# Introduction to Data Science

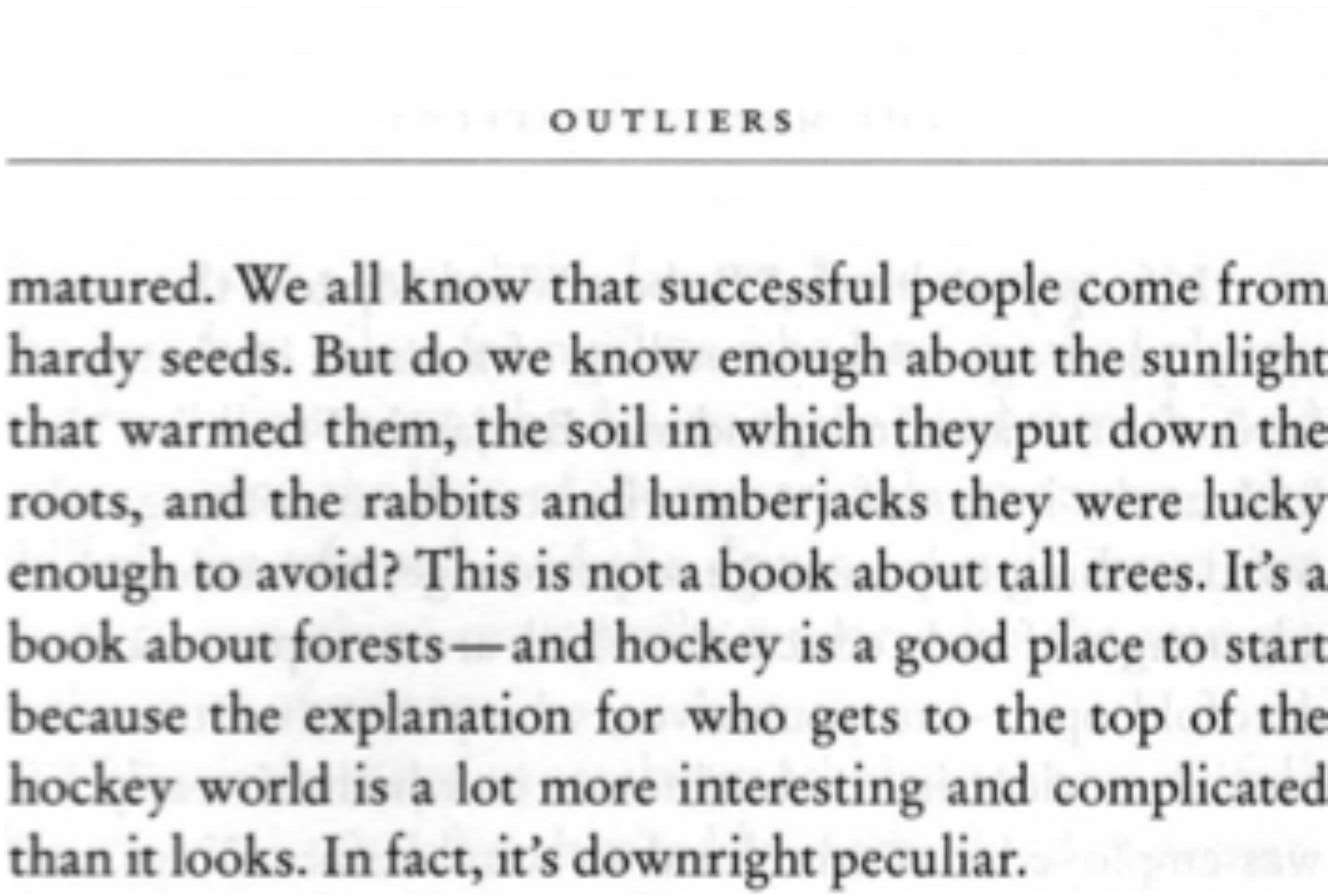
GIRI NARASIMHAN, SCIS, FIU

# Outliers

**FROM JOHNSON & WICHERN, *APPLIED MULTIVARIATE STATISTICAL ANALYSIS*, 6TH ED**

# Canadian Hockey

- Kids trained early; Leagues for age groups
- Most talented get on Major Jr A league team& compete for Memorial Cup
- What makes a top-notch hockey player?
- Soccer, Baseball, Cricket, Swimming, Gymnastics



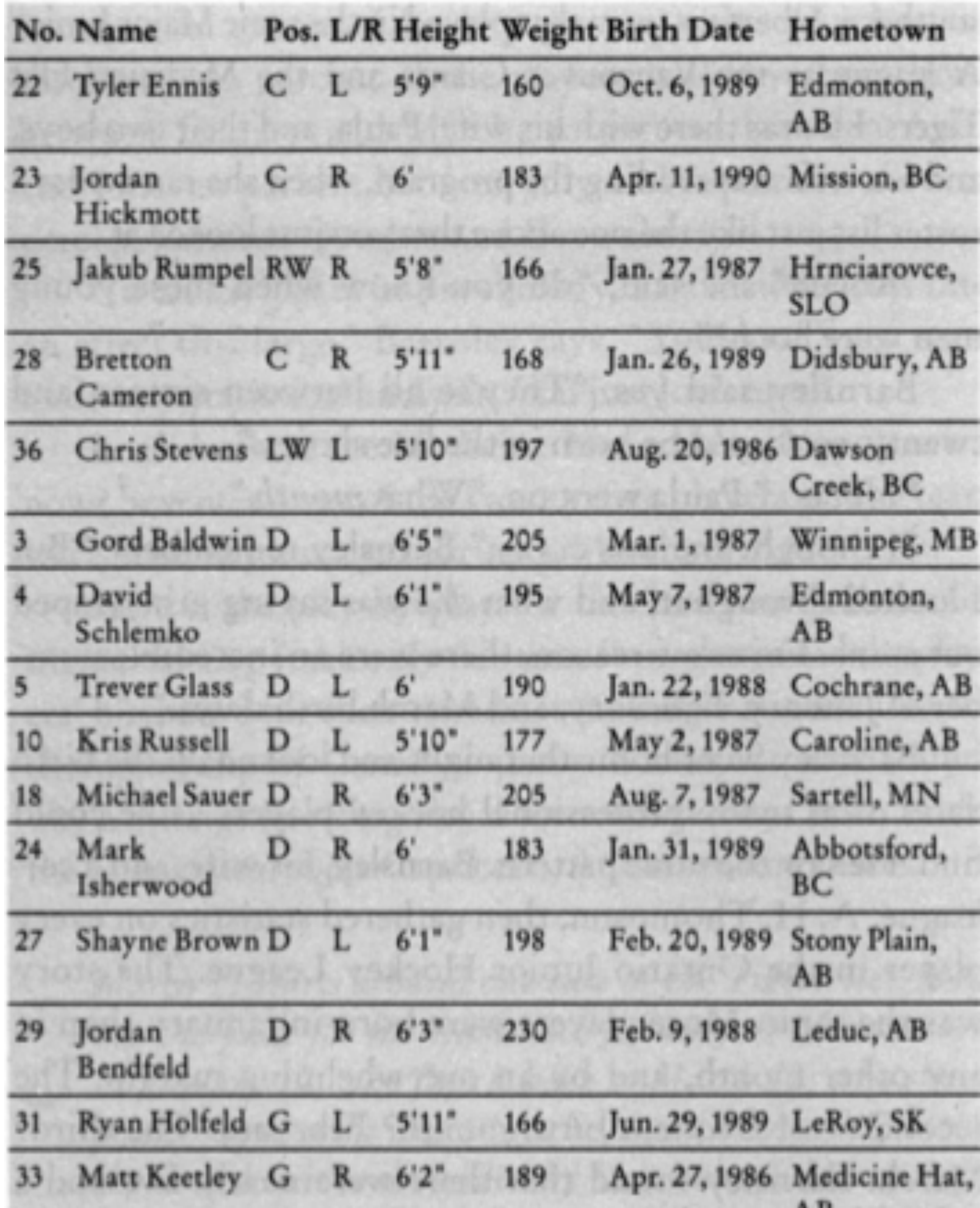
matured. We all know that successful people come from hardy seeds. But do we know enough about the sunlight that warmed them, the soil in which they put down the roots, and the rabbits and lumberjacks they were lucky enough to avoid? This is not a book about tall trees. It's a book about forests—and hockey is a good place to start because the explanation for who gets to the top of the hockey world is a lot more interesting and complicated than it looks. In fact, it's downright peculiar.

3

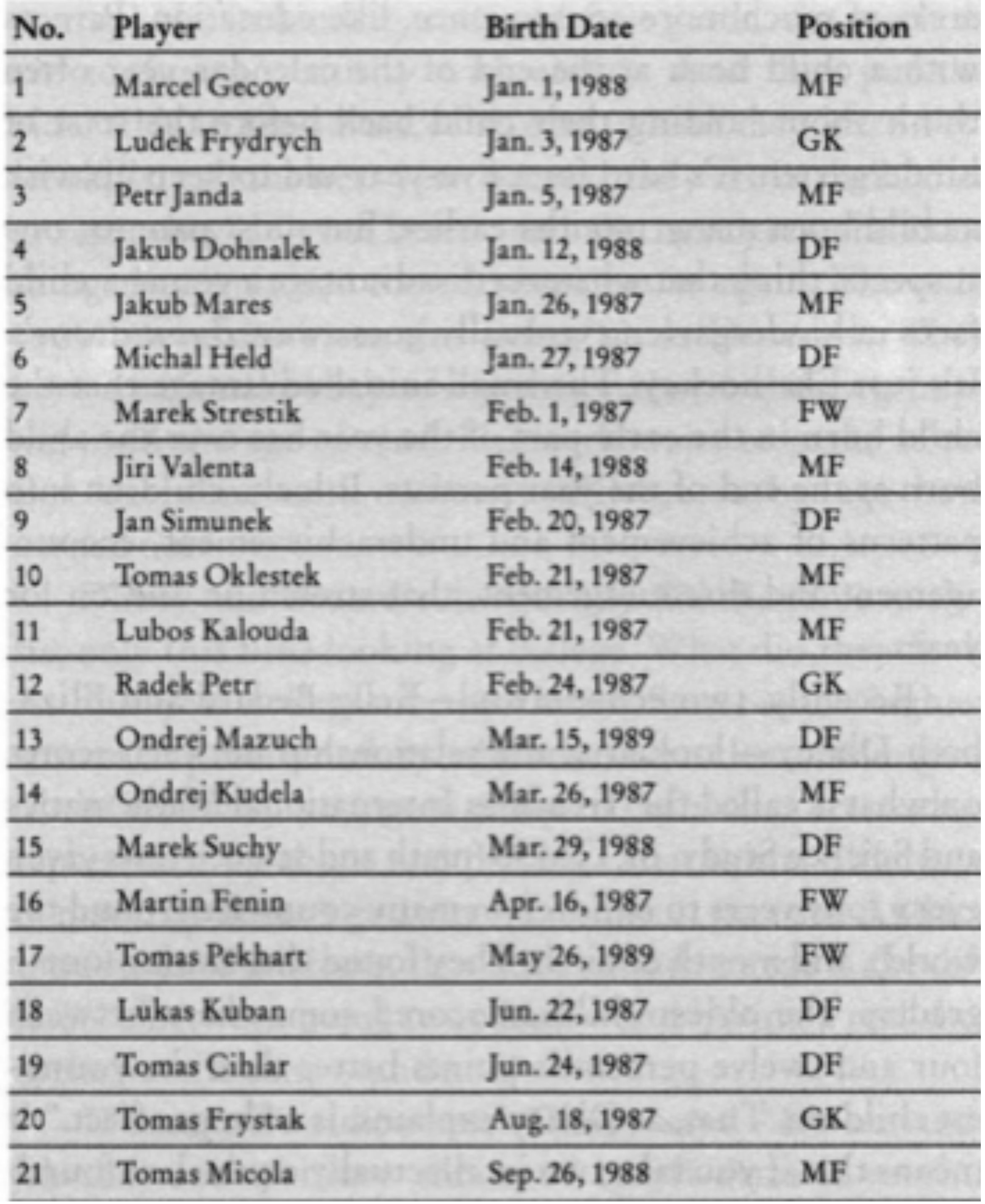
Jan
Feb
Mar
Apr

May 8 2
Jun 3
Jul 3
Aug 3 2

Sep
1 1
Oct
1
Nov
Dec 1



| No. | Name | Pos. | L/R | Height | Weight | Birth Date | Hometown |
|---|---|---|---|---|---|---|---|
| 22 | Tyler Ennis | C | L | 5'9" | 160 | Oct. 6, 1989 | Edmonton, AB |
| 23 | Jordan Hickmott | C | R | 6' | 183 | Apr. 11, 1990 | Mission, BC |
| 25 | Jakub Rumpel | RW | R | 5'8" | 166 | Jan. 27, 1987 | Hrnciarovce, SLO |
| 28 | Bretton Cameron | C | R | 5'11" | 168 | Jan. 26, 1989 | Didsbury, AB |
| 36 | Chris Stevens | LW | L | 5'10" | 197 | Aug. 20, 1986 | Dawson Creek, BC |
| 3 | Gord Baldwin | D | L | 6'5" | 205 | Mar. 1, 1987 | Winnipeg, MB |
| 4 | David Schlemko | D | L | 6'1" | 195 | May 7, 1987 | Edmonton, AB |
| 5 | Trever Glass | D | L | 6' | 190 | Jan. 22, 1988 | Cochrane, AB |
| 10 | Kris Russell | D | L | 5'10" | 177 | May 2, 1987 | Caroline, AB |
| 18 | Michael Sauer | D | R | 6'3" | 205 | Aug. 7, 1987 | Sartell, MN |
| 24 | Mark Isherwood | D | R | 6' | 183 | Jan. 31, 1989 | Abbotsford, BC |
| 27 | Shayne Brown | D | L | 6'1" | 198 | Feb. 20, 1989 | Stony Plain, AB |
| 29 | Jordan Bendfeld | D | R | 6'3" | 230 | Feb. 9, 1988 | Leduc, AB |
| 31 | Ryan Holfeld | G | L | 5'11" | 166 | Jun. 29, 1989 | LeRoy, SK |
| 33 | Matt Keetley | G | R | 6'2" | 189 | Apr. 27, 1986 | Medicine Hat, AB |



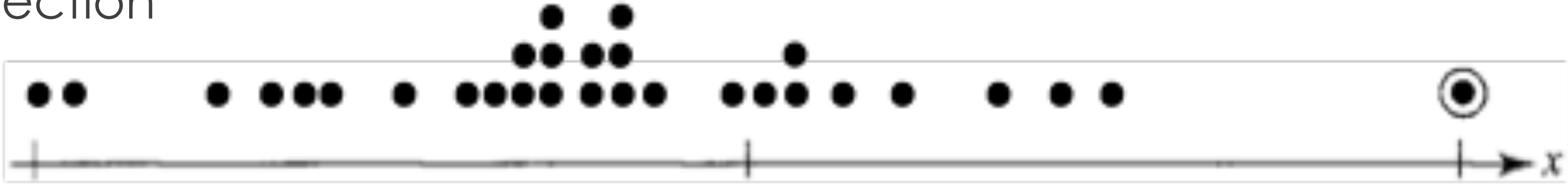
| No. | Player | Birth Date | Position |
|---|---|---|---|
| 1 | Marcel Gecov | Jan. 1, 1988 | MF |
| 2 | Ludek Frydrych | Jan. 3, 1987 | GK |
| 3 | Petr Janda | Jan. 5, 1987 | MF |
| 4 | Jakub Dohnalek | Jan. 12, 1988 | DF |
| 5 | Jakub Mares | Jan. 26, 1987 | MF |
| 6 | Michal Held | Jan. 27, 1987 | DF |
| 7 | Marek Strestik | Feb. 1, 1987 | FW |
| 8 | Jiri Valenta | Feb. 14, 1988 | MF |
| 9 | Jan Simunek | Feb. 20, 1987 | DF |
| 10 | Tomas Oklestek | Feb. 21, 1987 | MF |
| 11 | Lubos Kalouda | Feb. 21, 1987 | MF |
| 12 | Radek Petr | Feb. 24, 1987 | GK |
| 13 | Ondrej Mazuch | Mar. 15, 1989 | DF |
| 14 | Ondrej Kudela | Mar. 26, 1987 | MF |
| 15 | Marek Suchy | Mar. 29, 1988 | DF |
| 16 | Martin Fenin | Apr. 16, 1987 | FW |
| 17 | Tomas Pekhart | May 26, 1989 | FW |
| 18 | Lukas Kuban | Jun. 22, 1987 | DF |
| 19 | Tomas Cihlar | Jun. 24, 1987 | DF |
| 20 | Tomas Frystak | Aug. 18, 1987 | GK |
| 21 | Tomas Micola | Sep. 26, 1988 | MF |

# Canadian Hockey Players

- Cutoff birthdate is the key
- Only accept kids who are not yet 10 on Jan 1
- January Kids matured almost one extra year over December kids

# Detecting Outliers

- One SD variation should cover 68.3% in normal distribution
- Two SDs should cover 95.4% in normal distribution
- Visual detection



- Harder in multivariate case. Why?
  - ❑ May be univariate or multivariate outlier

# Bivariate Outliers

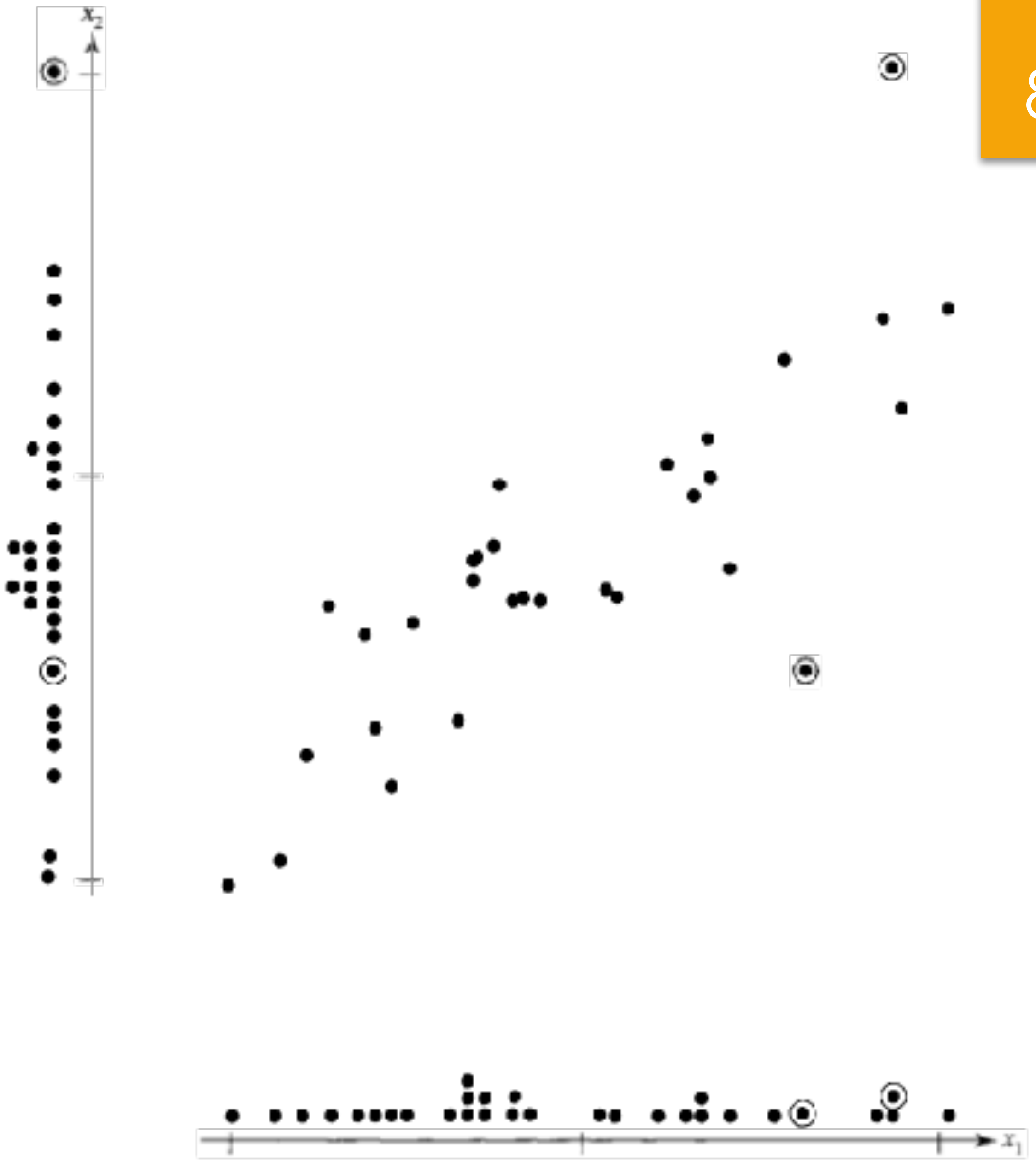


**Figure 4.10** Two outliers; one univariate and one bivariate.

# Multivariate Outliers

- Some outliers are hard to detect
- Look for large values of
  - ❑

$$(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}).$$

$$\mathbf{X} = \begin{bmatrix} & & \cdots & \\ x_{11} & x_{12} & x_{1k} & x_{1p} \\ & & \cdots & \\ x_{21} & x_{22} & x_{2k} & x_{2p} \\ & & \cdots & \\ & \vdots & & \vdots \\ x_{j1} & x_{j2} & x_{jk} & x_{jp} \\ & & \cdots & \\ & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{nk} & x_{np} \end{bmatrix}$$

- p variables
- n items/samples

# Sample Covariance & Correlation

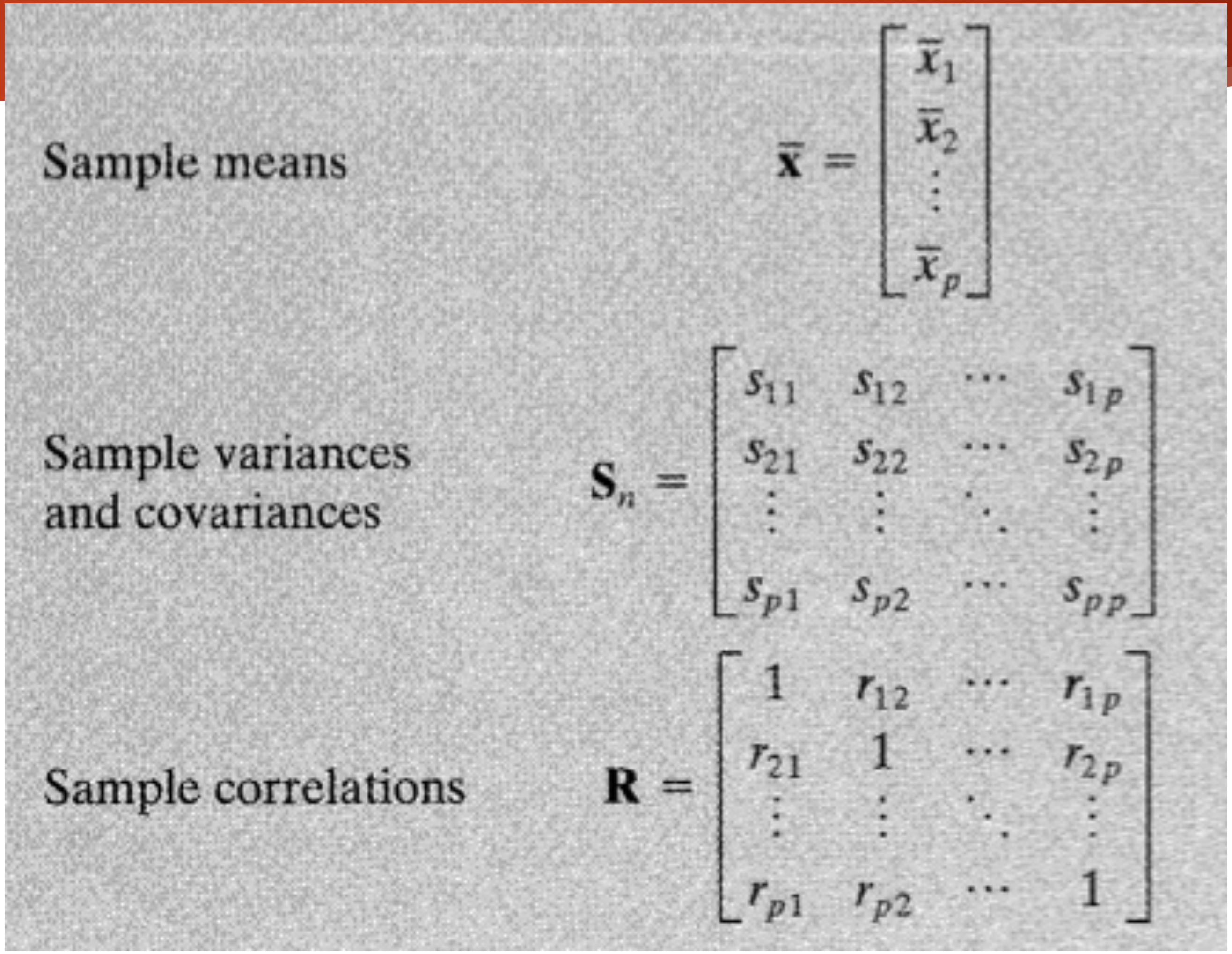
The *sample covariance*

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \qquad i = 1, 2, \ldots, p, \quad k = 1, 2, \ldots, p$$

The sample correlation coefficient for the $i$th and $k$th variables is defined as

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\,\sqrt{s_{kk}}} = \frac{\displaystyle\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\displaystyle\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2}\sqrt{\displaystyle\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2}}$$

# Basic Descriptive Statistics



Sample means
$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Sample variances and covariances
$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Sample correlations
$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

# Outlier detection

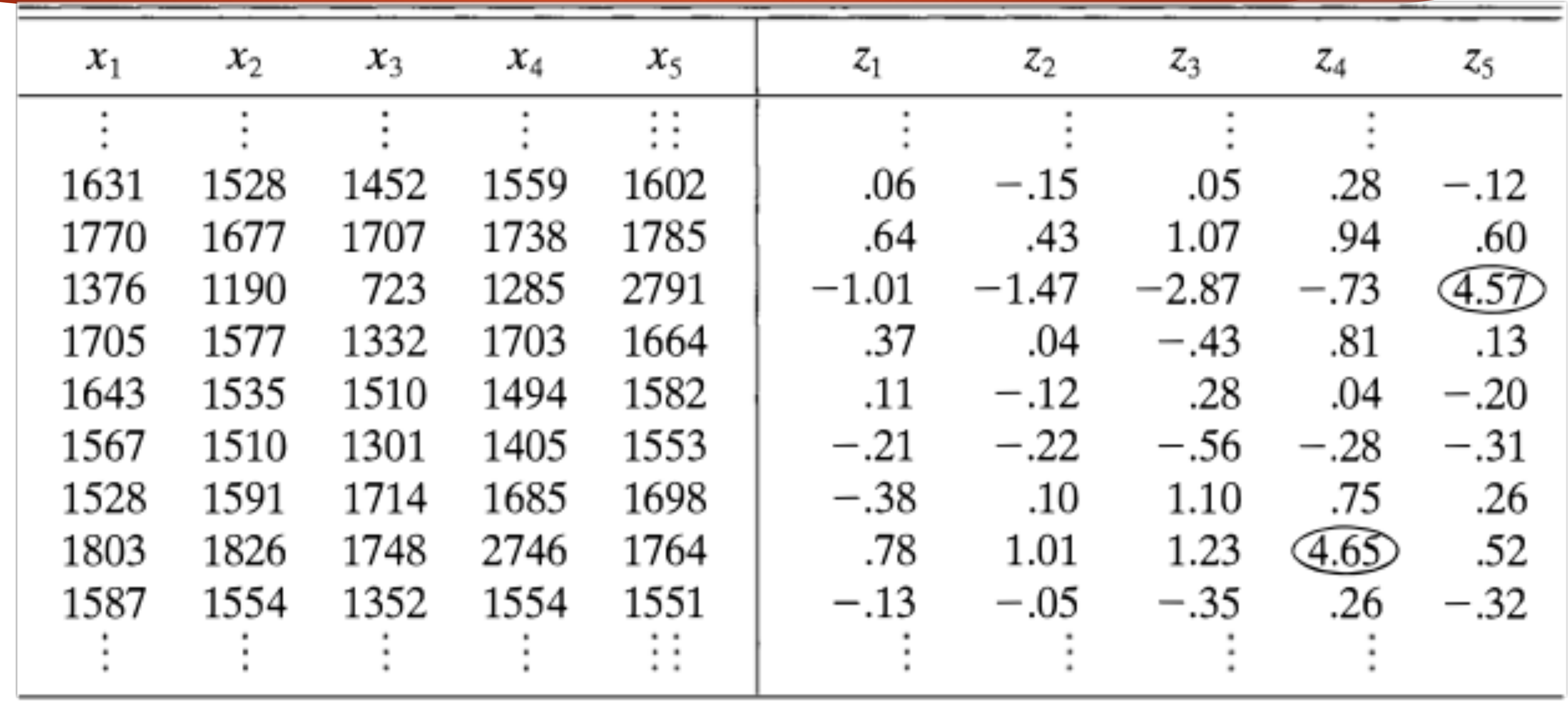
- Dot plots for each variable
- Scatter plot for each pair of variables
- Calculate z-values and examine for outliers

$$z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$$

- Calculate gen sq distances & look for outliers

$$(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}).$$

# Spotting outliers from z-values



| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $z_5$ |
|---|---|---|---|---|---|---|---|---|---|
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |
| 1631 | 1528 | 1452 | 1559 | 1602 | .06 | −.15 | .05 | .28 | −.12 |
| 1770 | 1677 | 1707 | 1738 | 1785 | .64 | .43 | 1.07 | .94 | .60 |
| 1376 | 1190 | 723 | 1285 | 2791 | −1.01 | −1.47 | −2.87 | −.73 | 4.57 |
| 1705 | 1577 | 1332 | 1703 | 1664 | .37 | .04 | −.43 | .81 | .13 |
| 1643 | 1535 | 1510 | 1494 | 1582 | .11 | −.12 | .28 | .04 | −.20 |
| 1567 | 1510 | 1301 | 1405 | 1553 | −.21 | −.22 | −.56 | −.28 | −.31 |
| 1528 | 1591 | 1714 | 1685 | 1698 | −.38 | .10 | 1.10 | .75 | .26 |
| 1803 | 1826 | 1748 | 2746 | 1764 | .78 | 1.01 | 1.23 | 4.65 | .52 |
| 1587 | 1554 | 1352 | 1554 | 1551 | −.13 | −.05 | −.35 | .26 | −.32 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |

# Spotting outliers from Gen. Distance values

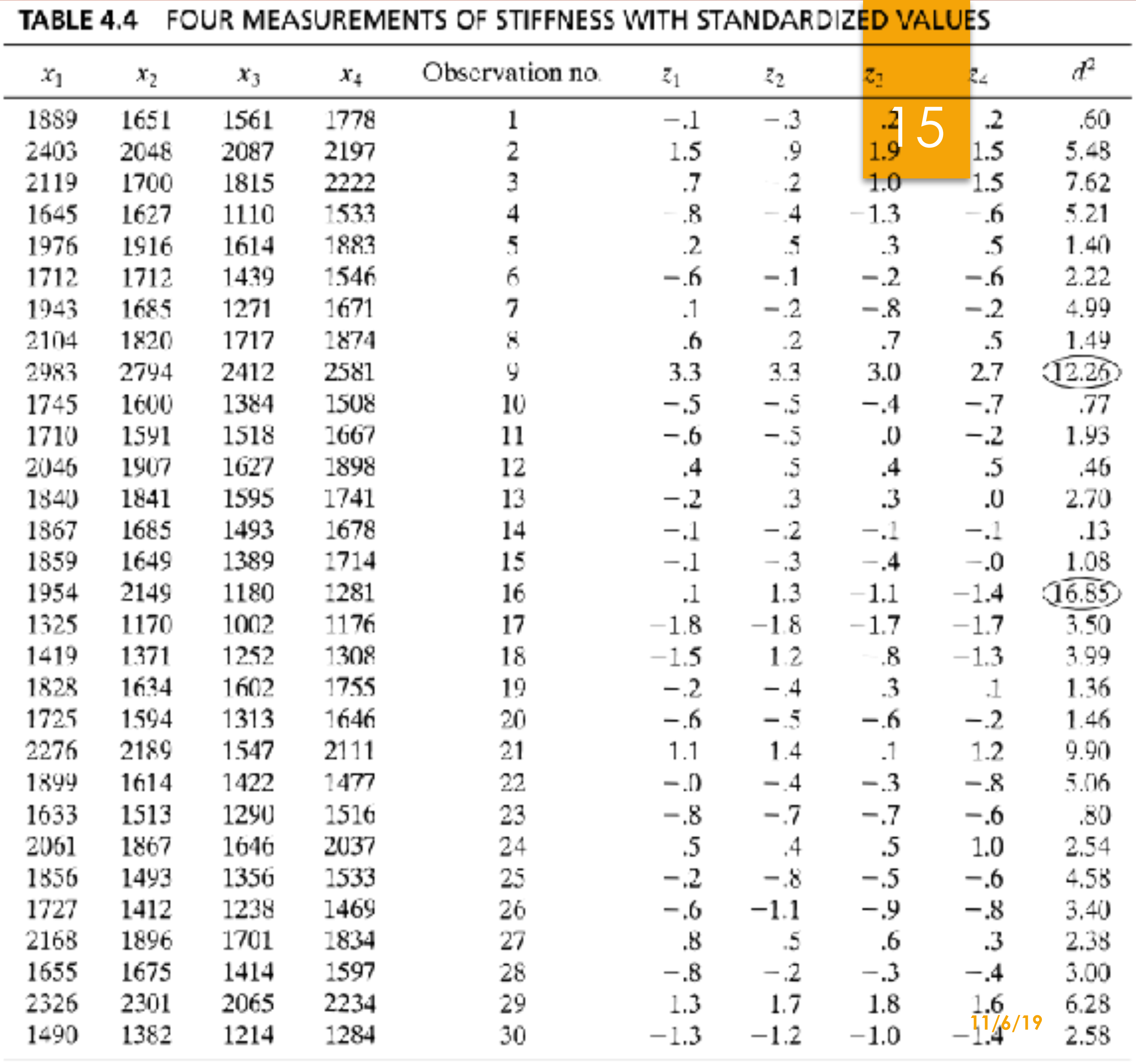
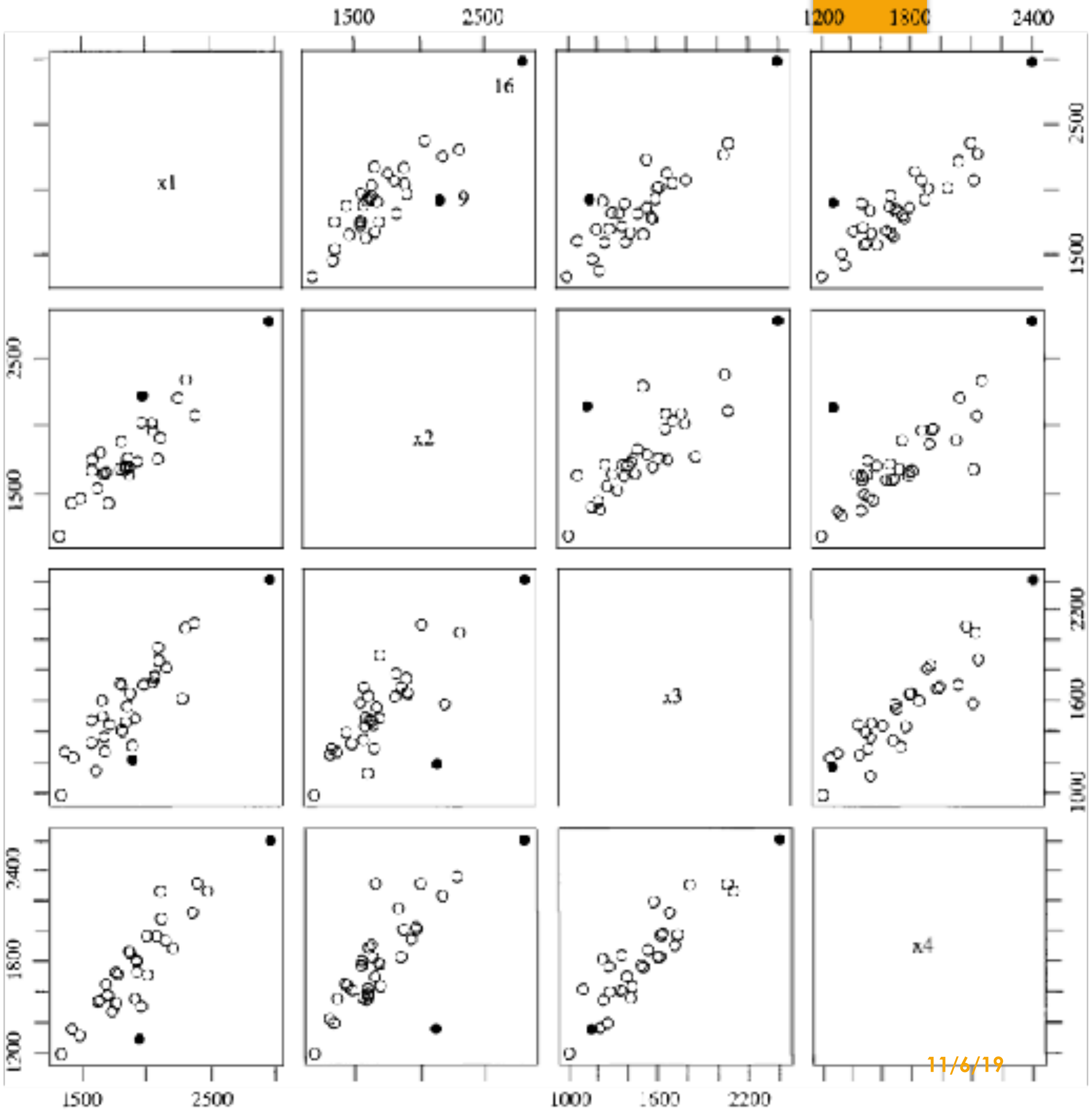


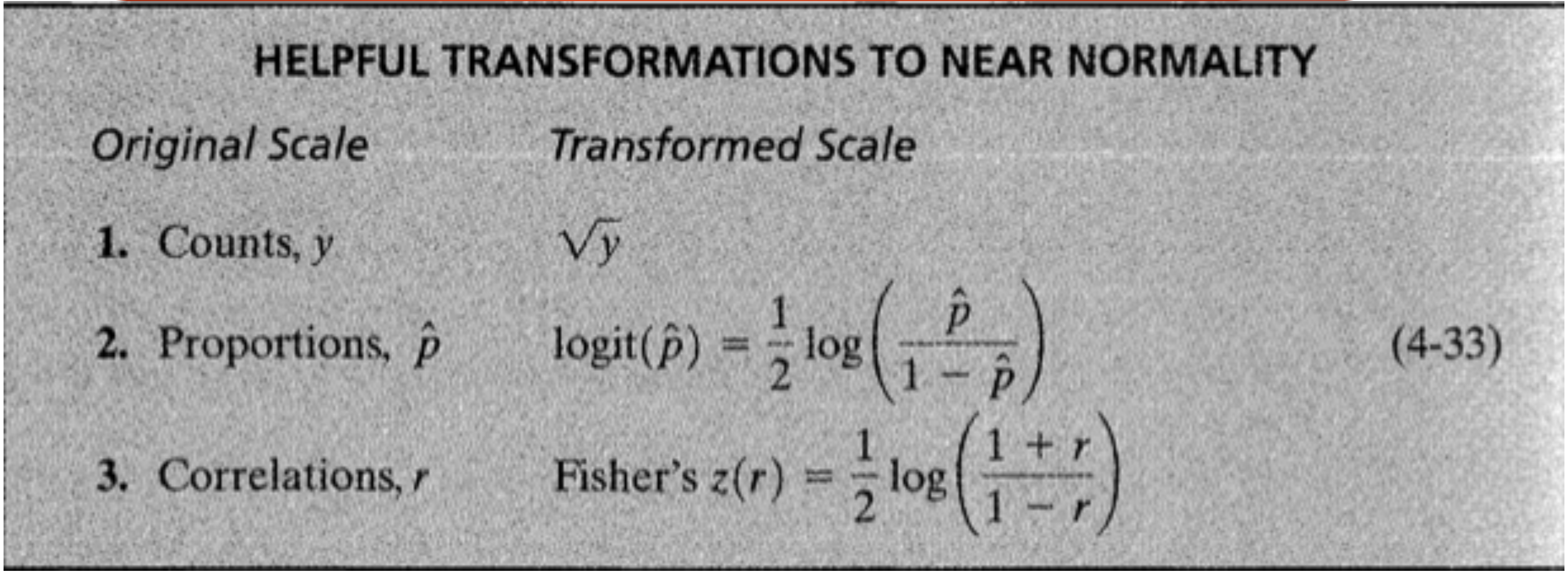
TABLE 4.4 FOUR MEASUREMENTS OF STIFFNESS WITH STANDARDIZED VALUES

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | Observation no. | $z_1$ | $z_2$ | $z_3$ | $z_4$ | $d^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 1889 | 1651 | 1561 | 1778 | 1 | −.1 | −.3 | .2 | .2 | .60 |
| 2403 | 2048 | 2087 | 2197 | 2 | 1.5 | .9 | 1.9 | 1.5 | 5.48 |
| 2119 | 1700 | 1815 | 2222 | 3 | .7 | −.2 | 1.0 | 1.5 | 7.62 |
| 1645 | 1627 | 1110 | 1533 | 4 | −.8 | −.4 | −1.3 | −.6 | 5.21 |
| 1976 | 1916 | 1614 | 1883 | 5 | .2 | .5 | .3 | .5 | 1.40 |
| 1712 | 1712 | 1439 | 1546 | 6 | −.6 | −.1 | −.2 | −.6 | 2.22 |
| 1943 | 1685 | 1271 | 1671 | 7 | .1 | −.2 | −.8 | −.2 | 4.99 |
| 2104 | 1820 | 1717 | 1874 | 8 | .6 | .2 | .7 | .5 | 1.49 |
| 2983 | 2794 | 2412 | 2581 | 9 | 3.3 | 3.3 | 3.0 | 2.7 | 12.26 |
| 1745 | 1600 | 1384 | 1508 | 10 | −.5 | −.5 | −.4 | −.7 | .77 |
| 1710 | 1591 | 1518 | 1667 | 11 | −.6 | −.5 | .0 | −.2 | 1.93 |
| 2046 | 1907 | 1627 | 1898 | 12 | .4 | .5 | .4 | .5 | .46 |
| 1840 | 1841 | 1595 | 1741 | 13 | −.2 | .3 | .3 | .0 | 2.70 |
| 1867 | 1685 | 1493 | 1678 | 14 | −.1 | −.2 | −.1 | −.1 | .13 |
| 1859 | 1649 | 1389 | 1714 | 15 | −.1 | −.3 | −.4 | −.0 | 1.08 |
| 1954 | 2149 | 1180 | 1281 | 16 | .1 | 1.3 | −1.1 | −1.4 | 16.85 |
| 1325 | 1170 | 1002 | 1176 | 17 | −1.8 | −1.8 | −1.7 | −1.7 | 3.50 |
| 1419 | 1371 | 1252 | 1308 | 18 | −1.5 | 1.2 | −.8 | −1.3 | 3.99 |
| 1828 | 1634 | 1602 | 1755 | 19 | −.2 | −.4 | .3 | .1 | 1.36 |
| 1725 | 1594 | 1313 | 1646 | 20 | −.6 | −.5 | −.6 | −.2 | 1.46 |
| 2276 | 2189 | 1547 | 2111 | 21 | 1.1 | 1.4 | .1 | 1.2 | 9.90 |
| 1899 | 1614 | 1422 | 1477 | 22 | −.0 | −.4 | −.3 | −.8 | 5.06 |
| 1633 | 1513 | 1290 | 1516 | 23 | −.8 | −.7 | −.7 | −.6 | .80 |
| 2061 | 1867 | 1646 | 2037 | 24 | .5 | .4 | .5 | 1.0 | 2.54 |
| 1856 | 1493 | 1356 | 1533 | 25 | −.2 | −.8 | −.5 | −.6 | 4.58 |
| 1727 | 1412 | 1238 | 1469 | 26 | −.6 | −1.1 | −.9 | −.8 | 3.40 |
| 2168 | 1896 | 1701 | 1834 | 27 | .8 | .5 | .6 | .3 | 2.38 |
| 1655 | 1675 | 1414 | 1597 | 28 | −.8 | −.2 | −.3 | −.4 | 3.00 |
| 2326 | 2301 | 2065 | 2234 | 29 | 1.3 | 1.7 | 1.8 | 1.6 | 6.28 |
| 1490 | 1382 | 1214 | 1284 | 30 | −1.3 | −1.2 | −1.0 | −1.4 | 2.58 |

# Harder to spot them on scatter plots!



# Other Transforms for Normality



**HELPFUL TRANSFORMATIONS TO NEAR NORMALITY**

| *Original Scale* | *Transformed Scale* | |
|---|---|---|
| **1.** Counts, $y$ | $\sqrt{y}$ | |
| **2.** Proportions, $\hat{p}$ | $\text{logit}(\hat{p}) = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ | (4-33) |
| **3.** Correlations, $r$ | Fisher's $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$ | |

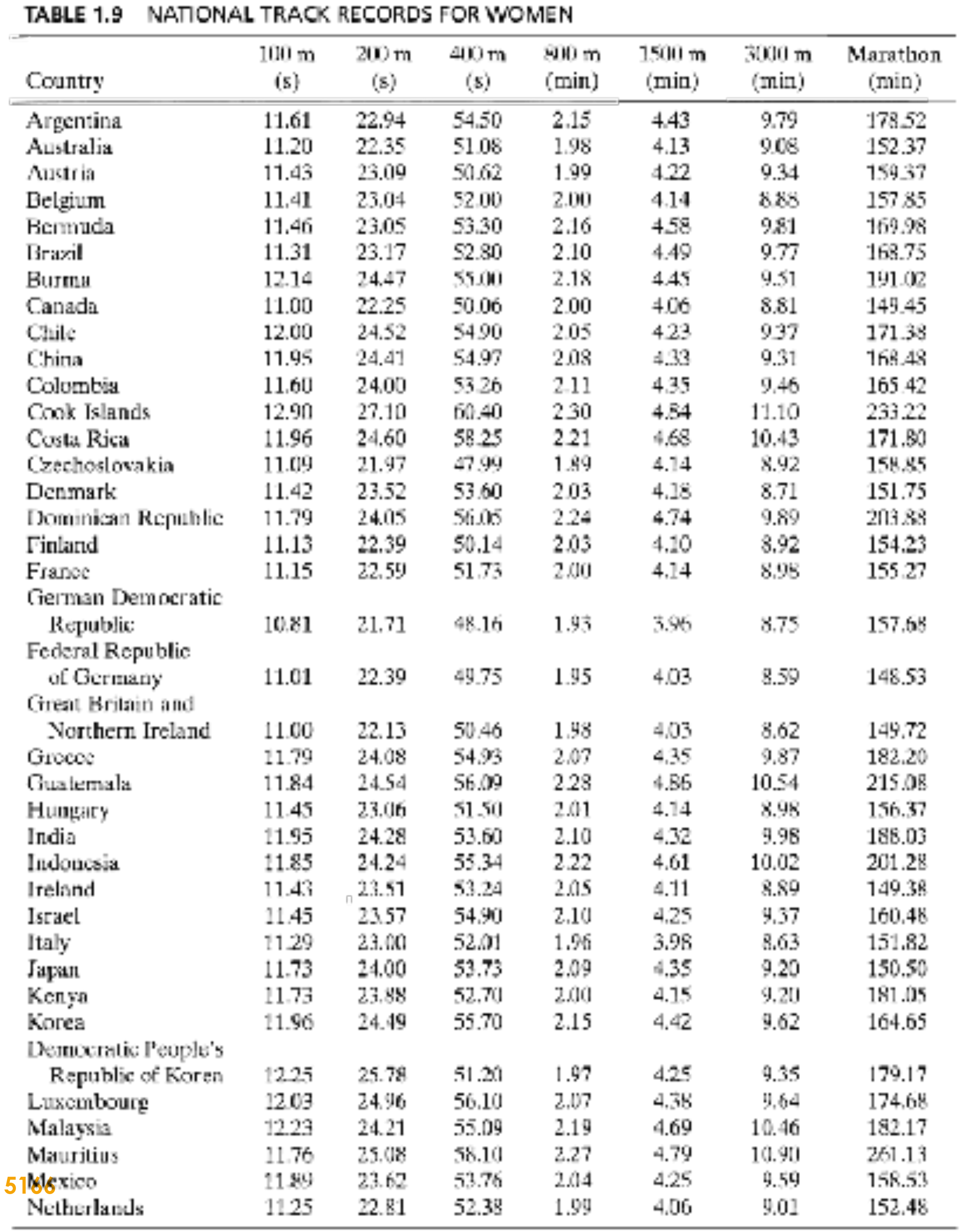


TABLE 1.9 NATIONAL TRACK RECORDS FOR WOMEN

| Country | 100 m<br>(s) | 200 m<br>(s) | 400 m<br>(s) | 800 m<br>(min) | 1500 m<br>(min) | 3000 m<br>(min) | Marathon<br>(min) |
|---|---|---|---|---|---|---|---|
| Argentina | 11.61 | 22.94 | 54.50 | 2.15 | 4.43 | 9.79 | 178.52 |
| Australia | 11.20 | 22.35 | 51.08 | 1.98 | 4.13 | 9.08 | 152.37 |
| Austria | 11.43 | 23.09 | 50.62 | 1.99 | 4.22 | 9.34 | 159.37 |
| Belgium | 11.41 | 23.04 | 52.00 | 2.00 | 4.14 | 8.88 | 157.85 |
| Bermuda | 11.46 | 23.05 | 53.30 | 2.16 | 4.58 | 9.81 | 169.98 |
| Brazil | 11.31 | 23.17 | 52.80 | 2.10 | 4.49 | 9.77 | 168.75 |
| Burma | 12.14 | 24.47 | 55.00 | 2.18 | 4.45 | 9.51 | 191.02 |
| Canada | 11.00 | 22.25 | 50.06 | 2.00 | 4.06 | 8.81 | 149.45 |
| Chile | 12.00 | 24.52 | 54.90 | 2.05 | 4.23 | 9.37 | 171.38 |
| China | 11.95 | 24.41 | 54.97 | 2.08 | 4.33 | 9.31 | 168.48 |
| Colombia | 11.60 | 24.00 | 53.26 | 2.11 | 4.35 | 9.46 | 165.42 |
| Cook Islands | 12.90 | 27.10 | 60.40 | 2.30 | 4.84 | 11.10 | 233.22 |
| Costa Rica | 11.96 | 24.60 | 58.25 | 2.21 | 4.68 | 10.43 | 171.80 |
| Czechoslovakia | 11.09 | 21.97 | 47.99 | 1.89 | 4.14 | 8.92 | 158.85 |
| Denmark | 11.42 | 23.52 | 53.60 | 2.03 | 4.18 | 8.71 | 151.75 |
| Dominican Republic | 11.79 | 24.05 | 56.05 | 2.24 | 4.74 | 9.89 | 203.88 |
| Finland | 11.13 | 22.39 | 50.14 | 2.03 | 4.10 | 8.92 | 154.23 |
| France | 11.15 | 22.59 | 51.73 | 2.00 | 4.14 | 8.98 | 155.27 |
| German Democratic Republic | 10.81 | 21.71 | 48.16 | 1.93 | 3.96 | 8.75 | 157.68 |
| Federal Republic of Germany | 11.01 | 22.39 | 49.75 | 1.95 | 4.03 | 8.59 | 148.53 |
| Great Britain and Northern Ireland | 11.00 | 22.13 | 50.46 | 1.98 | 4.03 | 8.62 | 149.72 |
| Greece | 11.79 | 24.08 | 54.93 | 2.07 | 4.35 | 9.87 | 182.20 |
| Guatemala | 11.84 | 24.54 | 56.09 | 2.28 | 4.86 | 10.54 | 215.08 |
| Hungary | 11.45 | 23.06 | 51.50 | 2.01 | 4.14 | 8.98 | 156.37 |
| India | 11.95 | 24.28 | 53.60 | 2.10 | 4.32 | 9.98 | 188.03 |
| Indonesia | 11.85 | 24.24 | 55.34 | 2.22 | 4.61 | 10.02 | 201.28 |
| Ireland | 11.43 | 23.51 | 53.24 | 2.05 | 4.11 | 8.89 | 149.38 |
| Israel | 11.45 | 23.57 | 54.90 | 2.10 | 4.25 | 9.37 | 160.48 |
| Italy | 11.29 | 23.00 | 52.01 | 1.96 | 3.98 | 8.63 | 151.82 |
| Japan | 11.73 | 24.00 | 53.73 | 2.09 | 4.35 | 9.20 | 150.50 |
| Kenya | 11.73 | 23.88 | 52.70 | 2.00 | 4.15 | 9.20 | 181.05 |
| Korea | 11.96 | 24.49 | 55.70 | 2.15 | 4.42 | 9.62 | 164.65 |
| Democratic People's Republic of Korea | 12.25 | 25.78 | 51.20 | 1.97 | 4.25 | 9.35 | 179.17 |
| Luxembourg | 12.03 | 24.96 | 56.10 | 2.07 | 4.38 | 9.64 | 174.68 |
| Malaysia | 12.23 | 24.21 | 55.09 | 2.19 | 4.69 | 10.46 | 182.17 |
| Mauritius | 11.76 | 25.08 | 58.10 | 2.27 | 4.79 | 10.90 | 261.13 |
| Mexico | 11.89 | 23.62 | 53.76 | 2.04 | 4.25 | 9.59 | 158.53 |
| Netherlands | 11.25 | 22.81 | 52.38 | 1.99 | 4.06 | 9.01 | 152.48 |

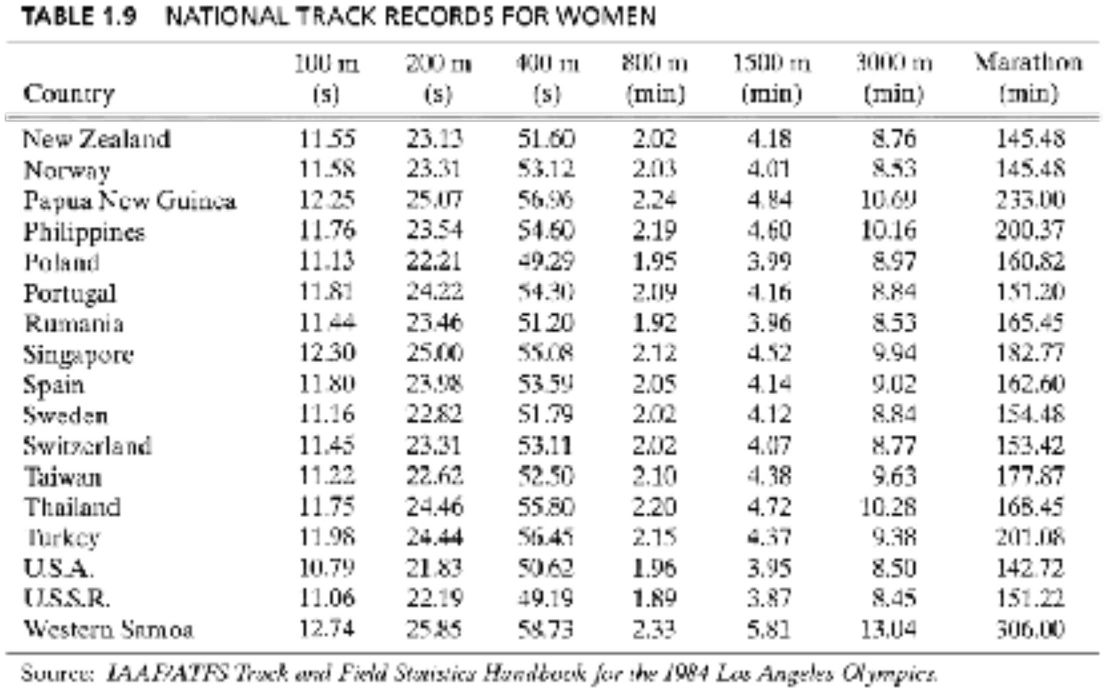


TABLE 1.9 NATIONAL TRACK RECORDS FOR WOMEN

| Country | 100 m<br>(s) | 200 m<br>(s) | 400 m<br>(s) | 800 m<br>(min) | 1500 m<br>(min) | 3000 m<br>(min) | Marathon<br>(min) |
|---|---|---|---|---|---|---|---|
| New Zealand | 11.55 | 23.13 | 51.60 | 2.02 | 4.18 | 8.76 | 145.48 |
| Norway | 11.58 | 23.31 | 53.12 | 2.03 | 4.01 | 8.53 | 145.48 |
| Papua New Guinea | 12.25 | 25.07 | 56.96 | 2.24 | 4.84 | 10.69 | 233.00 |
| Philippines | 11.76 | 23.54 | 54.60 | 2.19 | 4.60 | 10.16 | 200.37 |
| Poland | 11.13 | 22.21 | 49.29 | 1.95 | 3.99 | 8.97 | 160.82 |
| Portugal | 11.81 | 24.22 | 54.30 | 2.09 | 4.16 | 8.84 | 151.20 |
| Rumania | 11.44 | 23.46 | 51.20 | 1.92 | 3.96 | 8.53 | 165.45 |
| Singapore | 12.30 | 25.00 | 55.08 | 2.12 | 4.52 | 9.94 | 182.77 |
| Spain | 11.80 | 23.98 | 53.59 | 2.05 | 4.14 | 9.02 | 162.60 |
| Sweden | 11.16 | 22.82 | 51.79 | 2.02 | 4.12 | 8.84 | 154.48 |
| Switzerland | 11.45 | 23.31 | 53.11 | 2.02 | 4.07 | 8.77 | 153.42 |
| Taiwan | 11.22 | 22.62 | 52.50 | 2.10 | 4.38 | 9.63 | 177.87 |
| Thailand | 11.75 | 24.46 | 55.80 | 2.20 | 4.72 | 10.28 | 168.45 |
| Turkey | 11.98 | 24.44 | 56.45 | 2.15 | 4.37 | 9.38 | 201.08 |
| U.S.A. | 10.79 | 21.83 | 50.62 | 1.96 | 3.95 | 8.50 | 142.72 |
| U.S.S.R. | 11.06 | 22.19 | 49.19 | 1.89 | 3.87 | 8.45 | 151.22 |
| Western Samoa | 12.74 | 25.85 | 58.73 | 2.33 | 5.81 | 13.04 | 306.00 |

Source: *IAAF/ATFS Track and Field Statistics Handbook for the 1984 Los Angeles Olympics.*

# Hypothesis Testing

HTTPS://365DATASCIENCE.COM/WP-CONTENT/UPLOADS/2019/05/365-DATA-SCIENCE_HYPOTHESIS-TESTING-CHEAT-SHEET.PDF

# Regression Analysis

FORM OF PREDICTIVE MODELLING

- A **DEPENDENT** (TARGET) AND
- **INDEPENDENT VARIABLE (S)** (PREDICTOR).
- USED FOR FORECASTING, TIME SERIES MODELLING AND FINDING CAUSAL EFFECT RELATIONSHIP
  - E.g., relationship between driver age and # of road accidents by driver

# Regression Analysis

## Linear Regression

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma x y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Evaluate model using **R-squared measure**

$$R^2 = 1 - \frac{\mathrm{MSE(model)}}{\mathrm{MSE(baseline)}}$$

# Logistic Regression

- Logistic regression used to find probability of event=**Success** and event=**Failure**
- Use logistic regression when dependent variable is binary (0/ 1, True/ False, Yes/ No)
- widely used for classification problems

# Other types of regression

- Non-linear/polynomial regression
- Stepwise
- Ridge
- Lasso
- ElasticNet
- …

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/