

# CDF Bounds

Chris Calabro

August 30, 2003

Consider 2 random variables  $X, Y$  each of finite range and with the same mean. Is it possible for the cumulative distribution function of one to bound that of the other? Trivially, this is true if  $X, Y$  have the same distribution. But is it possible when they do not? We will show that the answer is no and then give 1 application.

## 1 Main theorem

**Theorem 1.** *Let  $p, q, c \in \mathbb{R}^n$  such that*

- (1)  $p_i, q_i \geq 0$
- (2)  $\sum p_i = \sum q_i = 1$
- (3)  $\sum c_i p_i = \sum c_i q_i$  where  $c_1 < \dots < c_n$
- (4)  $\forall a \in [1, n] \sum_{i=1}^a p_i \leq \sum_{i=1}^a q_i$ .

*Then  $p = q$ .*

This will be very hard to prove. Let us first note that condition (1) is not needed and that the 1 in condition (2) is also not needed. We prove the following generalization.

**Theorem 2.** *Let  $\delta, c \in \mathbb{R}^n$  such that*

- (1)  $\sum \delta_i = 0$
- (2)  $\sum c_i \delta_i = 0$  where  $c_1 < \dots < c_n$
- (3)  $\forall a \in [1, n] \sum_{i=1}^a \delta_i \geq 0$ .

*Then  $\delta = 0$ .*

This will prove theorem 1 as a corollary by setting  $\delta = q - p$ . That  $c$  is increasing is a crucial hypothesis, for without it,  $\delta = (1, 1, -2), c = (\frac{1}{3}, \frac{2}{3}, \frac{1}{2})$  would be a counterexample to theorem 2. We will first prove a special case to get the idea.

*Proof of theorem 2 assuming  $\delta \in \mathbb{Q}^n$ .* If the theorem fails, then wlog we may assume also that  $\delta_i \in \mathbb{Z} - \{0\}$  and  $\min_{i \neq j} |c_i - c_j| \geq 1$ . (here we need  $c$  to be strictly increasing)

**Lemma 3 (mountain).** *Suppose  $\delta \in \{\pm 1\}^n$ ,  $\sum \delta_i = 0$ ,  $\forall a \in [1, n]$   $\sum_{i=1}^a \delta_i \geq 0$ . Then there is a bijection*

$$f : \{i \in [1, n] \mid \delta_i > 0\} \rightarrow \{i \in [1, n] \mid \delta_i < 0\}$$

such that  $\forall i \in \text{dom } f$   $f(i) > i$ .

*Proof.* To fuel our intuition, note that the hypotheses show that the CDF function  $g(a) = \sum_{i=1}^a \delta_i$  is shaped like a mountain: it starts at 0, ends at 0, and is always  $\geq 0$ . The idea is that each upward segment can be bijectively matched to a later downward segment.

We prove the lemma by induction on  $n$ . If  $n = 0$ , we take  $f = \emptyset$ .  $n$  cannot be 1 since  $n$  is even. Now suppose that  $n \geq 2$ . Suppose that the mountain touches the base so that  $\exists a \in [2, n-1]$   $g(a) = 0$ . Then by the induction hypothesis, there are bijections

$$\begin{aligned} f_1 &: \{i \in [1, a] \mid \delta_i > 0\} \rightarrow \{i \in [1, a] \mid \delta_i < 0\} \\ f_2 &: \{i \in [a+1, n] \mid \delta_i > 0\} \rightarrow \{i \in [a+1, n] \mid \delta_i < 0\} \end{aligned}$$

such that  $f_1(i) > i$ ,  $f_2(i) > i$ . We take  $f = f_1 \cup f_2$ .

On the other hand, if the mountain does not touch the base so that  $\neg \exists a \in [2, n-1]$   $g(a) = 0$ . Then by the induction hypothesis, there is a bijection

$$f_3 : \{i \in [2, n-1] \mid \delta_i > 0\} \rightarrow \{i \in [2, n-1] \mid \delta_i < 0\}$$

such that  $f_3(i) > i$ . We take  $f = f_3 \cup \{(1, n)\}$ . □

We set

$$\begin{aligned} \delta_{i,j} &= \text{sgn } \delta_i && \text{for } j \in [1, |\delta_i|] \\ c_{i,j} &= c_i && \text{for } j \in [1, |\delta_i|]. \end{aligned}$$

and give these the lexicographical ordering. Then the hypotheses of lemma 3 hold. Of course, it is no longer true that  $c_{i,j} < c_{i',j'}$  for  $(i, j) < (i', j')$ , but the  $f$  constructed in lemma 3 must match  $(i, j)$  with  $f(i, j) = (i', j') > (i, j)$  such that  $\delta_{i,j} > 0$ ,  $\delta_{i',j'} < 0$ , which implies

$$\begin{aligned} \text{sgn } \delta_i \neq \text{sgn } \delta_{i'} &\Rightarrow i \neq i' \\ &\Rightarrow i' > i \quad \text{since } (i', j') > (i, j) \\ &\Rightarrow c_{i'} > c_i \\ &\Rightarrow c_{i',j'} > c_{i,j}. \end{aligned}$$

So

$$\sum_i c_i \delta_i = \sum_{i,j} c_{i,j} \delta_{i,j} = \sum_{\substack{i,j \\ \delta_{i,j} > 0}} (c_{i,j} - c_{f(i,j)}) < 0,$$

contradicting (2).  $\square$

*Proof of theorem 2.* This will be somewhat harder, but will use the same basic strategy. So suppose indirectly that there is a counterexample to the theorem with  $\delta_i \in \mathbb{R}$ . Wlog assume that  $\delta_i \neq 0$  and  $\min_{i \neq j} |c_i - c_j| \geq 1$ .

We approximate  $\delta_i$  by  $q_i \in \mathbb{Q} - \{0\}$  such that  $|\delta_i - q_i| \leq \frac{1}{d}$  and  $|\delta_i| \leq |q_i|$ . In other words,

$$q_i = \begin{cases} \frac{\lceil d\delta_i \rceil}{d} & \text{if } \delta_i > 0 \\ \frac{\lfloor d\delta_i \rfloor}{d} & \text{if } \delta_i < 0 \end{cases}.$$

Then  $|\sum q_i| \leq \frac{n}{d}$ ,  $\sum_{i=1}^a q_i \geq -\frac{n}{d}$ , and

$$\sum c_i q_i = \sum c_i (\delta_i + q_i - \delta_i) \geq \sum c_i (q_i - \delta_i) \geq -\frac{n \max |c_i|}{d}.$$

Let  $z_i = \begin{cases} \lceil d\delta_i \rceil & \text{if } \delta_i > 0 \\ \lfloor d\delta_i \rfloor & \text{if } \delta_i < 0 \end{cases}$ . Then  $|\sum z_i| \leq n$ ,  $\sum_{i=1}^a z_i \geq -n$ ,  $\sum c_i z_i \geq -n \max |c_i|$ . We need a generalization to lemma 3 that says all but  $n$  upward segments of the mountain can be bijectively matched with later downward segments.

**Lemma 4.** *Suppose that  $z \in \{\pm 1\}^n$ ,  $|\sum z_i| \leq b$ ,  $\forall a \in [1, n]$   $\sum_{i=1}^a z_i \geq -b$ . Then  $\exists U \subseteq \{i \in [1, n] \mid z_i > 0\}$   $|U| \geq \frac{n-3b}{2}$  and an injection*

$$f : U \rightarrow \{i \in [1, n] \mid z_i < 0\}$$

such that  $\forall i \in U$   $f(i) > i$ .

*Proof.* Now the mountain  $g(a) = \sum_{i=1}^a z_i$  begins at 0, ends in  $[-b, b]$ , and is always  $\geq -b$ . Let  $b' = \sum z_i$ . We add  $b$  extra segments to the beginning:  $z_{-(b-1)} = \dots = z_{-1} = z_0 = 1$ . We also add  $b + b'$  extra segments to the end:  $z_{n+1} = \dots = z_{n+b+b'} = -1$ . Then  $z_{-(b-1)}, \dots, z_{n+b+b'}$  satisfies the hypotheses of lemma 3. So there is a bijection

$$f : \{i \in [-(b-1), n+b+b'] \mid z_i > 0\} \rightarrow \{i \in [-(b-1), n+b+b'] \mid z_i < 0\}$$

such that  $f(i) > i$ . Let us now remove from  $f$  the images and preimages of the newly added segments. So let  $f' = f|([1, n] - f^{-1}[n+1, n+b+b'])$ ,  $U = \text{dom } f'$ . Then

$$\begin{aligned} |\text{dom}(f|[1, n]) - \text{dom } f'| &\leq b + b' \\ \Rightarrow |U| &\geq \frac{n+b'}{2} - (b+b') = \frac{n-b'}{2} - b \geq \frac{n-3b}{2}. \end{aligned}$$

$\square$

So now set  $N = \sum |z_i| \geq n \min |z_i| \geq nd \min |\delta_i|$  and let

$$\begin{aligned} z_{i,j} &= \operatorname{sgn} z_i && \text{for } j \in [1, |z_i|] \\ c_{i,j} &= c_i && \text{for } j \in [1, |z_i|] \end{aligned}$$

and give these the lexicographical ordering. Then the hypotheses of lemma 4 hold. (with lemma 4  $b = n$ , lemma 4  $n = N$ ) Again for  $f(i, j) = (i', j') > (i, j)$ , we have  $z_{i,j} > 0, z_{i',j'} < 0$ , which implies

$$\begin{aligned} \operatorname{sgn} z_i \neq \operatorname{sgn} z_{i'} &\Rightarrow i \neq i' \\ &\Rightarrow i' > i \quad \text{since } (i', j') > (i, j) \\ &\Rightarrow c_{i'} > c_i \\ &\Rightarrow c_{i',j'} > c_{i,j}. \end{aligned}$$

So

$$\begin{aligned} -n \max |c_i| &\leq \sum_i c_i z_i = \sum_{i,j} c_{i,j} z_{i,j} \\ &\leq 3n \max |c_i| + \sum_{(i,j) \in \operatorname{dom} f} (c_{i,j} - c_{f(i,j)}) \\ &\leq 3n \max |c_i| - |\operatorname{dom} f| \\ &\leq 3n \max |c_i| - \frac{N - 3n}{2}, \end{aligned}$$

and so

$$\begin{aligned} N &\leq n(8 \max |c_i| + 3) \\ \Rightarrow nd \min |\delta_i| &\leq N \leq n(8 \max |c_i| + 3) \\ \Rightarrow d &\leq \frac{8 \max |c_i| + 3}{\min |\delta_i|}. \end{aligned}$$

So choose  $d > \frac{8 \max |c_i| + 3}{\min |\delta_i|}$  to get a contradiction.  $\square$

## 2 An application

Let  $X$  be hypergeometric (good, out of, samples) =  $(s, n, i)$  and  $Y$  be binomial (samples, rate) =  $(i, \frac{s}{n})$ . Then  $E(X) = \sum P(X_j = 1) = \frac{si}{n}$  where

$$X_j = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ sample is good} \\ 0 & \text{else} \end{cases}$$

and  $E(Y) = \frac{si}{n}$ . So theorem 1 applies with  $c_j = j$ . i.e. conditions (1), (2), (3) hold, but the conclusion is false. So (4) must fail!

The moral of the story is that if you ever want to bound the CDF of a random variable by the CDF of another random variable (perhaps with a simpler CDF function), you must not choose them to have the same mean.