JANKOVY TABULKY — REVISITED

JAROMÍR ANTOCH

MFF UK, KPMS

Věnováno Ing. Josefu Machkovi, CSc., který více než podstatnou měrou přispěl k tomu, že se Jankovy tabulky staly vysoce ceněnou pomůckou v rukách mnoha statistiků, a památce Prof. Dr. Berndta Streitberga, který mi během své brilantní přednášky v létě 1991 pootevřel dveře k některým možnostem systému MATHEMATICA.

SHRNUTÍ: Cílem tohoto příspěvku je čtenářům přiblížit některé možnosti, jež při výpočtu kritických hodnot, kvantilů a dalších důležitých statistických charakteristik nabízí pokrok ve výpočetním prostředí pro statistiku, zvláště pak symbolické procedurální jazyky typu *MATHEMATICA*.

NĚKOLIK SLOV ÚVODEM

„... *Jak ukazuje každodenní život, s přechodem určité vědy od popisného stupně vývoje k metrickému stává se užívání matematicko-statistické theorie pro vědeckého pracovníka nutným. Následkem neobyčejně rychlého rozvoje statistické theorie a techniky se projevila v řadách odborných pracovníků, užívajících metod matematické statistiky, naléhavá potřeba sbírky nejdůležitějších tabulek vhodně upravených pro praxi. Tyto tabulky mají býti pomůckou, která mu má jeho práci usnadnit. Je tomu tak i při výuce studentů, kde vypomáhaly zatím litografované tabulky. Bylo proto usneseno na první konferenci matematických statistiků, pořádané Československou akademií věd, aby litografované tabulky byly rozšířeny a vydány tiskem* ...“[1].

Od vydání Jankových tabulek vyšly u nás tiskem v podstatě jediné rozsáhlé statistické tabulky, jež k tisku připravili Likeš a Laga (1978). Ty jsou však již dávno beznadějně vyprodány, nemluvě o tabulkách Jankových. Dlouho připravované nové vydání Jankových tabulek, které připravoval Josef Machek za pomoci Honzy Hanouska, patrně nikdy nevyjde. Osobně si myslím, že to je škoda. Cílem tohoto krátkého příspěvku je proto poukázat na některé možnosti, které pokrok výpočetních metod přináší i pro tuto oblast. Nesnažím se podat přesný návod na výpočet všech tabulek z citované knihy. Především jsem vypustil popis výpočtu tabulek založených na rozpětí, tj. tabulek 19–27. Hlavním důvodem pro mne byla skutečnost, že se nyní téměř nepoužívají. Na druhé straně jsem zařadil poznámky o výpočtu kritických hodnot některých více či méně běžných postupů, jež se v současné době používají častěji. Především pak odstavec o výpočtu přesných kritických hodnot pro neparametrické postupy. Čtenář může samozřejmě namítnout, že v leckterém statistickém programovém systému jsou výpočty potřebných kritických hodnot k dispozici. To je pravda, ale zpravidla jsou velmi úzce vázány na

[1] Citováno věrně dle Janko (1958).

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

implementované metody. V okamžiku, kdy potřebujeme jejich byť jenom malou modifikaci, rychle narážíme na problémy, a to problémy tím vážnější čím je naše modifikace nestandardnější.

VÝPOČET KVANTILŮ A KRITICKÝCH HODNOT VYBRANÝCH STANDARDNÍCH ROZDĚLENÍ

Se systémem *MATHEMATICA* uživatel obdrží též poměrně rozsáhlou knihovnu podprogramů napsaných ve formě makropříkazů v jejím jazyce. Mezi makropříkazy jsou k dispozici v direktoráři *dataanal*[2] též podprogramy připravené speciálně pro statistiku. Z nich nás budou pro potřeby tohoto příspěvku zajímat především následující dva, totiž soubory *continuo.m* a *discrete.m*. Podívejme se nejprve stručně na jejich obsah.

Typ rozdělení	Název v MATHEMATICE	v.
Beta	BetaDistribution	1.2
Cauchyho	CauchyDistribution	1.2
Exponenciální	ExponentialDistribution	1.2
Extrémních hodnot	ExtremeValueDistribution	1.2
Fisherovo – Snedecorovo F	FRatioDistribution	1.2
Gamma	GammaDistribution	1.2
χ	ChiDistribution	1.2
χ^2	ChiSquareDistribution	1.2
Laplaceoveo	LaplaceDistribution	1.2
Logaritmicko normální	LogNormalDistribution	1.2
Logistické	LogisticDistribution	1.2
Normální	NormalDistribution	1.2
Poloviční normální	HalfNormalDistribution	1.2
Raleyghovo	RayleighDistribution	1.2
Rovnoměrné	UniformDistribution	1.2
Studentovo t	StudentTDistribution	1.2
Weibullovo	WeibullDistribution	1.2
Necentrální χ^2	NoncentralChiSquareDistribution	2.0
Necentrální Studentovo t	NoncentralFRatioDistribution	2.0
Necentrální Fisherovo – Snedecorovo F	NoncentralStudentTDistribution	2.0

Tabulka 1. Přehled spojitých rozdělení jež jsou k disposici v systému *MATHEMATICA* v knihovně *continuo.m*.

[2] Celý tento příspěvek vychází z verze 1.2, není-li uvedeno jinak. Je dobré však mít na paměti, že počínajíc verzí 2.0 byly použity nové názvy pro jména některých direktorářů, knihoven či funkcí.

Typ rozdělení	Název v Mathematice	v.
Binomické	BernoulliDistribution	1.2
Beta binomické	BetaBinomialDistribution	1.2
Beta Pascalovo	BetaPascalDistribution	1.2
Binomické	BinomialDistribution	1.2
Diskrétné rovnoměrné	DiscreteUniformDistribution	1.2
Diskrétní Weibullovo	DiscreteWeibullDistribution	1.2
Geometrické	GeometricDistribution	1.2
Hypergeometrické	HypergeometricDistribution	1.2
Záporně binomické	NegativeBinomialDistribution	1.2
Poissonovo	PoissonDistribution	1.2
Diskrétní logaritmické	LogSeriesDistribution	2.0

Tabulka 2. Přehled diskrétních rozdělení jež jsou k disposici v systému *MATHEMATICA* v knihovně *discrete.m*.

Pro každé z těchto rozdělení jsou (zpravidla) k dispozici následující charakteristiky, jež si ukážeme na příkladě χ^2 rozdělení.

```
Domain[ChiSquareDistribution[___]] := {0, Infinity}
Density[ChiSquareDistribution[n_], x_] :=
        x^(n/2-1) / (Exp[x/2] Sqrt[2]^n Gamma[n/2])
CumulativeDensity[ChiSquareDistribution[n_], x_] :=
        GammaRegularized[n/2, 0, x/2]
Mean[ChiSquareDistribution[n_]] := n
Variance[ChiSquareDistribution[n_]] := 2 n
StandardDeviation[ChiSquareDistribution[n_]] :=
        Sqrt[2 n]
Skewness[ChiSquareDistribution[n_]] := 2 Sqrt[2/n]
Kurtosis[ChiSquareDistribution[n_]] := 3 + 12/n
KurtosisExcess[ChiSquareDistribution[n_]] := 12/n
CharacteristicFunction[ChiSquareDistribution[n_], t_] :=
        (1 - 2 I t)^(-n/2)
Quantile[ChiSquareDistribution[n_], q_] :=
        2 InverseGammaRegularized[n/2, 0, q]
Random[ChiSquareDistribution[_]] :=
        2 InverseGammaRegularized[n/2, 0, Random[]]
```

Je okamžitě vidět, že s jejich pomocí můžeme počítat nejen hodnoty kvantilů, nýbrž i hustot a distribučních, resp. charakteristických, funkcí. Je samozřejmostí, že všechny tyto funkce můžeme též nejen graficky zobrazovat, ale i animovat. Funkce *Random* nám navíc umožňuje z daného rozdělení též generovat.

Výpočet tabulek kvantilů všech těchto rozdělení, a tím pádem i různých typů kritických hodnot, je možné v zásadě zrealizovat pomocí jednoduchých analogií následujícího makropříkazu pro výpočet tabulky kvantilů rozdělení χ^2_n. Toto ukázka je příkladem toho, kdy naše makro netvoří tzv. *Package*.

```
TabulkaKvantiluChiSquare[SeznamN_List, SeznamAlpha_List]:=
Block[{Vysledek={}, MujKvantil, n, q, x, AplikaceNaSeznam,
 VypocetVlastnichHodnot, MaticeArgumentu, PridejHorniSloupec,
 PridejPrvniRadek},
 MujKvantil[n_,q_]       := Quantile[ChiSquareDistribution[n],q];
 AplikaceNaSeznam[x_]    := MujKvantil[Apply[Sequence,x]];
 MaticeArgumentu          = Outer[List,SeznamN,SeznamAlpha];
 VypocetHodnot            = N[Map[AplikaceNaSeznam,MaticeArgumentu,{2}]];
 PridejHorniRadek         = Partition[Flatten[Prepend[Flatten
                            [VypocetHodnot],SeznamAlpha]],Length[
                            SeznamAlpha]];
 PridejPrvniSloupec       = Transpose[Partition[Flatten[Prepend[Flatten
                            [Transpose[PridejHorniRadek]], Prepend
                            [Flatten[SeznamN],Null]]],Length[SeznamN]+1]];
 Vysledek                 = TableForm[PridejPrvniSloupec]]
End[]
```

Předpokládejme, že tento text máme uložen v souboru *chisq_kv.m* v direktoráři *dataanal*. Následující výpis ukazuje způsob použití při výpočtu tabulky 10%, 80% a 90%ních kvantilů χ^2_n pro $n = 1, \ldots, 5$. Chceme-li počítat kvantily pro větší rozsahy, respektive pro jiné hodnoty α, zřejmým způsobem změníme volání v řádku `In[3]`. Podobně lze postupovat při výpočtu kritických hodnot.

```
In[1]:= <<dataanal/continuo.m          % načtení souboru continuo.m
In[2]:= <<dataanal/chisq_kv.m          % načtení souboru chisq_kv.m
In[3]:= TabulkaKvantiluChiSquare[Range[ 1, 5],{0.1, 0.8, 0.95}]
                       Null      0.1         0.8        0.95
                       1         0.0157908   1.64237    3.84146
                       2         0.210721    3.21888    5.99146
                       3         0.584374    4.64163    7.81473
                       4         1.06362     5.98862    9.48773
Out[3]//TableForm =    5         1.61031     7.28928    11.0705
```

Analogicky „poměrně“ snadno spočteme tabulky číslo 1, 2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 29, 30.

Ne vždy nás však zajímají pouze prosté hodnoty kvantilů. Mohou to být například hodnoty podílu $\chi^2_{1-\alpha}/n$ jako je tomu v tabulce číslo 6. Ale i zde je řešení jednoduché, stačí pouze zaměnit řádky

```
MujKvantil[n_,q_]       := Quantile[ChiSquareDistribution[n],q];
AplikaceNaSeznam[x_]    := MujKvantil[Apply[Sequence,x]];
```

za řádky

```
Tabulka6[n_,q_]         := Quantile[ChiSquareDistribution[n],1-q]/n;
AplikaceNaSeznam[x_]    := Tabulka6[Apply[Sequence,x]];
```

Podobným způsobem můžeme postupovat při výpočtu tabulek 15, 16, 28, 31, 35 a 36.

Všimněme si, že analogicky lze tabelovat hodnoty hustot, resp. distribučních funkcí, všech výše zmíněných rozdělení, respektive hodnoty určité jiné funkce, jako je tomu například v tabulkách číslo 17, 18, 32, 37, 38, 39 či 40. Jak čtenář snadno nahlédne, stačí funkce zadefinovat a použít podobný obrat jako v předchozím případě.

VÝPOČET PŘESNÝCH KRITICKÝCH HODNOT PRO NĚKTERÉ NEPARAMETRICKÉ POSTUPY

Jedním ze známých problémů neparametrické statistiky je výpočet přesných kritických hodnot pro jednotlivé procedury. Pro některé postupy existuje více či méně jednoduché explicitní vyjádření, například pro mediánový test lze použít tabulky hypergeometrického rozdělení. Pro většinu postupů však takto jednoduché vyjádření nemáme a zpravidla jsme odkázáni na použití některého rekurentního postupu, např. dle Hájek a Šidák (1967). Na jedné straně je třeba říci, že rekurentní postupy je možné poměrně snadno v *MATHEMATICE* implementovat. Praktická zkušenost však v tomto případě ukázala, že se nejedná o přístup v převážné většině případů optimální. Důvodem je fakt, že vnitřní procedury systému zprostředkovávající rekurze jsou velice náročné na paměť a značně pomalé. V řadě případů však můžeme vyjít z „vytvořujících funkcí“ příslušných statistik a celou záležitost řešit kombinatoricky. I tento přístup v praxi naráží velmi rychle na problémy, především díky exponenciální složitosti odpovídajících algoritmů. Nicméně zvláště pro malé rozsahy výběru se jedná o postup velmi efektivní a názorný, který nám umožňuje relativně jednoduché výpočty hledaných kritických hodnot i v případě výskytů shodných pozorování. Následuje výpis makropříkazu, vycházející z tohoto přístupu, který slouží pro výpočet 5% kritických hodnot dvouvýběrového Wilcoxonova testu. Procedura je v ukázce aplikována na výpočet kritických hodnot pro $m = n = 24$.

```
Wi[m_,n_,alpha_]:=
Block[{ res = {}, res1, res2 = {}, junk, i, index, k = 0, koef,
    mpn = m + n},
    koef  = CoefficientList[Expand[Product[1+x^i y,{i,1,mpn}]],y];
    koef  = Drop[koef,1];
```

```
    junk  = koef[[m]];
    Do[k  = k + 1; AppendTo[res2,Exponent[junk[[k]],x]],
            {i,1,Length[junk]}];
    res1  = Select[CoefficientList[junk,x],Positive];
    junk  = N[res1/Apply[Plus,res1]];
    junk  = N[Accumulate[Plus,junk]];
    index = Length[Select[junk-alpha,Negative]];
    res   = {{res2[[index]],junk[[index]]},{res2[[index+1]],
            junk[[index+1]]}}
]
End[]

In[4]:= TableForm[Wi[24,24,.05]]
                      507        0.048602
Out[4]//TableForm= 508        0.050753
```

VÝPOČET NĚKTERÝCH SLOŽITĚJŠÍCH KRITICKÝCH HODNOT

Během své přednášky na ROBUSTu'92 jsem jako jednu z motivací, která mi k celé této záležitosti přivedla, uváděl domácí úlohu kterou mi svého času zadal Marek Malý. Jednalo se o výpočet kritických hodnot pro test hypotézy o rovnosti rozptylů několika výběrů navržený Davidem v šedesátých letech. Abych však zbytečně neprodlužoval svůj příspěvek popisem Markova problému, bez nějž by ukázka byla značně nesrozumitelnou, vybral jsem si pro ilustraci makropříkaz pro výpočet kritických hodnot pro proceduru podrobně popisovanou v tomto sborníku Danielu Jaruškovou. Procedura je příkladem jednoduchého *Package*.

```
Daniela::usage="Daniela[n,t,alpha] slouží pro výpočet kritických
hodnot pro příspěvek Daniely Jaruškové ve sborníku z ROBUSTu'92.
Vstupními parametry je rozsah výběru  n, počáteční odhad kritické
hodnoty  t  a zvolená hladina významnost  alpha. Výstupem je
hledaná kritická hodnota a skutečně dosažená hladina významnosti.
Pro větší čitelnost procedury je použito půlení intervalů."

Begin["Private`"]

Daniela[ n_, t_, alpha_] :=
Block[{ res={}, aknt, bknt, junk, k, tt = t, ttold, x, delta = 1,
    eps = 0.0001, eps1 = 10^(-10), eps2 = 1-10^(-10), k = 1,
    sumanew = 0, sumaold = 0},
    While[Abs[sumanew - alpha] > eps,
```

```
        sumanev = 0;
        While[k < n,
           aknt = x /. FindRoot[-(k Log[x (n/k)] +
                 (n-k) Log[(1-x) (n/(n-k))])==(tt/2), {x,eps1},
                  MaxIterations->30];
           bknt = x /. FindRoot[-(k Log[x (n/k)] +
                 (n-k) Log[(1-x) (n/(n-k))])==(tt/2), {x,eps2},
                  MaxIterations->30];
           junk = 1 - (BetaRegularized[bknt,k,n-k]-
                  BetaRegularized[aknt,k,n-k]);
           sumanev = sumanev + junk;
           k = k + 1];
        k = 1;
        ttold = tt;
        If [(sumanev - alpha) (sumaold - alpha) < 0,
             delta = delta/2];
        If [ sumanev < alpha, tt = tt - delta, tt = tt + delta];
        sumaold = sumanev]
    AppendTo[res,N[ttold]];
    AppendTo[res,sumanev]
]
End[]
```

Použití je jednoduché. Stačí makro načíst pomocí << a zavolat

```
In[5]:= Daniela[10, 8, .05]
```

Jako výsledek dostaneme

```
Out[5]={8.65625, 0.050002}
```

JANKOVY TABULKY NEJSOU POUZE ČÍSLA

Jankovy tabulky skutečně nejsou pouze tabulky čísel doprovázené ve zvláště zapeklitých případech nomogramy. Jejich nedílnou součástí je rozsáhlý průvodní text doprovázený obrázky. Vždy jsem litoval, že ilustrativních obrázků v textu není více. Jak si mi však nejednou Josef Machek stěžoval, problém nebyl pouze s jejich nesnadnou přípravou. Ještě větším problémem byla skutečnost, že nakladatelství Academia se zařazení obrázků velmi bránilo, a to i pro nové vydání. Hlavním argumentem bylo tvrzení, že obrázky jim přidělávají mnoho starostí a podstatně zdražují výrobní náklady.

Myslím si však, že v současné době by toto nemusilo být až takovým problémem, neboť tabulky by patrně byly vysázeny v TeXu.[3] Při doplňování obrázků do textu by šlo

[3] Pozor, *MATHEMATICA* umožňuje výstupy formátovat ve zdrojovém kódu TeXu.

využít grafické možnosti jež systém *MATHEMATICA* nabízí. Nejjednodušší cestou se mi zdá export obrázků do *PostScriptu*, neboť tento výstup již lze „poměrně snadno“ dle potřeby do TeXu integrovat. Použití *PostScriptu* není samozřejmě jedinou možností jak obrázky z *MATHEMATIKY* integrovat do TeXového zdrojového textu. Jedná se však o přístup nejpřirozenější, neboť veškerý grafický výstup v systému *MATHEMATICA* je interně realizován právě přes *PostScript*.

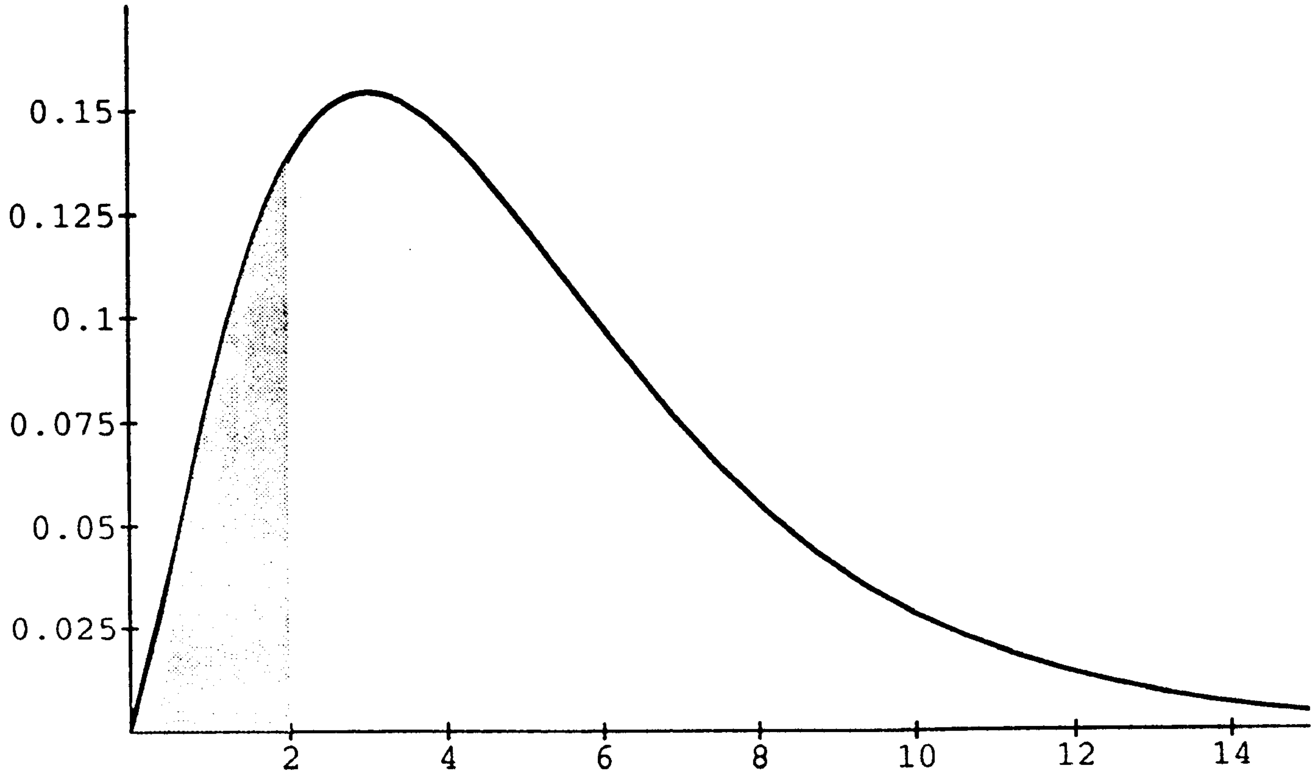
Následující makro umožňuje kreslit analogie obrázků číslo 4 *a* a 4 *b* z Jankových tabulek. Všimněme si, že nejvíce práce si vyžádalo „šrafování“, nikoliv kreslení vlastní hustoty.

```
Obrazek:usage="Obrazek[StupneVolnosti_, xmax_, ymax_,
OdkudVypln_, KamVypln_] umoznuje kreslit "


Begin["Private`"]


Obrazek[StupneVolnosti_, xmax_, ymax_, OdkudVypln_, KamVypln_]:=
Block[{AplikaceNaSeznam, junk, MojeHustota, polygon, srafovane,
      text, vypln, ZakladniObrazek},
 ZakladniObrazek          = Plot[Density[ChiSquareDistribution
                            [StupneVolnosti],x],{x,0,xmax},
                            PlotRange->{{0,xmax},{0,ymax}}];
 MojeHustota[n_,x_]      := Density[ChiSquareDistribution[n],x];
 AplikaceNaSeznam[y_]    := MojeHustota[Apply[Sequence,y]];
 vypln                    = Range[OdkudVypln,KamVypln,.01];
 junk                     = Outer[List,{StupneVolnosti},vypln];
 polygon                  = Join[{{OdkudVypln,0}},Transpose[{vypln,
                            Flatten[N[Map[AplikaceNaSeznam,junk,
                            {2}]]]}],{{KamVypln,0}}];
 srafovane                = Graphics[{GrayLevel[.9],Polygon[polygon]},
                            PlotRange->{{0,xmax},{0,ymax}}];
 text                     = Graphics[{Text[FontForm["Zde bude nadpis",
                            "Italic",24],{xmax/2,ymax-(ymax/10)}]}];
 Show[ZakladniObrazek,srafovane,text]]
End[]
```

Použití je opět velmi jednoduché. Po načtení souboru *dataanal.m* a souboru obsahující náš makropříkaz stačí napsat

```
In[7]:= Obrazek[5,15,.175,0,2]
```

abychom dostali obrázek 1.


0.15
0.125
0.1
0.075
0.05
0.025
2
4
6
8
10
12
14

NĚKOLIK SLOV ZÁVĚREM

Závěrem bych chtěl říci, že uvedená řešení nejsou vázána výhradně na program *MATHEMATICA*. Jsem přesvědčen, že řešení analogická těm jež jsou uvedena v tomto příspěvku by bylo možné napsat bez větších potíží při použití programů typu *AXIOM*, *MAPLE* či *REDUCE* apod. Hlavním důvodem k mé volbě byla především skutečnost, že v uplynulé době jsem měl k disposici právě *MATHEMATIC*u. Vedle toho je dobré si uvědomit, že výše popsané procedury byly již dávno implementovány v některém z klasických programovacích jazyků a s jejich pomocí tabelovány. A nejen to, tabulky z nejpoužívanějších byly již na počátku tohoto století spočteny kalkulátory na kličku a to s přesností, jež se dodnes těm počítačově zpracovaným vyrovná. Na druhé straně jsem však přesvědčen, že mnou ilustrovaný přístup k výpočtu kritických hodnot může v řadě případů statistikovi významně usnadnit práci.

REFERENCES

Hájek J., Šidák Z., *Theory of rank tests*, Academia, Praha, 1967.

Janko J., *Statistické tabulky*, Nakladatelství Československé akademie věd, Praha, 1958.

Likeš J., Laga J., *Základní statistické tabulky*, SNTL, Praha, 1978.

Maeder R., *Programming in MATHEMATICA*, Addison-Wesley, Redwood City, California, 1990.

Streitberg B., Tismer Ch., *Computer algebra and statistics*, Handout for Summer School on Computational Aspects of Model Choice, MFF UK, Praha, 1991.

Wolfram S., *MATHEMATICA. A System for Doing Mathematics by Computer*, Addison-Wesley, Redwood City, California, 1988.