6 Probabilistic forecasts slide 217

6.1 Introduction slide 218

A fundamental step in the scientific reasoning

In a discussion article on the application of mathematics in meteorology, Bigelow (1905) said

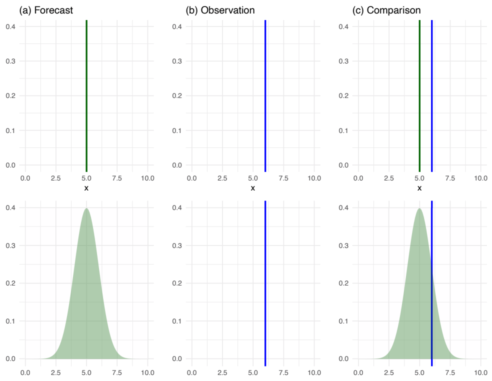
> There are three processes that are generally essential for the complete development of any branch of science, and they must be accurately applied before the subject can be considered to be satisfactorily explained. The first is the discovery of a mathematical analysis, the second is the discussion of numerous observations, and the third is a correct application of the mathematics to the observations, including a demonstration that these are in agreement.

The main topic for the rest of the semester is the last item on Bigelow's list, i.e., methods to demonstrate the agreement between a model and a set of observations

We will focus on forecast evaluation: out-of-sample model evaluation

Definition of forecast

A forecast is a prediction issued before the predicted quantity could be determined.



Use of probabilistic forecasts

- □ In finance: predictive distributions of assets;
- □ In marketing: predictive distributions of future sales and inventory;
- □ In economy: predictive distributions of inflation rates;
- □ Probabilistic population projections for the UN from predictive distributions of fertility and mortality rates, etc.
- □ In meteorology: predictive distributions of weather quantities;

The L’Aquila earthquake trial

In October 2012, an Italian court sentenced six leading scientists and a government official to six years of prison each for providing “incomplete, imprecise and contradictory information” (Hall 2011, p. 266) on the probability and risk of a major seismic event prior to the devastating earthquake that hit the city of L’Aquila on April 6, 2009.

□ statistical models can provide short-term probabilistic forecasts during periods of heightened seismic activity;

□ for very rare events, predicted probability is low and uncertainty is high;

□ communicating forecast uncertainty in low-probability high-risk environments is challenging.

Weather forecast

Weather forecast is a deterministic forecast relying on advanced numerical models of the atmosphere

□ set of initial conditions is the observed data;

□ one weather model simulation that provides a single prediction of the future weather;

□ does not explicitly quantify uncertainty—it assumes the single forecast is the best possible prediction;

□ limits the predictability of weather phenomena to few days.

Ensemble weather forecast

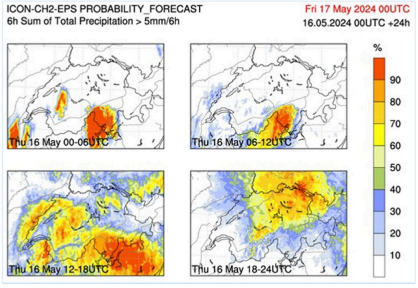
Ensemble weather forecast runs multiple simulations (typically, between 5 and 50) of the same numerical weather model

□ multiple sets of initial conditions and/or parameterised mathematical representation of the atmosphere (to capture uncertainty in observations and modelled physics of the atmosphere);

□ multiple simulations (ensemble members) to generate a range of possible outcomes;

□ provides a probabilistic forecast.

Ensemble weather forecast



MeteoSwiss probability maps generated with ICON-CH2-EPS (one of the two probabilistic forecasting system used by MeteoSwiss) indicating the probability that more than 1 mm of precipitation will fall in 6 hours within a given time period and at a particular location.

Ensemble weather forecast

Ensembles can be affected by biases and dispersion errors

□ If the numerical weather model consistently overestimates or underestimates a particular weather variable, the ensemble forecast will inherit this bias.

□ Ensembles often exhibit under-dispersion leading to overconfident forecasts.

Correction of these issues is done through calibration and post-processing techniques.

Ensemble weather forecast: Statistical post-processing

Consider an ensemble of point forecasts $x_1, \ldots, x_M$ for temperature at a given time and location

□ Bayesian model averaging (Raftery et al., 2005): mixture of parametric probability densities, each associated with a single ensemble member. The forecast density is

$$y \mid x_1, \ldots, x_M \sim \sum_{m=1}^{M} w_m \mathcal{N}(\mu_{0m} + \mu_{1m} x_m, \sigma^2)$$

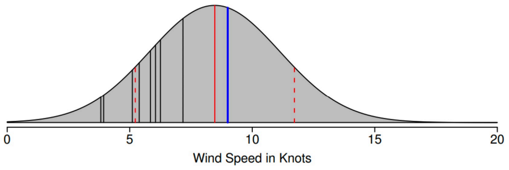
with non-negative weights $w_1, \ldots, w_M$, s.t. $\sum_{m=1}^M w_m = 1$, bias parameters $\mu_{01}, \ldots, \mu_{0M}$ and $\mu_{11}, \ldots, \mu_{1M}$, and a common variance σ^2.

□ Ensemble model output statistics (EMOS) of Gneiting et al. (2005): single predictive distribution with parameters depending on the ensemble values

$$y \mid x_1, \ldots, x_M \sim \mathcal{N}(a_0 + a_1 x_1 + \ldots + a_M x_M, b_0 + b_1 s^2)$$

with bias and spread parameters $a_0, a_1, \ldots, a_M$ and b_0 and b_1, respectively, and where s^2 is the empirical variance of the ensemble values.

Ensemble weather forecast: Statistical post-processing



48-hour ahead EMOS density forecast of maximum wind speed valid August 15, 2008 at The Dalles, Oregon. The black lines represent the M = 8 members of the University of Washington Mesoscale Ensemble (UWME). The red lines show the EMOS median forecast, at 8.5 knots, and the 77.8% central prediction interval for the EMOS density forecast, which ranges from 5.2 to 11.7 knots. The blue line represents the verifying observation, at 9 knots.

Outline

This chapter will discuss

□ point and probabilistic forecasts for continuous variables;

□ scoring rules for forecast evaluation;

□ testing predictive performance.

To answer the following questions

□ What constitutes a good forecast?

□ How do we evaluate the ”goodness” of a forecast?

Prediction space

Assume we issue a probabilistic forecast F for a real-valued observation Y.

- □ Denote by $\mathcal{A}$ the information set that contains the training data, expertise, theories, and assumptions at hand.
- □ $\mathcal{A}$ encodes the forecast's information set.
- □ Instead of one probabilistic forecast F, one can consider a collection of forecasts each described by an information set $\mathcal{A}_i$.

Definition 36 A prediction space is a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$ where elements of the sample space Ω are tuples (F, Y), with F a CDF-valued random quantity that is measurable with respect to $\mathcal{A}$ and Y a real-valued random variable. This probability space is equipped with a probability measure $\mathbb{Q}$ that specifies the joint distribution of the probabilistic forecast and the observation.

Prediction space

Definition 37 The CDF-valued random quantity F is ideal relative to the information set $\mathcal{A}$ if the distribution of $Y \mid \mathcal{A}$ is F, almost surely. Stated differently, the random quantity F is ideal when it makes the best possible use of available information.

Example 38 Let $Y \mid \mu \sim \mathcal{N}(\mu, 1)$, where $\mu \sim \mathcal{N}(0, 1)$. That is, nature draws a random number $\mu_t \sim \mathcal{N}(0, 1)$ that corresponds to the information at time t and picks the data-generating distribution $\mathcal{N}(\mu_t, 1)$. Then

- □ The perfect probabilistic forecast $\mathcal{N}(\mu, 1)$, i.e., with predictive distribution $F = \mathcal{N}(\mu, 1)$ is ideal w.r.t. the information set generated by μ.
- □ The climatological probabilistic forecast $\mathcal{N}(0, 2)$ (regardless of t) is ideal w.r.t. the trivial information set.
- □ The sign-reversed probabilistic forecast $\mathcal{N}(-\mu, 1)$ is not ideal.
- □ The unfocused forecast $\frac{1}{2}\{\mathcal{N}(\mu, 1) + \mathcal{N}(\mu + \xi, 1)\}$, for $\xi = \pm 1$ with equal probability, independent of Y, μ.

Note to Example 3

A CDF-valued random quantity F is ideal with respect to an information set $\mathcal{A}$ if the conditional distribution of Y given $\mathcal{A}$ is almost surely equal to F. We verify this property for the three forecasts:

□ The perfect probabilistic forecast $\mathcal{N}(\mu, 1)$ w.r.t. μ:

$$Y \mid \mu \sim \mathcal{N}(\mu, 1),$$

which is exactly the forecast F.

□ The climatological forecast $\mathcal{N}(0, 2)$ w.r.t. to the trivial information set:
We compute the marginal distribution of Y

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{Y|\mu}(y \mid \mu) f_\mu(\mu)\, d\mu \\ &= \int_{-\infty}^{\infty} \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{(y-\mu)^2}{2}\right) \tfrac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{\mu^2}{2}\right)\, d\mu \\ &= \tfrac{1}{2\pi} \exp\left(-\tfrac{y^2}{4}\right) \cdot \sqrt{\pi} = \tfrac{1}{2\sqrt{\pi}} \exp\left(-\tfrac{y^2}{4}\right). \end{aligned}$$

Thus, $Y \sim \mathcal{N}(0, 2)$ and the climatological forecast is ideal.

□ The sign-reversed forecast $\mathcal{N}(-\mu, 1)$ and the unfocused forecast are not ideal as they are different from the unconditional and conditional distributions of Y derived above.

Probability integral transform

To study the statistical compatibility between probabilistic forecasts and the corresponding realisations, we need the notion of randomised probability integral transform (PIT).

Definition 39 Let V be a standard uniformly distributed variable that is independent of the CDF-valued random quantity F and the observation Y. Then,

$$Z_F = F(Y-) + V\{F(Y) - F(Y-)\}$$

is the probability integral transform of F. Here, $F(y-) = \lim_{x \to y^-} F(x)$. If F is continuous, then

$$Z_F = F(Y).$$

The PIT is the value attained by the forecast at the observation, up to adjustments for discontinuity.

PIT: Example

Example 40 Let X be a real-valued random variable with fixed CDF F, and let V be uniformly distributed on the unit interval and independent of X (or everything else that is random). Consider the randomised PIT of X defined as

$$U = F(X-) + V\{F(X) - F(X-)\}.$$

Assuming continuity of F, U is uniformly distributed on the unit interval $[0, 1]$ and $X = F^{-1}(U)$ almost surely.

Calibration and dispersion of forecasts

Let F be a CDF-valued random quantity with PIT Z_F.

□ The forecast F is marginally calibrated if $\mathrm{E}_{\mathbb{Q}}\{F(y)\} = \mathbb{Q}(Y \leq y)$ for all $y \in \mathbb{R}$.

□ The forecast F is probabilistically calibrated if Z_F is uniformly distributed on the unit interval, i.e.,

$$\mathbb{Q}(Z_F) = \mathbb{Q}(F(Y-) + V\{F(Y) - F(Y-)\} \leq \tau) = \tau, \quad \text{for all } \tau \in [0, 1]$$

In terms of quantiles, it is equivalent to

$$\mathbb{Q}\{Y \leq F^{-1}(\tau)\} = \tau.$$

This is intuitive: the forecast's predicted quantile at level 95% should be the value below which the target Y lies 95% of the time.

□ The forecast F is overdispersed if $\mathrm{var}(Z_F) < \mathrm{var}(U) = \frac{1}{12}$, neutrally dispersed if $\mathrm{var}(Z_F) = \frac{1}{12}$, and underdispersed if $\mathrm{var}(Z_F) > \frac{1}{12}$.

Calibration of forecasts

□ Marginal and probabilistic calibration are not related (one does not imply the other).

□ Probabilistic calibration is a statement about the joint distribution of the forecast F and target Y.

□ Marginal calibration is not; it is only a statement, as its name suggests, about the marginal distributions of F and Y.

Theorem 41 (Ideal forecast) A forecast that is ideal relative to the information set $\mathcal{A}_0$ is both marginally and probabilistically calibrated.

Example 42

□ The perfect forecaster is both probabilistically and marginally calibrated;

□ The climatological forecaster is both probabilistically and marginally calibrated;

□ The sign-reversed forecaster is marginally but not probabilistically calibrated;

□ The unfocused forecaster is probabilistically but not marginally calibrated.

Note to Theorem 6

Consider the prediction space $(\Omega, \mathcal{A}_0, \mathbb{Q})$ and suppose that F is ideal with respect to $\mathcal{A}_0$. Then, $F(y) = \mathbb{Q}(Y \leq y \mid \mathcal{A}_0)$ almost surely $\forall y \in \mathbb{R}$ and

$$\begin{aligned} \mathrm{E}_{\mathbb{Q}}\{F(y)\} &= \mathrm{E}_{\mathbb{Q}}\{\mathbb{Q}(Y \leq y \mid \mathcal{A}_0)\} \\ &= \mathrm{E}_{\mathbb{Q}}\{\mathrm{E}_{\mathbb{Q}}(\mathbf{1}_{\{Y \leq y\}} \mid \mathcal{A}_0)\} \\ &= \mathrm{E}_{\mathbb{Q}}(\mathbf{1}_{\{Y \leq y\}}) = \mathbb{Q}(Y \leq y), \end{aligned}$$

which proves marginal calibration of F.
To prove probabilistic calibration, we first denote by $\mathbb{Q}_0$ the marginal law of Y under $\mathbb{Q}$. For simplicity of exposition, we assume the continuity of F. Then, as the forecast F is ideal, $Z_F = F(Y) = \mathbb{Q}_0(Y \mid \mathcal{A}_0)$ and

$$\mathbb{Q}(Z_F \leq z) = \mathrm{E}_{\mathbb{Q}}\mathrm{E}_{\mathbb{Q}}(\mathbf{1}_{\{Z_F \leq z\}} \mid \mathcal{A}_0) = \mathrm{E}_{\mathbb{Q}}\mathrm{E}_{\mathbb{Q}}(\mathbf{1}_{\{\mathbb{Q}_0(Y) \leq z\}} \mid \mathcal{A}_0) = z,$$

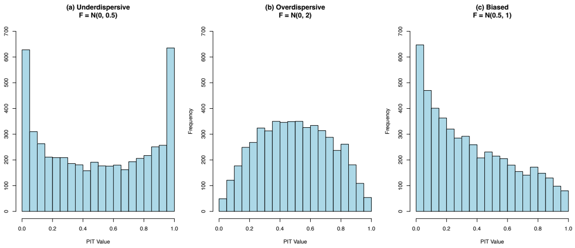
where the last equality follows from the uniformity of the traditional probability integral transform with fixed (not random) distribution, namely the marginal distribution $\mathbb{Q}_0$ of Y.

Calibration check

To check probabilistic calibration, we need to verify the uniformity of the PIT.
In practice, we observe a sample

$$\{(F_j, y_j) : j = 1, \ldots, J\}$$

from the joint distribution of the probabilistic forecasts and the observation.



The observation is $Y \sim \mathcal{N}(0, 1)$.

Calibration check

□ If F places too little mass in the tails, then the PIT will be U-shaped (prediction intervals too narrow).

□ If F places too much mass in the tails, then the PIT will be upside-down U-shaped.

□ Skewed histograms are seen when the predictive distributions are biased in their location ($\mathrm{E}_{\mathbb{Q}}\{F(Y)\} > 0.5$ or $\mathrm{E}_{\mathbb{Q}}\{F(Y)\} < 0.5$).

Formal tests of the hypothesis that a forecasting method is probabilistically calibrated can be used, provided they account for complex dependence structures, e.g., time series forecasts.

Is calibration enough?

Following Gneiting et al. (2007), the goal in probabilistic forecasting is to

> maximise the sharpness of the predictive distributions subject to calibration.

□ calibration is related to the joint distribution of the forecast and the observation (the observation is supposed to be a random draw from the predictive distribution of the forecast).

□ sharpness is a property of the forecast only and refers to the concentration of its predictive distribution. Thus, the more concentrated (the narrower are the prediction intervals), the better, subject to calibration.

If we have a forecast and an observation, can we measure the skill of the forecast using a score function? What are good properties of such function?

6.3 Scoring rules slide 240

Scoring rules

□ Scoring rules allow us to address calibration and sharpness simultaneously by providing summary measures for the evaluation of probabilistic forecasts.

□ A scoring rule assigns a numerical (real or infinite) score $S(F, y)$ to each pair (F, y), where $F \in \mathcal{F}$ is a probabilistic forecast and $y \in \mathbb{R}$ is the realised value. By convention, we always take the score to be negatively oriented.

□ In general, we use the notation $S(F, G)$ to denote the expectation of the score over draws $Y \sim G$

$$S(F, G) = \mathrm{E}_G\{S(F, Y)\}.$$

Proper scoring rules

Definition 43 If $\mathcal{F}$ denotes a class of probabilistic forecasts on $\mathbb{R}$, a proper scoring rule is any function

$$\mathrm{S} : \mathcal{F} \times \mathbb{R} \to \mathbb{R} \cup \{\pm\infty\}$$

such that

$$\mathrm{S}(G, G) := \mathrm{E}_G\{S(G, Y)\} \le E_G\{S(F, Y)\} =: S(F, G)$$

for all $F, G \in \mathcal{F}$. If inequality is strict for $F \neq G$, then the scoring rule is strictly proper.

Thus, using a proper scoring rule, an optimal strategy (in expectation) is to choose the true distribution as a forecast.

Proper scoring rules

Example 44 Discuss if the following scoring rules are proper.

□ For a forecast that has density p (its predictive density),

- the logarithmic score is defined by

$$S(p, y) = -\log\{p(y)\}.$$

  The log score very sharply penalizes forecasts that place insufficiently low probability on events that materialise (for small $p(y)$, the score $-\log\{p(y)\}$ is very large).

- the quadratic score (also known as Brier score) is defined by

$$S(p, y) = -2p(y) + \|p\|_2^2,$$

  where $\|p\|_2^2 = \int p(y)^2 dy$. This is more robust than the log score in the sense that it penalises less heavily forecasts placing low probability on events that materialise.

□ For a forecast with predictive CDF F, e.g., precipitation forecasts with point mass at zero,

- the continuous ranked probability score (CRPS) is defined by

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{+\infty} \{F(x) - I(y \le x)\}^2 dx.$$

Note to Example 9

□ We want to show that $S(p,q) - S(q,q)$ is greater or equal to zero with equality if and only is $p = q$.

$$\begin{aligned}
S(p,q) - S(q,q) &= \int -\log p(y)q(y)dy - \int -\log q(y)q(y)dy \\
&= -\int \log \left\{ \tfrac{p(y)}{q(y)} \right\} q(y)dy \\
&= \int \log \left\{ \tfrac{q(y)}{p(y)} \right\} q(y)dy \\
&= -\mathbb{E}_{Y\sim q} \left\{ \log \tfrac{p(Y)}{q(Y)} \right\} \\
&\geq -\log \mathbb{E}_{Y\sim q} \left\{ \log \tfrac{p(Y)}{q(Y)} \right\} = -\log \int p(y)dy = 0,
\end{aligned}$$

where the inequality in the last line follows from Jensen's inequality and concavity of log. Hence, $S(p,q) - S(q,q) \leq 0$, with equality if and only if $p = q$. Therefore, the log scoring rule is a strictly proper scoring rule.

Note that the quantity on the second line is known as the Kullback-Leibler divergence between q and p, often denoted $KL(q,p)$ and is known to be nonnegative, and positive for $p \neq q$.

□ We show that the Brier score is a strictly proper scoring rule.

$$\begin{aligned}
S(p,q) - S(q,q) &= \|p\|_2^2 - \|q\|_2^2 - 2\int p(y)q(y)dy + 2\int q^2(y)dy \\
&= \|p\|_2^2 + \|q\|_2^2 - 2\int p(y)q(y)dy \\
&= \|p - q\|_2^2
\end{aligned}$$

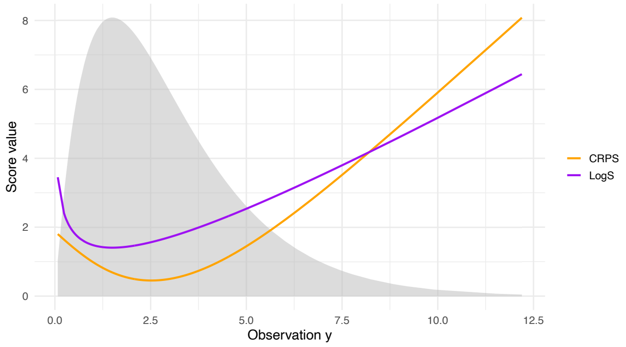
Thus, $S(p,q) - S(q,q)$ is nonnegative and positive for $p \neq q$.

□ We show that the CRPS is a strictly proper scoring rule.

$$\begin{aligned}
\text{CRPS}(F,G) - \text{CRPS}(G,G) &= \int \left\{ F(x)^2 - G(x)^2 - 2(F(x) - G(x))\text{E}_{Y\sim G}(I(Y \leq x)) \right\} dx \\
&= \int \left\{ F(x)^2 - G(x)^2 - 2(F(x) - G(x))G(x) \right\} dx \\
&= \int \{F(x) - G(x)\}^2 dx
\end{aligned}$$

This is the Cramér-von Mises distance between F and G. It is nonnegative, and positive for $F \neq G$.

Score behaviour for gamma truth



Score values for a gamma distribution with shape = 2 and scale = 1.5 as a function of the observation. A scaled version of the predictive density is shown in gray.

- □ the logarithmic score rapidly increases at the right-sided limit of zero, and the minimum score value is attained if the observation equals the predictive distribution's mode.
- □ the CRPS is more symmetric around the minimum that is attained at the median value of the predictive distribution.

What if we only care about extremes?

Consider the restricted logarithmic score

$$R^*(F, y) = -I\{y \geq t\} \log f(y).$$

However, if $g(y) > f(y)$ for all $y \geq t$, then

$$\mathrm{E}_H\{R^*(G, Y)\} < \mathrm{E}_H\{R^*(F, Y)\}$$

independent of the true sampling density H of Y.
Indeed, if the forecaster's belief is F, their best prediction under R^* is

$$g(y) = \frac{f(y)}{\int_t^\infty f(x)dx} I\{y \geq t\}$$

(Gneiting and Ranjan, JBES, 2011).

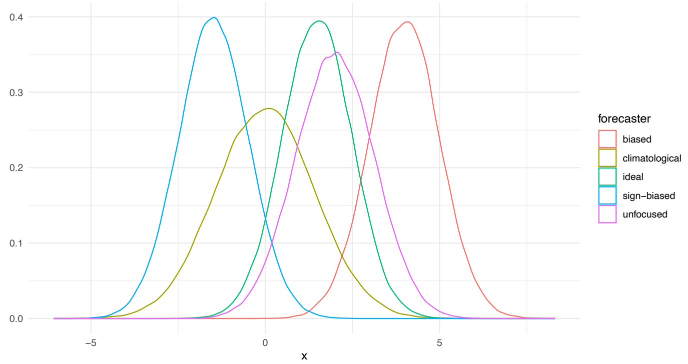
Demonstration by simulation

True data distribution: $G = \mathcal{N}(\mu, 1)$ with $\mu \sim \mathcal{N}(0, 1)$.

Consider the following forecasters

- □ The perfect forecast $\mathcal{N}(\mu, 1)$;
- □ The climatological forecast $\mathcal{N}(0, 2)$;
- □ The sign-reversed forecast $\mathcal{N}(-\mu, 1)$;
- □ The unfocused forecast $\frac{1}{2}\{\mathcal{N}(\mu, 1) + \mathcal{N}(\mu + \xi, 1)\}$, for $\xi = \pm 1$;
- □ The biased forecast $\mathcal{N}(\mu + 2.5, 1)$.

Demonstration by simulation



Forecaster	CRPS	LogS
Ideal	0.56	1.42
Sign-biased	2.43	5.84
Climatological	1.06	2.07
Unfocused	0.63	1.54
Biased	1.98	4.54

Demonstration by simulation

Comparing forecasts in the upper tail (above the 99% quantile of the true distribution)

Forecaster	CRPS*	LogS*
Ideal	2.09	4.47
Sign-biased	5.06	16.79
Climatological	3.34	5.57
Unfocused	1.60	3.27
Biased	0.28	0.98

Better option

Use threshold-weighted scoring rules such as the threshold weighted CRPS

$$\begin{aligned} R(F,y) &= \int_{-\infty}^{\infty} (F(x) - I\{y \le x\})^2 \omega(x)dx \\ &= \int_0^1 \left\{F^{-1}(\tau) - y\right\} \left(I\left\{y \le F^{-1}(\tau)\right\} - \tau\right) \omega(\tau)d\tau \end{aligned}$$

Here, we may set

$$\begin{aligned} w_1(x) &= I\{x \ge u\}; \\ w_2(x) &= 1 + I\{x \ge u\}; \\ w_3(x) &= 1 + I\{x \ge u\}u. \end{aligned}$$

Testing for equal predictive performance

Assume two forecasting methods compete. They issue probabilistic forecasts F_{1i} and F_{2i} with verifying observations y_i, at a finite set of times or locations $i = 1, \dots, n$.

□ In practice, forecasting procedures are ranked by their average score

$$\overline{S}_n^{F_1} = \frac{1}{n} \sum_{i=1}^{n} S(F_{1i}, y_i).$$

□ If the forecast cases $F_{11}, \dots, F_{1n}$ are independent (same goes for F_{2i}), a test of equal forecast performance can be based on the test statistic

$$t_n = \sqrt{n} \frac{\overline{S}_n^{F_1} - \overline{S}_n^{F_2}}{\widehat{\sigma}_n},$$

where the variance estimate of the score difference is

$$\widehat{\sigma}_n^2 = \tfrac{1}{n} \sum_{i=1}^{n} \{S(F_{1i}, y_i) - S(F_{2i}, y_i)\}^2.$$

Subject to weak regularity conditions, the statistic t_n is asymptotically standard normal under the null hypothesis of equal performance.

Testing in a sequential setup

In the case of sequential k-step ahead time series forecasts, the assumption of independence between the forecast cases is violated.

□ Diebold and Mariano (1995) generalise the variance estimate to

$$\widehat{\sigma}_n^2 = \frac{1}{n-k+1} \sum_{j=-(k-1)}^{k-1} \sum_{i=1}^{n-|j|} d_i d_{i+|j|},$$

where $d_i = S(F_{1i}, y_i) - S(F_{2i}, y_i)$.
Under regularity conditions, the resulting statistic is still asymptotically standard normal. This is known as the Diebold-Mariano test.

This is an appealing procedure to compare forecasts. However, in practice, ranking of competing forecasts usually depends on the choice of the scoring rule...

Forecast dominance

The concept of forecast dominance arises when one forecast performs at least as well as another across a range of scoring rules.

Definition 45 Let F_1 and F_2 be CDF-valued random quantities (forecasts). Then F_1 dominates F_2 with respect to a class $\mathcal{S}$ of proper scoring rules if:

$$\mathrm{E}_{\mathbb{Q}}\{S(F_1, Y)\} \leq \mathrm{E}_{\mathbb{Q}}\{S(F_2, Y)\},$$

for all scoring rules $S \in \mathcal{S}$, where the expectations are taken with respect to the joint distribution $\mathbb{Q}$ of the triple (F_1, F_2, Y).

Forecast dominance

Theorem 46 (Forecast dominance) Assume Y is the observation and let F_1 and F_2 be CDF-valued random quantities (forecasts) that rely on the sets of information $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. Furthermore, let $\mathcal{S}$ be a class of proper scoring rules.

□ If F_1 is ideal relative to $\mathcal{A}_1$, i.e., the distribution of $Y \mid \mathcal{A}_1$ is F_1, and F_2 is ideal relative to $\mathcal{A}_2 \subset \mathcal{A}_1$, then F_1 dominates F_2 with respect to $\mathcal{S}$, i.e.,

$$\mathrm{E}_{\mathbb{Q}}\{S(F_1, Y)\} \leq \mathrm{E}_{\mathbb{Q}}\{S(F_2, Y)\}, \quad \text{for all } S \in \mathcal{S}.$$

□ If F_1 is ideal relative to $\mathcal{A}_1$, and F_2 relies on the set of information $\mathcal{A}_1$, then

$$\mathrm{E}_{\mathbb{Q}}\{S(F_1, Y)\} \leq \mathrm{E}_{\mathbb{Q}}\{S(F_2, Y)\}, \quad \text{for all } S \in \mathcal{S}.$$

Note to Theorem 11

We first show the second part. If F_1 is the distribution of $Y \mid \mathcal{A}_1$ and $S \in \mathcal{S}$, then

$$\begin{aligned}
\mathrm{E}_Q\{S(F_1, Y)\} &= \mathrm{E}_Q \mathrm{E}_Q\left\{S(F_1, Y) \mid \mathcal{A}_1\right\} \\
&= \mathrm{E}_Q \mathrm{E}_{Y \sim F_1}\{S(F_1, Y)\} \\
&\leq \mathrm{E}_Q \mathrm{E}_{Y \sim F_1}\{S(F_2, Y)\}, \quad \text{since } S \text{ is a proper scoring rule} \\
&= \mathrm{E}_Q \mathrm{E}_Q\left\{S(F_2, Y) \mid \mathcal{A}_1\right\}, \quad \text{since } F_1 \text{ is ideal relative to } \mathcal{A}_1 \\
&= \mathrm{E}_Q\{S(F_2, Y)\}.
\end{aligned}$$

The statement of the first part is immediate from the second, as the fact that F_2 is based on the information set $\mathcal{A}_2$ together with $\mathcal{A}_2 \subseteq \mathcal{A}_1$ implies that F_2 is based on the information set $\mathcal{A}_1$. The fact that F_2 is idea relative to $\mathcal{A}_2$ is not relevant but emphasises the dominance of the ideal F_1 over all forecasts using the information set $\mathcal{A}_2$, even the one that makes the best use of that information.

Forecast dominance

- □ Graphical tools such as the Murphy diagrams (Ehm et al., 2016) can be used to check forecast dominance.
- □ Statistical tests for forecast dominance exist as well; see Ehm and Krüger (2018).
- □ More recently, Krüger and Ziegel (2021) derived a characterisation of dominance among forecasts of the mean.

There is more to it ...

- □ Tail calibration of forecasts (Allen et al., 2025)

$$F \text{ (not random) is } \mathcal{A}\text{-tail calibrated for } Y \Leftrightarrow \frac{P(Y > t \mid \mathcal{A})}{1 - F(t)} \to 1 \quad \text{a.s. as } t \to x_Y.$$

- □ Probabilistic forecasts of multivariate outcome (e,g., Gneiting et al., 2008).