The bootstrap

Let $\mathbf{X} = (X_1, \ldots, X_n)$ represent an independent sample from the loss distribution.

- We would like to describe some property θ of this distribution. θ might be the expectation (i.e. the *risk*), or it might be a quantile such as the 99.5th percentile.
- We estimate θ from the sample in some fashion, denoted $\hat{\theta} = t(\mathbf{X}^{\text{obs}})$, where $\mathbf{X}^{\text{obs}}$ are the observed values of our sample. *Because n is not infinitely large, our estimate $\hat{\theta}$ is unlikely to be exactly the same as the true value, θ.*
- The bootstrap is a simple empirical technique to assess the variability of our estimate. It comes in many forms, but we will use one of the simplest, the *basic bootstrap* (with transformation).
- The bootstrap is covered in most statistics textbooks, but the best reference is A.C. Davison and D. Hinkley, 1997, *Bootstrap Methods and their Application*, Cambridge University Press.

Underlying principle

Let $T = t(\mathbf{X})$. Suppose we can find values a and b such that

$$\Pr\{a \leq T - \theta \leq b\} = 1 - 2\alpha,$$

for some specified value α; typically $\alpha = 2.5\%$. By rearranging these inequalities we can see that

$$\Pr\{T - b \leq \theta \leq T - a\} = 1 - 2\alpha, \qquad (2)$$

or $[\hat{\theta} - b, \hat{\theta} - a]$ is a $(1 - 2\alpha)$ confidence interval for θ.

Basic bootstrap

In the basic bootstrap, we estimate a and b from the sample $\mathbf{X}^{\text{obs}}$.

- Let $\mathbf{X}^*$ denote a sample of size n taken from $\mathbf{X}^{\text{obs}}$ *with replacement*. Let $R = 999$ or same other large number, and $\mathbf{X}_1^*, \ldots, \mathbf{X}_R^*$ be R random samples.
- Let $T_{[i]}^*$ denote the ith ordered value from $t(\mathbf{X}_1^*), \ldots, t(\mathbf{X}_R^*)$. Then

$$a \approx T_{[(R+1)\alpha]}^* - \hat{\theta} \quad \text{and} \quad b \approx T_{[(R+1)(1-\alpha)]}^* - \hat{\theta},$$

  where $\hat{\theta}$ is standing in for θ.
- Plugging these into the formula for the confidence interval, eq. (2), we get

$$\left[2\hat{\theta} - T_{[(R+1)(1-\alpha)]}^*, 2\hat{\theta} - T_{[(R+1)\alpha]}^*\right]$$

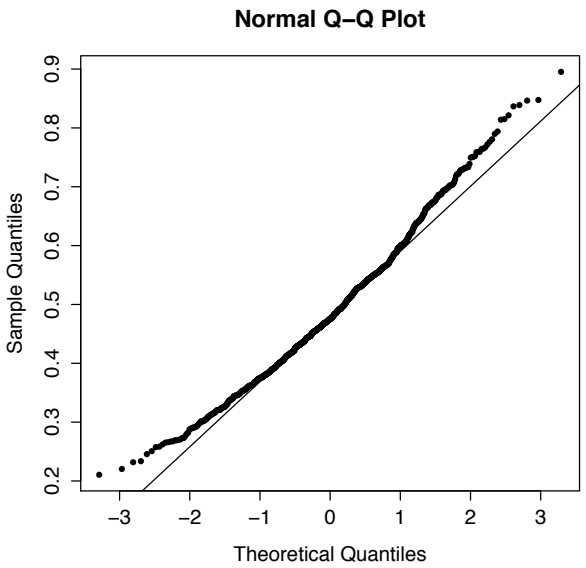
is an approximate $1 - 2\alpha$ confidence interval for θ.

```
#### code snippet for basic bootstrap to find
#### 95% CI for mean of Exp(2)

## here are the 'true' values

set.seed(101) # or some other value

n <- 20
X <- rexp(n, rate = 2)
that <- mean(X) # true mean is 1/rate = 0.5

## here is a bootstrap estimate of the 95% CI for the mean

alpha <- 0.025
R <- 999
Tstar <- sapply(1:R, function(i) {
  Xstar <- sample(X, size = n, replace = TRUE)
  mean(Xstar)
})
Tvals <- sort(Tstar)[c((R+1) * (1-alpha), (R+1) * alpha)]
CI <- 2 * that - Tvals
print(CI)
# [1] 0.2414974 0.6850869 -- my answer

## change the set.seed value or comment it out
## to generate a different random value to your
## neighbour's
```

Properties

The basic bootstrap works best if $t(X) - \theta$ has a Normal distribution that does not depend on the distribution of X. We can investigate using the normal quantile plot of T^*:

```
par(mfrow = c(1, 2), cex = 0.75)
qqnorm(Tstar, pch = 16, cex = 0.9); qqline(Tstar)
```



which is not very Normal.

Transformations

We would like to make the distribution of $t(X) - \theta$ more Normal.

- We can introduce a continuous and increasing transformation h, and then if we can find a and b such that

$$\Pr\{a \leq h(T) - h(\theta) \leq b\} = 1 - 2\alpha,$$

then, by similar reasoning to that in eq. (2)

$$\left[h^{-1}(h(\hat{\theta}) - b), h^{-1}(h(\hat{\theta}) - a)\right]$$

is a $(1 - 2\alpha)$ confidence interval for θ.

- For losses, which tend to have strong positive skewness, the logarithmic transformation $h(t) = \log(t)$ might work well.

Transformations (cont)

Here's the normal quantile plot, without and with a logarithmic transformation:

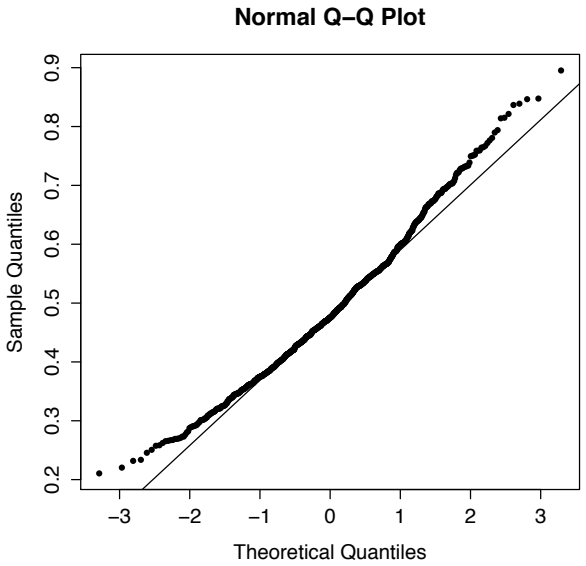
```
par(mfrow = c(1, 2), cex = 0.75) # start new pair of plots
qqnorm(Tstar, pch = 16, cex = 0.9, main = "Original scale")
qqline(Tstar)
qqnorm(log(Tstar), pch = 16, cex = 0.9,
       main = "Logarithmic scale"); qqline(log(Tstar))
```

which is more Normal.

Transformations (cont)

Here's the normal quantile plot, without and with a logarithmic transformation:

```
par(mfrow = c(1, 2), cex = 0.75) # start new pair of plots
qqnorm(Tstar, pch = 16, cex = 0.9, main = "Original scale")
qqline(Tstar)
qqnorm(log(Tstar), pch = 16, cex = 0.9,
       main = "Logarithmic scale"); qqline(log(Tstar))
```

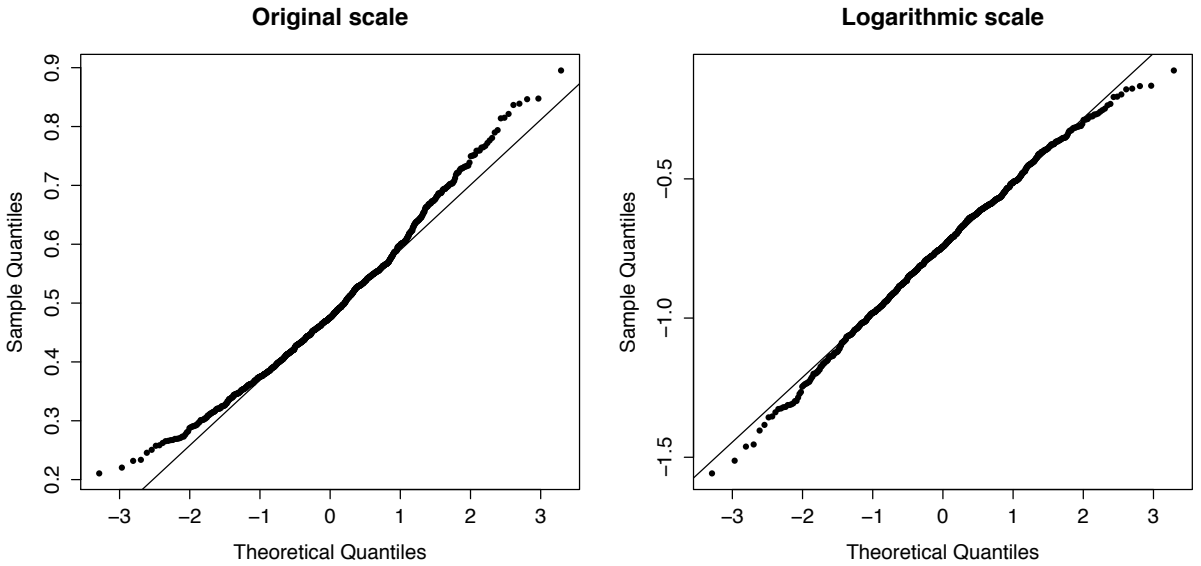


which is more Normal.

Transformations (cont)

Here's the normal quantile plot, without and with a logarithmic transformation:

```
par(mfrow = c(1, 2), cex = 0.75) # start new pair of plots
qqnorm(Tstar, pch = 16, cex = 0.9, main = "Original scale")
qqline(Tstar)
qqnorm(log(Tstar), pch = 16, cex = 0.9,
       main = "Logarithmic scale"); qqline(log(Tstar))
```



which is more Normal.

Transformations (cont)

Handling transformations is straightforward, because the ordering of T^* is the same as the ordering of $h(T^*)$.

```
#### ** with log transformation **

CI <- exp(2 * log(that) - log(Tvals))
print(CI)
# [1] 0.3240093 0.8197120 -- my answer
```

High percentiles

- In the Solvency II regulations for insurance and re-insurance, companies must have capital assests sufficient to meet a loss as big as the 99.5th percentile. High percentiles can be hard to estimate with moderately-sized n, and the bootstrap is a useful technique for quantifying uncertainty about the true percentiles.
- For the 99.5th percentile, one would want a sample of at least $n = 1000$, and preferably much more.
- More generally, hazard losses are often summarised in terms of, e.g., the 30 year return period, i.e. the 96.7th percentile.

```
#### code snippet for basic bootstrap to find
#### 95% CI for 99.5th percentile of Exp(2),
#### with logarithmic transformation

## 'true' values

set.seed(201)

n <- 10000 # ought to be large for high percentile
X <- rexp(n, rate = 2)
that <- quantile(X, prob = 0.995) # true value is 2.649159

## bootstrap, alpha and R as before

Tstar <- sapply(1:R, function(i) {
  Xstar <- sample(X, size = n, replace = TRUE)
  quantile(Xstar, prob = 0.995)
})
Tvals <- sort(Tstar)[c((R+1) * (1-alpha), (R+1) * alpha)]
CI <- exp(2 * log(that) - log(Tvals))
print(CI)
#    99.5%     99.5%
# 2.429986 2.690188 -- my answer, ignore the labels in 1st row
```

Exercises

1. Check the sampled values of Tstar from the previous snippet for Normality, both untransformed and with a log transformation. What do you notice?
2. Repeat the previous code snippet with `n <- 1000`. how much difference does this make to the width of the 95% CI for the 99.5th percentile?
3. Starting with the following set of values:

$$\mathbf{X} = (5, 3, 4, 1, 6, 3, 2, 4, 2, 4)$$

use the basic bootstrap to estimate the 95% CI for the median (hint, try the R function `median`). You may want to investigate transformations such as $h =$ square root.