

Forecast Similarity

Johannes Bracher, Nutch Wattanachit, Evan Ray, Nick Reich, Nutch Wattanachit

07/21/2021

Cramér distance

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

$$\text{CD}(F, G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx \quad (1)$$

$$= \mathbb{E}_{F,G}|X - Y| - 0.5 [\mathbb{E}_F|X - X'| + \mathbb{E}_G|Y - Y'|], \quad (2)$$

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 (Thorarinnsson 2013, Gneiting and Raftery 2007), which is defined by

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}(x \geq y))^2 dx \quad (3)$$

$$= 2 \int_0^1 \{(\mathbf{1}(y \leq q_k^F) - \tau_k)(q_k^F - y)\} d\tau_k \quad (4)$$

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 τ_1, \dots, τ_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):

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

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

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

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 $q_i^F > q_j^G$ despite $i < j$ or $q_i^F \neq q_j^G$ despite $i = j$), a penalty $|q_i^F - q_j^G|$ is added to the sum. This corresponds to the shift which would be necessary to make q_i^F and q_j^G compatible.

Cramér Distance Approximation for Any Intervals

Using the same notations as before, now assume we don't necessarily have equally spaced τ_1, \dots, τ_K from two distributions F and G . The CRPS can be approximated as follows

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

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

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

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 19.

Left-sided Riemann sum approximation

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \quad (9)$$

$$\approx \sum_{j=1}^{2K-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) \quad (10)$$

$$(11)$$

Since $q_j \in \{q_1, \dots, q_{2K}\}$ belongs to either q_1^F, \dots, q_K^F or q_1^G, \dots, q_K^G , we can rewrite the above approximation using τ_1, \dots, τ_K as follows

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \{\hat{F}(q_j) - \hat{G}(q_j)\}^2 (q_{j+1} - q_j) \quad (12)$$

$$= \sum_{j=1}^{2K-1} \{\tau_j^F - \tau_j^G\}^2 (q_{j+1} - q_j) \quad (13)$$

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

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

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 .

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

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

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \int_{q_j}^{q_{j+1}} F(x) - G(x)^2 \quad (14)$$

$$\approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) \quad (15)$$

$$(16)$$

Similarly, we can rewrite the above approximation using τ_1, \dots, τ_K as defined in the left-sided Riemann sum approximation as follows

$$\text{CD}(F, G) \approx \sum_{j=1}^{2K-1} \frac{\{\hat{F}(q_j) - \hat{G}(q_j)\}^2 + \{\hat{F}(q_{j+1}) - \hat{G}(q_{j+1})\}^2}{2} (q_{j+1} - q_j) \quad (17)$$

$$= \sum_{j=1}^{2K-1} \frac{\{\tau_j^F - \tau_j^G\}^2 + \{\tau_{j+1}^F - \tau_{j+1}^G\}^2}{2} (q_{j+1} - q_j). \quad (18)$$

Decomposition of Approximated Cramér Distance

The Cramér distance is commonly used to measure the similarity of forecast distributions (see Richardson et al 2020 for a recent application). 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. This CRPS approximation given by (??) 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

$$\text{CD}(F, G) \approx \frac{1}{K(K+1)} \sum_{i=1}^K \sum_{j=1}^K \mathbf{1}\{(i-j) \times (q_i^F - q_j^G) \leq 0\} \times |q_i^F - q_j^G|, \quad (19)$$

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 $q_i^F > q_j^G$ despite $i < j$ or $q_i^F \neq q_j^G$ despite $i = j$), a penalty $|q_i^F - q_j^G|$ is added to the sum. This corresponds to the shift which would be necessary to make q_i^F and q_j^G compatible.

A divergence measure for central prediction intervals with potentially different nominal coverages

Consider two central prediction intervals $[l^F, u^F]$ and $[l^G, u^G]$ with nominal levels α^F and α^G , respectively (meaning that l^F is the $(1 - \alpha^F)/2$ quantile of F etc). We can define an *interval divergence* measure by

comparing the two pairs of predictive quantiles and summing up the respective incompatibility penalties as in (19). Adapting notation to the interval formulation and structuring the sum slightly differently, this can be written as:

$$\begin{aligned} \text{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) = & \mathbf{1}(\alpha^F \leq \alpha^G) \times \{\max(l^G - l^F, 0) + \max(u^F - u^G, 0)\} + \\ & \mathbf{1}(\alpha^F \geq \alpha^G) \times \{\max(l^F - l^G, 0) + \max(u^G - u^F, 0)\} + \\ & \max(l^F - u^G, 0) + \\ & \max(l^G - u^F, 0) \end{aligned}$$

The first row adds penalties for the case where $[l^F, u^F]$ should be nested in $[l^G, u^G]$, but at least one of its ends is more extreme than the respective end of $[l^G, u^G]$. The second row covers the converse case. The last two rows add penalties if the lower end of one interval exceeds the upper end of the other, i.e. the intervals do not overlap.

This can be seen as a (scaled version of a) generalization of the interval score, but writing out the exact relationship is a bit tedious.

We now define four auxiliary terms with an intuitive interpretation which add up to the interval divergence:

- The term

$$D_F = \mathbf{1}(\alpha^F \leq \alpha^G) \times \max\{(u^F - l^F) - (u^G - l^G), 0\}$$

is the sum of penalties resulting from F being more dispersed than G . It is positive whenever the interval $[l^F, u^F]$ is longer than $[l^G, u^G]$, even though it should be nested in the latter. D_F then tells us by how much we would need to shorten $[l^F, u^F]$ so it could fit into $[l^G, u^G]$.

- The term

$$D_G = \mathbf{1}(\alpha^G \leq \alpha^F) \times \max\{(u^G - l^G) - (u^F - l^F), 0\}$$

measures the converse, i.e. overdispersion of G relative to F .

- The term

$$S^F = \max\{\mathbf{1}(\alpha^G \leq \alpha^F) \times \max(l^F - l^G, 0) + \mathbf{1}(\alpha^F \leq \alpha^G) \times \max(u^F - u^G, 0) + \max(l^F - u^G, 0) - D_F - D_G, 0\}$$

sums over penalties for values in $\{l^F, u^F\}$ exceeding those from $\{l^G, u^G\}$ where they should not (only counting penalties not already covered in D_F or D_G). It thus represents an *upward shift* of F relative to G .

- The term

$$S^G = \max\{\mathbf{1}(\alpha^F \leq \alpha^G) \times \max(l^G - l^F, 0) + \mathbf{1}(\alpha^G \leq \alpha^F) \times \max(u^G - u^F, 0) + \max(l^G - u^F, 0) - D_G - D_F, 0\}$$

accordingly represents an *upward shift* of G relative to F .

It can be shown that

$$\text{ID}([l^F, u^F], [l^G, u^G], \alpha^F, \alpha^G) = D_F + D_G + S^F + S^G$$

Intuitively the interval divergence measures by how much we need to move the quantiles of the interval with lower nominal coverage so it fits into the one with larger nominal coverage.

Approximating the Cramér distance using interval divergences

Assuming K is even, the K equally spaced predictive quantiles of each distribution can be seen as $L = K/2$ central prediction intervals with coverage levels $\alpha_i = 2i/(L+1)$, $i = 1, \dots, L$. Similarly to the definition of the WIS, the approximation (19) can also be expressed in terms of these intervals as

$$\text{CD}(F, G) \approx \frac{1}{2L(2L+1)} \sum_{k=1}^L \sum_{m=1}^L \text{ID}([l_k^F, u_k^F], [l_m^G, u_m^G], \alpha_k^F, \alpha_m^G).$$

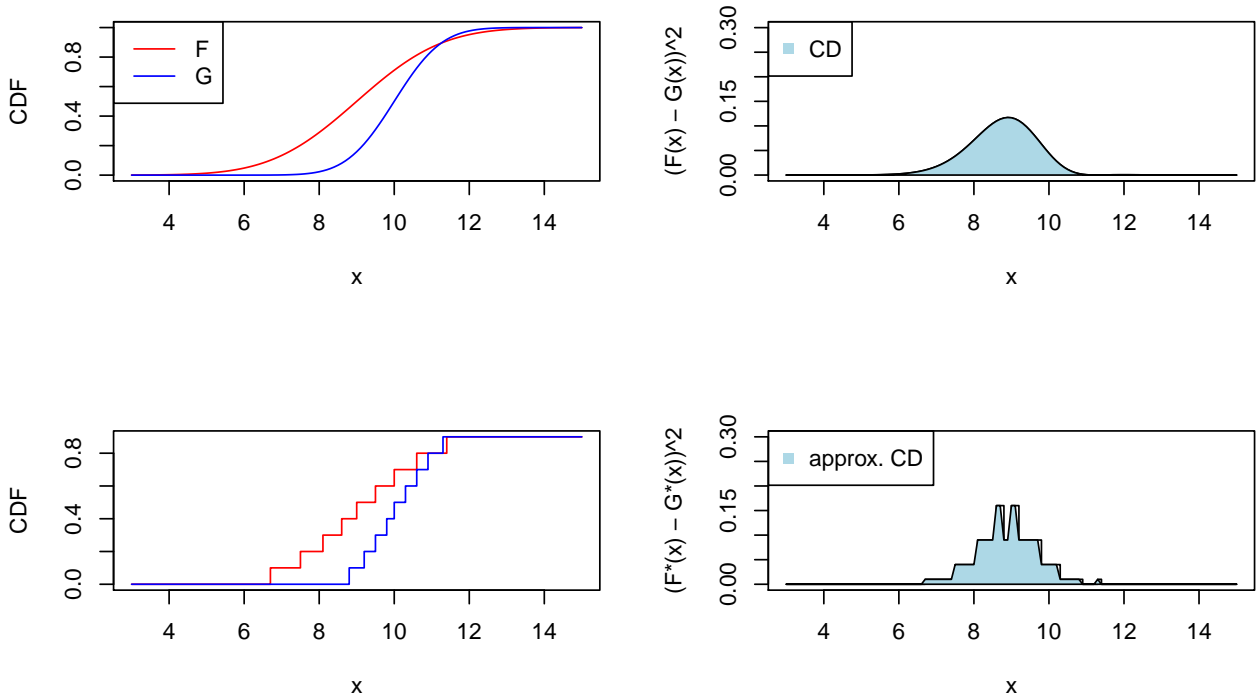
This implies a decomposition of the Cramér distance into the four interpretable components defined for the interval divergence in the previous section. If G is a one-point distribution, the CD reduces to the WIS and the proposed decomposition reduces to the well-known decomposition of the WIS into dispersion, overprediction and underprediction components.

Note that in practice we usually have an uneven rather than even number K of predictive quantiles. In this case the median needs to be treated separately (comparisons of the “0% prediction interval” need to be weighted down with a factor of 2; this is the same little quirk as the one identified by Ryan and Evan for the WIS a few months ago). The decomposition has the following properties:

- Additive shifts of the two distributions only affect the shift components, not the dispersion components.
- Consequently, if G and G are identical up to an additive shift, both dispersion components will be 0.
- If F and G are both symmetric and have the same median, the both shift components will be 0.
- I think that in general it is possible that both shift components or both dispersion components are greater than 0, which leads to a somewhat strange interpretation. But this should only concern constructed examples.

Examples

Equally-spaced intervals



In this example, six different approximations are applied to the distributions $F \sim N(9, 1.8)$ and $G \sim N(10, 1)$ 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
```

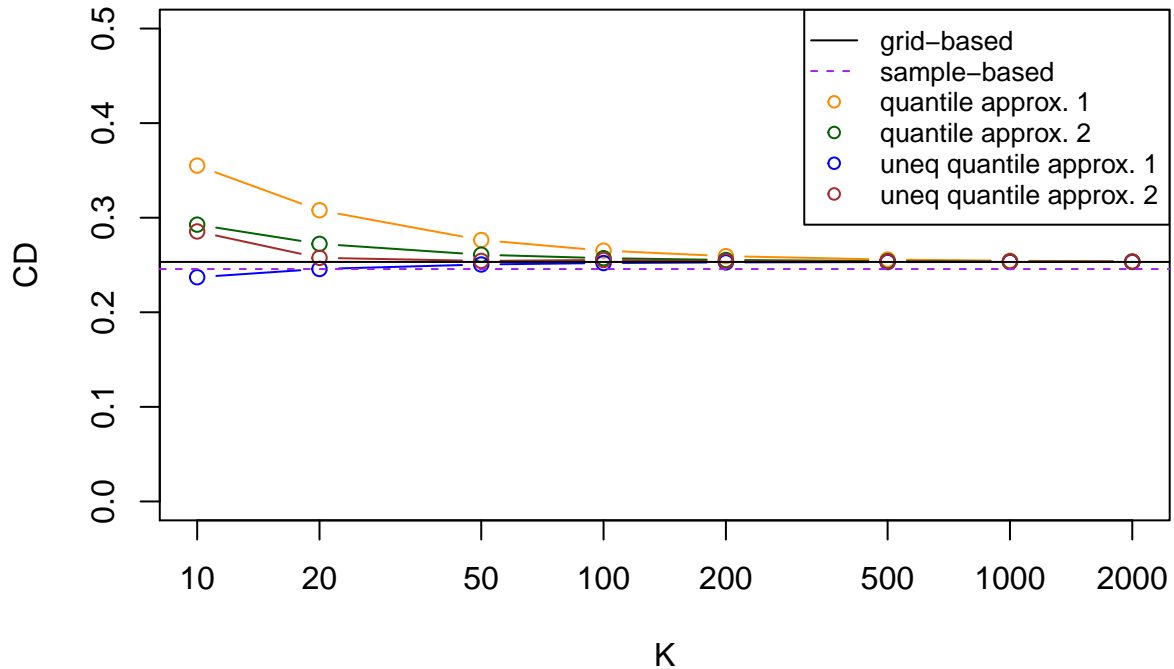
```
FALSE [8] 0.2532128
```

- 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.

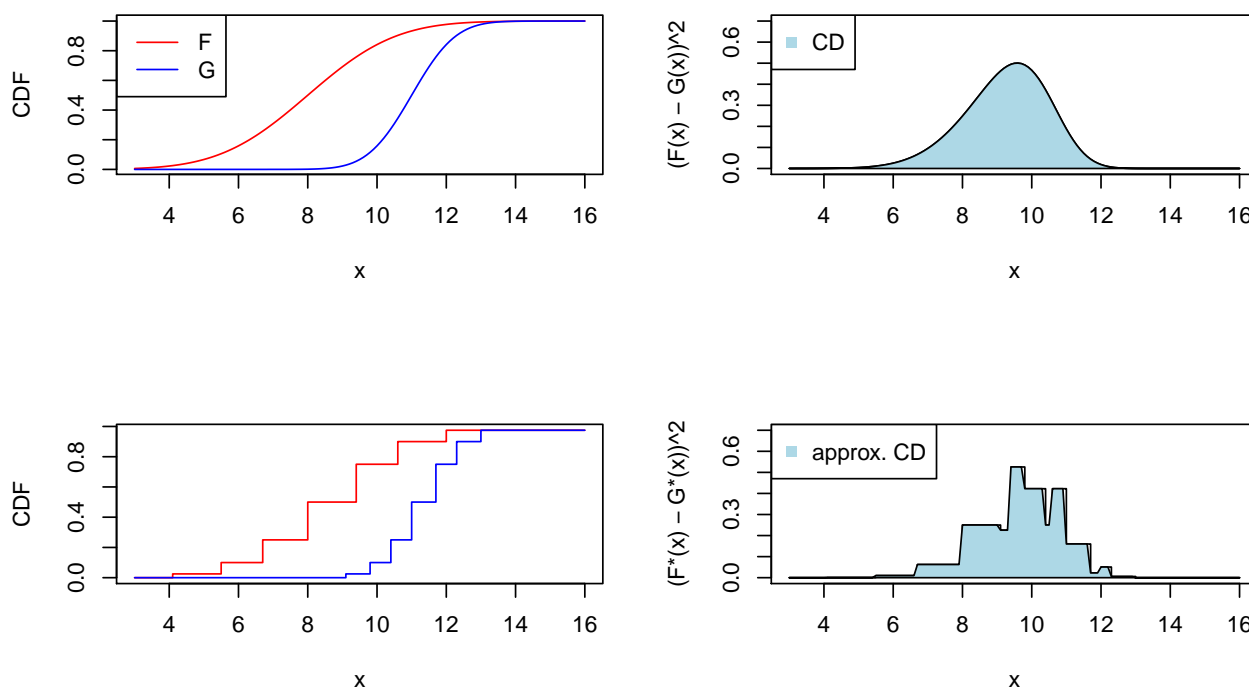
```
FALSE [1] "Grid-based approx.: 0.61599852942592"
```

Unequally-spaced 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
```

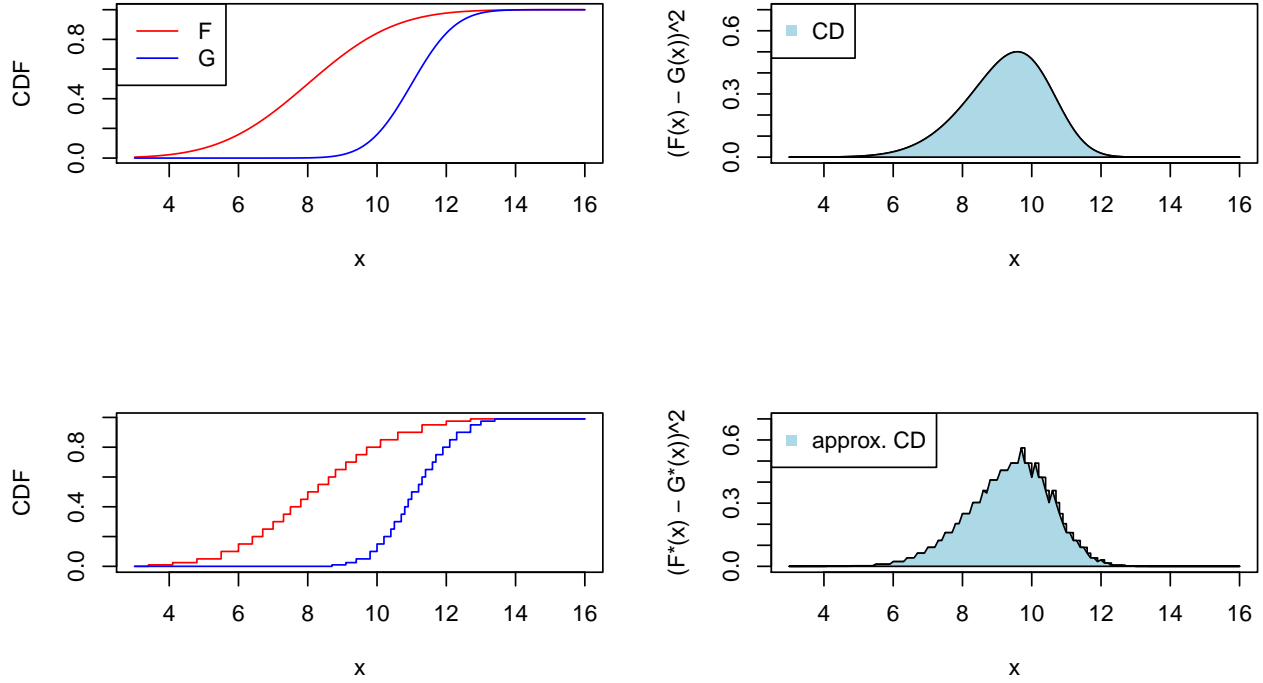
- Using the trapezoidal Riemann sum-based approximation:

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

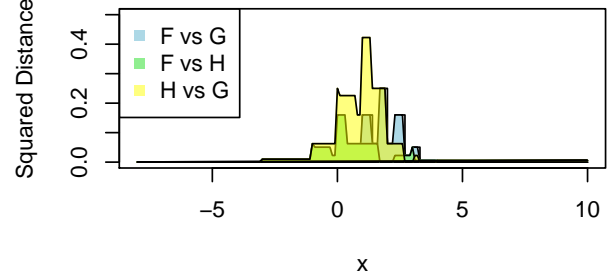
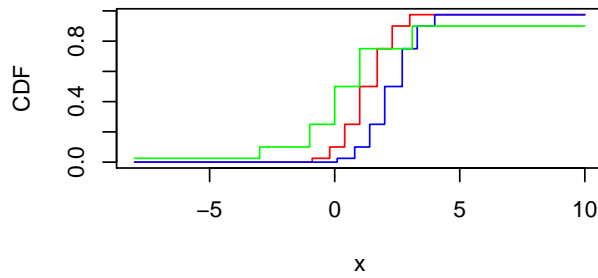
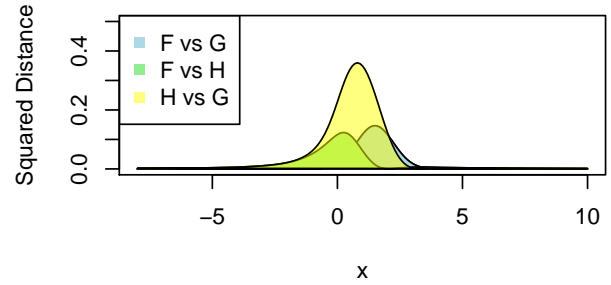
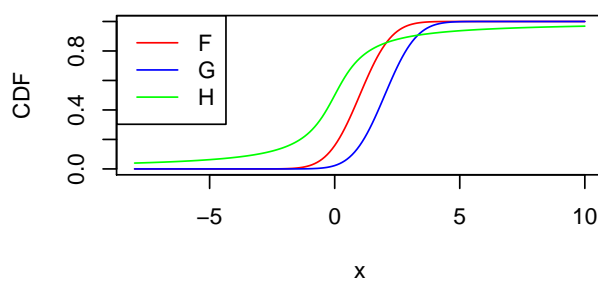
- Using the trapezoidal Riemann sum-based approximation:

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"