Empirical reference distributions – example

If a relatively large data set exists, then the distribution of a particular statistic (e.g. a 30-yr mean) can be created by repeatedly

- drawing a sample from the data set,
- calculating the value of the statistic for the sample,
- saving the statistic, and then

and, after a sufficient number of samples have been taken (the more the better),

- looking at the distribution of the sample statistics (via the histogram or other plot, or by direct inspection).

Use of an empirical reference distribution

The examples here use the 1189-year-long time series of summer temperature (TSum) for the Sierra Nevada, as reconstructed from tree-ring data (Graumlich, 1993). [**sierra.csv**] This data set is sufficiently long enough to illustrate the idea of having a large amount of empirical information to base judgements about the *significance* (or the relative unusualness) of individual values or statistics.

*How unusual is a particular observation of **TSum**?*

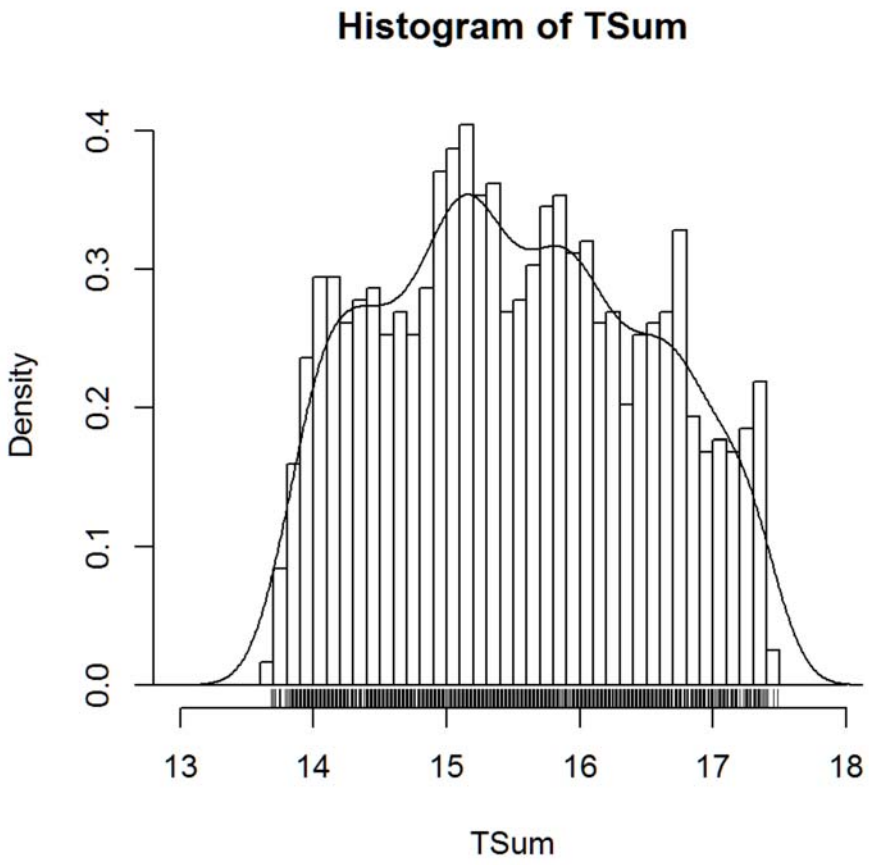
To answer this question, the appropriate reference distribution is the one formed by all of the individual observations of **TSum**. In other words, the entire data set should be used to answer questions about the significance (or relative size) of an individual observation, simply by comparing that observation to all other observations. This comparison is facilitated by first sorting the observations, and then creating a new variable that indicates the relative position within the sorted observations of each observation.

```
# Sierra Nevada tree-ring summer temperature reconstructions
attach(sierra)

# sort and rank TSum values
sort.TSum <- sort(TSum) # arranges values from low to high
rank.TSum <- rank(TSum) # replaces values with rank number (1 to n)

# inspect the sorted and ranked values
# close window to exit
data.entry(Year, TSum, rank(TSum), sort.TSum, rank(sort.TSum),
     Year[rank(TSum)])
```

From this table, for example, the value 13.79 C can be seen to have occurred three times, in 1101, 1479, and 1892, and is a relatively low value of temperature, in contrast to the highest temperature value, 17.48, which occurred in 1194.

Similar information can be gained from the histogram.

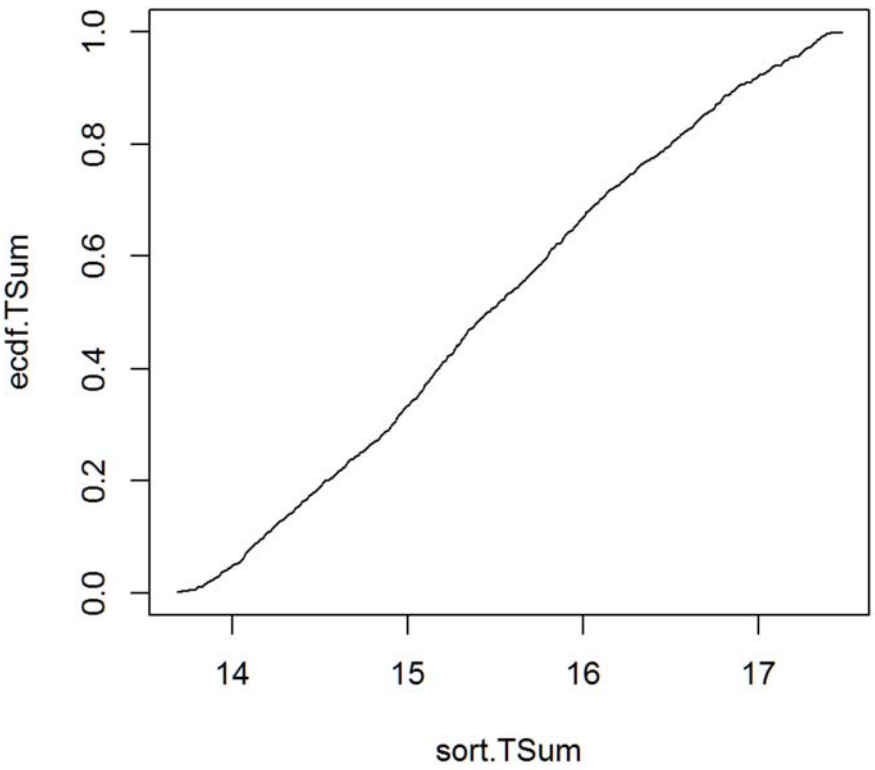
```
# nice histogram
cutpts <- seq(12.0, 18.0, by=0.1)
hist(TSum, breaks=cutpts, probability=T, xlim=c(13,18))
lines(density(TSum))
rug(TSum)
```



*What value of **TSum** is exceeded 10 percent of the time?*

Other information can be gained by examining the empirical cumulative density function, which is a plot of the sorted values of **TSum** and the probability of equaling or exceeding that value in the data set

```
# empirical cumulative density function
ecdf.TSum <- rank(sort.TSum)/length(TSum)
plot(sort.TSum, ecdf.TSum, type="l")
```




```
data.entry(Year, TSum, rank(TSum), sort.TSum, rank(sort.TSum),
     Year[rank(TSum)], ecdf.TSum)
```

Again, the specific answer can be obtained by inspection of the data sheet, the histogram, or a Line plot of ProbTSum vs SortTSum.

Alternatively, the values corresponding to specific quantiles can be gotten by the following command:

```
quantile(TSum, probs = c(0.0,.05,.10,.25,.50,.75,.90,.95, 1.0))
```

```
##     0%     5%    10%    25%    50%    75%    90%    95%   100%
## 13.680 14.014 14.180 14.730 15.450 16.290 16.870 17.160 17.480
```

How unusual is the mean of the last 30 years of the record?

The following command gets the mean of the last 30 yrs of data:

```
mean(TSum[(length(TSum)-29):(length(TSum))])
```

```
## [1] 15.848
```

To answer the above question, the appropriate reference distribution is no longer that formed from the whole data set, but instead is the distribution that would be formed by a set of mean values of 30-year long samples from the data. This set can be obtained by repeated sampling of the **TSum** data using the following:

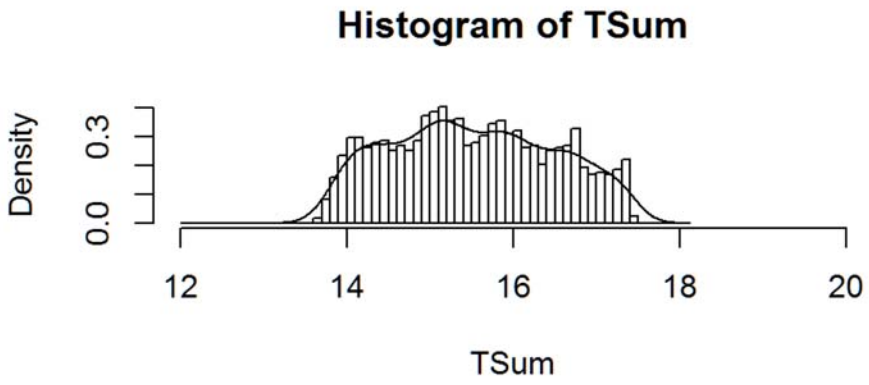
```
# repeated sampling and calculation of means
nsamp <- 200 # number of samples
Means30yr <- matrix(1:nsamp) # matrix to hold means
for (i in 1:nsamp) {
     samp <- sample(TSum, 30, replace=T)
          Means30yr[i] <- mean(samp)
     }
```

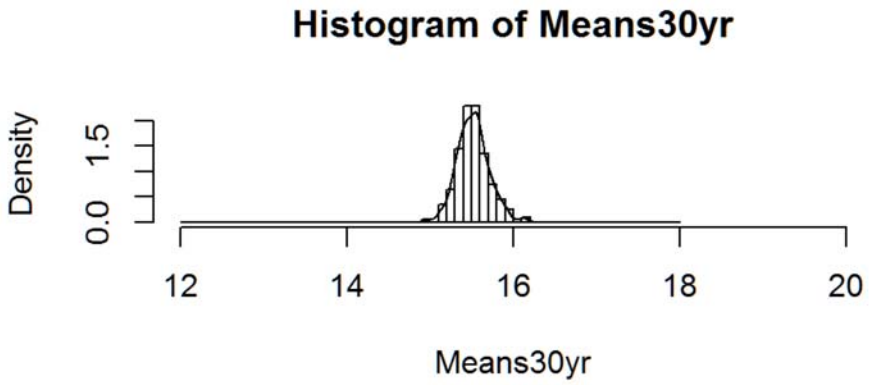
The script causes causes **nsamp** samples of 30 observations each to be drawn at random from the column of data containing **TSum**. For each

sample, the mean is calculated, and stored in the variable **Means30yr**. The distribution of this variable (**Means30yr**) rather than that of the original data is the appropriate one for answering the above question.

The two empirical reference distributions can be compared by looking at their histograms:

```
# histograms of data and of 30-Means
par(mfrow=c(2,1))
cutpts <- seq(12.0, 18.0, by=0.1)
hist(TSum, breaks=cutpts, probability=T, xlim=c(12,20))
lines(density(TSum))
hist(Means30yr, breaks=cutpts, probability=T, xlim=c(12,20))
lines(density(Means30yr))
```





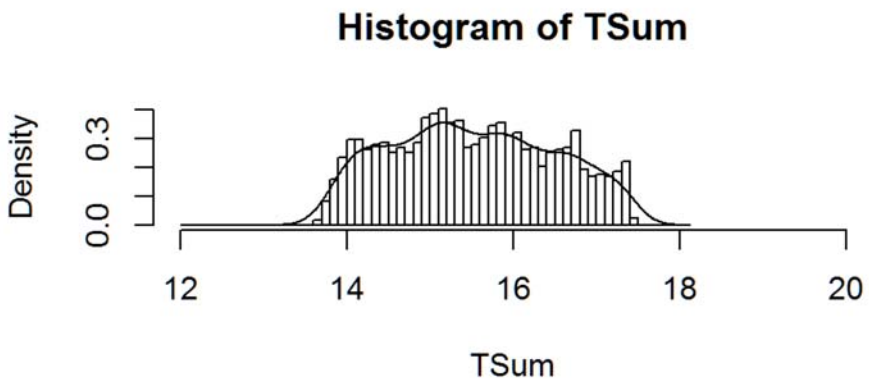
The above procedure draws 30 years at random from TSum. It might be more appropriate to draw 30-consecutive-year samples from TSum. This can be done as follows:

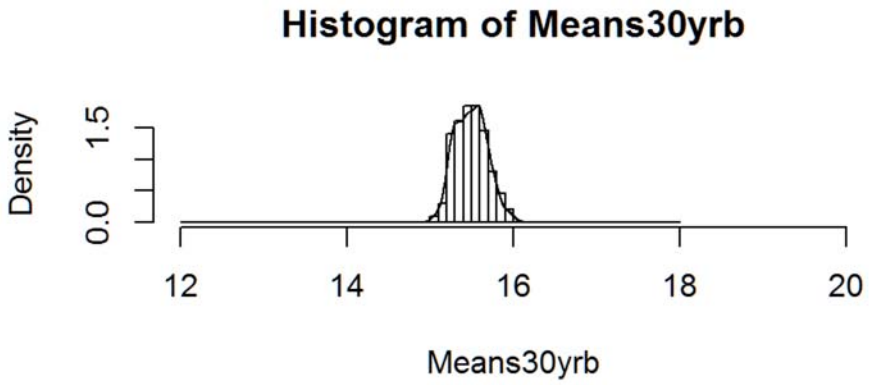
```
# repeated sampling and calculation of means
# consecutive samples
nsamp <- 200 # number of samples
Means30yrb <- matrix(1:nsamp) # matrix to hold means
for (i in 1:nsamp) {
     start <- runif(1, min=1, max=length(Year)-30)
     start <- round(start)
     sampmean <- 0.0
     for (j in 1:30) {
          sampmean <- sampmean + TSum[start+j-1]
     }
     Means30yrb[i] <- sampmean/30.0
}
```

And here are the histograms:

```
# histograms of data and of 30-Means from consecutive samples
par(mfrow=c(2,1))
cutpts <- seq(12.0, 18.0, by=0.1)
```

```
hist(TSum, breaks=cutpts, probability=T, xlim=c(12,20))
lines(density(TSum))
hist(Means30yrb, breaks=cutpts, probability=T, xlim=c(12,20))
lines(density(Means30yrb))
```





Note that this approach for using an empirical reference distribution is a general one–the significance of other statistics (e.g. the variance) can be evaluated the same way.