Index of 9-L

Practical Information

Today's lecture — another new method/analysis. . .

- main topic: one-way ANOVA (analysis of variance),[1]
  - the step from two to multiple (≥ 2) independent samples for quantitative data,
  - some similarity with 2-sample analysis, but also many new features,
- contrasts are not part of the course curriculum[2],
- additional topic: Kruskal-Wallis test (PSLS Chapter 27) — a non-parametric one-way ANOVA.

Schedule news:

- home assignment 3 has been/will be posted today: worth 15% of course mark, deadline Friday 12/11,
- optional project proposal: also due Friday 12/11,
- now jumping ahead in textbooks:
  - we will come back to regression and correlation,
  - skip over (for now) textbook references to residuals and R^2.

[1] PSLS 4e: Chapters 24 & 26; S: Section 11.3 (too briefly); IPS 7e: Chapter 12.
[2] Contrasts are useful!: Yossa & Verdegem (2015), *Aquaculture* 437, 344−350; see media page.

Data Example(s): Reading Scores

Teaching reading comprehension[3]:

- 66 students, randomly assigned to one of 3 teaching groups, with 22 students in each group,
  - Basal: traditional method,
  - DRTA and Strat: innovative methods,
- 2 pre-tests (before teaching) and 3 post-tests (after teaching),
- questions of interest: compare the 3 groups, in particular the new ones versus Basal, and also DRTA versus Strat,
- in the lecture, we look at `pre1` and `post3`:

Variable	Group	observations (22 in each row)							mean	sd
`pre1`	Basal	4	6	9	12	16	...	9	10.5	2.97
	DRTA	7	7	12	10	16	...	10	9.7	2.69
	Strat	11	7	4	7	7	...	8	9.1	3.34
`post3`	Basal	41	41	43	46	46	...	32	41.0	5.64
	DRTA	31	40	48	30	42	...	49	46.7	7.39
	Strat	53	47	41	49	43	...	42	44.3	5.77

[3] IPS textbook dataset, from a study conducted by Jim Baumann and Leah Jones at the Purdue University School of Education.

Notation for One-Way ANOVA

Data layout and notation:

group		observations (j)					number	mean	std.dev.
	1	X_{11}	X_{12}	$\ldots$	X_{1n_1}		n_1	$\bar{X}_1$	s_1
row	2	X_{21}	X_{22}	$\ldots$	$\ldots$	X_{2n_2}	n_2	$\bar{X}_2$	s_2
(i)	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\ldots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	I	X_{I1}	X_{I2}	$\ldots$	X_{In_I}		n_I	$\bar{X}_I$	s_I

- $X_{ij} = j$th observation in ith group (row), where
  - $i = 1, \ldots, I$, and $I =$ number of groups (rows),
  - $j = 1, \ldots, n_I$, and $n_i =$ number of observations in ith group,
- denote also by $N = n_1 + \ldots + n_I$ the total number of observations, and by $\bar{X} = \sum_{ij} X_{ij}/N$ the overall mean,
- the dataset/design is balanced, if all groups are equally large (i.e., $n_1 = \ldots = n_I$), otherwise unbalanced (some groups of different size) — balanced designs are "nice".[4]

The natural model would seem to be (assuming normals),

$$\begin{aligned} X_{11}, \ldots, X_{1n_1} &\quad \text{i.i.d. } \mathrm{N}(\mu_1, \sigma_1), \\ &\vdots \\ X_{I1}, \ldots, X_{In_I} &\quad \text{i.i.d. } \mathrm{N}(\mu_I, \sigma_I), \end{aligned}$$

but we make the additional assumption: $\sigma_1 = \ldots = \sigma_I$.

[4] Making some calculations and formulas simpler, but when using a computer unbalancedness is not a problem.

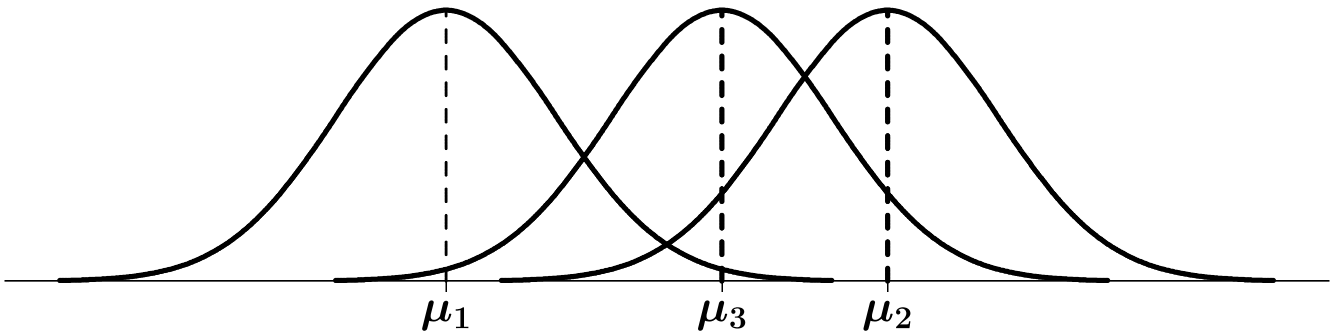
Statistical Model

Model: $$X_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, I; \; j = 1, \ldots, n_i,$$
where ε_{ij}'s are i.i.d. and $\sim \mathrm{N}(0, \sigma)$.

i	observations			
1	$X_{11} = \mu_1 + \varepsilon_{11}$	$X_{12} = \mu_1 + \varepsilon_{12}$	$\ldots$	$X_{1n_1} = \mu_1 + \varepsilon_{1n_1}$
2	$X_{21} = \mu_2 + \varepsilon_{21}$	$X_{22} = \mu_2 + \varepsilon_{22}$	$\ldots$	$X_{2n_2} = \mu_2 + \varepsilon_{2n_2}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
I	$X_{I1} = \mu_I + \varepsilon_{I1}$	$X_{I2} = \mu_I + \varepsilon_{I2}$	$\ldots$	$X_{In_I} = \mu_I + \varepsilon_{In_I}$

- parameters: $\mu_1, \ldots, \mu_I$ (group means) and σ (common standard deviation of all X's and ε's),
- $\varepsilon_{ij} = X_{ij} - \mu_i$, i.e. the deviation of X_{ij} from its mean $\Rightarrow$ ε's are interpreted as random errors / perturbations / noise,
- same model as on previous slide: $X_{ij} \sim \mathrm{N}(\mu_i, \sigma)$.

Normal distributions for 3 groups:



Estimation

Rules and formulas for estimation of model parameters:

- group sample means as estimates for μ's:

$$\begin{aligned} \hat{\mu}_i &= \bar{X}_i \sim \mathrm{N}(\mu_i, \sigma/\sqrt{n_i}), \quad i = 1, \ldots, I, \\ \mathrm{SE}(\hat{\mu}_i) &= s_p/\sqrt{n_i}, \end{aligned}$$

- pooled sample variance estimate for σ^2 (a weighted average of group variances s_i^2):

$$\begin{aligned} \hat{\sigma}^2 &= s_p^2 = \sum_i \frac{n_i - 1}{N - I} s_i^2 = \sum_{ij} \frac{(X_{ij} - \bar{X}_i)^2}{N - I} = \mathrm{SSE}/\mathrm{DFE}, \\ \hat{\sigma} &= s_p = \sqrt{s_p^2}, \end{aligned}$$

  where we have introduced the new notation:

  - $\mathrm{SSE} = \sum_{ij}(X_{ij} - \bar{X}_i)^2$ — the within-group sum of squares,
  - $\mathrm{DFE} = N - I = (n_1 - 1) + \ldots + (n_I - 1)$ — the degrees of freedom for s_p^2.

Confidence intervals for μ_i computed the "usual way", e.g.

$$(1-\alpha) \text{ CI for } \mu_i : \ \hat{\mu}_i \pm t^* \, \mathrm{SE}(\hat{\mu}_i), \ t^* = t_{1-\alpha/2}(\mathrm{DFE}),$$

note: using the pooled standard deviation s_p and its DF for the confidence intervals.

Model Checking I

Summary of model assumptions:

(1) all observations are independent,

(2) all observations are normally distributed,

(3) all observations have the same standard deviation (often called variance homogeneity or homoscedasticity),

(4) all observations within a group have the same mean.

Useful graphical displays:

- boxplots/stemplots/dotplots for all groups in same diagram (overview of data, assumptions (2) and (3)),
- normal probability plots for each of the groups (2).

Useful statistics:

- standard descriptive statistics for each group (3),
- normality tests for each of the groups (2), not overall.

Model Checking II

Extra tool for model checking: residuals — to be introduced in the context of regression (next lecture), but equally applicable to ANOVA.

Practical considerations:

- assumption of equal standard deviations:
  * textbook (PSLS/IPS) guideline based on group standard deviations s_i:
    - okay, if ratio of largest to smallest standard deviation less than 2,
    - may be too prescriptive for small group sizes (n_i) where the s_i are quite noisy,
  * possible to test for equal standard deviations (in software):
    - not (necessarily) a good idea, because the test is more sensitive to unequal standard deviations and non-normality than the ANOVA itself $\Rightarrow$ risk of stating a problem where there really is none. . . ,
    - however fine if such tests (Levene's test is preferable) are non-significant,
- dealing with violations of assumptions:
  * if all groups show the same non-normal pattern (e.g., skewness), transformation may be a solution,
  * in particular, if higher group means are associated with higher standard deviation, transformation with log or $\sqrt{\cdot}$ may work well.

Hypothesis and Test

Overall group equality hypothesis:

$$H_0\colon\ \mu_1 = \ldots = \mu_I \quad \text{(all groups equal, homogeneity between groups)},$$

- alternative hypothesis H_a: some μ's differ ("two-sided"),[5]
- test statistic calculated in several steps,
  - define group sum of squares: $\text{SSG} = \sum_i n_i(\bar{X}_i - \bar{X})^2$ — a weighted sum of squared deviations between group means and the overall mean
  - define group degrees of freedom: $\text{DFG} = I - 1$,
  - introduce group mean square[6]: $\text{MSG} = \text{SSG}/\text{DFG}$,
  - finally, the test statistic is: $F = \text{MSG}/s_p^2$,
- some "motivations":
  - F compares variation between groups with variation within groups,
  - nominator and denominator have similar forms: $F = \text{MSG}/\text{MSE}$,[7]
- under H_0: F-statistic $\sim F(\text{DFG}, \text{DFE})$, and large values are critical for $H_0 \Rightarrow$ P-value calculated as: $P = \text{P}(F(\text{DFG}, \text{DFE}) > F_{\text{obs}})$.

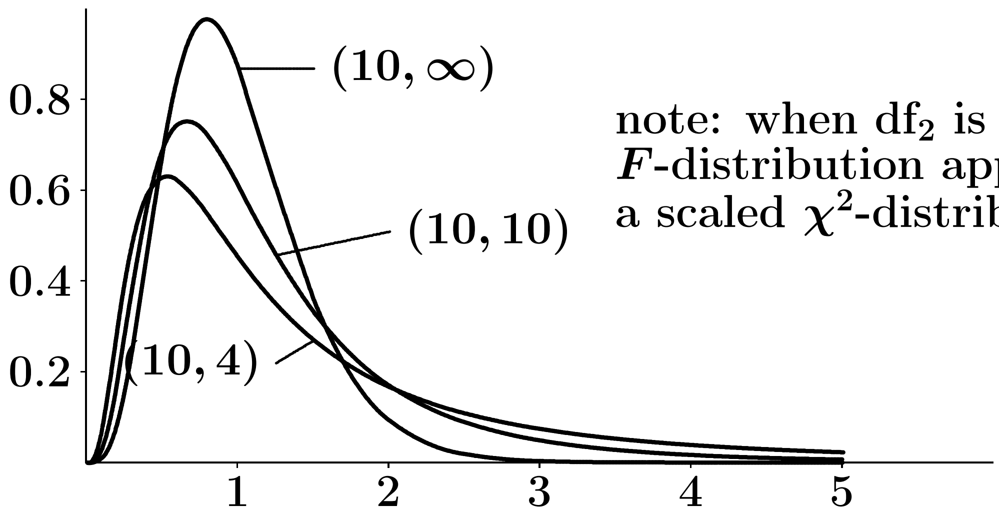
[5] It is a common misunderstanding that H_a implies all the μ's to be different, but the opposite of H_0 is just that the means are not all the same.
[6] Generally, a mean square is a sum of squares divided by the corresponding degrees of freedom.
[7] We previously (slide 9L–5) introduced notation to make $s_p^2 = \text{SSE}/\text{DFE} = \text{MSE}$.

$\boldsymbol{F}$-DISTRIBUTIONS

Another distribution — to be used for tests in normal models (that are more complex than one or two samples):

- $\boldsymbol{F}$-distributions have two parameters, and we write $\boldsymbol{F}(\mathrm{df}_1, \mathrm{df}_2)$ to indicate them:
  - $\mathrm{df}_1, \mathrm{df}_2$ are numbers in $\{1, 2, 3, \ldots\}$
  - called numerator (df_1) and denominator (df_2) degrees of freedom, because $\boldsymbol{F}$ variables are usually ratios,
  - given from the data, and not to be estimated,
  - the order of df_1 and df_2 is important: $\boldsymbol{F}(\mathrm{df}_1, \mathrm{df}_2) \neq \boldsymbol{F}(\mathrm{df}_2, \mathrm{df}_1)$,
- distributions on $(0, \infty)$ — only positive values,
- right skewed; decreasing mean and standard dev. with increasing df_2; Table F/E of of PSLS/IPS has percentiles.

Some $\boldsymbol{F}$ distributions:



note: when df_2 is large (∞), the $\boldsymbol{F}$-distribution approximates a scaled χ^2-distribution

ANOVA Table

Analysis of variance table = convenient layout for summarizing the analysis:

- idea: split variation in the data into parts:

$$\begin{aligned} \text{total variation} &= \text{variation between groups ("Groups")} \\ &\quad + \text{variation within groups ("Error")}, \end{aligned}$$

- collect quantities related to each source of variation on a separate line,
- provides good overview and eases computations (if done by hand),
- generalizes to models with more variables than one.

Source	Degrees of freedom	Sum of squares	Mean square	F	P
Groups	$\text{DFG} = I-1$	$\text{SSG} = \sum_i n_i(\bar{X}_i - \bar{X})^2$	$\text{MSG} = \text{SSG}/\text{DFG}$	MSG/MSE	$\mathrm{P}(F \geq F_{\text{obs}})$
Error	$\text{DFE} = N-I$	$\text{SSE} = \sum_{ij}(X_{ij} - \bar{X}_i)^2$	$\text{MSE} = \text{SSE}/\text{DFE}$	$F \sim F(\text{DFG,DFE})$	
Total	$\text{DFT} = N-1$	$\text{SST} = \sum_{ij}(X_{ij} - \bar{X})^2$	$(\text{MST} = \text{SST}/\text{DFT})$		

Notes:

- MST often omitted from table ($\neq$ MSG + MSE),
- always remember: $\hat{\sigma} = s_p = \sqrt{\text{MSE}}$.

Exercises 12.1, 12.9, 12.25

Exercise 12.1: (response, populations, I, n_i's and N)

(a) response = tomato yield, $I = 4$ varieties, all $n_i = 12$, and $N = 48$,

(b) response = rate of attractiveness, $I = 5$ types of packaging, all $n_i = 40$, and $N = 200$,

(c) response = weight loss, $I = 3$ programs, all $n_i = 20$, and $N = 60$.

Exercise 12.9: (degrees of freedom and hypotheses) For all settings, the hypotheses are:

$$H_0: \ \mu_1 = \mu_2 = \mu_3 = \mu_4, \quad H_a: \ \text{some } \mu\text{'s different},$$

where $I = 4, 5$ and 3, respectively.

(a) varieties (DF = 3), error (DF = 44), total (DF = 47), $F(3, 44)$,

(b) packagings (DF = 4), error (DF = 195), total (DF = 199), $F(4, 195)$,

(c) programs (DF = 2), error (DF = 57), total (DF = 59), $F(2, 57)$.

Exercise 12.25:

(a)

Source	Degrees of freedom	Sum of squares	Mean square	F
Groups	3	104,855.87		
Error	32	70,500.59		
Total				

(b) $H_0: \ \mu_1 = \mu_2 = \mu_3 = \mu_4, \quad H_a: \ \text{some } \mu\text{'s different},$

(c) $F \sim F(3, 32)$ under H_0, $P = P(F(3, 32) > 15.9) \ll 0.001 \Rightarrow$ reject H_0;
conclusion: there are some differences between groups,

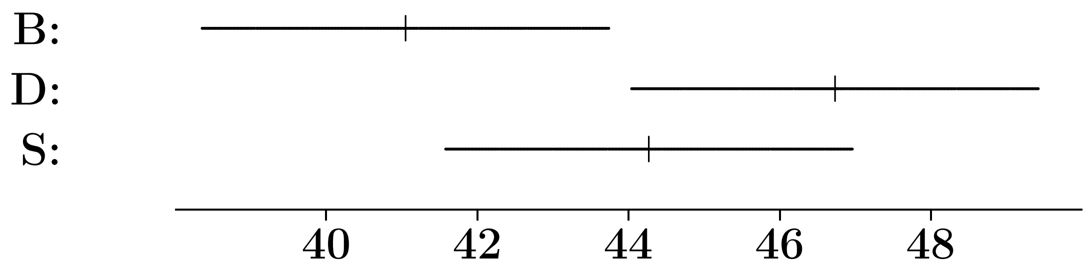
(d) $s_p^2 = \text{MSE} = 2203$, $s_p = \sqrt{2203} = 46.9$.

Comparing and Presenting Groups

Group comparisons based on confidence intervals for group means μ_i:

- principal method uses test/CI for difference between parameters (next slide),
- but conclusions available from group CIs in 2/3 cases (see figure):

  example: reading data — 95% CIs for `post3`:



  - B vs. D: disjoint (non-overlapping) CIs ⇒ significance ($P < 0.05$),
  - D vs. S: estimate inside another CI ⇒ no significance ($P > 0.05$),
  - B vs. S: need CI for difference ($\mu_B - \mu_S$) to assess significance,
- method assumes independent estimates, and is unadjusted for multiple testing (following slides),
- method is applicable to many other settings than one-way ANOVA.

Pairwise Comparisons

Generally, how to proceed after ANOVA? (and any pre-planned contrasts (appendix))

If overall H_0 is non-significant: no further analysis needed (relevant), report $\hat{\mu}$ and P.

If overall H_0 is significant:

- for illustration: plot of $\hat{\mu}_i$'s with error bars (previous page),
- for informal comparisons of group levels — LSD (least significant difference):
  - margin of error of CIs for pairwise difference $\mu_i - \mu_j$ based on $\bar{X}_i - \bar{X}_j$:

$$\text{LSD}_{1-\alpha} = t^* \, s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \quad t^* = t_{1-\alpha/2}(\text{DFE}),$$

  - most useful for *balanced* data (all n_i are equal), because one LSD-value applies to comparisons between any two groups,
  - example (reading data, `post3`):

$$\text{LSD}_{0.95} = t_{0.975}(63) \times 6.314 \times \sqrt{2/22} = 3.80,$$

  - interpretation: smallest distance between $\bar{X}_i$ and $\bar{X}_j$ so that CI for $\mu_i - \mu_j$ does not contain 0 ($\Rightarrow H_0 : \mu_i = \mu_j$ significant at level α, if pre-planned[8]),
  - same as pairwise 2-sample t-tests based on s_p (also called Fisher method),
- for formal comparisons of group levels we must take into account an increase in overall (simultaneous) error level for multiple unplanned[8] comparisons.

[8] The theory behind statistical hypothesis testing is based on pre-determined hypotheses, and does not apply to hypotheses suggested by the data!

Bonferroni Method for Multiple Tests

Basis: If A and B are events, it always holds that: $\mathrm{P}(A \text{ or } B) \leq \mathrm{P}(A) + \mathrm{P}(B)$.
In particular, in the context of performing several tests,

$$\mathrm{P}(\text{error in one or more tests}) \leq \text{ sum of error probabilities}$$

Therefore, if we make k tests/comparisons, we can achieve the simultaneous error probability for all tests to be $\leq \alpha$, by taking the error prob. for each test equal to α/k.

Adjustment of LSD method for k preplanned tests:

$$\mathrm{LSD}_{1-\alpha/k} = t_{1-\alpha/(2k)}(\mathrm{DFE})\, s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

Adjustment of LSD method for unplanned comparisons (suggested by the data):

- take k = total number of comparisons = "I choose 2" $= I(I-1)/2$,
- use above LSD-formula with that value of k.

Notes for Bonferroni method:

- is conservative (wider CIs and higher P-values),
- is flexible and applicable to many situations (not only one-way ANOVA).

Alternative Tukey method for multiple comparisons, described in Chapter 26 of PSLS:

- acceptable method, less conservative than Bonferroni,
- method difficult to explain, and less flexible than Bonferroni.

Summary of One-Way ANOVAs for Reading Scores

Statistical models (for `pre1` and `post3` variables):

$$X_{ij} = \mu_i + \varepsilon_{ij},$$

where $i = 1, 2, 3$ (B, D, S), $j = 1, \ldots, 22$ (students), and ε_{ij}'s i.i.d. and $\sim \mathrm{N}(0, \sigma)$.

ANOVA tables:

		pre1				post3			
Source	DF	SS	MS	F	P	SS	MS	F	P
Groups	2	20.58	10.29	1.13	0.33	357.3	178.7	4.48	0.015
Error	63	572.45	9.09			2511.7	39.9		
Total	65	593.03				2869.0			

Hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ (no differences), H_a: not H_0,
Test of $H_0 \rightarrow$ table F-tests: not significant for `pre1`, but significant for `pre3`.

Estimation/
Presentation:
(using $t^\star = t_{0.975}(63) = 2.00$)

	pre1	post3		
statistic	overall	Basal	DRTA	Strat
mean	9.79	41.05	46.73	44.27
std.dev. s_p	$\sqrt{9.09} = 3.01$	$\sqrt{39.9} = 6.31$		
SE(mean)	$s_p/\sqrt{66} = 0.37$	$s_p/\sqrt{22} = 1.35$		
95% CI	± 0.74	± 2.69		
$\mathrm{LSD}_{0.95}$	—	3.80		

Conclusions: (`pre1`): no differences before teaching,
(`post3`): no differences D/S or B/S; difference D/B.

Method for Presenting Group Comparisons

Issue: pairwise comparison results are not easy to present compactly, because the number of comparisons is larger than number of groups ($\boldsymbol{I}$), and growing fast with $\boldsymbol{I}$ (slide 9L−14).

Idea: indicate significance between groups by codes displayed in list of groups (e.g., table of group estimates or CIs).

Most common system[9] uses letter codes $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \ldots$, so that:

- groups with the same letter are not significantly different,
- for manual construction of codes, proceed as follows:
  * order group means from lowest to highest,
  * designate $\boldsymbol{a}$ to highest group + all groups not significantly different from it,
  * designate $\boldsymbol{b}$ to next group in the same way (but drop if same pattern as for $\boldsymbol{a}$),
  * continue through all groups,
- example: reading data with uncorrected 5% error coding:

$$\mathrm{B}^{b} \quad \mathrm{S}^{ab} \quad \mathrm{D}^{a} \; ,$$

  ⇒ only significant difference is between B and D (as previously shown).

[9] Available in both Minitab and Stata.

Kruskal-Wallis Test

- Model: I independent samples from different distributions:

$$\begin{array}{lll} X_{11},\dots,X_{1n_1} & \text{i.i.d.} & \text{with distribution Dist}_1, \\ \vdots & & \vdots \\ X_{I1},\dots,X_{In_I} & \text{i.i.d.} & \text{with distribution Dist}_I, \end{array}$$

- Hypotheses — two possibilities:
  (1) H_0: $\text{Dist}_1 = \dots = \text{Dist}_I$ (same distribution), H_a: not H_0,[10]
  (2) assuming "$\text{Dist}_i = \text{Dist}_0 + \Delta_i$" (distributions differ only in positions Δ_i)[11], H_0: Δ_i's$=0$ (corresponding to same medians) versus two-sided alternative H_a,

- Test procedure:
  * rank all observations as if a single sample, and compute rank averages $\bar{R}_i$ for each group i,
  * test statistic: sum of squares for ranks,

$$H = constant \times \textstyle\sum_i n_i(\bar{R}_i - \bar{R})^2, \quad \bar{R} = (N+1)/2,$$

    corresponding to SSG in a one-way ANOVA for ranks,
  * under H_0: distribution of Y has no easy form, and software use different approximations for the P-value (based on the $\chi^2(I-1)$-distribution).

[10] More specific wording of H_a: for some (i, i'), P_i is systematically larger than $P_{i'}$; see Chapter 27 of PSLS.
[11] The same assumption as for the Mann-Whitney-Wilcoxon test, implying all distributions to have the same shape.

Notes on Home Assignment 2

Some general comments:[12]

- models/assumptions and hypotheses often not stated (do start including those — it's not for me, it's for yourself. . .),
- for quantitative data, z^* only applies when the standard deviation is known (i.e., not estimated from data), so the general rule is to use t^*,
- t-distribution inference is exact if its assumptions (normally distributed data) are met (exact means that the long-run inclusion rate of the true value is exactly equal to the confidence level),
- with a 99% CI, the corresponding significance test ("check whether target value is inside") is at significance level $\alpha = 1 - 0.99 = 0.01 \Rightarrow$ can give: $P < 0.01$ or $P > 0.01$; with a 95% CI, we get $\alpha = 1 - 0.95 = 0.05 \Rightarrow$ can give: $P < 0.05$ or $P > 0.05$,
- a confidence interval tells us what we know about the population mean (or another parameter), not about the distribution of individual data points;
  an interval for individual points we should not label a confidence interval,
- when doing a statistical test, the target value must come from the context (here: $p = 0.05$, because there should be a 5% chance for values to not be in the interval) and not from the data.

[12] See also the solution at the homepage, slightly expanded with a few extra explanations.

Summary Notes

Key words and concepts for one-way ANOVA:

- comparison of multiple samples (SRS or i.i.d.), assumed normally distributed with separate means but same variance,
- estimation: sample means, pooled variance/standard deviation,
- model checking: normality in each sample, equal standard deviation rule,
- hypothesis and test: F-statistic and F-distribution, calculations organized in ANOVA table with:
  - sum of squares (SS), degrees of freedom (DF), mean square (MS),
- after ANOVA: significant test is followed by pairwise comparisons (or contrasts, not in course syllabus),
  - pre-planned or unplanned comparisons,
  - Bonferroni adjustment for multiple comparisons,
- nonparametric rank-based one-way ANOVA: Kruskal-Wallis test.

APPENDIX: CONTRASTS

Another type of null hypothesis:

- more specific than overall homogeneity of groups (illustrated by the reading data):
  - (1) involves two particular groups, e.g. H_0: $\mu_D = \mu_S$, or
  - (2) involves a linear combination of several group means, e.g.: $\mu_B = (\mu_D + \mu_S)/2$,
- often (not necessarily) considered after test of overall H_0,
- ideally decided prior to data collection/analysis (otherwise different methods needed).

Definition: A contrast is a linear combination of group mean parameters of the form,

$$L = \sum_i c_i \, \mu_i, \quad \text{(PSLS notation)}$$

where the c_i's are known constants with sum 0 ($\sum_i c_i = 0$).

Examples from above: (1): $c_D = 1$, $c_S = -1$, $c_B = 0$, and (2): $c_B = 1$, $c_D = -0.5$, $c_S = -0.5$.

Statistical inference for contrasts:

- estimate: $\hat{L} = \sum_i c_i \bar{X}_i$ (sample contrast),
- standard error: $\mathrm{SE}_{\hat{L}} = s_p \sqrt{\sum_i c_i^2 / n_i}$,
- $(1-\alpha)$ confidence interval for L: $\hat{L} \pm t^* \, \mathrm{SE}_{\hat{L}}$, $t^* = t_{1-\alpha/2}(\mathrm{DFE})$,
- test of H_0: $L = 0$ by: $t = \hat{L}/\mathrm{SE}_{\hat{L}} \sim t(\mathrm{DFE})$ under H_0.