STAT 212-501 (Fall 2022) Name ______________________
Instructor: Dr. Sharmistha Guha

Midterm I - Practice Exam

Instructions (please read the following carefully before starting the exam):

- **Do not open** the exam unless you are explicitly instructed to do so.
- You may use **one** formula sheet (with writing on front and back), a standard scientific calculator (with **no** internet access), and pencil(s) - all fully functioning (no sharing or borrowing is allowed). **No** cell phones or any other electronic devices are allowed.
- Copies of the Normal and t distribution tables are provided at the back of the exam in that order: Normal table (1 page) and t table (1 page). (These are Tables A.3 and A.4, respectively, from pages 495–496 of your textbook.)
- There is **NO partial credit. Only ONE answer is correct** for each of the questions. If you mark two or more options in the Scantron for any question, you will get 0 points for that question *irrespective* of whether one of those marked options is correct or not.
- **Mark your answers clearly** (and fully) with a number 2 pencil on your Scantron.
- **Mark which form** you have, **A or B**, on your Scantron.
- Also, **clearly bubble in the following on your Scantron:** *department* (**STAT**), the *course number* (**212**), the *section number* (**501**), your **name and UIN** on the Scantron.
- You may write on your exam (and use the blank side of each page for scratch work). But none of this work will count. Your score will be based solely on the Scantron.
- No cheating or discussing or helping others in any way will be tolerated. If detected, everyone involved will get a grade of 0 on the exam. You **must** work alone.
- **Turn in both your exam and Scantron** when you are done. Bring your **TAMU ID**.
- Good luck!

Note: *Throughout this exam, unless specifically stated otherwise, all notations/symbols and abbreviations used have their usual standard meanings, as used/defined in the lecture notes.*

Read the following information to answer **Questions 1–4**:

Polynomial models of degree 1 to 6 were fitted to a set of $n = 100$ observations of (x, y) pairs. Fitted models and a scatterplot of the data are shown on **pg. 8** and information obtained from `R` is given below. Use this information and the plots to answer **Questions 1–4** below.

Polynomial degree	SSR	AIC	BIC
1	149.85	478.36	486.18
2	156.43	479.36	489.78
3	401.20	434.34	447.36
4	401.64	436.23	451.86
5	418.89	433.90	452.14
6	418.92	435.90	456.74

$$\text{SST} = 808.9614$$

1. The value of adjusted R^2 for the fourth degree polynomial model is closest to:

[✓ (a)] 0.4753.

(b) 0.7572.

(c) 0.4826.

(d) 436.23.

(e) 0.3165.

The value of R_4^2 is 401.64/808.9614 = 0.4965. Using the formula on p. 60 of the notes,

$$R^2_{\text{adj},4} = \frac{99(0.4965) - 4}{99 - 4} = 0.4753.$$

2. Suppose you were presented with the plots of the third and fifth degree polynomials (as in **pg. 8**) and were asked to choose between the two, *strictly* on the basis of the plots. Then:

(a) The fifth degree polynomial is preferable because the two curves look almost the same, but the fifth degree model is more complex.

(b) The third degree polynomial is preferable because simpler models are always preferable.

[✓ (c)] The third degree polynomial is preferable because the two curves look almost the same, but the third degree model is simpler.

(d) The fifth degree model is preferable because more complex models are always preferable.

(e) Not even William of Occam could decide between the two.

When two fitted models (and the fitted curves) are virtually the same, but one is simpler, then the simpler model is always the better choice following the principle of Occam's Razor.

3. Based on the plots and all the other given information, the best model choice would be:

(a) Either the polynomial of degree 1 or degree 2.

[✓ (b)] The polynomial of degree 3.

(c) The polynomial of degree 4.

(d) The polynomial of degree 5.

(e) The polynomial of degree 6.

AIC and BIC prefer the fifth and third degree models, respectively. However, since the plots of those two models are very similar, it makes sense to choose the simpler model following the principle of Occam's Razor.

4. The estimate of the error standard deviation, σ, based on the third degree model is:

(a) $\sqrt{401.20/96}$.

(b) $\sqrt{808.9614/96}$.

[✓ (c)] $\sqrt{(808.9614 - 401.20)/96}$.

(d) $401.20/96$.

(e) $(808.9614 - 401.20)/96$.

For a kth degree polynomial model, the estimate of σ is:

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-k-1}}.$$

The answer follows from the fact that $SSE = SST - SSR = 808.9614 - 401.20$.

5. Suppose that x is the length and Y is the weight of smallmouth bass (a type of fish) in a large Texas lake. A sample of smallmouth bass was caught at this lake and a simple linear regression model for Y versus x was fitted using the observed data. Using all the results we studied in class regarding inference for simple linear regression, the following two intervals were produced, one of which is a 95% confidence interval for the average weight (in pounds) of smallmouth bass that are 14 inches long, and the other is a 95% prediction interval for the weight (in pounds) of a particular smallmouth bass that is 14 inches long:

$$1.50 \pm 0.15 \quad \text{and} \quad 1.50 \pm 0.50.$$

Which of the following is correct?

(a) This information implies that the average weight of a 21 inch long smallmouth bass is about 2.25 lbs.

(b) The interval 1.50±0.15 is the prediction interval and the other is the confidence interval.

[✓ (c)] The interval 1.50±0.15 is the confidence interval and the other is the prediction interval.

(d) Both (a) and (b) are correct.

(e) Both (a) and (c) are correct.

The prediction interval for Y given x is *always* wider than the confidence interval (of the same confidence level) for the average of Y given x, since the former accounts for more uncertainty, the additional variability in Y (or the noise ϵ) given x. This is also seen from pg. 37 of Chapter 1A notes and was discussed at length in class. Hence, option (c) is correct.

Read the following information to answer **Questions 6–9**:

The following information is from a report on the determination of silver content of galena crystals grown in a closed hydrothermal system over a range of temperatures. Let x denote the crystallization temperature in degrees centigrade, and Y the silver content in mol%. A simple linear regression was fitted for Y versus x. The observed data are as follows:

x	398	292	352	575	568	450	550	408	484	350	503	600
y	0.15	0.05	0.23	0.43	0.23	0.4	0.44	0.44	0.45	0.09	0.59	0.63

Some summary statistics from the regression analysis are as follows:

$$\sum_{i=1}^{12} x_i = 5530, \qquad \sum_{i=1}^{12} y_i = 4.13, \qquad \sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = 155.4983,$$

$$\sum_{i=1}^{12}(x_i - \bar{x})^2 = 113642, \qquad \sum_{i=1}^{12}(y_i - \bar{y})^2 = 0.39709, \qquad \sum_{i=1}^{12}(y_i - \hat{y}_i)^2 = 0.18432.$$

Use this information to answer **Questions 6-9** below.

6. The least-squares estimates of the intercept and slope parameters, respectively, are:

(a) $\hat{\beta}_0 = 0.2164$ and $\hat{\beta}_1 = 0.224$.

(b) $\hat{\beta}_0 = 0.42$ and $\hat{\beta}_1 = 0.05$.

(c) $\hat{\beta}_0 = 0.2164$ and $\hat{\beta}_1 = -0.224$.

[✓ (d)] $\hat{\beta}_0 = -0.2864$ and $\hat{\beta}_1 = 0.001368$.

(e) $\hat{\beta}_0 = -0.80$ and $\hat{\beta}_1 = 0.001368$.

The slope estimate is:

$$\hat{\beta}_1 = \frac{155.4983}{113642} = 0.001368,$$

and the intercept is:

$$\hat{\beta}_0 = 4.13/12 - \hat{\beta}_1(5530/12) = -0.2864.$$

7. To test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$, the test statistic to be used is:

(a) $\hat{\beta}_1/(\hat{\sigma}\sqrt{1/12 + 1/113642})$

(b) $(\hat{\beta}_0 + \hat{\beta}_1)/\hat{\sigma}$.

(c) $\hat{\beta}_1/\hat{\sigma}$.

(d) $\sqrt{12}\,\hat{\beta}_1/\hat{\sigma}$.

[✓ (e)] $(337.108)\hat{\beta}_1/\hat{\sigma}$.

From pg. 32 of the Chapter 1A notes the test statistic T is:

$$T = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/\sqrt{\sum_{i=1}^{n}(x - \bar{x})^2}} = \frac{(337.108)\hat{\beta}_1}{\hat{\sigma}}.$$

8. The proportion of variance in silver content that is explained by temperature is:

(a) 0.668.

(b) 0.956.

(c) 0.810.

(d) 0.464.

[✓ (e)] 0.536.

We know that this proportion is R^2, or SSR/SST. We are given $SSE = 0.18432$. So

$$R^2 = SSR/SST = (SST - SSE)/SST = (0.39709 - 0.18432)/0.39709 = 0.536.$$

9. Given that the quantity: $1/12 + (500 - 5530/12)^2/113642$ equals 0.0968, a 95% confidence interval for the *average* silver content of galena crystals grown in a closed hydrothermal system at 500 degrees centigrade is:

(a) $(\hat{\beta}_0 + 500\hat{\beta}_1) \pm 1.96\hat{\sigma}\sqrt{0.0968}$.

(b) $(\hat{\beta}_0 + 500\hat{\beta}_1) \pm 1.96\hat{\sigma}\sqrt{1 + 0.0968}$.

[✓ (c)] $(\hat{\beta}_0 + 500\hat{\beta}_1) \pm 2.228\hat{\sigma}\sqrt{0.0968}$.

(d) $(\hat{\beta}_0 + 500\hat{\beta}_1) \pm 2.228\hat{\sigma}\sqrt{1 + 0.0968}$.

(e) Cannot be determined from the information given.

Just use the formula for the confidence interval on p. 35 of the notes (Chapter 1-Part A).

10. A simple linear regression model is fitted to $n = 20$ observations of (x, y) pairs, and the following summary statistics were computed:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 202.71 \qquad \text{and} \qquad \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 157.25.$$

In this case, the estimated error variance $\hat{\sigma}^2$ is:

(a) 157.25/202.71.

(b) 157.25/18.

[✓ (c)] (202.71 – 157.25)/18.

(d) 1 – 157.25/202.71.

(e) Cannot be determined from the information given.

Using the summary statistics, $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 157.25$ and $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 202.71$. Hence, $SSE = SST - SSR = 202.71 - 157.25$. The estimate of the error variance is given by: $\hat{\sigma}^2 = SSE/(n-2)$ which therefore equals $(202.71 - 157.25)/18$ since $n = 20$.

11. Two random variables X and Y have a population correlation coefficient, ρ, that equals 0. We may thus conclude that:

(a) X and Y are independent.

(b) X and Y are not independent.

[✓ (c)] X and Y could be independent.

(d) $E(X) = E(Y) = 0$.

(e) Donkeys can fly.

See pg. 43 of the Chapter 1A notes. Two random variables may be uncorrelated (i.e. $\rho = 0$) and still may be dependent!

12. A researcher has data $(x_1, y_1), \ldots, (x_n, y_n)$ and wants to fit the following linear model:

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

i.e. a linear regression model *without* an intercept parameter. Here, $\epsilon_1, \ldots, \epsilon_n$ are independent and each $\epsilon_i \sim N(0, \sigma^2)$ as usual. Define $\bar{y} = \sum_{i=1}^{n} y_i/n$ and $\bar{x} = \sum_{i=1}^{n} x_i/n$. Then, the correct expression for the least-squares estimate $\hat{\beta}$ of β for this model is given by:

(a) $\hat{\beta} = \bar{x}/\bar{y}$.

(b) $\hat{\beta} = \bar{y}/\bar{x}$.

(c) $\hat{\beta} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/\sum_{i=1}^{n}(x_i - \bar{x})^2$.

[✓ (d)] $\hat{\beta} = \sum_{i=1}^{n} x_i y_i/\sum_{i=1}^{n} x_i^2$.

(e) Cannot be determined from the information given.

The appropriate error sum of squares for the model being considered is given by:

$$f(\beta) = \sum_{i=1}^{n}(y_i - \beta x_i)^2.$$

We need to find the value of β that minimizes $f(\beta)$ and this minimizer will be the required least-squares estimate $\hat{\beta}$ of β. Taking the derivative of $f(\beta)$ with respect to β, we have:

$$f'(\beta) = \frac{d}{d\beta}f(\beta) = 2\sum_{i=1}^{n}(y_i - \beta x_i)(-x_i).$$

Setting this derivative equal to 0 and solving for β gives the reqd. solution/minimizer $\hat{\beta}$ as:

$$-2\sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0 \;\Rightarrow\; \sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0 \;\Rightarrow\; \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \beta x_i^2 \;\Rightarrow\; \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

13. Suppose that a response variable Y is such that $Y = \alpha e^{\beta x}\epsilon$, where ϵ is a random variable and α, β and x are constants. Then, the correct expressions for $\mathrm{Var}(Y)$ and $\mathrm{Var}(\log Y)$ are:

(a) $\mathrm{Var}(Y) = \mathrm{Var}(\epsilon)$ and $\mathrm{Var}(\log Y) = \mathrm{Var}(\log \epsilon)$.

(b) $\mathrm{Var}(Y) = \alpha e^{\beta x}\mathrm{Var}(\epsilon)$ and $\mathrm{Var}(\log Y) = \mathrm{Var}(\log \epsilon)$.

(c) $\mathrm{Var}(Y) = \alpha^2 e^{2\beta x}\mathrm{Var}(\epsilon)$ and $\mathrm{Var}(\log Y) = \log\alpha + \beta x + \mathrm{Var}(\log \epsilon)$.

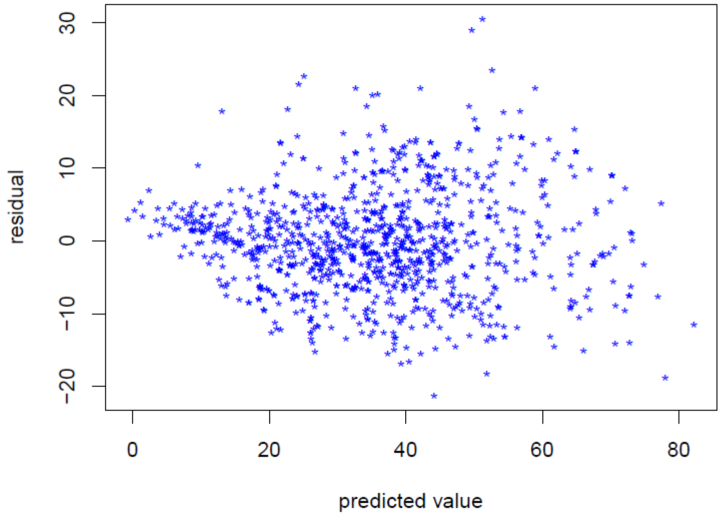
(d) $\mathrm{Var}(Y) = \alpha^2 e^{2\beta x}\mathrm{Var}(\epsilon)$ and $\mathrm{Var}(\log Y) = \log\alpha + \log(\beta x) + \mathrm{Var}(\log \epsilon)$.

[✓ (e)] $\mathrm{Var}(Y) = \alpha^2 e^{2\beta x}\mathrm{Var}(\epsilon)$ and $\mathrm{Var}(\log Y) = \mathrm{Var}(\log \epsilon)$.

Since α, β and x are constants,

$$\begin{aligned}\mathrm{Var}(Y) &= \left(\alpha e^{\beta x}\right)^2 \mathrm{Var}(\epsilon) &= \alpha^2 e^{2\beta x}\mathrm{Var}(\epsilon) \quad \text{and}\\ \mathrm{Var}(\log Y) &= \mathrm{Var}(\log\alpha + \beta x + \log\epsilon) &= \mathrm{Var}(\log\epsilon).\end{aligned}$$

14. The residual plot below was obtained from a polynomial regression analysis.



From this plot we can see that:

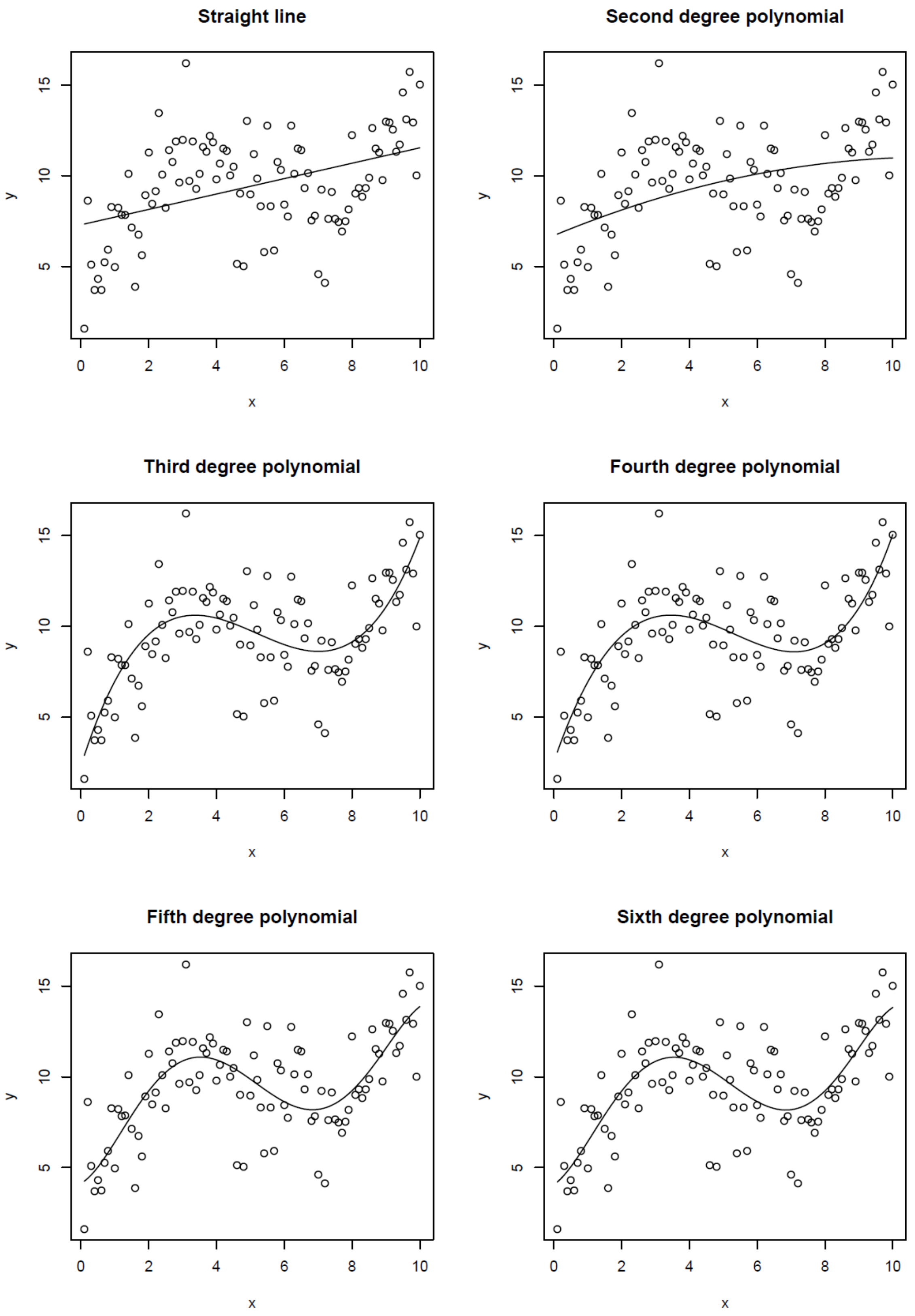
(a) The residuals show no pattern.

(b) The residuals are clearly Normally distributed.

[✓ (c)] The residual variance increases somewhat as the predicted value increases.

(d) Both (b) and (c) are true.

(e) The residuals were obviously computed incorrectly.


Straight line

Second degree polynomial

Third degree polynomial

Fourth degree polynomial

Fifth degree polynomial

Sixth degree polynomial

Table A.3 The Cumulative Distribution Function for the Standard Normal Distribution: Values of $\Phi(z)$ for Nonnegative z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table A.4 Percentiles of the *T* Distribution

df	90%	95%	97.5%	99%	99.5%	99.9%
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.183	4.541	5.841	10.215
4	1.533	2.132	2.777	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.708	5.208
7	1.415	1.895	2.365	2.998	3.500	4.785
8	1.397	1.860	2.306	2.897	3.355	4.501
9	1.383	1.833	2.262	2.822	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.625	2.977	3.787
15	1.341	1.753	2.132	2.603	2.947	3.733
16	1.337	1.746	2.120	2.584	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.879	3.611
19	1.328	1.729	2.093	2.540	2.861	3.580
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.788	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.705	3.307
80	1.292	1.664	1.990	2.374	2.639	3.195
∞	1.282	1.645	1.960	2.326	2.576	3.090