Machine Translation Marathon 2016: MT Significance Testing

Yvette Graham

Talk Outline

Sampling Distribution & Confidence Intervals

MT System Significance Tests

MT Metric Significance Tests

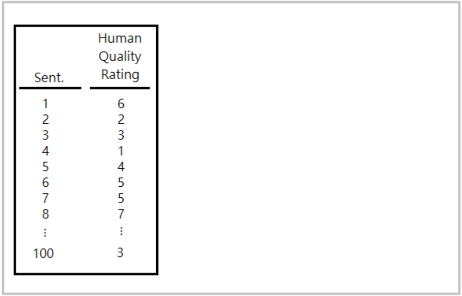
Sampling Distributions

Commonly need to measure the performance of an MT system or metric using a statistic, such as the average score attributed to sentences in a test set (by a human assessor) or the correlation between metric scores and human assessment scores for a a number of MT systems.

In the case of evaluating an MT system:

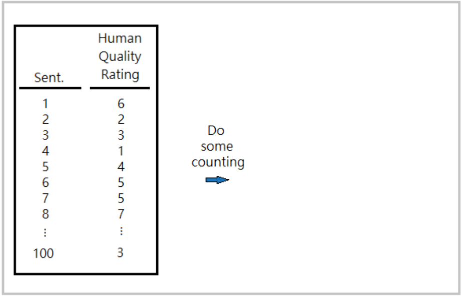
- We can't evaluate a translation of every possible input sentence;
- Instead, we rely on an evaluation based on a sample of translations;
- To understand the uncertainty in the statistic we use to evaluate systems or metrics, we need to think about the sampling distribution of the statistic;
- Confidence intervals are estimated from the sampling distribution of a specific statistic.

Sample of Translations



Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

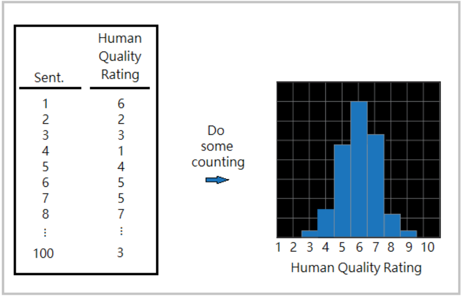
Sample of Translations



Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

Do some counting

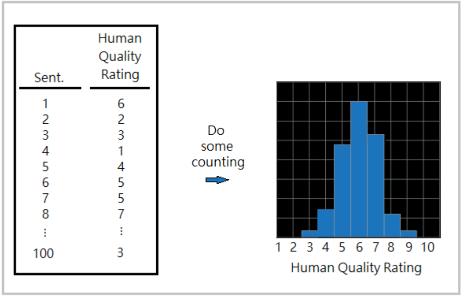
Sample of Translations



Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

Do some counting →

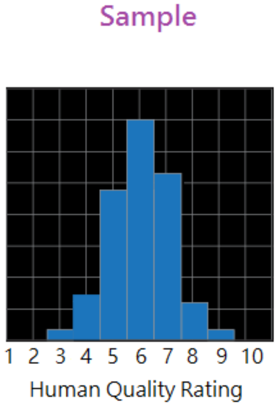
Sample of Translations



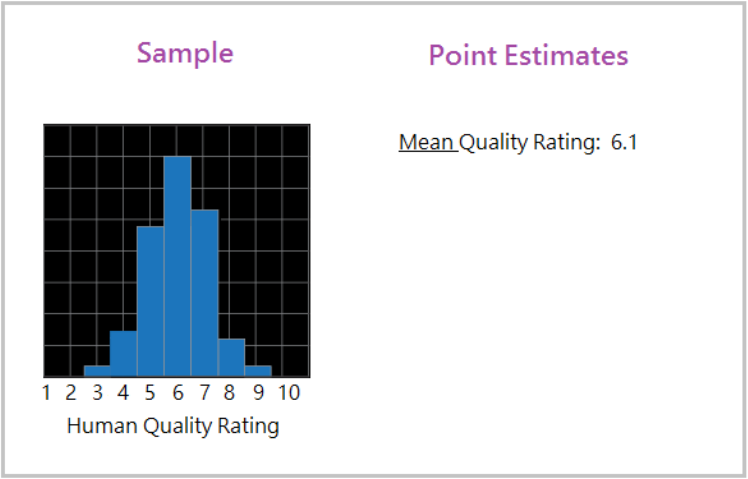
Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

- We have run a human evaluation on a sample 100 translations;
- We don't want to report the entire distribution of human scores;
- Instead report a statistic, like average human score, median, variance, correlation, ...

Point Estimates



Point Estimates

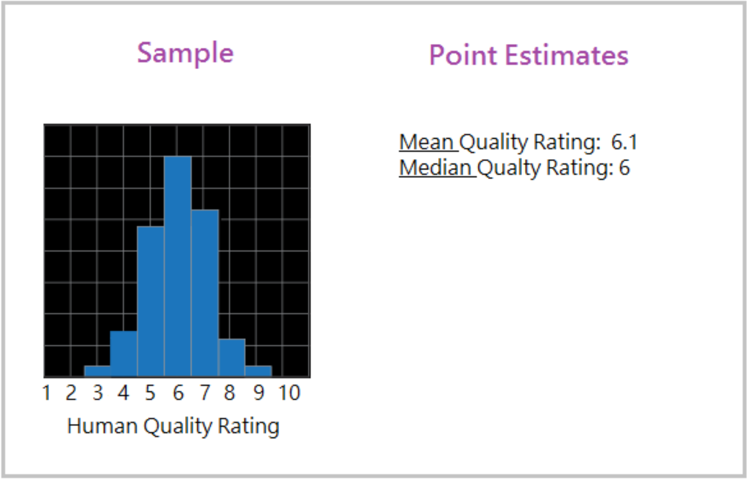


Sample

Point Estimates

Mean Quality Rating: 6.1

Point Estimates

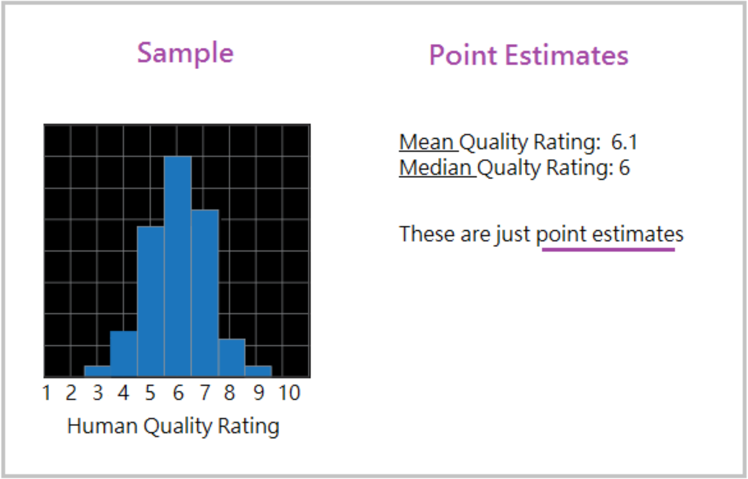


Sample

Point Estimates

Mean Quality Rating: 6.1
Median Qualty Rating: 6

Point Estimates



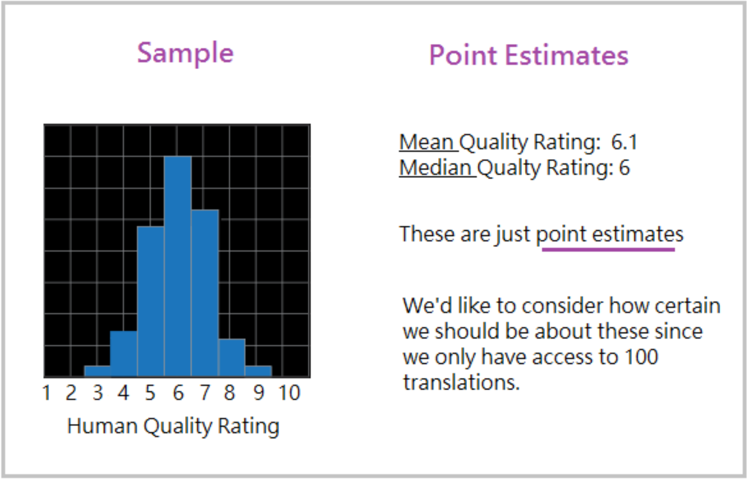
Sample

Point Estimates

Mean Quality Rating: 6.1
Median Qualty Rating: 6

These are just point estimates

Point Estimates



Sample

Point Estimates

Mean Quality Rating: 6.1
Median Qualty Rating: 6

These are just point estimates

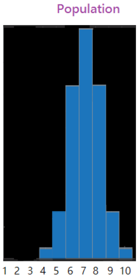
We'd like to consider how certain we should be about these since we only have access to 100 translations.

Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?

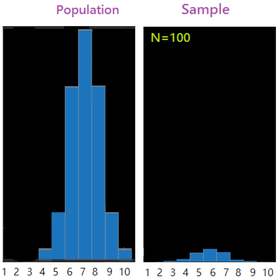
Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?



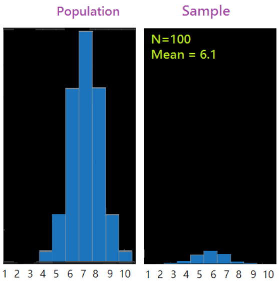
Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?



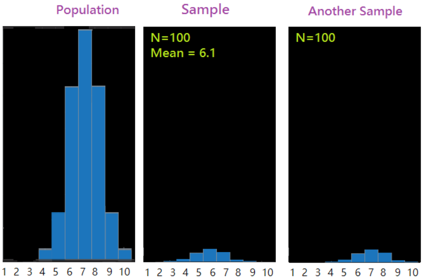
Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?



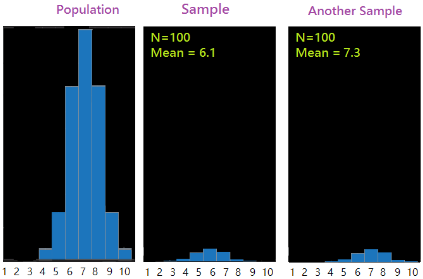
Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?



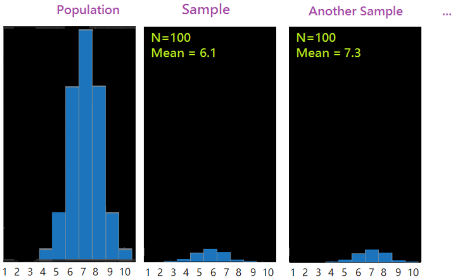
Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?



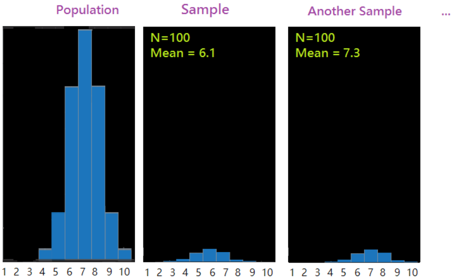
Sampling Distribution

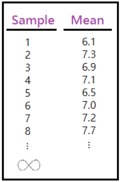
Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?



Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?

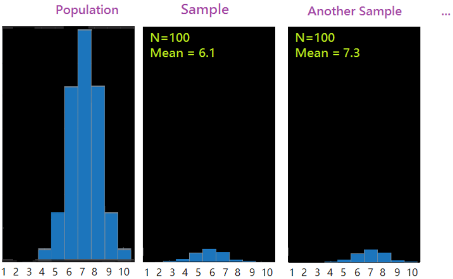


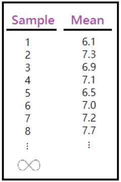


Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

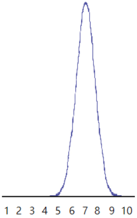
Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?



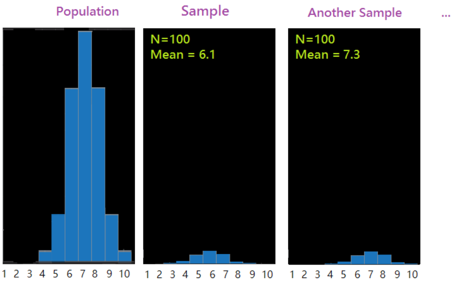


Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

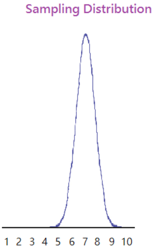


Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?

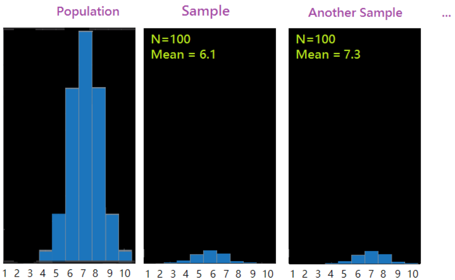


Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

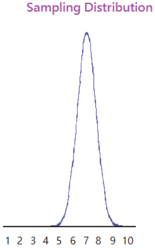


Sampling Distribution

Population: all posible input sentences translated by our MT system, what would the human assessment distribution for the population look like?

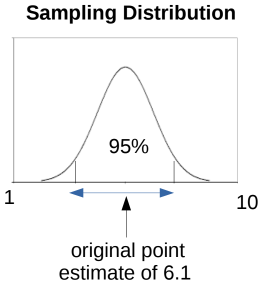


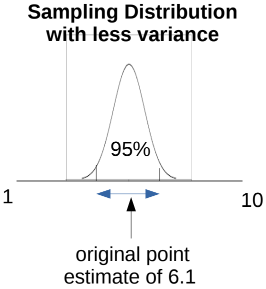
Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	



The spread of this is what we try to estimate with confidence intervals: within what interval would 95% of our means be if we were to sample with size=N from the population an infinite no. of times?

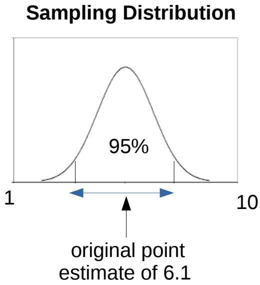
Reliability of Point Estimates

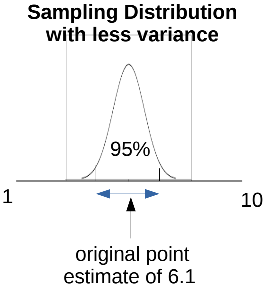

Sampling Distribution


Sampling Distribution
with less variance

- By estimating the spread of the sampling distribution, this allows us to judge how reliable a point estimate is.
- We can estimate the sampling distribution / confidence interval for a <u>specific statistic</u> (eg mean) at a <u>specific sample size</u> (eg N=100)

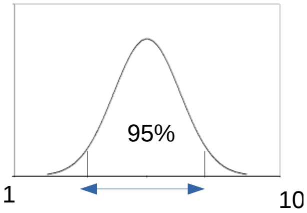
Confidence Intervals for Point Estimates





- Confidence intervals estimate the range of values within which eg. 95% of point estimates would lie if that statistic were computed from an infinite number of random samples of size N drawn from the population.

Confidence Intervals:

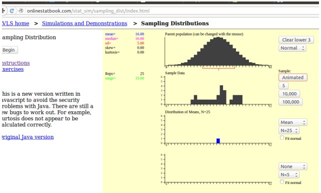


Wide confidence intervals are generally not what we want.

Two things that affect confidence interval width:

- Variance: all else being equal, more variance in the population should result in a wider confidence interval (we can't know this variance, we estimate it);
- Sample size: All else being equal, a smaller sample size should result in wider confidence intervals.

Sampling Distribution Demo



onlinestatbook.com/stat_sim/sampling_dist/index.html

VLS home > Simulations and Demonstrations > **Sampling Distributions**

ampling Distribution

istructions
xercises

his is a new version written in avascript to avoid the security roblems with Java. There are still a ew bugs to work out. For example, urtosis does not appear to be alculated correctly.

riginal Java version

mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00

Reps= 25
range= 15.00

Try this fun demo:
`http://onlinestatbook.com/stat_sim/sampling_dist/index.html`
(David Lane, Rice University , University of Houston Clear Lake, and Tufts University)

Look at the samlping distribution for the mean

- At different N's
- With different population variance

Bootstrap Resampling

There are pretty straight-forward formulas that can be used to estimate the spread of the sampling distribution of the mean.

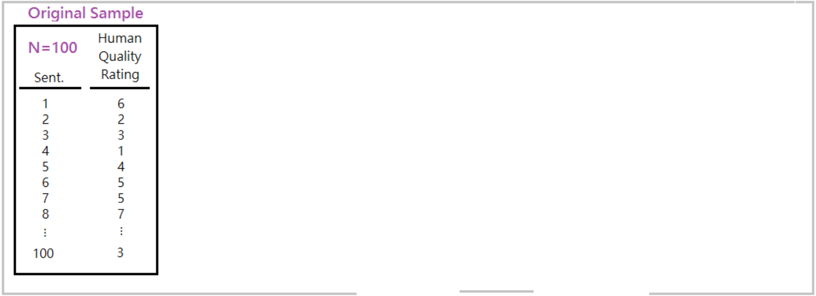
Equivalent formulas don't exist for other estimators.

Bootstrap resampling provides an alternative.

Bootstrap (Efron, 1979) is based on the following:

- Substituting the population with the empirical distribution (sample) to compute the spread of the sampling distribution results in a reasonable estimate of this spread for any estimator.

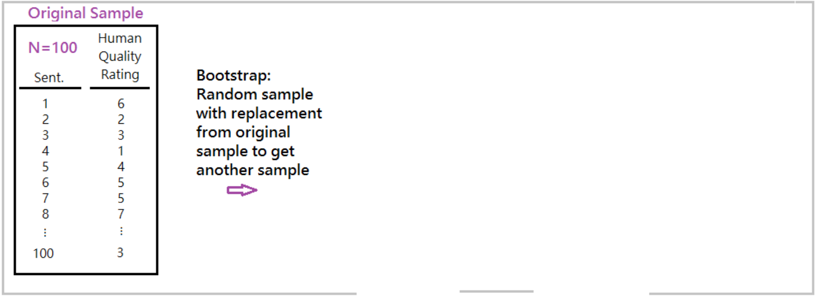
Bootstrap Example



Original Sample

N=100 Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

Bootstrap Example



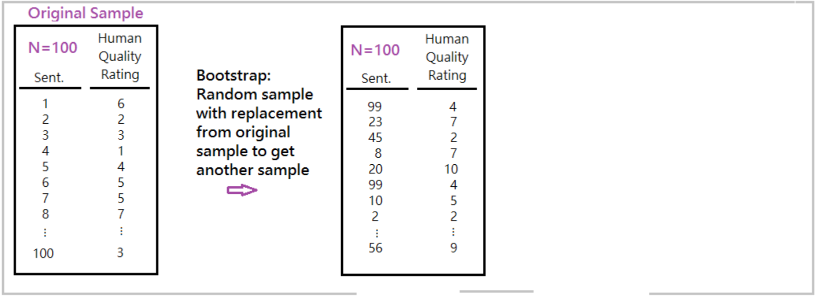
Original Sample

N=100

Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

Bootstrap:
Random sample with replacement from original sample to get another sample ⇨

Bootstrap Example



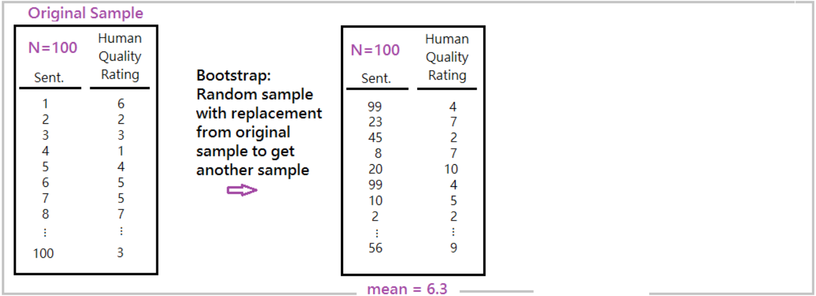
Original Sample

N=100 Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

Bootstrap: Random sample with replacement from original sample to get another sample ⇨

N=100 Sent.	Human Quality Rating
99	4
23	7
45	2
8	7
20	10
99	4
10	5
2	2
⋮	⋮
56	9

Bootstrap Example



Original Sample

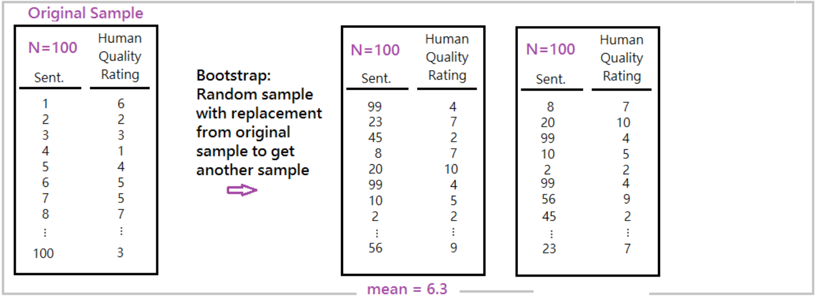
N=100 Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

Bootstrap: Random sample with replacement from original sample to get another sample ⇨

N=100 Sent.	Human Quality Rating
99	4
23	7
45	2
8	7
20	10
99	4
10	5
2	2
⋮	⋮
56	9

mean = 6.3

Bootstrap Example



Original Sample

N=100 Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

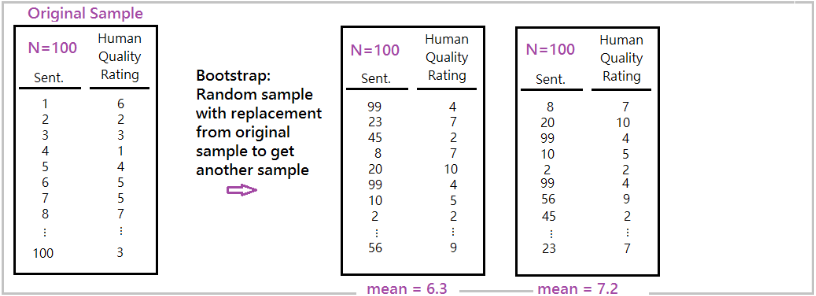
Bootstrap: Random sample with replacement from original sample to get another sample ⇨

N=100 Sent.	Human Quality Rating
99	4
23	7
45	2
8	7
20	10
99	4
10	5
2	2
⋮	⋮
56	9

mean = 6.3

N=100 Sent.	Human Quality Rating
8	7
20	10
99	4
10	5
2	2
99	4
56	9
45	2
⋮	⋮
23	7

Bootstrap Example



Original Sample

N=100 Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

Bootstrap: Random sample with replacement from original sample to get another sample ⇨

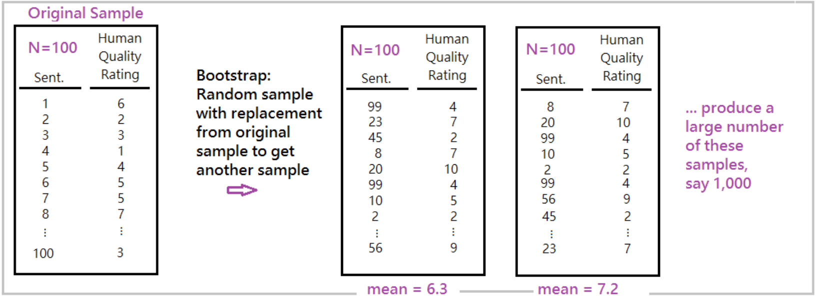
N=100 Sent.	Human Quality Rating
99	4
23	7
45	2
8	7
20	10
99	4
10	5
2	2
⋮	⋮
56	9

mean = 6.3

N=100 Sent.	Human Quality Rating
8	7
20	10
99	4
10	5
2	2
99	4
56	9
45	2
⋮	⋮
23	7

mean = 7.2

Bootstrap Example



Original Sample

N=100 Sent.	Human Quality Rating
1	6
2	2
3	3
4	1
5	4
6	5
7	5
8	7
⋮	⋮
100	3

Bootstrap: Random sample with replacement from original sample to get another sample ⇨

N=100 Sent.	Human Quality Rating
99	4
23	7
45	2
8	7
20	10
99	4
10	5
2	2
⋮	⋮
56	9

mean = 6.3

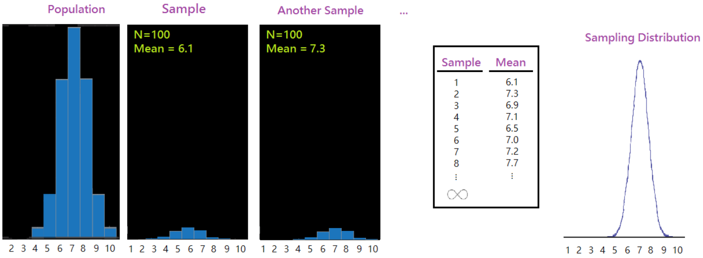
N=100 Sent.	Human Quality Rating
8	7
20	10
99	4
10	5
2	2
99	4
56	9
45	2
⋮	⋮
23	7

mean = 7.2

... produce a large number of these samples, say 1,000

Bootstrap Example

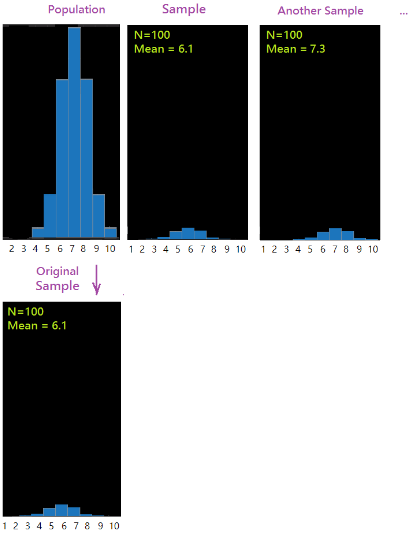
From earlier:



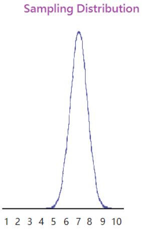
Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

Bootstrap Example

From earlier:

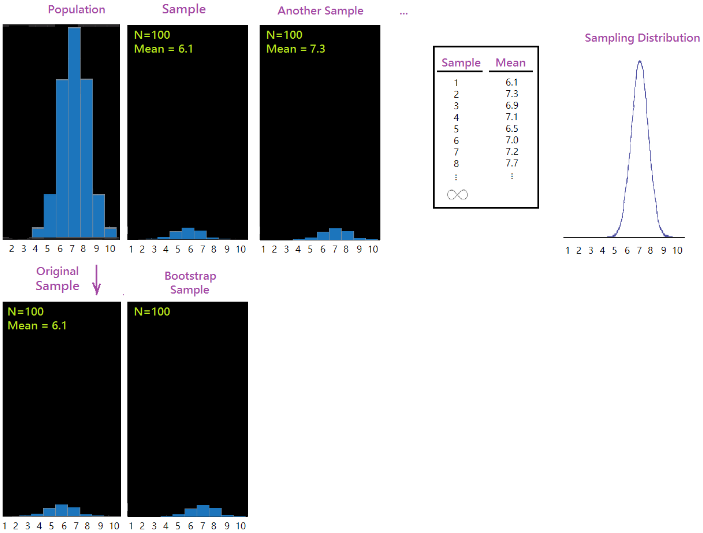


Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	



Bootstrap Example

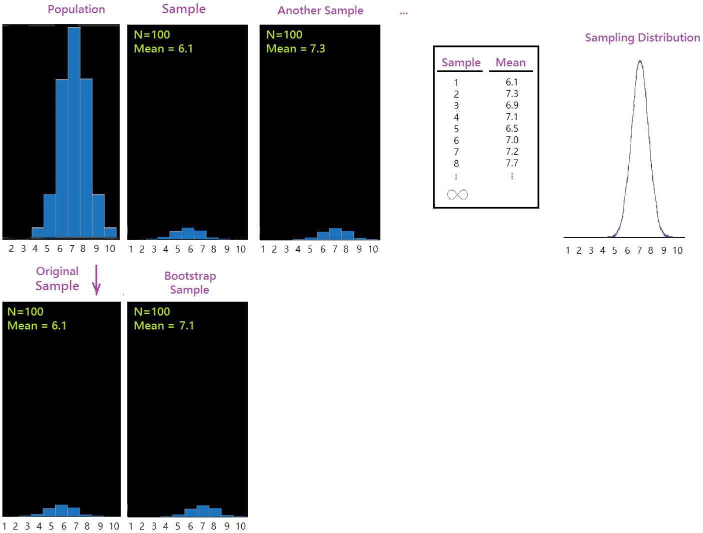
From earlier:



Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

Bootstrap Example

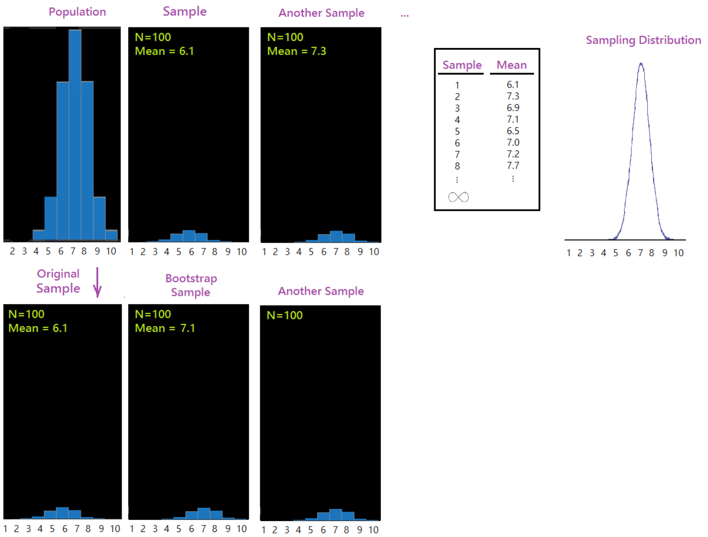
From earlier:



Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

Bootstrap Example

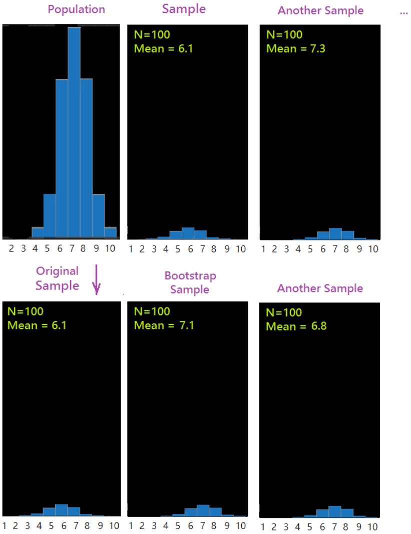
From earlier:



Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

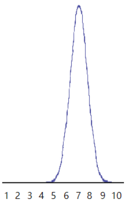
Bootstrap Example

From earlier:



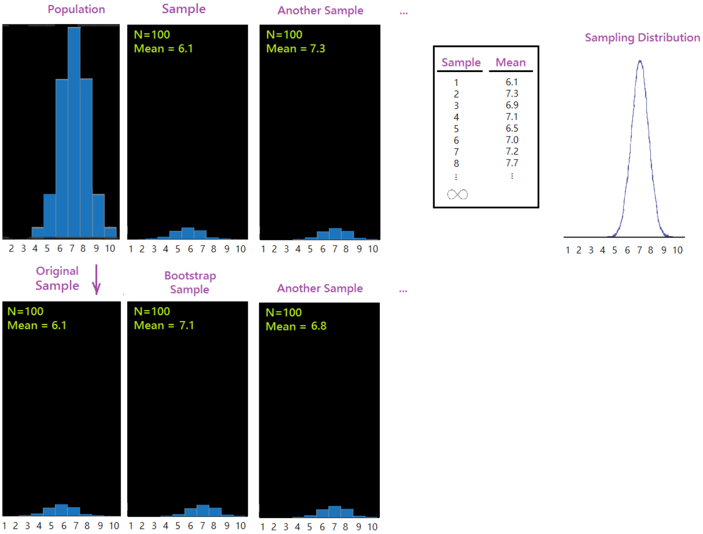
Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

Sampling Distribution



Bootstrap Example

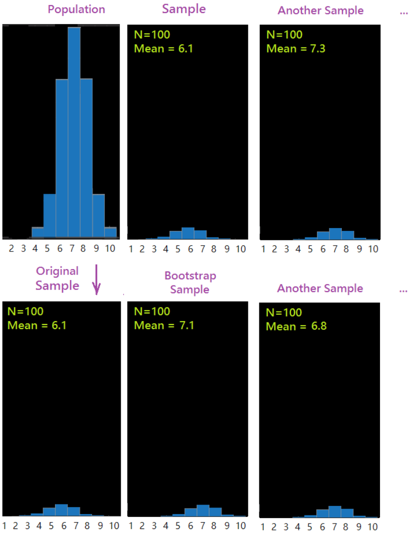
From earlier:



Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

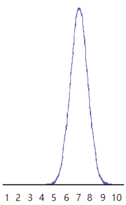
Bootstrap Example

From earlier:



Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

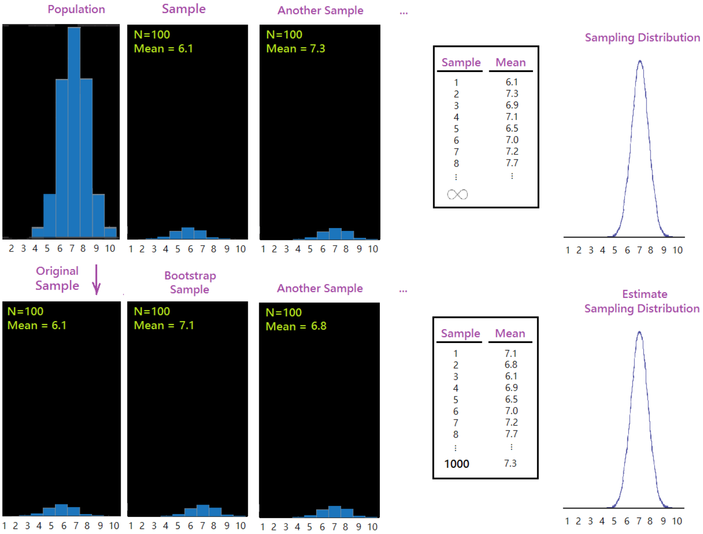
Sampling Distribution



Sample	Mean
1	7.1
2	6.8
3	6.1
4	6.9
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
1000	7.3

Bootstrap Example

From earlier:



Sample	Mean
1	6.1
2	7.3
3	6.9
4	7.1
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
∞	

Sample	Mean
1	7.1
2	6.8
3	6.1
4	6.9
5	6.5
6	7.0
7	7.2
8	7.7
⋮	⋮
1000	7.3

Bootstrap: Important Points

- Random sample with replacement;
- Use at least 1,000 bootstrap samples;
- Make sure the sampling distribution simulates the sampling distribution for the correct sample size, data and statistic.

Example: Measure the performance of an MT system with the mean of 200 human judgments

- Size of each of the 1K bootstrap samples: 200;
- Data to resample: human judgments;
- Statistic to compute 1K times: mean.

Example: Measure the performance of a MT metric with the correlation of 300 pairs of human judgment and metric scores.

- Size of each of the 1K bootstrap samples: 300;
- Data to resample: human judgment and metric score pairs;
- Statistic to compute 1K times: correlation.

Confidence Intervals and Significance

When comparing two systems, if data (test sentences) are not paired and the confidence intervals of the means for two systems do not overlap, we can conclude from that a significant difference in performance;

But data in MT evaluation is mostly paired;

For paired data, instead of estimating the sampling ditribution for the mean of two separate systems and seeing if they overlap, instead we should compute a single sampling distribution for differences in mean scores;

Signficance can be concluded if the 95% (eg) confidence interval of the difference in means does not include zero;

The point estimate for the mean difference in performance of the two systems is significantly different from zero.

In Summary – Sampling Distribution

- The sampling distribution of a statistic is what try to estimate when computing confidence intervals for point estimates;
- Sampling distribution is estimated for a specific statistic, and a specific sample size;
- Smaller sample size generally means wider confidence intervals;
- Larger sample size generally means less spread in sampling distribution and narrower confidence intervals.

- The sampling distribution of the mean is normal, making estimation of confidence intervals for a difference in means pretty straight-forward.
- The sampling distribution of the Pearson correlation is skewed, making significance of differences in correlations less straight-forward.

Talk Outline

Sampling Distribution & Confidence Intervals

MT System Significance Tests

MT Metric Significance Tests

Significance Tests with Bleu

Problem: automatic MT metrics such as Bleu are calculated at the *document-level*, over the totality of translations, and return a single aggregated score, not segment-level scores – we can't do e.g. difference of means significance test for distributions of Bleu scores, since these would be required for individual sentences.

Solution: *randomised significance tests* for Bleu where we apply bootstrap resampling to significance of differences in Bleu scores for a pair of MT systems.

MT System Significance Testing

Three potential randomized tests for significance testing differences in MT metric scores:

1. Paired bootstrap resampling [Koehn, 2004]
2. Approximate randomization [Riezler and Maxwell, 2005]
3. Bootstrap Resampling [Graham et al., 2014]

Criticisms:

- Criticism of (3) bootstrap resampling: S_{H_0} has the same shape but a different mean than S_{boot} (does not happen with (1) or (2));
- Other problems can arise for (2).

Paired Bootstrap Resampling

```
Set $c = 0$
For bootstrap samples $b = 1, ..., B$
    If $S_{X_b} < S_{Y_b}$
        $c = c + 1$
If $c/B \leq \alpha$
    Reject the null hypothesis
```

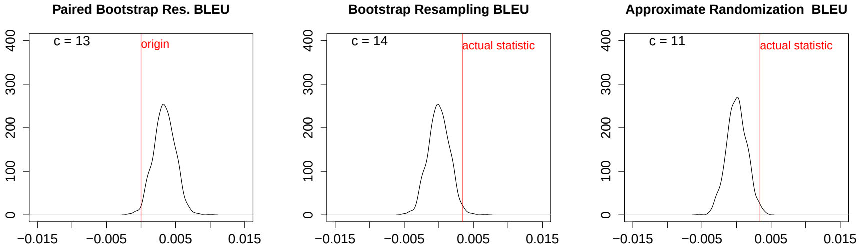
Bootstrap Resampling

```
Set $c = 0$
Compute actual statistic of score differences $S_X - S_Y$
on test data
Calculate sample mean $\tau_B = \frac{1}{B} \sum_{b=1}^{B} S_{X_b} - S_{Y_b}$ over
bootstrap samples $b = 1, \ldots, B$
For bootstrap samples $b = 1, \ldots, B$
    Sample with replacement from variable tuples test
    sentences for systems $X$ and $Y$
    Compute pseudo-statistic $S_{X_b} - S_{Y_b}$ on bootstrap data
    If $S_{X_b} - S_{Y_b} - \tau_B \geq S_X - S_Y$
        $c = c + 1$
If $c/B \leq \alpha$
    Reject the null hypothesis
```

Approximate Randomization

```
Set $c = 0$
Compute actual statistic of score differences $S_X - S_Y$
on test data
For random shuffles $r = 1, ..., R$
    For sentences in test set
        Shuffle variable tuples between systems $X$ and $Y$
        with probability 0.5
    Compute pseudo-statistic $S_{X_r} - S_{Y_r}$ on shuffled data
    If $S_{X_r} - S_{Y_r} \geq S_X - S_Y$
        $c = c + 1$
If $c/R \leq \alpha$
    Reject the null hypothesis
```

Example Pseudo-statistic Distributions



MT Significance Test Comparison

- Use translations from all participating WMT12 ES–EN and EN–ES systems (12 and 11 systems, resp.)
- Use AMT to manually annotate each translation for fluency and adequacy based on a continuous (Likert) scale, with strict annotator-level quality controls [Graham et al., 2013]
- Standardize the scores from a given annotator according to mean and standard deviation
- Final dataset: average of 1,483 (1,280) adequacy and 1,534 (1,013) fluency assessments per ES–EN (EN–ES) system

Evaluation Methodology

- Evaluate each pair of systems separately for:

  1. adequacy
  2. fluency
  3. combined adequacy–fluency (if no significant difference in adequacy, use fluency as fallback)

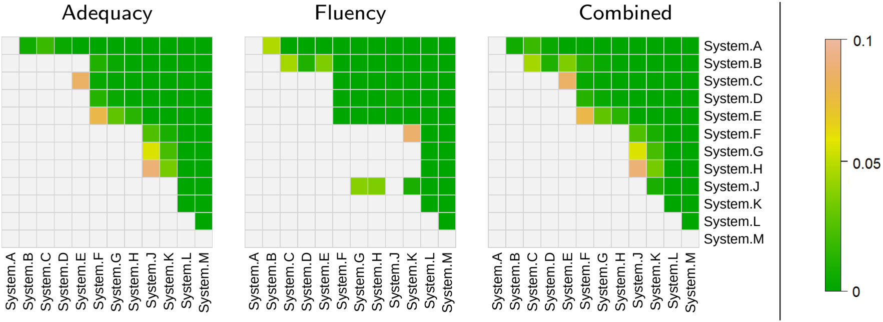
  based on the Wilcoxon rank-sum test

- Score each translation sample based on:

  1. BLEU [Papineni et al., 2002]
  2. NIST [NIST, 2002]
  3. TER [Snover et al., 2005]
  4. METEOR [Banerjee and Lavie, 2005]

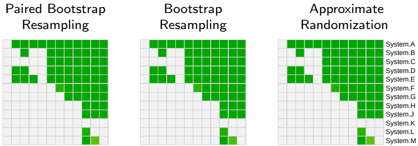
Reference Results (ES–EN)

- System comparison based on the segment-level human assessments (ES–EN):



Pairwise Significance Tests (ES–EN)

Bleu

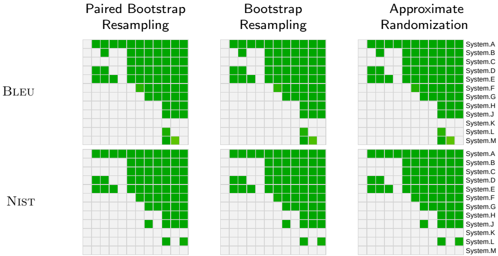


Paired Bootstrap Resampling

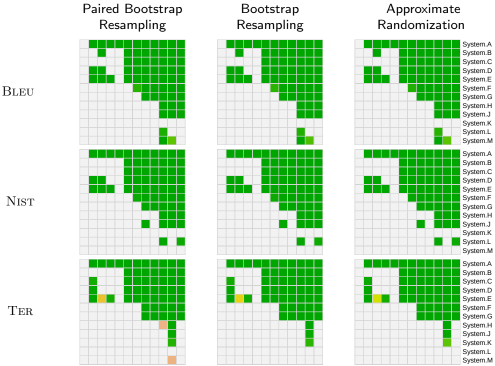
Bootstrap Resampling

Approximate Randomization

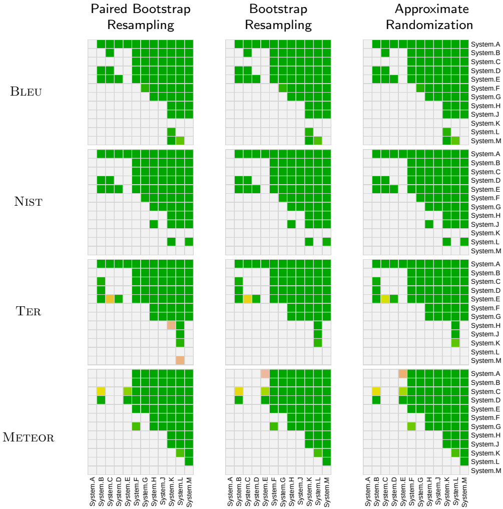
Pairwise Significance Tests (ES–EN)



Pairwise Significance Tests (ES–EN)



Pairwise Significance Tests (ES–EN)



Accuracy (%) for ES–EN

p		Paired Bootstrap		Bootstrap		Approx. Rand.	
0.05	BLEU	80.3	[68.7, 89.1]	80.3	[68.7, 89.1]	80.3	[68.7, 89.1]
	NIST	**81.8**	[70.4, 90.2]	**81.8**	[70.4, 90.2]	**81.8**	[70.4, 90.2]
	TER	78.8	[67.0, 87.9]	78.8	[67.0, 87.9]	78.8	[67.0, 87.9]
	METEOR	78.8	[67.0, 87.9]	78.8	[67.0, 87.9]	78.8	[67.0, 87.9]

Accuracy (%) for ES–EN

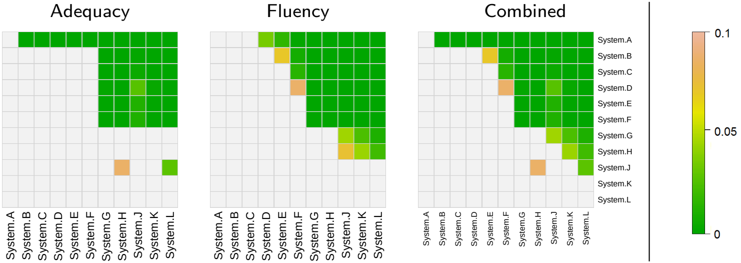
p		Paired Bootstrap	Bootstrap	Approx. Rand.
0.05	Bleu	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]
	Nist	**81.8** [70.4, 90.2]	**81.8** [70.4, 90.2]	**81.8** [70.4, 90.2]
	Ter	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]
	Meteor	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]	78.8 [67.0, 87.9]
0.01	Bleu	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	Nist	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	Ter	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]	77.3 [65.3, 86.7]
	Meteor	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]	80.3 [68.7, 89.1]

Accuracy (%) for ES–EN

p		Paired Bootstrap		Bootstrap		Approx. Rand.	
0.05	Bleu	80.3	[68.7, 89.1]	80.3	[68.7, 89.1]	80.3	[68.7, 89.1]
	Nist	**81.8**	[70.4, 90.2]	**81.8**	[70.4, 90.2]	**81.8**	[70.4, 90.2]
	Ter	78.8	[67.0, 87.9]	78.8	[67.0, 87.9]	78.8	[67.0, 87.9]
	Meteor	78.8	[67.0, 87.9]	78.8	[67.0, 87.9]	78.8	[67.0, 87.9]
0.01	Bleu	77.3	[65.3, 86.7]	77.3	[65.3, 86.7]	77.3	[65.3, 86.7]
	Nist	77.3	[65.3, 86.7]	77.3	[65.3, 86.7]	77.3	[65.3, 86.7]
	Ter	77.3	[65.3, 86.7]	77.3	[65.3, 86.7]	77.3	[65.3, 86.7]
	Meteor	80.3	[68.7, 89.1]	80.3	[68.7, 89.1]	80.3	[68.7, 89.1]
0.001	Bleu	72.7	[60.4, 83.0]	72.7	[60.4, 83.0]	72.7	[60.4, 83.0]
	Nist	72.7	[60.4, 83.0]	72.7	[60.4, 83.0]	72.7	[60.4, 83.0]
	Ter	75.8	[63.6, 85.5]	77.3	[65.3, 86.7]	78.8	[67.0, 87.9]
	Meteor	80.3	[68.7, 89.1]	80.3	[68.7, 89.1]	78.8	[67.0, 87.9]

Human Assessment Ranking (EN–ES)

System comparison based on human assessments (EN–ES):



Pairwise Significance Tests (EN–ES)

Bleu

Paired Bootstrap Resampling

	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
			🟩	🟩		🟩	🟩	🟩	🟩	🟩
			🟩	🟩		🟩	🟩	🟩	🟩	🟩
				🟩		🟩	🟩			🟩
						🟩	🟩			🟩
	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩
							🟩			🟩
			🟩	🟩		🟩	🟩			🟩
			🟩	🟩		🟩	🟩	🟩		🟩

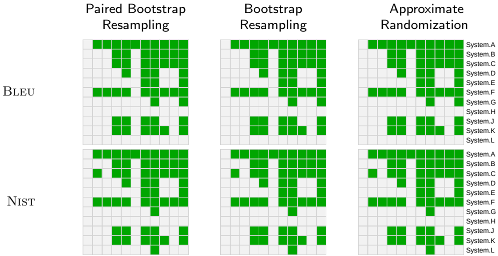
Bootstrap Resampling

	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
			🟩	🟩		🟩	🟩	🟩	🟩	🟩
			🟩	🟩		🟩	🟩	🟩	🟩	🟩
				🟩		🟩	🟩			🟩
						🟩	🟩			🟩
	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩
							🟩			🟩
			🟩	🟩		🟩	🟩			🟩
			🟩	🟩		🟩	🟩	🟩		🟩

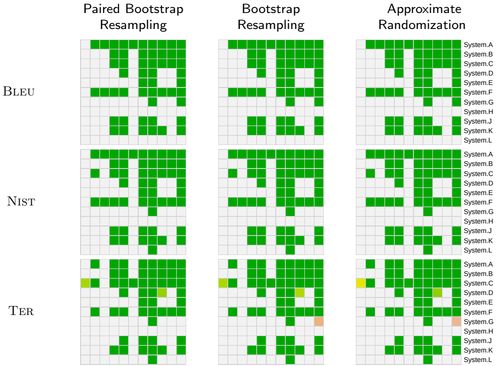
Approximate Randomization

	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	System.A
			🟩	🟩		🟩	🟩	🟩	🟩	🟩	System.B
			🟩	🟩		🟩	🟩	🟩	🟩	🟩	System.C
				🟩		🟩	🟩			🟩	System.D
						🟩	🟩			🟩	System.E
	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	System.F
							🟩			🟩	System.G
											System.H
			🟩	🟩		🟩	🟩			🟩	System.J
			🟩	🟩		🟩	🟩	🟩		🟩	System.K
											System.L

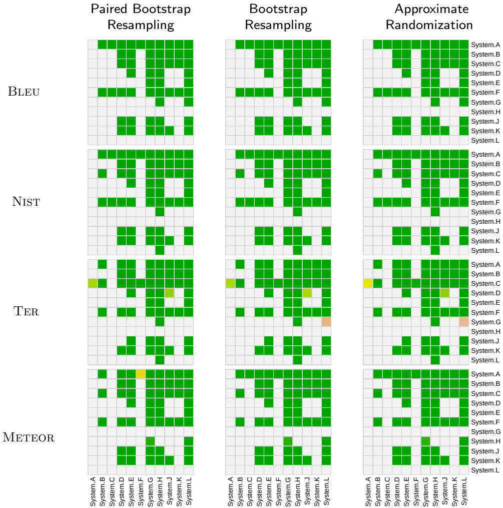
Pairwise Significance Tests (EN–ES)



Pairwise Significance Tests (EN–ES)



Pairwise Significance Tests (EN–ES)



Paired Bootstrap Resampling

Bootstrap Resampling

Approximate Randomization

BLEU

NIST

TER

METEOR

MT Significance Test Summary

- Very little difference between the three significance tests for either grouping of systems/language pair
- Differences between MT evaluation metrics, but within metric, very little difference across tests
- In terms of agreement with the human evaluations at $p < 0.05$:
  - for ES–EN, Nist the most accurate (82% agreement)
  - for EN–ES, Bleu the most accurate (62%(!) agreement)

Talk Outline

Sampling Distribution & Confidence Intervals

MT System Significance Tests

MT Metric Significance Tests

Significance Testing I

Common evaluation setting: Assess performance of Metric$_{new}$, we compare

- Correlation achieved by new metric $r(Metric_{new}, Human) = 0.9$ with
- That of the baseline metric $r(Metric_{Baseline}, Human) = 0.8$

A common mistake:

- Apply an individual significance test to each correlation;
- Conclude $r(Metric_{Baseline}, Human)$ is significant;
- And probably that $r(Metric_{Baseline}, Human)$ is also significant;
- Significant result – yay! (hmmmmm...)

Significance testing individual correlations only tells you if the correlation is significantly different from zero, not the correct question!

Significance Testing II

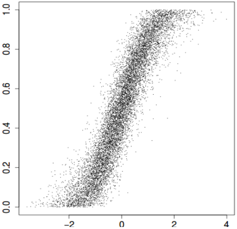
Much more meaningful: Test the significance of the **difference** in correlation!

- New Metric: $r(Metric_{new}, Human) = 0.9$
- Baseline Metric: $r(Metric_{Baseline}, Human) = 0.8$
- Instead: test if 0.1 for significance!
- (In the correct context for the data we are dealing with!)

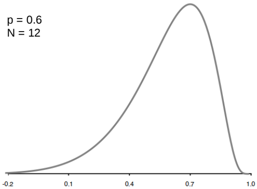
Pearson Correlation Sampling Distribution

The sampling distribution can be dramatically different from the sampling distribution of the mean (the usual example), depending on the statistic.

Population



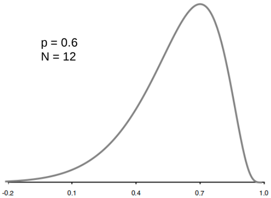
Sampling Distribution

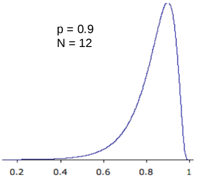


Pearson correlation:

- Bivariate data analysis: two variables
- Example from MT: Metrics are evaluated by correlation of eg. BLEU scores with human assessment.

Pearson Correlation Sampling Distribution for different ρ





- Sampling distribution is skewed due to the correlation coefficient not being able to exceed +1 or fall below -1;
- The closer ρ is to 1 or -1, the more extreme the skew;
- Further complications arise from correlation coefficients not being additive;
- All of this, makes dealing with significance of differences in correlations (MT metrics) much less straight-forward than difference in means (MT systems).

Independent Samples? I

- In medicine and psychology – often the case that data is independent – specific test for correlations in this case;
- Example: Two separate samples, interested in two relationships:
  - Hours of education received by mother and salary of child
    r(mom_ed,kid_salary) = 0.8
  - Hours of education recede by father and salary of child
    r(dad_ed,kid_salary) = 0.7

Is the correlation between mom's higher education and child's salary significantly stronger than that of father's?

Independent Samples? II

Is the correlation between mom's education and child's salary significantly stronger than that of father's?

Subject	Educ. (hrs)	Kid Salary ($)
Mary	21k	65k
Alice	50k	55k
Maria	60k	53k
.		
.		
.		

Subject	Educ. (hrs)	Kid Salary ($)
Jim	22k	57k
Jack	51k	55k
Fred	69k	53k
.		
.		
.		

Since we don't have any correspondence between the subjects – independent data.

If we had the following instead for both parents of the same child – dependent (or paired) data:

Subject	Mom Educ. (hrs)	Dad Educ. (hrs)	Kid Salary ($)
Kid 1 (Mary & Joe's Kid)	21k	27k	65k
Kid 2 (Alice & Mick's Kid)	50k	48k	55k
Kid 3 (Maria & Graham's Kid)	60k	66k	53k
.			
.			
.			

Metric Evaluation: Dependent Data

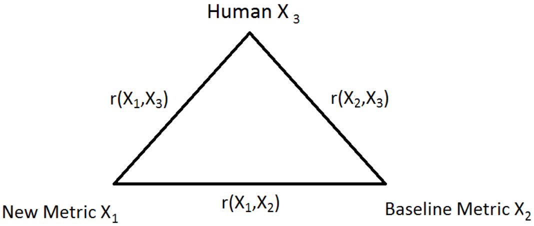
Nearly all cases metric evaluation data will be dependent data.

Example, Machine Translation:

Subject	pink	Bleu	Human Assessment ($)
MT System 1 (Mary & Joe's MT System)	21.45	27.8	65%
MT System 2 (Alice & Mick's MT System)	50.23	48.5	55%
MT System 3 (Maria & Graham's MT System)	60.12	66.0	53%
.			
.			
.			

- Dependent data!
- Question: is the difference between r(pink,human) significantly greater than r(Bleu,human)?

Williams Test (Hotelling-Williams) I



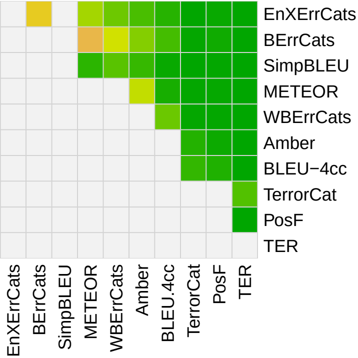
- Williams: test for difference in dependent correlations;
- Suited to many kinds of Metric evaluation & MT Quality Estimation evaluation;
- Takes dependent nature of data into account;
- Test is more powerful test when r(X_1,X_2) stronger;
- See Graham and Baldwin (2014) EMNLP paper for more details.

Williams Test (Hotelling-Williams) II

How to run this test?

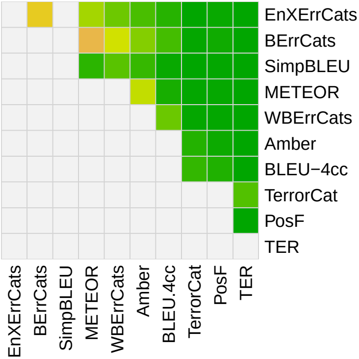
- See `https://github.com/ygraham/nlp-williams` for R code;

Example: MT Metrics Correlation with Human Judgment



WMT-13 English to French

Example: MT Metrics Correlation with Human Judgment



WMT-13 English to French

Questions!

References I

S. Banerjee and A. Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgements. In *Proc. Wkshp. Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–73, Ann Arbor, MI, 2005.

Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. Continuous measurement scales in human evaluation of machine translation. In *Proc. 7th Linguistic Annotation Wkshp. & Interoperability with Discourse*, pages 33—-41, Sofia, Bulgaria, 2013. ACL.

Y. Graham, N. Mathur, and T. Baldwin. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–74, Baltimore, MA, 2014.

P. Koehn. Statistical significance tests for machine translation evaluation. In *Proc. of Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, 2004. ACL.

NIST. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report, 2002.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. A Method for Automatic Evaluation of Machine Translation. In *Proc. 40th Ann. Meeting of the Assoc. Computational Linguistics*, pages 311–318, Philadelphia, PA, 2002.

S. Riezler and J.T. Maxwell. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 57–64, Ann Arbor, MI, 2005. ACL.

Mathew Snover, Bonnie Dorr, Richard Scwartz, John Makhoul, Linnea Micciula, and Ralphe Weischeidel. A Study of Translation Error Rate with Targeted Human Annotation. Technical report, College Park, MD, 2005.