Hypothesis testing, comparison of 2 independent groups

- All analyses shown today except likelihood test are executable by EZR menu.
- The difference between SD and SE.
  - SD is the variation of actual data
    - SD = √Σ (xi – x)^2/(n-1)
  - SE is the variation of estimated values
    - SEM (standard error of mean) = SD/√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)
- In statistical hypothesis testing, usually the relationships among estimated values are tested, and thus SE is used for calculation.

The logic of 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.
- There are 2 ways of consideration
  - Fisher's: Under the null-hypothesis, how large the probability of current data is obtained by chance?
  - Neyman-Pearson's: Which of null-hypothesis and alternative hypothesis is more plausible when the sampling would be repeated?

Comparison of means

- Between 2 independent groups
  - In general, the analysis of variance (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”.
  - However, t-test is usually applied for this test. Null hypothesis is E(X)=E(Y). $t0=|E(X)-E(Y)|/\sqrt{(Sx/nx+Sy/ny)}$ and $\varphi=(Sx/nx+Sy/ny)^2/\{(Sx/nx)^2/(nx-1)+(Sy/ny)^2/(ny-1)\}$ are applied (φ is degree of freedom of t-distribution). In R, form of script is `t.test(x, y)` or `t.test(y ~ x, data=z)`
- Between paired samples
  - Null hypothesis is E(X-Y)=0. In R, form of script is `t.test(x, y, paired=TRUE)`

Comparison of location of distribution in 2 groups by non-parametric test

- If the data are not normally distributed (especially with outliers), the statistical power of t-test will decrease. In such case, non-parametric tests are applicable, where the test results should be shown with median [Q1, Q3] instead of mean±SD.
- Between 2 independent groups
  - Wilcoxon's rank sum test should be used. At first, assign the rank for whole pooled data. Then sum up ranks for either group (the rank-sum). Compare the rank-sum with expected rank-sum if the locations of 2 distributions are not different. In R, form of script is
    `wilcox.test(x, y)` or **`wilcox.test(y ~ x, data=z)`**
  - Another method is Brunner-Munzel test. It's available by **`brunner.munzel.test()`** of **`lawstat`** package
- Between 2 paired samples
  - Wilcoxon's signed rank test should be used. In R, **`wilcox.test(x, y, paired=TRUE)`**

Comparison of ordered categories between independent 2 groups

- Sometimes less adequate 3 methods are used
  - Assigning arbitrary values to each category and apply t-test
  - Applying Wilcoxon's rank-sum test (but less accurate because of many ties)
  - Ignoring the order information and apply chi-square test
- Applying likelihood ratio test (likelihood of model explaining order by group divided by likelihood of model explaining order by constant) is recommended.
- Example R code

  -
    ```
    set.seed(123)
    y1 <- c(sample(1:3, 20, rep=TRUE), sample(4:6, 30, rep=TRUE))
    y2 <- c(sample(1:3, 30, rep=TRUE), sample(4:6, 20, rep=TRUE))
    y <- as.ordered(c(y1, y2))
    x <- as.factor(c(rep("A", 50), rep("B", 50)))
    table(y, x)
    wilcox.test(y1, y2)
    library(MASS)
    anova(polr(y ~ x), polr(y ~ 1)) # Likelihood ratio test
    chisq.test(table(y, x))
    ```

Comparison of proportions between 2 independent groups

- Among n1 patients, r1 smoker, among n2 controls, r2 smoker was counted.
- Compare smokers' proportion between patients and controls
  - Null hypothesis: p1 = p2 (=p)
  - ^p1 = r1/n1, ^p2 = r2/n2, ^p = (r1+r2)/(n1+n2)
  - E(^p1 – ^p2) = p1 – p2
  - V(^p1 – ^p2) = p1(1 – p1)/n1+p2(1 – p2)/n2
  - V(p1 – p2) = p(1 – p)(1/n1+1/n2)
  - Z = (^p1 – ^p2)/√(^p(1 – ^p)(1/n1+1/n2) ~ N(0, 1)
  - * Usually continuity correction is applied
- Example
  - Among 100 lung cancer patient, 40 smokers and among 100 healthy control, 20 smokers were found. Compare smoking proportion
  - R code

```
smoker <- c(40, 20)
pop <- c(100, 100)
prop.test(smoker, pop)
```