

Math 5060: Exploratory Data Analysis
Karen Kafadar (kk@math.cudenver.edu)

Textbooks:

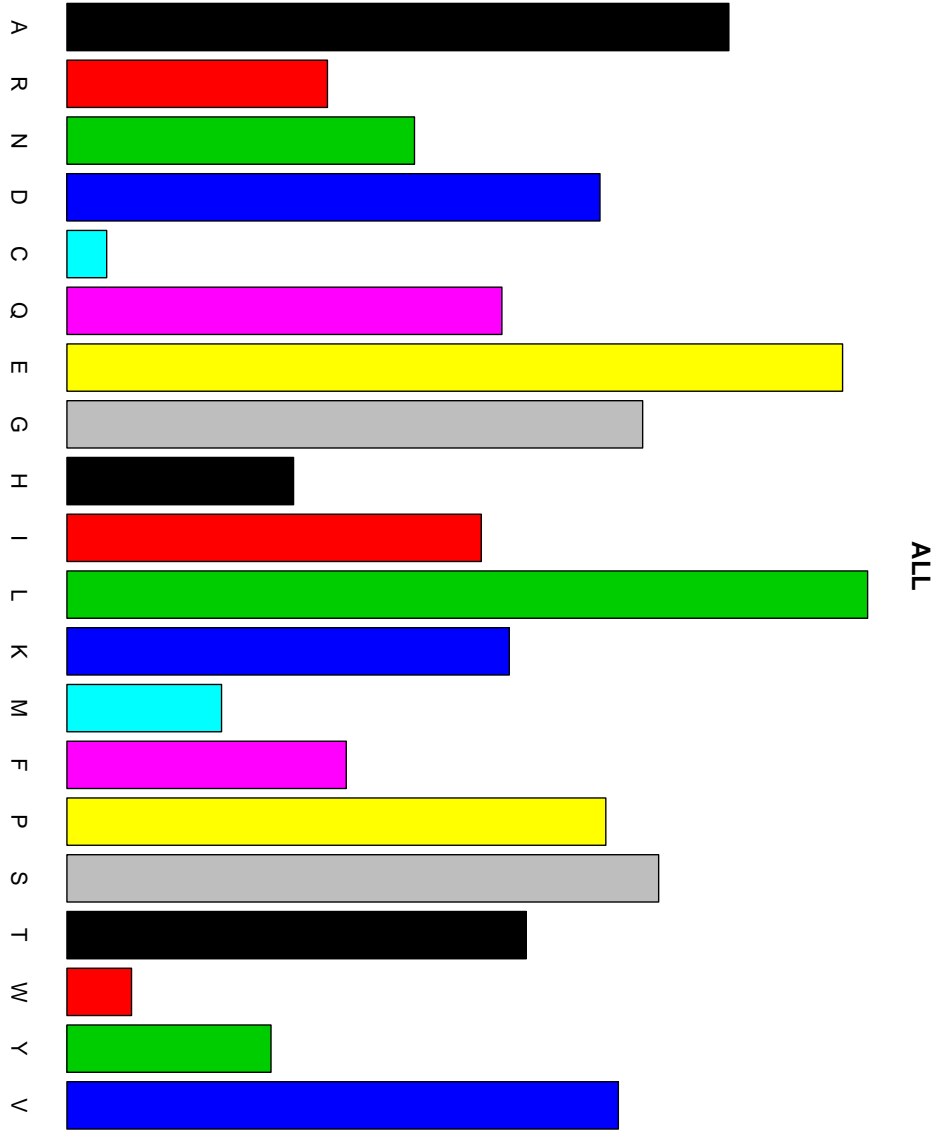
- **URED**A: DC Hoaglin, F Mosteller, JW Tukey (1983), *Understanding Robust and Exploratory Data Analysis*
- **EDT**TS: DC Hoaglin, F Mosteller, JW Tukey (1985), *Exploring Data Tables, Trends, Shapes*
- **FEA**V: DC Hoaglin, F Mosteller, JW Tukey (1991), *Fundamentals of Exploratory Analysis of Variance*
- **EDA**: JW Tukey (1977), *Exploratory Data Analysis*

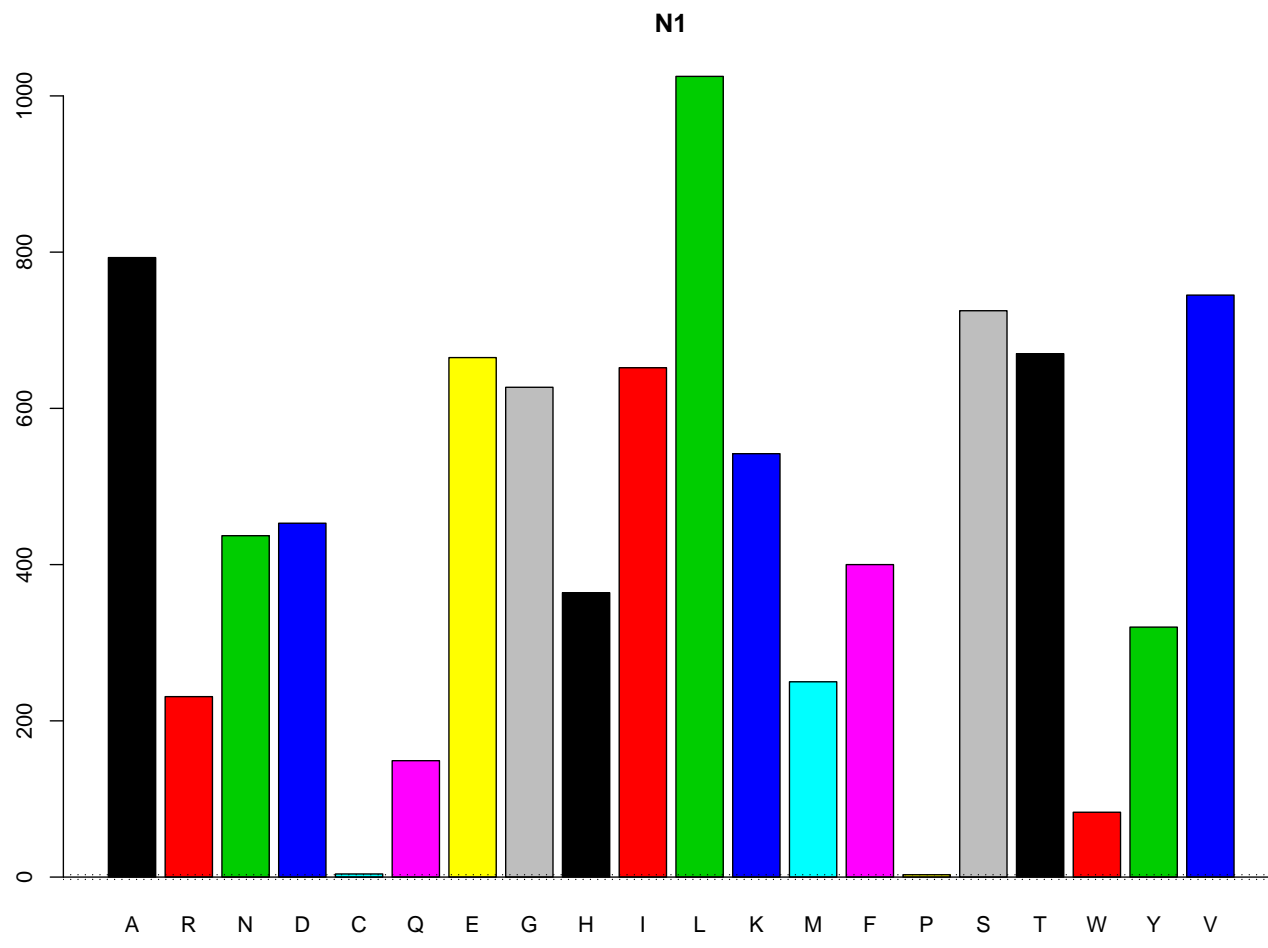
Example: Compare frequency distributions of 20 amino acids in 137,870 peptides with those of 9,138 peptides at specific sites

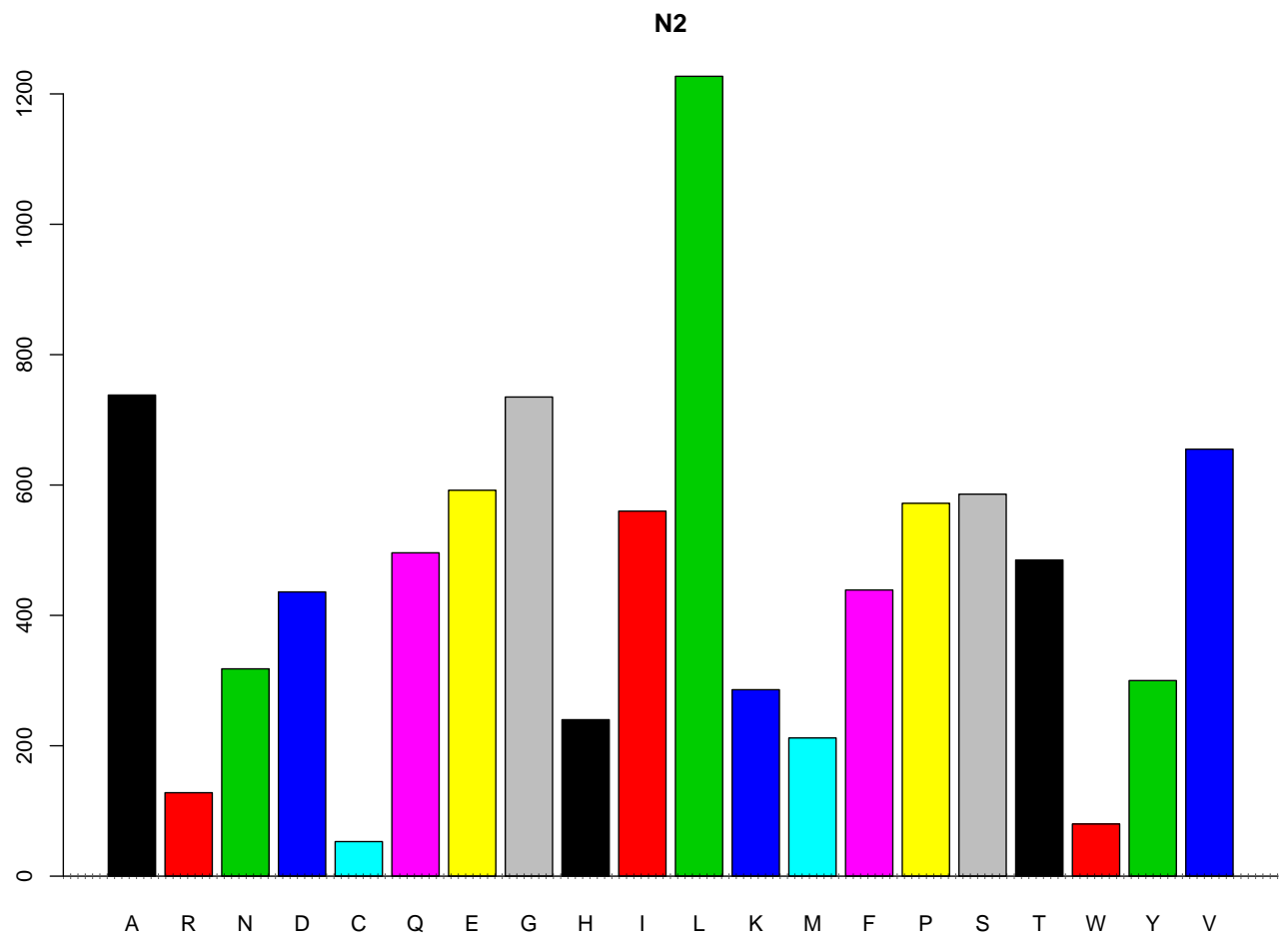
| | ALL | N1 | N2 | N3 | N4 | N5 | C2 | C1 |
|---|-------|------|------|------|------|-----|------|------|
| A | 10920 | 793 | 738 | 765 | 699 | 704 | 795 | 5 |
| R | 4298 | 231 | 128 | 43 | 35 | 25 | 13 | 3655 |
| N | 5734 | 437 | 318 | 395 | 384 | 445 | 392 | 6 |
| D | 8794 | 453 | 436 | 745 | 690 | 695 | 419 | 9 |
| C | 656 | 4 | 53 | 57 | 42 | 48 | 56 | 0 |
| Q | 7176 | 149 | 496 | 594 | 535 | 522 | 556 | 6 |
| E | 12796 | 665 | 592 | 1012 | 1040 | 957 | 959 | 8 |
| G | 9498 | 627 | 735 | 697 | 639 | 624 | 629 | 1 |
| H | 3739 | 364 | 240 | 260 | 283 | 269 | 207 | 3 |
| I | 6835 | 652 | 560 | 472 | 478 | 476 | 518 | 3 |
| L | 13209 | 1025 | 1227 | 878 | 921 | 969 | 1027 | 5 |

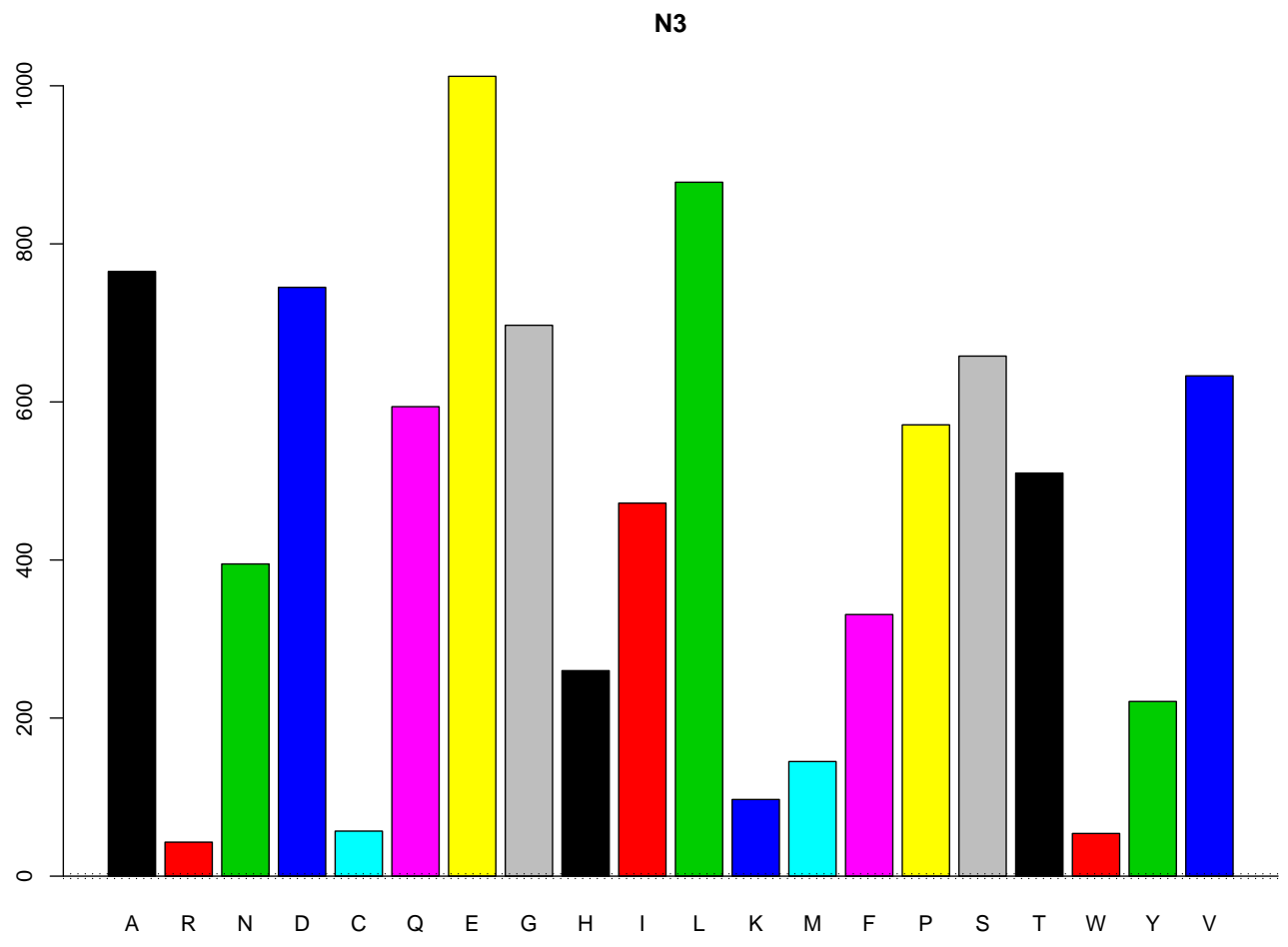
| | | | | | | | | |
|---|------|-----|-----|-----|-----|-----|-----|------|
| K | 7297 | 542 | 286 | 97 | 87 | 75 | 135 | 5424 |
| M | 2551 | 250 | 212 | 145 | 175 | 162 | 205 | 1 |
| F | 4608 | 400 | 439 | 331 | 326 | 338 | 347 | 1 |
| P | 8890 | 3 | 572 | 571 | 666 | 670 | 553 | 1 |
| S | 9762 | 725 | 586 | 658 | 700 | 653 | 737 | 2 |
| T | 7577 | 670 | 485 | 510 | 520 | 473 | 542 | 6 |
| W | 1066 | 83 | 80 | 54 | 83 | 87 | 91 | 0 |
| Y | 3367 | 320 | 300 | 221 | 198 | 254 | 264 | 0 |
| V | 9097 | 745 | 655 | 633 | 637 | 692 | 693 | 2 |

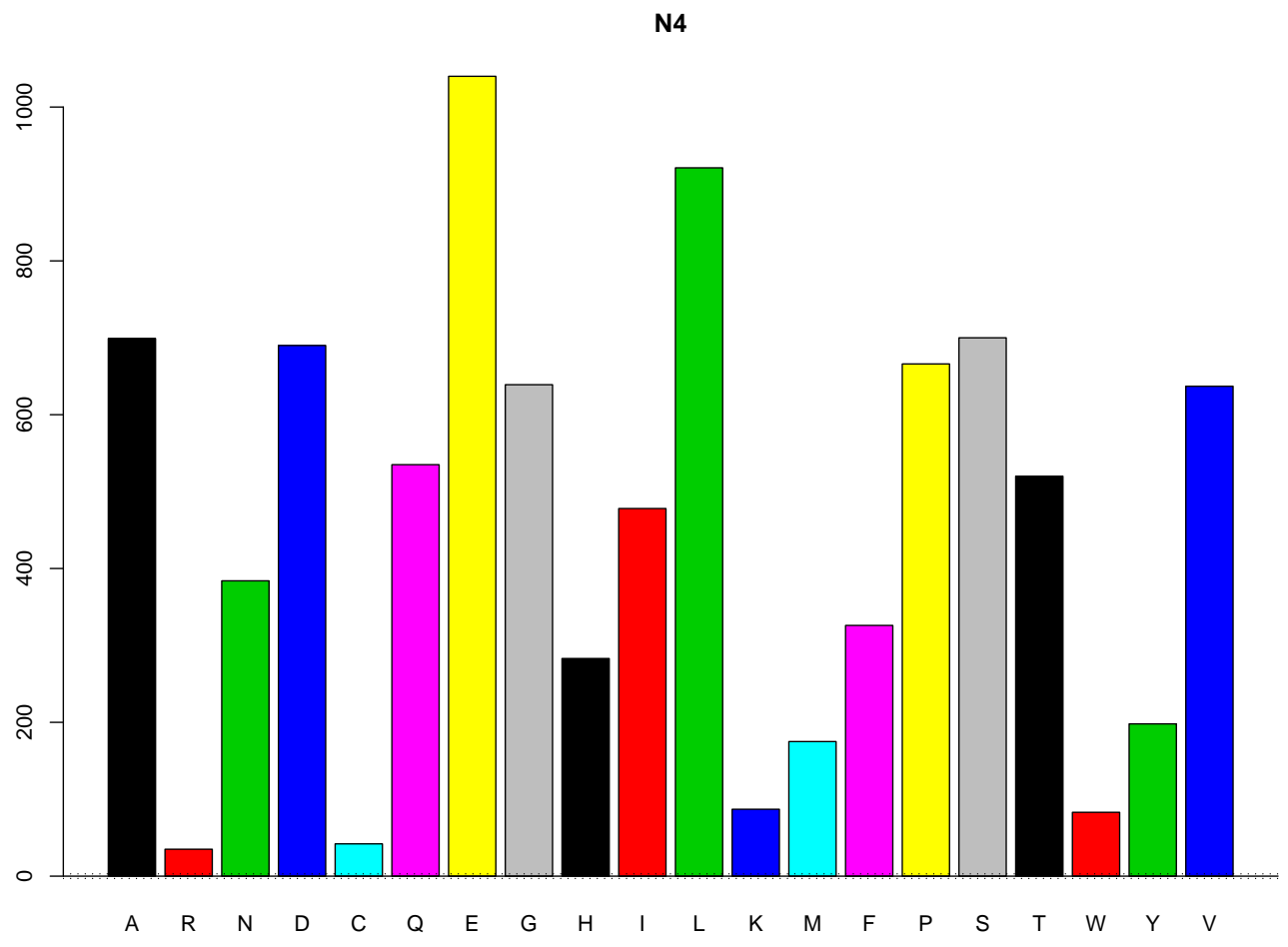
0 2000 4000 6000 8000 10000 12000

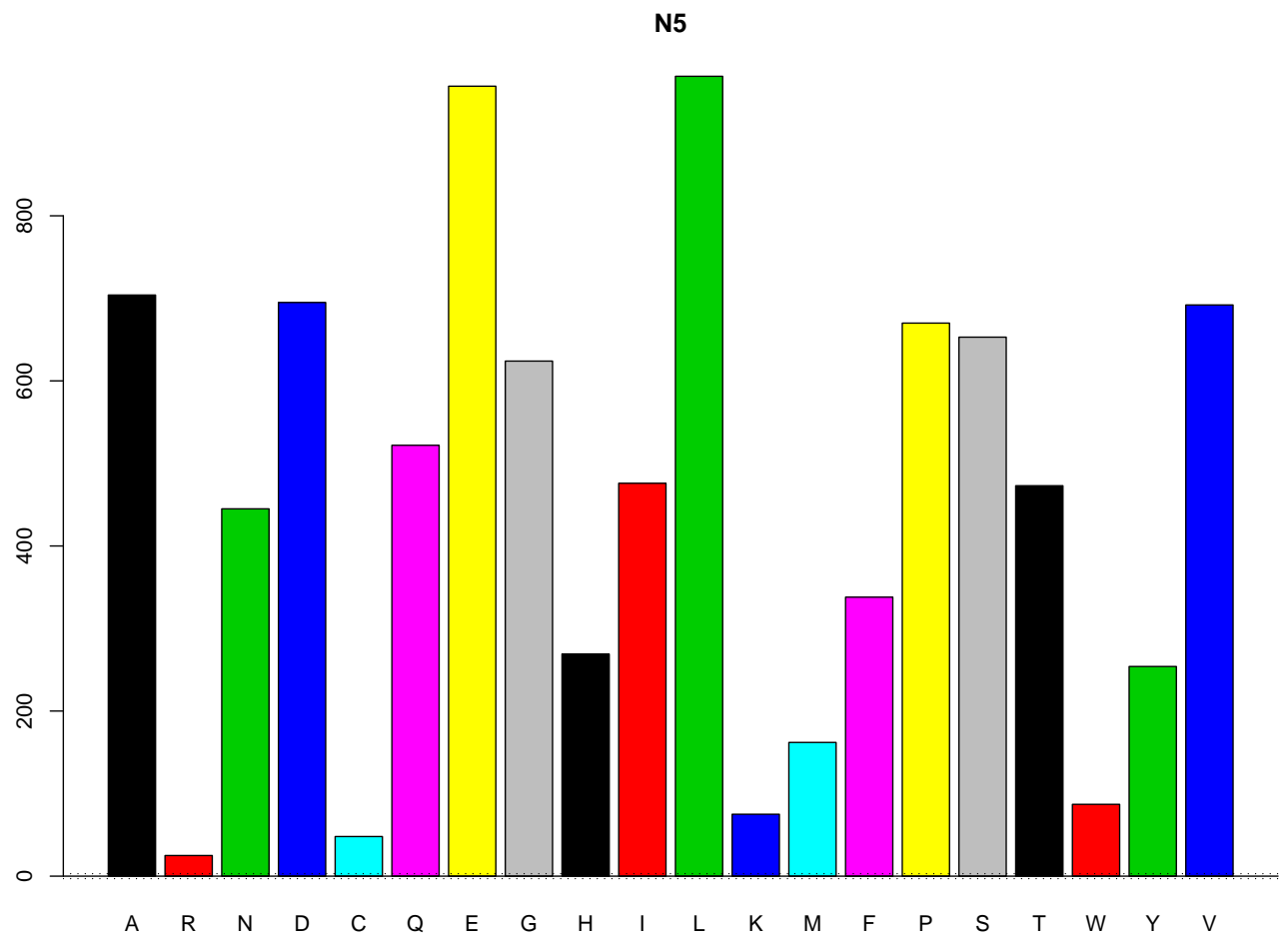


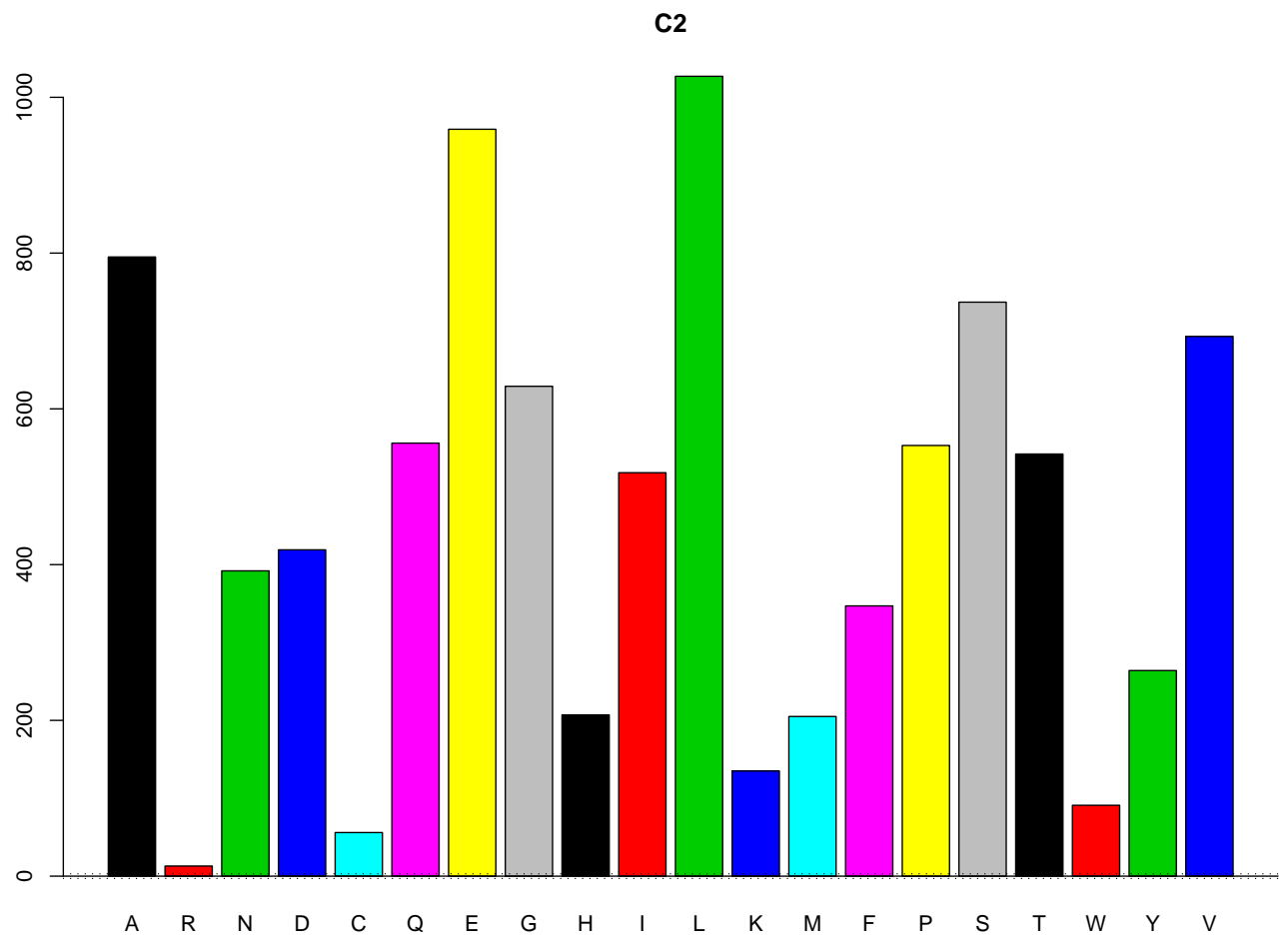












Better: Compare **Standardized** counts!

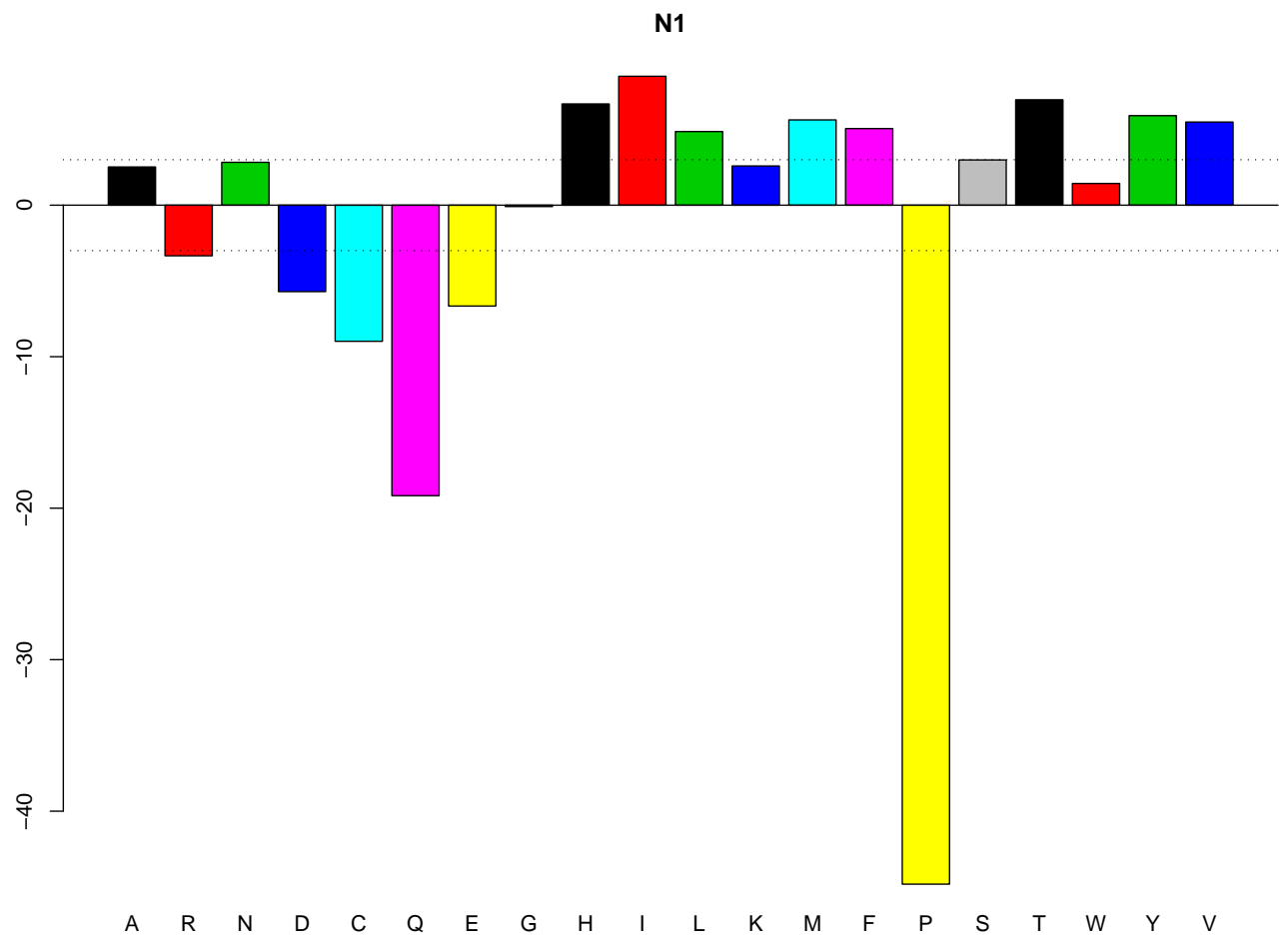
Expected count of amino acid A:

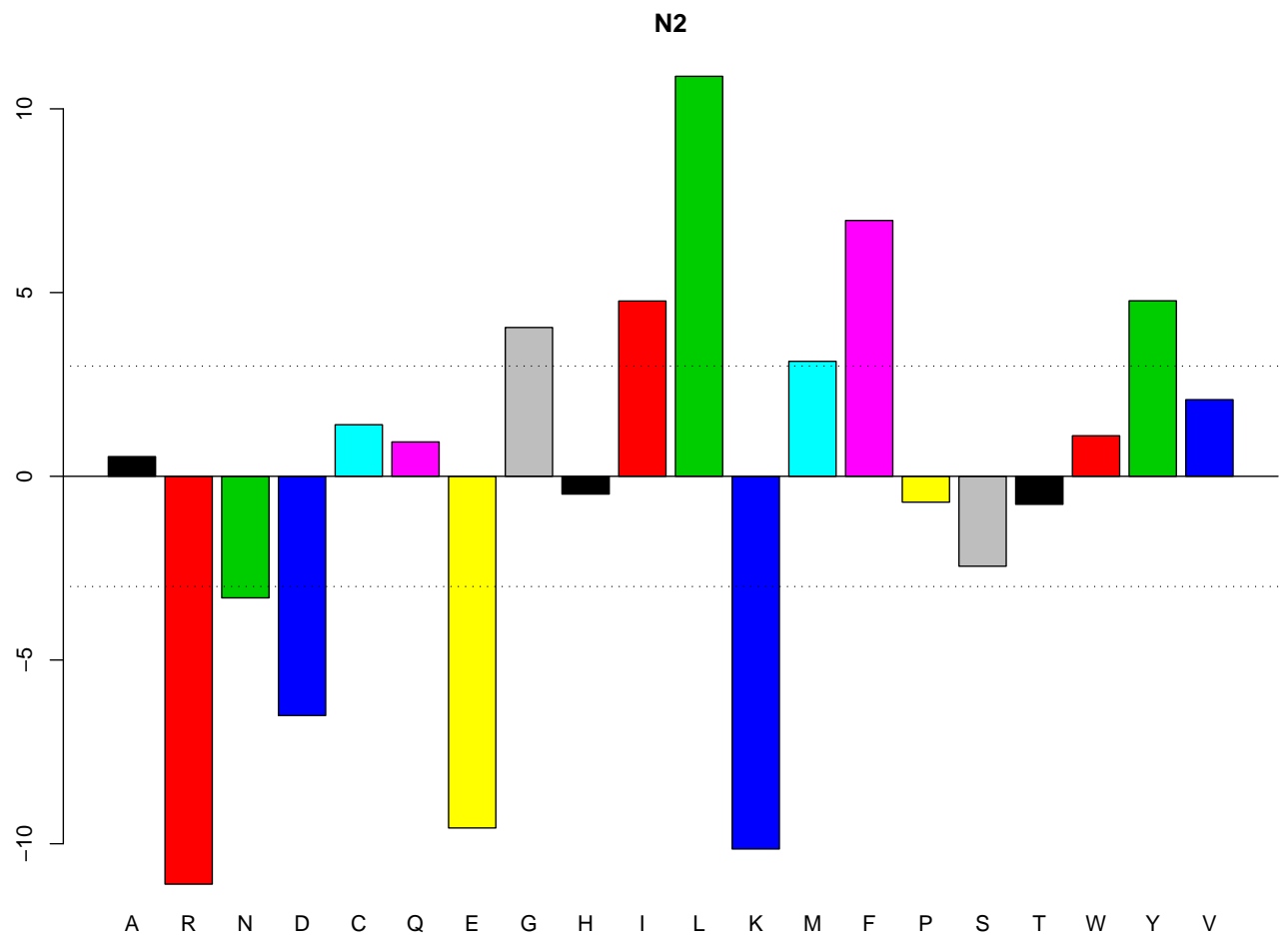
$$\begin{aligned} & (\text{N1 total count}) \times (\text{proportion of A's in ALL}) \\ & (9138) \times (10920/137870) = 723.8 \end{aligned}$$

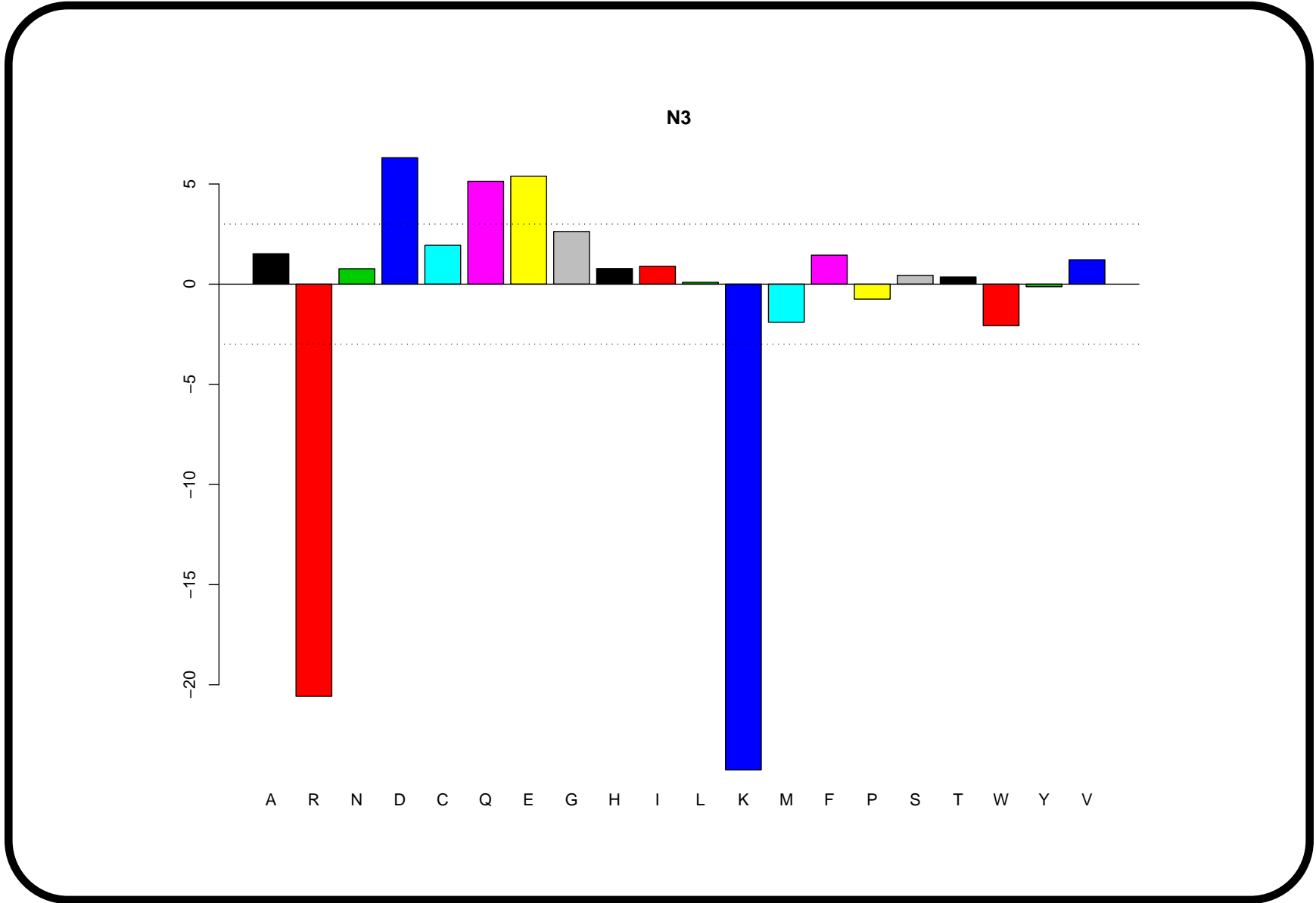
Compare with *Observed* N1 count of A (793):

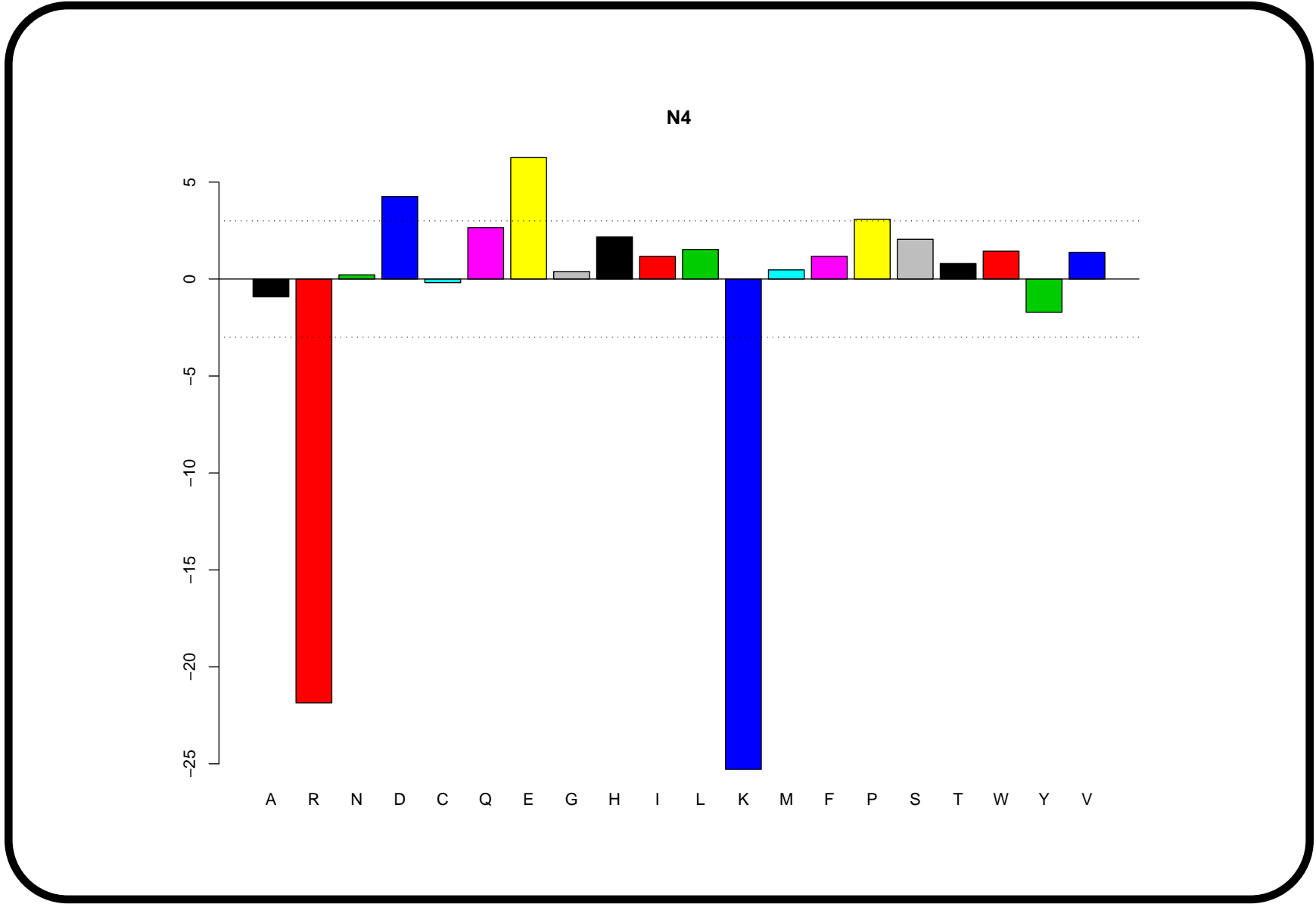
$$\begin{aligned} & \sqrt{4(\text{observed}) + 2} - \sqrt{4(\text{expected}) + 1} \sim N(0, 1) \\ & (56.34) - (53.82) = 2.52 \end{aligned}$$

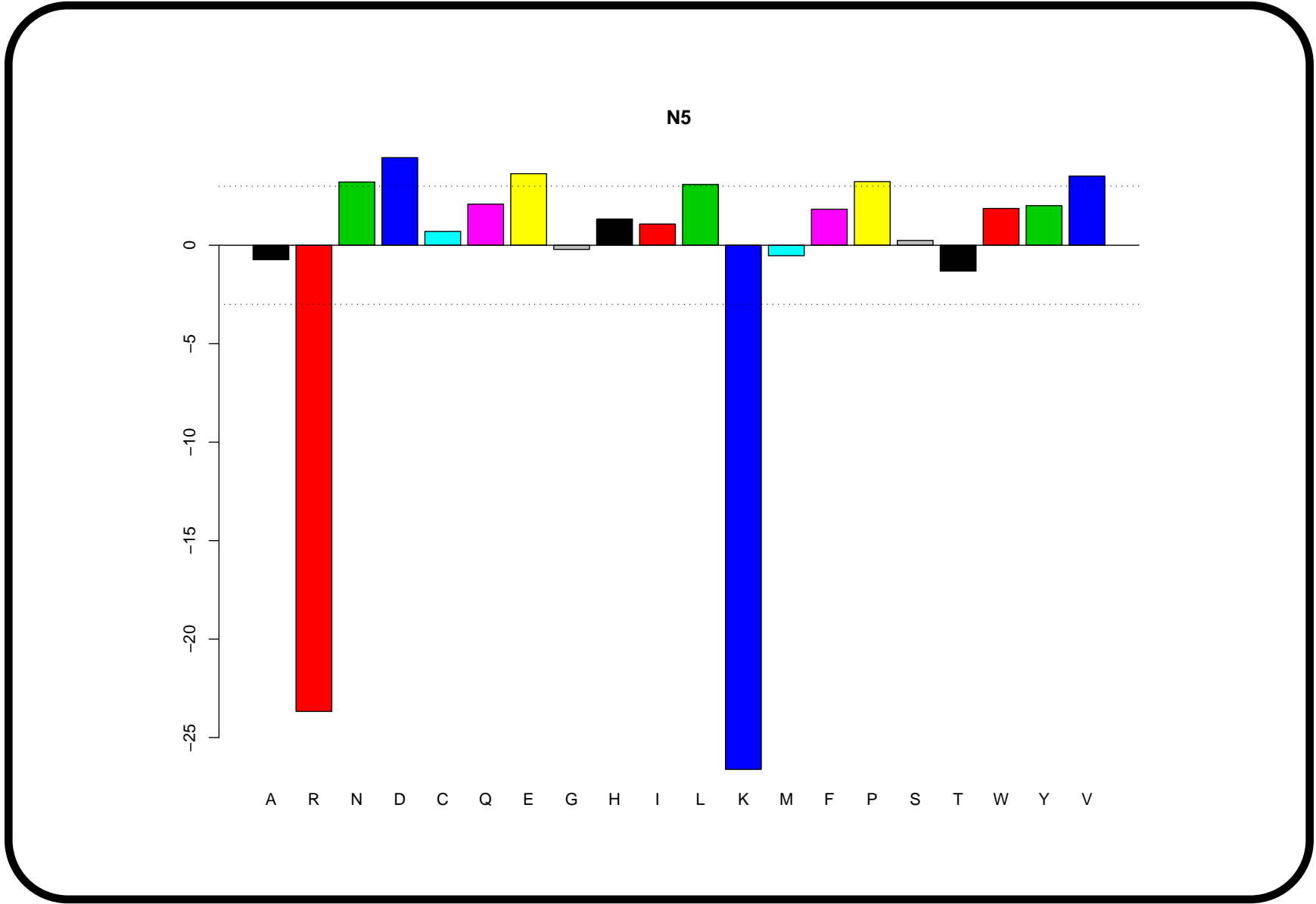
Colors distinguish amino acids consistently for all plots

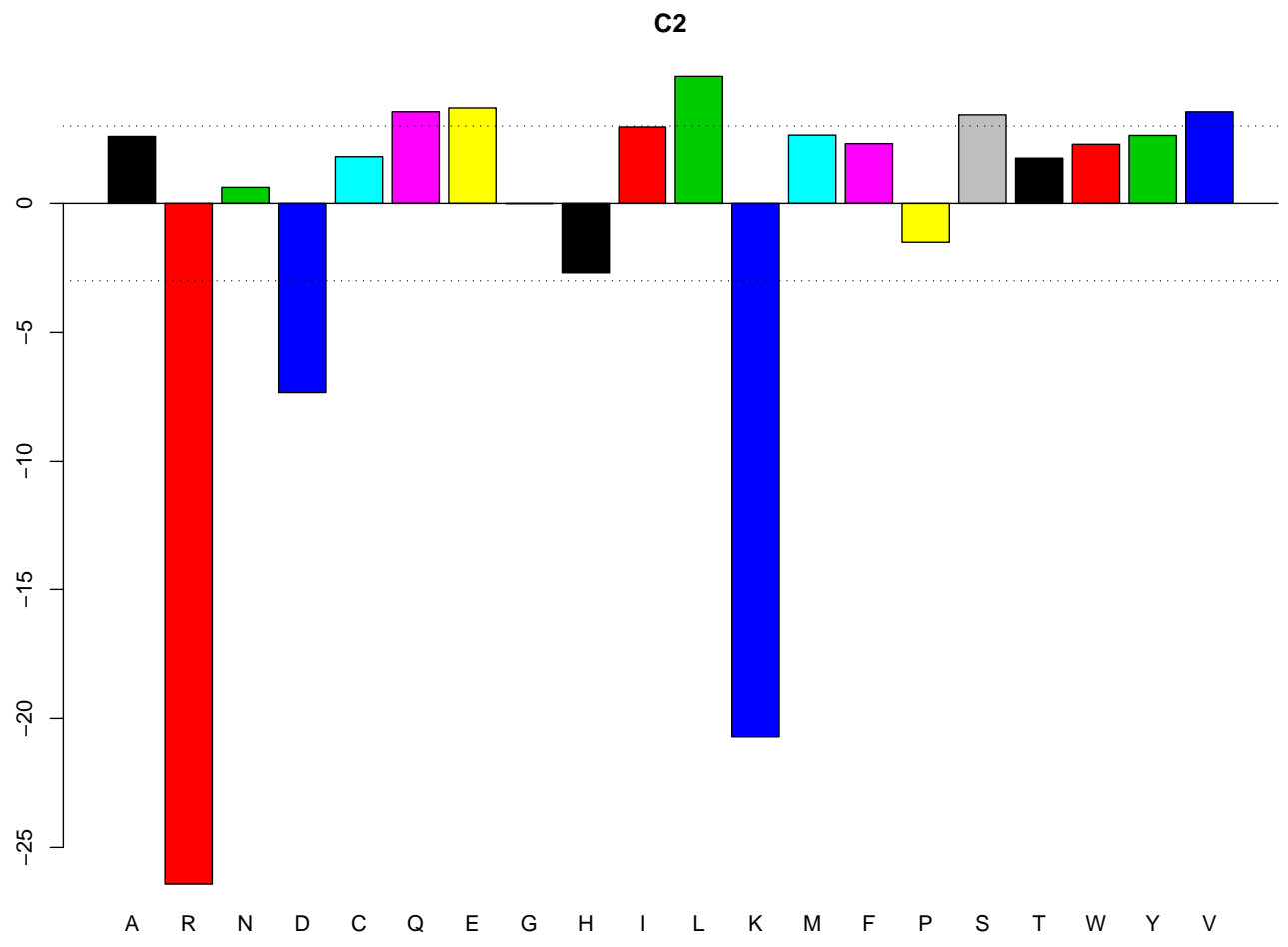












Example (Cuthbert Daniel): Prevalence rates of hearing loss among males aged 55–64 ($\geq 25\%$ loss of hearing)

| hz | profl | farm | sales | crafts | oper | serv | labor |
|------|-------|------|-------|--------|------|------|-------|
| 500 | 2.1 | 6.8 | 8.4 | 1.4 | 14.6 | 7.9 | 4.8 |
| 1000 | 1.7 | 8.1 | 8.4 | 1.4 | 12.0 | 3.7 | 4.5 |
| 2000 | 14.4 | 14.8 | 27.0 | 30.9 | 36.5 | 36.4 | 31.4 |
| 3000 | 57.4 | 62.4 | 37.4 | 63.3 | 65.5 | 65.6 | 59.8 |
| 4000 | 66.2 | 81.7 | 53.3 | 80.7 | 79.7 | 80.8 | 82.4 |
| 6000 | 75.2 | 94.0 | 74.3 | 87.9 | 93.3 | 87.8 | 80.5 |
| norm | 4.1 | 10.2 | 10.7 | 5.5 | 18.1 | 11.4 | 6.1 |

Comments:

First: It is a lot easier to type in the data without decimal points.

Later: $\text{data} < - \text{data}/10$

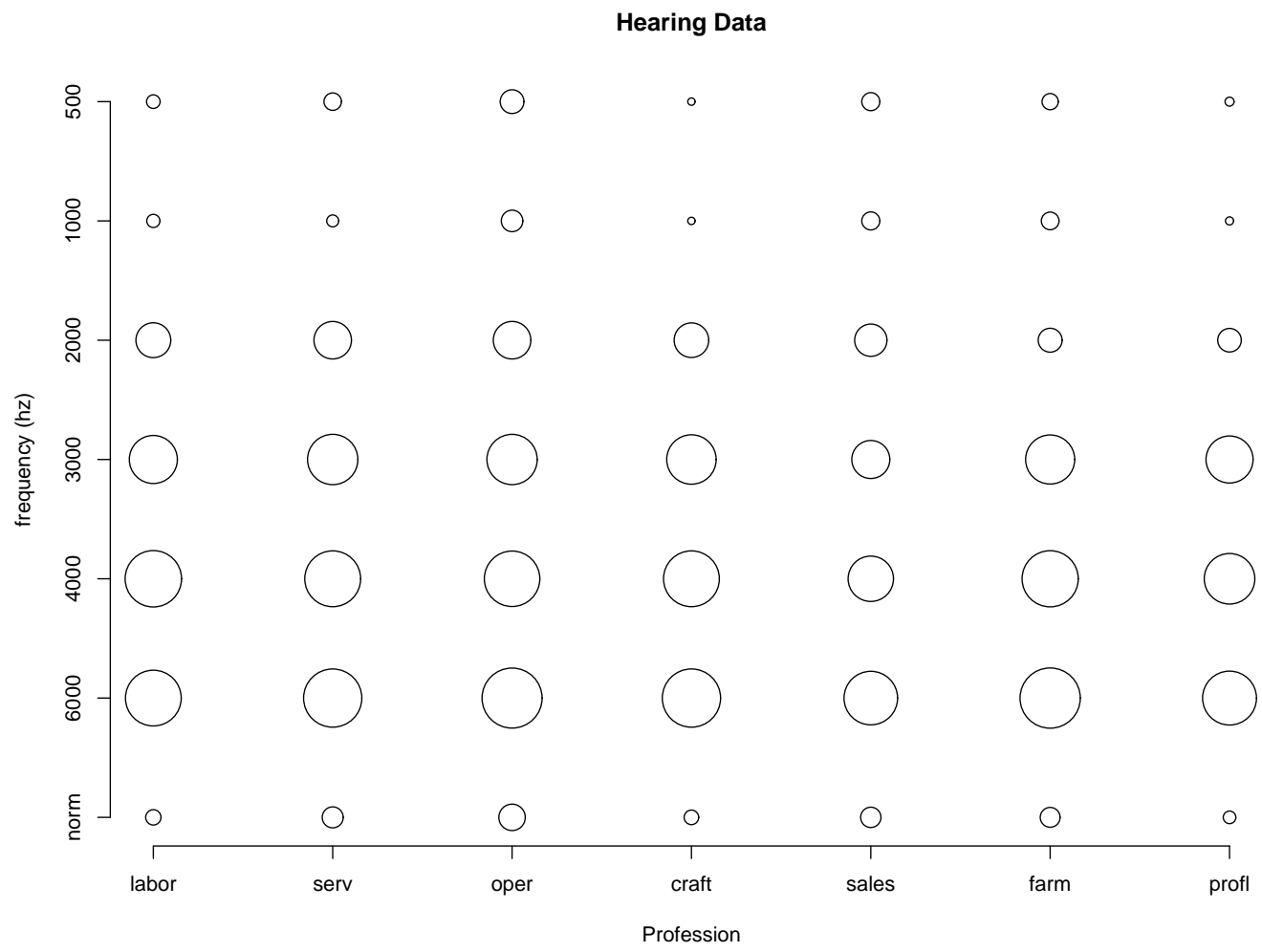
“Confirmatory” approach: Analysis of variance

- What is the A/V table?
- Which factors are significant ($\alpha = 0.05, 0.01$)?
- Is the interaction term “significant”?
- How much variation (%) is explained by each factor?
- What is the RMS (standard error)?

“Exploratory”: Fit the data

- What trends appear in the data?
- Will a simple fit $\hat{data} = common + row + column$ suffice?
- If not, will the simple “PLUS fit” (additive model) work if we transform the data?
- Are there outliers, without which the “PLUS fit” suffices?
- Which factor affects hearing loss more, profession or increased frequency?
- Which profession suffers greatest/least hearing loss, on average (across all frequency levels)?

- Roughly how much additional loss is suffered, in general, from each 1000-hz increase in frequency?
- Can we fit a simple expression to the effects of “frequency”?
- Where does “normal” fall in the frequency range, in terms of hearing loss? (Same for all professions?)
- Do the residuals show any unusual patterns?
- Can we fit the patterns in the residuals with a “PLUS” fit; i.e. $residual = fit_2 + rough_2$? If so, *data* can be expressed as:
$$data = common + row + column + fit_2 + rough_2$$
- What pictures can be made to display our findings (residuals, magnitude of effects, etc.)?



R code:

```
hear <- scan() # data lines here
hear <- matrix(hear,byrow=T,nrow=7)
dimnames(hear) <-
list( c("500",format(1000*(1:4)),"6000", "norm"),
      c("profl", "farm", "sales", "craft", "oper", "serv", "labor"))

plot(1:7,1:7,axes=F,type="n",main="Hearing Data",
      xlab="Profession", ylab="frequency (hz)")
axis(side=1,at=1:7,labels=rev(dimnames(hear)[[2]]))
axis(side=2,at=1:7,labels=rev(dimnames(hear)[[1]]))
symbols(rep(1:7,rep(7,7)),rep(1:7,7),add=T,
        circles=sqrt(c(rev(hear))),inches=.3)
```

Exploratory versus Confirmatory

Confirmatory

- Formulate model **before** seeing the data
- Analyze data
- Assess “significance” (inference) based on model

Exploratory

- Look at the data — plots, trends, outliers, ...
- What do the data indicate?
- Can we fit the data with simple expressions?
- Can we go further?

BOTH: What questions do we hope the data can answer?

“Exploratory data analysis is detective work — numerical detective work — or counting detective work — or graphical detective work” (EDA, p.1)

“The Future of Data Analysis” (*Ann. Math. Stat.* 1962, 1–67):
Necessary attitudes:

- awareness of “*more realistic problems*” (p. 61) (potentially non-normal data);
- “*necessarily approximate nature of useful results*” (p. 61);
- “*need for iterative procedures*” (p. 62);
- “*both indication and conclusion in the same analysis*” (p. 62);
- “*free use of ad hoc and informal procedures in seeking indications*” (p. 63);

- recognition for “*indication procedures to grow up before the corresponding conclusion procedures*” (p. 63);
- acknowledgement that a ‘conclusion procedure’ (i.e. 1%- or 5%-level test) need not be precise to be useful (p. 64);
- “*data analysis is intrinsically an empirical science*” (p. 64).

“Churchill Eisenhart ... defined practical power as the product of the mathematical power by the probability that the procedure will be used. A compact procedure may well be used so much more often as to more than compensate for its loss of mathematical power.”

— J.W. Tukey (1959), “A Quick, Compact, Two-Sample Test to Duckworth’s Specifications,” *Technometrics* 1(1), p.32

Discussion of *JASA* article by Parzen (p.121):

“...‘exploratory data analysis’ is an attitude, a state of flexibility, a willingness to look for things that we believe are not there, as well as for those we believe might be there. Except for emphasis on graphs, its tools are secondary to its purposes.”

“Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider”

“Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone — as the first step” (*EDA*, p.3)

EDA is an *approach* to analyzing data, whose main objective is to find patterns and clues in the data (“detective work”)

But: “We need BOTH exploratory AND confirmatory” (*TAS*)

EDA: tools for exploring, investigating data

CDA: tools for validating hypotheses (prevent us from jumping at each pattern we might see)

Data analysis: back-and-forth between the two approaches:

1. Formulate 1–2 specific questions or hypotheses (scientist)
2. Conduct appropriate hypothesis test, p-value, model-based uncertainty (statistician: CDA)
3. Explore data, note unexpected patterns (statistician: EDA)
4. Design future investigations to confirm patterns found through EDA (statistician + scientist communication)

Four (five) R's of EDA:

1. *Resistance*:

insensitivity to “small” changes in the data

- LARGE change in small fraction of data (e.g., 3.53 miscoded as 353)
- small changes in LARGE fraction of data (e.g., rounding)

How resistant are: sample mean, trimmed mean, median?

Sample standard deviation, sample IQR?

Other resistant measures of location? of spread?

2. *Residuals*:

Usually data cannot be expressed perfectly by a simple fit (line, two-way analysis,):

$$data = fitted\ value + residual$$

$$data = smooth + rough$$

Do the residuals show structure? (F.R. Anscombe: “The examination of residuals”, *Technometrics* 1963?)

3. *Re-expression*:

Two philosophies for fitting data:

(a) $data = g(fit, residual)$; g is nonlinear combination of fit and $residual$; residual has (xxx) distribution: model parameters are fit via nonlinear least squares or computational maximum likelihood

(b) $data = g(fit, residual)$, g nonlinear, but
 $t(data) = simple\ fit + residual$
i.e., transform data and the fitting process is simpler.

EDA tends to choose (b): Much more is known about

- effects of departures from normality
- methods that are less sensitive (resistant) to Gaussian assumption
- methods to diagnose nonlinearity
- methods to detect outliers (“1–10% errors in data”)
- methods for characterizing the uncertainty in estimates obtained on transformed data (propagation of uncertainty formulae, jackknife, bootstrap)
- etc etc etc

4. *Revelation*:

Reveal patterns (microscope): varied

displays/graphs/pictures “**The greatest value of a picture is when it *forces* us to notice what we never expected to**

see” (EDA p.vi).

5. *Re-iteration:*

Sequential fitting:

$$data = fit_1 + rough_1$$

$$rough_1 = fit_2 + rough_2$$

$$rough_2 = fit_3 + rough_3$$

$$\dots \Rightarrow data = fit_1 + fit_2 + fit_3 + \dots + final\ rough$$

Also, many resistant procedures require iteration (resistant line, median polish, biweight, ...) Computers are good at repetition.

Four useful initial exploratory data plots:

1. Time plot: plot data values in order they appear
2. histogram
3. Lag-1 plot (plot of x_i vs x_{i-1})
4. QQ-plot (“straightened” histogram)

Always plot your data!

“There is no excuse for failing to plot and look” (JWT)

First looks at data: Stem-and-leaf displays

Emerson & Hoaglin (Ch 1): early versions

Objectives: Illustrate **shape** of data

1. Symmetry or lack thereof
2. Especially popular or wildly aberrant values
3. Possible rounding (a.0, b.0, c.5, d.5, e.5, f.5, g.0)
4. clusters, gaps
5. approximate center & spread of data

Related to histograms but retains numerical information

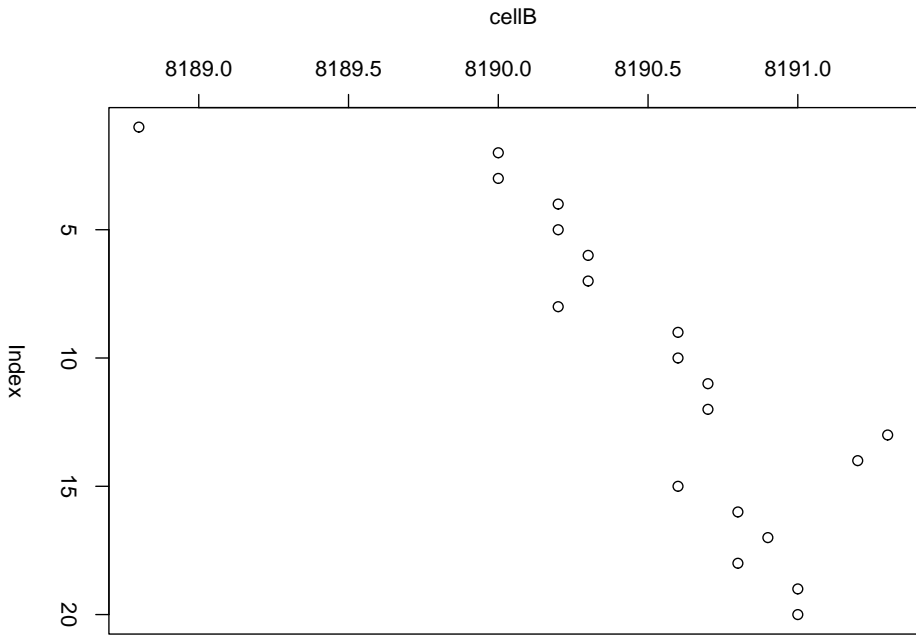
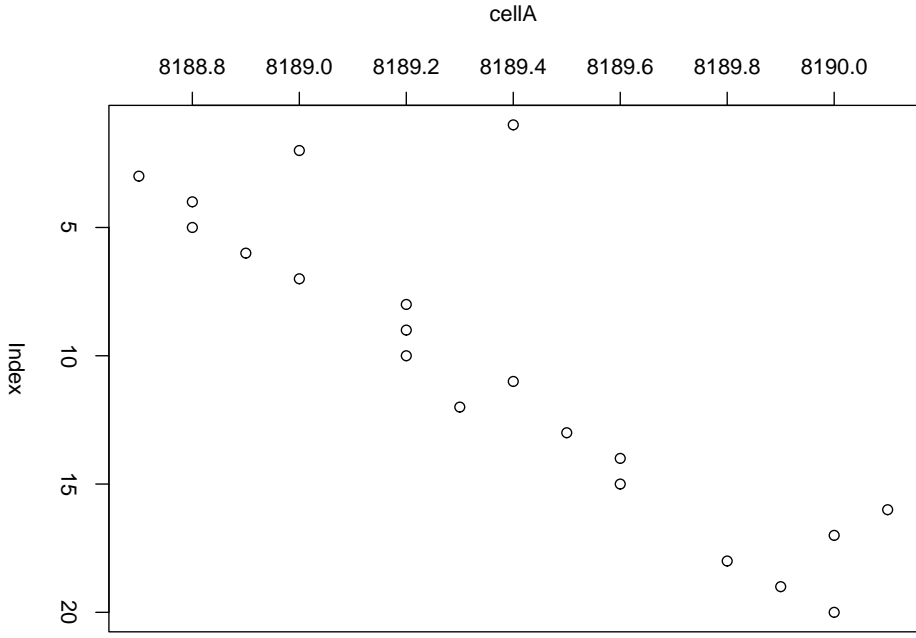
Minutes to get from home to campus

NIST voltage transfer cells (data in millivolts): $cellA < -scan()$

1: 8189.4 8189.0 8188.7 8188.8 8188.8 8188.9
7: 8189.0 8189.2 8189.2 8189.2 8189.4 8189.3
13: 8189.5 8189.6 8189.6 8190.1 8190.0 8189.8
19: 8189.9 8190.0

$cellB < -scan()$

1: 8188.8 8190.0 8190.0 8190.2 8190.2 8190.3
7: 8190.3 8190.2 8190.6 8190.6 8190.7 8190.7
13: 8191.3 8191.2 8190.6 8190.8 8190.9 8190.8
19: 8191.0 8191.0



stem(cellA)

The decimal point is at the —

8188 — 7889

8189 — 00222344

8189 — 56689

8190 — 001

stem(cellB)

The decimal point is at the “—

8188 — 8

8189 —

8190 — 002223366677889

8191 — 0023

stem(cellB,scale=2) The decimal point is at the —

8188 — 8

8189 —

8189 —

8190 — 0022233

8190 — 66677889

8191 — 0023

“Back-to-back” stem-and-leaf display:

9887 — 8188 — 8

44322200 — 8189 —

98665 — 8189 —

100 — 8190 — 0022233

— 8190 — 66677889

— 8191 — 0023

Displays show that:

- Measurement process on both cells drifted over time;
- Cell A holds $\approx 8189 \mu\text{volts}$, Cell B $\approx 1.5 \mu\text{volts}$ more;
- The spread in measurements of both cells is similar;
- One measurement on Cell B is about $2 \mu\text{volts}$ less than the other 19 measurements (see time plot: this outlier was first measurement)

Mechanics of stem-and-leaf displays:

Separate number into its stem (the part that needn't be repeated each time) and the leaf (extra digit of detail):

- Single line per decade: 2 — 0123456789
- Two lines per decade:

2 — 01234

2+ — 56789

- Five lines per decade:

2 O — 01

T — 23

F — 45

S — 67

* — 89

Ex: Tumor progression in patients with glioblastoma (p.14–15)

Ex: Temperature readings (p.31):

88 66 71 63 101 55 76

49 63 38 91 79 41 36

73 55 42 49 50 90 51

Ex: Hardness of aluminum die castings ($\times 10$)

530 702 843 553 785 635 714 534

825 673 695 730 557 858 954 511

744 541 778 524 691 535 643 827

557 705 875 507 723 595

One line per decade:

5 — 01233345559

6 — 34799

7 — 00123478

8 — 22457

9 — 5

Two lines per decade:

5 — 0123334

5+— 5559

6 — 34

6+— 799

7 — 001234

7+— 78

8 — 224

8+— 57

9 —

9+— 5

0.12 0.15 0.15 0.10 0.13 0.15 0.14
0.08 0.11 0.09 0.14 0.09 0.13 0.14
0.12 0.16 0.15 0.13 0.12 0.12 0.09

8 — 0000

10 — 00

12 — 0000000

14 — 0000000

16 — 0

0 * — 8999

1 O — 01

T — 2332322

F — 5554445

S — 6

Connection between stem-and-leaf and histogram:

$L = \#$ of lines in stem-and-leaf display \approx

$h =$ histogram bin size $\approx 3.5\hat{\sigma}/n^{\frac{1}{3}}$

Rough rule of thumb: $L = 10 \cdot \log_{10} n$

| | | | | | | |
|-------|----|----|----|----|----|----|
| $n =$ | 10 | 20 | 30 | 40 | 50 | 60 |
|-------|----|----|----|----|----|----|

| | | | | | | |
|-------|----|----|----|----|----|----|
| $L =$ | 10 | 13 | 15 | 16 | 17 | 18 |
|-------|----|----|----|----|----|----|

$2\sqrt{n}$ is OK for $n \leq 100$; then it grows too fast

Sturges' rule: $1 + \log_2 n$ is way too small

D.W. Scott (1979): Estimate pdf f via \hat{f} using mean squared error (MSE) and integrated MSE (IMSE):

$$MSE(u) = E[\hat{f}(u) - f(u)]^2$$

$$IMSE(\hat{f}) = \int MSE(u) du$$

Minimize IMSE as a function of bin width used in \hat{f} :

$$h_n = [6/(n \int_{-\infty}^{\infty} (f'(x))^2 dx)]^{\frac{1}{3}} \propto n^{-\frac{1}{3}}$$

Problem: Need to know f and f' ! For $f = \phi$ (Gaussian),

$h_n \approx 3.5\sigma/n^{\frac{1}{3}}$; Estimate σ for Gaussian by $\frac{3}{4}IQR \Rightarrow 2.6IQR/n^{\frac{1}{3}}$

Freedman and Diaconis: How far might \hat{f}_h (estimate of f based on bin width h) be from f : $D(h) = \max_x |\hat{f}_h(x) - f(x)|$; find h to minimize $D(h) \Rightarrow h_n \approx c(f)(\log_e n/n)^{\frac{1}{3}} \approx 2(IQR)/n^{\frac{1}{3}}$ if $f = \phi$