Multiple samples: Pairwise comparisons and categorical outcomes

Patrick Breheny

April 26

Introduction

- In the previous lecture, we saw how one could use ANOVA with the tailgating study to test the hypothesis that the average following distances in all four of the groups were the same
- There was strong evidence ($p = 0.006$ using the rank transformation) that this was not the case
- We then looked at the estimated means and confidence intervals for the average quantile in each group and remarked that it looked like the MDMA group had much lower following distances than the other three groups, but that the alcohol, marijuana, drug-free groups were about the same

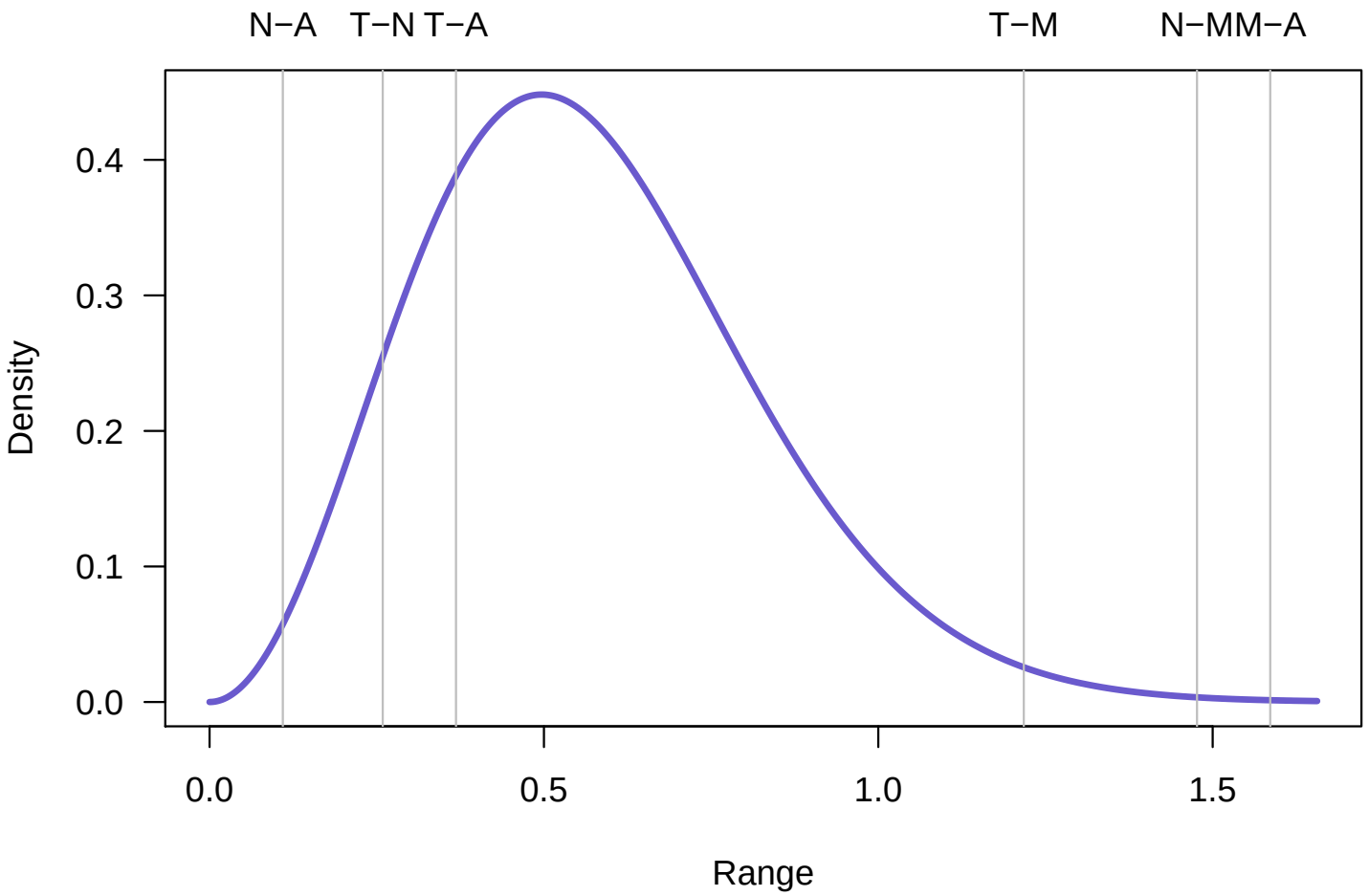
Pairwise comparisons

- In many ways, this is fine – our primary analysis determined that there was a difference among the means, and the rest is just commentary about which of those differences are most substantial
- However, it is sometimes desirable to have a formal, objective criterion for deciding which pairs are significantly different from each other
- One approach would be to carry out all 6 pairwise comparisons with a Bonferroni correction to the significance levels

Tukey's range distribution

- A somewhat more powerful approach, however, is to use Tukey's range test (also known as Tukey's "Honest Significant Difference" test and the Tukey-Kramer method)
- Tukey's idea was to focus on the distribution of the largest difference (i.e., the range) in the means of multiple groups, and he worked out mathematical expressions for the distribution of the range when comparing K sample means, all of which have the same population mean
- With these expressions, we can set a threshold of, say, the 95th percentile of the range distribution; doing so directly controls the family-wise error rate as there is only a 5% chance that a single pairwise difference will exceed this bound

Tukey's range distribution



Adjusted p-values

Calculating tail areas, we arrive at adjusted p-values[1]:

	Difference in means[2]	Adjusted p-value
MDMA-ALC	0.29	0.01
MDMA-NODRUG	0.27	0.01
MDMA-THC	0.22	0.04
THC-ALC	0.07	0.78
THC-NODRUG	0.05	0.87
ALC-NODRUG	-0.02	0.99

[1] The range distribution on the previous page was for $n = 30$ in each group. In the actual study, some groups had more than 30 and others less than 30, which is properly accounted for in the above p-values

[2] Difference in average quantiles, since we carried out a rank transformation

Tukey's HSD in R

To obtain the output on the previous page (plus confidence intervals) in R, we can submit

```
fit <- aov(y ~ Group, Data)
TukeyHSD(fit)
```

Bonferroni, Tukey, and Unadjusted p-values

	Difference in means	p	Adjusted p-values Bonferroni	Tukey
MDMA-ALC	0.29	0.002	0.01	0.01
MDMA-NODRUG	0.27	0.001	0.01	0.01
MDMA-THC	0.22	0.008	0.05	0.04
THC-ALC	0.07	0.349	1.00	0.78
THC-NODRUG	0.05	0.448	1.00	0.87
NODRUG-ALC	-0.02	0.780	1.00	0.99

Testosterone treatment for low libido

- Let's look at another example of a multiple group study, this one from 2008 and examining the use of testosterone as a therapy for diminished libido in postmenopausal women
- In the study, women were randomly assigned to three groups: one group received a patch delivering 300 μg of testosterone per day, one group received a patch delivering 150 μg of testosterone per day, and one group was assigned to placebo (a patch delivering no testosterone)
- Potentially, we are interested in three different comparisons here: comparing each of the two treatments to the placebo, as well as comparing the two testosterone treatments to each other

Outcome: Number of satisfying episodes

As a primary outcome, the authors looked at the number of satisfying sexual episodes over the final four-week-period of the study:

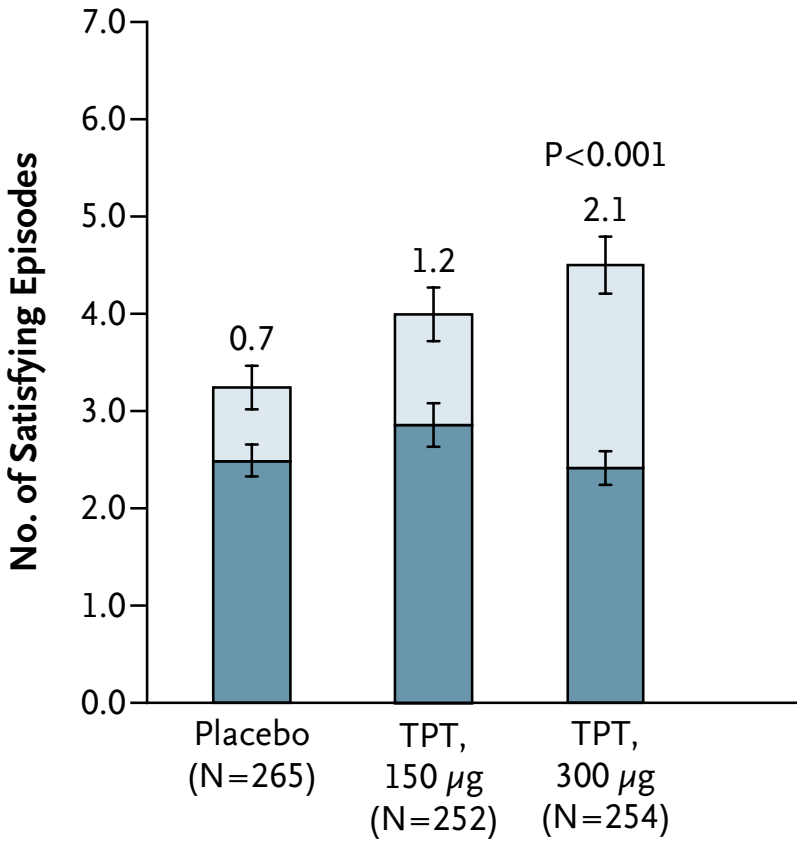


Figure from Davis, et al. (2008), "Testosterone for Low Libido in Postmenopausal Women Not Taking Estrogen", *N Engl J Med*, **359**:2005-17:

ANOVA

- Performing an ANOVA comparing the three-mean model and the common-mean model:

$$\text{RSS}_0 = 9,668$$
$$\text{RSS}_1 = 9,408$$

- The three-mean model explains 2.7% of the variability in frequency of satisfying episodes:

$$\frac{9,668 - 9,408}{9,668} = .027$$

- A total of 771 women completed the study, so $\hat{\sigma}^2 = 12.25$, $F = 10.6$, and $p = 0.00003$

Estimates (and CIs) for difference from baseline

Difference from baseline four-week frequency of satisfying sexual episodes:

Significance of pairwise comparisons

	Difference	p	p_{Tukey}
High-Placebo	1.4	< 0.0001	< 0.0001
Low-Placebo	0.5	0.11	0.17
High-Low	0.9	0.01	0.01

Complications from testosterone therapy

- The study also looked at complications arising from testosterone therapy
- We'll take a closer look at two binary (yes/no) complications to get a sense of how to analyze multiple-sample studies involving categorical outcomes

Complications from acne

	Acne	
	Yes	No
Placebo	14	263
Low dose (150 μg)	15	252
High dose (300 μg)	16	251

Expected counts

- We can perform a test of the overall hypothesis that the complication rate is the same in all three groups by using a χ^2 test, which proceeds exactly as in the 2×2 case
- We begin by calculating expected counts under the null:

	Acne	
	Yes	No
Placebo	15.4	261.6
Low dose (150 μg)	14.8	252.2
High dose (300 μg)	14.8	251.2

χ^2 and Fisher tests

- The χ^2 test statistic here is 0.23
- For tables bigger than 2×2, we can still use a χ^2 distribution, but the degrees of freedom change; specifically, $df = (I-1)(J-1)$ where I and J are the number of rows and columns of the table
- Comparing $X^2 = 0.23$ to a χ^2 distribution with 2 df, we obtain $p = 0.89$
- Fisher's Exact Test can also be applied to larger tables, and yields a very similar result: 0.91

Increased hair growth

So there's no evidence that testosterone therapy leads to acne, but what about increased hair growth?

	Increased hair growth	
	Yes	No
Placebo	29	248
Low dose (150 μg)	31	236
High dose (300 μg)	53	214

Expected counts

	Increased hair growth	
	Yes	No
Placebo	38.6	238.4
Low dose (150 μg)	37.2	229.8
High dose (300 μg)	37.2	229.8

$X^2 = 11.8$; $p = 0.003$

Odds ratios

- So there is clear evidence that testosterone therapy increases hair growth
- As for continuous outcomes, we often wish to follow up on a significant finding with pairwise comparisons:
  - High dose vs. placebo: OR=2.1 (1.3, 3.6)
  - Low dose vs. placebo: OR=1.1 (0.6, 2.0)
  - High dose vs. low dose: OR=1.9 (1.1, 3.2)
- So in conclusion, the 300 μg dose of testosterone definitely improves sexual desire in postmenopausal women, but leads to increased hair growth
- There is no evidence that the 150 μg dose leads to any complications, but it seems to have little benefit