The Jackknife

Resources:
Shao, Tu: The Jackknife and Bootstrap
Efron: The Jackknife, Bootstrap and other resampling plans

wjack.pdf Formulae for weighted jackknife
jackcorr2.pdf Sketch of an idea for improvement

Workhorse for ADMIXTOOLS as a way to estimate standard error in the presence of LD.

Simple jackknife. Estimate stanndard error. Example mean

We have a probability distribution F and want to estimate a parameter θ.
Estimator $T_n = \hat{\theta}(x_1, \ldots, x_n)$.
Want to know bias and standard error of T_n.

Bias Correction

Define

$$\begin{aligned} T_{n-1,i} &= \hat{\theta}(x_1, \ldots, x_{i-1}, x_{i+1} \ldots x_n) \\ \overline{T_n} &= (\sum_i T_{n-1,i})/n \\ b_{JACK} &= (n-1)(T_n - \overline{T_n}) \\ T_{JACK} &= T_n - b_{JACK} = T_n - (n-1)\overline{T_n} \\ &= 1/n \sum_i [nT_n - (n-1)T_{n-1,i}] \end{aligned}$$

Estimate of standard error and variance

$$
\begin{aligned}
Q &= \sum_i (T_{n,i} - \overline{T_n})^2 \\
\sigma^2_{JACK} &= \frac{n-1}{n} Q
\end{aligned}
$$

Alternative formulae:

$$
\begin{aligned}
P_i &= nT_n - (n-1)T_{n-1,i} \\
\overline{T_n} &= 1/n \sum_i P_i \\
\sigma^2_{JACK} &= \frac{1}{n(n-1)} \sum_i (P_i - \overline{T_n})^2
\end{aligned}
$$

P_i were called by Tukey *pseudovalues*

Exercises:

1. Show that the two formulae for jackknifed mean and variance are equivalent.
2. Let θ be the mean. So

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) = 1/n \sum_i x_i$$

the sample mean. Show

(a)

$$b_{JACK} = 0$$

(b)

$$\sigma^2_{JACK} = \frac{1}{n(n-1)})(x_i - \theta)^2$$

This last is of course the usual unbiased estimate of error variance

Bias correction in my experience not very relevant in Genetics
Efron: σ^2_{JACK} more reliable than b_{JACK}.

We don't discuss here exact conditions for jackknife standard error to be meaningful.
But for instance won't work for expected value of median.

The Block Jackknife

LD is of course the problem here. Nearby SNPs are not independent. My experience suggests that ignoring LD usually inflates Z -scores by a factor of about 2.

Idea is to divide genome into blocks. There is theory for the *weighted block jackknife (see wjack.pdf for formulae)*
Basic idea is to delete a block, and as with the ordinary jackknife compute an estimate from the deleted data. Estimate is weighted. Usual weight is number of SNPs used in block.

Almost all programs of ADMIXTOOLS use this.

Issues:

1. Ignores LD at block boundaries. Minor problem (?) usually.
   But some populations (recently admixed or deeply bottlenecked) may be trouble.

2. What should the weights be? Not always clear.
   Experience shows choice usually not crucial
   Want weights proportional to variance of estimate from block deletion.
   Number of SNPs in block usually reasonable.

Simple version: useful if you need to roll your own:
Chromosomal Jackknife.

I think users of ADMIXTOOLS should experiment more with blocksize.
(set blgsize: $< value >$
I'd be interested in cases where a modest change (factor of 2?)
makes big difference to Z-scores.

A possible improvement – the Fourier Jackknife

We expect that the jackknife estimates from each block should be uncorrelated, except at lag 1. Let the jackknifed estimate for block i be $j(i)$, with weight $w(i)$. Set m be mean of $j(i)$. Set

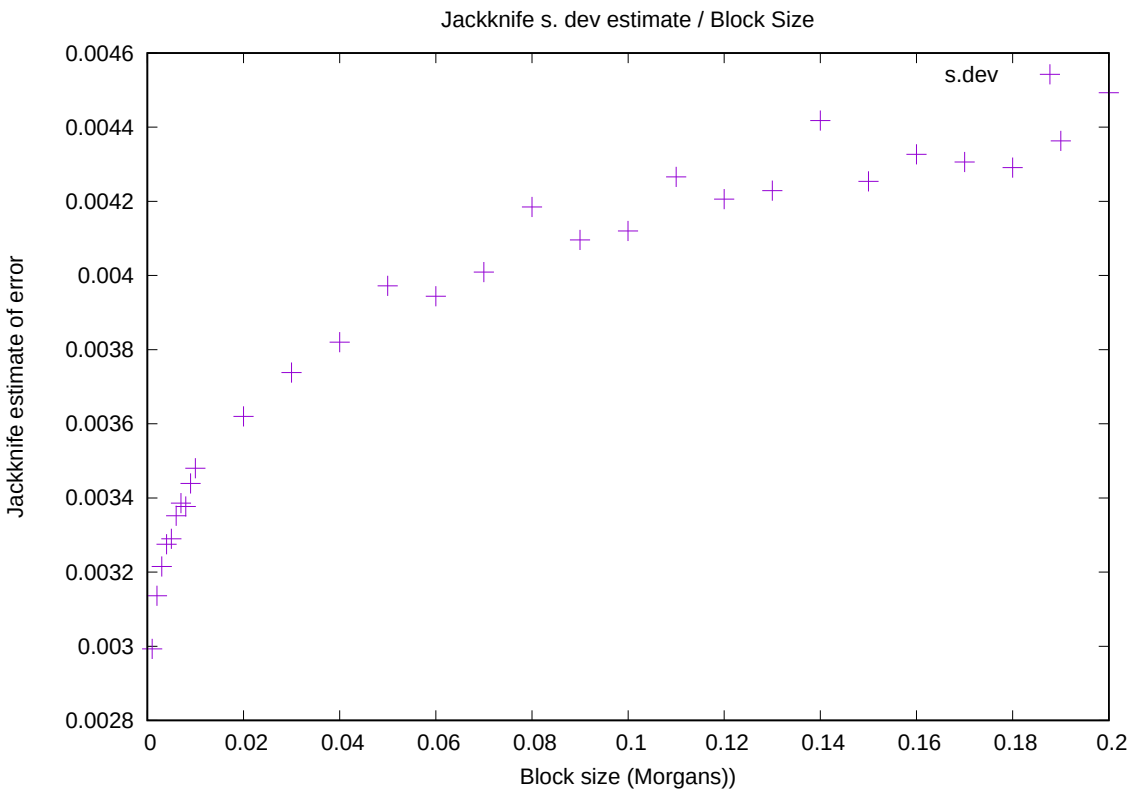
$$z(i) = \frac{j(i) - m}{\sqrt{w(i)}}$$

Lag-1 correlation $z(i)$ be ρ. Construct $c(k)$ so that circular convolution:

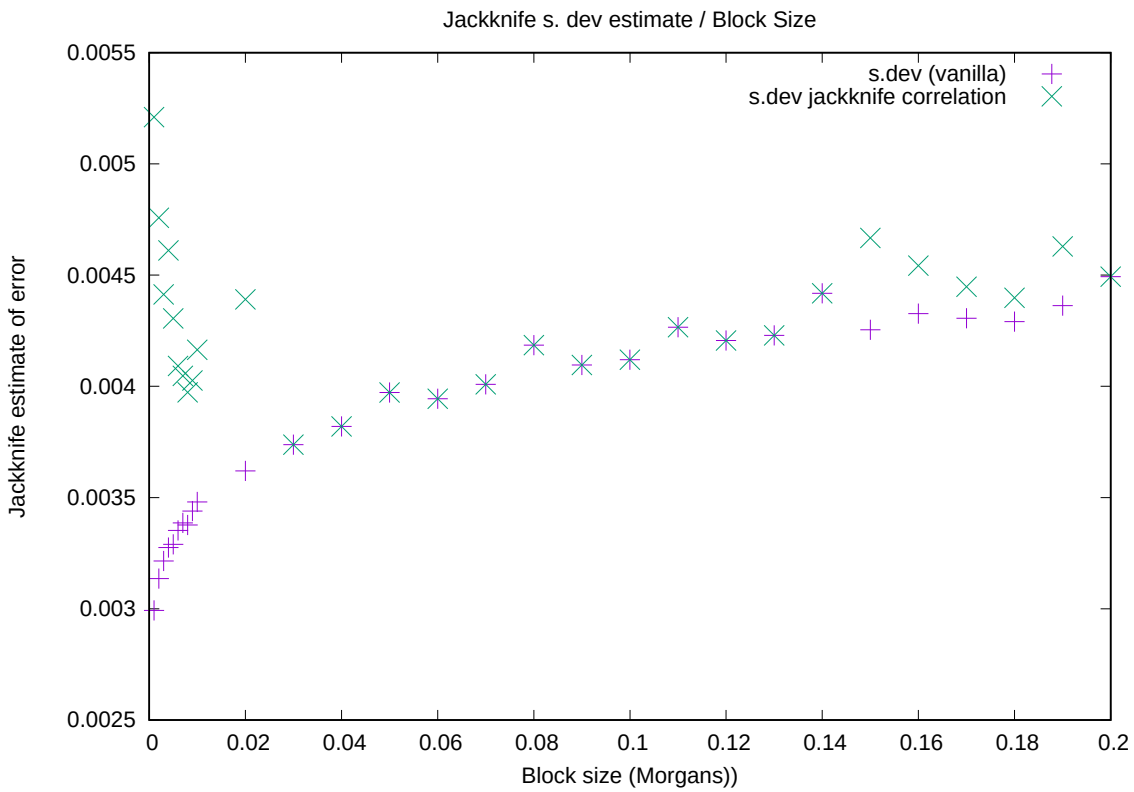
$$z'(i) = \sum_k c(k) z(i-k)$$

is (approximately) uncorrelated. Worked out in *jackcorr2.pdf*
(Uses Fourier Transform) Palamera lab. has experimented with this.

Standard error of $f_4(Chimp, Altai; Yoruba, Han)$



.


Jackknife s. dev estimate / Block Size
s.dev (vanilla)
s.dev jackknife correlation
Jackknife estimate of error
Block size (Morgans))
0.0055
0.005
0.0045
0.004
0.0035
0.003
0.0025
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2

Stay safe and well.
See you in the Fall!