Categorical Verification

Tara Jensen

Contributions from Matt Pocernich, Eric Gilleland, Tressa Fowler, Barbara Brown and others
Finley Tornado Data (1884)

Forecast answering the question:

Will there be a tornado?

YES

NO

Observation answering the question:

Did a tornado occur?

YES

NO

Answers fall into 1 of 2 categories

Forecasts and Obs are Binary
## Finley Tornado Data (1884)

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

Contingency Table
# A Success?

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>28</td>
<td>72</td>
<td>100</td>
</tr>
<tr>
<td>No</td>
<td>23</td>
<td>2680</td>
<td>2703</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>2752</td>
<td>2803</td>
</tr>
</tbody>
</table>

Percent Correct = \(\frac{28+2680}{2803} = 96.6\% \) !!!!
What if forecaster never forecasted a tornado?

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>51</td>
<td>2752</td>
<td>2803</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>2752</td>
<td>2803</td>
</tr>
</tbody>
</table>

Percent Correct = (0 + 2752) / 2803 = 98.2% !!!!
maybe Accuracy is not the most informative statistic

But the contingency table concept is good…
### 2 x 2 Contingency Table

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>Hit</td>
<td>False Alarm</td>
<td>Forecast Yes</td>
</tr>
<tr>
<td>No</td>
<td>Miss</td>
<td>Correct Negative</td>
<td>Forecast No</td>
</tr>
<tr>
<td>Total</td>
<td>Obs. Yes</td>
<td>Obs. No</td>
<td>Total</td>
</tr>
</tbody>
</table>

**Example:** Accuracy = \((\text{Hits} + \text{Correct Negs})/\text{Total}\)

MET supports both 2x2 and NxN Contingency Tables
Common Notation
(However not universal notation)

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>n</td>
</tr>
</tbody>
</table>

**Example:** Accuracy = (a+d)/n
What if data are not binary?

**Threshold**

- Temperature $< 0$ C
- Precipitation $> 1$ inch
- CAPE $> 1000$ J/kg
- Ozone $> 20$ µg/m³
- Winds at 80 m $> 24$ m/s
- 500 mb HGTS $< 5520$ m
- Radar Reflectivity $> 40$ dBZ
- MSLP $< 990$ hPa
- LCL $< 1000$ ft
- Cloud Droplet Concentration $> 500$/cc

**Hint:** Pick a threshold that is meaningful to your end-user.

Copyright 2015, University Corporation for Atmospheric Research, all rights reserved
Contingency Table for Freezing Temps (i.e. $T \leq 0$ C)

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\leq 0$C</td>
<td>$&gt; 0$C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\leq 0$C</td>
<td>a</td>
<td>b</td>
<td></td>
<td>a+b</td>
</tr>
<tr>
<td>$&gt; 0$C</td>
<td>c</td>
<td>d</td>
<td></td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td></td>
<td>n</td>
</tr>
</tbody>
</table>

Another Example:
Base Rate (aka sample climatology) = $\frac{(a+c)}{n}$
Alternative Perspective on Contingency Table

\[ \begin{align*}
\text{Correct Negatives} & \quad d \\
\text{False Alarms} & \quad b \\
\text{Hits} & \quad a \\
\text{Misses} & \quad c
\end{align*} \]

Copyright 2015, University Corporation for Atmospheric Research, all rights reserved
Conditioning to form a statistic

- Considers the probability of one event given another event
- Notation: $p(X|Y=1)$ is probability of $X$ occurring given $Y=1$ or in other words $Y=yes$

**Conditioning on Fcst provides:**
- Info about how your forecast is performing
- Apples-to-Oranges comparison if comparing stats from 2 models

**Conditioning on Obs provides:**
- Info about ability of forecast to discriminate between event and non-event - also called Conditional Probability or “Likelihood”
- Apples-to-Apples comparison if comparing stats from 2 models
Conditioning on forecasts

Forecast = yes

\[ f=1 \]

Observed = yes

\[ x=1 \]

\[ p(x|f=1) = \frac{a}{a+b} = \text{Fraction of Hits} \]

\[ p(x=0|f=1) = \frac{b}{a+b} = \text{False Alarm Ratio} \]
Conditioning on observations

Forecast = yes
f = 1

Observed = yes
x = 1

\[ p(f=1|x=1) = \frac{a}{a+c} = \text{Hit Rate} \]

\[ p(f=0|x=1) = \frac{c}{a+c} = \text{Fraction of Misses} \]
What’s considered good?

**Conditioning on Forecast**

Fraction of hits - \( p(x=1|f=1) = \frac{a}{a+b} \): close to 1

False Alarm Ratio - \( p(x=0|f=1) = \frac{b}{a+b} \): close to 0

**Conditioning on Observations**

Hit Rate - \( p(f=1|x=1) = \frac{a}{a+c} \): close to 1

\[aka\ Probability\ of\ Detection\ Yes\ (PODy)\]\n
Fraction of misses \( p(f=0|x=1) = \frac{a}{a+c} \): close to 0
Examples of Categorical Scores
(most based on conditioning)

- Hit Rate (PODy) = \(a/(a+c)\)
- PODn = \(d/(b+d) = (1 - POFD)\)
- False Alarm Rate (POFD) = \(b/(b+d)\)
- False Alarm Ratio (FAR) = \(b/(a+b)\)
- (Frequency) Bias (FBIAS) = \((a+b)/(a+c)\)
- Threat Score or Critical Success Index = \(a/(a+b+c)\)

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>
Examples of CTC calculations

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>28</td>
<td>72</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>23</td>
<td>2680</td>
<td>2703</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>2752</td>
<td>2803</td>
<td></td>
</tr>
</tbody>
</table>

Threat Score = 28 / (28 + 72 + 23) = 0.228
Probability of Detection = 28 / (28 + 23) = 0.55
False Alarm Ratio = 72 / (28 + 72) = 0.720
Example

Timeseries of PODY and FAR for 20% Ramp Events during daylight hours

Perfect

No Skill

Neutral Skill
50% Prob of Detection
60% False Alarm Ratio

Good Skill
80% Prob of Detection
30% False Alarm Ratio

High Skill
100% Prob of Detection
0% False Alarm Ratio

Copyright 2015, University Corporation for Atmospheric Research, all rights reserved
Relationships among scores

- CSI is a *nonlinear* function of POD and FAR
- CSI depends on base rate (event frequency) and Bias

\[
\text{CSI} = \frac{1}{1 + \frac{1}{\text{POD}} + \frac{1}{1 - \text{FAR}}} - 1
\]

\[
\text{Bias} = \frac{\text{POD}}{1 - \text{FAR}}
\]

Very different combinations of FAR and POD lead to the same CSI value
HMT Performance Diagram

All on same plot
- POD
- 1-FAR (aka Success Ratio)
- CSI
- Freq Bias

Dots: Scores Aggregated Over Lead Time
Colors: Different Thresholds

Here we see:
- Decreasing skill with higher thresholds even with multiple metrics
- Highest skill at 18-24h leads

Roberts et al. (2011), Roebber (WAF, 2009), Wilson (presentation, 2008)
Skill Scores

How do you compare the skill of easy to predict events with difficult to predict events?

• Provides a single value to summarize performance.
• Reference forecast - best naive guess; persistence; climatology.
• Reference forecast must be comparable.
• Perfect forecast implies that the object can be perfectly observed.
Generic Skill Score

$$SS = \frac{(A - A_{ref})}{(A_{perf} - A_{ref})}$$  

where \( A = \) any measure
\( ref = \) reference
\( perf = \) perfect

Example:  
$$MSESS = 1 - \frac{MSE}{MSE_{climo}}$$  

where \( MSE = \) Mean Square Error

Interpreted as fractional improvement over reference forecast

Reference could be: Climatology, Persistence, your baseline forecast, etc..

Climatology could be a separate forecast or a gridded forecast sample climatology

SS typically positively oriented with 1 as optimal
Commonly Used Skill Scores

- **Gilbert Skill Score** - based on the CSI corrected for the number of hits that would be expected by chance.

- **Heidke Skill Score** - based on Accuracy corrected by the number of hits that would be expected by chance.

- **Hanssen-Kuipers Discriminant** – (Pierce Skill Score) measures the ability of the forecast to discriminate between (or correctly classify) events and non-events. H-K=POD-POFD

- **Brier Skill Score** for probabilistic forecasts

- **Fractional Skill Score** for neighborhood methods

- **Intensity-Scale Skill Score** for wavelet methods
Example

Timeseries of CSI, GSS and Base Rate for 20% Ramp Events during daylight hours

- **Optimal**
- **High Skill**
  - 100% of events were forecasted properly
- **Moderate Skill**
  - 50% of all events were forecasted properly
- **Lower Skill**
  - 30% of all events were forecasted properly
- **Moderate Base Rate/Climatology**
  - 20-30% of all observations times were events
- **Low Base Rate/Sample Climatology**
  - 10% of all observations times were events

LOCAL TIME

Copyright 2015, University Corporation for Atmospheric Research, all rights reserved
Thank you!

References:


WMO Verification working group forecast verification web page,