Unit 2: Some Basics

On the accuracy of statistical procedures in Microsoft Excel 97

B. D. McCullough, and **Berry Wilson**

Federal Communications Commission, 445 12th St. SW Room 2C-134, Washington, DC 20554, USA

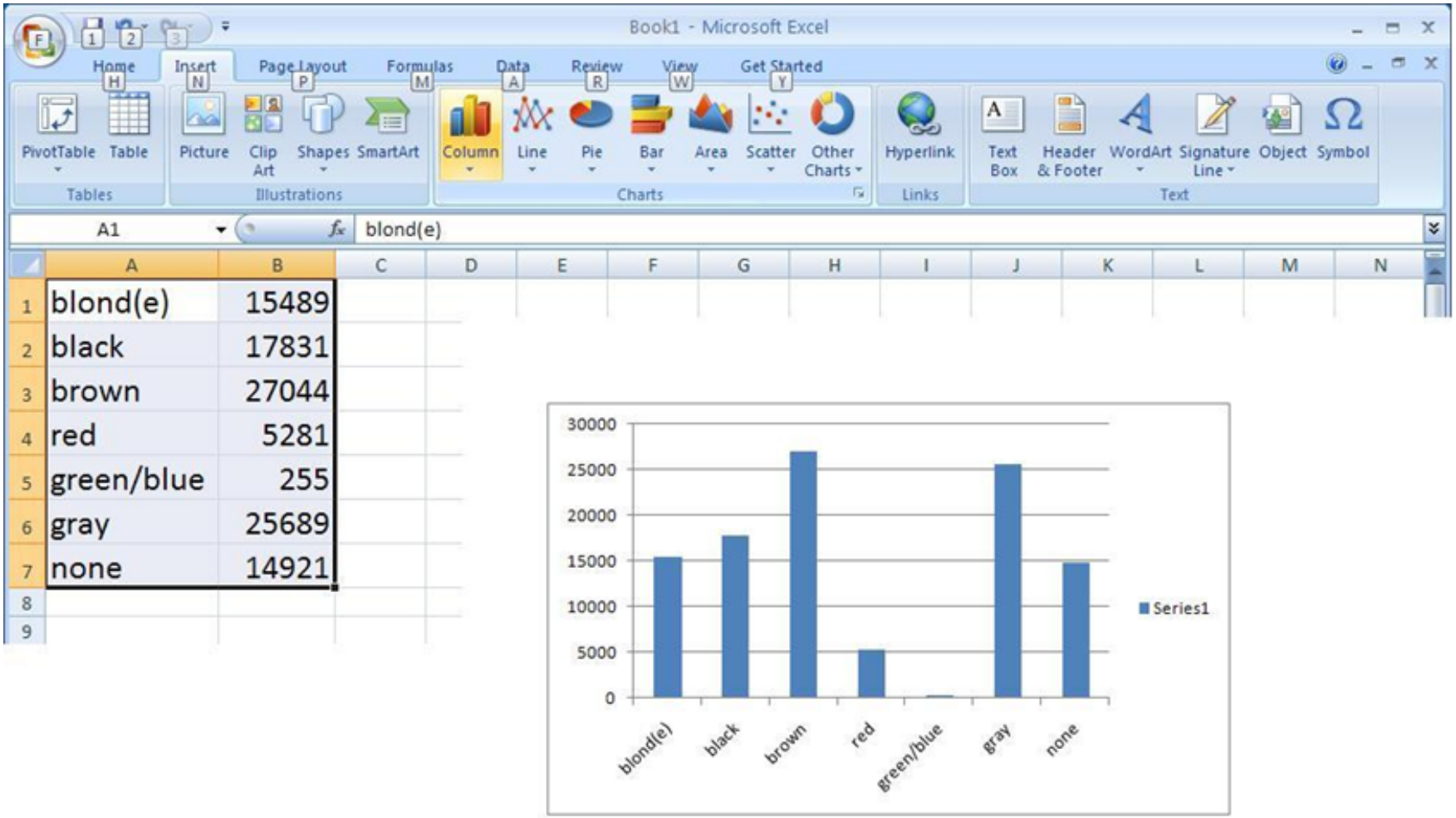
Received 1 June 1998; revised 1 December 1998. Available online 10 August 1999.

Abstract

The reliability of statistical procedures in Excel are assessed in three areas: estimation (both linear and nonlinear); random number generation; and statistical distributions (e.g., for calculating *p*-values). Excel's performance in all three areas is found to be inadequate. Persons desiring to conduct statistical analyses of data are advised not to use Excel.

Author Keywords: DIEHARD; ELV; Numerical accuracy; Software testing; StRD

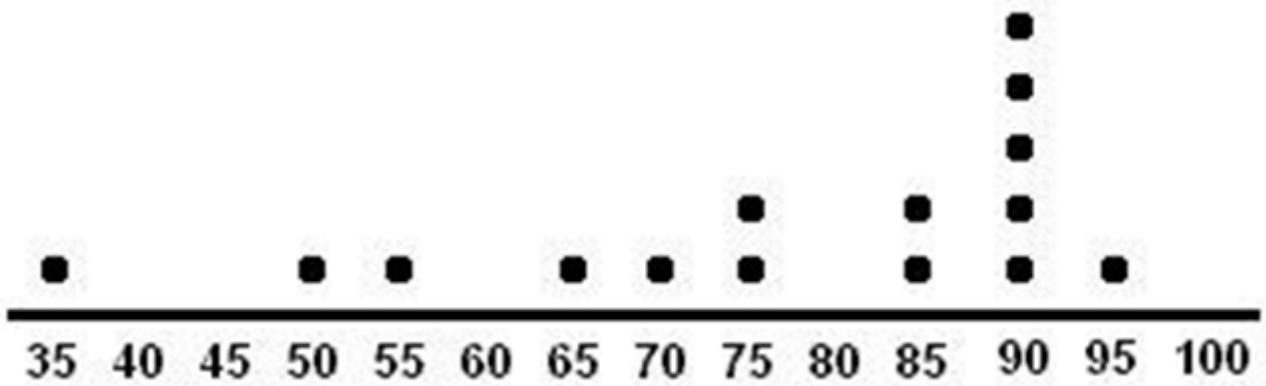
Example: Hair color at NYS Fair



Numerical variables: Graphs of frequencies I

dot plots
stem-and-leaf plots

Ex: 79, 91, 59, 52, 94, 74, 75, 87, 67, 35, 91, 89, 96, 92, 92



3	
3	5
4	
4	
5	2
5	9
6	
6	7
7	4
7	5 9
8	
8	7 9
9	1 1 2 2 4
9	6

Match Variable to Dotplot

A. jersey numbers of 2011 Colgate football players

B. annual snowfall amounts for a sample of U.S. cities

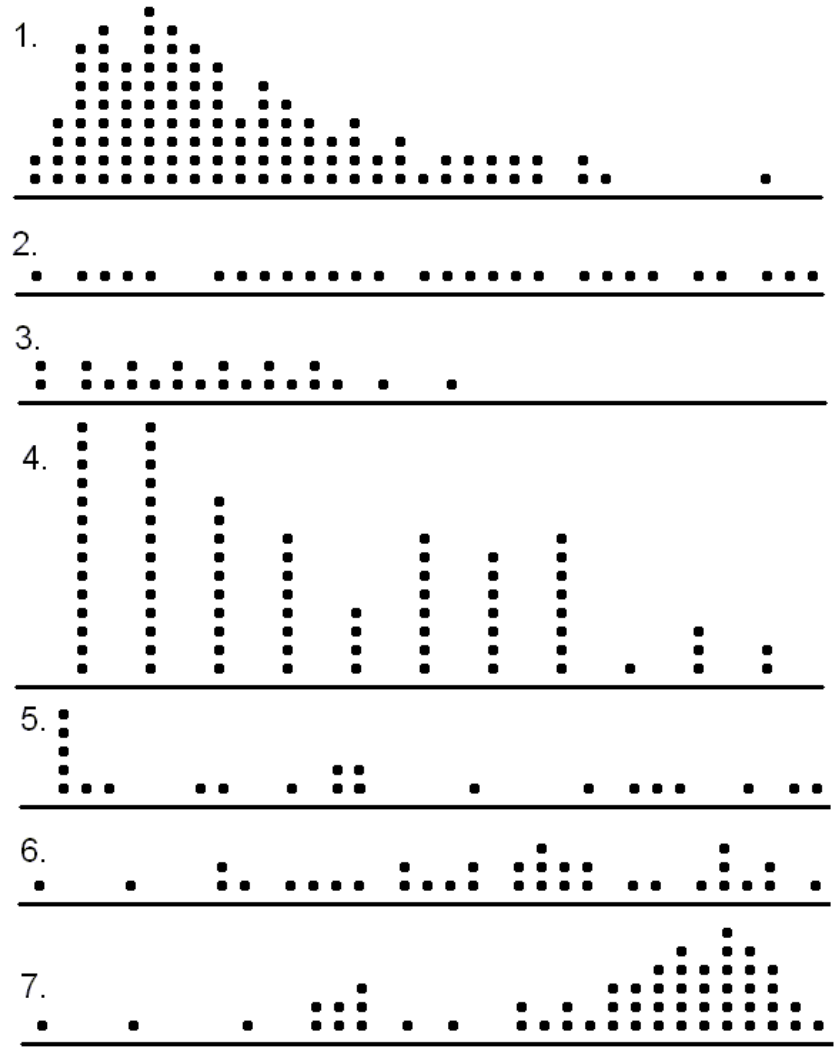
C. margins of victory in a sample of Major League Baseball games

D. prices of properties in the Monopoly board game

E. ages at which a sample of mothers had their first child

F. scores on a statistics exam

G. weights of 1999 cars



Match Variable to Dotplot

A. jersey numbers of 2011 Colgate football players

B. annual snowfall amounts for a sample of U.S. cities

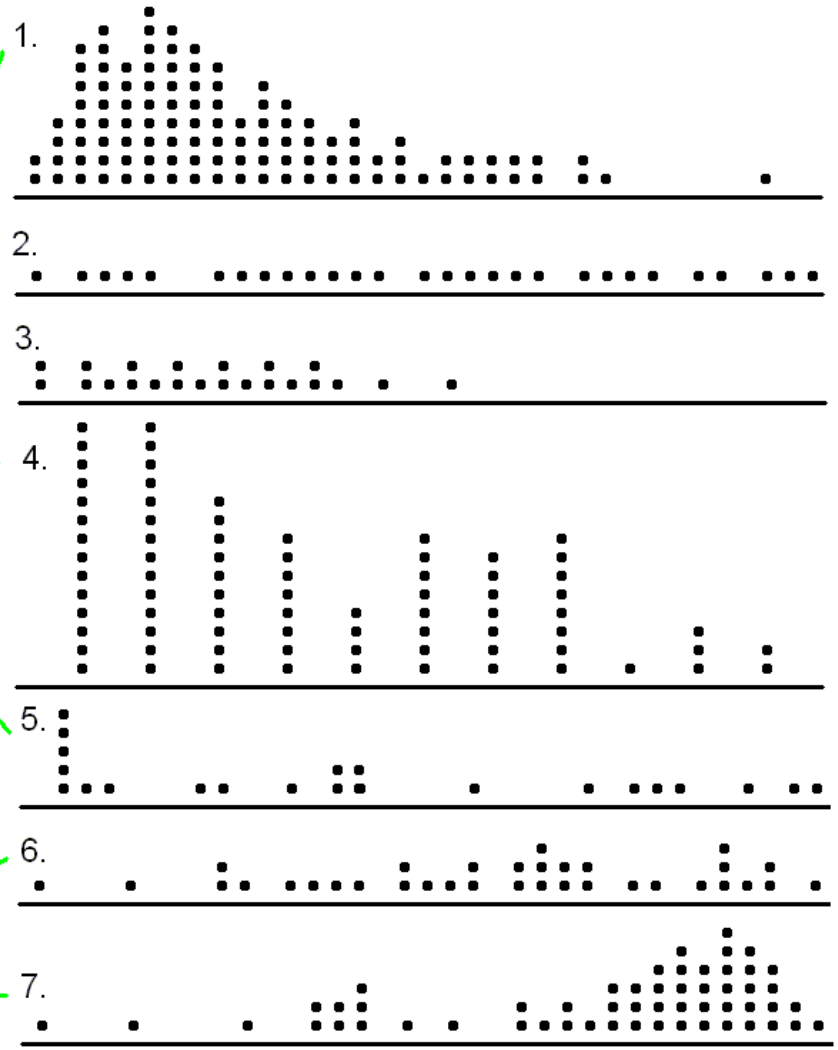
C. margins of victory in a sample of Major League Baseball games

D. prices of properties in the Monopoly board game

E. ages at which a sample of mothers had their first child

F. scores on a statistics exam

G. weights of 1999 cars

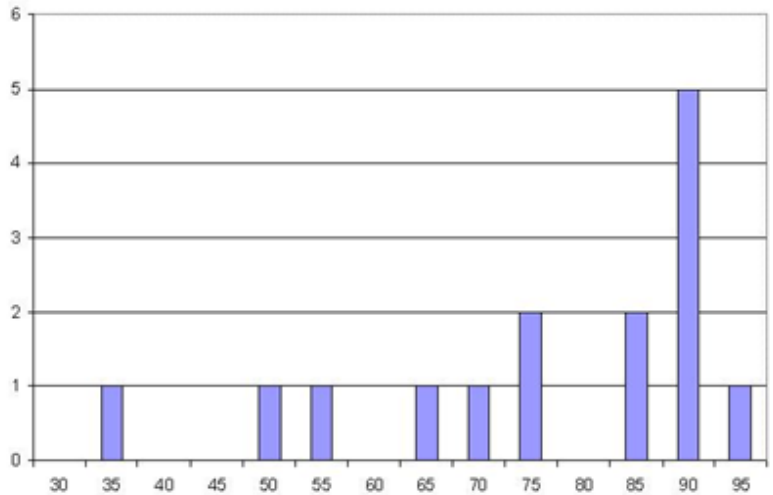


Numerical variables: Graphs of frequencies II

histogram

Same data: 79, 91, 59, 52, 94, 74, 75, 87, 67, 35, 91, 89, 96, 92, 92

(Note similarity to dotplot and stem-and-leaf plot)



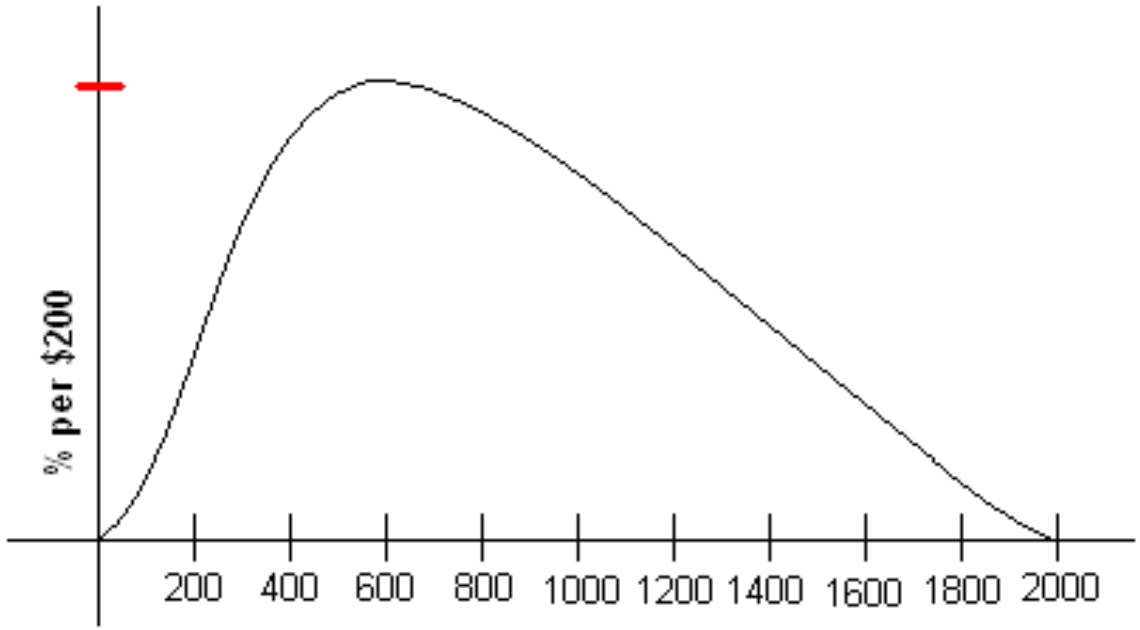
Excel's "histograms" aren't, unless bar widths are equal.

Density scale

Weekly salaries in a company:

Vertical axis is in % per \$200.

Where is the high point?



Reading a histogram

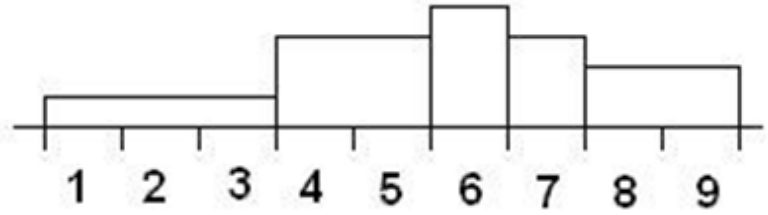
Hrs slept by CU students

(Data from questionnaire)

Hrs	Areas	
1,2,3	$3 \times 1 = 3$	$3/20 = 15\%$
4,5	$2 \times 3 = 6$	$6/20 = 30\%$
6	$1 \times 4 = 4$	$4/20 = 20\%$
7	$1 \times 3 = 3$	$3/20 = 15\%$
8,9	$2 \times 2 = 4$	$4/20 = 20\%$
	Sum $= 20$	

So maybe 5% slept 1hr, 5% 2hr and 5% 3hr;

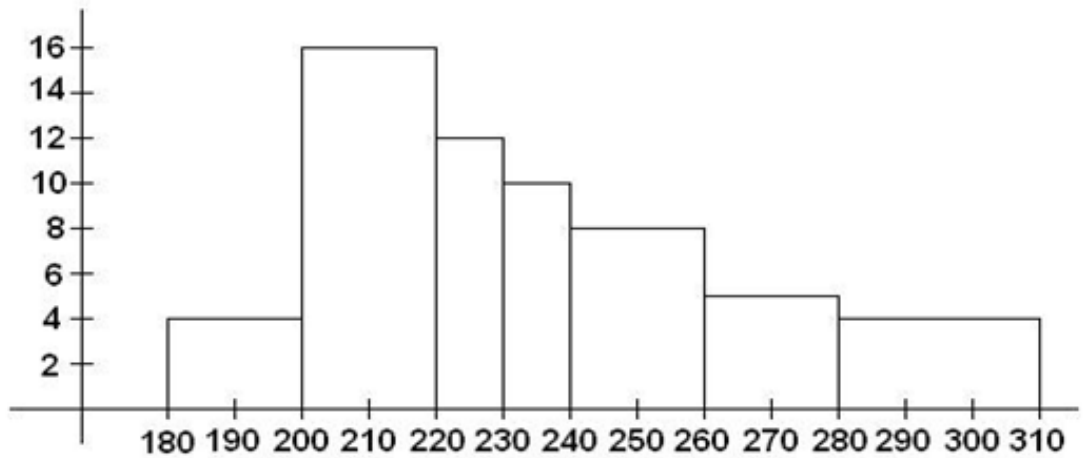
or maybe 0% slept 1hr, 7% 2hr, 8% 3hr



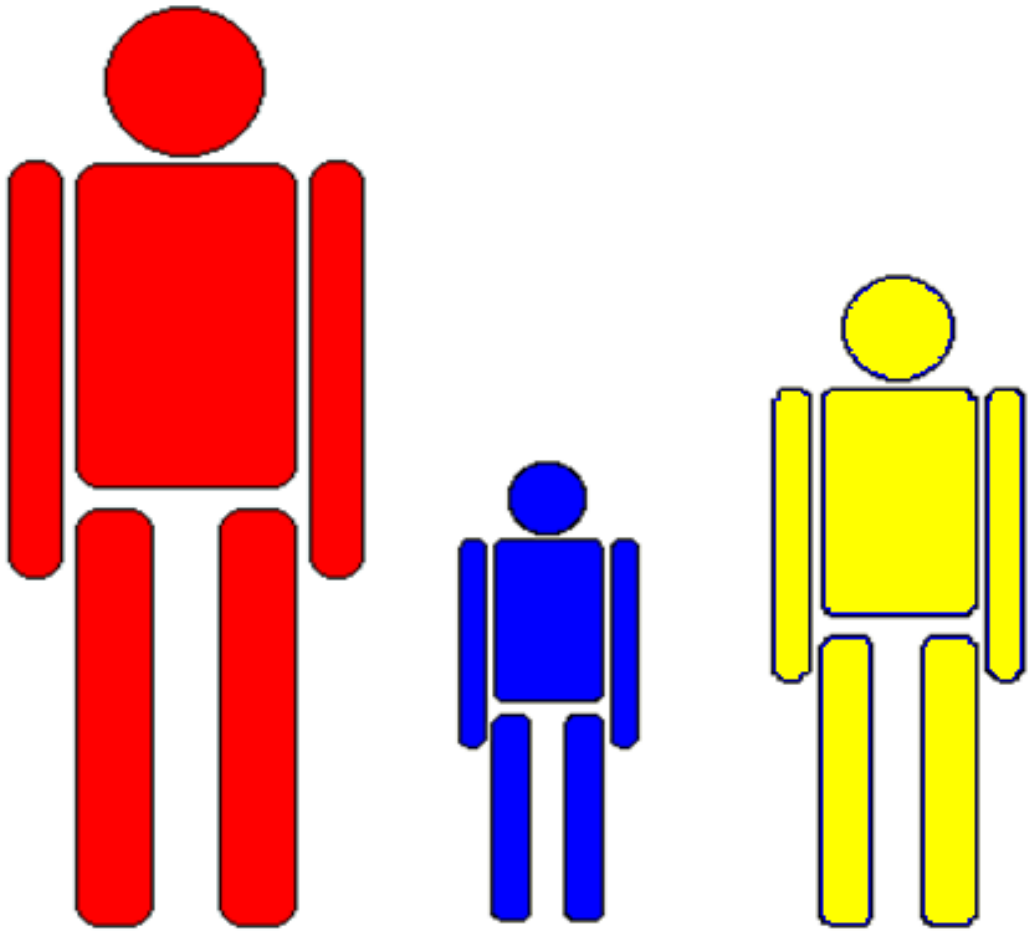
1. Which given interval contains the most students?
2. Which 1-hr period contains the most students (i.e., is most "crowded")?
3. About what % slept 8 hr?
4. About what % slept 3 or 4 hr?
5. If there were 240 surveyed, about how many slept 6–7 hr?

Drawing a histogram: Horses' weights in kg

Wts (widths)	Counts	%	%/10 kg
180–200 (20)	20	8	4
200–220 (20)	80	32	16
220–230 (10)	30	12	12
230–240 (10)	25	10	10
240–260 (20)	40	16	8
260–280 (20)	25	10	5
280–310 (30)	30	12	4
Sum	250		



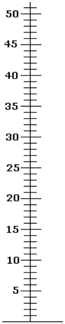
Frequencies by area?



Computing your "digital index"

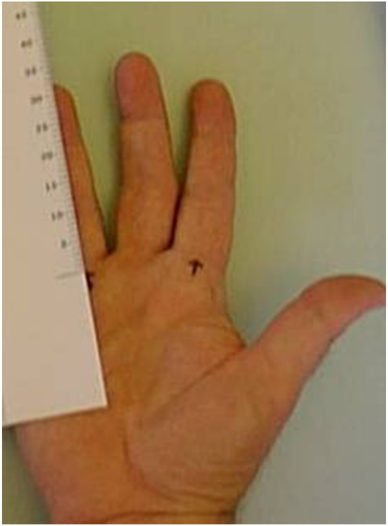
length of index finger / length of ring finger

(dominant hand)



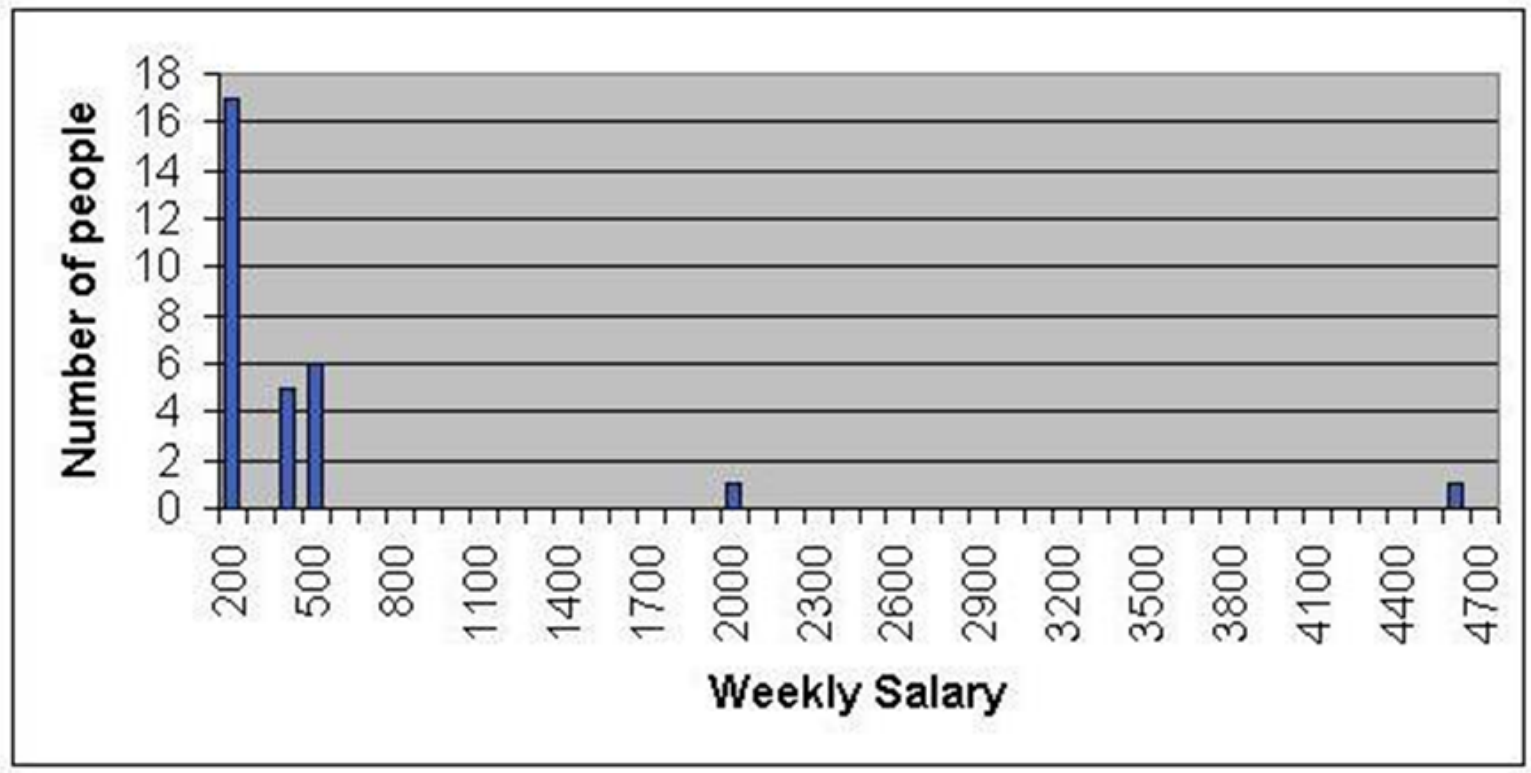
INDEX / RING

20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
21	1.00	1.05	1.10	1.14	1.19	1.24	1.29	1.33	1.38	1.43	1.48	1.52	1.57	1.62	1.67	1.71	1.76	1.81
22	0.95	1.00	1.05	1.09	1.14	1.18	1.23	1.27	1.32	1.36	1.41	1.45	1.50	1.55	1.59	1.64	1.68	1.73
23	0.91	0.96	1.00	1.04	1.09	1.13	1.17	1.22	1.26	1.30	1.35	1.39	1.43	1.48	1.52	1.57	1.61	1.65
24	0.88	0.92	0.96	1.00	1.04	1.08	1.13	1.17	1.21	1.25	1.29	1.33	1.38	1.42	1.46	1.50	1.54	1.58
25	0.84	0.88	0.92	0.96	1.00	1.04	1.08	1.12	1.16	1.20	1.24	1.28	1.32	1.36	1.40	1.44	1.48	1.52
26	0.81	0.85	0.88	0.92	0.96	1.00	1.04	1.08	1.12	1.15	1.19	1.23	1.27	1.31	1.35	1.38	1.42	1.46
27	0.78	0.81	0.85	0.89	0.93	0.96	1.00	1.04	1.07	1.11	1.15	1.19	1.22	1.26	1.30	1.33	1.37	1.41
28	0.75	0.79	0.82	0.86	0.89	0.93	0.96	1.00	1.04	1.07	1.11	1.14	1.18	1.21	1.25	1.29	1.32	1.36
29	0.72	0.76	0.79	0.83	0.86	0.90	0.93	0.97	1.00	1.03	1.07	1.10	1.14	1.17	1.21	1.24	1.28	1.31
30	0.70	0.73	0.77	0.80	0.83	0.87	0.90	0.93	0.97	1.00	1.03	1.07	1.10	1.13	1.17	1.20	1.23	1.27
31	0.68	0.71	0.74	0.77	0.81	0.84	0.87	0.90	0.94	0.97	1.00	1.03	1.06	1.10	1.13	1.16	1.19	1.23
32	0.66	0.69	0.72	0.75	0.78	0.81	0.84	0.88	0.91	0.94	0.97	1.00	1.03	1.06	1.09	1.13	1.16	1.19
33	0.64	0.67	0.70	0.73	0.76	0.79	0.82	0.85	0.88	0.91	0.94	0.97	1.00	1.03	1.06	1.09	1.12	1.15
34	0.62	0.65	0.68	0.71	0.74	0.76	0.79	0.82	0.85	0.88	0.91	0.94	0.97	1.00	1.03	1.06	1.09	1.12
35	0.60	0.63	0.66	0.69	0.71	0.74	0.77	0.80	0.83	0.86	0.89	0.91	0.94	0.97	1.00	1.03	1.06	1.09
36	0.58	0.61	0.64	0.67	0.69	0.72	0.75	0.78	0.81	0.83	0.86	0.89	0.92	0.94	0.97	1.00	1.03	1.06
37	0.57	0.59	0.62	0.65	0.68	0.70	0.73	0.76	0.78	0.81	0.84	0.86	0.89	0.92	0.95	0.97	1.00	1.03
38	0.55	0.58	0.61	0.63	0.66	0.68	0.71	0.74	0.76	0.79	0.82	0.84	0.87	0.89	0.92	0.95	0.97	1.00
39	0.54	0.56	0.59	0.62	0.64	0.67	0.69	0.72	0.74	0.77	0.79	0.82	0.85	0.87	0.90	0.92	0.95	0.97
40	0.53	0.55	0.58	0.60	0.63	0.65	0.68	0.70	0.73	0.75	0.78	0.80	0.83	0.85	0.88	0.90	0.93	0.95
41	0.51	0.54	0.56	0.59	0.61	0.63	0.66	0.68	0.71	0.73	0.76	0.78	0.80	0.83	0.85	0.88	0.90	0.93
42	0.50	0.52	0.55	0.57	0.60	0.62	0.64	0.67	0.69	0.71	0.74	0.76	0.79	0.81	0.83	0.86	0.88	0.90
43	0.49	0.51	0.53	0.56	0.58	0.60	0.63	0.65	0.67	0.70	0.72	0.74	0.77	0.79	0.81	0.84	0.86	0.88
44	0.48	0.50	0.52	0.55	0.57	0.59	0.61	0.64	0.66	0.68	0.70	0.73	0.75	0.77	0.80	0.82	0.84	0.86
45	0.47	0.49	0.51	0.53	0.56	0.58	0.60	0.62	0.64	0.67	0.69	0.71	0.73	0.76	0.78	0.80	0.82	0.84



Digital ratios spreadsheet

Weekly salaries in a small factory: 30 workers
17 get \$200, 5 \$400, 6 \$500, 1 \$2000, 1 \$4600
avg = \$500, median = \$200



Test for skew

median $<$ avg : skewed to right
median $>$ avg : skewed to left
(These can be used as the def of left or right skew.)

Changing the list (I)

"Linear" changes of variable (i.e., changing units of measure)

Basic list: 6, 9, 15

$$\mu = \frac{6+9+15}{3} = 10$$

$$\sigma = \sqrt{\frac{(6-10)^2+(9-10)^2+(15-10)^2}{3}} = \sqrt{14}$$

Add 5: 11, 14, 20

$$\mu = \frac{11+14+20}{3} = 15$$

$$\sigma = \sqrt{\frac{(11-15)^2+(14-15)^2+(20-15)^2}{3}} = \sqrt{14}$$

Multiply by 12: 72, 108, 180

$$\mu = \frac{72+108+180}{3} = 120$$

$$\sigma = \sqrt{\frac{(72-120)^2+(108-120)^2+(180-120)^2}{3}} = 12\sqrt{14}$$

Changing the list (II):

Adjoining copies of the average: 6, 9, 15, 10, 10, 10, 10

$$\mu = \frac{6+9+15+4(10)}{7} = 10$$

$$\sigma = \sqrt{\frac{(6-10)^2+(9-10)^2+(15-10)^2+4(10-10)^2}{7}}$$

$$= \sqrt{\frac{3}{7}} \cdot \sqrt{14} < \sqrt{14}$$

Combining two lists:

A: 25, 35 B: 40, 40, 50, 50

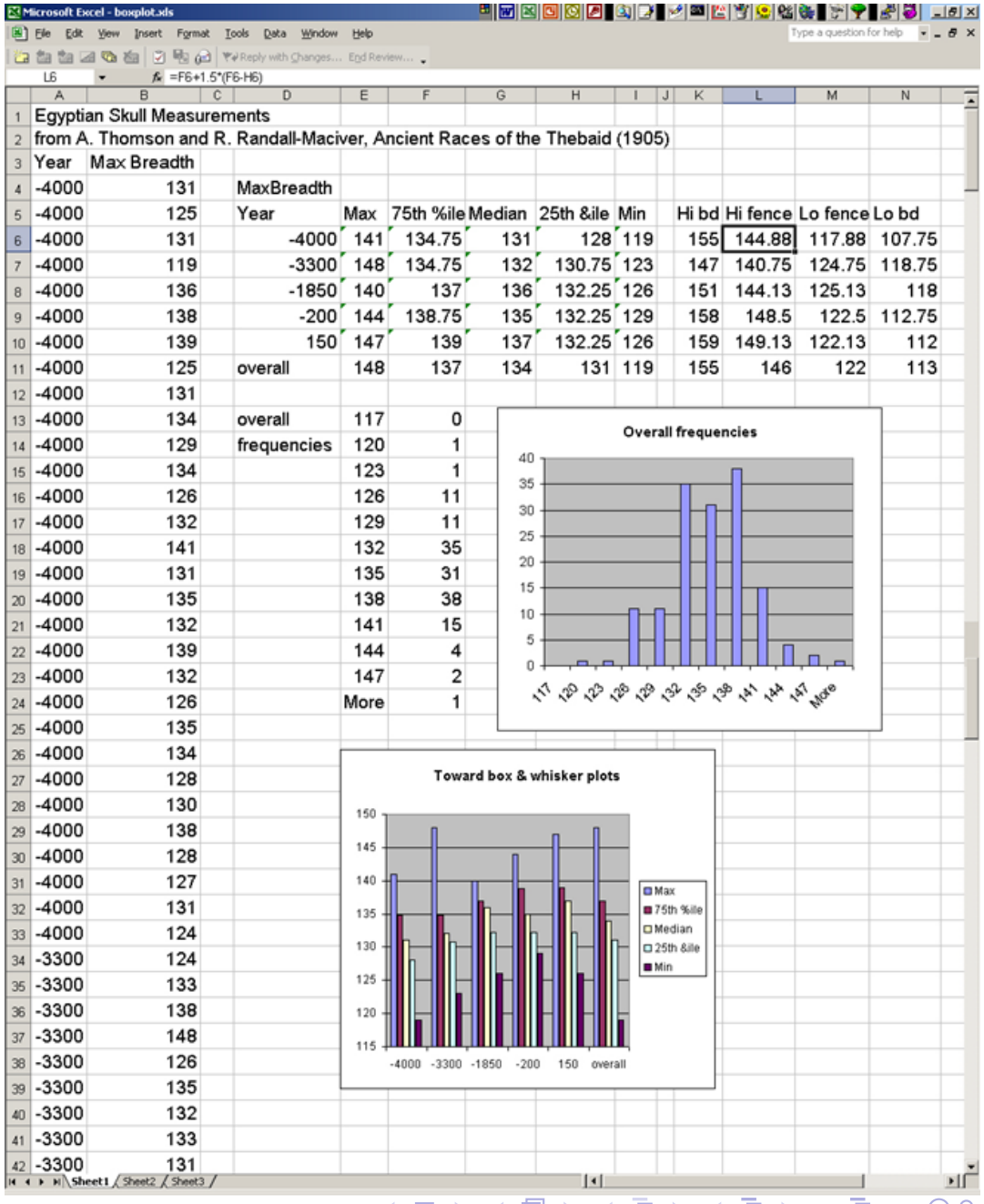
$$\mu_A = 30, \quad \sigma_A = 5, \quad \mu_B = 45, \quad \sigma_B = 5$$

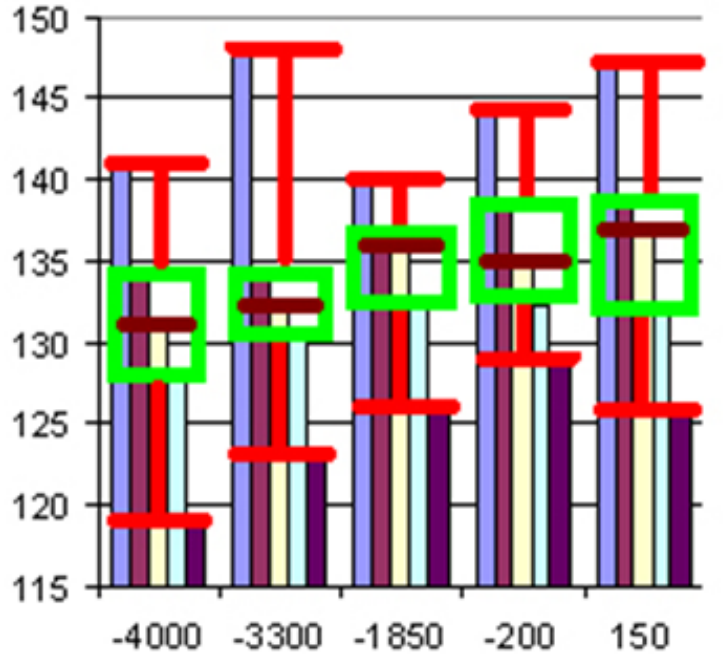
A and B: 25, 35, 40, 40, 50, 50

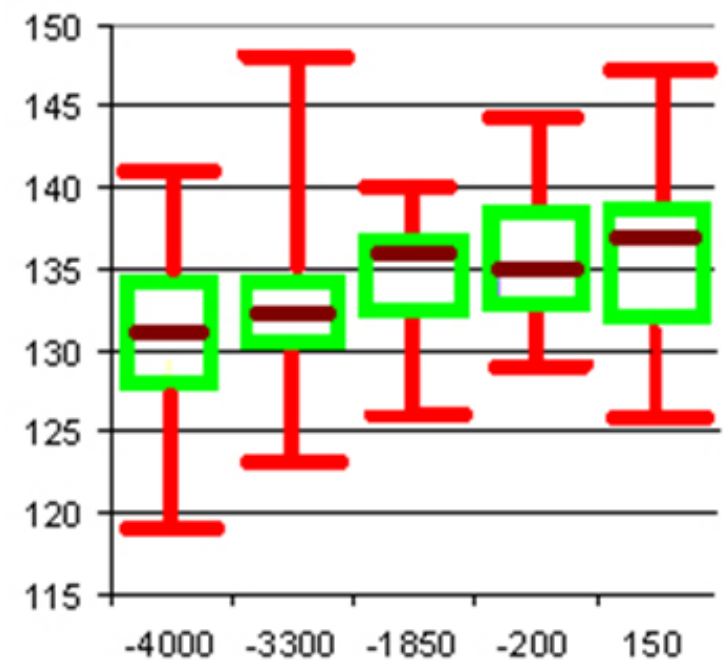
$$\mu = 40, \quad \sigma = 5\sqrt{3} > 5$$

Box-&-whisker plots

Include min, first quartile, median, third quartile, max — "normal" box-&-whisker plot just shows them all.

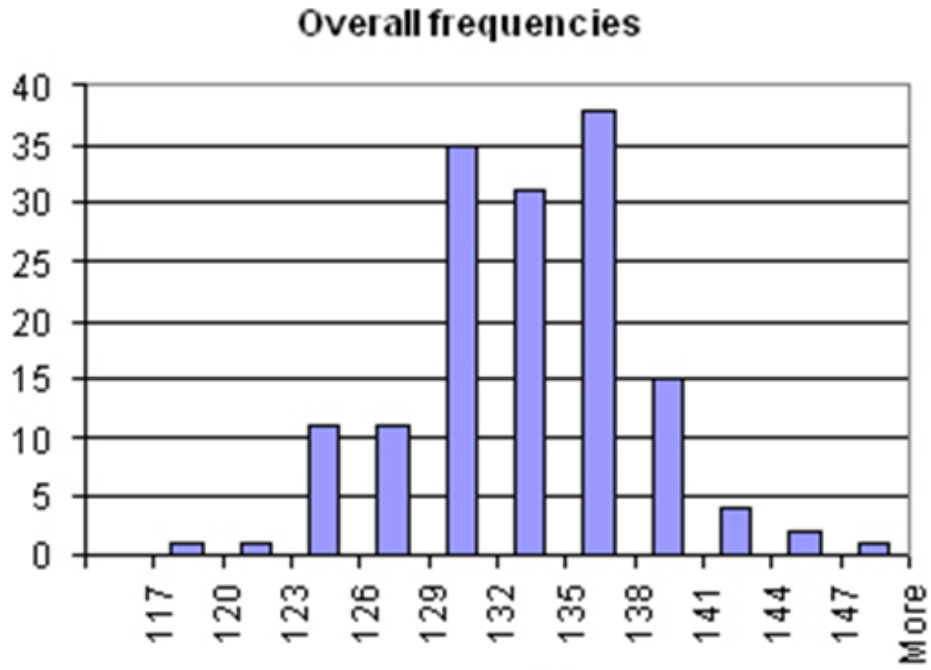



150
145
140
135
130
125
120
115
-4000
-3300
-1850
-200
150

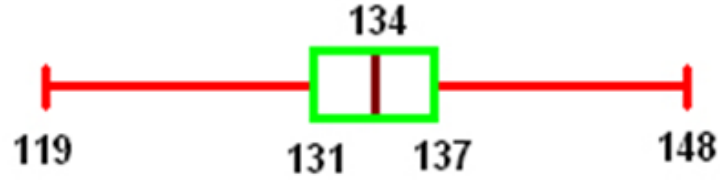

150
145
140
135
130
125
120
115
-4000
-3300
-1850
-200
150

"Modified" plot puts limit on whisker length: 1.5 IQR.

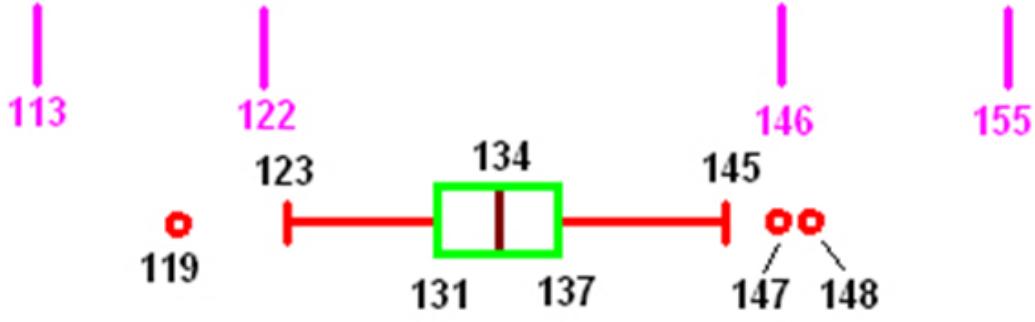
- 3rd quartile $+1.5$ IQR $=$ "(inner) fence"
- whisker ends at last value before or on fence
- datum beyond the fence is an "outlier" [which we reject only "for cause" (?)]
- beyond 3rd quartile $+$ 3 IQR ("outer fence" or "bound") is "extreme outlier"

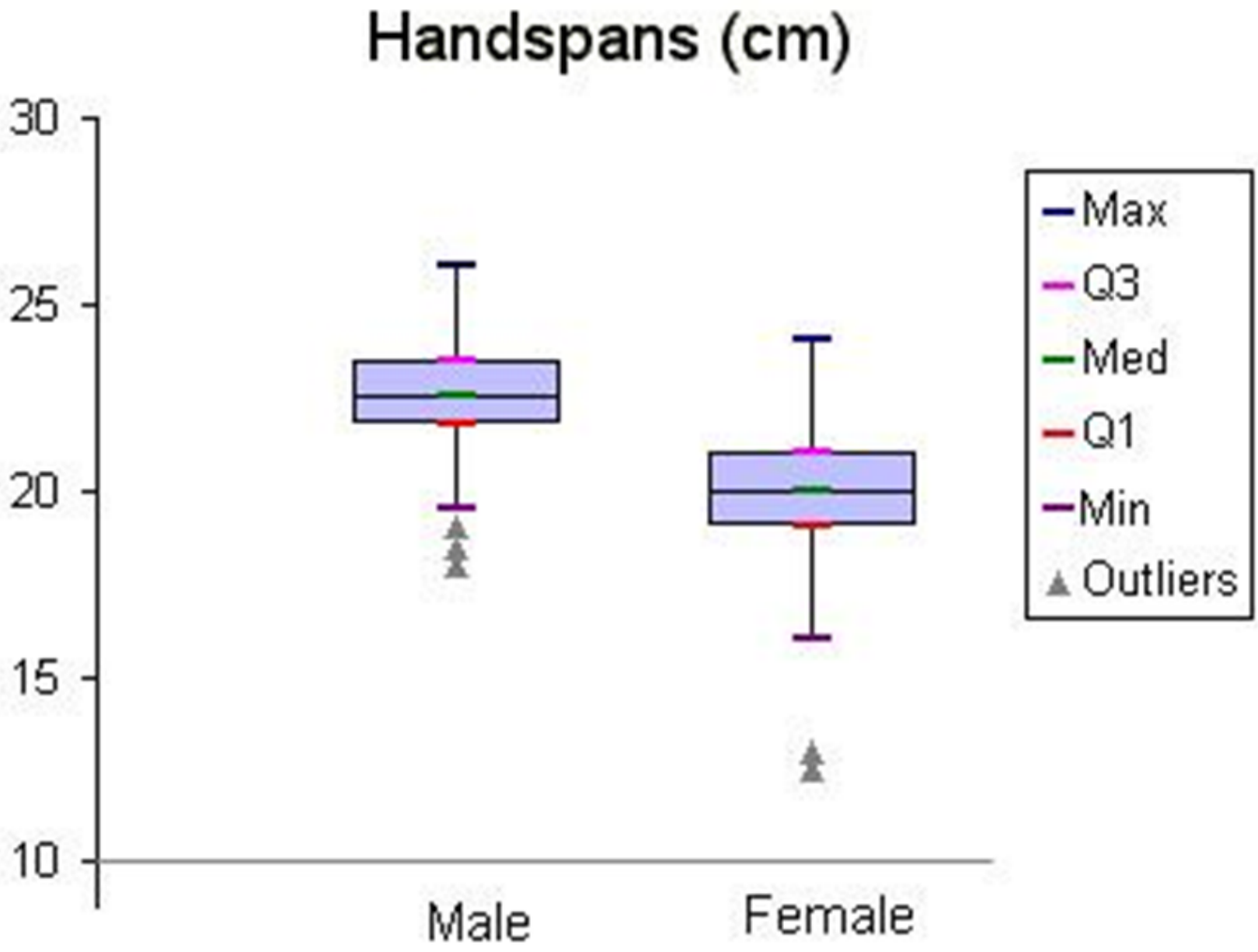


Normal



Modified




Handspans (cm)
30
25
20
15
10
Male
Female
Max
Q3
Med
Q1
Min
Outliers