COVID-19 Forecast Similarity Analysis

Johannes Bracher, Aaron Gerding, Evan Ray, Nick Reich, Nutcha Wattanachit

07/21/2021

Cramér distance

Consider two predictive distributions F and G. Their *Cramér distance* or *integrated quadratic distance* is defined as

$$\begin{aligned} \mathrm{CD}(F,G) &= \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx &(1)\\ &= \mathbb{E}_{F,G}|X-Y| - 0.5\left[\mathbb{E}_F|X-X'| + \mathbb{E}_G|Y-Y'|\right], &(2)\end{aligned}$$

where $F(x)$ and $G(x)$ denote the cumulative distribution functions, X, X' are independent random variables following F, and Y, Y' are independent random variables following G. This formulation in (2) illustrates that the Cramér distance depends on the shift between F and G (first term) and the variability of both F and G (of which the two last expectations in above equation are a measure).

The Cramér distance is the divergence associated with the continuous ranked probability score (Thorarinsdottir 2013, Gneiting and Raftery 2007), which is defined by

$$\begin{aligned} \mathrm{CRPS}(F,y) &= \int_{-\infty}^{\infty} (F(x) - \mathbf{1}(x \geq y))^2 dx &(3)\\ &= 2\int_0^1 \{(\mathbf{1}(y \leq q_k^F) - \tau_k\}(q_k^F - y) d\tau_k &(4)\end{aligned}$$

Cramér Distance Approximation for Equally-Spaced Intervals

Assume that for each of the distributions F and G we only know K quantiles at equally spaced levels $1/(K+1), 2/(K+1), \dots, K/(K+1)$. Denote these quantiles by $q_1^F, \dots, q_K^F$ and $q_1^G, \dots, q_K^G$, respectively. Let $\tau_1, ..., \tau_K$ be the probability levels corresponding to those quantiles. It is well known that the CRPS can be approximated by an average of linear quantile scores (Laio and Tamea 2007, Gneiting and Raftery 2007):

$$\mathrm{CRPS}(F,y) \approx \frac{1}{K} \times \sum_{k=1}^{K} 2\{\mathbf{1}(y \leq q_k^F) - \tau_k\} \times (q_k^F - y). \quad (5)$$

This approximation is equivalent to the weighted interval score (WIS) which is in use for evaluation of quantile forecasts at the Forecast Hub, see Section 2.2 of Bracher et al (2021). This approximation can be generalized to the Cramér distance as

$$\mathrm{CD}(F,G) \approx \frac{1}{K(K+1)} \sum_{i=1}^{K}\sum_{j=1}^{K} \mathbf{1}\{(i-j) \times (q_i^F - q_j^G) \leq 0\} \times \left|q_i^F - q_j^G\right|, \quad (6)$$

This can be seen as a sum of penalties for *incompatibility* of predictive quantiles. Whenever the predictive quantiles q_i^F and q_j^G are incompatible in the sense that they imply F and G are different distributions (e.g. because $q_F^i > q_G^j$ despite $i < j$ or $q_F^i \neq q_G^j$ despite $i = j$), a penalty $\left|q_i^F - q_j^G\right|$ is added to the sum. This corresponds to the shift which would be necessary to make q_F^i and q_G^j compatible.

Cramér Distance Approximation for Any Intervals

Using the same notations as before, now assume we don't necessarily have equally spaced $\tau_1, ..., \tau_K$ from two distributions F and G. The CRPS can be approximated as follows

$$\text{CRPS}(F, y) \approx 2 \sum_{k=1}^{K} \{\mathbf{1}(y \leq q_k^F) - \tau_k\} \times (q_k^F - y) \times (\tau_k - \tau_{k-1}). \tag{7}$$

with $\tau_0 = 0$. We can see that with the increment $\tau_k - \tau_{k-1} = 1/K$, this equation generalizes to 5.

This approximation of CRPS can be generalized to the Cramér distance as

$$\text{CD}(F, G) \approx \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{1}\{(i - j) \times (q_i^F - q_j^G) \leq 0\} \times \left|q_i^F - q_j^G\right| \times (\tau_i^F - \tau_j^G), \tag{8}$$

The same interpretation of this quantity as a sum of penalties as in the previous section also applies. With $\frac{1}{K(K+1)}$, we can see that this equation generalizes to 6.

Left-sided Riemann sum approximation

$$\begin{align}
\text{CD}(F, G) &\approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \tag{9} \\
&\approx \sum_{j=1}^{2K-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) \tag{10} \\
& \tag{11}
\end{align}$$

Since $q_j \in \{q_1, ..., q_{2K}\}$ belongs to either $q_1^F, ..., q_K^F$ or $q_1^G, ..., q_K^G$, we can rewrite the above approximation using $\tau_1, ..., \tau_K$ as follows

$$\begin{align}
\text{CD}(F, G) &\approx \sum_{j=1}^{2K-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) \tag{12} \\
&= \sum_{j=1}^{2K-1} \{\tau_j^F - \tau_j^G\}^2 (q_{j+1} - q_j) \tag{13}
\end{align}$$

where $\tau_j^F \in \tau_F$ and $\tau_j^G \in \tau_G$. τ_F and τ_G are vectors of length $2K - 1$ with elements

$$\tau_j^F = \begin{cases} I(q_1 = q_1^F) \times \tau_{q_1}^F & \text{for } j = 1 \\ I(q_j \in \{q_1^F, ..., q_K^F\}) \times \tau_{q_j}^F + I(q_j \in \{q_1^G, ..., q_K^G\}) \times \tau_{j-1}^F & \text{for } j > 1 \end{cases}$$

where $\tau_{q_j}^F$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from F, and τ_{j-1}^F is the $(j-1)^{th}$ probability level in τ_F.

$$\tau_j^G = \begin{cases} I(q_1 = q_1^G) \times \tau_{q_1}^G & \text{for } j = 1 \\ I(q_j \in \{q_1^G, ..., q_K^G\}) \times \tau_{q_j}^G + I(q_j \in \{q_1^F, ..., q_K^F\}) \times \tau_{j-1}^G & \text{for } j > 1 \end{cases}$$

where $\tau_{q_j}^G$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from G, and τ_{j-1}^G is the $(j-1)^{th}$ probability level in τ_G.

Trapezoidal rule

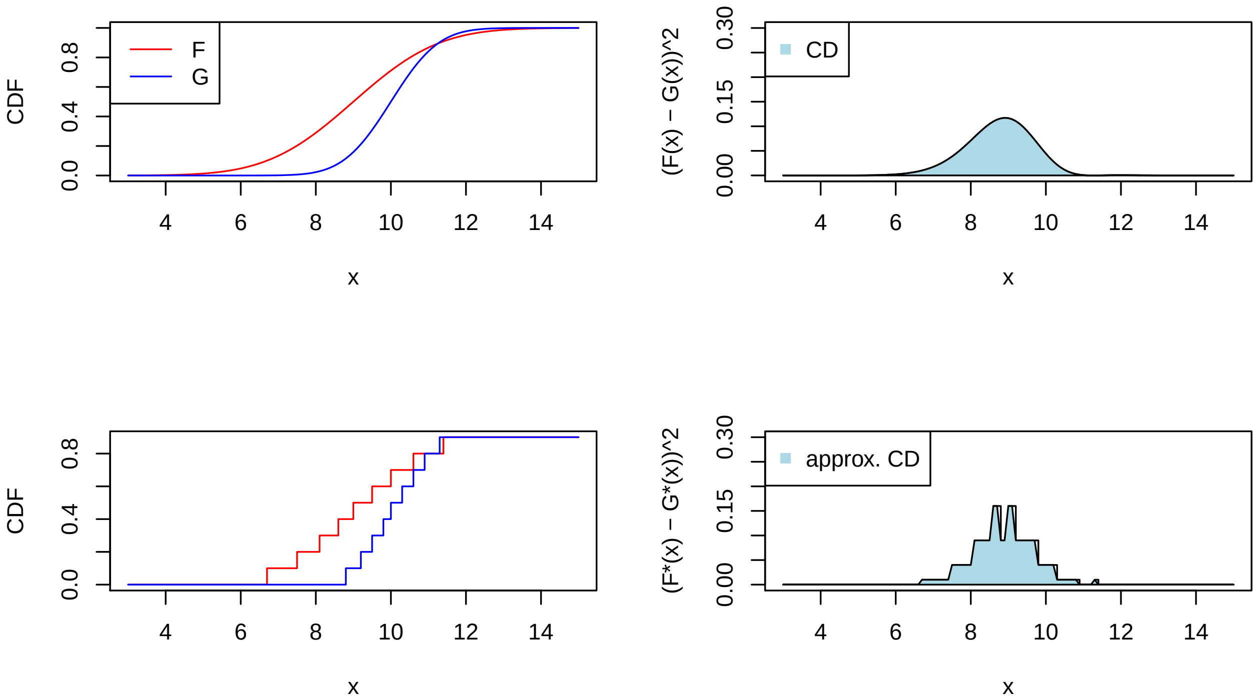
$$\begin{aligned} \mathrm{CD}(F,G) &\approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 && (14) \\ &\approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) && (15) \\ & && (16) \end{aligned}$$

Similarly, we can rewrite the above approximation using $\tau_1, ..., \tau_K$ as defined in the left-sided Riemann sum approximation as follows

$$\begin{aligned} \mathrm{CD}(F,G) &\approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) && (17) \\ &= \sum_{j=1}^{2K-1} \frac{\{\tau_j^F - \tau_j^G\}^2 + \{\tau_{j+1}^F - \tau_{j+1}^G\}^2}{2} (q_{j+1} - q_j). && (18) \end{aligned}$$

Examples

Equally-spaced intervals



In this example, six different approximations are applied to the distributions F $N(9, 1.8)$ and G $N(10, 1)$ in the figures above.

- Using direct numerical integration based on a fine grid of values for x:

```
FALSE [1] 0.2532376
```

- Using sampling and the alternative expression (**??**) of the CD from above:

```
FALSE [1] 0.2457156
```

- Using the first quantile-based approximation (**??**) and various values of K:

```
FALSE [1] 0.3550788 0.3078906 0.2764153 0.2652018 0.2593619 0.2557450 0.2545077
FALSE [8] 0.2538792
```

- Using the second quantile-based approximation (**??**) and various values of K:

```
FALSE [1] 0.2926809 0.2723571 0.2608768 0.2572045 0.2552998 0.2541028 0.2536835
FALSE [8] 0.2534662
```

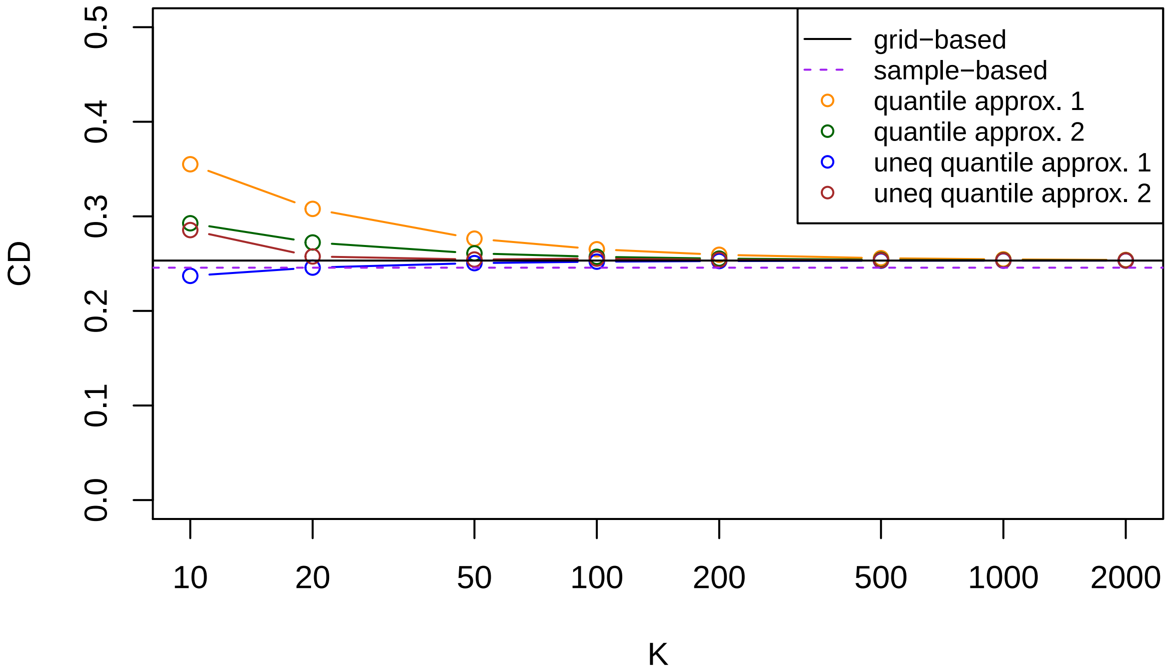
- Using the left-sided Riemann sum approximation and various values of K:

```
FALSE [1] 0.2370715 0.2458022 0.2505461 0.2520862 0.2527531 0.2530874 0.2531764
FALSE [8] 0.2532128
```

- Using the trapezoidal Riemann sum approximation and various values of K:

```
FALSE [1] 0.2854597 0.2575762 0.2543386 0.2552775 0.2540318 0.2535609 0.2534094
FALSE [8] 0.2533309
```

The below plot shows the results from the different computations.



In the case that G is a point mass at $y = 10$, approximation (**??**) indeed coincides with (5).

```
FALSE [1] "Quantile approx. 1: 0.688567227886639"
```

```
FALSE [1] "Quantile approx. 2: 0.608983067073759"
```

```
FALSE [1] "Uneq quantile approx. 1: 1.03814992169128"
```

```
FALSE [1] "Uneq quantile approx. 2: 1.24791020451193"
```

```
FALSE [1] "Quantile score WIS: 0.688567227886639"
```

The approximation (**??**) is closer to the grid-based direct evaluation of the integral. Since the unequally-spaced approximations were not formulated from (equally-spaced) WIS, it may be expected.

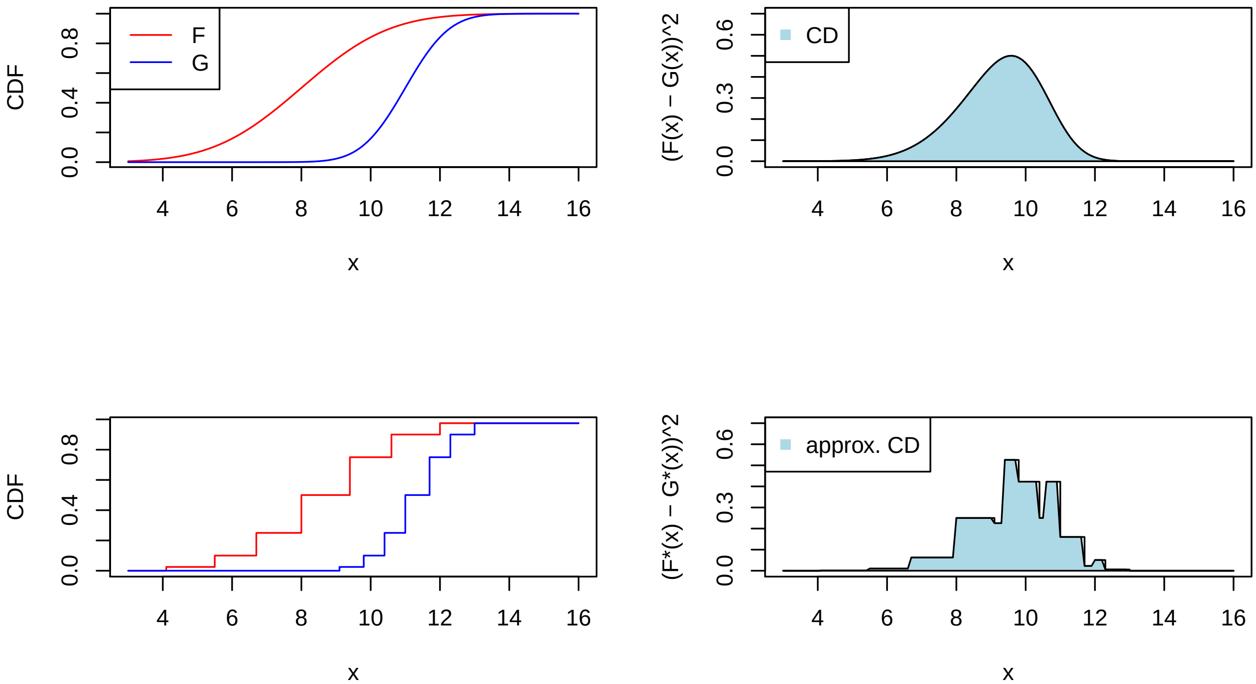
```
FALSE [1] "Grid-based approx.: 0.61599852942592"
```

Unequally-spaceed intervals

We apply the same six approximations as in the previous example to the two distributions $F \sim N(8, 2)$ and $G \sim N(11, 1)$ whose quantiles correspond to unequally-spaced probability levels.

7 quantiles with unequally-spaced intervals

The probability levels corresponding to the given set of quantiles in this example is $0.025, 0.1, 0.25, 0.5, 0.75, 0.9, 0.975$, which is the same probability levels provided by the COVID-hub case forecasts.



- Using direct numerical integration based on a fine grid of values for x.

```
FALSE [1] 1.493653
```

- Using sampling and the alternative expression (**??**) of the CD from above:

```
FALSE [1] 1.483948
```

- Using the first quantile-based approximation:

```
FALSE [1] 1.919252
```

- Using the second quantile-based approximation:

```
FALSE [1] 1.764859
```

- Using the left-sided Riemann sum-based approximation:

```
FALSE [1] 1.35122
```

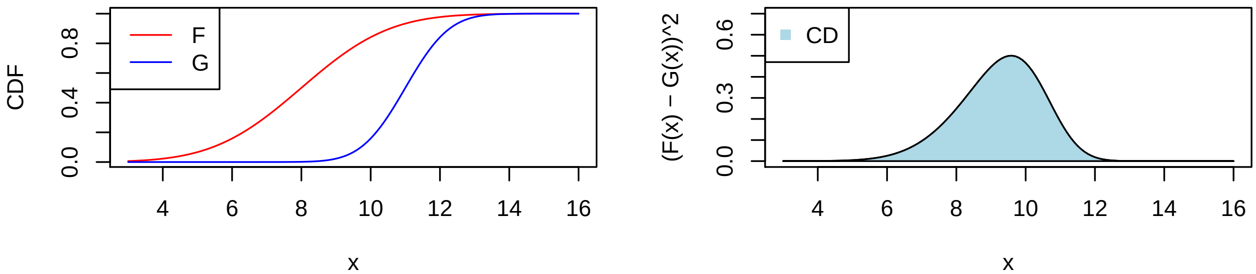
- Using the trapezoidal Riemann sum-based approximation:

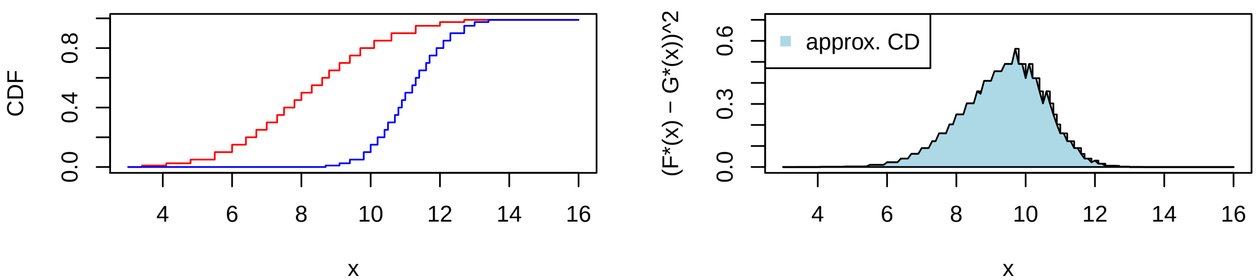
```
FALSE [1] 1.468801
```

Out of all four quantile-based approximation, the trapezoidal Riemann sum-based approximation is closest to the grid-based integral evaluation.

23 quantiles with 2 unequally-spaced probability levels at the tails

Using the same F and G, the probability levels corresponding to the given set of quantiles in this example is the same probability levels provided by the COVID-hub death forecasts. They are almost equally-spaced, except at the tails.





- Using the first quantile-based approximation:

```
FALSE [1] 1.640408
```

- Using the second quantile-based approximation:

```
FALSE [1] 1.581296
```

- Using the left-sided Riemann sum-based approximation:

```
FALSE [1] 1.452266
```

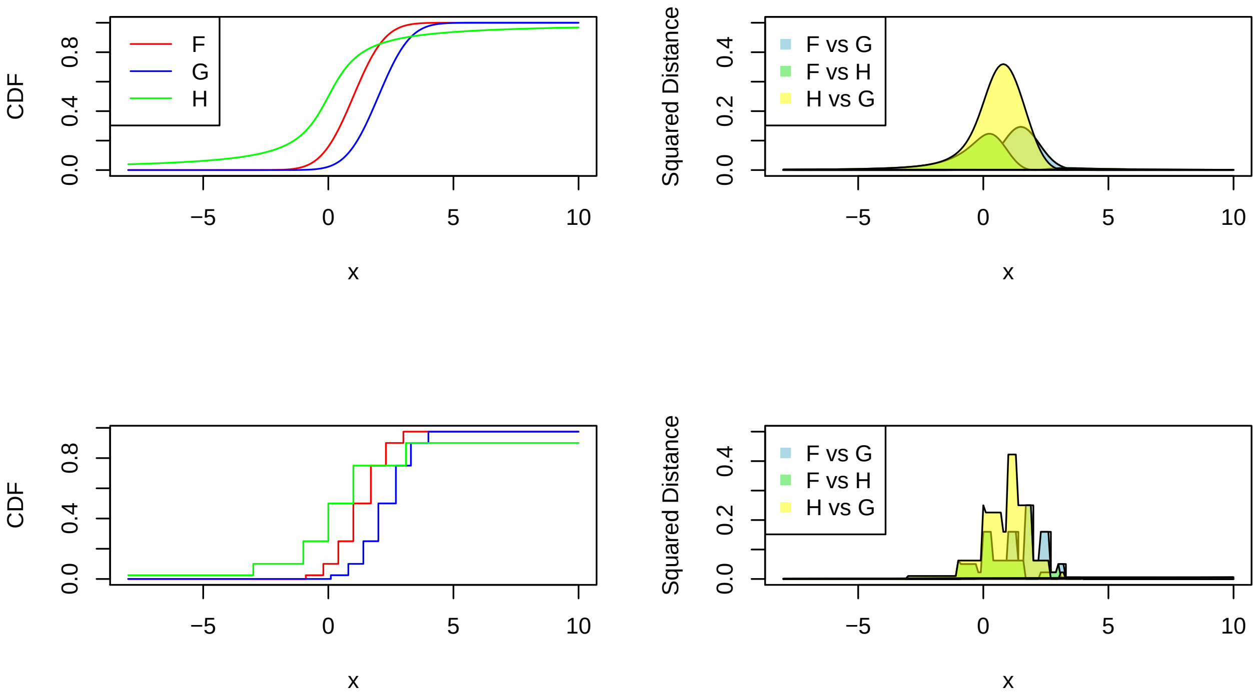
- Using the trapezoidal Riemann sum-based approximation:

```
FALSE [1] 1.470718
```

Again, the trapezoidal Riemann sum-based approximation is closest to the grid-based integral evaluation of 1.493653.

Examples of Disagreement Between Equally- and Unequally-spaced Interval Methods

Heavy tails Suppose we have three cumulative distributions, $F \sim N(1,1)$, $G \sim N(2,1)$ and $H \sim T_1$, represented by 7 unequally-spaced quantiles. The probability levels corresponding to the given set of quantiles in this example is $0.025, 0.1, 0.25, 0.5, 0.75, 0.9, 0.975$.



- Using direct numerical integration based on a fine grid of values for x.

```
FALSE [1] "CD of F vs G: 0.270903289652979"
FALSE [1] "CD of F vs H: 0.303008857878541"
FALSE [1] "CD of H vs G: 0.830227986212452"
```

- Using the first quantile-based approximation:

```
FALSE [1] "Approx. CD of F vs G: 0.430834455349389"
FALSE [1] "Approx. CD of F vs H: 0.412836097817344"
FALSE [1] " Approx. CD of H vs G: 1.0891147790307"
```

- Using the second quantile-based approximation:

```
FALSE [1] "Approx. CD of F vs G: 0.349525091827873"
FALSE [1] "Approx. CD of F vs H: 0.318185573750552"
FALSE [1] " Approx. CD of H vs G: 0.958988318892232"
```

- Using the left-sided Riemann sum-based approximation:

```
FALSE [1] "Approx. CD of F vs G: 0.267605148430715"
FALSE [1] "Approx. CD of F vs H: 0.243610829902767"
FALSE [1] " Approx. CD of H vs G: 0.734225431651865"
```

- Using the trapezoidal Riemann sum-based approximation:

```
FALSE [1] "Approx. CD of F vs G: 0.302511061162121"
FALSE [1] "Approx. CD of F vs H: 0.266926890705267"
FALSE [1] " Approx. CD of H vs G: 0.752143988834321"
```

Forecast inclusion criteria

- Models: All models with complete submissions for the following criteria
- Targets: 1-4 wk ahead inc death and inc case
- Target end dates: Oct 19th, 2020 - May 24th,2021
- Probability levels: All
- Locations:
  - 5 states with highest cumulative deaths by February 27th, 2021: CA, FL, NY, PA, TX
  - 5 states with highest cumulative cases by February 27th, 2021: CA, FL, IL, NY, TX
  - 5 states with lowest cumulative deaths by February 27th, 2021: AK, HI, ME, VT, WY
  - 5 states with lowest cumulative cases by February 27th, 2021: DC, HI, ME, VT, WY