Checking for Fit

In "Hype and Heavy Tails: A Closer Look at Data Breaches", 2015, the authors use the Kolmogorov-Smirnov test to make sure that their model selection should not be immediately rejected. They mention thick tails, but they don't use Student's t distribution for log breach size. They use the Gaussian. But the t distribution has a higher p-value in the K-S test.

Total Records

Although I can't find a definition for the fields in the data from privacyrights.org, it looks like query results have an equal number of "Total.Records" and "Records.Breached". The second has more prose inserted in the value, e.g., "1.2 million" and parenthetical notes. So I'm using total records below.

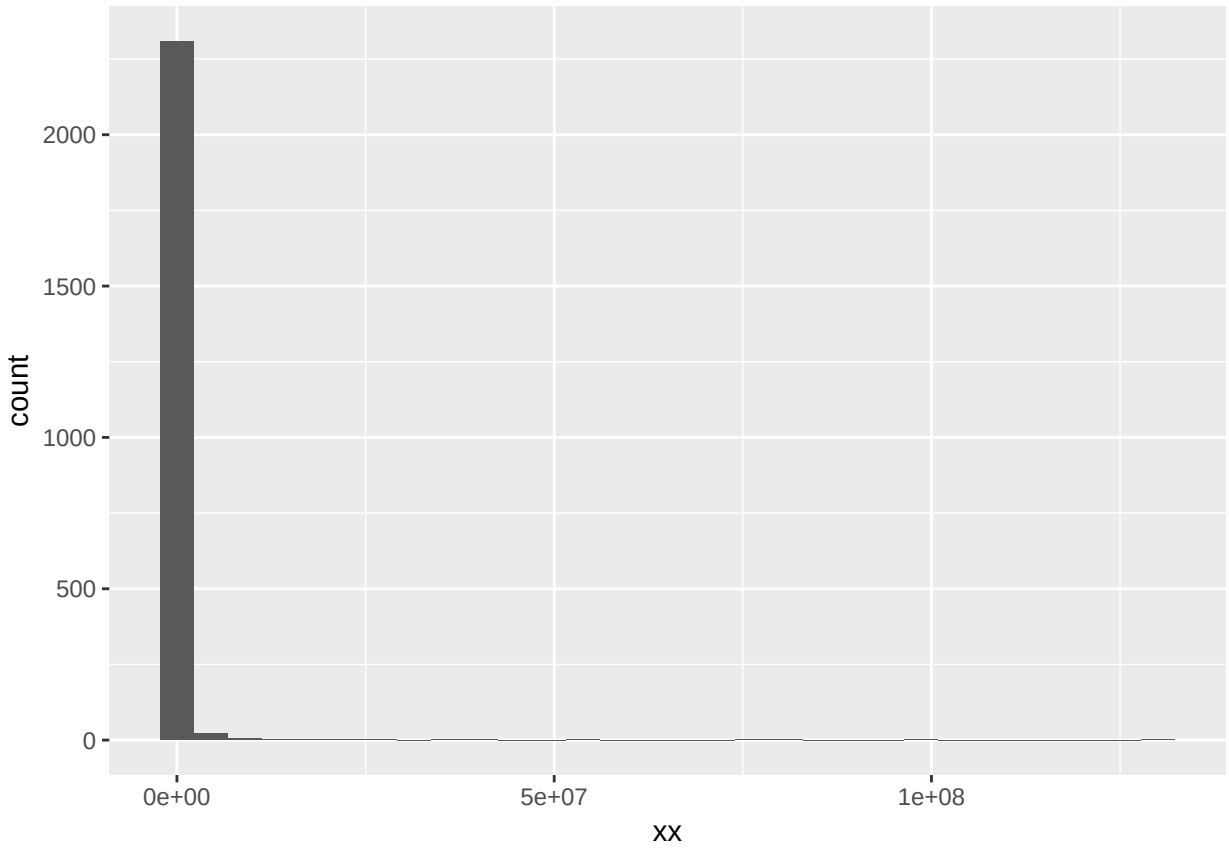
```
source("prep.R")
dd <- prep.data("~/data/privacyrights.org/select-all.csv")
d <- dd$prepped
xx <- as.numeric(d$Total.Records)
summary(xx)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##         2       340      2500    388045     19950 130000000
```

How far out is that max?

```
library(ggplot2)
qplot(xx)
```

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




```
tail(sort(xx))
```

```
## [1] 4.0e+07 5.6e+07 7.6e+07 8.0e+07 1.0e+08 1.3e+08
```

It's pretty far out there, but not *too* crazy. Can we rule out the log normal distribution as a model for this data?

```
y <- log(xx)
normal <- rnorm(length(y), mean(y), sd(y))
ks.test(y, normal)
```

```
## Warning in ks.test(y, normal): p-value will be approximate in the presence
## of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  y and normal
## D = 0.03234, p-value = 0.1711
## alternative hypothesis: two-sided
```

No, Edwards et al. don't want to rule out the model when the p-value is above 0.05.

Just out of curiousity, how's the t distribution with—say—five degrees of freedom?

```r
t_dist <- rt(length(y), 5) * sd(y) + mean(y)
ks.test(y, t_dist)
```

```
## Warning in ks.test(y, t_dist): p-value will be approximate in the presence
## of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  y and t_dist
## D = 0.054468, p-value = 0.001876
## alternative hypothesis: two-sided
```

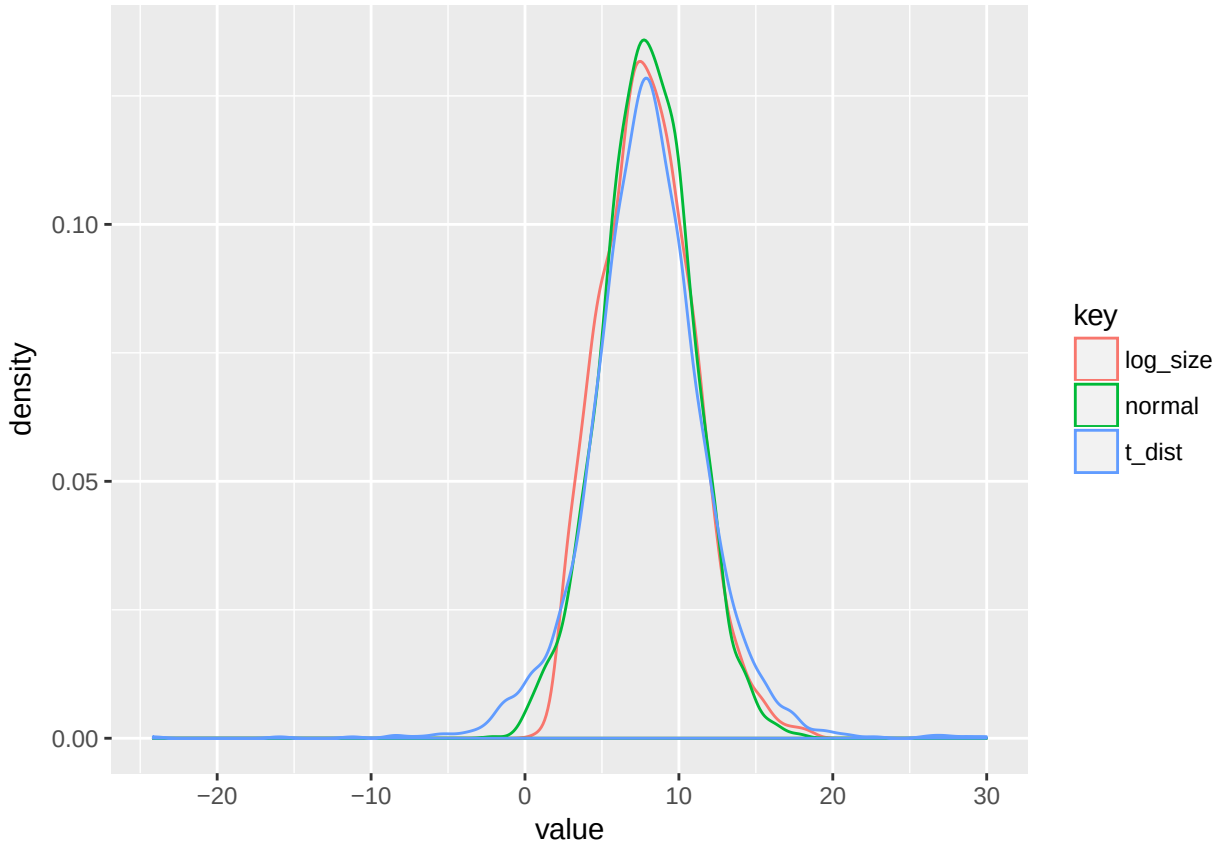
Not as good.

So I think we can go ahead with log-normal as a model for breach size.

```r
library(tidyr)

df <- data.frame(
    log_size = y,
    t_dist = t_dist,
    normal = normal
) %>% gather()

ggplot(df, aes(value, color=key)) + geom_density()
```



They both look OK, but `ks.test` likes log-normal best.