Lecture 32 ANOVA & *Post hoc* Multiple Comparisons

BIO210 Biostatistics

Xi Chen

Fall, 2024

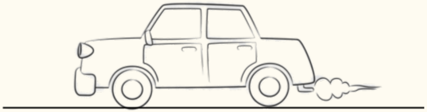
School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Stopping Distance of A Car - Data

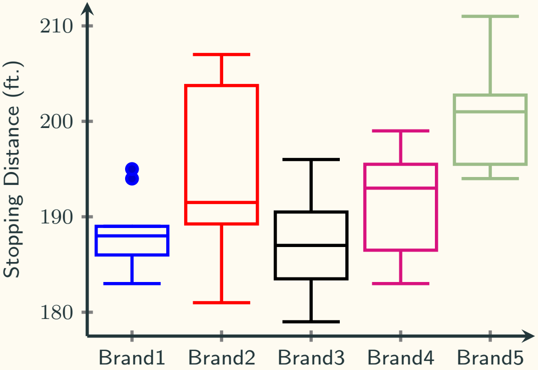
A researcher for an automobile safety institute was interested in determining whether or not the distance that it takes to stop a car going 60 miles per hour depends on the brand of the tire. The researcher measured the stopping distance (in feet) of ten randomly selected cars for each of five different brands. The researcher arbitrarily labeled the brands of the tires as Brand1, Brand2, Brand3, Brand4, and Brand5, so that he and his assistants would remain blinded. Here are the data resulting from his experiment:

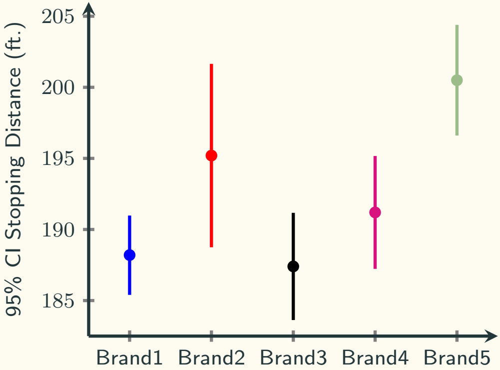
Brand1	Brand2	Brand3	Brand4	Brand5
194	189	185	183	195
184	204	183	193	197
189	190	186	184	194
189	190	183	186	202
188	189	179	194	200
186	207	191	199	211
195	203	188	196	203
186	193	196	188	206
183	181	189	193	202
188	206	194	196	195



Stopping Distance of A Car - Descriptive Stats

	Brand1	Brand2	Brand3	Brand4	Brand5
n	10	10	10	10	10
Mean	188.2	195.2	187.4	191.2	200.5
Var	15.06	81.29	27.82	30.84	29.61





Stopping Distance of A Car - The ANOVA Table

	Brand1	Brand2	Brand3	Brand4	Brand5
n	10	10	10	10	10
Mean	188.2	195.2	187.4	191.2	200.5
Var	15.06	81.29	27.82	30.84	29.61

Source of Variation	SS	*df*	MS	F	*p*-value
Between	1174.8	4	293.7		
Within	1161.7	45	36.9	7.95	6.17×10^{-5}
Total	2836.5	49	[redacted]		

Assumptions When Using ANOVA

- Randomness, Independence
- Population normally distributed $\left(F = \dfrac{\mathrm{MSB}}{\mathrm{MSW}}\right)$
- Different groups have equal variance (classical ANOVA)

$$\mathrm{MSW} = \frac{\mathrm{SSW}}{n-k} = \frac{df_1 \cdot s_1^2 + df_2 \cdot s_2^2 + \cdots + df_k \cdot s_k^2}{n-k} = \frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2 + \cdots + (n_k-1)\cdot s_k^2}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)}$$

- Unequal variance: Welch's ANOVA

The Relation Between F-test and t-test

- **Think:** What if the ANOVA method, i.e. using SSB, SSW and the F statistic, is used to compare means from two groups? Valid, or not ?

- $\boldsymbol{t}$-test statistic with equal variance:

$$\boldsymbol{t} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \ \nu = n_1 + n_2 - 2\,, \ s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- The ANOVA Table When $k = 2$

Source of Variation	**SS**	***df***	**MS**	**F**
Between	$n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2$	1	SSB	$\frac{SS_B(n_1 + n_2 - 2)}{SS_W}$
Within	$(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$	$n_1 - 1 + n_2 - 1$	$\frac{\text{SSW}}{n_1 + n_2 - 2}$	
Total	SSB + SSW	$n - 1$		

F-test vs t-test When There Are Two Groups

- **Example:** Brand 3 ($\bar{x}_1 = 187.4$, $s_1^2 = 27.82$) vs. Brand 4 ($\bar{x}_2 = 191.2$, $s_2^2 = 30.84$)
- $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -1.57$, $p = \mathbb{P}\left(|t| \geqslant 1.57\right) = 2 \times \mathbb{P}\left(t \leqslant -1.57\right) = 0.134$
- $F_{1,18} = \dfrac{\text{MSB}}{\text{MSW}} = 2.46$, $p = \mathbb{P}\left(F \geqslant 2.46\right) = 0.134$

Post hoc Tests

- ANOVA test tells me to reject $H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k$, so what ?

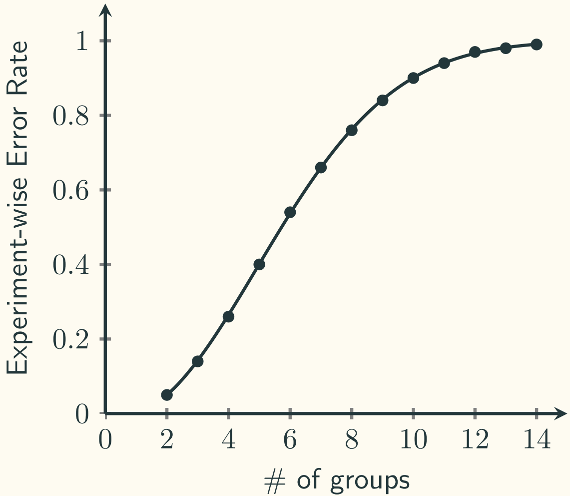
- *Post hoc* tests - multiple pairwise comparisons. The following commonly-used tests have different ways of controlling type I error rate:
  - Bonferroni Procedure
  - Duncan's new multiple range test (MRT)
  - Dunn's Multiple Comparison Test
  - Fisher's Least Significant Difference (LSD)
  - Holm-Bonferroni Procedure
  - Newman-Keuls
  - Rodger's Method
  - Scheffé's Method
  - Tukey's Test (often used in classical ANOVA in stats software)
  - Dunnett's correction
  - Benjamini-Hochberg (BH) procedure

Post hoc Tests

Pairwise comparison $\alpha = 0.05$

# of groups	# of comparisons	Probability of making at least one type I error
2	1	0.05
3	3	0.14
4	6	0.26
5	10	0.4
6	15	0.54
7	21	0.66
8	28	0.76
9	36	0.84
10	45	0.9
11	55	0.94
12	66	0.97
13	78	0.98
14	91	0.99

$$1 - (1 - \alpha)^c$$



The Bonferroni Procedure

Pairwise comparison $\alpha = 0.05$: not good enough!

Goal: when doing many comparisons, we want the overall error rate to be α, meaning that the probability of making at least one type I error after performing all the comparisons is α.

$$1 - (1 - \alpha^*)^c = \alpha\,, \text{ where } c = \binom{k}{2}$$

Note, when α^* is small: $(1 - \alpha^*)^c \approx 1 - c\alpha^*$. We have:

$$1 - (1 - c\alpha^*) \approx \alpha \quad \Rightarrow \quad c\alpha^* \approx \alpha \quad \Rightarrow \quad \alpha^* \approx \frac{\alpha}{c} = \frac{\alpha}{\binom{k}{2}}$$

Bonferroni correction

Named after Carlo Emilio Bonferroni

The Bonferroni Procedure

To control the experiment-wise error rate to be α, we need to let the significance level α^* in each of the pairwise comparison to be α/c, where c is the # of comparison.

For each comparison, if the $p < \alpha^*$, then H_0 is rejected.	$\Leftrightarrow$	If $p < \alpha/c$, then H_0 is rejected.	$\Leftrightarrow$	If $p \times c < \alpha$, then H_0 is rejected.

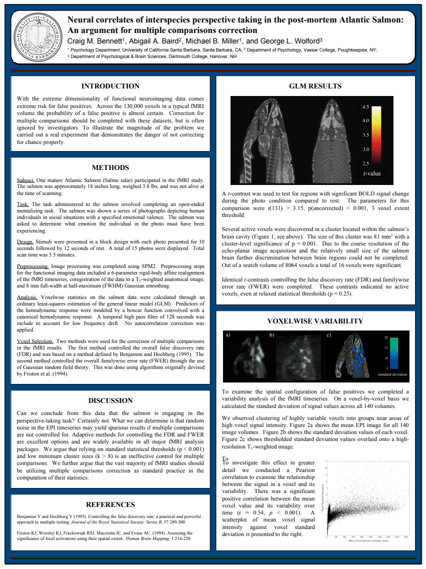
Corrected p-value

$padj$

$p.adj$

$$p.adj = min\left[p \times \binom{k}{2},\ 1\right]$$, if $p.adj < \alpha$, then H_0 is rejected.

Multiple Comparisons - The Salmon Test



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]
[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY;
[3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a T_1-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

Analysis. Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was applied.

Voxel Selection. Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

DISCUSSION

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

GLM RESULTS

A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

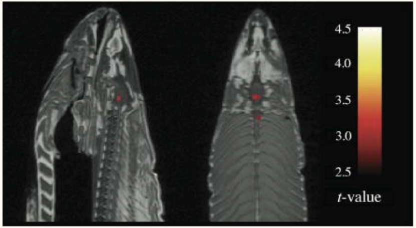
Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

VOXELWISE VARIABILITY

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T_1-weighted image.

To
To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.



Multiple Comparisons - Significant

https://xkcd.com/882/