DMC | INSTITUTE

Dip. Data Sciencie

Curso:

Estadística Descriptiva

Sesión 03

Docente: Nilton Yanac

Enero, 2026

REGLAS

Se requiere **puntualidad** para un mejor desarrollo del curso.

Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.

Las preguntas se realizarán **a través del chat** y en caso de que lo requieran **podrán activar el micrófono**.

Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.

Identificarse en la sala Zoom con el primer nombre y primer apellido.

ITINERARIO

07:00 PM – 07:30 PM	***Soporte técnico DMC***
07:30 PM – 08:50 PM	***Agenda***
08:50 PM – 09:00 PM	***Pausa Activa***
09:00 PM – 10:30 PM	***Agenda***

Horario de Atención Área Académica y Soporte

Lunes a Viernes 09:00 am a 10:30 pm / Sábado 09:00 am a 02:00pm

SILABO

Objetivo del curso:

Conocer las principales metodologías de análisis de datos para la toma de decisiones en los negocios

Agenda de la sesión 03:

- *Tema 01: Repaso de la sesión 02*
- *Tema 02: Estandarización y Normalización de Datos*
- *Tema 03: Estimación puntual y por intervalos*
- *Tema 04: Correlación y Regresión*
- *Tema 05: Práctica en Python*

TEMA 01:
REPASO DE LA SESIÓN 02

Proceso del análisis exploratorio de datos

TEMA 02: ESTANDARIZACIÓN Y NORMALIZACIÓN DE DATOS

Ley de los Grandes Números

Si se obtiene una muestra aleatoria de una población que obedece a cualquier modelo probabilístico, discreto o continuo, entonces si la muestra es suficientemente grande, el promedio muestral se aproxima al promedio poblacional.

La ley de los grandes números nos indica que, si se toman muchas observaciones, a la larga los errores tienden a compensarse.

Ley de los Grandes Números

Concepto

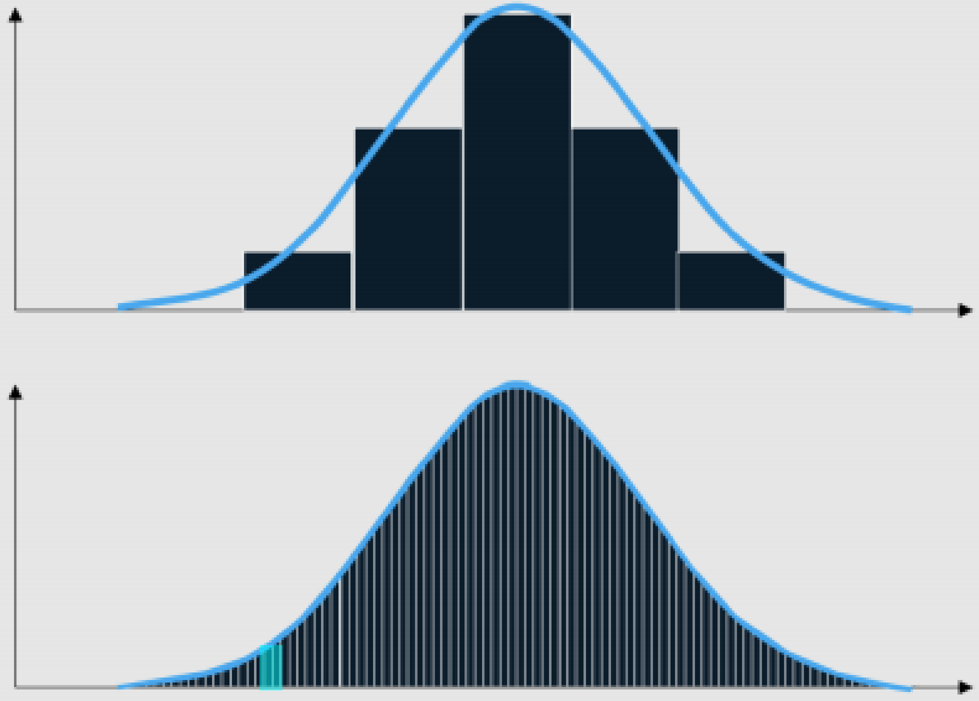
"un promedio de muchas mediciones es más preciso que una sola medición"

Formalmente: $X_1, X_2, \ldots$ son variables aleatorias *(i.i.d)* con igual media y varianza

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum X_i$$

Entonces podemos saber que $\overline{X_n}$ converge en probabilidad a la media de X

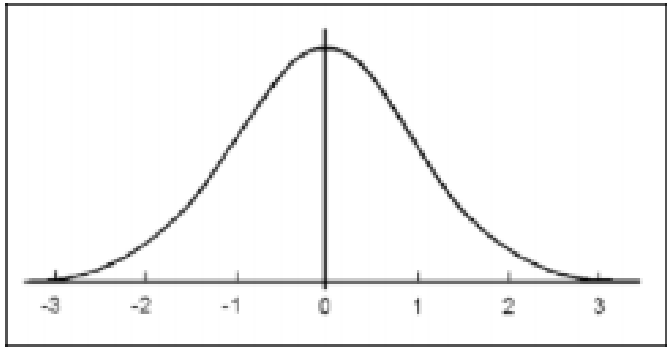
Implica que el histograma converge a la función *pdf*



Estandarización de Datos

Transformación que expresa la variable en términos de la media (promedio) y la desviación estándar (medida de variación).

$$z = (X - \mu) / \sigma$$



Para crear los números aleatorios en base a la distribución normal, usaremos el paquete NumPy.

•¿Cuándo usarlo?:

- Cuando los datos tienen una distribución aproximadamente normal.
- Cuando trabajas con algoritmos que **suponen normalidad en los datos**, como **regresión lineal, regresión logística, SVM, PCA y redes neuronales profundas**.

Normalización de Datos

La normalización transforma los valores para que estén dentro de un rango específico, típicamente [0,1] o [−1,1]. La fórmula es:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

¿Cuándo usarlo?:

- Cuando los datos no siguen una distribución normal.
- Cuando los valores tienen diferentes escalas y quieres llevarlos a un mismo rango (ejemplo: precios en miles y edades en años).
- Algoritmos sensibles a la escala de datos como **KNN, redes neuronales y SVM con kernel basado en distancia**.

Normalidad de Datos

La distribución normal es una distribución de probabilidad que es simétrica y tiene forma de campana.

En una distribución normal:

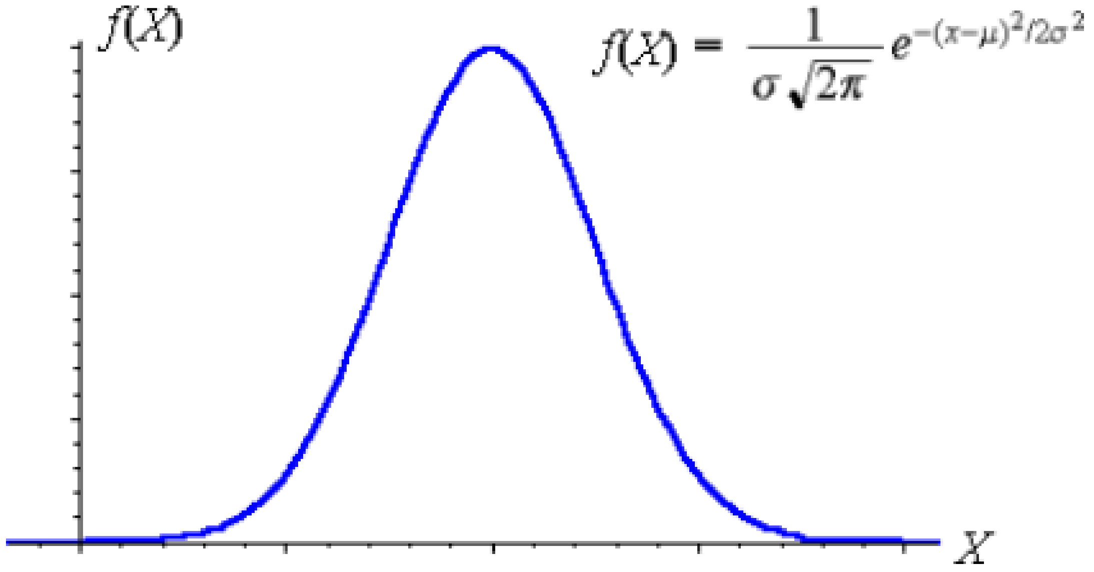
- La media, la mediana y la moda son iguales.
- La curva es simétrica alrededor de la media.
- Aproximadamente el 68% de los datos se encuentran dentro de un desvío estándar de la media.

Distribución Normal

Conocida como la distribución gaussiana, es la distribución continua de uso más común en la estadística, debido a que la inferencia estadística clásica se soporta mucho sobre ella.

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \mu &= \text{ Mean} \\ \sigma &= \text{ Standard Deviation} \\ \pi &\approx 3.14159\cdots \\ e &\approx 2.71828\cdots \end{aligned}$$



TALLER:
APLICACIÓN DE EDA Y PREPARACIÓN PARA REGRESIÓN LINEAL PARA DETERMINAR SALARIOS:

Descarguen el notebook colab para ejecutarlo en clase

TEMA 03: ESTIMACIÓN

Estimación: Concepto

Proceso de utilizar información de una muestra para extraer conclusiones acerca de toda la población

Se utiliza la información para estimar un valor

Tipos de Estimación

PUNTUAL: Se obtiene un único número al que se le puede asignar un punto de la recta

POR INTERVALOS: Se obtienen dos puntos que representan un límite inferior y superior (li, ls)

Estimación: Propiedades

- No tener sesgos (Término asociado al error)
- Poca variabilidad de una muestra a otra

En una Estimación de debe considerar...

Un intervalo

Espacio que tiene una cierta probabilidad de contener el verdadero valor del parámetro desconocido

Una medida de confianza

Coeficiente de confianza $= 1 - \alpha$

Nivel de confianza $= 100*(1 - \alpha)\%$

Intervalos de Confianza

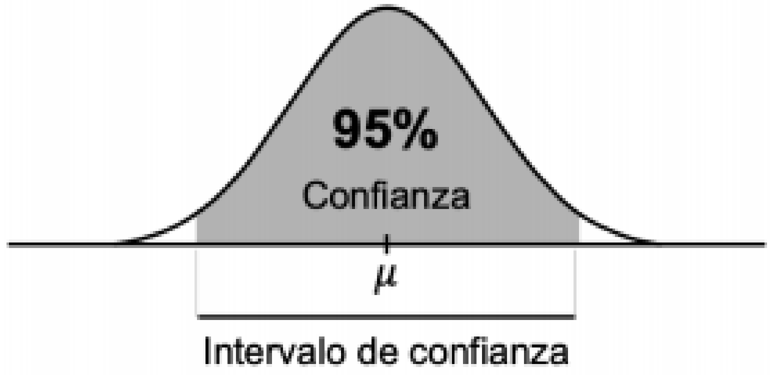
Concepto

Debido a que la estadística frecuentista trata la estimación de parámetros como una estimación puntual, no tiene una aproximación inmediata al concepto de incertidumbre en la estimación.

Como solución a esto surge el concepto de intervalo de confianza. Este concepto está atado a la variabilidad (σ-desviación estándar) de nuestros datos muestrales y el nivel de confianza (α) que queremos tener.

$$\text{Inter. Confianza} = \left(\mu - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \mu + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$

El intervalo de confianza resultante es:



La desviación estándar σ es usada como una medida de probabilidad

Reglas simples:

$P(-1 \leq Z \leq 1) \approx 68$

$P(-2 \leq Z \leq 2) \approx 95$

$P(-3 \leq Z \leq 3) \approx 99.7$

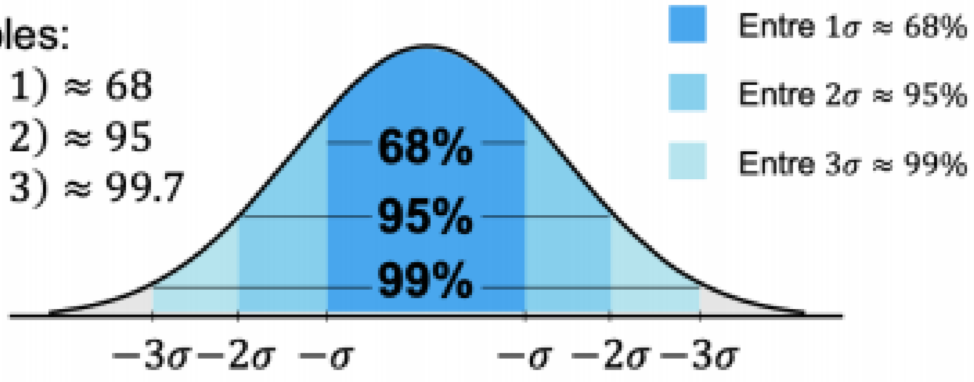


Tabla de Probabilidades

http://www.est.uc3m.es/esp/nueva_docencia/comp_col_leg/ing_tec_inf_gestion/estadistica/Documentacion/Tablas/tablas2caras.pdf

El valor de la tabla para z es el área bajo la curva de la normal estándar a la izquierda de z

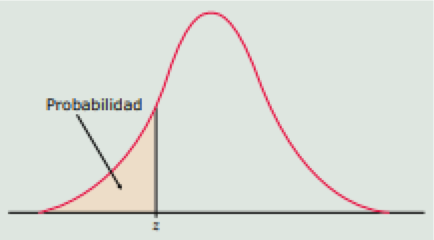


TABLA A: Probabilidades de la normal estándar

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

El valor de la tabla para z es el área bajo la curva de la normal estándar a la izquierda de z

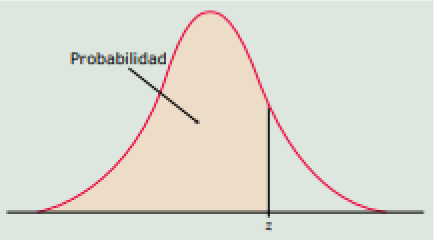


TABLA A: Probabilidades de la normal estándar (cont.)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Intervalos de Confianza

- Se pueden crear para cualquier parámetro de la población.

EJEMPLOS

- Media: tiempo medio de recuperación
- Proporción: de niños que sufren apendicitis
- Desviación estándar: del error de medida de un aparato médico

Intervalos de Confianza para la Media

Factor relacionado con la confianza

Parámetro: Media Poblacional

Error Estándar

Estimo

$$IC\,95\% = \left[\ \bar{x} - t_{n-1}\frac{s_{n-1}}{\sqrt{n}} \ \leq\ \mu \ \leq\ \bar{x} + t_{n-1}\frac{s_{n-1}}{\sqrt{n}}\ \right]$$

Nivel de confianza

Límites de confianza

Intervalos de Confianza para Proporciones

Parámetro: Prevalencia Poblacional

Error Estándar

$$IC\,95\% = \left[\, p - z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \;\le\; \pi \;\le\; p + z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}} \,\right]$$

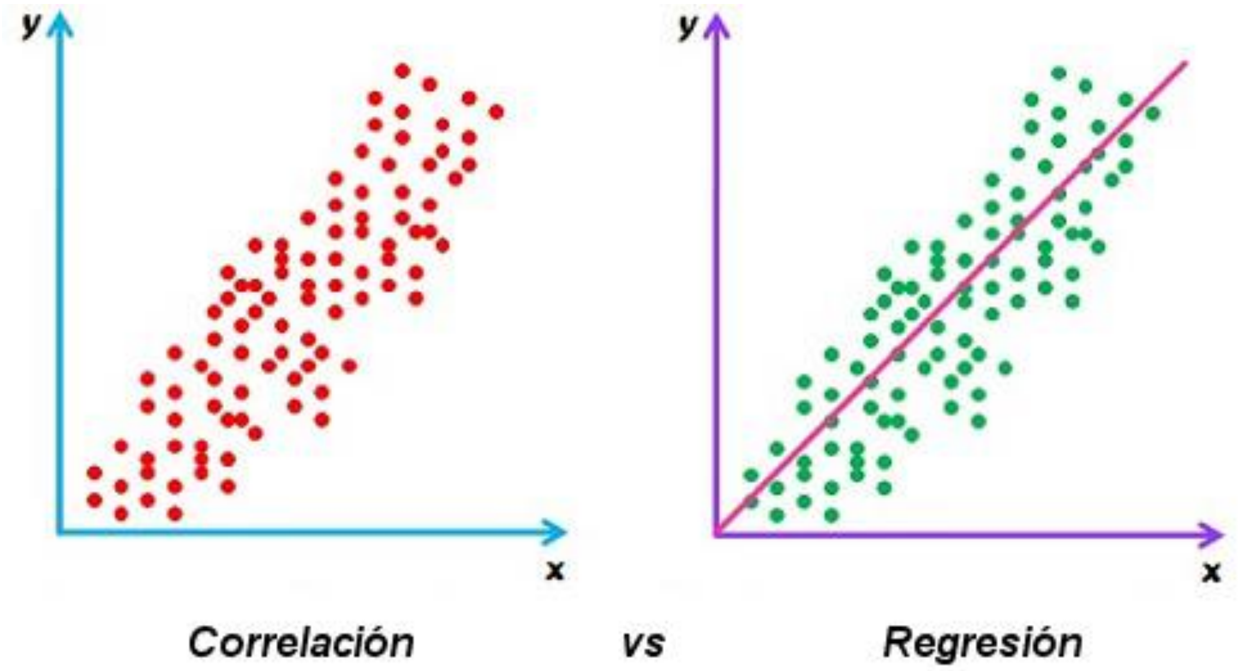
Nivel de confianza

Límites de confianza

TEMA 04:
CORRELACIÓN Y REGRESIÓN

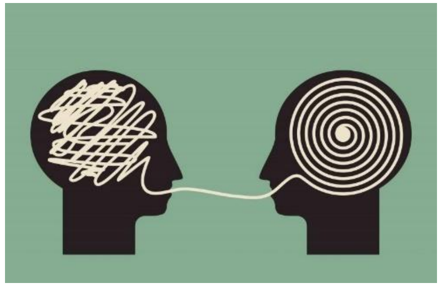
Regresión y Correlación

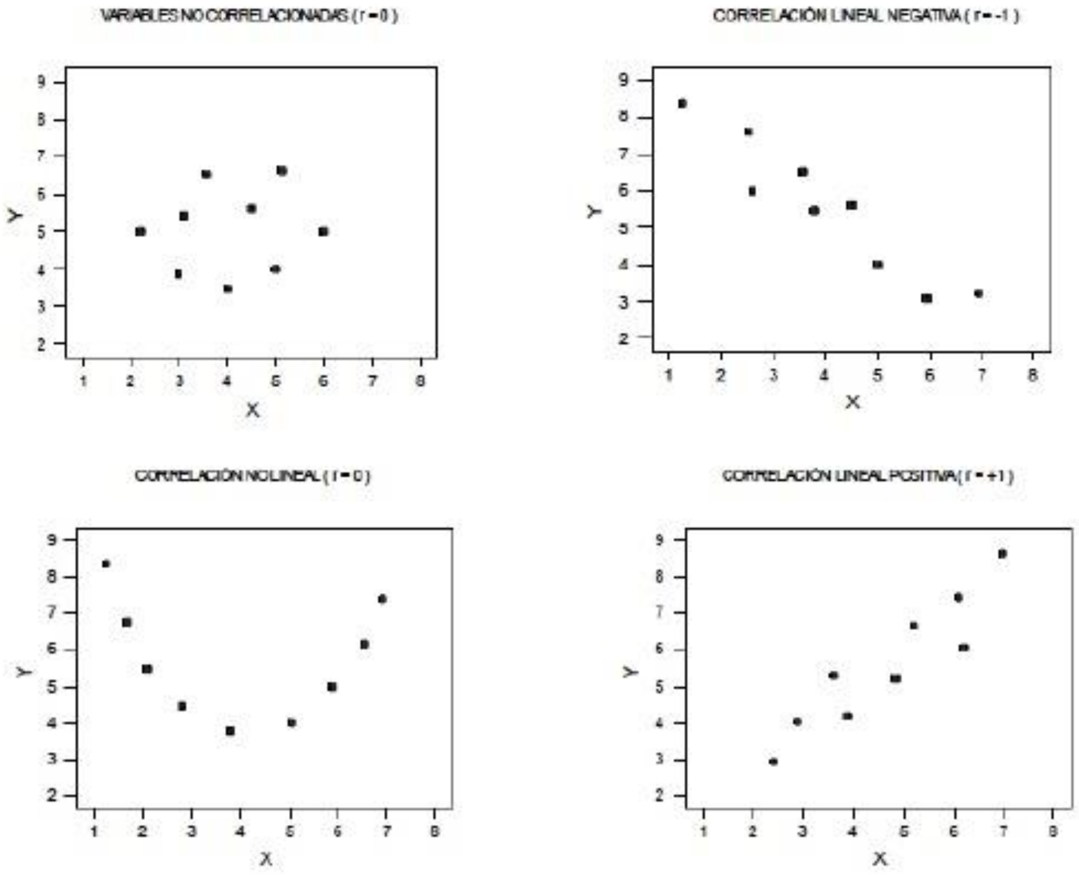
La regresión y la correlación son dos técnicas estrechamente relacionadas y comprenden una forma de estimación

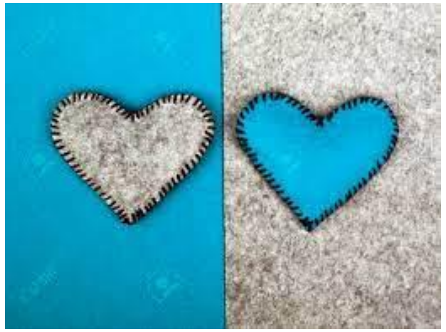


INSTITUTE

Tipos de Correlación







Correlación lineal: Consideraciones

- La correlación cuantifica cuan relacionadas están dos variables
- El cálculo de la correlación entre dos variables es independiente del orden o asignación de cada variable a XX e YY, mide únicamente la relación entre ambas sin considerar dependencias.
- A nivel experimental, la correlación se suele emplear cuando ninguna de las variables se ha controlado, simplemente se han medido ambas y se desea saber si están relacionadas

Interpretación del Coeficiente de Correlación

Coeficiente	Interpretación
r = 1	Correlación perfecta
0.80 < r < 1	Muy alta
0.60 < r < 0.80	Alta
0.40 < r < 0.60	Moderada
0.20 < r < 0.40	Baja
0 < r < 0.20	Muy baja
r = 0	Nula

Coeficientes de Correlación:

Coeficiente de correlación de Pearson

Tiene el objetivo de explicar la asociación que existen entre variables cuantitativas

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

- σ_{xy} es la covarianza de (X,Y)
- σ_x es la desviación estándar de la variable X
- σ_y es la desviación estándar de la variable Y

¿Cuál uso?

Pearson → Normalidad

Spearman → No Normalidad

Kendall → No Normalidad*

Regresión Lineal

El análisis de regresión lineal múltiple es el primer modelo en el cuál pensar para predecir una variable en función de otra o ver pesos e impacto, no es efectivo en todos los casos, pero si es una propuesta a evaluar casi siempre. Los coeficientes de cada aspecto nos darán su importancia en la satisfacción general

Ecuación de Regresión

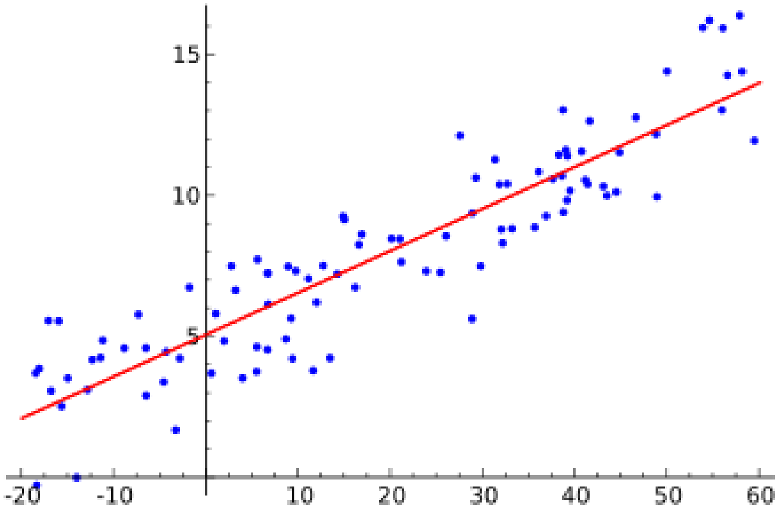
$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

Variable Dependiente

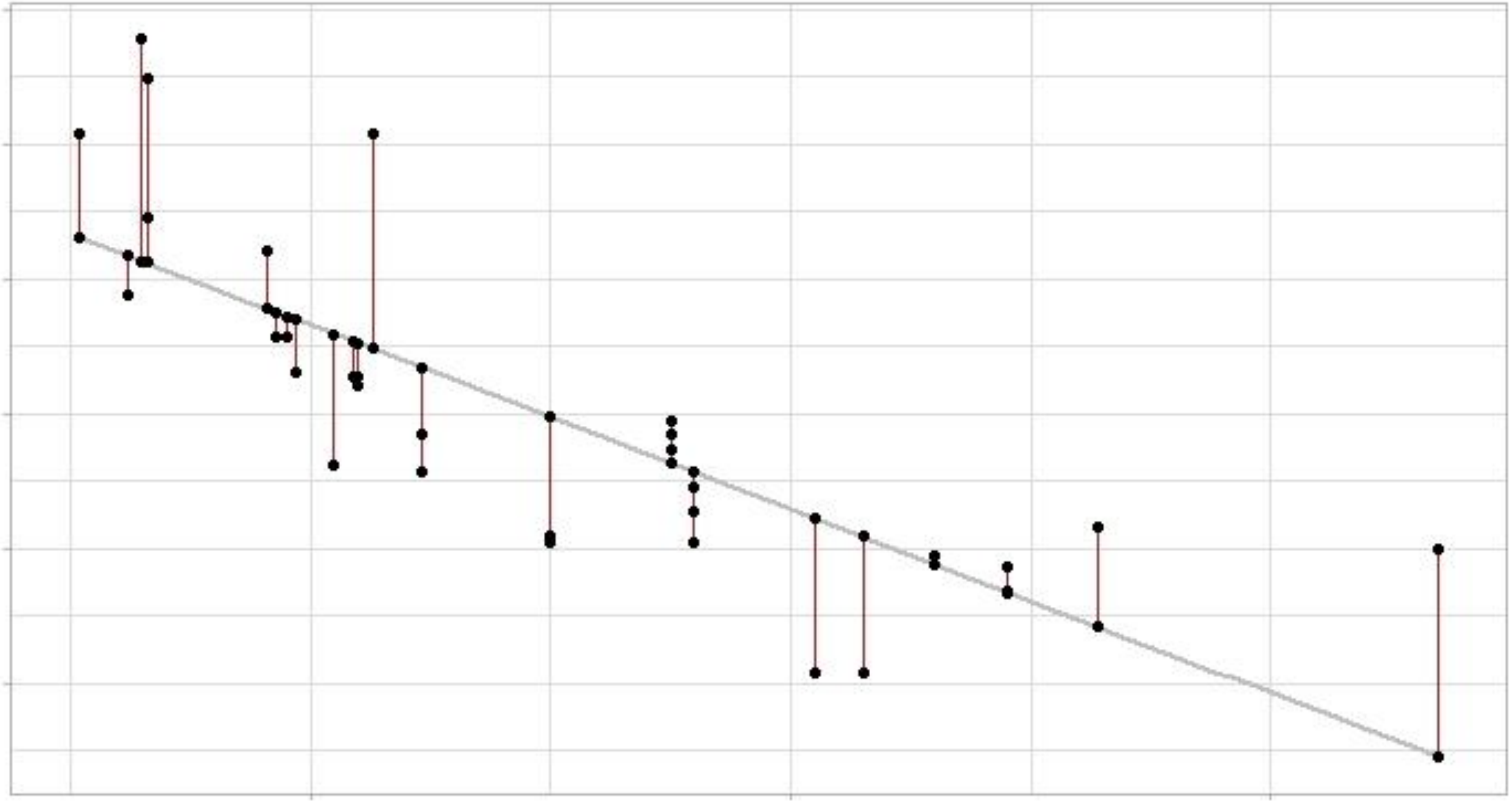
Coeficiente que suple aspectos no medidos

Coeficientes de cada aspecto medido

Error aleatorio en la predicción del modelo



Regresión: Importancia de los errores



Coeficiente de Determinación R2

$$r^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{SCR}{SCtot}$$

El coeficiente de determinación puede interpretarse como la **proporción de variabilidad de *Y* que es explicada por *X***. Mide la proximidad de la recta ajustada a los valores observados de Y.

- **R2 = 0 → El modelo no explica nada. X no puede explicar Y.**
- **R2 cerca a 1 → El modelo es apropiado y X explica a Y.**
- **R2 cerca a 0 → El modelo es débil o X no explica del todo a Y.**

Condiciones para la Regresión Lineal

- **Multicolinealidad (No colinealidad):** no debe haber relación lineal entre los predictores.
- **Relación Lineal entre los predictores numéricos y la variable respuesta:** debe existir una relación lineal entre la variable respuesta “Y” y cada uno de los predictores sin afectar al resto.
- **Distribución normal de la variable respuesta**: debe tener una distribución normal bajo test de hipótesis de normalidad y gráficas como histogramas.
- **Homocedasticidad (Varianza constante de la variable respuesta):** se grafican los residuos para identificar si la variable respuesta es constante en todo el rango de los predictores.
- **Independencia (No autocorrelación):** principalmente cuando se trabaja con series de tiempo, los valores de cada observación deben ser independientes de los otros.
- **Valores Atípicos:** identificarlos a través de los residuos y excluirlos.
- **Tamaño de la muestra:** usar la regla práctica, por cada variable predictora se debe tener como mínimo 20 casos.

Intentar entender al otro significa destruir los clichés que lo rodean, sin negar ni borrar su alteridad.

Umberto Eco

INSTITUTE

¡Gracias… Totales!

Docente: Nilton Yanac
Enero, 2026