Part A Statistics HT 2004 Exercises 1 Dr. Gesine Reinert

1. The following data are errors (measured in degrees towards a 90 degree angle) made by 36 children when trying to draw a 45 degree angle to a horizontal line. The range of the data is from -180 degrees to 180 degrees.

16	-34	2	18	0	25	-20	3	3	19	55	121
-14	-25	7	114	50	-9	12	10	-10	7	13	7
7	3	8	4	-15	-4	15	17	6	8	32	28

   (a) Make a histogram of the data. Would you conjecture the data to be normally distributed?

   (b) Make a boxplot of the data. Are there any outliers?

   (c) Calculate mean and standard deviation for the sample.

   (d) Would there be any indication of a tendency for the children to make a particular type of error?

2. When analysing poetry writing styles, one approach is to count the frequency distribution of the number of words required to reach the next word not already used before in the poem. For instance, in the sentence *There is gold in there but it is too deep, too deep to dig it out* the first four words all have a delay (waiting time) of 0 until a new word is reached, but there is a delay of at the second *there*, which is not new in the sentence. There is a similar delay of 1 at the second *is*. So the sequence of delays for this sentence is 0,0,0,0,1,0,1,0,2,0,0,1. For the following poem by Robert Frost (1874-1963), *The Road Not Taken*, count the frequencies of waiting times until new words, and plot the histogram. Compute the average waiting time, and use its reciprocal as an estimate of the parameter of an exponential distribution. Construct an exponential quantile-quantile plot (probability plot) using this parameter estimate.

   The Road Not Taken

   Two roads diverged in a yellow wood,
   And sorry I could not travel both
   And be one traveler, long I stood
   And looked down one as far as I could
   To where it bent in the undergrowth;

   Then took the other, as just as fair,
   And having perhaps the better claim,
   Because it was grassy and wanted wear;
   Though as for that the passing there
   Had worn them really about the same,

   And both that morning equally lay
   In leaves no step had trodden black.
   Oh, I kept the first for another day!
   Yet knowing how way leads on to way,
   I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I-
I took the one less traveled by,
And that has made all the difference.

3. (a) Suppose that $U_1, \ldots, U_n$ is a random sample from the uniform distribution on $(0,1)$. Let $U_{k:n}$ denote the kth order statistics in the sample; in other words, $U_{k:n}$ is the kth smallest value. Show that

$$EU_{k:n} = \frac{k}{n+1}$$

and

$$Var(U_{k:n}) = \frac{k}{(n+1)(n+2)}\left(1 - \frac{k}{n+1}\right).$$

(b) Find the approximate mean and variance of $X_{k:n}$, the kth order statistic of a sample of size n from a distribution with strictly increasing c.d.f. F. To do this, let $U_i = F(X_i)$ and use Taylor expansion for $X_{k:n} = F^{-1}(U_{k:n})$ around $\frac{k}{n+1}$ to argue that

$$EX_{k:n} \approx F^{-1}\left(\frac{k}{n+1}\right)$$

and

$$Var(X_{k:n}) \approx \frac{k}{(n+1)(n+2)}\left(1 - \frac{k}{n+1}\right)\frac{1}{(f\{F^{-1}[k/(n+1)]\})^2}.$$

(c) Show that the variance of the pth sample quantile is approximately

$$\frac{1}{nf^2(x_p)}p(1-p),$$

where x_p is the pth quantile.

(d) Find the approximate variance of the median of a sample of size n from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Compare to the variance of the mean.

4. Let $X_1, \ldots, X_n$ be a sample from a distribution F, and let F_n denote the e.c.d.f. Show that

$$Cov(F_n(u), F_n(v)) = \frac{1}{n}(F(m) - F(u)F(v)),$$

where $m = min(u,v)$. Conclude that $F_n(u)$ and $F_n(v)$ are positively correlated.