

EL PLAN DE ANALISIS

Patricio Suárez Gil

Julio Alonso Lorenzo

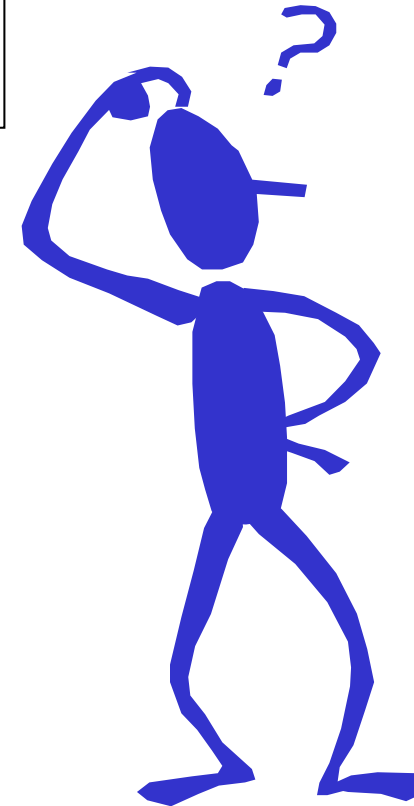
La Fresneda, Asturias 2011

Programa MFyC

¿Qué es investigar?

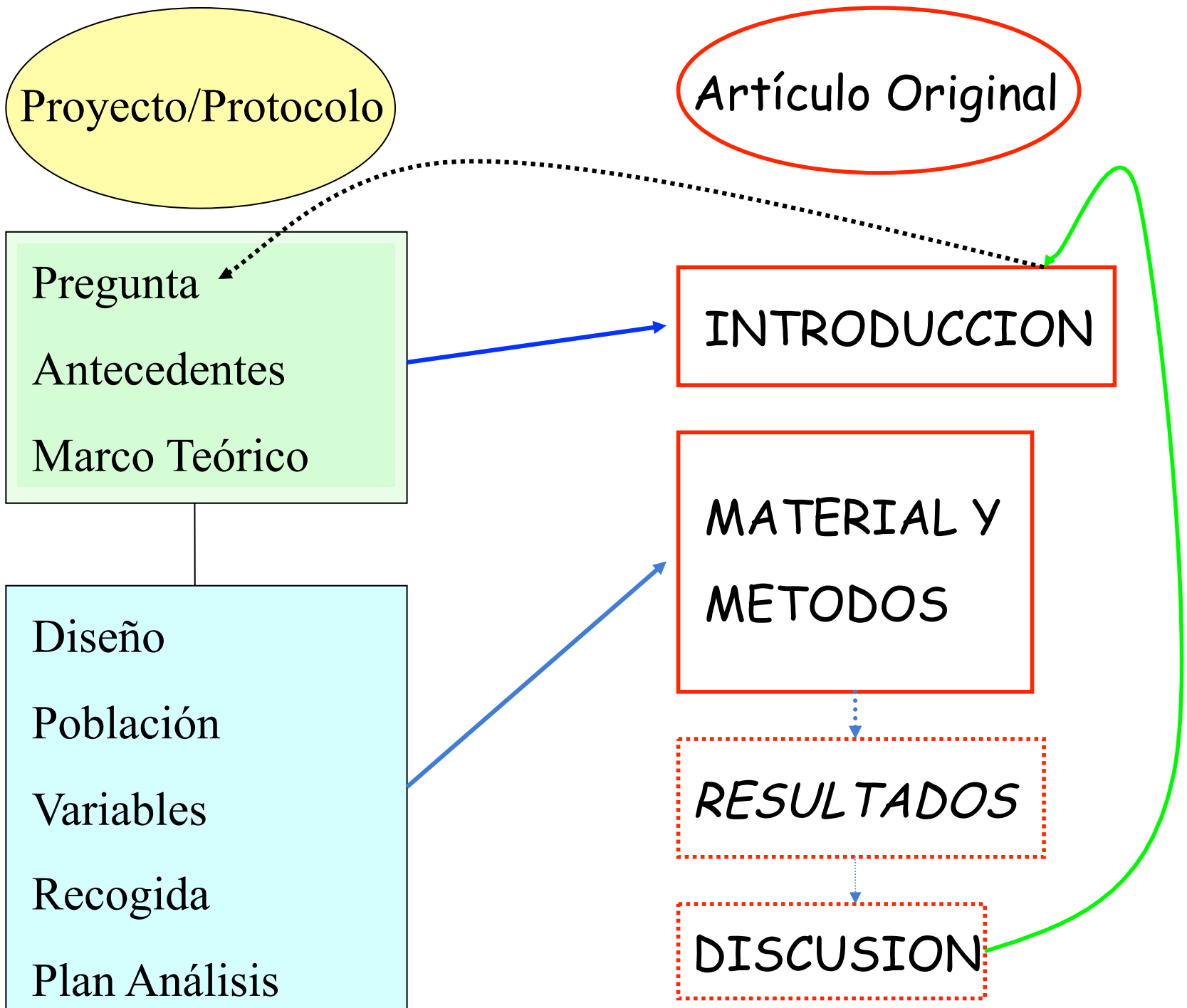
En esencia es

RESPONDER PREGUNTAS



FASES DE UN PROYECTO

- Formular una PREGUNTA
- Antecedente y Justificación
- Marco Teórico
- Diseño del estudio
- Población de estudio
- Variables o Mediciones
- Recogida de Información
- Plan de Análisis
- Plan de Trabajo, organización, recursos y cronograma



Plan de análisis

- Técnicas que *ayudan* a responder las preguntas formuladas.
- Ha de definirse *antes* de recoger los datos.
- En investigación cuantitativa (datos numéricos) las técnicas serán esencialmente *estadísticas*

Plan de Análisis: Condicionantes

- Preguntas
- Tipo de Variables
- Diseño del estudio
- Tipo de muestreo
- Software disponible y “dominable”
- Estudios previos

Plan de Análisis. Enfoque Básico

- Descriptivo (vs. Contemplativo)
- Inferencial:
 - A población muestreada
 - Otros marcos inferenciales
 - Problema de estimación/contraste de hipótesis.

Pasos

- Matriz de datos ($n \times k$)
- Depuración (técnicas)
- Transformación de variables
- Análisis univariante
- Análisis bivariante
- Análisis multivariante

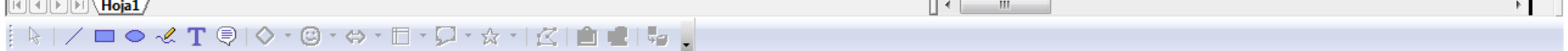
¿Qué es una matriz de datos?

- Ordenación de la información en una disposición de filas y columnas, donde
 - **Filas**: Individuos/unidades de observación (n)
 - **Columnas**: Datos o variables (k, kt)



E112 f(x) Σ = 2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Number	School	Form	Age	Sex	Relig	Know1	Know2	Know3	Know4	Know5	Know6	Know7	Know8	Knowtot	Faith1	Stigma1	Assert1	Condom1
2	1	5	3	19	1	1	1	1	1	1	1	1	0	1		3	0	4	1
3	2	5	2	15	1	1	1	1	1	1	1	1	0	0		4	0	0	0
4	3	5	2	16	1	1	1	1	1	1	1	1	0	0		4	4	4	3
5	4	5	2	17	2	1	1	0	1	1	1	0	0	1		4	0	4	4
6	5	5	2	17	2	1	1	1	1	1	1	1	0	1		4	1	2	2
7	6	5	2	16	1	1	0	0	1	0	1	0	0	0		4	0	3	4
8	7	5	2	16	1	1	1	0	1	1	1	1	1	0		99	0	4	2
9	8	5	2	17	1	1	1	1	1	1	1	0	0	0		4	0	99	0
10	9	5	2	16	1	1	1	0	1	1	1	0	0	1		4	0	4	3
11	10	5	2	16	1	1	1	1	1	1	1	0	0	1		4	0	3	2
12	11	5	1	16	2	1	1	1	1	1	1	0	0	1		4	0	4	4
13	12	5	1	14	1	1	1	1	1	1	0	1	1	0		4	0	0	4
14	13	5	1	17	2	99	1	1	1	1	1	1	0	0		4	3	99	99
15	14	5	1	18	2	1	0	1	1	0	0	1	1	1		4	3	3	4
16	15	5	1	15	1	1	1	1	1	1	1	1	0	1		4	0	1	4
17	16	5	1	14	1	1	1	1	1	0	0	1	0	1		2	0	4	0
18	17	4	1	17	1	1	1	0	1	1	1	1	1	1		4	0	0	0
19	18	4	1	15	1	1	1	1	1	1	1	1	0	0		4	0	4	0
20	19	4	1	15	1	1	1	0	1	1	1	0	1	1		4	4	1	0
21	20	4	1	17	2	1	1	0	1	1	1	0	1	1		2	4	0	4
22	21	3	2	16	1	1	1	1	1	1	1	1	1	1		4	0	4	1
23	22	3	2	16	1	1	1	1	1	1	1	1	0	1		4	0	4	4
24	23	6	3	18	2	1	1	1	1	1	1	0	1	1		4	0	0	4
25	24	3	2	17	2	1	1	1	1	1	1	1	1	1		4	0	99	1
26	25	3	2	16	1	1	1	0	1	1	1	0	0	1		4	0	0	4
27	26	3	2	17	1	1	1	1	1	1	1	1	1	1		4	0	0	4
28	27	3	2	17	2	1	1	1	1	1	1	1	1	1		4	0	4	3
29	28	6	2	16	1	1	1	1	1	1	1	0	0	1		3	0	4	2
30	29	6	2	17	2	1	1	1	1	1	1	1	0	1		4	4	4	0
31	30	6	1	15	1	1	1	1	1	0	1	0	0	1		4	0	2	4
32	31	3	2	15	2	1	1	1	1	0	1	1	1	1		4	0	4	4
33	32	6	3	17	2	1	1	1	1	1	1	0	0	1		4	4	4	4



COD	AREA	PERSONA	MUNICIP	MUNICIP2	HOSPITAL	COD_HOSP	W	DIAS	MES	PESO	LACTA	ALIMEN	LACTAEXC	LACTPREV					
1	1	1	VALDES	JARRIO	1	1,00	76,00	3	3660	1	0	#;NULO!	#;NULO!	1	8	0	2		
2	3	1	CORVERA DE ASTURIAS	SAN AGUSTIN	3	2,99	84,00	3	2570	1	1	0	#;NULO!	#;NULO!	1	10	0	1	
3	1	1	VEGADEO	JARRIO	1	1,00	73,00	3	3240	1	0	1	#;NULO!	#;NULO!	1	10	0	2	
4	6	1	LLANES	ORIENTE	6	1,00	147,00	5	3450	0	1	0	1	18	1	5	0	2	
5	1	1	FRANCO (EL)	JARRIO	1	1,00	87,00	3	3330	1	0	1	#;NULO!	#;NULO!	1	1	6	0	
6	2	1	TINEO	CANGAS NARCEA	2	1,00	62,00	3	3210	1	0	1	#;NULO!	#;NULO!	1	1	4	0	
7	6	1	VILLAMAYOR (PILOÑA)	ORIENTE	6	1,00	101,00	4	3340	0	1	0	1	0	1	4	1	3	
8	1	1	NAVIA	JARRIO	1	1,00	76,00	3	2840	1	0	1	#;NULO!	#;NULO!	1	7	0	2	
9	1	1	VALDES	JARRIO	1	1,00	81,00	3	3480	1	0	1	#;NULO!	#;NULO!	1	2	0	2	
10	6	1	ARRIONDAS	ORIENTE	6	1,00	190,00	6	3350	0	1	0	1	16	1	4	0		
11	1	1	NAVIA	JARRIO	1	1,00	34,00	2	4190	1	0	1	#;NULO!	#;NULO!	1	13	1	2	
12	1	1	NAVIA PUERTO DE VEGA	JARRIO	1	1,00	125,00	5	3700	1	1	0	#;NULO!	#;NULO!	1	1	4	0	
13	3	1	CASTRILLON	SAN AGUSTIN	3	2,99	68,00	3	2750	1	1	0	#;NULO!	#;NULO!	1	1	5	0	
14	1	1	VALDES	JARRIO	1	1,00	31,00	2	3840	1	0	1	#;NULO!	#;NULO!	1	8	0	2	
15	1	1	GRANDAS DE SALIME	JARRIO	1	1,00	46,00	2	3860	0	1	0	1	0	1	1	4	0	
16	2	1	CANGAS DEL NARCEA	CANGAS NARCEA	2	1,00	118,00	4	2410	0	1	0	1	0	1	9	1	1	
17	1	1	NAVIA NAVIA	JARRIO	1	1,00	104,00	4	3050	0	1	0	1	4	1	8	0	2	
18	1	1	TAPIA DE CASARIEGO	JARRIO	1	1,00	41,00	2	3080	0	1	0	1	0	1	1	1	5	
19	8	1	LANGREO	VALLE NALON	8	1,00	10,00	1	3430	0	1	0	1	1	1	1	2	0	
20	3	1	CORVERA DE ASTURIAS	SAN AGUSTIN	3	2,99	163,00	6	3340	0	1	0	1	0	0	#;NULO!	0	0	
21	3	1	AVILES	SAN AGUSTIN	3	2,99	168,00	6	3040	0	1	0	0	0	#;NULO!	0	4	0	
22	1	1	FRANCO (EL)	JARRIO	1	1,00	82,00	3	3120	1	1	0	#;NULO!	#;NULO!	1	5	0	0	
23	1	1	NAVIA PUERTO DE VEGA	JARRIO	1	1,00	46,00	2	3800	1	0	1	#;NULO!	#;NULO!	1	6	0	0	
24	1	1	FRANCO (EL)	JARRIO	1	1,00	122,00	5	2860	1	0	1	#;NULO!	#;NULO!	1	6	0	0	
25	2	1	TINEO	CANGAS NARCEA	2	1,00	33,00	2	3120	1	0	1	#;NULO!	#;NULO!	1	2	0	0	
26	1	1	VALDES	JARRIO	1	1,00	55,00	2	3200	0	1	0	1	1	1	2	0	2	
27	1	1	NAVIA	JARRIO	1	1,00	99,00	4	3075	0	1	0	1	2	1	2	0	1	
28	1	1	VALDES	JARRIO	1	1,00	56,00	2	2600	1	1	0	#;NULO!	#;NULO!	1	1	0	1	
29	1	1	VALDES	JARRIO	1	1,00	22,00	1	3470	1	0	1	#;NULO!	#;NULO!	1	4	0	3	
30	3	1	AVILES	SAN AGUSTIN	3	2,99	30,00	1	3010	1	1	0	#;NULO!	#;NULO!	1	4	0	0	
31	3	1	GOZON	SAN AGUSTIN	3	2,99	12,00	1	3400	0	1	0	0	0	#;NULO!	0	6	0	
32	3	1	GOZON	SAN AGUSTIN	3	2,99	20,00	1	3640	1	0	1	#;NULO!	#;NULO!	1	7	0	0	
33	5	1	GIJON	NAVIA	JARRIO	1	1,00	88,00	3	2700	1	0	1	#;NULO!	#;NULO!	1	6	0	1
34	6	1	LLANES	ORIENTE	6	1,00	54,00	2	3500	1	0	1	#;NULO!	#;NULO!	1	4	0	1	
35	1	1	COAÑA	JARRIO	1	1,00	94,00	4	3070	0	1	0	1	2	1	6	0	4	
36	1	1	VALDES	JARRIO	1	1,00	164,00	6	3020	0	1	0	1	8	1	4	0	1	
37	1	1	VEGADEO	JARRIO	1	1,00	160,00	6	3940	1	0	1	#;NULO!	#;NULO!	1	3	0	1	
38	1	1	NAVIA	JARRIO	1	1,00	113,00	4	2870	0	1	0	1	12	1	2	0	1	
39	3	1	AVILES	SAN AGUSTIN	3	2,99	180,00	6	3280	0	1	0	0	0	#;NULO!	0	5	0	
40	3	1	AVILES	SAN AGUSTIN	3	2,99	120,00	4	3650	1	1	0	#;NULO!	#;NULO!	1	6	0	0	
41	1	1	VEGADEO	JARRIO	1	1,00	130,00	5	3580	0	1	0	1	8	1	10	0	1	
42	3	1	CASTRILLON	SAN AGUSTIN	3	2,99	171,00	6	3950	0	1	0	1	0	1	10	1	1	
43	3	1	AVILES	SAN AGUSTIN	3	2,99	187,00	6	4260	1	1	0	#;NULO!	#;NULO!	1	4	0	0	
44	3	1	AVILES	SAN AGUSTIN	3	2,99	164,00	6	3040	0	1	0	1	10	1	3	0	0	
45	4	1	OVIEDO	CANGAS NARCEA	2	1,00	118,00	4	2800	1	1	0	#;NULO!	#;NULO!	1	6	0	0	
46	3	1	PRAVIA	SAN AGUSTIN	3	2,99	5,00	1	3750	0	1	0	0	0	#;NULO!	0	8	0	
47	6	1	CANGAS DE ONIS	ORIENTE	6	1,00	51,00	2	3545	1	0	1	#;NULO!	#;NULO!	1	12	0	0	
48	1	1	COAÑA	JARRIO	1	1,00	148,00	5	3670	0	1	0	0	#;NULO!	#;NULO!	0	4	0	1
49	1	1	COAÑA	JARRIO	1	1,00	106,00	4	3790	1	1	0	#;NULO!	#;NULO!	1	8	0	2	
50	3	1	CASTRILLON	SAN AGUSTIN	3	2,99	122,00	5	3600	1	1	0	#;NULO!	#;NULO!	1	#;NULO!	1	5	
51	3	1	CASTRILLON	SAN AGUSTIN	3	2,99	20,00	1	2950	0	1	0	0	0	#;NULO!	0	6	0	
52	3	1	AVILES	SAN AGUSTIN	3	2,99	38,00	2	3750	1	1	0	#;NULO!	#;NULO!	1	4	0	0	
53	3	1	CASTRILLON	MIRANDA(AVILES)	SAN AGUSTIN	3	2,99	28,00	1	3890	1	0	1	0	1	#;NULO!	#;NULO!	0	0

Depuración

- Listar variables
- Descripción univariante
- Valores aberrantes, fuera de recorrido, fechas, duplicados,...
- ¿Qué hacer?
 - Ir al registro original
 - Reemplazar
 - Eliminar

Transformación de variables

- Cuantitativas \longrightarrow Categóricas
- Reducción de categorías
- Variables *dummy*
- Estabilización de la varianza y normalización: logaritmos, potencias, inversos, etc.

Análisis Univariante (i)

- V. Cualitativas/categóricas

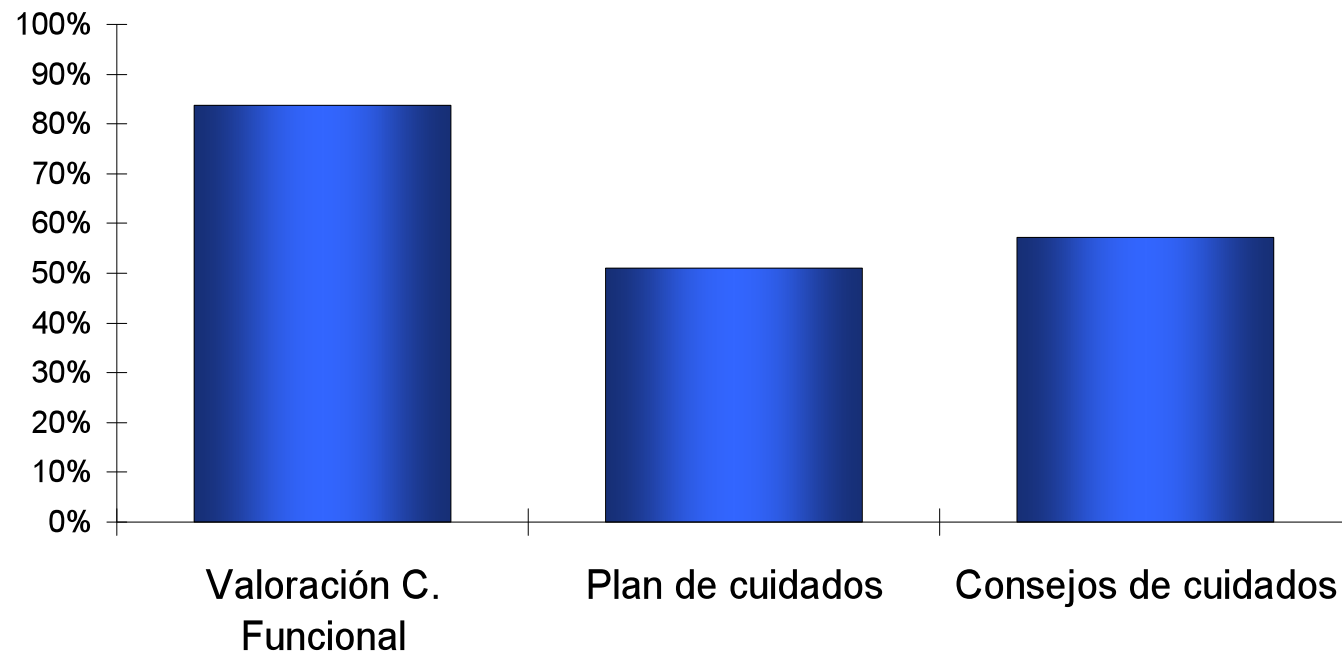
- Distribución de frecuencias
- Gráficos (barras, sectores, etc.)

- V. Cuantitativas

- Medidas de tendencia central y dispersión
- Histogramas

Gráficos

Cumplimiento de ... y



Análisis Univariante (ii)

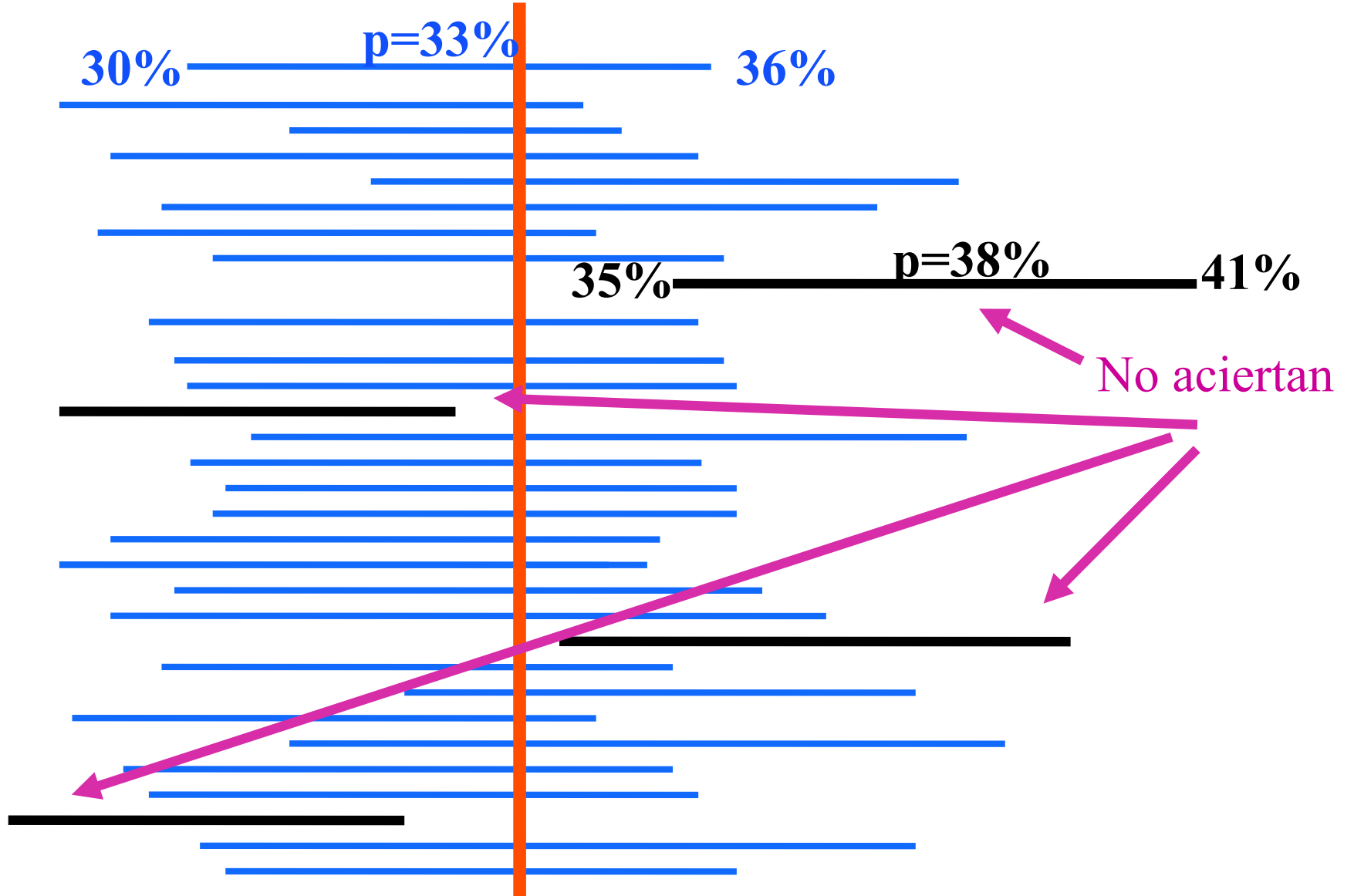
■ Intervalo de Confianza (*media, proporción,...*)

– ¿Cómo se interpreta?

– ¿Cómo se calcula?

– ¿Cómo se presenta?

P=34%



Algunos disparates:

- a. “Que 3% no podemos estar seguros que sean fumadores”
- b. “que el 95% de asturianos están entre el 30% y el 36% como fumadores habituales”
- c. “que tenemos una probabilidad del 95% de encontrar al 34% de los fumadores”
- d. “Entre 30-36% de una muestra de asturianos adultos son fumadores habituales con una probabilidad del 95%”
- e. “Hay un 95% probabilidades de que el % de fumadores habituales de nuestra muestra esté entre el 30 y el 36%”
- f. “Hay un 95% de posibilidades de que los fumadores habituales se encuentren entre el 30 y el 36%, de nuestra muestra”
- g. “Que la proporción de fumadores estimada en este estudio es del 33% y que el 95% de los valores se encuentran entre 30 y 36%”
- h. “Que con el 95% de confianza incluye al 33%”

Análisis bivariante (i):

¿Asociación o independencia?

- **2 V. Cuantitativas:**

Regresión y Correlación

- **1 Cuanti/1 Categórica (k=2):**

Comparación de medias

(t Student o U de McWitney)

- **1 Cuanti/1 Categórica (k>2): Anova**

- **2 Categóricas: z, Ji-cuadrado,**

McNemar, etc.

Análisis bivariante (ii)

- ¿Qué es un valor P?

INFERENCIA ESTADÍSTICA

Fisher (década de los 20)

$H_0: d=0$

Se observa d_0

$$p = P(d \geq d_0 | H_0)$$

Medida de la discrepancia de los datos con la hipótesis, llamada a tener un papel informal (**no especificado**), junto con el resto de la información, en el flujo inferencial

INFERENCIA ESTADÍSTICA

Neyman y Pearson
(década de los 30)

$H_0: d=0$

$H_1: d \neq 0$

Se observa d_0

se computa

$$p = P(D > d_0 | H_0)$$

Y se adopta una **decisión**

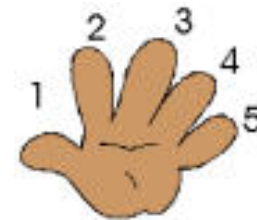
Si $p \leq \alpha$ se rechaza

$H_0: d=0$

¿ $\alpha = 0,05$?

Si $p > \alpha$ no se rechaza

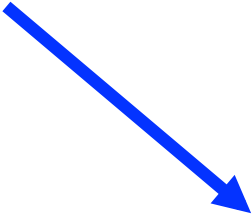
$H_0: d=0$




Crítica principal para los valores p

d grande
n pequeña

d pequeña
n grande


$$p = P(d > d_0 | H_0)$$


$$p = P(d > d_0 | H_0)$$

$$p = P(d > d_0 | H_0)$$

Con mucha frecuencia se sabe de antemano que las hipótesis de nulidad son falsas incluso antes de recoger los datos; en ese caso el rechazo o la aceptación, simplemente es un reflejo del tamaño de la muestra y no hace, por tanto, contribución alguna a la ciencia. *Savage (1957)* ↑



Resulta preferible omitir los valores p de los informes sobre estudios epidemiológicos puesto que pueden emplearse las estimaciones puntuales y los intervalos de confianza (además de otras alternativas)



Análisis bivariante (ii)

**Comité Internacional de Directores de
Revistas Médicas
(*Grupo de Vancouver*)**

Se evitará la dependencia de las pruebas estadísticas de verificación de hipótesis, tal como el uso de los valores P , que no aportan ninguna información importante.

Análisis bivariante (iii)

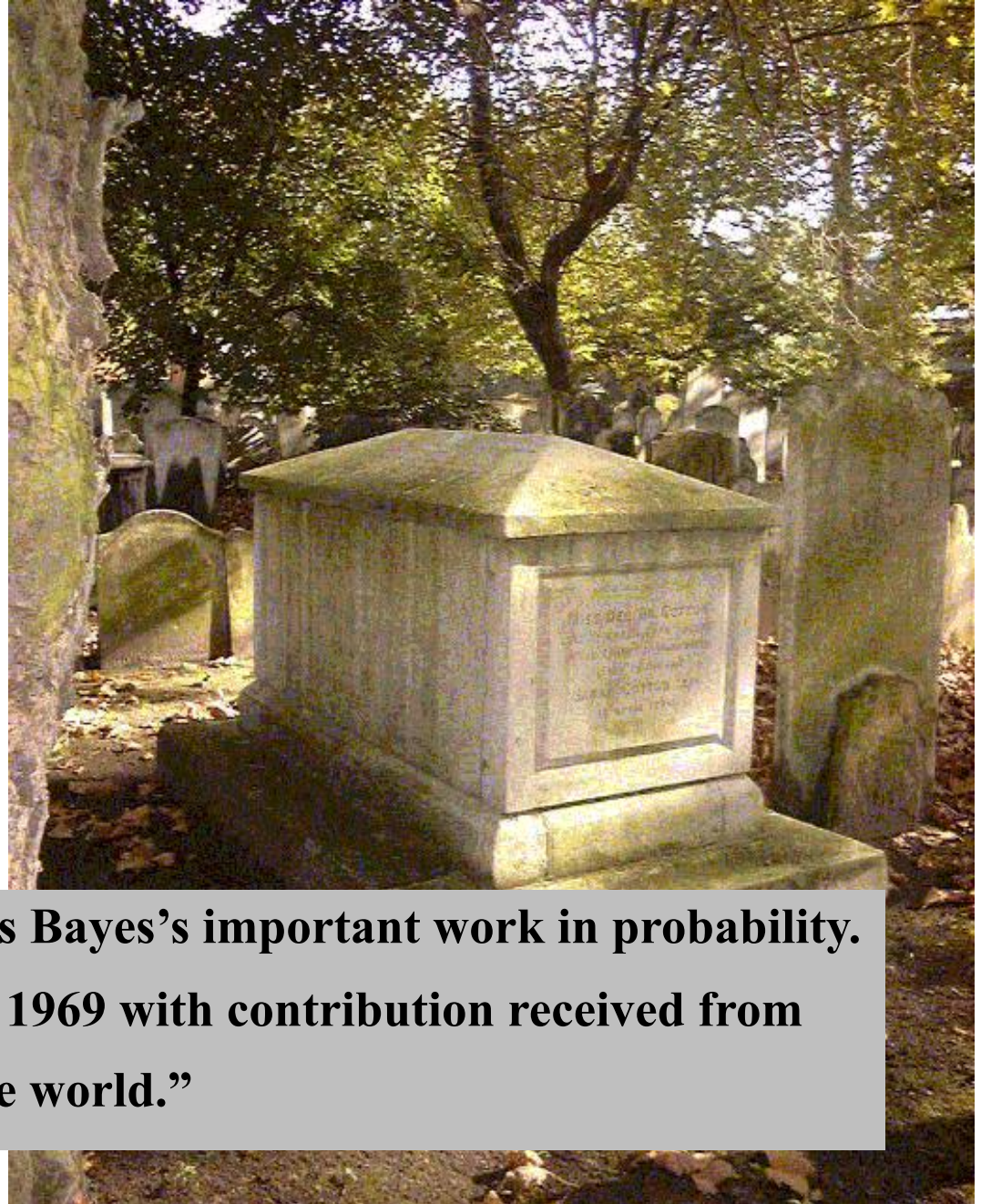
1047 artículos de Atención Primaria(1992-2000)

- **812 usan el valor p (78%)**
- **155 usan Int de Conf (15%)**
- **182 usan términos como “muestra significativa”, “las probabilidades de enfermar” (17%)**
- **79 concluyen que “hallamos significación” (8%)**

**Intervalos de confianza
como alternativa**



Rev Thomas Bayes
1701-1761



“In recognition of Thomas Bayes’s important work in probability. The vault was restored in 1969 with contribution received from statisticians throughout the world.”

Análisis Multivariante (i)

- Diversas técnicas capaces de controlar la **confusión** y de detectar **interacciones**.
- Variable respuesta (dependiente)
- Variables exposición (independientes)
- Variables de control

Análisis Multivariante (ii)

- Regresión lineal múltiple ($Y = B_0 + B_1X_1 + \dots$)
- (Regresión no lineal múltiple)
- Regresión logística
- Regresión de Poisson
- Modelos lineales generalizados
- Regresión de Cox

Para qué NO sirve el análisis de los datos

- No controla los Sesgos de selección:

- Autoselección

- Berkson

- Pérdidas en el seguimiento

- Supervivencia relativa

Para qué NO sirve el análisis de los datos

■ No controla los sesgos de información:

Memoria

Diagnóstico

Veracidad

Tipo de cuestionario

Distinta cumplimentación según edad...

Para qué SI sirve el análisis de los datos

- Control de Confusión....
- Detección de Interacción

X está asociado con Y

	Explicación	Tipo	Fuente	Realidad	Modelo Causal
A	Azar	Espuria	Error Aleatorio	No hay relación causal	NINGUNO
B	Sesgo	Espuria	Error Sistemático	No hay relación causal	NINGUNO
C	Efecto-causa	Verdadera	Medición Inversa	La causa es realmente efecto	$Y \rightarrow X$
D	Efecto-efecto	Verdadera	Factores Confusores	Otros factores explican la asociación	$Z \rightarrow X$ $Z \rightarrow Y$
E	Causa-efecto	Verdadera	Relación Causal	$X \rightarrow Y$	$X \rightarrow Y$

SOFTWARE (1)

- Grandes paquetes estadísticos:
 - SAS
 - SPSS
 - BMDP
 - MINITAB
 - STATA
 - R

SOFTWARE (2)

■ Libre distribución:

- Epiinfo (www.cdc.gov)
- Epidat 3.1 (en breve, versión 4.0)
- R (www.r-project.org)

Plan de implementación

Plan de implementación

- Cronograma
- Cuantificación de recursos y costes
 - Recursos físicos
 - Costes
 - Personal
 - Capacitación
- Supervisión y coordinación interna
 - Reuniones de coordinación, controles de calidad
- Presupuesto

Cronograma

- Enunciar y listar las tareas
- Ordenar lógicamente
- Marcar para cada una las forzosamente precedentes
- Tiempo previsto de cada una de ellas (máximo y mínimo)
- Diagramas de Gantt o PERT
- Revisar periódicamente

Cronograma

	Noviembre		Diciembre		
	22-23	26-30	3-7	10-12	13-14
Preparación de listados y bases de datos					
Llamadas telefónicas					
Procesamiento de la información					
Elaboración del informe					

Tareas

- 1) Plan de análisis: Al menos el estadístico/os para la pregunta
- 2) Limitaciones del estudio para responder a la pregunta.
- 3) Cronograma