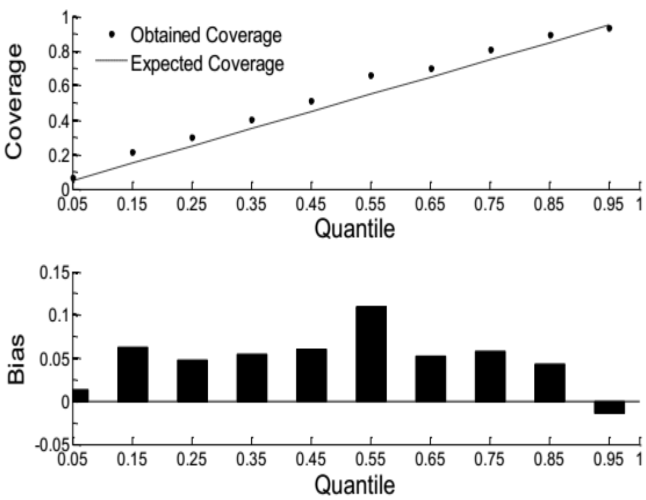
- The quality of quantile forecasts can be assessed using reliability diagrams (*coverage* and *bias*). It can be useful to gauge the accuracy of forecasts at different key quantiles of interest (extremes – 5%, 95%; median– 50% etc.)

- Unconditional coverage - measures the degree to which the actual observations lie within the different quantile ranges of the forecast distribution. Note that each quantile range of density forecast is expected to encompass a specific proportion of the actual observations, as quantified by coverage

- Bias – difference between the expected and obtained coverage. Bias would be zero across all quantiles for a perfect forecast

- For perfect reliability, the forecast probability and the frequency of occurrence should be equal. For example, if a forecast is made that a crash would occur with a probability of 5%, then one would expect the crash to occur on 5% of the actual events. Basically, the obtained and expected coverage would lie on a diagonal, and so the bias would be zero for a perfect forecast



# Density forecast evaluation – CRPS

- The Continuous Rank Probability Score (CRPS) was proposed by Gneiting and Raftery (2007)
- The advantage of using CRPS for quantifying density forecast performance is that it takes into account both *sharpness* and *calibration*
- Sharpness – reward the model if the forecast density has a small spread and is highly peaked around the actual observation
- Calibration – reward the model if the agreement between the forecast density and actual data density is good
- The CRPS can be viewed as the distributional analogue of the mean absolute error. Note that for point forecasts, the CRPS is the same as MAE

# Density forecast evaluation – Brier score

- Brier (1950) introduced the Brier score to measure the accuracy of a set of probability assessments
- It quantifies the average deviation between predicted probabilities for a set of events and their outcomes, so a lower score represents higher accuracy
- Suppose a forecaster gives a probability $p_i$ of a particular event occurring
- Let $x_i = 1$ if the event occurs and $x_i = 0$ otherwise
- The Brier score is given by

$$BrierScore_i = (p_i - x_i)^2$$

- A Brier score of 0 indicates a perfect forecast and one indicates poor performance
- Ex. If we forecast the probability that India will win the cricket world cup,

If the forecast is $p$=1, and India wins, then the Brier score is 0 (perfect score)

If the forecast is $p$=0.8, and India wins, then the Brier score is $(0.8 - 1)^2$=0.04

If the forecast is $p$=0, and India wins, then the Brier score is 1 (bad score, but good outcome)

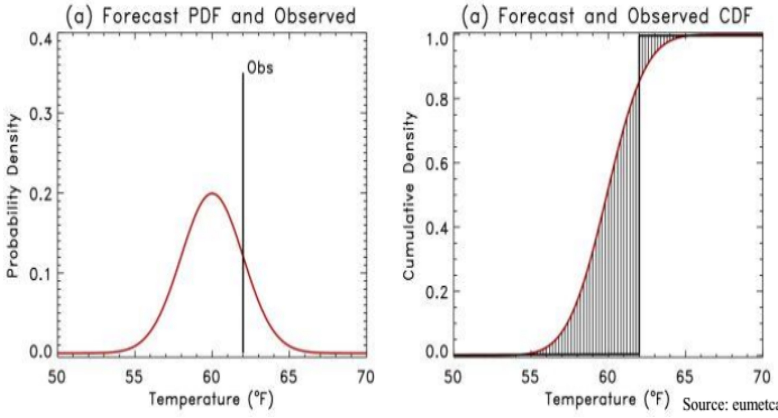
# Density forecast evaluation – CRPS

- The CRPS is given by

$$CRPS(F, y) = \int_{z=-\infty}^{z=+\infty} \{F(z) - \mathbf{1}(z \geq y)\}^2 dz$$

where $y$ is the actual observation, $F$ is the cumulative distribution function of the forecast distribution, and $\mathbf{1}$ is an indicator variable that equals one for $z \geq y$

- CRPS is the area between the CDF of the forecast and CDF of the observation (shaded region in plot below)



Source: eumetcal.com

# Classification

- Consider an early warning system for predicting financial crashes. A perfect predictor will correctly identify all crashes (100% sensitivity) and also accurately classify all normal activity periods as non-crash events (100% specificity)

- This two class prediction is also known as binary classification. There are two classes, labelled either as the positive (*p*, crash) or negative (*n*, normal/no-crash) class

- There are four possible outcomes for this two-class prediction problem. If the prediction is *p* and the actual value is also *p*, then it is called a *true positive* (TP); however if the actual value is *n* then it is a *false positive* (FP). In contrast, if the prediction is *n* and the actual value is *n*, it is called a *true negative* (TN); however if the actual value is *p* and the prediction is *n,* it is a *false negative* (FN)

# Classification

- The four possible outcomes for the two-class prediction problem give a 2 × 2 *contingency table* or *confusion matrix*

| | True Condition | |
|---|---|---|
| Predicted Condition | **Crash** | **No Crash** |
| **Predicted Crash** | True Positive | False Positive |
| **Predicted No-Crash** | False Negative | True Negative |

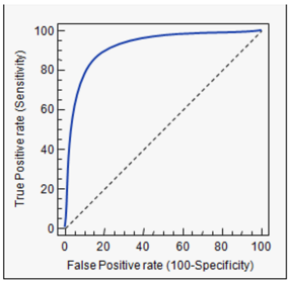
- $Sensitivity = \frac{true\ positives}{(true\ positives + false\ negatives)}$
- $Specificity = \frac{true\ negatives}{(true\ negatives + false\ positives)}$
- *For example:* True condition – {1, 1, 0, 0, 1}

  Predicted condition – {1, 0, 1, 0, 1}

$TP = 2\ (index\ 1,5);\ FN = 1\ (index\ 2);\ FP = 1(index\ 3); TN = 1\ (index\ 4)$

Sensitivity = 2/(2+1)=66.67%; Specificity = 1/(1+1)=0.5

# ROC Analysis




- A Receiver Operating Characteristic (ROC), or simply ROC curve, is a graphical plot of the True Positive Rate (TPR) vs False Positive Rate (FPR)
- Note that FPR(%)=100-Specificity and TPR=Sensitivity
- The ROC space displays the relative trade-offs between true positive (benefits) and false positive (costs)
- The ROC curve can be summarised using the area under the curve (AUC), which is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one

- A good classifier will have an AUC value greater than ½

# Model evaluation – Summary

- **Point forecasts** – $R^2$, adjusted $R^2$, RMSE (sensitive to outliers), MAE, MAD (robust estimator), MAPE ($y$ to be non-zero)
- **Quantile forecasts** – Coverage (degree to which actual observations lie within the different quantile ranges of the forecast distribution) and Bias (difference between actual and obtained coverage)
- **Density forecasts** – Brier score and CRPS (sharpness and calibration)
- **Classification** – Sensitivity and Specificity, ROC curve

# Model evaluation – points to consider

- Always use **multiple performance scores** when comparing models
- Consider **estimating model parameters using the CRPS** – encapsulate in-sample forecast sharpness and calibration during the estimation process
- It is common to first generate density forecasts for the post-sample observations, and then issue either the **mean or median of the forecast distribution** as a point forecast
- Compare models with **naïve benchmarks**