Forecast Similarity Using Cramer Distance Approximation

Johannes Bracher, Evan Ray, Nick Reich, Nutcha Wattanachit

06/24/2021

Cramer Distance

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

$$\text{CD}(F, G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx$$

where $F(x)$ and $G(x)$ denote the cumulative distribution functions. The Cramer distance is the divergence associated with the continuous ranked probability score (Thorarinsdottir 2013, Gneiting and Raftery 2007).

$$\text{CD}(F, G) = \mathbb{E}_{F,G}|x - y| - 0.5 \left[\mathbb{E}_F|x - x'| + \mathbb{E}_G|y - y'|\right], \tag{1}$$

where x, x' are independent random variables following F and y, y' are independent random variables following G. This formulation illustrates that the Cramer 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).

Cramer Distance Approximation for Equally-Spaced Intervals

Now 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. 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):

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

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 Cramer distance as

$$\text{CD}(F, G) \approx \frac{1}{K(K+1)} \times \sum_{i=1}^{2K-1} b_i(b_i + 1)(q_{i+1} - q_i), \tag{3}$$

where we use the following notation:

- $\mathbf{q}$ is a vector of length $2K$. It is obtained by pooling the $q_k^F, q_k^G, k = 1, \dots, K$ and ordering them in increasing order (ties can be ordered in an arbitrary manner).
- $\mathbf{a}$ is a vector of length $2K$ containing the value 1 wherever $\mathbf{q}$ contains a quantile of F and -1 wherever it contains a value of G.
- $\mathbf{b}$ is a vector of length $2K$ containing the absolute cumulative sums of $\mathbf{a}$, i.e. $b_i = \left|\sum_{j=1}^{i} a_j\right|$.

For small K it seems that the slightly different approximation

$$\mathrm{CD}(F,G) \approx \frac{1}{(K+1)^2} \times \sum_{i=1}^{2K-1} b_i^2 (q_{i+1} - q_i), \tag{4}$$

actually works better. This just corresponds to the integrated squared difference between two step functions F^* and G^* with $F^*(x) = 0$ for $x < q_1^F$, $F^*(x) = k/(K+1)$ for $q_k^F \leq x < q_k^F$, $F^*(x) = K/(K+1)$ for $x \geq x_K^F$ and G^* defined accordingly. We illustrate this in the figure below, with light blue areas representing the CD and approximated CD.

Cramer Distance Approximation for Unequally-Spaced Intervals

Suppose we have quantiles $q_1^F, ..., q_K^F$ and $q_1^G, ..., q_K^G$ at K probability levels $\tau_1, ..., \tau_K$ from two distributions F and G. Define the combined vector of quantiles $q_1, ..., q_{2K}$ by combining the vectors $q_1^F, ..., q_K^F$ and $q_1^G, ..., q_K^G$ and sorting them in an ascending order. Essentially, we can approximate the Cramer distance by eliminating the tails of the integral to the left of q_1 and the right of q_{2K}, and approximating the center via a Riemann sum:

$$\begin{aligned}
\mathrm{CD}(F,G) &= \int_{-\infty}^{\infty} F(x) - G(x)^2 dx && (5)\\
&\approx \int_{q_1}^{q_{2K}} F(x) - G(x)^2 dx && (6)\\
&= \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 dx && (7)
\end{aligned}$$

There are a variety of options that can be used for each term in this sum, for instance:

Left-sided Riemann sum approximation

$$\begin{aligned}
\mathrm{CD}(F,G) &\approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 && (8)\\
&\approx \sum_{j=1}^{2K-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) && (9)\\
& && (10)
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{CD}(F,G) &\approx \sum_{j=1}^{2K-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) && (11)\\
&= \sum_{j=1}^{2K-1} \{\tau_j^F - \tau_j^G\}^2 (q_{j+1} - q_j) && (12)
\end{aligned}$$

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

$$\begin{aligned} \text{CD}(F, G) &\approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 && (13) \\ &\approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2}(q_{j+1} - q_j) && (14) \\ & && (15) \end{aligned}$$

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

$$\begin{aligned} \text{CD}(F, G) &\approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2}(q_{j+1} - q_j) && (16) \\ &= \sum_{j=1}^{2K-1} \frac{\{\tau_j^F - \tau_j^G\}^2 + \{\tau_{j+1}^F - \tau_{j+1}^G\}^2}{2}(q_{j+1} - q_j). && (17) \end{aligned}$$

Cramer Distance Approximation for Unequally-Spaced Intervals and Different Probability Levels

We (probably) can further modify the formula of the Cramer distance approximation for unequally-spaced intervals to accommodate different probability levels from F and G. Suppose we have quantiles $q_1^F, ..., q_N^F$ at K probability levels $\tau_1^F, ..., \tau_N^F$ from the distribution F, and $q_1^G, ..., q_M^G$ at M probability levels $\tau_1^G, ..., \tau_M^G$ from the distribution G. Define the combined vector of quantiles $q_1, ..., q_{N+M}$ by combining the vectors $q_1^F, ..., q_N^F$ and $q_1^G, ..., q_M^G$ and again sorting them in an ascending order. Using the same definitions as previously defined, we can approximate the Cramer distance via a Riemann sum as follows:

Left-sided Riemann sum approximation

$$\mathrm{CD}(F,G) \approx \sum_{j=1}^{N+M-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \tag{18}$$

$$\approx \sum_{j=1}^{N+M-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j), \tag{19}$$

$$\tag{20}$$

which we can rewrite using $\tau_1^F, ..., \tau_N^F$ and $\tau_1^G, ..., \tau_M^G$ as follows

$$\mathrm{CD}(F,G) \approx \sum_{j=1}^{N+M-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) \tag{21}$$

$$= \sum_{j=1}^{N+M-1} \{\tau_j^F - \tau_j^G\}^2 (q_{j+1} - q_j) \tag{22}$$

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

$$\tau_j^F = \begin{cases} \tau_{q_j}^F & \text{if } q_j \in \{q_1^F, ..., q_N^F\} \\ \tau_{q_{j-1}}^F & \text{if } q_j \notin \{q_1^F, ..., q_N^F\} \end{cases}$$

where $\tau_{q_j}^F$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from F.

$$\tau_j^G = \begin{cases} \tau_{q_j}^G & \text{if } q_j \in \{q_1^G, ..., q_M^G\} \\ \tau_{q_{j-1}}^G & \text{if } q_j \notin \{q_1^G, ..., q_M^G\} \end{cases}$$

where $\tau_{q_j}^G$ is the probability level corresponding to q_j given q_j in the pooled quantiles comes from G.

Trapezoidal rule

$$\mathrm{CD}(F,G) \approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \tag{23}$$

$$\approx \sum_{j=1}^{N+M-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j), \tag{24}$$

$$\tag{25}$$

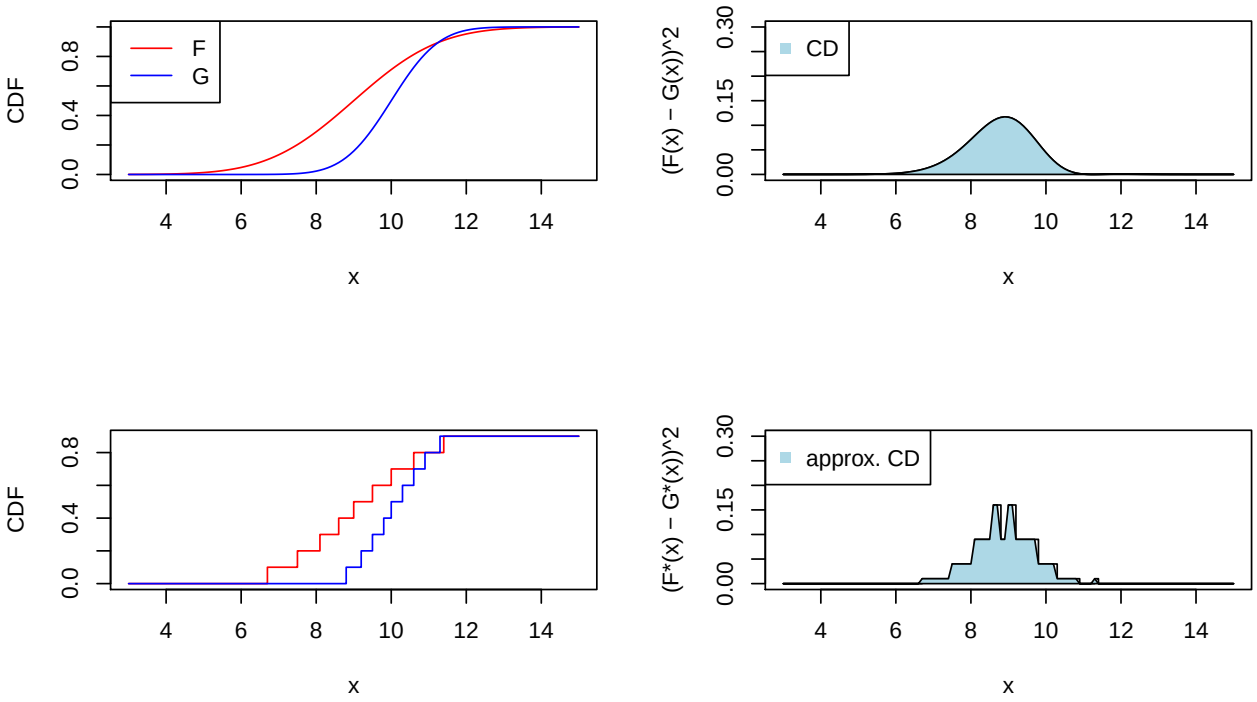
which we can rewrite as follows

$$\mathrm{CD}(F,G) \approx \sum_{j=1}^{N+M-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) \tag{26}$$

$$= \sum_{j=1}^{N+M-1} \frac{\{\tau_j^F - \tau_j^G\}^2 + \{\tau_{j+1}^F - \tau_{j+1}^G\}^2}{2} (q_{j+1} - q_j). \tag{27}$$

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:

```
## [1] 0.2532376
```

- Using sampling and the alternative expression (1) of the CD from above:

```
## [1] 0.2457156
```

- Using the first quantile-based approximation (4) and various values of K:

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

- Using the second quantile-based approximation (4) and various values of K:

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

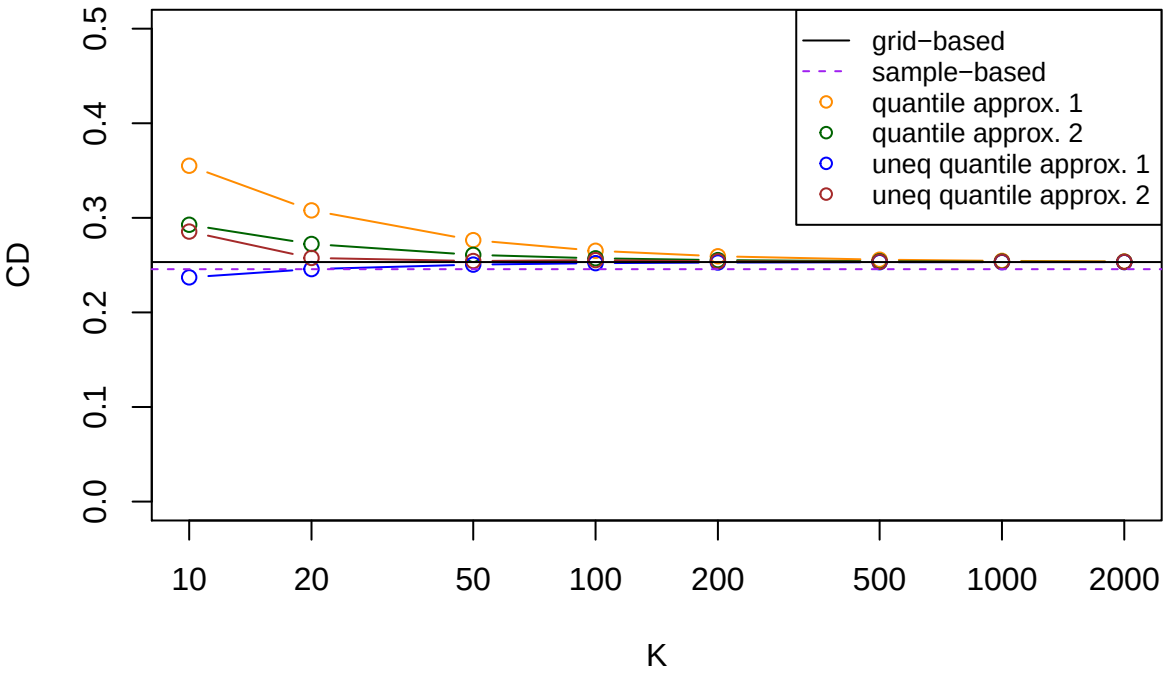
- Using the left-sided Riemann sum approximation and various values of K:

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

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

```
## [1] 0.2854597 0.2575762 0.2543386 0.2552775 0.2540318 0.2535609 0.2534094
## [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 (3) indeed coincides with (2).

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

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

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

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

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

The approximation (3) 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.

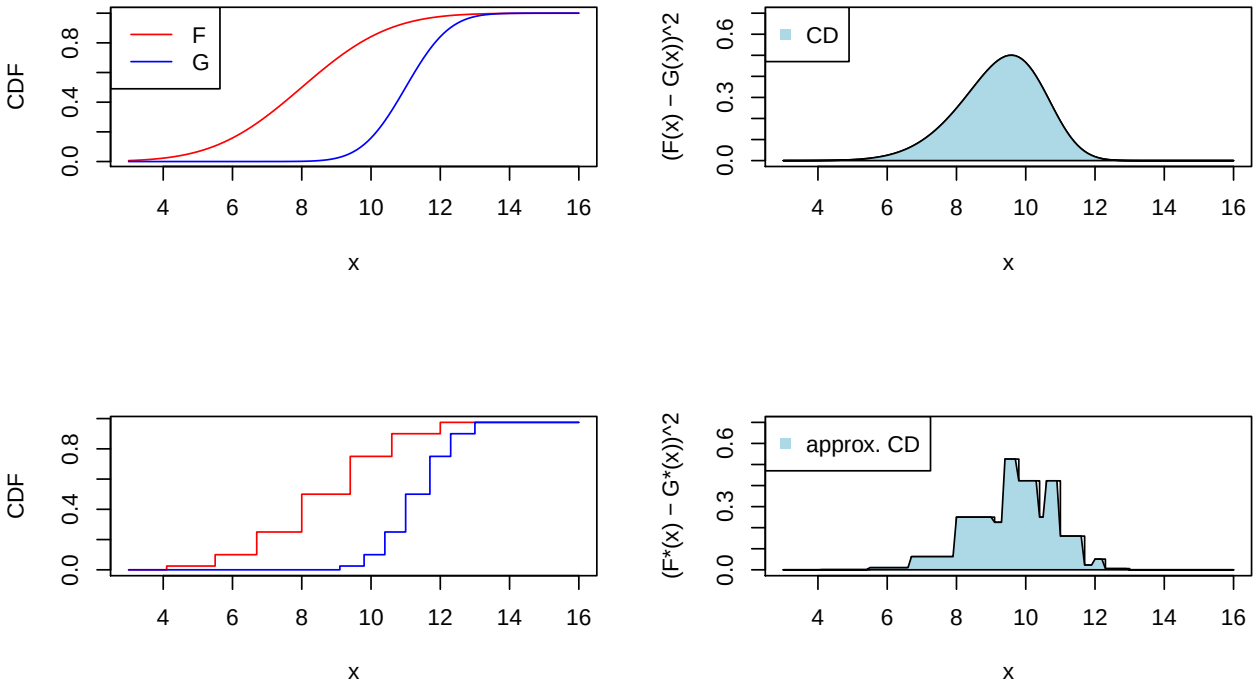
```
## [1] "Grid-based approx.: 0.61599852942592"
```

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.

```
## [1] 1.493653
```

- Using sampling and the alternative expression (1) of the CD from above:

```
## [1] 1.483948
```

- Using the first quantile-based approximation:

```
## [1] 1.919252
```

- Using the second quantile-based approximation:

```
## [1] 1.764859
```

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

```
## [1] 1.35122
```

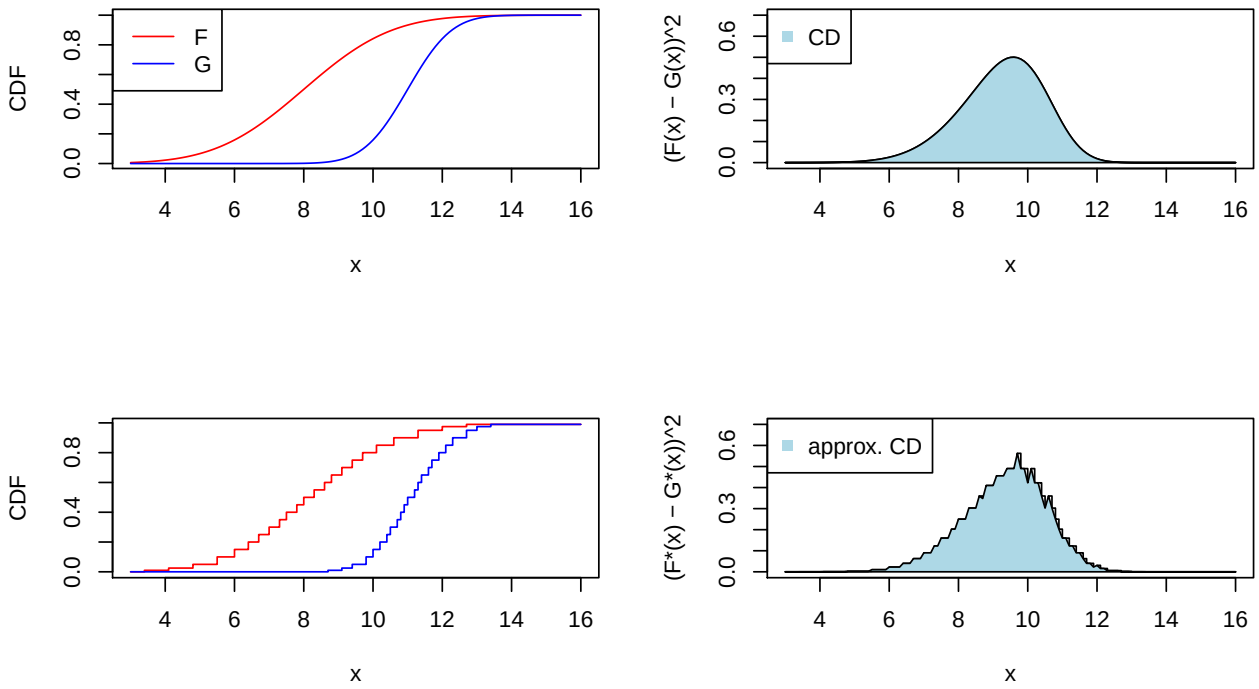
- Using the trapezoidal Riemann sum-based approximation:

```
## [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:

```
## [1] 1.640408
```

- Using the second quantile-based approximation:

```
## [1] 1.581296
```

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

```
## [1] 1.452266
```

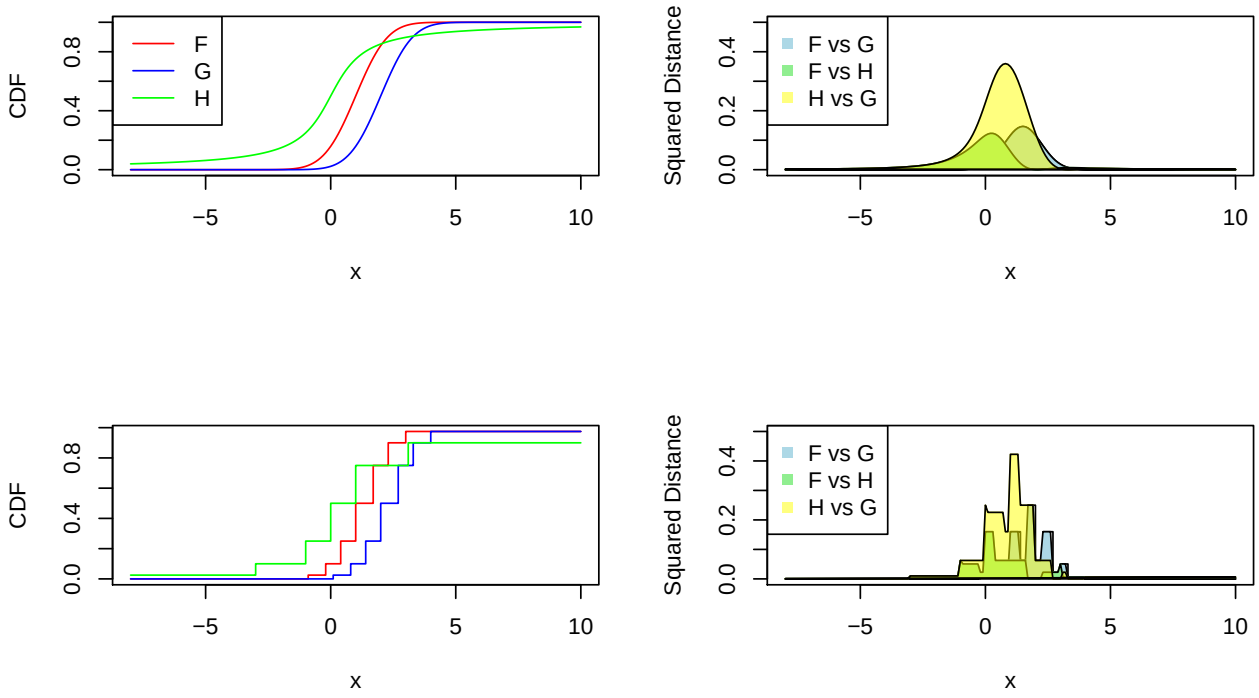
- Using the trapezoidal Riemann sum-based approximation:

```
## [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.

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

- Using the first quantile-based approximation:

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

- Using the second quantile-based approximation:

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

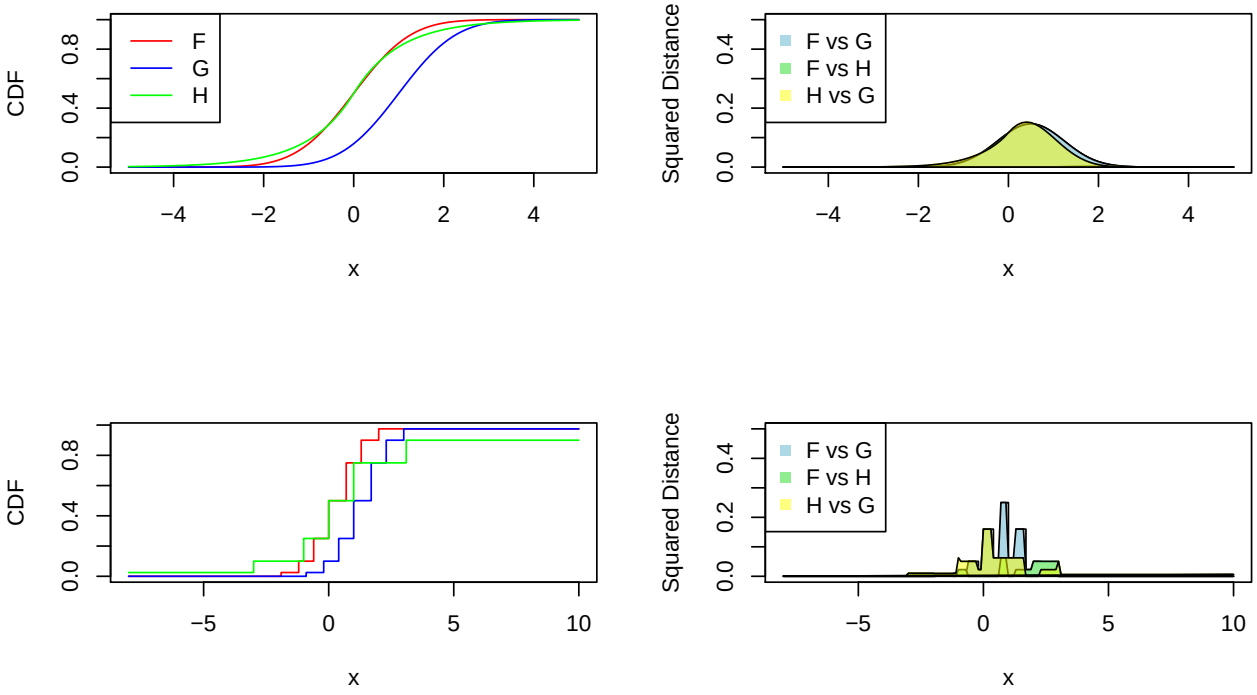
- Using the left-sided Riemann sum-based approximation:

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

- Using the trapezoidal Riemann sum-based approximation:

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

Long tails Suppose we have three cumulative distributions, $F \sim N(0,1)$, $G \sim N(1,1)$ and $H \sim$ Laplace$(0,1)$, 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.

```
## [1] "CD of F vs G: 0.270903289581517"
## [1] "CD of F vs H: 0.0068412250422997"
## [1] "CD of H vs G: 0.257655267503305"
```

- Using the first quantile-based approximation:

```
## [1] "Approx. CD of F vs G: 0.430834455349389"
## [1] "Approx. CD of F vs H: 0.0203971216157386"
## [1] " Approx. CD of H vs G: 0.416591384312851"
```

- Using the second quantile-based approximation:

```
## [1] "Approx. CD of F vs G: 0.349525091827873"
## [1] "Approx. CD of F vs H: 0.0116554980661363"
## [1] " Approx. CD of H vs G: 0.333247296357544"
```

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

```
## [1] "Approx. CD of F vs G: 0.267605148430715"
## [1] "Approx. CD of F vs H: 0.00892374070688564"
## [1] " Approx. CD of H vs G: 0.255142461273745"
```

- Using the trapezoidal Riemann sum-based approximation:

```
## [1] "Approx. CD of F vs G: 0.302511061162121"
## [1] "Approx. CD of F vs H: 0.0194133332014688"
```

```
## [1] " Approx. CD of H vs G: 0.259617817413175"
```