2 Accuracy, Precision, Sources of Data, Representing Data

2.1 Accuracy, Precision, and Being Correct

In statistics and science in general, *accuracy* is the degree of closeness to reality or truth. Thus if a restaurant has sold exactly 244 pizzas today, the accurate number is 244; 243 and 245 are less accurate approximations or can be simply described as *inaccurate* or *wrong*.

Precision is the degree of uncertainty in a measure; in our restaurant example, 243, 244 and 245 are equally precise even though 243 and 245 are inaccurate.

Continuing our pizza example, if we decide to count the 244 pizzas in dozens, describing the sales as *20 ⅓ dozen pizzas* is perfectly accurate and perfectly precise. Describing the sales as *20.3333* pizzas is less precise and slightly less accurate: one-third is actually an infinitely repeating decimal, so 0.3333 is a little less than 0.3333333333….. Estimates of *20.333*, *20.33*, *20.3 20* or "*more than 10 and less than 30*" dozen are all decreasingly precise and also decreasingly accurate because of the decreasing precision. However, if we imagine that there are 30 pizzas in an order, the values 30., 30.0, 30.00, 30.000 and so on may have increasing precision, but they all have equal accuracy.

A different issue is *being correct* in the *choice* of statistical methods and in the *execution* of those methods. Many applications of statistics provide significant benefits to users – if the analyses and calculations are correct. If the wrong statistical procedure is applied to a problem or if a valid procedure is performed sloppily, resulting in incorrect results, the consequences can be severe. To emphasize the importance of being careful, precise and correct, I have consistently given a zero grade to incorrect responses to statistical assignments regardless of the pathetic mewling of students used to getting partial credit for trying to apply the right technique.

In the real world of work, no one legitimately does statistical analysis just to look good. Presenting a mistaken analysis of which marketing campaign had the best results in selling a new product can result in collapse of a new company. Make a mistake in a critical calculation and you could be fired. But making a mistake in the calculations of how much of a new drug should be administered to patients as a function of their weight could make sick people even sicker – or kill them. Miscalculating the effects of temperature on an O-ring on a space shuttle booster actually did kill astronauts.[24, 25]

Students must get used to *checking their work* as professionals do. Make a habit of checking what you've done before you move on to the next step. When you are using computer programs that you have written or entering values into routines from statistical packages, you will find it helpful to try your calculations with data for which you know the answers.

---

[24] (Stathopoulos 2012)

[25] **Personal note:** When I was teaching applied statistics in Rwanda, Africa from 1976 through 1978, my students were astonished at being given zeros for having made arithmetic errors. However, almost all of the students in my classes in the Faculté des sciences économiques et sociales at the Université nationale du Rwanda were headed for government jobs, so I put the issue in terms that made sense to them. I told them to imagine that they were responsible for estimating how much of the international development aid should be spent on Hepatitis C vaccination programs in the struggling nation. They could base their estimates of appropriate costs on knowledge of the unit costs of the vaccine, the costs of training and paying health workers, the number of people to be vaccinated, and the statistics on average losses of the vaccines through accident and errors. So what would happen if they made an arithmetic mistake in their calculations and failed to notice that they were, say, underestimating the appropriate cost by an order of magnitude? Imagine that instead of allocating the equivalent of U$100,000 to the project, they erroneously reserved only U$10,000 – and 790 men, women and children succumbed to Hepatitis C? Would they argue that they should get partial credit despite the mistake in their calculations? Would the victims' families agree? There was silence in the classroom that day, and no one ever complained again about having to check their work step by step as they went through their problem solving.

2.2 Significant Figures

Significant figures reflect the degree of uncertainty in a measurement.[26] If we count 29 Ping-Pong balls in a bag, the number 29 has 2 significant figures and is exact: there is no uncertainty about it. However, if we weigh the 29 Ping-Pong balls and have a total weight of 115.2 g, the weight has four significant figures and represents a value that could actually be between 115.150 g and 115.249 g. This weight to four significant figures is expressed in *scientific notation* as 1.152×10^2 g or as 1.152e2 g.

Here are some examples of constants that can be expressed with different numbers of significant figures:

- π represents the ratio of the circumference of a circle to its diameter. It can be expressed as, say, 3.14159 using six significant figures but has been calculated to more than a million decimal digits;[27]
- *e*, the limit of $(1 + 1/n)^n$ and the base of natural logarithms, to nine significant figures is 2.71828183 but continues without limit to the number of digits that can theoretically be computed;[28]
- *c*, the speed of light in a vacuum and the maximum speed of anything in the known universe except science-fiction ships using faster-than-light travel, can be given as 299,792,458 meters per second (a number with nine significant figures also); if it is expressed as 186,282 miles per second then it has six significant figures; and if it is expressed as 1.079e9 km/hour then it has four significant figures;[29]
- *Avogadro's number* to seven significant figures is 6.022142×10^{23} particles /mole (with many more significant digits available – but not an infinite number of significant figures because molecular weight involves discrete particles and therefore can in theory be counted to an exact integer). This number is the number of atoms in 12 grams of pure carbon-12 and defines a *mole* of a material.[30]

Most other numbers we work with are naturally variables: salaries, returns on investment, quality percentages, customer satisfaction responses, and contaminant measurements. So to how many significant figures should we express them?

The issue depends on how variable a measure is compared with its size in the scale of measurement we want to use. For example, suppose we are looking at the number of Internet Protocol (IP) packets arriving every second at a specific port in a firewall; this number fluctuates from moment to moment, but we can actually know exactly how many packets there are from the log files kept by the firewall software. However, if someone asks, "How many packets per second were reaching port 25 during the last hour?" answering with a list of 3,600 numbers – one total per second for each of the 3,600 seconds of the hour – is likely to be met with astonishment, if not hostility. We naturally want to express the number using a summary, so we compute an average (we will study those in detail later) and answer using the average.

INSTANT TEST Page 2

Suppose a cubic nanometer of pure Ultronium is estimated to have 17,298 atoms – which has 5 significant figures.

Express this number in scientific notation with the e format (e.g., 542.3 = 5.423e2) to 1 significant figure, then to 2 significant figures, then to 3 significant figures and finally to 4 significant figures.

Notice also what happens to 17,298 when you express it to 4 and then to 3 significant figures. . . .

[26] (Morgan 2010)
[27] (University of Exeter 2012)
[28] (Wolfram Mathworld 2012)
[29] (Fowler 2009)
[30] (Cambell 2011)

2.3 Determining Suitable Precision for Statistics

How precisely should we express derived statistics such as an average? Consider the following two cases:

- Case 1: there were a total of 129,357,389 packets received over the hour for an average of exactly 49,752.8419230769 packets per second. The lowest rate was 7,288 per second and the highest rate was 92,223 per second.
- Case 2: the same number of packets was received, so the average was the same as in Case 1, but this time the lowest rate was 49,708 packets in a second and the highest rate was 49,793 packets in a single second.

So how should we express the data and the averages in these two cases?

Precision refers to the "closeness of repeated measurements of the same quantity" as Robert R. Sokal and F. James Rohlf put it in their classic statistics textbook[31] which influenced generations of biologists.[32,33,34]

In both of our examples, the 15 significant figures of 49,752.8419230769 seem pointlessly and incorrectly precise. That number implies that the real average value is approximately between 49,752.84192307685 and 49,752.84192307695 but such precision is misleading when the actual observed variation is between 7,288 and 92,223 packets per second in the first case and 49,708 and 49,793 packets per second in the second case.

A general approach for deciding on appropriate precision is that we should express results for original observations so that the *range* (difference between the smallest and the largest observed values) is *divided into roughly 30 to 300 steps or divisions*. In addition, *derived statistics* based on our data typically have *one more significant digit* than the original data.

The easiest approach to deciding on significant figures is to estimate the number of steps at different levels of precision. After a while, one becomes good at guessing a reasonable degree of precision. Using our two cases,

- Case 1: a range from a minimum of 7,288 per second up to a maximum of 92,223 per second means that
  - One significant figure would result in a value of 90,000 packets per second (i.e., 9 units of 10,000 packets) for the upper limit and thus the lower limit would be 10,000 packets per second (1 unit of 10,000 packets – not 7,000, which would imply a precision of thousands rather than of tens of thousands).
  - Reporting to *two* significant figures would give us values of 92,000 and 7,000; the number of thousands would be 92 – 7 +1 = 86 steps or divisions (thousands).[35] That would be perfect: right in the middle of the 30-300 steps rule. Notice that the minimum has *fewer significant figures* in this case to maintain consistency with the upper value.
  - To estimate how many steps there would be if you increased the number of significant figures by one (e.g., 92,200 in which we tally the number of hundreds), you can either multiply the previous number of divisions at the current level (86) by 10 for a rough approximation (~860 divisions) or you can insist on being precise and do the arithmetic (922 hundreds – 73 hundreds + 1 = 850 divisions) to see how many steps there would be with *three* significant figures. That's too many divisions for the raw data.
  - However, the average *would* be reported with one more significant figure than the two used for the data, thus giving *three* significant figures for derived statistics. Suppose the average were, say, 49,758.268. This average would thus be reported reasonably with three significant

[31] (Sokal and Rohlf, Biometry: The Principles and Practice of Statistics in Biological Research 1981)

[32] I studied the First Edition in 1969 and helped proofread the Second Edition in 1980, getting mentioned in the acknowledgements [beams proudly] and also receiving a surprise check for U$400 (about U$1,200 in 2012 dollars) that enabled me to buy an excellent touring bicycle!

[33] (I'm really old.)

[34] (Compared to you).

[35] Why is there a +1? Because if you subtract a lower number from a higher number, you don't get the total number of values including them; e.g., 6 – 3 = 3 but there are actually 4 values included in that range: *3*, *4*, *5* & *6*.

figures as 49,800 packets per second. Note that there cannot be a decimal point there because it would incorrectly suggest that we were reporting using five significant figures.

- Case 2: the lowest rate was 49,708 packets in a second and the highest rate was 49,793. Let's also suppose that the average was as above: 49,758.268.
    - Using what we found for Case 1 (*three* significant figures) as a guess, we see that the range based on the minimum and maximum would be reported as 49,700 – 49,800 = only *two* steps.
    - It's easy to see that if we use four significant figures, there would be 49,790 - 49,710 + 1 = 81 steps which is perfectly within the 30-300 guidelines.
    - So four significant figures would be great for these raw data and therefore the average would be reported with one more significant figure = *5*. The average (e.g., 49,758.268) would thus be reported reasonably as 49,753 packets per second. It would be OK to use a decimal point (49,753.) If such a number ended in a zero (e.g., some other average computed as, say, 49,750) it would be reported with a terminal decimal point (47,750.) to be clear that it was expressed with 5 significant figures.

To repeat this last note on significant figures of numbers ending in zero that are expressed in ordinary notation: be careful about the use of a decimal point when you have relatively few significant figures. For example, if a number expressed to 3 significant figures is *12,300* it would be a mistake to write it as *12,300.* (note the period) because that would imply that we were reporting to 5 significant figures. This kind of nuisance explains why we often choose to use scientific notation when correct communication of numerical values is paramount.

Scientific notation uses an integer between 1 and 9 and is followed by an appropriate number of decimal digits. For example,

- 1,234.56789 expressed to 6 significant figures would be shown as 1.23457×10^3 or as 1.23457e3 (this latter format is particularly useful because it works in Excel)
- The same number reduced to 4 significant figures would be shown as 1.235×10^3 or 1.235e3
- With only significant figures, the number would be shown as 1.23×10^3 or 1.23e3
- For numbers smaller than 1, count the number of places you have to move the decimal point to arrive at the integer portion of the scientific notation. Thus a value of
    - 0.0001234 to 3 significant figures would be 1.23×10^{-4} or 1.23e-4
    - 0.00456789 to 5 significant figures would be 4.5679×10^{-3} or 4.5679e-3

INSTANT TEST Page 2-5

(1) The average cost of shares for Urgonium Corporation over the last month is calculated as 123.456789 credits. Express the average in non-scientific notation using

- 2 significant figures
- 3 significant figures
- 4 significant figures
- 8 significant figures

(2) Express the average above in scientific notation using

- 2 significant figures
- 3 significant figures
- 4 significant figures
- 8 significant figures

(cont'd)

(3) The smallest cost observed (the minimum) for Urgonium Corporation shares was 118 credits and the largest value observed (the maximum) was 128 credits. How many steps would there be in the range if you used 2, 3, 4, or 8 significant figures for these limits?

(4) The minimum number of retaining bolts per strut on the Garagano Narrows Bridge is 17,826 and the maximum number is 22,012. The average number is 19,943.24 Which of the following is an appropriate number of significant figures for these data? Why?

a) 17,826.0 | 19,943.24 | 22,012.0

b) 17,826. | 19,943.2 | 22,012.

c) 17,830. | 19,943. | 22,010

d) 17,800 | 19,940 | 22,000

e) 18,000 | 19,900 | 20,000

f) 20,000 | 20,000 | 20,000

2.4 Sources of Real Statistical Data

At many points in this course, you will be given the opportunity to apply newly learned methods to real-world data. There are many sources of such data; e.g., looking at United States resources first,

- US Bureau of Labor Statistics[36] has a wealth of economic data such as
    - Inflation & Prices
    - Unemployment
    - Employment
    - Spending & Time Use
    - Pay & Benefits
    - Productivity
    - Workplace injuries
    - International economic comparisons
- US Bureau of Justice Statistics[37] provides data about crime and justice such as
    - Corrections
    - Courts & Sentencing
    - Crime Type
    - Criminal Justice Data Improvement Program
    - Employment & Expenditure
    - Federal
    - Law Enforcement
    - Victims
- US Census Bureau[38] provides information about
    - People & Households Business & Industry Geography
    - Fraudulent Activity & Scams
    - Census Bureau Data & Emergency Preparedness
    - The Statistical Abstract of the United States, "the authoritative and comprehensive summary of statistics on the social, political, and economic organization of the United States. Use the Abstract as a convenient volume for statistical reference, and as a guide to sources of more information both in print and on the Web."
- US Bureau of Transportation Statistics[39] has the "mission is to create, manage, and share transportation statistical knowledge with public and private transportation communities and the Nation."
- US Centers for Disease Control and Prevention (CDC)[40] provide verified data about health issues such as
    - Diseases & Conditions
    - Emergency Preparedness & Response
    - Environmental Health
    - Life States & Populations

---

[36] < http://www.bls.gov/ >
[37] < http://bjs.ojp.usdoj.gov/ >
[38] < http://www.census.gov/ >
[39] < http://www.bts.gov/ >
[40] < http://www.cdc.gov/ >

- Healthy Living
- Injury, Violence & Safety
- Travelers' Health
- Workplace Safety & Health

- US Washington Headquarters Services (WHS)[41] whose Pentagon Library offers links to many sources of statistical information, especially focused on military matters.
- World Health Organization (WHO)[42] of the United Nations provides extensive information (in many languages) about health and wellbeing around the world such as
  - Mortality & Health Status
  - Diseases
  - Coverage of health services
  - Risk factors (alcohol, nutrition, overweight & obesity, tobacco)
  - Health Systems
- Pew Research Center covers a vast array of topics[43] comprising
  - Demography
  - Domestic Policy
  - Economics
  - Election '08
  - Election '12
  - Energy and Environment
  - Global Attitudes/Foreign Affairs
  - Immigration
  - Internet and Technology
  - Legal
  - News Media
  - Politics and Elections
  - Polling History
  - Public Opinion
  - Religion
  - Research Methodology
  - Social Trends
- Sports statistics sites[44]
- Internet World Stats[45] provides global information and details broken down by region
  - Africa
  - America
  - Asia
  - Europe

---

[41] < http://www.whs.mil/ >
[42] < http://www.who.int/research/en/ >
[43] < http://pewresearch.org/topics/ >
[44] Use a search engine with the search term *sports statistics* to find many Websites providing such information
[45] < http://www.internetworldstats.com/stats.htm >

  - European Union
  - Mid-East
  - Oceania
- Scientific and Professional Journals
  - Visit your university or local library to enquire about how to access to thousands of scientific and professional journals available online that often include detailed statistical information
  - At Norwich University, the Kreitzberg Library Databases[46] are available online for all members of the Norwich community and for all disciplines.

In general, most government and survey sources provide direct access to downloadable data in various formats. If you need to work with the original data used in a scientific or professional publication, you may be able to ask the analysts for access to raw data files; although you may not always gain access, it's worth trying. You should explain exactly what you intend to do with the data and should offer to provide the original analysts with the results of your work. If the data are not in the public domain (e.g., published by the US government) then you must explicitly ask for permission to use copyright material in your work.

Failing direct access to usable data formats, you can use conversion programs; for example, Adobe Acrobat Professional includes tools for recognizing and interpreting text from scanned documents or other PDF files. However, optical character recognition (OCR) of all types is never perfect; you should check the conversions with minute attention to detail. Publishing an analysis that uses incorrect values is a career-killer.

Finally, you may simply have to re-enter data; again, be sure to double-check the accuracy of the transcription.

Remember, in addition to ruining your reputation, appropriating data without permission or misstating the data from other people's work could lead to lawsuits for copyright infringement and for defamation (putting someone in false light).

[46] < http://www.norwich.edu/academics/library/databases.html >

INSTANT TEST P 2-9

Using your own particular interests, look up some interesting statistical information from among the resources listed or others and report on what you find by posting a note in the appropriate discussion group on NUoodle.

Examples of potentially interesting statistics:

- Examine the changing distributions of wealth and income in the USA over the last several decades
- Compile information about educational attainment in different demographic sectors
- Evaluate the US position on health care delivery and population health compared with other countries' records
- Gather information about construction trends – what kinds of building materials are being used – and how have these patterns changed over time?
- Evaluate fuel efficiency statistics for different kinds of vehicles; e.g., various types of cars, trains, ships and airplanes
- Track the popularity of different sports or teams over time and in different regions of the USA or of the world
- Learn about the changing patterns of Internet, cell phone, computer usage

2.5 Representing Data

Both to clarify our own understanding of data and to help others, we often show data in various forms – tables, charts, graphs and other formats. The most important principle in choosing a data representation is clarity: the format must communicate our intention with a minimum of ambiguity.

2.6 Presenting Raw Data

There are many situations in which the specific sequence of data collection is valuable for our research – either because it reflects a time sequence relating the observations to the external world (for example, stock-market price fluctuations that may be related to news events).

Figure 2-1. Sample of raw data in Excel.

	A
1	Raw Data
2	282
3	211
4	433
5	229
6	126
7	254
8	334
9	373
10	462
11	170
48	241
49	428
50	151

There may also be interest in internal data sequences such as the likelihood that a particular pattern will follow a specific sequence of data (this kind of analysis includes what is called *Markov chain* analysis) such as machine learning algorithms to identify customer buying patterns (more likely to buy a certain brand if the customer has just bought two other products in a row) or network attack patterns (e.g., perhaps probes on ports 80 and 25 in the last five minutes may increase the likelihood of probes on port 20 within the next 10 minutes).

Although we rarely publish thousands of individual observations in a table, it is frequently the case that we must produce the original data for inspection. A simple approach is to list the observations in a single column or row, but with thousands of values this method is impractical for publication. Figure 2-1shows the top and bottom parts of a simple table of 50 rows (a label plus 49 values).

Creating multi-column tables from data in a single column is not trivial in EXCEL; there are no facilities in that program for automatically showing parts of the single column in side-by-side segments. However, word-processing packages such as MS-WORD make it easy to convert raw data into a useful and professionally presentable format. These packages include facilities for creating columns, so it becomes trivially easy to convert the data.

We can import the data into a column and then format them as text into several columns, as shown in Figure 2-2. Be careful not to make the columns so narrow that numbers begin breaking into different lines. Reduce the number of columns if that happens.

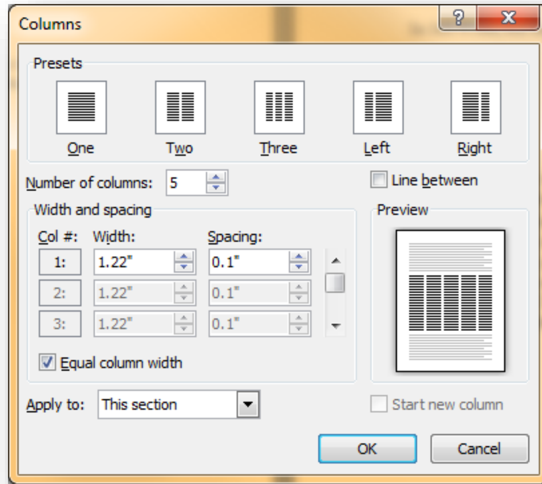
Figure 2-2. Simple presentation of data in columns.

Raw Data:

282	345	428	224	141
211	472	176	295	127
433	351	148	358	233
229	269	159	162	322
126	492	471	139	306
254	192	390	115	263
334	460	303	352	241
373	419	374	223	428
462	287	161	495	151
170	222	452	260	

In MS-WORD, the dialog for Columns offers many options, as shown in Figure 2-3. It's easy to determine the number of columns and the spacing between each pair. There's also an option for adding a vertical line between columns.

Figure 2-3. MS-WORD dialog box for defining columns.



Enhancing a multi-column table of data is easy in WORD. For example, Figure 2-4 shows a few rows of data in an EXCEL table.

Figure 2-4. Multi-column table in Excel.

	A	B	C
1	Variable 1	Variable 2	Variable 3
2	282	6.5	8.0E+03
3	211	9.4	3.4E+04
4	433	9.0	3.2E+04
5	229	6.6	4.3E+04
6	126	7.9	2.4E+04
7	254	5.3	1.3E+04
8	334	6.6	3.4E+04
9	373	9.1	2.4E+04
10	462	8.9	1.2E+04
11	170	6.2	1.7E+04

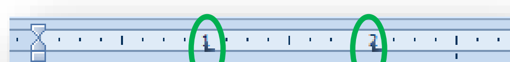
Once the data are pasted into WORD, columns are separated using the Tab character. In Figure 2-6, the data are shown using the Show Paragraph function (¶), which is a toggle that is also accessible using the key-combination Shift-Ctl-*. The arrows represent the Tab character and the paragraph symbols show the end-of-line character. These characters are not normally visible – they are in the snapshot to show how columns are converted from EXCEL into WORD –columns are separated by Tab characters in every row.

The simple approach to aligning the labels with the columns is to highlight the data and insert appropriately seven spaced Tab marks in the WORD ruler, as shown in Figure 2-5. We'll come back to cosmetic enhancements of tables in WORD in a few pages.

Figure 2-5. Data pasted into WORD showing tab and paragraph marks.

Variable·1	Variable·2	Variable·3¶
282	6.5	8.0E+03¶
211	9.4	3.4E+04¶
433	9.0	3.2E+04¶
229	6.6	4.3E+04¶
126	7.9	2.4E+04¶
254	5.3	1.3E+04¶
334	6.6	3.4E+04¶
373	9.1	2.4E+04¶
462	8.9	1.2E+04¶
170	6.2	1.7E+04¶

Figure 2-6. Tab marks in ruler used to align labels with imported data.

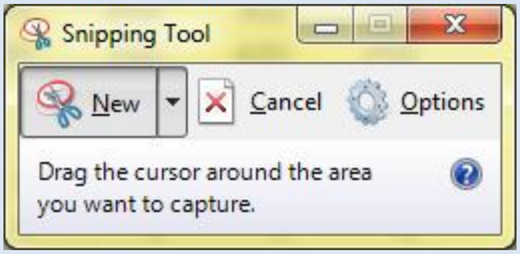


Variable 1	Variable 2	Variable 3
229	6.5	8.0E+03
433	5.3	8.0E+03
462	8.9	1.1E+04
229	7.9	1.1E+04
126	9.1	8.0E+03
126	9.1	1.1E+04

INSTANT TEST P 2-12

Using data that you find on the Web to reflect your own interests, practice the following skills:

- Capturing a *screenshot* of an interesting table (e.g., using Windows 7 "Snipping Tool" or clicking on a window and pressing Alt-PrtScn)



- Pasting an image into a WORD document using Ctl-Alt-V
- Copying data from a Web page by highlighting them and then pasting them into a WORD document in various formats (Ctl-Alt-V again)
- Taking the same data from the Web page you found and pasting them into an Excel spreadsheet
- Copying a table created in Word and pasting the data into Excel in various formats
- Copying a table created in Excel and pasting the data into Word in various formats
- Writing out a list of numbers with spaces between them and pasting them into Excel (not pretty, eh?)
- Converting the data in your Word file into the same list with TAB characters between them instead of spaces and then pasting the new list into Excel
- Creating a list of e-mail addresses in Excel, then copying the list into Word and replacing the TAB characters by the string <;^P > (not including < and >) to created a semicolon-delimited list with a space after every semicolon so you can paste the list into an e-mail BCC field.