HARALD SCHMIDBAUER İstanbul Bilgi Üniversitesi

# Bus 701: Advanced Statistics

Fall 2011

## SOME SIMPLE PROBLEMS

**Problem 1:** Obtaining a sound data set is often a laborious process. File `insurance_data_w_errors.xls` contains (simulated, but realistic) data on 1000 insurants: date of birth, the date they entered the insurance contract, gender, the number of damages they incurred in 2008, and the corresponding total damage amount. — Unfortunately, some typical errors have crept into this data set: Data of nine cases need to be fixed before the data set can be analyzed.

a) Find the errors and fix them in a plausible way.

b) How many people with at least one damage in 2008 are there?

c) Compute the average damage amount (the arithmetic mean) among those who incurred a damage.

**Problem 2:** This exercise shows that care must be taken when we translate numerical information into a verbal formulation. Consider two statements:

$A$: "34% of all households are single-person households."
$B$: "34% of all people live in single-person households."

a) Are $A$ and $B$ equivalent?
b) Assuming $A$ is true, make an approximate statement about the share of people living in a single-person household.

**Problem 3:** Suppose you want to buy a used car, and you prefer a certain model. You find the following offers in newspapers:

| no. | year | 1000 kms | price |
|---|---|---|---|
| 1 | 1994 | 100 | 31 |
| 2 | 1992 | 127 | 28 |
| 3 | 1993 | 82 | 29 |
| 4 | 1992 | 101 | 27 |
| 5 | 1992 | 125 | 23 |
| 6 | 1992 | 103 | 29 |
| 7 | 1994 | 94 | 30 |
| 8 | 1994 | 149 | 26 |
| 9 | 1992 | 81 | 24 |
| 10 | 1994 | 89 | 29 |
| 11 | 1992 | 97 | 25 |
| 12 | 1994 | 93 | 35 |
| 13 | 1994 | 59 | 33 |

| no. | year | 1000 kms | price |
|---|---|---|---|
| 14 | 1992 | 106 | 27 |
| 15 | 1992 | 93 | 27 |
| 16 | 1993 | 90 | 32 |
| 17 | 1994 | 170 | 24 |
| 18 | 1992 | 160 | 19 |
| 19 | 1994 | 140 | 35 |
| 20 | 1992 | 118 | 28 |
| 21 | 1993 | 127 | 23 |
| 22 | 1992 | 80 | 27 |
| 23 | 1994 | 60 | 39 |
| 24 | 1992 | 55 | 31 |
| 25 | 1993 | 115 | 29 |
| 26 | 1994 | 145 | 27 |

Here, the price is in DM 1000. (This was the situation of the BMW 325i coupé model in Germany in August 1998.)

a) Draw a stemplot of these data with the price as variable. (Hint: Use categories 1•, 2⋆, 2•, 3⋆, 3•; 1•: price from DM 15000 up to below 20000 etc.)

b) In your stemplot, mark all cars which were built in 1992 with a bar below the "leaf" (e.g., $\underline{7}$).

c) In the same stemplot, mark all cars which have run 125000 or more kilometers with a bar above the "leaf" (e.g., $\overline{7}$).

d) You are offered a 1993 model of this type in good condition, 100000 kilometers, at DM 25000. Is this a good offer? Use the stemplot to answer this question.

**Problem 4:** An automobile repair garage analyzes regularly the daily number of customers (that is, the daily number of cars brought for inspection), in order to be able to react to new circumstances immediately. Car arrivals in November were:

| date | weekday | number | date | weekday | number | date | weekday | number |
|---|---|---|---|---|---|---|---|---|
| Nov 01 | Friday | 43 | Nov 11 | Monday | 45 | Nov 21 | Thursday | 28 |
| Nov 02 | Saturday | 23 | Nov 12 | Tuesday | 39 | Nov 22 | Friday | 26 |
| Nov 03 | Sunday | 0 | Nov 13 | Wednesday | 28 | Nov 23 | Saturday | 26 |
| Nov 04 | Monday | 42 | Nov 14 | Thursday | 35 | Nov 24 | Sunday | 3 |
| Nov 05 | Tuesday | 30 | Nov 15 | Friday | 22 | Nov 25 | Monday | 27 |
| Nov 06 | Wednesday | 22 | Nov 16 | Saturday | 34 | Nov 26 | Tuesday | 26 |
| Nov 07 | Thursday | 30 | Nov 17 | Sunday | 11 | Nov 27 | Wednesday | 34 |
| Nov 08 | Friday | 30 | Nov 18 | Monday | 42 | Nov 28 | Thursday | 31 |
| Nov 09 | Saturday | 34 | Nov 19 | Tuesday | 22 | Nov 29 | Friday | 28 |
| Nov 10 | Sunday | 16 | Nov 20 | Wednesday | 34 | Nov 30 | Saturday | 31 |

a) Draw a stemplot of the variable *number of arrivals*.
b) Compute the arithmetic mean.
c) Compute the median.

(Hints: The ordered set of observations is: 0, 3, 11, 16, 22, 22, 22, 23, 26, 26, 26, 27, 28, 28, 28, 30, 30, 30, 31, 31, 34, 34, 34, 34, 35, 39, 42, 42, 43, 45. Furthermore, $\sum x_i = 842$, $\sum x_i^2 = 26\,870$.)

**Problem 5:** Consider the variable $X$ = age (completed years) of insurants of a car insurance company. A sample of size $n = 20$ is:

25, 27, 30, 37, 38, 40, 41, 42, 43, 44, 46, 47, 49, 50, 53, 57, 62, 65, 65, 66

a) Draw a stem-and-leaf display of these data.

b) Determine the median.

c) Will this median equal the median of *all* insurants? (Give reasons for your answer.)

d) What is the scaling of the variable $X$?

**Problem 6:** Several times a year, OPEC (Organization of the Petroleum Exporting Countries) hosts meetings among its members to agree on further oil production policies. OPEC then announces which decision concerning world oil production levels (no change, increase, or cut) was made. (For those of you who are interested in crude oil prices and the OPEC, please find a list of announcement dates and other details on the next page. See also http://opec.org.)
The chart below shows a stemplot (with 50 observations) of the variable

$X$ = number of days between two successive announcements, in the time period 2000 through 2009.

```
 1 | 5
 2 | 002589
 3 | 1123334557899
 4 | 23334559
 5 | 24458889
 6 | 001349
 7 | 11
 8 | 18
 9 |
10 |
11 | 5
12 | 49
13 | 1
```

The mean of $X$ is 51.96 and the standard deviation of $X$ is 26.76.

a) Explain what the row `8 | 18` means.

b) Determine the median of this distribution.

c) Determine the lower (first) and upper (third) quartile of this distribution.

d) Make a boxplot of this distribution.

e) Make a histogram of this distribution, based on the breakpoints 0, 20, 30, 40, 50, 60, 70, 140. (The first interval is $[0, 20)$, the second $[20, 30)$, etc.)

f) What can you say about the skewness of this distribution? Explain also in terms of median and arithmetic mean.

g) Does the one-sigma rule hold in this example?

h) Using this data set, estimate the probability that it takes more than 100 days until OPEC makes the next announcement.

**Problem 7:** Consider the following variables: population of a country; political party a person voted for in the last general elections; letter grade of a student in an exam; total time a student spent on studying for BUS 273; temperature in °F. Determine the scaling of each of these variables.

**Problem 8:** This example makes strong model assumptions, but contributes to understanding the length sampling bias. Two types, $A$ and $B$, of items are stored in an inventory. Their sojourn times are

type $A$: 10 minutes,
type $B$: 20 minutes.

The interarrival time (i.e. the time interval between two successive arrivals) is 5 minutes. The two types arrive alternatingly, so that a type $B$ item arrives after a type $A$ item, and vice versa.

a) Draw a Becker diagram which represents this situation.

b) For each type, find the number of items present at an arbitrary instant.

c) On length sampling, what is the average sojourn time of an item in the inventory? (In other words, what is the average sojourn time of the items which we encounter at an arbitrary instant?)

d) What is the average sojourn time of all incoming items?

**Problem 9:** A tourist inquiry was carried out in Istanbul at Easter: Interviewers asked 183 tourists whom they met randomly in the streets of Sultanahmet how long they would stay in Istanbul. It turned out that their mean sojourn time was 1.75 days, that is, the arithmetic mean of the 183 tourists was 1.75 days. The number of tourists who said they would stay at least three days was 17.

a) Is 1.75 a reliable estimate for the average sojourn time of a tourist in Istanbul at Easter?

b) Is 17/183=9.3% a reliable estimate for the share of tourists who stay at least three days?

c) How should the data be collected in order to estimate the average sojourn time and the share of tourists who stay at least three days reliably?

**Problem 10:** Suppose that you are zapping through TV channels. Is it true that you are more likely to encounter a long TV commercial than a short one? Explain.

**Problem 11:** A researcher wants to carry out an inquiry about the motivation and expectations of tourists visiting Istanbul. A questionnaire is to be developed, which will finally be used in face-to-face interviews with tourists.

a) A first draft of the questionnaire has as first question:

> "Do you like spending your holiday in a country with an Islam-based culture?"

Do you think this is a good idea for a first question in the questionnaire? Do you have a better idea? Discuss briefly.

b) Should the question

> "Why did you choose Turkey for your holiday?"

be included in the questionnaire? Is there a better way to investigate why tourists choose Turkey? Develop some ideas as to how a professional questionnaire should be structured in this case.

**Problem 12:** File `bank_customers.xls` contains (simulated, but realistic) data concerning 5868 customers of a bank. Each row corresponds to one customer. The variables and their values are:

| symbol | description | values |
|---|---|---|
| deficit | Credit limit exceeded within last 6 months? | 0: no<br>1: yes |
| age | age in completed years | positive integer number |
| m.status | marital status | 0: single or not specified<br>1: married |
| edu | highest level of education achieved | 1: no graduation or general-education graduation<br>2: vocational school, apprenticeship<br>3: university of applied sciences, university |
| econ.act | economic activity | 0: unemployed or inactive<br>1: employee or public servant<br>2: self-employed (or family member) |
| urban | Is place of residence in urban area? | 0: no<br>1: yes |
| stability | Is residential area stable? (Many moves during recent period indicate "no".) | 0: no<br>1: yes |
| cellphone | number of cellular phone contracts | 0: no contract<br>1: one contract<br>2: two or more contracts |

a) Determine the scaling of each variable.

b) Find the percentage of customers with `deficit = 1`.

c) Compute the average age (here: the arithmetic mean) of customers with `deficit = 0`.

d) Compute the average age (here: the arithmetic mean) of customers with `deficit = 1`.

e) Compute the percentage of customers with `deficit = 1` among those with no (one, two or more) cellular phone contract(s).

**Problem 13:** The distribution of new passenger vehicles registered in Germany in August of 2009 by colour is:

| colour | number | share in % |
|---|---|---|
| grey | 86973 | 31.54 |
| black | 73782 | 26.76 |
| blue | 37890 | 13.74 |
| white | 27661 | 10.03 |
| red | 27540 | 9.99 |
| yellow | 5887 | 2.14 |
| green | 5047 | 1.83 |
| others | 10934 | 3.97 |
| total | 275714 | 100.00 |

(Source: www.kba.de. This distribution is also given in file `registration_new_cars_by_colour.xls`.)

a) Draw a pie chart of this distribution. Make sure the colours in your pie chart match the colour names.

b) The R help text for the command `pie`, which draws a pie chart, states: "Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas." Make a plot of the distribution which incorporates this criticism.

**Problem 14:** Of the 100 observations on a metric variable $X$, 50 are in the interval $[0, 10)$, 25 are in $[10, 20)$, and another 25 are in $[20, 40]$. Plot a histogram of the distribution of $X$.

**Problem 15:** Three companies $A$, $B$, $C$ were asked the percentage of waste paper they recycled. The result was:

| company | percentage recycled | total quantity of waste paper (per year) |
|---|---|---|
| $A$ | 10% | 10000 kg |
| $B$ | 30% | 900 kg |
| $C$ | 50% | 100 kg |

a) What does the arithmetic mean $\frac{1}{3}(10\% + 30\% + 50\%) = 30\%$ tell us?

b) Compute the average percentage of waste paper recycled in the companies $A$, $B$ and $C$.

**Problem 16:** The developing department of an automotive supplier is testing a new direction sensor. One crucial quality indicator is the offset (the bias) of the sensor, which is an angle measured in degrees. A series of sixty measurements of the offset had mean 0.02 and standard deviation 0.03.

a) How many observations would you expect to be in the one-sigma interval? (Give reasons for your answer.)

b) Compute the "six-sigma interval". Would you expect any measurements to lie outside this interval? Explain briefly how the six-sigma interval could be used in quality management.

**Problem 17:** The following table gives the annual real changes of GDP in Turkey:

| year | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|
| change in GDP | 8.4% | 6.9% | 4.7% | 1.1% | −5.8% |

a) The average annual change in GDP for this period is 2.93%. Write down the *exact expression*, using the values in the above table, which yields this average annual change in percent.

b) Now suppose the annual change equals 2.93% for each year in the period 2010–2014. What would be the *total* change in percent from 2010 through 2014? Write down the correct expression.

**Problem 18:** A car travels as follows:

| distance | speed |
|---|---|
| 10 km | 100 km/h |
| 20 km | 40 km/h |

What is the average speed of the car over the whole distance of 30 km?

**Problem 19:** The result of an elementary analysis of monthly returns (end of month closing quotations) in percent on the Chinese stock index SSEC (Shanghai Securities Exchange Center) in the period form January 2000 through October 2008 (105 observations) is:

| | | | |
|---|---|---|---|
| arithmetic mean, $\bar{r}$ | 0.83 | minimum | −20.31 |
| variance, $s^2$ | 67.96 | lower quartile | −4.68 |
| standard deviation, $s$ | 8.24 | median | 0.81 |
| skewness, $\gamma_1$ | 0.03 | upper quartile | 5.57 |
| kurtosis, $\gamma_2$ | 0.76 | maximum | 27.45 |

a) What does the three-sigma-rule say in this case?

b) Give a rough sketch of the distribution of monthly returns, which reflects the parameters given in the table.