Results Notebook

Zipf

Read data

```
z <- read.delim('results.dat', header=TRUE)
head(z)
```

```
##                                  Language Alpha Sigma Xmin vsExp.R vsExp.P
## 1                                 Amarasi 1.316 0.074    1  -4.375   0.000
## 2 American English - Southern Michigan 1.406 0.062    1  -2.139   0.032
## 3                                 Amharic 1.421 0.060    1  -0.443   0.658
## 4                                  Arabic 1.387 0.064    1  -2.036   0.042
## 5                                Assamese 1.318 0.059    1  -4.339   0.000
## 6                      Australian English 1.439 0.065    1  -1.038   0.299
##   vsLN.R vsLN.P vsTPL.R vsTPL.P
## 1 -3.028  0.002  -9.215       0
## 2 -3.522  0.000  -6.543       0
## 3 -3.341  0.001  -4.604       0
## 4 -3.354  0.001  -6.456       0
## 5 -3.653  0.000 -10.125       0
## 6 -2.967  0.003  -4.824       0
```

plot alpha parameter

```
library(ggplot2)

p <- ggplot(z, aes(x=Alpha)) + geom_histogram() + xlim(0, 6) + theme_classic()
p <- p + geom_vline(xintercept=1, color="tomato")
p
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
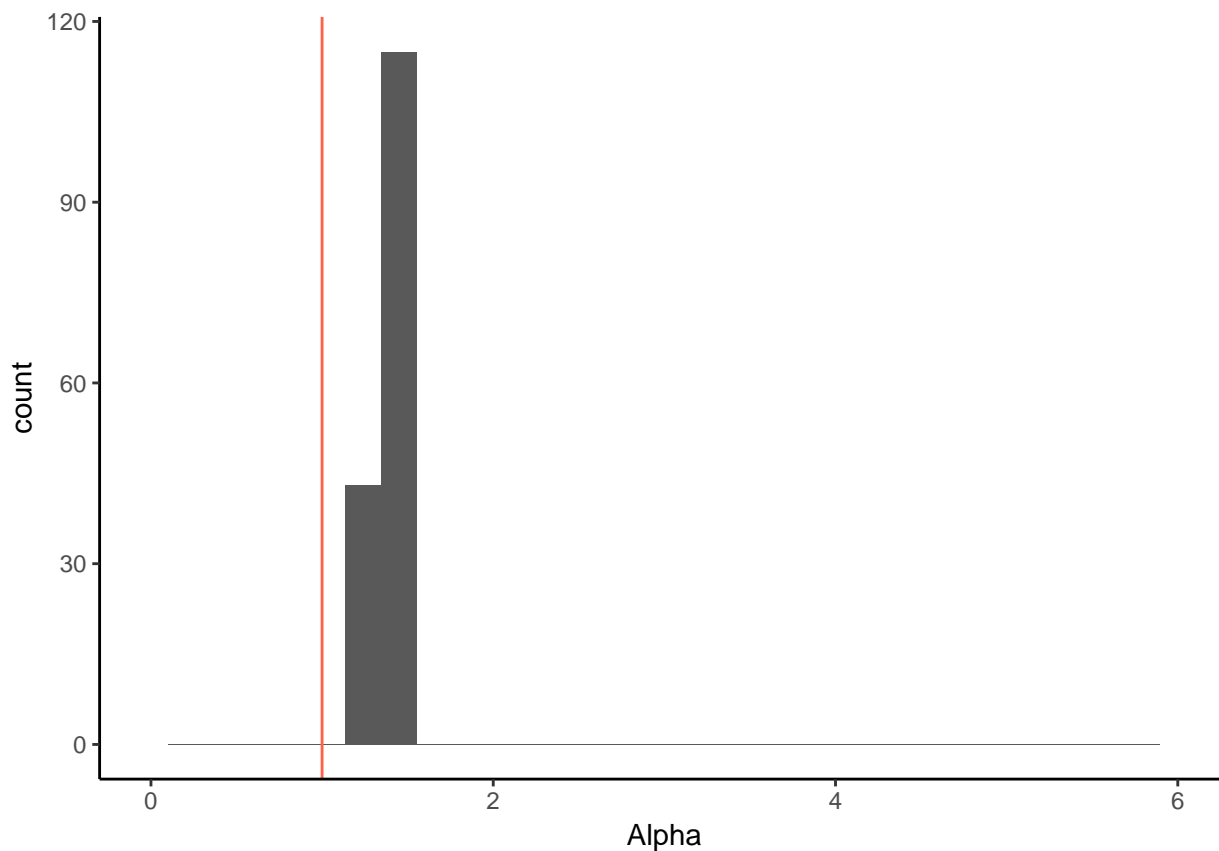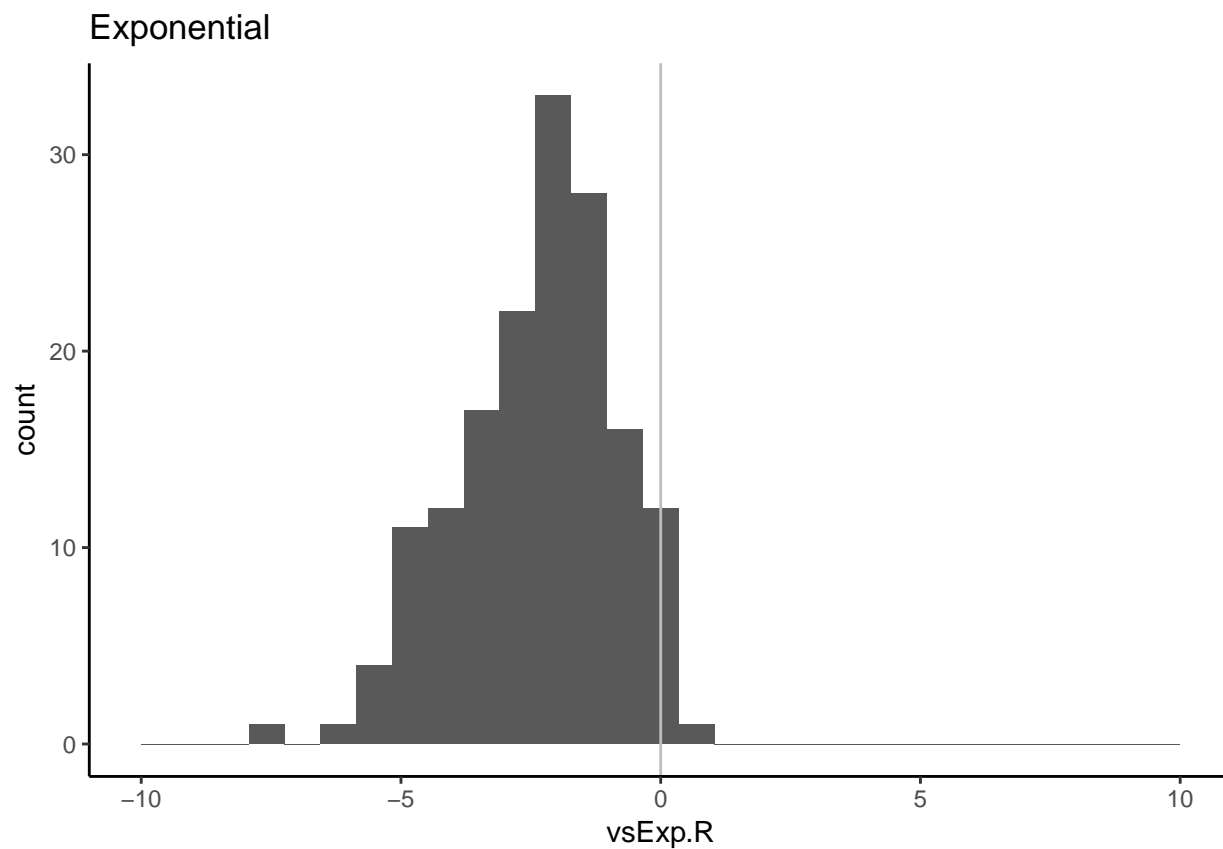
Hmm. Not much around expected Zipf alpha of 1.

compare models: Exponential

Negative means preference for the named model.

```
p <- ggplot(z, aes(x=vsExp.R)) + geom_histogram() + xlim(-10, 10) + theme_classic()
p <- p + geom_vline(xintercept=0, color="gray")
p <- p + ggtitle('Exponential')
p
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
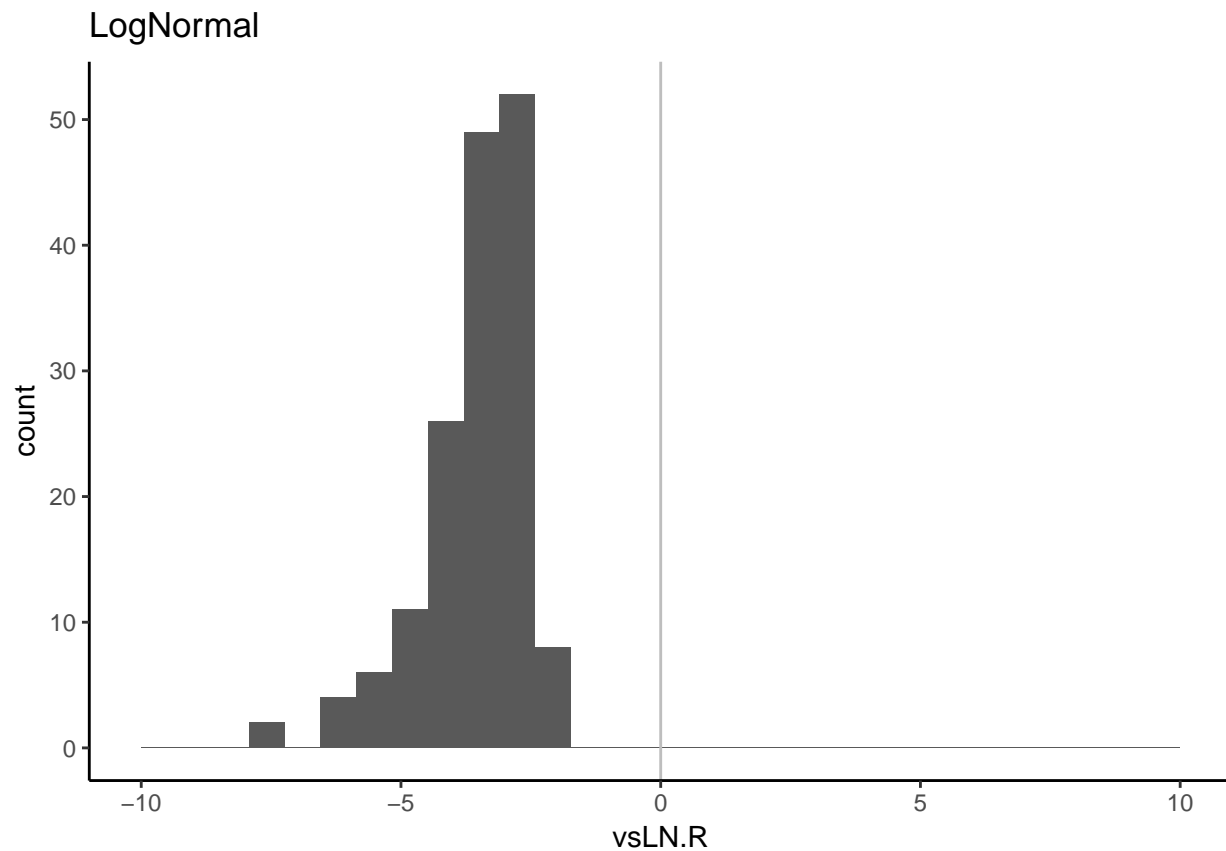
Exponential

compare models: Lognormal

Negative means preference for the named model.

```r
p <- ggplot(z, aes(x=vsLN.R)) + geom_histogram() + xlim(-10, 10) + theme_classic()
p <- p + geom_vline(xintercept=0, color="gray")
p <- p + ggtitle('LogNormal')
p
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

compare models: Truncated Power law

```r
p <- ggplot(z, aes(x=vsTPL.R)) + geom_histogram() + xlim(-10, 10) + theme_classic()
p <- p + geom_vline(xintercept=0, color="gray")
p <- p + ggtitle('LogNormal')
p
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 25 rows containing non-finite values (stat_bin).
```