# Statistics 2000, Section 001, Quiz 1 (200 Points)

September 28, 2001, Dr. Jürgen Symanzik

**Your Name:** _____

**First look at all 4 questions. Then start with the question that looks easiest to you. Continue with a more difficult question. Try to answer as many questions as possible in these 50 minutes.**

**Note that you will obtain at least partial credit if you indicate a correct formula but your final result is incorrect. If you just rely on your calculator without indicating the formula that should be used and your result is incorrect, you will obtain no credit at all for this part of a question.**

**Question 1:** Numbers and Graphs (**60 Points**)

After Homework 4, we are somewhat suspicious that the circulation of 1.1 million for "Shape" may have been obtained by some manipulation, e.g., by providing free issues as an advertisement. Therefore, we discard this number as an outlier and redo parts of our analysis of the "Weider Empire" magazine circulation based on the following 9 magazines:

| Magazine Name | Circulation |
|---|---|
| Muscle & Fitness | 450,000 |
| Living Fit | 320,000 |
| Men's Fitness | 300,000 |
| Jump | 300,000 |
| Senior Golfer | 240,000 |
| Fit Pregnancy | 200,000 |
| Prime Health & Fitness | 175,000 |
| Flex | 150,000 |
| Shape Cooks | 130,000 |

Please answer the following questions:

1. Determine the mean magazine circulation for the "Weider" empire based on these 9 magazines. (**10 Points**)

$n = 9$;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9}\sum_{i=1}^{n} x_i$$

$$= \frac{1}{9}\left(0.45 + 0.32 + \ldots + 0.15 + 0.13\right) \cdot 10^6$$

$$= \frac{1}{9} \cdot 2.265 \cdot 10^6 \approx 251,667$$

2. Determine the median magazine circulation for the "Weider" empire based on these 9 magazines. (**10 Points**)

$n = 9$, $n$ odd:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{9+1}{2}\right)} = x_{(5)}$$

$$= 240,000$$

3. Calculate the range, the quartiles, and the interquartile range of the magazine circulation for the "Weider" empire based on these 9 magazines. **(10 Points)**

$R = x_{(9)} - x_{(1)} = 450,000 - 130,000 = 320,000$ (3)

$Q_1 = \frac{150,000 + 175,000}{2} = 162,500$ (2)

$Q_3 = \frac{300,000 + 320,000}{2} = 310,000$ (2)

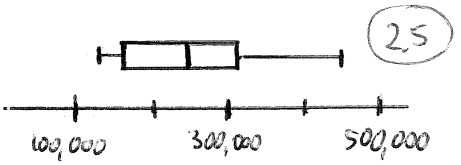
$IQR = Q_3 - Q_1 = 310,000 - 162,500 = 147,500$ (3)

4. Indicate the 5 number summary and construct a boxplot based on your results from (3). Make a clear statement whether the data (i.e., the remaining 9 magazines) contains another outlier. **(10 Points)**

5 number summary: 130,000 162,500 240,000 310,000 450,000 (2.5)

$1.5 \times IQR = 1.5 \times 147,500 = 221,250$ (1)

(2) $Q_1 - 1.5 \times IQR = -58,750$
$Q_3 + 1.5 \times IQR = 531,250$
$\Rightarrow$ no outliers! (2)



(2.5)

5. Using your results from (3) and (4), is the 1.1 million circulation for "Shape" really an outlier? Explain your answer. **(10 Points)**

Suspected outliers would be values greater than 531,250
(in this case, it is impossible to obtain an outlier on the lower end since magazine circulation is positive, but suspected outliers would be $< -58,750$).
So, 1.1 million has a good chance to be in fact an outlier.

6. We have learned by now that the "Weider" group possesses far more than the 9 magazines listed above. Calculate the variance and make a clear statement whether you are calculating a population or a sample variance. **(10 Points)**

The 9 magazines are a subset of all the magazines of the "Weider" group, so we have to calculate the sample variance.

$n = 9, \sum_{i=1}^{n} x_i = 2.265 \cdot 10^6$

$\sum_{i=1}^{n} x_i^2 = 0.652525 \cdot 10^{12}$

$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$

$= 0.652525 \cdot 10^{12} - \frac{(2.265 \cdot 10^6)^2}{9}$

$= 0.0825 \cdot 10^{12}$

$s^2 = \frac{1}{n-1} SS(x)$

$= \frac{1}{8} \cdot 0.0825 \cdot 10^{12}$

$\approx 1.03125 \cdot 10^{10}$ sample variance (3)

$s = \sqrt{s^2}$

$\approx 101,550$ sample standard deviation

**Question 2:** Normal Distribution (**50 Points**)

Part I:
Let $Z$ be a standard Normal variable, i.e., $Z \sim N(0,1)$, and $X$ be a Normal variable with mean $\mu = 3$ and variance $\sigma^2 = 25$, i.e., $X \sim N(3, 5^2)$. Determine the following:
(**5 Points each**)

1. $P(Z < -0.54) =$ $0.2946$

2. $P(X < -0.54) =$ $P\left(\frac{X-3}{5} < \frac{-0.54-3}{5}\right)$
$= P(Z < -0.708)$
$\approx 0.24$

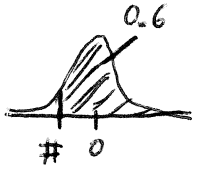
$\Phi(-0.70) = 0.2420$
$\Phi(-0.71) = 0.2389$

3. $P(-2.0 < Z < 1.5) =$ $\Phi(1.5) - \Phi(-2.0)$
$= 0.9332 - 0.0228$
$= 0.9104$

4. $P(-2.0 < X < 1.5) =$ $P\left(\frac{-2.0-3}{5} < \frac{X-3}{5} < \frac{1.5-3}{5}\right)$
$= P(-1.0 < Z < -0.3)$
$= \Phi(-0.3) - \Phi(-1.0) = 0.3821 - 0.1587 = 0.2234$

5. Find a number # such that
$P(Z > \#) = 0.60$ which is equivalent to $P(Z \le \#) = 0.40$.

-2 if sides flipped




Starting at the body of the table, we get
$\Phi(-0.26) \approx 0.3974$
$\Phi(-0.25) \approx 0.4013$
$\Rightarrow \# \approx -0.255$

6. Find a number # such that
$P(X > \#) = 0.60$ which is equivalent to $P(X \le \#) = 0.40$.

-1 if a follow-up to part 5)

Since $Z = \frac{X-3}{5}$, we can solve this equation for $X$ and get
$X = 5Z + 3 \approx 5 \cdot (-0.255) + 3 \approx 1.725$
(i.e., somewhere between 1.70 and 1.75)

Part II:

A college that has an excellent track–and–field athletics program runs short on scholarships and cannot further support all of its 100m track athletes. The athletics director wants to make a decision which athletes to support in the future based on their athletic capabilities. Based on the athletes performance over the last few years, it is known that the distribution of running times is approximately Normal with mean $\mu = 10.8$ sec and standard deviation $\sigma = 0.2$ sec. Answer the following 2 questions:

1. Which qualifying time should the athletics director request such that only 70% of the athletes will be able to achieve this (or a better) time? (**10 Points**)

$X \sim N(10.8, 0.2^2)$

Find # such that

$$P(X \leq \#) = 0.7 = P\left(\frac{X-10.8}{0.2} \leq \frac{\#-10.8}{0.2}\right) = P\left(Z \leq \frac{\#-10.8}{0.2}\right)$$

Starting at the body of the table, we get

$\left.\begin{array}{l} \Phi(0.52) = 0.6985 \\ \Phi(0.53) = 0.7019 \end{array}\right\} \Rightarrow \frac{\#-10.8}{0.2} \approx 0.525$

$\Rightarrow \# \approx 0.2 \cdot 0.525 + 10.8$

$\approx 10.905$

2. What are the chances that any athlete from this college will set a new world record of 9.7 sec or better? (**10 Points**)

almost 0

We do not need to calculate anything here since the the empirical rule indicates that 99.7% of the observations fall in the interval $\mu - 3\sigma$ to $\mu + 3\sigma$, i.e., 10.2 sec to 11.4 sec.

Note that 9.7 sec is $5.5 \times \sigma$ away from $\mu = 10.8$ sec.

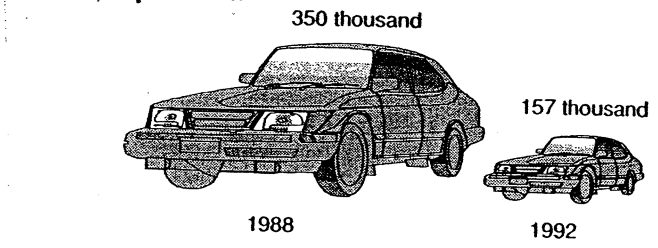
Therefore, almost 100% of the observations fall in the interval $\mu - 5.5\sigma$ to $\mu + 5.5\sigma$, i.e., 9.7 sec to 11.9 sec.

The chance to obtain a time of 9.7 sec or better is minimal.

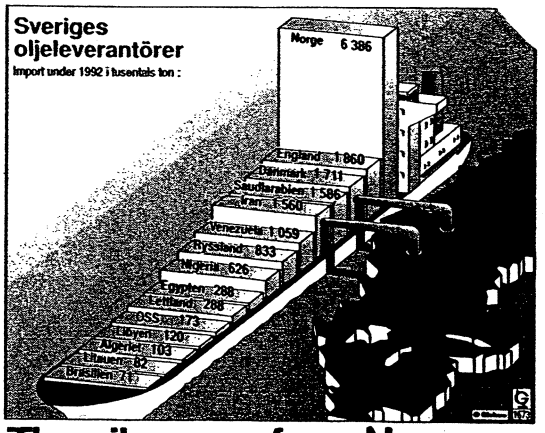
**Question 3:** Newspaper Graphics **(40 Points)**

The following two graphics have been taken from Wallgren et al. (1996) "Graphing Statistics & Data".

A




B




1. For each of the two graphics, determine if there is something wrong with it. If so, carefully explain what is wrong. **(30 Points)**

Graphic A: The numbers 157,000 and 350,000 are represented by area (or even by volume). The larger number should be represented by a symbol that is $\frac{350,000}{157,000} \approx 2.2$ times larger than the smaller symbol.
However, in this graphic the larger car symbol is about 4 to 5 times larger than the smaller car symbol when we compare area (and perhaps 10 to 12 times larger when we see these symbols in 3D, i.e., as volume).
The graphic gives a completely misleading comparison of the 2 numbers.

Graphic B: As a pictogram, this graphic looks perfect. Below the comments from Wallgren.

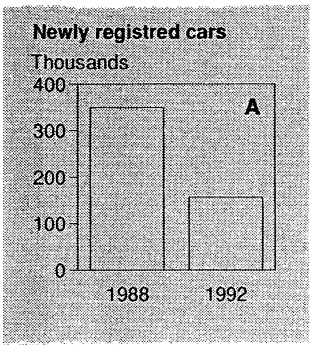
**Chart B: Spot on!**
The picture attracts attention, has a well-founded association with the subject and also gives a good picture of the data since the heights of the bars give an approximate picture of the size of imports. The three-dimensional bars give a natural impression since they are included in the picture as ship's cargo.

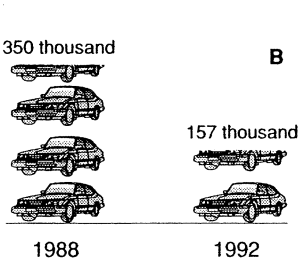
If one wants to complain about anything at all, it is the 3D impression. There is no perspective in this graphic, i.e., the numbers directly relate to height no matter where the boxes are located.

2. For each of the graphics from part (1) above that contains something wrong, draw a sketch how a corrected graphic should look like. **(10 Points)**

Graphic A: Wallgren suggests two alternatives:

- a bar chart
- a pictogram where the numbers are represented by multiple symbols (i.e., cars) of the same size, where each symbol represents 100,000 cars sold




Graphic B: If we want to redraw this graphic, a bar chart certainly would be the best solution. Countries could be listed alphabetically or with respect to increasing (or decreasing) amount of oil imported.

**Question 4:** Micromaps (**50 Points**)

The 2 micromap displays on the next 2 pages have been taken from the U.S. Department of Agriculture, National Agricultural Statistics Service (Research and Development Division) Web site at

`http://www.nass.usda.gov/research/gmcrnyap.htm`

and

`http://www.nass.usda.gov/research/gmcrnapy.htm`.

In map A, the states have been arranged from highest to lowest "Acreage" (in millions of acres) and in map B, the states have been arranged from highest to lowest "Yield" (in bushels per acre). Please answer the following questions:

1. Describe each of the 2 micromap displays (in **5 sentences or less**) with respect to the variable by which the states have been sorted. (**25 Points**)

   Map A (by Acreage): This micromap display emphasizes the "Acreage" of corn grown (in millions of acres) by state. Most acreage of corn exists in the central (Midwestern) states such as IA, IL, NE, MI & IN. Least acreage of corn exists in the western states and the New England states. Overall, corn is grown in 45 of the 50 states.

   Map B (by Yield): This micromap display emphasizes the "Yield" of corn grown (in bushels per acre) by state. Highest yields were obtained in the western states such as WA, OR, NM, CA & AZ. Lowest yields were obtained in the south eastern states such as WV, MD, NC, AL & FL. There is an overall spatial trend from high to low yield of corn from the west/north west to the east/south east

2. Recall how we described the relationship between "Average HAP" and "% Urban Census Tracts" for the "Hazardous Air Pollutants" micromap discussed in class. Now make similar statements (if possible) regarding the relationship between "Acreage" and "Production" as well as "Acreage" and "Yield". (**25 Points**)

   "Acreage" and "Production": There is almost a perfect relationship between "Acreage" and "Production". High "Acreage" results in high "Production", low "Acreage" in low "Production".

   "Acreage" and "Yield": There is no obvious relationship between "Acreage" and "Yield". States with the highest "Acreage" (IA, IL, NE, MI & IN) have a "Yield" that is not very different from the "Yield" for the states with the lowest "Acreage" (MT, VT, CT, MA, NH). Also, WA & FL have almost the same "Acreage" - however WA has the highest "Yield" while FL has the lowest "Yield" from all 45 states. It appears that geographic location mostly determines "Yield" (in bushels per acre!) and not "Acreage".