# Estadísticas: una (re-)introducción

Prof. Rashid C.J. Marcano Rivera

SOCI 6015 - Métodos cuantitativos
Departamento de Sociología y Antropología
Facultad de Ciencias Sociales
Universidad de Puerto Rico
Recinto de Río Piedras

2 de septiembre de 2025

# Estadísticas

- **Estadísticas**: Una rama de las matemáticas que se enfoca en la organización, análisis e interpretación de un grupo de números.
- **Estadísticas descriptivas**: Procedimientos para resumir un grupo de puntuaciones u otras formas de hacerlas más comprensibles.
- **Estadísticas inferenciales**: Procedimientos para extraer conclusiones (inferencias) basadas en las puntuaciones recolectadas en un estudio de investigación, y que van más allá de esas puntuaciones.

# Conceptos Básicos

- **Variable**: Característica que puede tener diferentes valores.

- **Variable**: Característica que puede tener diferentes valores.
- **Valor**: Número o categoría que una puntuación puede tener.

# Conceptos Básicos

- **Variable**: Característica que puede tener diferentes valores.
- **Valor**: Número o categoría que una puntuación puede tener.
- **Puntuación**: El valor particular de una persona en una variable.

# Ejemplo de Valores

- Un valor es simplemente un número, como 9, -81, 0.1, o 367.12.
- Un valor también puede ser una categoría, como femenino, masculino, o el país en del que procedas (Cuba, Puerto Rico, Rep. Dominicana, Venezuela, Colombia, etc.).

# Tipos de Variables

Los valores que toman las variables pueden ser cualitativos (categorías), como sucede con la variable color, o cuantitativos (números), como ocurre con la variable edad.

| **Tipo de variable** | **Definición** | **Ejemplo** |
|---|---|---|
| **Numérica** | Variable cuantitativa o cuyo valor es un número general. | Edad, salario, temperatura |
| **Intervalo** | Sus valores tienen un orden natural, es posible cuantificar la diferencia entre dos valores de intervalo. Suelen tener unidad de medida. | Puntuación en un examen, grados Celsius, distancia en metros. |
| **Ordinal** | Variable numérica o categórica con valores en rangos. | Posición en una carrera, rango militar, nivel socioeconómico, nivel educativo. |
| **Nominal** | Variable cuyos valores son categorías. | Género, grupo racial, país de origen, partido por el cual se vota |

En la práctica la numérica es una generalización. Dentro de ella hay de razón o proporción, que permiten realizar operaciones de multiplicación y división sobre sus valores, además de las de suma y resta. La mayoría de variables cuantitativas son de razón. Dónde se ubiquen variables dependerá del nivel de medida: ¿cómo decide uno medir una variable? Es probable que haya varias opciones.

# Tabla de frecuencia: calificación por servicio

| **Calificación** | **Frecuencia** | **Porcentaje ( %)** |
|---|---|---|
| 1 | 3 | 15 % |
| 2 | 5 | 25 % |
| 3 | 6 | 30 % |
| 4 | 4 | 20 % |
| 5 | 2 | 10 % |
| **Total** | **20** | **100 %** |

Tabla: Distribución de calificaciones por servicio

En este caso, la frecuencia se asigna a una variable numérica. Si las tablas fueran de demasiados grupos o categorías, estos se pueden combinar en intervalos.

# Histograma de calificaciones por servicio

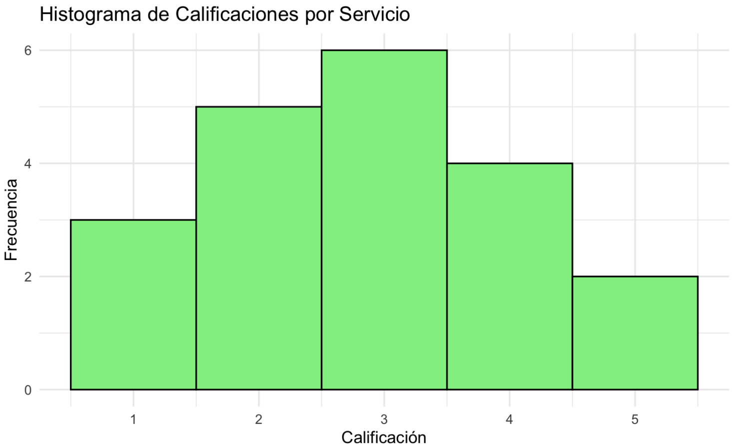
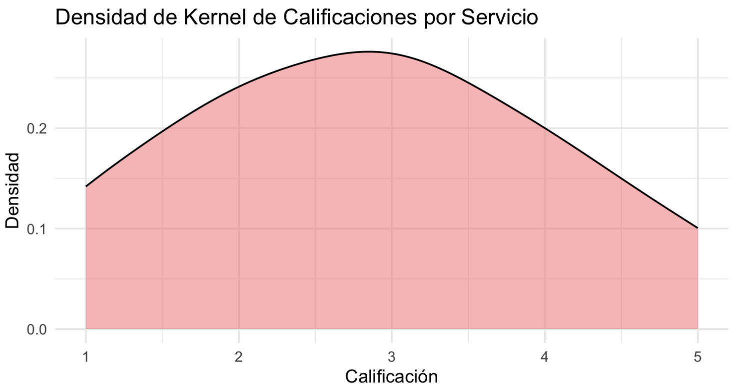



Figura: Este histograma muestra la distribución de las calificaciones que los usuarios dieron por el servicio.

# Densidad de kernel de calificaciones por servicio



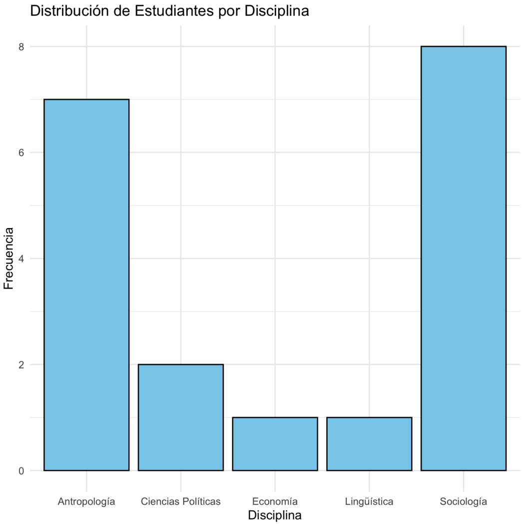
# Tabla de frecuencia: distribución de estudiantes

| **Disciplina** | **Frecuencia** | **Porcentaje ( %)** |
|---|---|---|
| Sociología | 8 | 40 % |
| Antropología | 7 | 35 % |
| Ciencias Políticas | 2 | 10 % |
| Economía | 1 | 5 % |
| Lingüística | 1 | 5 % |
| **Total** | **20** | **100 %** |

Tabla: Distribución de estudiantes por disciplina en una clase subgraduada

Este caso es de variables categóricas.

# Distribución de estudiantes por disciplina



# Promedios

Usualmente, la mejor medida representativa de un grupo de puntuaciones es el promedio, que es la suma de todas las puntuaciones dividida por el número total. En estadística, esto se conoce como la media.

Ejemplo: Un politólogo estudia los años de experiencia en cargos electos de los alcaldes de las 10 ciudades más grandes de una región. Los años servidos fueron:
**7, 8, 8, 7, 3, 1, 6, 9, 3, 8**

La media de estos 10 valores es 6, es decir, en promedio, estos alcaldes han servido 6 años en el cargo.

# Promedios (cont.)

**Definición**

El promedio (media) de un conjunto de puntuaciones se define matemáticamente como:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

Donde:

- $\bar{\mathbf{x}}$ es la media del conjunto de datos.
- $n$ es el número total de puntuaciones (siendo aquí un escalar).
- $\mathbf{x}_i$ representa cada una de las puntuaciones individuales.

**Moda**
La moda es otra medida del valor representativo (o típico) en un grupo de puntuaciones. La moda es el valor único más común en una distribución.

En el ejemplo de los alcaldes, la moda es 8, ya que tres alcaldes han servido 8 años en el cargo y no hay otro número de años con tantos alcaldes.

Otra forma de pensar en la moda es que es el valor con la mayor frecuencia en una tabla de frecuencias, o el punto más alto o pico de un histograma de la distribución.

**Mediana**
Otra medida diferente del valor representativo de un grupo de puntuaciones es la mediana. Si alineas todas las puntuaciones de menor a mayor, la puntuación que queda en el medio es la mediana.

En el ejemplo de los alcaldes, si ordenas los años en el cargo de menor a mayor, la quinta y sexta puntuaciones (las dos del medio) son ambas 7. Por lo tanto, la mediana es 7 años.

# Comparando Valores Representativos de una Distribución

A veces, la mediana es mejor que la media (o la moda) como valor representativo para un grupo de puntuaciones. Esto ocurre cuando unas pocas puntuaciones extremas afectarían fuertemente la media, pero no afectarían la mediana.

Por ejemplo, supongamos que en una comunidad, 99 ciudadanos tienen un ingreso anual de 500$USD y un ciudadano (el dueño de una gran empresa) tiene un ingreso anual de 450,500$USD. El ingreso promedio (digamos, PNB per cápita) de este país sería 5,000$USD ($99 \times 500 = 49,500$; $49,500 + 450,500 = 500,000$; $500,000/100 = 5,000$).

En este caso, el PNB per cápita (media) no refleja adecuadamente el ingreso típico de los ciudadanos, mientras que la mediana, que sería 500$USD, sí lo haría.

# Mediana vs. Media

Los científicos sociales y del comportamiento utilizan la mediana cuando hay puntuaciones extremas (en inglés, *outliers*) que harían que la media no sea representativa.

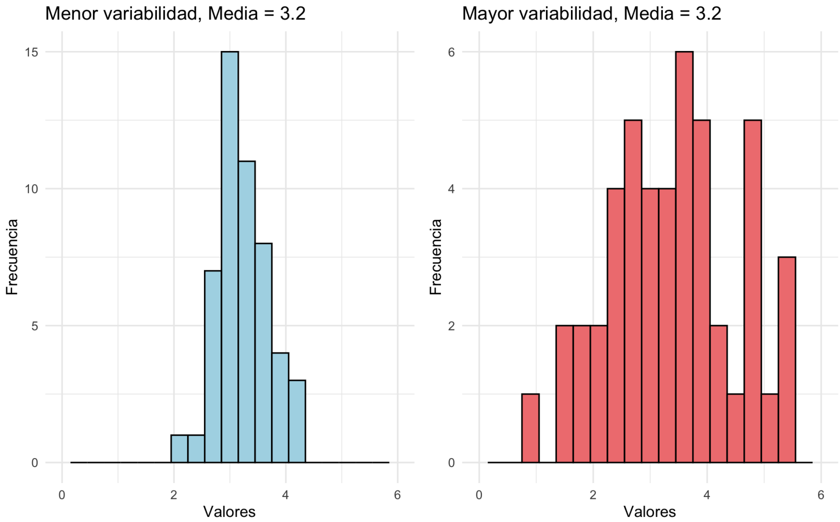
Un outlier es una puntuación extremadamente alta o baja en comparación con el resto de la distribución. Sin embargo, en ausencia de outliers, la media es casi siempre la medida preferida por su importancia en métodos estadísticos más complejos.

**Variabilidad**
Los investigadores también desean saber qué tan dispersas están las puntuaciones en una distribución. Esto muestra la cantidad de variabilidad en la distribución.

Por ejemplo, supongamos que te preguntan: “¿Qué edad tienen los estudiantes en tu clase de estadística?” En una universidad en una ciudad con muchos estudiantes a tiempo parcial, la edad promedio podría ser 34 años.

Sin embargo, esta media no cuenta toda la historia. Podrías tener un promedio de 34 si todos los estudiantes tienen exactamente 34 años (sin variabilidad). O bien, podrías tener un promedio de 34 porque la mitad de los estudiantes tiene 24 y la otra mitad 44 (con alta variabilidad).

# Variabilidad con la misma media



**Varianza**
La varianza de un grupo de puntuaciones te indica qué tan dispersas están las puntuaciones alrededor de la media. Para ser precisos, la varianza es el promedio de las diferencias al cuadrado de cada puntuación respecto a la media.

**Definición:**

- **Varianza:** Medida de cuán disperso está un conjunto de puntuaciones; promedio de las desviaciones al cuadrado respecto a la media.

# Cálculo de la Varianza

## Fórmula de la varianza

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

## Pasos para calcular la varianza:

1. **Restar la media de cada puntuación:** $(x_i - \mu)$. Esto te da la *desviación* de cada puntuación respecto a la media.
2. **Elevar al cuadrado cada una de estas desviaciones:** $(x_i - \mu)^2$. Esto elimina los signos negativos y amplifica las diferencias.
3. **Sumar todas las desviaciones al cuadrado:** $\sum(x_i - \mu)^2$. Este total se llama la *suma de desviaciones al cuadrado*.
4. **Dividir la suma de desviaciones al cuadrado entre el número de puntuaciones:** $\frac{1}{N}\sum(x_i - \mu)^2$. Esto te da la media de las desviaciones al cuadrado, conocida como *varianza*.

**Desviación estándar**
La forma más utilizada de describir la dispersión de un grupo de puntuaciones es la desviación estándar. La desviación estándar está directamente relacionada con la varianza y se calcula tomando la raíz cuadrada de la varianza.

**Pasos para calcular la desviación estándar:**

1. **Calcular la varianza:** $\sigma^2$.
2. **Tomar la raíz cuadrada:** La desviación estándar es la raíz cuadrada positiva de la varianza. (Cualquier número tiene tanto una raíz cuadrada positiva como una negativa. Por ejemplo, la raíz cuadrada de 9 es tanto +3 como -3, pero aquí usamos la raíz positiva).

**Fórmula:**

$$\sigma = \sqrt{\sigma^2}$$

### Desviaciones y Desviación Estándar

- **Puntuación de desviación:** Puntuación menos la media.
- **Puntuación de desviación al cuadrado:** Cuadrado de la diferencia entre una puntuación y la media.
- **Suma de las desviaciones al cuadrado:** Total de todas las puntuaciones de la diferencia al cuadrado de cada puntuación respecto a la media.
- **Desviación estándar:** Raíz cuadrada del promedio de las desviaciones al cuadrado respecto a la media; la estadística descriptiva más común para la variación; aproximadamente la cantidad promedio en que las puntuaciones en una distribución varían respecto a la media.

# ¿Por qué importa la variabilidad en las ciencias sociales y del comportamiento?

**Importancia de la Variabilidad**
La variabilidad es un tema crucial en la investigación en ciencias sociales y del comportamiento porque gran parte de esta investigación se centra en explicar la variabilidad.

**Ejemplo:** Diferentes estudiantes experimentan diferentes niveles de estrés al aprender estadísticas. Explicar la variabilidad significa identificar los factores que explican por qué los estudiantes difieren en la cantidad de estrés que experimentan.

**Posibles factores:**

- La cantidad de experiencia previa con matemáticas podría explicar parte de la variabilidad en el estrés.
- El grado en que los estudiantes generalmente experimentan estrés en sus vidas también podría explicar las diferencias en el estrés relacionado con el aprendizaje de estadísticas.

# Varianza como suma de desviaciones al cuadrado dividida por $N-1$

**Varianza:** $\frac{\sum(x_i-\mu)^2}{N-1}$ Los investigadores a menudo utilizan una definición ligeramente diferente de la varianza. Anteriormente, definimos la varianza como el promedio de las desviaciones al cuadrado, dividiendo la suma de las desviaciones al cuadrado entre el número de puntuaciones ($N$). Sin embargo, en muchos casos, es preferible dividir la suma de las desviaciones al cuadrado entre $N-1$.

**¿Por qué $N-1$?**

- Este enfoque se utiliza cuando tienes puntuaciones de un grupo específico (muestra) y deseas estimar la varianza para un grupo más amplio (población) que estos individuos representan.
- Las varianzas y desviaciones estándar en artículos de investigación suelen calcularse utilizando el enfoque de $N-1$.

**Importante:** Para propósitos descriptivos y el material de esta sesión (como los puntajes Z), dividir por $N$ está bien *por ahora*.

# ¿Qué es un Puntaje Z?

**Puntaje Z**

Un puntaje Z utiliza la media y la desviación estándar para describir una puntuación en particular. Específicamente, un puntaje Z es el número de desviaciones estándar que la puntuación real está por encima o por debajo de la media.

- Si la puntuación real está por encima de la media, el puntaje Z es positivo.
- Si la puntuación real está por debajo de la media, el puntaje Z es negativo.

**Definiciones:**

- **Puntaje Z:** El número de desviaciones estándar que una puntuación está por encima (o por debajo, si es negativo) de la media de su distribución.
- **Puntuación bruta:** Puntuación ordinaria (o cualquier otro número en una distribución antes de que se convierta en un puntaje Z o se transforme de otra manera).

# Ejemplo de Puntaje Z

**Ejemplo 1: Número de veces que Juan habló**
Supongamos que un especialista en desarrollo observó a Juan, de 3 años, en una situación de laboratorio estandarizada jugando con otros niños de la misma edad. Durante la observación, el especialista contó el número de veces que Juan habló con los otros niños. El resultado, tras varias observaciones, fue que Juan habló con otros niños unas 8 veces por hora de juego.

Sin un estándar de comparación, sería difícil sacar conclusiones. Sin embargo, se sabe por investigaciones previas que en condiciones similares el número promedio de veces que los niños hablan es 12, con una desviación estándar de 4. Claramente, Juan habló menos que los otros niños en general, pero no mucho menos. Juan tendría un puntaje Z de -1 ($M = 12$ y $SD = 4$, por lo tanto, una puntuación de 8 está 1 SD por debajo de la media).

**Ejemplo 2: Comparación de puntajes**
Otra ventaja de los puntajes Z es que se pueden convertir puntuaciones de variables completamente diferentes en puntajes Z y compararlos. En nuestro ejemplo, si los mismos niños también fueron evaluados en una prueba de habilidad lingüística, podríamos comparar directamente los puntajes Z de habilidad lingüística con los puntajes Z de las veces que hablaron con otros niños.
Supongamos que Juan obtuvo una puntuación de 100 en la prueba de habilidad lingüística. Si la media en esa prueba era 82 y la desviación estándar era 6, entonces Juan es mucho mejor que el promedio en habilidad lingüística, con un puntaje Z de +3.

**Fórmula Matemática del Puntaje Z**

$$Z = \frac{x_i - \mu}{\sigma}$$

**Donde:**

- $Z$ es el puntaje Z.
- $x_i$ es la puntuación individual (puntuación bruta).
- $\mu$ es la media de la distribución.
- $\sigma$ es la desviación estándar de la distribución.

**Interpretación:** El puntaje Z indica cuántas desviaciones estándar está una puntuación por encima o por debajo de la media.

# Pasos para convertir una puntuación bruta a un puntaje Z

**Pasos para calcular un puntaje Z**

1. **Calcular la puntuación de desviación:** Resta la media de la puntuación bruta.
2. **Calcular el puntaje Z:** Divide la puntuación de desviación por la desviación estándar.

**Ejemplo con Juan:**

1. **Calcular la puntuación de desviación:** Resta la media de la puntuación bruta. $8 - 12 = -4$.
2. **Calcular el puntaje Z:** Divide la puntuación de desviación por la desviación estándar. $-4/4 = -1$.

# Revertir un puntaje Z a una puntuación bruta

**Fórmula para revertir un puntaje Z**

$$x_i = Z \cdot \sigma + \mu$$

**Pasos para convertir un puntaje Z a una puntuación bruta**

1. **Calcular la puntuación de desviación:** Multiplica el puntaje Z por la desviación estándar.
2. **Calcular la puntuación bruta:** Suma la media a la puntuación de desviación.

**Ejemplo con Juan:**

1. **Calcular la puntuación de desviación:** Multiplica el puntaje Z por la desviación estándar. $3 \times 6 = 18$.
2. **Calcular la puntuación bruta:** Suma la media a la puntuación de desviación. $18 + 82 = 100$.

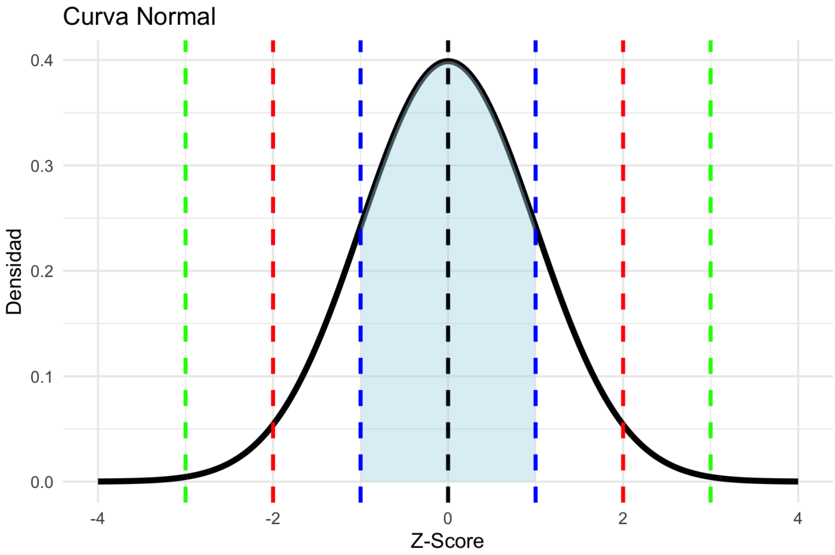
**La curva normal**

Las gráficas de muchas distribuciones de variables que los científicos del comportamiento y sociales estudian (así como muchas otras distribuciones en la naturaleza) siguen una distribución unimodal, aproximadamente simétrica y con forma de campana.

Estos histogramas en forma de campana o polígonos de frecuencia se aproximan a una distribución precisa e importante llamada distribución normal o, más simplemente, la curva normal. La curva normal es una distribución matemática (o teórica).

Los investigadores a menudo comparan las distribuciones reales de las variables que estudian con la curva normal. Aunque no esperan que las distribuciones de sus variables coincidan perfectamente con la curva normal (ya que la curva normal es una distribución teórica), los investigadores suelen verificar si sus variables aproximadamente siguen una curva normal.

# Curva normal



**Distribución Normal**

- **Distribución de Frecuencia Normal:** Una distribución de frecuencias que sigue una curva normal.
- **Curva Normal:** Una distribución de frecuencias específica, definida matemáticamente, con forma de campana que es simétrica y unimodal; las distribuciones observadas en la naturaleza y en la investigación a menudo se aproximan a ella.

**Características Clave:**

- Simétrica respecto a su media.
- Tiene una sola moda (unimodal).
- La mayoría de las puntuaciones se agrupan cerca de la media, con menos puntuaciones en los extremos.

# Por qué la Curva Normal es común en la naturaleza

- Las influencias aleatorias positivas y negativas tienden a cancelarse entre sí, haciendo que la mayoría de los valores se agrupen cerca de la media en una distribución unimodal y simétrica.
- La curva normal describe cómo muchas variables naturales se distribuyen debido al principio del teorema del límite central, donde la suma de una gran cantidad de influencias aleatorias produce una distribución normal aproximada.
- En una curva normal, aproximadamente el 68 % de las puntuaciones se encuentran dentro de 1 desviación estándar de la media, y aproximadamente el 95 % se encuentran dentro de 2 desviaciones estándar.

# La tabla de la Curva Normal y Puntajes Z

- Los valores del 50 %, 34 % y 14 % son útiles para trabajar con distribuciones normales, pero los científicos sociales a menudo necesitan información más precisa.
- La curva normal es una curva matemática precisa, lo que permite calcular el porcentaje exacto de puntuaciones entre cualesquiera dos puntos, no sólo en 1 o 2 desviaciones estándar.
- Las tablas de la curva normal proporcionan los porcentajes de puntuaciones entre la media ($Z = 0$) y cualquier otro puntaje Z, simplificando los cálculos.

# Curva normal

**Table 1** Normal Curve Areas: Percentage of the Normal Curve Between the Mean and the $Z$ Scores Shown and Percentage of Scores in the Tail for the $Z$ Scores Shown. (First part of table only. Highlighted values are examples from the text.)

| $Z$ | % Mean to $Z$ | % in Tail | $Z$ | % Mean to $Z$ | % in Tail |
|---|---|---|---|---|---|
| .00 | .00 | 50.00 | .45 | 17.36 | 32.64 |
| .01 | .40 | 49.60 | .46 | 17.72 | 32.28 |
| .02 | .80 | 49.20 | .47 | 18.08 | 31.92 |
| .03 | 1.20 | 48.80 | .48 | 18.44 | 31.56 |
| .04 | 1.60 | 48.40 | .49 | 18.79 | 31.21 |
| .05 | 1.99 | 48.01 | .50 | 19.15 | 30.85 |
| .06 | 2.39 | 47.61 | .51 | 19.50 | 30.50 |
| .07 | 2.79 | 47.21 | .52 | 19.85 | 30.15 |
| .08 | 3.19 | 46.81 | .53 | 20.19 | 29.81 |
| .09 | 3.59 | 46.41 | .54 | 20.54 | 29.46 |
| .10 | 3.98 | 46.02 | .55 | 20.88 | 29.12 |
| .11 | 4.38 | 45.62 | .56 | 21.23 | 28.77 |
| .12 | 4.78 | 45.22 | .57 | 21.57 | 28.43 |
| .13 | 5.17 | 44.83 | .58 | 21.90 | 28.10 |
| .14 | 5.57 | 44.43 | .59 | 22.24 | 27.76 |
| .15 | 5.96 | 44.04 | .60 | 22.57 | 27.43 |
| .16 | 6.36 | 43.64 | .61 | 22.91 | 27.09 |
| .17 | 6.75 | 43.25 | .62 | 23.24 | 26.76 |
| .18 | 7.14 | 42.86 | .63 | 23.57 | 26.43 |
| .19 | 7.53 | 42.47 | .64 | 23.89 | 26.11 |
| .20 | 7.93 | 42.07 | .65 | 24.22 | 25.78 |
| .21 | 8.32 | 41.68 | .66 | 24.54 | 25.46 |

La tabla de la curva normal muestra los puntajes Z, el porcentaje de puntuaciones entre la media y ese puntaje Z, y el porcentaje de puntuaciones en la cola. Dado que la curva normal es simétrica, los valores para los puntajes Z positivos y negativos son idénticos, y los porcentajes de “Media a Z” y “En la cola” siempre suman 50 %.

# Pasos para Calcular el Porcentaje de Puntuaciones con la Tabla de la Curva Normal

1. Si comienzas con una puntuación en bruto, conviértela primero a un puntaje Z usando la fórmula $Z = \frac{x_i - \mu}{\sigma}$.
2. Dibuja una curva normal, ubica el puntaje Z y sombrea el área cuya proporción deseas encontrar.
3. Haz una estimación aproximada del porcentaje del área sombreada:
   - Recuerda que el 50 % del área está a la izquierda de la media (Z=0).
   - El 34 % del área está entre la media y Z=1, y un 14 % adicional está entre Z=1 y Z=2.
4. Encuentra el porcentaje exacto usando la tabla de la curva normal, sumando 50 % si es necesario.
5. Verifica que el porcentaje exacto esté dentro del rango estimado en el paso 3.

# Empezando con Muestra y Población

## Muestra y Población

- **Población:** Es el conjunto completo de elementos de interés. Por ejemplo, en una olla de habichuelas, la población sería todas las habichuelas en la olla.
- **Muestra:** Es una parte de la población de la cual obtenemos información. En el ejemplo, una cucharada de habichuelas representa la muestra.
- En la investigación en ciencias sociales y del comportamiento, trabajamos con muestras (como 50 estudiantes de escuela elemental en un experimento) para hacer inferencias sobre una población más grande (como todos los estudiantes de escuela elemental del país).

## Ejemplo de Encuesta:

- Una muestra podría ser 1,000 personas seleccionadas para una encuesta de opinión.
- La población sería el público votante en general, al que los encuestadores esperan aplicar los resultados.

Continuaremos de lleno con este tema el martes que viene.