Delft University of Technology
Faculty EWI

Mekelweg 4
2628 CD Delft

**Exam wi4301IN**

**1 November 2010**
**14:00 - 17:00**

It is allowed to use a pocket calculator and one self-made A4-sheet that may contain notes on both sides. It is *not* allowed to use a book or other notes during the exam.

1. (a) In the appendix to this question, a figure is displayed with two normal QQ-plots of two datasets of size 100 that have been generated from two unknown distributions. What can you conclude about the thickness of the left and the right tail of the two distributions compared to that of the normal distribution?

   (b) The table below lists the median and the lower- and upper hinges of several Box-Cox power transformations of a skewed dataset:

| Power | $-1$ | $-0.5$ | $0$ | $0.5$ | $1$ |
|---|---|---|---|---|---|
| $H_L$ | $-9.25$ | $-4.39$ | $-2.32$ | $-1.37$ | $-0.90$ |
| Median | $-0.73$ | $-0.63$ | $-0.55$ | $-0.48$ | $-0.42$ |
| $H_U$ | $0.37$ | $0.41$ | $0.45$ | $0.51$ | $0.58$ |

   Give the definition of the Box-Cox family of powers and decide which of the transformations is most suitable to correct the skewness of the original dataset. Moreover, argue whether the original dataset is skewed to the right (or left), i.e., has a modus right (or left) of the center.

   (c) In the figure given in the appendix to this question, a spread-level plot is displayed for the population (in 1000's) in 51 states in the US divided over 9 regions. Decide on the basis of this plot, which region has the largest variation in population size and which power transformation is suggested to correct the nonconstant spread over the regions.

2. In an experiment one sets the values of two variables $X_1$ and $X_2$ and measures the value of a response variable $Y$. The settings and measurements are listed below:

| | | | | | | |
|---|---|---|---|---|---|---|
| $X_1$ | $-3$ | $2$ | $2$ | $1$ | $0$ | $-2$ |
| $X_2$ | $1$ | $-1$ | $-2$ | $1$ | $-1$ | $2$ |
| $Y$ | $2$ | $1$ | $-2$ | $5$ | $-1$ | $5$ |

   The corresponding regression equation is $Y_i = A + B_1 X_{i1} + B_2 X_{i2} + E_i$.

   (a) Determine the least squares estimators $A$, $B_1$ and $B_2$.

   (b) One performs two more experiments with settings

| | | |
|---|---|---|
| $X_1$ | $1$ | $-1$ |
| $X_2$ | $a$ | $-a$ |

   for some $a \in \mathbb{R}$, measures the corresponding responses and determines the least squares estimators on the basis of all 8 records. For which value $a \in \mathbb{R}$ are $A$, $B_1$ and $B_2$ uncorrelated?

3. In the appendix to this question part of the output is listed of a regression analysis concerning the variables `fconvict`, `year`, `tfr`, `partic`, and `degrees`.

   (a) Investigate whether `tfr` contributes significantly to the full model by testing an appropriate null hypothesis at significance level 1%. Report the value of the test statistic and give your conclusion about the null hypothesis.

   (b) Investigate whether `partic` and `degrees` can be left out of the model simultaneously by testing an appropriate null hypothesis at significance level 5%. Report the value of the test statistic and give your conclusion about the null hypothesis.

4. Continue with the model and the output of question 3.

   (a) Determine a 95% confidence interval for the coefficient of the variable `degrees` in the full model.

   (b) The 95% confidence interval for the expected value of `fconvict` for the `year` 2011, with `tfr`=500, `partic`=800 and `degrees`=300, has width 283.28. Determine the 98% prediction interval for `fconvict` for this case.

5. In the appendix to this question the output is listed from regressing the variable `infantMortality` (infant deaths per 1000 live births) on `lifeFemale` (life expectancy at birth for women), `educationFemale` (average number of years of education for women) and the factor `region`, which is coded by four dummy-variables as follows:

| Region | $D_1$ | $D_2$ | $D_3$ | $D_4$ |
|---|---|---|---|---|
| Africa | 1 | 0 | 0 | 0 |
| America | 0 | 1 | 0 | 0 |
| Asia | 0 | 0 | 1 | 0 |
| Europe | 0 | 0 | 0 | 1 |
| Oceania | 0 | 0 | 0 | 0 |

   Answer the following questions on the basis of this output.

   (a) Specify the regression equation for Africa and Europe separately.

   (b) Investigate whether the interaction between region and the other two regressors is significant by testing an appropriate null hypothesis at level 5%. Report the value of the test statistic and give your conclusion about the null hypothesis.

   (c) Consider the model with `lifeFemale`, `educationFemale`, `region`, and *only* the interaction between `region` and `lifeFemale`. Investigate whether this interaction is significant by testing the appropriate null hypothesis at level 5%. Report the value of the test statistic and give your conclusion about the null hypothesis.

6. In a psychological experiment two response variables of a personality questionnaire are recorded for 1421 individuals, i.e.,

   - `neuroticism`: degree of being emotional
   - `extraversion`: degree of being talkative and outgoing

The individuals are divided by sex (`Female` - `Male`) and being a volunteer (`Yes` - `No`). In the appendix to this question the averages of `extraversion` are displayed graphically over the two factors and part of the output is listed of a 2-way ANOVA for `neuroticism` by `sex` and `volunteer`. Answer the following questions on the basis of this output.

**NB:** You may use $F_{0.05}(\nu_1, \nu_2) \approx F_{0.05}(\nu_1, 120)$ for $\nu_2 > 120$.

(a) What sort of main and/or interaction effects on `extraversion` do you expect to find on the basis of the plot? Motivate your conclusions clearly.

(b) Test whether the interaction between `sex` and `volunteer` is significant for `neuroticism` at level 5%. Report the value of the test statistic and motivate your conclusion.

(c) In the model without interaction, test whether `sex` has a significant effect on `neuroticism` (after `volunteer` has been added to the model). Report the value of the test statistic and motivate your conclusion.

7. (a) In the appendix to this question part of the output is listed of a principal component analysis based on the correlation matrix $\mathbf{R}$ of three explanatory variables $X_1$, $X_2$ and $X_3$, as well as the multiple correlation coefficients of three partial regressions. On the basis of this output, compute the variance inflation factors and determine what proportion of variation is accounted for by the first principal component.

(b) During the course we have introduced a number of diagnostics, such as (in alphabetical order)

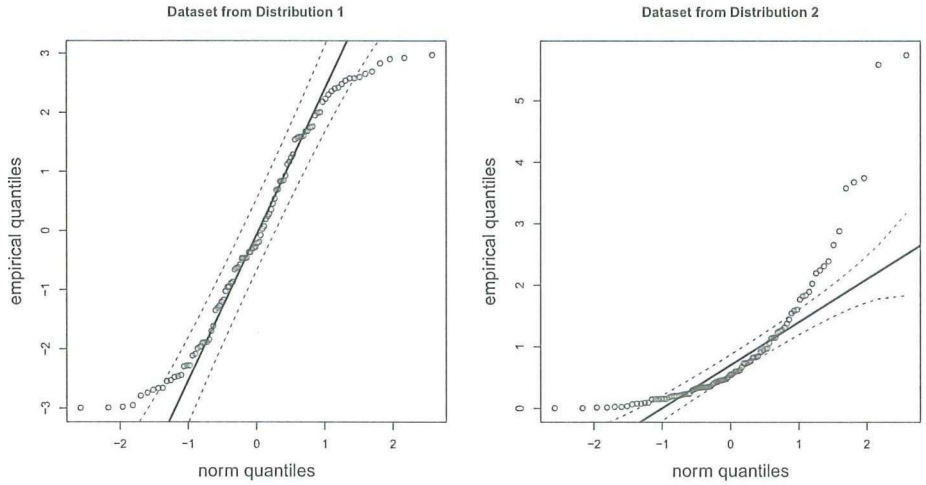
1. added-variable plot
2. component-plus-residual plot
3. Cook's distance
4. COVRATIO
5. DFBETAS
6. hat-values
7. studentized residuals

Report which of the above diagnostics can be used to detect the following anomalies in the data:

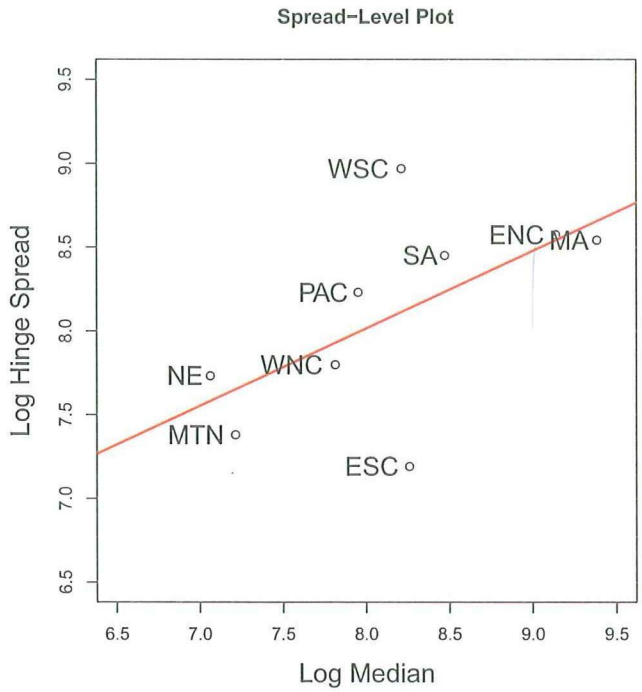
a. high leverage
b. outliers
c. influential observations
d. nonnormal errors
e. nonconstant error variance
f. nonlinearity

## Appendix to question 1.

- Two normal QQ-plots:




- Spread-level plot over regions `ENC`, East North Central; `ESC`, East South Central; `MA`, Mid-Atlantic; `MTN`, Mountain; `NE`, New England; `PAC`, Pacific; `SA`, South Atlantic; `WNC`, West North Central; `WSC`, West South Central [1].




[1] Data from United States (1992) *Statistical Abstract of the United States.* Bureau of the Census.

**Appendix to questions 3 and 4.**[2] Output of a linear regression concerning the variables

`fconvict`: Female indictable-offense conviction rate per 100 000.

`year`: 1931 – 1968.

`tfr`: Total fertility rate per 1000 women.

`partic`: Women's labor-force participation rate per 1000.

`degrees`: Women's post-secondary degree rate per 10 000.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2255.2066  1962.3582 -1.1492   0.2587
year            1.3015     1.0452  1.2452   0.2218
tfr            -0.0697     0.0157  ......   ......
partic          0.1890     0.1344  1.4062   ......
degrees        -0.7866     ......  ......   ......

Residual standard error: 19.23 on 33 degrees of freedom
Multiple R-squared: 0.6933,     Adjusted R-squared: 0.6562
F-statistic: 18.65 on 4 and 33 DF,  p-value: 4.213e-08

Analysis of Variance Table

Response: fconvict
          Df  Sum Sq Mean Sq F value    Pr(>F)
year       1  2841.6 .......  7.6823  0.009093
tfr        1 22381.6 ....... 60.5082  ........
partic     1  1285.6 .......  3.4757  0.071192
degrees    1  1089.5 .......  2.9454  0.095497
Residuals 33 ....... .......
```

[2]Data from Fox, J., and Hartnagel, T. F (1979) *Changing social roles and female crime in Canada: A time series analysis.* Canadian Review of Sociology and Anthroplogy, 16, 96104.

## Appendix to question 5.[3]

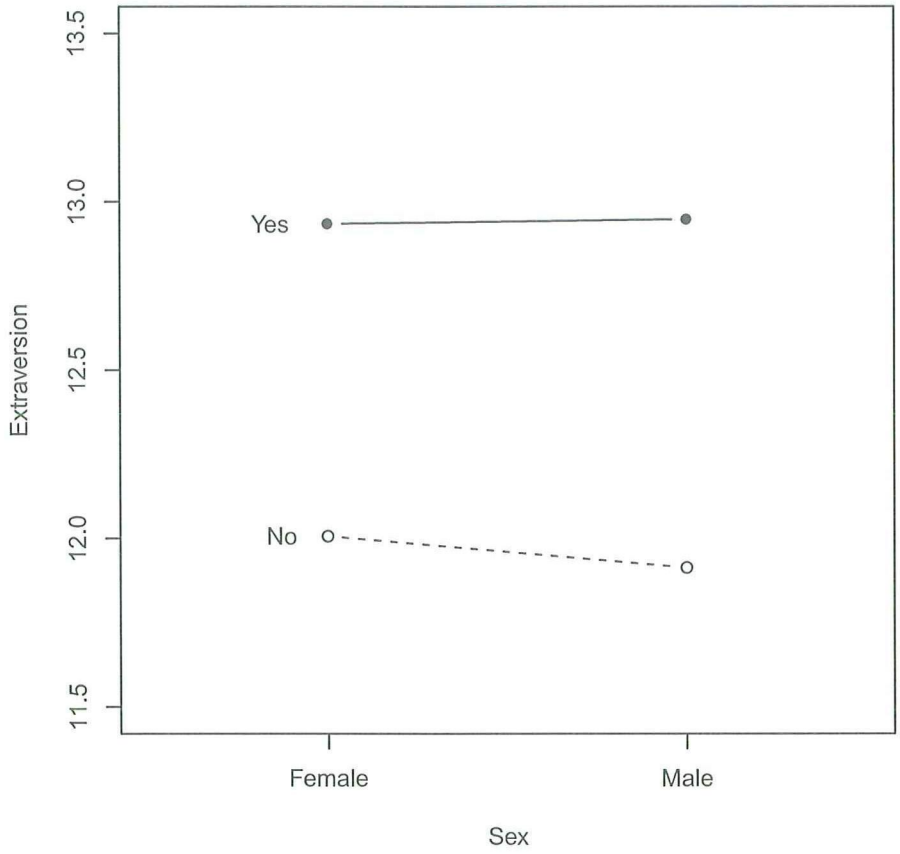
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         320.6848   222.7359   1.440    0.155
lifeFemale           -2.8120     4.5288  -0.621    0.537
educationFemale      -5.3633     9.0410  -0.593    0.555
d1                  -79.6390   223.1558  -0.357    0.722
d2                   46.2252   230.1219   0.201    0.841
d3                  -44.3573   224.3209  -0.198    0.844
d4                 -212.7342   231.0068  -0.921    0.361
lifeFemale:d1         0.1315     4.5379   0.029    0.977
lifeFemale:d2        -1.9200     4.6239  -0.415    0.679
lifeFemale:d3        -0.3177     4.5553  -0.070    0.945
lifeFemale:d4         1.7728     4.6404   0.382    0.704
educationFemale:d1    4.1495     9.0759   0.457    0.649
educationFemale:d2    6.5886     9.1616   0.719    0.475
educationFemale:d3    3.6802     9.1445   0.402    0.689
educationFemale:d4    4.1396     9.2184   0.449    0.655

Analysis of Variance Table

Response: infantMortality
                   Df Sum Sq Mean Sq   F value    Pr(>F)
lifeFemale          1  78587   78587 1069.0836 < 2.2e-16
educationFemale     1    317     317    4.3090  0.042134
d1                  1     22      22    0.2954  0.588739
d2                  1    103     103    1.3987  0.241539
d3                  1      7       7    0.1016  0.750950
d4                  1    384     384    5.2280  0.025711
lifeFemale:d1       1     96      96    1.3086  0.257119
lifeFemale:d2       1     83      83    1.1330  0.291341
lifeFemale:d3       1    314     314    4.2650  0.043163
lifeFemale:d4       1    657     657    8.9384  0.004022
educationFemale:d1  1     20      20    0.2730  0.603220
educationFemale:d2  1    170     170    2.3173  0.133113
educationFemale:d3  1      1       1    0.0182  0.893158
educationFemale:d4  1     15      15    0.2017  0.654983
Residuals          61   4484    ....
```

[3]U.N. social indicators data taken from J. Fox (2008) *Applied Regression, Generalized Linear Models, and Related Methods*, Second Edition, Sage Publications.

**Appendix to question 6.** Group averages of `Extraversion` over `sex` and `volunteer`[4]




Output from fitting a two-way ANOVA model with interaction:

```
Analysis of Variance Table

Response: neuroticism
               Df Sum Sq Mean Sq F value    Pr(>F)
volunteer       1      5       5  ......    0.6290
sex             1   1124    1124  ......    ......
volunteer:sex   1      2       2  ......    ......
Residuals     ...  32950     ...
```

[4]Data from Cowles, M. and C. Davis (1987) The subject matter of psychology: Volunteers. *British Journal of Social Psychology* **26**, 97102.

**Appendix to question 7.**

- Loadings (eigenvectors of $\mathbf{R}$):

| $\mathbf{a}_1$ | $\mathbf{a}_2$ | $\mathbf{a}_3$ |
|---|---|---|
| 0.5442 | 0.8380 | 0.0405 |
| 0.5953 | −0.3517 | −0.7224 |
| 0.5911 | −0.4172 | 0.6903 |

- Eigenvalues of $\mathbf{R}$

| $L_1$ | $L_2$ | $L_3$ |
|---|---|---|
| ... | 0.292 | 0.024 |

- Squared multiple correlation coefficients $R_j^2$ for regressing $X_j$ on other $X$'s:

| $j$ | Partial regression | $R_j^2$ |
|---|---|---|
| 1 | $X_1$ on $X_2, X_3$ | 0.6128 |
| 2 | $X_2$ on $X_1, X_3$ | ... |
| 3 | $X_3$ on $X_1, X_2$ | 0.9518 |

| | $\longleftarrow \nu_1 \longrightarrow$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\nu_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 | 238.88 | 240.54 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 |
| 50 | 4.03 | 3.18 | 2.79 | 2.56 | 2.40 | 2.29 | 2.20 | 2.13 | 2.07 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 |
| 70 | 3.98 | 3.13 | 2.74 | 2.50 | 2.35 | 2.23 | 2.14 | 2.07 | 2.02 |
| 80 | 3.96 | 3.11 | 2.72 | 2.49 | 2.33 | 2.21 | 2.13 | 2.06 | 2.00 |
| 90 | 3.95 | 3.10 | 2.71 | 2.47 | 2.32 | 2.20 | 2.11 | 2.04 | 1.99 |
| 100 | 3.94 | 3.09 | 2.70 | 2.46 | 2.31 | 2.19 | 2.10 | 2.03 | 1.97 |
| 110 | 3.93 | 3.08 | 2.69 | 2.45 | 2.30 | 2.18 | 2.09 | 2.02 | 1.97 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 |

Tabel 1: Rechter kritieke waarden $F_\alpha(\nu_1, \nu_2)$ van de $F(\nu_1, \nu_2)$ verdeling bij $\alpha = 0.05$.

| $m \backslash p$ | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0025 | 0.001 | 0.0005 |
|---|---|---|---|---|---|---|---|---|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 127.321 | 318.309 | 636.619 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 14.089 | 22.327 | 31.599 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 7.453 | 10.215 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.222 | 3.646 | 3.965 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.197 | 3.610 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.135 | 3.527 | 3.819 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.119 | 3.505 | 3.792 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.104 | 3.485 | 3.768 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.091 | 3.467 | 3.745 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.078 | 3.450 | 3.725 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.067 | 3.435 | 3.707 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.057 | 3.421 | 3.690 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.047 | 3.408 | 3.674 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.038 | 3.396 | 3.659 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.030 | 3.385 | 3.646 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 2.971 | 3.307 | 3.551 |
| 50 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 | 2.937 | 3.261 | 3.496 |
| $\infty$ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 2.807 | 3.090 | 3.291 |

Tabel 2: Right critical values $t_{m,p}$ of the $t$-distribution with $m$ degrees of freedom corresponding to right tail probability $p$: $\mathrm{P}\left(T_m \geq t_{m,p}\right) = p$. The last row in the table are right critical values of the $N(0,1)$ distribution: $t_{\infty,p} = z_p$