Example

Max. wind speed (m/s)

95% probability of wind speed below the corresponding value

Norwegian Meteorological Institute  met.no
Overview

- Reliability
- Sharpness
- Refinement

- Ranking

An example of forecasting wind speed is used throughout
  - One station
  - 239 cases for the lead time used
Reliability (calibration)

Are the quantile probabilities proper/valid?

Statistic
Fractions of observations below each quantile

Example

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Torungen lighthouse</td>
<td>.083</td>
<td>.227</td>
<td>.541</td>
<td>.762</td>
<td>.923</td>
</tr>
</tbody>
</table>
Hypothesis tests

Does the forecast model produce reliable quantiles?

Each quantile separately

\[ H_0: \ p_{\text{true}} = p \quad (\text{true prob.} = \text{quantile prob.}) \]

- binomial test (preferable) or \( \chi^2 \)-test
  - R: binom.test(), chisq.test(), prop.test()

\( p \)-values are appropriate for presenting results
  - the smaller the \( p \)-value, the stronger the evidence that the model is unreliable
All quantiles simultaneously

I. \( H_0: p_{\text{true},i} = p_i \) for all quantiles \( i \)

II. \( H_0: p_{\text{true},i} = p_i \) for all intervals \( i \)
   - The \( p \)'s are interval probabilities
   - Intervals formed by (between) the quantiles
   - Number of intervals = number of quantiles + 1

- \( \chi^2 \)-tests appropriate for both tests
- 2nd test preferable (no overlapping classes)
- Different \( p \)-values!
Example

Statistics

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Torungen lighthouse</td>
<td>.083</td>
<td>.227</td>
<td>.541</td>
<td>.762</td>
<td>.923</td>
</tr>
</tbody>
</table>

P-values

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
<th>ALL*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Torungen lighthouse</td>
<td>.057</td>
<td>.493</td>
<td>.298</td>
<td>.732</td>
<td>.120</td>
<td>.008</td>
</tr>
</tbody>
</table>

*) p-value based on intervals

Norwegian Meteorological Institute  met.no
The lead time used above
Remarks

- Decision making
  - Choose test(s)
  - Fix significance level
  - Require p-value(s) above this

- $\chi^2$-tests are approximate
  - Thumb rule: expected counts in each cell should be greater than 5 (conservative)
Conditional reliability

- Previous methods only check *overall* reliability (unconditional reliability)

- Quantile probabilities should be valid for *all* forecasts

- Does the reliability depend on
  - Forecasted value?
  - Lead time, time, season, ...?
Stratification of data

- Sort forecasts by e.g. value (for each quantile prob.)
- Group data (e.g. roughly equal sizes)
- Compute statistics for each group

Example

<table>
<thead>
<tr>
<th>Quantile value</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>.117</td>
<td>.333</td>
<td>.550</td>
<td>.767</td>
<td>.950</td>
</tr>
<tr>
<td>Medium</td>
<td>.049</td>
<td>.197</td>
<td>.541</td>
<td>.754</td>
<td>.934</td>
</tr>
<tr>
<td>High</td>
<td>.083</td>
<td>.150</td>
<td>.533</td>
<td>.767</td>
<td>.883</td>
</tr>
</tbody>
</table>
Hypothesis tests

Each quantile separately (by value)

I. \( H_0: p_{\text{low}} = p_{\text{med}} = p_{\text{high}} \)

II. \( H_0: p_{\text{low}} = p_{\text{med}} = p_{\text{high}} = p \) (quantile prob.)

- \( \chi^2 \)-tests can be used in both cases
  - R: prop.test()
- Test II is most complete/relevant
- Same principle for testing all quantiles simultaneously
Example

Statistics

<table>
<thead>
<tr>
<th>Quantile value</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>.117</td>
<td>.333</td>
<td>.550</td>
<td>.767</td>
<td>.950</td>
</tr>
<tr>
<td>Medium</td>
<td>.049</td>
<td>.197</td>
<td>.541</td>
<td>.754</td>
<td>.934</td>
</tr>
<tr>
<td>High</td>
<td>.083</td>
<td>.150</td>
<td>.533</td>
<td>.767</td>
<td>.883</td>
</tr>
</tbody>
</table>

P-values

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (homogenity)</td>
<td>.404</td>
<td>.045</td>
<td>.983</td>
<td>.983</td>
<td>.359</td>
</tr>
<tr>
<td>II (joint)</td>
<td>.071</td>
<td>.096</td>
<td>.735</td>
<td>.980</td>
<td>.115</td>
</tr>
<tr>
<td>unconditional</td>
<td>.057</td>
<td>.493</td>
<td>.298</td>
<td>.732</td>
<td>.120</td>
</tr>
</tbody>
</table>
Regression based tests (conditional reliability)

- Logistic regression for each quantile prob.

\[
\log\left(\frac{p_{true}}{1-p_{true}}\right) = \alpha_0 + \alpha_1 q_p
\]

\(H_0: \alpha_1 = 0\) (no trend)

\(H_0: \alpha_1 = 0\) and \(\alpha_0 = \log(p/1-p)\) (no trend and proper prob.)

Likelihood ratio tests?
Sharpness

Probability mass should be distributed on short interval(s)

Several quantiles
- Average length(s) of intervals formed by pair(s) of quantiles
  - Ex.: average length of 50% and 90% intervals
  - Bimodality is often penalised too much
  - Empirical distributions provide additional information
  - Single number would be useful for decision making

Single quantile
- Variation as measured by standard deviation or range (as for deterministic forecasts)
Refinement / Variation

Information about uncertainty is less important if it is constant

Measures
- Standard deviation (or range) of interval lengths
- Deviation from climate quantiles?
Ranking quantile forecasts

Score functions
- Discrete ranked probability score (RPS) is not suitable
  • Sharpness is not given credit
- Make complete CDFs (and PDFs?) of the quantiles and use CRPS or other scoring rules (not easy)
  • Approximate CRPS by integrating only over the range of the quantiles

Reliability and sharpness
- Require reliability at a given significance level and rank reliable models by average interval length(s)
- Most suitable in the process of making forecast models
Summary

Reliability
- Hypothesis tests useful
- Important to also assess cond. reliability

Sharpness
- Length of forecast intervals

Ranking models important problem
- Scoring rules useful