

Quality Assurance & Quality Control

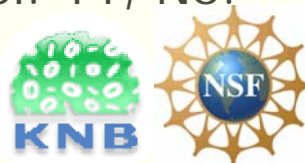




References

□ Primary References

- Michener and Brunt (2000) Ecological Data: Design, Management and Processing. Blackwell Science.
 - Edwards (2000) Ch. 4
 - Brunt (2000) Ch. 2
 - Michener (2000) Ch. 7
- Dux, J.P. 1986. Handbook of Quality Assurance for the Analytical Chemistry Laboratory. Van Nostrand Reinhold Company
- Mullins, E. 1994. Introduction to Control Charts in the Analytical Laboratory: Tutorial Review. Analyst 119: 369 – 375.
- Grubbs, Frank (February 1969), Procedures for Detecting Outlying Observations in Samples, Technometrics, Vol. 11, No. 1, pp. 1-21.





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



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 in a dataset
 - Can be easy to find
 - Malfunctioning instrumentation
 - Sensor drift
 - Low batteries
 - Damage
 - Animal mischief
 - Data entry errors
- **Omission**
 - 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, entered in a computer and analyzed 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





Flowering Plant Phenology – Data Entry Form Design

- ❑ Four sites, each with 3 transects
- ❑ Each species will have phenological class recorded



What's wrong with this data sheet?

Plant

Life Stage

PHENOLOGY DATA SHEET

Collectors: _____

Date: _____ Time: _____

Location: black butte, 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

Collectors Troy Maddux

Date: 16 May 1991 **Time:** 13:12

Location: Deep Well

Transect: 1

Notes: Cloudy day, 3 gopher burrows on transect

ardi P/G V B FL FR M S D NP

Y N Y N Y N Y N Y N Y N Y N Y N Y N

[illegible][illegible][illegible]



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





Validation Rules:

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

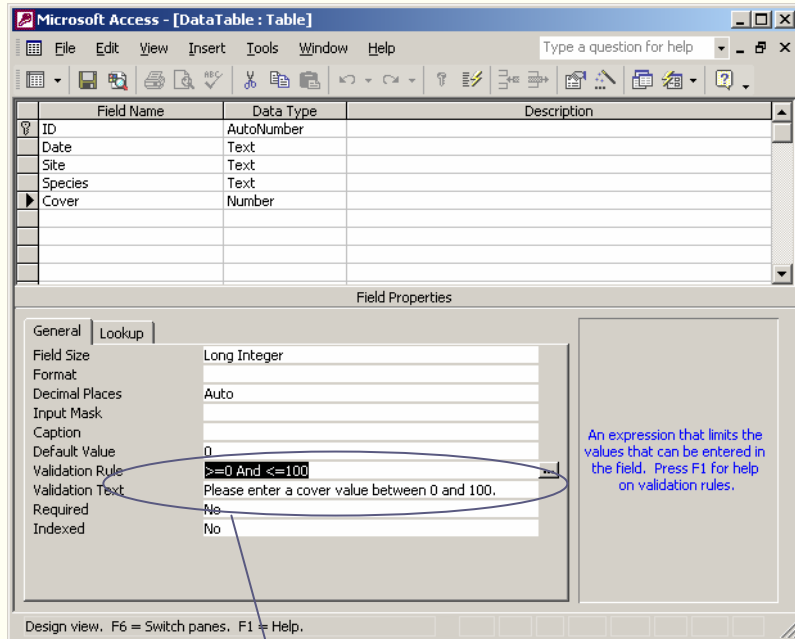


Validation Rule Examples:

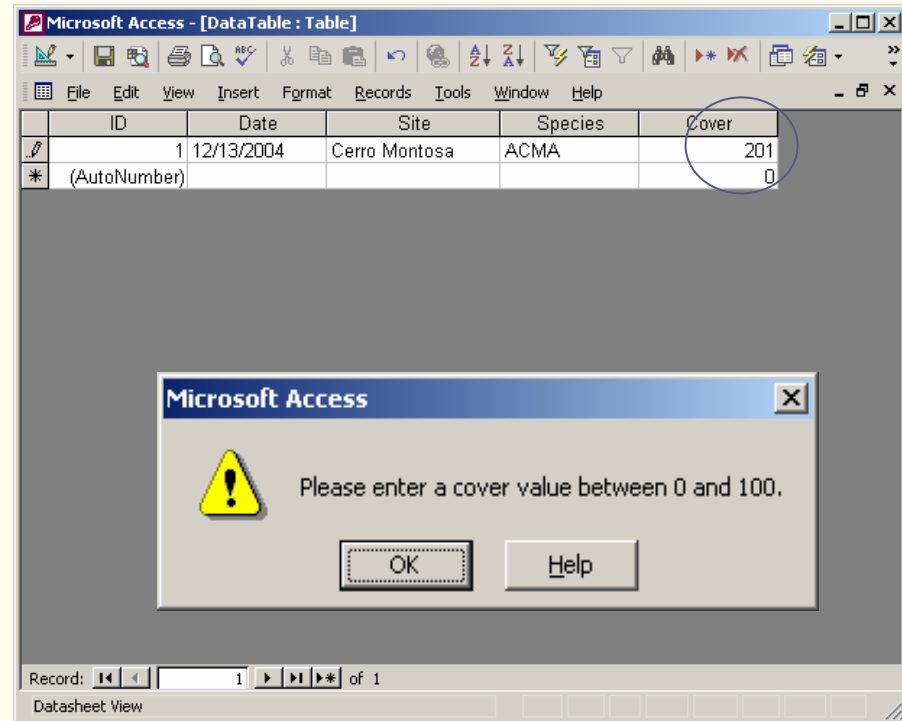
- ≥ 10
- Between 0 and 100
- Between #1/1/70# and Date()



Validation rules in MS Access: Enter in Table Design View



>=0 and <= 100



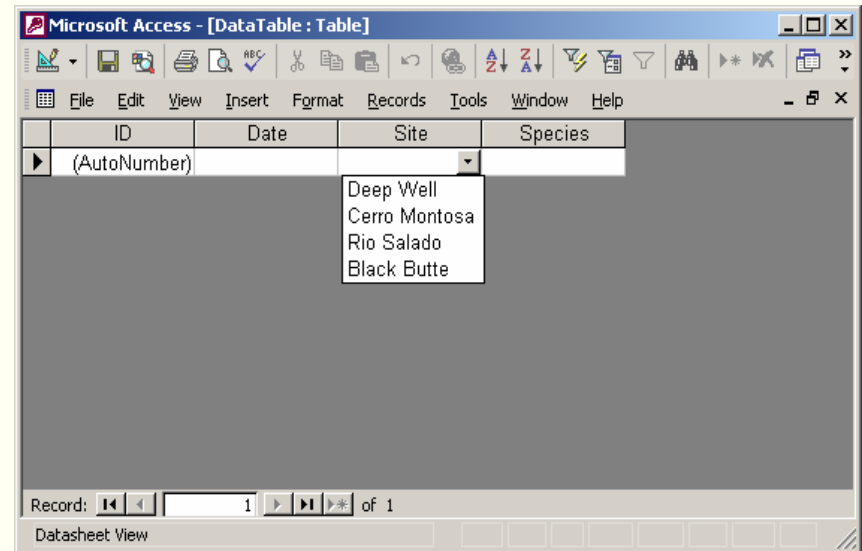
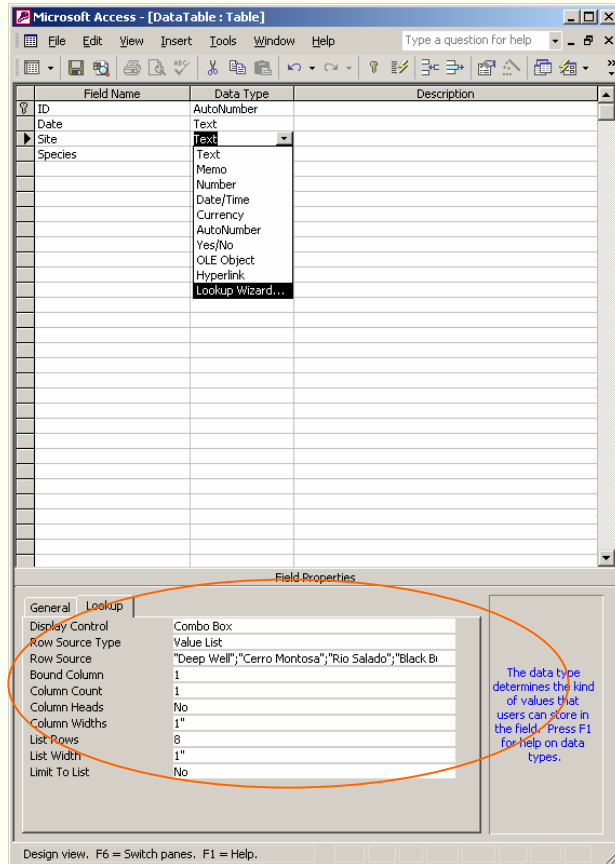


Look-up Fields

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



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





Macros

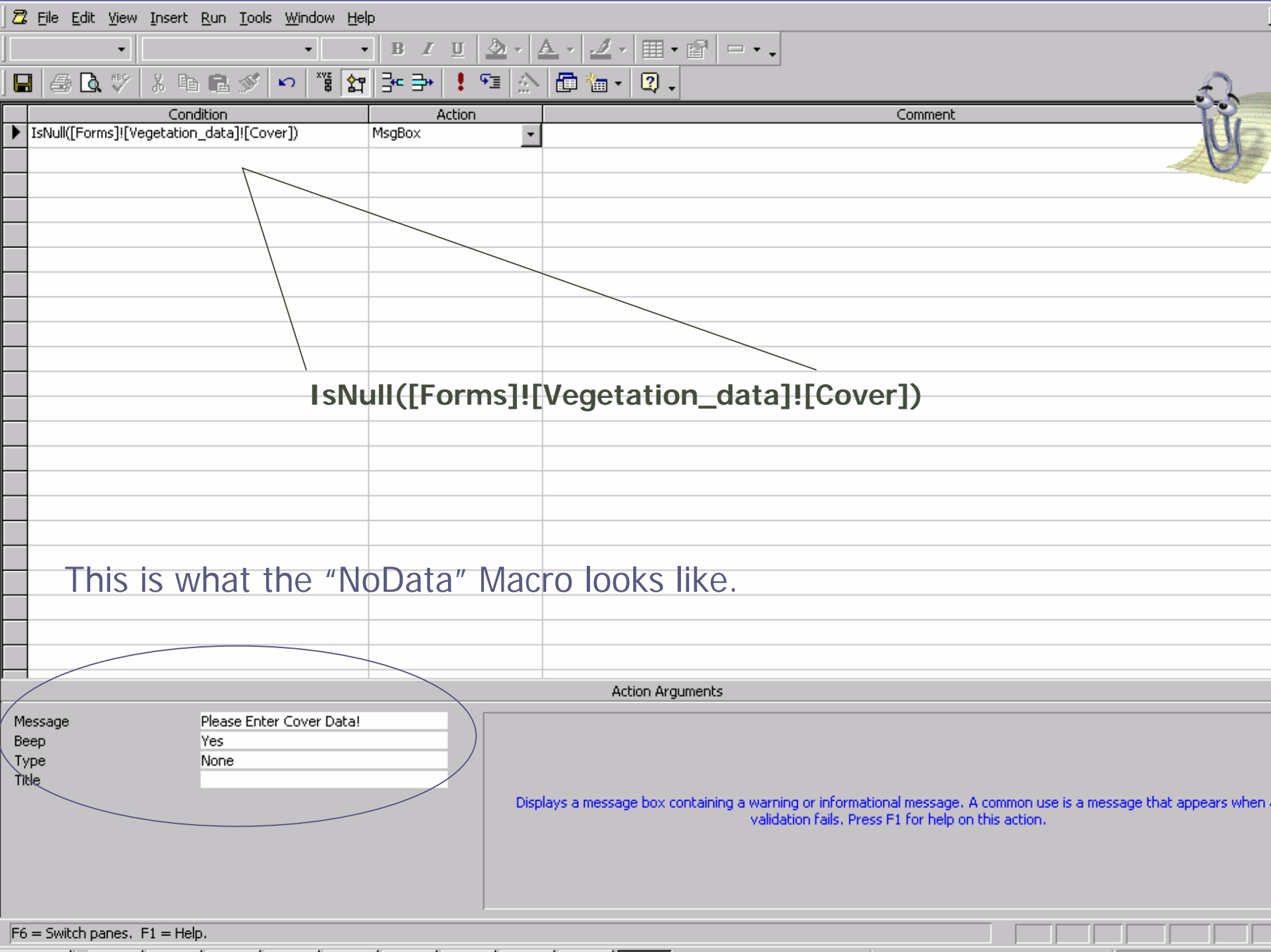
- ❑ Validate data based on conditional statements



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.

The screenshot shows the Microsoft Access interface for a database named [Vegetation_data]. The menu bar includes File, Edit, View, Insert, Format, Records, Tools, Window, and Help. The toolbar shows various formatting and editing options. The data entry form contains the following fields:

- Date: Text box
- Location: Text box with a dropdown arrow
- Plot: Text box
- Plant_Species_Code: Text box with a dropdown arrow
- Cover: Text box, circled in pink
- Observation_Number: Text box with the value [AutoNumber]



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

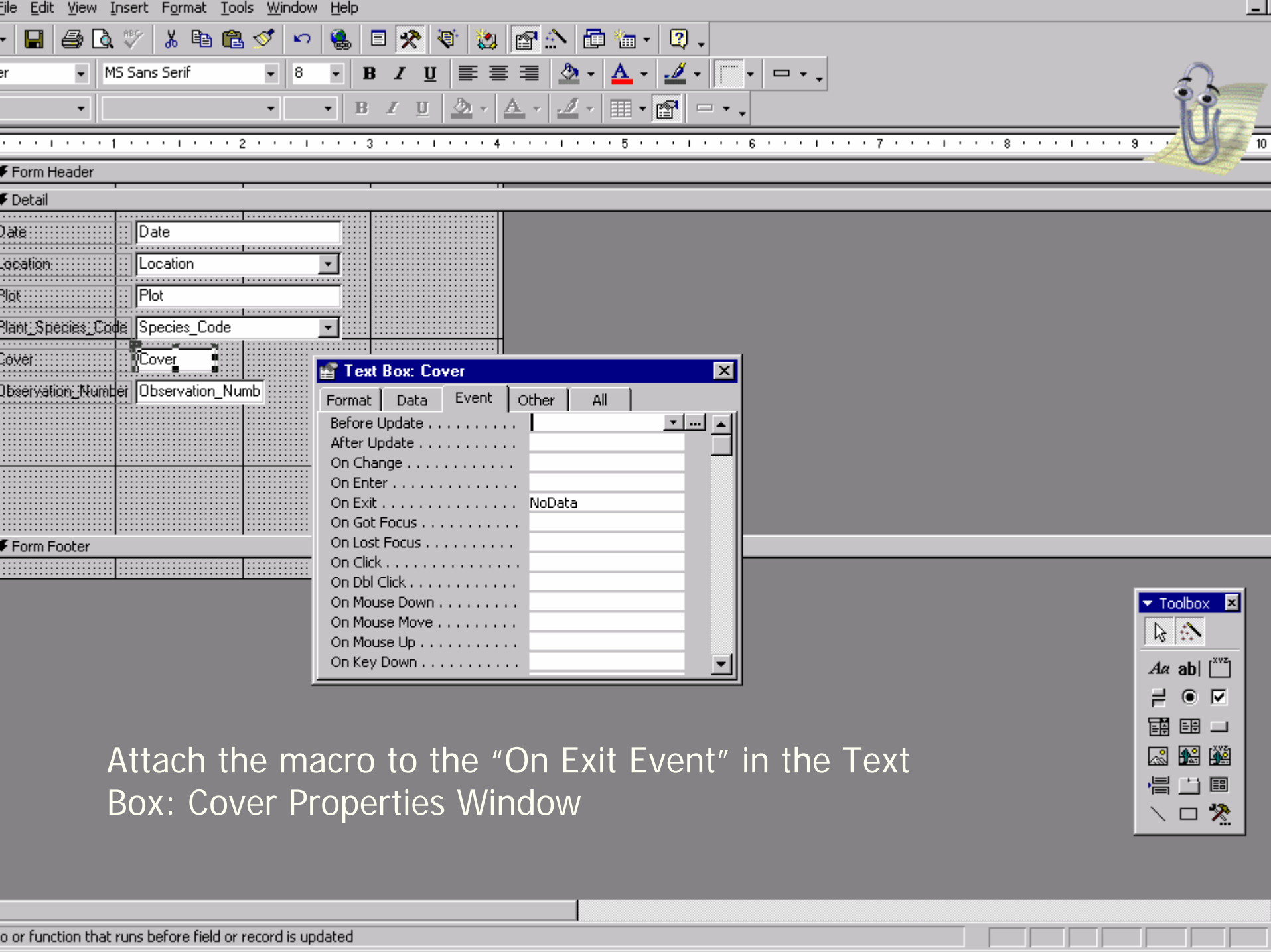
This is what the "NoData" Macro looks like.

Action Arguments

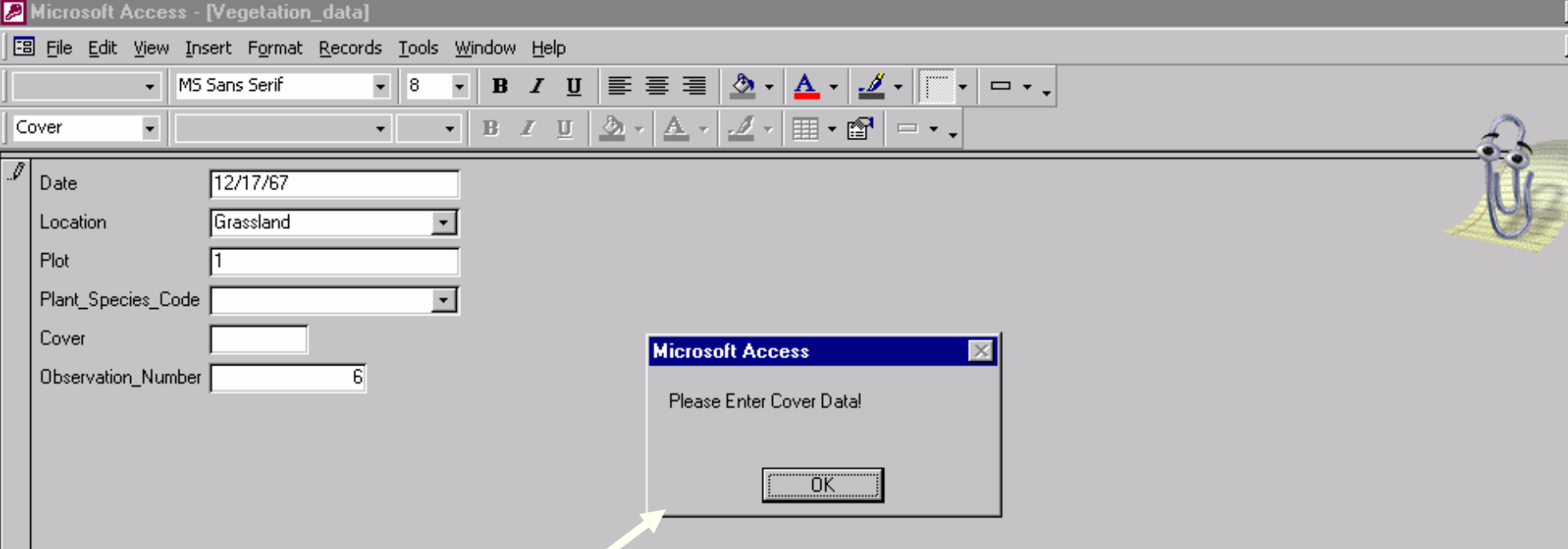
Message
Beep
Type
Title

Please Enter Cover Data!
Yes
None

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

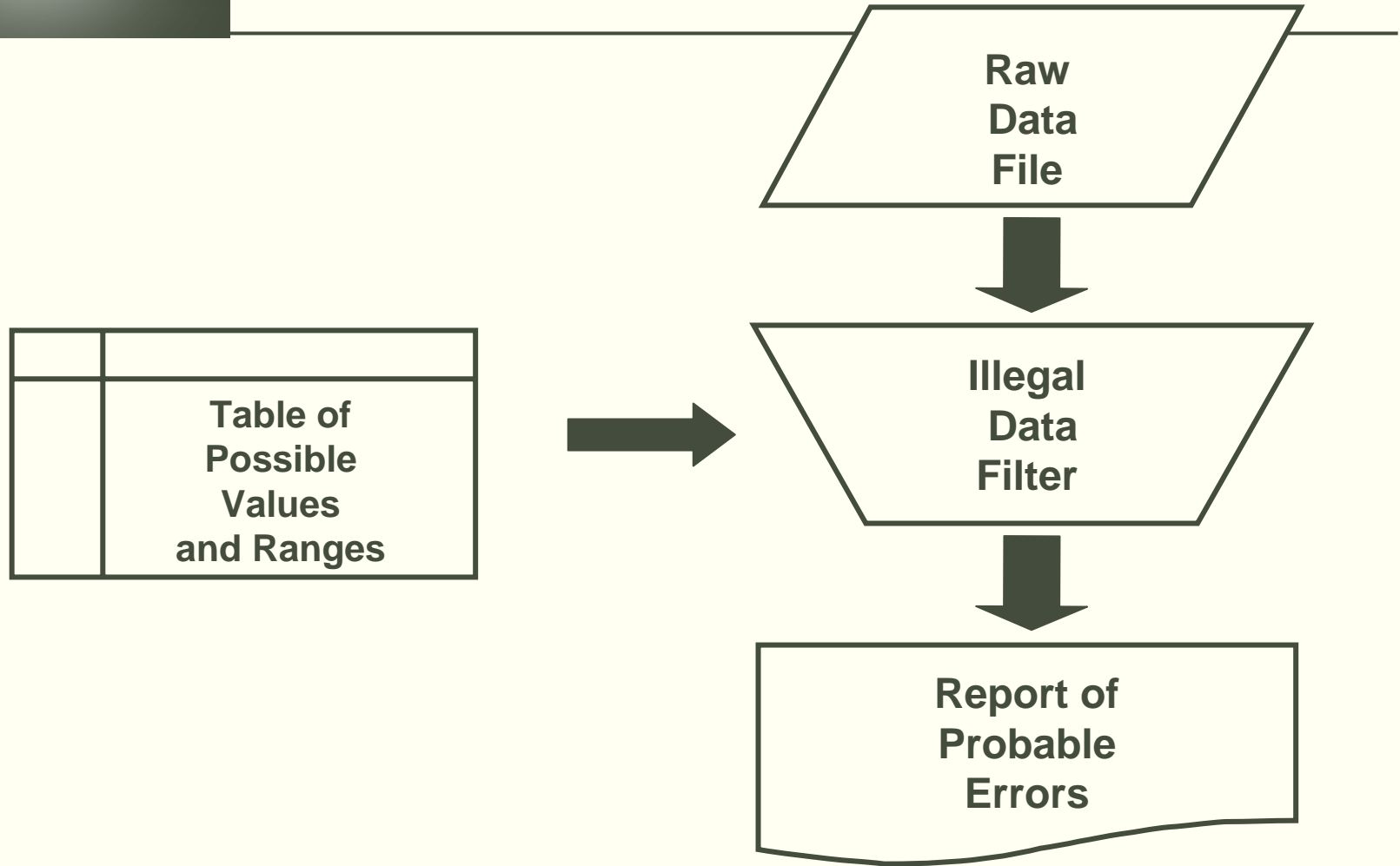
- ❑ Double-keying of data by independent data entry technicians followed by computer verification for agreement
- ❑ Use text-to-speech program to read data back
- ❑ Filters for “illegal” data
 - Statistical/spreadsheet/database programs
 - ❑ Legal range of values
 - ❑ Sanity checks

Edwards 2000



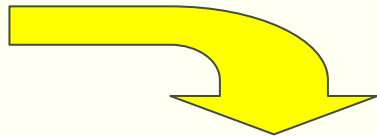


Flow of Information in Filtering For “Illegal” Data



Data set 'tree'

ID	cover (%)	1998 (cm)	1999 (cm)
a	43	3	2
b	133	3	4
c	16	5	4
d	230	2	3



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;
```

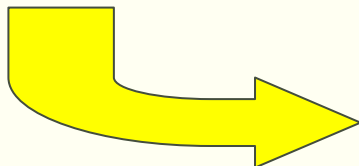
```
If dbh_1998>dbh_1999 then do; message="dbh_1998 is  
larger than dbh_1999"; output; end;
```

```
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

<u>id#</u>	<u>mass</u>	<u>body</u>	<u>tail</u>	<u>foot</u>	<u>rangecheck</u> <u>mass</u>
1	22.5	81.0	85.6	23.00	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

=if(mass>50,1,0)

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:

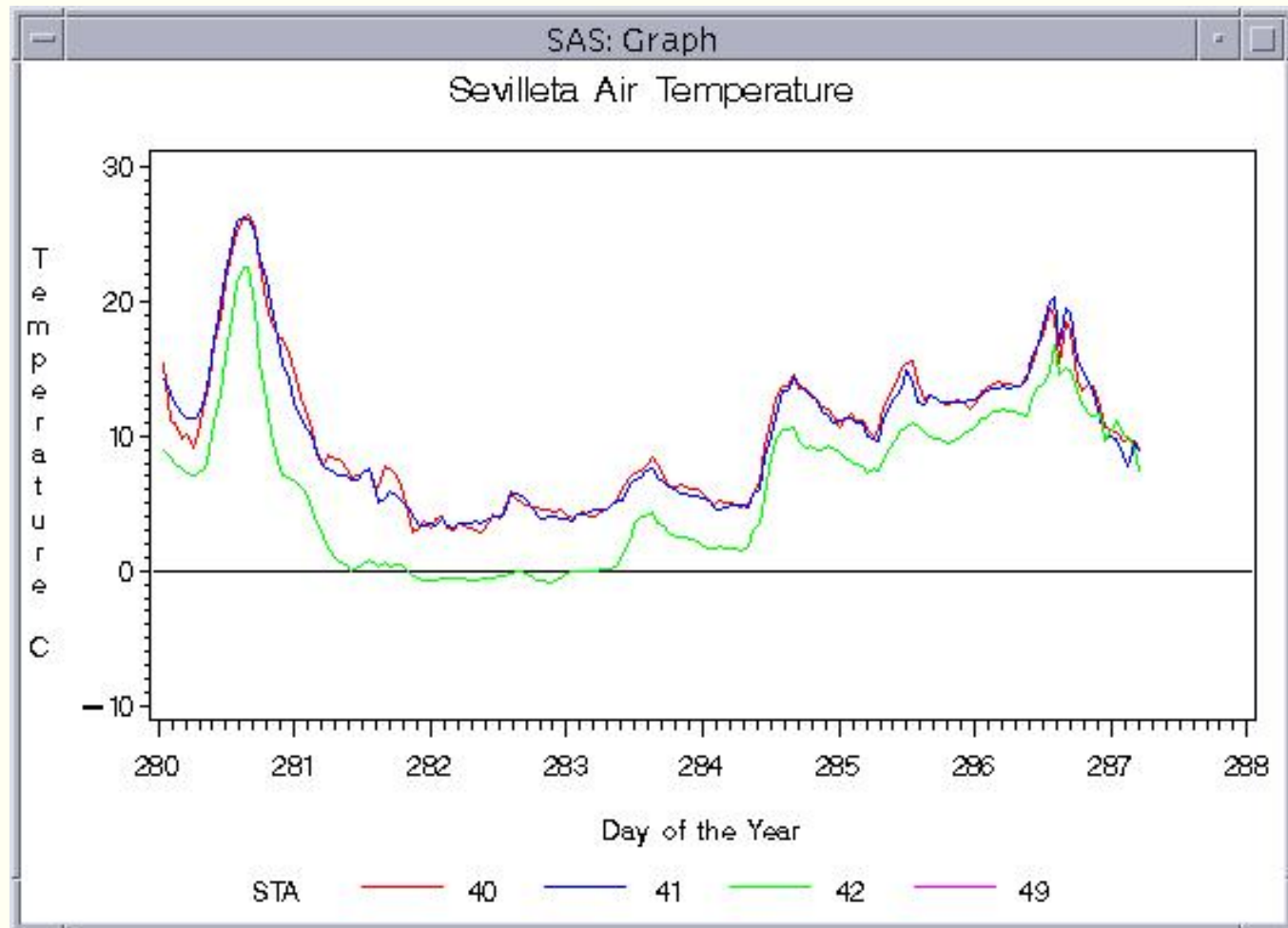
- Define QA/QC
- QC procedures
 - Designing data sheets
 - Data entry using validation rules, filters, lookup tables
- QA procedures
 - Graphics and Statistics
 - Outlier detection
 - Samples
 - Simple linear regression





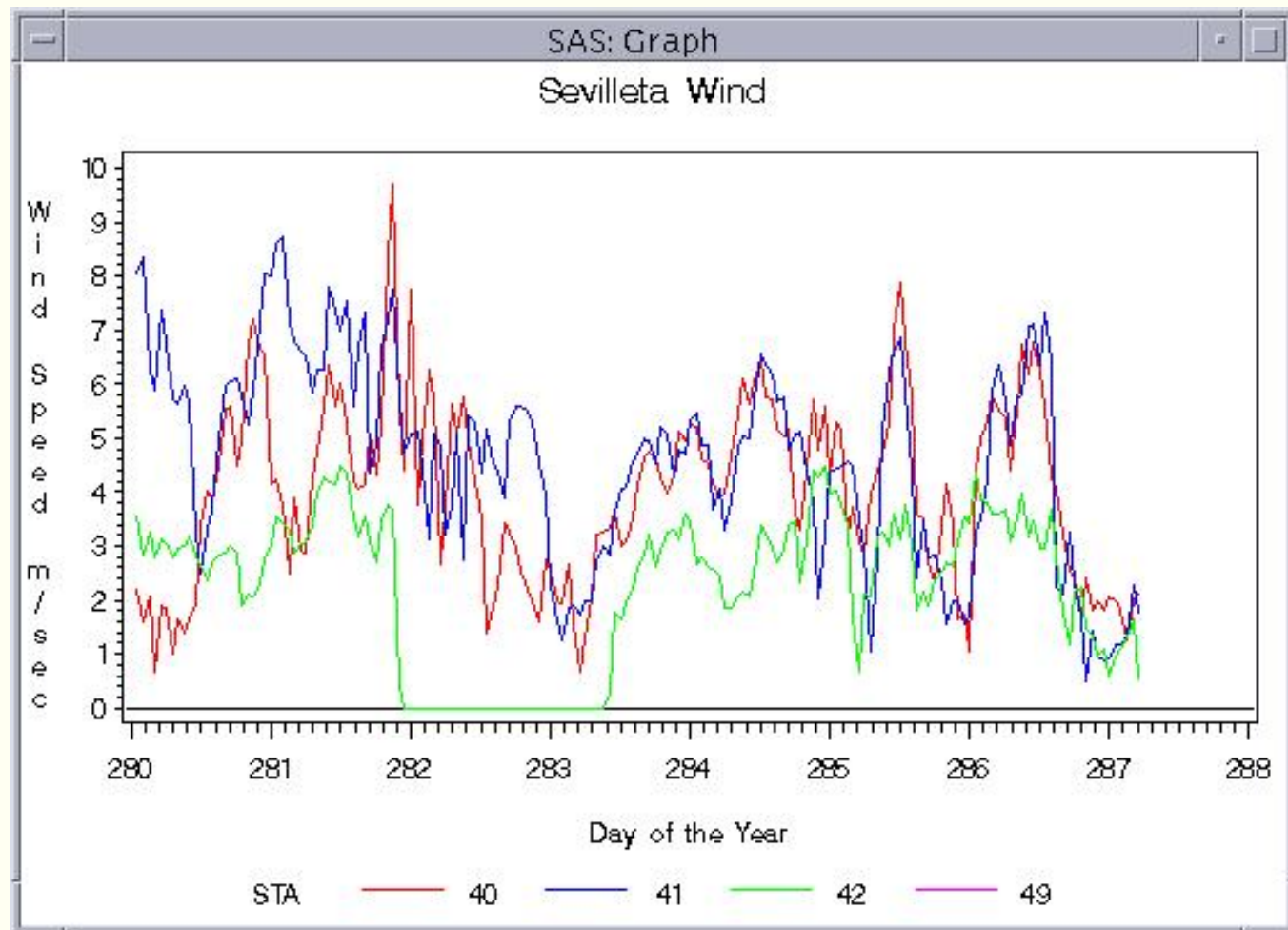
Identifying Sensor Errors:

Comparison of data from three Met stations, Sevilleta LTER





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



QA/QC in the Lab: Using Control Charts





Laboratory quality control using statistical process control charts:

- Determine whether analytical system is “in control” by examining
 - Mean
 - Variability (range)



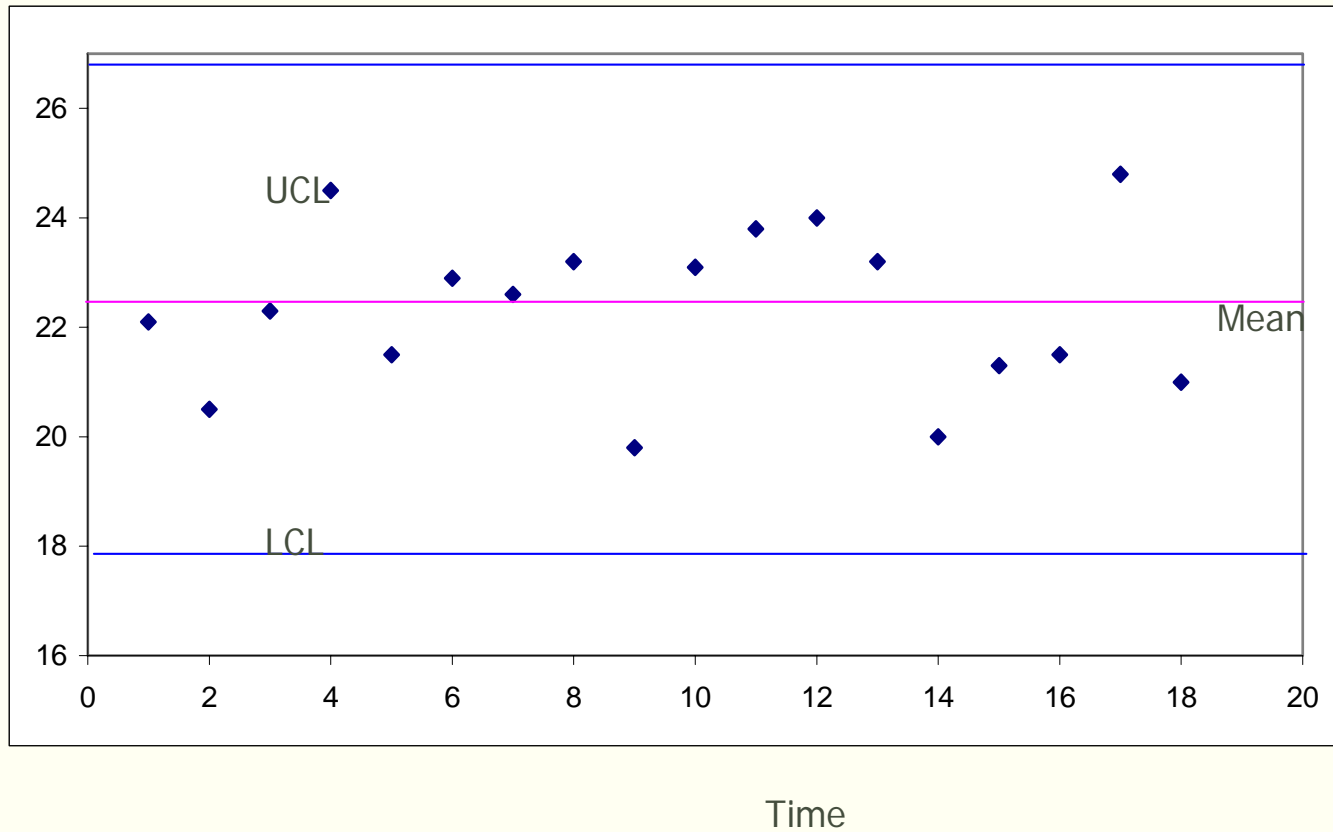
All control charts have three basic components:

- a centerline, usually the mathematical average of all the samples plotted.
- upper and lower statistical control limits that define the constraints of common cause variations.
- performance data plotted over time.





X-Bar Control Chart





Constructing an X-Bar control chart

- Each point represents a check standard run with each group of 20 samples, for example
- $UCL = \text{mean} + 3 * \text{standard deviation}$

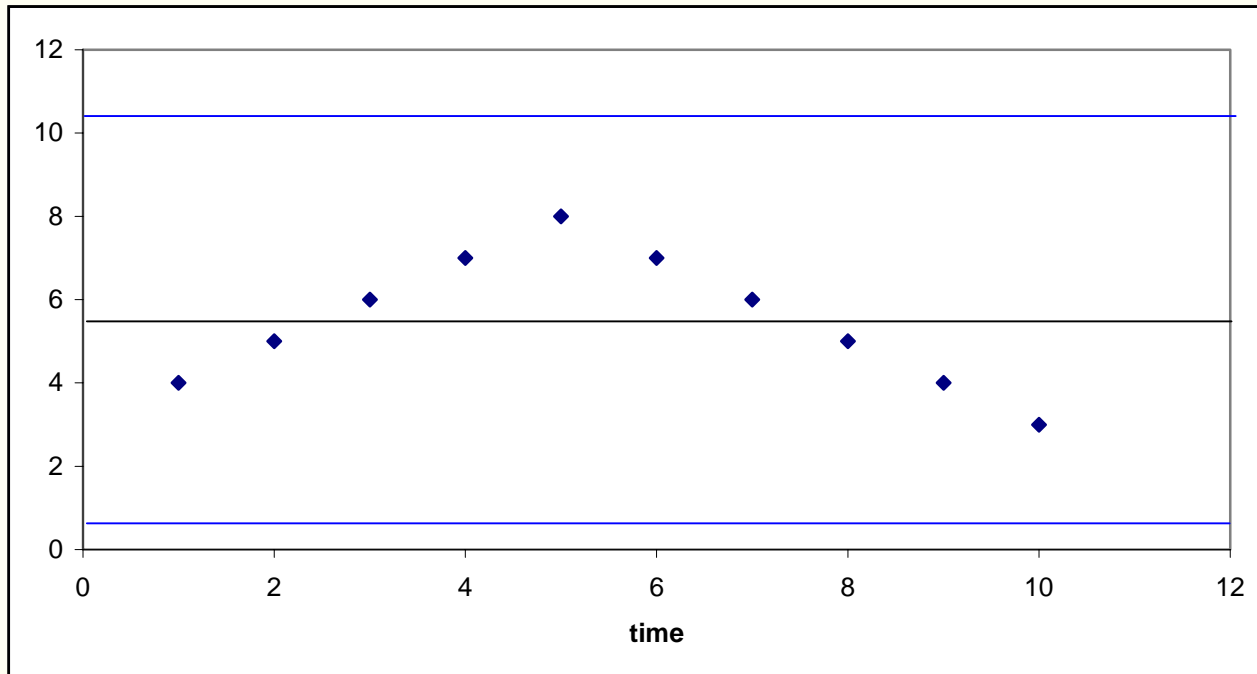


Things to look for in a control chart:

The point of making control charts is to look at variation, seeking patterns or statistically unusual values. Look for:

- 1 data point falling outside the control limits
- 6 or more points in a row steadily increasing or decreasing
- 8 or more points in a row on one side of the centerline
- 14 or more points alternating up and down

Linear trend





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





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.



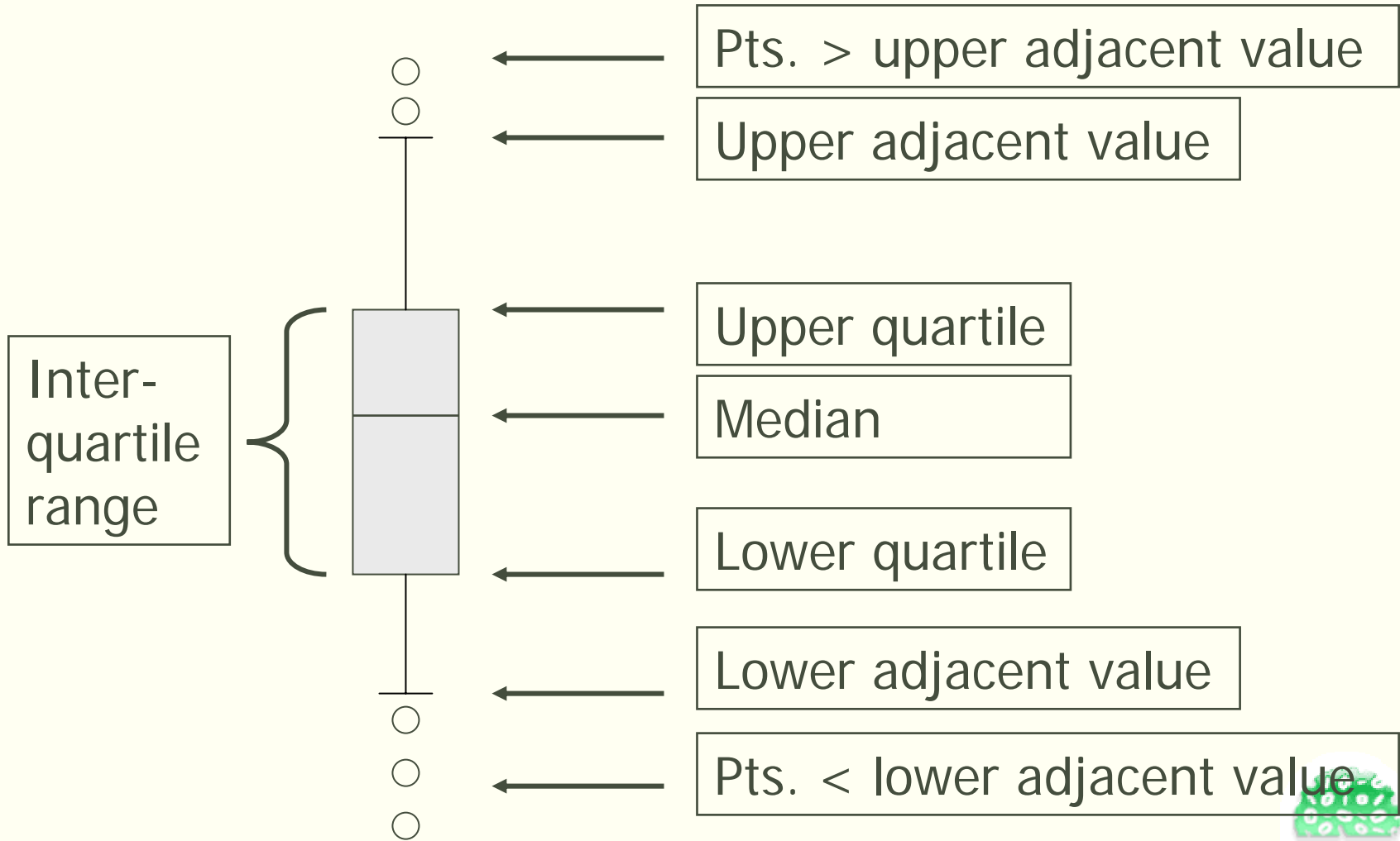
Methods for Detecting Outliers

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

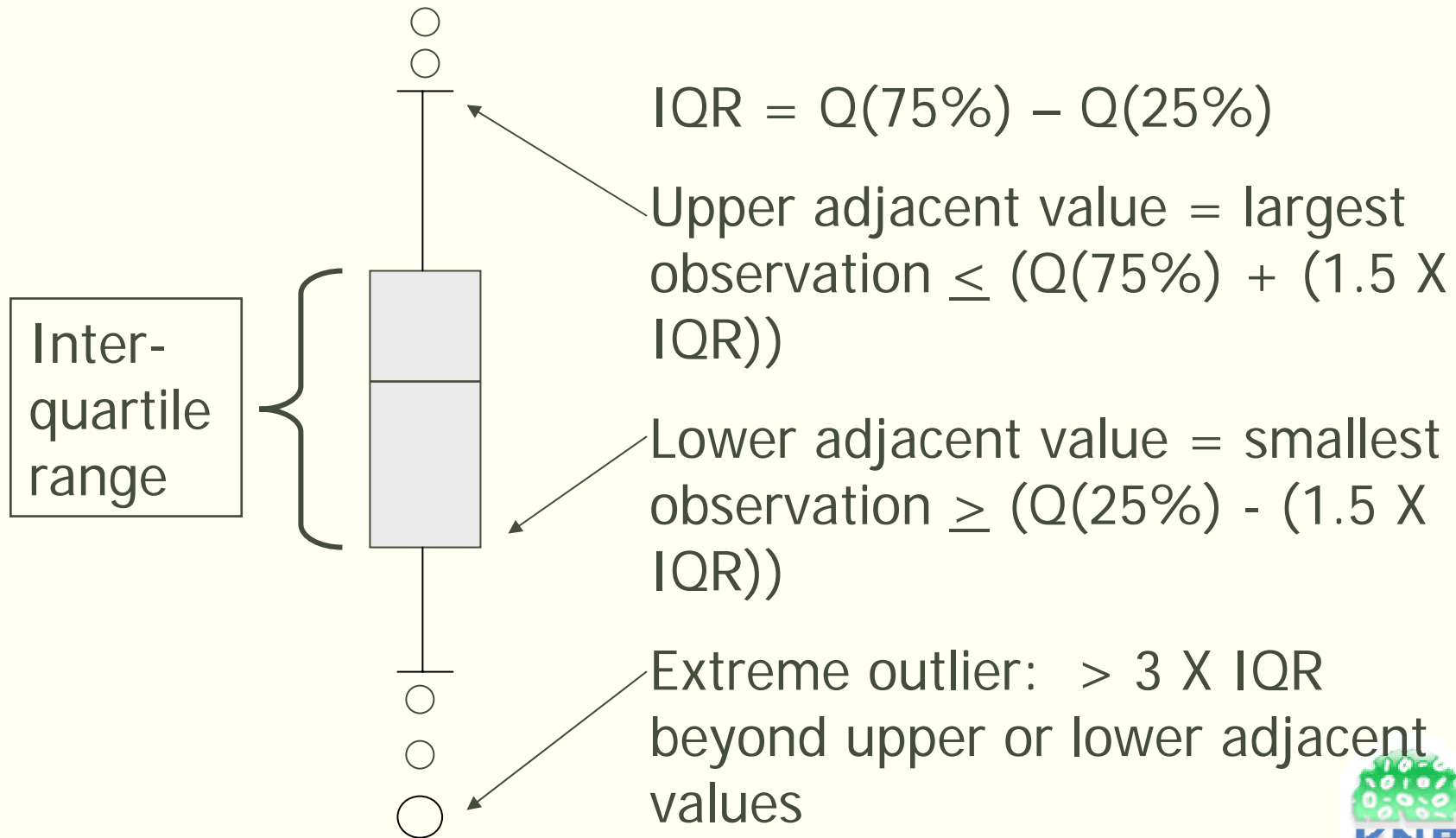
Edwards 2000



Box Plot Interpretation

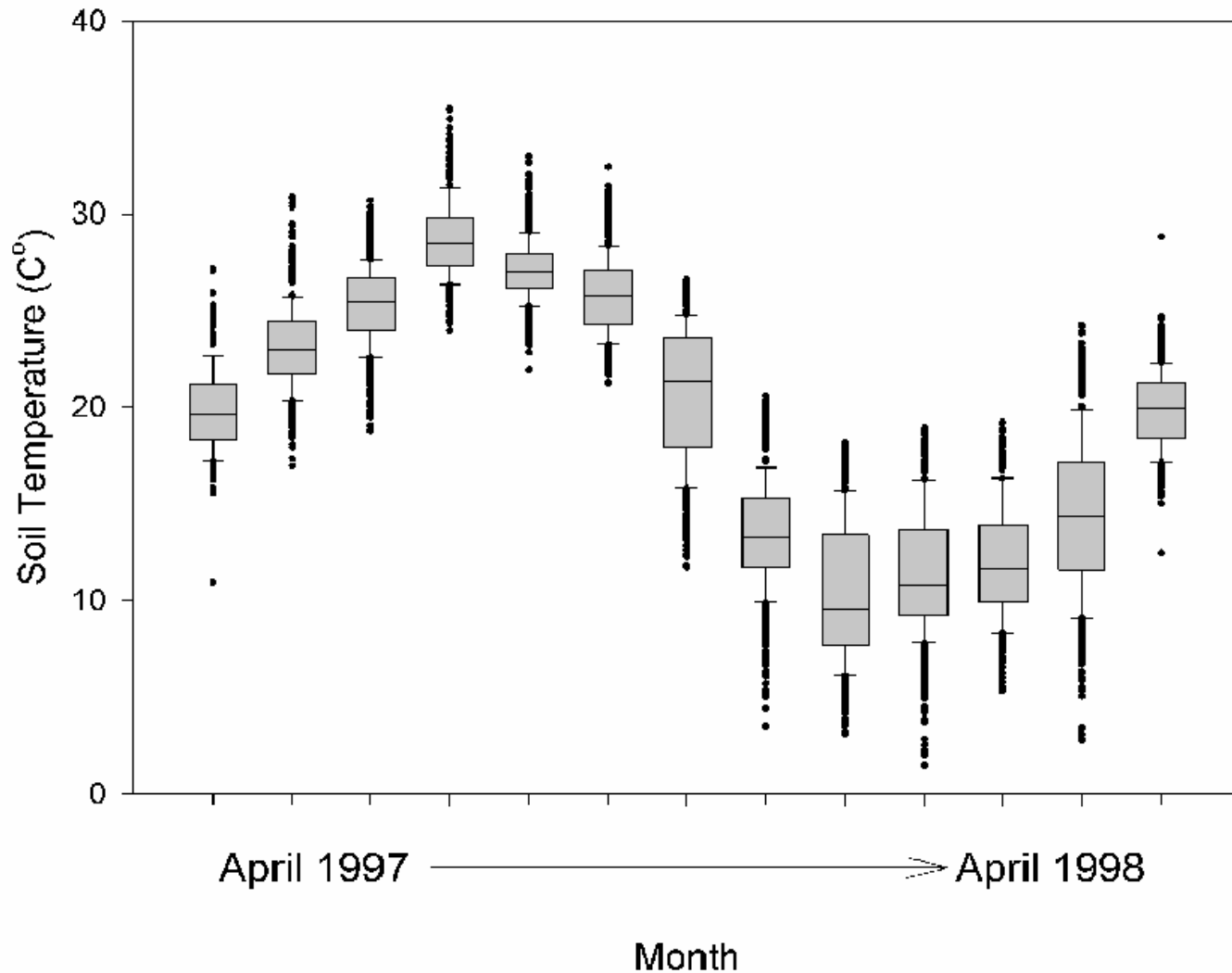


Box Plot Interpretation





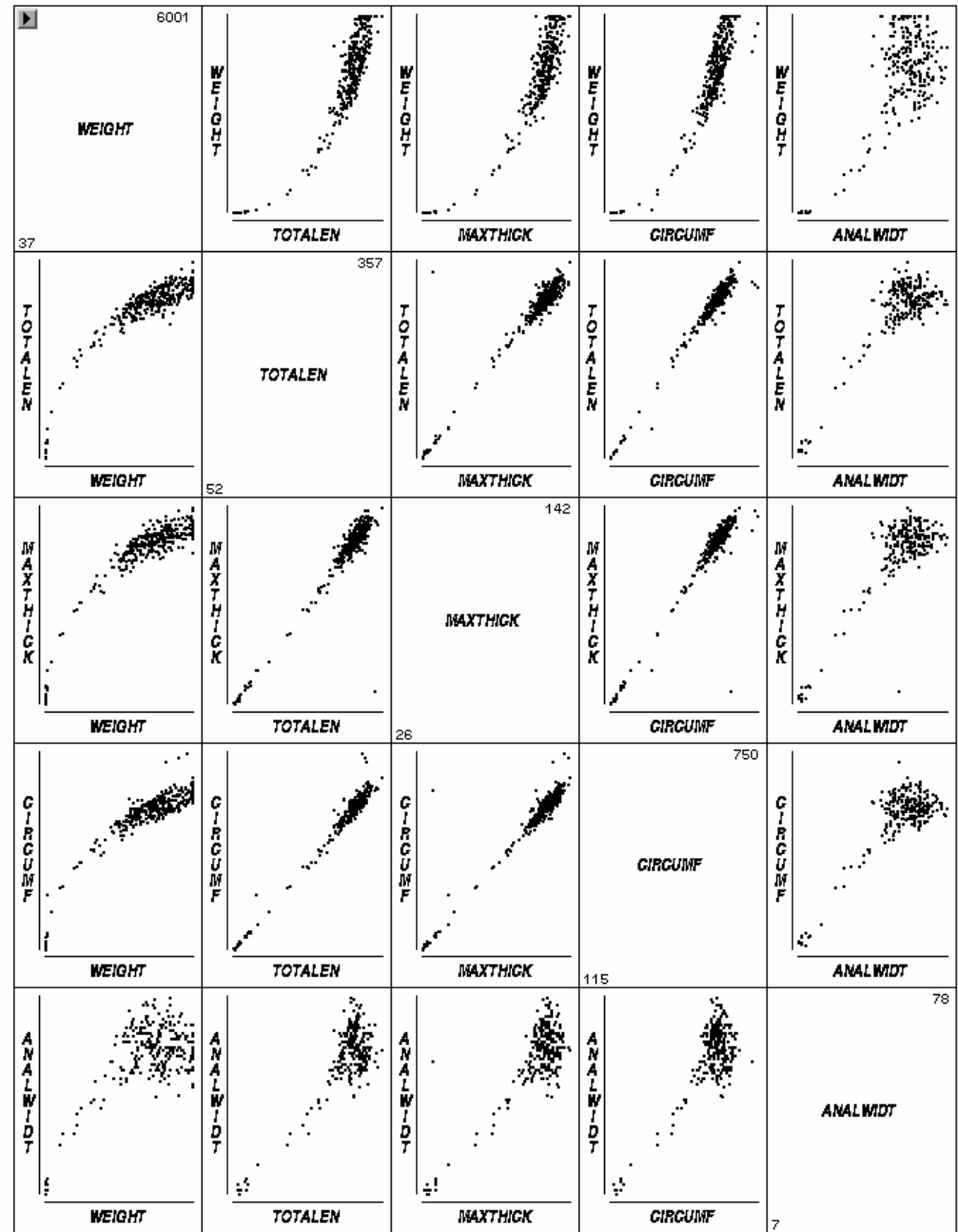
Box Plots Depicting Statistical Distribution of Soil Temperature



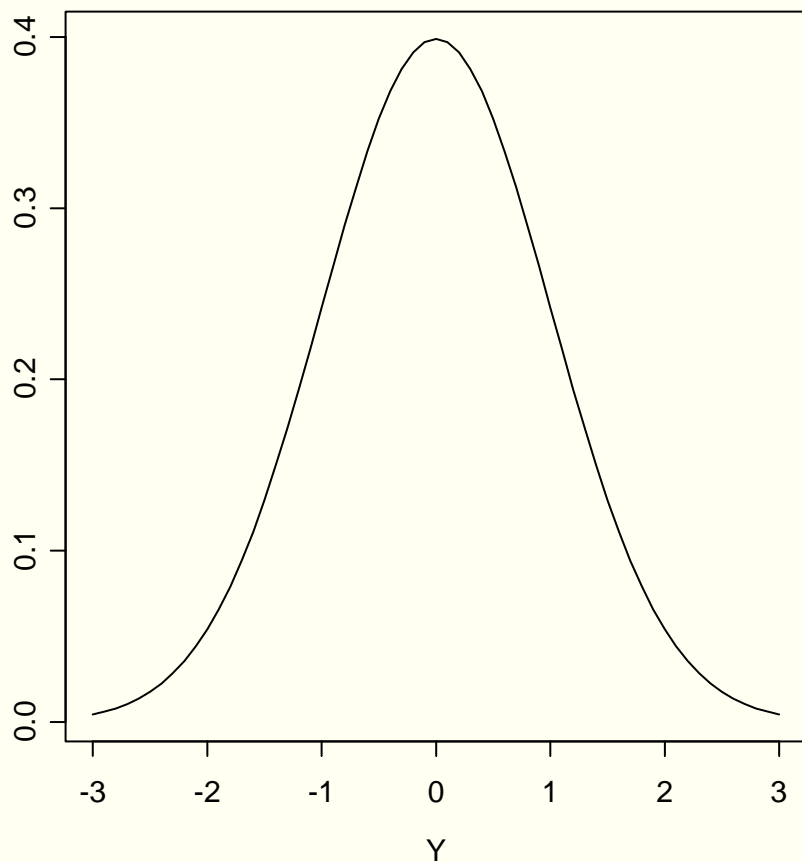


X-Y scatter plots of gopher tortoise morphometrics

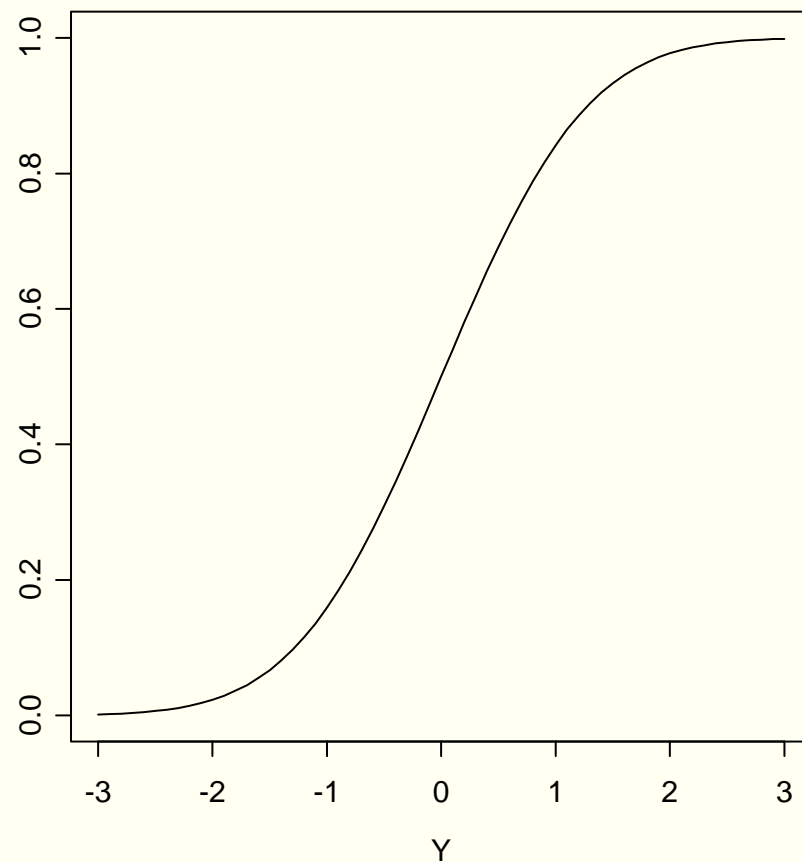
Michener 2000



a. The Normal Frequency Curve

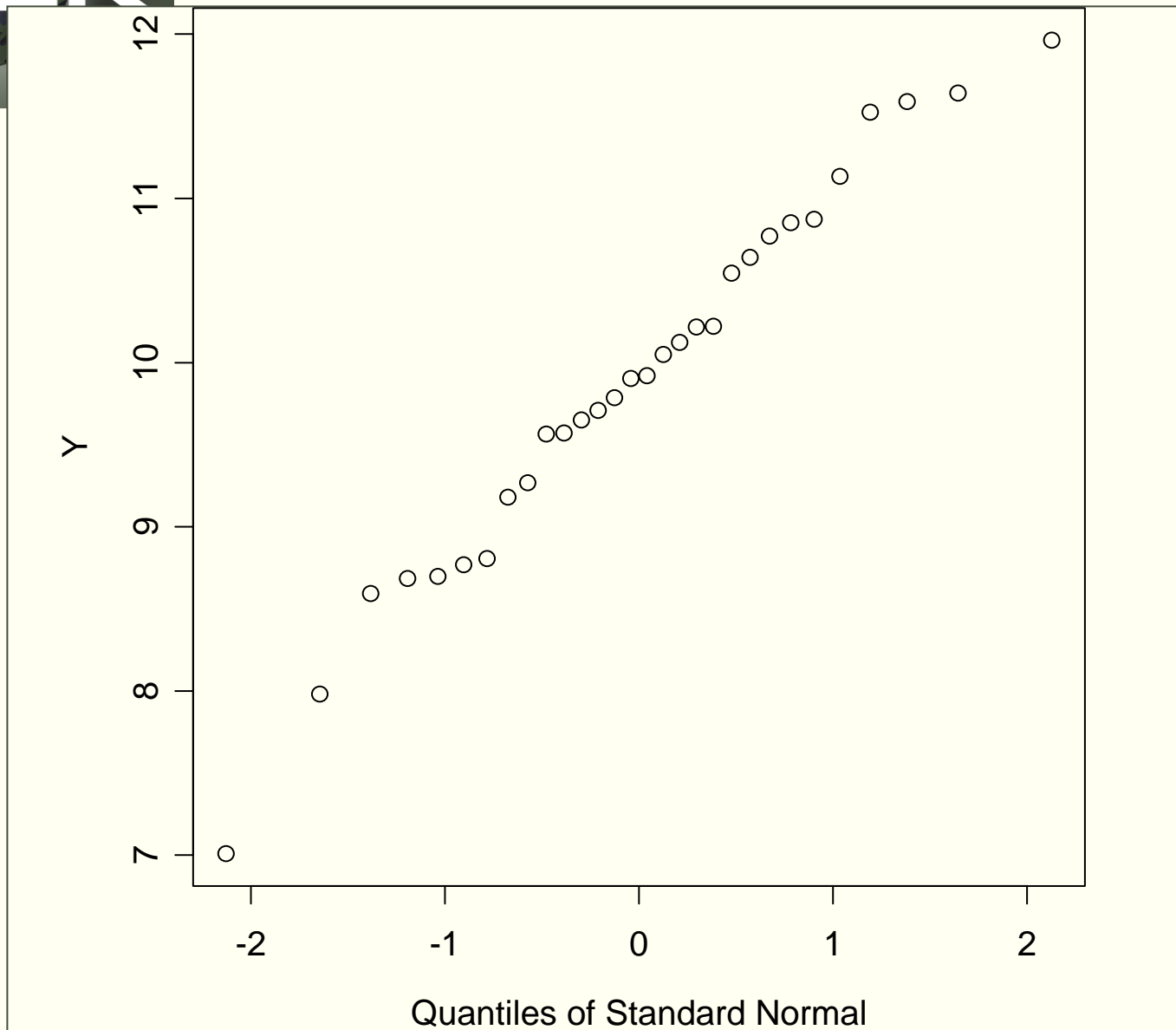


b. Cumulative Normal Curve Areas

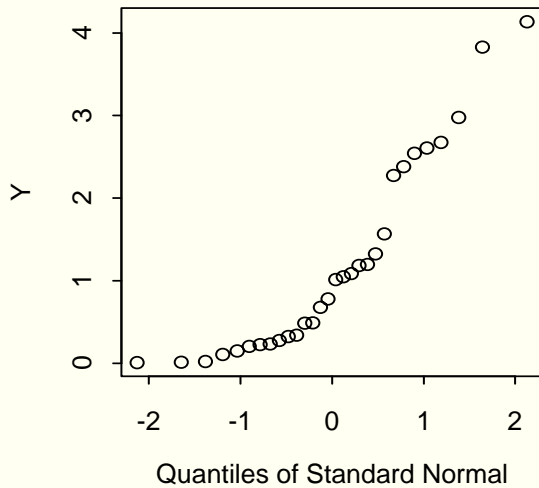




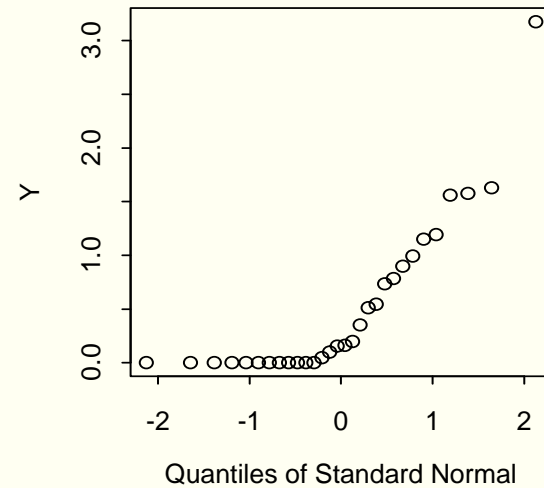
Normal Plot of 30 Observations from a Normal Distribution



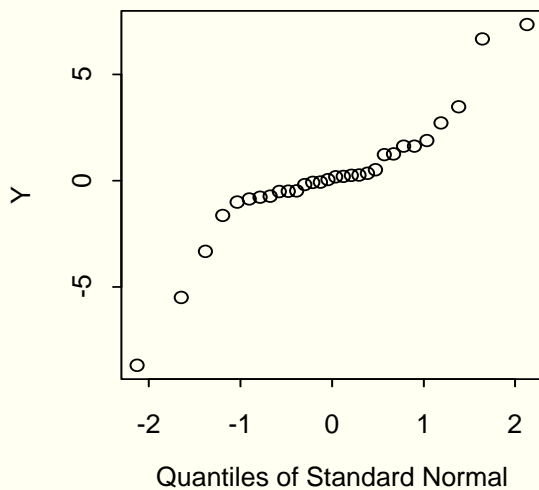
a. a right skewed distribution



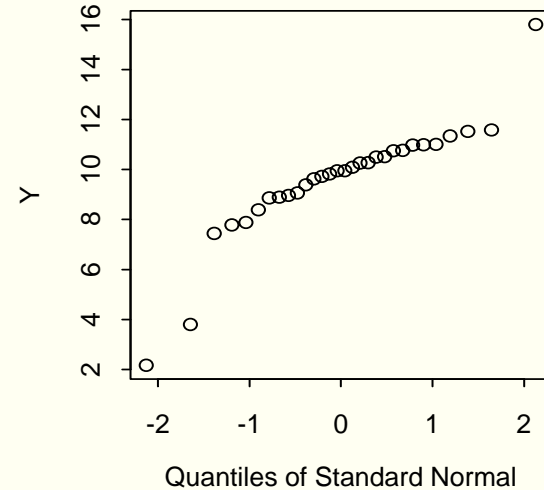
b. a left-truncated Normal distribution



c. a heavy-tailed distribution



d. a contaminated Normal distribution





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 $T_{.01[n]}$

This test assumes that data are normally distributed!!





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

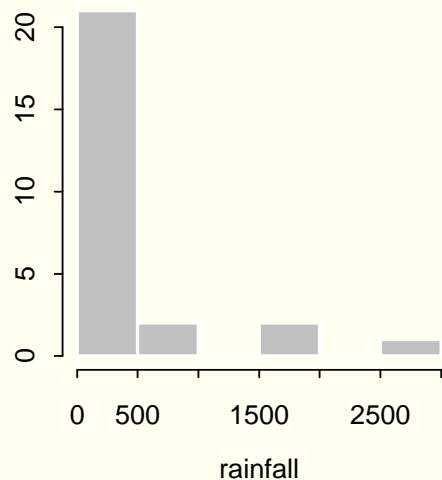
4.1	7.7	17.5	31.4
32.7	40.6	92.4	115.3
118.3	119.0	129.6	198.6
200.7	242.5	255.0	274.7
274.7	302.8	334.1	430.0
489.1	703.4	978.0	
1656.0	1697.8	2745.6	

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

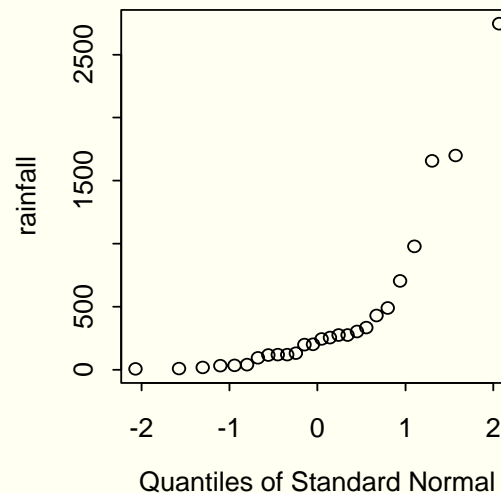
Edwards 2000

Checking Assumptions on Rainfall Data

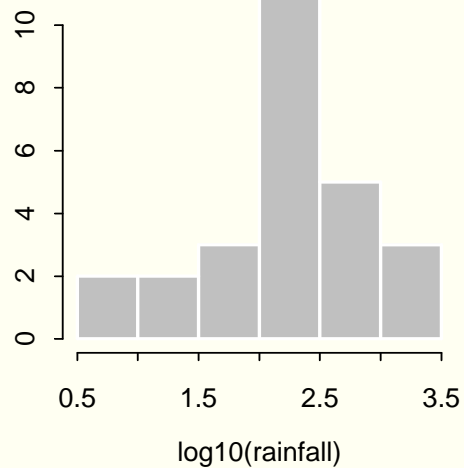
a. rainfall



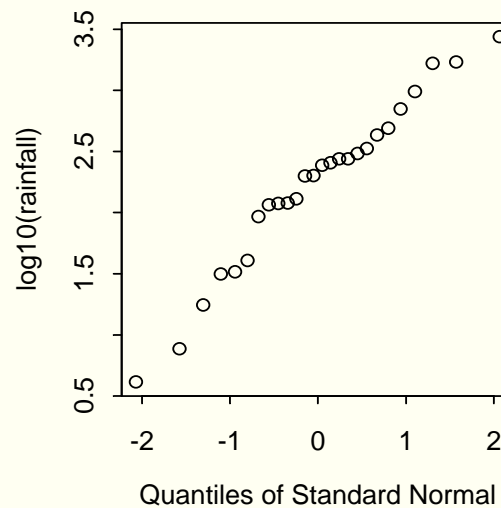
b. Normal plot



c. log10(rainfall)



d. Normal plot





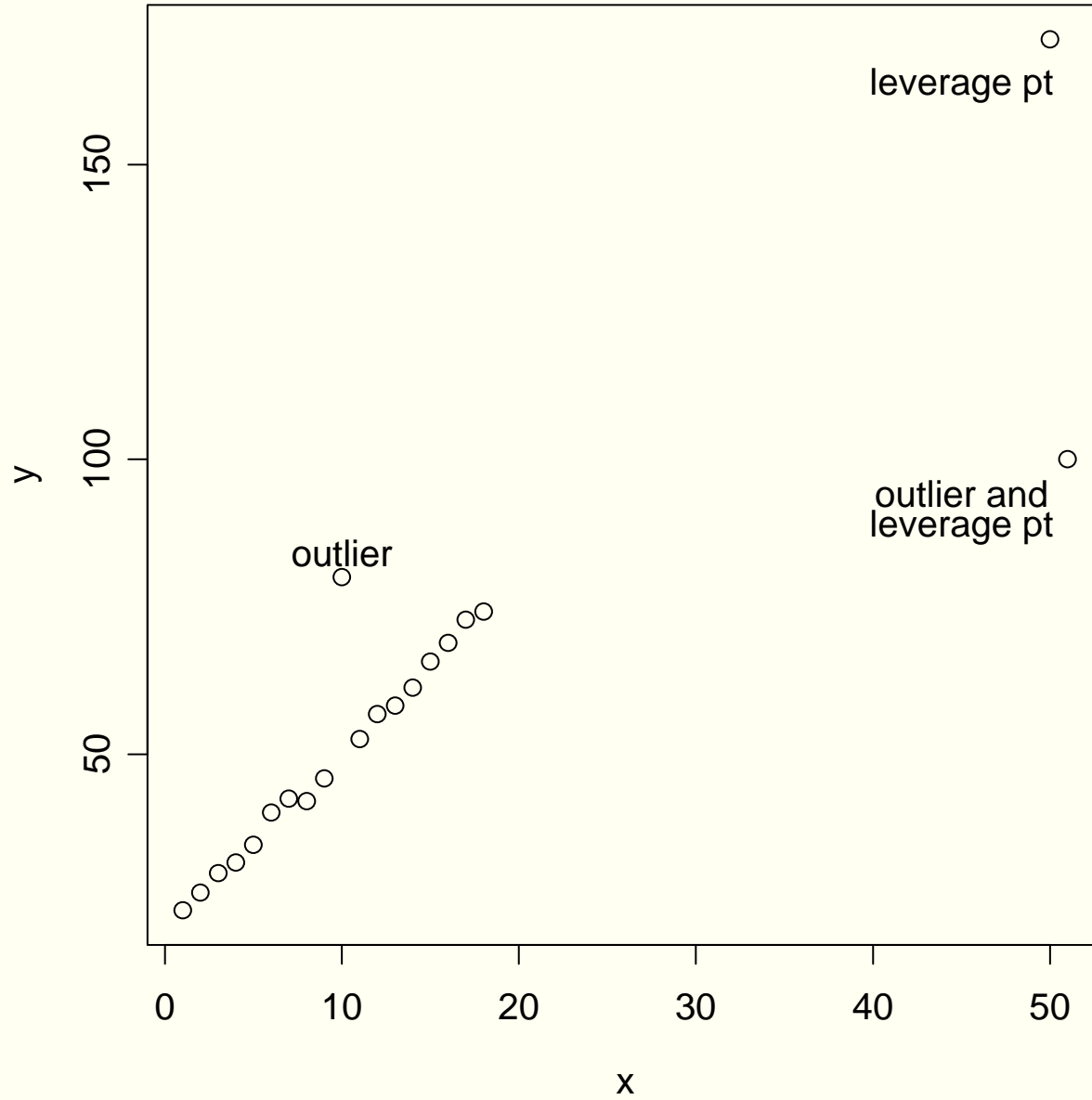
Simple Linear Regression: check for model-based.....

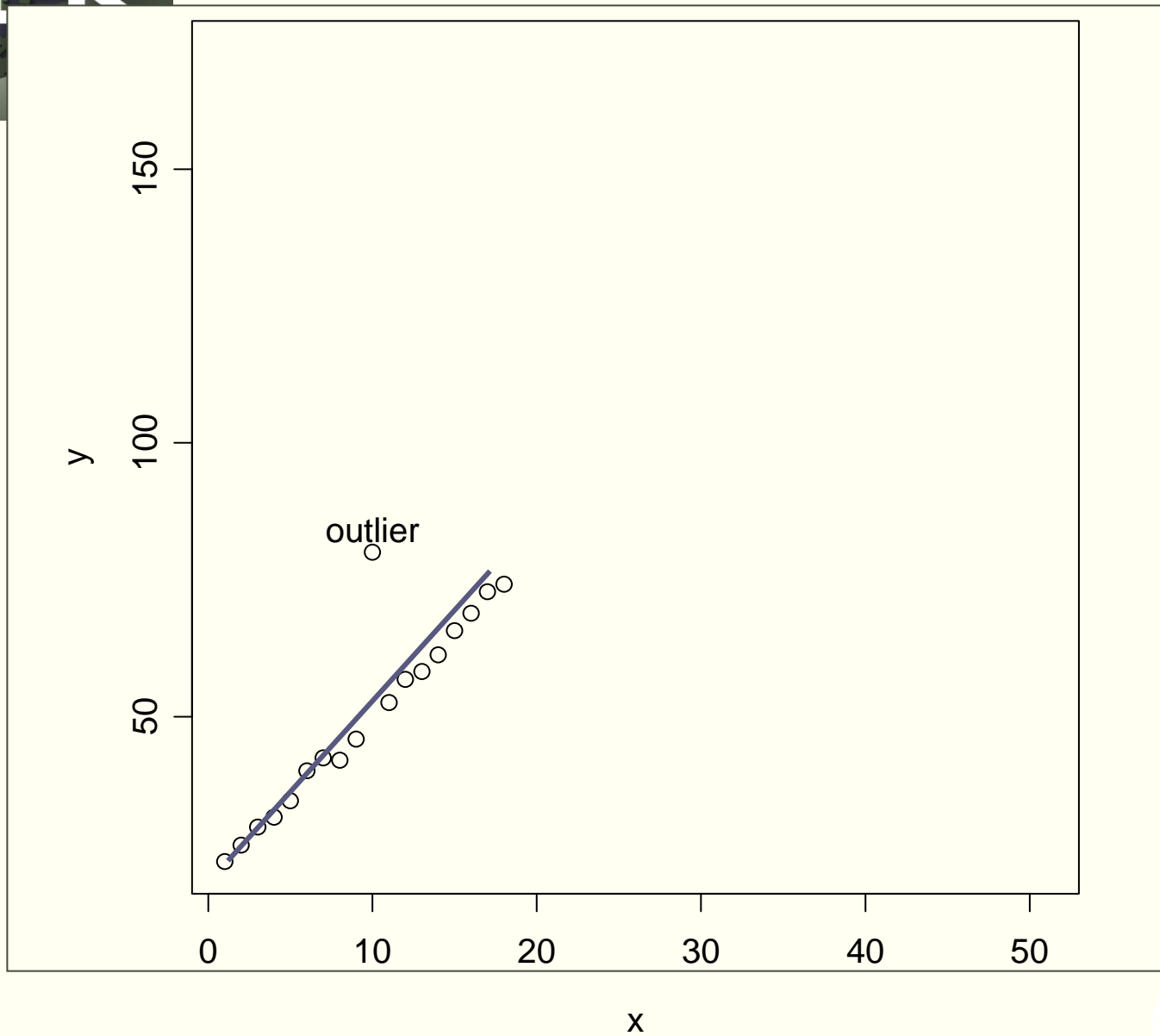
- ❑ Outliers
- ❑ Influential (leverage) points

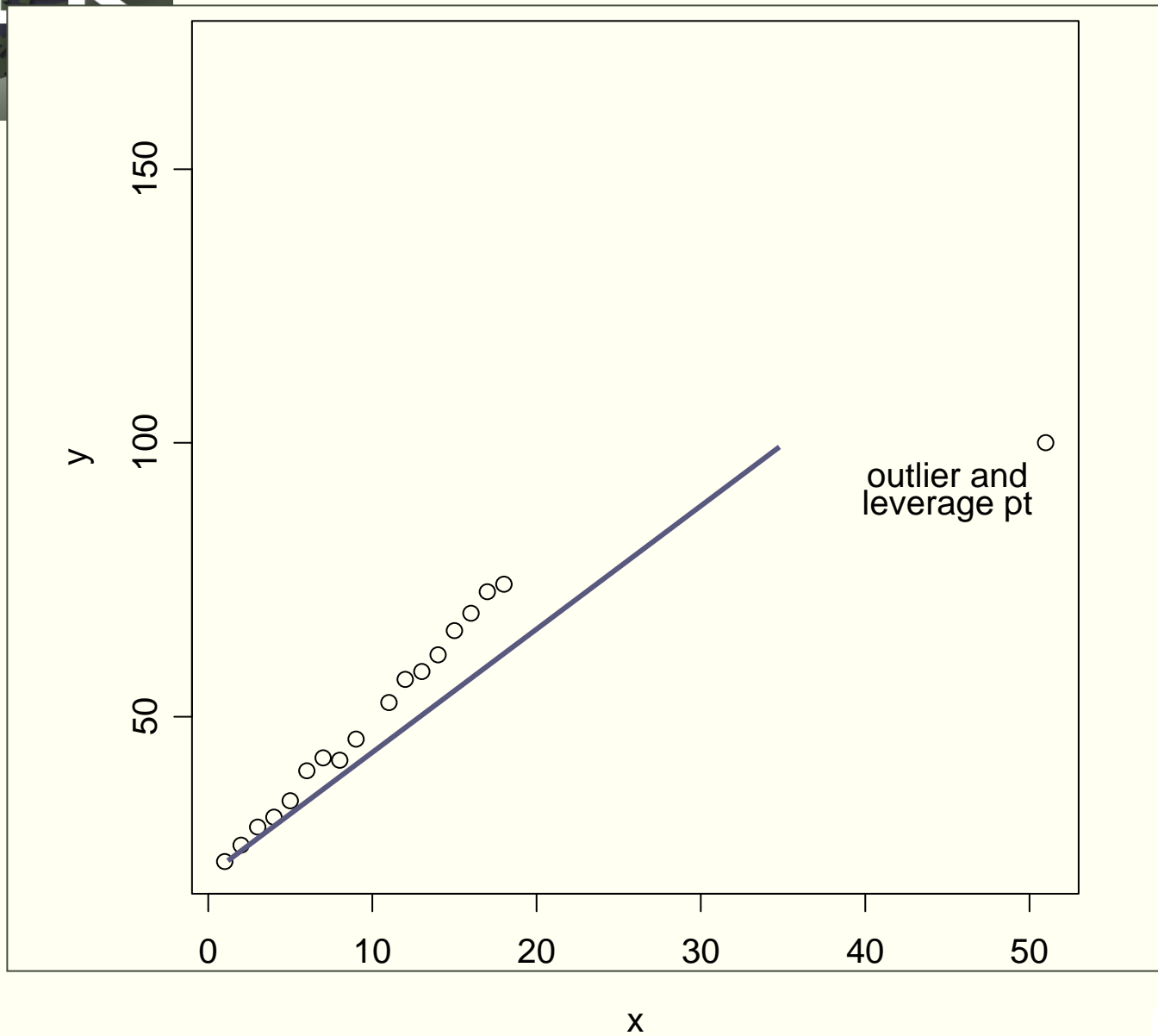


Influential points in simple linear regression:

- ❑ 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.

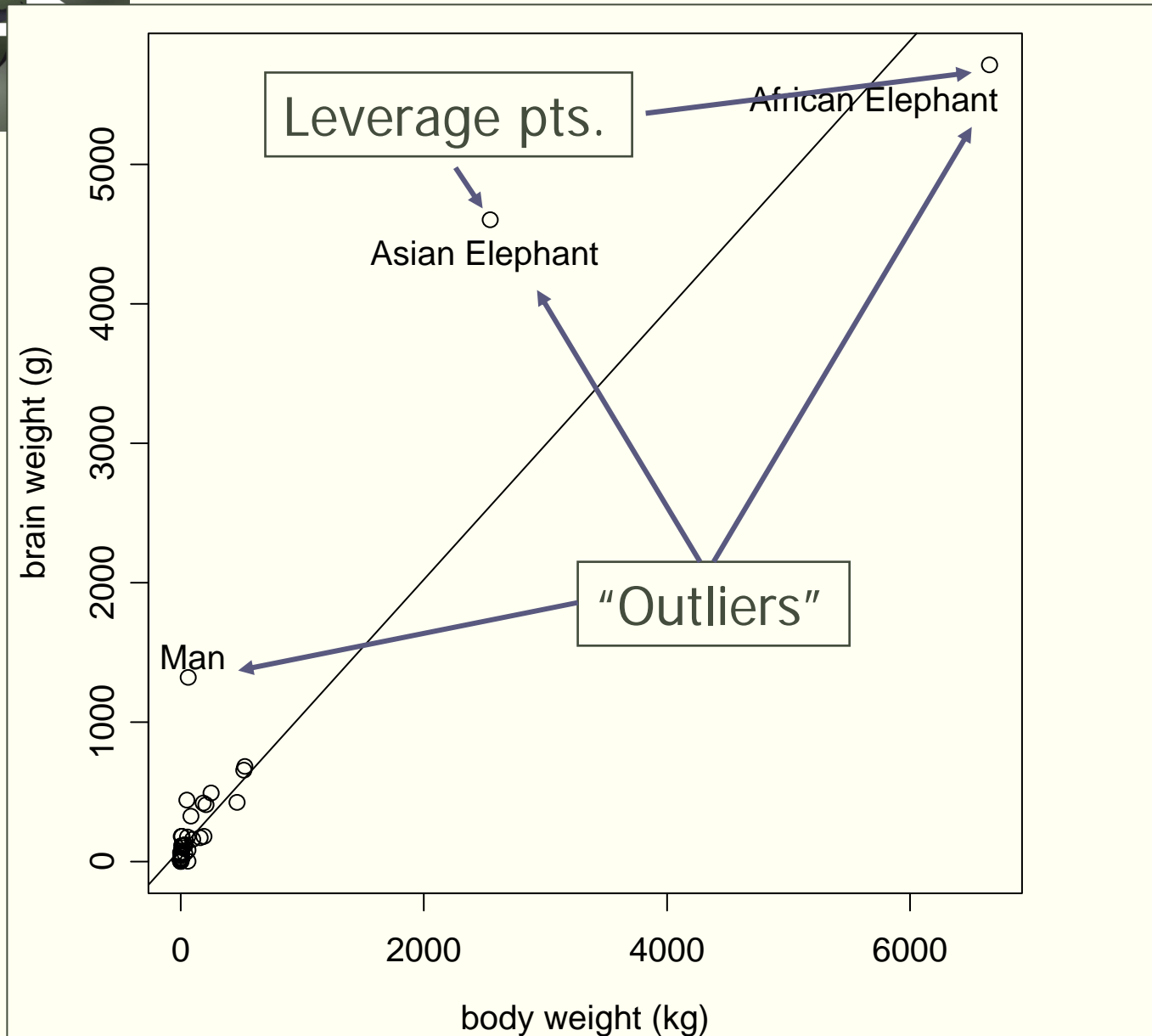


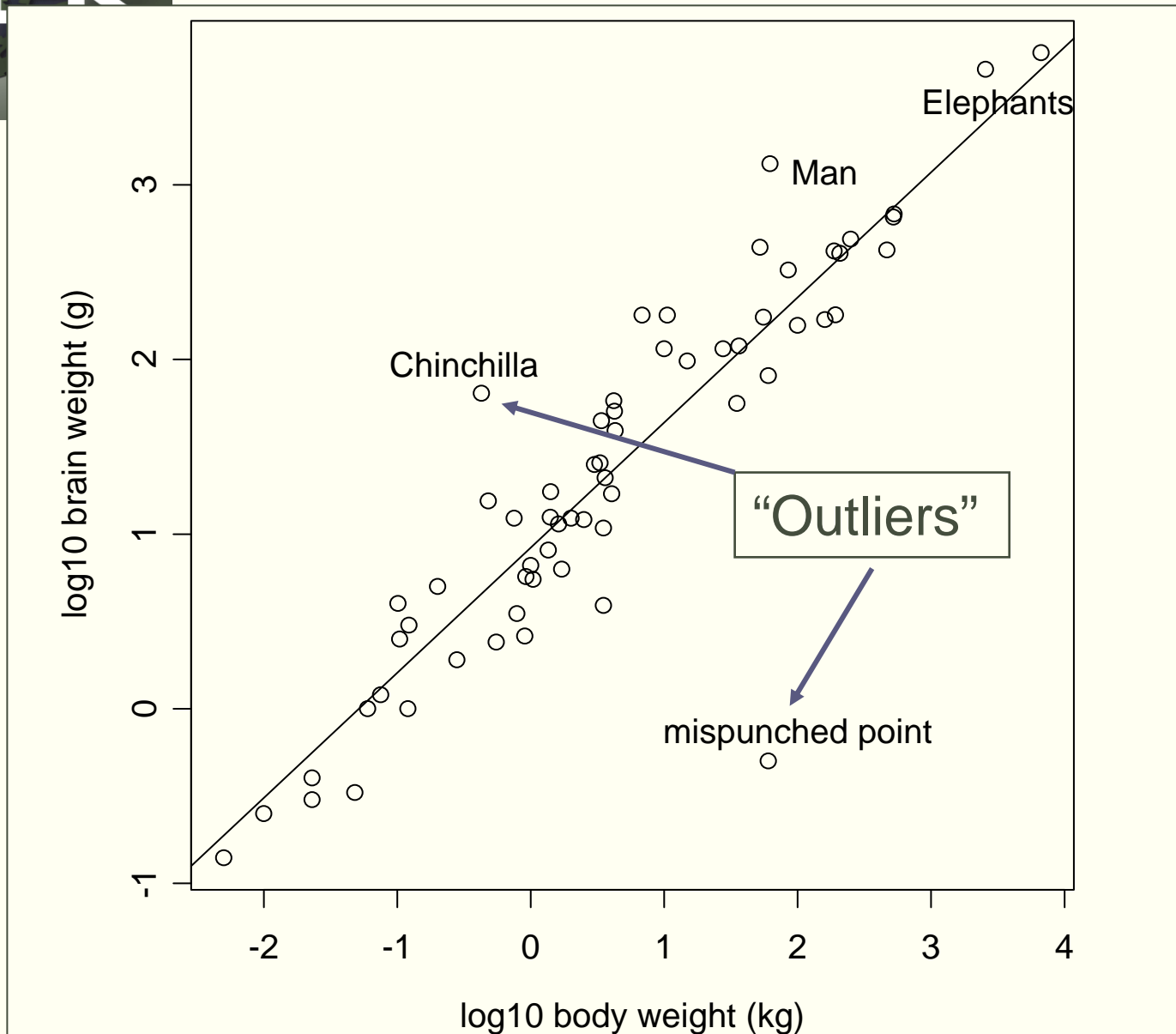






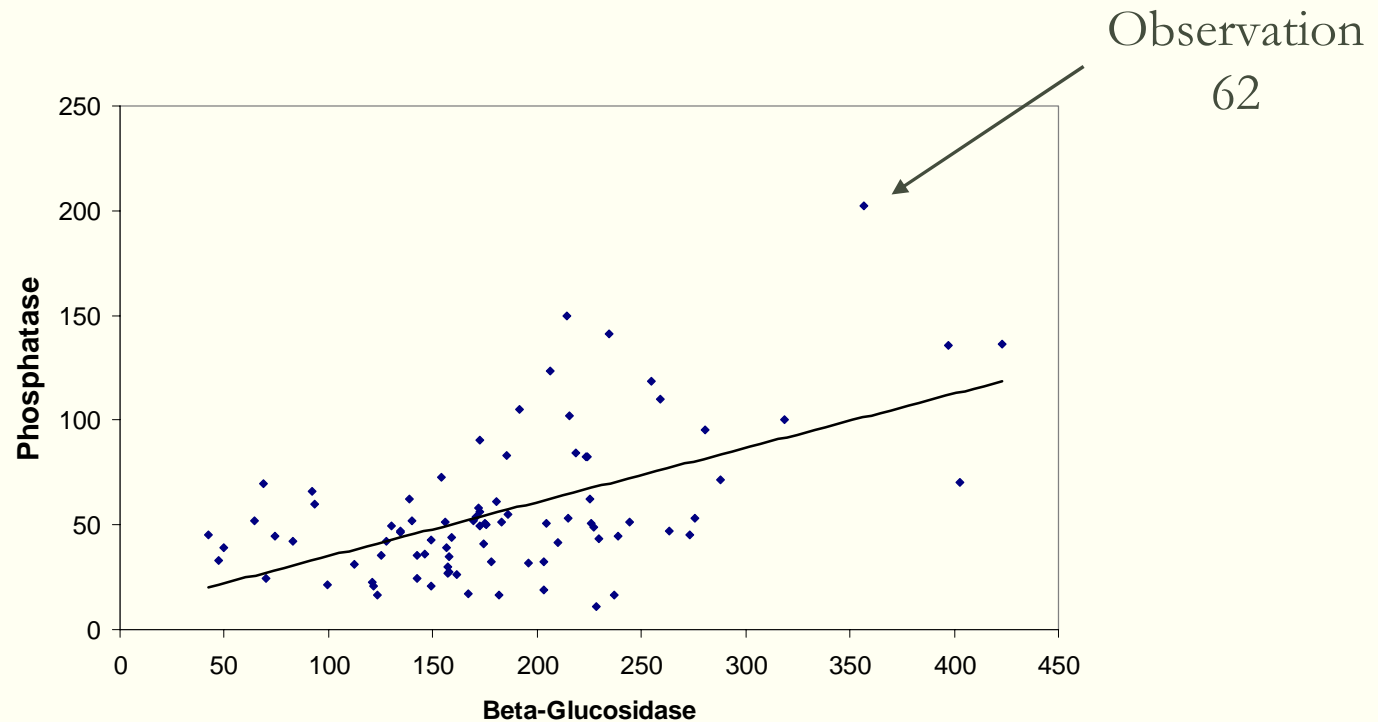
Brain weight vs. body weight, 63 species of terrestrial mammals







Outliers in simple linear regression





Outliers: identify using studentized residuals

- Contamination may exist if

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

$$\alpha = 0.01$$



Simple linear regression:

Outlier identification

The SAS System

14:12 Thursday, October 12, 2000 92

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

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

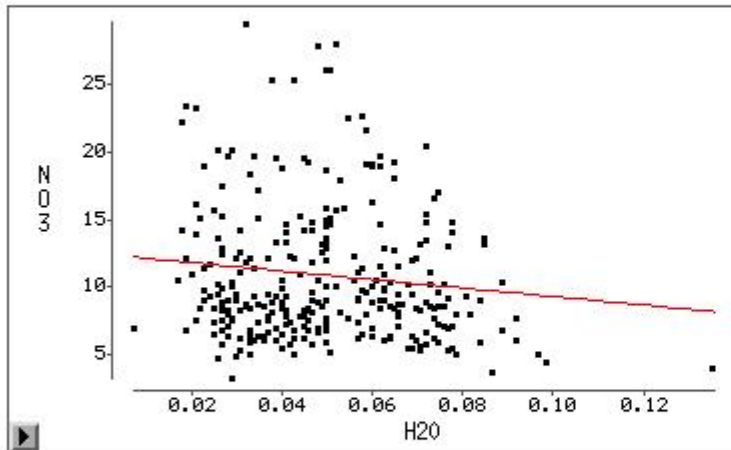
A point is a leverage point if $h_i > 4/n$, where n is the number of points used to fit the regression



Regression with leverage point: Soil nitrate vs. soil moisture

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

► Model Equation
NO3 = 12.3348 - 31.3432 H2O



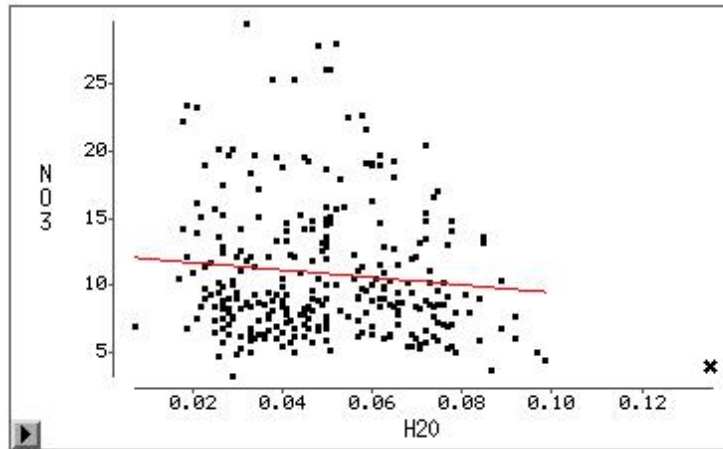
Parametric Regression Fit									
		Model		Error		R-Square	F Stat	Prob > F	
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square				
—	1	1	107.9918	290	24.2399	0.0151	4.4551	0.0357	

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								
Model			Error					
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

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	<u>0.0711</u>	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