Exam Applied Statistics in Transport WS 2010/2011

name (in block letters):	**matriculation number:**

(Duration: 60 minutes and 60 points)

(1) (6 points) Please explain the main characteristics of the scales of variables you know.

Nominal (yes, no)
Ordinal (ranking)
Interval (can be negative [temperature])
Ratio (true zero point)

(2) (5 points) The following numbers of inhabitants were registered in the region A:

Year	Inhabitants
1	24,500
2	26,210
3	28,780
4	30,500
5	34,420
6	35,520

Compute the middle growth factor and explain which measure of central tendency you use.

Solution:

We use the geometric mean, because we have here a growth phenomenon, where the future development depends on the previous development.

The formula for the geometric mean is $\overline{x_g} = (\prod_{i=1}^{n} x_i)^{1/n}$

We are interested in the middle growth factor. Hence we need the geometric mean of the growth factors: $\overline{x_g} = \sqrt[5]{1.070 * 1.098 * 1.060 * 1.129 * 1.032} = 1.077$. (we take the 5th root as we have five growth factors).

The middle growth factor is 7.7 percent.

We can also calculate: $\overline{x_g} = \sqrt[5]{\frac{Inhabitants_{n+1}}{Inhabitants_1}} = \sqrt[5]{\frac{35,520}{24,500}} = 1.077$

(3) (7 points) A voice communication system for a business contains 16 external lines. At a particular time, the system is observed, and some of the lines are being used. Let the random variable X denote the number of the voice lines that are in use at a particular time. Assume that X is a discrete uniform random variable. Compute the expected value and the variance for the random variable X. Name two main characteristics of the discrete uniform distribution.

Two main characteristics: discrete variables and all values of a finite set of possible values are equally probable.

$\mu = \frac{b+a}{2}$, $\sigma^2 = \frac{(b-a+1)^2-1}{12}$

(4) (9 points) Speed measurements at a section of a highway show an average speed $\overline{x} = 80km/h$ and a standard deviation of s=10km/h (parameters for the sample). The empirical frequencies show that

the speed is normally distributed, hence, this distribution can be described by a normal distribution with the parameters μ=80km/h and σ=10km/h.

a) Which speed x_1 is fallen below with a probability of 0.15?
b) Which speed x_2 is exceeded with a probability of 0.15?
c) The fastest and the slowest cars (with highest and lowest speed) should not be considered. Which range of speed is covered by the remaining 95% of speed?

To a) z-value 0.15, F(z<=0.15) → z==-1.0365
Transform the z-value:
$z_1 = \frac{x_1-\mu}{\sigma}, x_1 = z_1 * \sigma + \mu = -1.0365 * 10 + 80 = 69.6350$: The speed of 69.6km/h is fallen below with a probability of 0.15/15%.

To b)
This is the so-called 85% speed, for reasons of symmetry (compared to a):
$abs(\mu - x_1) = abs(\mu - x_2) \rightarrow \mu - x_1 = x_2 - \mu \rightarrow\ x_2 = 2\mu - x_1 = 160 - 69.6 = 90.4km/h$
P(X<=x2)=F(x2)=0.85, Standardisation: F(z<=0.85) → z=1.04.
$z_2 = \frac{x_2-\mu}{\sigma}, x_2 = z_2 * \sigma + \mu = 1.04 * 10 + 80 = 90.4$
The speed of 90.4km/h is exceeded with a probability of 0.15/15%.

To c)
P(X<=x3)=F(x3)=0.025, Standardisation: F(z<=0.025) → z=-1.96.
$z_3 = \frac{x_3-\mu}{\sigma}, x_3 = z_3 * \sigma + \mu = -1.96 * 10 + 80 = 60.4$
P(X<=x3)=F(x3)=0.975, Standardisation: F(z<=0.975) → z=1.96.
$z_3 = \frac{x_3-\mu}{\sigma}, x_3 = z_3 * \sigma + \mu = 1.96 * 10 + 80 = 99.6$
This is $\mu \pm 19.6$
The range of 60.4 to 99.6 km/h is covered by the middle 95% of speed.

(5) (5 points) The speed limit of 80km/h was introduced on a rural road. Now the average speed should be measured in order to evaluate the effects of the speed limit. The following requirements were determined: 95% confidence interval, relative error E_r=2%. A preliminary study (n=150 passenger cars) showed: $\overline{x} = 72.8km/h; s = 13.2km/h$. What is the minimal sample size for meeting the requirements for accuracy?

$E_r = \frac{E_a}{\overline{x}}$; $E_a = 0.02 * 72.8 = 1.4560$; $n \geq \frac{z_{(\alpha/2)}{}^2 * \hat{\sigma}^2}{E_a{}^2} \geq \frac{z_{(\alpha/2)}{}^2 * \hat{\sigma}^2}{E_r{}^2 * \overline{x}^2} = \frac{1.96^2 * 13.2^2}{0.02^2 * 72.8^2} = 315.7456$
The minimal sample size is 316 cars. Hence, the preliminary study was not sufficient for bounding the average speed.

(6) (9 points) Suppose that 75 percent of people use seatbelts regularly. Give the R-command for determining the probability that in 100 randomly chosen cars with the same number of passengers, in 70 or less of the cars the people do use the seatbelt. Give the R-command for the binomial and the normal distribution.

pbinom(70,100,0.75)= 0.1495
$\mu = n * p = 100 * 0.75 = 75;\ \sigma = \sqrt{n * p * q} = \sqrt{100 * 0.75 * 0.25} = 4.33,$
pnorm(70,75,4.33) = 0.1493

(7) (10 points) Pencils produced in firm A should have an average length of 17 cm. Assume the length of the pencils to be normally distributed with unknown variance σ^2. A sample of 5 pencils is taken to check whether the pencils have the required average length. The following lengths are measured for the 5 pencils: 19.2 cm, 17.4 cm, 18.5 cm, 16.5 cm, 18.9 cm. Formulate the null and the alternative hypothesis for this test problem. Does the average length of the pencils in the sample

significantly differ from the required mean of 17 cm at a 1%-level of significance (two-sided)? Formulate your conclusions in a final sentence.

t-test
mean=(19.2+17.4+18.5+16.5+18.9)/5=18.1
variance: ((19.2-18.1)^2+(17.4-18.1)^2+(18.5-18.1)^2+(16.5-18.1)^2+(18.9-18.1)^2)/(5-1)=1.265
x<-c(19.2,17.4,18.5,16.5,18.9)
var(x) #1.265
z=(18.1-17)/sqrt(1.265)*sqrt(5)= 2.186918
t0.005,df=4=4.604
abs(2.187)<4.604 – We do not reject Ho, there are not significant differences between the two means.

(8) (3 points)

a) Please explain briefly the difference between the covariance and the coefficient of correlation.
b) Give one example for a negative correlation.

Covariance: A measure of association between two random variables obtained as the expected value of the product of the two random variables around their means.

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n} = \frac{\sum_{i=1}^{n} x_i * y_i - \frac{\sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i}{n}}{n}$$

An important problem with the covariance is that it depends on the size/units of the variables

The **correlation coefficient** r considers this problem by dividing the covariance by the product of the standard deviations, with this the correlation coefficient compensates for the differences in dispersion for the two variables.

Correlation coefficient: A dimensionless measure of the interdependence between two variables, usually lying in the interval from -1 to +1, with zero indicating the absence of correlation (but not necessarily the independence of the two variables).

$$r = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}, \; r = \frac{cov(x,y)}{s_x * s_y}$$

(9) (6 points) Explain briefly how to read data into R and give two examples for R-commands that can be used for exploratory data analysis.

Read.table(file="",header=T)
str(), summary(), head(), dim(), table, hist, boxplot,...