UNIVERSITY OF TWENTE.

Course	: Statistical Techniques for TCS/BIT – Exam	Date	: Thursday 19 December, 2024
Module	: 06	Time	: 13:45 – 16:00 (2 hours and 15 minutes)
Course code	: 202001033	Reference	: Intelligent Interaction Design TCS/BIT (2024-1B)

D

Statistical Techniques for TCS/BIT – Exam

Instructions

This exam contains 11 exercises. Only use the provided *answer form* to submit your answers.

- For PART 1, exercises 1–7, you are only required to fill in the **final answer** on the answer form.
- For PART 2, exercises 8–11, you are required to write down a **full calculation and argumentation**.

Only hand in your answer form. **Any text outside the answer form will not be considered.**
If you run out of space, use the extra space at the end of the form and reference it in your original answer.
Do not use red pen or pencil. Do not use correction fluid or tape
Electronic devices are prohibited. Only simple calculators (non-graphing) are allowed.

PART 1: Final answer questions

Report all your answers on the answer form.

1. Answer the following statements **(A)** to **(I)** with either true (T) or false (F). [3 pt]

(**A**) The range is calculated by subtracting the smallest value in the data from the largest one.

(**B**) A histogram is used to display the relationship between two variables.

(**C**) If a data point has a z-score of -2, it is 2 standard deviations below the mean.

(**D**) A bar chart is suitable for displaying the frequency distribution of categorical data.

(**E**) In a box plot, the interquartile range (IQR) represents the middle 75% of the data.

(**F**) The median is less sensitive to outliers compared to the mean.

(**G**) The mean of a dataset is always greater than or equal to the median.

(**H**) If the standard deviation of a dataset is zero, all the values in the dataset are identical.

(**I**) Skewness measures the "tailedness" of a distribution, while kurtosis measures the symmetry of a distribution.

2. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2. [3 pt]
Two point estimators of μ are proposed:

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{and} \quad T_2 = \frac{1}{n+1}\sum_{i=1}^{n} X_i.$$

I. Which of the following is true?

(**A**) Both T_1 and T_2 are unbiased estimators of μ.

(**B**) T_2 is a biased and T_1 is an unbiased estimator of μ.

(**C**) T_1 is a biased and T_2 is an unbiased estimator of μ.

(**D**) We cannot tell about the bias of T_1 and T_2 with the given information.

(**E**) Both T_1 and T_2 are biased estimators of μ.

II. Which of the following is true?

(A) $Var(T_1) < Var(T_2)$ **(B)** $Var(T_1) > Var(T_2)$ **(C)** $Var(T_1) = Var(T_2)$
(D) We cannot tell the relationship between $Var(T_1)$ and $Var(T_2)$ with the given information.

III. Which of the following is true regarding the Mean Squared Error (MSE)?

(A) T_2 is a better estimator of μ than T_1.

(B) We cannot tell whether T_1 or T_2 is a better estimator of μ with the given information.

(C) T_1 is a better estimator of μ than T_2.

3. In a baseline study, the average error rate for completing tasks with an old interface was 10%. A new interface is developed, and an experiment is conducted where 50 users achieve an average error rate of 8% with a population standard deviation of $\sigma = 2\%$. [3 pt]

I. Suppose that the hypotheses of interest are: $H_0 : \mu = 10\%$, $H_1 : \mu \neq 10\%$. If the test rejects if either $\bar{X} > 10.233$ or $\bar{X} < 9.767$, which of the following statements about the level α of the test is true?

(A) $\alpha = P\left(|\bar{X} - 10| > 0.465 \,|\, H_0\right) = 0.1$
(B) $\alpha = P\left(|\bar{X} - 10| < 0.465 \,|\, H_0\right) = 0.1$
(C) $\alpha = P\left(|\bar{X} - 10| < 0.465 \,|\, H_0\right) = 0.05$
(D) $\alpha = P\left(|\bar{X} - 10| > 0.465 \,|\, H_0\right) = 0.05$

II. Suppose that the hypotheses of interest are: $H_0 : \mu = 10\%$, $H_1 : \mu \neq 10\%$. About the p-value, which of the following statements is true?

(A) The p-value is $2 \times P\left(Z > \frac{8-10}{2/\sqrt{50}}\right)$
(B) The p-value is $P\left(Z > \frac{8-10}{2/\sqrt{50}}\right) + P\left(Z < \frac{8-10}{2/\sqrt{50}}\right)$
(C) The p-value is $P\left(Z < \frac{8-10}{2/\sqrt{50}}\right)$
(D) None of them
(E) The p-value is $P\left(Z > \frac{8-10}{2/\sqrt{50}}\right)$
(F) The p-value is $2 \times P\left(Z < \frac{8-10}{2/\sqrt{50}}\right)$

4. A study evaluates the typing speeds (words per minute) of 15 individuals using a new ergonomic keyboard. The sample variance of typing speeds is found to be $s^2 = 25\,\text{words}^2/\text{minute}^2$. Assume that typing speeds are normally distributed. [3 pt]

I. Let $\chi^2_{n-1,\alpha}$ denote the α-quantile of the χ^2_{n-1}-distribution, i.e., $\chi^2_{n-1,\alpha} = P(\chi^2_{n-1} > \alpha)$. Which of the following statements is true?

(A) The formula for a $(1-\alpha)$-confidence interval for σ^2 is $\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha}}\right)$.

(B) The formula for a $(1-\alpha)$-confidence interval for σ^2 is $\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right)$.

(C) The formula for a $(1-\alpha)$-confidence interval for σ^2 is $\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}}\right)$.

(D) The formula for a $(1-\alpha)$-confidence interval for σ^2 is $\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right)$.

II. Based on the confidence interval for σ^2, denoted by (L, U), we aim to decide whether the new keyboard leads to highly variable typing speeds compared to a known acceptable variability of $\sigma^2 = 20\,\text{words}^2/\text{minute}^2$. Which of the following statements is true?

(A) If $20 > U$, at a confidence level of $(1-\alpha)100\%$, the variability is too high.

(B) If $20 \in (L, U)$, at a confidence level of $(1-\alpha)100\%$, the variability is too high.

(C) If $20 < L$, at a confidence level of $(1-\alpha)100\%$, the variability is too high.

(D) We cannot say anything about this question based on the confidence interval.

5. A researcher is studying the amount of snowfall (in cm) before and after a winter storm in a specific region. The data below shows the snowfall amounts measured at various locations before and after the storm: [2 pt]

Location	1	2	3	4	5	6	7
Snowfall Before Storm (X)	10	8	6	5	7	9	8
Snowfall After Storm (Y)	12	10	9	7	11	13	10

I. You want to conduct a test on the researcher's claim that the snowfall increases after a winter storm under the assumption that the amounts of snowfall follows a normal distribution with unknown mean. Which of the following represents the suitable test for the hypothesis testing mentioned?

(A) Independent samples t-test

(B) None of them

(C) Chi-Squared test on independence

(D) Test on the difference of two population proportions

(E) Shapiro Wilk's test

(F) Paired samples t-test

II. Which of the following represents the distribution for the proper test statistic?

(A) t_{14} Under H_0 **(B)** t_{12} Under H_1 **(C)** t_{12} Under H_0 **(D)** t_6 Under H_0 **(E)** t_7 Under H_0

6. [2 pt]

I. Which of the following is **NOT** an assumption of simple linear regression?

(A) The independent variable x is normally distributed.

(B) The relationship between x and y is linear.

(C) The errors (residuals) are normally distributed.

(D) The variance of the errors (residuals) is constant (homoscedasticity).

II. In a simple linear regression modelWhat does the slope β_1 represent?

(A) The value of y that minimizes the residual sum of squares.

(B) The predicted value of y when $x = 0$.

(C) The strength of the linear relationship between x and y.

(D) The change in y for a one-unit increase in x.

7. [2 pt]

I. What is the primary purpose of the Shapiro-Wilk test?

(A) To test if two samples have equal means.

(B) To test if the variance of a sample equals a specified value.

(C) To test if a sample comes from a normally distributed population.

(D) To test for independence between two variables.

II. If the p-value of a Shapiro-Wilk test is 0.03, what conclusion can be drawn at a significance level of $\alpha = 0.05$?

(A) The sample data does not come from a normal distribution.

(B) The test is inconclusive because $p > \alpha$.

(C) The sample size is too small to make a conclusion.

(D) The sample data is consistent with a normal distribution.

Continues on the following page.

PART 2: Open questions

The full solutions to exercises 8—11 must be clearly written down on the answer form, including calculations and argumentations.

Points will not be awarded for reaching a correct result if this is not supported by a correct procedure and by a sound and clear argumentation.

8. The following table shows the number of coding errors detected during debugging sessions by a group of developers:

Developer	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
#Errors	5	6	7	8	9	9	9	9	10	11	12	12	13	13	15	16	17	18	21	22

Tab-10 In the table you find for $n = 3, 4, \ldots, 50$ and specific value of α the critical value c such that $P(W \leq c \mid \text{normal distribution}) = \alpha$.
Example
$n = 30$, $\alpha = 0.05$: $P(W \leq 0.927 \mid \text{normal distribution}) = 0.05$.

$n \backslash \alpha$	0.01	0.02	0.05	0.10	0.50	0.90	0.95	0.98	0.99
20	.868	.884	.905	.920	.959	.979	.983	.986	.988

Statistical Measure	Value
Sample size	20
Sample mean	12.1
Sample standard deviation	4.79

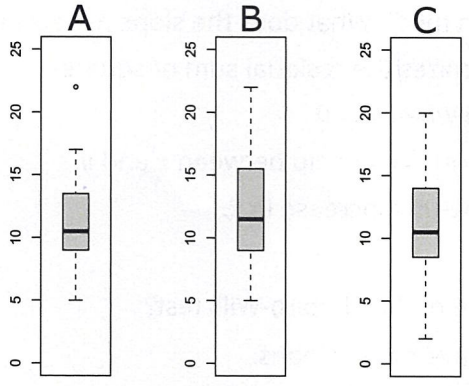
I. We now perform a first analysis of the data. Motivate all answers with statistical arguments. [4 pt]

(a) Determine median and range of these data.

(b) Determine the z-scores for the minimal and maximal number of errors. You may use the information in the above table.

(c) Decide whether the maximum is an outlier according to the empirical rule.

(d) In the figure below you find three box plots (A, B, and C). Decide whether either of them represents the data given above.



II. To evaluate the applicability of the normal distribution to the observed numbers of errors, a Shapiro Wilk test has been employed with $W = 0.94859$ [2 pt]

(a) On the left of the above figure you find an excerpt of the Shapiro Wilk table from the reader. Based on the table, decide whether the p-value is $0.0317, 0.3462$, or 0.92. Show all steps to your solution for this decision.

(b) Based on part (a), decide if normality is a reasonable assumption.

III. Determine a 99%-confidence interval for the expected number of errors, assuming a normal model. Present all relevant calculations. You find all relevant table excerpts on the last page. [3 pt]

9. The elves are comparing two different teams (Team Red and Team Green) for consistency in decorating Christmas cookies. The number of cookies decorated by each team in 10-minute intervals is recorded as follows: [7 pt]

Team Red: $20, 22, 19, 21, 20, 22, 19, 20$; $s_{\text{Red}} = 1.188$, $\bar{x}_{\text{Red}} = 20.375$.

Team Green: $18, 17, 19, 20, 18, 17, 19, 18$; $s_{\text{Green}} = 1.035$, $\bar{x}_{\text{Green}} = 18.250$.

I. State the null and alternative hypotheses for testing if the variances of cookie decoration speeds are the same for the two teams.

II. Calculate the sample variances for Team Red and Team Green.

III. Assess if the variances are significantly different at the $\alpha = 0.01$ significance level using an appropriate statistical test. You may assume normality of the data. Follow the eight step protocol.

IV. Explain the results to Santa in terms of the elves' consistency.

10. A robotics company has developed a new type of robot and tested its performance in three different conditions: *Factory*, *Laboratory*, and *Outdoor Field*. The following table shows the results: [5 pt]

	Success	Failure
Factory	55	28
Laboratory	43	16
Outdoor Field	64	12

I. The company wants to know if the testing condition has any effect on the robot's performance. Which test should we use in this case: a test on independence or on homogeneity? Why?

II. Conduct the hypothesis test above at significance level 10%. Follow the 8 step protocol.

11. Santa has noticed that the number of gifts he delivers depends on the hours he spends preparing with the elves in his workshop. He collects data for 10 days: [7 pt]

Time (X)	2	3	5	4	6	7	8	3	5	4
Gifts (Y)	50	70	120	100	150	180	200	80	130	110

	Value
$\sum_{i=1}^{n} x_i$	47
$\sum_{i=1}^{n} y_i$	1190
$\sum_{i=1}^{n} (x_i - \bar{x})^2$	32.1
$\sum_{i=1}^{n} (y_i - \bar{y})^2$	20490.0
$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$	807.0

I. Use simple linear regression to determine the relationship between the hours of preparation (X) and the number of gifts delivered (Y).

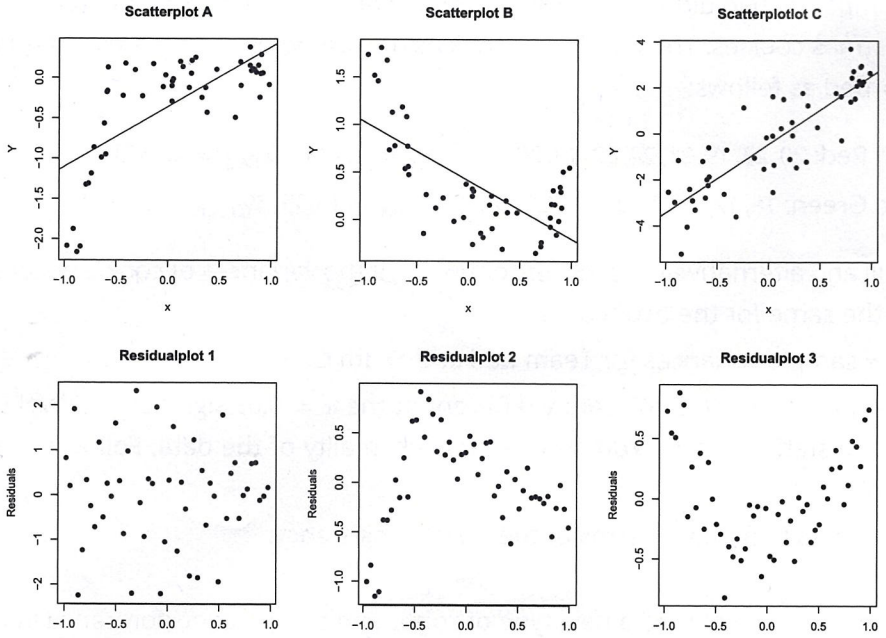
(a) Identify the dependent and independent variables.

(b) Write down the linear regression model.

(c) Calculate the regression coefficients (slope and intercept).

(d) Predict the number of gifts Santa will deliver if he spends 9 hours preparing with the elves.

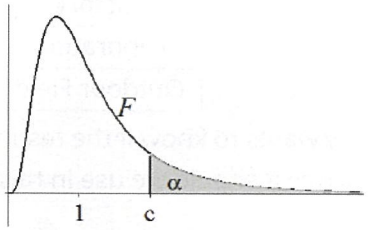
II. The graph below shows three scatterplots of data sets together with the fitted regression lines (A, B, and C; top row) and three residual plots (1, 2, and 3; bottom row). Which residual plot corresponds to which data set? Explain.



Tab 4

Table for the F-distribution, $\alpha = 0.05$

In the table you will find critical values c such that $P(F \geq c) = \alpha$

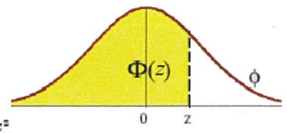


f_2 \ f_1	Number of degrees of freedom in the numerator										
	1	2	3	4	5	6	7	8	9	10	11
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26

degrees of freedom in the denominator

Tab-1

Standard normal probabilities



The table gives the distribution function Φ for a N(0,1)-variable Z

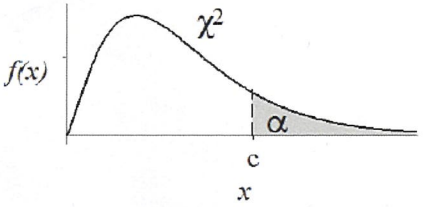
$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}}\, dx$$

Last column: N(0,1)-density function (z in 1 dec.): $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

	Second decimal of z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	φ(z)
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	0.1109
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	0.0940
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	0.0790
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	0.0656
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	0.0540
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	0.0440
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	0.0355
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	0.0283
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	0.0224
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	0.0175

Tab-3

Table Chi-square distribution



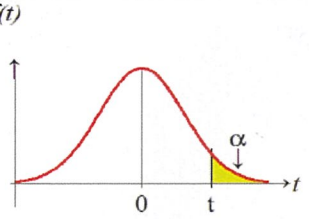
In the table you will find critical values c for the upper-tailed probabilities

$$P(\chi^2 \ge c) = \alpha$$

df = number of degrees of freedom

	α											
df	0.995	0.990	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.10	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	18.25	22.31	25.00	27.49	30.58	32.80

Table t-distribution



In the table you find the critical values t for the upper-tailed probabilities such that

$$P(T \ge t) = \alpha$$

Number of degrees of freedom	α							
	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
35	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	0.678	1.294	1.667	1.994	2.381	2.648	3.211	3.435

Total: 46 pt