Modelos estadísticos de medición:

Teoría clásica del test

Dr. Héctor Nájera
PUED-UNAM

¿Cómo llegamos aquí?

Progresión no lineal de la historia, acuerdos y desacuerdos de lo que significa medir en ciencias

- Siglo XX: Teoría representacional
- Siglo XXI: Medición basada en modelos

Mientras tanto…

La clase pasada Medición basada en modelos

Medición basada en modelos

Fenómenos
(ante los ojos)

≠

Objetos científicos

Fenómenos
(ante los ojos)

≠

Observación
(codificada)

Puntajes

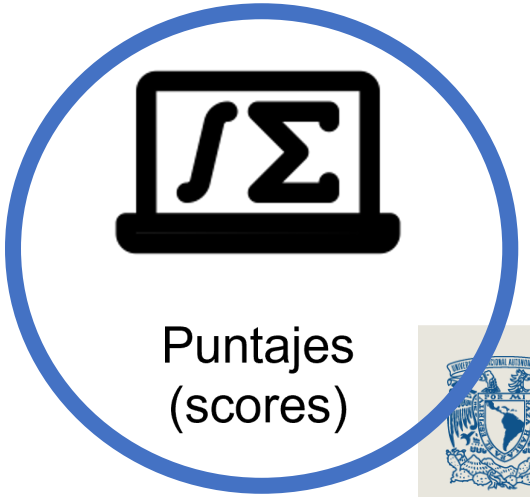
≠

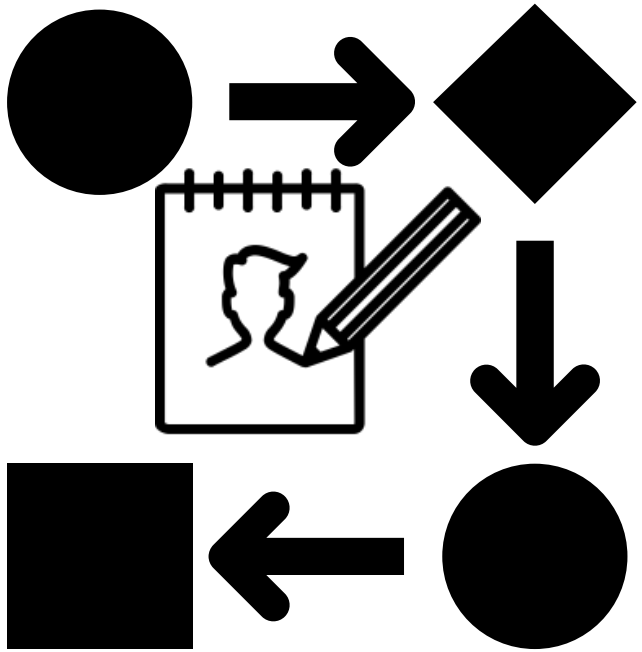
Objetos científicos

Datos

≠

Estimadores





Modelo de medición

- Una representación abstracta y local construida a partir de supuestos simplificadores
- Hipótesis teóricas sobre las relaciones que guardan los instrumentos con aquello que se quiere medir y con el ambiente ([DAG] sobre cómo fueron producidos los datos)
- Modelo teórico y estadístico del proceso de medición mismo
- Permite la rastreabilidad/trazabilidad de la generación de los resultados de la medición (a lo largo de cada eslabón de la cadena) en su relación con aquello que se quiere medir
- Establece relaciones **cuantitativas** entre aquello que se quiere medir y el resultado de su medición
- Generativos: genera instancias de datos (input-output de acuerdo con el proceso de medición idealizado)

Modelo de medición

Sólo bajo el modelo es posible evaluar la *interpretabilidad representacional* de los puntajes (su validez)

- Coherencia de los supuestos con las teorías contextuales relevantes
- Consistencia mutua de resultados con diferentes instrumentos, ambientes y modelos

Sin modelo no hay medición

Modelo teórico

Definición

Relación con otros conceptos

Hipótesis

Modelo estadístico

Hipótesis pasadas a parámetros

Métodos

Este curso es sobre SEM como modelo estadístico

Orígenes de la teoría clásica como modelo estadístico de medición

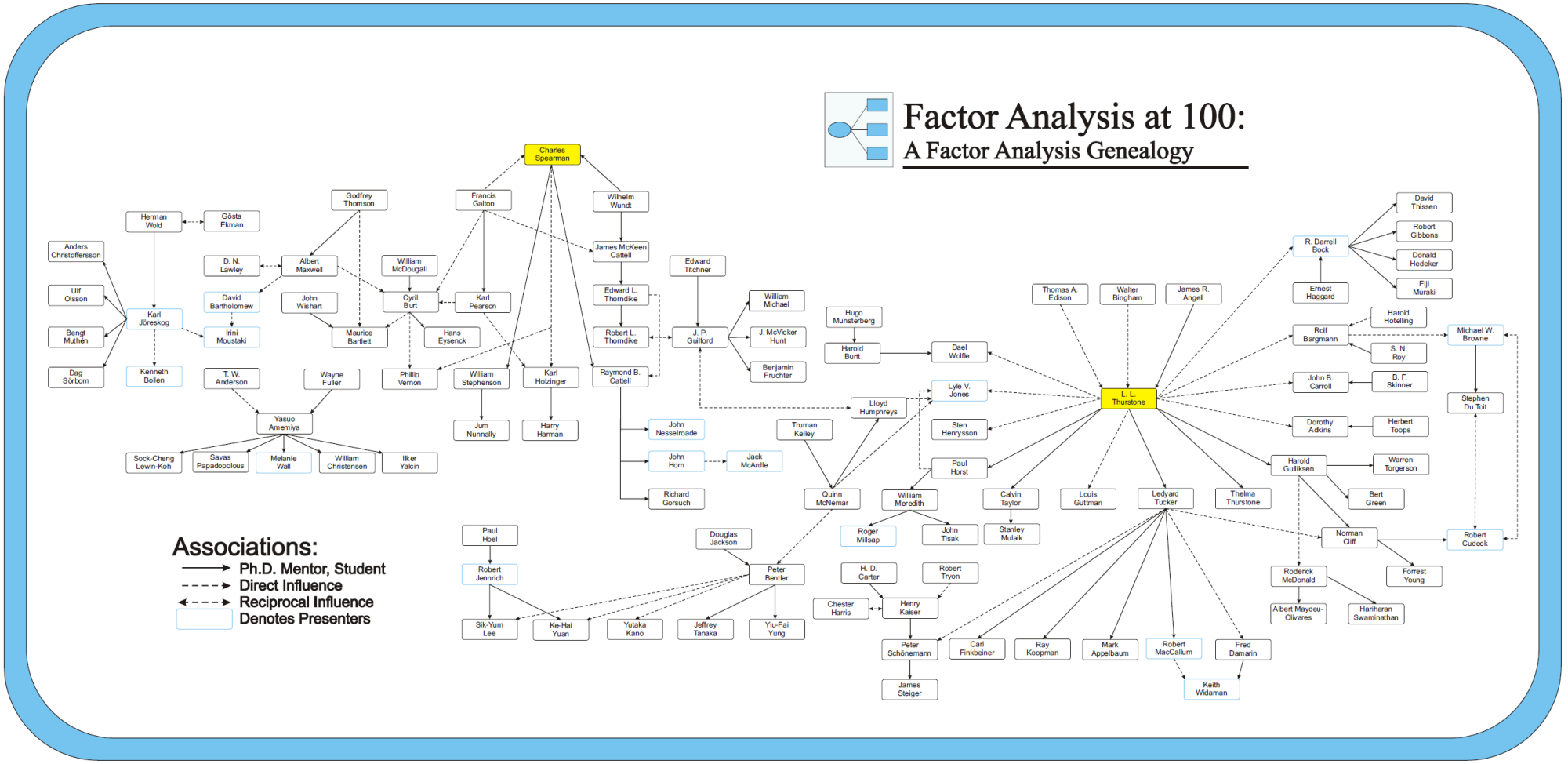
La historia de esta teoría comienza a la vuelta del siglo XX (1904)

- Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-292.
- Spearman, C. (1904). The proof and measurement of association between two things. *American journal of Psychology*, *15*(1), 72-101.

Charles Edward Spearman
1863-1945

15,000 citas en GS!

Origen la teoría clásica del test y SEM como modelos estadísticos de medición

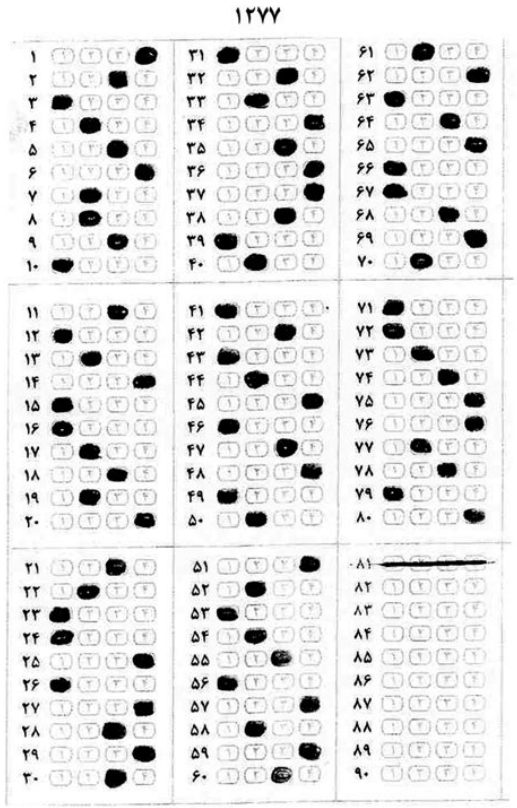


Tres ideas

- Atenuación de correlación por error
- El error como variable aletoria
- Distinción entre variables observables (con error) y variable latente

Pilares de Teoría clásica del test y eventualmente de ecuaciones estructurales

Habilidad
A

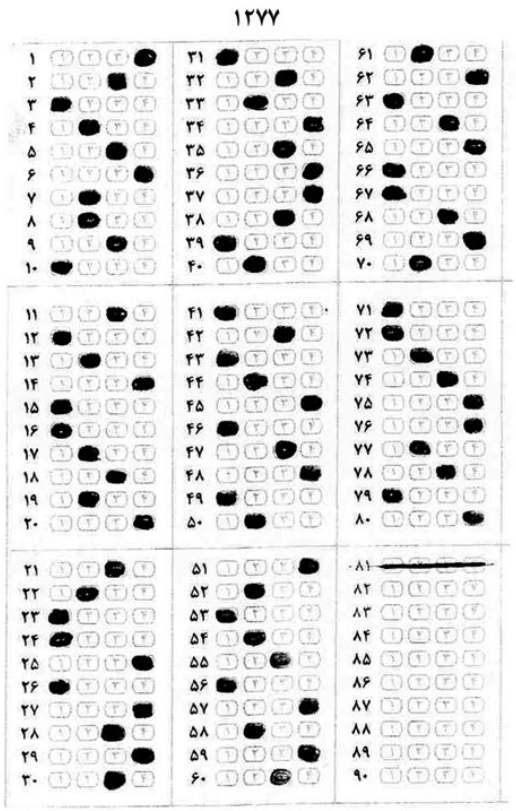


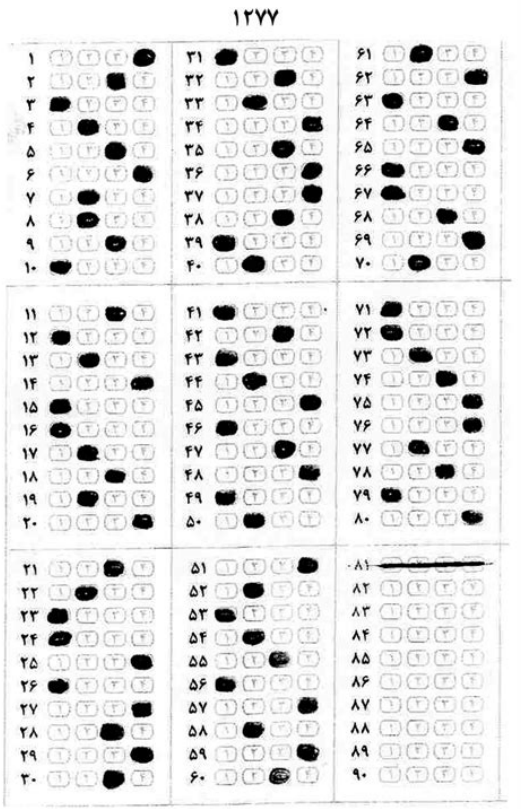
Si las respuestas rastrean/son causa de la misma “señal”, deberían estar correlacionadas “intra-sujetos”

Habilidad A

≈

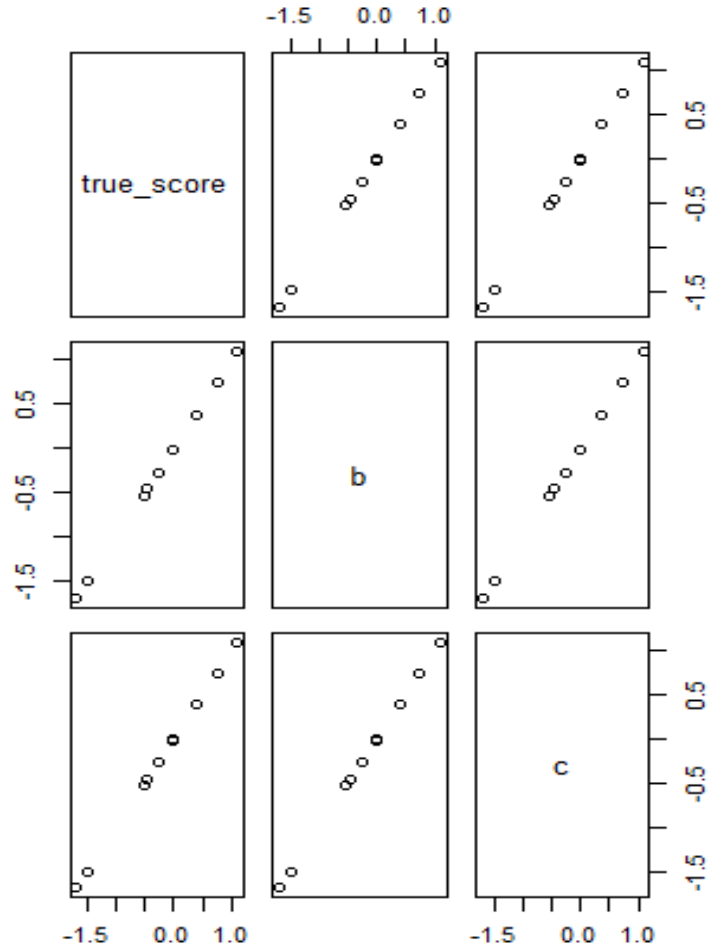
Habilidad B



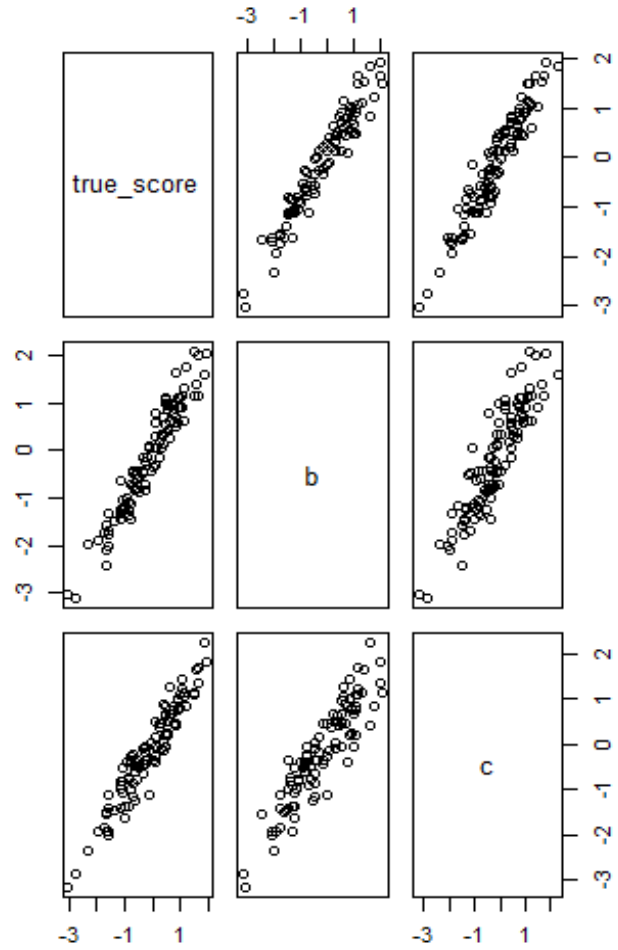


Si las respuestas rastrean/son causa de la misma “señal”, deberían estar correlacionadas “intra-sujetos” pero también “inter-sujetos”

Introducción de la idea de variable latente



Suponemos que b y c son consecuencias/reflejo/producto/resultado de T



b = T + e … e=0

b = T + e … e>0

Teoría clásica del test

Score verdadero: Lo que quisieras medir/representar bajo un modelo/Resultado

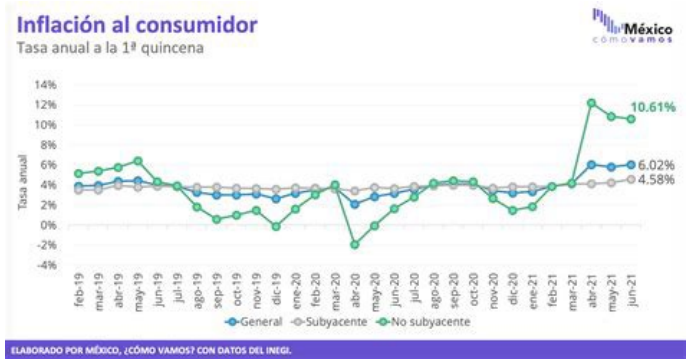
$$X_1 = T + e_1$$

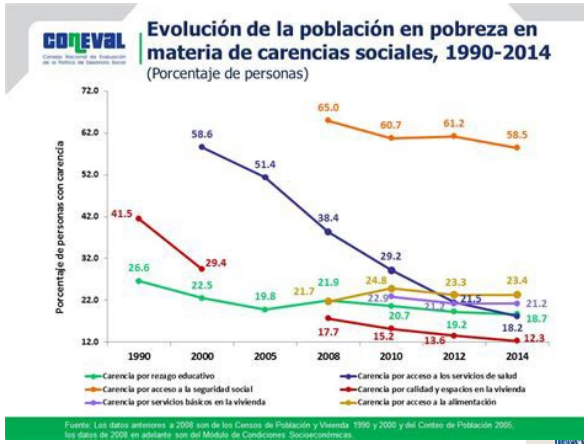
Score observado:

Indicaciones

Indicadores

Lo que no te interesa y distorsiona lo que concluyes a partir de X





Teoría clásica del test

Score verdadero: Lo que quisieras medir/representar bajo un modelo/Resultado

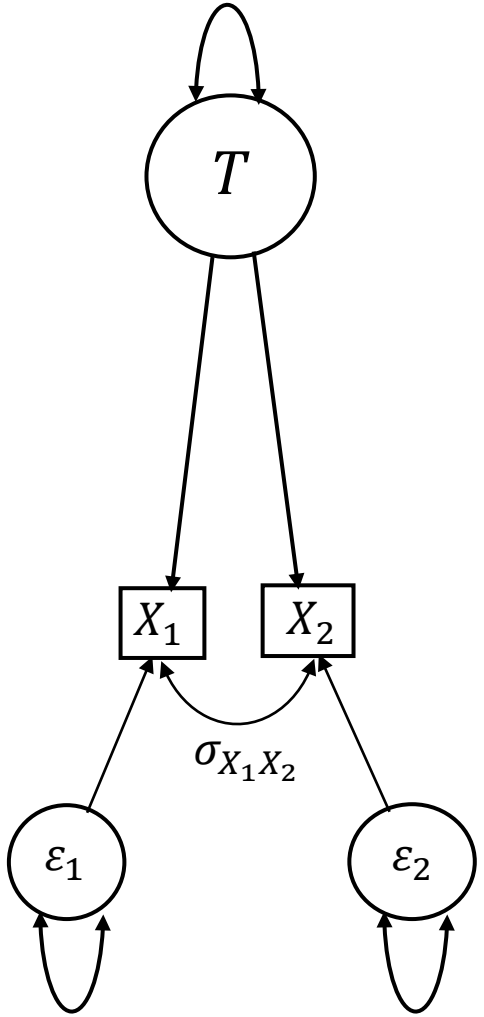
$$X_1 = T + e_1$$

Score observado:

Indicaciones

Indicadores

Lo que no te interesa y distorsiona lo que concluyes a partir de X



¿De qué forma añade esa variabilidad?

Segun la TCT se traduce en lo siguiente

Para la primera medida tenemos poco error:

$$e_1 \sim N(0,.1)$$

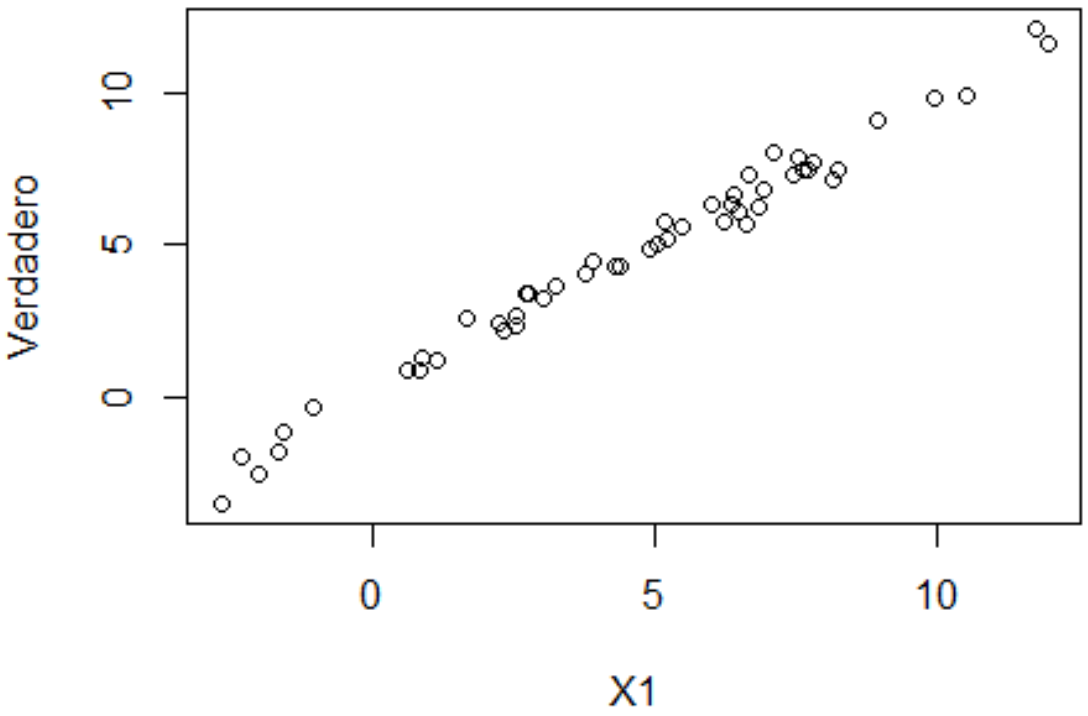
$$X_1 = T + e_1$$

Las respuestas de la estudiante "A" tienen poca variabilidad.

Es muy consistente

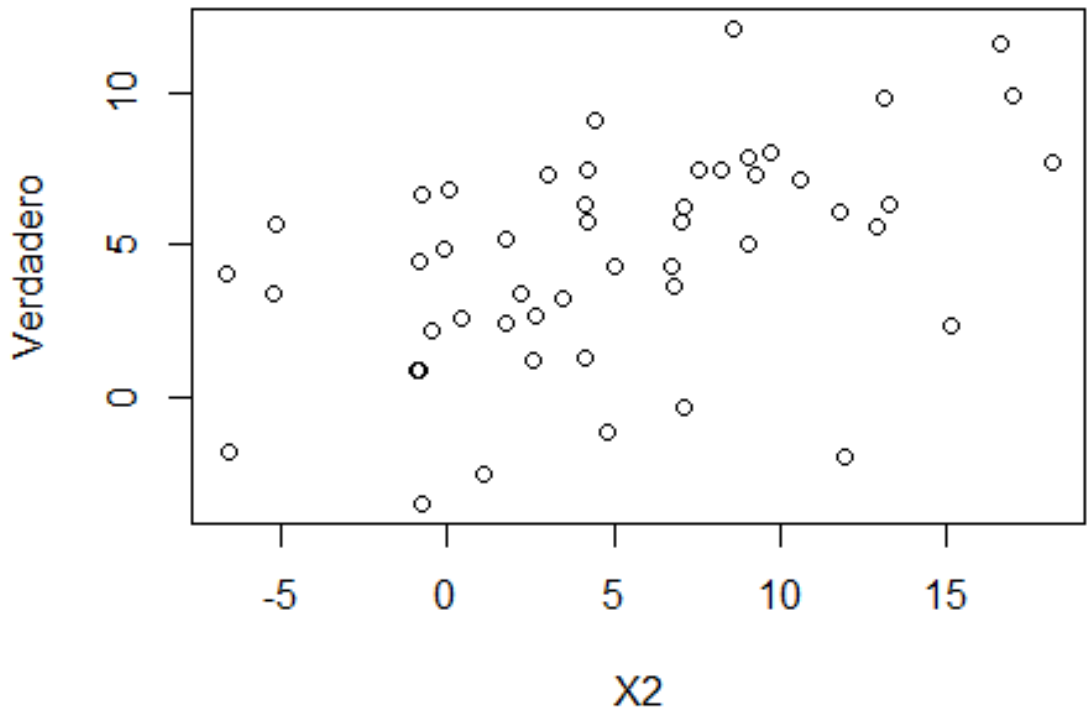
Si esto se extiende al resto del grupo, estamos haciendo una buena medición

Si graficamos: $X_1 = T + e_1$



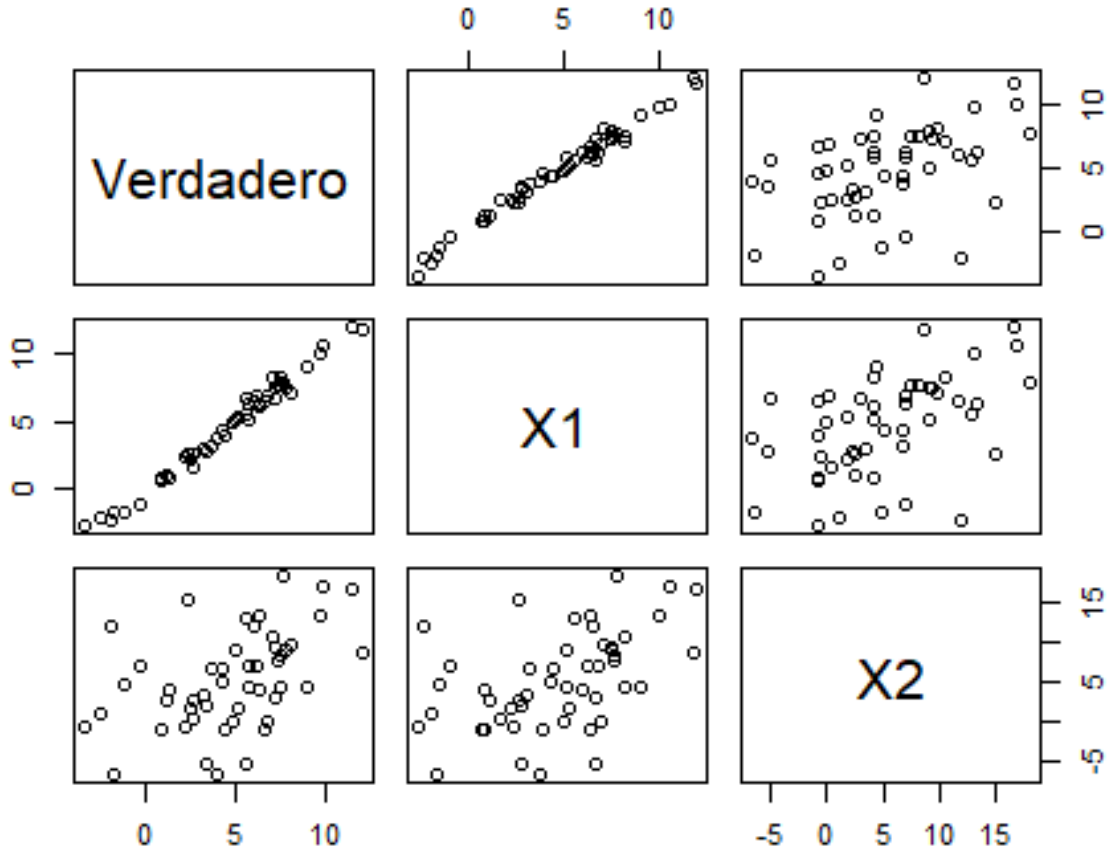
1. X1 y T ya no forman una línea recta
2. El error mueve los puntos en torno a la recta "latente" (o sea T)
3. Pero al ser e1 pequeño, los movimientos no son tan bruscos
4. La posición relativa de la persona/país/hogar con mayor score se mantiene

Error de medición



1. X2 y T no parecen moverse igual
2. El error mueve mucho los puntos en torno a la recta "latente" (o sea T)
3. Pero al ser e2 grande, los movimientos son bruscos
4. La posición relativa de la persona/país/hogar con mayor score NO se mantiene

¿Qué diría Spearman de lo siguiente?



Persona, hogar, estado, país	Score verdadero	X1	X2
1	-4.4	-4.2	-3.5
2	-10.8	-11.2	-12
3	1.1	0.8	-0.3
4	2.1	2.3	-2.1
5	-6.6…	-6.6	-3
6…			

Atenuación de coeficientes de correlación

Cuadro 1.5. Preguntas de acceso a la alimentación en los hogares, ENIGH 2008

Acceso a la alimentación en los hogares	
	Si en el hogar no hay personas menores de 18 años pase a la sección V. Equipamiento del hogar
1. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez usted o algún adulto en su hogar tuvo una alimentación basada en muy poca variedad de alimentos? Sí..... |1| No.....|2|	7. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez algún menor de 18 años en su hogar tuvo una alimentación basada en muy poca variedad de alimentos? Sí..... |1| No.....|2|
2. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez usted o algún adulto en su hogar dejó de desayunar, comer o cenar? Sí..... |1| No.....|2|	8. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez algún menor de 18 años en su hogar comió menos de lo que debía? Sí..... |1| No.....|2|
3. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez usted o algún adulto en su hogar comió menos de lo que usted piensa debía comer Sí..... |1| No.....|2|	9. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez tuvieron que disminuir la cantidad servida en la comida a algún menor de 18 años del hogar? Sí..... |1| No.....|2|
4. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez se quedaron sin comida? Sí..... |1| No.....|2|	10. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez algún menor de 18 años sintió hambre pero no comió? Sí..... |1| No.....|2|
5. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez usted o algún adulto de este hogar sintió hambre pero no comió? Sí..... |1| No.....|2|	11. En los últimos tres meses, por falta de dinero o recursos ¿algún menor de 18 años se acostó con hambre? Sí..... |1| No.....|2|
6. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez usted o algún adulto en su hogar sólo comió una vez al día o dejó de comer todo un día? Sí..... |1| No.....|2|	12. En los últimos tres meses, por falta de dinero o recursos ¿alguna vez algún menor de 18 años comió una vez al día o dejó de comer todo un día? Sí..... |1| No.....|2|


Charles Spearman

Como en un examen, lo que esperamos es que las respuestas sean un reflejo de inseguridad alimentaria.

Si las respuestas (1, 0) no reflejan el fenómeno, entonces los resultados serán aleatorios.

Si son aleatorios, la correlación se atenúa por el “ruido” en la medición

De nuevo

Mismo nivel latente

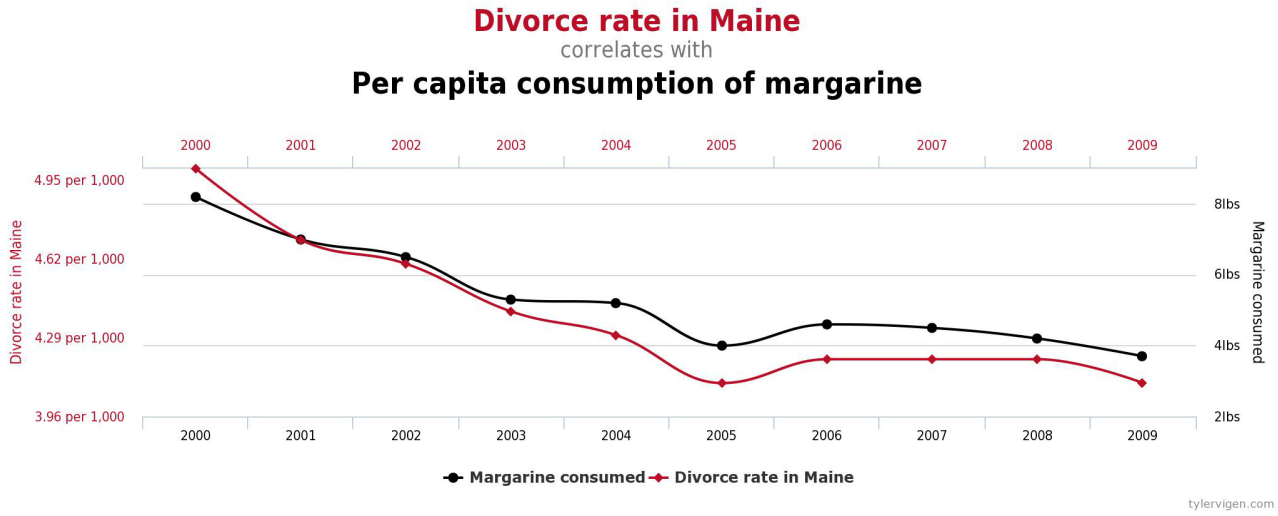
Persona A ≡ Persona B

Ítems

Score observado A ≈ Score observado B

Atenuación por error aleatorio

Componentes del modelo

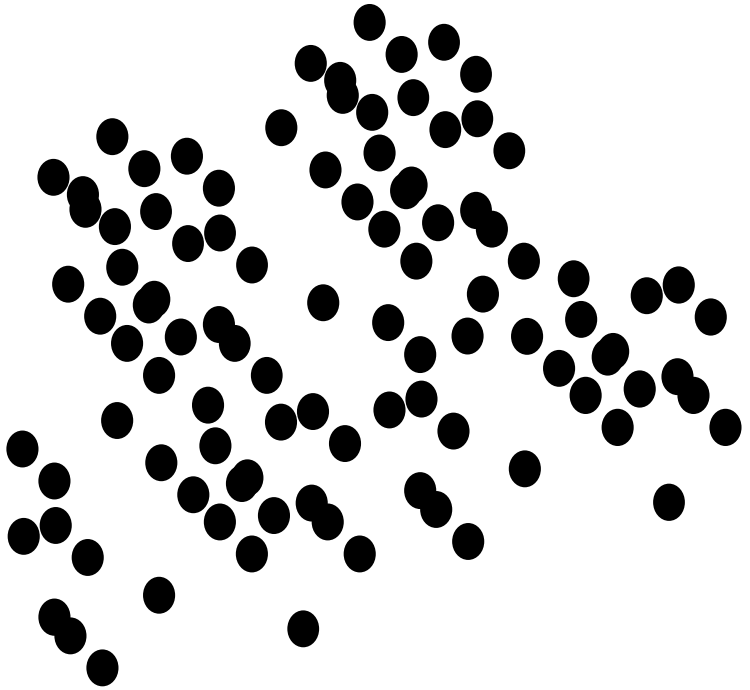


Sin el modelo teórico las correlaciones (parámetro) pierden sentido y utilidad

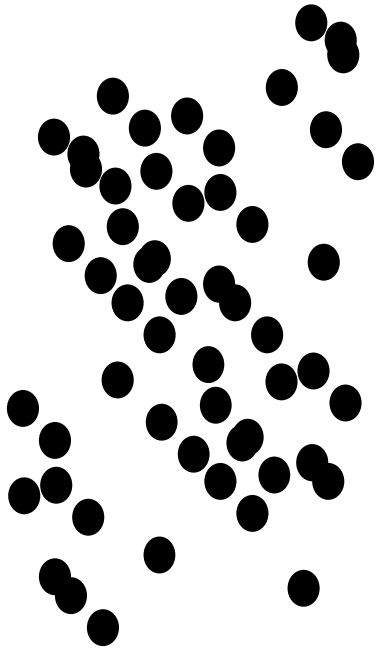
¿Cómo saber que dos correlaciones rastrean la misma señal?

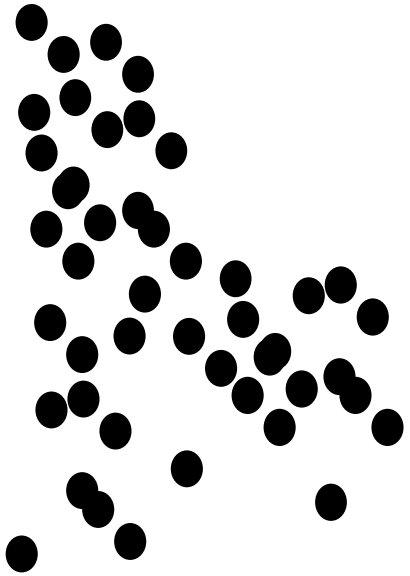
Sin modelo NO hay medición

Consecuencias de la falta de confiabilidad (i.e. alto error de medición)

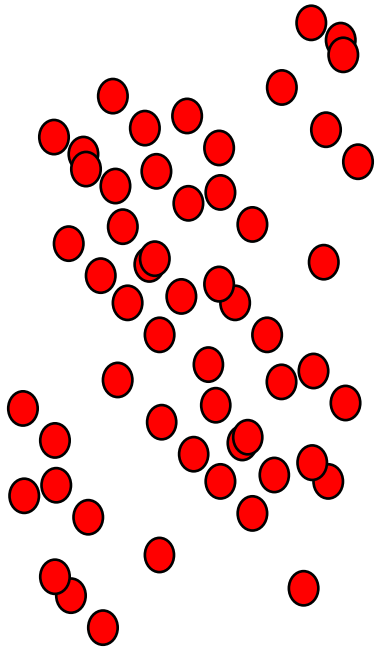


Clasificación

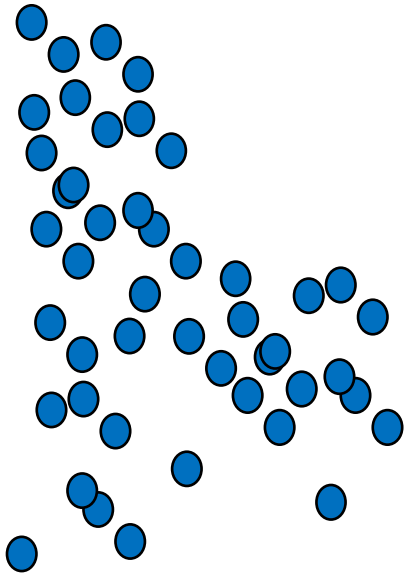




Clasificación ideal

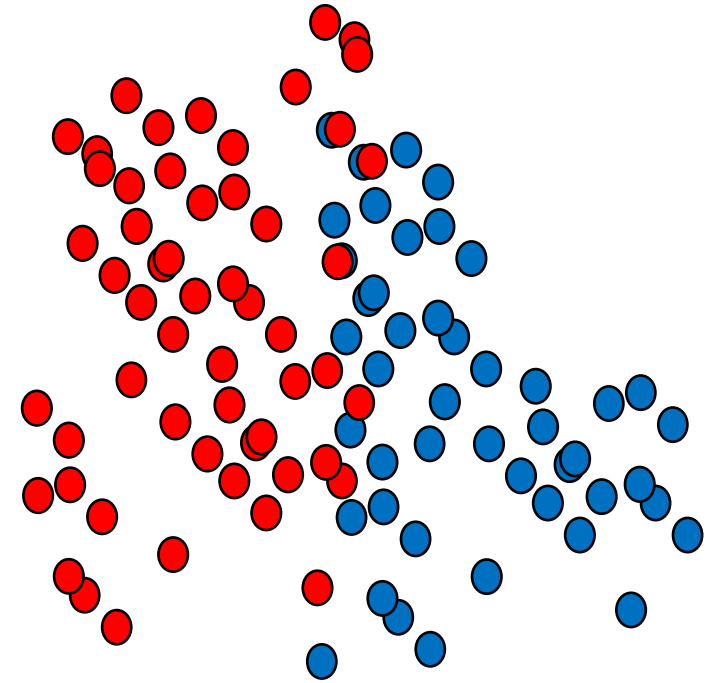


Grupo A



Grupo B

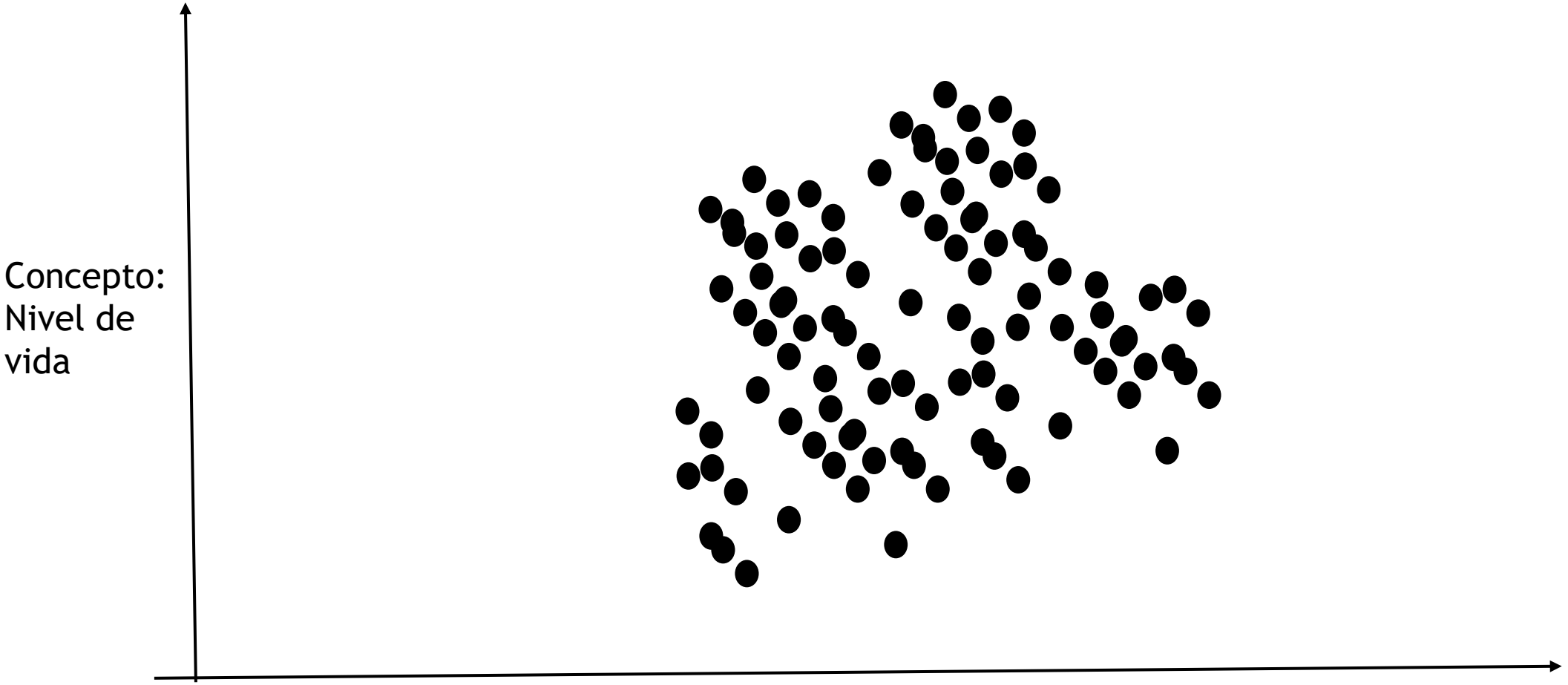
Clasificación factible



Grupo A

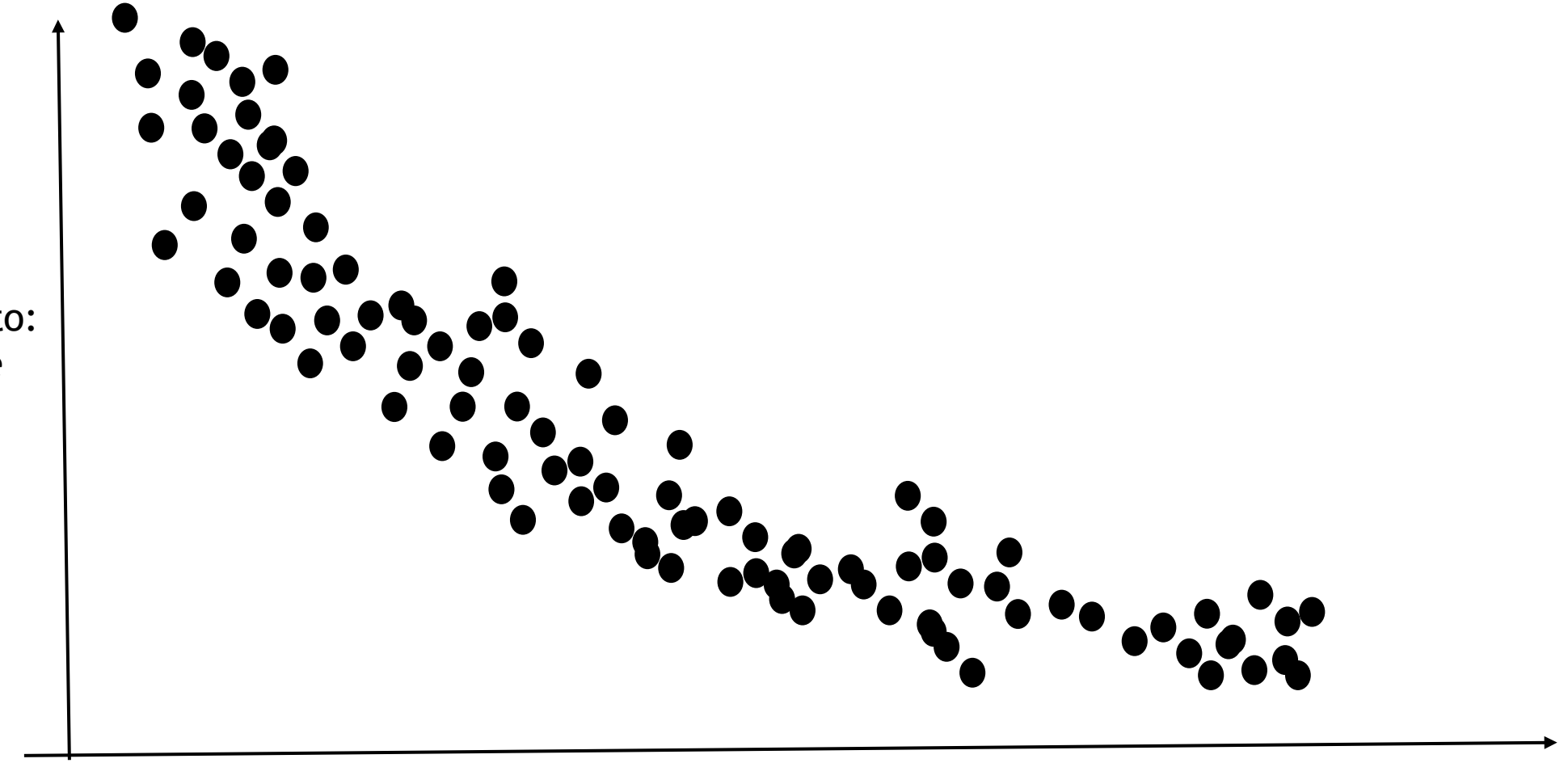
Grupo B

Concepto y observaciones



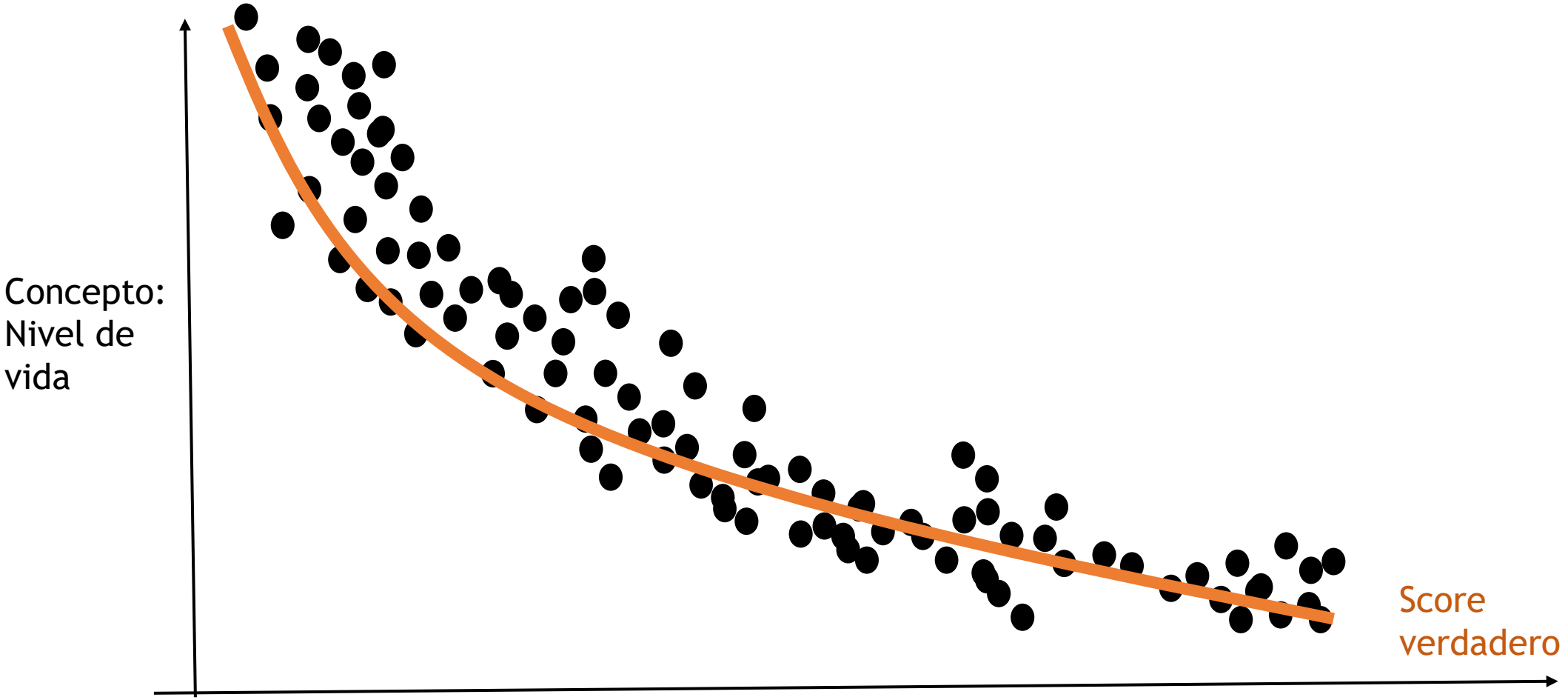
Manifestaciones (Observables): Bienes durables, ingreso, activos, vivienda, nivel educativo, etc.

Concepto y observaciones



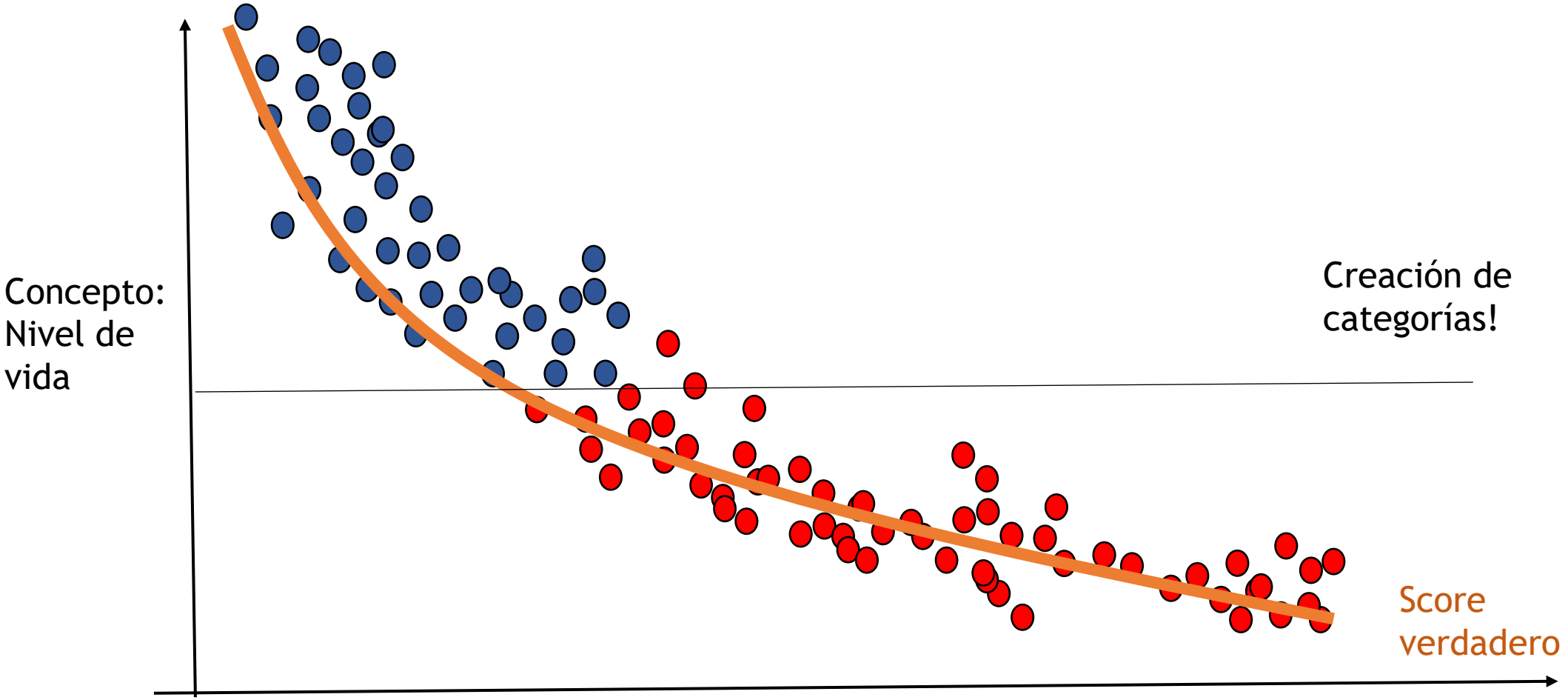
Manifestaciones (Observables): Bienes durables, ingreso, activos, vivienda, nivel educativo, etc.

Concepto y observaciones



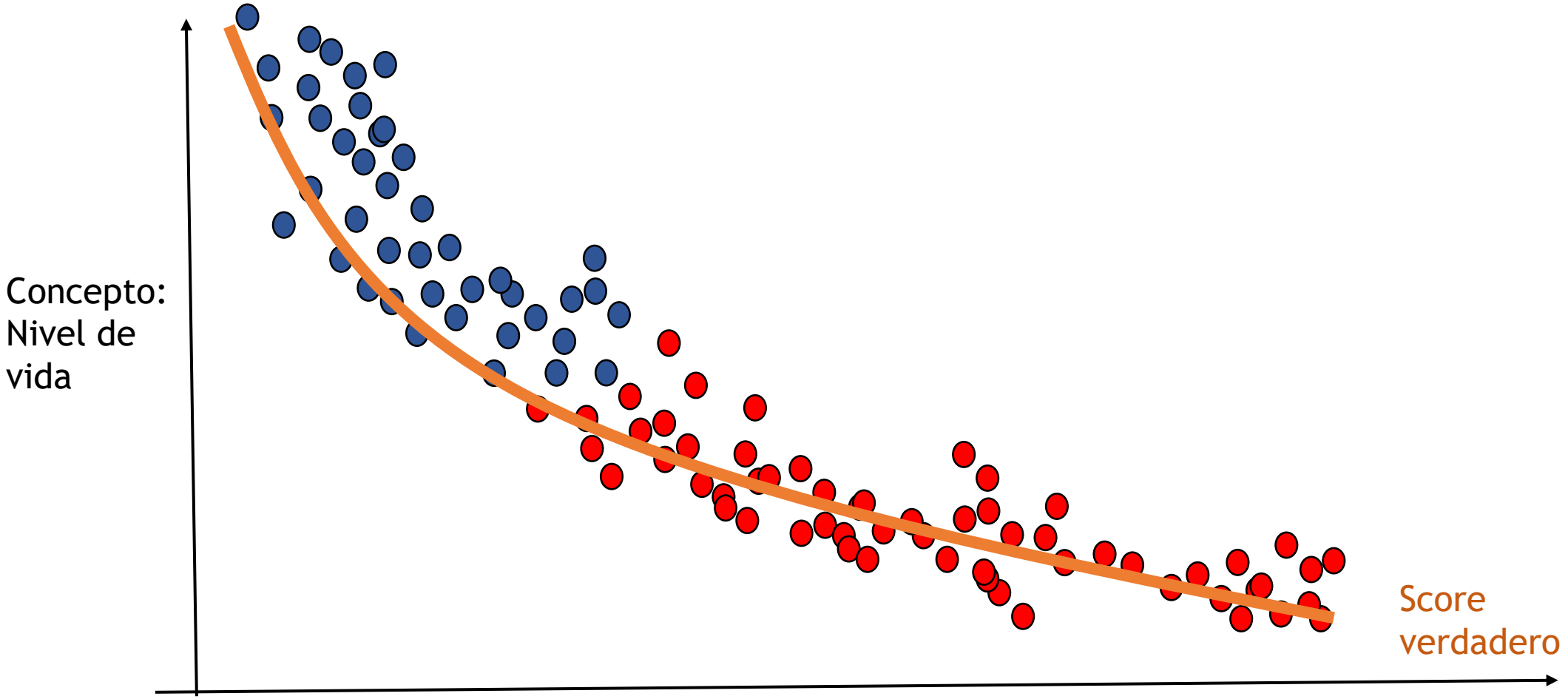
Manifestaciones (Observables): Bienes durables, ingreso, activos, vivienda, nivel educativo, etc.

Concepto y observaciones



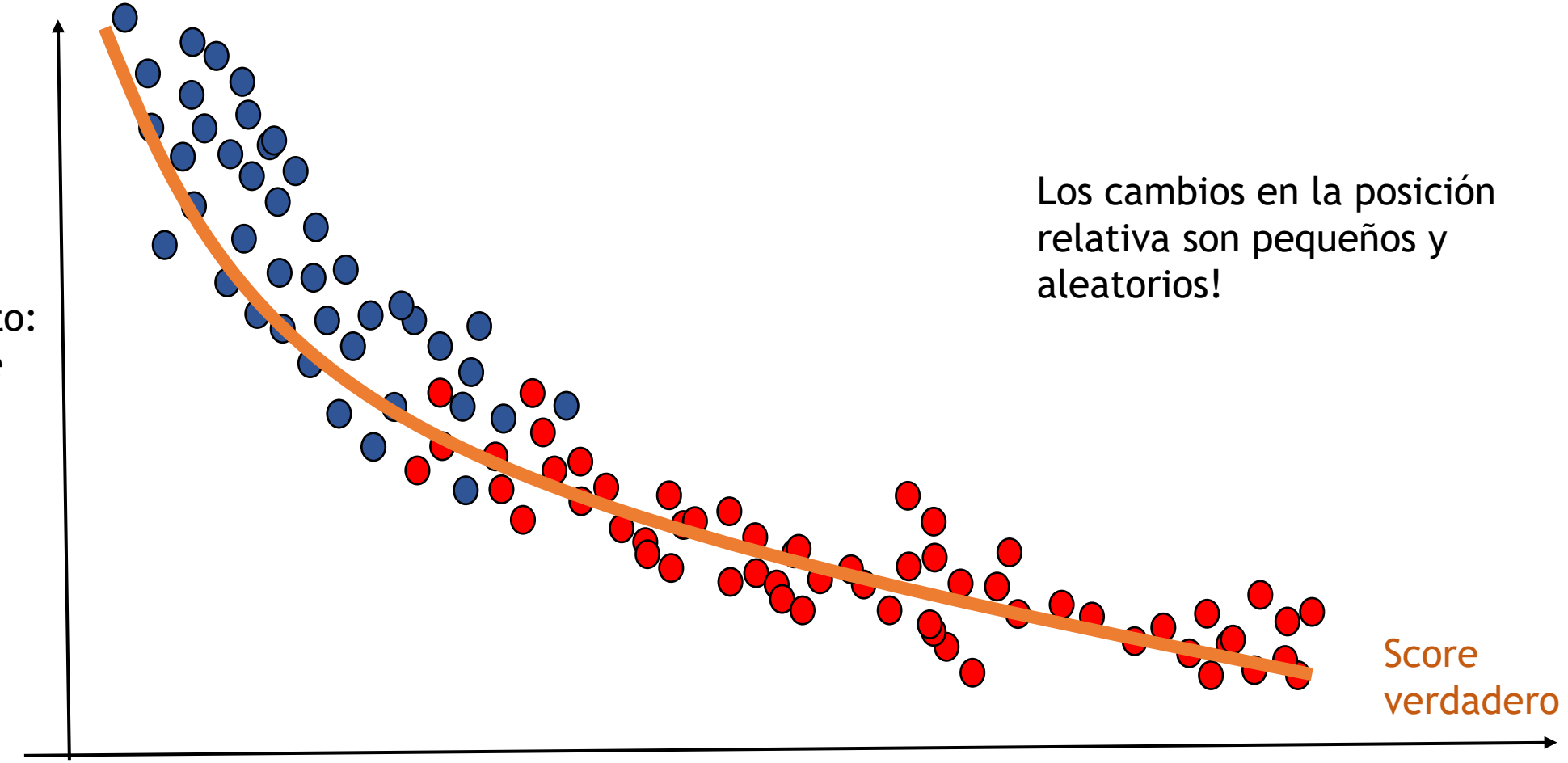
Manifestaciones (Observables): Bienes durables, ingreso, activos, vivienda, nivel educativo, etc.

Ideal en medición: Tiempo 1



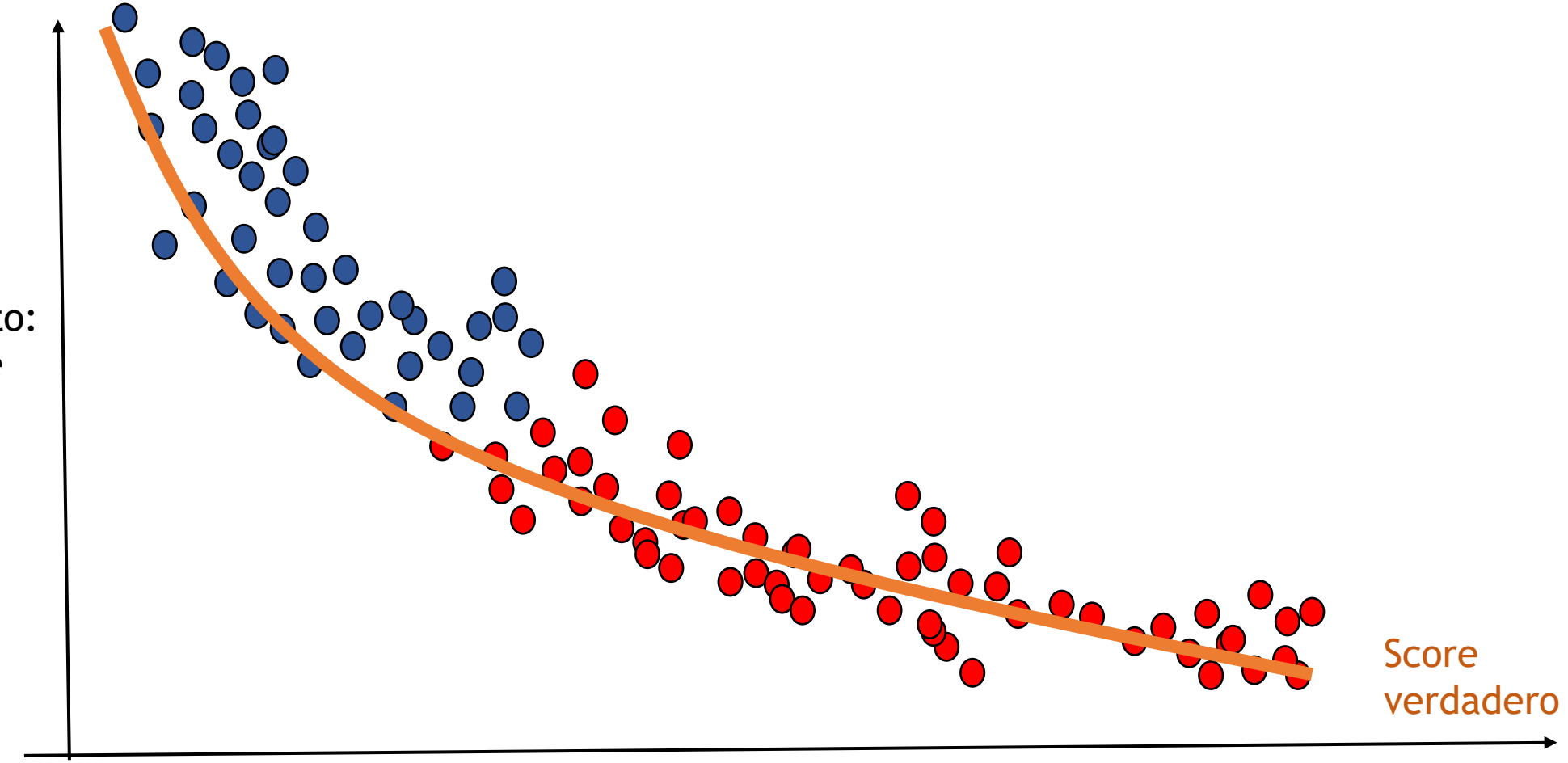
Manifestaciones (Observables): Bienes durables, ingreso, activos, vivienda, nivel educativo, etc.

Ideal en medición: Tiempo 2



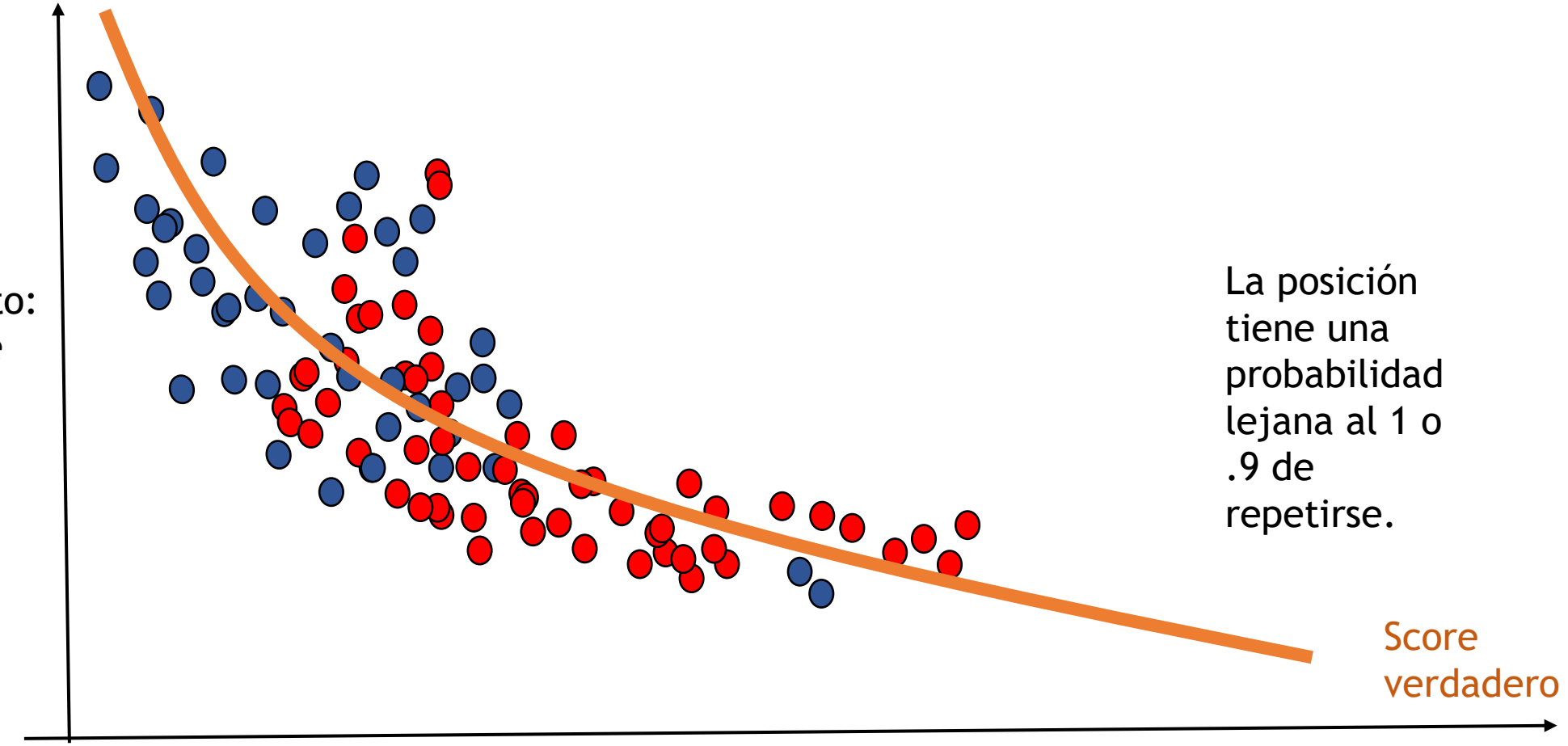
Manifestaciones (Observables): Bienes durables, ingreso, activos, vivienda, nivel educativo, etc.

Lejos del ideal en medición: Tiempo 1



Manifestaciones (Observables): Bienes durables, ingreso, activos, vivienda, nivel educativo, etc.

Lejos del ideal en medición: Tiempo 2



Manifestaciones (Observables): Bienes durables, ingreso, activos, vivienda, nivel educativo, etc.

Error de medición y confiabilidad

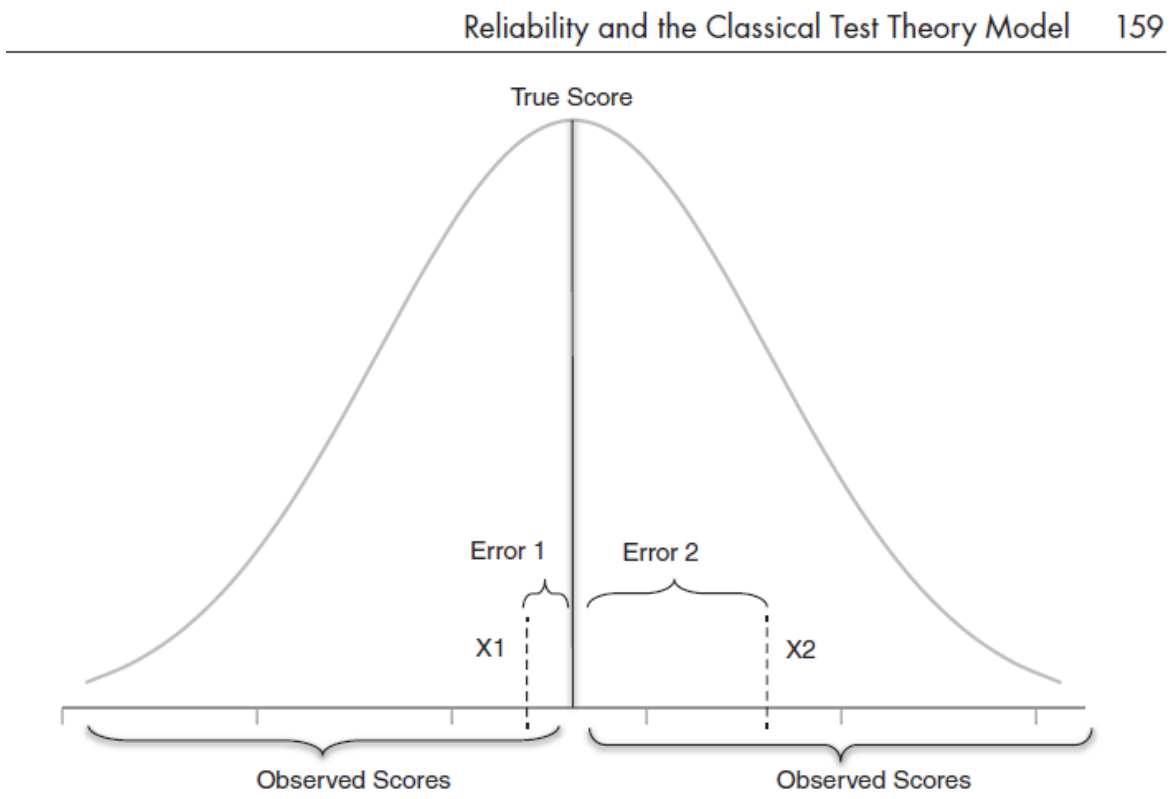


FIGURE 7.1. The distribution of observed scores around the true score.

Como decrece la confiabilidad de nuestras medidas y la suerte se vuelve más importante, las magnitudes de los coeficientes de correlación se acercan a cero (las correlaciones positivas se vuelven menos positivas y las correlaciones negativas se vuelven menos negativas)

$$R_{XY} = \frac{r_{XY}}{\sqrt{reliability_X reliability_Y}}$$

$$¿confiable \approx aleatorio^{-1}?$$

Confiabilidad (*reliability*)

- Implicaciones de un concepto relativo de confiabilidad para la lógica de la medición.

$$Confiabilidad = \frac{Variabilidad\ individual}{Variabilidad\ individual + Error\ de\ medición} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}$$

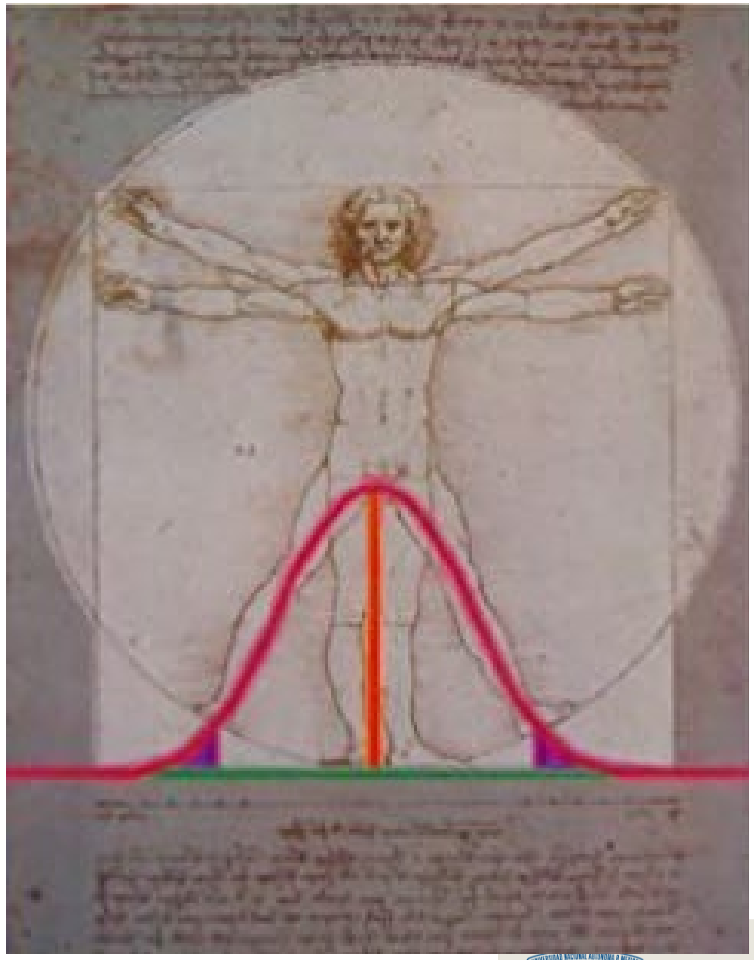
  - El *coeficiente de confiabilidad* refleja el grado en que un *instrumento* (artefacto) de medición es capaz de diferenciar entre individuos/objetos de estudio/unidades de observación (sujetos/hogares/familias/familias/escuelas/municipios/países/estados de la naturaleza).
  - La confiabilidad de una medida está íntimamente ligada a la población sobre la cual se quiere aplicar la medición.
    - No existe tal cosa como la confiabilidad de un instrumento/artefacto (a secas); el coeficiente sólo tiene significado cuando es aplicado a poblaciones específicas.
    - Es más difícil distinguir entre estados de la naturaleza (personas/hogares/municipios) si éstos son relativamente homogéneos que si éstos son muy diferentes.

Confiabilidad (*reliability*)

- Confiabilidad es un término relativo (¿confiable para qué?)
  - Nuestra comodidad con un determinado error de medición depende de que éste sea una fracción pequeña del rango en las observaciones
  - Para proveer información útil acerca de un error de medición, siempre debe contrastarse con la variación esperada entre las observaciones a llevar a cabo.
    - Una función del cociente entre la señal y el ruido.
    - La proporción entre lo relevante y lo irrelevante de nuestras observaciones empíricas (puntajes observados o medidas).
    - La fracción de nuestras mediciones que **no** es irrelevante.
    - La razón entre la varianza de nuestro interés y la varianza total de nuestras mediciones.
    - El porcentaje de la variación de nuestras mediciones que no es error.
    - La proporción de la varianza de nuestras mediciones que se debe a diferencias entre los individuos/objetos de estudio (“el-mundo-allá-afuera”).

Confiabilidad (*reliability*)

- No tiene sentido hablar de la confiabilidad (a secas) de un termómetro sin saber el rango de las temperaturas para las que va a ser utilizado.
- ES UN ERROR HABLAR DE LA CONFIABILIDAD DE UN TEST (ARTEFACTO).
  - L a confiabilidad NO es una propiedad inherente e inmutable de una escala.
  - la confiabilidad se refiere al RESULTADO obtenido con un instrumento (artefacto) y no al instrumento (artefacto) mismo.
  - La confiabilidad ES el resultado de la interacción entre el instrumento/artefacto y el sistema empírico al que éste es aplicado (objetos/individuos y su situación).

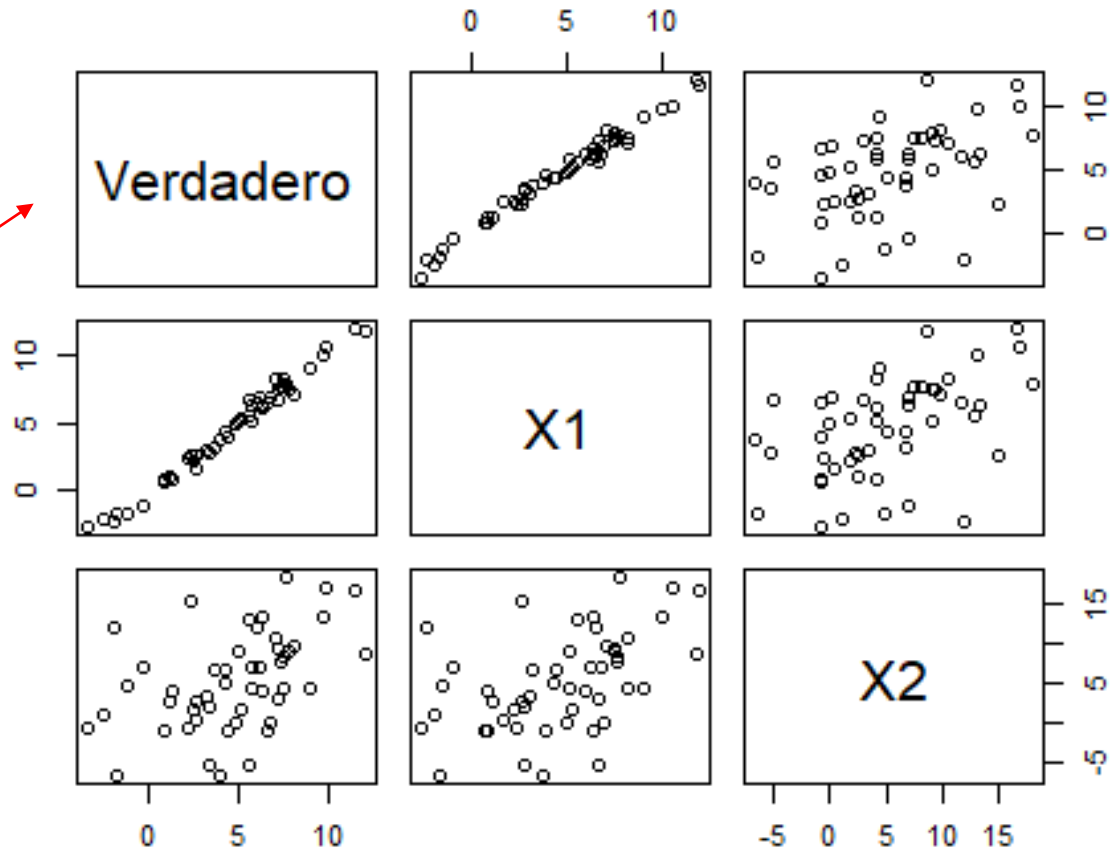


¿Cómo se estima la confiabilidad?

Teoría clásica del test

Scores verdaderos

No lo conocemos



Usar la información que tenemos para poder estimar el error

La gran dificultad

Esto está muy bien porque conocemos el valor del score verdadero.

Pero nunca lo conocemos

Sólo tenemos X1 y X2.

¿Cómo sabemos cuál tiene menos ruido respecto al valor verdadero?

Información y supuestos de la TCT

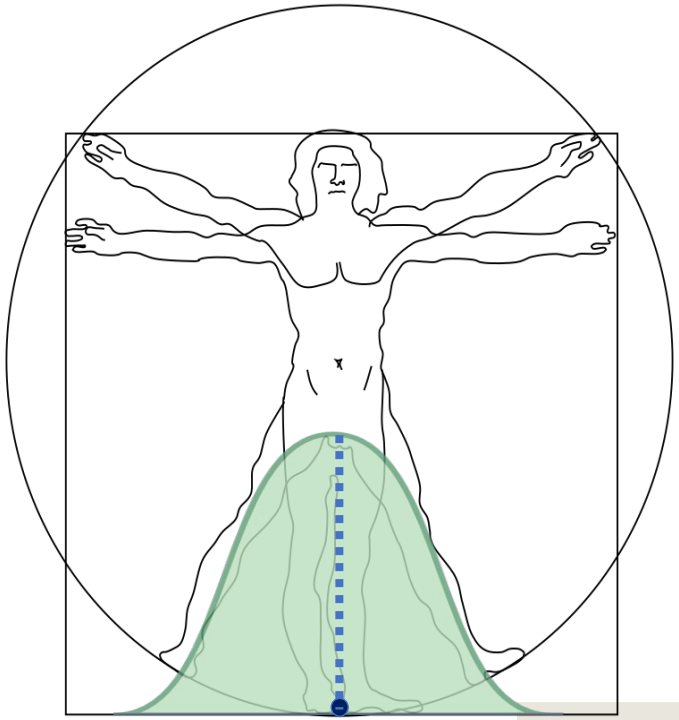
Matriz de correlación			
	Verdadero	X1	X2
Verdadero	1	0.9	0.4
X1	0.9	1	0.4
X2	0.4	0.4	1

El cálculo de confiabilidad en la TCT gira en torno a los supuestos que nos permiten utilizar los valores de X1 y X2 (*Test paralelos, equivalencia tau, … , half-Split reliability*)

¿Cómo estimar confiabilidad?

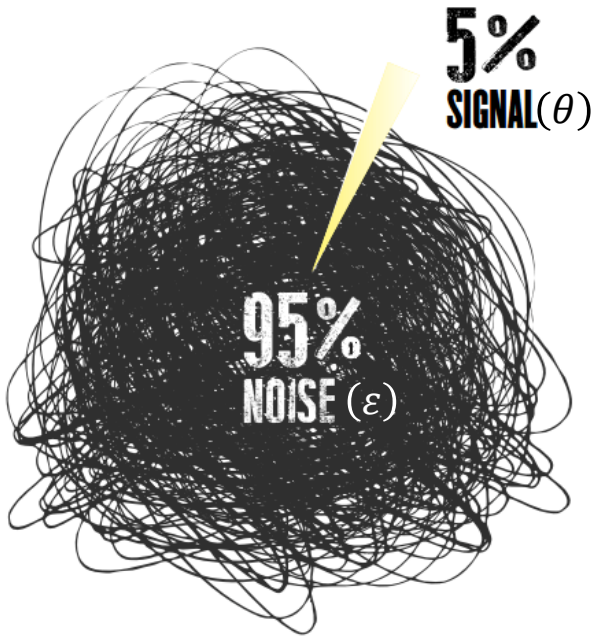
$$Confiabilidad = \frac{Variabilidad\ individual}{Variabilidad\ individual + Error\ de\ medición} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$$

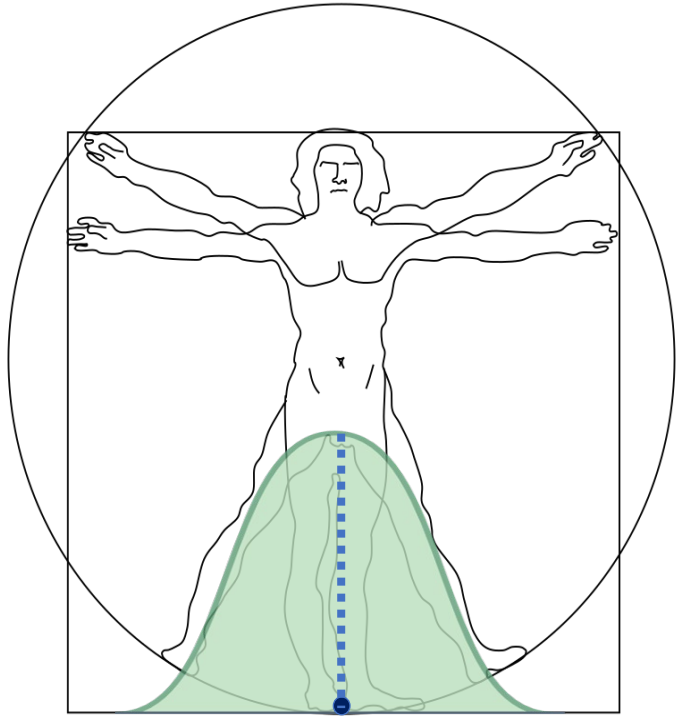
Desafortunadamente, todo lo discutido hasta ahora no sirve de nada si no podemos estimar σ_s^2 y σ_e^2



¿Cómo estimar confiabilidad?

- ¿Es possible deducir la composición de señal y ruido a partir de UNA observación (por objeto de estudio)?





Supuestos

- Test paralelos: Primera mitad del Siglo XX
- Equivalencia Tau: Mitad del Siglo XX
- Medidas congéneres: Finales del XX

- Ecuaciones estructurales y variables latentes: Presente

Pensemos estos supuestos

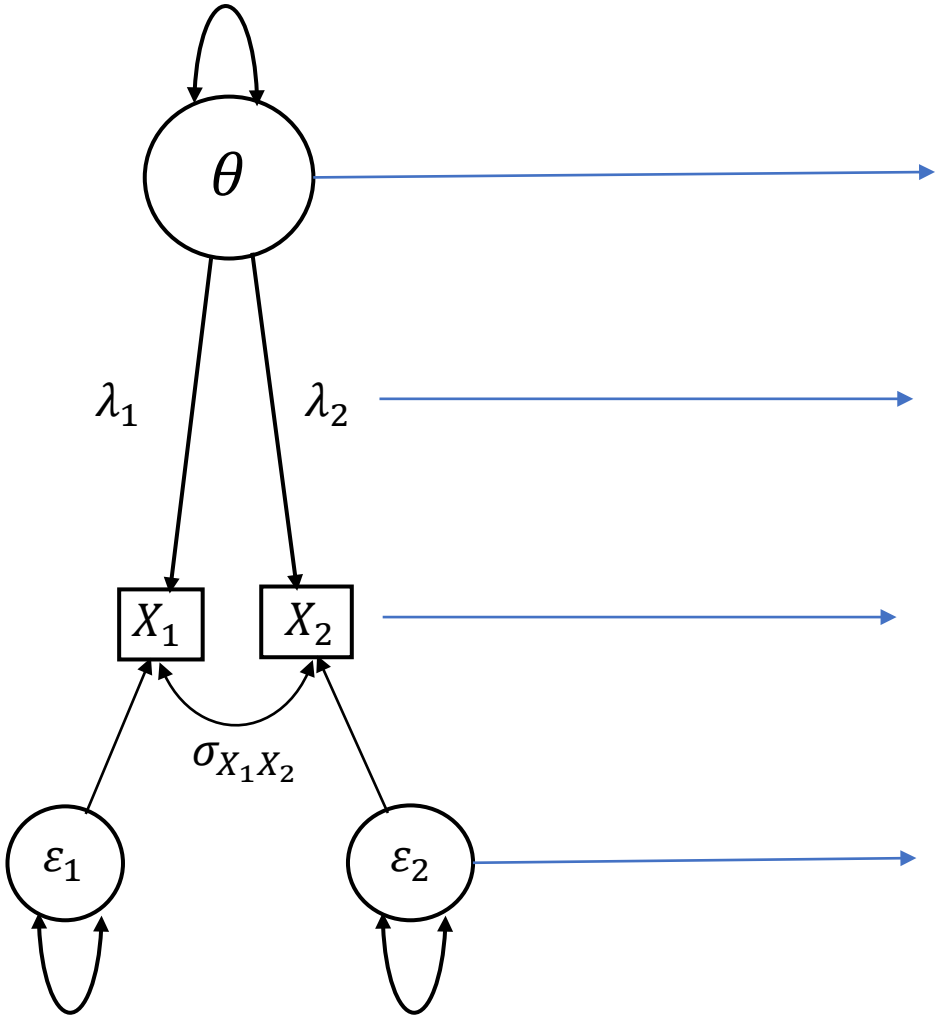
Cuadro C.1. Porcentaje de no especificados por entidad federativa según indicador socioeconómico, 2015

Clave de la entidad federativa	Entidad federativa	% Población de 15 años o más analfabeta	% Población de 15 años o más sin primaria completa	% Ocupantes en viviendas sin drenaje ni sanitario	% Ocupantes en viviendas sin energía eléctrica	% Ocupantes en viviendas sin agua entubada	% Viviendas con algún nivel de hacinamiento	% Ocupantes en viviendas con piso de tierra	% Población en localidades con menos de 5 000 habitantes	% Población ocupada con ingreso de hasta 2 salarios mínimos
	Nacional	**0.90**	**0.41**	**0.36**	**0.24**	**0.30**	**0.31**	**0.57**	—	**9.69**
01	Aguascalientes	0.38	0.12	0.04	0.03	0.05	0.05	0.09	—	7.28
02	Baja California	0.47	0.26	0.07	0.02	0.05	0.06	0.16	—	10.89
03	Baja California Sur	0.72	0.28	0.22	0.14	0.17	0.17	0.54	—	10.06
04	Campeche	0.47	0.14	0.04	0.03	0.02	0.07	0.18	—	6.83
05	Coahuila de Zaragoza	0.93	0.44	0.22	0.05	0.14	0.11	0.32	—	9.01
06	Colima	0.54	0.12	0.09	0.05	0.07	0.06	0.26	—	6.85
07	Chiapas	0.96	0.19	0.18	0.13	0.11	0.22	0.36	—	9.85
08	Chihuahua	2.41	2.08	2.02	1.90	1.93	1.97	2.07	—	9.74

Partimos de que son tests del mismos fenómeno…
¿Estas 8 variables se relacionarán de igual manera con la marginación?
¿Tendrán la misma varianza?

¿Si ninguna de esas condiciones se cumple, puedo estimar el error?

Lenguaje SEM



Fenómeno de interés (No lo vemos)

Relación del score observado con el fenómenos de interés / Variabilidad causada por θ (variabilidad de interés)

Indicador/variable/ítem/test observada

Término de error. Lo que no se explica por θ

Paralelas, tau-equivalent and congeneric

- Si los indicadores son reflejo del mismo fenómeno, estos pueden clasificarse de acuerdo a su grado de similaridad:

BOX 7.1

Properties of Parallel, Tau-Equivalent, Essentially Tau-Equivalent, and Congeneric Measures

Type of measure	μ_X	σ_X^2	σ_T^2	σ_E^2	$\sigma_{X_1X_2}$	$\rho_{X_1X_2}$	Relationship between true scores
Parallel	Must be equal	Must be equal	Must be equal	Must be equal	Must be equal	Must be equal	$t_i = 0 + 1 * t_j$
Tau-equivalent	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	$t_i = 0 + 1 * t_j$
Essentially tau-equivalent	May be equal or unequal	May be equal or unequal	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	$t_i = a_{ij} + 1 * t_j$
Congeneric	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	$t_i = a_{ij} + b_{ij} * t_j$

En los tres primeros casos la relación con la variable latente (t) es igual

Es factible estimar la confiabilidiad como vimos ayer (una de las x juega el papel de t). (Bandalos p. 167)

Noten que las condiciones van de más estrictas a menos estrictas

¿Cómo podemos saber que esto esta pasando?

Supuestos:

- Test paralelos: Primera mitad del Siglo XX
- Equivalencia Tau: Mitad del Siglo XX
- Medidas congéneres: Finales del XX

- Ecuaciones estructurales y variables latentes: Presente

Test paralelos

BOX 7.1

Properties of Parallel, Tau-Equivalent, Essentially Tau-Equivalent, and Congeneric Measures

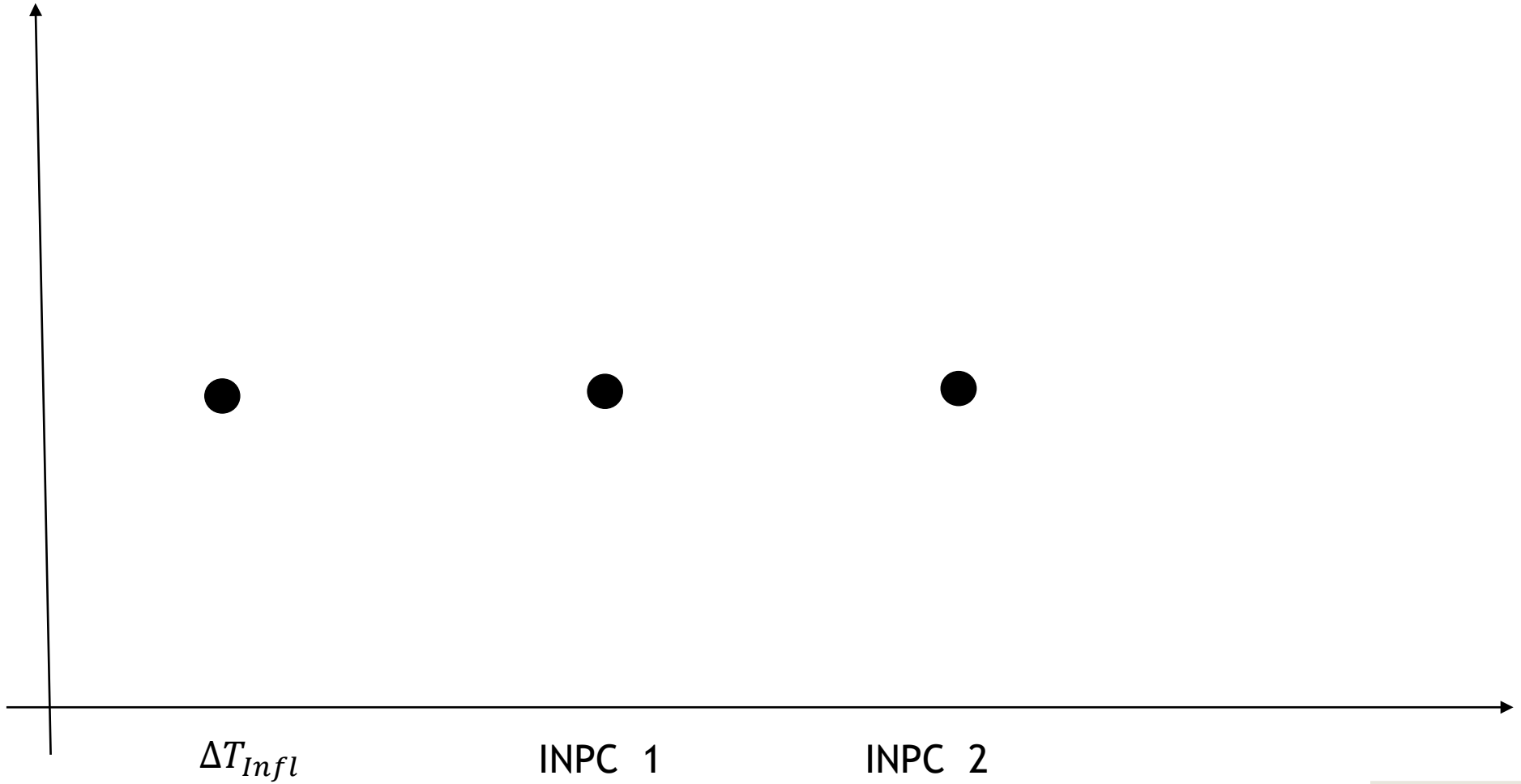
Type of measure	μ_X	σ_X^2	σ_T^2	σ_E^2	$\sigma_{X_1X_2}$	$\rho_{X_1X_2}$	Relationship between true scores
Parallel	Must be equal	Must be equal	Must be equal	Must be equal	Must be equal	Must be equal	$t_i = 0 + 1 * t_j$
Tau-equivalent	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	$t_i = 0 + 1 * t_j$
Essentially tau-equivalent	May be equal or unequal	May be equal or unequal	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	$t_i = a_{ij} + 1 * t_j$
Congeneric	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	$t_i = a_{ij} + b_{ij} * t_j$

$Y_1 = 0 + 1 * Y_2$

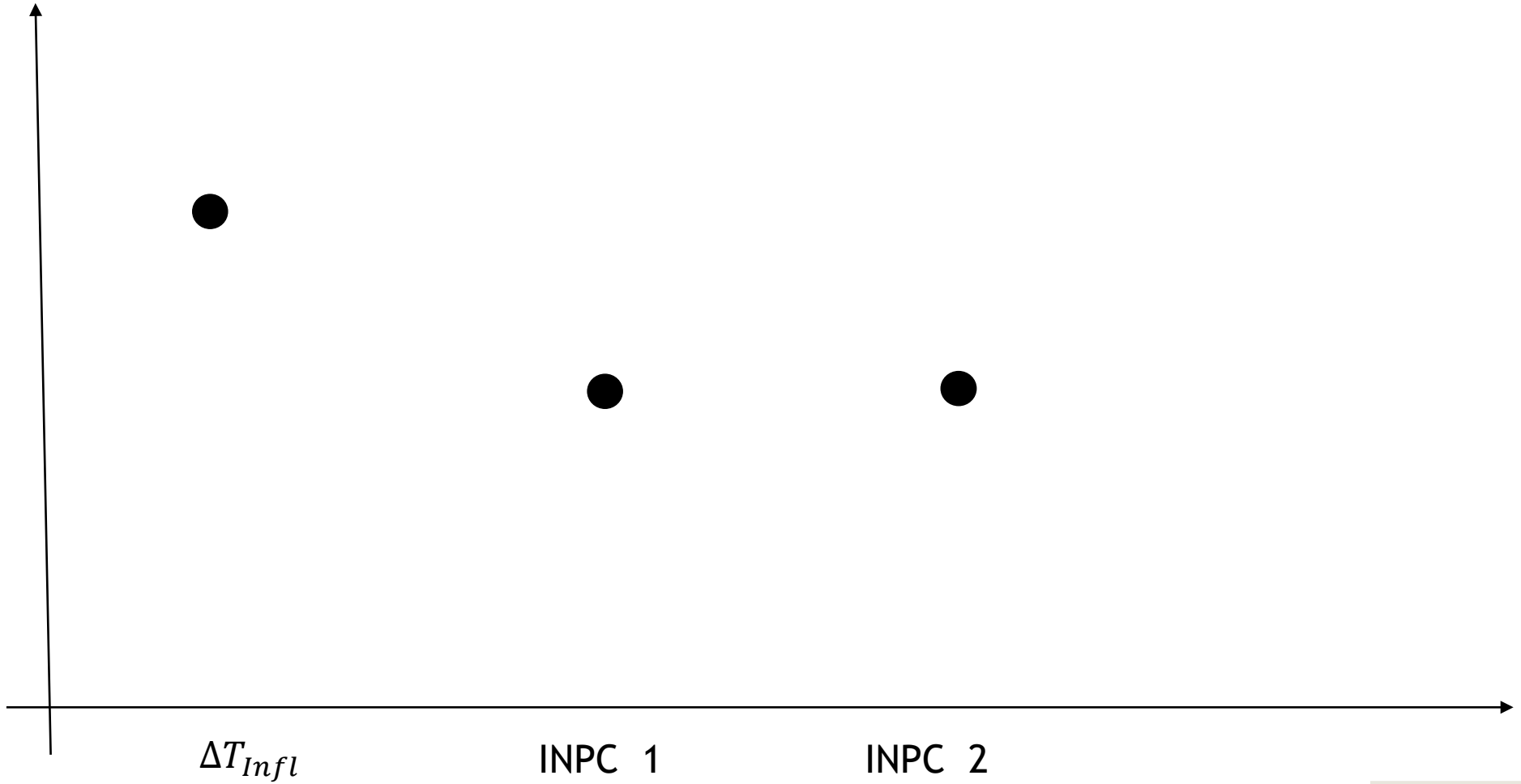
La medias son iguales

Las varianzas tambien

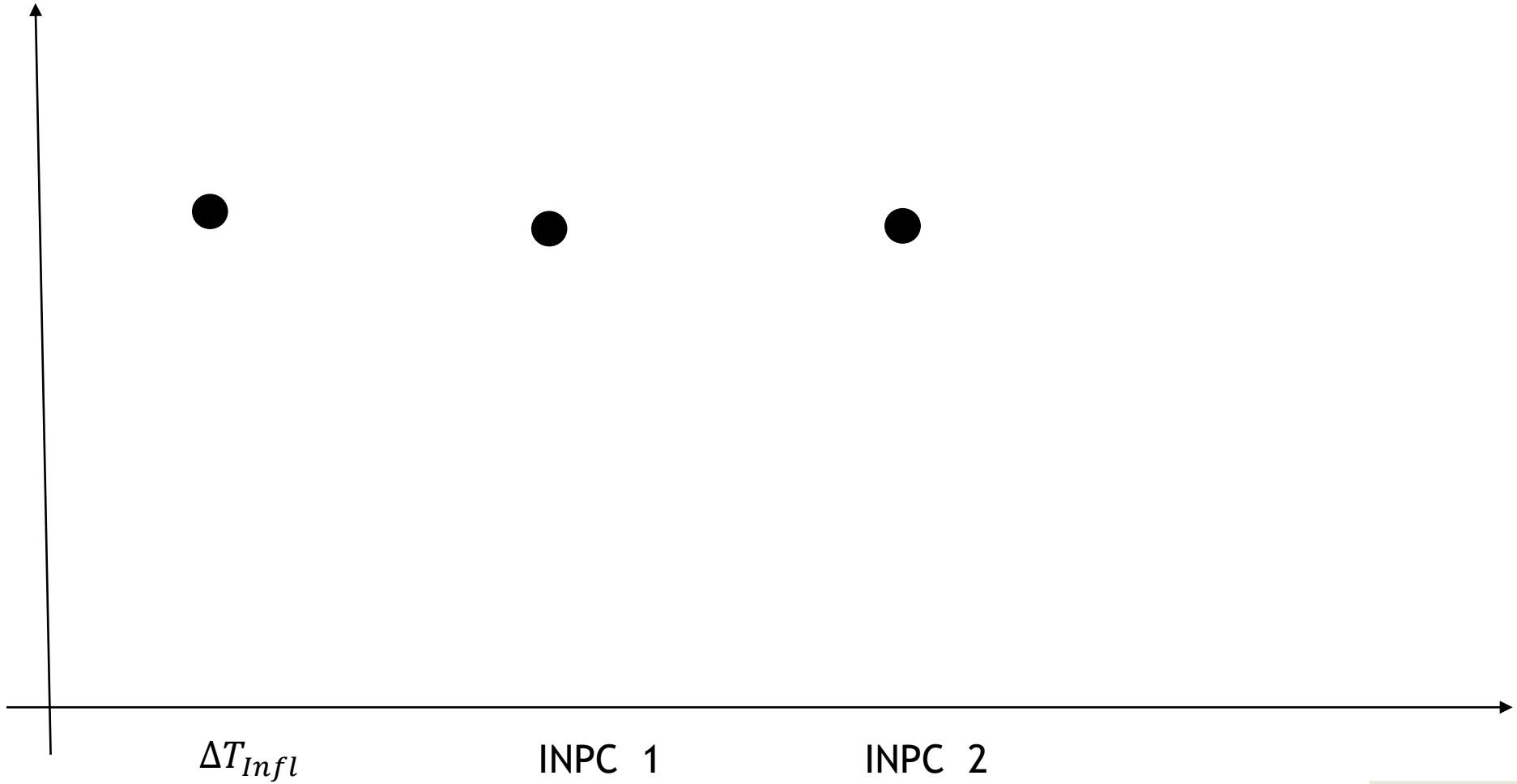
Test paralelos



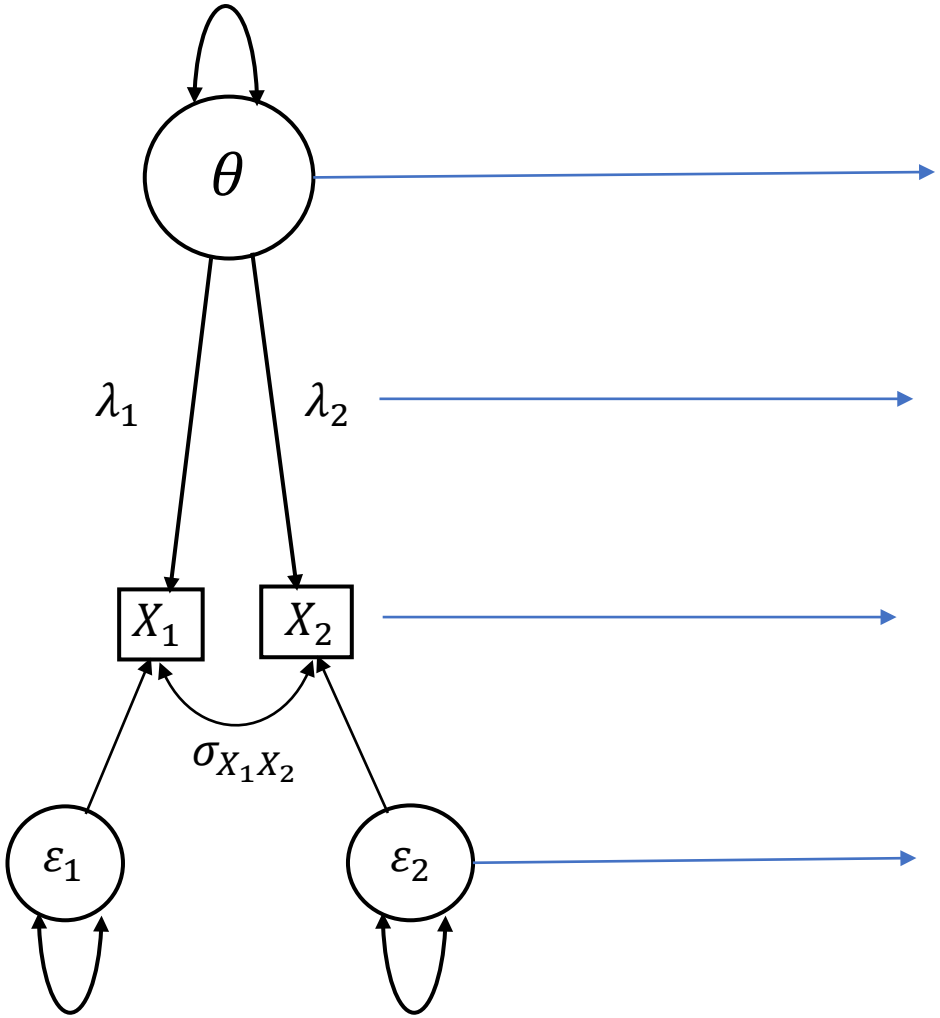
Test paralelos



Test paralelos



Test paralelos



Cambios fenómeno, con misma media respecto a X´s

Las lambdas son iguales. Misma reacción

Indicador/variable/ítem/test observada

Los errores son los mismos

Tests (medidas/indicadores) paralelos

No siempre puedo repetir y además en medición derivada busco ampliar el espectro de información

Incluyo más ítems y supongo que son paralelos

```
load("Data")
head(D)
```

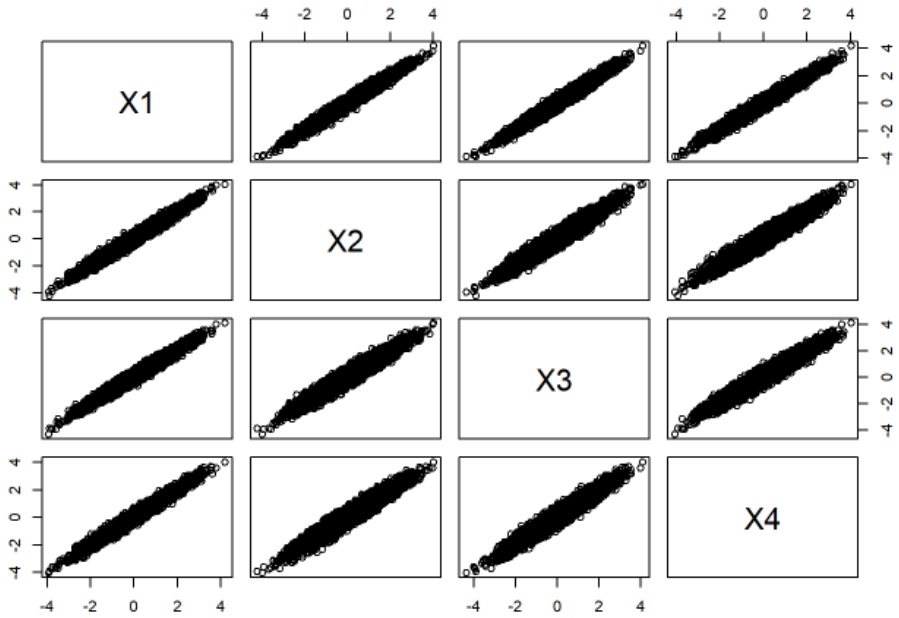
```
##           X1          X2          X3          X4
## 1  0.1079993  0.1263256  0.1835274  0.1068131
## 2 -2.1913068 -2.4435596 -2.0918066 -2.3461443
## 3 -0.7024488 -0.6761810 -0.8373665 -0.5846087
## 4 -0.4593898 -0.6938371 -0.3483979 -0.2026447
## 5  0.6445464  0.5939276  0.9114592  0.4786136
## 6 -0.9880279 -0.8805711 -0.9995512 -1.1631823
```

Tests (medidas/indicadores) paralelos

```
cor(D)
```

```
##           X1        X2        X3        X4
## X1 1.0000000 0.9836211 0.9837738 0.9834104
## X2 0.9836211 1.0000000 0.9676541 0.9673263
## X3 0.9837738 0.9676541 1.0000000 0.9670745
## X4 0.9834104 0.9673263 0.9670745 1.0000000
```

```
plot(D)
```



¿Cómo podemos saber que esto esta pasando?

Supuestos:

- Test paralelos: Primera mitad del Siglo XX
- Equivalencia Tau: Mitad del Siglo XX
- Medidas congéneres: Finales del XX

- Ecuaciones estructurales y variables latentes: Presente

Equivalencia Tau

BOX 7.1

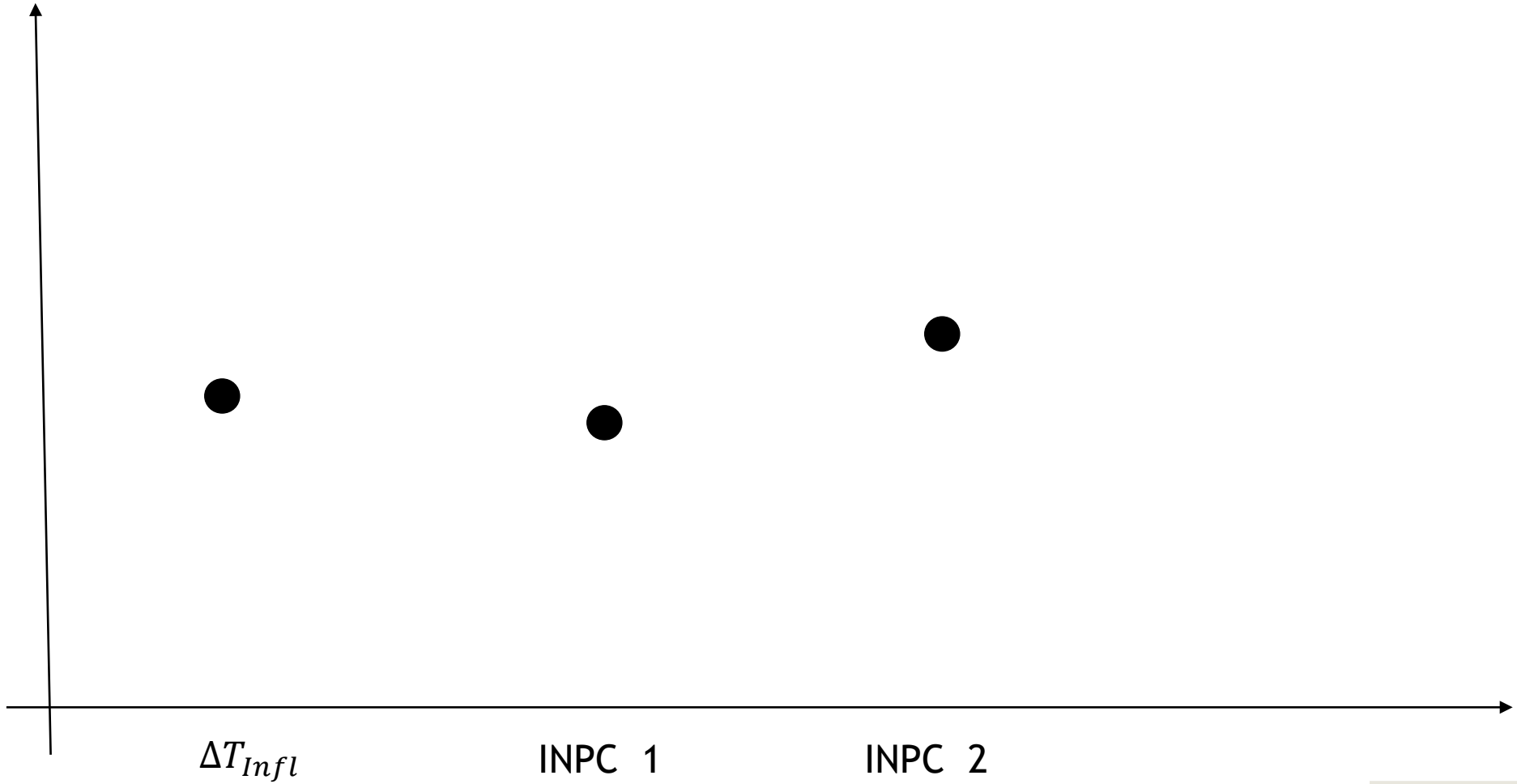
Properties of Parallel, Tau-Equivalent, Essentially Tau-Equivalent, and Congeneric Measures

Type of measure	μ_X	σ_X^2	σ_T^2	σ_E^2	$\sigma_{X_1X_2}$	$\rho_{X_1X_2}$	Relationship between true scores
Parallel	Must be equal	Must be equal	Must be equal	Must be equal	Must be equal	Must be equal	$t_i = 0 + 1 * t_j$
Tau-equivalent	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	$t_i = 0 + 1 * t_j$
Essentially tau-equivalent	May be equal or unequal	May be equal or unequal	Must be equal	May be equal or unequal	Must be equal	May be equal or unequal	$t_i = a_{ij} + 1 * t_j$
Congeneric	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	May be equal or unequal	$t_i = a_{ij} + b_{ij} * t_j$

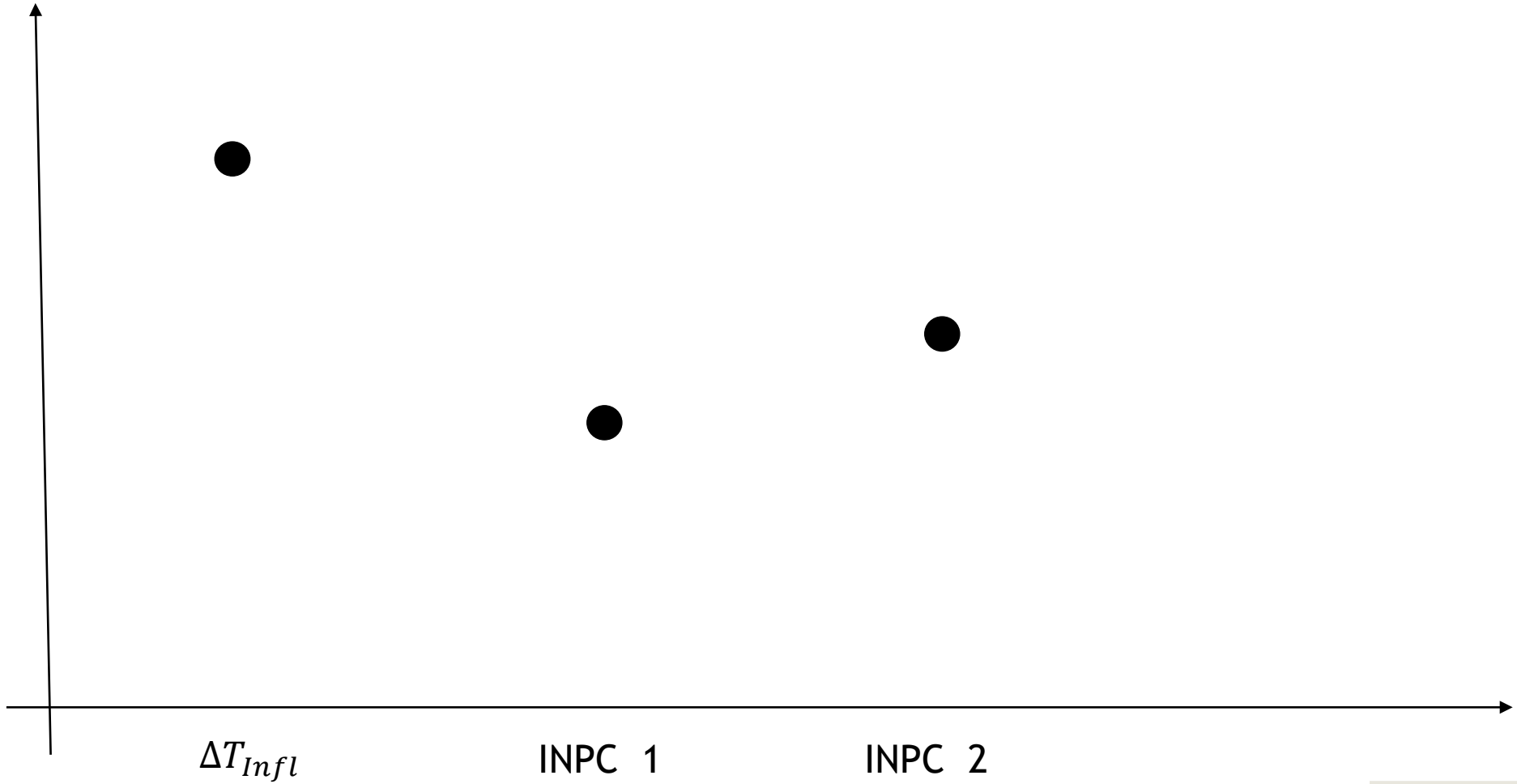
$Y_1 = a + 1 * Y_2$

La medias no tienen que ser iguales

Equivalencia Tau



Equivalencia Tau



Equivalencia Tau