Corpus Statistics
2008/12/11
Baayen (2008: pp. 105–111)

Toru Hitomi

4.4 A numerical vector and a factor: analysis of variance (*continued*)

- 因子が 3 つ以上の水準をもつ場合→**多重比較(multiple comparisons)**

コメント [T1]: p 値は 3 つのうちのどこか 1 つで surprise

e.g. 3 つのグループの平均を比較したい時、t 検定(α = 0.05)を 3 回繰り返しては？

```
> 1 - pbinom (0, 3, 0.05)        #起こる確率が 0.05 のものを 3 回試行し、少なくとも 1 回は起こる確率
[1] 0.142625          ← inflation in surprise
```

それに対して、、、

- **Bonferroni correction(ボンフェローニ補正)**

n 回の比較に対して、単純に有意確率 α を n で割る → $\frac{\alpha}{n}$

e.g. データセット Aux は 3 水準、組み合わせ 3 回の比較(4 水準なら、組み合わせ 6 回 $p = 0.0083$)

  - n = 3, $p = \frac{0.05}{3} = 0.0167$

- **Tukey honestly significant difference(テューキーの HSD 検定)**
  - R では、TukeyHSD()を用いる。
  - 利点:Bonferroni 法よりも、検定力が強い。
  - 欠点:比較するサンプルサイズが等しいという前提条件がある。
    →多少の違いなら、等価することで意味のある結果を出すことが可能

コメント [T2]: これでは、意味のない結果になってしまう。

e.g. データセットAuxの場合は、水準間のサンプル数に大きな差異があるためにTukey HSDではなく、Bonferroni法を用いる。

e.g. データセット warpbreaks を用いて、一元配置分散分析を行う。

因子:breaks(numerical), wool(2 水準), & tension(3 水準)

```
> warpbreaks.lm = lm(breaks ~ tension, data = warpbreaks)
> anova(warpbreaks.lm)
Analysis of Variance Table
```

Response: breaks

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tension	2	2034.3	1017.1	7.2061	0.001753 **
Residuals	51	7198.6	141.1		

```
> summary(warpbreaks.lm)
lm(formula = breaks ~ tension, data = warpbreaks)
```

Residuals:

Min	1Q	Median	3Q	Max
−22.389	−8.139	−2.667	6.333	33.611

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.39	2.80	12.995	< 2e−16 ***
tensionM	−10.00	3.96	−2.525	0.014717 *
tensionH	−14.72	3.96	−3.718	0.000501 ***

```
Residual standard error: 11.88 on 51 degrees of freedom
Multiple R-squared: 0.2203,     Adjusted R-squared: 0.1898
F-statistic: 7.206 on 2 and 51 DF,  p-value: 0.001753
```

- 糸の張り具合の medium・high と low との間で有意な差がある。

- aov()を用いて、Tukey HSD を利用してみる。

```
> warpbreaks.aov = aov (breaks ~ tension, data = warpbreaks)
> summary (warpbreaks.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tension	2	2034.3	1017.1	7.2061	0.001753 **
Residuals	51	7198.6	141.1		

```
> 1 - pf (7.206, 2, 51)
 [1] 0.001752972
```

- 結果は、summary(warpbreaks.lm)と同じ。

```
> TukeyHSD(warpbreaks.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = breaks ~ tension, data = warpbreaks)
$tension
```

	diff	lwr	upr	p adj
M-L	-10.000000	-19.55982	-0.4401756	0.0384598
H-L	-14.722222	-24.28205	-5.1623978	0.0014315
H-M	-4.722222	-14.28205	4.8376022	0.4630831

summary(warpbreaks.lm)
よりも保守的！

```
> plot(TukeyHSD(warpbreaks.aov))        #作図 (Figure 4.15, p. 108)
```

- lm()と aov()は根底では同じなので、両方使うことに意味はなく、さらに、Aux のサンプルの等分散性がそもそも満たされていないので、問題！

```
> tapply(auxiliaries$VerbalSynsets, auxiliaries$Aux, var)
```

hebben	zijn	zijnheb
5.994165	18.066667	11.503932

- ノンパラメトリックである、**Kruskal-Wallis rank sum test(クラスカル・ウォリス検定)**を用いる。

```
> kruskal.test(auxiliaries$VerbalSynsets, auxiliaries$Aux)

        Kruskal-Wallis rank sum test

data:  auxiliaries$VerbalSynsets and auxiliaries$Aux
Kruskal-Wallis chi-squared = 11.7206, df = 2, p-value = 0.002850
```

4.4.1 Two numerical vectors and a factor: Analysis of covariance

これまでは：
線型回帰 → numerical vector が予測変数
分散分析 → 因子(複数の水準)が、予測変数

2 つの numerical vectors と 1 つの因子で予測＝**analysis of covariance**

- 線型回帰、分散分析、共分散分析(analysis of covariance, ANCOVA)はすべて、lm()を用いる。

コメント [T3]: 根底は同じ原理

```
> ratings.lm = lm(meanSizeRating ~ meanFamiliarity * Class + I(meanFamiliarity^2), data = ratings)
> summary (ratings.lm)

lm(formula = meanSizeRating ~ meanFamiliarity * Class + I(meanFamiliarity^2), data = ratings)

Residuals:
      Min        1Q    Median        3Q       Max
-0.97325 -0.19146  0.03541  0.19628  0.90686
```

Corpus Statistics
2008/12/11
Baayen (2008: pp. 105–111)

Toru Hitomi

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)		説明
(Intercept)	4.42894	0.54787	8.084	7.6e−12 ***	…	modified group mean
meanFamiliarity	−0.63131	0.29540	−2.137	0.03580 *	…	linear term coefficient
I(meanFamiliarity^2)	0.10971	0.03801	2.886	0.00508 **	…	quadric term coefficient
Classplant	−1.01248	0.41530	−2.438	0.01711 *	…	plant の group mean にするために coefficient 分引く →−0.631−0.212=−0.843
meanFamiliarity:Classplant	−0.21179	0.09779	−2.166	0.03346 *	…	interaction

コメント [T4]: 両クラスでシェア＝傾き同じで、分布の高さが違う

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3424 on 76 degrees of freedom
Multiple R-squared: 0.8805, Adjusted R-squared: 0.8742
F-statistic: 140 on 4 and 76 DF, p-value: < 2.2e-16

- Figure 4.16 (p. 111) を作図する。

```
> ratings$fitted = fitted (ratings.lm)        #適合値を各名詞に列として付与

> plot(ratings$meanFamiliarity, ratings$meanSizeRating, xlab="mean familiarity", ylab="mean size rating",
+ type="n")              #枠作り
> text (ratings$meanFamiliarity, ratings$meanSizeRating, substr(as.character(ratings$Class), 1, 1),
+ col="darkgrey")     #頭文字 (i.e., a と p) の挿入

> plants = ratings[ratings$Class == "plant",]
> animals = ratings[ratings$Class == "animal",]

> plants = plants[order(plants$meanFamiliarity),]
> animals = animals[order(animals$meanFamiliarity),]

> lines(plants$meanFamiliarity, plants$fitted)
> lines(animals$meanFamiliarity, animals$fitted)
```

Class それぞれのデータセットに分割し、meanFamiliarity 順に並べ替える。