Interpretation of Environmental Statistics

Instructors

John Warren
Office of Environmental Information, EPA

Barry Nussbaum
Office of Environmental Information, EPA

2

Schedule

1:00	**Introduction**
1:15	**Where does data come from?**
1:45	**Making the numbers talk**
3:00	***Break***
3:15	**What data distributions look like**
3:45	**Seeing data (class exercise)**
4:15	**Estimation, precision, bias**
4:45	***Conclude***

3

Data Quality

- **Meaningful only when "data quality" relates to intended use of data.**
- **Some data are good ("high quality") for some purposes but are bad ("low quality") for others**

4

Visual Basic Statistics

- Histogram
- Scatterplot
- Boxplot
- Stem-and-leaf
- Data distributions
- Estimation
- Precision and Bias

5

Course Objectives

At the end of this course, you should be able to:

- **Explain why statistics are important and how they can be applied to your projects.**
- **Interpret basic statistics and simple graphs.**

6

Where Does Data Come From?

Data is data, is data, is data

It is very common to think there are two types of data:

- **Data collected for or by you**
- **Somebody else's data**

Are they the same?

From a data analyst's point of view…Yes

But from a project manager's point of view…Maybe

2

How was data collected?

Not statistical sample selection scheme but more to do with the regime under which the data were collected

- **Data collected by or for you:**
  - ***Use of systematic planning***
  - ***Data Quality Objectives***
  - ***Sampling and Analysis Plans***
  - ***QA Project Plans***

- **Someone else's data:**
  - ***Use of systematic planning***
  - ***Performance Criteria (for new data)***
  - ***Acceptance Criteria (for existing data)***

3

Commonality: Systematic Planning

The use of systematic planning is both good common sense and proper scientific practice

- **EPA Order 5360.1 A2 (2000) Section 6.a.(6) "Use of a systematic planning approach to develop acceptance or performance criteria for all work covered by this Order."**

- **The Data Quality Objectives Process is the Agency's recommended approach when data are being used in decision-making or deriving an estimate**

4

Unknown quality of data can affect results

Consider the potential effects of data being analyzed when it is thought that all went well during its collection:

- **Suppose there has been a significant departure from the Sampling and Analysis Plan:**
  - ***samples not taken where they were supposed to be***
  - ***improper mixing of samples in the field***

- **Suppose there has been a serious departure from the QA Project Plan:**
  - ***failure to calibrate equipment correctly***
  - ***samples held longer than the holding time***

5

Types of Data Gathering

- ***Survey Data***
  - **National Agricultural Statistics Survey**
  - **Bureau of Labor Statistics**
- ***Administrative Data***
  - **EPA Discharge Permits**
  - **Toxic Release Inventory**
- ***Surveillance Data***
  - **Passports**
  - **Credit Cards**
- ***Scientific Data***
  - **Systematically planned investigations**
  - **Scientific research experiments**

6

Even the easy datasets are not

- **Duplication of entries:**
  - **Denise Wise**
  - **Denice Wise**
  - **D'Nise Wise**
- **Manipulation of data:**
  - **All less than detection entries suppressed**
  - **Data thought to be outliers discarded**
  - **Several data sets combined to make one large one**

…these can make interpretation difficult

7

The Literary Digest survey 1936

In 1936, Franklin Delano Roosevelt had been President for one term. The magazine, *The Literary Digest*, predicted that Alf Landon would beat him in that year's election by 57% to 43%.

The Digest mailed over 10 million questionnaires to names drawn from lists of automobile and telephone owners, and over 2.3 million people responded - a huge sample.

8

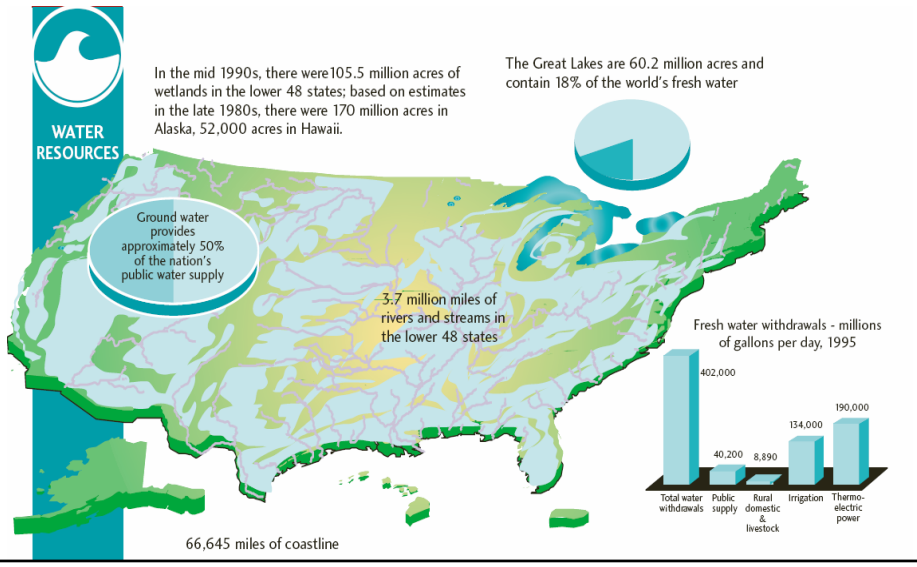
The Literary Digest survey 1936

At the same time, a survey researcher named George Gallup sampled only 50,000 people and predicted that Roosevelt would win. Gallup's prediction was ridiculed as naive. After all, *The Literary Digest* had predicted the winner in every election since 1916, and had based its predictions on the largest response to any poll in history.

Roosevelt won with 62% of the vote. What went wrong?

9

EPA Draft Report on the Environment 2003



But if you google: "U.S. Coastline"

Teachervision: 88,633 miles

EPA 2003 Report: 66,645 miles

CIA Factbook: 12,383 miles

Infoplease: 12,383 miles General Coastline
88,633 miles Tidal Shoreline

Fractal Geometry: Infinite miles because the smaller the measuring device, the more intricate details can be measured.

11

Data must have integrity

"The Government are very keen on amassing statistics. They collect them, add them, raise them to the n^{th} power, take the cube root and prepare wonderful diagrams.

But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn pleases."

Sir Josiah Stamp
Inland Revenue
1896 - 1919

12

Looking at the Data: Making the Numbers Talk

Calibration Problem

- Four Technicians doing same analysis
- X = Controlled variable
- Y = Measured variable

A		B		C		D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

2

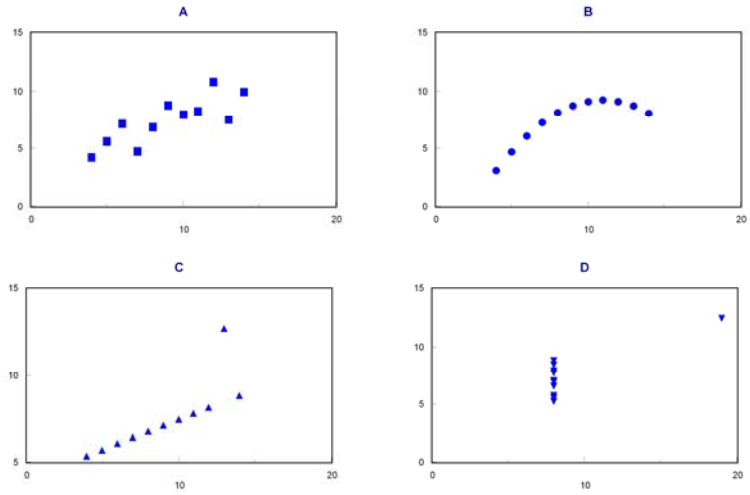
Computer Printout of Summary Statistics

For each technician:

- **n = 11**
- **mean of X_s = 9.0**
- **mean of Y_s = 7.5**
- **equation of regression line: Y = 3+0.5X**
- **s.e. estimate of slope = 0.118**
- ***t* statistic = 4.24 (significant)**
- **sum of squares = 110.00**
- **regression sum of squares = 27.50**
- **residual sum of squares of Y = 13.75**
- **correlation coefficient = 0.82**

Within rounding error, each technician had identical summary statistics . . . or did they?

Actual Results in Graphical Form



Moral: Numerical statistics alone don't tell everything!

What is "Statistics"?

- A *statistic* is a numerical summary
- *Statistics* is the science of data collection, analysis and interpretation
- *Statistical methods* present information to a manager in a useful form

Applying Statistics

- Objective is to learn from the data:
  - Analyze the data to examine features
  - Interpret these features
- Interpretation depends on:
  - Method used to summarize
  - Assumptions about how data came about
  - Context in which inference will be made

Importance of Context

- Interpretation requires context
- The context should include:
  - Description of the big picture (conceptual model)
  - Description of the methods used
  - Definitions, units of measurements, etc.
- Without context, interpretation loses its scientific and realistic basis

Displaying Data

- Histogram
- Scatterplot
- Boxplot
- Stem-and-Leaf

8

Histogram: Initial Data

Cadmium concentration in 100 random samples from the Midway Municipal Site (ppm)

10.46	10.06	11.49	9.47	11.02
11.39	10.91	11.18	8.50	9.31
11.37	9.52	8.62	11.01	9.99
11.39	11.79	9.89	8.66	11.04
9.72	8.81	12.27	9.56	11.40
10.20	10.16	9.49	10.04	8.87
10.77	10.38	10.16	10.29	11.03
9.67	9.71	8.58	8.65	11.25
10.42	10.38	10.86	9.45	9.69
12.46	10.59	9.65	10.24	9.15
9.49	7.47	9.51	9.53	9.44
11.68	8.96	10.60	10.76	10.23
9.74	9.85	11.83	9.10	8.84
7.99	9.64	8.86	10.54	7.94
10.21	11.18	9.66	10.36	9.77
10.08	10.27	9.11	9.69	7.90
11.28	8.36	10.49	9.48	12.99
9.46	9.86	9.11	10.19	9.80
9.56	8.06	7.13	9.76	10.53
8.31	10.66	8.35	9.37	10.40

9

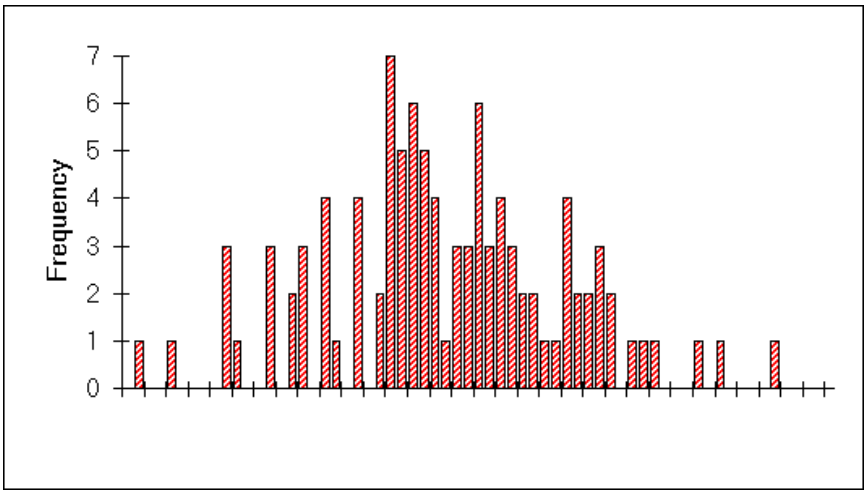
Difficult to comprehend unless we "trim away" some decimals.

Rounded to Nearest Tenth

10.5	10.1	11.5	9.5	11.0
11.4	10.9	11.2	8.5	9.3
11.4	9.5	8.6	11.0	10.0
11.4	11.8	9.9	8.7	11.0
9.8	8.8	12.3	9.6	11.4
10.2	10.2	9.5	10.0	8.9
10.8	10.4	10.2	10.3	11.0
9.7	9.7	8.6	8.7	11.3
10.4	10.4	10.9	9.5	9.7
12.5	10.6	9.6	10.2	9.3
9.5	7.5	9.5	9.5	9.4
11.7	9.0	10.6	10.8	10.2
9.7	9.9	11.8	9.1	8.8
8.0	9.6	8.9	10.5	7.9
10.2	11.2	9.7	10.4	9.8
10.2	10.3	9.1	9.7	7.9
11.3	8.4	10.5	9.5	13.0
9.5	9.9	9.1	10.2	9.8
9.6	8.1	7.1	9.8	10.5
8.3	10.7	8.4	9.4	10.4

Getting any clearer?

Histogram of Data Rounded to Nearest Tenth

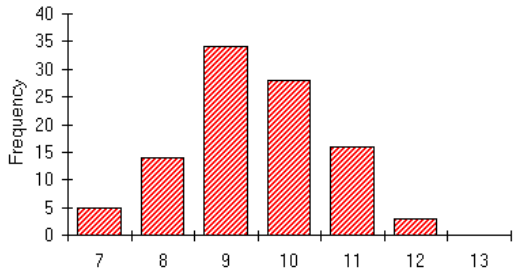


Rounded to Nearest Whole Number

10	10	11	9	11
11	11	11	9	9
11	10	9	11	10
11	12	10	9	11
10	9	12	10	11
10	10	9	10	9
11	10	10	10	11
10	10	9	9	11
10	10	11	9	10
12	11	10	10	9
9	7	10	10	9
12	9	11	11	10
10	10	12	9	9
8	10	9	11	8
10	11	10	10	10
10	10	9	10	8
11	8	10	9	13
9	10	9	10	10
10	8	7	10	11
8	11	8	9	10

Better?

Histogram of Data Rounded to the Nearest Whole Number

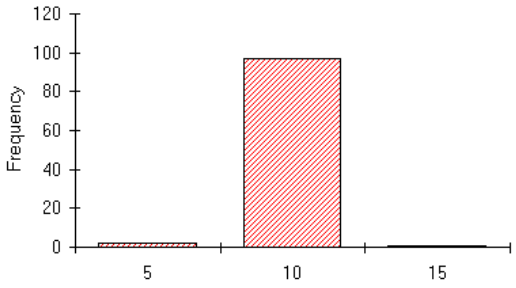


Rounded to Nearest 5

10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	5	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	10
10	10	10	10	15
10	10	10	10	10
10	10	5	10	10
10	10	10	10	10

Even better?

Histogram of Data Rounded to Nearest 5



Conclusions for Histograms

- Different groupings (box or bin sizes) can lead to different conclusions.
  - Try several bin-sizes and "feel out" the data
- Different scales can change conclusions.
  - Make the bins equal in width and contiguous to each other.
- Histograms are used to assess fit for different theoretical distributions, for example, Normal or Lognormal
- Histograms have good visual impact and are useful in expressing probabilities and error distributions

16

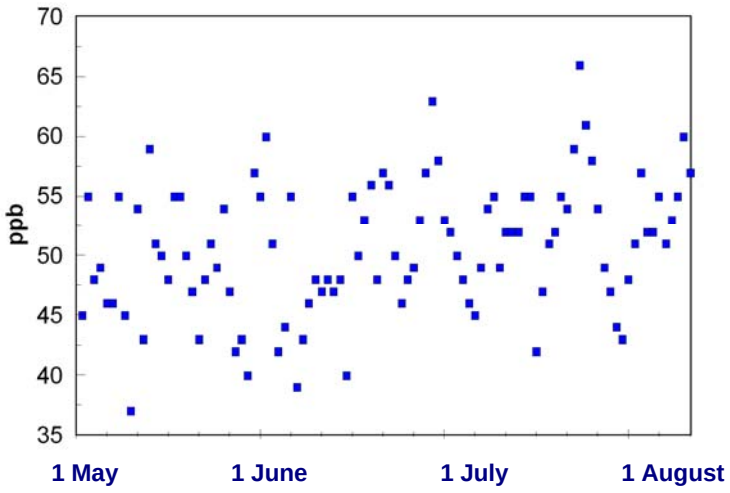
Scatterplots: Comstock Air Monitoring Station

Ozone ppb

May 1	45	48	48	52	59
	55	51	47	Jul 1	50	66
	48	49	48		48	61
	49	54	40		46	58
	46	47	55		45	54
	46	42	50		49	49
	55	43	53		54	47
	45	40	56		55	44
	37	57	48		49	43
	54	55	57		52	48
	43	60	56		52	51
	59	Jun 1	51	50		52	57
	51	42	46		55	Aug 1	52
	50	44	48		55	52
	48	55	49		42	55
	55	39	53		47	51
	55	43	57		51	53
	50	46	63		52	55
	47	48	58		55	60
	43	47	53		54	57

***100 consecutive days in Summer 2005**

Comstock Ozone Readings



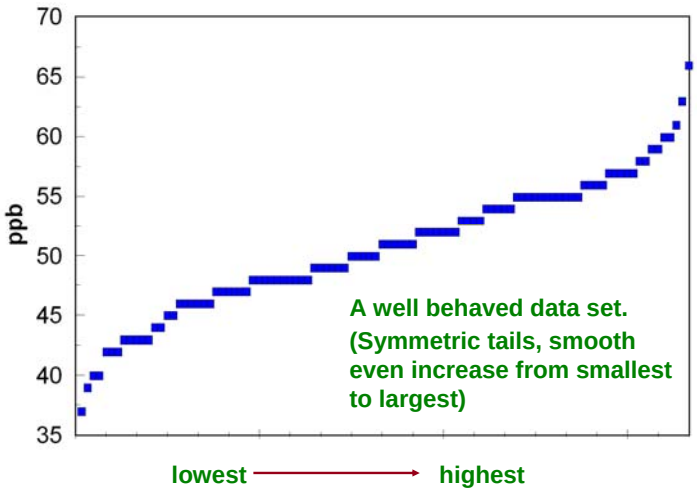
Rank Ordered Data: Smallest to Largest

Comstock Air Monitoring Station Ozone ppb

37, 39, 40, 40, 42, 42, 42, 43, 43, 43, 43, 43, 44, 44, 45, 45, 46, 46, 46, 46, 46, 46, 47, 47, 47, 47, 47, 47, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 49, 49, 49, 49, 49, 49, 50, 50, 50, 50, 50, 51, 51, 51, 51, 51, 51, 52, 52, 52, 52, 52, 52, 52, 53, 53, 53, 53, 54, 54, 54, 54, 54, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 56, 56, 56, 56, 57, 57, 57, 57, 57, 58, 58, 59, 59, 60, 60, 61, 63, 66

- **100 Observations**
- **Minimum = 37**
- **Maximum = 66**
- **Range = 29**

Comstock Ozone



Lode Air Monitoring Station

May 1	37		48		65		45		47
	48		58		44	July 1	45		46
	48		64		45		43		52
	47		54		47		48		46
	46		48		47		48		44
	47		48		44		45		48
	48		45		43		46		47
	49		43		47		46		47
	48		47		42		47		45
	48		48		43		47		45
	54		44		43		47		48
	45	June 1	61		43		49		46
	45		56		45		56	August 1	45
	45		64		44		63		50
	47		45		45		50		45
	47		45		44		47		45
	46		52		47		45		50
	44		63		47		48		46
	45		66		45		44		44
	48		59		44		45		45

***100 Consecutive Days in Summer 2005**

Rank Ordered Data: Smallest to Largest

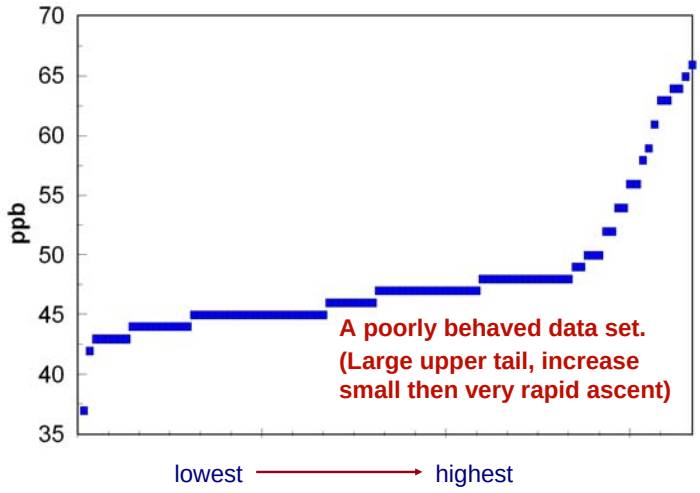
Lode Air Monitoring Station Ozone ppb

37, 42, 43, 43, 43, 43, 43, 43, 44, 44, 44, 44, 44, 44, 44, 44, 44, 44, 45, 45,
45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45,
46, 46, 46, 46, 46, 46, 46, 46, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47,
47, 47, 47, 47, 47, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48,
49, 49, 50, 50, 50, 52, 52, 54, 54, 56, 56, 58, 59, 61, 63, 63, 64, 64, 65, 66

- 100 Observations
- Minimum = 37
- Maximum = 66
- Range = 29

22

Lode Ozone



Conclusions for Scatterplots

- Simple graph over time can show potential trends and relationships clearly
- Potential anomalies are easy to identify
- Plotting the ordered data should give a smooth sinuous display to indicate approximate Normal distribution:
  - Clearly defined roughly equal tails
  - Steady rise from smallest to largest
  - Fairly large plateau in the center
- Unknown distribution indicated by:
  - Unequal tails
  - Abrupt changes in the rise of values
  - Poorly defined or unequal tails

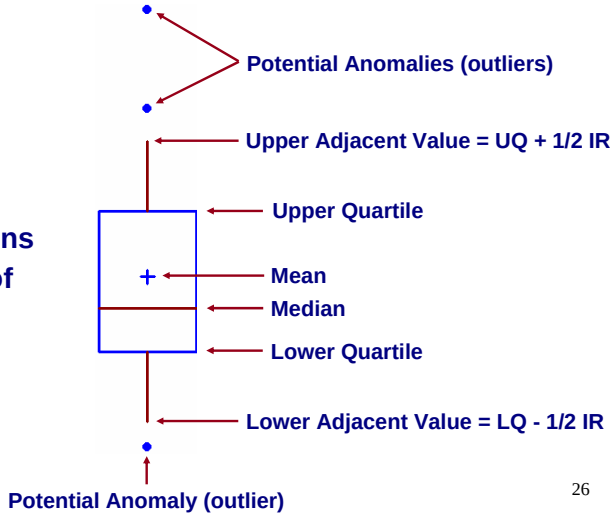
24

Boxplots

- Puts into visual form data percentiles using:
- Median: the value such that half of all values are larger, half are smaller = 50th percentile
- Upper quartile (UQ): the value such that 25% of all values are larger =75th percentile
- Lower quartile(LQ): the value such that only 25% of all values are smaller = 25th percentile
- Interquartile Range (IR): IR = UQ - LQ

A Boxplot Actually Uses a Box

The "Box" contains the central 50% of the data.



26

Comstock Ozone (Well Behaved Data)

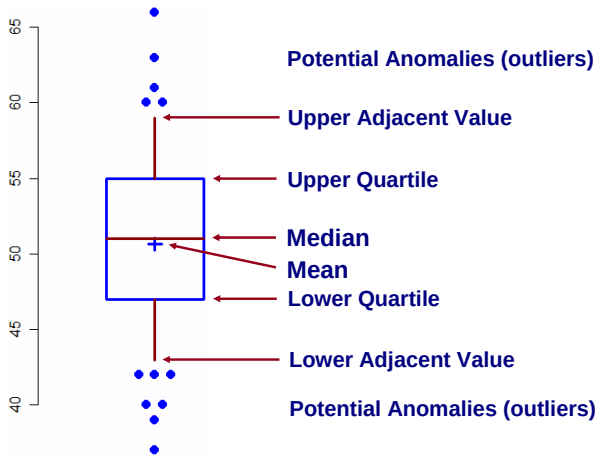
Comstock Air Monitoring Station Ozone ppb

37, 39, 40, 40, 42, 42, 42, 43, 43, 43, 43, 43, 44, 44, 45, 45, 46, 46, 46, 46, 46, 46, 47, 47, 47, 47, 47, 47, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 49, 49, 49, 49, 49, 49, 50, 50, 50, 50, 50, 51, 51, 51, 51, 51, 51, 52, 52, 52, 52, 52, 52, 52, 53, 53, 53, 53, 54, 54, 54, 54, 54, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 55, 56, 56, 56, 56, 57, 57, 57, 57, 57, 58, 58, 59, 59, 60, 60, 61, 63, 66

LQ is between 25th & 26th data points = 47
Median is between 50th & 51st data points = 51
UQ is between 75th & 76th data points = 55

IR = UQ - LQ = 55 - 47 = 8

Comstock Ozone - Well Behaved Data



28

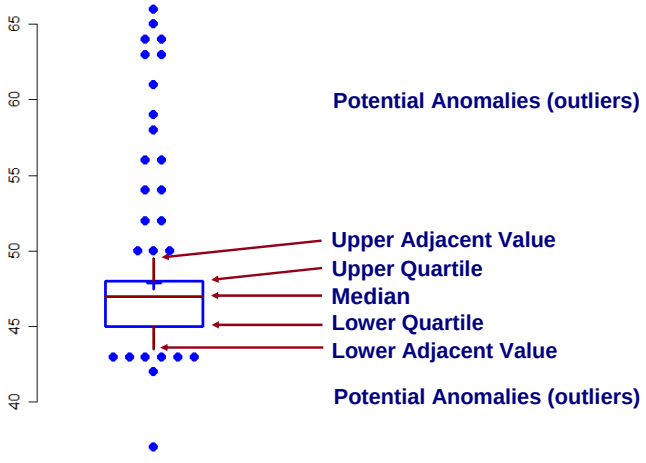
Lode Ozone - Poorly Behaved Data

37, 42, 43, 43, 43, 43, 43, 43, 44, 44, 44, 44, 44, 44, 44, 44, 44, 44, 45, 45,
45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45, 45,
46, 46, 46, 46, 46, 46, 46, 46, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47,
47, 47, 47, 47, 47, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48,
49, 49, 50, 50, 50, 52, 52, 54, 54, 56, 56, 58, 59, 61, 63, 63, 64, 64, 65, 66

LQ is between 25th & 26th data points = 45
Median is between 50th & 51st data points = 47
UQ is between 75th & 76th data points = 48

IR = UQ - LQ = 48 - 45 = 3

Lode Ozone (Poorly Behaved Data)



30

Conclusions for Boxplots

- Few potential anomalies should be observed
- Many potential anomalies makes identification of distribution difficult
- "Eliminating" the central mass of data by replacing it with a "box" facilitates an understanding of the data distribution

Stem and Leaf Displays

Major value = stem
Minor value = leaf

Comstock Ozone June Data

Stem	Leaf
60	3
55	5 5 5 6 7 7 8
50	0 0 1 2 3 3 3
45	5 6 7 7 8 8 8 8 8 9
40	0 2 3 4
35	9

Stem and Leaf Comparison

Lode Ozone June		Comstock Ozone June
6 5 4 3 1	60	3
9 6	55	5 5 5 6 7 7 8
2	50	0 0 1 2 3 3 3
7 7 7 7 7 5 5 5 5 5 5	45	5 6 7 7 8 8 8 8 8 9
4 4 4 4 4 4 3 3 3 3 2	40	0 2 3 4
	35	9

The imbalance between the data sets shows they are probably from different distributions

Other Graphical Methods

- **Frequency graphs**
- **Parallel coordinate plots**
- **Empirical Quantile-Quantile plots**
- **Normal plots**

Bizarre: Unaccountable Gaps

pH of effluent discharge reported fom Hoffner Plant

1 pm	2 pm	3 pm	5 pm	6 pm	7 pm	9 pm	10 pm
4.6	4.5	4.1	4.5	4.4	4.0	4.8	4.7

- Why are 4 pm and 8 pm missing?
- If deliberate, could they be extreme values?
- If accidental, can we impute a value?

Bizarre: Preponderance of Values

Opacity reading from the Churchman smoke stack
(0 = clear, 1.0 = opaque)

0.5, 0.1, 0.5, 0.2, 0.5, 0.8, 1.0, 1.0,
0.5, 0.5, 0.1, 0.5, 1.0, 0.8, 0.5, 0.5,
0.5, 0.5, 0.1, 0.5, 1.0, 0.5, 0.8, 0.5,
1.0, 0.5, 0.5, 0.5, 0.1, 0.5, 0.5, 0.8.

- Where are values like 0.6 or 0.4?
- Why so many 0.5 values?
- Can 0.1 really be distinguished from 0.2?

Bizarre: Too Many Decimals

Arsenic in soil (ppm)

March - April: 12.0, 13.0, 12.0, 12.0,
May - June: 12.5632, 13.1129, 13.0076, 12.9665

- How much accuracy? Four decimal places or one?
- Different methods or different analysts?
- Data rounded off prior to recording?
- Spurious decimals to give illusion of precision?

Conclusions

"A picture is worth a _ _ _ _ _ _ _ words"

What Data Distributions Look Like

Raw data must be grouped in order to see patterns

- Data in numerical form are difficult to visualize directly
- Identification of patterns in data help us use the information from the sample in an efficient manner
- The most obvious pattern in everyday data is the way in which the data values group together
  - Clustering of values round an average
  - Predominance of very small values
  - Occurrence of a few high values with mostly low values

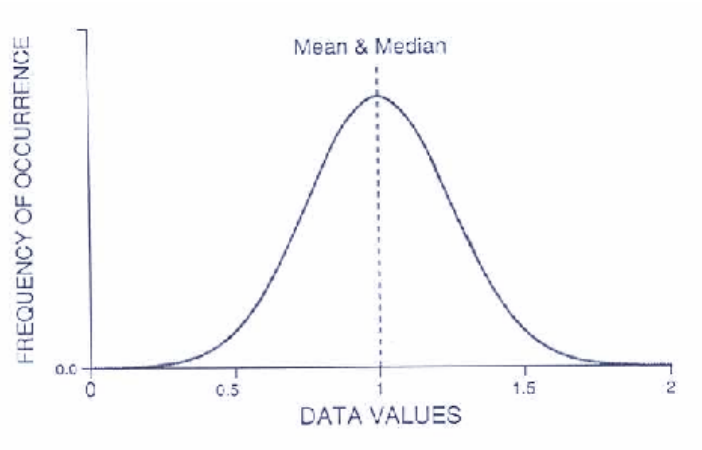
Environmental data comes as two types

- **Environmental data comes in one of two forms; continuous (measurement type data) and discrete (counting type data)**
- **Continuous data requires the calculus of integration, discrete requires the use of summation techniques**
- **Continuous data can be mistaken for discrete data due to the fact we must round off some of the decimal places**
  - **e.g. A measurement of 8.23411527873458734590963… is recorded as 8.23 or even as just 8 depending on what is the final disposition of the values**

The most commonly encountered types

- **Continuous:**
  - **Normal**
    - ***bell-shaped curve***
  - **Lognormal**
    - ***Logarithm of the values are Normal***
- **Discrete:**
  - **Binomial**
    - ***Everything is one thing or another***
  - **Poisson**
    - ***Extreme case of a Binomial***

The Normal Distribution



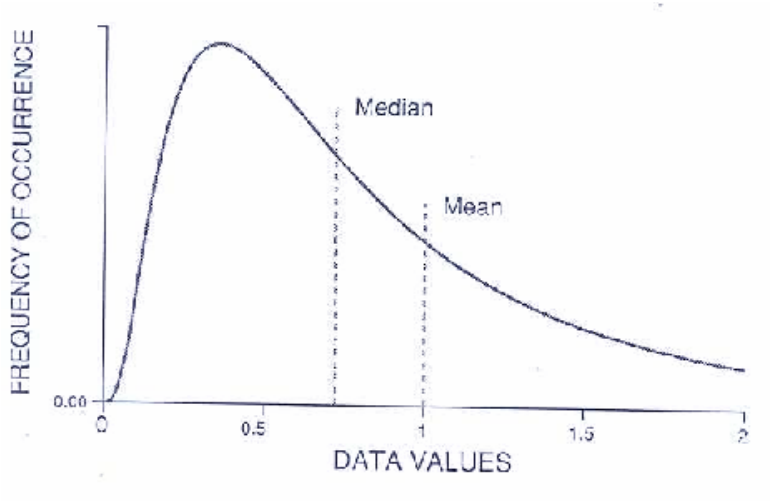
Normal data

- **Normal data are continuous data**
- **Often "taken for granted" with data sets (see the Central Limit Theorem later)**
- **Measurements of arsenic at Royal Smelting (in ppm)**

  1.251, 1.423, 1.323, 0.789, 0.429,
  3.033, 2.131, 2.055, 1.001, 1.488

- **Note roughly symmetric around a mean of roughly 1.5**
- **No really extreme values**

The Lognormal Distribution



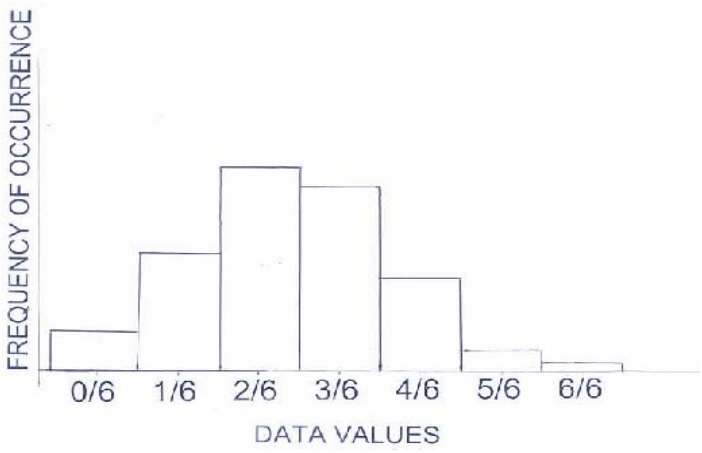
Lognormal data

- **Lognormal data are continuous data**
- **Not always obvious when dealing with small data sets**
- **Measurements of selenium in Melrose Lake (in ppb):**

3.16, 4.15, 3.75, 2.20, 1.53,
20.76, 8.42, 7.81, 2.72, 4.43

- **Note the suspicious high value**
- **Is this an outlier or genuine value?**
- **If data are transformed by taking logarithms, does approximate normality result?**

The Binomial Distribution



Binomial data

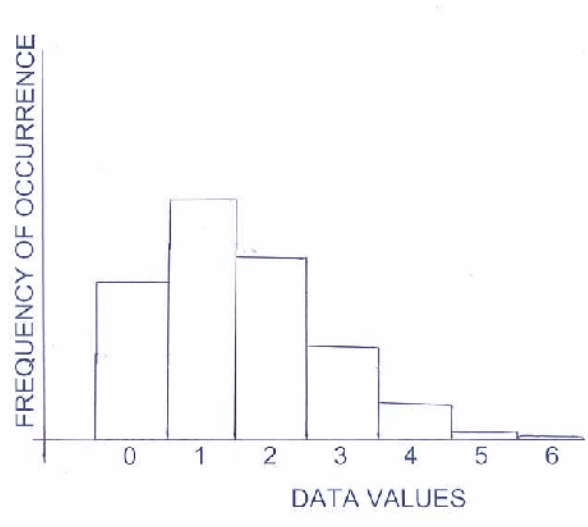
- **Binomial data are discrete data**
- **Need a fixed (albeit unknown) probability of an occurrence**
- **Need a fixed number of possibilities**
- **Infected cells (out of 6 possibilities) in a biometer used 8 times**

  4/6, 0/6, 3/6, 3/6

  6/6, 3/6, 4/6, 2/6
- **Total number of infections = 25 out of 48 cells**

  assuming independence, chance is 25/48 i.e. roughly 0.5
- **Each time the biometer was used, any out of a fixed number of 6 cells could be infected**

The Poisson Distribution



Poisson data

- **Poisson data are discrete data**
- **Usually arises in the investigation of rare events**
- **Number of non-compliance-weeks per year at Butte River:**
  0, 0, 1, 1, 0, 3, 0, 0

- **Assumes a non-compliance-week is a rare event**
- **Total number of non-compliance-weeks = 5 out of 416 weeks**
  assuming independence, chance is 5/416 i.e roughly 0.12
- **Can be used to approximate a Binomial**

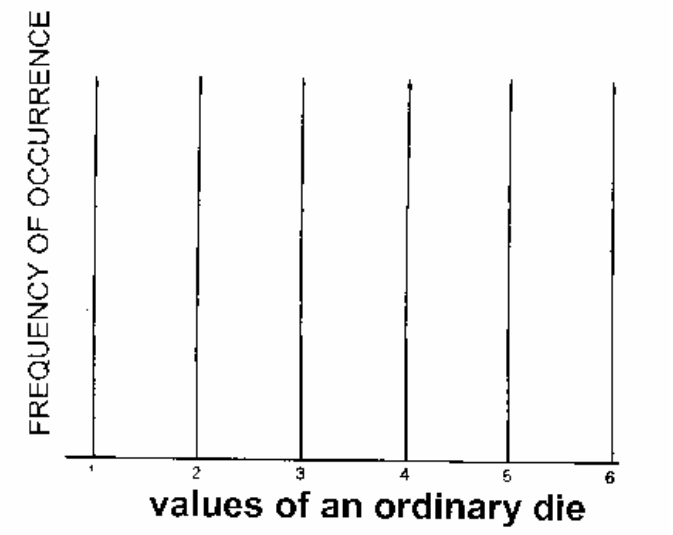
Identification of "outlier" depends on the assumed distribution

- Recall the Melrose Lake lognormal data:
  3.16, 4.15, 3.75, 2.20, 1.53,
  20.76, 8.42, 7.81, 2.72, 4.43
- 20.76 was not an outlier as this was lognormal data. However, suppose it was assumed that it was normal data, what then?
- Statistical outlier tests are easy to apply but all of them assume that the distribution of all the data other than the suspected outlier is known. Is this true in practice?
- Could the Melrose Lake data be roughly normal when the 20.76 value is omitted? Very difficult for small data sets.

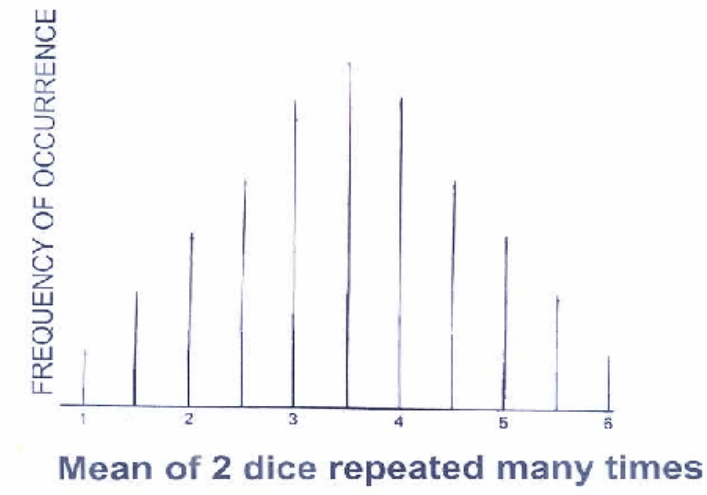
Why is everybody so concerned with the mean of a sample?

- The power of the Central Limit Theorem (CLT)
- In everyday words, the CLT says: As the sample size becomes large, the mean of that random sample will behave as if it came from a normal even though the original data does not.
- In practical terms:
  Take as large a sample as you possibly can
  find the average and hope the number in the
  sample is large enough for the CLT to hold
- Why? Normal data are nice and easy to deal with!

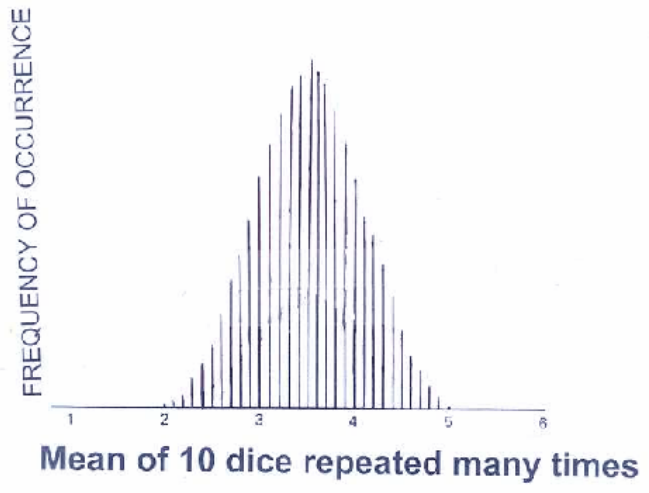
CLT demonstrated when the original data was from a rectangular distribution



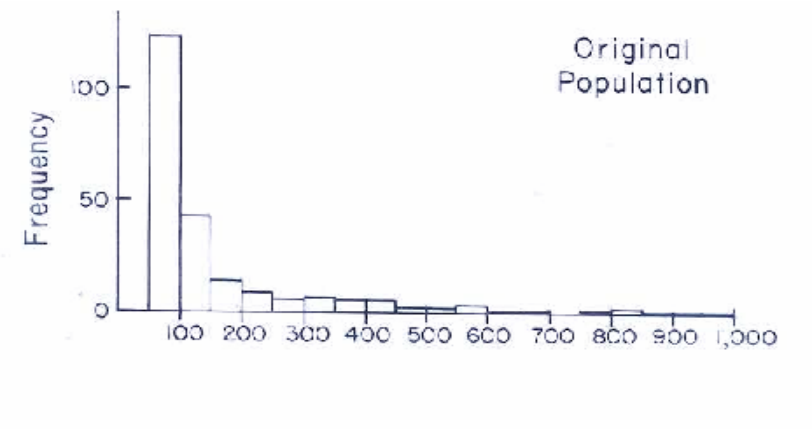
CLT demonstrated when the original data was from a rectangular distribution



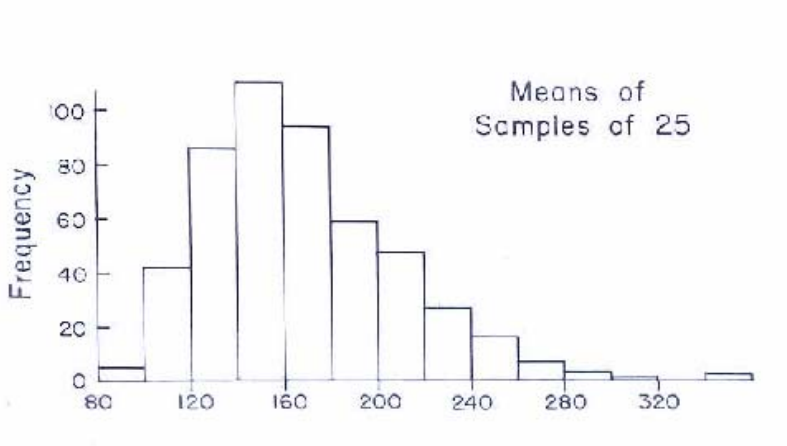
CLT demonstrated when the original data was from a rectangular distribution



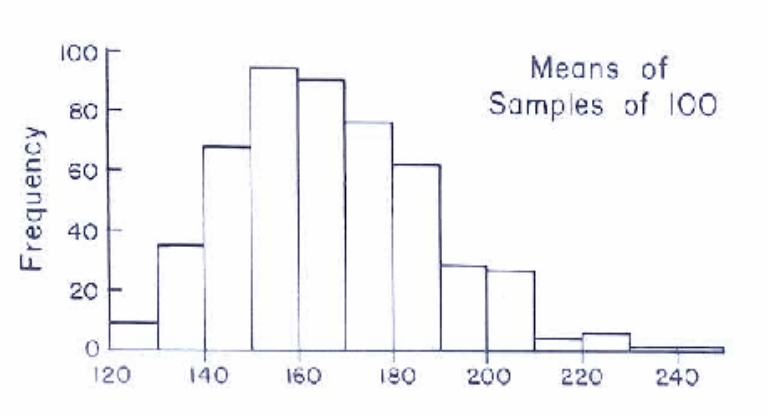
CLT demonstrated when the original distribution is unknown



CLT demonstrated when the original distribution is unknown



CLT demonstrated when the original distribution is unknown



Distributions and statistical tests

Does it really matter if the distribution is unknown?

YES:
- **+ The appropriate statistical test will be more powerful**
- **+ Enables quantifiable estimates of decision errors**
- **+ Better understanding of the problem**

NO:
- **- Need nonparametric statistical tests (not as powerful)**
- **- Need large sample sizes (CLT)**
- **- Not so great understanding of the problem**

Good news

- **Most of the standard statistical tests designed with the assumption of normality are actually quite robust (strong) to departures from normality**
- **Approximate normality or even the assurance of approximate symmetry about the average is sufficient for the test to work quite well and the assumed "level of significance" and associated "statistical power" are denigrated only slightly**
- **If large sample sizes are available the CLT can apply and also nonparametric tests can out perform standard statistical tests**

Seeing Data (Class Exercise)

There are three different scenarios

- **Identify the probable distribution of data for each scenario**
- **Use only common sense, a pencil, and paper**
- **Complicated calculations are not necessary**
- **Potential choices for each scenario:**

  - Approximate normal distribution
  - Approximate lognormal distribution
  - Possible Binomial distribution
  - Possible Poisson distribution
  - Approximate normal with outlier
  - Approximate lognormal with outlier
  - Possible Poisson with outlier
  - Some unknown distribution

Estimation, Precision, Bias, Types of Intervals

Estimation

- **Estimation is the process of extrapolating information from a sample to a much larger universe**
- **"Sample" (for an analytical chemist) is that actual physical specimen which will be chemically analyzed. "Size of a sample" refers to physical dimensions.**
- **"Sample" (for a statistician) is that whole group of individual physical specimens. "Size of a sample" refers to how many is in that group.**

Point and Interval Estimation

- **Point Estimate: A single summary value derived from a sample**
  - **Mean**
  - **Median**
  - **Variance or Standard deviation**
- **Interval Estimate: A point estimate combined with a probability statement of containing the true (population) value**
  - **Confidence interval**
  - **Prediction interval**
  - **Tolerance Interval**

Point Estimation

- **A single summary value derived from a sample**
  - **Mean: the arithmetic average**

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

  - **Median: that value where half the data are larger, half the data are smaller (i.e. the one in the middle)**
  - **Variance (equals the square of the Standard Deviation)**

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\left(\sum_{i=1}^{n} x_i / n\right)^2}{n - 1}$$

Population and Sample

- **Population: the entire universe of possible values that could be measured. Often exists only in theory or concept. Characteristics are given Greek letters.**
- **Sample: A very small part of the population that is actually obtained. Often assumed to have some understanding that it is representative of the population. Characteristics are given Latin (ordinary) letters.**

Population, Sample, and Inference

- **Take a small sample size "n" from the population size "N"**
- **We hope the mean of the sample "$\bar{X}$" is a reasonable estimate of the population mean "μ"**
- **We hope the variance of the sample "s^2" is a reasonable estimate of the population variance "σ^2"**
- **The key lies in the necessity that we have a truly representative sample**

Precision

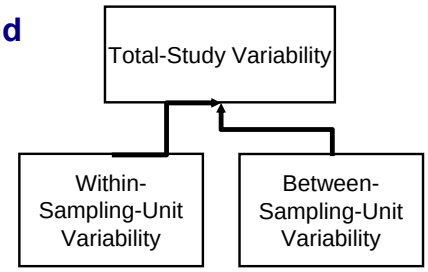
Common definition: Precision is the measure of agreement among repeated measurements of the same property under identical or substantially similar conditions.

Common Indicators of Precision

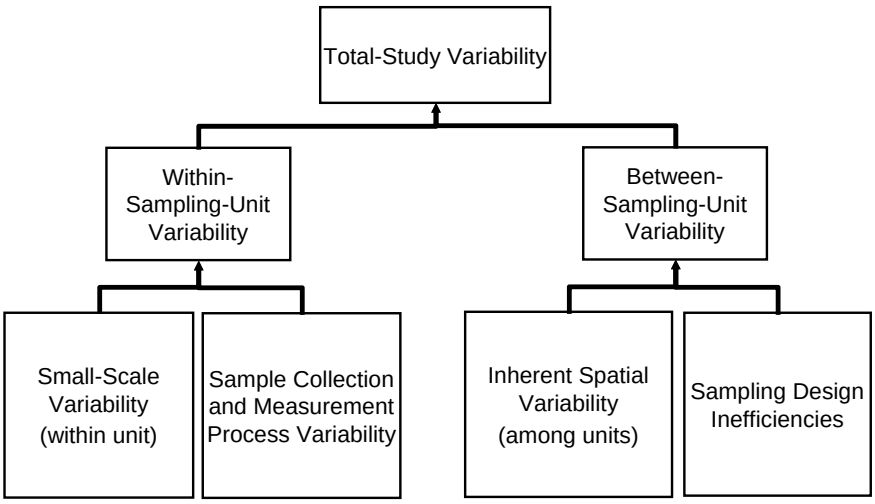
- **Range**
  - –difference between largest and smallest values
- **Variance or standard deviation**
  - –a statistical measure of the spread of data calculated from two or more measured values
  - –the standard deviation is the square root of the variance
- **Relative standard deviation (CV)**
  - –the standard deviation calculated from two or more values divided by the mean of those values

Framework for Evaluating Indicators of Precision

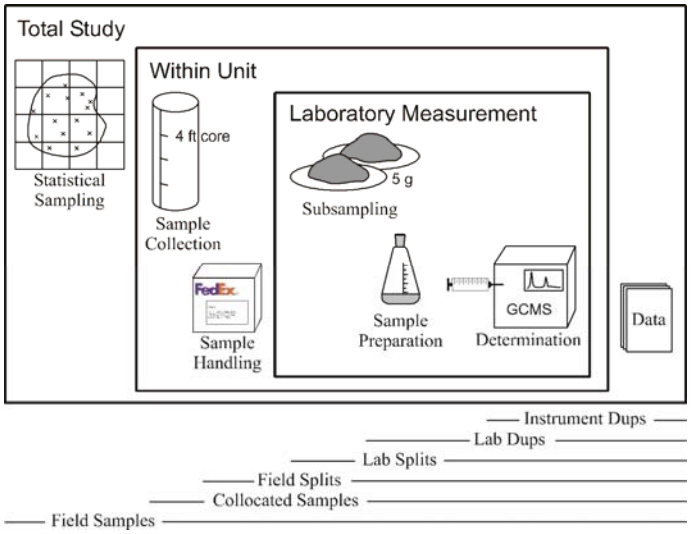
- **A simple model allows us to evaluate the components and indicators of total-study variability**
  - **within-sampling-unit (the physical samples) variability:**
    - **measurement process**
    - **small-scale variability**
    - **sample acquisition**
  - **between-sampling-unit (among the physical samples in the group) variability:**
    - **inherent spatial variability**
    - **sampling design error**



Simple Total-Study Variability Model



QA Samples Used to Evaluate Components of Total-Study Variability



Bias

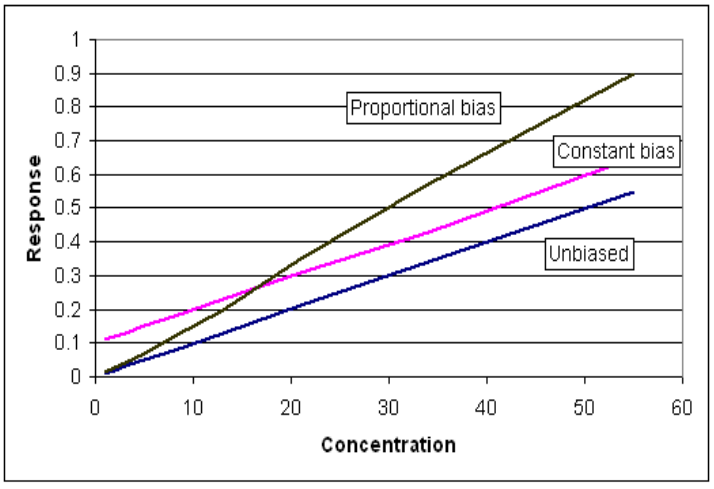
- Bias = measured result - true value
- Relative bias = $\frac{\text{measured result - true value}}{\text{true value}}$
- When dealing with recovery rates:
- Recovery = 1 + $\frac{\text{measured result - true value}}{\text{true value}}$

and expressed as a percentage

Principal Causes of Bias

- Analytical aspects
  - calibration error
  - sample contamination
  - matrix effects
  - interferences
- Sample collection
  - incorrect location identification
  - Incomplete collection of samples
  - Use of a judgmental sampling scheme

Calibration Errors Leading to Bias



Sample Contamination Leading to Bias

- Field and laboratory
  - contamination of volatile organics with vehicle exhaust
  - contamination of metals with materials used in sample collection
  - incorrect sample containers and unlined sample tops
  - reagent water, standards contamination from laboratory solvents (methanol and methylene chloride), and spent membranes (periodic maintenance)
  - contamination from other samples in storage (refrigerators)

Matrix Effects Leading to Bias

- The composition of the matrix can interfere and influence both preparation and analysis
- Non-ideal chemical behavior influences samples differently than standards
  - high ionic strength water enhances purging of volatile organic chemicals (VOCs) (bias high)
  - natural buffering in soil influences leaching of lead in TCLP extraction
  - x-ray fluorescence can result in high (secondary excitation) or low (matrix absorbs greater than analyte) bias

Why Bias, Why Not Accuracy? Mean Square Error

- Accuracy includes both precision (random error that could be positive or negative for each individual reading) and bias (systematic error that is either positive or negative for all readings)
- Accuracy (mean square error) = variance + bias^2
- Precision is estimated through replicate measurements
- Bias is estimated by comparison of the mean of replicate measurements to a known standard
- Without standards bias cannot be estimated with confidence, only a reduction in bias is possible

Conclusions

- Precision involves random error and is estimated by variance in a set of data.
- Bias is made as small as possible by good QC and adherence to good laboratory protocols.
- By making bias very small, the Accuracy (Mean Squared Error) essentially becomes almost indistinguishable from precision and all the standard statistical tests and methods apply.

Conclusions

Basic References

- **Data Quality Assessment: Reviewer's Guide (QA/G-9R)**
  www.epa.gov/quality/qa_docs.html
- **Data Quality Assessment: Statistical Tools for Practitioners (QA/G-9S)**
  www.epa.gov/quality/qa_docs.html
- ***Statistical Methods for Environmental Pollution Monitoring*** **by Richard O. Gilbert, John Wiley & Sons**
- ***Statistical Tools for Environmental Quality Measurement*** **by Michael E. Ginevan & Douglas E. Splitstone, CRC Press**

Advanced References

- ***Environmental Statistics with S-Plus*** **by Steven P. Millard & Nagaraj K. Neerchal, CRC Press**
- ***Statistics for Environmental Science and Management*** **by Bryan F. J. Manly, CRC Press**
- ***Statistical Methods for Detection and Quantification of Environmental Contamination*** **by Robert D. Gibbons & David E. Coleman, John Wiley & Sons**
- ***Nondetects and Data Analysis*** **by Dennis R. Helsel, John Wiley & Sons**