Chapter IR:VIII

VIII. Evaluation

- ❑ Laboratory Experiments
- ❑ Performance Measures
- ❑ Training and Testing
- ❑ Logging

Training and Testing

Statistical Hypothesis Testing

Claim:

- System 1 is better than System 2 because it achieves 0.66 MAP, 0.16 more than System 2.

What would you reply to this claim?

Training and Testing

Statistical Hypothesis Testing

Claim:

- System 1 is better than System 2 because it achieves 0.66 MAP, 0.16 more than System 2.

Supporting data:

	Average Precision		Mean
	Topic 1	Topic 2	
System 1	0.78	0.44	0.61
System 2	0.52	0.44	0.48
Difference	$+0.26$	± 0.00	$+0.13$

What would you reply to this data?

Training and Testing

Statistical Hypothesis Testing

Claim:

- System 1 is better than System 2 because it achieves 0.66 MAP, 0.16 more than System 2.

Supporting data:

	Average Precision		Mean
	Topic 1	Topic 2	
System 1	0.78	0.44	0.66
System 2	0.52	0.44	0.48
Difference	$+0.26$	± 0.00	$+0.16$

Rebuttal:

- That was just luck.
- With more topics, the gains and losses may even out.

→ Although better on a specific topic, System 1 is not really shown more effective than System 2.

Training and Testing

Statistical Hypothesis Testing

	Average Precision						Mean
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35
Difference	$+0.26$	± 0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$

Given these results, determine whether they have been obtained by chance.

Training and Testing

Statistical Hypothesis Testing

	Average Precision						Mean
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35
Difference	$+0.26$	± 0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$

Given these results, determine whether they have been obtained by chance.

Null hypothesis:

- Average precision values of both systems are drawn from the same underlying probability distribution.
- The differences observed arise from the natural variation of that distribution.

→ The differences are randomly distributed.

Employ a test statistic to compute the probability p of observing the differences if the null hypothesis were true. If the p value is small, the null hypothesis may be false.

Typically, $p < 0.05$ suffices to claim that the differences are statistically significant.

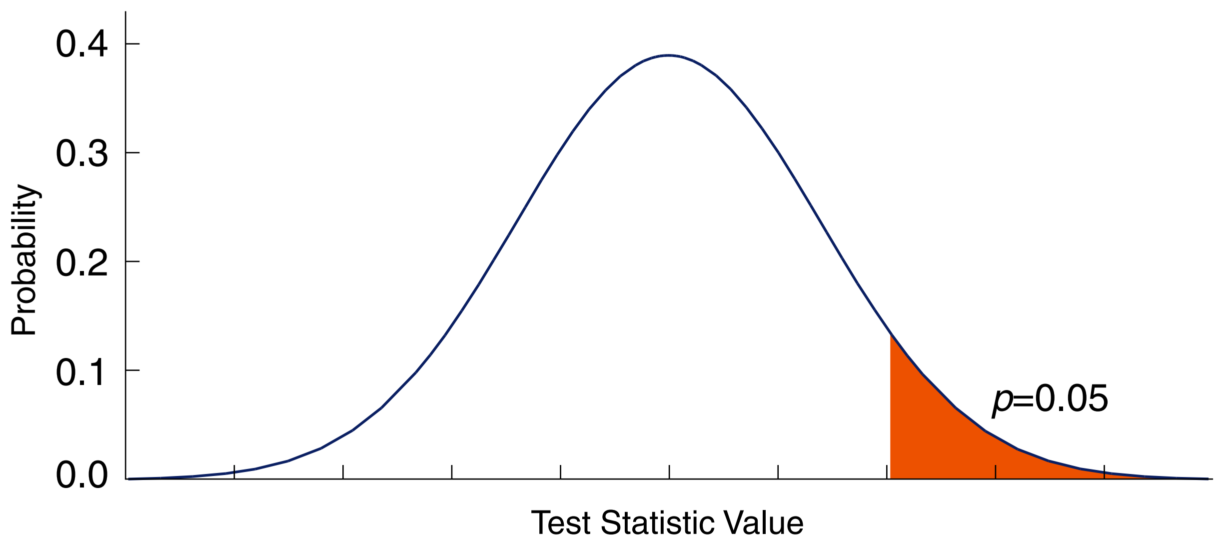
Training and Testing

Statistical Hypothesis Testing

	Average Precision						Mean
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35
Difference	$+0.26$	± 0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$

Given these results, determine whether they have been obtained by chance.

Illustration:



Remarks:

- Rejecting the null hypothesis based on a small p value does not necessarily mean we can accept the opposing hypothesis as true.

Training and Testing

Statistical Hypothesis Testing: Sign Test

	Average Precision						Mean
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35
Difference	$+0.26$	± 0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$
Sign	$+$	$=$	$-$	$+$	$+$	$+$	n/a

Training and Testing

Statistical Hypothesis Testing: Sign Test

	Average Precision						Mean
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35
Difference	$+0.26$	±0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$
Sign	$+$	$=$	$-$	$+$	$+$	$+$	n/a

Procedure:

- Sign $+$ denotes System 1 $>$ System 2, $-$ the opposite, and $=$ a tie.
- Test statistic: number m of $+$ signs.

Null hypothesis:

- Disregarding $=$, the probability of $+$ and $-$ is equal: $P(+) = P(-) = 0.5$.

Assumptions:

- The topics are independent of each other.
- The differences are drawn from the same distribution.
- The individual scores for each topic can be meaningfully compared.

Training and Testing

Statistical Hypothesis Testing: Sign Test

	Average Precision						Mean
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35
Difference	$+0.26$	±0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$
Sign	$+$	$=$	$-$	$+$	$+$	$+$	n/a

If the null hypothesis were true, what is the probability of observing at least $m = 4$ times $+$ out of $n = 5$ experiments?

If $P(+) = P(-) = 0.5$ holds, the test statistic is binomially distributed:

$$p \;=\; P(+ \geq m) \;=\; \sum_{k=m}^{n} \frac{n!}{k!(n-k)!} \cdot P(+)^n \;=\; \frac{5+1}{32} = 0.1875$$

Conclusions:

- The differences of Systems 1 and 2 are not statistically significant as $p > 0.05$.
- We cannot reject the null hypothesis.
- Under the sign test, Systems 1 and 2 must be presumed equally effective.

Training and Testing

Statistical Hypothesis Testing: Student's t-test

	Average Precision						Mean	s
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6		
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51	0.19
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35	0.19
Difference	$+0.26$	± 0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$	0.15

Procedure:

- Compute the score differences of the scores of Systems 1 and 2.
- Test statistic: $t = (\bar{d} - \mu_0)/(s_d/\sqrt{n})$ for n topics, the average $\bar{d}$ of the differences between Systems 1 and 2, and their observed standard deviation s_d.

Null hypothesis:

- The average difference $\bar{d}$ is at most μ_0.

Assumptions:

- The topics are independent of each other.
- The differences are approximately normally distributed.

Training and Testing

Statistical Hypothesis Testing: Student's t-test

	Average Precision						Mean	s
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6		
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51	0.19
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35	0.19
Difference	$+0.26$	±0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$	0.15

If the null hypothesis were true, what is the probability of observing $\bar{d} = 0.16$ and $s_d = 0.15$ for $n = 6$ at an expected $\mu_0 = 0$?

The test statistic is t-distributed with $n - 1$ degrees of freedom:

$$t = \frac{0.16 - 0}{0.15/\sqrt{6}} = 2.613 \quad \leadsto \quad t(0.975; n-1) \; < \; 1 - p \; < \; t(0.99; n-1)$$

t-distribution table [Wikipedia]:

n	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
⋮											
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
⋮											

Training and Testing

Statistical Hypothesis Testing: Student's t-test

	Average Precision						Mean	s
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6		
System 1	0.78	0.44	0.54	0.62	0.45	0.22	0.51	0.19
System 2	0.52	0.44	0.55	0.32	0.12	0.13	0.35	0.19
Difference	$+0.26$	± 0.00	-0.01	$+0.30$	$+0.33$	$+0.09$	$+0.16$	0.15

If the null hypothesis were true, what is the probability of observing $\bar{d} = 0.16$ and $s_d = 0.15$ for $n = 6$ at an expected $\mu_0 = 0$?

The test statistic is t-distributed:

$$t = \frac{0.16 - 0}{0.15/\sqrt{6}} = 2.613 \quad \rightsquigarrow \quad p = 0.025$$

where p has been computed precisely using an implementation of the t-distribution.

Conclusions:

- The differences of Systems 1 and 2 are statistically significant as $p < 0.05$.
- We can reject the null hypothesis.
- Under the Student's t-test, System 1 may be better than System 2.

Training and Testing

Statistical Hypothesis Testing: Power and Effect Size

Two statistical analysis tools complement statistical hypothesis testing:

- Power Analysis [Wikipedia] [G*Power]

  The power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis when a specific alternative hypothesis is true. Power analysis is applied before conducting an experiment to determine the sample size (=number of topics) required to obtain a certain statistical power.

  Furthermore, the aforementioned statistical tests are said to have more or less power, dependent on their ability to correctly reject the null hypothesis, given the alternative hypothesis is true. For instance, a low-power test (e.g., sign test) may lead to falsely accepting the null hypothesis.

- Effect Size Estimation [Wikipedia]

  An effect size is a quantitative measure of the magnitude of a phenomenon. The effect size does not directly determine the significance level, or vice versa. Given a sufficiently large sample size, a non-null statistical comparison will always show a statistically significant result unless the population effect size is exactly zero. The reporting of effect sizes facilitates the interpretation of the substantive, as opposed to the statistical, significance of a research result. About 50 to 100 different measures of effect size are known.

  For the Student's t-test, Cohen's d is a well-known effect size estimator.

Remarks:

- For the example above, Cohen's $d = 0.84$.

  Common interpretation:

Effect size	d
Very small	0.01
Small	0.20
Medium	0.50
Large	0.80
Very large	1.20
Huge	2.00

Training and Testing

Hyperparameter Optimization

Search engines possess many parameters, many of which affect retrieval effectivenss. Examples: Algorithm parameters, alternative algorithms for a subtask, weights of document fields.

In IR, hyperparameter optimization often boils down to trial and error:

- Grid Search.

  Systematic trials of all parameter combinations from pre-specified value ranges and steps for each parameter.

- Random Search.

  Selection of a random subset of all parameter combinations of pre-specified value ranges and steps for each parameter.

Ideally, parameters are optimized based on a 3-way split of the available data into subsets used for training, validation, and test.

Training data are used to fine-tune learning algorithms. Validation data are used to repeatedly check a search engine's performance trajectory during optimization. Test data are used once at the end as a final check.