Computer algebra and rank statistics.

Alessandro Di Bucchianico

HCM Workshop
Coimbra
November 5, 1997

Outline

- General remarks on rank statistics
- What is computer algebra?
- Critical values of rank statistics
- Moments of the Mann-Whitney statistic
- Conclusions

General remarks on rank statistics

- Practical problems
  - tables (limited, errors, not exact,...)
  - limited availability in statistical software
  - procedures in statistical software based on asymptotics

- Mathematical problems
  - in general no closed expression for distribution function
  - direct enumeration only feasible for small sample sizes
  - recurrences time-consuming

What is computer algebra?

```
In[1]:= Expand[(1+x)*(1+x^2)]

                    2    3
Out[1]= 1 + x + x  + x

In[2]:= <<DiscreteMath`RSolve`

In[3]:= n /: Info[n] = (n >= 0);

In[4]:=  SeriesTerm[1/(1-x^4),{x,0,n}]

                 n                    n/2
        1   (-1)                 (-1)
Out[4]= - + ----- + If[Even[n], -------, 0]
        4     4                     2
```

Case study: Mann-Whitney statistic

independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$

continuous distribution functions F, G resp. (hence no ties with probability one)

order the pooled sample from small to large

Wilcoxon: $W_{m,n} = \sum_{i=1}^{m} \text{rank}(X_i)$

Mann-Whitney: $M_{m,n} = \# \{(i,j) \mid (Y_j < X_i)\}$

$W_{m,n} = M_{m,n} + \frac{1}{2} m(m+1)$

What is the distribution of $M_{m,n}$ under $H_0 : F = G$?

Mann-Whitney: generating function

$$\sum_{k=0}^{mn} P(M_{m,n} = k)\, x^k = \frac{1}{\binom{m+n}{n}} \frac{\prod_{i=n+1}^{m+n} (1 - x^i)}{\prod_{j=1}^{m} (1 - x^j)}$$

```
In[1]:= Expand[Simplify
    [Product[1-x^i,{i,3,5}]/
     Product[1-x^j,{j,1,3}]/Binomial[5,3]]]

                     2     3     4     5     6
        1     x     x     x     x     x     x
Out[1]:= -- + -- + -- + -- + -- + -- + --
        10    10    5     5     5     10    10

In[2]:= CoefficientList[%,x]

        1    1    1   1   1   1    1
Out[2]= {--, --, -, -, -, --, --}
        10   10   5   5   5   10   10
```

Mathematica implementation

```
MannWhitneyCumFreq[m_,n_] :=
Module[{x,i,j},
FoldList[Plus,0,
CoefficientList[Expand[Factor[
Product[1-x^i,{i,n+1,m+n}]/
Product[1-x^j,{j,1,m}]/Binomial[m+n,n]]],x]
       ]
       ]
```

```
MannWhitneyLeftCritValue[m_,n_,alpha_] :=
Module[{v= Length[Select[
MannWhitneyCumFreq[m,n], # <= alpha&]]-2},
If[NonNegative[N[value]],v,"*"]
      ]
```

```
MannWhitneyLeftSigValue[m_,n_,k_] :=
MannWhitneyCumFreq[[k+2]]
```

Computatations (Pentium 133 MHz)

Example 1:

exact: $P(M_{5,5} \leq 4) = \frac{1}{21} \approx 0.0476$

computing time: 0.05 sec (gen. function: degree 25)

$P(M_{5,5} \leq 4) \approx P(Z \leq -1.77) = 0.0384$

Example 2:

exact: $P(M_{20,20} \leq 138) = 0.0482$ (rounded)

computing time: 8.5 sec (gen. function polynomial of degree 400)

$P(M_{20,20} \leq 138) \approx P(Z \leq -1.67) = 0.0475$

Asymptotics and exact calculations are both useful!

Other test statistics

Other examples of nonparametric statistics with closed form for generating function:

- Wilcoxon signed rank statistic
- Kendall rank correlation statistic
- Kolmogorov one-sample statistic
- Smirnov two-sample statistic (for many combinations of sample sizes)
- Jonckheere-Terpstra statistic

Consult the combinatorial literature (e.g.,

'Combinatorial chance' by Barton and David)!

What to do if there is no generating function?

Linear rank statistics

Linear rank statistics:

$$T = \sum_{\ell=1}^{m+n} a(\ell)\, Z_\ell$$

$Z_\ell = 1$ if ℓ^{th} order statistic in the pooled sample is an X-observation, and zero otherwise ($\ell = 1, \ldots, m+n$).

Streitberg & Röhmel 1986 (cf. Euler 1748):

$$\sum_{m+n=N} \sum_k \binom{N}{m} \Pr(T_{m,n} = k)\, x^k\, y^m =$$

$$(1 + x^{a(1)}\, y) \ldots (1 + x^{a(N)}\, y),$$

where $T_{m,n} = \sum_{\ell=1}^{m+n} a(\ell)\, Z_\ell$.

Branch-and-bound algorithm (Van de Wiel)

Moments of Mann-Whitney statistic

Mann and Whitney (1947) calculated 4th central moment of $M_{m,n}$.

Fix and Hodges (1955) calculated 6th central moment of $M_{m,n}$.

Computations are based on recurrences.

Can we improve?

'21th century' solution: computer algebra and generating functions.

Computing moments of $M_{m,n}$

$$E\,X_{(n)} = \left(\frac{d^n}{dx^n} \sum_k P(X=k)\,x^k \right) \Bigg|_{x=1}$$

recompute $E(M_{m,n})$ (following René Swarttouw)

$$G_{m,n}(x) := \frac{(1-x^{n+1})\dots(1-x^{n+m})}{(1-x)\dots(1-x^m)}$$

$$\log G_{n,m}(x) = \sum_{k=1}^{m} \log(1-x^{n+k}) - \sum_{k=1}^{m} \log(1-x^k)$$

$$\frac{d}{dx} \log G_{n,m}(x) = \sum_{k=1}^{m} \frac{k\,x^{k-1}}{1-x^k} - \sum_{k=1}^{m} \frac{(n+k)\,x^{n+k-1}}{1-x^{n+k}}$$

$$\frac{\frac{d}{dx} G_{n,m}(x)}{G_{n,m}(x)} = \frac{d}{dx} \log G_{n,m}(x) =$$

$$\sum_{k=1}^{m} \frac{k\,x^{k-1}(1-x^{n+k}) - (n+k)\,x^{n+k-1}\,(1-x^k)}{(1-x^k)\,(1-x^{n+k})}$$

Computing $E(M_{m,n})$: continued

$G_{m,n}$ is a polynomial, hence $G'_{m,n}(1) = \lim_{x\to 1} G'_{m,n}(x)$

Fact: $\lim_{x\to 1} G_{m,n}(x) = \binom{n+m}{n}$

Hence, it remains to calculate for $1 \le k \le m$:

$$\lim_{x\to 1} \frac{k\,x^{k-1}\,(1-x^{n+k}) - (n+k)\,x^{n+k-1}(1-x^k)}{(1-x^k)\,(1-x^{n+k})} =$$

$$\lim_{x\to 1} \frac{k(1-x^n) - nx^n(1-x^k)}{(1-x^k)\,(1-x^{n+k})}$$

After some further simplifications:

$$\lim_{x\to 1} \frac{k\left(1+\ldots+x^{n-1}\right) - n\left(x^n+\ldots+x^{n+k-1}\right)}{k\,(n+k)\,(1-x)}$$

L' Hôpital's rule yields that the limit equals

$$\frac{-1}{k(n+k)}\left(\frac{1}{2}kn(n-1) - n(kn+\frac{1}{2}k(k-1)\right) = \frac{n}{2}$$

Mathematica implementation

It is tedious to perform these computations by hand.

Alternative: Mathematica procedure

```
LogG[k_,n_,x_] := Log[1-x^(n+k)] - Log[1-x^k]

DerivativeOfLogG[r_] := Module[{j,der},
Sum[Simplify[r! Coefficient[
Normal[Series[LogG[k,n,x],{x,1,r+1}]],x-1,r]],
{k,1,m}]]

FactorialMoments[r_] := Module[{j,equations},
equations := Table[
ReplaceAll[Simplify[Together[D[Log[G[z]],{z,j}]]],
G[z]->1]== DerivativeOfLogG[j],{j,1,r}];
Flatten[
ReplaceAll[Table[D[G[z],{z,i}],{i,1,r}],
Solve[equations,Table[D[G[z],{z,i}],{i,1,r}]]]
       ]
                                    ]
```

8th central moment of $M_{m,n}$

$$\frac{mn(1+m+n)}{34560}\,P(m,n) \text{ with } P(m,n) =$$

$$-96\,n+96\,n^2+240\,n^3-240\,n^4-432\,n^5-144\,n^6+$$

$$\left(-96+192\,n+224\,n^2-540\,n^3-100\,n^4+780\,n^5+\right.$$

$$\left.404\,n^6\right)m+\left(96+224\,n-600\,n^2-200\,n^3+\right.$$

$$\left.900\,n^4-48\,n^5-420\,n^6\right)m^2+(240-540\,n-$$

$$\left.200\,n^2+1095\,n^3-395\,n^4-735\,n^5+175\,n^6\right)m^3+$$

$$\left(-240-100\,n+900\,n^2-395\,n^3-630\,n^4+\right.$$

$$\left.525\,n^5\right)m^4+\left(-432+780\,n-48\,n^2-735\,n^3+\right.$$

$$\left.525\,n^4\right)m^5+\left(-144+404\,n-420\,n^2+175\,n^3\right)m^6$$

Conclusions

1. generating functions are also useful in non-parametric statistics

2. computer algebra is a natural tool for mathematicians

3. asymptotics and exact calculations complement each other

Topics under investigation

- tests for censored data
- power calculations
- analysis of variance
- Spearman's rho (next talk)
- multimedia/ World Wide Web implementation

References

- A. Di Bucchianico, Combinatorics, computer algebra and the Wilcoxon-Mann-Whitney test, submitted.

- B. Streitberg and J. Röhmel, Exact distributions for permutation and rank tests: An introduction to some recently published algorithms, Stat. Software Newsletter 12 (1986), 10-18.

- M.A. van de Wiel, Exact distributions of nonparametric statistics using computer algebra, Masters' Thesis, TUE, 1996.

- M.A. van de Wiel and A. Di Bucchianico, The exact distribution of Spearman's rho, in progress.

- M.A. van de Wiel, A. Di Bucchianico and P. van der Laan, Exact distributions of nonparametric test statistics using computer algebra, submitted.