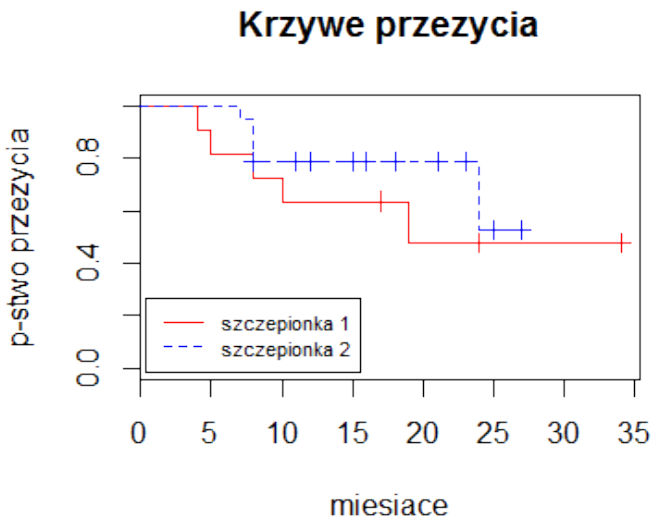
BIOSTATYSTYKA – PRACA DOMOWA 1
MARTA SOMMER – BSMAD

Zadanie 1.

Nasze dane wyglądają następująco: (zmienna *czas* to czas przeżycia, zmienna *zdarzenie* jest równa 1, gdy nastąpił zgon, a 0, gdy obserwacje były cenzurowane, zmienna *rodzaj.terapii* mówi o dwóch różnych rodzajach leczenia immunologicznego, zaś zmienna *wiek* odpowiada za grupę wiekową pacjenta: 1 dla grupy $21-40$ lat, 2 dla grupy $41-60$ lat, a 3 dla grupy 60 lat i więcej).

```
##   czas zdarzenie rodzaj.terapii wiek
## 1    4         1              1    2
## 2    5         1              1    3
## 3    8         1              1    1
## 4   10         1              1    3
## 5   17         0              1    2
## 6   17         0              1    1
```

Stosując metodę Kaplana-Meiera, otrzymujemy następujący wykres krzywych przeżycia dla dwóch rodzajów leczenia immunologicznego:



Czy istnieje istotna różnica między krzywymi przeżycia dla różnych rodzajów leczenia? Przekonajmy się stosując test $log-rank$:

```
## Call:
## survdiff(formula = Surv(czas, zdarzenie) ~ rodzaj.terapii, data = sz)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## rodzaj.terapii=1 11        5     3.71     0.448     0.756
## rodzaj.terapii=2 19        5     6.29     0.264     0.756
##
##  Chisq= 0.8  on 1 degrees of freedom, p= 0.385
```

Widać więc, że całkowita liczba oczekiwanych zgonów dla pierwszej grupy wynosi $3,71$, zaś dla drugiej grupy $6,29$. Oszacowanie wariancji wynosi zaś $2,2$. Wartość statystyki testowej dla testu $log-rank$, opartej na powyższym oszacowaniu wariancji, wynosi $0,76 \simeq 0,8$, a $p-value$ tej statystyki odpowiednio $0,385$, czyli nie mamy podstaw do odrzucenia hipotezy o tym, że funkcje przeżycia w obu grupach są takie same. Innymi słowy, metody leczenia nie różnią się znacznie między sobą. Wartość statystyki wyznaczona przy pomocy „prostszych obliczeń" wynosi $0,71$, a odpowiadające jej $p-value$ $0,399$, zatem wnioski dla tej uproszczonej statystyki będą takie same.

Zadanie 2.

Bazujemy na danych z Zadania 1. Tym razem jednak zastosujemy warstwowy (ze względu na *wiek*) test $log - rank$ do oszacowania, czy krzywe przeżycia dla różnych metod leczenia różnią się istotnie.

Oczekiwana liczba zgonów dla każdej z grup w każdej z warstw wynosi:

```
##               wiek 21-40 wiek 41-60 wiek 60+
## szczepionka 1      2.257     0.5222   0.9833
## szczepionka 2      2.743     1.4778   2.0167
```

Reszty dowiemy się patrząc na wydruk R-a:

```
## Call:
## survdiff(formula = Surv(czas, zdarzenie) ~ rodzaj.terapii + strata(wiek),
##     data = sz)
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## rodzaj.terapii=1 11        5     3.76     0.407     0.688
## rodzaj.terapii=2 19        5     6.24     0.245     0.688
##
##  Chisq= 0.7  on 1 degrees of freedom, p= 0.407
```

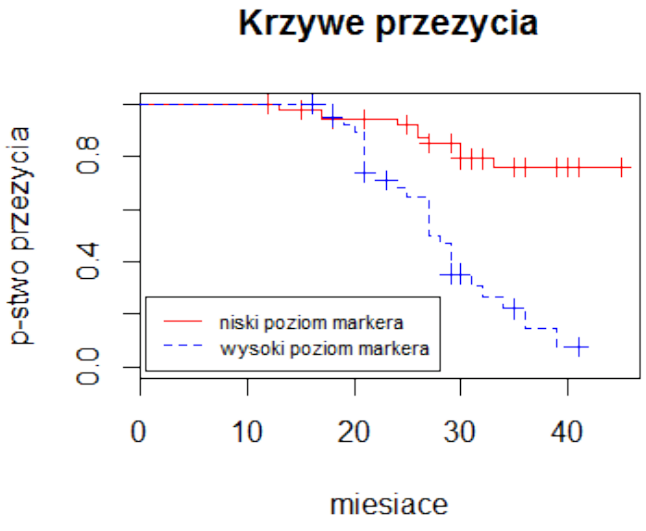
Widać więc, że sumaryczna oczekiwana liczba zgonów dla szczepionki pierwszej wynosi $3,76$, zaś dla szczopionki drugiej $6,24$. Wartość statystyki testowej wyznaczona przy pomocy wariancji wynosi $0,7$, a odpowiadające jej $p - value$ $0,407$. Zaś statystyka testowa wyznaczona przy pomocy „prostszych obliczeń" jest równa $0,65$, a jej $p - value$ $0,419$. Zatem w obu przypadkach nie mamy podstaw do odrzucenia hipotezy. Czyli, mimo wprowadzenia warstwowania, krzywe i tak istotnie nie różnią się między sobą.

Zadanie 3.

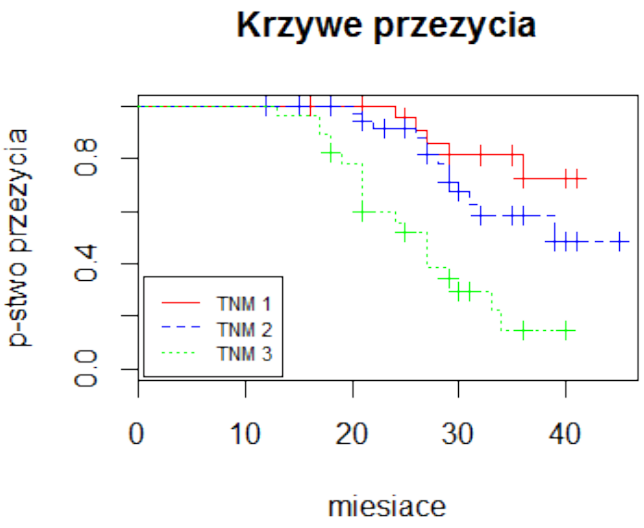
Zobaczmy, jak wyglądają nasze dane:

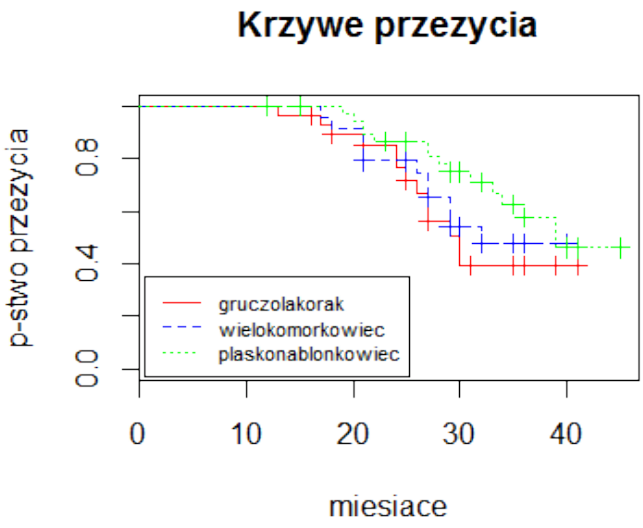
```
##   patient histpat survtime survind level stage
## 1      91       1       12       0     0     2
## 2      89       2       12       0     0     1
## 3      51       0       13       1     0     3
```

Dopasujmy najpierw zwykły model Kaplana-Meiera w zależności od zmiennej *level* odpowiadającej za poziom markera.



Z rysunku widać rzeczywiście, że krzywe raczej różnią się między sobą. Zanim jednak przeprowadzimy test formalny, przyjrzyjmy się analogicznej zależności, ale od zmiennej *stage* odpowiadającej za stopień zaawansowania nowotworu oraz od zmiennej *histpat* odpowiadającej za typ histologiczny nowotworu.





Wyraźnie widać, że zmienna *histpat* istotna nie jest, mamy zaś podejrzenie, że zmienna *stage* istotna być już może. Żeby nie popełnić błędu wielokrotnego testowania, na podstawie powyższych przypuszczeń, przeprowadzę warstwowy (ze względu na zmienną *stage*) test $log-rank$ dla dwóch rodzajów markera. Oto wydruk R-a dla naszego testu:

```
## Call:
## survdiff(formula = Surv(survtime, survind) ~ level + strata(stage),
##     data = a)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## level=0 54       10     21.9      6.43      17.8
## level=1 40       28     16.1      8.70      17.8
##
##  Chisq= 17.8  on 1 degrees of freedom, p= 2.44e-05
```

$P-value$ testu (dla statystyki testowej wynoszącej $17,8$) jest równe $2,44e-05$, zatem mamy podstawy do odrzucenia hipotezy o równości funkcji przeżycia. Czyli rzeczywiście jest różnica w długości przeżycia w zależności od poziomu markera CYFRA-21 i ze względu na warstwowanie stopniem zaawansowania nowotworu. Oczekiwana liczba zgonów dla każdej z grup w każdej z warstw, będzie równa:

```
##                       TNM 1 TNM 2  TNM 3
## niski poziom markera   3.58  9.01  9.262
## wysoki poziom markera  1.42  3.99 10.738
```

Syntaks R-a

```
# zad.1

sz <- matrix(c(4, 5, 8, 10, 17, 17, 17, 19, 24, 34, 34, 7, 8, 8, 8, 8, 8, 8,
    11, 11, 12, 12, 15, 16, 18, 21, 23, 24, 25, 27, 1, 1, 1, 1, 0, 0, 0, 1,
    0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, rep(1,
        11), rep(2, 19), 2, 3, 1, 3, 2, 1, 1, 1, 1, 1, 2, 1, 1, 2, 3, 1, 2,
    2, 3, 2, 1, 2, 2, 1, 1, 1, 2, 1, 3, 1), nrow = 30)
sz <- as.data.frame(sz)
names(sz) <- c("czas", "zdarzenie", "rodzaj.terapii", "wiek")
sz.KM <- survfit(Surv(czas, zdarzenie) ~ rodzaj.terapii, data = sz, conf.type = "none")
plot(sz.KM, col = c("red", "blue"), lty = 1:2, main = "Krzywe przezycia", xlab = "miesiace",
    ylab = "p-stwo przezycia")
legend(0.4, 0.27, c("szczepionka 1", "szczepionka 2"), col = c("red", "blue"),
    lty = 1:2, cex = 0.7)
summary(sz.KM)
sz.test <- survdiff(Surv(czas, zdarzenie) ~ rodzaj.terapii, data = sz)
sz.test$exp
sz.test$var
stat_prost <- sum((sz.test$obs - sz.test$exp)^2/sz.test$exp)
p.val <- 1 - pchisq(stat_prost, 1)

# zad.2

sz.strata <- survdiff(Surv(czas, zdarzenie) ~ rodzaj.terapii + strata(wiek),
    data = sz)
sz.strata
sz.strata$exp
e <- rowSums(sz.strata$exp)
o <- rowSums(sz.strata$obs)
stat_prost <- sum((e - o)^2/e)
p.val <- 1 - pchisq(stat_prost, 1)

# zad.3

a <- read.csv2("C:\\Users\\Marta\\Desktop\\Marta\\studia\\rok4\\Biostatystyka\\1\\cyfra_short.csv",
    sep = ",")
head(a)
a.KM_level <- survfit(Surv(survtime, survind) ~ level, data = a, conf.type = "none")
plot(a.KM_level, col = c("red", "blue"), lty = 1:2, main = "Krzywe przezycia",
    xlab = "miesiace", ylab = "p-stwo przezycia")
legend(0.4, 0.27, c("niski poziom markera", "wysoki poziom markera"), col = c("red",
    "blue"), lty = 1:2, cex = 0.7)
a.KM_stage <- survfit(Surv(survtime, survind) ~ stage, data = a, conf.type = "none")
plot(a.KM_stage, col = c("red", "blue", "green"), lty = 1:3, main = "Krzywe przezycia",
    xlab = "miesiace", ylab = "p-stwo przezycia")
legend(0.4, 0.35, c("TNM 1", "TNM 2", "TNM 3"), col = c("red", "blue", "green"),
    lty = 1:3, cex = 0.7)
a.KM_histpat <- survfit(Surv(survtime, survind) ~ histpat, data = a, conf.type = "none")
plot(a.KM_histpat, col = c("red", "blue", "green"), lty = 1:3, main = "Krzywe przezycia",
    xlab = "miesiace", ylab = "p-stwo przezycia")
legend(0.4, 0.35, c("gruczolakorak", "wielokomorkowiec", "plaskonablonkowiec"),
    col = c("red", "blue", "green"), lty = 1:3, cex = 0.7)
a.KM <- survdiff(Surv(survtime, survind) ~ level + strata(stage), data = a)
a.KM
tt <- a.KM$exp
rownames(tt) <- c("niski poziom markera", "wysoki poziom markera")
colnames(tt) <- c("TNM 1", "TNM 2", "TNM 3")
```