Statistical Methods for Data Science

Lesson 23 - Two sample testing of the mean, and F-test.

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

Tests and confidence intervals for classifier performance

The Caret package

```
Define sets of model parameter values to evaluate
for each parameter set do
    for each resampling iteration do
        Hold–out specific samples
        [Optional] Pre–process the data
        Fit the model on the remainder
        Predict the hold–out samples
    end
    Calculate the average performance across hold–out predictions
end
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set
```

The binomial test

- Dataset $x_1, \ldots, x_n$ realization of $X_1, \ldots, X_n \sim Ber(\theta)$
- $H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0$
- Test statistics: $B = \sum_{i=1}^{n} X_i \sim Bin(n, \theta_0)$ *[Asymmetric distribution]*
- b-value is $\sum_{i=1}^{n} x_i$
- Critical values (exact test):

$$P(B \leq l) = \sum_{i=0}^{l} \binom{n}{i} \theta_0^i (1-\theta_0)^{n-1} = P(B \geq u) = \sum_{i=u}^{n} \binom{n}{i} \theta_0^i (1-\theta_0)^{n-1} = \alpha/2$$

- Normal approximation $Bin(n, \theta_0) \approx N(n\theta_0, n\theta_0(1-\theta_0))$
  - scaled test statistics:

$$B^\star = \frac{B - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \sim N(0,1)$$

  - use z-test with $\sigma^2 = \theta_0(1-\theta_0)$ because $B^\star = \frac{B/n - \theta_0}{\sqrt{\theta_0(1-\theta_0)}/\sqrt{n}} = \frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}}$
  - or even t-test for large samples

See R script

Two sample test of the mean

- Dataset $x_1, \dots, x_n$ realization of $X_1, \dots, X_n \sim F_1$ with $E[X_i] = \mu_1$ and $Var(X_i) = \sigma_X^2$
- Dataset $y_1, \dots, y_m$ realization of $Y_1, \dots, Y_m \sim F_2$ with $E[Y_i] = \mu_2$ and $Var(Y_i) = \sigma_Y^2$
  - measurements for control and (medical) treatment groups of patients
  - performances on benchmark datasets/folds of two different classifiers
- $H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$
- Test statistics: $T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{Var(\bar{X}_n - \bar{Y}_m)}} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$
- We distinguish a few cases:
  - F_1, F_2 are normal distributions
    - σ_X^2 and σ_Y^2 are known *[z-test]*
    - σ_X^2 and σ_Y^2 are unknown and $\sigma_X^2 = \sigma_Y^2$ *[t-test]*
    - σ_X^2 and σ_Y^2 are unknown and $\sigma_X^2 \neq \sigma_Y^2$ *[Welch test]*
  - F_1, F_2 are general distributions
    - Large sample *[t-test]*
    - $F_1(x - \Delta) = F_2(x)$ location-shift *[Wilcoxon test]*
    - Bootstrap two sample test
  - Paired data *[paired t-test]*

Normal data with known σ_X^2 and σ_Y^2: z-test

- $X_1, \ldots, X_n \sim N(\mu_1, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_Y^2)$
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $Z = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$ test statistics when H_0 is true
- z value is $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$ and p-value $p = P(|Z| \geq |z|) = 2(1 - \Phi(|z|))$
- $P(Z \leq -z_{\alpha/2}) = \alpha/2$ and $P(Z \geq z_{\alpha/2}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|z| \geq z_{\alpha/2}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

Unknown $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ and pooled variance

- We need to estimate $Var(\bar{X}_n - \bar{Y}_m) = \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)$
- Recall

$$S_X = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{and} \quad S_Y = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2$$

  are unbiased estimators of σ_X^2 and σ_Y^2
- The *pooled variance*:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right) = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)$$

  is an unbiased estimator of $\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)$

Testing variances for normal data: F-test

- $X_1, \ldots, X_n \sim N(\mu_1, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_Y^2)$
- $H_0 : \sigma_X^2 = \sigma_Y^2$
- $H_1 : \sigma_X^2 \neq \sigma_Y^2$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $F = \frac{S_X^2}{S_Y^2} \sim F(n-1, m-1)$ test statistics when H_0 is true *[Fisher-Snedecor distribution]*
- f value is $\frac{s_X^2}{s_Y^2}$ and p-value is $p = 2\min\{P(F \leq f), 1 - P(F \leq f)\}$ *[Asymmetric]*
- $P(F \leq l) = \alpha/2$ and $P(F \geq u) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $t \leq l$ or $t \geq u$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

Normal data with unknown $\sigma_X^2 = \sigma_Y^2 = \sigma^2$: t-test

- $X_1, \ldots, X_n \sim N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma^2)$
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T_p = \frac{\bar{X}_n - \bar{Y}_m}{S_p} \sim t(n+m-2)$ test statistics when H_0 is true
- t value is $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)}}$ and p-value $p = P(|T_p| \geq |t|)$
- $P(T_p \leq -t_{n+m-2,\alpha/2}) = \alpha/2$ and $P(T_p \geq t_{n+m-2,\alpha/2}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|t| \geq t_{n+m-2,\alpha/2}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

Normal data with unknown $\sigma_X^2 \neq \sigma_Y^2$

- The *nonpooled variance*:
$$S_d^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$$
is an unbiased estimator of $Var(\bar{X}_n - \bar{Y}_m) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$
- The test statistics $T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d}$ is not t-distributed!
- Possible solution: empirical bootstrap (see textbook Sect. 28.3)
- Another solution: Welch t-test

Normal data with unknown $\sigma_X^2 \neq \sigma_Y^2$: Welch t-test

- $X_1, \ldots, X_n \sim N(\mu_1, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_Y^2)$
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d} \approx t(\upsilon)$ test statistics when H_0 is true, with $\upsilon = \frac{\left(\frac{1}{n}+\frac{u}{m}\right)^2}{\frac{1}{n^2(n-1)}+\frac{u^2}{m^2(m-1)}}$ and $u = \frac{s_Y^2}{s_X^2}$
- t value is $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{m}}}$ and p-value $p = P(|T_d| \geq |t|)$
- $P(T_d \leq -t_{\upsilon,\alpha/2}) = \alpha/2$ and $P(T_d \geq t_{\upsilon,\alpha/2}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|t| \geq t_{\upsilon,\alpha/2}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

General data, large sample: t-test

- $X_1, \ldots, X_n \sim F_1$ and $Y_1, \ldots, Y_m \sim F_2$
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d} \approx N(0,1)$
- t value is $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$ and p-value $p = P(|T_d| \geq |t|)$
- $P(T_d \leq -z_{\alpha/2}) = \alpha/2$ and $P(T_d \geq z_{\alpha/2}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|t| \geq z_{\alpha/2}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

General data, location-shift: Wilcoxon rank-sum test

- Also called as: **Mann–Whitney U test**, Mann–Whitney–Wilcoxon (MWW), or Wilcoxon–Mann–Whitney test
- $X_1, \ldots, X_n \sim F_1$ and $Y_1, \ldots, Y_m \sim F_2$
- $H_0 : \mu_1 = \mu_2$
  - actually, $H_0 : F_1(x - \Delta) = F_2(x)$ where $\Delta = \mu_2 - \mu_1$ *[Location-shift model]*
  - we should test that empirical distributions have **the same shape**
- $H_1 : \mu_1 \neq \mu_2$ *[Two-tailed test]*
- $W = \sum_{i=1}^{n} S_i \sim W(n, m)$ when H_0 is true
  - where S_i is the rank of X_i in sorted$(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$
  - `pwilcox` in R, or large sample Normal approx, or $U = W - \frac{n(n+1)}{2}$ statistic
- w value is $\sum_{i=1}^{n} s_i$ and p-value $p = P(|W| \geq |w|)$
- $P(W \leq -w_{\alpha/2}) = \alpha/2$ and $P(T_p \geq w_{\alpha/2}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|w| \geq w_{\alpha/2}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

General data: bootstrap test

- Equal variance ($\sigma_X^2 = \sigma_Y^2$)
  - bootstrap of pooled studentized mean difference

$$t_p^\star = \frac{(\bar{x}_n^\star - \bar{y}_m^\star) - (\bar{x}_n - \bar{y}_m)}{s_p^\star}$$

- Non-equal variance ($\sigma_X^2 \neq \sigma_Y^2$)
  - bootstrap of nonpooled studentized mean difference

$$t_d^\star = \frac{(\bar{x}_n^\star - \bar{y}_m^\star) - (\bar{x}_n - \bar{y}_m)}{s_d^\star}$$

See R script

Paired data

- Datasets $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are measurement for the same experimental unit
  - unit: a person before and after a (medical) treatment
  - unit: a dataset/fold used to train two different classifiers
- The theory is essentially based on taking differences $x_1 - y_1, \ldots, x_n - y_n$ and thus reducing the problem to that of a one-sample test.
- $H_0 : \mu_1 = \mu_2 \Rightarrow H_0 : \mu_1 - \mu_2 = 0$
- Advantage: better power / lower Type II risk of the test w.r.t. unpaired version
  - $P_{paired}(p \leq \alpha | H_1) \geq P_{unpaired}(p \leq \alpha | H_1)$

See R script

Reference handbook (with R code)