Numerical Summary Measures

With thanks to Dr. Rino Bellocco and Prof. Marcello Pagano

1

Let x represent FEV_1 in liters

$x_1 = 2.30$
$x_2 = 2.15$
$x_3 = 3.50$
$x_4 = 2.60$
$x_5 = 2.75$
$x_6 = 2.82$
$x_7 = 4.05$
$x_8 = 2.25$
$x_9 = 2.68$
$x_{10} = 3.00$
$x_{11} = 4.02$
$x_{12} = 2.85$
$x_{13} = 3.38$

2

Measures of central tendency

Numbers: $x_1, x_2, \ldots, x_n$

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2.3, 2.15, 3.50, 2.60, 2.75, 2.82, 4.05, 2.25, 2.68, 3.00, 4.02, 2.85 (n=13)

$$\text{Sum} = \sum_{i=1}^{n} x_i = 38.5$$

Mean = $\bar{x} = \frac{38.35}{13} = 2.95$

3

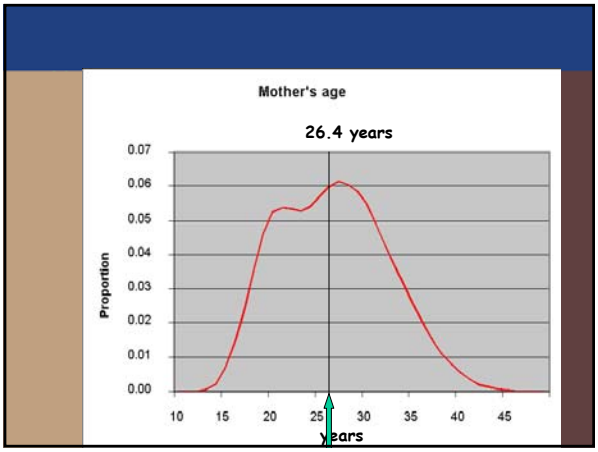
Approx 4 million singleton births, 1991 :

Variable	Mean
Mother's age	26.4 years
Gestational age	39.15 weeks
Birth weight	3358.6 grams
Weight gain*	30.4 lbs

Among those who died within 1 year:

Survival	49.4 days

4



Note what happens when one number, 4.02 say, becomes large, say 40.2 :

2.3, 2.15, 3.50, 2.60, 2.75, 2.82, 4.05, 2.25, 2.68, 3.00, 40.2, 2.85

Mean = $\bar{x}$ = 5.73

(versus 2.95, from before)

Mean is sensitive to every observation, it is not robust.

6

Median

More robust, but not sensitive enough.

Definition: At least 50% of the observations are greater than or equal to the *median*, and at least 50% of the observations are less than or equal to the *median*.

2.15, 2.25, 2.30 --- median = 2.25

2.15, 2.25, 2.30, 2.60 ---

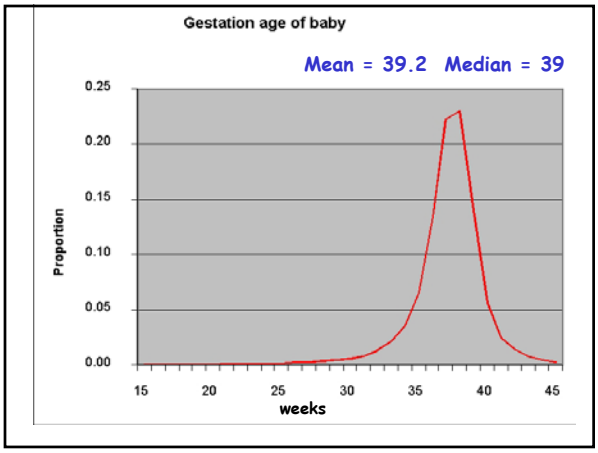
median = $\frac{1}{2}$ (2.25 + 2.30) = 2.275

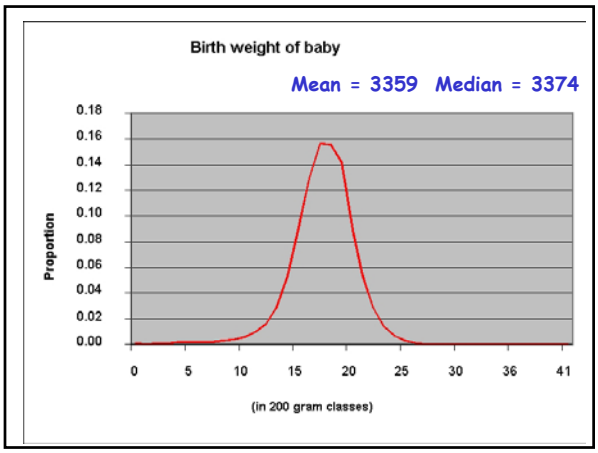
7

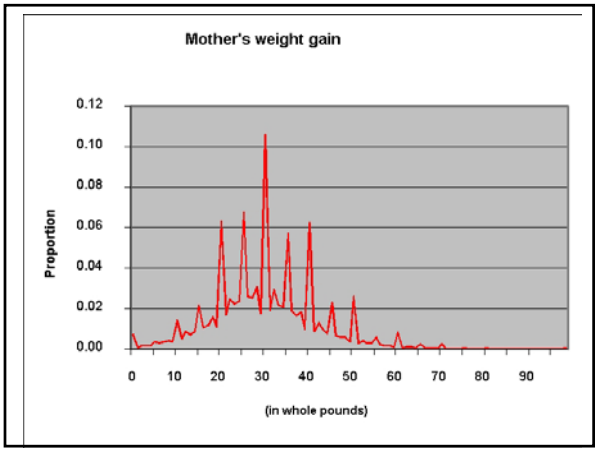
Singleton births, 1991 :

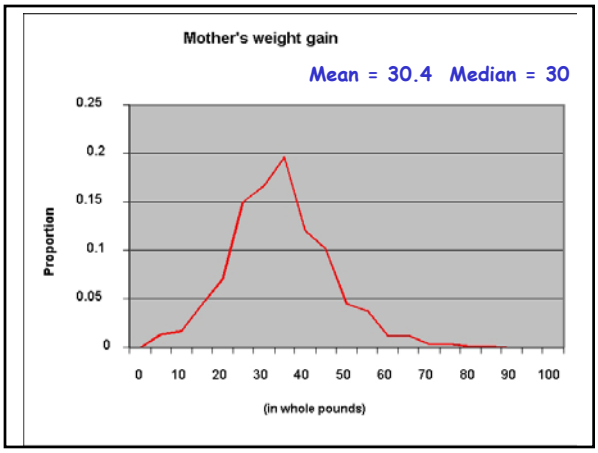
Variable	Mean	Median
Mom's age (yrs)	26.4	25
Gest. Age (wks)	39.2	39
Birth weight (gms)	3359	3374
Weight gain (lbs)	30.4	30
Survival (days)	49.4	7

8




Birth weight of baby
Mean = 3359 Median = 3374
Proportion
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0 5 10 15 20 25 30 36 41
(in 200 gram classes)


Mother's weight gain
Proportion
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0 10 20 30 40 50 60 70 80 90
(in whole pounds)


Mother's weight gain
Mean = 30.4 Median = 30
Proportion
0.25
0.2
0.15
0.1
0.05
0
0 10 20 30 40 50 60 70 80 90 100
(in whole pounds)

Mean or median?

Mean performs best when we have a symmetric distribution with thin tails.

If data are skew, use the median.

Can also report both the mean and median if space permits.

13

Spread:

Singleton births, 1991 :

Variable	Min	Max	Range
Mom's age	10	49	39
Gest. Age	17	47	30
Birth weight	227	8164	7937
Weight gain	0	98	98
Survival	0	363	363

14

FEV_1	$(x_i - \bar{x})$
2.30	-0.65
2.15	-0.80
3.50	0.55
2.60	-0.35
2.75	-0.20
2.82	-0.13
4.05	1.10
2.25	-0.70
2.68	-0.27
3.00	0.05
4.02	1.07
2.85	-0.10
3.38	0.43

15

FEV_1	$(x_i - \bar{x})$
2.30	-0.65
2.15	-0.80
3.50	0.55
2.60	-0.35
2.75	-0.20
2.82	-0.13
4.05	1.10
2.25	-0.70
2.68	-0.27
3.00	0.05
4.02	1.07
2.85	-0.10
3.38	0.43
Total	0.00

16

FEV_1	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2.30	-0.65	0.423
2.15	-0.80	0.640
3.50	0.55	0.303
2.60	-0.35	0.123
2.75	-0.20	0.040
2.82	-0.13	0.169
4.05	1.10	1.210
2.25	-0.70	0.490
2.68	-0.27	0.073
3.00	0.05	0.003
4.02	1.07	1.145
2.85	-0.10	0.010
3.38	0.43	0.185
Total	0.00	4.66

17

$$Variance = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

e.g.

$$= \frac{4.66}{12} = 0.39 \text{ liters}^2$$

18

Standard deviation = $+\sqrt{\text{Variance}}$

e.g.

$= \sqrt{0.39}$

$= 0.62$ liters

19

Singleton births, 1991 :

Variable	Mean	Std dev
Mom's age (yrs)	26.4	5.84
Gest. Age (wks)	39.2	2.61
Birth weight (gms)	3359	227
Weight gain (lbs)	30.4	12.13
Survival (days)	49.4	76.1

20

Empirical Rule:

If dealing with a *unimodal* and *symmetric* distribution, then

Mean ± 1 sd covers approx 67% obs.

Mean ± 2 sd covers approx 95% obs

Mean ± 3 sd covers approx all obs

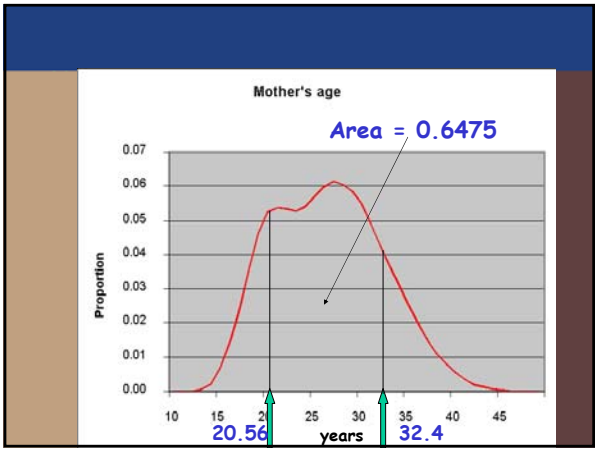
21

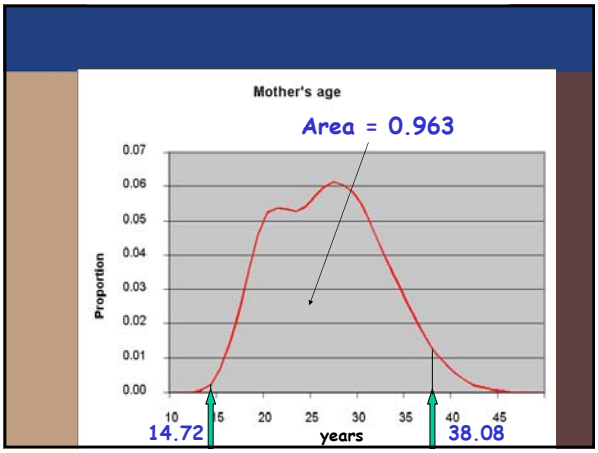
Mother's age: mean = 26.4 yrs
s.d. = 5.84 yrs

Table of $\bar{x}$ ± k s.d.s

k	left limit	right limit	Emp.
1	20.56	32.24	67%
2	14.72	38.08	95%
3	8.88	43.92	all

22





Mother's age: mean = 26.4 yrs
s.d. = 5.84 yrs

Table of $\bar{x}$ ± k s.d.s

k	left limit	right limit	Emp.	Actual
1	20.56	32.24	67%	64.75%
2	14.72	38.08	95%	96.3%
3	8.88	43.92	all	99.89%

25

InterQuantile Range

IQR= Q-q

Q = 75th percentile

q = 25th percentile

26

Table 1. Baseline Descriptive Statistics*

Variable	CHA Cohort (Aged 18-39 y)	PG Cohort (Aged 25-39 y)	MRFIT Cohort (Aged 35-39 y)
No. of men	11 017	1266	69 205
Age, y	29.7 (5.5)	32.5 (3.9)	37.0 (1.4)
Serum cholesterol level, mg/dL†	189.9 (36.2)	209.0 (38.7)	205.7 (38.7)
Systolic blood pressure, mm Hg	134.6 (15.4)	125.3 (14.6)	126.4 (13.3)
Diastolic blood pressure, mm Hg	78.2 (10.5)	77.8 (8.8)	81.7 (10.1)
Cigarette smokers, %	47.4	65.6	39.6
Cigarettes/d, No. (smokers)	21.2 (10.5)	21.7 (7.5)	25.4 (13.1)
Cigarettes/d, No. (all)	10.0 (12.8)	14.3 (12.0)	10.0 (14.9)
African American, %	8.1	8.0	6.7
Education, y	13.8 (2.6)	11.8 (2.4)	14.2 (0.8)
ECG abnormality			
Major, %	3.3	1.8‡	NA§
Minor, %	5.5	4.1‡	NA§
Body mass index, kg/m²	26.0 (3.6)	26.2 (3.4)	NA§

*CHA indicates Chicago Heart Association Detection Project in Industry; PG, Peoples Gas Company Study; MRFIT, Multiple Risk Factor Intervention Trial; ECG, electrocardiogram; and NA, not available. Data are presented as mean (SD) unless otherwise indicated.
†To convert mg/dL to mmol/L, multiply by 0.0259.
‡Data are not available for 428 men.
§Data were not collected on these variables at first screening for recruitment for MRFIT.