An introduction to statistical analysis

Overheads

D G Rossiter
Department of Earth Systems Analysis
International Institute for Geo-information Science & Earth Observation (ITC)
<http://www.itc.nl/personal/rossiter>

January 9, 2006

Topic: Motivation

- Why is statistics important?

It is part of the **quantitative approach** to knowledge:

> “In physical science the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it.
>
> “I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it;
>
> “. . . but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind;
>
> “. . . it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be.”
>
> – Lord Kelvin (William Thomson), *Popular Lectures and Addresses* 1:73

A simple definition

Statistics: "The determination of the **probable** from the **possible**"

– Davis, *Statistics and data analysis in geology*, p. 6

...which implies the **rigorous definition** and then **quantification** of "probable".

- Probable **causes** of **past** events or observations
- Probable **occurrence** of **future events** or observations

This is a definition of **inferential** statistics:

Observations $\Longrightarrow$ **Inferences**

What is "statistics"?

Two common use of the word:

1. **Descriptive** statistics: numerical summaries of **samples**;
   - (what was observed)
2. **Inferential** statistics: from samples to **populations**.
   - (what could have been or will be observed)

Example:

Descriptive "The adjustments of 14 GPS control points for this orthorectification ranged from 3.63 to 8.36 m with an arithmetic mean of 5.145"

Inferential "The mean adjustment for any set of GPS points used for orthorectification is no less than 4.3 and no more than 6.1 m; this statement has a 5% probability of being wrong."

Why use statistical analysis?

1. **Descriptive**: we want to summarize some data in a shorter form

2. **Inferential**: We are trying to **understand** some process and possible **predict** based on this understanding

   - So we need **model** it, i.e. make a conceptual or mathematical representation, from which we **infer** the process.
   - But how do we know if the model is "correct"?
     - Are we imagining relations where there are none?
     - Are there true relations we haven't found?
   - Statistical analysis gives us a way to **quantify the confidence** we can have in our inferences.

Topic: Introduction

1. Outline of statistical analysis

2. Types of variables

3. Statistical inference

4. Data analysis strategy

5. Univariate analysis

6. Bivariate analysis; correlation; linear regression

7. Analysis of variance

8. Non-parametric methods

Reference web pages

- **Electronic Statistics Textbook**: [StatSoft] `http://www.statsoftinc.com/textbook/stathome.html`
- **NIST/SEMATECH e-Handbook of Statistical Methods**: `http://www.itl.nist.gov/div898/handbook/`
- **HyperStat Online Textbook**: `http://davidmlane.com/hyperstat/`
- The **R** environment for statistical computing and graphics: `http://www.r-project.org/`
- **StatLib**: "a system for distributing statistical software, datasets, and information" `http://lib.stat.cmu.edu/`

Texts

There are hundreds of texts at every level and for every application. Here are a few I have found useful.

Elementary:

- Bulmer, M.G., 1979. *Principles of statistics*. Dover Publications, New York.
- Dalgaard, P., 2002. *Introductory Statistics with R*. Springer-Verlag.

Advanced:

- Venables, W.N. and Ripley, B.D., 2002. *Modern applied statistics with S*. Springer-Verlag.
- Fox, J., 1997. *Applied regression, linear models, and related methods*. Sage, Newbury Park.

Applications:

- Davis, J.C., 2002. *Statistics and data analysis in geology*. John Wiley & Sons, New York.
  - Website: `http://www.kgs.ku.edu/Mathgeo/Books/Stat/index.html`
- Webster, R. and Oliver, M.A., 1990. *Statistical methods in soil and land resource survey*. Oxford University Press, Oxford.

Topic: Outline of statistical analysis

- What is statistical analysis?
- Populations, samples, outliers
- Steps in statistical analysis

What is “statistical analysis”?

This term refers to a wide range of techniques to. . .

1. (Describe)

2. Explore

3. Understand

4. Prove

5. Predict

. . . based on **sample datasets** collected from **populations**, using some **sampling strategy**.

Why use statistical analysis?

1. We want to summarize some data in a shorter form

2. We are trying to **understand** some process and possible **predict** based on this understanding

   • So we need **model** it, i.e. make a conceptual or mathematical representation, from which we **infer** the process.
   • But how do we know if the model is "correct"?
     * Are we imagining relations where there are none?
     * Are there true relations we haven't found?
   • Statistical analysis gives us a way to **quantify the confidence** we can have in our inferences.

Populations and samples

- **Population**: a set of elements (individuals)
  - * Finite vs. "infinite"
- **Sample**: a subset of elements taken from a population
  - * Representative vs. biased
- We make **inferences** about a population from a sample taken from it.
- In some situations we can examine the entire population; then there is no inference from a sample. Example: all pixels in an image.

Step 1: Outliers

Three uses of this word:

1. An observation that is some defined distance away from the sample mean (an **empirical** outlier;

2. An extreme member of a **population**;

3. An observation in the **sample** that is *not* part of the population of interest.

Example: In a set of soil samples, one has an order of magnitude greater level of heavy metals (Cd, Pb, Cu etc.) than all the others.

1. The sample is an empirical outlier because it is more than 1.5 times the inter-quartile range from the 3rd quartile;

2. This is an extreme value but is included in our analysis of soil contamination;

3. This sample comes from an industrial site and is not important for our target population of agricultural soils.

Step 1: Explore & Describe

- Questions
  - What is the **nature of the dataset** (lineage, variables . . .)?
  - What is the relation of the dataset to the underlying **population(s)**?
- Techniques
  - **Graphical** (visualisation): humans are usually good at picking out patterns
  - **Numerical**: summaries of outstanding features (**descriptive** statistics)
  - These may suggest **hypotheses** and appropriate **analytical techniques**

Step 2: Understand

- **If** there is an underlying **process** of which the sampled data are a **representative sample** . . .
- . . . **then** the data allow us to **infer** the nature of the process
- Example: the distribution of heavy metals in soil is the result of:
  - Parent material
  - Pollutants transported by wind, water, or humans
  - Transformations in the soil since deposition
  - Movement of materials within and through the soil
  - . . .
- Summarize the understanding with a **model**

What is a statistical model?

- A mathematical representation of a process or its outcome . . .
- . . . with a computable level of **uncertainty**
- . . . according to **assumptions** (more or less plausible or proveable)

This is an example of an **empirical** model. It may **imply** the underlying process, but need not. It might be useful for **prediction**, even if it's a "black box".

Assumptions: not part of the model, but must be true for the model to be correct.

(Note: A **process** model explicitly represents the underlying process and tries to **simulate** it.)

Step 3: "Prove"

A further step is to **"prove"**, in some sense, a statement about nature.

E.g. "Soil pollution in this area is caused by river flooding; pollutants originate upstream in industrial areas."

- The model must be **plausible** → evidence of **causation**
- With what **confidence** can we state that our understanding (model) is correct?
- Nothing can be proved absolutely; statistics allows us to **accumulate evidence**
- We can determine sampling strategies to achieve a given confidence level
- Underlying **assumptions** may not be proveable, only plausible

Step 4: Predict

- The model can be applied to **unsampled entities** in the underlying population
  - **Interpolation**: within the range of the original sample
  - **Extrapolation**: outside this range

- The model can be applied to future events; this assumes that future conditions (the **context** in which the events will take place) is the same as past conditions (c.f. "uniformitarianism" of Hutton and Playfair)

- A *geo*-statistical model can be applied to **unsampled locations**; this assumes that the process at these locations is the same as at the sample locations.

Key point: we must **assume** that the **sample** on which the model is based is **representative** of the **population** in which the predictions are made. We argue for this with **meta-statistical analysis** (outside of statistics itself).

Topic: Types of variables

In order of **information content** (least to most):

1. **Nominal**

2. **Ordinal**

3. **Interval**

4. **Ratio**

Nominal variables

- Values are from a set of **classes** with **no natural ordering**
- Example: Land uses (agriculture, forestry, residential . . .)
- Can determine **equality**, but *not* rank
- Meaningful sample statistics: mode (class with most observations); frequency distribution (how many observations in each class)
- Numbers may be used to label the classes but these are arbitrary and have no numeric meaning (the “first” class could just as well be the “third”); ordering is by convenience (e.g. alphabetic)
- R: “unordered factors”

Ordinal variables

- Values are from a set of **naturally ordered classes** with **no meaningful units of measurement**
- Example: Soil structural grade (0 = structureless, 1 = very weak, 2 = weak, 3 = medium, 4 = strong, 5= very strong)
- N.b. This ordering is an *intrinsic part of the class definition*
- Can determine **rank** (greater, less than)
- Meaningful sample statistics: mode; frequency distribution
- Numbers may be used to label the classes; their **order** is meaningful, but not the **intervals** between adjacent classes are not defined (e.g. the interval from 1 to 2 vs. that from 2 to 3)
- R: “ordered factors”

Interval variables

- Values are measured on a **continuous scale** with **well-defined units of measurement** but **no natural origin of the scale**, i.e. the zero is arbitrary, so that differences are meaningful but not ratios

- Example: Temperature in $^\circ\mathrm{C}$.

- "It is twice as warm yesterday as today" is meaningless, even though "Today it is $20^\circ\mathrm{C}$ and yesterday it was $10^\circ\mathrm{C}$" may be true.
  - (To see this, try the same statement with Farenheit temperatures)

- Meaningful statistics: quantiles, mean, variance

Ratio variables

- Values are measured on a **continuous scale** with **well-defined units of measurement** and a **natural origin of the scale**, i.e. the zero is meaningful
- Examples: Temperature in $^\circ\mathrm{K}$; concentration of a chemical in solution
- "There is twice a much heat in this system as that" is meaningful, if one system is at $300^\circ\mathrm{K}$ and the other at $150^\circ\mathrm{K}$
- Meaningful statistics: quantiles, mean, variance; also the coefficient of variation. (Recall: CV = SD / Mean; this is a ratio).

Continuous vs. discrete

Interval and ratio variables can be either:

Discrete Taking one of a limited set of discrete values, e.g. integers

Continuous Can take any value (limited by precision) in a defined *range*

- Not “continuous” in the strict mathematical sense (because the computer can only represent rational numbers)

Topic: Statistical Inference

One of the main uses of statistics is to **infer** from a sample to a population, e.g.

- the “true” value of some parameter of interest (e.g. mean)
- the degree of support for or against a hypothesis

This is a contentious subject; here we use simple “frequentist” notions.

Statistical inference

- Using the **sample** to infer facts about the underlying **population** of which (we hope) it is representative

- Example: true value of a **population mean**, estimated from **sample mean** and its **standard error**
  - **confidence intervals**: having a **known probability of containing the true value**
  - For a sample from a normally-distributed variate, 95% probability ($\alpha = 0.05$):

$$\bar{x} - 1.96 \cdot s_{\overline{X}} \leq \mu \leq \bar{x} + 1.96 \cdot s_{\overline{X}}$$

  - The standard error is estimated from the sample variance:

$$s_{\overline{X}} = \sqrt{s_X^2 / n}$$

Inference from small samples

- Probabilities are referred to the t (Student's) distribution, rather than the z (Normal) distribution

- This corrects for the fact that we are estimating both the mean and variance from the same sample, and the variance is difficult to estimate from small samples

$$(\overline{x} - t_{\alpha=0.05,n-1} \cdot s_{\overline{X}}) \ \leq \ \mu \ \leq (\overline{x} + t_{\alpha=0.05,n-1} \cdot s_{\overline{X}})$$

- t from tables; $t \rightarrow z$ as $n \rightarrow \infty$

- $t_{\alpha=0.05,10} = 2.228$, $t_{\alpha=0.05,30} = 2.042$, $t_{\alpha=0.05,120} = 1.980$

What does this really mean?

- "There is only a 1 in 20 chance that the true value of the population mean is outside this interval"
  - **If** the sample is representative of the population
  - **If** the distribution of values in the sample satisfies the requirements of the inferential method

- "If we repeat the same sampling strategy again (collecting a new sample), there is only a 1 in 20 chance that the confidence interval constructed from that sample will not contain the mean value from this first sample"

- This does *not* mean that 95% of the sample or population is within this interval!

The null and alternate hypotheses

- **Null** hypothesis H_0: Accepted until proved otherwise ("innocent until proven guilty")

- **Alternate hypothesis** H_1: Something we'd like to prove, but we want to be fairly sure

- In the absence of prior information, the null hypothesis is that there is no relation

  * Classic example: a new crop variety does not (null) have a higher yield than the current variety (note one-tailed hypothesis in this case)

- But may use prior information for an 'informative' null hypothesis

Significance levels and types of error

- α is the risk of a **false positive** (rejecting the null hypothesis when it is in fact true), the **Type I** error
  * "The probability of convicting an innocent person" (null hypothesis: innocent until proven guilty)

- β is the risk of a **false negative** (accepting the null hypothesis when it is in fact false), the **Type II** error.
  * "The probability of freeing a guilty person"

- α set by analyst, β depends on the form of the test

Selecting a confidence level

These must be balanced depending on the **consequences** of making each kind of error. For example:

- The cost of introducing a new crop variety if it's not really better (Type I), vs.
- The lost income by not using the truly better variety (Type II)
- The British legal system is heavily weighted towards low Type I errors (i.e. keep innocent people out of prison)
- The Napoleonic system accepts more Type I error in order to lower Type II error (i.e. keep criminals off the street)

(Or, the British and Napoleonic systems may have opposite null hypotheses.)

Topic: Data anlysis strategy

1. Posing the research questions
2. Examining data items and their support
3. Exploratory non-spatial data analysis
4. Non-spatial modelling
5. Exploratory spatial data analysis
6. Spatial modelling
7. Prediction
8. Answering the research questions

Research questions

- What research questions are supposed to be answered with the help of these data?

Data items and their support

- How were the data collected (sampling plan)?
- What are the variables and what do they represent?
- What are the units of measure?
- What kind of variables are these (nominal, ordinal, interval, or ratio)?
- Which data items could be used to stratify the population?
- Which data items are intended as response variables, and which as predictors?

Non-spatial modelling

- Univariate descriptions: normality tests, summary statistics
- Transformations as necessary and justified
- Bivariate relations between variables (correlation)
- Multivariate relations between variables
- Analysis of Variance (ANOVA) on predictive factors (confirms subpopulations)

Exploratory spatial data analysis

If the data were collected at known points in geographic space, we should visualise them in that space.

- Postplots: where are which values?
- Geographic postplots: with images, landuse maps etc. as background: do there appear to be any explanation for the distribution of values?
- Spatial structure: range, direction, strength . . .
- Is there anisotropy? In what direction(s)?
- Populations: one or many?

Spatial modelling

If the data were collected at known points in geographic space, it may be possible to model this.

- Model the spatial structure
  * Local models (spatial dependence)
  * Global models (geographic trends, feature space predictors)
  * Mixed models

Prediction

- Values at points or blocks
- Summary values (e.g. regional averages)
- Uncertainty of predictions

Answer the research questions

- How do the data answer the research question?
- Are more data needed? If so, how many and where?

Topic: The Meuse soil pollution data set

This will be used as a **running example** for the following lectures.

It is an example of an “environmental” dataset which can be used to answer a variety of practical and theoretical research question.

Source

Rikken, M.G.J. & Van Rijn, R.P.G., 1993.

Soil pollution with heavy metals – An inquiry into spatial variation, cost of mapping and the risk evaluation of copper, cadmium, lead and zinc in the floodplains of the Meuse west of Stein, the Netherlands.

Doctoraalveldwerkverslag, Dept. of Physical Geography, Utrecht University

This data set is also used as an example in `gstat` and in the GIS text of Burrough & McDonnell.

Variables

155 samples taken on a support of 10x10 m from the top 0-20 cm of alluvial soils in a 5x2 km part the floodplain of the Maas (Meuse) near Stein (NL).

`id`	point number
`x, y`	coordinates E and N in Dutch national grid coordinates, in meters
`cadmium`	concentration in the soil, in mg kg^{-1}
`copper`	concentration in the soil, in mg kg^{-1}
`lead`	concentration in the soil, in mg kg^{-1}
`zinc`	concentration in the soil, in mg kg^{-1}
`elev`	elevation above local reference level, in meters
`om`	organic matter loss on ignition, in percent
`ffreq`	flood frequency class, 1: annual, 2: 2-5 years, 3: every 5 years
`soil`	soil class, coded
`lime`	has the land here been limed? 0 or 1 = F or T
`landuse`	land use, coded
`dist.m`	distance from main River Maas channel, in meters

Accessing the Meuse data set

In R:

```
> library(gstat)
> data(meuse)
> str(meuse)
```

To import in other programs: comma-separated value (CSV) file `meuse.csv`.

```
"x","y","cadmium","copper","lead","zinc","elev","dist","om","ffreq","soil","lime","landuse",
181072,333611,11.7, 85,299,1022, 7.909,0.00135803,13.6,"1","1","1","Ah",  50
181025,333558, 8.6, 81,277,1141, 6.983,0.01222430,14.0,"1","1","1","Ah",  30
...
```

Topic: Probability

1. probability

2. discrete and continuous probability distributions

3. normality, transformations

Probability

- A very controversial topic, deep relation to philosophy;
- Two major concepts: **Bayesian** and **Frequentist**;
- The second can model the first, but not vice-versa;
- Most elementary statistics courses and computer programs take the frequentist point of view.

The **probability of an event** is:

Bayesian **degree of rational belief** that the event will occur, from 0 (**impossible**) to 1 (**certain**)

Frequentist the **proportion of time** the event would occur, should the "experiment" that gives rise to the event be **repeated a large number of times**

Frequentist concept of probability

- Intutively-appealing if an experiment can **easily be repeated** or another sample easily be taken, under the same conditions.
  - Often the case in earth sciences: if we have 100 soil samples we could (if budget allows) take 100 more; effectively there are an infinite number of possible samples

- Not so helpful with rare events
  - What is the frequentist “probability” of a major earthquake in the next year?
  - This is why Bayesian methods (e.g. **weights of evidence**) are often used in risk assessment.

Probability distributions

- A complete account of the **probability of each possible outcome** . . .
- **assuming** some **underlying process**
- n.b. the **sum** of the probabilities of all events is by definition 1 (it's certain that *something* will happen!)

Examples:

- Number of radioactive decays in a given time period: **Poisson**
  - *assuming* exponential decay with constant half-life, independent events.
- Number of successes in a given number of binary ("Bernoulli") trials (e.g. finding water within a fixed depth): **Binomial**
  - *assuming* constant probability of success, independent trials

Probability vs. reality

- Frequentist probability refers to an **idealized** world with perfect mathematical properties;
- It is useful **if we can argue that the assumptions are met**;
- This is a **meta-statistical** argument.

Example: to describe a well-drilling programme with the binomial distribution, we must argue that:

1. An attempt can be unambiguously classified as a success or failure;
2. Every attempt to drill a well is independent;
3. Every attempt to drill a well has the same (possibly unknown) probability of success.

Only then can we **model** the campaign with a binomial distribution.

The Binomial distribution

This is a **discrete** probability distribution:

- Probability Density Function (PDF) of x successes in n trials, each with probability p of success:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where

$$\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}$$

is the **binomial coefficient**, i.e. the number of different ways of selecting x distinct items out of n total items.

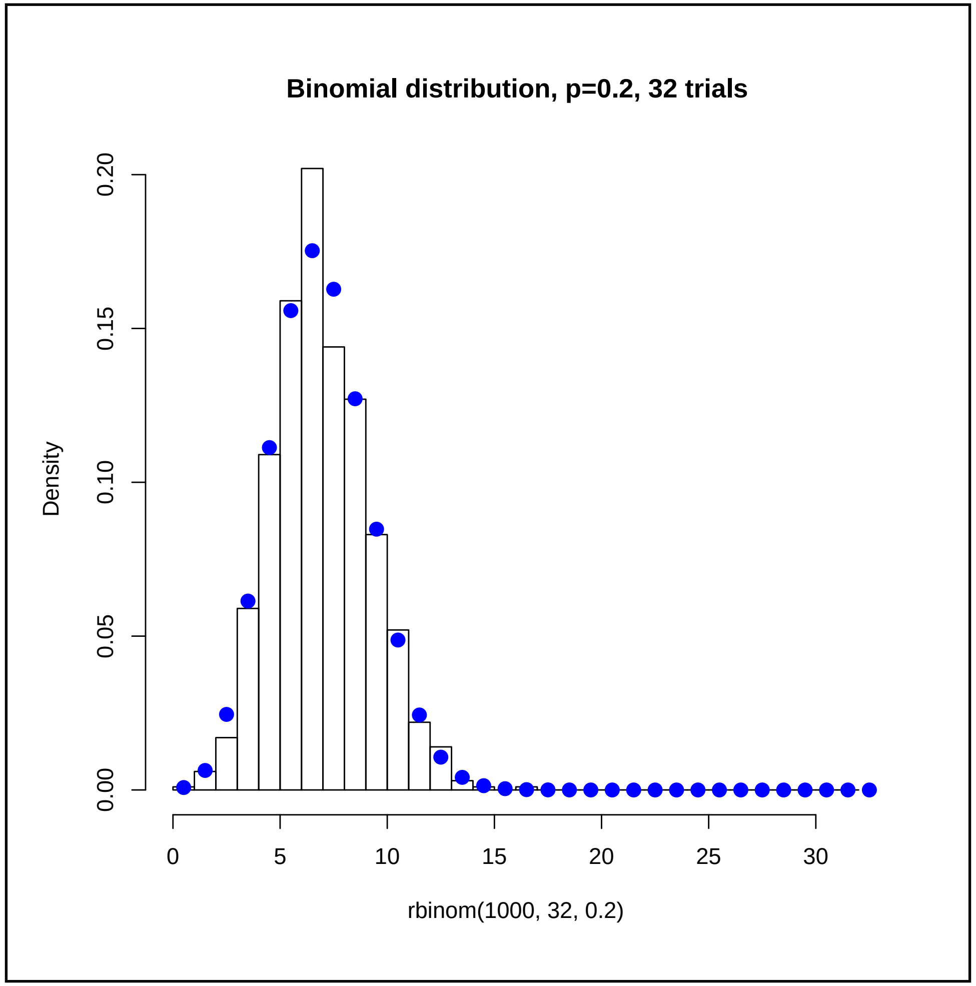
Mean and variance:

$$\mu = np; \sigma^2 = np(1-p)$$

Example computation in R

```
> # number of distinct ways of selecting 2 from 16
> (f2 <- factorial(16)/(factorial(2)*factorial(16-2)))
[1] 120
> # direct computation of a single binomial density
> #   for prob(success) = 0.2
> p <- 0.2; n <- 16; x <- 2
> f2 * p^x * (1-p)^(n-x)
[1] 0.21111
> # probability of 0..16 productive wells if prob(success) = 0.2
> round(dbinom(0:16, 16, 0.2),3)
 [1] 0.028 0.113 0.211 0.246 0.200 0.120 0.055 0.020
 [9] 0.006 0.001 0.000 0.000 0.000 0.000 0.000 0.000
[17] 0.000
> # simulate 20 drilling campaigns of 16 wells, prob(success) = 0.2
> trials <- rbinom(20, 16, .2)
> summary(trials)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.75    3.00    3.45    4.00    8.00
> # compare with theoretical mean and variance
> (mu <- n * p)
[1] 3.2
> (var <- n * p * (1-p)); var(trials)
[1] 2.56
[1] 2.2605
> sort(trials)
 [1] 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 5 5 8
```

Graph of an empirical vs. theoretical binomial distribution




```
> hist(rbinom(1000, 32, .2), breaks=(0:32), right=F, freq=F,
+         main="Binomial distribution, p=0.2, 32 trials")
> points(cbind((0:32)+0.5,dbinom(0:32, 32, 0.2)), col="blue",
+         pch=20, cex=2)
```

The Normal (Gaussian) probability distribution

This is a **continuous** probability distribution.

- Arises naturally in many processes: a variables that can be modelled as a **sum of many small contributions**, each with the same distribution of errors (**central limit theorem**)
- Easy mathematical manipulation
- Fits many observed distributions of **errors** or **random effects**
- Some statistical procedures require that a variable be at least approximately normally distributed
- Note: even if a variable itself is not normally distributed, its *mean* may be, since the deviations from the mean may be the "sum of many small errors".

Mathematical form of the Normal distribution

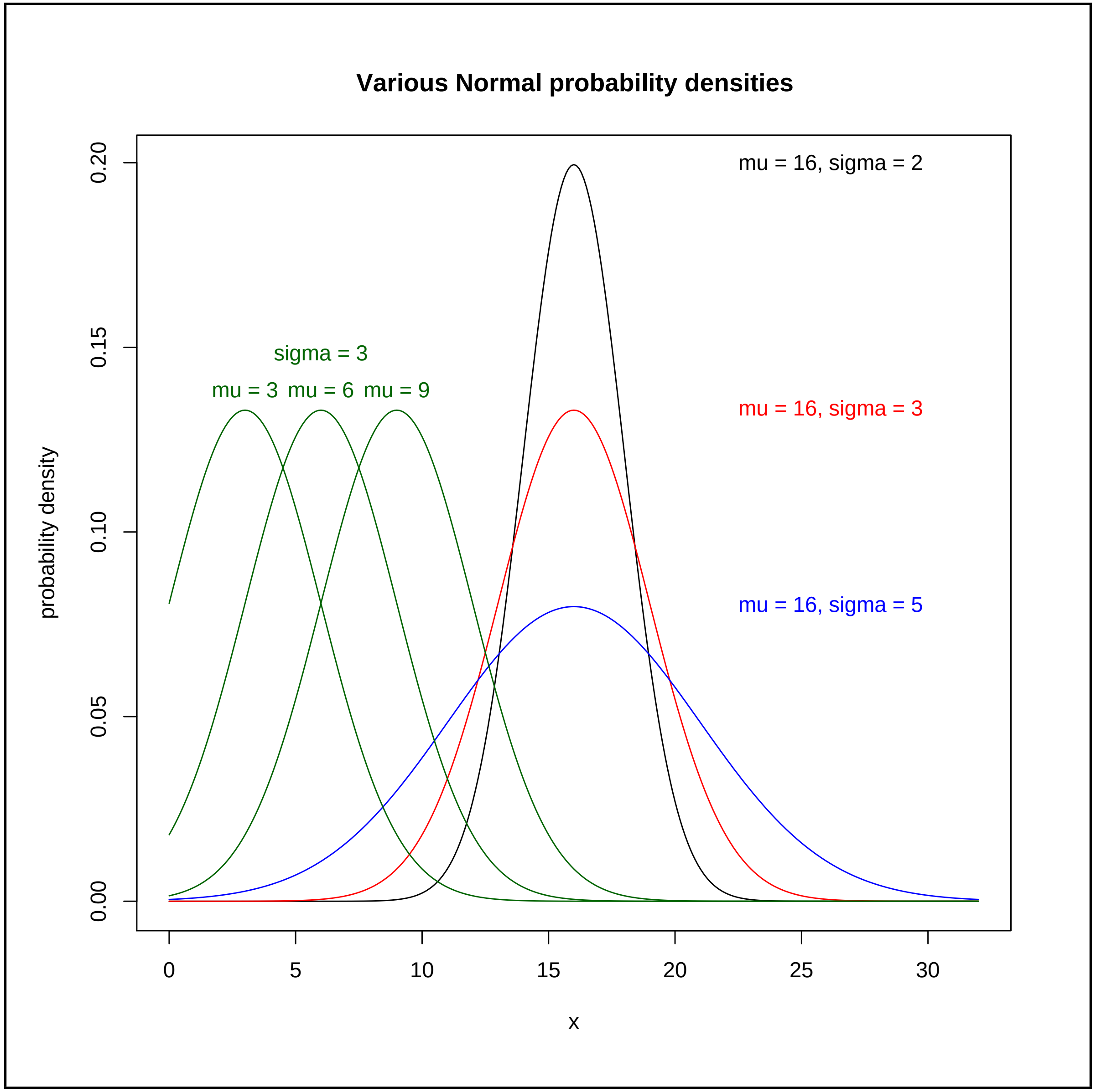
- Probability Density Function (pdf) with mean μ, standard deviation σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \int_{x=-\infty}^{\infty} f(x) = 1$$

- Cumulative Density Function (cdf)

$$F(z) = \int_{x=-\infty}^{z} f(x)$$

```
> # 8 normal variates with mean 1.6, var .2
> rnorm(8, 1.6, .2)
[1] 1.771682 1.910130 1.518092 1.712963 1.365242 1.837332 1.777395 1.749878
> # z-values for some common probabilities
> qnorm(seq(0.80,0.95, by=.05),1.6,.2)
[1] 1.768324 1.807287 1.856310 1.928971
```




```
> range <- seq(0,32, by=.1)
> plot(range, dnorm(range, 16, 2), type="l") # etc.
```

Standardization

- All normally-distributed variates can be directly compared by **standardization**: subtract μ, divide by σ

- **Standardized** normal: all variables have the same scale and deviation: $\mu = 0, \sigma = 1$

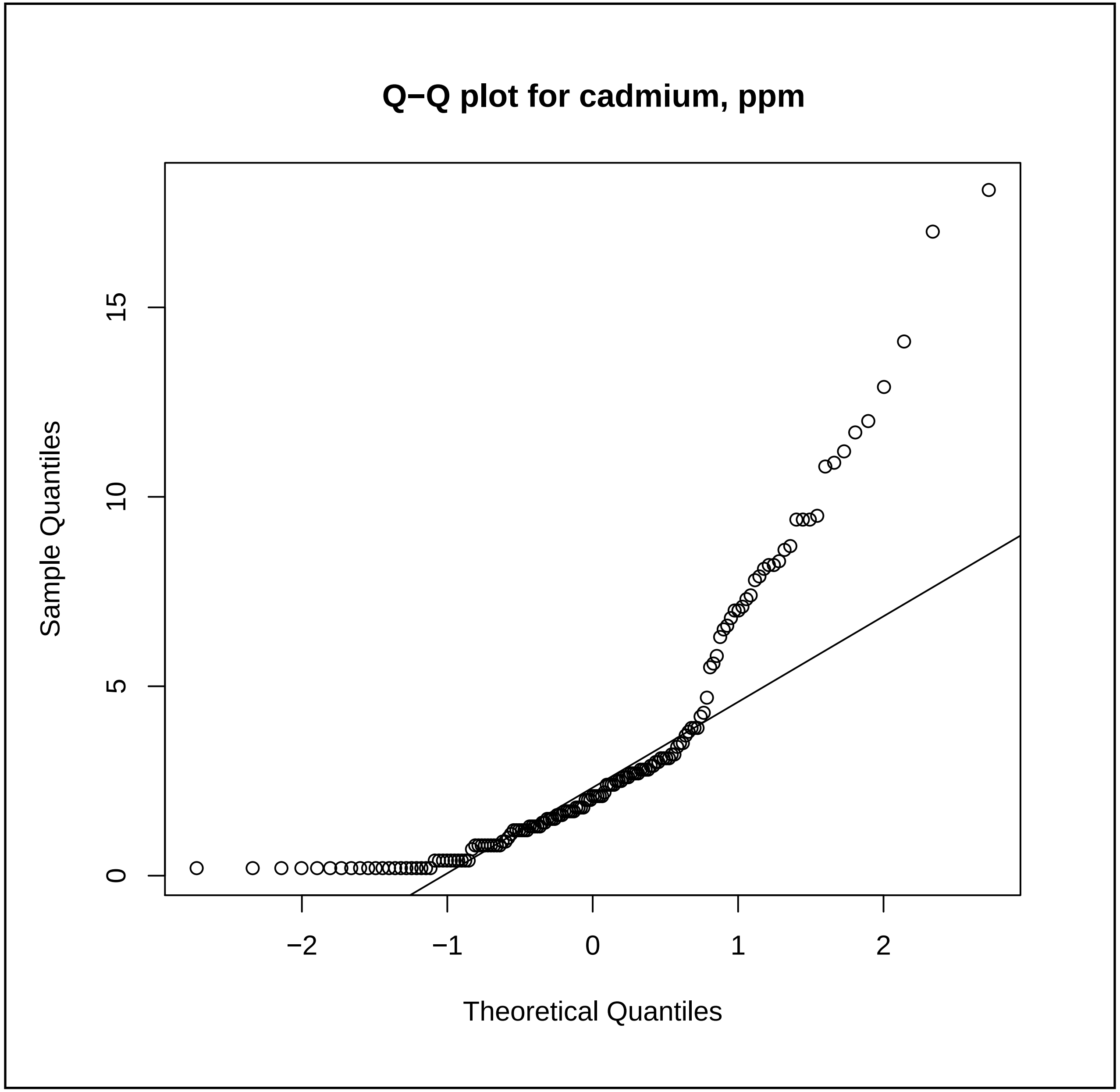
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

```
> sdze<-function(x) { (x-mean(x))/sd(x) }
```

Evaluating Normality

- Graphical
  - Histograms
  - Quantile-Quantile plots (normal probability plots)
- Numerical
  - Various tests including Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk
  - These all work by compare the **observed** distribution with the **theoretical** normal distribution having parameters estimated from the observed, and computing the **probability** that the observed is a realisation of the theoretical

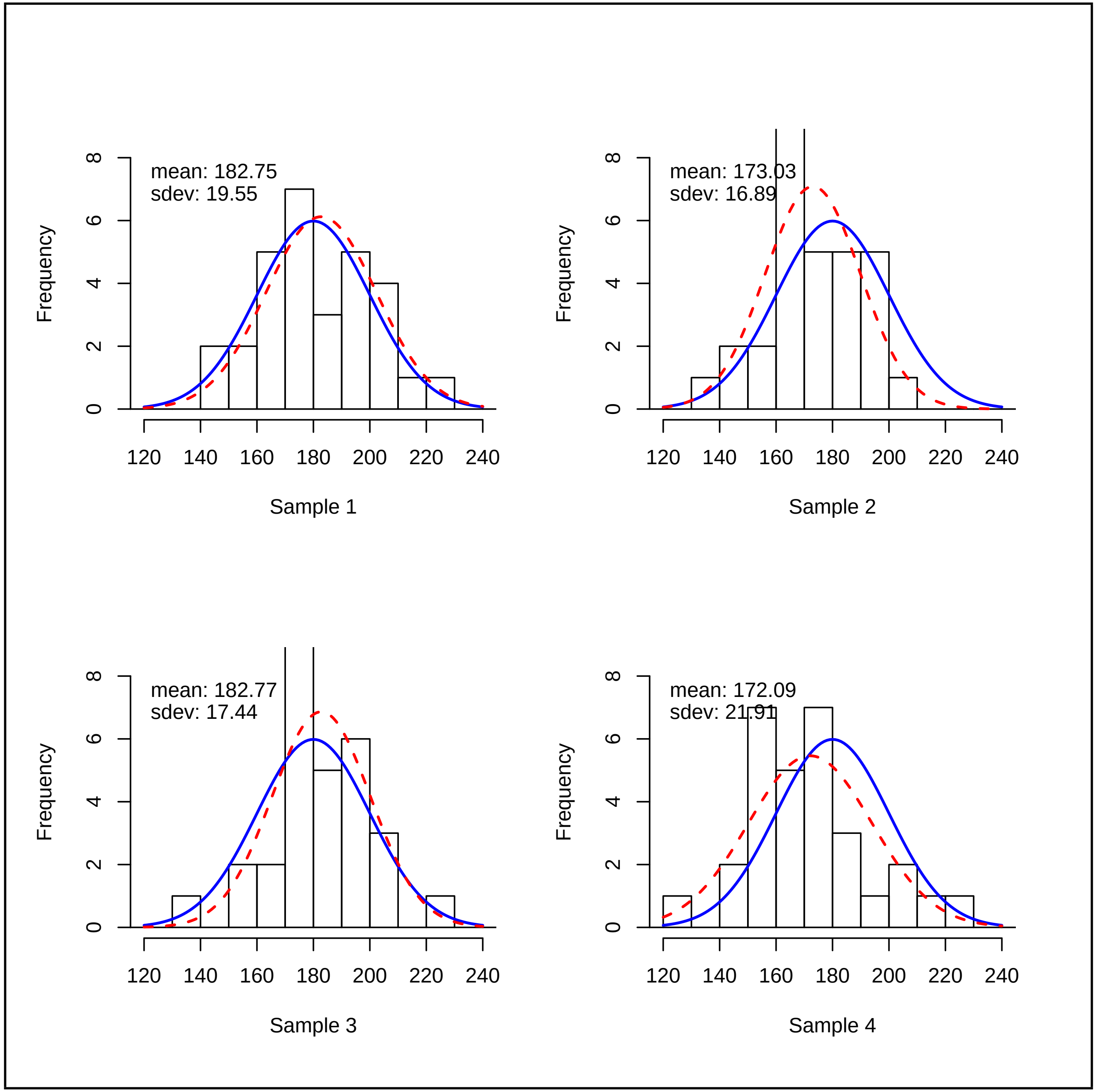
```
> qqnorm(cadmium);  qqline(cadmium)
> shapiro.test(cadmium)
Shapiro-Wilk normality test
W = 0.7856, p-value = 8.601e-14
```


Q−Q plot for cadmium, ppm
Sample Quantiles
0
5
10
15
−2
−1
0
1
2
Theoretical Quantiles

Variability of small samples from a normal distribution

Can we infer that the population is normal from a small sample?

```
> for (i in 1:r) v[,i]<-rnorm(4, 180, 20)
> for (i in 1:r) {
+    hist(v[,i], xlim=c(120, 240), ylim=c(0, 4/3.5),
+    breaks=seq(100, 260, by=10),
+    main="", xlab=paste("Sample", i)) ;
+    x<-seq(120, 240, by=1)
+    points(x,dnorm(x, 180, 20)*4*10,  type="l"",
+         col="blue, lty=1, lwd=1.8)
+    points(x,dnorm(x, mean(v[,i]), sd(v[,i]))*4*10, type="l",
+         col="red", lty=2, lwd=1.8)
+ }
```


mean: 182.75
sdev: 19.55
Frequency
Sample 1
mean: 173.03
sdev: 16.89
Frequency
Sample 2
mean: 182.77
sdev: 17.44
Frequency
Sample 3
mean: 172.09
sdev: 21.91
Frequency
Sample 4

Transforming to Normality: Based on what criteria?

These are listed in order of preference:

1. *A priori* understanding of the process
   - e.g. lognormal arises if contributing variables multiply, rather than add
2. EDA: visual impression of what should be done
3. Results: transformed variable appears and tests normal

Transforming to Normality: Which transformation?

- $x' = \ln(x+a)$: **logarithmic**; removes positive skew
  note: must add a small adjustment to zeroes

- $x' = \sqrt{x}$: **square root**: removes moderate skew

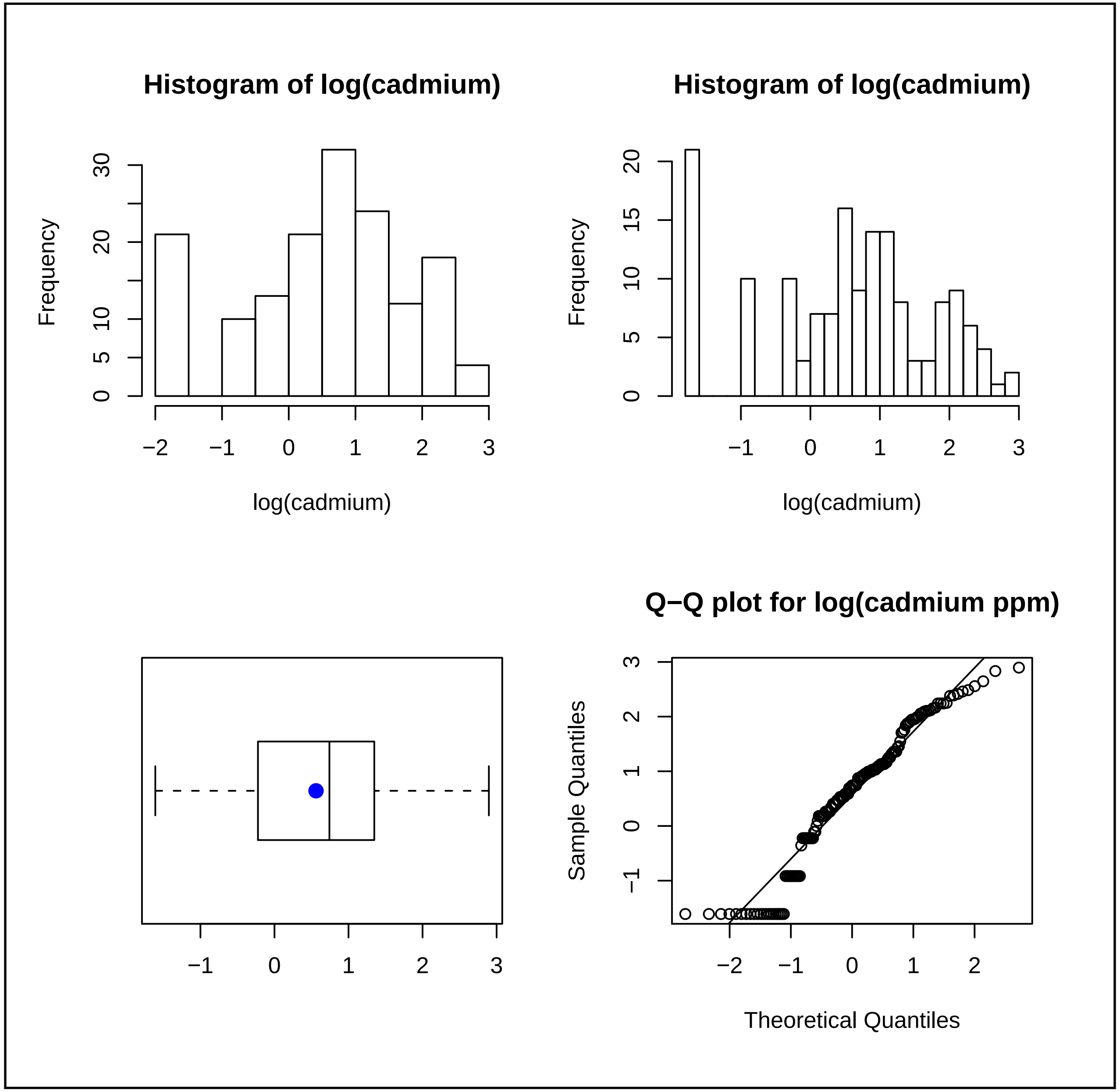
- $x' = \sin^{-1} x$: **arcsine**: for proportions $x \in [0 \ldots 1]$
  spreads the distribution near the tails

- $x' = \ln[x/(1-x)]$: **logit** (logistic) for proportions $x \in [0 \ldots 1]$
  note: must add a small adjustment to zeroes

Example: log transform of a variable with positive skew

```
> summary(log(cadmium))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.6090 -0.2231  0.7419  0.5611  1.3480  2.8960
> stem(logcad)
> hist(log(cadmium))
> hist(log(cadmium), n=20)
> plot(ecdf(log(cadmium)))
> boxplot(log(cadmium), horizontal=T)
> points(mean(log(cadmium)),1, pch=20, cex=2, col="blue")
> qqnorm(log(cadmium), main="Q-Q plot for log(cadmium ppm)")
> qqline(log(cadmium))
> shapiro.test(log(cadmium))
Shapiro-Wilk normality test
W = 0.9462, p-value = 1.18e-05
```

This is still not normal, but much more symmetric


Histogram of log(cadmium)
Frequency
log(cadmium)
Histogram of log(cadmium)
Frequency
log(cadmium)
Q−Q plot for log(cadmium ppm)
Sample Quantiles
Theoretical Quantiles

Topic: Non-spatial univariate Exploratory Data Analysis (EDA)

1. exploratory data analysis

2. descriptive statistics

Exploratory Data Analysis (EDA)

- Statistical analysis should lead to understanding, not confusion . . .

- . . . so it makes sense to **examine** and **visualise** the data with a **critical eye** to see:

  1. Patterns; outstanding features
  2. **unusual** data items (not fitting a pattern); **blunders**? from a **different population**?
  3. Promising analyses

- Reconaissance before the battle

- Draw obvious conclusions with a minimum of analysis

Graphical Univariate EDA

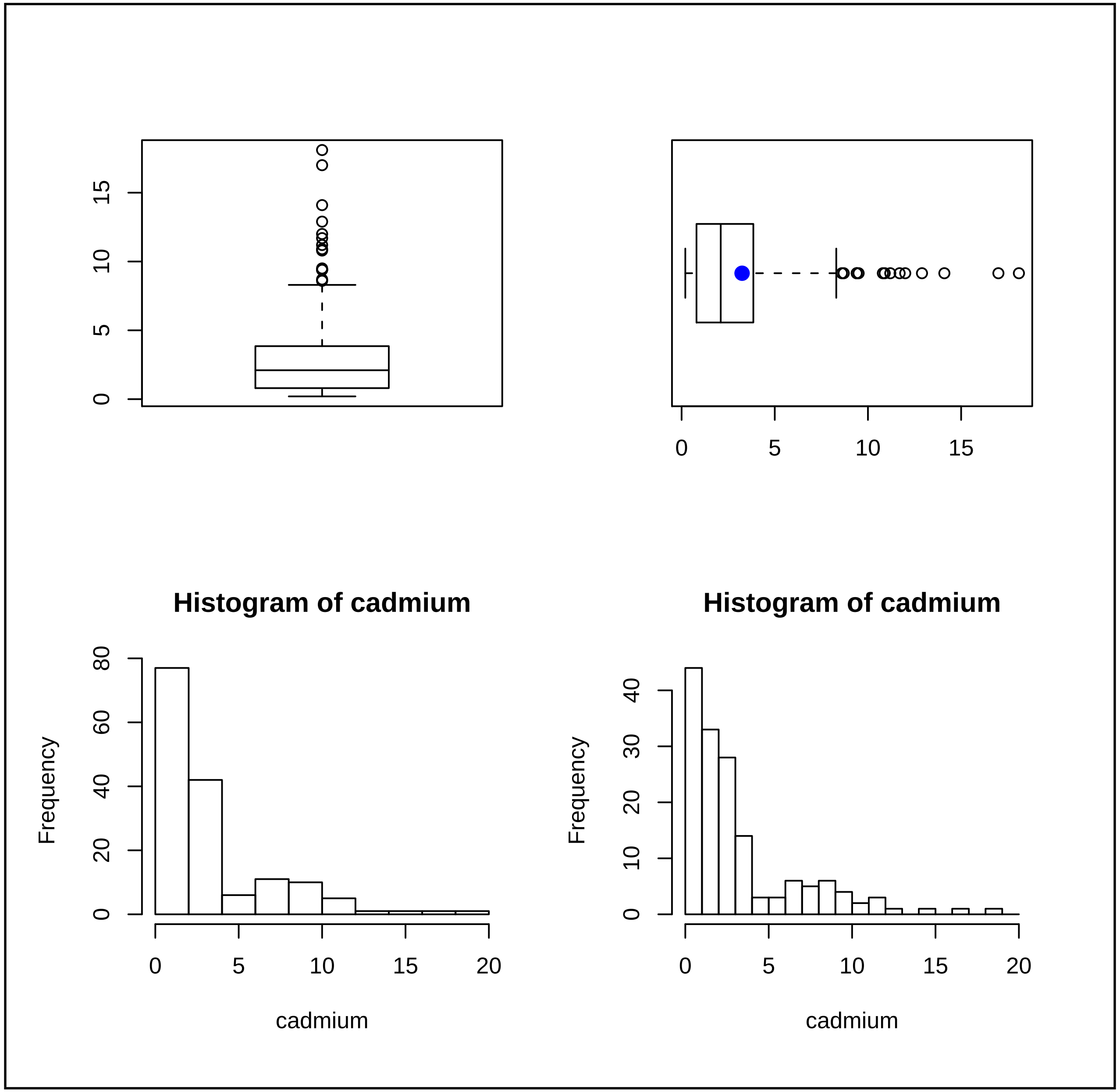
- Boxplot, stem-and-leaf plot, histogram, empirical CDF
- Questions
  - One population or several?
  - Outliers?
  - Centered or skewed (mean vs. median)?
  - “Heavy” or “light” tails (kurtosis)?

```
> stem(cadmium)
> boxplot(cadmium)
> boxplot(cadmium, horizontal = T)
> points(mean(cadmium),1, pch=20, cex=2, col="blue")
> hist(cadmium)                              #automatic bin selection
> hist(cadmium, n=16)                     #specify the number of bins
> hist(cadmium, breaks=seq(0,20, by=1))    #specify breakpoints
> plot(ecdf(cadmium))
```

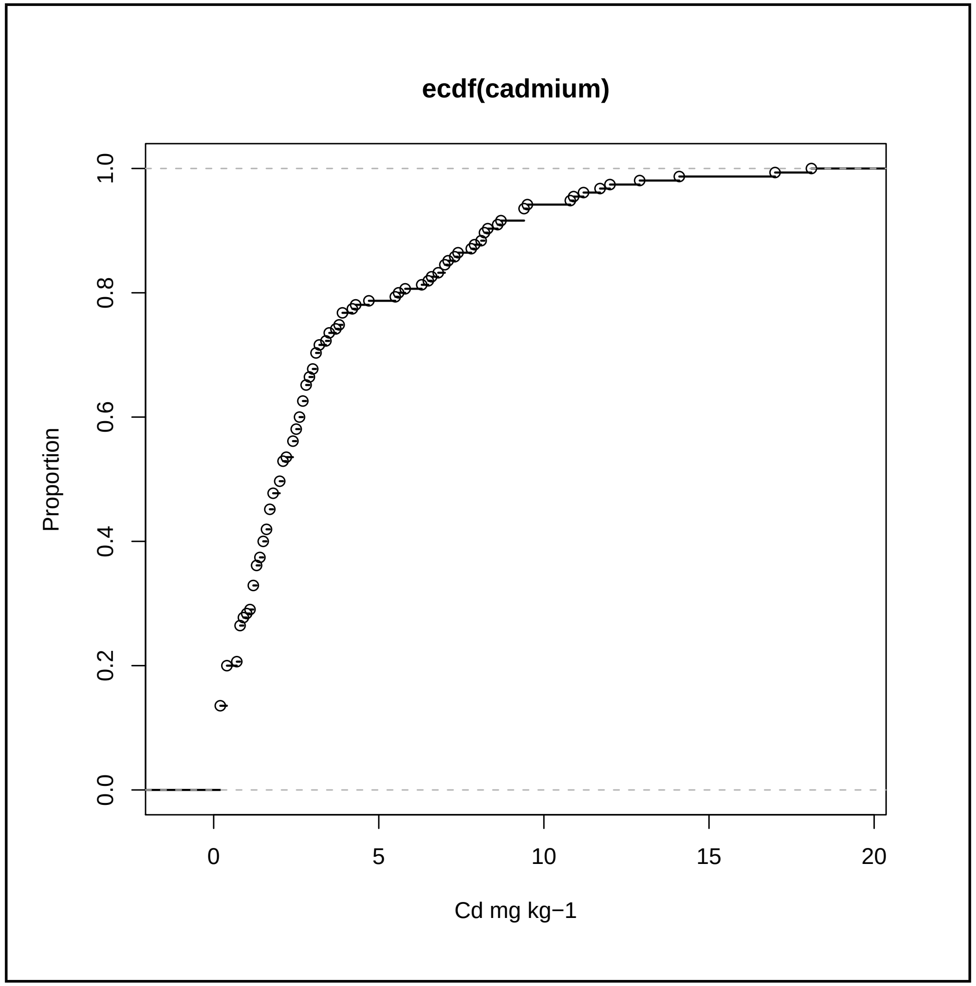
Example stem plot

```
> stem(cadmium)
  The decimal point is at the |
   0 | 2222222222222222222224444444444788888888899
   1 | 0122222233333445555666777778888
   2 | 00011111244445556667777888899
   3 | 0011112245578999
   4 | 237
   5 | 568
   6 | 3568
   7 | 0013489
   8 | 122367
   9 | 4445
  10 | 89
  11 | 27
  12 | 09
  13 |
  14 | 1
  15 |
  16 |
  17 | 0
  18 | 1
```

Example boxplots and histograms



Example empirical cumulative distribution plot



Summary statistics (1)

These summarize a single sample of a single variable

- 5-number summary (min, 1st Q, median, 3rd Q, max)
- Sample mean and variance

$$\overline{x} = \sum_{i=1}^{n} x_i \qquad s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

```
> summary(cadmium)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.200   0.800   2.100   3.246   3.850  18.100
> var(cadmium)
[1] 12.41678
```

Summary statistics (2)

- Sample standard deviation (same units as mean), CV

$$s_X = \sqrt{s_X^2} \qquad CV = \frac{s_X}{\overline{x}}$$

```
> sd(cadmium)
[1] 3.523746
> sqrt(var(cadmium))
[1] 3.523746
> round((sqrt(var(cadmium))/mean(cadmium))*100,0)
[1] 109
```

Cautions

- The quantiles, including the median, are always meaningful
- The mean and variance are mathemtically meaningful, but not so useful unless the sample is “approximately” normal
- This imples one population (**unimodal**)

```
> quantile(cadmium, probs=seq(0, 1, .1))
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
 0.20  0.20  0.64  1.20  1.56  2.10  2.64  3.10  5.64  8.26 18.10
```

Precision of the sample mean

- Standard error of the mean: standard deviation adjusted by sample size

$$s_e = \frac{s_X}{\sqrt{n}}$$

- This is also written as $s_{\overline{X}}$

- Note that increasing sample size increases precision of the estimate (but as $\sqrt{n}$, not n)

```
> sd(cadmium)/sqrt(length(cadmium))
[1] 0.2830341
```

Confidence interval of the sample mean

- Estimated from sample mean and standard error, using the t distribution.
- Distribution the estimates of the *mean* is normal, even if the distribution of the *variable* isn't.

Test against null hypothesis of 0 (*usually not very interesting*):

```
> t.test(cadmium)
t = 11.4679, df = 154, p-value = < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.68668 3.80494
sample estimates:
mean of x
  3.24581
```

Test whether *less* than a **target value**; user must set α (confidence level):

```
> t.test(cadmium, alt="less", mu=3, conf.level = .99)
t = 0.8685, df = 154, p-value = 0.8068
alternative hypothesis: true mean is less than 3
99 percent confidence interval:
    -Inf 3.91116
sample estimates:
mean of x
  3.24581
```

Note that in this case the confidence interval is *one sided*: from $3 \ldots 3.91116$; we don't care what the mean is if it's less than 3.

Populations & Outliers

- Most samples from "nature" are quite small
- Even if the assumption of one population with a normal distribution is true, by chance we can get extreme values
- How can we determine whether an "unusual" value is an **outlier**?
- How can we determine whether we have several populations?
- Answer: look for an underlying factor (**co-variate**), separate into sub-populations and test their difference

Topic: Bivariate EDA and correlation analysis

- "**Bi**variate": **two** variables which we suspect are related
- Question: what is the nature of the relation?
- Question: how strong is the relation?

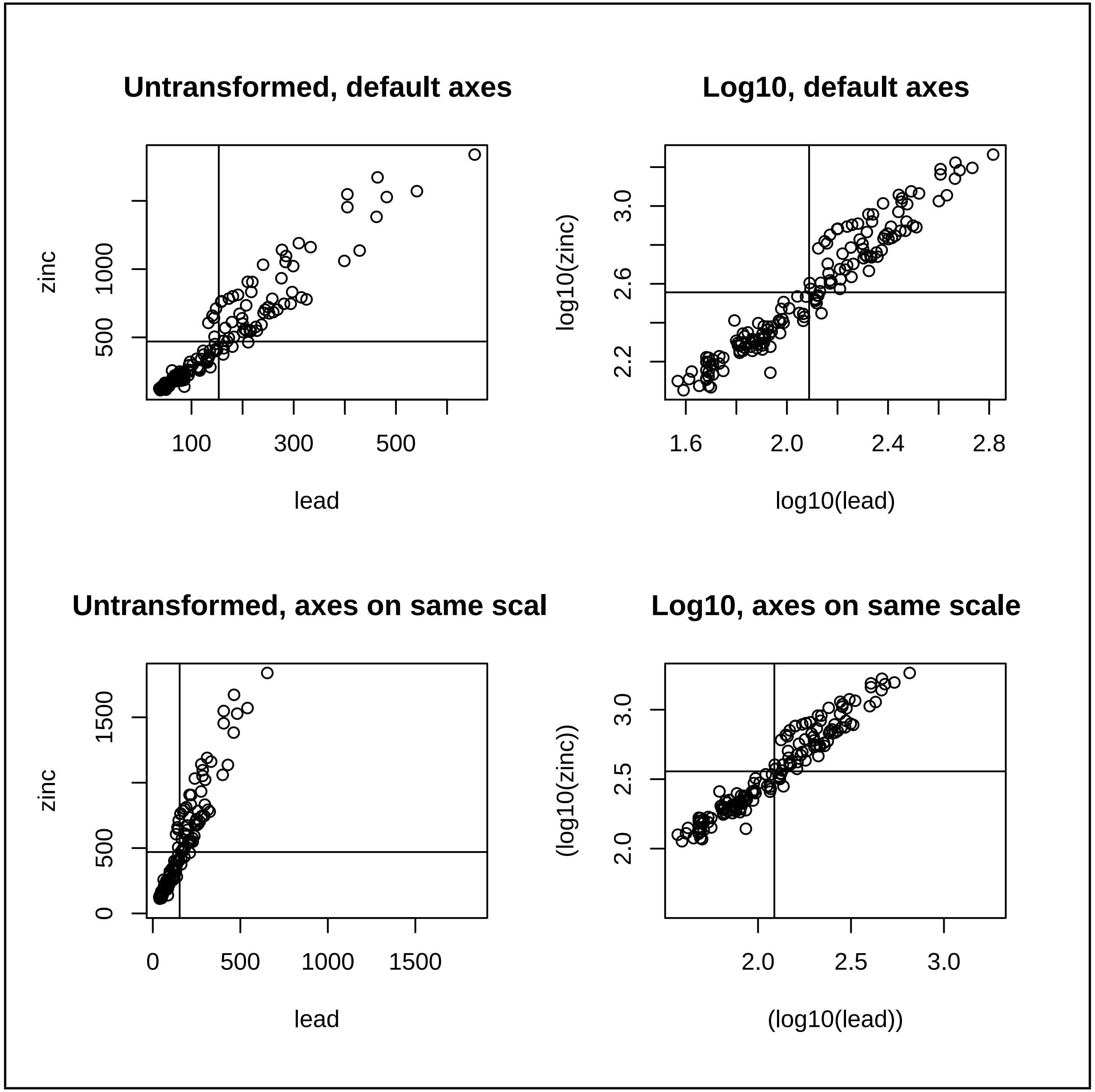
Bivariate scatterplot

- Shows the relation of two variates in **feature space** (a plane made up of the two variables' ranges)

- Display two ways:
  - **Non-standardized**: with original values on the axes (and same zero); shows relative *magnitudes*
  - **Standardized** to zero sample means and unit variances: shows relative *spreads*
  - Note: some displays automatically scale the axes, so that non-standardized looks like standardized

Scatterplots of two heavy metals; automatic vs. same scales; also log-transformed; standardized and not.

```
> plot(lead,zinc)
> abline(v=mean(lead)); abline(h=mean(zinc))
> lim<-c(min(min(lead,zinc)), max(max(lead,zinc)))
> plot(lead, zinc, xlim=lim, ylim=lim)
> abline(v=mean(lead)); abline(h=mean(zinc))
> plot(log(lead), log(zinc))
> abline(v=mean(log(lead))); abline(h=mean(log(zinc)))
> plot(log(lead), log(zinc), xlim=log(lim), ylim=log(lim))
> abline(v=mean(log(lead))); abline(h=mean(log(zinc)))
> sdze<-function(x) { (x-mean(x))/sd(x) }
> plot(sdze(lead), sdze(zinc)); abline(h=0);abline(v=0)
> plot(sdze(log(lead)), sdze(log(zinc))); abline(h=0); abline(v=0)
```


Untransformed, default axes
zinc
lead
100
300
500
1000
Log10, default axes
log10(zinc)
log10(lead)
1.6
2.0
2.4
2.8
2.2
2.6
3.0
Untransformed, axes on same scal
zinc
lead
0
500
1000
1500
Log10, axes on same scale
(log10(zinc))
(log10(lead))
2.0
2.5
3.0

Measuring the strength of a bivariate relation: theoretical

- The *theoretical covariance* of two variables X and Y

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E\{(X-\mu_X)(Y-\mu_Y)\} \\ &= \sigma_{XY}\end{aligned}$$

- The *theoretical correlation coefficient*: covariance normalized by population standard deviations; range $[-1 \ldots 1]$:

$$\begin{aligned}\rho_{XY} &= \frac{\mathrm{Cov}(XY)}{\sigma_X \cdot \sigma_Y} \\ &= \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}\end{aligned}$$

Measuring the strength of a bivariate relation: estimate from sample

In practice, we estimate **population** covariance and correlation from a **sample**:

$$
\begin{aligned}
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y}) \\
r_{xy} &= \frac{s_{xy}}{s_x \cdot s_y} \\
&= \frac{\sum(x_i-\bar{x})\cdot\sum(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\cdot\sum(y_i-\bar{y})^2}}
\end{aligned}
$$

Sample vs. population covariance and correlation

- Sample $\bar{x}$ estimates population μ_X
- Sample s_x estimates population σ_X
- Sample r_{xy} estimates population ρ_{XY}

Example of correlation & confidence interval: positive, strong

```
> cor.test(lead,zinc)

        Pearson's product-moment correlation

data:  lead and zinc
t = 39.6807, df = 153, p-value = < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9382556 0.9668269
sample estimates:
      cor
0.9546913
```

This explains $0.955^2 = 0.912$ of the total variance.

Example of correlation & confidence interval: negative, weak

```
> cor.test(lead,dist.m)

        Pearson's product-moment correlation

data:  lead and dist.m
t = -8.9269, df = 153, p-value = 1.279e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6801118 -0.4710150
sample estimates:
       cor
-0.5852087
```

This explains $-0.585^2 = 0.342$ of the total variance.

Topic: Regression

- A general term for *modelling* the distribution of one variable (the **response** or **dependent**) from ("on") another (the **predictor** or **independent**)

- This is only logical if we have *a priori* (non-statistical) reasons to believe in a *causal relation*

- Correlation: makes no assumptions about causation; both variables have the same logical status

- Regression: assumes one variable is the predictor and the other the response

Actual vs. fictional ‘causality’

- Example: proportion of fine sand in a topsoil and subsoil layer
- Does one “cause” the other?
- Do they have a common cause?
- Can one be used to **predict** the other?
- Why would this be useful?

Simple linear regression (one predictor)

- Model: $y = \beta_0 + \beta_1 x + \varepsilon$
- β_0: **intercept**, constant shift from $\bar{x}$ to $\bar{y}$
- β_1: **slope**, change in y for an equivalent change in x
- ε: **error**, or better, unexplained variation
- The parameters β_0 and β_1 are selected to minimize some summary measure of ε over all sample points

Simple regression (continued)

- Given the fitted model, we can **predict** at the original data points: $\hat{y}_i$; these are called the **fitted values**

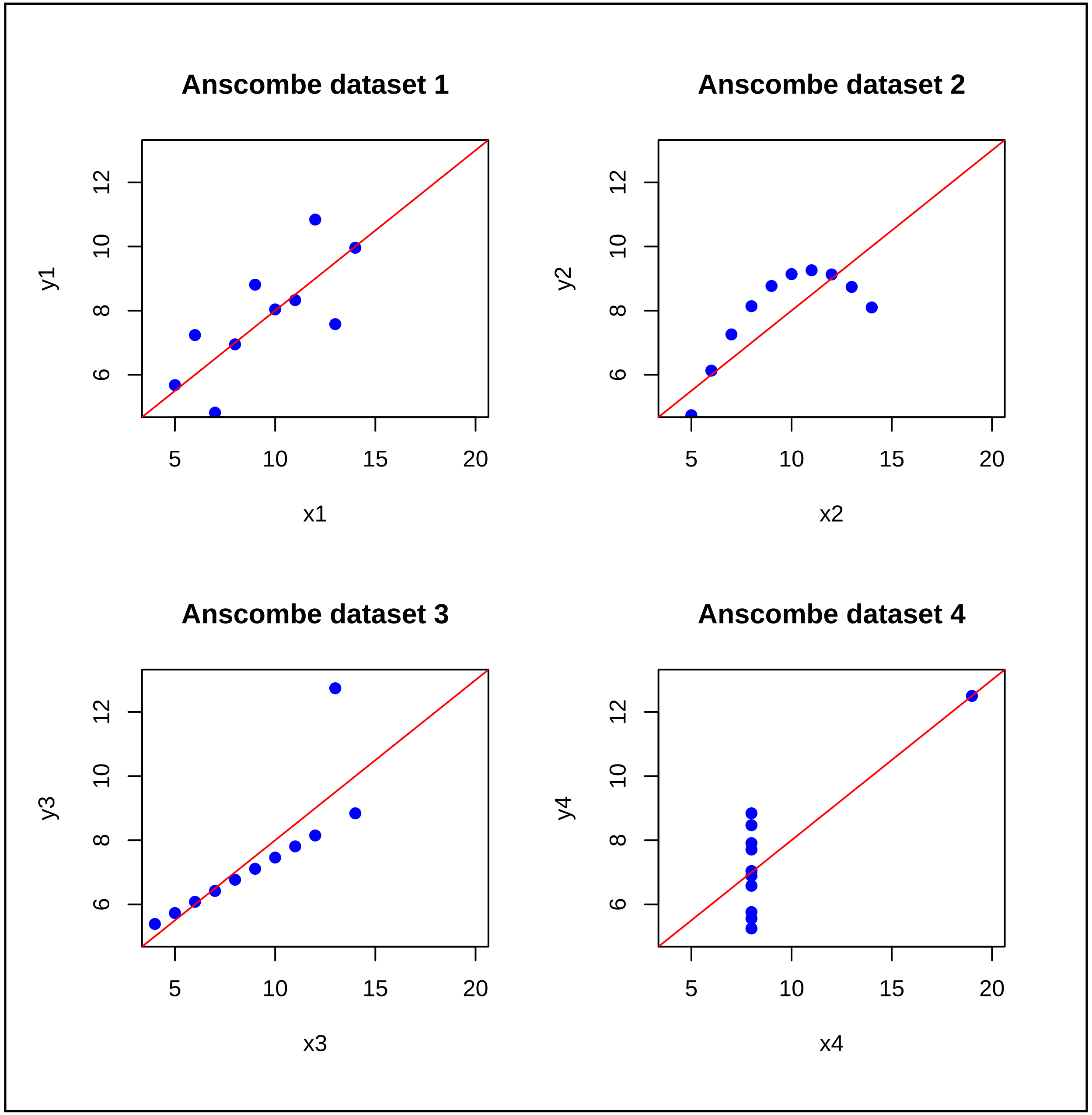
- Then we can compute the **deviations** of the fitted from the measured values: $\hat{e}_i = (\hat{y}_i - y_i)$; these are called the **residuals**

- The deviations can be summarized to give an overall **goodness-of-fit** measure

Look before you leap!

Anscombe developed four different bivariate datasets, all with the exact same correlation $r = 0.81$ and linear regression $y = 3 + 0.5x$:

1. bi-variate normal
2. quadratic
3. bi-variate normal with one outlier
4. one high-leverage point


Anscombe dataset 1
Anscombe dataset 2
Anscombe dataset 3
Anscombe dataset 4
y1
y2
y3
y4
x1
x2
x3
x4
6
8
10
12
5
10
15
20

Least squares estimates

- Compute the parameters to *minimize* the *sum* of the *squared* deviations

- **Slope**: $\beta_1 = s_{XY}/s_X^2$

- Note the similarly with covariance, except here we standardize by the *predictor* only, so the regression of x on y gives a different slope than that of y on x

- **Intercept**: To make the fitted and sample means co-incide: $\beta_0 = \bar{y} - \beta_1 \bar{x}$

Sums of squares

- The regression **partitions the variability** in the sample into two parts:

  1. **explained by the model**
  2. **not explained**, left over, i.e. **residual**

- Note we always know the **mean**, so the **total** variability refers to the variability *around the mean*

- Question: **how much more of the variability is explained by the model?**

- Total SS = Regression SS + Residual SS

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- The least squares estimate maximizes the Regression SS and minimizes the Residual SS

Analysis of Variance (ANOVA)

- Partition the *total* variance in a population into the *model* and *residual*
- If the model has more than one term, also partition the model variance into *components* due to each term
- Can be applied to any linear additive *design* specified by a *model*
- Each component can be tested for *signficance* vs. the null hypothesis that it does not contribute to the model fit

ANOVA for simple linear regression

- *total* sum of squared deviations is divided into *model* (regression) and *error* (residual) sums of squares

- Their ratio is the *coefficient of determination* R^2

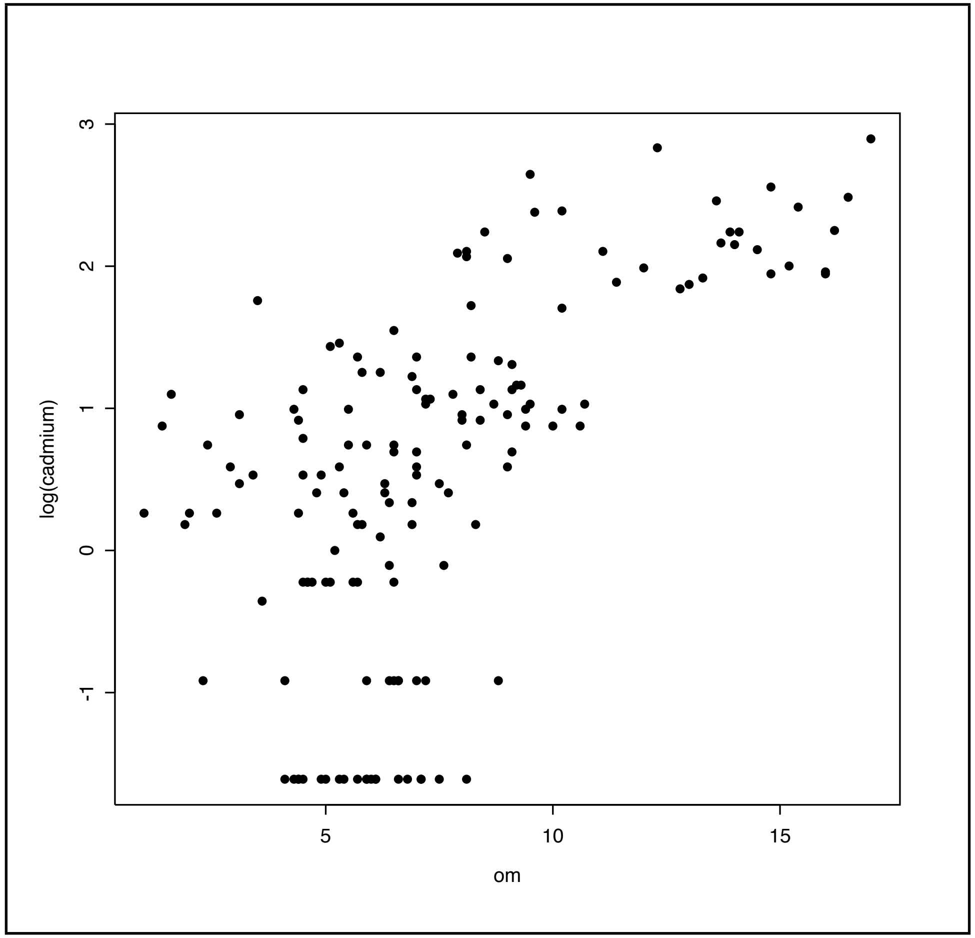
- These are each divided by their *degrees of freedom* to obtain the *mean* SS

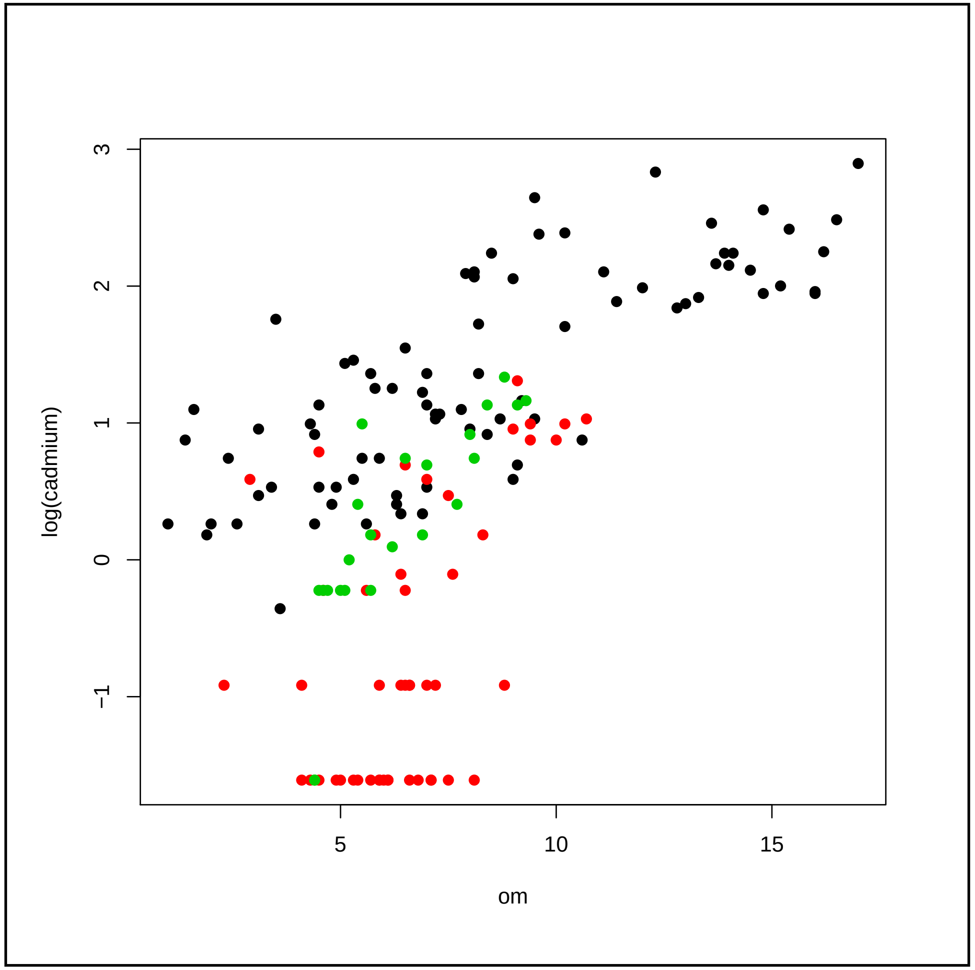
- Their ratio is distributed as F and can be tested for significance

Bivariate analysis: heavy metals vs. organic matter

- Scatterplot
- Scatterplot by flood frequency
- Regression of metal on organic matter (why this order?)
- Same, including flood frequency in the model

```
> plot(om,log(cadmium))
> plot(om, log(cadmium), col=as.numeric(ffreq), cex=1.5, pch=20)
```





Note the additional information we get from visualising the flood frequency class.

Model: Regression of metal on organic matter

```
> m<-lm(log(cadmium) ~ om)
> summary(m)
Residuals:
    Min      1Q  Median      3Q     Max
-2.3070 -0.3655  0.1270  0.6079  2.0503
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.04574    0.19234  -5.437 2.13e-07 ***
om           0.21522    0.02339   9.202 2.70e-16 ***
---
Residual standard error: 0.9899 on 151 degrees of freedom
Multiple R-Squared: 0.3593,     Adjusted R-squared: 0.3551
F-statistic: 84.68 on 1 and 151 DF,  p-value: 2.703e-16
```

Highly-significant model, but organic matter content explains only about 35% of the variability of log(Cd).

Good fit vs. significant fit

- R^2 can be highly significant (reject null hypothesis of no relation), but . . .
- . . . the prediction can be poor
- In other words, only a “small” portion of the variance is explained by the model
- Two possiblities

1. **incorrect** or **incomplete model**
   (a) other factors are more predictive
   (b) other factors can be included to improve the model
   (c) form of the model is wrong
2. correct model, **noisy data**
   (a) imprecise **measurement** method . . .
   (b) . . . or just an inherently variable **process**

Regression diagnostics

- Objective: to see if the regression truly represents the presumed relation
- Objective: to see if the computational methods are adequate
- Main tool: plot of *standardized* **residuals** vs. **fitted** values
- Numerical measures: *leverage*, *large residuals*

Examining the scatterplot with the fitted line

- Is there a **trend in lack of fit**? (further away in part of range)
  - * $\longrightarrow$ a non-linear model

- Is there a **trend in the spread**?
  - * $\longrightarrow$ *heteroscedasity* (unequal variances) so linear modelling is invalid

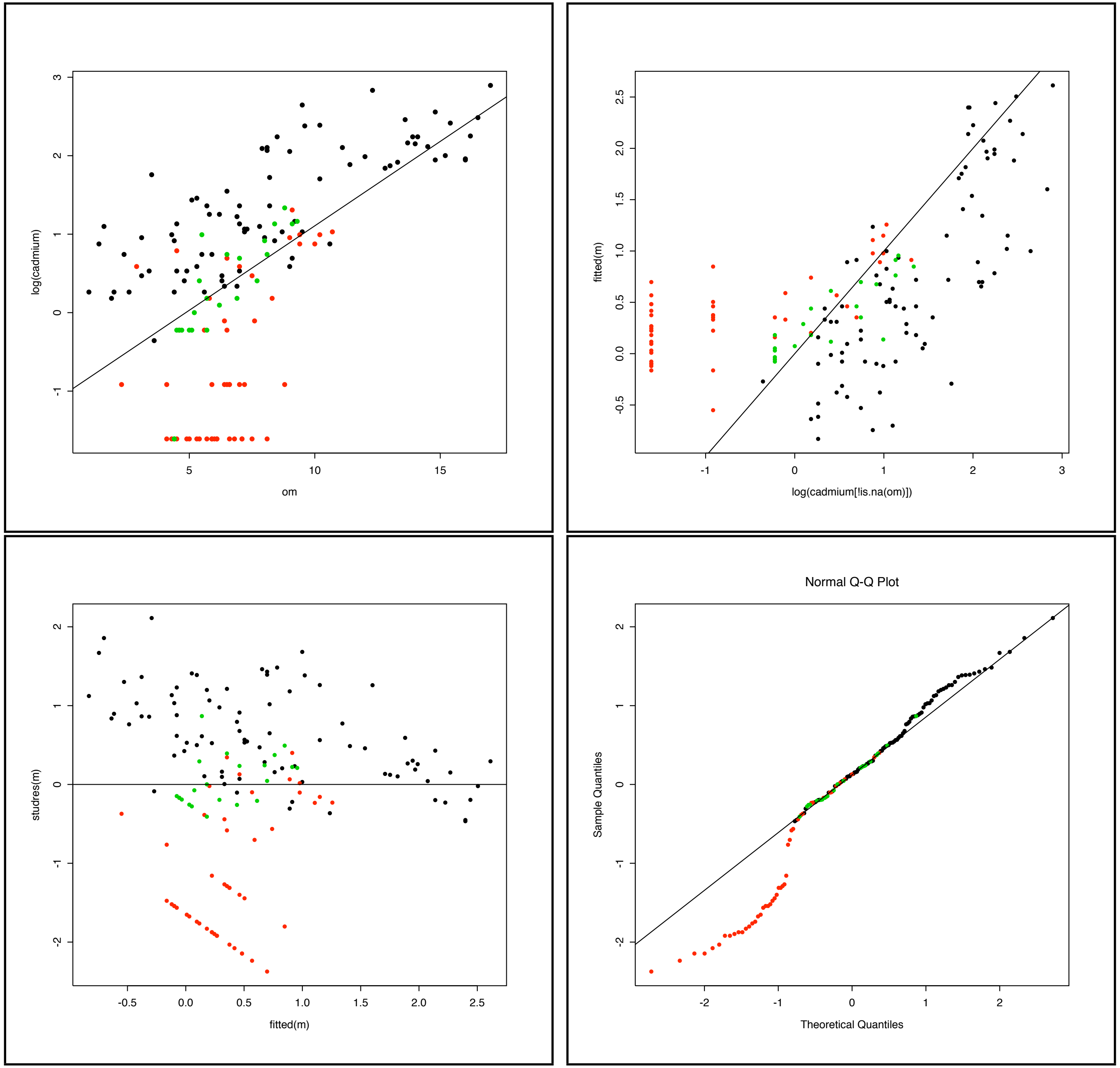
- Are there **high-leverage observations** that, if eliminated, would **substantially change the fit**?
  - * $\longrightarrow$ high *leverage*, isolated in the range and far from other points

Model diagnostics: regression of metal on organic matter

```
> m<-lm(log(cadmium) ~ om)
> plot(om, log(cadmium), col=as.numeric(ffreq), cex=1.5, pch=20); abline(m)
> plot(log(cadmium[!is.na(om)]),fitted(m), col=as.numeric(ffreq), pch=20)
> abline(0,1)
> plot(fitted(m),studres(m), col=as.numeric(ffreq), pch=20)
> abline(h=0)
> qqnorm(studres(m), col=as.numeric(ffreq), pch=20);qqline(studres(m))
```

- We can see problems at the low metal concentrations. This is probably an artifact of the measurement precision at these levels (near or below the detection limit).

- These are almost all in flood frequency class 3 (rarely flooded).



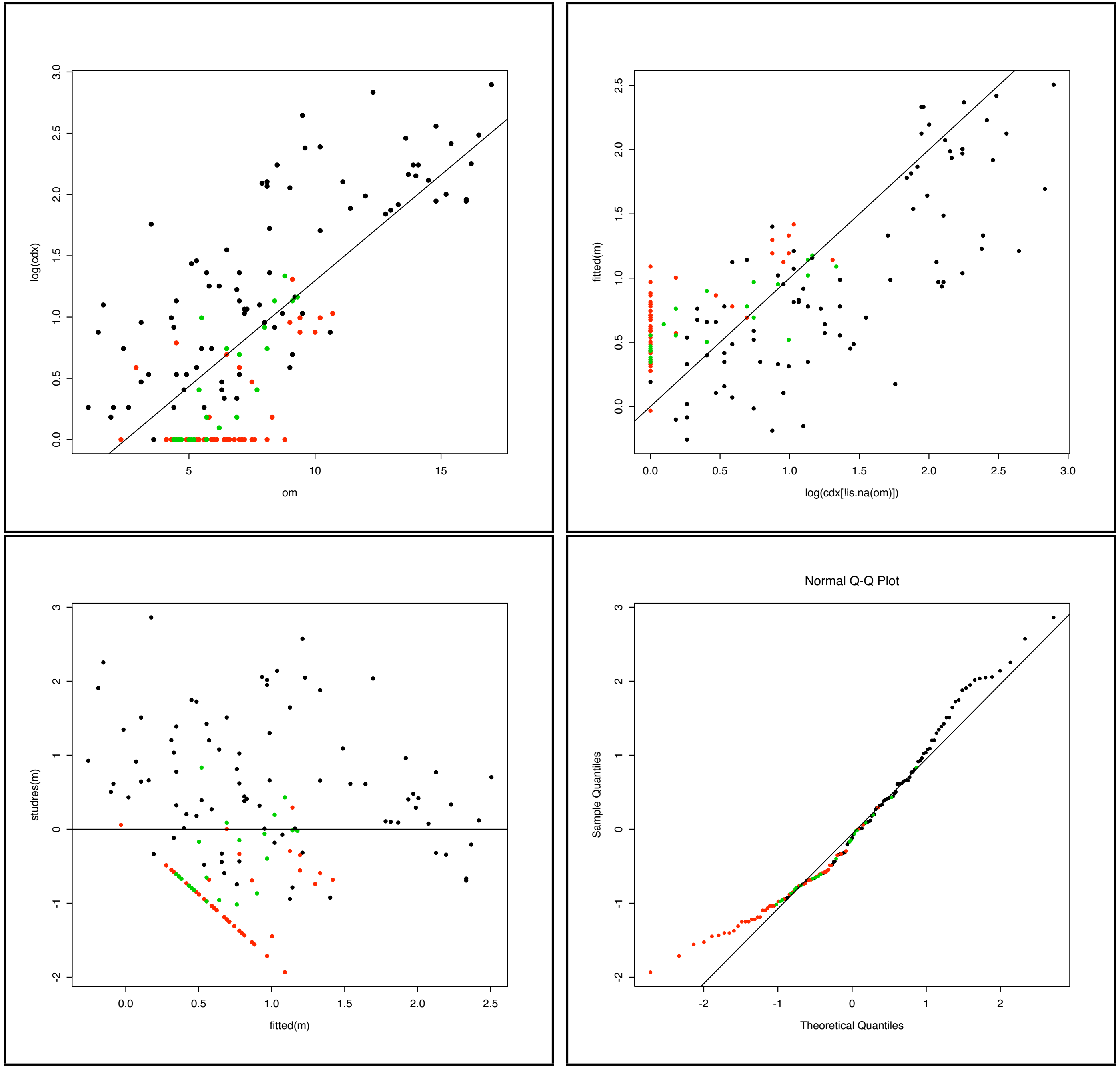
log(cadmium)
om
fitted(m)
log(cadmium[!is.na(om)])
studres(m)
fitted(m)
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

Revised model: Cd detection limit

Values of Cd below 1mg kg^{-1} are unreliable; replace them all with 1mg kg^{-1} and re-analyze:

```
> cdx<-ifelse(cadmium>1, cadmium, 1)
> plot(om, log(cdx), col=as.numeric(ffreq), cex=1.5, pch=20)
> m<-lm(log(cdx) ~ om); summary(m)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0896 -0.4250 -0.0673  0.3527  1.5836
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.43030    0.11092  -3.879 0.000156 ***
om           0.17272    0.01349  12.806  < 2e-16 ***
---
Residual standard error: 0.5709 on 151 degrees of freedom
Multiple R-Squared: 0.5206,Adjusted R-squared: 0.5174
F-statistic:   164 on 1 and 151 DF,  p-value: < 2.2e-16
> abline(m)
> plot(log(cdx[!is.na(om)]),fitted(m),col=as.numeric(ffreq),pch=20); abline(0,1)
> plot(fitted(m),studres(m),col=as.numeric(ffreq),pch=20); abline(h=0)
> qqnorm(studres(m),col=as.numeric(ffreq),pch=20); qqline(studres(m))
```

Much higher R^2 and better diagnostics. Still, there is a lot of spread at any value of the predictor (organic matter).

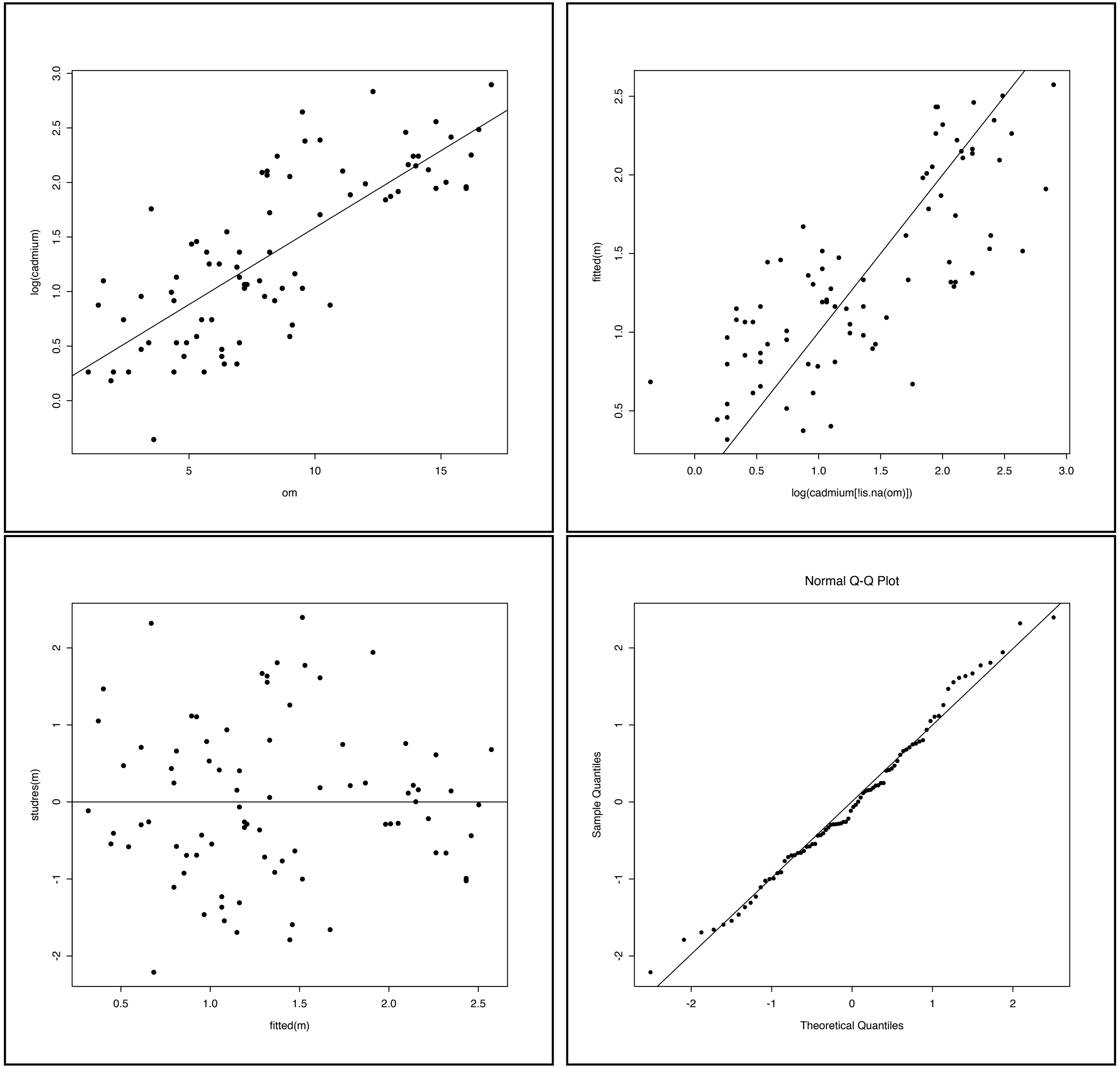

log(cdx)
om
fitted(m)
log(cdx[!is.na(om)])
studres(m)
fitted(m)
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

Revised model: flood class 1

The relation looks more consistent in the frequently-flooded soils; re-analyze this subset.

```
> meuse.1<-meuse[ffreq==1,]; attach(meuse.1)
> plot(om, log(cadmium), cex=1.6, pch=20)
> m<-lm(log(cadmium) ~ om); summary(m)
Residuals:
     Min       1Q   Median       3Q      Max
-1.04064 -0.31782 -0.04348  0.32210  1.13034
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17639    0.11999    1.47    0.145
om           0.14099    0.01286   10.96   <2e-16 ***
---
Residual standard error: 0.4888 on 80 degrees of freedom
Multiple R-Squared: 0.6003,Adjusted R-squared: 0.5954
F-statistic: 120.2 on 1 and 80 DF,  p-value: < 2.2e-16
> abline(m)
> plot(log(cadmium[!is.na(om)]),fitted(m)); abline(0,1)
> plot(fitted(m),studres(m)); abline(h=0)
> qqnorm(studres(m)); qqline(studres(m))
```

Still higher R^2 and excellent diagnostics. There is still a lot of spread at any value of the predictor (organic matter), so OM is not an efficient predictor of Cd.


log(cadmium)
om
0.0
0.5
1.0
1.5
2.0
2.5
3.0
5
10
15
fitted(m)
log(cadmium[!is.na(om)])
0.5
1.0
1.5
2.0
2.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
studres(m)
fitted(m)
-2
-1
0
1
2
0.5
1.0
1.5
2.0
2.5
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
-2
-1
0
1
2

Categorical ANOVA

- Model the *response* by a *categorical* variable (nominal); ordinal variables are treated as nominal
- Model: $y = \beta_0 + \beta_j x + \varepsilon$; where each observation x is multiplied by the $beta_j$ corresponding to the class to which it belongs (of n classes)
- The β_j represent the deviations of each class mean from the grand mean

Example: Meuse soil pollution

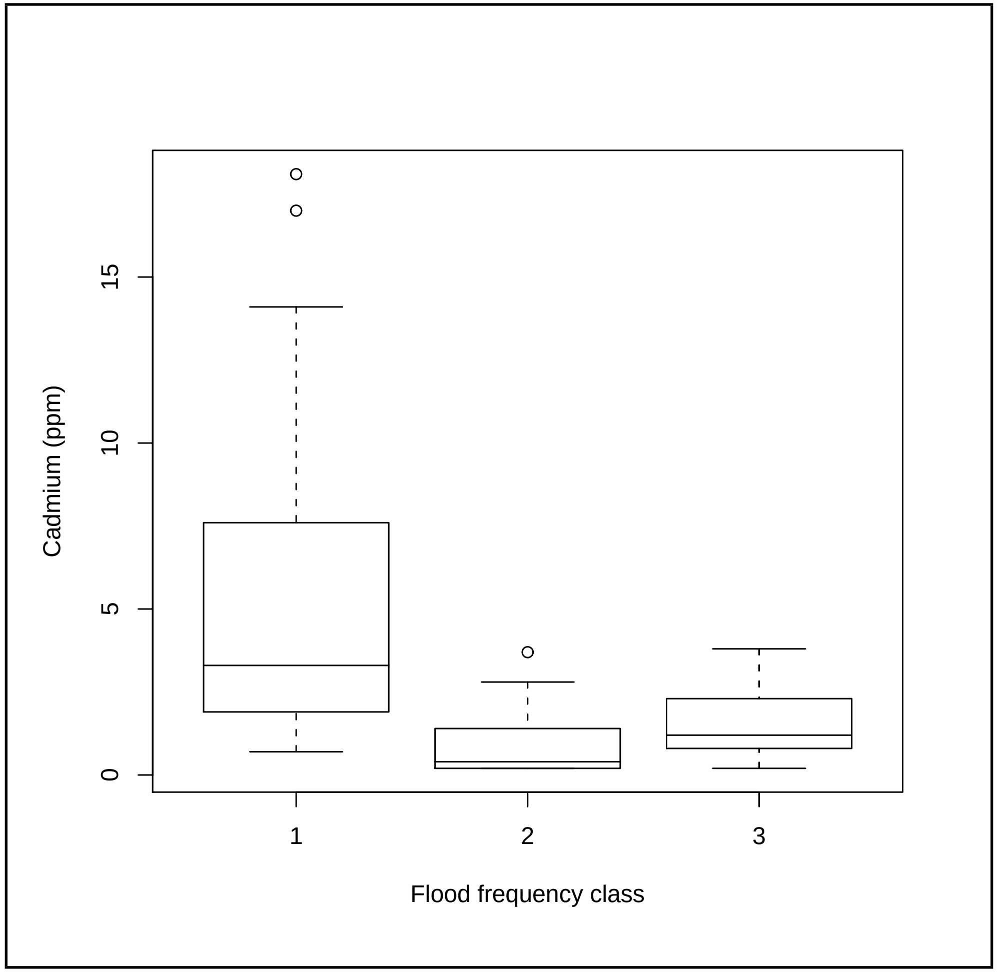
- Question: do metals depend on flood frequency (3 of these)

- EDA: categorical boxplots

- Analysis: one-way ANOVA on the frequency

Categorical EDA

```
> boxplot(cadmium ~ ffreq,xlab="Flood frequency class",ylab="Cadmium (ppm)")
```



Example ANOVA

```
> m<-lm(log(cadmium) ~ ffreq)
> summary(m)
Residuals:
    Min      1Q  Median      3Q     Max
-1.8512 -0.7968 -0.1960  0.7331  1.9354
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.32743    0.09351  14.196  < 2e-16 ***
ffreq2      -1.95451    0.15506 -12.605  < 2e-16 ***
ffreq3      -1.08566    0.20168  -5.383 2.72e-07 ***

Residual standard error: 0.857 on 152 degrees of freedom
Multiple R-Squared: 0.5169,     Adjusted R-squared: 0.5105
F-statistic: 81.31 on 2 and 152 DF,  p-value: < 2.2e-16
```

Difference between classes

```
> TukeyHSD(aov(log(cadmium) ~ ffreq))
  Tukey multiple comparisons of means,
  95% family-wise confidence level

Fit: aov(formula = log(cadmium) ~ ffreq)

$ffreq
          diff        lwr        upr
2-1 -1.9545070 -2.3215258 -1.5874882
3-1 -1.0856629 -1.5630272 -0.6082986
3-2  0.8688442  0.3544379  1.3832504
```

All per-pair class differences are significant (confidence interval does not include zero).

Non-parametric statistics

A **non-parametric** statistic is one that does not assume any underlying data distribution.

For example:

- a **mean** is an estimate of a parameter of location of some **assumed distribution** (e.g.mid-point of normal, expected proportion of success in a binomial, . . .)
- a **median** is simply the value at which half the samples are smaller and half larger, without knowing anything about the distribution underlying the process which produced the sample.

So “non-parametric” inferential methods are those that make no assumptions about the distribution of the data values, only their **order** (rank).

Non-parametric statistics: Correlation

As an example of the non-parameteric approach, consider the measure of **association between two variables**, commonly called *correlation* ('co-relation').

The standard measure is *parametric*, i.e. the Pearson's Product Moment Correlation (PPMC); this is computed from the sample covariance of the two variables:

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\{(x_i-\bar{x})(y_i-\bar{y})\}$$

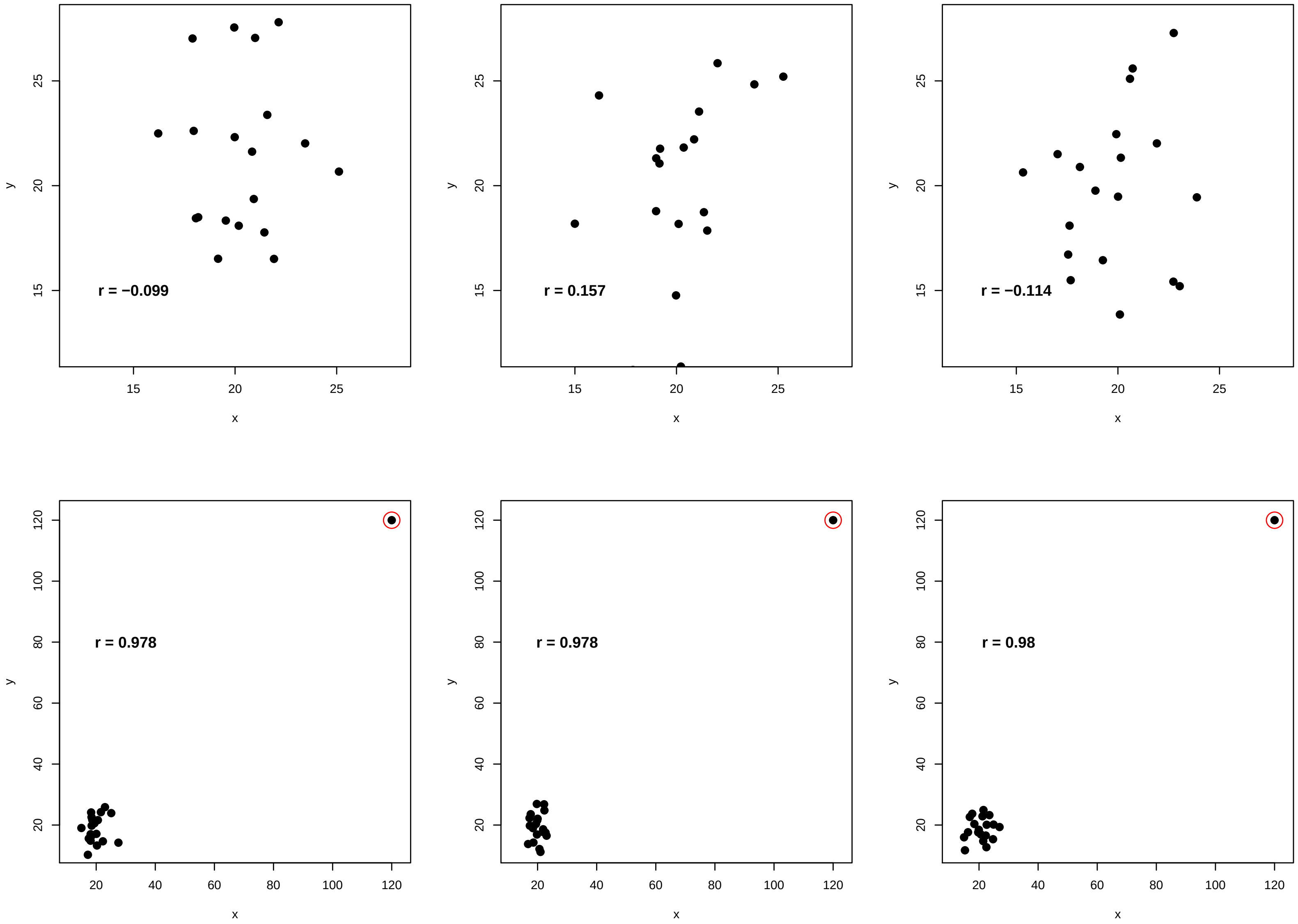
Then the *sample Pearson's correlation coefficient* is computed as:

$$r_{XY} = \mathrm{Cov}(X,Y)/s_X \cdot s_Y$$

Parametric correlation – example of inappropriate use

Consider the following two cases: (1) 20 bivariate normal samples that should be uncorrelated; (2) same, but with one value replaced by a very high value (no longer a normal distribution).

```
n<-20
par(mfrow=c(2,3))
for (i in 1:3)
  { x<-rnorm(n, 20, 3); y<-rnorm(n, 20, 4);
    plot(x,y, pch=20, cex=2, xlim=c(12,28), ylim=c(12,28));
    text(15,15, paste("r =",round(cor(x,y),3)), font=2, cex=1.2)
  }
for (i in 1:3)
  { x<-c(rnorm((n-1), 20, 3), 120); y<-c(rnorm((n-1), 20, 4), 120);
    plot(x,y, pch=20, cex=2, xlim=c(12, 122), ylim=c(12, 122));
    points(120, 120, col="red", cex=3);
    text(30,80, paste("r =",round(cor(x,y),3)), font=2, cex=1.2)
  }
```



r = −0.099
r = 0.157
r = −0.114
r = 0.978
r = 0.978
r = 0.98
x
y

Non-parametric correlation

The solution here is to use a method such as *Spearman's* correlation, which correlates the **ranks**, not the **values**; therefore the distribution ("gaps between values") has no influence.

From numbers to ranks:

```
> n<-10
> (x<-rnorm(n, 20, 4))
 [1] 15.1179 23.7801 21.2801 21.5191 23.0096 18.5065 19.1448 24.9254 29.3211
[10] 14.1453
> (ix<-(sort(x, index=T)$ix))
 [1] 10  1  6  7  3  4  5  2  8  9
```

If we change the largest of these to any large value, the rank does not change:

```
> x[ix[n]]<-120; x
 [1]  15.1179  23.7801  21.2801  21.5191  23.0096  18.5065  19.1448  24.9254
 [9] 120.0000  14.1453
> (ix<-(sort(x, index=T)$ix))
 [1] 10  1  6  7  3  4  5  2  8  9
```

Compare the two correlation coefficients:

```
pearsons<-vector(); spearmans<-vector()
> n<-10
> for (i in 1:n)
+  { x<-rnorm(n, 20, 4); y<-rnorm(n, 20, 4);
+    pearsons[i]<-cor(x,y);
+    spearmans[i]<-cor(x,y, method="spearman")}
> round(pearsons, 2); round(spearmans, 2)
 [1] -0.29 -0.02 -0.49 -0.01 -0.17  0.16  0.06 -0.07 -0.11  0.37
 [1]  0.32  0.16 -0.25  0.01  0.35 -0.42  0.03 -0.33  0.68 -0.12

> for (i in 1:n)
+  { x<-c(rnorm((n-1), 20, 4), 120); y<-c(rnorm((n-1), 20, 4), 120);
+    pearsons[i]<-cor(x,y);
+    spearmans[i]<-cor(x,y, method="spearman") }
> round(pearsons, 2); round(spearmans, 2)
 [1] 0.98 0.99 0.98 0.99 0.98 0.98 0.99 0.99 0.99 0.99
 [1]  0.25  0.08  0.49  0.03  0.61 -0.04  0.36  0.26 -0.25  0.36
```

The Pearson's (parametric) coefficient is completely changed by the one high-valued pair, whereas the Spearman's is unaffected.

Other non-parametric methods

- t-test for equivalence of *means* $\longrightarrow$ **Mann-Whitney** test for equivalence of *medians*

- One-way ANOVA $\longrightarrow$ **Kruskal-Wallis**

- χ^2 goodness-of-fit $\longrightarrow$ **Kolmogorov-Smirnov** goodness-of-fit