IN4080 – 2019 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning

Evaluation and variation

Lecture 10, 17 Oct

Why statistics in evaluation?

Task1

- You know the best classifier on a task has 0.8 (80%) accuracy (baseline).
- You have made a classifier which classify 85 items correctly on a test set of 100 items.
- Can you conclude your classifier is better than the baseline?

Task 2

- You have made a classifier. You test it on 500 items. It classifies 375 correctly.
- What is the accuracy of your classifier?

Why?

(Task 3)

- (You have two different classifiers, one with accuracy 0.89 and one with accuracy 0.91on 1000 test items.
- Can you conclude that one is better than the other?)

Task 4

- The two classifiers from task 3 agree on 870 items.
- One is doing better on 20 items, the other is doing better on 40 items.
- Can we draw conclusions from this?

Why?

Selecting test items corresponds to

- drawing balls from a urn:
  - The ones we classify correctly
    - Blue balls
  - The ones we classify incorrectly
    - Red balls
- This corresponds nearly to flipping a coin, or throwing a dice
  - Given:
    - We choose randomly
    - Each choice is done independently
    - The urn is large (corresponding to drawing with replacement)



Vancouver Sun, «IKEA ballroom»

Task1

- Task1:
  - You know the best classifier on a task has 0.8 (80%) accuracy (baseline).
  - You have made a classifier which classify 85 items correctly on a test set of 100 items.
  - Can you conclude your classifier is better than the baseline?

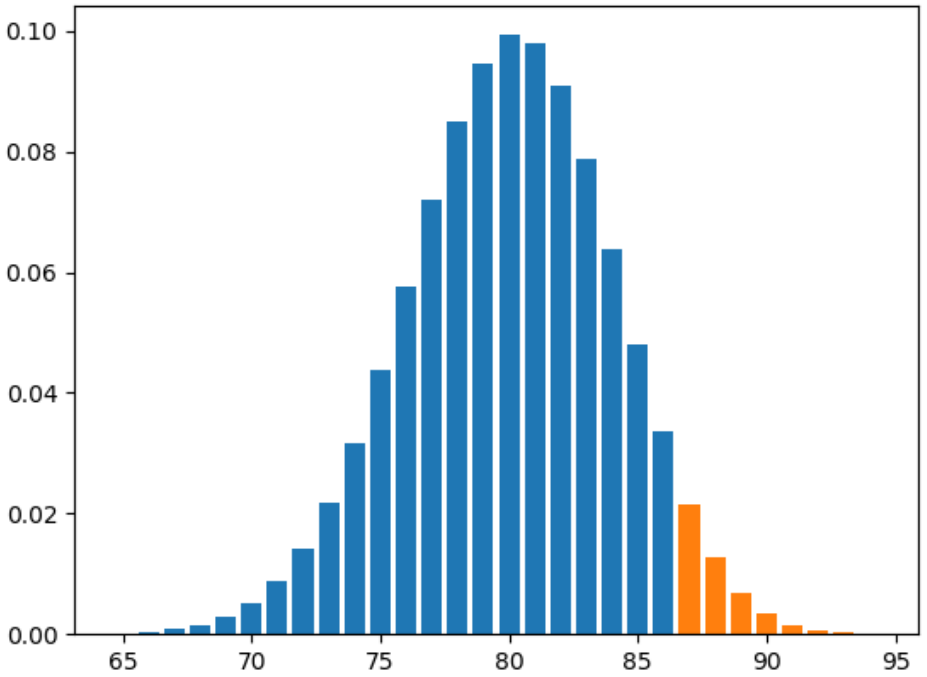
Formulate the question:
"Given that my classifier has an accuracy of 0.8 how likely is it to get a result as extreme or more extreme than getting 85 out of 100"



Vancouver Sun, «IKEA ballroom»

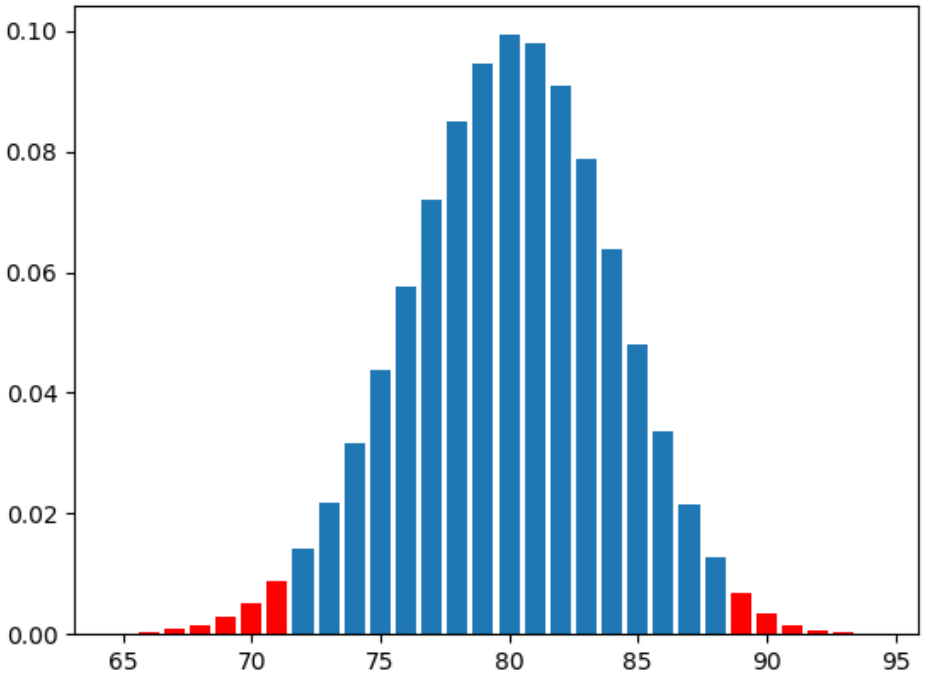
Task 1

- This follows a binomial distribution B(100, 0.8)
  - $\sigma = \sqrt{n * p(1-p)} = \sqrt{100 * 0.8 * (1-0.8)} = 4$
- The likelihood of getting 85 or more:
  $$pmf(85) + pmf(86) + \cdots pmf(100) = 1 - cdf(84) = 0.1285$$
- This is called the P-value of the test
- Since it is more than 0.05 we will not conclude that new classifier is better than the baseline.
- To have a result which is significant at the $\alpha = 0.05$-level, i.e., $p < \alpha$, the classifier must be correct at least 87 times.



One-sided and two-sided tests

- There is a difference between asking whether
  - our classifier is better than baseline (=one-sided)?
  - or different from baseline (=two-sided)?
- For two-sided test, $\alpha = 0.05$, ask whether our observation has a P-value below $\alpha/2$ or above $1 - \alpha/2$

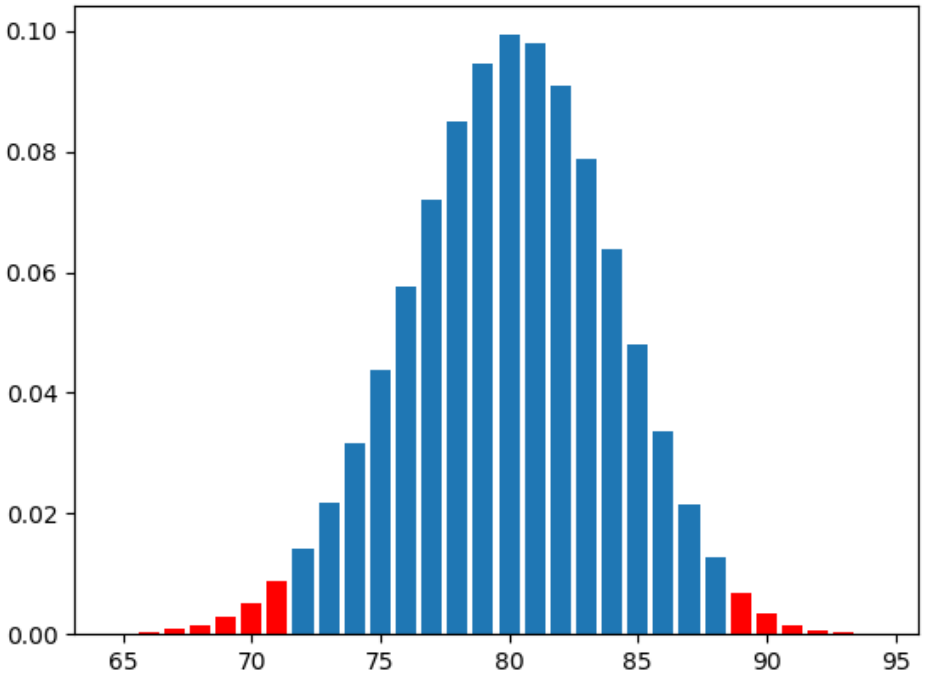


One-sided and two-sided tests

- In [437]:
  stats.binom.ppf(.975,100,0.8)
- Out[437]: 88.0

- In [438]:
  stats.binom.ppf(.025,100,0.8)
- Out[438]: 72.0

- Hence our statistics must be 71 or less, or 88 or more



Larger numbers

- What if the baseline is 0.8 and 1000 test items?
- What if the baseline is 0.8 and 10000 test items?

```
In [429]: stats.binom.ppf(.95,100,0.8)
Out[429]: 86.0
In [430]: stats.binom.ppf(.95,200,0.8)
Out[430]: 169.0
In [433]: stats.binom.ppf(.95,400,0.8)/400
Out[433]: 0.8325
In [435]: stats.binom.ppf(.95,10000,0.8)
Out[435]: 8066.0
```

Sample size	100	200	400	1000	10000	
Number of correct items to beat baseline	87	170	334	822	8067	
Proportion of correct items to beat baseline	0.87	0.85	0.835	0.822	0.8067	

So far

- Be careful to conclude from observed numbers
- Binomial distribution can be used for determining whether an observed accuracy beats an established baseline
- But don't trust "statistical significance" blindly:
  - We do not always fulfill the requirements for an experiment.
  - We must specify the hypothesis before we conduct the experiment
  - If we conduct more than one experiment, we must adjust our confidence
    - If there is a chance of 5% mistakes in one test,
    - There is a chance of 40% mistakes in conducting 10 tests.

Estimation: Task 2

Example

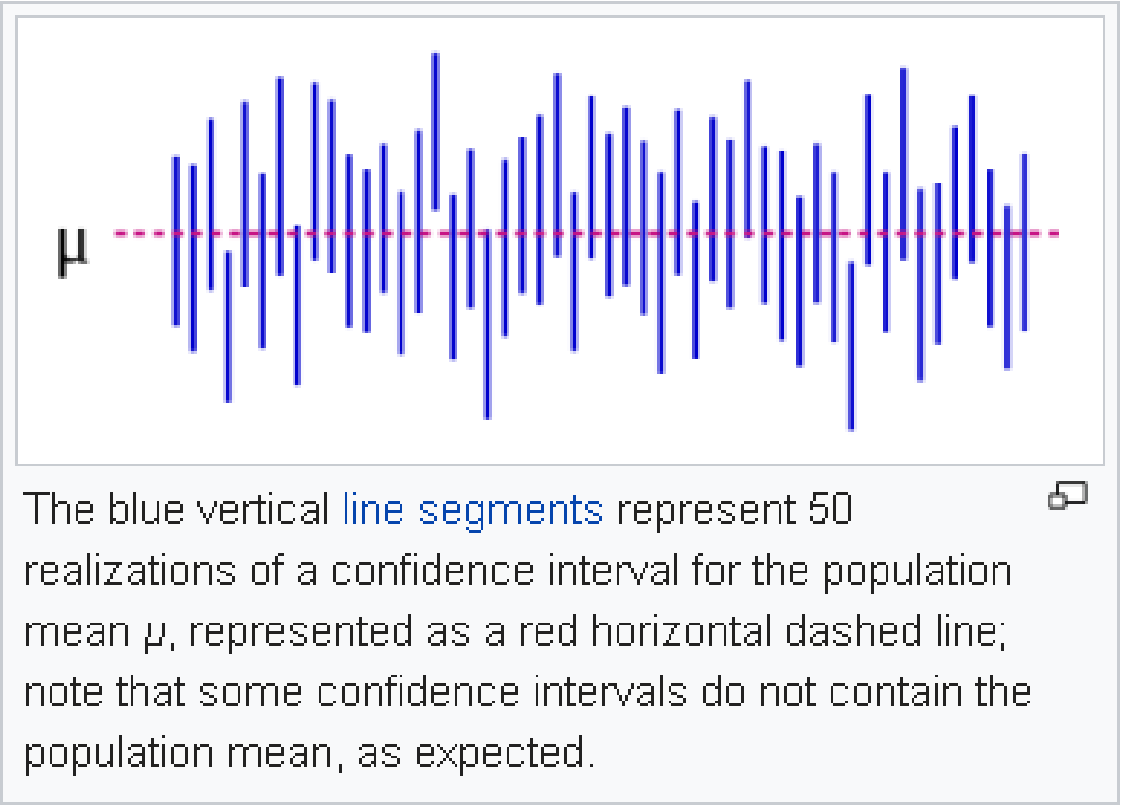
Task 2:

- You have made a classifier. You test it on 500 items. It classifies 375 correctly.
- What is the accuracy of your classifier on the whole population?

- Our best guess is $\frac{375}{500} = 0.75$
- This is probably not accurate
- Can we be 95% sure that the true accuracy is between 0.7 and 0.8?
- What about 0.72 and 0.78?

Confidence interval

- We decide on a confidence level, C, e.g., 0.95
- We are looking for an interval [*p_low, p_high*] for which we can be confident that the true *p* is within the interval.



The blue vertical line segments represent 50 realizations of a confidence interval for the population mean μ, represented as a red horizontal dashed line; note that some confidence intervals do not contain the population mean, as expected.

Estimating a confidence interval

- Suppose we have a population with an unknown proportion p.
- We select a random sample of n individuals from this population and measure its proportion $\hat{p}$.
- If we had known p, we know how to find an ε such that
  - $P(p - \varepsilon < \hat{p} < p + \varepsilon) > C$

  for a given C.
  - We could use software for the binomial distribution as we did in task 1.
  - Or, if we know some statistics, we could find it by using the normal approximation

Estimating a confidence interval

- As programmers, we could go along, try succesive potential values for p and check whether they fullfill the criteria, and we are done.
- But we can be smarter than that.
- Observe that
  - $P(p - \varepsilon < \hat{p} < p + \varepsilon) > C$
  - if and only if
  - $P(\hat{p} - \varepsilon < p < \hat{p} + \varepsilon) > C$
- One problem: ε depends on the distribution of the population, i.e., on p
- But what we can do, is to use the best approximation we have for p which is $\hat{p}$.

Using the binomial distribution

- We use the binomial distribution B(n, $\hat{p}$) and ask for its .025- and .975-percentiles
- p_low is the 0.025 percentile
- p_high is the 0.975 percentile

- In other words, we assume that p is below the .975-percentile for B(n, $\hat{p}$) iff $\hat{p}$ is above the .025-percentile of B(n, p) and similarly from the other side
- This assumption works well if
  - $n*\hat{p}>10$ and $n*(1-\hat{p})>10$

(Using the normal approximation)

- Instead of $\sigma = \sqrt{\frac{p(1-p)}{n}}$ we use the best estimate we have.
- The best estimate we have for p is $\hat{p}$
- The best estimate we have for σ is $s = \mathrm{SE}(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$
- Hence the confidence interval is $[\hat{p} - Z * s, \hat{p} - Z * s]$ where Z depends on the chosen confidence level C.

- Footnote: Normally we compensate for the difference between the true standard deviation and the estimated standard error, and use t-distribution instead of normal distribution, but not when we consider proportions. We can stick to the normal distribution.

Example

- Using binomial distribution:
  - In [288]: stats.binom.ppf([0.025, 0.975], 500, 0.75)/500
  - Out[288]: array([0.712, 0.788])

- Using normal distribution approximation:
  - In [284]: stats.norm.ppf([0.025, 0.975],0.75, np.sqrt(0.75*0.25/500))
  - Out[284]: array([0.71204546, 0.78795454])

- $\hat{p} = \frac{375}{500} = 0.75$, $s = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \left(= \sqrt{\frac{0.75(1-0.75)}{500}} = 0.01936 \right)$
- $[\hat{p} - Z * s, \hat{p} + Z * s] = [0.75 - 1.96 * 0.01936, 0.75 + 1.96 * 0.01936] = [0.0712, 0.0788]$

Comparing populations/proportions

Why?

(Task 3)

- (You have two different classifiers, one with accuracy 0.89 and one with accuracy 0.91on 1000 test items.
- Can you conclude that one is better than the other?)

Task 4

- The two classifiers from task 3 agree on 870 items.
- One is doing better on 20 items, the other is doing better on 40 items.
- Can we draw conclusions from this?

Paired data

- We are considering the second question first.
- It is more useful.
- There is a simple test for it.

Example

		Classifier 2	
		correct	incorrect
Classifier 1	correct	870	40
	incorrect	20	70

- We can focus on the items were they disagree.
- We expect (C1-correct & C2-incorrect) and (C2-correct and C1-incorrect) to be equally likely
- How unlikely is it that from the 60 items where they disagree, C1 wins 20 (or fewer) times?

Sign test


```
In [69]: stats.binom.cdf(20,60,0.5)
Out[69]: 0.006744646865595934
```

- We see that this is unlikely
- We can reject the
  - null-hypothesis: classifier 2 is not better than classifier 1 (p= 0.01)

- This test is called the sign-test
- Can be used to numerical data:
  - Count the instances where x>y or y>x
- When used on Boolean values, {0, 1}, often called McNemar's test (with some variation)
- We may compare item to item, or by folds in an n-fold cross-val.

A few reamining items in evaluation

Macro- and micro-averaging

- We have earlier seen how we may compute precision and recall in a multi-class setting.
- There are two ways to take the average of these measures
  - Macro-averaing
  - Micro-averaging

Macro-averaging

- Calculate the value for each class
- Sum these and divide by the number of classes
- Small classes get same weight as big classes
  - (The American senate)

$$Pr^M = \frac{1}{L}\sum_{i=1}^{L} pr_i$$

$$Re^M = \frac{1}{L}\sum_{i=1}^{L} re_i$$

$$F_1^M = \frac{1}{L}\sum_{i=1}^{L} F_{1i}$$

Micro-averaging

- Calculate item by item
- All items get equal weight
- Large classes get higher weight than smaller classes

$$Pr^{\mu} = \frac{\sum_{i=1}^{L} N_{TP_i}}{\sum_{i=1}^{L} (N_{TP_i} + N_{FP_i})}$$

$$Re^{\mu} = \frac{\sum_{i=1}^{L} N_{TP_i}}{\sum_{i=1}^{L} (N_{TP_i} + N_{FN_i})}$$

$$F_1^{\mu} = \frac{2 \times Pr^{\mu} \times Re^{\mu}}{Pr^{\mu} + Re^{\mu}}$$

Learning curves

- It is useful to see the effect of the size of your training data, e.g.
  - split your training data into 10 equally sized blocks
  - Train on 1 block, than on 2 and so on up to 10.
  - Measure the effect.

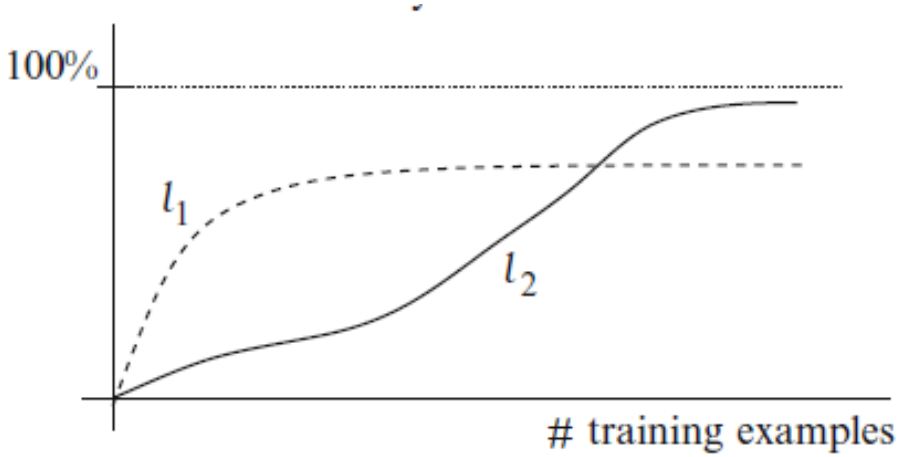


Fig. 11.3 A learning curve shows how the achieved classification accuracy depends on the number of examples used during learning

What we have done

1. How likely is a sample given a known population?
    1. How good is a classifier compared to a baseline?
2. From a sample, estimate an interval for the true mean value.
3. Comparing paired data.
4. Micro- and macro-average
5. Learning curves