# BIOSTAT III: Survival Analysis

# Examination

November 18, 2011

Time: 9:00–11.30

Exam room location: Sal Jacob Berzelius (aka Adam),
Berzelius väg 3, Karolinska Institutet

Code (please do not write your name):

- Time allowed is 2 1/2 hours.
- Please try and write your answers on the exam sheet. You may use separate paper if absolutely necessary. Your working and motivation for your answer, not just the final answer, will be assessed when grading the examination.
- The exam contains 2 sections; the first section tests your knowledge in general epidemiological concepts in a survival analysis framework whereas the second section focusses on more specific topics in survival analysis. Each section contains 3 questions (with several parts). The marks available for each part are indicated.
- A score of 8 marks or more out of a possible 15 in each of the two sections will be required to obtain a passing grade.
- The questions may be answered in English or Swedish (or a combination thereof).
- A non-programmable scientific calculator (i.e., with ln() and exp() functions) will most probably be useful. You may not use a mobile phone or other communication device as a calculator or for any other purpose.
- The exam is not 'open book' but each student will be allowed to bring one A4 sheet of paper into the exam room which may contain, for example, hand-written notes or photocopies from textbooks/lecture notes etc. Both sides of the page may be used.
- The exam supervisors have been advised not to answer any questions you may have regarding the content of the exam. If you believe a question contains an error or is ambiguous then please write a note with your answer indicating how you have interpreted the question.
- Tables of critical values of the $\chi^2$ distribution are provided on the last page.

# Description of the data

In Sweden, every physician and pathologist/cytologist is obliged by law to report each occurrence of cancer to the population-based nationwide Swedish Cancer Registry. Some of the questions in this examination are based on a statistical analysis, performed using Stata 11, of the survival of women diagnosed with ovarian cancer in Sweden between 1993 and 2009. Women were followed up from the date of diagnosis until death, first emigration or 31 December 2010, whichever occurred first. The outcome of interest is death due to any cause. The variable `dead` was coded as `1` (one) for women who died during follow-up and `0` (zero) for women who did not die.

The following Stata output shows output from the `stset` command and frequency tables for some of the variables used in the analysis.

```
. /** stset the data using time since diagnosis as the timescale **/

. stset exitdate, failure(dead == 1) enter(diagdate) ///
                  origin(diagdate) scale(365.24)

     failure event:  dead == 1
obs. time interval:  (origin, exitdate]
 enter on or after:  time diagdate
 exit on or before:  failure
    t for analysis:  (time-origin)/365.24
            origin:  time diagdate

--------------------------------------------------------------------------
     9078  total obs.
        0  exclusions
--------------------------------------------------------------------------
     9078  obs. remaining, representing
     5748  failures in single record/single failure data
 42686.91  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  17.98817


. tab agediag_cat (Age at diagnosis)

agediag_cat     |      Freq.     Percent        Cum.
----------------+-----------------------------------
0 = 16-44 years |        725        7.99        7.99
1 = 45-54 years |      1,654       18.22       26.21
2 = 55-64 years |      2,377       26.18       52.39
3 = 65-74 years |      2,404       26.48       78.87
4 = >74 years   |      1,918       21.13      100.00
----------------+-----------------------------------
          Total |      9,078      100.00
```

```
. tab period_cat

period_cat  |      Freq.     Percent        Cum.
------------+-----------------------------------
0 = 1993-1998 |      3,373       37.16       37.16
1 = 1999-2004 |      3,136       34.55       71.70
2 = 2005-2009 |      2,569       28.30      100.00
------------+-----------------------------------
        Total |      9,078      100.00

. tab histology

      Histology of tumour |      Freq.     Percent        Cum.
--------------------------+-----------------------------------
 1 =       Serous tumours |      5,667       62.43       62.43
 2 =     Mucinous tumours |      1,185       13.05       75.48
 3 = Endometrioid tumours |      1,639       18.05       93.53
 4 =   Clear cell tumours |        587        6.47      100.00
--------------------------+-----------------------------------
                    Total |      9,078      100.00

/** split the data according to follow-up time **/
. stsplit fup, at(1 5 10)
(11956 observations (episodes) created)

. tab fup

       timeband |      Freq.     Percent        Cum.
----------------+-----------------------------------
0 =    0-1 year |      9,078       43.16       43.16
1 =   1-5 years |      7,588       36.07       79.23
5 =  5-10 years |      3,022       14.37       93.60
10 = > 10 years |      1,346        6.40      100.00
----------------+-----------------------------------
          Total |     21,034      100.00
```

# Section 1

The following questions test your knowledge of general concepts in statistical modelling of epidemiological data.

1. We first fit a Cox regression model adjusted for age at diagnosis, calendar period of diagnosis, histology and time since diagnosis.

```
MODEL A
stcox i.agediag_cat i.period_cat i.histology


Cox regression -- Breslow method for ties

No. of subjects =         9078                     Number of obs   =      9078
No. of failures =         5748
Time at risk    =  42686.90724
                                                   LR chi2(9)      =   1308.23
Log likelihood  =   -48263.538                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 agediag_cat |
           1 |   1.524186   .1053999     6.09   0.000     1.330994     1.74542
           2 |   2.015157   .1327233    10.64   0.000     1.771114    2.292826
           3 |   2.627054   .1706197    14.87   0.000     2.313055    2.983679
           4 |   4.201458   .2745555    21.97   0.000     3.696375    4.775557
             |
period_cat   |
           1 |   .8920521   .0269964    -3.77   0.000     .8406787    .9465649
           2 |   .9135931   .0338503    -2.44   0.015     .8495995    .9824068
             |
   histology |
           2 |   .6644437   .0285876    -9.50   0.000     .6107104    .7229047
           3 |   .5740802   .0218874   -14.56   0.000     .5327454    .6186221
           4 |   .6082746   .0367545    -8.23   0.000     .5403393    .6847512
------------------------------------------------------------------------------
```

   (a) From model A, can you assess if the effect of calendar period is confounded by age at diagnosis? Motivate your answer. (1 mark)

   (b) From model A, can you assess if the effect of calendar period is modified by age at diagnosis? If your answer is no, motivate why. If your answer if yes, assess this formally. Remember to state the null hypothesis, alternative hypothesis, value of the test statistic, assumed distribution of the test statistic under the null hypothesis, and a comment on statistical significance. (2 marks)

   (c) For each calendar period, provide an estimate of the hazard ratio that compares patients in the oldest age group to patients in the youngest age group. (2 marks)

2. We now also include interaction terms between the variables age at diagnosis and period of diagnosis.

```
  MODEL B
. stcox i.agediag_cat##i.period_cat i.histology


Cox regression -- Breslow method for ties

No. of subjects =         9078          Number of obs   =      9078
No. of failures =         5748
Time at risk    =  42686.90724
                                         LR chi2(17)     =   1335.57
Log likelihood  =   -48249.869           Prob > chi2     =    0.0000

------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|
-------------+----------------------------------------------
 agediag_cat |
           1 |   2.024925   .2152258     6.64   0.000
           2 |   2.662411   .2765082     9.43   0.000
           3 |    3.69056   .3744086    12.87   0.000
           4 |    5.72524   .5902736    16.92   0.000
             |
period_cat   |
           1 |   1.416403    .189393     2.60   0.009
           2 |   1.902441   .3128409     3.91   0.000
             |
 agediag_cat#|
period_cat   |
         1 1 |   .6546404   .1001043    -2.77   0.006
         1 2 |   .4585966   .0898603    -3.98   0.000
         2 1 |    .642826   .0943323    -3.01   0.003
         2 2 |   .4933915   .0879671    -3.96   0.000
         3 1 |   .5767046   .0834117    -3.81   0.000
         3 2 |   .4397882   .0776431    -4.65   0.000
         4 1 |   .6122446   .0890029    -3.37   0.001
         4 2 |   .4677025   .0826364    -4.30   0.000
             |
   histology |
           2 |   .6659783   .0286743    -9.44   0.000
           3 |   .5726746   .0218448   -14.61   0.000
           4 |   .6088562   .0368052    -8.21   0.000
------------------------------------------------------------
```

(a) From model A and/or B, can you assess if the effect of calendar period is confounded by age at diagnosis? Motivate your answer. (1 mark)

(b) From model A and/or B, can you assess if the effect of calendar period is modified by age at diagnosis? If your answer is no, motivate why. If your answer if yes, assess this formally. Remember to state the null hypothesis, alternative hypothesis, value of the test statistic, assumed distribution of the test statistic under the null hypothesis, and a comment on statistical significance. (2 marks)

(c) For each calendar period, provide an estimate of the hazard ratio that compares patients in the oldest age group to patients in the youngest age group. (2 marks)

3. We now instead present the parameter estimates of model B on the original scale (i.e., the scale on which the model is estimated).

```
. stcox i.agediag_cat##i.period_cat i.histology, nohr

Cox regression -- Breslow method for ties

No. of subjects =         9078        Number of obs   =      9078
No. of failures =         5748
Time at risk    =  42686.90724
                                      LR chi2(17)     =   1335.57
Log likelihood  =   -48249.869        Prob > chi2     =    0.0000

------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------------
 agediag_cat |
           1 |   .7055328   .1062883     6.64   0.000
           2 |   .9792323   .1038563     9.43   0.000
           3 |   1.305778   .1014503    12.87   0.000
           4 |   1.744884   .1031002    16.92   0.000
             |
period_cat   |
           1 |   .3481208    .133714     2.60   0.009
           2 |   .6431378   .1644418     3.91   0.000
             |
 agediag_cat#|
period_cat   |
         1 1 |  -.4236692   .1529149    -2.77   0.006
         1 2 |  -.7795844   .1959464    -3.98   0.000
         2 1 |  -.4418812   .1467463    -3.01   0.003
         2 2 |  -.7064523   .1782906    -3.96   0.000
         3 1 |  -.5504251   .1446351    -3.81   0.000
         3 2 |  -.8214621   .1765465    -4.65   0.000
         4 1 |  -.4906234   .1453715    -3.37   0.001
         4 2 |   -.759923   .1766859    -4.30   0.000
             |
   histology |
           2 |  -.4064982   .0430559    -9.44   0.000
           3 |  -.5574377   .0381453   -14.61   0.000
           4 |  -.4961731   .0604498    -8.21   0.000
------------------------------------------------------------
```

(a) Interpret the coefficient for histology 4 (i.e, -.4961731). (1 mark)

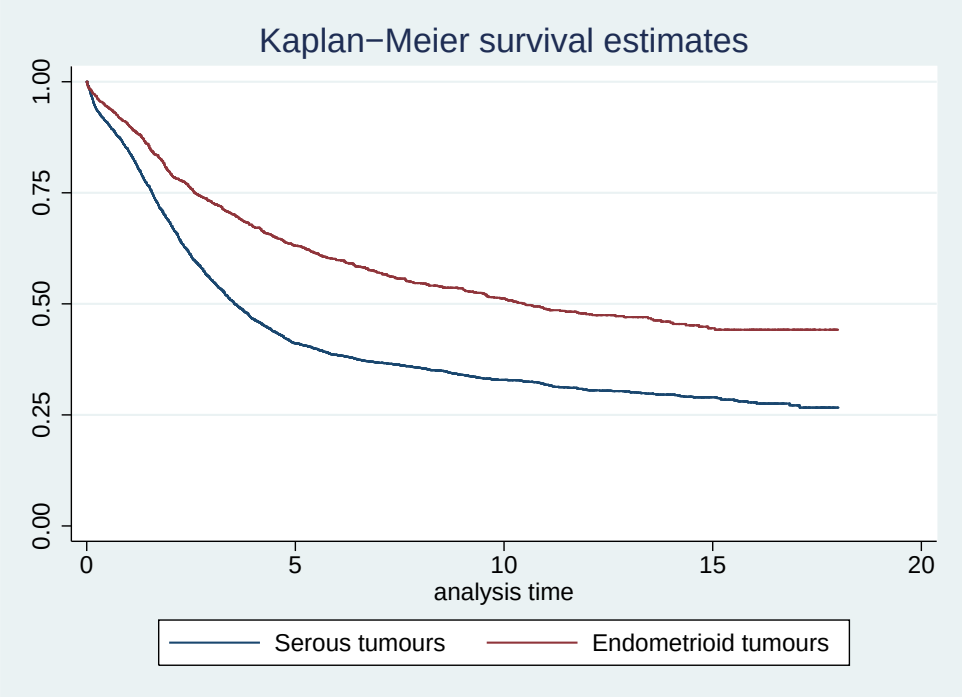
(b) Provide a 95% confidence interval for the hazard ratio that compares patients with endometriod tumours (histology = 3) to patients with serous tumours (histology = 1). (2 marks)

(c) What is the hazard ratio for comparing patients aged 65-74 to those aged 45-54 for patients diagnosed with mucinous tumours in 2003? (2 marks)

## Section 2

The following questions test your knowledge of concepts that are of special interest in survival analysis.

1. Below is a Kaplan-Meier graph showing the survival curves for two of the four groups of histology.




(a) Which histology group has highest survival? (0.5 mark)

(b) What is the 10-year survival for patients with serous tumours? (0.5 mark)

(c) What is the median survival time for patients with endometrioid tumours? (1 mark)

(d) During which years following diagnosis is the mortality higher for patients with serous tumours compared to those with endometrioid tumours? (1 mark)

2. This question tests your understanding of the proportional hazards assumption.

   (a) Model A, in section 1 of this exam, assumes proportional hazards for all covariate effects. What does this mean? (1 mark)

   (b) The Stata output below provides formal tests of the proportional hazards assumtion for each covariate effect in model A. For which covariate/covariates does the assumption not seem to be satisfied? (1 mark)

   (c) For a specific parameter (e.g. 4.histology), state the formal hypothesis for the test and comment on the statistical significance. (1 mark)

```
. estat phtest, detail

      Test of proportional-hazards assumption

      Time:  Time
      ------------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+-----------------------------------------------------
      1.agediag_cat|     -0.01632          1.54        1          0.2147
      2.agediag_cat|     -0.00056          0.00        1          0.9659
      3.agediag_cat|      0.01039          0.63        1          0.4290
      4.agediag_cat|     -0.00152          0.01        1          0.9079
      1.period_cat |      0.03275          6.21        1          0.0127
      2.period_cat |      0.02697          4.27        1          0.0389
      2.histology  |     -0.09485         52.96        1          0.0000
      3.histology  |     -0.02238          2.89        1          0.0890
      4.histology  |     -0.04535         11.89        1          0.0006
      ------------+-----------------------------------------------------
```

   (d) Explain two ways of modifying model A in section 1 so that it allows for non-proportional hazards. You can choose whichever variable you find most relevant to use for illustration. (2 marks)

3. This question tests your understanding of the use of regression models in a survival analysis framework.

   (a) We first fit a Poisson regression model adjusted for age at diagnosis, calendar period of diagnosis and histology. Use the Stata output below to draw estimates of the log hazard rates (natural logarithm, i.e. ln) for the youngest and the oldest age groups respectively, where histology and calendar period of diagnosis are at their reference levels. Use the blank graph provided below the Stata output. (2 marks)

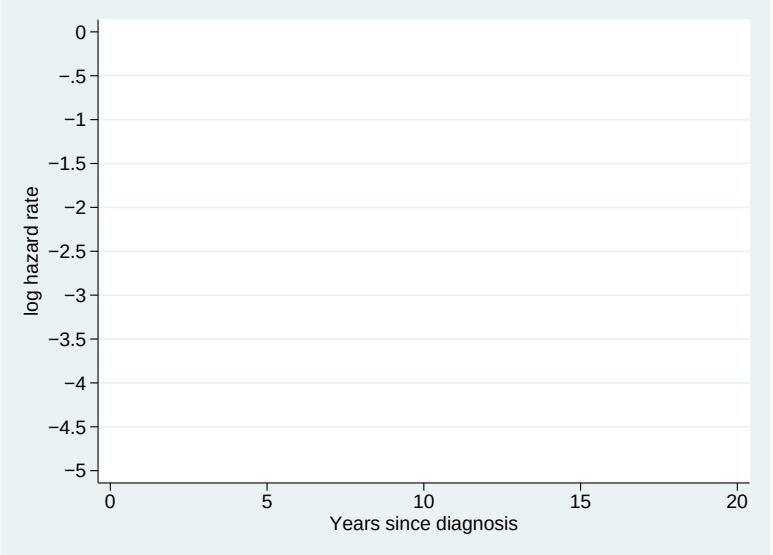
MODEL D

```
. streg i.agediag_cat i.period_cat i.histology, distribution(exponential) nohr

Exponential regression -- log relative-hazard form

Log likelihood  =   -13681.354                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 agediag_cat |
           1 |   .4662855   .0691347     6.74   0.000     .3307839     .601787
           2 |   .7873142   .0658051    11.96   0.000     .6583386    .9162897
           3 |   1.083855    .064833    16.72   0.000     .9567848    1.210926
           4 |   1.605046   .0651814    24.62   0.000     1.477293    1.732799
             |
period_cat   |
           1 |   .0401145   .0298635     1.34   0.179    -.0184169     .098646
           2 |   .2375498   .0359327     6.61   0.000     .1671231    .3079766
             |
   histology |
           2 |  -.4982549   .0429393   -11.60   0.000    -.5824143   -.4140955
           3 |  -.6308649   .0380513   -16.58   0.000     -.705444   -.5562858
           4 |  -.5842384   .0603669    -9.68   0.000    -.7025553   -.4659215
             |
       _cons |  -2.738888   .0623139   -43.95   0.000    -2.861021   -2.616755
------------------------------------------------------------------------------
```

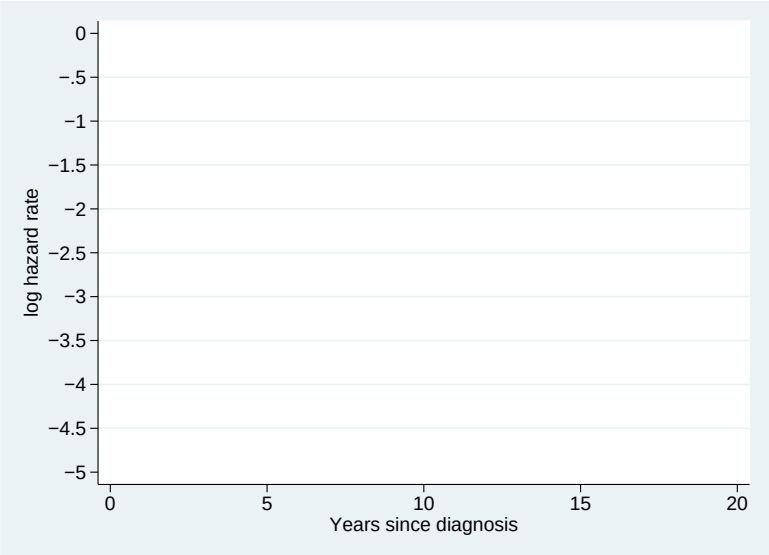


(b) We have split the data into timebands (variable name = fup) representing the underlying time scale. The time bands have been added to the Poisson regression model in part a. Again, use the Stata output below to draw estimates of the log hazard rates (natural logarithm, i.e. ln) for the youngest and the oldest age groups respectively, where histology and calendar period of diagnosis are at their reference levels. Use the blank graph provided below the Stata output. (2 marks)

```
Model E
. streg i.fup i.agediag_cat i.period_cat i.histology, distribut(exponential) nohr

Exponential regression -- log relative-hazard form

No. of subjects =         9078                 Number of obs   =     21034
No. of failures =         5748
Time at risk    =  42686.90724
                                               LR chi2(12)     =   2541.69
Log likelihood  =   -13343.707                 Prob > chi2     =    0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fup |
           1 |   .0516576   .0314384     1.64   0.100    -.0099604    .1132757
           5 |  -.6919949   .0462981   -14.95   0.000    -.7827375   -.6012523
          10 |  -1.257165    .079354   -15.84   0.000    -1.412696   -1.101634
             |
 agediag_cat |
           1 |   .4301627   .0691434     6.22   0.000     .2946441    .5656813
           2 |   .7146502   .0658464    10.85   0.000     .5855937    .8437068
           3 |   .9853482   .0649176    15.18   0.000      .858112    1.112584
           4 |   1.464699    .065304    22.43   0.000     1.336705    1.592692
             |
period_cat |
           1 |  -.1064008   .0302174    -3.52   0.000    -.1656259   -.0471758
           2 |  -.0465054   .0369093    -1.26   0.208    -.1188463    .0258355
             |
   histology |
           2 |  -.4250272   .0429919    -9.89   0.000    -.5092897   -.3407647
           3 |   -.570369   .0381005   -14.97   0.000    -.6450446   -.4956935
           4 |  -.5134653   .0604064    -8.50   0.000    -.6318596   -.3950709
             |
       _cons |  -2.398814   .0676798   -35.44   0.000    -2.531464   -2.266164
------------------------------------------------------------------------------
```



(c) If we would split the data at each event time (i.e., each time a death occurs) model E would be theoretically identical to model A (see section 1 of this exam). In the blank graph below draw an approximate estimate of what the log hazard rates (natural logarithm, i.e. ln) would look like for such model for the youngest and the oldest age groups respectively, where histology and calendar period of diagnosis are at their reference levels. (3 marks)

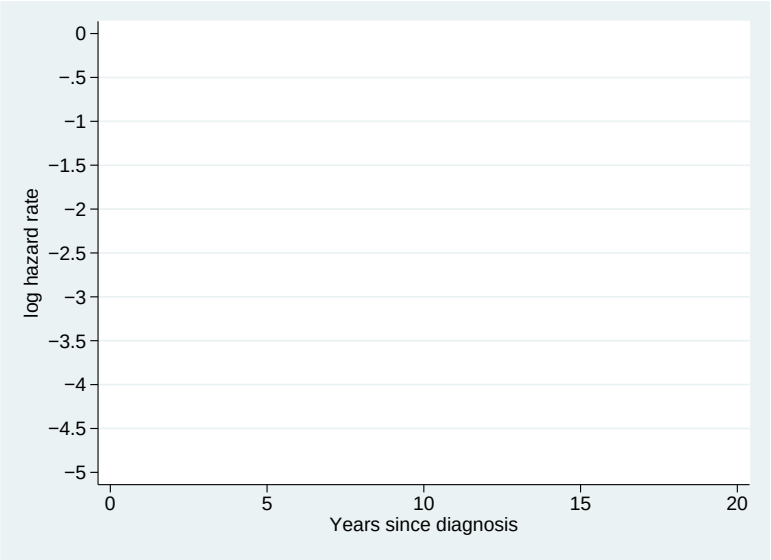


**Table A3** Critical Values of Chi-Square

| df | $\alpha = 0.10$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
|---|---|---|---|
| 1 | 2.706 | 3.841 | 6.635 |
| 2 | 4.605 | 5.991 | 9.210 |
| 3 | 6.251 | 7.815 | 11.345 |
| 4 | 7.779 | 9.488 | 13.277 |
| 5 | 9.236 | 11.070 | 15.086 |
| 6 | 10.645 | 12.592 | 16.812 |
| 7 | 12.017 | 14.067 | 18.475 |
| 8 | 13.362 | 15.507 | 20.090 |
| 9 | 14.684 | 16.919 | 21.666 |
| 10 | 15.987 | 18.307 | 23.209 |
| 11 | 17.275 | 19.675 | 24.725 |
| 12 | 18.549 | 21.026 | 26.217 |
| 13 | 19.812 | 22.362 | 27.688 |
| 14 | 21.064 | 23.685 | 29.141 |
| 15 | 22.307 | 24.996 | 30.578 |
| 16 | 23.542 | 26.296 | 32.000 |
| 17 | 24.769 | 27.587 | 33.409 |
| 18 | 25.989 | 28.869 | 34.805 |
| 19 | 27.204 | 30.144 | 36.191 |
| 20 | 28.412 | 31.410 | 37.566 |
| 21 | 29.615 | 32.671 | 38.932 |
| 22 | 30.813 | 33.924 | 40.289 |
| 23 | 32.007 | 35.172 | 41.638 |
| 24 | 33.196 | 36.415 | 42.980 |
| 25 | 34.382 | 37.652 | 44.314 |
| 30 | 40.256 | 43.773 | 50.892 |
| 35 | 46.059 | 49.802 | 57.342 |
| 40 | 51.805 | 55.758 | 63.691 |
| 45 | 57.505 | 61.656 | 69.957 |
| 50 | 63.167 | 67.505 | 76.154 |
| 60 | 74.397 | 79.082 | 88.379 |
| 70 | 85.527 | 90.531 | 100.425 |
| 80 | 96.578 | 101.879 | 112.329 |
| 90 | 107.565 | 113.145 | 124.116 |
| 100 | 118.498 | 124.432 | 135.807 |

The value tabulated is $c$ such that $P(\chi^2 \geq c) = \alpha$.