MATH274001

This question paper consists of 10 printed pages, each of which is identified by the reference **MATH274001**.

Statistical tables are attached.
A formulae sheet is attached.
Only approved basic scientific calculators may be used.

Examination for the Module MATH2740
(May-June 2004)

Environmental Statistics

Time allowed: **2 hours**

Do not answer more than **four** questions.
All questions carry equal marks.

1. (a) Explain, in terms of the mean number of points per unit area and the variance in the number of points per unit area, what is meant by: (i) a random point pattern; (ii) a clustered point pattern; (iii) a regular point pattern.

(b) The number of clusters, M, in a given area follows a Poisson distribution with probability generating function

$$G_M(s) = \exp[\lambda_1(s-1)], \quad (\lambda_1 > 0).$$

The number of points in the jth cluster is denoted as Y_j for $j = 1, 2, \ldots, M$, each following an independent logarithmic distribution with probability generating function

$$G_Y(s) = -c\ln(1-qs), \quad c = \frac{-1}{\ln(1-q)}, \quad (0 < q < 1).$$

The total number of points in the given area is calculated as $X = Y_1 + \ldots + Y_M$.

(i) Show that the probability generating function of X, $G_X(s)$, is given by:

$$G_X(s) = \left(\frac{1-q}{1-qs}\right)^{\lambda_1 c}.$$

(ii) Let $r = \lambda_1 c$ and $p = 1-q$. Use the probability generating function, $G_X(s)$, to show that:

$$\mathrm{P}(X=0) = p^r.$$

(c) The table below gives the distribution of the number of *Urtica Dioica* plants in 400 1 square metre quadrats:

Number of plants per quadrat	0	1	2	3	4	5+	Total
Number of quadrats	235	82	41	23	19	0	400

It has been suggested that the Negative binomial distribution with parameters r and p be used to model the data.

(i) Show that the method of moments estimates of r and p are $\widehat{r} = 1.1405$ and $\widehat{p} = 0.5962$ (correct to 4 decimal places).

(ii) Use the recursion:

$$\mathrm{P}(X = x) = \left(\frac{r + x - 1}{x}\right)(1 - p)\mathrm{P}(X = x - 1) \quad (x \geq 1),$$

to determine the expected frequencies under a Negative binomial model.

(iii) Perform a χ^2 goodness-of-fit test to assess the proposed model and comment on the result.

2. (a) Explain how departure from a spatially random pattern affects the distance from a randomly chosen plant to the nearest neighbouring plant assuming: (i) a clustered pattern; (ii) a regular pattern.

(b) Assume a Poisson forest of plants, with intensity λ plants per unit area. Let the random variable X denote the number of plants in a circle of area $A = \pi r^2$. Let the random variable R denote the distance from a randomly selected point to the nearest plant.

(i) Calculate $\mathrm{P}(X = 0)$. Hence or otherwise derive the cumulative distribution function of R, and show that the probability density function is given by:

$$\mathrm{f}(r) = 2\lambda\pi r \exp(-\lambda\pi r^2) \quad (r \geq 0).$$

(ii) Using a substitution $y = \lambda\pi r^2$, or otherwise, show that:

$$\mathrm{E}(R) = \frac{1}{2\sqrt{\lambda}} \quad \mathrm{Var}(R) = \frac{4 - \pi}{4\lambda\pi}.$$

You may quote, without proof, the results:

$$\begin{aligned}
\Gamma(k) &= \int_{z=0}^{\infty} z^{k-1}\exp(-z)\mathrm{d}z \quad k > 0, \\
\Gamma(k) &= (k-1)\Gamma(k-1) \quad k > 0, \\
\Gamma(k) &= (k-1)! \quad k = 1, 2, \ldots, \\
\Gamma(1/2) &= \sqrt{\pi}.
\end{aligned}$$

(c) Consider the following set of nearest neighbour distances, $\{r_i\}$, between the 8 *Digitalis Purpurea* plants in a 25 square metre study region.

$$\{r_i\} = \{1.87, 1.78, 0.64, 4.31, 1.67, 2.24, 0.64, 1.78\}$$

Use the Clark-Evans test to determine whether the locations of these *Digitalis Purpurea* plants can be considered spatially random.

3. (a) Explain what is meant by the terms: (i) positive spatial autocorrelation; (ii) negative spatial autocorrelation.

(b) (i) Justify the use of the formula:

$$BW = \frac{1}{2}\sum_i \sum_j \delta_{ij}(x_i - x_j)^2,$$

for the number of black/white (i.e. BW) joins in a black/white map of spatial pattern, where:

$$x_i = \begin{cases} 1 & \text{if cell } i \text{ is } B \\ 0 & \text{if cell } i \text{ is } W \end{cases}$$

and δ_{ij} denotes the contiguity matrix defined as:

$$\delta_{ij} = \begin{cases} 1 & \text{if cells } i \text{ and } j \text{ are joined} \\ 0 & \text{otherwise} \end{cases}$$

with $\delta_{ii} = 0$ for all i.

(ii) Let L denote the total number of joins in the map. Let p denote the probability that a cell in the map is labelled B (independent of all other cells). Show that, assuming free-sampling:

$$\mathrm{E}(BW) = 2Lpq, \quad (q = 1 - p).$$

(c) Consider the following data which refer to the number of *Bellis Perennis* plants in 9 contiguous quadrats over a 9 square metre study region.

19	16	10
8	20	17
11	16	20

In a previous study of *Bellis Perennis* plants, the probability that a 1 square metre quadrat contained more than 15 plants was found to be 0.55.

(i) By considering, for each quadrat, whether the number of *Bellis Perennis* plants is greater than 15, or at most 15, construct a black/white map of the data.

(ii) Use the information from the previous study to calculate the test statistic Z_{BW}, and hence assess the evidence for the presence of positive spatial autocorrelation in the data. Comment on the validity of the assumptions made in performing the test.

4. (a) Explain the advantage of the likelihood-ratio statistic, Y^2, over the chi-square statistic, X^2, when testing the addition of a parameter to a log-linear model.

(b) Consider the saturated model for a 2×2 contingency table with observed frequencies $\{f_{ij}\}$:

$$y_{ij} = \ln(f_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \quad (i = 1, 2), \quad (j = 1, 2),$$

subject to centered constraints:

$$\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0.$$

 QUESTION 4 CONTINUED...

Show that:

$$\mu = \frac{y_{11} + y_{12} + y_{21} + y_{22}}{4}, \quad \lambda_1^A = \frac{y_{11} + y_{12} - y_{21} - y_{22}}{4},$$
$$\lambda_1^B = \frac{y_{11} - y_{12} + y_{21} - y_{22}}{4}, \quad \lambda_{11}^{AB} = \frac{y_{11} - y_{12} - y_{21} + y_{22}}{4}.$$

(c) The results of a survey into the accuracy of rainfall forecasts are summarised in the following table. The variables are A (1=no rain forecast, 2=rain forecast) and B (1=no rain observed, 2=rain observed).

	A_1	A_2
B_1	41	34
B_2	22	44

(i) Fit the saturated model to these data, obtaining parameter estimates $\widehat{\mu}$, $\widehat{\lambda}_1^A$, $\widehat{\lambda}_1^B$ and $\widehat{\lambda}_{11}^{AB}$.

(ii) Interpret the sign of the parameter estimate $\widehat{\lambda}_{11}^{AB}$, with reference to the accuracy of rainfall forecasts.

(iii) Using the likelihood ratio statistic, Y^2, test whether the independence model, A/B, is a good fit to the data. Interpret the result with reference to the accuracy of rainfall forecasts.

5. A study was conducted on the association between age, schooling and attitude towards recycling. Data from the study were analysed using R. The variables were `age` (1=15-39 years, 2=40+ years), `school` (1=elementary only, 2=beyond elementary) and `recycle` (1=recycling important, 2=recycling not important).

The aim of the analysis was to build a log-linear model which explained the significant factors and interactions present in the data. Consider the following extract from the R session:

```
>  options(contrasts=c("contr.sum","contr.poly"))
>  saturated.model <- glm(count~age*school*recycle,family=poisson)
>  summary(saturated.model)

         Coefficients:
                          Estimate  Std.  Error  z value  Pr(>|z|)
             (Intercept)  4.942872     0.032688  151.212   < 2e-16  ***
                    age1  0.005773     0.032688    0.177     0.860
                 school1  0.294629     0.032688    9.013   < 2e-16  ***
                recycle1 -0.480302     0.032688  -14.693   < 2e-16  ***
            age1:school1 -0.271343     0.032688   -8.301   < 2e-16  ***
           age1:recycle1 -0.036677     0.032688   -1.122     0.262
        school1:recycle1 -0.161926     0.032688   -4.954  7.28e-07  ***
   age1:school1:recycle1  0.037707     0.032688    1.154     0.249
                     ---
   Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1

>  forward.start <- glm(count~age+school+recycle,family=poisson)
>  try.models <- list(upper=~age*school*recycle,lower=~age+school+recycle)
>  step(forward.start,scope=try.models,direction="forward")

   Step:  AIC= 89.23
   count ~ age + school + recycle + age:school
                               Df     Deviance     AIC
        + school:recycle        1        2.113  68.370
                  <none>                24.975  89.232
           + age:recycle        1       24.808  91.065

   Step:  AIC= 68.37
   count ~ age + school + recycle + age:school + school:recycle
                               Df     Deviance     AIC
                  <none>                 2.113  68.370
           + age:recycle        1        1.325  69.582
```

(a) Describe the purpose of the `options()` command used above.

(b) Consider the output of the `summary()` command. Using A to denote the variable `age`, B to denote the variable `school`, and C to denote the variable `recycle`, list the factors and interactions in the saturated model which are significant at the 5% level. Interpret any significant interactions with reference to the association between age, schooling and attitude towards recycling.

(c) Explain in detail the purpose of the `step()` command used above.

(d) Using hierarchy notation, write down the model selected by the `step()` command. Interpret the model with reference to the association between age, schooling and attitude towards recycling.

Formulae Sheet

Poisson

$$\mathrm{P}(X=x)=\frac{\exp(-\lambda)\lambda^x}{x!},\quad (x=0,1,\ldots).$$
$$\mathrm{E}(X)=\mathrm{Var}(X)=\lambda,\quad G_X(s)=\exp[\lambda(s-1)].$$

Logarithmic

$$\mathrm{P}(X=x)=c\frac{q^x}{x},\quad (x=1,2,\ldots),\quad (0<q<1),\quad c=\frac{-1}{\ln(1-q)}.$$
$$\mathrm{E}(X)=\frac{cq}{(1-q)},\quad G_X(s)=-c\ln(1-qs).$$

Neyman Type A

$$\mathrm{E}(X)=\lambda_1\lambda_2,\quad \mathrm{Var}(X)=\lambda_1\lambda_2(1+\lambda_2),\quad G_X(s)=\exp\left[\lambda_1\{\exp[\lambda_2(s-1)]-1\}\right].$$

Negative Binomial

$$\mathrm{P}(X=x)=\binom{r+x-1}{x}p^r(1-p)^x,\quad (x=0,1,\ldots),\quad (0<p<1).$$
$$\mathrm{E}(X)=\frac{r(1-p)}{p},\quad \mathrm{Var}(X)=\frac{r(1-p)}{p^2},\quad G_X(s)=\left(\frac{p}{1-qs}\right)^r,\quad (q=1-p).$$

Gamma

$$\mathrm{f}(x)=\frac{\alpha^r}{\Gamma(r)}x^{r-1}\exp(-\alpha x),\quad (x\ge 0),\quad (r,\alpha>0).$$
$$\mathrm{E}(X)=\frac{r}{\alpha},\quad \mathrm{Var}(X)=\frac{r}{\alpha^2}.$$

Weibull

$$\mathrm{f}(r)=2\lambda\pi r\exp(-\lambda\pi r^2),\quad (r\ge 0),\quad (\lambda>0).$$
$$\mathrm{E}(R)=\frac{1}{2\sqrt{\lambda}},\quad \mathrm{Var}(R)=\frac{4-\pi}{4\lambda\pi}.$$

Testing for spatial randomness

$$Z=\frac{(n-1)s^2}{\bar{x}}\sim\chi^2_{n-1}.$$
$$\mathrm{E}(\bar{r})=\mathrm{E}(R)=\frac{1}{2\sqrt{\lambda}}.$$
$$\mathrm{Var}(\bar{r})=\frac{\mathrm{Var}(R)}{m}=\frac{4-\pi}{4\lambda\pi m}.$$
$$S=2m\widehat{\lambda}\pi\bar{u}\sim\chi^2_{2m}.$$
$$H=\frac{\sum r_{1i}^2}{\sum r_{2i}^2}\sim\mathrm{F}_{2m,2m}.$$

CONTINUED...

Spatial autocorrelation

$$L = \frac{1}{2}\sum_i L_i \text{ where } L_i \text{ is the number of cells joined to cell } i, \quad K = \frac{1}{2}\sum_i L_i(L_i - 1).$$

Free-sampling

$$\mathrm{E}(BB) = Lp^2, \quad \mathrm{E}(BW) = 2Lpq, \quad \mathrm{E}(WW) = Lq^2$$

$$\begin{aligned}
\mathrm{Var}(BB) &= Lp^2 + 2Kp^3 - (L+2K)p^4. \\
\mathrm{Var}(BW) &= 2(L+K)pq - 4(L+2K)p^2q^2. \\
\mathrm{Var}(WW) &= Lq^2 + 2Kq^3 - (L+2K)q^4.
\end{aligned}$$

Non-free sampling

$$\begin{aligned}
\mathrm{E}(BB) &= L\frac{n_1(n_1-1)}{n(n-1)}. \\
\mathrm{E}(BW) &= 2L\frac{n_1n_2}{n(n-1)}. \\
\mathrm{E}(WW) &= L\frac{n_2(n_2-1)}{n(n-1)}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(BB) &= L\frac{n_1(n_1-1)}{n(n-1)} + 2K\frac{n_1(n_1-1)(n_1-2)}{n(n-1)(n-2)} \\
&+ [L(L-1)-2K]\,\frac{n_1(n_1-1)(n_1-2)(n_1-3)}{n(n-1)(n-2)(n-3)} - \left[L\frac{n_1(n_1-1)}{n(n-1)}\right]^2.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(BW) &= \frac{2(L+K)n_1n_2}{n(n-1)} + 4\,[L(L-1)-2K]\,\frac{n_1(n_1-1)n_2(n_2-1)}{n(n-1)(n-2)(n-3)} \\
&- 4\left[\frac{Ln_1n_2}{n(n-1)}\right]^2.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(WW) &= L\frac{n_2(n_2-1)}{n(n-1)} + 2K\frac{n_2(n_2-1)(n_2-2)}{n(n-1)(n-2)} \\
&+ [L(L-1)-2K]\,\frac{n_2(n_2-1)(n_2-2)(n_2-3)}{n(n-1)(n-2)(n-3)} - \left[L\frac{n_2(n_2-1)}{n(n-1)}\right]^2.
\end{aligned}$$

Categorical data

For a 2×2 contingency table: $Y^2 = 2\sum_{ij} f_{ij}\ln(f_{ij}/e_{ij})$.

For the A/B model: $e_{ij} = \dfrac{f_{i0}f_{0j}}{f_{00}}$.

For the A model: $e_{ij} = \dfrac{f_{i0}}{2}$.

For the B model: $e_{ij} = \dfrac{f_{0j}}{2}$.

CONTINUED...

Normal Distribution Function Tables

The first table gives

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$$

and this corresponds to the shaded area in the figure to the right. $\Phi(x)$ is the probability that a random variable, normally distributed with zero mean amd unit variance, will be less than or equal to x. When $x < 0$ use $\Phi(x) = 1 - \Phi(-x)$, as the normal distribution with mean zero is symmetric about zero.

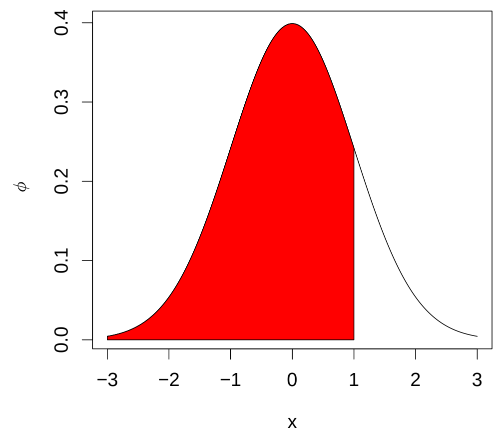


Table 1

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.5000	**0.50**	0.6915	**1.00**	0.8413	**1.50**	0.9332	**2.00**	0.9772	**2.50**	0.9938
0.05	0.5199	**0.55**	0.7088	**1.05**	0.8531	**1.55**	0.9394	**2.05**	0.9798	**2.55**	0.9946
0.10	0.5398	**0.60**	0.7257	**1.10**	0.8643	**1.60**	0.9452	**2.10**	0.9821	**2.60**	0.9953
0.15	0.5596	**0.65**	0.7422	**1.15**	0.8749	**1.65**	0.9505	**2.15**	0.9842	**2.65**	0.9960
0.20	0.5793	**0.70**	0.7580	**1.20**	0.8849	**1.70**	0.9554	**2.20**	0.9861	**2.70**	0.9965
0.25	0.5987	**0.75**	0.7734	**1.25**	0.8944	**1.75**	0.9599	**2.25**	0.9878	**2.75**	0.9970
0.30	0.6179	**0.80**	0.7881	**1.30**	0.9032	**1.80**	0.9641	**2.30**	0.9893	**2.80**	0.9974
0.35	0.6368	**0.85**	0.8023	**1.35**	0.9115	**1.85**	0.9678	**2.35**	0.9906	**2.85**	0.9978
0.40	0.6554	**0.90**	0.8159	**1.40**	0.9192	**1.90**	0.9713	**2.40**	0.9918	**2.90**	0.9981
0.45	0.6736	**0.95**	0.8289	**1.45**	0.9265	**1.95**	0.9744	**2.45**	0.9929	**2.95**	0.9984
0.50	0.6915	**1.00**	0.8413	**1.50**	0.9332	**2.00**	0.9772	**2.50**	0.9938	**3.00**	0.9987

The inverse function $\Phi^{-1}(p)$ is tabulated below for various values of p.

Table 2

p	**0.900**	**0.950**	**0.975**	**0.990**	**0.995**	**0.999**	**0.9995**
$\Phi^{-1}(p)$	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

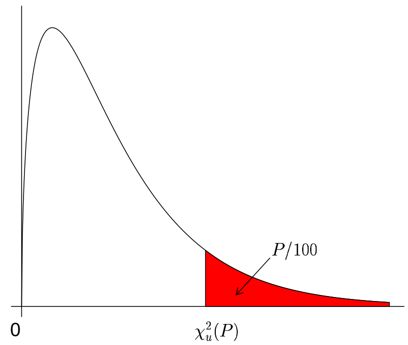
CONTINUED...

Percentage Points of the χ^2-Distribution

This table gives the percentage points $\chi^2_\nu(P)$ for various values of P and degrees of freedom ν, as indicated by the figure to the right.

If X is a variable distributed as χ^2 with ν degrees of freedom, $P/100$ is the probability that $X \geq \chi^2_\nu(P)$.

For $\nu > 100$, $\sqrt{2X}$ is approximately normally distributed with mean $\sqrt{2\nu - 1}$ and unit variance.



	Percentage points P					
ν	**99**	**97.5**	**95**	**5**	**2.5**	**1**
1	0.000	0.001	0.004	3.841	5.024	6.635
2	0.020	0.051	0.103	5.992	7.378	9.210
3	0.115	0.216	0.352	7.815	9.348	11.345
4	0.297	0.484	0.711	9.488	11.143	13.277
5	0.554	0.831	1.145	11.070	12.833	15.086
6	0.872	1.237	1.635	12.592	14.449	16.812
7	1.239	1.690	2.167	14.067	16.013	18.475
8	1.646	2.180	2.733	15.507	17.535	20.090
9	2.088	2.700	3.325	16.919	19.023	21.666
10	2.558	3.247	3.940	18.307	20.483	23.209
11	3.053	3.816	4.575	19.675	21.920	24.725
12	3.571	4.404	5.226	21.026	23.337	26.217
13	4.107	5.009	5.892	22.362	24.736	27.688
14	4.660	5.629	6.571	23.685	26.119	29.141
15	5.229	6.262	7.261	24.996	27.488	30.578
16	5.812	6.908	7.962	26.296	28.845	32.000
17	6.408	7.564	8.672	27.587	30.191	33.409
18	7.015	8.231	9.390	28.869	31.526	34.805
19	7.633	8.907	10.117	30.144	32.852	36.191
20	8.260	9.591	10.851	31.410	34.170	37.566
25	11.524	13.120	14.611	37.652	40.646	44.314
30	14.953	16.791	18.493	43.773	46.979	50.892
40	22.164	24.433	26.509	55.758	59.342	63.691
50	29.707	32.357	34.764	67.505	71.420	76.154
80	53.540	57.153	60.391	101.879	106.629	112.329

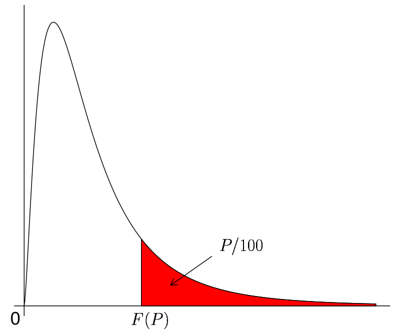
CONTINUED...

2.5 Percent Points of the F-Distribution

This table gives the percentage points $F_{\nu_1,\nu_2}(P)$ for $P = 0.025$ and degrees of freedom ν_1, ν_2, as indicated by the figure to the right.

The lower percentage points, that is the values $F'_{\nu_1,\nu_2}(P)$ such that the probability that $F \leq F'_{\nu_1,\nu_2}(P)$ is equal to $P/100$, may be found using the formula

$$F'_{\nu_1,\nu_2}(P) = 1/F_{\nu_1,\nu_2}(P)$$



	ν_1								
ν_2	**1**	**2**	**3**	**4**	**5**	**6**	**12**	**24**	**∞**
2	38.506	39.000	39.165	39.248	39.298	39.331	39.415	39.456	39.498
3	17.443	16.044	15.439	15.101	14.885	14.735	14.337	14.124	13.902
4	12.218	10.649	9.979	9.605	9.364	9.197	8.751	8.511	8.257
5	10.007	8.434	7.764	7.388	7.146	6.978	6.525	6.278	6.015
6	8.813	7.260	6.599	6.227	5.988	5.820	5.366	5.117	4.849
7	8.073	6.542	5.890	5.523	5.285	5.119	4.666	4.415	4.142
8	7.571	6.059	5.416	5.053	4.817	4.652	4.200	3.947	3.670
9	7.209	5.715	5.078	4.718	4.484	4.320	3.868	3.614	3.333
10	6.937	5.456	4.826	4.468	4.236	4.072	3.621	3.365	3.080
11	6.724	5.256	4.630	4.275	4.044	3.881	3.430	3.173	2.883
12	6.554	5.096	4.474	4.121	3.891	3.728	3.277	3.019	2.725
13	6.414	4.965	4.347	3.996	3.767	3.604	3.153	2.893	2.595
14	6.298	4.857	4.242	3.892	3.663	3.501	3.050	2.789	2.487
15	6.200	4.765	4.153	3.804	3.576	3.415	2.963	2.701	2.395
16	6.115	4.687	4.077	3.729	3.502	3.341	2.889	2.625	2.316
17	6.042	4.619	4.011	3.665	3.438	3.277	2.825	2.560	2.247
18	5.978	4.560	3.954	3.608	3.382	3.221	2.769	2.503	2.187
19	5.922	4.508	3.903	3.559	3.333	3.172	2.720	2.452	2.133
20	5.871	4.461	3.859	3.515	3.289	3.128	2.676	2.408	2.085
25	5.686	4.291	3.694	3.353	3.129	2.969	2.515	2.242	1.906
30	5.568	4.182	3.589	3.250	3.026	2.867	2.412	2.136	1.787
40	5.424	4.051	3.463	3.126	2.904	2.744	2.288	2.007	1.637
50	5.340	3.975	3.390	3.054	2.833	2.674	2.216	1.931	1.545
100	5.179	3.828	3.250	2.917	2.696	2.537	2.077	1.784	1.347
∞	5.024	3.689	3.116	2.786	2.567	2.408	1.945	1.640	1.003

END