Lecture 3: Statistics and Graphics

Math 586

Data = $\{x_1, x_2, ..., x_n\}$ (usually a sample from population). Goal: explore data to produce numerical summaries and graphs (hopefully, reveal some structure). For now, assume that x_i's are independent realizations from a random variable X. Multivariate data: random vectors $\mathbf{x}_i$ are observed (several variables per observation).

Statistics

- **Sample variance** is an estimate for $\sigma^2 = Var(X)$.

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

  Sample standard deviation s. Note: denominator of $(n-1)$ makes the estimate unbiased, that is, $\mathbb{E}\left(s^2\right) = \sigma^2$.

- **Sample correlation coefficient**, bivariate data $\mathbf{x}_i = (x_i, y_i)$.

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}$$

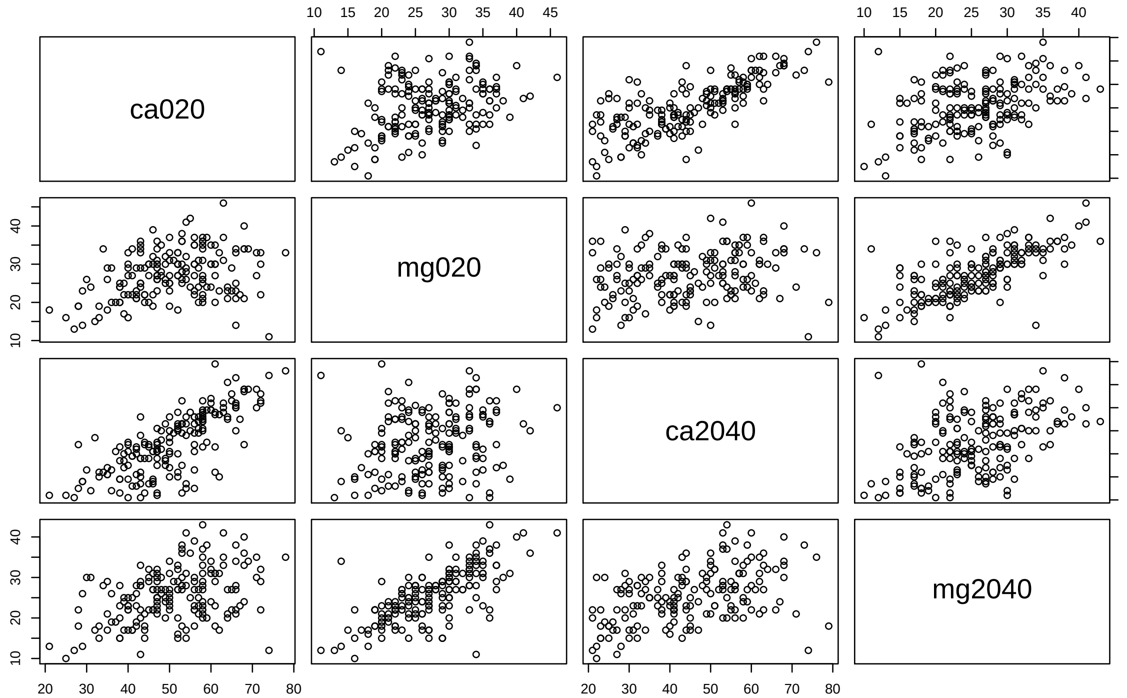
  $-1 \le r \le 1$, shows the direction and strength of *linear* relationship between X, Y.

Bivariate graphs

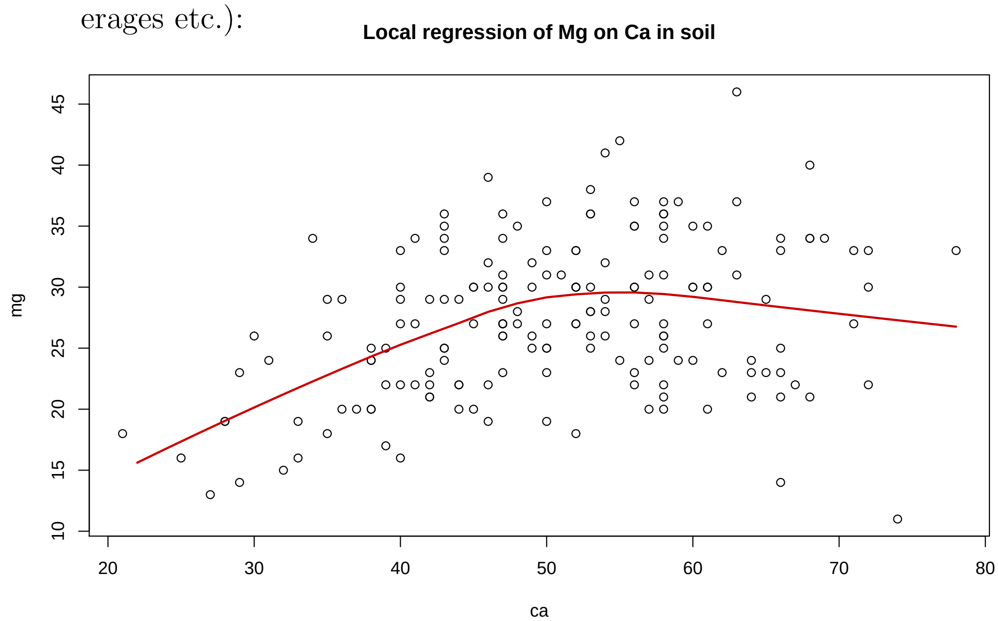
- Scatterplots.

  Scatterplot matrices are useful when exploring several variables.
  Example: data set `camg` from `geoR` library

  `ca020` = calcium content in the 0-20cm soil layer, measured in $mmol_c/dm^3$.
  `mg020` = magnesium content in the 0-20cm soil layer, measured in $mmol_c/dm^3$.
  `ca2040, mg2040` = same in the 20-40cm soil layer



- Local Trend estimates (e.g. local regression `loess`, splines, moving averages etc.):



Univariate graphs

** boxplots, histograms, QQ plots, Norm. Prob. plots

Example: Islands - areas of biggest land masses

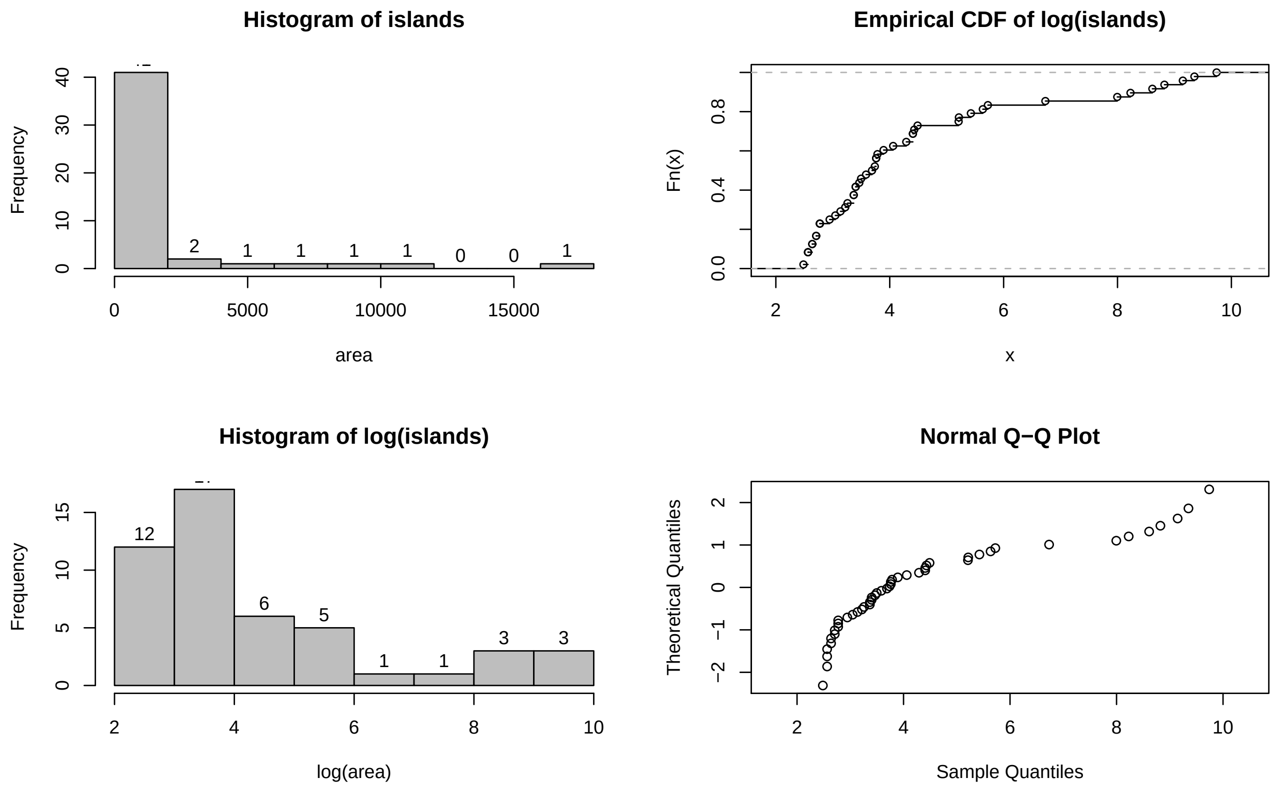
Africa	Antarctica	Asia	Australia
11506	5500	16988	2968
Axel Heiberg	Baffin	Banks	Borneo
16	184	23	280
Britain	Celebes	Ceylon	Cuba
84	73	25	43
Devon	Ellesmere	Europe	Greenland
21	82	3745	840
Hainan	Hispaniola	Hokkaido	Honshu
13	30	30	89
Iceland	Ireland	Java	Kyushu
40	33	49	14
Luzon	Madagascar	Melville	Mindanao
42	227	16	36
Moluccas	New Britain	New Guinea	New Zealand (N)
29	15	306	44
New Zealand (S)	Newfoundland	North America	Novaya Zemlya
58	43	9390	32
Prince of Wales	Sakhalin	South America	Southampton
13	29	6795	16
Spitsbergen	Sumatra	Taiwan	Tasmania
15	183	14	26
Tierra del Fuego	Timor	Vancouver	Victoria
19	13	12	82

- Histograms
- **Quantiles**

Definition
p^{th} **quantile** of a distribution (or a sample) is the point with $p\%$ of data below it, or a solution to the equation $F(x) = p$.

Important quantiles are: median (50th), first and third quartiles (25th and 75th).

Algorithm for computation from a sample:
a) order the data
b) take the $p(n+1)/100\%$-th smallest observation as your quantile (may or may not interpolate).
May also estimate graphically, looking at CDF graph.



- **Transforms**

  When the data are non-normal, we may wish to change them to normal (e.g. for kriging - later).
  Most popular transforms are *log* and square root - useful for positive right-skewed data.

  Normal score transform.

  Let $\Phi(x) = \int_{-\infty}^{x} \frac{exp(-z^2/2)\, dz}{\sqrt{2\pi}}$ standard normal CDF.
  Then normal quantiles are given by inverse function $\Phi^{-1}(p)$.
  Suppose we have ordered data $x_1 \leq x_2 \leq ... \leq x_n$ we'd like to transform to normal. Then

  $$y_i = \Phi^{-1}[i/(n+1)] \quad i = 1, ..., n$$

  are approximately normal. (To see this, consider $P(Y \leq y_i) = \Phi(y)$.)
  Why divide by $(n+1)$?

- Normal quantile plots (Q-Q plots) to assess normality.

Also may check for log-normality by running a normal plot on logged data.
Plot sample quantiles on one axis and standard normal (theoretical) quantiles on the other. For example, log island data (n=48): suppose the following classes are given:

breaks	3	4	5	6	7	8	9	10
counts	12	17	6	5	1	1	3	3
cumulative counts	12	29	35	40	41	42	45	48
percent(rounded)	24	59	71	82	84	86	92	98
normal quantile	-0.71	0.23	0.55	0.92	0.99	1.08	1.41	2.05

Plot `breaks` against `normal quantile`. Looks somewhat similar to empirical CDF, but with y-axis distorted. Normal distribution will correspond to a straight line on the Normal quantile plot. The island data does not seem to fit the Normal distribution.

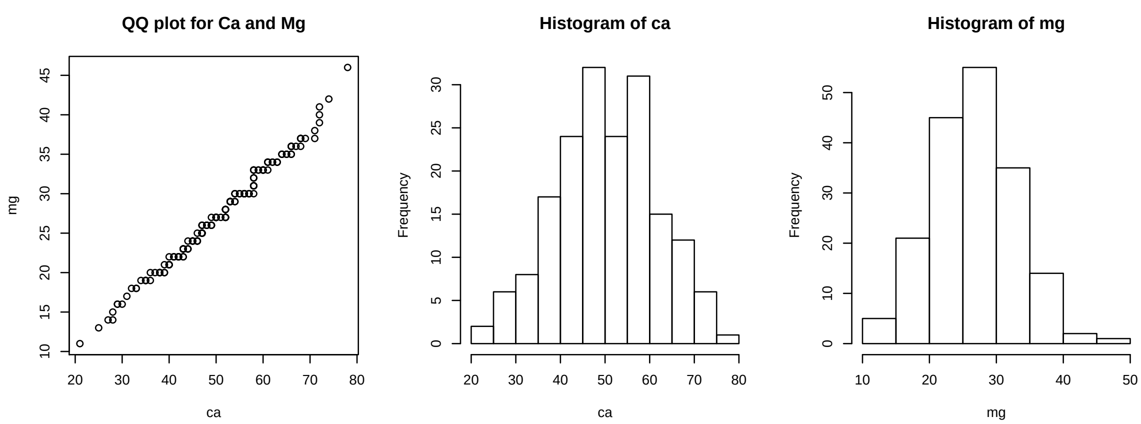
- Q-Q plots for comparing distributions.

  Let us now have two samples: from X and Y variables. (The samples are *independent*: they don't have to be from the same observational units!) We'd like to compare distributions of the two variables. Are they the same? Are they similar in some way?
  We can make a Q-Q plot similar to above, computing and plotting select quantiles.

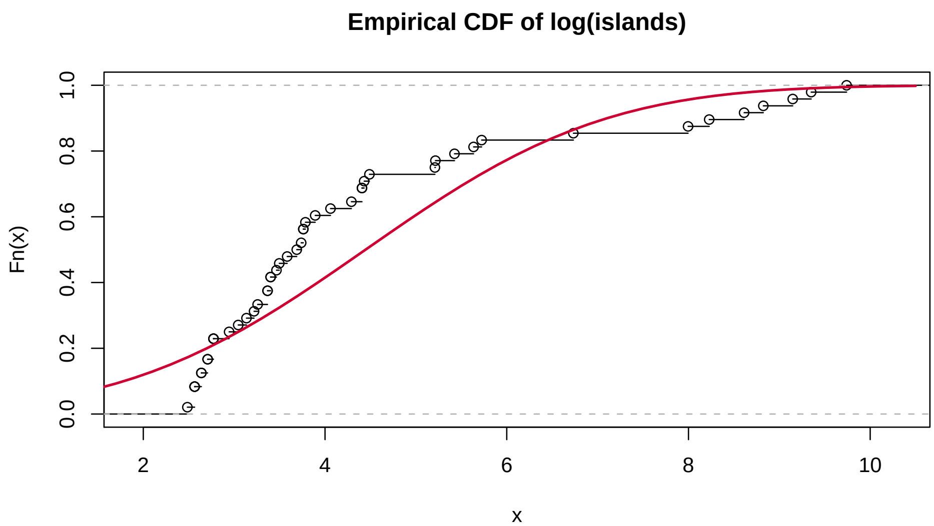
  Example (Ca/Mg data): X = ca020, Y = mg020.

percent	10%	20%	30%	40%	50%	60%	70%	80%	90%
Ca quantile	37	41	45	47	50	54	57	60	66
Mg quantile	20	22	24	26	27	29	30	33	35

  The distributions are of course not the same, but they seem to follow a straight line $\Rightarrow$ the same up to a linear transformation. In fact, they both seem normal.



- <u>Kolmogorov-Smirnov test</u> for distributions.
  Compare the empirical CDF, $\hat{F}_n(x)$ with a given CDF $F_0(x)$.
  Hypothesis test H_0 : sample $x_1, ..., x_n$ comes from a population with distribution $F_0(x)$.
  Test statistic: $D = \max_x |\hat{F}_n(x) - F_0(x)|$ = largest absolute difference between the two functions.
  Reject H_0 at 5% level if $D > 1.36/\sqrt{n}$.



E.g. for the islands example, we may compare $\hat{F}_n(x)$ with a normal CDF (sample mean and st.dev. are used as its parameters - a rather crude technique). Computed $D = 0.221 > 1.36/\sqrt{48} = 0.196$ thus we reject H_0 and take it as evidence that the distribution is **not** Normal.

Sofware also reports p-value = 0.01837. Based on it, we would also reject H_0, since p-value < 0.05.

(Note: K-S test is not the best option to test for normality, however. Anderson-Darling and Shapiro-Wilk tests are more respected in the statistical community.)

There is also a version of K-S test based on the same idea (maximum separation between CDF's) for comparing two empirical CDF's.