Lesson 5:
Statistical Testing

Recap: Bivariate Statistics

- correlation vs. linear regression

Important commands (recap)

COMMAND	EFFECT
plot(var1, var2)	plot scatterplot of two continuous variables
scatterplot()	package `car`, advances scatterplot
cor(var1, var2)	package `stats`, calculate correlation between two variables
cor.test()	package `stats`, test the correlation for significance
rcorr()	package `Hmisc`, calculates correlation also for more than one variable, adds p values
abline()	add linear regression line to scatterplot
lm()	package `stats`, calculates linear regression model
summary()	summarizes linear regression model

Statistical Testing

1. continuous variables

- difference between 2 groups: t Test
- difference between > 2 groups: ANOVA

2. categorical variables

- chi-square Test / Fisher exact Test

3. survival data: log-rank Test

1. continuous: 2 groups – t Test

assumptions:

- is data normally distributed?
- load package `stats`

```
> shapiro.test(data$variable)
```

```
        Shapiro-Wilk normality test

data:  data$cariable
W = 0.92449, p-value < 2.2e-16
```

if p-value below 0.05 → data is NOT normally distributed!

1. continuous: 2 groups – t Test

assumptions:

- is data normally distributed?

→ YES: t Test (package `stats`)

```
> t.test(group_1$var, group_2$var)

        Welch Two Sample t-test

data: group_1$var and group_2$var
t = 0.14342, df = 915.79, p-value = 0.886
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -104.3638  120.8202
sample estimates:
mean of x mean of y
 1674.243  1666.014
```

1. continuous: 2 groups – Wilcoxon Test

assumptions:

- is data normally distributed?

→ NO: Wilcoxon Test(package `stats`)

```
> wilcox.test(group_1$var, group_2$var)

	Wilcoxon rank sum test with continuity correction

data:  group_1$var and group_2$var
W = 108542, p-value = 0.8349
alternative hypothesis: true location shift is not equal to 0
```

1. continuous: > 2 groups – ANOVA

assumptions:

- is data normally distributed? (`shapiro.test()`)
- are variances equal? → Levene Test (package `car`)

```
> leveneTest(data$variable, data$categories, center = mean)

Levene's Test for Homogeneity of Variance (center = mean)
       Df F value  Pr(>F)
group   2  2.7311 0.06567 .
      926
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

if p-value below 0.05 → groups do NOT have same variance!

1. continuous: > 2 groups – ANOVA

assumptions:

- is data normally distributed? → YES
- are variances equal? → YES

→ ANOVA

```
> summary(aov(variable ~ categories, data = data))

              Df    Sum Sq Mean Sq F value   Pr(>F)
categories     2   7540308 3770154       5 0.00692 **
Residuals    926 698256213  754056
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. continuous: > 2 groups – ANOVA

assumptions:

- is data normally distributed? → NO

→ Kruskal-Wallis Test

```
> kruskal.test(variable ~ categories, data = data)

        Kruskal-Wallis rank sum test

data:  variable by categories
Kruskal-Wallis chi-squared = 9.2014, df = 2, p-value = 0.01004
```

1. continuous: > 2 groups – ANOVA

assumptions:

- are variances equal? → NO

→ Welch Test

```
> oneway.test(variable ~ categories, data = data)

        One-way analysis of means (not assuming equal variances)

data:  variable and categories
F = 5.0458, num df = 2.0, denom df = 616.8, p-value = 0.006704
```

2. categorical: χ^2 Test

- use `chisq.test()` from package `stats`

```
> chisq.test(data$cat_1, data$cat_2)

        Pearson's Chi-squared test with Yates' continuity correction

data:  data$cat_1 and data$cat_2
X-squared = 1.511, df = 1, p-value = 0.219
```

2. categorical (small samples): Fisher Test

- for very small samples use Fisher-exact Test:

```
fisher.test(data$cat_1, data$cat_2)
```

3. survival data: Log-rank Test

- use `survdiff()` from package `survival`

```
> p=survdiff(Surv(time, status == 1) ~ sex, data=test1)
> p
Call:
survdiff(formula = Surv(time, status == 1) ~ sex, data = test1)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=0 445      215      217   0.01049    0.0202
sex=1 484      237      235   0.00965    0.0202

 Chisq= 0  on 1 degrees of freedom, p= 0.9
> pchisq(p$chisq, length(p$n)-1, lower.tail = FALSE)
[1] 0.8871178
> 
```

3. survival data: Log-rank Test

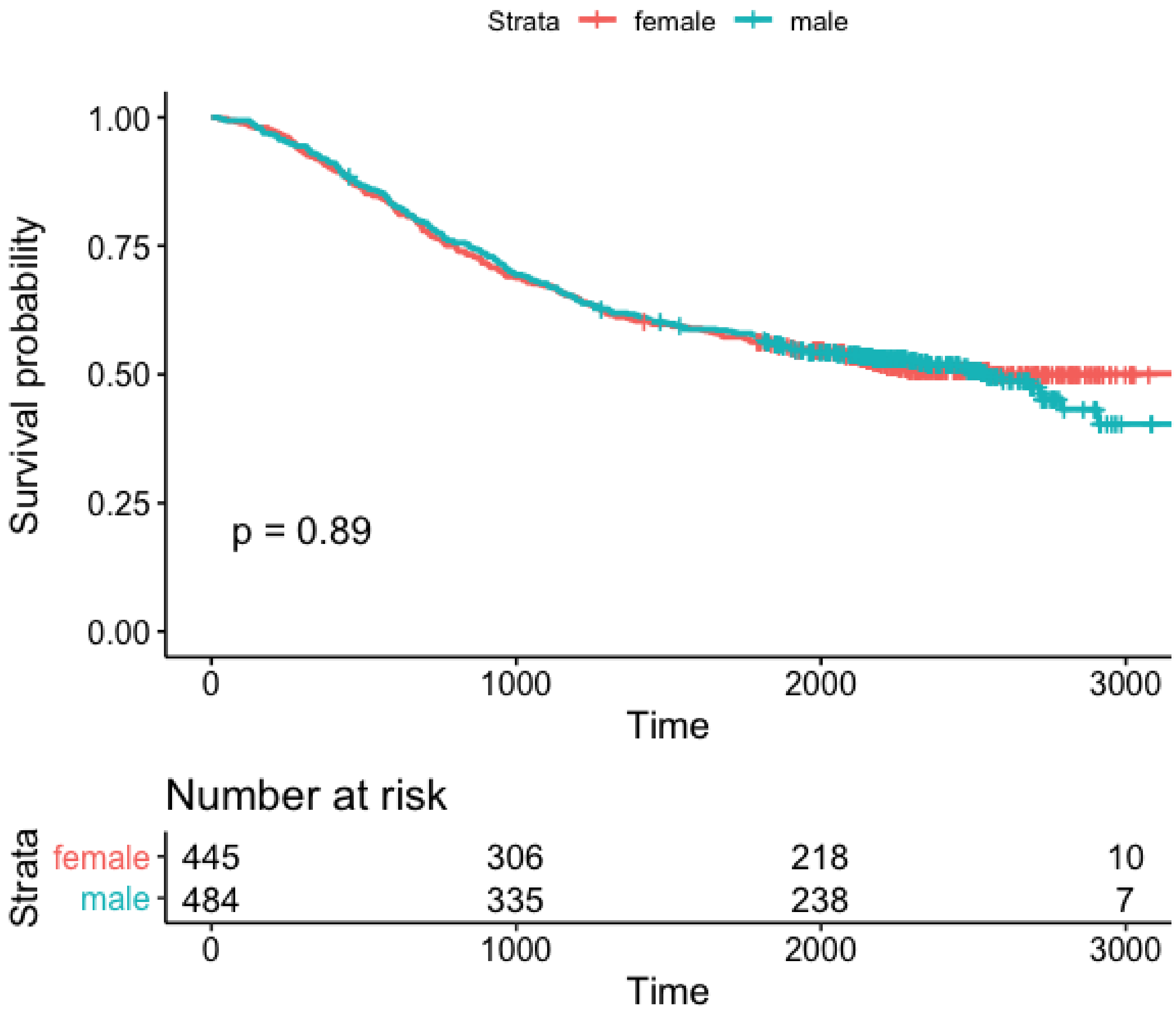
- requires packages `survival` and `survminer`

```
> fit <- survfit(Surv(data$cont_var, data$status == 1) ~ cat_var, data = data)
> summary(fit)

Call: survfit(formula = Surv(colon$time, colon$status == 1) ~ sex, data = colon)

                cat_var=0
 cont_var n.risk n.event survival std.err lower 95% CI upper 95% CI
   23      445       1    0.998   0.00224      0.993        1.000
   52      444       1    0.996   0.00317      0.989        1.000
   56      443       1    0.993   0.00388      0.986        1.000

   ...
```

3. survival data: Log-rank Test

- requires packages `survival` and `survminer`

```
> fit <- survfit(Surv(data$cont_var, data$status == 1) ~ cat_var, data = data)
> ggsurvplot(fit, data = colon,
             pval = TRUE,
             risk.table = TRUE,
             legend.labs = c("female", "male"))
```



Strata
female
male
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.89
0
1000
2000
3000
Time
Number at risk
Strata
female
male
445
306
218
10
484
335
238
7
0
1000
2000
3000
Time

Important commands

COMMAND	PACKAGE	EFFECT
shapiro.test(data$variable)	stats	tests variable for normal distribution
t.test(group_1$var, group_2$var)	stats	tests two groups with normally distributed data for equality
wilcox.test(group_1$var, group_2$var)	stats	tests two groups with not normally distributed data for equality
summary(**aov**(variable ~ categories, data = data))	stats	tests more than two groups split by `categories` for equality
leveneTest(data$variable, data$categories, center = mean)	car	tests a `variable` split into groups by `categories` for variance homogeneity
oneway.test(variable ~ categories, data = data)	stats	tests more than two groups where variances are unequal for equality
kruskal.test(variable ~ categories, data = data)	stats	tests more than two groups that are not normally distributed for equality
chisq.test(data$cat_1, data$cat_2)	stats	tests two categorical variables
fisher.test(data$cat_1, data$cat_2)	stats	tests two categorical variables with small n
Surv(data$cont_var, data$status == 1)	survival	
survfit(**survfunction** ~ cat_var, data = data)	survival	
ggsurvplot(fit, data = colon, pval = TRUE, risk.table = TRUE, legend.labs = c("female", "male"))	survminer	