetro poblacional y cuál el estadístico muestral en este ejemplo?

- El parámetro poblacional es la edad media, μ, de comienzo a fumar de todos los hombres de 18 años o más (en 2005) que vivían en conglomerados de 5000 o más habitantes. El estadístico muestral es la media de edad de comienzo a fumar de los hombres en la muestra $\bar{x} = 16.7$.

Estimación Puntual

En general, para contestar una pregunta acerca de un parámetro poblacional necesitaríamos recolectar datos de todos los individuos de esa población. Esto puede resultar difícil o imposible debido a distintos motivos. En vez de esto, se puede recolectar la información de una muestra de esta población, calcular la cantidad de interés con los casos de la muestra y usar ese estadístico muestral para *estimar* el valor para toda la población. El valor del estadístico para una muestra particular nos da un *estimador puntual* (o *estimador*) del parámetro poblacional. Si solo tenemos la muestra y no conocemos el valor del parámetro poblacional, este estimador puntual, es lo mejor que tenemos para aproximar el valor del parámetro.

Ejemplo

En 2014, el INDEC, en conjunto con las Direcciones Provinciales de Estadística realizaron la Encuesta Nacional de Jóvenes 2014 (ENJ). Se entrevistaron 6340 jóvenes en todo el territorio nacional de 15 a 29 años residentes en centros urbanos de 2000 o más habitantes. Entre los resultados de esta encuesta, se expone que el 58.5% de jóvenes asiste a educación formal y/o no formal. ¿Cuál es el parámetro poblacional de interés? ¿Cuál es un buen estimador para ese parámetro?

- La población son jóvenes de 15 a 29 años residentes en centros urbanos de 2000 o más habitantes. El parámetro poblacional es p, la proporción de jóvenes que asisten a educación formal y/o formal. Para esta muestra $\hat{p}= 0.585$. Salvo que tengamos más información 0.585 es la mejor estimación del parámetro poblacional. Si quisiéramos encontrar p exactamente, deberíamos entrevistar a todos los jóvenes de la población de estudio.

  Definición
  Usamos el estadístico calculado de una muestra como un *estimador puntual* del parámetro poblacional.

Variabilidad en los estadísticos muestrales

Los parámetros poblacionales son números fijos, pero los estadísticos muestrales varían de una muestra a otra, dependiendo de los casos que fueron seleccionados para la muestra. Pero, ¿cuánto puede variar un estadístico de una muestra a otra? Esa es una de las preguntas fundamentales de la inferencia estadística, porque de esa variabilidad depende cuán preciso es un estadístico como estimador de un parámetro.

Ejemplo

Consideremos las Universidades Nacionales Argentinas. Del INDEC podemos obtener los datos de cantidad de egresados por Universidad en 2015. La tabla se puede bajar de la página del INDEC y se puede encontrar como anexo en este apunte. Hay 47 Universidades que tuvieron egresados ese año.

Según los datos, la media de egresados por universidad en 2015 es de 2097.06. Como estos datos corresponden al total de la población de Universidades Nacionales, tenemos que el parámetro poblacional μ es conocido y $\mu = 2097.06$.

Hagamos el siguiente experimento, tomemos una muestra tamaño 10 de las Universidades Nacionales y calculemos el promedio de egresados en 2015. La siguiente tabla tenemos una muestra aleatoria tamaño 10.

Universidad	Número de egresados en 2015
Universidad Nacional de Córdoba	7213
Universidad Nacional de La Pampa	418
Universidad Nacional de la Patagonia Austral	93
Universidad Nacional de Lanús	460
Universidad Nacional de Lomas de Zamora	2906
Universidad Nacional de Luján	843
Universidad Nacional de Mar del Plata	1101
Universidad Nacional de Quilmes	1143
Universidad Nacional del Comahue	984
Universidad Tecnológica Nacional	5144

El promedio de egresados por Universidad para esta muestra resulta $\bar{x} = 2030.50$. Este número es similar a μ pero no es exactamente igual. Observemos otra posible muestra:

Universidad	Número de egresados en 2015
Universidad Nacional de la Defensa	763
Universidad Nacional de Entre Ríos	1382
Universidad Nacional de Formosa	445
Universidad Nacional de Jujuy	268
Universidad Nacional de La Pampa	418
Universidad Nacional de la Patagonia San Juan Bosco	494
Universidad Nacional de Lomas de Zamora	2906
Universidad Nacional de San Juan	677
Universidad Nacional de San Luis	508
Universidad Nacional de Santiago del Estero	1205

Para esta muestra, $\bar{x} = 906.6$. Esta tampoco es igual a μ y tampoco es igual a la media muestral de la primera muestra encontrada. Si cada uno de nosotros elegimos una muestra tamaño 10 del total de la población de Universidades Nacionales, y calculáramos la media muestral, tendremos tantos resultados distintos como personas participantes, de hecho, ¡hay 5178066751 muestras tamaño 10 posibles! Esperamos que esas medias muestrales se encuentren alrededor de la verdadera media. Para ver que eso ocurre, generamos 1000 muestras de tamaño 10 y calculamos las medias muestrales, estas medias se pueden observar en la Figura 5. Observar que las medias muestrales se encuentran alrededor de $\mu = 2097.06$.

Figura 5. 1000 medias para muestras tamaño $n = 10$ de Egresados Universidades Nacionales 2015.

La distribución de los estadísticos muestrales como la de la Figura 5, se llama *distribución muestral*.

> Definición
> La *distribución muestral* es la distribución del estadístico muestral calculado para muchas muestras del mismo tamaño de la misma población. Expone como varía el estadístico de muestra a muestra.

¿Qué sucede si en vez de tomar muestras de tamaño 10, las tomamos de tamaño $n = 20$ o $n = 30$? La Figura 6 muestra las distribuciones muestrales para estos distintos tamaños de muestra.

Figura 6. Egresados de Universidades Nacionales 2015, 1000 muestras para cada tamaño muestral.

Podemos observar que en los tres gráficos, las medias están centradas en μ y se puede decir que tienen forma de campana, sobre todo para los tamaños muestrales más grandes. Lo que cambia de una a otra es la dispersión alrededor de ese valor. Aquí podemos sacar algunas conclusiones y dar algunas definiciones:

> *Forma y centro de una distribución muestral*
> Para casi todos los parámetros que podemos considerar, si las muestras están elegidas aleatoriamente y n es suficientemente grande, la distribución muestral será simétrica, con forma de campana y centrada en el valor del parámetro poblacional.

Una medida asociada a la variabilidad del estadístico muestral se puede encontrar calculando el desvío estándar de los estadísticos muestrales en una distribución muestral. Aunque este se calcula igual que el desvío estándar de los valores de una muestra, el desvío estándar de un estadístico muestral tiene su propio nombre: *error estándar*. Esto es por su gran importancia a la hora de saber cuán preciso es un estadístico a la hora de usarlo como estimador de un parámetro. Interpretamos al error estándar como la distancia típica entre los estadísticos muestrales y el parámetro poblacional.

> **Definición**
> El *error estandar* de un estadístico, denotado se, es el desvío estándar del estimador muestral.

Ejemplo

Según el censo de 2010, el 0.68% de los jóvenes de entre 15 y 17 años asiste al colegio secundario (secundario o polimodal). Podemos investigar el comportamiento de las proporciones muestrales usando algún programa que nos deje generar muestras aleatorias cuando la proporción poblacional en $p=0.68$. En la Figura 7 se observan las proporciones muestrales después de generar 1000 muestras tamaño $n=200$. Observemos que la distribución muestral tiene forma de campana, es simétrica y está centrada alrededor de 0.68.

Figura 7. Proporciones muestrales cuando $n=200$ y $p=0.68$.

En la Figura 8. observamos las proporciones muestrales para distintos valores de n.

Figura 8. Proporciones muestrales cuando $p=0.68$. Efecto del tamaño muestral. Regla del 95%.

Vemos que las distribuciones son simétricas, centradas en 0.68 y con forma de campana. Si calculamos los errores estándar obtenemos 0.066, 0.0323 y 0.0147 para $n = 50, 200$ y 1000 respectivamente. Y si calculamos el porcentaje de estadísticos muestrales que caen a una distancia menor a dos desviaciones estándar en cada caso, obtenemos 94.7%, 94.6% y 96%. Esto es coherente con la regla del 95%. Vemos que mientras más grande es n más disminuye se, disminuyendo la distancia típica entre los estadísticos y el parámetro y, por lo tanto, haciendo al estadístico más preciso. Esto nos indica que ¡el tamaño de la muestra es importante!

> *El tamaño muestral importa*
> Cuando el tamaño muestral crece, la variabilidad del estadístico muestral tiende a disminuir y, por lo tanto, los estadísticos muestrales tienden a estar más cerca del verdadero valor del parámetro.

Contrastes de Hipótesis

Ejemplo

Un estudio famoso de los años 60 explora si los delfines pueden comunicar ideas abstractas. En este estudio un investigador entrenó (durante mucho tiempo) a dos delfines, Buzz y Doris, para que tocaran, a cambio de un premio, uno de dos botones cuando veían una señal de luz. Cuando la luz no parpadeaba, los delfines debían tocar el botón de la derecha y, si parpadeaba, el de la izquierda. Luego, el investigador separó la pileta en dos, de manera tal que Buzz no viera la luz. Al encender la luz sin parpadear, el investigador observó que Doris se acercó a la separación, silbó y entonces Buzz se acercó al botón de la derecha, lo tocó y así ambos recibieron su premio. ¿Es suficiente esto para decir que los delfines se comunicaron? Podría ser una casualidad. Para definir si los delfines se comunicaban o no, el investigador repitió el experimento varias veces. En este escenario, aunque los delfines se estuvieran comunicando, no esperamos que

Buzz toque el botón correcto todas las veces, esperamos algún tipo de "aleatoriedad". También podría ser que los delfines no se estuvieran comunicando. ¿Cómo podemos decidir cuál de los casos es? La idea es *estimar* cuantas veces Buzz tocaría bien el botón si repitiéramos el experimento infinitas veces.

El parámetro poblacional en este caso es esa *proporción a largo plazo*, de veces que Buzz tocaría el botón correcto, es decir la *probabilidad* de que Buzz toque el botón correcto. Llamaremos a esa probabilidad π.

La conjetura del experimento es que Buzz y Doris se comunican y, por lo tanto, esperamos que esa probabilidad sea grande. Si Buzz y Doris no se comunican entonces Buzz estaría adivinando cuál botón tocar (y π sería 0.5). Como no se puede repetir el experimento infinitas veces, se utiliza una muestra para estimar el parámetro π.

En el estudio, el investigador repite en experimento 16 veces, de las cuales Buzz toca el botón correcto 15 veces. Para resumir este resultado podemos usar el estadístico muestral $\hat{p} = 15/16 = 0.9375$ como estimador puntual del parámetro π. Pueden estar ocurriendo dos cosas:

- Buzz y Doris se comunican y por lo tanto $\pi > 0.5$.
- Buzz está adivinando, $\pi = 0.5$ y tuvo (mucha) suerte en esos 16 intentos.

¿Cuál parece la explicación más razonable? Es esta evidencia suficiente para decir que los delfines se están comunicando? ¿Cómo podríamos justificarlo?

Basándonos en los datos de la muestra queremos evaluar una afirmación acerca de una población. Esta es la esencia de un *contraste de hipótesis*: determinar si los resultados de una muestra son lo suficientemente convincentes para concluir algo sobre la población.

> Contraste de hipótesis
> Un *contraste de hipótesis* usa los datos de una muestra para evaluar una afirmación sobre una población.

Un contraste de hipótesis consta de dos afirmaciones contrarias, las hipótesis, a las que llamamos *hipótesis nula*, denotada por H_0, e hipótesis alternativa, denotada por H_a. La afirmación para la que buscamos evidencia convincente, es asignada a la hipótesis alternativa (Buzz y Doris se comunican). Generalmente, la hipótesis nula es la afirmación de que *no hay efecto* o *no hay diferencia* (Buzz está adivinando). La hipótesis alternativa se establece observando la evidencia (datos) que contradice la hipótesis nula.

> **Definición**
> Hipótesis nula (H_0): afirma que no hay efecto o diferencia.
> Hipótesis alternativa (H_a): afirmación para la cual buscamos evidencia significativa.

Para que las hipótesis sean concretas, se escriben en función de un parámetro poblacional. En el ejemplo de Doris y Buzz, escribimos:

- $H_0:\,\, \pi = 0.5$ (Buzz está adivinando).
- $H_a: \,\, \pi > 0.5$ (Buzz y Doris se comunican).

En el caso del ejemplo, sabemos que $\hat{p} = 0.9375$, es mayor a 0.5. La pregunta es si ese estadístico nos da evidencia suficiente para decir que la probabilidad de que Buzz elija correctamente es mayor a 0.5.

Si Buzz estuviera adivinando, es decir, eligiendo el cada botón con una probabilidad de 0.5, el proceso de elección se puede comparar con el proceso de tirar una moneda al aire. Podemos tirar una moneda al aire para **simular** o representar la elección de Buzz asumiendo que él simplemente adivinaba qué botón apretar. Para generar estos datos artificiales, supongamos que *cara* representa el hecho de que Buzz adivina correctamente y *seca* que se equivoca. Veamos la relación entre el problema de los delfines y la simulación:

Simulación	Real
tirar una moneda	Buzz elige un botón
cara	Buzz adivina correctamente
seca	Buzz se equivoca
probabilidad de sacar cara	0.5, probabilidad de que Buzz elija el botón correcto en el contexto de que este simplemente adivinando
una repetición	un conjunto de 16 intentos de Buzz

Simulemos un experimento. Tiramos la moneda 16 veces y, por ejemplo, obtenemos 11 caras y 5 secas, como en la Figura 9.

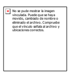


Figura 9. Monedas lanzadas. monedas

Esto simula la situación en la que Buzz se le hace elegir 16 veces un botón y en cada una de ellas, Buzz simplemente elige al azar uno de ellos. En este caso obtenemos el estadístico simulado de $11/16 = 0.6875$. Si repetimos el procedimiento otra vez, ¿obtendremos el mismo resultado? Como tirar una moneda al aire es un proceso aleatorio, sabemos que no obtendremos siempre la misma cantidad de caras en cada conjunto de 16 tiradas. Si continuamos arrojando la moneda en tandas de 16 y calculando la proporción de caras, podemos obtener la distribución muestral del estadístico $\hat{p}$ bajo la suposición de que $\pi=0.5$, esto es, la distribución muestral de $\hat{p}$ asumiendo la hipótesis de que Buzz elige al azar (y por lo tanto no se comunica con Doris). Sabemos que esa distribución está centrada en el parámetro poblacional, que en este caso es 0.5. Esa distribución nos permite contestar preguntas como, ¿qué proporciones de *cara* son las más probables de obtener? o, ¿cuánta variabilidad hay en nuestro estadístico muestral simulado?

En la Figura 10 se muestran los resultados para 1000 repeticiones del experimento simulado (16 lanzamientos de una moneda).

Figura 10. 1000 repeticiones de tirar una moneda 16 veces y contar el número de caras.

Según el gráfico, los resultados más probables son 7,8 y 9. El 6 y el 10 también ocurren bastantes veces, el 15 ocurrió una sola vez y el 0,1 y 16 ninguna. Podríamos considerar cualquier resultado entre 5 y 11 como típico, pero sacar menos de 5 o más de 11 ocurrieron muy raramente, por lo que los podemos considerar inusuales. Nos referimos a esos resultados inusuales como a que están en las *colas* de la distribución muestral.

En el estudio real, Buzz apretó el botón correcto 15 de 16 veces. Este resultado, acabamos de ver que es un resultado muy poco esperable si Buzz estaba eligiendo un botón al azar. Entonces, el modelo de la moneda nos dice que hay una fuerte evidencia de que Buzz no estaba simplemente adivinando. Los resultados nos dicen que la evidencia es lo suficientemente fuerte como ser considerada *estadísticamente significativa*.

> Significación Estadística
> Cuando un resultado como el estadístico muestral observado es muy poco probable de obtener por mero azar cuando se asume que la hipótesis nula es verdadera, decimos que el resultado es *estadísticamente significativo*.
> Si el resultado es estadísticamente significativo, tenemos evidencia suficiente en contra de H_0 y a favor de H_a.

Si en el ejemplo de Buzz y Doris, Buzz hubiera elegido correctamente 11 de las 16 veces. ¿Qué conclusión podríamos sacar? ¿Cómo decidimos qué resultados son estadísticamente significativos y cuáles no?

Ejemplo

El juego *piedra, papel o tijera* (o *Rochambeau*) es considerado un juego justo. Es decir, un juego en el que ambos jugadores tienen la misma probabilidad de ganar. El juego consiste en lo siguiente. Dos jugadores muestran a la vez una de tres señas de mano (*piedra*, *papel* o *tijera*) y el objetivo es mostrar una seña que le gane al otro jugador, donde el *papel* envuelve la *piedra*, la *piedra* aplasta la *tijera* y la *tijera* corta al papel. Las reglas se pueden encontrar aquí. Pero, ¿es un juego realmente justo? En un artículo publicado en College mathematics Journal en 2009[2], se encontró que los jugadores, particularmente los nuevos, tienden a **no** elegir *tijera*. ¿Cómo podemos hacer un estudio para ver si esto es cierto?

Un estudio posible sería el siguiente. Elegimos una persona que no haya jugado nunca, jugamos el juego 12 veces y contamos cuántas veces esa persona eligió *tijera*. ¿Cuáles son las unidades de observación y el tamaño muestral? ¿Cuál es el proceso subyacente? ¿Cuál es el parámetro de interés? ¿Cuál es el estadístico muestral?

- Cada partida es una unidad de observación y el tamaño muestral es 12. El proceso subyacente es la elección de qué seña mostrar. El parámetro de interés (poblacional) es la proporción a largo plazo de veces que se elige *tijera* o la probabilidad de elegir *tijera*. El estadístico muestral es $\hat{p}$, la proporción de veces que la persona eligió *tijera*.

¿Cuáles son las hipótesis de interés, H_0 y H_a?

- Queremos estudiar si realmente la gente nueva tiende a elegir menos de 1/3 de las veces la seña correspndiente a *tijera*. Luego nuestra hipótesis nula es la de que no hay diferencias entre las señas:

$H_0: \pi = 1/3$
$H_a: \pi < 1/3$

Supongamos que hacemos este experimento con un amigo nuestro que nunca jugó. Le enseñamos el juego, jugamos 12 veces, de las cuales solo elige tijera 2 veces. Es decir, la $\hat{p} = 2/12 = 0.167$. ¿Cómo podemos hacer para ver si el estadístico muestral observado es típico o está en las colas de la distribución muestral cuando asumimos H_0 verdadera? Otra forma de formular la pregunta es, ¿cómo hacemos para evaluar si este valor es evidencia suficiente para rechaza nuestra hipótesis nula a favor de la alternativa?

Como en el ejemplo anterior pensemos en construir una distribución muestral para $\hat{p}$ bajo H_0. Ahora ya no podemos tirar una moneda, dado que bajo hipótesis nula, $\pi = 1/3$ no $1/2$. Podríamos por ejemplo, tirar un dado de seis caras y decidir que tijera equivale a sacar 1 o 2 en la tirada del dado.

La Figura 11 muestra la distribución de $\hat{p}$ construida simulando 1000 veces el experimento de tirar el dado 12 veces y contar cuantas veces salió 1 o 2 (entre las 12).

Figura 11. La distribución bajo H_0 de la proporción de veces que se elige tijera en 12 partidas de piedra, papel o tijera.

Lo que tenemos que ver ahora es si el estadístico muestral ($2/12 \sim 0.167$) es consistente con la distribución muestral en el proceso donde $\pi =1/3$. Es decir, tenemos que decidir si 0.167 es un valor típico en esa distribución o está en las colas. En la Figura 11, vemos que no es tan fácil decidir cómo en el ejemplo anterior. El valor 0.17 no está tan alejado en la cola inferior de la distribución, pero tampoco está en el medio. Lo que necesitamos es una forma sistemática de decidir cuan inusual es un valor en la distribución nula. La forma de hacer esto es observar la proporción de la distribución que cae en o por debajo bajo del valor observado (o en y por encima).

> **Definición**
> La probabilidad de obtener un valor a lo sumo tan extremo como el observado (en la dirección de la hipótesis alternativa), cuando la hipótesis nula es verdadera, se llama *p-valor*.

Podemos estimar esta probabilidad contando cuántos (y la proporción) valores de la distribución simulada son tan o más extremos que el estadístico muestral observado. El p-valor nos provee de una medida directa de la fuerza de la evidencia en contra de la hipótesis nula. Mientras más chico es el p-valor, más improbable es observar un valor como el del estadístico muestral, bajo la hipótesis nula, por mero chance. Es decir, mientras más chico el p-valor, más estadísticamente significativa es la evidencia en contra de la hipótesis nula.

Calculemos el p-valor en este caso. Hay 172 muestras en las que salieron 2 o menos tijeras. Por lo tanto $$p\text{-valor} = 172/1000 = 0.172$$

¿Es este p-valor chico o grande? ¿Cuál es el límite para decir que tenemos evidencia suficiente para rechazar H_0? Dejemos esto para un poco más adelante. Antes, veamos unos ejemplos con variables cuantitativas.

Ejemplo

Si le preguntamos a cualquier persona cuál es la temperatura normal del cuerpo humano nos responderá, probablemente, $37^\circ C$. Este es un *hecho* que todos creemos verdad desde hace mucho tiempo. ¿Será posible que eso haya cambiado o que sea una creencia incorrecta? En un estudio publicado en el Journal of Statistics Education en 1996[3], los autores presentan datos de $n= 50$ personas sanas en los que se incluye la temperatura corporal. La temperatura media en esta muestra resulta $\bar{x} = 36.811$. Un histograma de los datos lo vemos en la Figura 12.

Figura 12. Temperatura corporal de 50 personas sanas.

Por supuesto, no esperamos que la media de la temperatura corporal de toda muestra tamaño 50 sea $37^\circ C$, habrá variabilidad de una muestra a otra, pero, nos preguntamos si la media observada en esta muestra, 36.811, está alejada más de lo que uno esperaría por mero azar si la verdadera media es $\mu = 37^\circ C$. En este caso, entonces, las hipótesis de interés son:

- $H_0: \mu = 37^\circ C$
- $H_a: \mu \neq 37^\circ C$

Como la media muestral es 36.811, podríamos elegir H_a como $\mu < 37^\circ C$, sin embargo, elegimos $\mu \neq 37^\circ C$ ya que en principio queremos saber si la media a cambiado, no importa en qué sentido. Este tipo de prueba se llama a *dos colas* mientras que las de los ejemplos anteriores se llaman a *una cola*. Como se plantea la hipótesis alternativa, afecta en la forma que concluimos la significación estadística.

Para evaluar la significación de la evidencia en esta muestra, necesitamos aproximar la distribución muestral del estadístico $\bar{x}$. Para hacer esto utilizaremos un método de aleatorización o remuestreo. Como nos interesa el tipo de estadísticos que se obtienen cuando la hipótesis nula es verdadera, simulamos datos de manera tal que sean consistentes con la hipótesis nula y con el proceso que generó la muestra original. Llamamos a estas muestras, muestras aleatorizadas.

Generando muestras aleatorizadas
Los criterios más importantes a tener en cuenta para generar muestras aleatorizadas para un contraste de hipótesis son:
- ser consistente con la hipótesis nula.
- Usar los datos de la muestra original.
- Imitar la forma en la que fue generada la muestra original.

En el ejemplo, esto significa que debemos generar muestras que sean consistentes con la hipótesis de que la media es $37^\circ C$. Y, además, es necesario que las muestras generadas reflejen la estructura de la muestra original. En particular, que mantengan la variabilidad de los datos. La forma más sencilla de hacer esto, en este caso, es sumar a cada dato de la muestra original una constante, de manera tal que los datos trasladados tengan media $37^\circ C$ y luego tomar muestras aleatorias con reemplazo tamaño $n =50$ de la muestra original trasladada.

¿Cuál es la constante que debemos sumar a los datos de la muestra original? Si sumamos la constante $\mu - \bar{x}$ a cada dato, la media de los nuevos datos será μ (la cuenta es sencilla). En este caso, esa diferencia es 0.189.

La Figura 13 muestra la distribución muestral para $\bar{x}$ bajo la hipótesis nula. Observemos que está centrada en $\mu = 37$. Ahora, lo importante es ver dónde cae el valor observado del estadístico muestral y decidir si es un valor típico o está en las colas de la distribución.

No se pude mostrar la imagen vinculada. Puede que se haya movido, cambiado de nombre o eliminado el archivo. Compruebe que el vínculo señala al archivo y ubicaciones correctos.

Figura 13. Distribución de la media de las temperaturas corporales cuando $\mu =37 ^\circ C$.

¿Es esta evidencia suficiente para decir que la temperatura corporal media no es $37^\circ C$? Solo hay 4 medias muestrales menores a 36.811, esto nos dice que el valor del estadístico muestral es un valor no muy típico bajo hipótesis nula. Calculemos el p-valor. En este caso como la hipótesis alternativa es $\mu \neq 37^\circ C$, el p-valor corresponde a la probabilidad de observar valores en ambas colas tan raros como el observado, luego

$$p\text{-valor} = 2*(4/1000) = 0.008$$ Este es un valor muy pequeño, lo que nos deja afirmar que hay evidencia suficiente de que la media corporal en humanos sanos es distinta a $37^\circ C$.

Siempre es esencial que la forma de remuestrear imite la forma de tomar las muestras aleatorias originales, pero asumiendo que la hipótesis nula es verdadera. En este caso lo logramos generando muestras aleatorias con la misma estructura que la original pero desplazadas en media. Veamos un ejemplo más, en el que la forma de aleatorizar es distinta.

Ejemplo

Es de conocimiento popular que la cafeína nos despierta, nos pone atentos. ¿O es mito? Los efectos de la cafeína en el cuerpo han sido muy estudiados. En un estudio en particular[4], los investigadores entrenaron a un grupo de hombres para golpear un dedo contra la mesa a un ritmo rápido. La muestra fue dividida en 2 grupos de 10 individuos cada una. Cada persona tomó 2 tazas de café, con 200 mg de cafeína para un grupo y descafeinado para el otro. Después de dos horas, cada individuo fue testeado para medir la cantidad de golpes por minuto que podían dar. El objetivo del estudio era ver si la cafeína aumentaba la cantidad de golpes por minuto. Los datos son los siguientes:

											$\bar{x}$
Cafeína	246	248	250	252	248	250	246	248	245	250	248.3
No cafeína	242	245	244	248	247	248	242	244	246	242	244.8

Aquí tenemos datos cuantitativos y estamos interesados en el promedio de golpes por minuto para los dos grupos distintos. Podemos denotar por μ_c y μ_n las medias poblacionales con y sin cafeína, respectivamente. Los investigadores buscan por diferencias en las medias. Vemos de la tabla que la media muestral para el grupo con cafeína es mayor que la media muestral para el grupo sin cafeína, la pregunta es si esa diferencia es estadísticamente significativa o puede haber ocurrido por mero chance. Entonces las hipótesis las podemos plantear como

- $H_0: \mu_c = \mu_n$
- $H_a: \mu_c > \mu_n$

La Figura 14 muestra los datos para los dos grupos, las flechas muestran donde están las medias muestrales.

Figura 14. Datos de golpes por minuto para los grupos con y sin cafeína.

Es notable que todos los individuos del grupo con cafeína tienen un ritmo mayor a la media del grupo sin cafeína, mientras que ninguno del grupo sin cafeína está por encima de la media del grupo con cafeína. Esto apunta a apoyar la hipótesis alternativa, de que la cafeína aumenta el ritmo. Por otro lado, hay mucha superposición entre las dos distribuciones, luego es posible que haya sido el azar que asignó a varios de los individuos con menor ritmo al grupo sin cafeína.

El estadístico muestral que consideraremos en este caso es la diferencia en medias. Para estos datos, la diferencia observada es

$$D = \bar{x}_c - \bar{x}_n = 248.3 -244.8 = 3.5$$

Para determinar si este valor es significativamente distinto de cero, necesitamos encontrar la probabilidad de haber obtenido un valor como $D = 3.5$ si la cafeína no tiene efecto en el ritmo de golpeo, es decir, asumiendo que la hipótesis nula es verdadera. Para esto, necesitamos aproximar la distribución muestral de D bajo H_0 para calcular un el p-valor.

Este caso es diferente al anterior. Asumir que la hipótesis nula es verdadera es decir que $\mu_c=\mu_n$. Esto es, que la cafeína no tiene efecto en el ritmo de los golpes que da una persona. Si H_0 es verdadera, entonces, implica que el ritmo de golpeo de una persona hubiera sido el mismo, habiendo tomado cafeína o no. Luego, cualquiera de los valores observados en el grupo con cafeína podría haber sido observados en el grupo sin cafeína si la asignación (aleatoria) de los individuos a los grupos hubiera sido otra. Esto nos da una idea de cómo generar nuestras muestras aleatorizadas para aproximar la distribución muestral de D. Podemos asignar aleatoriamente los 20 valores observados a dos grupos: con y sin cafeína y calcular D con la nueva asignación. ¿Cómo podemos simular este proceso? Podríamos por ejemplo, tomar 20 cartas, en cada una escribir un valor de la muestra, mezclarlas y luego armar dos pilas de diez, una para cada grupo. Por ejemplo, en la Figura 15, se muestra una de estas asignaciones aleatorias, las flechas señalan las medias muestrales. En esta muestra artificial, D resultó $\bar{x}_c - \bar{x}_n = 246.9 - 246.2 = 0.7$.

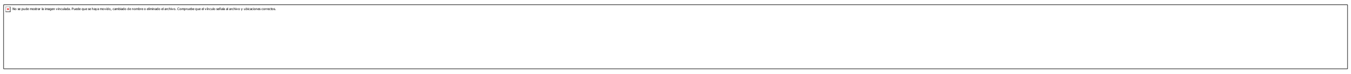


Figura 15. Muestra aleatorizada para el experimento del ritmo de golpeo para dos grupos: con y sin cafeína.

Cada vez que mezclamos y separamos en dos la pila de cartas, simulamos una repetición del experimento de separar aleatoriamente los individuos en dos grupos, asumiendo que la cafeína no influye en el ritmo de golpeo. Repitiendo muchas veces este proceso aproximamos la distribución de D y podemos ver dónde cae el estadístico observado $D=3.5$.

La Figura 16 muestra las diferencias en media de 1000 de estas muestras aleatorizadas. En rojo se muestran las medias mayores a 3.5. Como la hipótesis nula es la de igualdad de medias, o lo que es lo mismo $\mu_c -\mu_n =0$, no nos sorprende que la distribución muestral de la diferencia esté centrada en cero.

Figura 16. Diferencias en media para 1000 muestras aleatorizadas.

¿Es entonces 3.5 un valor típico bajo la hipótesis de que la cafeína no afecta al ritmo de golpeo? ¿Cuán poco probable es obtener este valor bajo H_0? De las 1000 muestras aleatorizadas, en 24 se obtuvo un valor igual o mayor a 3.5, luego

$$p\text{-valor} = \frac{24}{1000} = 0.024$$

¿Hay evidencia suficiente para decir que la cafeína aumenta el ritmo de golpeo? Nuevamente, nos preguntamos si este p-valor es suficientemente chico para considerar que la evidencia sea estadísticamente significativa.

Hemos dicho que mientras más chico el p-valor, mayor la evidencia en contra de la hipótesis nula, pero, si tenemos que tomar una decisión entre rechazar o no rechazar una hipótesis nula, cuán chico tiene que ser el p-valor para rechazar? Podemos estar de acuerdo en que un p-valor de 0.0001 es suficientemente evidencia para rechazar H_0 y que un p-valor de 0.5 es claramente insuficiente para rechazar. Pero, dónde está el corte entre suficiente e insuficiente evidencia? Este punto de corte, puede cambiar de contexto en contexto pero hay que elegirlo de antemano. Los estadísticos llamamos a ese punto *nivel de significación* y generalmente se denota con la letra griega α. Por ejemplo, si $\alpha = 0.5$, decimos que estamos llevando a cabo un contraste de hipótesis del 5%, rechazaremos la hipótesis nula si el p-valor resulta menor a 0.05 y diremos que se rechaza la hipótesis a nivel 5%.

> **Definición**
> El nivel de significación, α, para un contraste de hipótesis en un umbral por debajo del cual concluiremos que un p-valor muestra evidencia estadísticamente significativa en contra de la hipótesis nula.
>
> Decisión estadística formal, basada en un nivel de significación
> Dado un nivel de significación α y el p-valor obtenido de una muestra:
> - Rechazamos H_0 si p-valor $< \alpha$.
> - No rechazamos H_0 si p-valor $\geq \alpha$.

Valores de significación comunes son $\alpha = 0.05$, $\alpha = 0.01$, o $\alpha = 0.10$. Si bien esto puede cambiar de un contexto a otro, hay bastante consenso en las siguientes conclusiones:

$\quad \quad \quad$	$\quad\quad\quad$
0.10 < p - valor	no mucha evidencia en contra de H_0.
0.05 $\leq$ p - valor $\leq$ 0.10	evidencia moderada en contra de H_0
0.01 < p - valor $\leq$ 0.05	fuerte evidencia en contra de H_0
p - valor $\leq$ 0.01	evidencia muy fuerte en contra de H_0

Ahora somos capaces de concluir en el ejemplo del juego *piedra*, *papel* o *tijera* y el del ritmo de golpeo. En el primer caso, obtuvimos un p-valor de 0.172, luego la evidencia es no significativa a nivel 5% y 10%. En el segundo, el p-valor resultó 0.024, luego hay evidencia suficiente, a nivel 5% de que la cafeína aumenta el ritmo de golpeo.

Última consideración, ¿cómo elegimos el nivel de significación para nuestro problema? Al realizar un contraste, existen dos tipos de errores que uno puede cometer:

Error	
Tipo I	Rechazar H_0 cuando es verdadera
Tipo II	No rechazar H_0 cuando es falsa

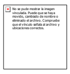
El nivel de significación, α, representa la máxima probabilidad tolerable de cometer el error tipo I. Pero esto quedará para otra oportunidad.

Sobre dónde hacer las simulaciones…

Hay muchas opciones:

- GeoGebra:

Por ejemplo en el ejemplo de Buzz y Doris, donde necesitamos simular el experimento de lanzar una moneda 16 veces:



GeoGebra

- Excel (Datos -> Análisis de datos -> Generación de números aleatorios)
- Infostat (Aplicaciones -> Didácticas -> Remuestreo)
- R

Hay aplicaciones online también (éstas en inglés):

- StatKey - del liro “Unlocking the power of data”
- Apps del Libro “Introduction to Statistical Investigations”

Por ejemplo, aquí se puede simular el ejemplo de los confites: ReesesPieces

Bibliogrfía

Tintle2015. Tintle, Nathan y Chance, Beth L y Cobb, George W y Rossman, Allan J y Roy, Soma y Swanson, Todd y VanderStoep, Jill. (2015) Introduction to Statistical Investigations: High School Binding. **Wiley Online Library.**

Lock2013. Lock, Robin H y Lock, Patti Frazer y Morgan, Kari Lock y Lock, Eric F y Lock, Dennis F. (2013) Statistics: Unlocking the power of data. **Wiley Hoboken, NJ**

1: Fleischer, Nancy L., et al. “Socioeconomic patterning in tobacco use in Argentina, 2005” (2011). Nicotine & Tobacco Research 13.10: 894-902.

2: Eyler, Shalla, Doumaux, and McDevitt. “Winning at Rock-Paper-Scissors” (2009). The College Mathematics Journal 40: 125–128.

3: Shoemaker, A., ’’What’s Normal?-Temperature, Gender and Heartrate" (1996). Journal of Statistics Education; 4(2).

4: Hand, A.J., Daly, F., Lund, A.D., McConway, K.J., and Ostrowski, E., Handbook of Small Data Sets, Chapman and Hall, London, 1994, p. 40.