



Quality Assurance & Quality Control





References

- Primary Reference
 - Michener and Brunt (2000) Ecological Data: Design, Management and Processing. Blackwell Science.
 - Edwards (2000), "Data Quality Assurance"
 - Brunt (2000) Ch. 2, "Data Management Principles, Implementation, and Administration"
 - Michener (2000) Ch. 7: "Transforming Data into Information and Knowledge"





Outline:

- Define QA/QC
- QC procedures
 - Designing data sheets
 - Data entry using validation rules, filters, lookup tables
- QA procedures
 - Graphics and Statistics
 - Outlier detection
 - Samples
 - Simple linear regression
- Archiving data





QA/QC

- “mechanisms [that] are designed to prevent the introduction of errors into a data set, a process known as data contamination”

Brunt 2000





Errors (2 types)

- **Commission:** Incorrect or inaccurate data are entered into a dataset
 - Can be easy to find
 - Malfunctioning instrumentation
 - Sensor drift
 - Low batteries
 - Damage
 - Animal mischief
 - Data entry errors
- **Omission:** Data or metadata are not recorded
 - Difficult or impossible to find
 - Inadequate documentation of data values, sampling methods, anomalies in field, human errors





Quality Control

- “mechanisms that are applied in advance, with *a priori* knowledge to ‘control’ data quality during the data acquisition process”

Brunt 2000





Quality Assurance

- “mechanisms [that] can be applied after the data have been collected and entered in a computer to identify errors of omission and commission”
 - graphics
 - statistics

Brunt 2000





QA/QC Activities

- Defining and enforcing standards for formats, codes, measurement units and metadata.
- Checking for unusual or unreasonable patterns in data.
- Checking for comparability of values between data sets.

Brunt 2000





Outline:

- Define QA/QC
- QC procedures
 - Designing data sheets
 - Data entry using validation rules, filters, lookup tables
- QA procedures
 - Graphics and Statistics
 - Outlier detection
 - Samples
 - Simple linear regression
- Archiving data





Flowering Plant Phenology – Data Collection Form Design

- Three sites, each with 3 transects
- On each transect, every species will have its phenological class recorded

Deep Well

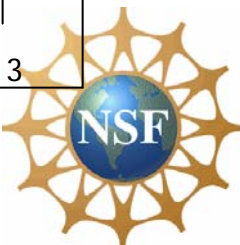
1	2	3

Five Points

1	2	3

Goat Draw

1	2	3



Data Collection Form Development:

What's wrong with this data sheet?

Plant

Life Stage

PHENOLOGY DATA SHEET

Collectors: _____

Date: _____ Time: _____

Location: deep well, five points, goat draw

Transect: 1 2 3

Notes: _____

Plant

Life Stage

<u>ardi</u>	P/G	V	B	FL	FR	M	S	D	NP
<u>arpu</u>	P/G	V	B	FL	FR	M	S	D	NP
<u>atca</u>	P/G	V	B	FL	FR	M	S	D	NP
<u>bamu</u>	P/G	V	B	FL	FR	M	S	D	NP
<u>zigr</u>	P/G	V	B	FL	FR	M	S	D	NP
_____	P/G	V	B	FL	FR	M	S	D	NP
_____	P/G	V	B	FL	FR	M	S	D	NP

P/G = perennating or germinating

V = vegetating

B = budding

FL = flowering

FR = fruiting

M = dispersing

S = senescing

D = dead

NP = not present



Outline:

- Define QA/QC
- QC procedures
 - Designing data sheets
 - Data entry using validation rules, filters and lookup tables
- QA procedures
 - Graphics and Statistics
 - Outlier detection
 - Samples
 - Simple linear regression
- Archiving data





Validation Rules:

- Control the values that a user can enter into a field





Validation Rule

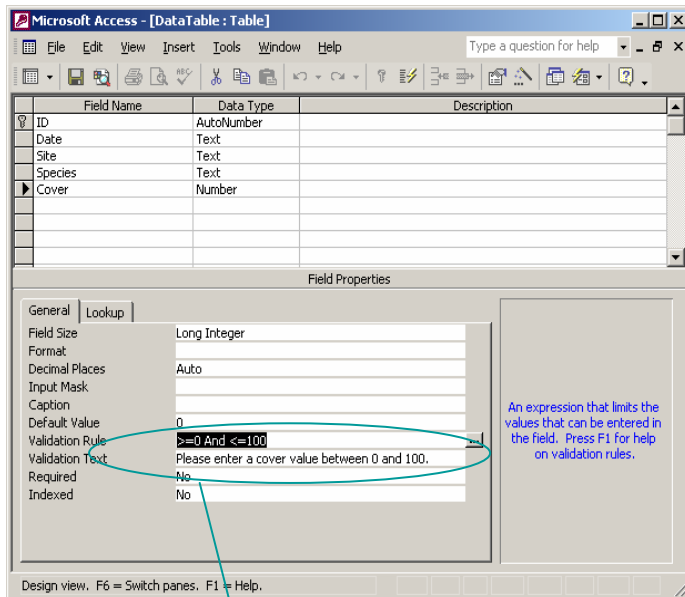
Examples in Access:

- ≥ 10
- Between 0 and 100
- Between #1/1/70# and Date()

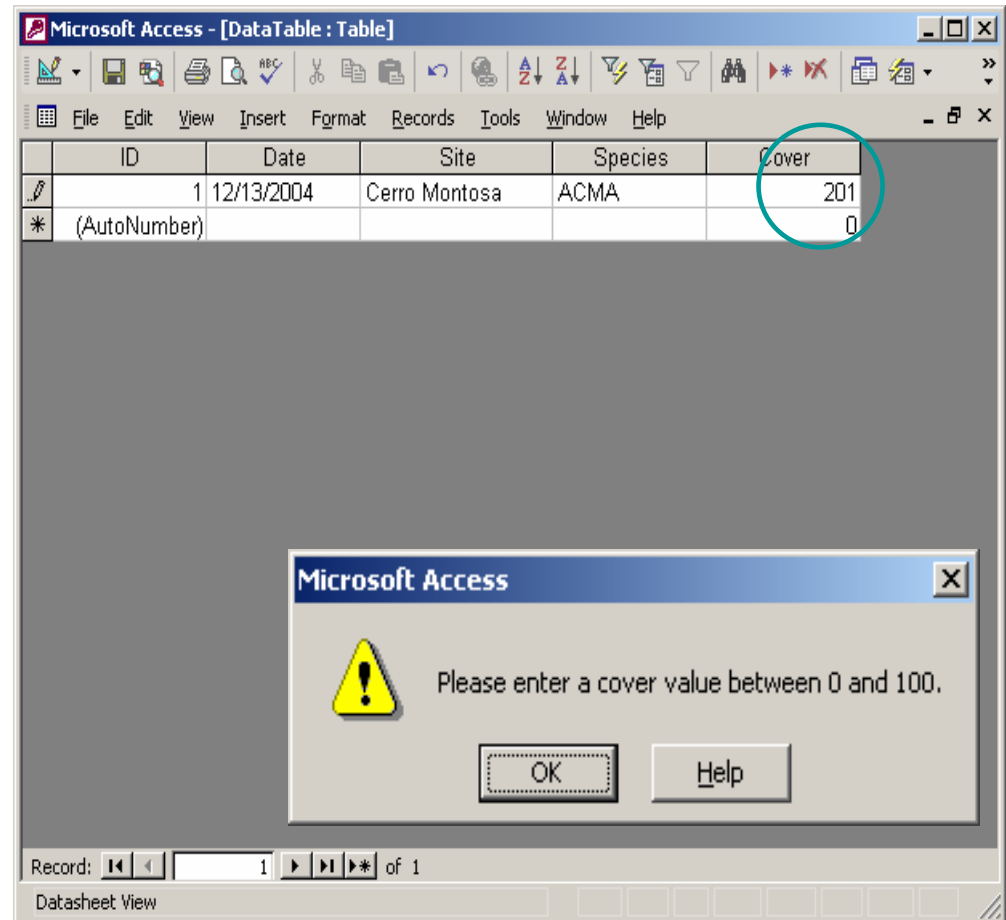




Validation rules in MS Access: Enter in Table Design View



≥ 0 and ≤ 100





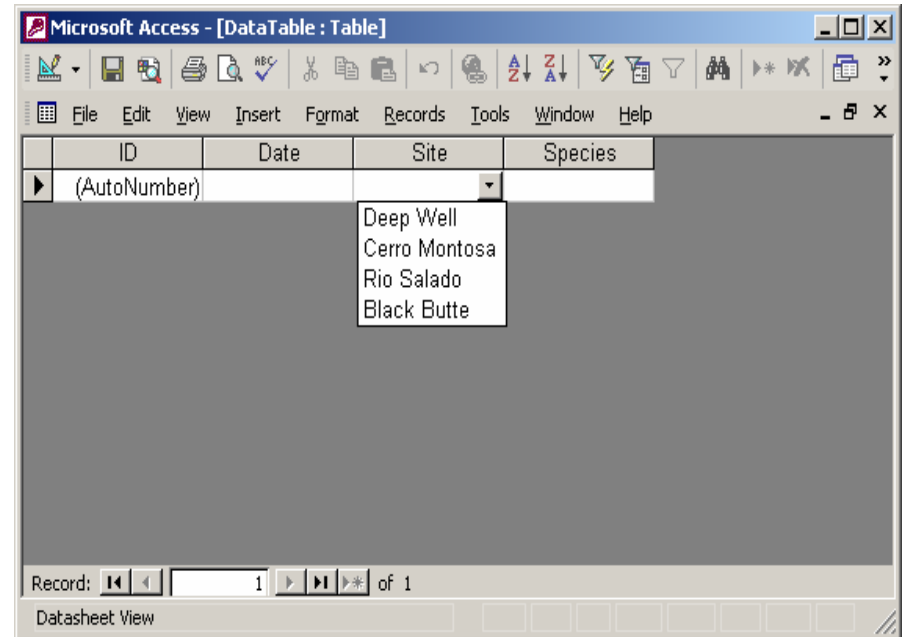
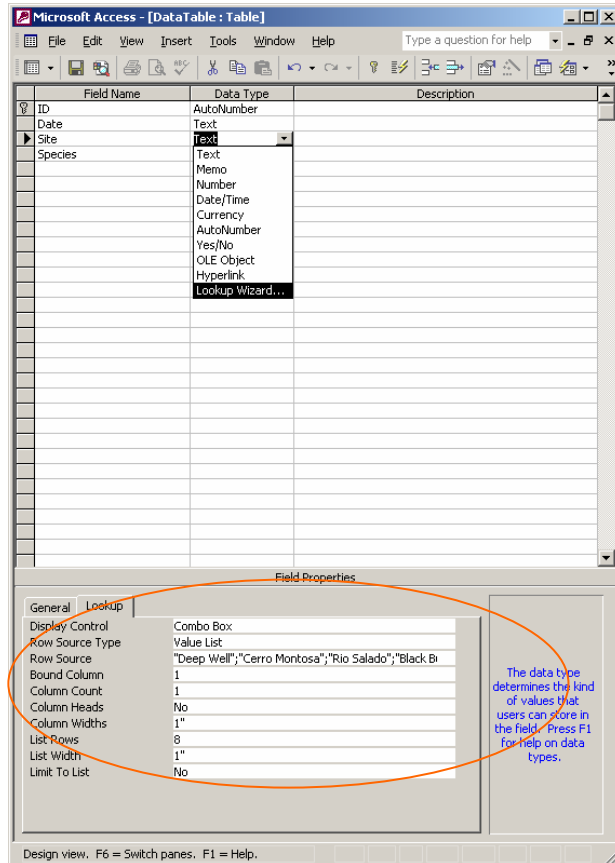
Look-up Fields

- Display a list of values from which entry can be selected





Look-up Tables in MS Access: Enter in Table Design View





Macros

- Validate data based on conditional statements





You want to make sure that a value for vegetation cover is entered in every record. To do this, create a macro called "NoData" that will examine the contents of the cover field whenever the field is exited.

Microsoft Access - [Vegetation_data]

File Edit View Insert Format Records Tools Window Help

MS Sans Serif 8 B I U

Observation_Nr

Date

Location

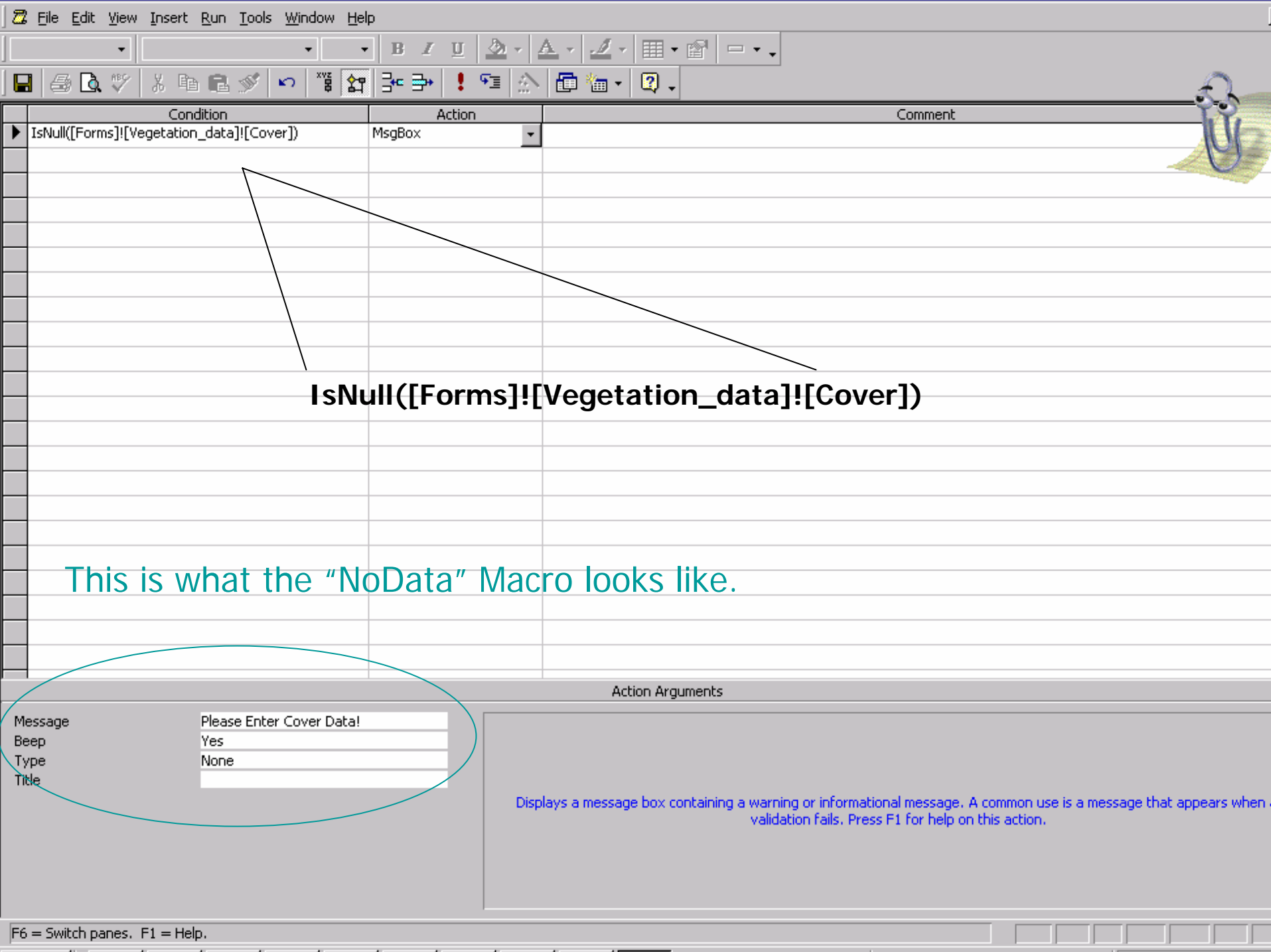
Plot

Plant_Species_Code

Cover

Observation_Number [AutoNumber]





IsNull([Forms]![Vegetation_data]![Cover])

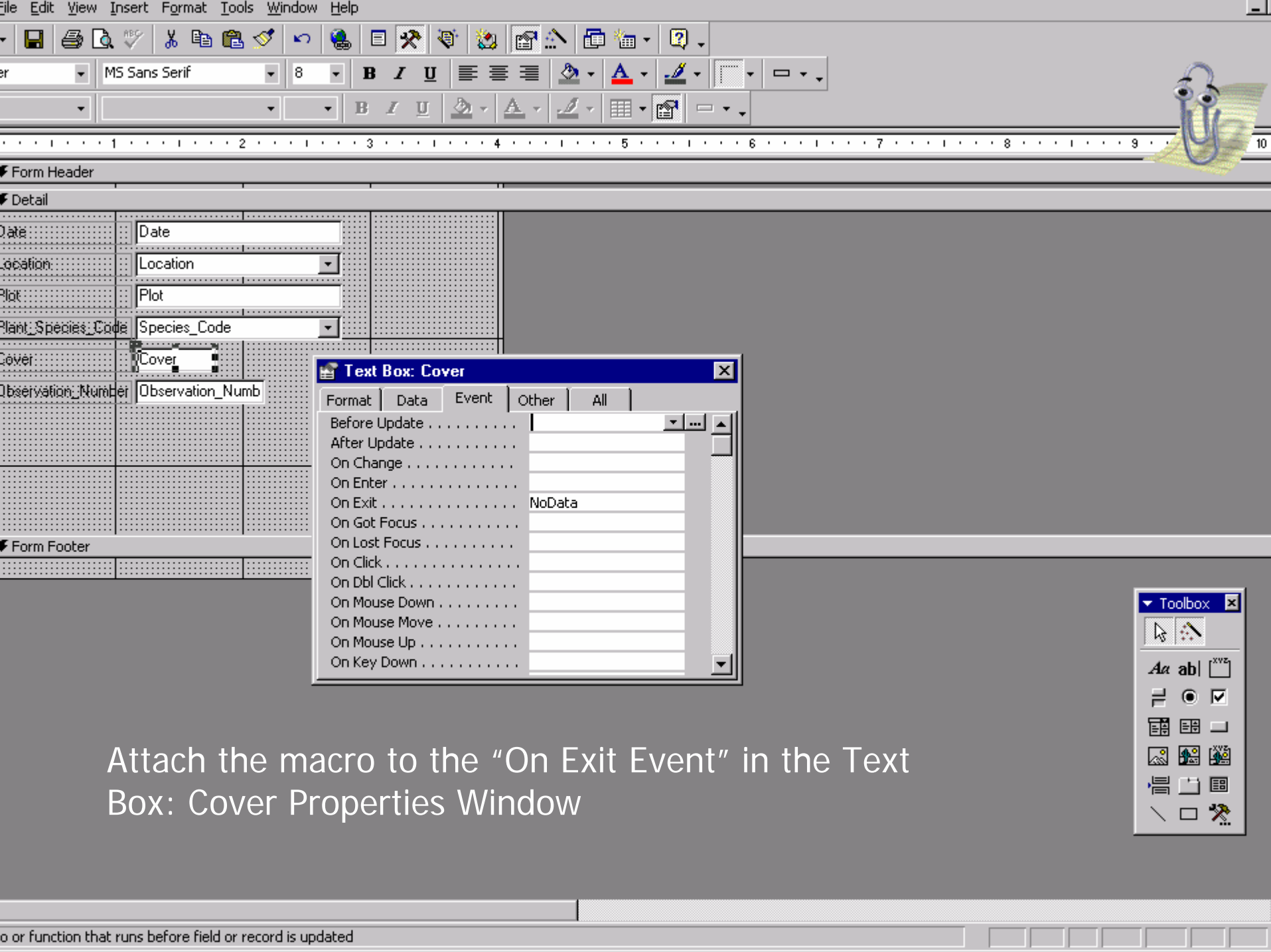
This is what the "NoData" Macro looks like.

Message
Beep
Type
Title

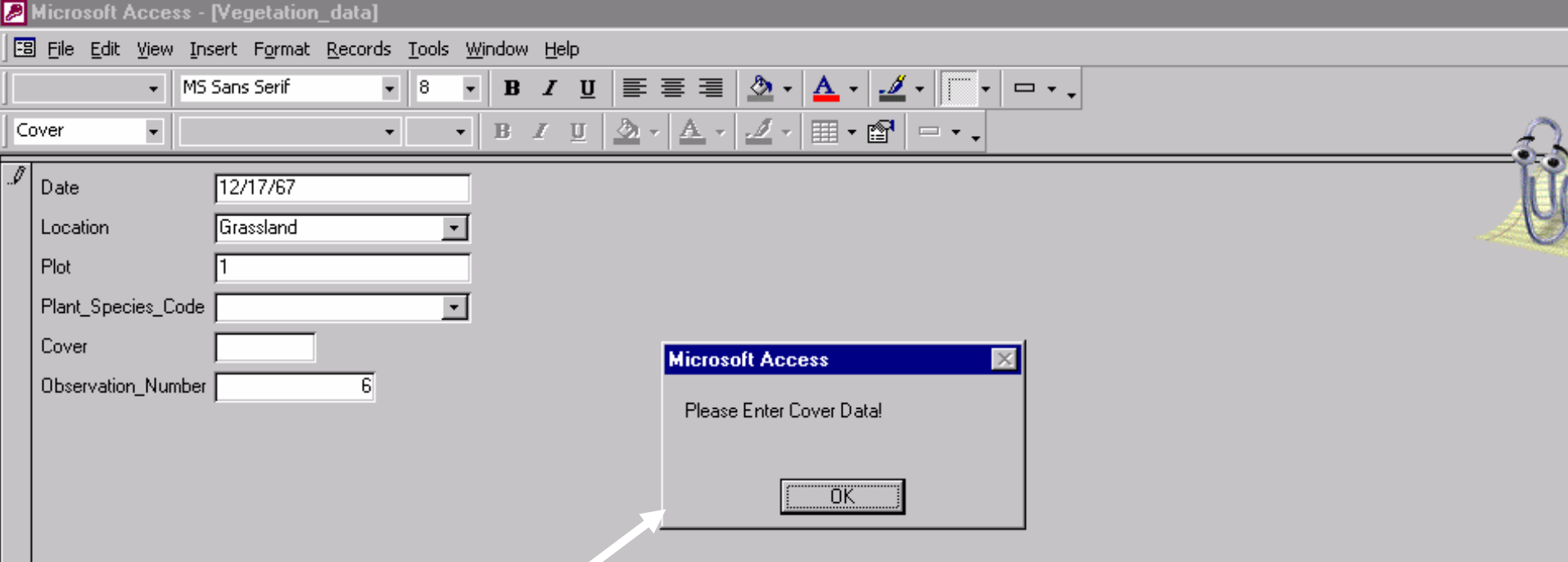
Please Enter Cover Data!
Yes
None

Action Arguments

Displays a message box containing a warning or informational message. A common use is a message that appears when validation fails. Press F1 for help on this action.



Attach the macro to the “On Exit Event” in the Text Box: Cover Properties Window



When the user tabs out of the "cover" field without entering data, this message box flashes to the screen.



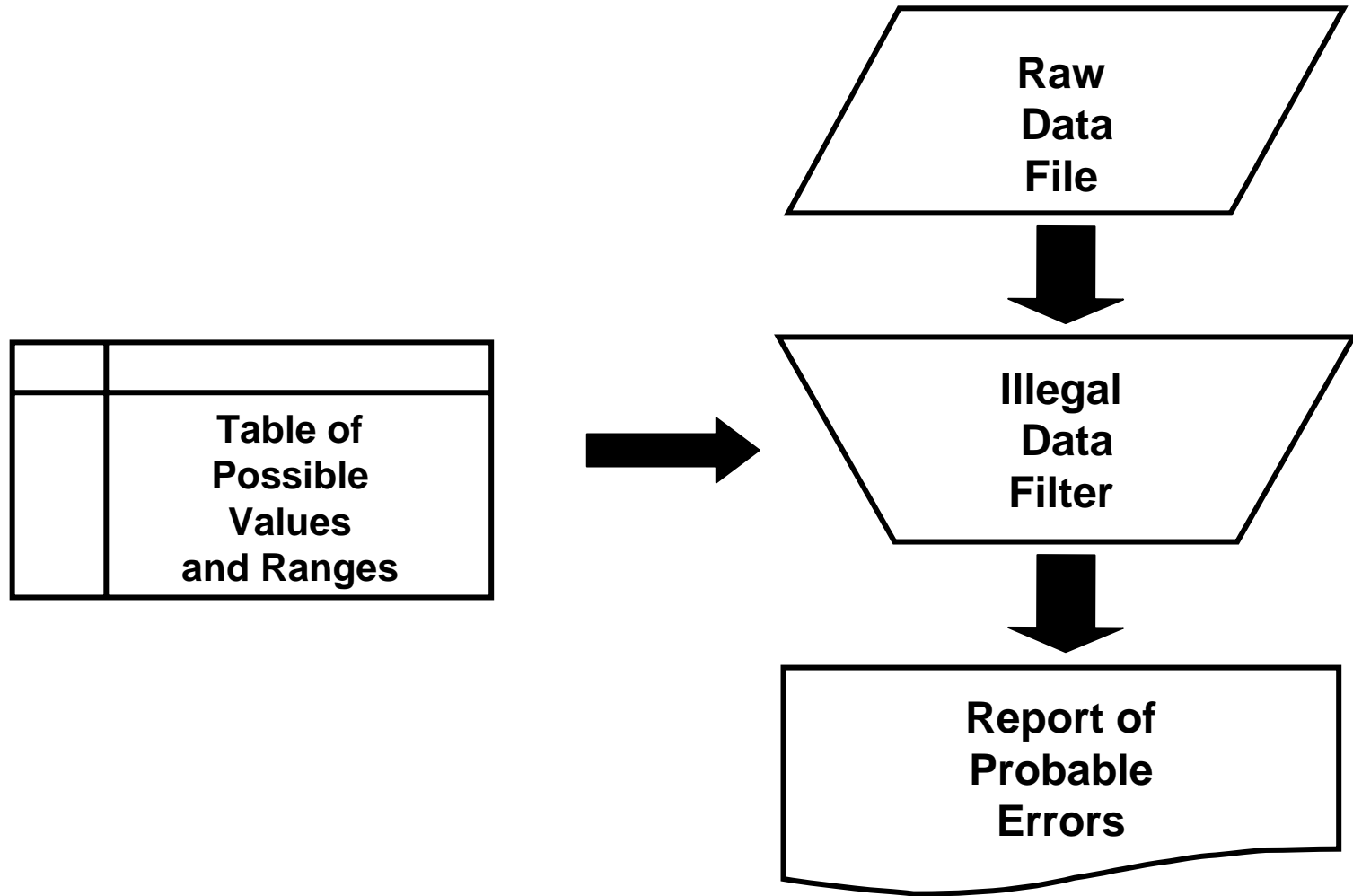
Other methods for preventing data contamination

- Double-keying of data by independent data entry technicians followed by computer verification for agreement
- Use text-to-speech program to read data back
- Filters for “illegal” data
 - Statistical/database/spreadsheet programs
 - Legal range of values
 - Sanity checks

Edwards 2000

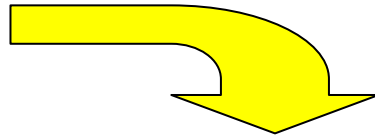


Flow of Information in Filtering “Illegal” Data



Tree Growth Data

Tree_ID	Cover (%)	DBH_1998 (cm)	DBH_1999 (cm)
a	43	300	200
b	231	300	400
c	46	530	480
d	109	200	300



A filter written in SAS

```
Data Checkum; Set tree;
```

```
Message=repeat(" ",39);
```

```
If cover<0 or cover>100 then do; message="cover is not in the interval  
[0,100]"; output; end;
```

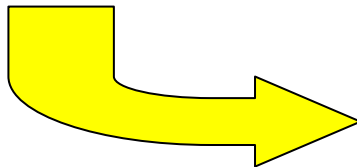
```
If dbh_1998>dbh_1999 then do; message="dbh_1998 is larger than  
dbh_1999"; output; end;
```

```
If message NE repeat(" ", 39);
```

```
Keep ID message;
```

```
Proc Print Data=Checkum;
```

Error report



Obs	ID	message
1	a	dbh_1998 is larger than dbh_1999
2	b	cover is not in the interval [0,100]
3	c	dbh_1998 is larger than dbh_1999
4	d	cover is not in the interval [0,100]



Spreadsheet column statistics: *Peromyscus truei* example

<u>id#</u>	<u>mass</u>	<u>body</u>	<u>tail</u>	<u>foot</u>
1	22.5	81.0	85.6	23.0
2	25.0	81.0	90.0	23.0
3	21.0	76.0	80.0	22.0
4	26.0	86.0	90.2	22.0
5	23.6	101.0	92.0	22.5
6	22.0	86.0	87.0	23.0
7	23.0	94.0	83.0	23.0
8	23.0	89.0	82.0	22.0
9	30.0	9.9	91.0	23.0
10	28.1	97.0	100.0	23.0
11	22.0	96.0	81.0	24.0
12	25.0	98.0	91.0	23.0
13	29.0	97.0	102.0	22.0
14	51.0	75.0	90.0	21.5
15	23.0	81.3	89.0	21.0
16	29.0	85.0	88.0	23.0
17	33.0	89.0	84.3	21.0
18	67.0	90.0	85.0	22.0
19	21.0	86.2	87.0	24.0
20	25.0	78.0	91.7	22.5
21	32.0	80.0	90.0	22.0

mean	28.6	83.6	88.6	22.5
median	25.0	86.0	89.0	22.5
standard deviation	11.0	18.5	5.5	0.8
variance	120.7	343.9	30.2	0.7
minimum	21.0	9.9	80.0	21.0
maximum	67.0	101.0	102.0	24.0





Spreadsheet range checks

id#	mass	body	tail	foot	rangecheck	mass
1	22.5	81.0	85.6	23.00	0	=if(mass>50,1,0)
2	25.0	81.0	90.0	23.00	0	
3	21.0	76.0	80.0	22.00	0	
4	26.0	86.0	90.2	22.00	0	
5	23.6	101.0	92.0	22.50	0	
6	22.0	86.0	87.0	23.00	0	
7	23.0	94.0	83.0	23.00	0	
8	23.0	89.0	82.0	22.00	0	
9	30.0	9.9	91.0	23.50	0	
10	28.1	97.0	100.0	23.00	0	
11	22.0	96.0	81.0	24.00	0	
12	25.0	98.0	91.0	23.00	0	
13	29.0	97.0	102.0	22.00	0	
14	51.0	75.0	90.0	21.50	1	
15	23.0	81.3	89.0	21.00	0	
16	29.0	85.0	88.0	23.00	0	
17	33.0	89.0	84.3	21.00	0	
18	67.0	90.0	85.0	22.00	1	
19	21.0	86.2	87.0	24.00	0	
20	25.0	78.0	91.7	22.50	0	
21	32.0	80.0	90.0	22.00	0	

mean	28.6	83.6	88.6	22.52
median	25.0	86.0	89.0	22.50
standard deviation	11.0	18.5	5.5	0.84
variance	120.7	343.8	30.2	0.71
minimum	21.0	9.9	80.0	21.00
maximum	67.0	101.0	102.0	24.00





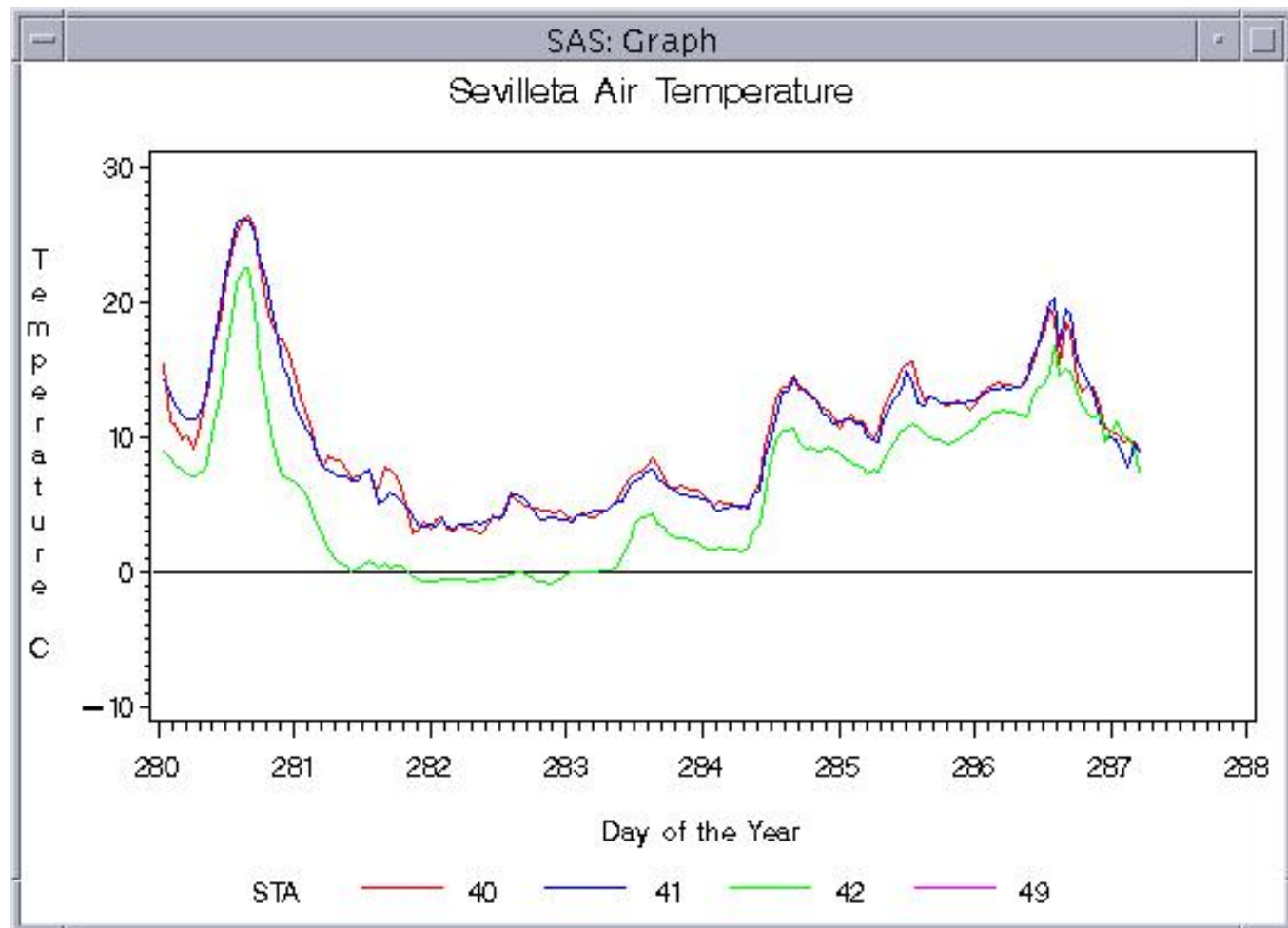
Outline:

- Define QA/QC
- QC procedures
 - Designing data sheets
 - Data entry using validation rules, filters, lookup tables
- QA procedures
 - Graphics and Statistics to find:
 - Unusual patterns
 - Outliers
- Archiving data



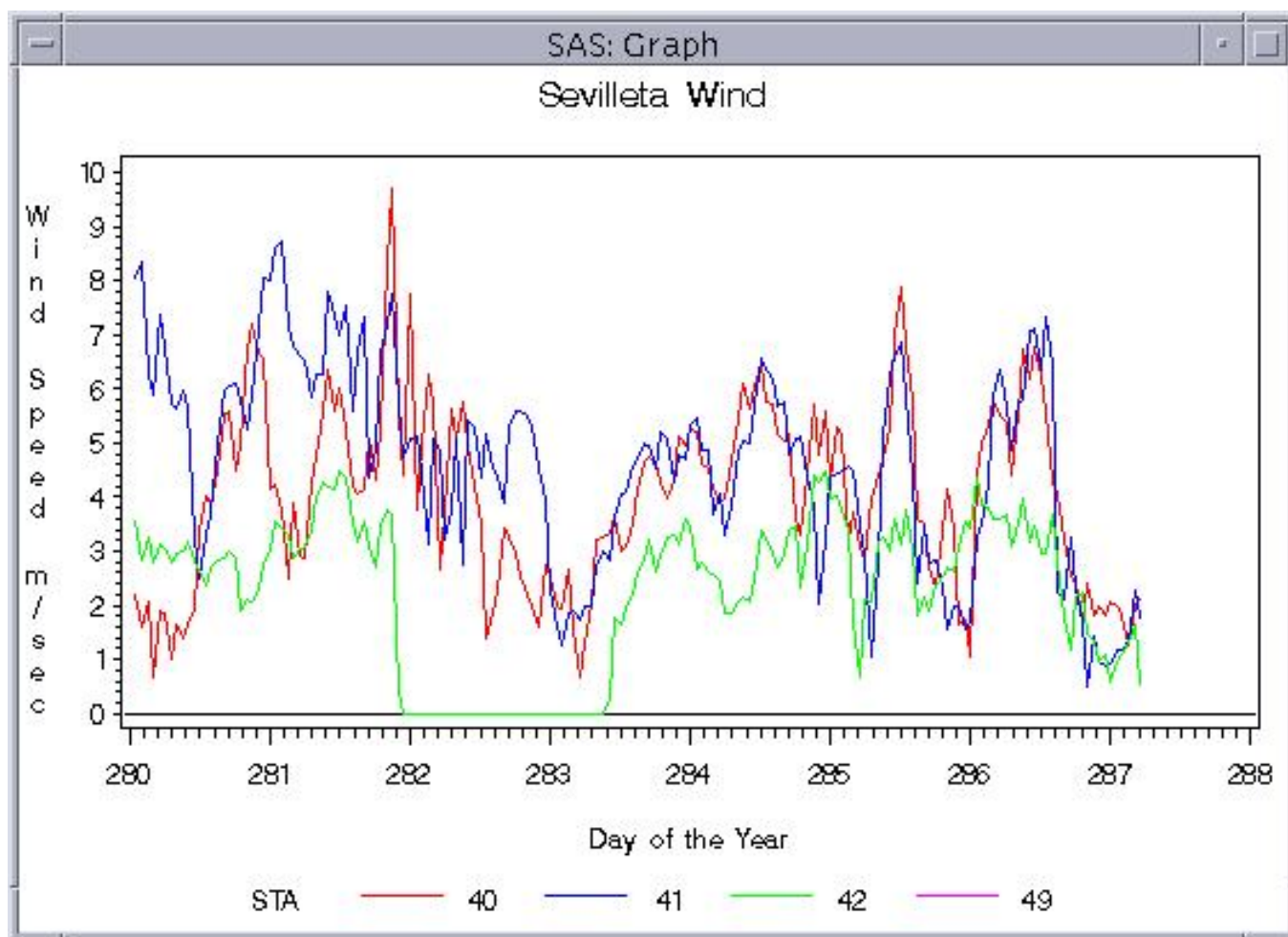


Identifying Sensor Errors: Comparison of data from three Met stations, Sevilleta LTER





Identification of Sensor Errors: Comparison of data from three Met stations, Seville LTER





Metadata for 'bad' data

Variable 9* Name: Average Wind Speed

- * Label: Avg_Windspeed

- * Definition: Average wind speed during the hour at 3 m

- * Units of Measure: meters/second

- * Precision of Measurements: .11 m/s

- * Range or List of Values: 0-50

- * Data Type: Real

- * Column Format: #.###

- * Field Position: Columns 51-58

- * Missing Data Code: -999 (bad) -888 (not measured)

- * Computational Method for Derived Data: na



Flagging Data Values

Microsoft Excel - WNDCE01

File Edit View Insert Format Tools Data Window Help Adobe PDF

Type a question for help

10

L3

	A	B	C	D	E	F	G	H	I	J	K
							fw1	windmag	fw3	winddir	fw4
								0.3		179.9	
								0.3		294.7	
								0.8		189.3	
								0.9		191.7	
								0.4		186.9	
								0.7		137	
								0.6		162.7	
								0.4		165.9	
								0.4		180.2	
								0.4		199	
								0.3		161.8	
								0.1		40.9	
								0.3		174.9	
								0.1		64.3	
						M			M		M
						M			M		M
						M			M		M
						M			M		M
						M			M		M
						M			M		M
						M			M		M
								0.3		106.3	
								0.1		90.8	
								0.8		70.5	
								0.3		142.1	

Andrews Experimental Forest LTER: HJA data domain details MS001 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.fsl.orst.edu/lter/data/domains.cfm?domain=enul

Customize Links Free Hotmail Windows Marketplace Windows Media Windows

HJ Andrews Experimental Forest

Long Term Ecological Research




What's New Links Opportunities LTER5 Search

home page > Data > Data Catalog > Research Categories

MS001

Attribute: FW1

code	description
E	Good value
E	Estimated value
M	Sensor most likely frozen (values are questionable)
M	Missing
Q	Questionable
S	Daily value based on sunrise to sunrise

[Home](#)
[Top of Page](#)
[Web-site Map](#)
[Contacts](#)
[Addresses](#)
[Disclaimers](#)
[Credits](#)

Done





Outliers

- An outlier is “an unusually extreme value for a variable, given the statistical model in use”
- The goal of QA is NOT to eliminate outliers! Rather, we wish to detect unusually extreme values.

Edwards 2000





Outlier Detection

- “the detection of outliers is an intermediate step in the elimination of [data] contamination”
 - Attempt to determine if contamination is responsible and, if so, flag the contaminated value.
 - If not, formally analyse with and without “outlier(s)” and see if results differ. Or use robust statistical methods.





Methods for Detecting Outliers

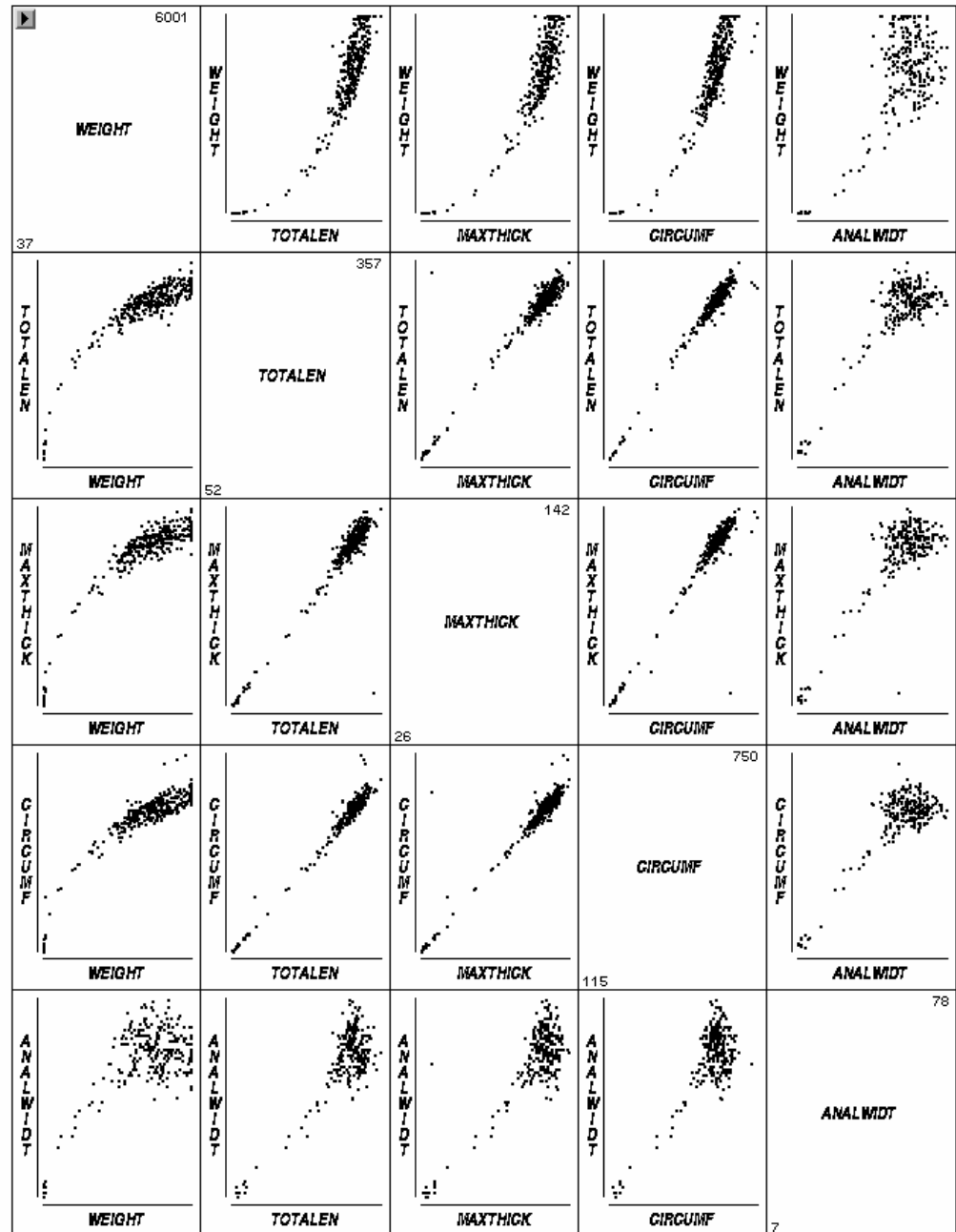
- Graphics
 - Scatter plots
 - Box plots
 - Histograms
 - Normal probability plots
- Formal statistical methods
 - Grubbs' test

Edwards 2000



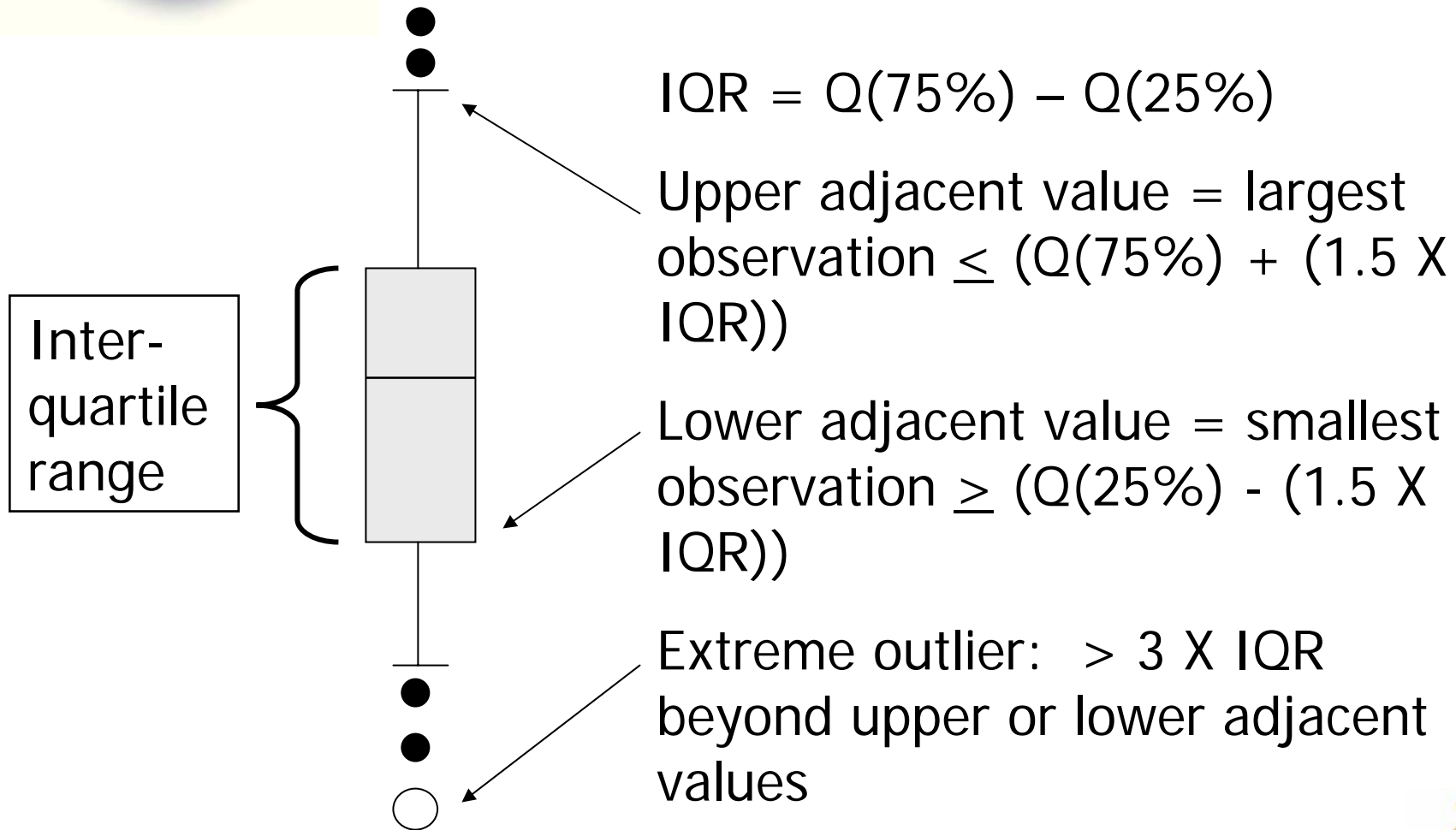
X-Y scatter plots of gopher tortoise morphometrics

Michener 2000

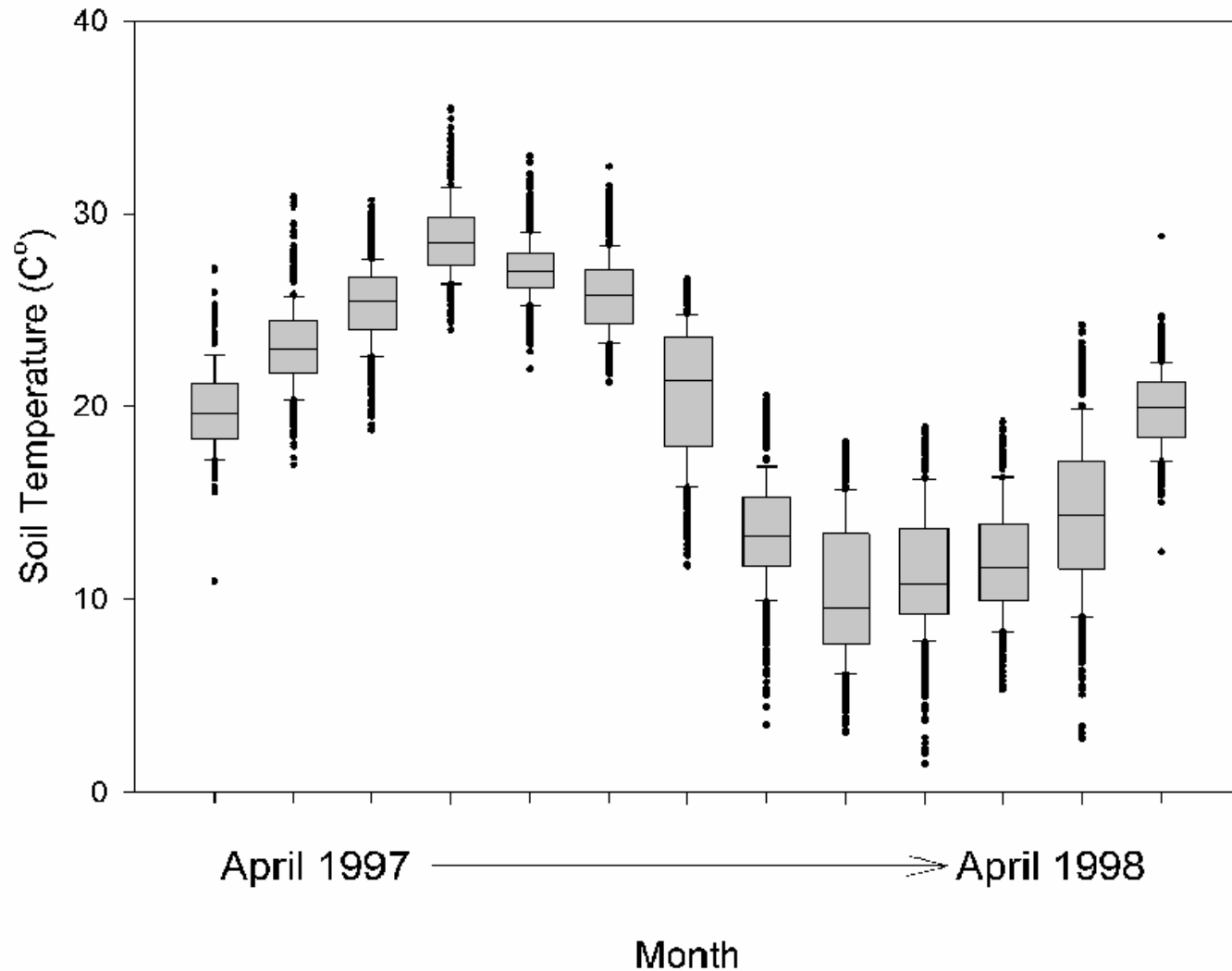




Box Plot Interpretation

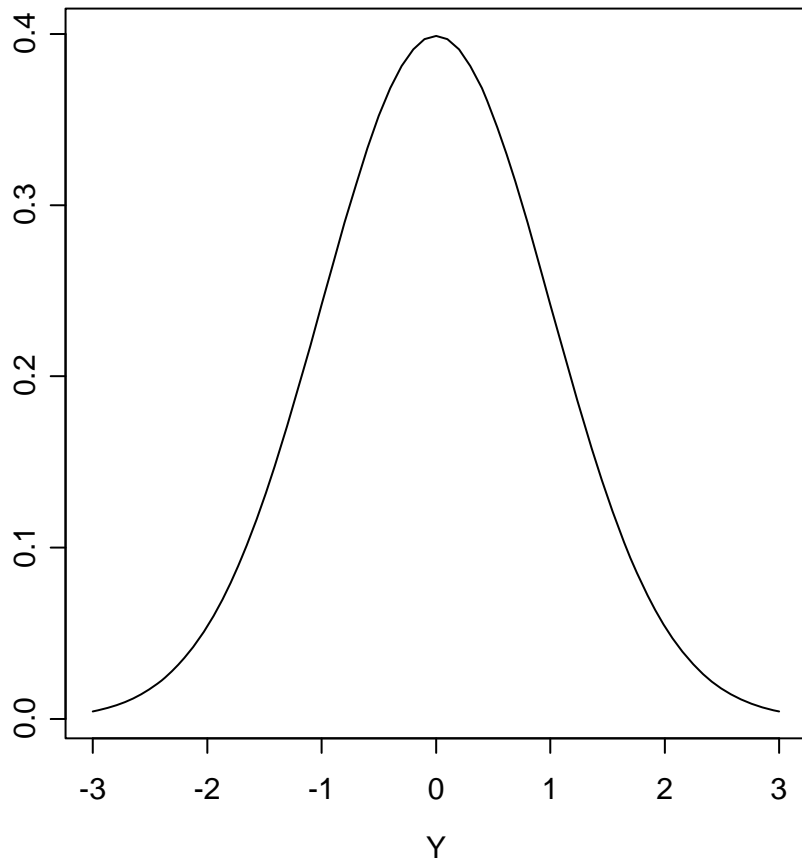


Box Plots Depicting Statistical Distribution of Soil Temperature

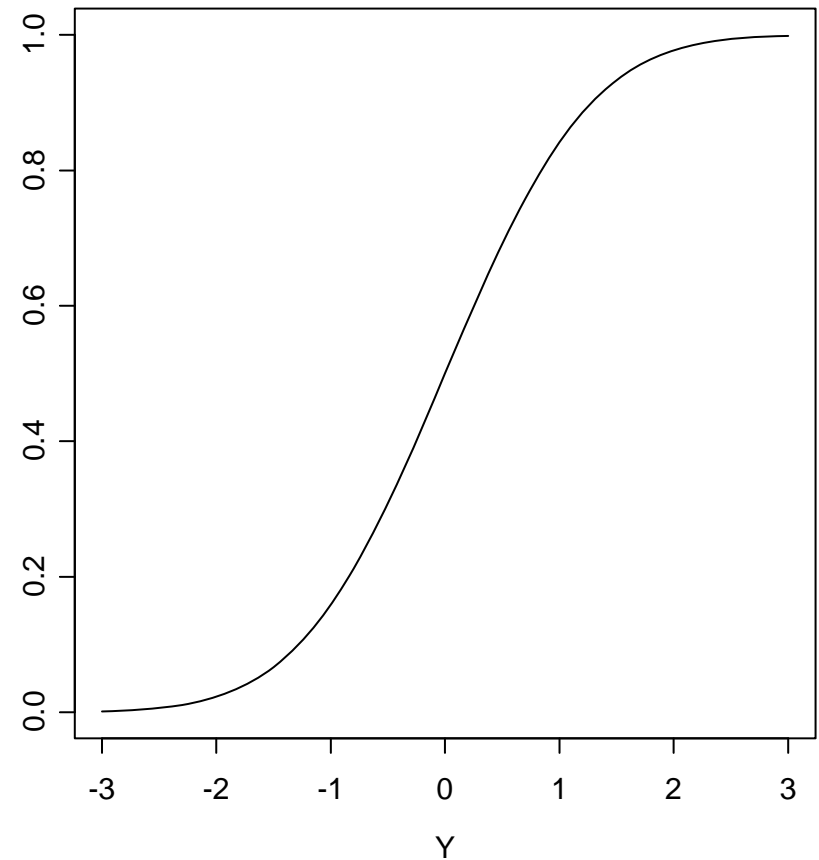


Normal density and Cumulative Distribution Functions

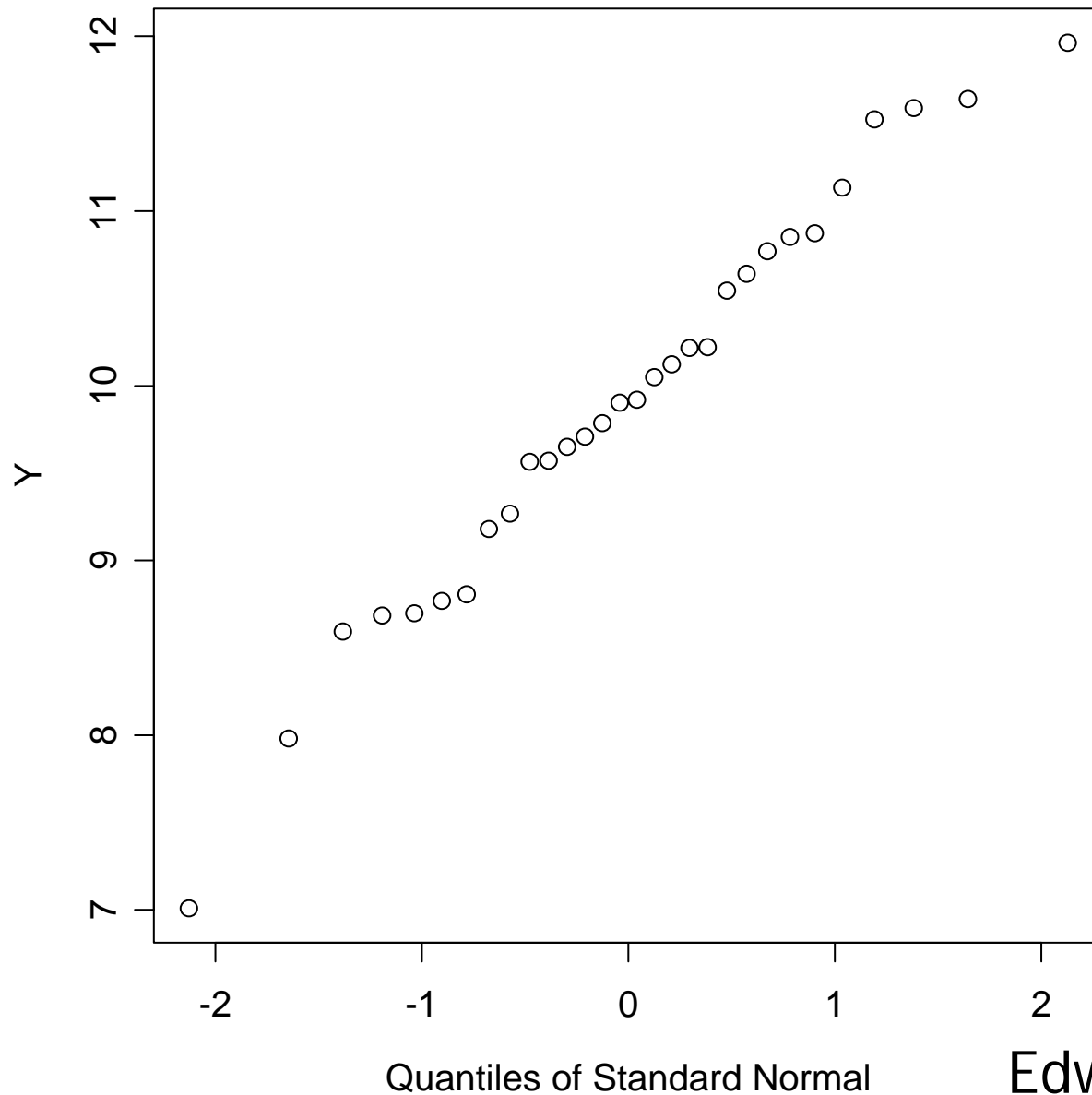
a. The Normal Frequency Curve



b. Cumulative Normal Curve Areas



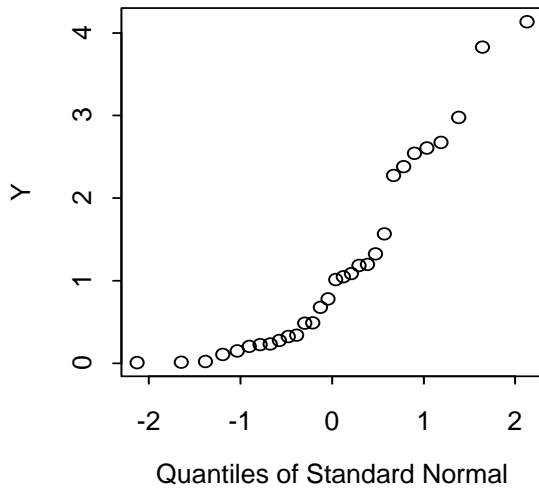
Normal Plot of 30 Observations from a Normal Distribution



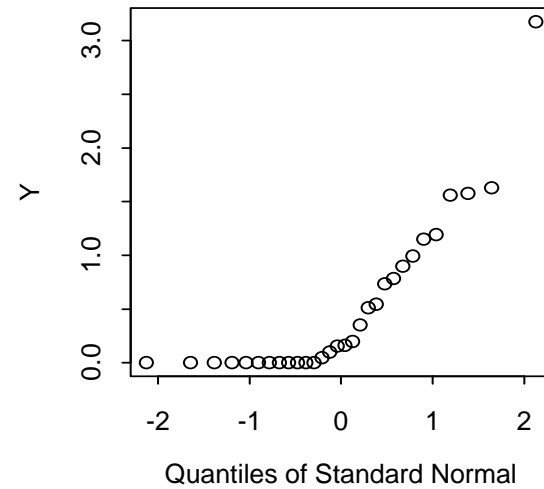
Edwards 2000

Normal Plots from Non-normally Distributed Data

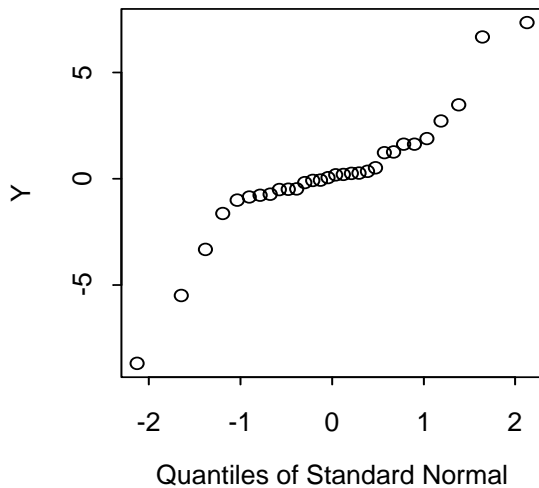
a. a right skewed distribution



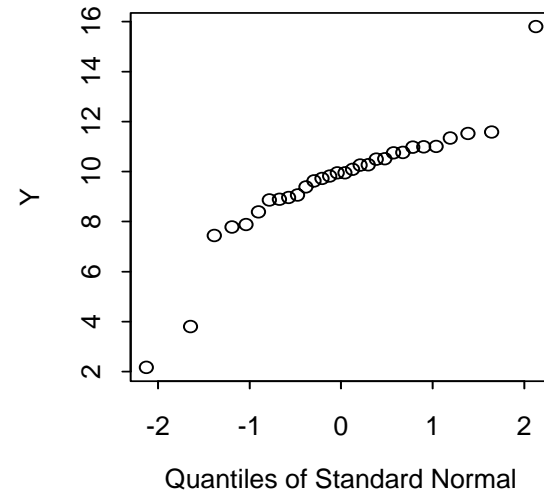
b. a left-truncated Normal distribution



c. a heavy-tailed distribution



d. a contaminated Normal distribution





Statistical tests for outliers assume that the data are normally distributed....

CHECK THIS ASSUMPTION!





Grubbs' test for outlier detection in a univariate data set:

$$T_n = (Y_n - Y_{\text{bar}})/S$$

where Y_n is the possible outlier,

Y_{bar} is the mean of the sample, and

S is the standard deviation of the sample

Contamination exists if T_n is greater than $T_{.01[n]}$

Grubbs, Frank (February 1969), Procedures for Detecting Outlying Observations in Samples, Technometrics, Vol. 11, No. 1, pp. 1-21.





Example of Grubbs' test for outliers rainfall in acre-feet from seeded clouds (Simpson et al. 1975)

4.1	7.7	17.5	31.4
32.7	40.6	92.4	115.3
118.3	119.0	129.6	198.6
200.7	242.5	255.0	274.7
274.7	302.8	334.1	430.0
489.1	703.4	978.0	
1656.0	1697.8	2745.6	

$T_{26} = 3.539 > 3.029$; Contaminated

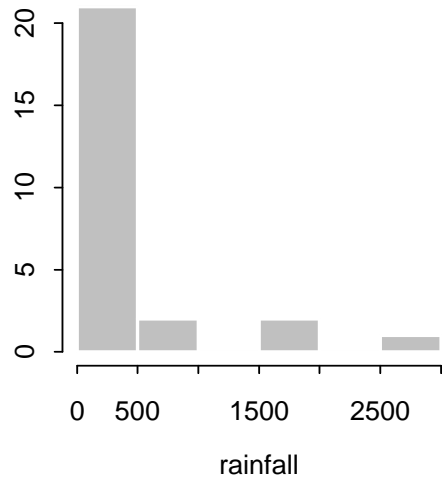
But Grubbs' test is sensitive
to non-normality.....

Edwards 2000



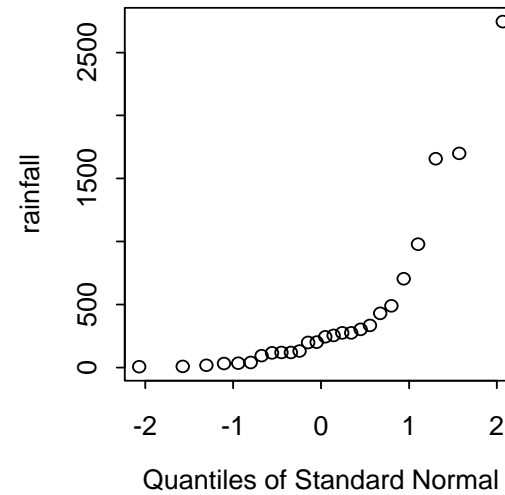
Checking Assumptions on Rainfall Data

a. rainfall

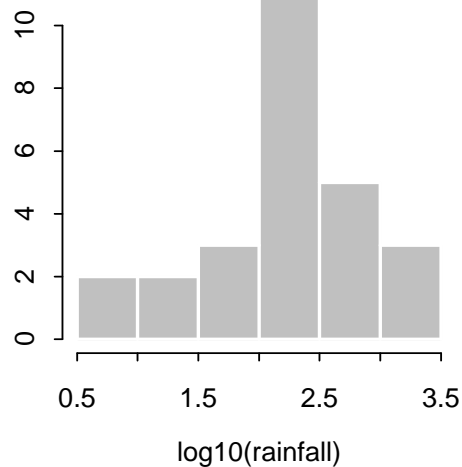


Skewed
distribution:
Grubbs' Test
detects
contaminating
points

b. Normal plot

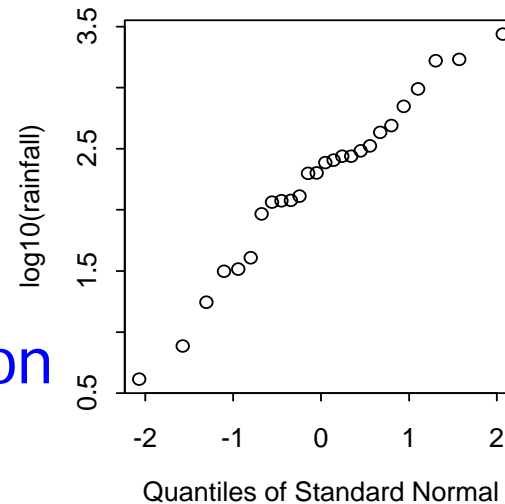


c. $\log_{10}(\text{rainfall})$



Normal
Distribution:
Grubbs' test
detects no
contamination

d. Normal plot





References about outliers:

- Barnett, V. and Lewis, T.: 1994, *Outliers in Statistical Data*, John Wiley & Sons, New York
- Iglewicz, B. and Hoaglin, D. C.: 1993 *How to Detect and Handle Outliers*, American Society for Quality Control, Milwaukee, WI.





Simple Linear Regression: check for model-based.....

- Outliers
- Influential (leverage) points





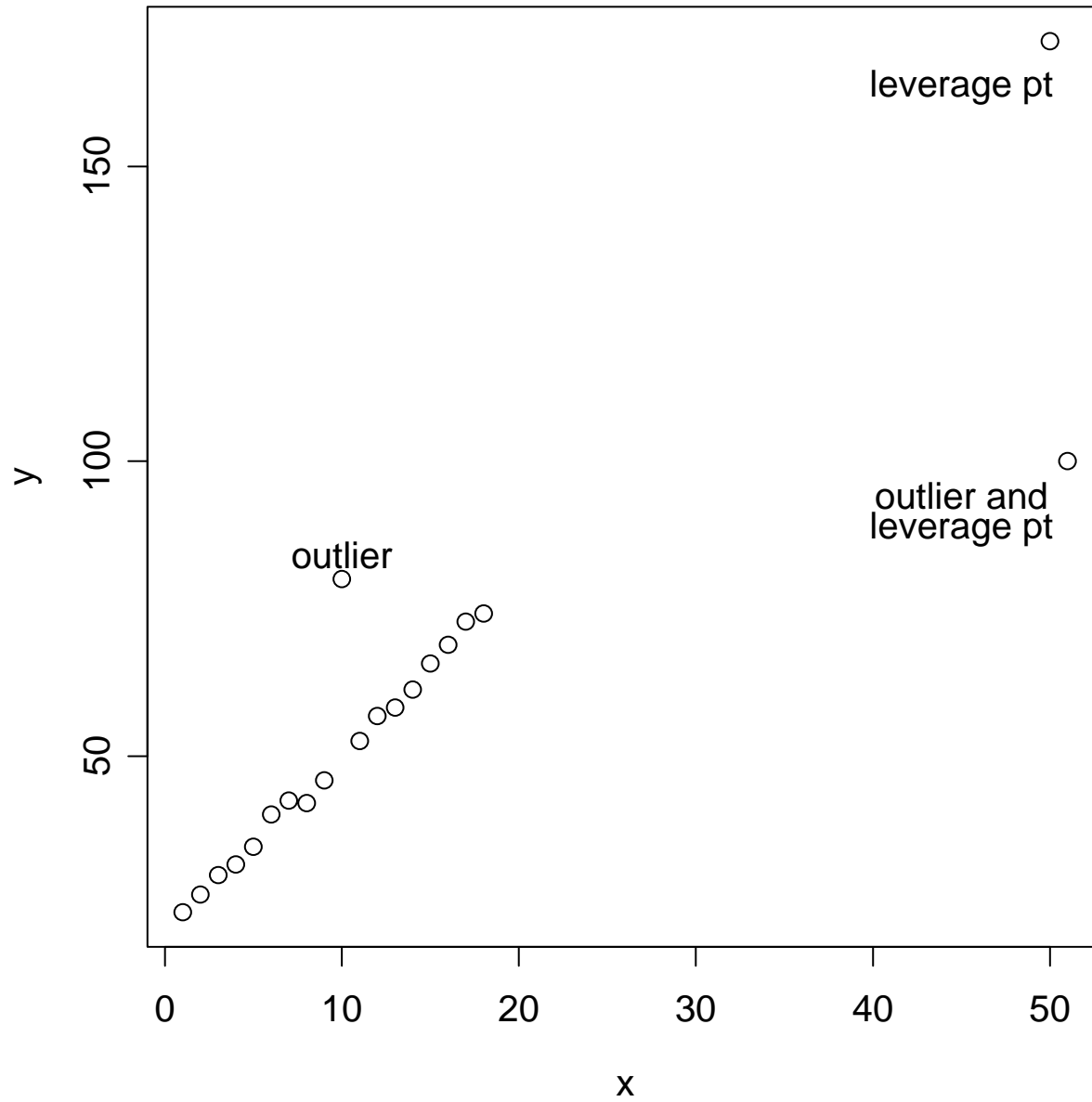
Influential points in simple linear regression:

- A “leverage” point is a point with an unusual regressor value that has more weight in determining regression coefficients than the other data values.
- An “outlier” is an observation with a response value that does not fit the X-Y pattern found in the rest of the data.

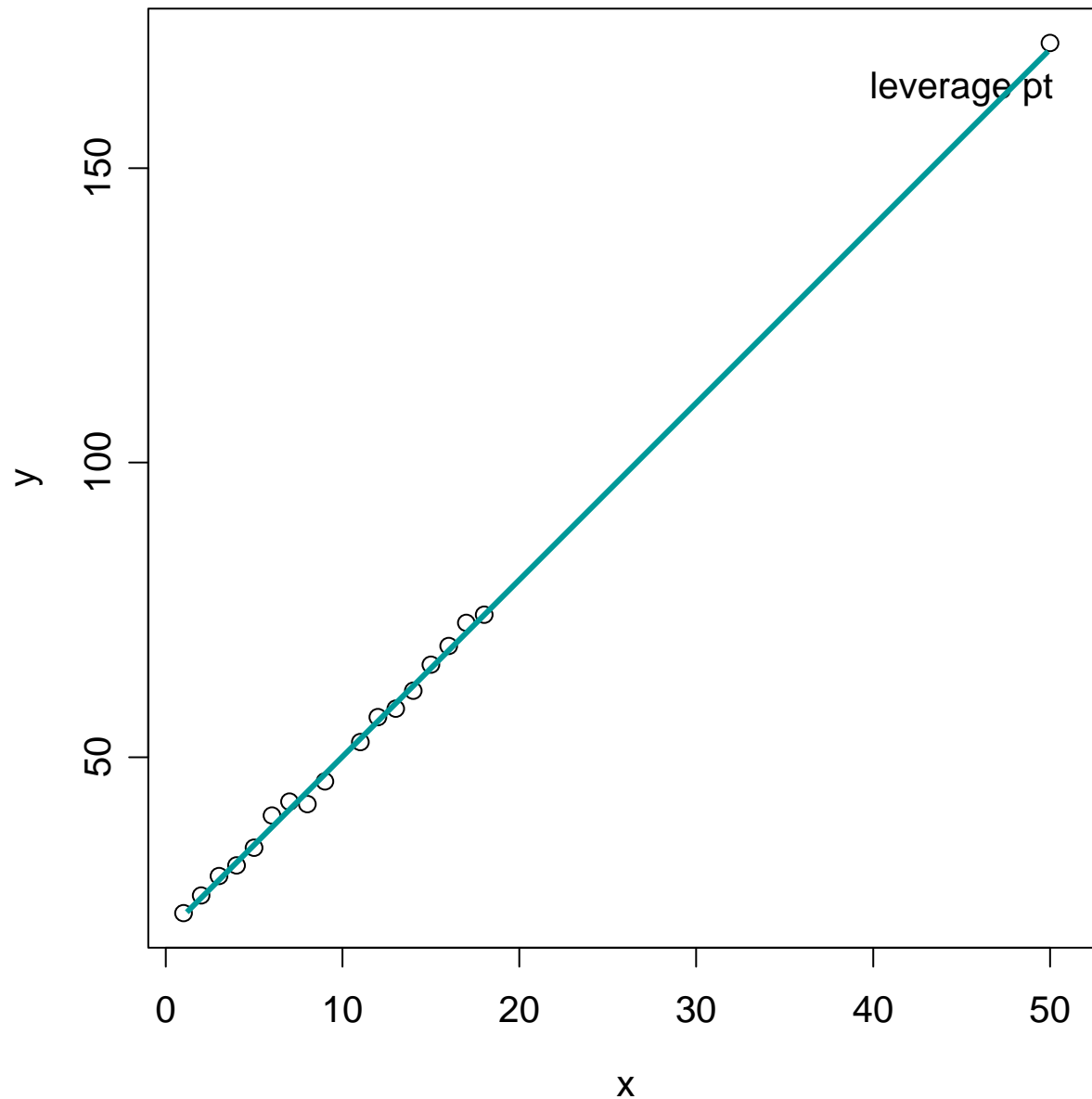
Edwards 2000



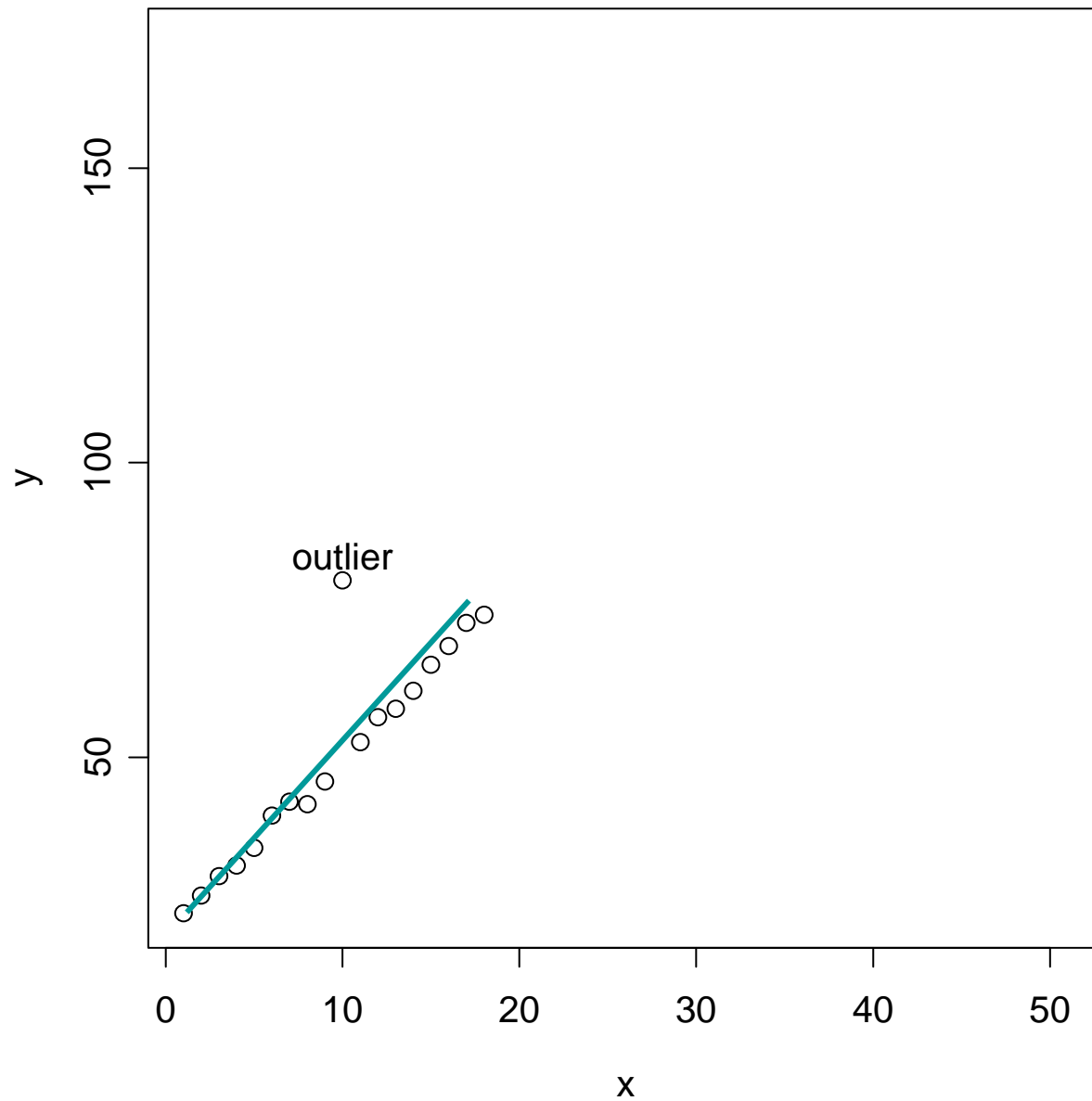
Influential Data Points in a Simple Linear Regression



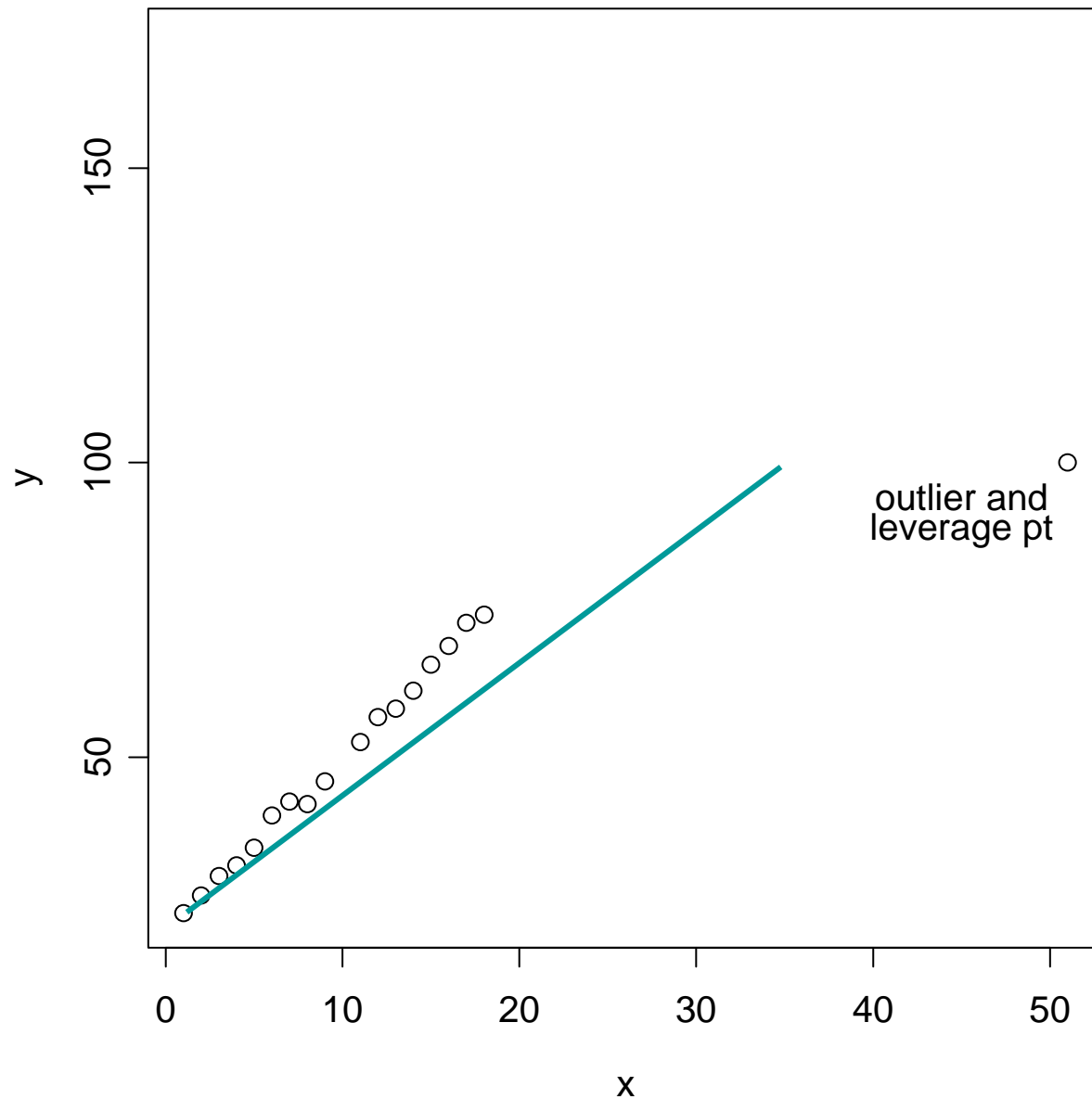
Influential Data Points in a Simple Linear Regression



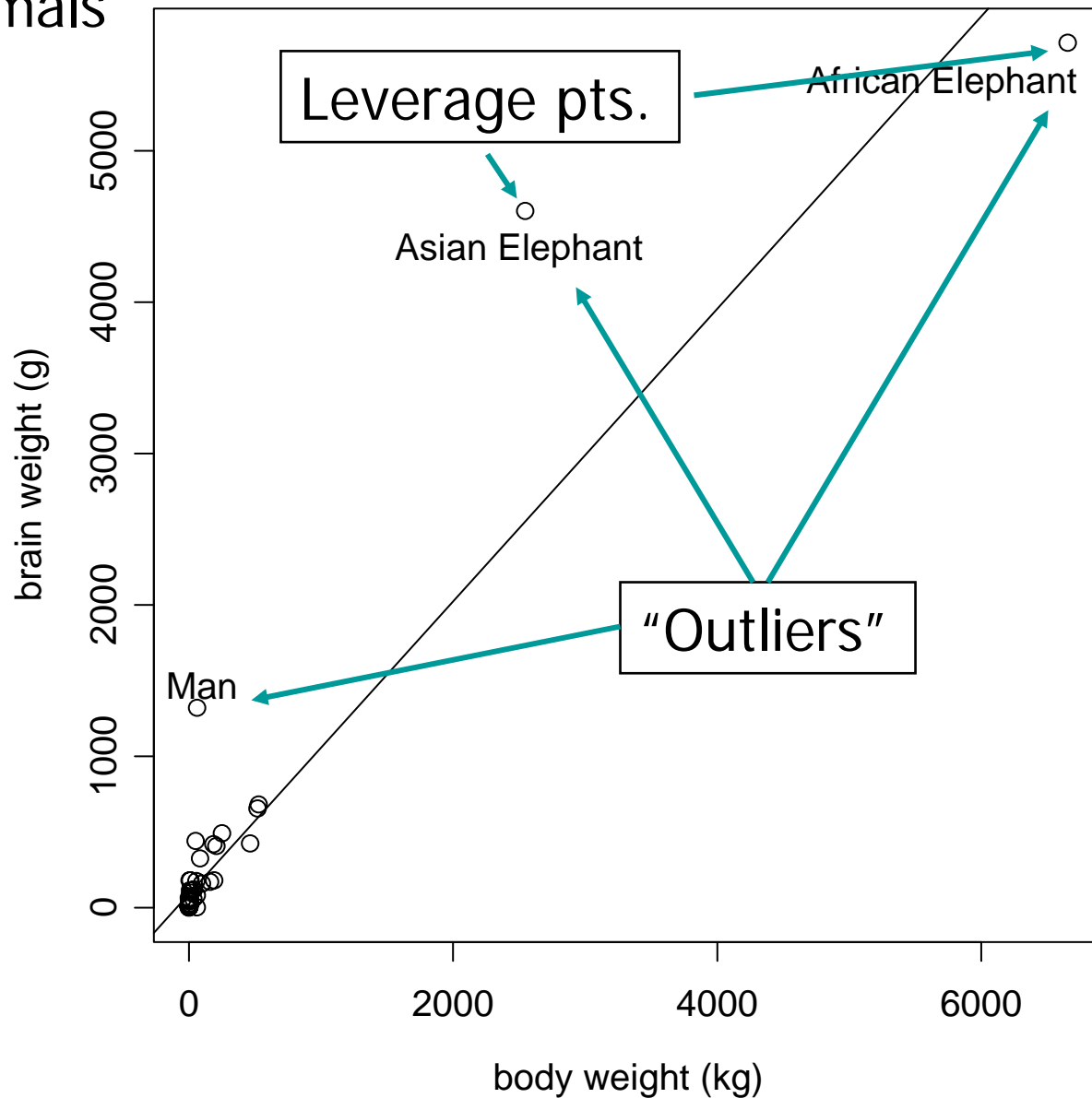
Influential Data Points in a Simple Linear Regression



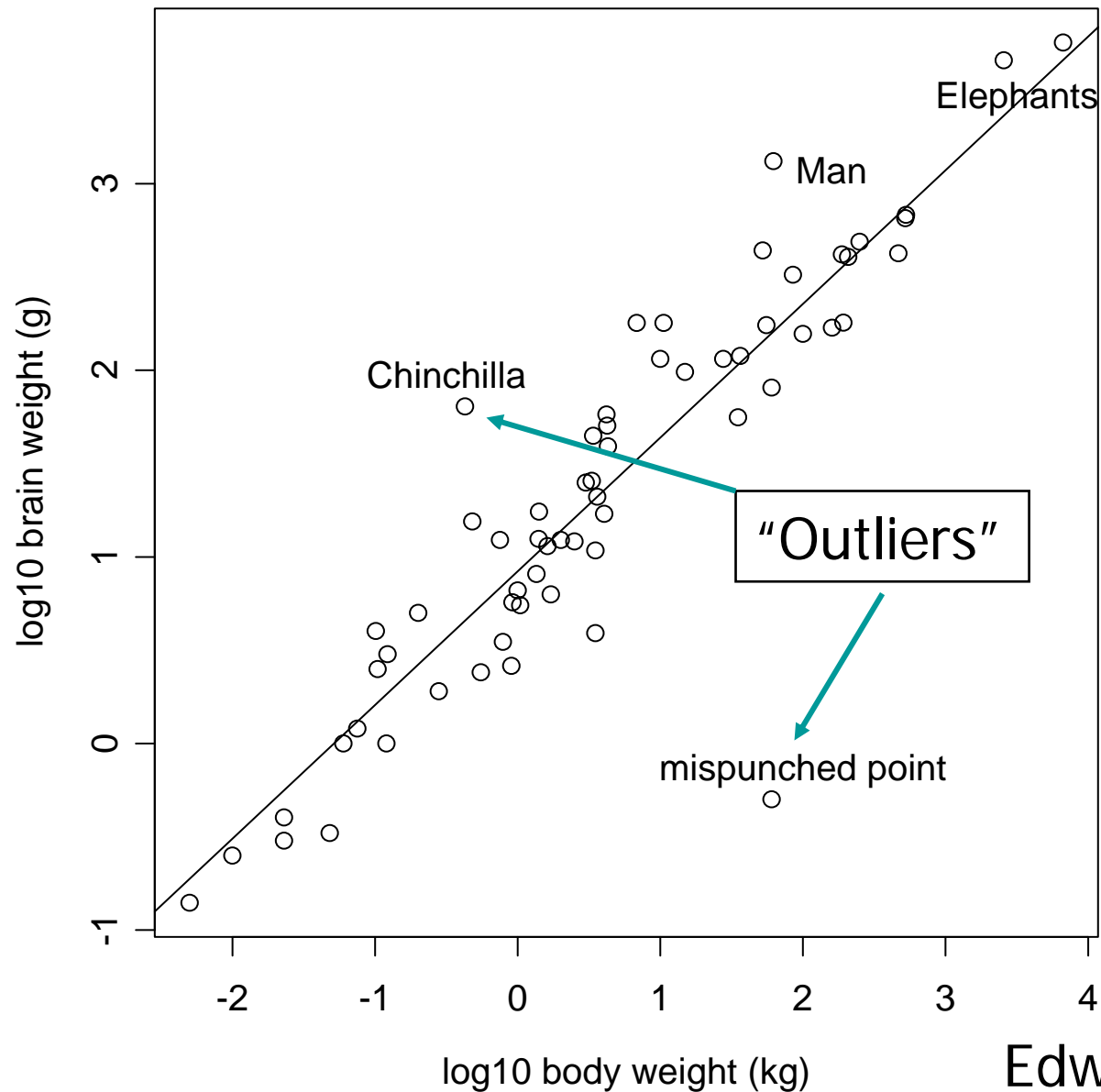
Influential Data Points in a Simple Linear Regression



Brain weight vs. body weight, 63 species of terrestrial mammals



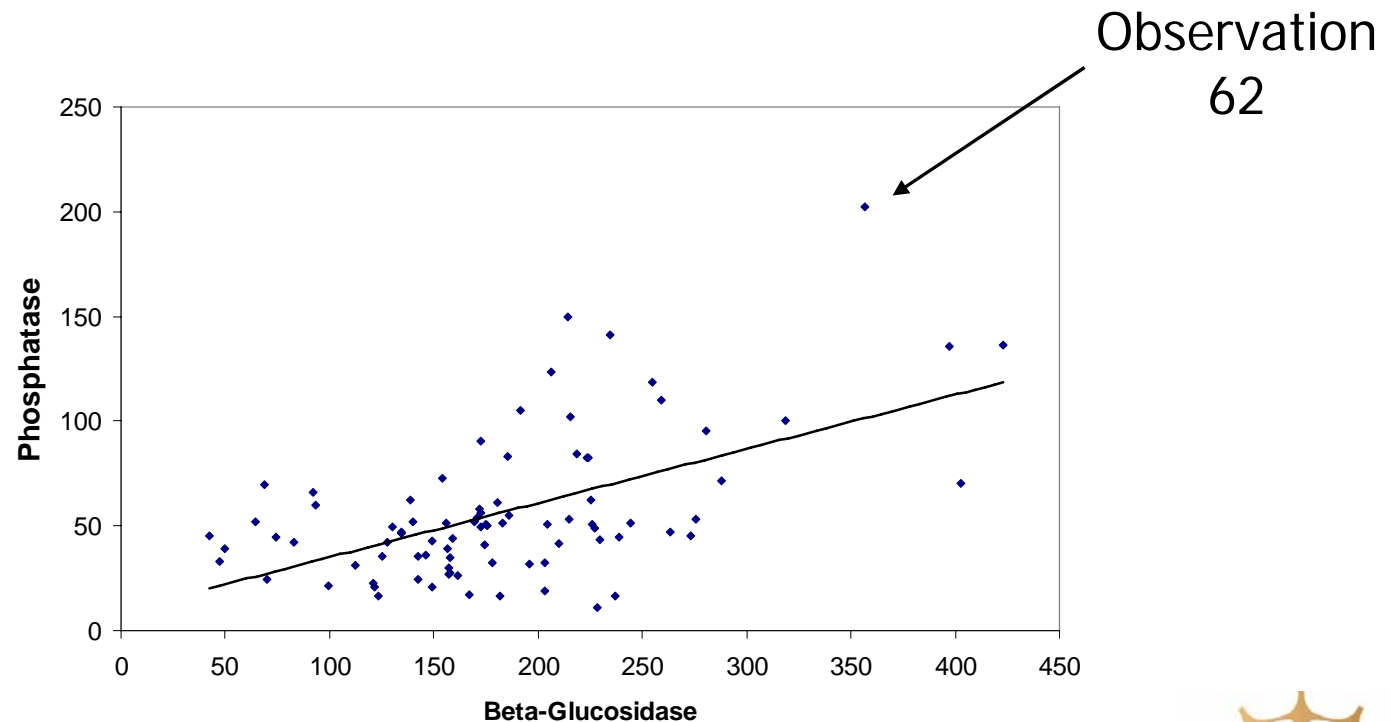
Logged brain weight vs. logged body weight



Edwards 2000



Outliers in simple linear regression





Outliers: identify using studentized residuals

- Contamination may exist if

$$|r_i| > t_{\alpha/2, n-3}$$

$$\alpha = 0.01$$

Where r_i is a studentized residual





Simple linear regression: Outlier identification

$n = 86$

$t_{\alpha/2, 83} = 1.98$

The SAS System

92

14:12 Thursday, October 12, 2000

Obs	Dep Var PHOS	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual
49	82.9000	166.1	7.290	-83.1835	62.045	-1.341
50	238.6	169.2	7.114	69.4246	62.066	1.119
51	215.1	179.5	6.765	35.6463	62.105	0.574
52	166.8	136.0	10.205	30.8387	61.633	0.500
53	157.0	151.0	8.531	6.0259	61.887	0.097
54	142.2	157.4	7.932	-15.2357	61.966	-0.246
55	202.9	154.3	8.215	48.6407	61.929	0.785
56	228.4	128.2	11.183	100.2	61.463	1.631
57	145.9	158.2	7.866	-12.3208	61.975	-0.199
58	218.4	216.5	8.624	1.8552	61.874	0.030
59	156.8	162.5	7.532	-5.6964	62.016	-0.092
60	210.2	164.9	7.364	45.2760	62.036	0.730
61	225.7	176.2	6.833	49.4711	62.097	0.797
62	402.9	199.8	7.250	203.1	62.050	3.273
63	161.3	146.4	9.001	14.8672	61.820	0.240
64	227.3	174.2	6.896	53.1002	62.090	0.855





Simple linear regression: detecting leverage points

$$h_i = (1/n) + (x_i - \bar{x})^2 / ((n-1)S_x^2)$$

A point is a leverage point if $h_i > 4/n$,
where n is the number of points used to fit
the regression

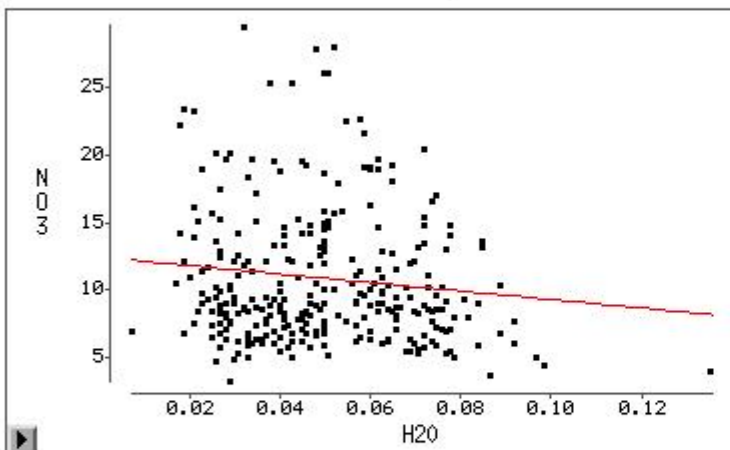




Regression with leverage point: Soil nitrate vs. soil moisture

► N03 = H2O
Response Distribution: Normal
Link Function: Identity

► Model Equation
N03 = 12.3348 - 31.3432 H2O



Parametric Regression Fit								
Curve	Degree(Polynomial)	DF	Model Mean Square	DF	Error Mean Square	R-Square	F Stat	Prob > F
—	1	1	107.9918	290	24.2399	0.0151	4.4551	0.0357

Summary of Fit			
Mean of Response	10.7858	R-Square	0.0151
Root MSE	4.9234	Adj R-Sq	0.0117

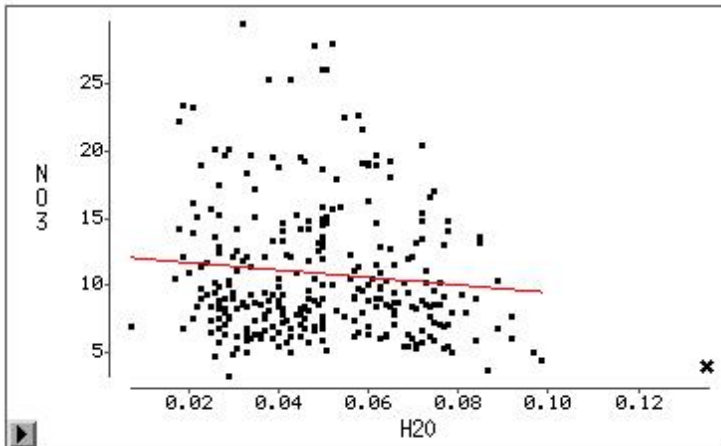




Regression without leverage point

► NO3 = H2O
 Response Distribution: Normal
 Link Function: Identity

► Model Equation
 NO3 = 12.1791 - 27.8885 H2O



Observation 46

Parametric Regression Fit								
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
—	1	1	79.6479	289	24.2620	0.0112	3.2828	0.0710

Summary of Fit			
Mean of Response	10.8091	R-Square	0.0112
Root MSE	4.9256	Adj R-Sq	0.0078





Output from SAS:

Leverage points

The SAS System

155

16:12 Wednesday, October 11, 2000

n = 336

h_i cutoff
value:

$$4/336=0.012$$

Obs	Residual	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	H2O Dfbetas
33	-3.8319	-0.7795	0.0068	1.0095	-0.0645	0.0264	-0.0460
34	-4.8160	-0.9806	0.0076	1.0079	-0.0855	0.0387	-0.0638
35	-4.9625	-1.0199	0.0256	1.0260	-0.1652	0.1217	-0.1539
36	-0.7222	-0.1473	0.0141	1.0211	-0.0176	0.0112	-0.0154
37	-0.7019	-0.1426	0.0068	1.0136	-0.0118	0.0048	-0.0084
38	-1.6159	-0.3283	0.0058	1.0120	-0.0251	0.0083	-0.0164
39	0.6302	0.1279	0.0050	1.0118	0.0091	-0.0022	0.0052
40	-3.4884	-0.7141	0.0197	1.0235	-0.1013	0.0708	-0.0923
41	-4.8933	-0.9952	0.0052	1.0053	-0.0723	-0.0614	0.0434
42	-5.6034	-1.1396	0.0042	1.0022	-0.0737	-0.0543	0.0325
43	-4.4392	-0.9029	0.0058	1.0071	-0.0690	-0.0608	0.0449
44	-4.8692	-0.9906	0.0058	1.0060	-0.0757	-0.0667	0.0492
45	-3.8633	-0.7852	0.0052	1.0079	-0.0570	-0.0485	0.0343
46	2.2328	0.4693	0.0711	1.0823	0.1299	-0.1079	0.1268
47	-3.3892	-0.6889	0.0058	1.0094	-0.0527	-0.0464	0.0342
48	-0.9312	-0.1893	0.0071	1.0138	-0.0161	-0.0149	0.0117





References:

- Rousseeuw, P.J. and Leroy, A.M.:1987 *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.
- Cook, R. D. (1977). "Detection of influential observations in linear regression" *Technometrics* 19, 15-18





Outline:

- Define QA/QC
- QC procedures
 - Designing data sheets
 - Data entry using validation rules, filters, lookup tables
- QA procedures
 - Graphics and Statistics
 - Outlier detection
 - Samples
 - Simple linear regression
- Archiving data





Archiving high quality data for easy reuse

- Avoid inconsistencies (e.g. different date ranges in title vs. the data)
- Avoid using the same column title more than once

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
4		All long-term data are from W6 (weighted averages of all gauges in the watershed), from 1964-2000													
5		published in USFS-USDA tech report NE-305													
6															
7					PRECIPITATION (mm)						TEMPERATURE (deg C)				
8	Year	Total	% of avg	June/July	% of avg	Sept/Oct	% of avg	June/July	% of avg	Sept/Oct	% of avg				
9	1994	1246.3	88	239.2	98	169	72	18.7	107	10.25	100				
10	1995	1483.5	104	217.2	89	361.2	154	18.6	107	10.25	100				
11	1996	1823.7	128	416.3	171	243.2	104	17.15	98	10.1	99				
12	1997	1326.7	93	253	104	119.1	51	17.75	102	10.4	101				
13	1998	1474.9	104	432.9	178	211	90	16.9	97	10.55	103				
14	1999	1495.5	105	199.2	82	440	188	19.15	110	11.05	108				
15	2000	1499.7	105	194.7	80	180.8	77	16.85	97	10.15	99				
16	2001	1042	73	201.6	83	230.4	98	17.85	102	11.75	115				
17	2002	1394.8	98	183.9	76	235.3	101	17.5	100	10.5	102				
18	2003	1674.6	118	187.6	77	396.4	169	18	103	10.75	105				
19	2004	1313.5	92	215.2	89	153.7	66	16.6	95	10.9	106				
20															
21	avg total		1423	June/July avg	243	Sept/Oct avg	234	June/July avg	17.45	Sept/Oct avg	10.25				
22															

Figure courtesy of Christine Laney, JRN LTER





Avoid formatting errors, cryptic data, and metadata interspersed with the data

Microsoft Excel - dataformatproblems_examples.xls [Read-Only]

Type a question for help

122

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	year	julian day	date	station	GMT (hour)	GMT (min)	lat (deg)	lat (min)	long (deg)	long (min)	irradiance	depth (m)	NO2- (uM)		
2	2001	313	11/9/2001	B	13	55	-64	46.685	-64	4.161	100	1	0.84		
3	2001	313	11/9/2001	B	13	55	-64	46.685	-64	4.161	50	3	0.53		
4	2001	313	11/9/2001	B	13	55	-64	46.685	-64	4.161	25	10	0.502		
5	2001	313	11/9/2001	SWI	19	0	-64	46.442	-64	3.29	0	0	0.809		
6	2001	319	11/15/2001	SWI	17	40	-64	46.442	-64	3.29	0	0	0.573		
7	2001	320	11/16/2001	B	15	50	-64	46.784	-64	4.456	100	0	0.46		
8	2001	320	11/16/2001	B	15	50	-64	46.784	-64	4.456	50	6	0.5		
9	2001	320	11/16/2001	B	15	50	-64	46.784	-64	4.456	25	15	1.125		
10	2001	320	11/16/2001	B	15	50	-64	46.784	-64	4.456	10	28	0.281		
11	2001	320	11/16/2001	B	15	50	-64	46.784	-64	4.456	5	41	0.423		
12	Probe broke, replaced 11/17/2001														
13	2001	320	11/16/2001	B	15	50	-64	46.784	-64	4.456	~1	45	0.321		
14	2001	323	11/19/2001	e	13	13	-64	48.933	-64	2.273	100	0	0.227		
15	2001	323	11/19/2001	e	13	13	-64	48.933	-64	2.273	50	5	0.197		
16	2001	323	11/19/2001	e	13	13	-64	48.933	-64	2.273	25	14.5	0.198		
17	2001	323	11/19/2001	E	13	13	-64	48.933	-64	2.273	10	27	0.148		
18	2001	323	11/19/2001	E	13	13	-64	48.933	-64	2.273	5	38	0.195		
19	2001	323	11/19/2001	E	13	13	-64	48.933	-64	2.273	1	72	0.1		
20	2001	324	11/20/2001	b	14	12	-64	46.784	-64	4.34	Deep	0	0.208		
21															

Figure courtesy of Christine Laney, JRN LTER





The nit-picky details

- Dates as an example:
 - 2-digit years
 - range of dates in single cell (e.g., 02/01-03/2006 or 02/01/2006,02/03/2006)
 - date with a letter appended to the end (ex: 02/01/1999A)
 - single digit day and month, especially when there are no delimiters between month, day, year. (e.g., 1212005)

Figure courtesy of Christine Laney, JRN LTER





Preferred data formats for synthesis

- Simple ascii delimited with commas, spaces, tabs, etc. with headers, or very simple excel spreadsheets. If fixed-width, give widths and spaces.
- Metadata in separate file
- All data in single file, not separated by year. If not possible, each file in exactly the same format.

Complex formatting systems, like multisheets & several tables in one sheet, are more difficult to interpret and extract information.





Best practices reference:

- Cook, R. B., R. J. Olson, P. Kanciruk, and L. A. Hook. 2001. Best practices for preparing ecological and ground-based data sets to share and archive. Ecol. Bulletins 82:138-141.



Questions?

