Estadísticos y Descripciones de Datos

UCR – ECCI

CI-0115 Probabilidad y Estadística

Prof. Kryscia Daviana Ramírez Benavides

Muestreo Aleatorio

- En este tipo de muestreo, todos los individuos de la población pueden formar parte de la muestra, tienen una probabilidad positiva.
- El resultado de un experimento estadístico se puede registrar como un valor numérico o como una representación descriptiva.
  - Cuando se lanza un par de dados y el total es el resultado de interés, se registra un valor numérico.
  - Cuando a los estudiantes de cierta escuela se les hace pruebas de sangre y el tipo sanguíneo es de interés, se registra una representación descriptiva.
- En cualquier estudio, el número de observaciones posibles puede ser pequeño, grande pero finito o infinito.

Muestreo Aleatorio (cont.)

- Una **población** consiste en la totalidad de las observaciones en las que se está interesado.
  - Conjunto de todos los elementos que cumplen una determinada característica.
  - Conjunto de todos los valores de una variable aleatoria.
- Los elementos de la población se llaman observaciones, individuos o unidades estadísticas.
- El número de observaciones en la población se define como el **tamaño de la población**.
  - El número total de observaciones puede ser finito o infinito.

Muestreo Aleatorio (cont.)

- La **variable estadística** es una propiedad característica de la población que estamos interesados en estudiar.
- Tipos de variables estadísticas:
  - **Cualitativa:** No se expresa mediante un número. Por ejemplo, el tipo sanguíneo de los estudiantes de cierta escuela.
  - **Cuantitativa:** Se expresa mediante un número, hay dos tipos:
    - **Cuantitativa Discreta:** Sólo admite valores aislados, toma un número determinado de valores. Por ejemplo, el resultado total que se obtiene a lanzar dos dados.
    - **Cuantitativa Continua:** Puede admitir cualquier valor dentro de un intervalo, puede tomar cualquier valor entre los valores dados. Por ejemplo, medir la presión atmosférica cada día del pasado al futuro.

Muestreo Aleatorio (cont.)

- Una variable estadística cualitativa se puede convertir a una variable aleatoria discreta, para poner realizar su estudio y análisis.
- Cada observación en una población es un valor de una variable aleatoria X que tiene alguna distribución de probabilidad $f(x)$.
  - Se puede hablar de población binomial, población normal, o en general, **la población $f(x)$**, para referirse a una población cuyas observaciones son valores de una variable aleatoria que tiene una distribución binomial, una distribución normal o una distribución $f(x)$.
  - Por lo tanto, la media y la varianza de una variable aleatoria o distribución de probabilidad también se les denomina la media y la varianza de la población correspondiente.

Muestreo Aleatorio (cont.)

- En el campo de la inferencia estadística el estadístico se interesa en llegar a conclusiones con respecto a la población cuando es imposible o poco práctico observar todo el conjunto de observaciones que constituyen la población.
  - La población de una producción de cierto producto, sería imposible probar toda la producción si se tienen que vender.
  - Los costos exorbitantes también pueden ser un factor prohibitivo para estudiar toda la población.
- Por lo que se depende de un subconjunto de observaciones para hacer inferencias con respecto a la población.
- Una **muestra** es un subconjunto de una población.

Muestreo Aleatorio (cont.)

- Si se quiere inferencias válidas a partir de la muestra para la población, se debe obtener muestras que sean representativas de la población.
- Cualquier procedimiento de muestreo que produzca inferencias que sobreestimen o subestimen de forma consistente alguna característica de la población se dice que está **sesgado**.
- Para evitar cualquier posibilidad de sesgo en el procedimiento de muestreo, es deseado elegir una **muestra aleatoria** en el sentido de que las observaciones se realizan de forma independiente y al azar.

Muestreo Aleatorio (cont.)

- Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes, cada una con la misma distribución de probabilidad $f(x)$. Se define entonces a $X_1, X_2, \ldots, X_n$ como una muestra aleatoria de tamaño n de la población $f(x)$ y se escribe su distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \ldots f(x_n)$$

Algunos Estadísticos Importantes

- El propósito principal al seleccionar muestras aleatorias es obtener información acerca de los parámetros desconocidos de la población.
- Por ejemplo, se quiere saber la proporción de una población que toman una marca de café determinada.
  - Aquí se podría preguntar a cada uno de los bebedores de café de la población en cuestión, si toman la marca de café.
  - En su lugar, se selecciona una muestra aleatoria grande y se calcula la proporción $\hat{p}$ de personas que prefieren la marcad de café.
  - El valor $\hat{p}$ se utiliza ahora para hacer una inferencia con respecto a la proporción *p* verdadera.

Algunos Estadísticos Importantes (cont.)

- Ahora, $\hat{p}$ es una función de los valores observados en la muestra aleatoria; como son posibles muchas muestras aleatorias a partir de la misma población, se espera que $\hat{p}$ variara algo de una muestra a otra.
- Es decir, $\hat{p}$ es un valor de una variable aleatoria que representamos con P.
- Tal variable aleatoria se llama **estadístico**, la cual se puede definir como cualquier función de las variables aleatorias que forman una muestra aleatoria.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de tamaño n, entonces la **media de la muestra** se define mediante el estadístico

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- Si el estadístico $\overline{X}$ toma el valor $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

  cuando X_1 toma el valor de x_1, X_2 toma el valor de x_2, y así sucesivamente.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de tamaño n, acomodada en orden creciente de magnitud, entonces la **mediana de la muestra** se define mediante el estadístico

$$\tilde{X} = \begin{cases} X_{(n+1)/2} & \text{si } n \text{ es impar} \\ \dfrac{X_{n/2} + X_{(n/2)+1}}{2} & \text{si } n \text{ es par} \end{cases}$$

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- Si $X_1, X_2, \ldots, X_n$, no necesariamente diferentes, representan una muestra aleatoria de tamaño n, entonces la **moda de la muestra *M*** es aquel valor de la muestra que ocurre más a menudo o con mayor frecuencia.
- La moda puede no existir, y cuando existe no necesariamente es única.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- La media de la muestra:
  - Es la medida de localización central más comúnmente utilizada en estadística.
  - Emplea toda la información disponible.
  - Las distribuciones de medias que se obtienen en muestreos repetidos de una población son bien conocidos, y en consecuencia los métodos que se utilizan en la inferencia estadística para estimar μ se basan en la media de la muestra.
  - La única desventaja real, es que puede resultar afectada de manera adversa por valores extremos.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- La mediana de la muestra:
  - Es fácil de calcular si el número de observaciones es relativamente pequeño.
  - No resulta influida por valores extremos.
  - Al tratar con muestras que se seleccionan de poblaciones, las medias de las muestras por lo general no variarán tanto de una muestra a otra como las medianas. Por lo tanto, si se desea estimar el centro de una población con base en un valor de la muestra, la media es más estable que la mediana.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- La moda de la muestra:
  - Es la menos utilizada de las tres.
  - Para conjuntos pequeños su valor casi no tiene utilidad, si es que existe.
  - Sólo tiene sentido significativo en una gran cantidad de datos.
  - No requiere cálculo, lo que se considera una ventaja.
  - Se puede usar para datos cualitativos como cuantitativos, lo que se considera una ventaja.

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- Las medidas de localización central o posición no dan por sí mismas una descripción adecuada de los datos. Es importante conocer cómo se dispersan las observaciones del promedio.
- La variabilidad de una muestra juega un papel muy importante en el análisis de datos.
  - La variabilidad de un proceso y de un producto es un hecho real en los sistemas científicos y de ingeniería.
  - La variabilidad en valores de población y datos de una muestra es un hecho real.

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- El **rango** (recorrido o amplitud) de una muestra aleatoria X_1, X_2, …, X_n, se define con el estadístico $X_{\max} - X_{\min}$, donde $X_{\min}$ y $X_{\max}$ son, respectivamente, las observaciones más grande y más pequeña de la muestra.
- El rango falla al medir la variabilidad entre la observación superior y la inferior, pero tiene algunas aplicaciones útiles.
- En la industria, el rango se puede determinar al especificar por adelantado que una medición particular de los artículos que salen de una línea de producción deba caer dentro de cierto intervalo.

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de tamaño n, entonces la **varianza de la muestra** se define mediante el estadístico

$$S^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{(n-1)}$$

- El valor calculado de S^2 para una muestra dada se denota con s^2.
- La varianza se define, esencialmente, como el promedio de los cuadrados de las desviaciones de las observaciones de su media.

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- **Teorema.** Si S^2 es la varianza de una muestra aleatoria de tamaño n, se puede escribir como

$$S^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- La **desviación estándar de la muestra**, que se denota con *S*, es la raíz cuadrada positiva de la varianza de la muestra.

$$S = \sqrt{\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{(n-1)}}$$

- La cantidad $n - 1$ a menudo se denomina **grados de libertad asociados con la varianza** estimada. Los grados de libertad representan el número de piezas de información independientes disponibles para calcular la variabilidad.

Presentaciones de Datos y Métodos Gráficos

- En la estadística, con frecuencia se hace la suposición de que la distribución es normal.
- La información gráfica con respecta a la validez de esta suposición se puede obtener de presentaciones como los diagramas de tronco y hojas, y los histogramas de frecuencias.
- A continuación se introduce la noción de gráficas de probabilidad normal y gráficas de cuantiles.
  - Estas gráficas se utilizan en estudios que tienen grados de complejidad que varían, con el objetivo principal de que las gráficas proporcionen una verificación diagnóstica de la suposición de que los datos vienen de una distribución normal.

Presentaciones de Datos y Métodos Gráficos (cont.)

- Los estadísticos vistas anteriormente proporcionan medidas simples, mientras que una representación gráfica agrega información adicional en términos de una imagen.
  - Las muestras múltiples se pueden comparar de forma gráfica.
  - Las gráficas de datos pueden sugerir relaciones entre variables.
  - Las gráficas pueden ayudar en la detección de anomalías o de observaciones de datos apartados en las muestras.

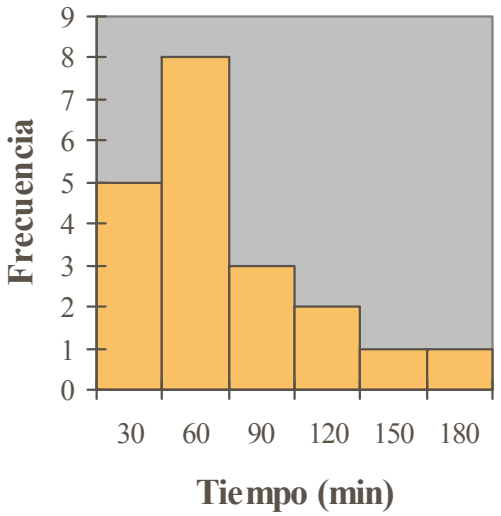
Presentaciones de Datos y Métodos Gráficos Histogramas (cont.)

- Los histogramas de frecuencia son similares a un diagrama de barras, sólo que en este caso, las barras ocupan todo el ancho del intervalo al que van asociadas, pudiendo estar pegadas unas con otras (algo que nunca podía pasar en un diagrama de barras).
  - Realmente, a diferencia de lo que pasaba en un diagrama de barras, en los histogramas de frecuencias, el área de cada rectángulo debe ser proporcional a la frecuencia relativa, lo que pasa es que si tomamos la precaución de tomar todos los intervalos con la misma amplitud, entonces no tendremos que preocuparnos, por ser la razón de proporcionalidad siempre la misma: la amplitud del intervalo.

Presentaciones de Datos y Métodos Gráficos
Histogramas (cont.)

- Tiempo que emplean los alumnos del curso en ir desde casa a la universidad se distribuye de la siguiente manera:

Tiempo (min)	Frecuencia
[0,30[	5
[30,60[	8
[60,90[	3
[90,120[	2
[120,150[	1
[150,180[	1



Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)

- Esta gráfica encierra el **rango intercuartil** de los datos en una caja que tiene la mediana representada dentro.
- El rango intercuartil tiene como extremos el percentil 75 (cuartil superior) y el percentil 25 (cuartil inferior).
- Además, de la caja se prolongan extensiones, que muestran las **observaciones extremas** en la muestra.
- Para muestras razonablemente grandes, la presentación muestra el centro de la localización, la variabilidad y el grado de asimetría.

Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)

- Una variación que se llama **gráfica de caja** puede proporcionar a quien la ve información con respecto a cuales observaciones **son datos apartados**.
  - Los datos apartados son observaciones que se consideran inusualmente alejadas de la masa de datos.
  - Técnicamente, se puede considerar un dato apartado como una observación que representa un "evento raro"; es decir, existe una probabilidad pequeña de obtener un valor tan alejado de la masa de datos.

Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)

- **Ejemplo.** Los valores de nicotina de 40 cigarrillos son:

1,09	1,92	2,31	1,79	2,28
1,74	1,47	1,97	0,85	1,24
1,58	2,03	1,70	2,17	2,55
2,11	1,86	1,90	1,68	1,51
1,64	0,72	1,69	1,85	1,82
1,79	2,46	1,88	2,08	1,67
1,37	1,93	1,40	1,64	2,09
1,75	1,63	2,37	1,75	1,69

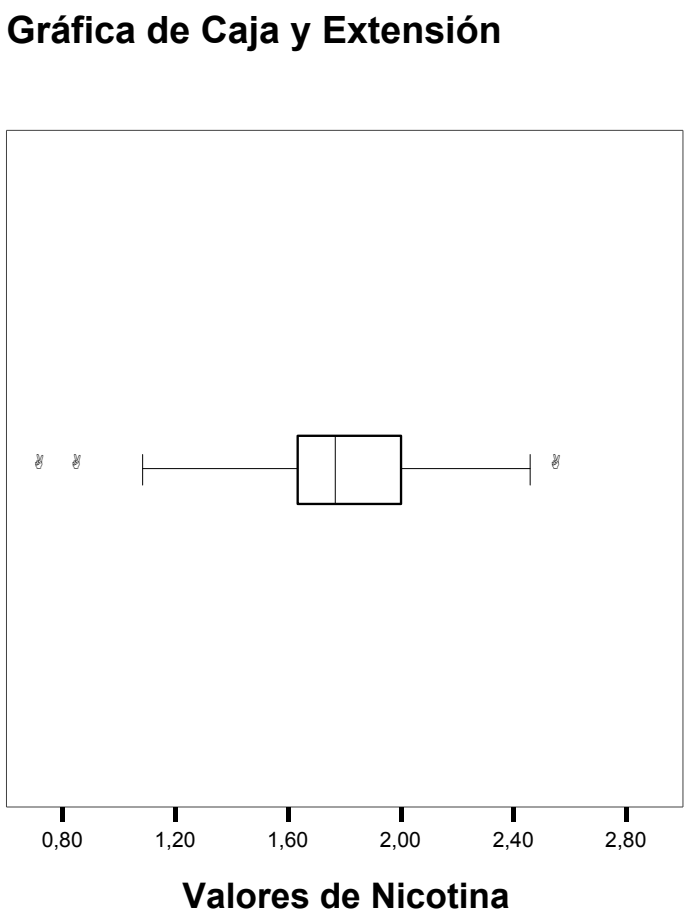
Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)

- Se tienen las siguientes estadísticas:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Valores de Nicotina	40	1,83	,72	2,55	1,7743	,39046	,152

Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)



Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles (cont.)

- El propósito de estas gráficas es describir, en forma de muestra, la función de distribución acumulada que se presentó en capítulos anteriores.
- Un **cuantil** de una muestra, $q(f)$, es un valor para el que una fracción específica f de los valores de los datos es menor que o igual a $q(f)$.
- Un cuantil representa una estimación de una característica de una población, o más bien, la distribución teórica.
- La mediana de la muestra es $q(0.5)$, el cuartil superior (percentil 75) es $q(0.75)$ y el cuartil inferior (percentil 25) es $q(0.25)$.

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles (cont.)

- Una gráfica de cuantiles simplemente grafica los valores de los datos en el eje vertical contra una evaluación empírica de la fracción de observaciones excedidas por los valores de los datos.
- Para la propósitos teóricos esta fracción se calcula con

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$$

donde i es el orden de las observaciones cuando se clasifican de inferior a superior.

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles (cont.)

- A diferencia de la gráfica de caja y extensión, la gráfica de cuantiles realmente muestra todas las observaciones.
- Todos los cuantiles, incluida la mediana y los cuantiles inferior y superior, se pueden aproximar de forma visual.
- Las indicaciones de agrupaciones relativamente grandes alrededor de valores específicos se indican por pendientes cercanas a cero, mientras que los datos dispersos en ciertas áreas producen pendientes más abruptas.

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles-Cuantiles Normales (cont.)

- La gráfica de cuantiles-cuantiles normales toma ventaja de lo que se conoce acerca de los cuantiles de la distribución normal.
- La metodología incluye una gráfica de los cuantiles empíricos recién presentados contra el cuantil correspondiente de la distribución normal.

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles-Cuantiles Normales (cont.)

- La expresión para un cuantil de una variable aleatoria $N(\mu,\sigma)$ es muy complicada. Una buena aproximación está dada por:

$$q_{\mu,\sigma}(f) = \mu + \sigma\left\{4.91\left[f^{0.14} - (1-f)^{0.14}\right]\right\}$$

- La expresión para un cuantil de una variable aleatoria $N(0,1)$ es:

$$q_{0,1}(f) = \left\{4.91\left[f^{0.14} - (1-f)^{0.14}\right]\right\}$$

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles (cont.)

- La gráfica de cuantiles-cuantiles normales es una gráfica de $y_{(i)}$ (observaciones ordenadas) contra $q_{0,1}(f_i)$, donde

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$$

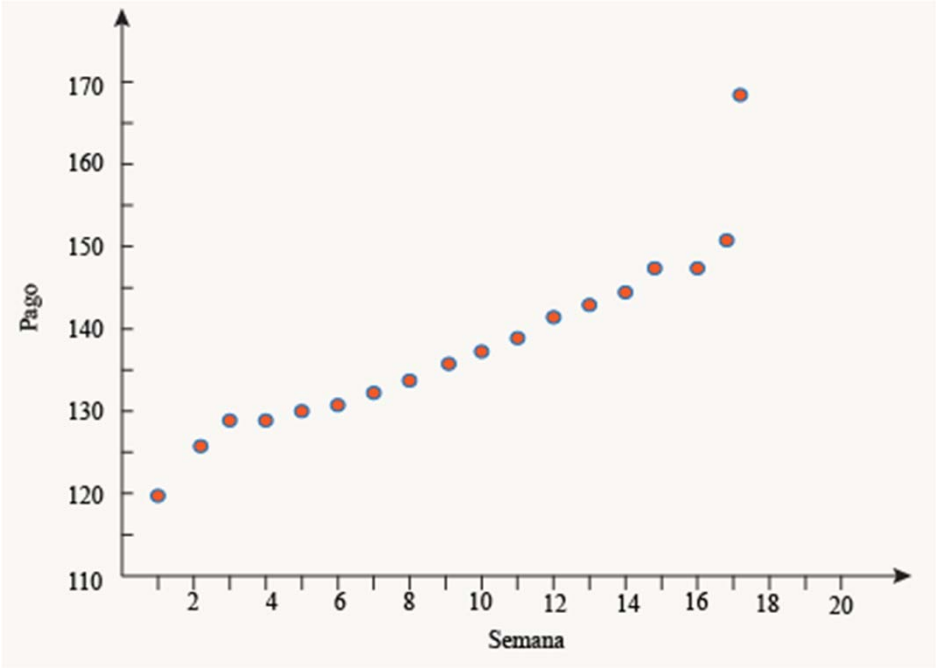
- Una relación cercana a una línea recta sugiere que los datos provienen de una distribución normal.
- La intersección en el eje vertical es una estimación de la media de la población y la pendiente es una estimación de la desviación estándar.

Presentaciones de Datos y Métodos Gráficos – Diagramas de Dispersión y Correlación (cont.)

- Los diagramas de dispersión son una forma de expresar datos de dos variables, y hacer predicciones basadas en los datos.
- Al contrario de los histogramas y los diagramas de caja, los de dispersión muestran valores de datos individuales.

Presentaciones de Datos y Métodos Gráficos – Diagramas de Dispersión y Correlación (cont.)

- El diagrama de dispersión que expresa la cantidad de dinero que se ganó Mateo cada semana trabajando en la tienda de su padre.



Presentaciones de Datos y Métodos Gráficos – Diagramas de Dispersión y Correlación (cont.)

- Las semanas están diagramadas en el eje *x*, y la cantidad de dinero que se ganó en esa semana en el eje *y*.
  - En general, la variable independiente (la variable que no está influenciada por nada) está en el eje *x* y la variable dependiente (la que es modificada por la variable independiente) está en el eje *y*.
- En este diagrama se puede ver que en la semana 2 Mateo se ganó alrededor de $125, y en la semana 18 estuvo cerca de los $165. Pero más importante aún es la tendencia.
  - Por ejemplo, con estos datos podemos ver que Mateo gana cada vez más según pasan las semanas. Quizá su padre le da más horas a la semana o más responsabilidades.

Presentaciones de Datos y Métodos Gráficos – Diagramas de Dispersión y Correlación (cont.)

- Con los diagramas de dispersión se puede ver cómo se relacionan ambas variables entre sí.
  - Esto es lo que se conoce como **correlación**.
- Hay tres tipos de correlación:
  - Correlación positiva.
  - Correlación negativa.
  - Sin correlación (nula).
- El diagrama de dispersión que se analizó tiene una fuerte correlación positiva: a medida que las semanas aumentan, su pago también.

Presentaciones de Datos y Métodos Gráficos – Diagramas de Dispersión y Correlación (cont.)

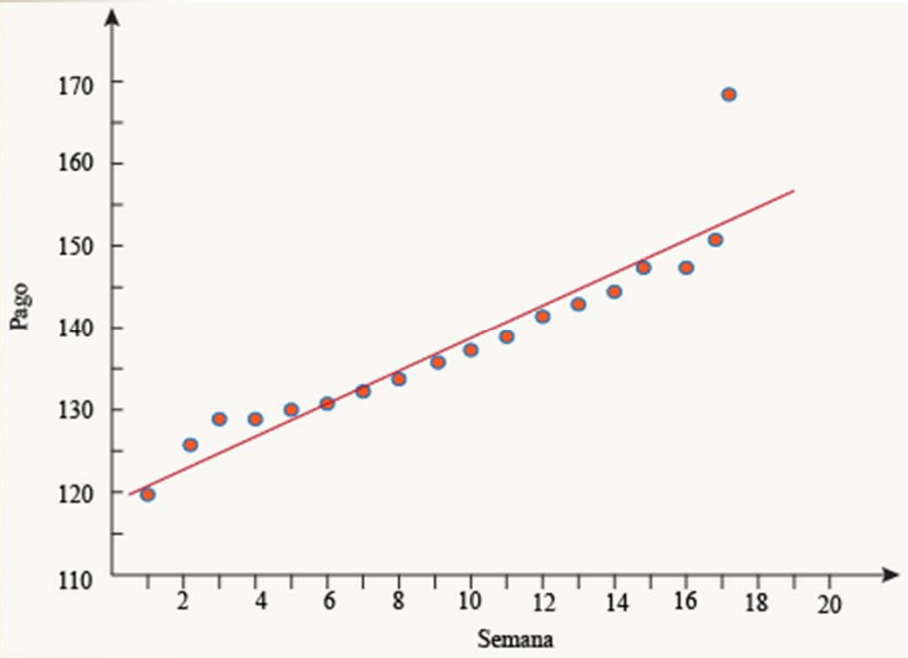
- Hay tres tipos de correlación:
  - **Correlación positiva.** Ocurre cuando una variable aumenta y la otra también. Por ejemplo: la altura de una persona y el tamaño de su pie; mientras aumenta la altura, el pie también.
  - **Correlación negativa.** Ocurre cuando una variable aumenta y la otra disminuye. Por ejemplo: el tiempo de estudio y el tiempo que pasas jugando videojuegos, tienen una correlación negativa, ya que cuando tu tiempo de estudio aumenta, no te queda tanto tiempo para jugar videojuegos.
  - **Sin correlación (nula).** No hay una relación aparente entre las variables. Por ejemplo: los puntos en tus videojuegos y tu talla de zapato no parece tener ninguna correlación; mientras una aumenta, la otra no tiene ningún efecto.

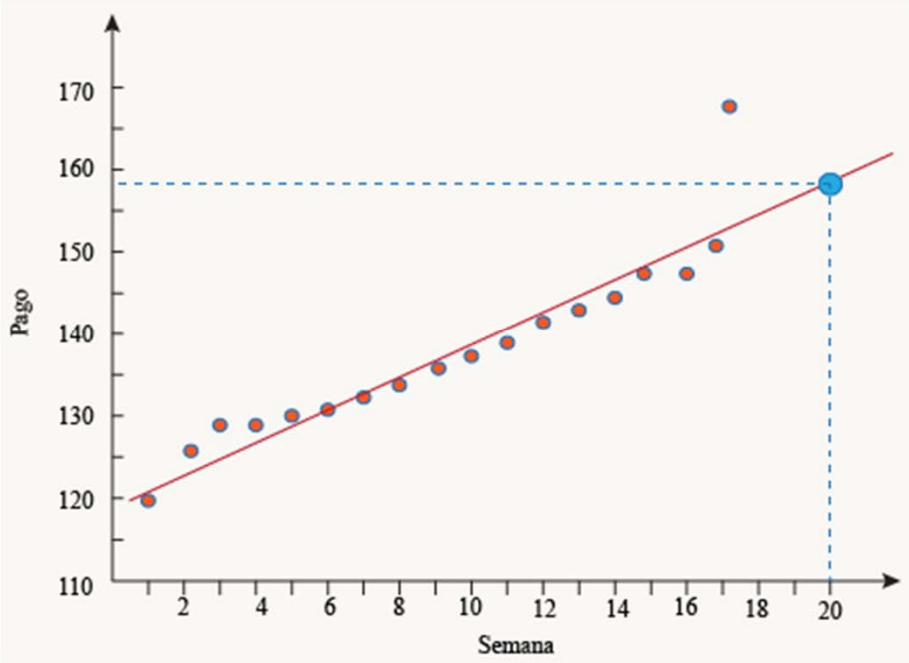
Presentaciones de Datos y Métodos Gráficos – Diagramas de Dispersión y Correlación (cont.)

- Se usa la **línea de ajuste** para hacer predicciones basándose en datos pasados.
- Hay muchas y muy complicadas fórmulas para encontrar esta recta, pero por ahora solo la dibujaremos a través de los puntos en la gráfica para que se ajuste a la tendencia que marcan los datos.
- Cuando se dibuja la recta, se debe asegurar de que encaje con la mayor parte de los datos.
  - Si hay un punto que está muy por encima o muy por debajo con respecto al resto (los atípicos) déjalo fuera de la recta.

Presentaciones de Datos y Métodos Gráficos – Diagramas de Dispersión y Correlación (cont.)

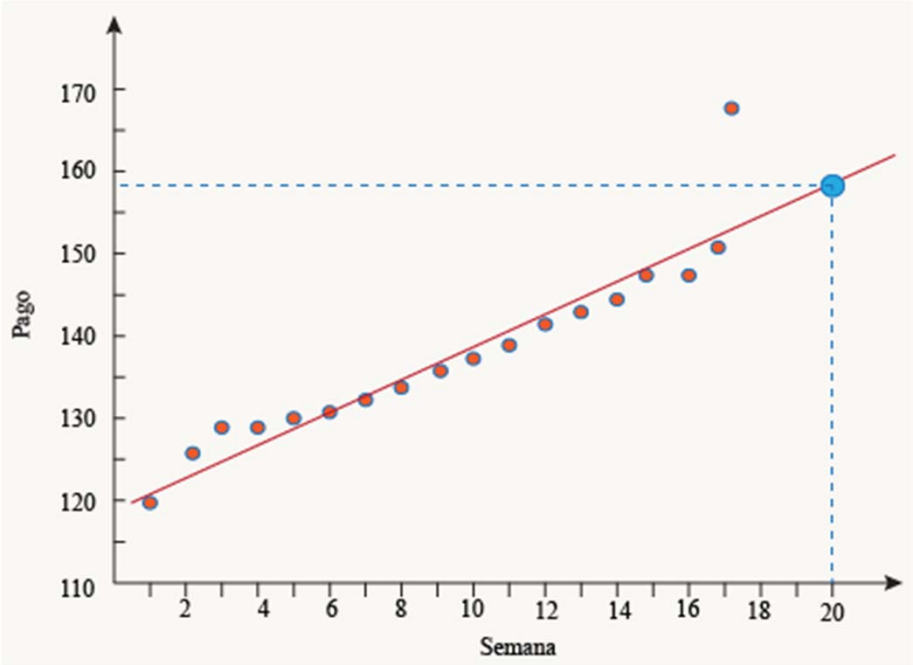
- Al usar la línea de ajuste se puede predecir cuánto dinero se ganará Mateo en 20 semanas de trabajo (asumiendo que el patrón continua.





Presentaciones de Datos y Métodos Gráficos – Diagramas de Dispersión y Correlación (cont.)

- En el ejemplo, Mateo se ganará, aproximadamente, $157 en la semana 20.



Referencias Bibliográficas

- Walpole, R.E.; Myers, R.H.; Myers, S.L. & Ye, K. “Probabilidad y estadística para ingeniería y ciencias”. Octava Edición. Pearson Prentice-Hall. México, 2007.