Sister telomeres: Basic Data Analysis*

Miljenko Huzak, Petra Lazić, and Ivica Rubelj

November 22, 2020

1 Introduction

Telomeres are specialized structures at the ends of linear chromosomes, composed of repetitive DNA and an associated protein complex called shelterin. They are dynamic structures, continuously losing their repeats with increasing cell divisions. In normal somatic cells critically short telomeres signal growth arrest which is considered to be the main mechanism of senescence and consequently the process of aging. Telomere length is widely used as a reliable biomarker for longevity and aging related diseases, both at the individual and population level. Since telomere length is associated with age related pathologies, including cardiovascular disease and cancer, accurate measuring and interpretation of measured results are necessities. The telomere Q-PNA-FISH technique has been widely used for this purpose.

The sensitivity level of Q-PNA-FISH is about 200 base pairs. Various techniques demonstrated that telomeres lose only about 50-150 base pairs per cell division which is below the detection level of Q-PNA-FISH. Thus, when metaphase chromosomes are analyzed, a time when sister telomeres are still together following replication, one could expect that their Q-PNA-FISH signal intensities will be about the same. However great differences in Q-PNA-FISH signal intensities between sister telomere pairs in normal cells were observed previously. This discrepancy is not a real biological phenomenon but the result of inefficient labeling of telomere repeat sequences by the PNA probe. This inefficiency in labeling results in, to some extent, random distribution of analyzed telomere Q-PNA-FISH signals. Hence an analysis of the relationship between Q-PNA-FISH signal intensities among sister telomeres are of great importance (see A. Čukušić, N. Š. Vidaček, M. Huzak, M. Ivanković, I. Rubelj, Telomere Q-PNA-FISH - Reliable results from stochastic signals, *PLOS ONE* **9** (3) (2014) e92559, 1-10).

*Based on paper A. Čukušić, N. Š. Vidaček, M. Huzak, M. Ivanković, I. Rubelj, Telomere Q-PNA-FISH - Reliable results from stochastic signals, *PLOS ONE* **9** (3) (2014) e92559, 1-10) and its supplement materials

2 Data

Three data samples of human fibroblast cultures were obtained: one that is 32 population doublings (PDs) old, one 42 PDs old, and one 52 PDs old. We shortly name these samples: "PD 32", "PD 42", and "PD 52" respectively.

Statistical units in all samples are sister chromatides.

Measured variables are Q-PNA-FISH signal intensities of sister telomeres (3-dimensional value): the first component is signal intensity of longer telomere sister, the second one is signal intensity of shorter telomere sister, and third component is a difference between the first two component (it is derived from the first two components).

How to read data from file to R?

```
> data30 <- read.table("PD30.txt")
> data40 <- read.table("PD40.txt")
> data50 <- read.table("PD50.txt")
```

3 Statistical model for the data

Let (X, Y) represent a pair of sister chromatids, where X is the signal intensity of the longer telomere (longer sister), and Y is the difference between signal intensities of the longer and shorter telomere sisters. For every pair, the relative difference in telomere lengths between the sisters is the number $Z = Y/X$, with $0 < \mathrm{Z} < 1$ necessary. More precisely, let

$$(X_1, Y_1),\ (X_2, Y_2), \ldots,\ (X_n, Y_n) \tag{1}$$

be n pairs of chromosome sisters in the sample ("PD 32", "PD 42", or "PD 52"), (X_i, Y_i) representing the ith pair, $i = 1, 2, , n$, and let

$$Z_1 = \frac{Y_1}{X_1},\ Z_2 = \frac{Y_2}{X_2}, \ldots,\ Z_n = \frac{Y_n}{X_n} \tag{2}$$

be their relative differences in telomere lengths. We assume that the signal intensity and the length of a telomere are proportional. Notice that is reasonable to assume that the coefficients of proportionality of a sister chromatids are the same, but there is no reason to assume the same for the coefficients across different chromatid pairs and different samples. By deriving a new variable, the *relative* difference in telomere lengths between the sisters (Z), we are avoiding the problem of coefficient heterogeneity.

Since the origin cell of any pair of sisters is chosen randomly and the splitting process resulting in a pair of chromosome sisters is independent of the other cells and the other chromosomes in the cell, we may think of the sample (1) as a sequence of independent random vectors. This implies that random variables (2) are independent too. Is sample (2) a sequence of *identicaly* distributed random

variables (r.vs.), or are r.vs. (2) *homogenuous*? We can check this for all three samples.

We can quicly check homogeneity graphically by comparing box plots of relative differences (Z) drawing on subsamples classified with respect to the signal intensity of the longer sister (X).

How to draw and graphically compare box plots of the subsamples in R?

First, we need to create subsamples.

```
> k <- 10
> c <- 0.22
> borders <- numeric(k+1)
> borders[1] <- 0.2
> for(i in 2:(k+1)) borders[i] <- borders[i-1] + c
> borders

 [1] 0.20 0.42 0.64 0.86 1.08 1.30 1.52 1.74 1.96 2.18 2.40

> n30 <- length(data30[,1])
> data30[,1] <- data30[,1]/10000
> labels30 <- numeric(n30)
> j <- 1
> for(i in 1:k){
+   while( (data30[j,1] >= borders[i]) && (data30[j,1] < borders[i+1]) ){
+     labels30[j] <- i
+     j <- j+1
+     if(j > n30) break
+   }
+         if(j > n30) break
+ }
> labels30 <- factor(labels30, levels = c(1:k))
> d30 <- data.frame(labels30, X30 = data30[,1], Z30 = (data30[,3]/10000)/data30[,1])
```

Now, we can plot boxplots.

Figure 1: Box plots of relative difference in telomere lengths between the sisters for sample "PD 32"

```
> categories <- paste(borders[-(k+1)], borders[-1], sep="-")
> categories

 [1] "0.2-0.42"  "0.42-0.64" "0.64-0.86" "0.86-1.08" "1.08-1.3"  "1.3-1.52"
 [7] "1.52-1.74" "1.74-1.96" "1.96-2.18" "2.18-2.4"

> par(mar = c(5,4,1,1))
> boxplot(Z30 ~ labels30, data = d30, names = categories, las = 2, xlab = NULL,
+         ylab = "Z_PD32")
```

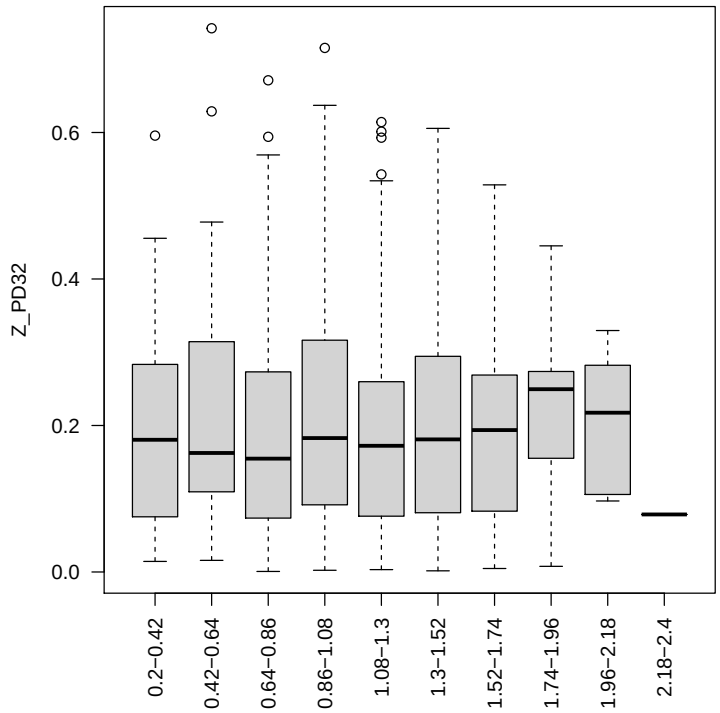


Figure 2: Box plots of relative difference in telomere lengths between the sisters for sample “PD 42”

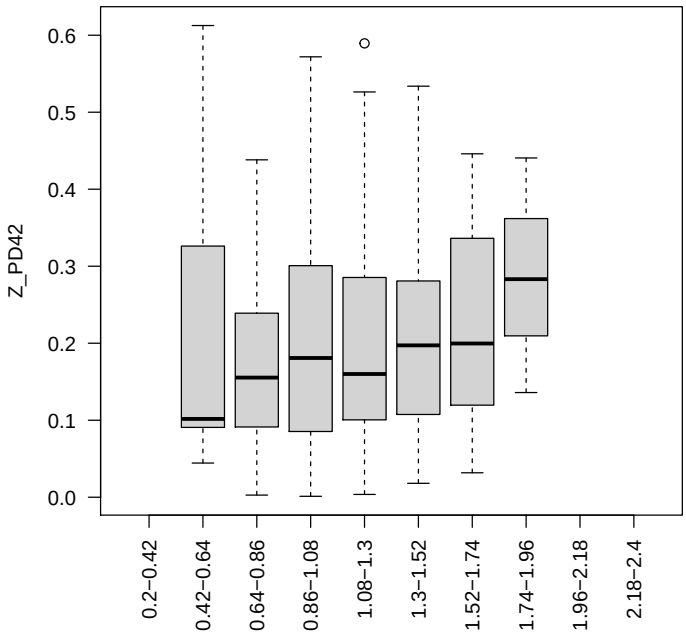
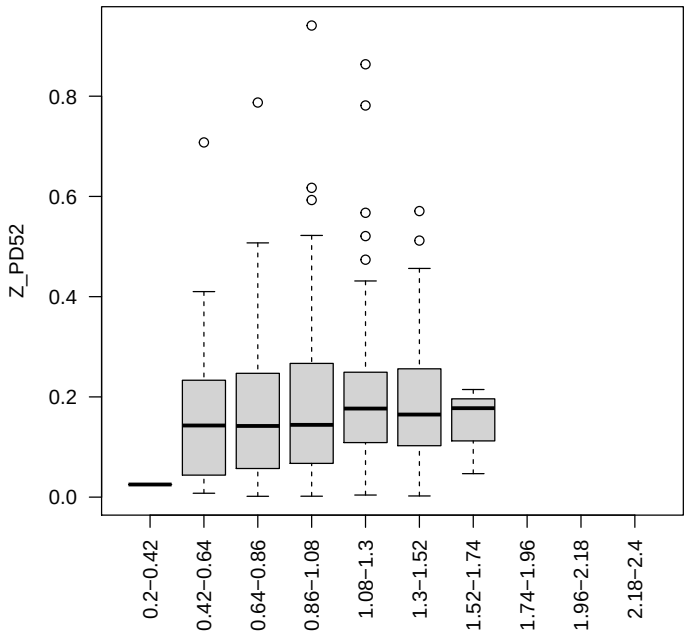


Figure 3: Box plots of relative difference in telomere lengths between the sisters for sample "PD 52"



Based on those graphs, we cannot reject assumption of homogeneity of distribution of random variables (2). Hence we assume that (2) are independent and identically distributed (i.i.d.) random variables.

Can we assume anything about the common population law of random variables (2)? Let us examine histograms of the samples "PD 32", "PD 42", and "PD 52".

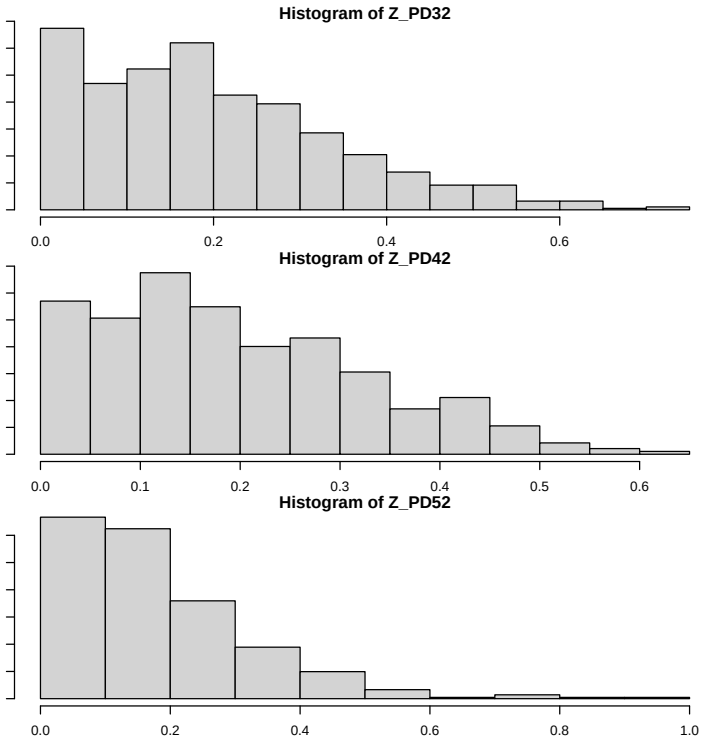
How to draw histograms of the samples in R?

In case of each sample we see that histogram, as graphical estimators of the density od the common law, are highly skewed so we cannot assume that the cammon law comes from normal laws family of distributions for example. So far there are no evidence to assume that the common law follows any of the known parametric distributions. Hence in the best we can and will assume that the common law has finite the first four moments: mean (μ), variance (σ^2), skewness (α) and kurtosis (κ).

The parameters of the model are μ, σ^2, α, and κ. Sometimes for us to be able to differ among the samples (and corresponding populations), we will put subscripts such as, e.g. μ_{PD32}.

Figure 4: Histogram of relative difference in telomere lengths for all three samples

```
> par(mar=c(2,1,1,1))
> par(mfrow = c(3,1))
> hist(d30$Z30, probability = T, ylab = NULL, xlab = NULL,
+      main = "Histogram of Z_PD32")
> hist(d40$Z40, probability = T, ylab = NULL, xlab = NULL,
+      main = "Histogram of Z_PD42")
> hist(d50$Z50, probability = T, ylab = NULL, xlab = NULL,
+      main = "Histogram of Z_PD52")
> par(mfrow = c(1,1))
>
```



4 Mathematical analysis of the model

Since model parameters are population moments we use *the method of moments* to estimate them from the data:

$$\begin{aligned}
\hat{\mu} &= \overline{Z}_n := \frac{1}{n}\sum_{i=1}^{n} Z_i \\
\hat{\sigma}^2 &= S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2 \\
\hat{\alpha} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Z_i - \overline{Z}_n}{S_n}\right)^3 \\
\hat{\kappa} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Z_i - \overline{Z}_n}{S_n}\right)^4 .
\end{aligned}$$

Since we assumed that the first four moments of the population distribution were finite, we can apply *The Strong Law of Large Numers* to show that the above stimators are *strongly consistent*, i.e.

$$\begin{aligned}
\hat{\mu} &= \overline{Z}_n \xrightarrow{\text{a.s.}} \mu, \quad n \to +\infty \\
\hat{\sigma}^2 &= S_n^2 \xrightarrow{\text{a.s.}} \sigma^2, \quad n \to +\infty \\
\hat{\alpha} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Z_i - \overline{Z}_n}{S_n}\right)^3 \xrightarrow{\text{a.s.}} \alpha, \quad n \to +\infty \\
\hat{\kappa} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Z_i - \overline{Z}_n}{S_n}\right)^4 \xrightarrow{\text{a.s.}} \kappa, \quad n \to +\infty.
\end{aligned}$$

We will compare populations by the first two moments (μ and σ^2). Kurtosis (κ) plays an auxiliary role.

First, let us estimate 95% *confidence intervals* for μ and σ^2 as their interval estimates. To do this we need *sample distributions* of their point estimators $\hat{\mu}$ and $\hat{\sigma}^2$, or, at least, their *approximate* sample distributions.

Since we assumed that r.vs. (2) are i.i.d. with a finite positive variance σ^2, we can applay *The Central Limit Theorem* and conclude that *for large* n,

$$\frac{\hat{\mu} - \mu}{\sigma}\sqrt{n} = \frac{\overline{Z}_n - \mu}{\sigma}\sqrt{n} \sim AN(0,1)$$

where 'A' before '$N(0,1)$' stands for word 'approximate'. In words: 'For large n r.v. $\frac{\hat{\mu}-\mu}{\sigma}\sqrt{n}$ has approximate standard normal law'. Since $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ is an consistent estimator of standard deviation σ, the following holds too: For large n,

$$\frac{\hat{\mu} - \mu}{\hat{\sigma}}\sqrt{n} \sim AN(0,1). \tag{3}$$

If $z_{0.025} = 1.96$ denotes 0.975-*quantile* of $N(0,1)$-distribution then

$$\begin{aligned} 0.95 &\approx \mathrm{P}\left(\left|\frac{\hat{\mu}-\mu}{\hat{\sigma}}\sqrt{n}\right| \leq 1.96\right) = \\ &= \mathrm{P}\left(\hat{\mu} - 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) \end{aligned}$$

implying that

$$\left[\hat{\mu} - 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}},\ \hat{\mu} + 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right] \tag{4}$$

is 95% confidence interval for μ.

Similarly, for large n

$$\frac{\log \hat{\sigma}^2 - \log \sigma^2}{\sqrt{\hat{\kappa}-1}}\sqrt{n} \sim AN(0,1) \tag{5}$$

by The Central Limit Theorem and *the delta method*. Hence 95% confidence interval for σ^2 is random interval

$$\left[\hat{\sigma}^2 \cdot \exp\left(-1.96 \cdot \sqrt{\frac{\hat{\kappa}-1}{n}}\right),\ \hat{\sigma}^2 \cdot \exp\left(1.96 \cdot \sqrt{\frac{\hat{\kappa}-1}{n}}\right)\right]. \tag{6}$$

At the second step, we need to test several hypothesis about comparison of mean and variance parameters.

For testing nul-hypothesis

$$H_0 : \ \mu_1 = \mu_2$$

where μ_j referes to to the sample from jth population for $j = 1$ or $j = 2$. Main assumption is that the both samples were taken independently of each other. In this case we will use the following test statistics:

$$Z_\mu = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \overset{H_0}{\sim} AN(0,1) \tag{7}$$

for large n_1 and n_2 where n_j is a length od the sample j for $j = 1$ or $j = 2$. In case of testing nul-hypothesis

$$H_0 : \ \sigma_1^2 = \sigma_2^2$$

we will use test-statistics

$$Z_\sigma = \frac{\log \hat{\sigma}_1^2 - \log \hat{\sigma}_2^2}{\sqrt{\frac{\hat{\kappa}_1 - 1}{n_1} + \frac{\hat{\kappa}_2 - 1}{n_2}}} \overset{H_0}{\sim} AN(0,1) \tag{8}$$

for large n_1 and n_2. Notice that the both statistics (7) and (8) are derived from (3) and (5) respectively, and from the assumptions that the samples were taken independently of each other and that nul-hypotheses hold.

For the reference, see e.g. Ferguson, T. S. (1996) *A course in large sample theory*, London; New York, Chapman & Hall.

5 Statistical inference

Populations about we would like to infer from the samples are all sister telomeres taken from the same type of cells but with different PDs obtained. These populations we will call briefly: "PD 32", "PD 42", and "PD 52", i.e. in the same way as we call appropriate samples.

How to calculate point and interval estimates from the samples in R?

```
> c(length(d30$Z30), length(d40$Z40), length(d50$Z50))

[1] 742 379 423

> c(mean(d30$Z30), mean(d40$Z40), mean(d50$Z50))

[1] 0.2007552 0.1971519 0.1809918

> c(var(d30$Z30), var(d40$Z40), var(d50$Z50))

[1] 0.02097411 0.01760579 0.02065258

> kapa <- function(x){
+   sum( ((x - mean(x))/sd(x))^4 / (length(x) - 1) )
+ }
> c(kapa(d30$Z30), kapa(d40$Z40), kapa(d50$Z50))

[1] 3.327851 2.739020 6.839382

> ci_mu <- function(x){
+   l <- mean(x) - 1.96 * sd(x) / length(x)
+   d <- mean(x) + 1.96 * sd(x) / length(x)
+   return( c(l,d) )
+ }
> c(ci_mu(d30$Z30), ci_mu(d40$Z40), ci_mu(d50$Z50))

[1] 0.2003727 0.2011378 0.1964657 0.1978381 0.1803259 0.1816577

> ci_sigma <- function(x){
+   l <- var(x) * exp( -1.96 * sqrt( (kapa(x) - 1) / length(x)))
+   d <- var(x) * exp( 1.96 * sqrt( (kapa(x) - 1) / length(x)))
+   return( c(l,d) )
+ }
> c(ci_sigma(d30$Z30), ci_sigma(d40$Z40), ci_sigma(d50$Z50))

[1] 0.01879342 0.02340784 0.01541685 0.02010552 0.01640446 0.02600080
```

Point estimates of the parameters and 95% confidence intervals (CI) for μ and σ^2 are presented in the following table.

sample	n	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\kappa}$	95% CI for μ	95% CI for σ^2
PD 32	742	0.201	0.0210	3.338	$[0.200, 0.201]$	$[0.0188, 0.0234]$
PD 42	379	0.197	0.0176	2.739	$[0.196, 0.198]$	$[0.0154, 0.0201]$
PD 52	423	0.181	0.0207	6.839	$[0.180, 0.181]$	$[0.0164, 0.0260]$

We can see that the estimates of mean values slightly decrese with increasing PDs, while there is no such (and any) monotonic trend in the estimates of variances. Let us test the significances of these observations.

How to calculate test statistics and p-values in R?

```
> mu.test <- function(x, y){
+   mu1 = mean(x)
+   mu2 = mean(y)
+   sigma1 = var(x)
+   sigma2 = var(y)
+   n1 = length(x)
+   n2 = length(y)
+
+   z = (mu1 - mu2)/ sqrt(sigma1/n1 + sigma2/n2)
+
+   p = pnorm(z, lower.tail = F)
+
+   return( c(z,p))
+ }
> c(mu.test(d30$Z30, d40$Z40), mu.test(d40$Z40, d50$Z50),
+   mu.test(d30$Z30, d50$Z50))

[1] 0.41685677 0.33839159 1.65557950 0.04890350 2.25092658 0.01219509

> sigma.test <- function(x, y){
+   sigma1 = var(x)
+   sigma2 = var(y)
+   kapa1 = kapa(x)
+   kapa2 = kapa(y)
+   n1 = length(x)
+   n2 = length(y)
+
+   z = (log(sigma1) - log(sigma2))/ sqrt((kapa1 - 1)/n1 + (kapa2 - 1)/n2)
+
+   p = 2 * pnorm(-abs(z))
+
+   return( c(z,p))
+ }
```

```
> c(sigma.test(d30$Z30, d40$Z40), sigma.test(d40$Z40, d50$Z50),
+   sigma.test(d30$Z30, d50$Z50))

[1]  1.99168363  0.04640578 -1.17689893  0.23923585  0.11868744  0.90552299
```

Results of testing hypothesis about means are presented in the first table:

H_0	H_a	z_μ	p-value
$\mu_{\text{PD32}} = \mu_{\text{PD42}}$	$\mu_{\text{PD32}} > \mu_{\text{PD42}}$	0.417	0.338
$\mu_{\text{PD42}} = \mu_{\text{PD52}}$	$\mu_{\text{PD42}} > \mu_{\text{PD52}}$	1.656	0.049
$\mu_{\text{PD32}} = \mu_{\text{PD52}}$	$\mu_{\text{PD32}} > \mu_{\text{PD52}}$	2.251	0.012

Here H_0 represents nul-hypothesis and H_a its alternative. There is no significant decreasing in mean values of relative difference in telomere lengths between chromatide sisters from 32 to 42 PDs old culture while the same trend between means of 42 and 52 PDs old culture is hardly significant (with 5% level of sigificance).

The second table represents results of testing hypothesis about variance:

H_0	H_a	z_σ	p-value
$\sigma^2_{\text{PD32}} = \sigma^2_{\text{PD42}}$	$\sigma^2_{\text{PD32}} \neq \sigma^2_{\text{PD42}}$	1.991	0.046
$\sigma^2_{\text{PD42}} = \sigma^2_{\text{PD52}}$	$\sigma^2_{\text{PD42}} \neq \sigma^2_{\text{PD52}}$	-1.177	0.239
$\sigma^2_{\text{PD32}} = \sigma^2_{\text{PD52}}$	$\sigma^2_{\text{PD32}} \neq \sigma^2_{\text{PD52}}$	0.119	0.905

We can conclude that there are no significant differences among variances.

Problems

1. Are sample lengths large enough that we can use normal approximation for statistics (3) and (5)?

2. Estimate 95% confidence intervals of μ and σ^2 by using bootstrap and Monte Carlo methods and compare them with 95% CIs from the first table.

3. Estimate p-values in testing hypothesis about variances from the last table by using bootstrap and Monte Carlo methods.