

I10 - Multiple comparisons

STAT 587 (Engineering)
Iowa State University

March 30, 2021

Mice diet effect on lifetimes

Female mice were randomly assigned to six treatment groups to investigate whether restricting dietary intake increases life expectancy. Diet treatments were:

- NP - mice ate unlimited amount of nonpurified, standard diet
- N/N85 - mice fed normally before and after weaning. After weaning, ration was controlled at 85 kcal/wk
- N/R50 - normal diet before weaning and reduced calorie diet (50 kcal/wk) after weaning
- R/R50 - reduced calorie diet of 50 kcal/wk both before and after weaning
- N/R50 lopro - normal diet before weaning, restricted diet (50 kcal/wk) after weaning and dietary protein content decreased with advancing age
- N/R40 - normal diet before weaning and reduced diet (40 Kcal/wk) after weaning.

Exploratory analysis

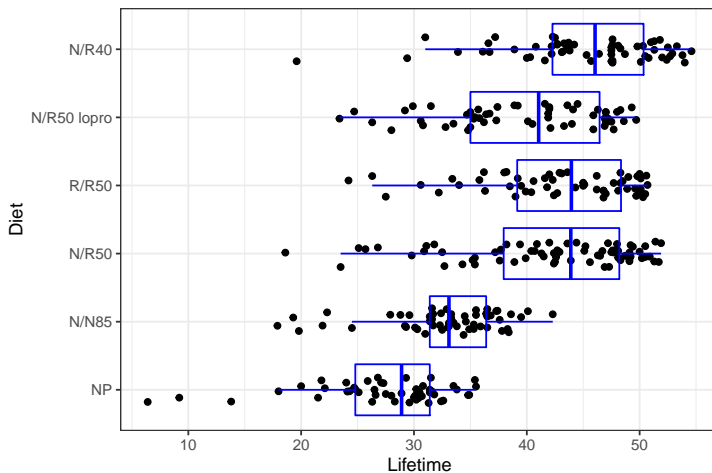
```
library("Sleuth3")
# head(case0501)
summary(case0501)
```

	Lifetime	Diet
Min.	: 6.4	N/N85:57
1st Qu.:	31.8	N/R40:60
Median :	39.5	N/R50:71
Mean	:38.8	NP :49
3rd Qu.:	46.9	R/R50:56
Max.	:54.6	lopro:56

```
case0501 <- case0501 %>%
  mutate(Diet = factor(Diet, c("NP", "N/N85", "N/R50", "R/R50", "lopro", "N/R40")),
         Diet = recode(Diet, lopro = "N/R50 lopro"))
case0501 %>% group_by(Diet) %>% summarize(n=n(), mean = mean(Lifetime), sd = sd(Lifetime))
```

```
# A tibble: 6 x 4
  Diet      n mean  sd
<fct>   <int> <dbl> <dbl>
1 NP         49  27.4  6.13
2 N/N85       57  32.7  5.13
3 N/R50       71  42.3  7.77
4 R/R50       56  42.9  6.68
5 N/R50 lopro  56  39.7  6.99
6 N/R40       60  45.1  6.70
```

```
ggplot(case0501, aes(x=Diet, y=Lifetime)) +
  geom_jitter(width=0.2, height=0) +
  geom_boxplot(fill=NA, color='blue', outlier.color = NA) +
  coord_flip() +
  theme_bw()
```



Are the data compatible with a common mean?

Let Y_{ij} represent the lifetime of mouse j in diet i for $i = 1, \dots, I$ and $j = 1, \dots, n_i$. Assume $Y_{ij} \stackrel{\text{ind}}{\sim} N(\mu_i, \sigma^2)$ and calculate a pvalue for $H_0 : \mu_i = \mu$ for all i .

```
bartlett.test(Lifetime ~ Diet, data = case0501)
```

Bartlett test of homogeneity of variances

data: Lifetime by Diet

Bartlett's K-squared = 10.996, df = 5, p-value = 0.05146

```
oneway.test(Lifetime ~ Diet, data = case0501, var.equal = TRUE)
```

One-way analysis of means

data: Lifetime and Diet

F = 57.104, num df = 5, denom df = 343, p-value < 2.2e-16

```
oneway.test(Lifetime ~ Diet, data = case0501, var.equal = FALSE)
```

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

data: Lifetime and Diet

F = 64.726, num df = 5.00, denom df = 157.84, p-value < 2.2e-16

Statistical testing errors

Definition

A **type I error** occurs when a true null hypothesis is rejected.

Definition

A **type II error** occurs when a false null hypothesis is not rejected. **Power** is one minus the type II error probability.

We set our significance level α to control the type I error probability. If we set $\alpha = 0.05$, then we will incorrectly reject a true null hypothesis 5% of the time.

Statistical testing errors

Decision	Truth	
	H_0 true	H_0 false
H_0 not true	Type I error	Correct (power)
H_0 true	Correct	Type II error

Definition

The **familywise error rate** is the probability of rejecting at least one true null hypothesis.

Type I error for all pairwise comparisons of J groups

How many combinations when choosing 2 items out of J ?

$$\binom{J}{2} = \frac{J!}{2!(J-2)!}.$$

If $J = 6$, then there are 15 different comparison of means. If we set $\alpha = 0.05$ as our significance level, then individually each test will only incorrectly reject 5% of the time.

If we have 15 tests and use $\alpha = 0.05$, what is the familywise error rate?

$$1 - (1 - 0.05)^{15} = 1 - (0.95)^{15} = 1 - 0.46 = 0.54$$

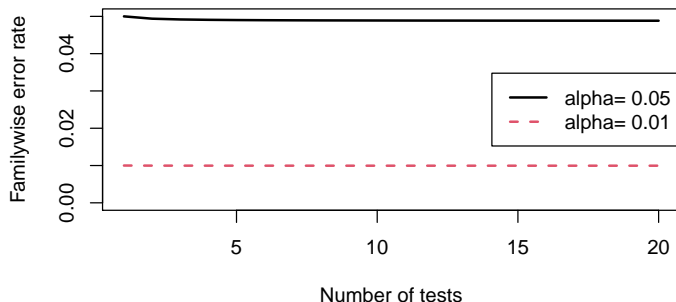
So there is a greater than 50% probability of falsely rejecting at least one true null hypothesis!

Bonferroni correction

Definition

If we do m tests and want the familywise error rate to be α , the **Bonferroni correction** uses α/m for each individual test. The familywise error rate, for independent tests, is $1 - (1 - \alpha/m)^m$.

Bonferroni familywise error rate



Pairwise comparisons

If we want to consider all pairwise comparisons of the average lifetimes on the 6 diets, we have 15 tests. In order to maintain a familywise error rate of 0.05, we need a significance level of $0.05/15 = 0.0033333$.

```
pairwise.t.test(case0501$Lifetime, case0501$Diet, p.adjust.method = "none")
```

Pairwise comparisons using t tests with pooled SD

data: case0501\$Lifetime and case0501\$Diet

	NP	N/N85	N/R50	R/R50	N/R50	lopro
N/N85	5.9e-05	-	-	-	-	-
N/R50	< 2e-16	1.1e-14	-	-	-	-
R/R50	< 2e-16	8.9e-15	0.622	-	-	-
N/R50 lopro	< 2e-16	5.2e-08	0.029	0.012	-	-
N/R40	< 2e-16	< 2e-16	0.017	0.073	1.6e-05	-

P value adjustment method: none

Pairwise comparisons

If we want to consider all pairwise comparisons of the average lifetimes on the 6 diets, we have 15 tests. Alternatively, you can let R do the adjusting for you, but now you need to compare with the original significance level α .

```
pairwise.t.test(case0501$Lifetime, case0501$Diet, p.adjust.method = "bonferroni")
```

Pairwise comparisons using t tests with pooled SD

data: case0501\$Lifetime and case0501\$Diet

	NP	N/N85	N/R50	R/R50	N/R50 lopro
N/N85	0.00089	-	-	-	-
N/R50	< 2e-16	1.6e-13	-	-	-
R/R50	< 2e-16	1.3e-13	1.00000	-	-
N/R50 lopro	< 2e-16	7.9e-07	0.44018	0.17507	-
N/R40	< 2e-16	< 2e-16	0.24881	1.00000	0.00024

P value adjustment method: bonferroni

Comments on the Bonferroni correction

The Bonferroni correction can be used in any situation. In particular, it can be used on unadjusted pvalues reported in an article that has many tests by comparing their pvalues to α/m where m is the number of tests they perform.

The Bonferroni correction is (in general) the **most** conservative multiple comparison adjustment, i.e. it will lead to the least null hypothesis rejections.

Constructing multiple confidence intervals

A $100(1 - \alpha)\%$ confidence interval should contain the true value $100(1 - \alpha)\%$ of the time when used with different data sets.

An error occurs if the confidence interval does not contain the true value.

Just like the Type I error and familywise error rate, we can ask what is the probability at least one confidence interval does not cover the true value.

The procedures we will talk about for confidence intervals have equivalent approaches for hypothesis testing (pvalues). Within these procedures we still have the equivalence between pvalues and CIs.

Constructing multiple confidence intervals

Confidence interval for the difference between group j and group j' :

$$\bar{Y}_j - \bar{Y}_{j'} \pm M s_p \sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}$$

where M is a multiplier that depends on the adjustment procedure:

Procedure	M	Use
LSD	$t_{n-J}(1 - \alpha/2)$	After significant F -test (no adjustment)
Dunnett	multivariate t	Compare all groups to control
Tukey-Kramer	$q_{J,n-J}(1 - \alpha)/\sqrt{2}$	All pairwise comparisons
Scheffé	$\sqrt{(J-1)F_{(J-1,n-J)}(1 - \alpha)}$	All contrasts
Bonferroni	$t_{n-J}(1 - (\alpha/m)/2)$	m tests (most generic)

Tukey for all pairwise comparisons

```
TukeyHSD(aov(Lifetime ~ Diet, data = case0501))
```

```
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = Lifetime ~ Diet, data = case0501)
```

```
$Diet
```

	diff	lwr	upr	p adj
N/N85-NP	5.2891873	1.5606269	9.0177476	0.0008380
N/R50-NP	14.8951423	11.3405719	18.4497127	0.0000000
R/R50-NP	15.4836735	11.7397556	19.2275913	0.0000000
N/R50 lopro-NP	12.2836735	8.5397556	16.0275913	0.0000000
N/R40-NP	17.7146259	14.0294069	21.3998448	0.0000000
N/R50-N/N85	9.6059550	6.2021702	13.0097399	0.0000000
R/R50-N/N85	10.1944862	6.5934168	13.7955556	0.0000000
N/R50 lopro-N/N85	6.9944862	3.3934168	10.5955556	0.0000008
N/R40-N/N85	12.4254386	8.8854359	15.9654413	0.0000000
R/R50-N/R50	0.5885312	-2.8320696	4.0091319	0.9963976
N/R50 lopro-N/R50	-2.6114688	-6.0320696	0.8091319	0.2460200
N/R40-N/R50	2.8194836	-0.5367684	6.1757356	0.1564608
N/R50 lopro-R/R50	-3.2000000	-6.8169683	0.4169683	0.1167873
N/R40-R/R50	2.2309524	-1.3252222	5.7871269	0.4684413
N/R40-N/R50 lopro	5.4309524	1.8747778	8.9871269	0.0002306

False Discovery Rate

Not wanting to make a single mistake is pretty conservative.
In high-throughput fields a more common multiple comparison adjustment is false discovery rate.

Definition

False discovery rate procedures try to control the expected proportion of incorrectly rejected null hypotheses.

How to incorporate multiple comparison adjustments

1. Determine what tests are going to be run (before looking at the data) or what confidence intervals are going to be constructed.
2. Determine which multiple comparison adjustment is the most relevant.
3. Use/state that adjustment and interpret your results.