# A short vignette showing the use of the **sumbin** package for evaluating the distribution of a sum of binomial random variables

Jerome V. Braun

## 1 Installing the sumbin R package

**sumbin** requires R version 3.1.1 or greater. The package was developed on the Windows operating system. The CRAN version of **sumbin** can be installed in an R console with:

```
install.packages("sumbin")
```

You can then load the package with:

```
library(sumbin)
```

You can read the help files and access this vignette again with:

```
?"sumbin-package"
?sumbin
help(package = "sumbin")
vignette("sumbin-vignette")
```

## 2 Yolo County grand jury dataset

Yolo County grand jurors are selected in a multi-step process from the population of eligible residents of Yolo county. Eligible residents must first apply (and are then part of the grand jury applicant pool). A set of grand jury nominees are chosen from this applicant pool upon vetting by the Court (and are then part of the grand jury nominee pool). Finally, a set of grand jurors are chosen from the grand jury nominee pool upon further examination by the Court.

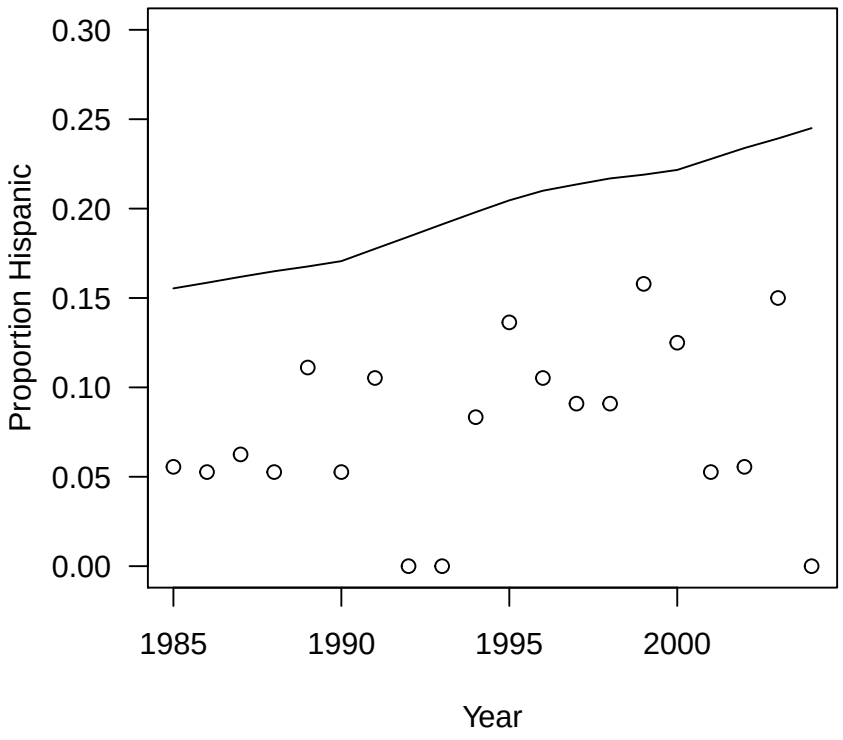
The composition of the Yolo County grand jury was challenged in court in 2007 on Sixth and Fourteenth Amendment grounds. Specifically, it was alleged that Hispanics were not fairly represented in the Yolo County grand jury.

The sample dataset includes information over time about the population of eligible residents of Yolo count, the composition of the grand jury nominee pools, and the grand jury pools. Estimation of the eligible population, the proportion of population that was Hispanic, and the composition of the grand jury nominee and grand jury pools is discussed in detail in [Braun, 2007].

You can load the included sample data set and access the included help using:

```
data(YoloGrandJury)
```

The following plot displays the proportion of Hispanics among the eligible population of Yolo County over time (solid line), as well as the composition of the Yolo County grand jury over time (open circles):




If the Yolo County grand jury is fair and representative, then we can consider the Hispanic composition of the grand jury each year to be a draw from a

binomial distribution. Both the size of the binomial distribution and the success probabilities vary from year to year in this case.

```
# Calculate the probability of a random draw of a Hispanic grand
# juror from the population of eligible residents of Yolo County.
prob <- with(YoloGrandJury$Eligible, Hispanic / Total)

# Set the number of trials each year for the grand jury pools.
size  <- YoloGrandJury$Juror$Total

# Finally, calculate the total number of Hispanic grand jurors
# seated over the time frame considered.
Hispanics <- sum(YoloGrandJury$Juror$Hispanic)
```

The total number of Hispanic grand jurors seated over the time frame of 20 years was 31 out of a total of 394 grand jurors seated. The distribution of this sum has no simple analytical form, since the success probabilities differ over time.

We may be interested in the probability under this model of seeing 31 or fewer Hispanics across the time frame. This can be obtained using **psumbin** by:

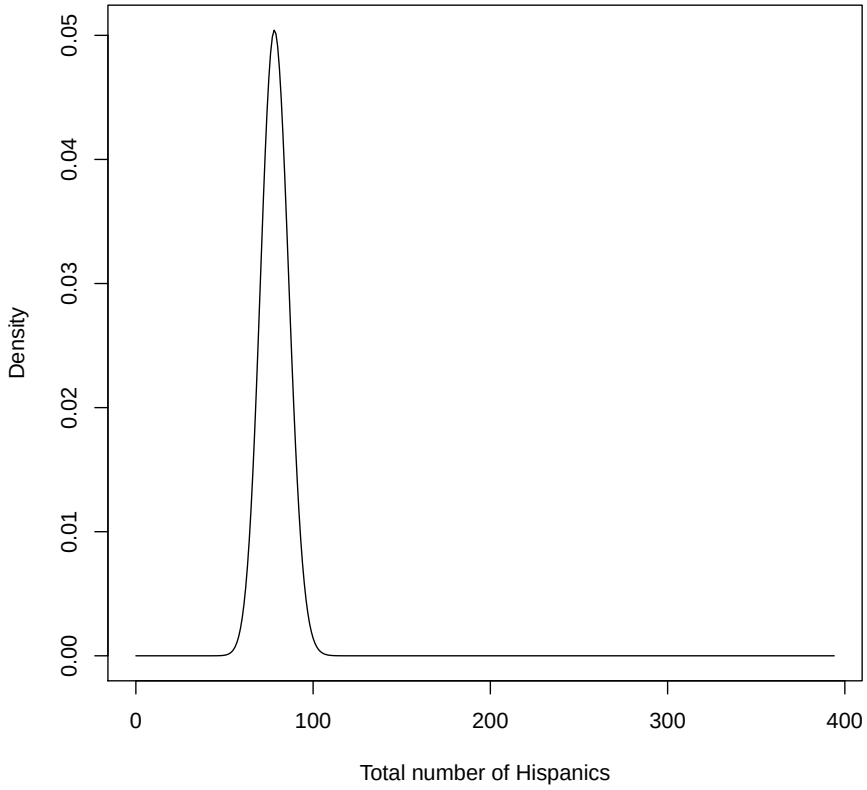
```
# The one-sided p-value under this model.
psumbin(Hispanics, size=size, prob=prob)

## [1] 2.649e-11
```

Note that this is an aggregate measure which disregards the actual configuration of the grand jury pools over time. However, it seems that there is some evidence that the representation over this time frame of Hispanics in the Yolo County grand jury is not representative of the underlying population of eligible residents. A more detailed analysis was carried out in [Braun, 2007].

Depending upon the situation, it might be of interest to plot the distribution of the total number of Hispanics expected under fair and representative selection using **dsumbin**:

```
plot(0:394, dsumbin(0:394, size=size, prob=prob), type='l',
     xlab='Total number of Hispanics', ylab='Density')
```




For simulation purposes, it might be useful to generate samples from the distribution using **rsumbin**. This could be done as follows:

```
# Generate a sample of size 1000 from the distribution.
rsumbin(1000, size=size, prob=prob)
```

# References

Braun, J. V. (2007). Exhibit C: Statistical and probabilistic analysis of representation of Hispanics and Asian & Pacific Islanders in the Yolo County grand jury from 1985 to 2004. *The People of the State of California vs. Michael Raquel, et al., Case No. 01-1577, Superior Court of the State of California in and for the County of Yolo.*