

Master Program in *Data Science and Business Informatics*

# Statistics for Data Science

Lesson 32 - Two sample testing of the mean, and  $F$ -test

Salvatore Ruggieri

Department of Computer Science

University of Pisa

[salvatore.ruggieri@unipi.it](mailto:salvatore.ruggieri@unipi.it)

# Tests and confidence intervals for classifier performance

## The Caret package

- 1 Define sets of model parameter values to evaluate
- 2 **for** *each parameter set* **do**
- 3 | **for** *each resampling iteration* **do**
- 4 | | Hold-out specific samples
- 5 | | [Optional] Pre-process the data
- 6 | | Fit the model on the remainder
- 7 | | Predict the hold-out samples
- 8 | **end**
- 9 | Calculate the average performance across hold-out predictions
- 10 **end**
- 11 Determine the optimal parameter set
- 12 Fit the final model to all the training data using the optimal parameter set

**See R script**

# Binary classifier performance metrics

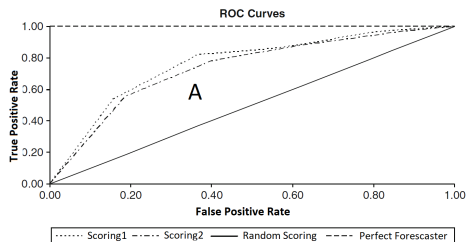
**Confusion matrix** (in R packages, it is transposed)

		Predicted condition			
		Positive (PP)	Negative (PN)		
Actual condition	Total population $= P + N$			Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
	Positive (P)	<b>True positive (TP),</b> hit	<b>False negative (FN),</b> type II error, miss, underestimation	<b>True positive rate (TPR), recall, sensitivity (SEN),</b> probability of detection, hit rate, <i>power</i> $= \frac{TP}{P} = 1 - FNR$	<b>False negative rate (FNR),</b> miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	<b>False positive (FP),</b> type I error, false alarm, overestimation	<b>True negative (TN),</b> correct rejection	<b>False positive rate (FPR),</b> probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	<b>True negative rate (TNR),</b> specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	<b>Positive predictive value (PPV),</b> <i>precision</i> $= \frac{TP}{PP} = 1 - FDR$	<b>False omission rate (FOR)</b> $= \frac{FN}{PN} = 1 - NPV$	<b>Positive likelihood ratio (LR+)</b> $= \frac{TPR}{FPR}$	<b>Negative likelihood ratio (LR-)</b> $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	<b>False discovery rate (FDR)</b> $= \frac{FP}{PP} = 1 - PPV$	<b>Negative predictive value (NPV)</b> $= \frac{TN}{PN} = 1 - FOR$	<b>Markedness (MK), deltaP (<math>\Delta p</math>)</b> $= PPV + NPV - 1$	<b>Diagnostic odds ratio (DOR) = <math>\frac{LR+}{LR-}</math></b>
	Balanced accuracy (BA) $(BA) = \frac{TPR + TNR}{2}$	<b>F<sub>1</sub> score</b> $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	<b>Fowlkes–Mallows index (FM)</b> $= \sqrt{PPV \times TPR}$	<b>Matthews correlation coefficient (MCC)</b> $= \frac{\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}}$	Threat score (TS), critical success index (CSI), <b>Jaccard index</b> = $\frac{TP}{TP + FN + FP}$

Metrics computed on a test set are intended to estimate some parameter over the general distribution.

- $X = (W, C) \sim F$ , i.e.,  $F$  is the (unknown) multivariate distribution of predictive features and class
- Accuracy  $ACC$  of a classifier  $y_{\theta}^+$  is a point estimate of  $E_F[\mathbb{1}_{y_{\theta}^+(W)=C}] = P_F(y_{\theta}^+(W) = C)$

# Binary classifier performance metrics



- Binary classifier score  $s_\theta(w) \in [0, 1]$  where  $s_\theta(w)$  estimates  $\eta(w) = P_{\theta_{TRUE}}(C = 1 | W = w)$
- ROC Curve
  - ▶  $TPR(p) = P(s_\theta(w) \geq p | C = 1)$  and  $FPR(p) = P(s_\theta(w) | C = 0)$
  - ▶ ROC Curve is the scatter plot  $TPR(p)$  over  $FPR(p)$  for  $p$  ranging from 1 down to 0
  - ▶ AUC-ROC is the area below the curve **What does AUC-ROC estimate?**
- Squared error loss or  $L_2$  loss or Brier score:  $\frac{1}{n} \sum_i (s_\theta(w_i) - c_i)^2$
- Classifier is calibrated if  $P(C = 1 | s_\theta(w) = p) = p$  [classifier-calibration.github.io](https://github.com/classifier-calibration)
  - ▶ Binary Expected Calibration Error (binary-ECE):  $\sum_b \frac{|B_b|}{n} |Y_b - S_b|$ 
    - $B_b$  is the set of  $i$ 's in the  $b^{th}$  bin,  $Y_b = |\{i | i \in B_b, c_i = 1\}| / |B_b|$ ,  $S_b = (\sum_{i \in B_b} s_\theta(w_i)) / |B_b|$

# 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  $\text{Var}(X_i) = \sigma_X^2$
- Dataset  $y_1, \dots, y_m$  realization of  $Y_1, \dots, Y_m \sim F_2$  with  $E[Y_i] = \mu_2$  and  $\text{Var}(Y_i) = \sigma_Y^2$ 
  - ▶ measurements for control and (medical) treatment groups of patients
  - ▶ performances on benchmark datasets/folds of two different classifiers
- $H_0 : \mu_1 = \mu_2$      $H_1 : \mu_1 \neq \mu_2$
- Wald test statistics: 
$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\text{Var}(\bar{X}_n - \bar{Y}_m)}} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$
- We distinguish a few cases:
  - ▶  $F_1, F_2$  are normal distributions
    - $\sigma_X^2$  and  $\sigma_Y^2$  are known [z-test]
    - $\sigma_X^2$  and  $\sigma_Y^2$  are unknown and  $\sigma_X^2 = \sigma_Y^2$  [t-test]
    - $\sigma_X^2$  and  $\sigma_Y^2$  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 $\sigma_X^2$ and $\sigma_Y^2$ : z-test

•  $X_1, \dots, X_n \sim N(\mu_1, \sigma_X^2)$  and  $Y_1, \dots, Y_m \sim N(\mu_2, \sigma_Y^2)$

•  $H_0 : \mu_1 = \mu_2$

•  $H_1 : \mu_1 \neq \mu_2$

•  $100(1 - \alpha)\%$ , e.g., 95% or 99% or 99.9%

▶ i.e.,  $\alpha = 0.05$  or  $\alpha = 0.01$  or  $\alpha = 0.001$

•  $Z = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1)$  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  $P(Z \geq z_{\alpha/2}) = \alpha/2$

• Output of the test at confidence level  $100(1 - \alpha)\%$  using critical values

▶  $|z| \geq z_{\alpha/2}$ :  $H_0$  is rejected

▶ otherwise:  $H_0$  cannot be rejected

*[Two-tailed test]*

*[Confidence level]*

*[Significance level]*

*[Critical values]*

*[Critical region]*

**See R script**

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

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

- 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 equal variances for normal data: $F$ -test

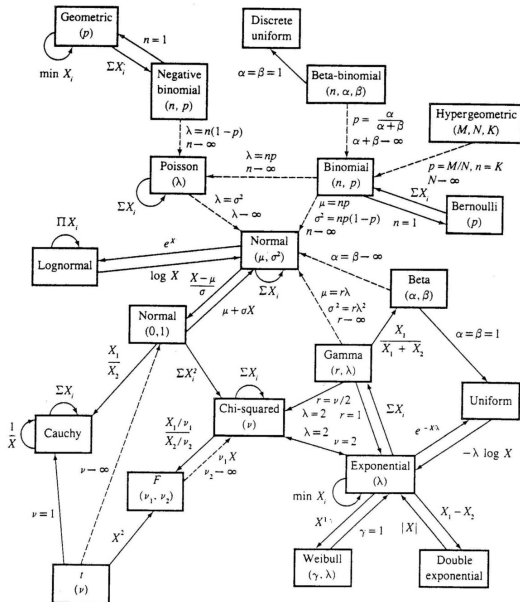
- $X_1, \dots, X_n \sim N(\mu_1, \sigma_X^2)$  and  $Y_1, \dots, Y_m \sim N(\mu_2, \sigma_Y^2)$
- $H_0 : \sigma_X^2 = \sigma_Y^2$
- $H_1 : \sigma_X^2 \neq \sigma_Y^2$  *[Two-tailed test]*
- $100(1 - \alpha)\%$ , e.g., 95% or 99% or 99.9% *[Confidence level]*
  - ▶ i.e.,  $\alpha = 0.05$  or  $\alpha = 0.01$  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 *[Critical region]*
  - ▶  $f \leq l$  or  $f \geq u$  :  $H_0$  is rejected
  - ▶ otherwise:  $H_0$  cannot be rejected

See R script



# Common distributions

- Probability distributions at Wikipedia
- Probability distributions in R
-  C. Forbes, M. Evans, N. Hastings, B. Peacock (2010) Statistical Distributions, 4th Edition Wiley



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

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

- $X_1, \dots, X_n \sim N(\mu_1, \sigma^2)$  and  $Y_1, \dots, Y_m \sim N(\mu_2, \sigma^2)$
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$  *[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 *[Critical region]*
  - ▶  $|t| \geq t_{n+m-2, \alpha/2}$ :  $H_0$  is rejected
  - ▶ 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  $\text{Var}(\bar{X}_n - \bar{Y}_m) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$

- The test statistics  $T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d}$  is not  $t$ -distributed!
- Possible solution: empirical bootstrap (see textbook Section 28.3)
- Another solution: Welch  $t$ -test

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

- $X_1, \dots, X_n \sim N(\mu_1, \sigma_X^2)$  and  $Y_1, \dots, Y_m \sim N(\mu_2, \sigma_Y^2)$

- $H_0 : \mu_1 = \mu_2$

- $H_1 : \mu_1 \neq \mu_2$

- $100(1 - \alpha)\%$ , e.g., 95% or 99% or 99.9%

  - ▶ i.e.,  $\alpha = 0.05$  or  $\alpha = 0.01$  or  $\alpha = 0.001$

*[Two-tailed test]*

*[Confidence level]*

*[Significance level]*

- $T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d} \approx t(v)$  test statistics when  $H_0$  is true, with  $v = \frac{(\frac{1}{n} + \frac{1}{m})^2}{\frac{1}{n^2(n-1)} + \frac{1}{m^2(m-1)}}$  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  $p = P(|T_d| \geq |t|)$

- $P(T_d \leq -t_{v,\alpha/2}) = \alpha/2$  and  $P(T_d \geq t_{v,\alpha/2}) = \alpha/2$

*[Critical values]*

- Output of the test at confidence level  $100(1 - \alpha)\%$  using critical values

  - ▶  $|t| \geq t_{v,\alpha/2}$ :  $H_0$  is rejected

*[Critical region]*

  - ▶ otherwise:  $H_0$  cannot be rejected

**See R script**

# General data, large sample: t-test

- $X_1, \dots, X_n \sim F_1$  and  $Y_1, \dots, Y_m \sim F_2$
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$
- $100(1 - \alpha)\%$ , e.g., 95% or 99% or 99.9%
  - ▶ i.e.,  $\alpha = 0.05$  or  $\alpha = 0.01$  or  $\alpha = 0.001$
- $T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d} \approx N(0, 1)$
- $t$  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  $P(T_d \geq z_{\alpha/2}) = \alpha/2$
- Output of the test at confidence level  $100(1 - \alpha)\%$  using critical values
  - ▶  $|t| \geq z_{\alpha/2}$ :  $H_0$  is rejected
  - ▶ otherwise:  $H_0$  cannot be rejected

*[Two-tailed test]*

*[Confidence level]*

*[Significance level]*

*[Critical values]*

*[Critical region]*

**See R script**

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

- Also called as: **Mann–Whitney  $U$  test** or Mann–Whitney–Wilcoxon (MWW)
- $X_1, \dots, X_n \sim F_1$  and  $Y_1, \dots, Y_m \sim F_2$
- $H_0 : \mu_1 = \mu_2$ 
  - ▶ 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( $X_1, \dots, X_n, Y_1, \dots, Y_m$ )
  - ▶ `pwilcox` in R, or large sample Normal approx
- $w$  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

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

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

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

**See R script**

# Paired data


- Datasets  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$  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  $x_1 - y_1, \dots, x_n - y_n$  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**



# Optional reference

- On confidence intervals and statistical tests (with R code)

 Myles Hollander, Douglas A. Wolfe, and Eric Chicken (2014)  
Nonparametric Statistical Methods.  
3rd edition, *John Wiley & Sons, Inc.*