Explanation of the homework and some basic ideas of statitics

- Hypothesis testing
- Comparison of proportions
- Comparison of means
- Standard Deviation and Standard Error

Hypothesis testing

- Under the null hypothesis (the hypothesis should be rejected), calculate probability that the current data or more apart data from expectation are obtained by chance
- If the probability is less than the significance level (which must be determined in advance), we can judge “the null hypothesis is wrong”, otherwise we suspend the judgement about the null hypothesis.

Comparison of proportions (1)

- The proportion of drop-out in each treatment
  A: 0/15, B: 2/13, C: 1/14
  Under the null hypothesis, the possible proportion of drop-out is (0+2+1)/(15+13+14) =3/42. If it is correct, expected cross-table is

	A	B	C
Dropout	15*(3/42)	13*(3/42)	14*(3/42)
Complete	15*(39/42)	13*(39/42)	14*(39/42)

- Calculate X= (0-15*(3/42))^2/(15*(3/42)) + …
  X obeys chi-square distribution with d.f.=2.

Comparison of proportions (2)

- The assumption of common proportion is exactly same as “the event is independent from the treatment”.
- Fisher's exact test can test this.

	A	B	C	Total
Dropout	0	2	1	3
Complete	15	11	13	39
Total	15	13	14	42

- We can calculate the probabilities that the other cross-tables with the same “total” (peripherals), for example,

	A	B	C	Total
Dropout	1	1	1	3
Complete	14	12	13	39
Total	15	13	14	42

- Summing up the probabilities equal to or less than that of actual table → exact probability

Comparison of means

- Anova table is the decomposition of variances into inter-class variance and intra-class variance.
- If the inter-class variance is much greater than intra-class variance (i.e. F ratio is much greater than 1), we can judge “the class significantly affects the values”.

Standard deviation vs standard error

- SD is the variation of actual data
  - SD = $\sqrt{\Sigma (xi - x)^2/(n-1)}$
- SE is the variation of estimated values
  - SEM (standard error of mean) = $SD/\sqrt{n}$
  - If the research is repeated many times, the variation of the estimated value may range within SE at certain probability.
  - eg. In regression analysis, we can calculate the standard error of the slope (regression coefficient)