# Conditioning on and controlling for variates via cumulative differences

## Measuring calibration, reliability, biases, and other treatment effects

Mark Tygert

Meta

mark@tygert.com

# Outline

## Calibration and reliability

# Calibration and reliability

# Probabilistic predictions

# Probabilistic predictions

# Predictions with a given probability

1. Consider an $S = 30\%$ chance of snow.
2. The reality will be either $R = 1$ (it actually did snow) or $R = 0$ (it did not snow).
3. A random variable $R$ which follows the Bernoulli distribution with expected value $S$ takes $R = 1$ with probability $S$ and $R = 0$ with probability $1 - S$.
4. Given $n$ independent random results, also known as "responses," $R_1, R_2, \ldots, R_n$, the prediction is perfectly calibrated if $S = 30\%$ of $R_1, R_2, \ldots, R_n$ are equal to 1, and the rest are equal to 0.

# Predictions with different probabilities

1. Probabilities of success $S_1, S_2, \ldots, S_n$; the probabilities are the expected values of Bernoulli distributions.
2. Responses $R_1, R_2, \ldots, R_n$; the responses are independent Bernoulli random variables ($R_k$ and $S_k$ come as a pair).
3. Probabilities of success are also known as "scores" and will be viewed as deterministic during this tutorial.
4. Responses are also known as "results" or "outcomes" and viewed as random and probabilistically independent.
5. Goal: gauge deviation of the responses from the scores.

# Without loss of generality

1. Re-order the indices such that $S_1 \leq S_2 \leq \cdots \leq S_n$, preserving the pairing of $R_k$ with $S_k$ for every $k = 1, 2, \ldots, n$.
2. Randomly perturb the scores such that $S_1 < S_2 < \cdots < S_n$ if the inequalities were not already strict.
3. There exists an alternative to this random perturbation that is more complicated but that avoids randomization altogether. Anyone interested can consult the later section, "Extensions," specifically its subsection, "Avoiding randomization."

# Cumulative differences

# Cumulative differences

# Cumulative aggregation

The cumulative difference of responses from scores is

$$C_k = \frac{1}{n} \sum_{j=1}^{k} (R_j - S_j) \tag{1}$$

for $k = 1, 2, \ldots, n$.

The expected slope of a graph of $C_k$ versus $k/n$ from $j = k - 1$ to $j = k$ is

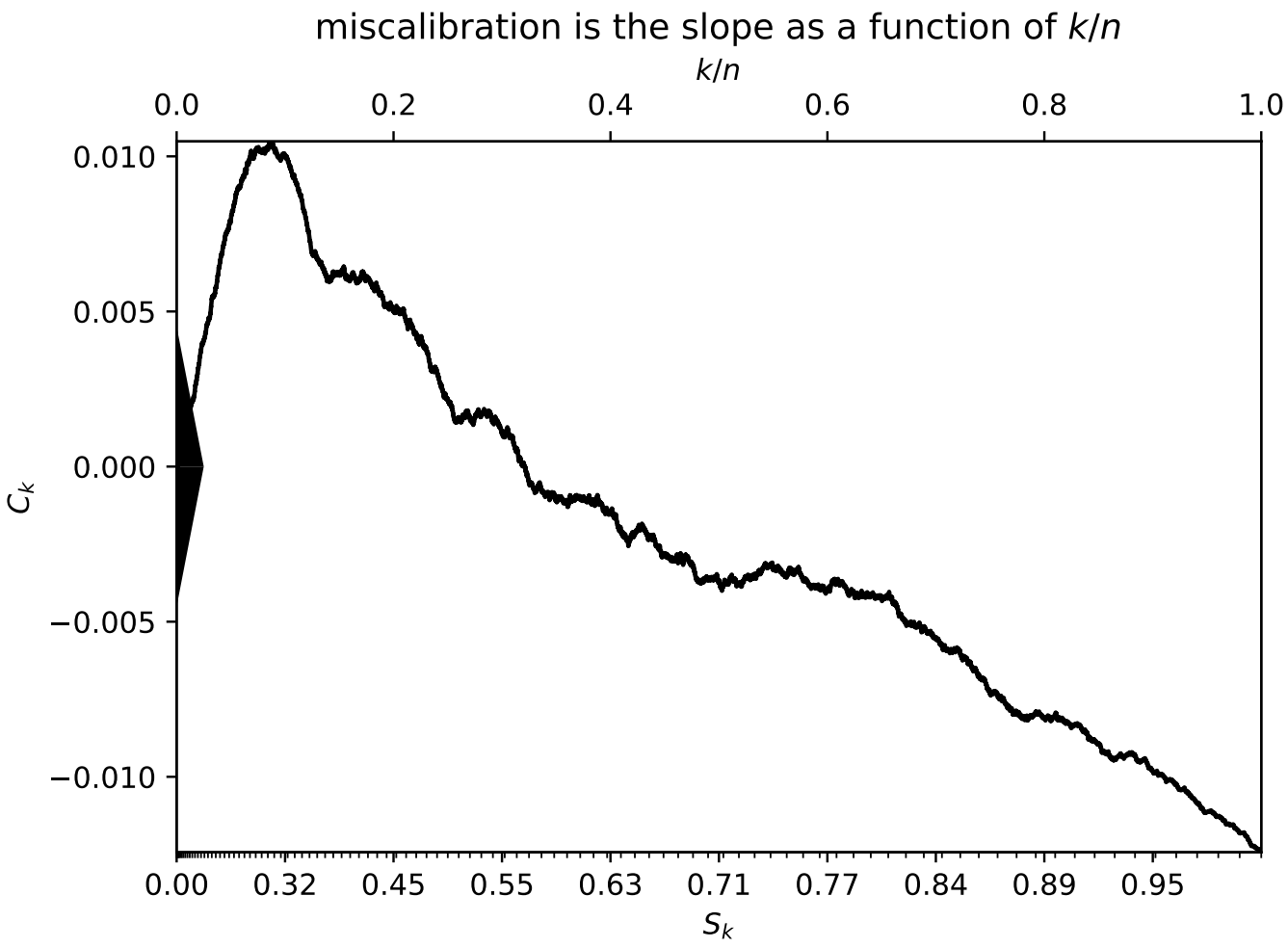
$$\frac{\mathbb{E}[C_k - C_{k-1}]}{k/n - (k-1)/n} = \mathbb{E}[R_k] - S_k \tag{2}$$

for $k = 1, 2, \ldots, n$; this $\mathbb{E}[R_k] - S_k$ is the expected difference from perfect calibration.

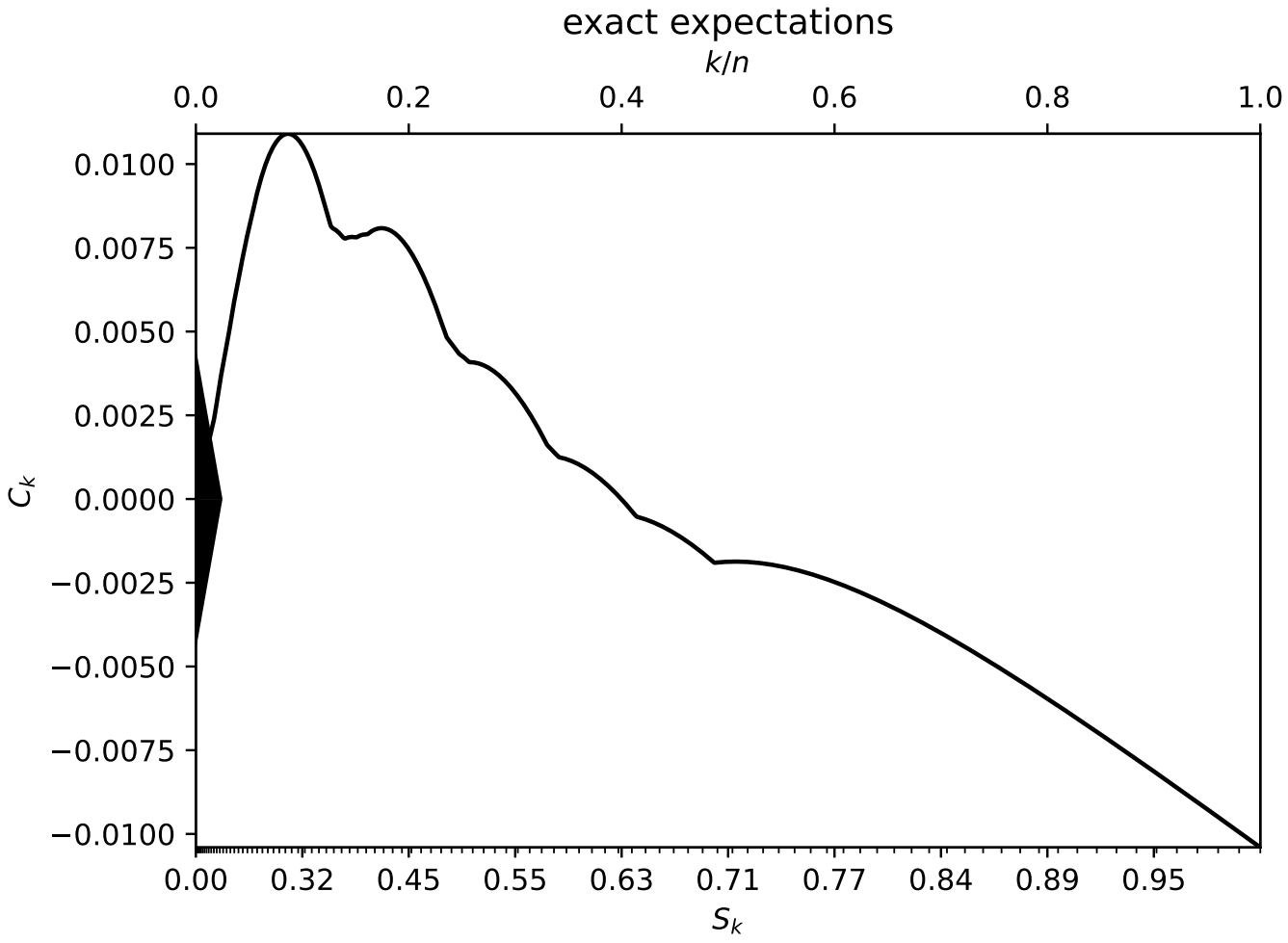
# Miscalibration is slope

1. The slope of a secant line connecting two points on the graph of $C_k$ versus $k/n$ becomes the average miscalibration over the long range of $k$ between those points. This follows from combining the law of large numbers with the last formula on the previous slide.
2. Slope is easy to perceive with quantitative precision even when the constant offsets of the secant lines are irrelevant.
3. Miscalibration at an index $k$ of interest determines the slope, unpolluted by accumulation from indices other than that of interest (the other indices affect only the constant offset).

# Empirical plot with $n = 2^{15} = 32{,}768$



# Plot of the ground truth



# Kolmogorov-Smirnov and Kuiper metrics

1. Good calibration corresponds to a flat, horizontal graph.
2. Deviation from flat and horizontal measures miscalibration.
3. Two scalar statistics which summarize the deviations from 0 are the max. absolute deviation and the range of deviations.
4. The max. absolute deviation is Kolmogorov's and Smirnov's

$$C_{\mathrm{MAD}} = \max_{1 \leq k \leq n} |C_k|. \tag{3}$$

5. The range of deviations is Kuiper's

$$C_{\mathrm{range}} = \max_{0 \leq k \leq n} C_k - \min_{0 \leq k \leq n} C_k, \tag{4}$$

where $C_0 = 0$.

# Another expression for the Kuiper metric

The absolute value of the total miscalibration $\sum_{j \in I}(R_j - S_j)/n$ over any interval $I$ of indices is less than or equal to $C_{\text{range}}$; indeed, the Kuiper metric $C_{\text{range}}$ is equal to the maximum of the absolute value of the total miscalibration over any interval of indices:

$$C_{\text{range}} = \max_I \left| \frac{1}{n} \sum_{j \in I} (R_j - S_j) \right|, \tag{5}$$

where the maximum is taken over every interval $I$ of indices; summing over the index $j \in I$ means summing over the indices $j = \min I,\ 1 + \min I,\ 2 + \min I,\ 3 + \min I,\ \ldots,\ \max I$.

# Significance testing

1. The P-value for a given assertion asserting a null hypothesis is the probability that a specified measure of deviation from the assertion is greater than or equal to that measure for the actual observations. The metric could be $C_{\text{range}}$, for example.
2. The null hypothesis relevant here is of perfect calibration, namely that every response $R_k$ comes from the Bernoulli distribution whose expected value is the associated score $S_k$.
3. A small P-value indicates that the observations are inconsistent with the assertion (the null hypothesis).
4. If $P$ is the P-value, then $1 - P$ is the attained confidence level for believing the assertion (the null hypothesis). This value $1 - P$ is not exactly the probability of the null hypothesis being true, but is a monotonically increasing function of such a probability, for most formalizations of such a probability.

# P-values

1. Under the null hypothesis of perfect calibration $\mathbb{E}[R_k] = S_k$ for $k = 1, 2, \ldots, n$, $C_{\mathrm{MAD}}/\sigma$ converges in distribution to the max. absolute value of the standard Brownian motion over $[0, 1]$, and $C_{\mathrm{range}}/\sigma$ converges in distribution to the Brownian motion's range, where $\sigma^2$ is the total expected variance

$$\sigma^2 = \sum_{k=1}^{n} \frac{S_k(1 - S_k)}{n^2}, \tag{6}$$

assuming that $\max_{1 \leq k \leq n} S_k(1 - S_k) / \sum_{j=1}^{n} S_j(1 - S_j)$ converges to 0 as $n$ becomes arbitrarily large.

2. $C_{\mathrm{MAD}}$ is $0.01243/\sigma = 5.512$, $C_{\mathrm{range}}$ is $0.02291/\sigma = 10.16$, and the corresponding asymptotic P-values are 7.1E–08 and less than the machine precision, for the earlier example.

# Why Brownian motion?

1. Under the null hypothesis of perfect calibration, the graph of the line segments connecting $(S_0, C_0)$, $(S_1, C_1)$, ..., $(S_n, C_n)$ are the definition of a driftless random walk (driftless means that the expected value of $C_k - C_{k-1} = (R_k - S_k)/n$ is 0 for all $k = 1, 2, \ldots, n$, which follows from the null hypothesis that the expected value of $R_k$ is $S_k$). The graph of the line segments converges in distribution to the graph of Brownian motion, under the condition on the previous slide (as $n \to \infty$).
2. The variance of a Bernoulli distribution whose expected value is $S_k$ is $S_k(1 - S_k)$. The variance of a sum of independent random variables is the sum of the variances of the random variables. Therefore, the variance of $C_n = \sum_{k=1}^{n}(R_k - S_k)/n$ is $\sigma^2$ defined on the previous slide. Dividing by $\sigma$ standardizes the Brownian motion to be over the unit interval $[0, 1]$.

# Error bars

1. The standard deviation of the standard Brownian motion over the unit interval $[0, 1]$ ranges from 0 at 0 to 1 at 1.
2. The cumulative graphs plot the cumulative differences without normalizing by $\sigma$; standardizing the Brownian motion associated with the cumulative graphs would require normalization by $\sigma$.
3. The cumulative plots (including the two from earlier slides) display a triangle at the origin whose tip-to-tip height is $4\sigma$, corresponding to a roughly 95% confidence interval.
4. The triangle at the origin indicates the length scale of statistically significant deviations from zero, effectively showing the effect of normalization by $\sigma$.

# Binned responses

Binned responses

# Binning (instead of the newer cumulative approach)

$S_1^1 < S_1^2 < \cdots < S_1^{n_1}$ will be the least $n_1$ of the scores,
$S_2^1 < S_2^2 < \cdots < S_2^{n_2}$ will be the next $n_2$ of the scores, ...,
$S_m^1 < S_m^2 < \cdots < S_m^{n_m}$ will be the greatest $n_m$ of the scores.
Maintain $n = \sum_{j=1}^{m} n_j$ and calculate the average score

$$\tilde{S}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} S_j^k \tag{7}$$

and the average response

$$\tilde{R}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} R_j^k \tag{8}$$

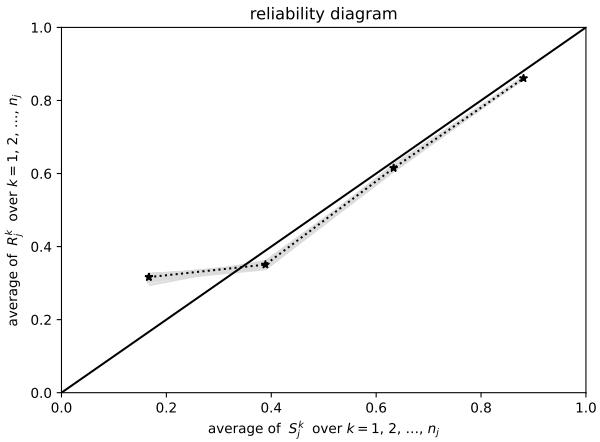
for $j = 1, 2, \ldots, m$.

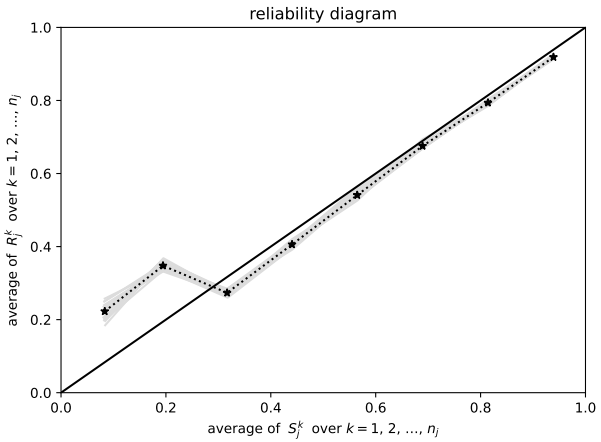
# Reliability diagram

The "reliability diagram" is the graph of the average response $\tilde{R}_j$ versus the average score $\tilde{S}_j$, together with a diagonal line from the origin $(0, 0)$ to the point $(1, 1)$; the diagonal line corresponds to perfect calibration, for which the response is equal to the score. The reliability diagram is the traditional graphical method for assessing calibration.

# Same range of probabilities for each bin, 4–32 bins

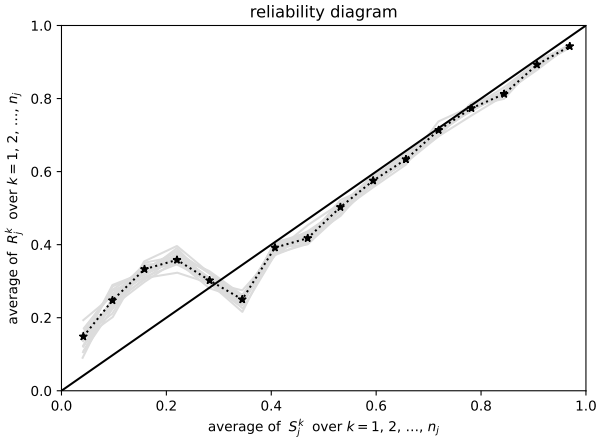
$m = 2^2$

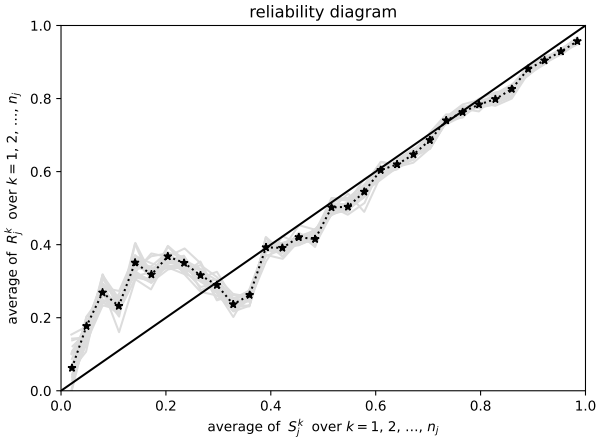



$m = 2^3$

$m = 2^4$

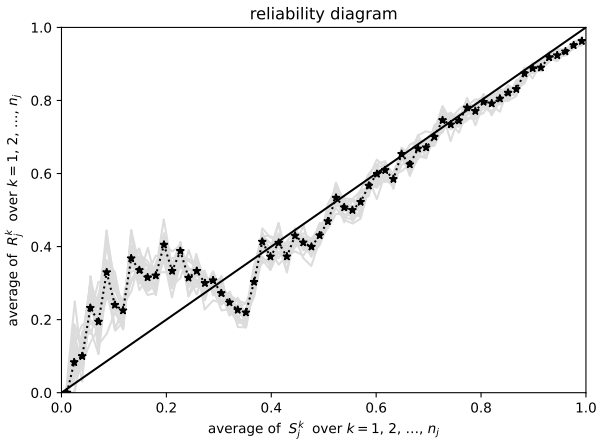


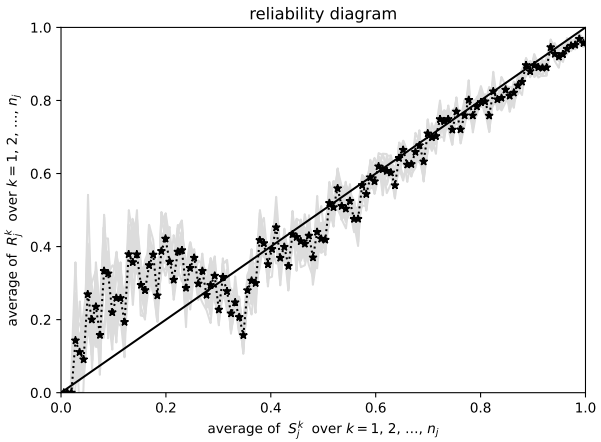




$m = 2^5$

# Same range of probabilities for each bin, 64–256 bins

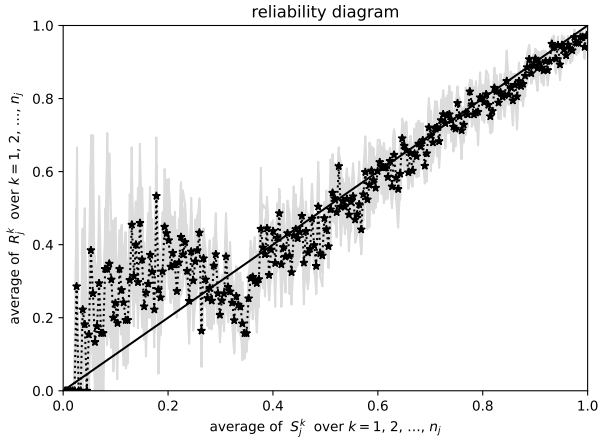
$m = 2^6$

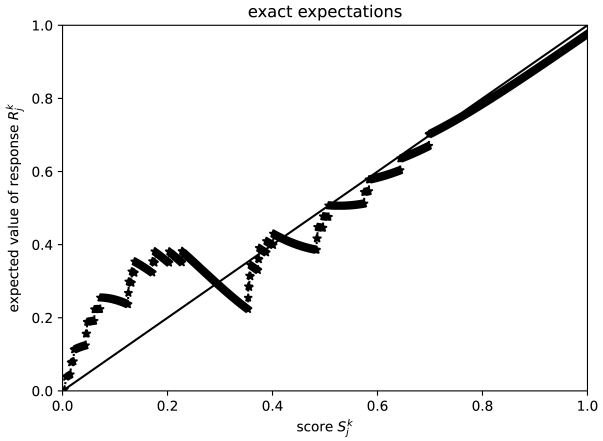



$m = 2^7$

$m = 2^8$

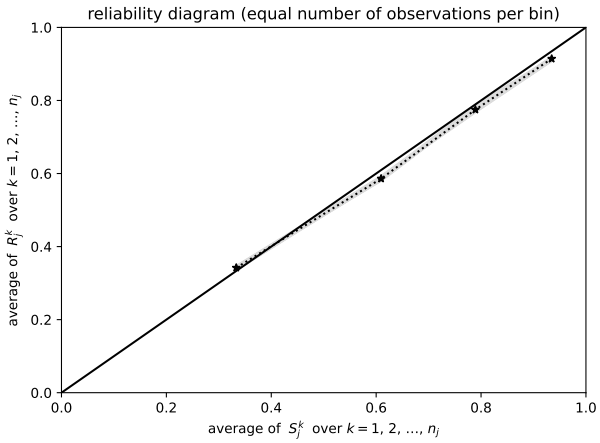


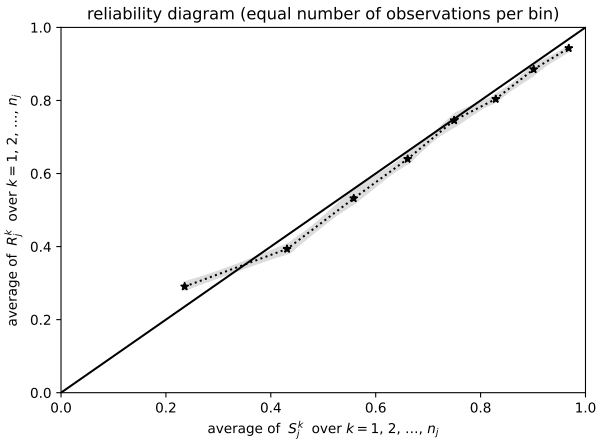




$n = 2^{15}$

# Same number of responses for each bin, 4–32 bins

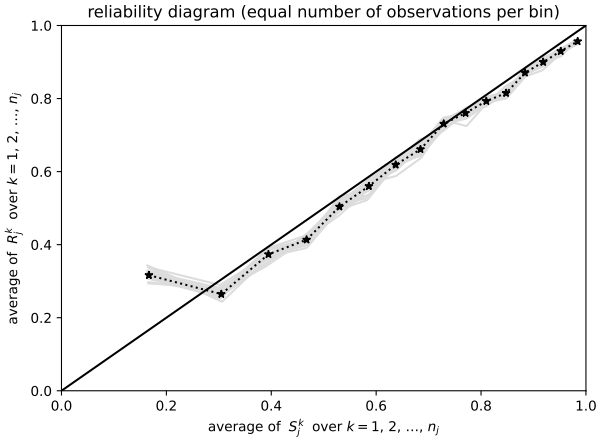
$m = 2^2$

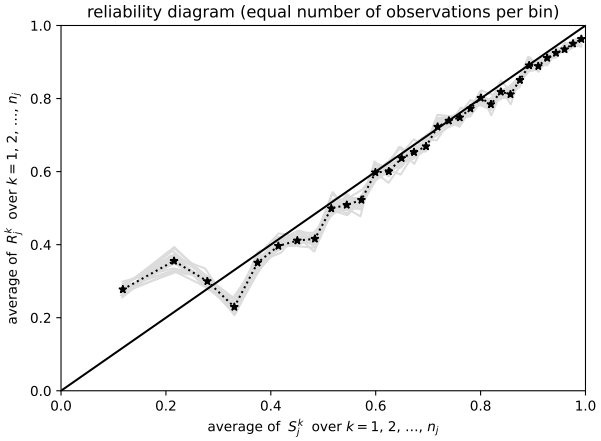



$m = 2^3$

$m = 2^4$

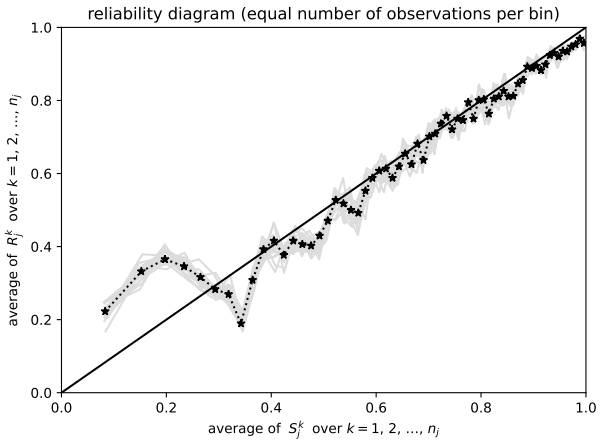


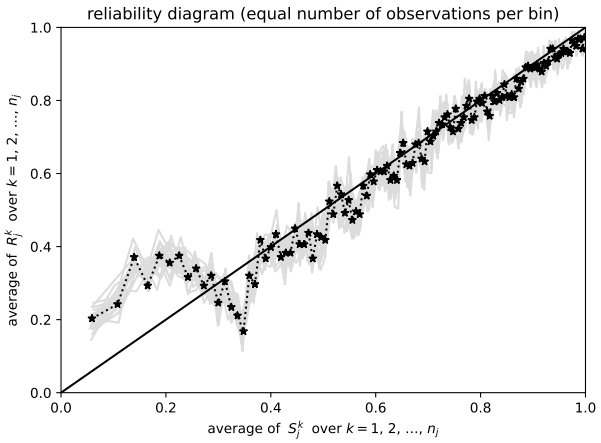




$m = 2^5$

# Same number of responses for each bin, 64–256 bins

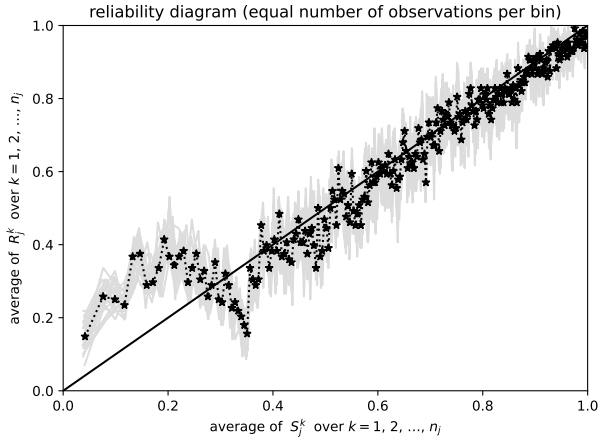
$m = 2^6$

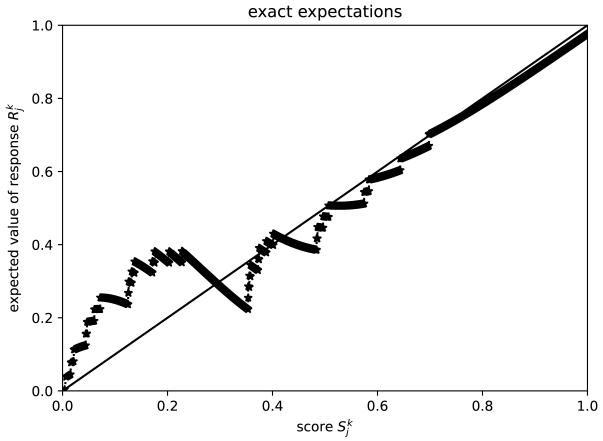



$m = 2^7$

$m = 2^8$




$n = 2^{15}$

# Empirical calibration errors

The empirical calibration errors (ECEs) are the Riemann sums

$$\mathrm{ECE}^1 = \sum_{j=1}^{m} (S_{j+1}^1 - S_j^1) \left| \tilde{R}_j - \tilde{S}_j \right| = \sum_{j=1}^{m} (S_{j+1}^1 - S_j^1) \left| \sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j} \right| \tag{9}$$

and

$$\mathrm{ECE}^2 = \sum_{j=1}^{m} (S_{j+1}^1 - S_j^1) \left| \tilde{R}_j - \tilde{S}_j \right|^2 = \sum_{j=1}^{m} (S_{j+1}^1 - S_j^1) \left| \sum_{k=1}^{n_j} \frac{R_j^k - S_j^k}{n_j} \right|^2 \tag{10}$$

where $S_{m+1}^1 = 1$, and the bin width $(S_{j+1}^1 - S_j^1)$ can be replaced with $1/m$ when the average scores are roughly equispaced on $[0, 1]$. The ECEs are the traditional metrics for assessing calibration.

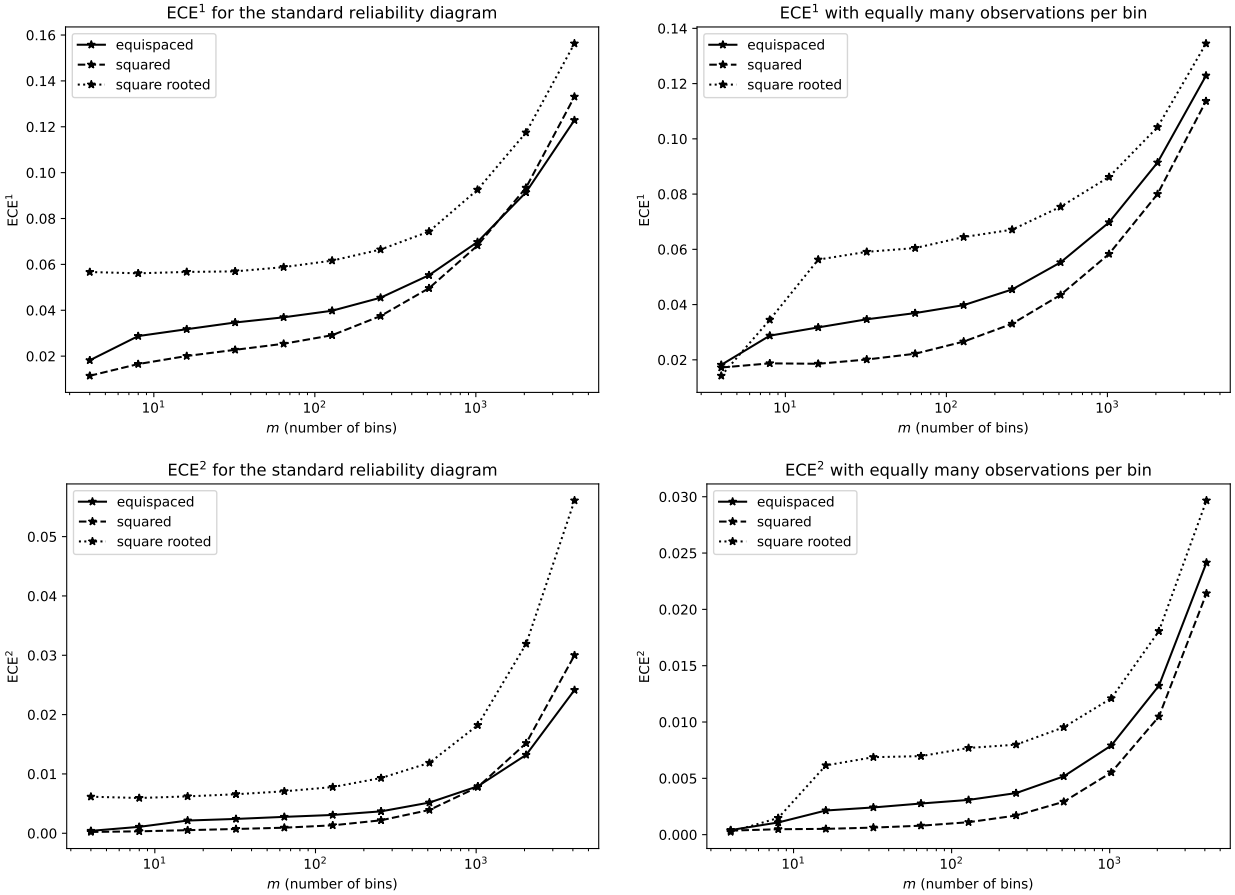
# Integrated calibration index (ICI) = ECE

While "ECE" is the accepted standard terminology, biostatisticians have recently re-introduced the ECE, referring to the ECE as the "integrated calibration index" (ICI). There is now a large literature that uses "ICI" rather than "ECE."

Beware, too, that the abbreviation, "ECE," can refer to "empirical," "estimated," "expected," or "experimental" calibration errors; all these possibilities for the first letter in "ECE" pertain to the same mathematical formulas from the previous slide.

# ECEs vary wildly with the choice of bins

$n = 2^{15}$; earlier, the scores were square rooted from equispaced.



# Comparison of ECEs to cumulative summary statistics

1. If the number of observations per bin stays bounded on a set interval of scores as the max. width of a bin converges to 0, then the ECEs hit a noise floor, which prevents the ECEs from distinguishing a fixed imperfectly calibrated distribution from perfect calibration, even in the limit of infinite sample size $n$.
2. This highlights a trade-off inherent to binning (or kernel density estimation) — decreasing the width of bins or kernels resolves finer variations as a function of score at the expense of averaging away less noise, while increasing the width of bins sacrifices resolving power but boosts statistical confidence.
3. The cumulative statistics $C_{\mathrm{range}}$ and $C_{\mathrm{MAD}}$ of Kuiper and of Kolmogorov and Smirnov have no such explicit trade-off.

## Fairness and matching scores

# Fairness and matching scores

# Deviation of a subpopulation from the full pop.

Deviation of a subpopulation from the full pop.

# Fairness for a subpopulation

1. Goal: gauge deviation of the responses of the subpopulation from those of the full population.
2. Compare only individuals who are comparable.
3. Use the scores to match up individuals being compared.
4. Deviation of the subpopulation from the full population may indicate inequities in or other effects on outcomes or treatment of members of the subpopulation.
5. Comparing individuals conditional on their scores ensures that the score variate is not the cause of differences detected.
6. Examples of scores in medicine and the social sciences include age, income, or predicted probability (as with calibration).

# Respecting the subpopulation's sampling

1. In addition to the scores $S_1 < S_2 < \cdots < S_p$ and corresponding responses $R_1, R_2, \ldots, R_p$ for a full population of $p$ individuals, consider a subpopulation specified by $n$ indices $i_1 < i_2 < \cdots < i_n$.
2. Define $\tilde{R}_{i_k}$ to be the average of the full population's responses whose corresponding scores are closer to $S_{i_k}$ than to any other of the subpopulation's scores $S_{i_1}, S_{i_2}, \ldots, S_{i_n}$.
3. Calculate the cumulative differences

$$C_j = \frac{1}{n} \sum_{k=1}^{j} \left( R_{i_k} - \tilde{R}_{i_k} \right) \tag{11}$$

for $j = 1, 2, \ldots, n$.

# Slope is deviation of the subpop. from the full pop.

The expected slope of a graph of $C_k$ versus $k/n$ from $j = k - 1$ to $j = k$ is

$$\frac{\mathbb{E}[C_k - C_{k-1}]}{k/n - (k-1)/n} = \mathbb{E}[R_{i_k}] - \mathbb{E}[\tilde{R}_{i_k}] \tag{12}$$

for $k = 1, 2, \ldots, n$; this $\mathbb{E}[R_{i_k}] - \mathbb{E}[\tilde{R}_{i_k}]$ is the expected deviation of the subpopulation from the full population.

The slope of a secant line connecting two distant points on the graph is the average deviation of the subpopulation from the full population over the subpopulation's indices between those points.

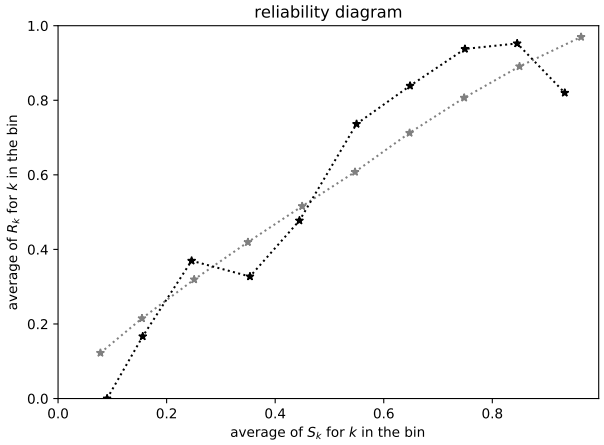
This analyzes the difference of the subpopulation's responses from the full population's while "controlling for" (also known as "conditioning on") scores, matching individuals with similar scores.

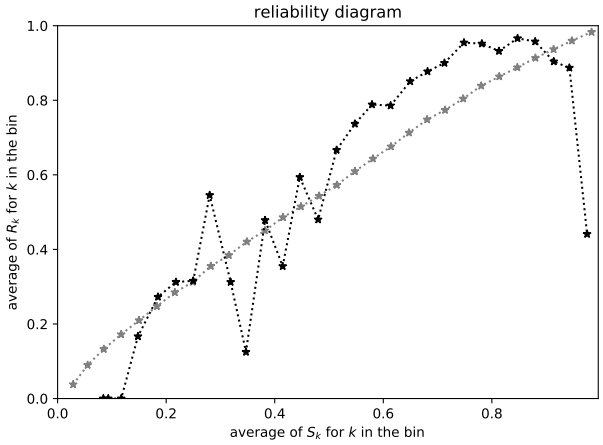
# Wild boars vs. full ImageNet

The next two slides display reliability diagrams and cumulative graphs for the difference in accuracy of classification for a subpopulation from the full population in popular training data from computer vision known as ImageNet. The black points and lines correspond to the images of wild boars in ImageNet, while the gray points and lines correspond to the full data set. The score is the predicted probability corresponding to the class predicted to be most likely; the score is a measure of confidence in the prediction. The reliability diagrams vary significantly as the number of bins varies, while the cumulative graphs are reasonably easy to interpret; in particular, the drop-off for the highest scores is easy to quantify in the cumulative graphs. The reliability diagrams might even look inconsistent with each other without looking at the cumulative graphs (the latter resolve the inconsistencies at a glance).

# Wild boars vs. full ImageNet (same width for each bin)

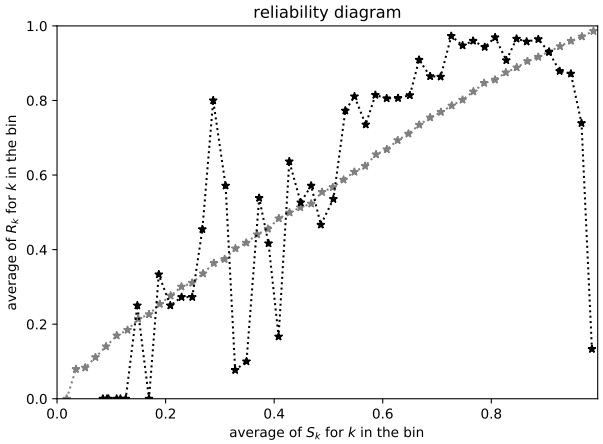
$m = 10$

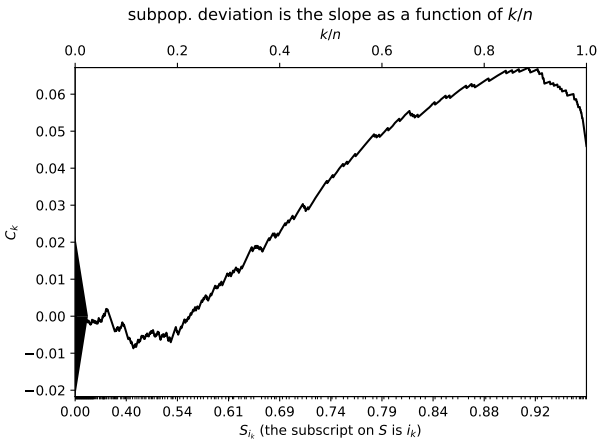



$m = 30$

$m = 50$

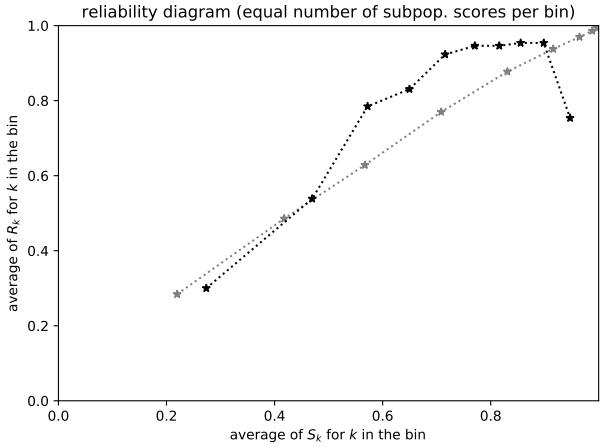


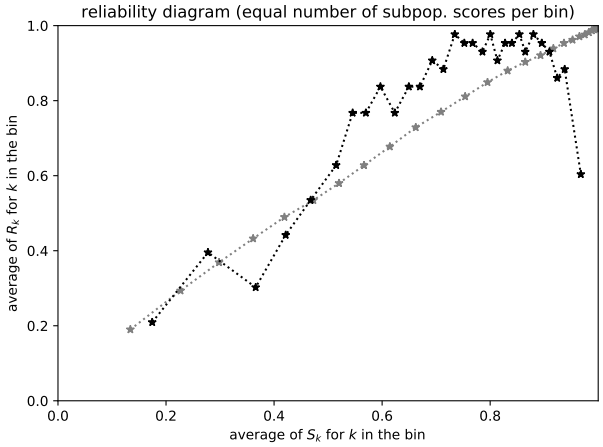




$n$=1300

# Wild boars vs. full ImageNet (adaptive widths of bins)

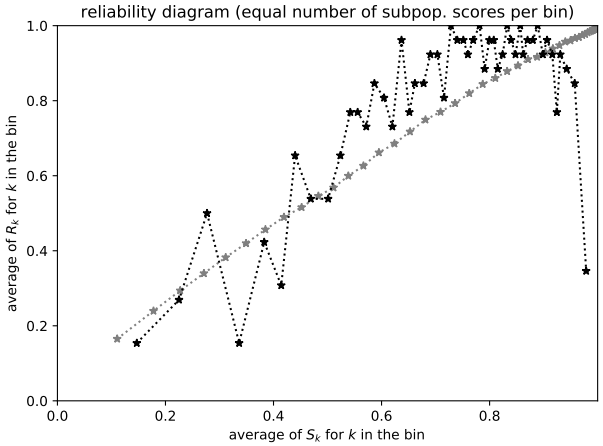
$m = 10$

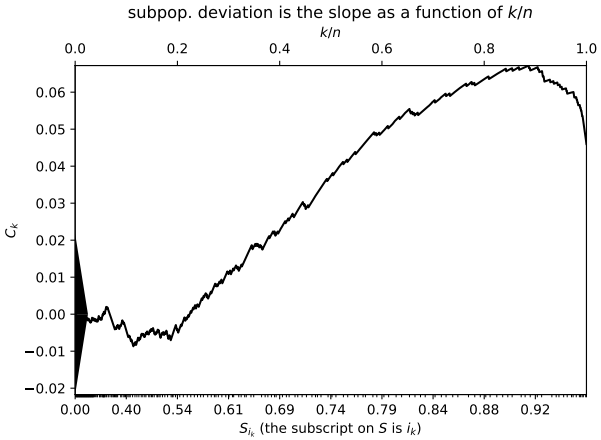



$m = 30$

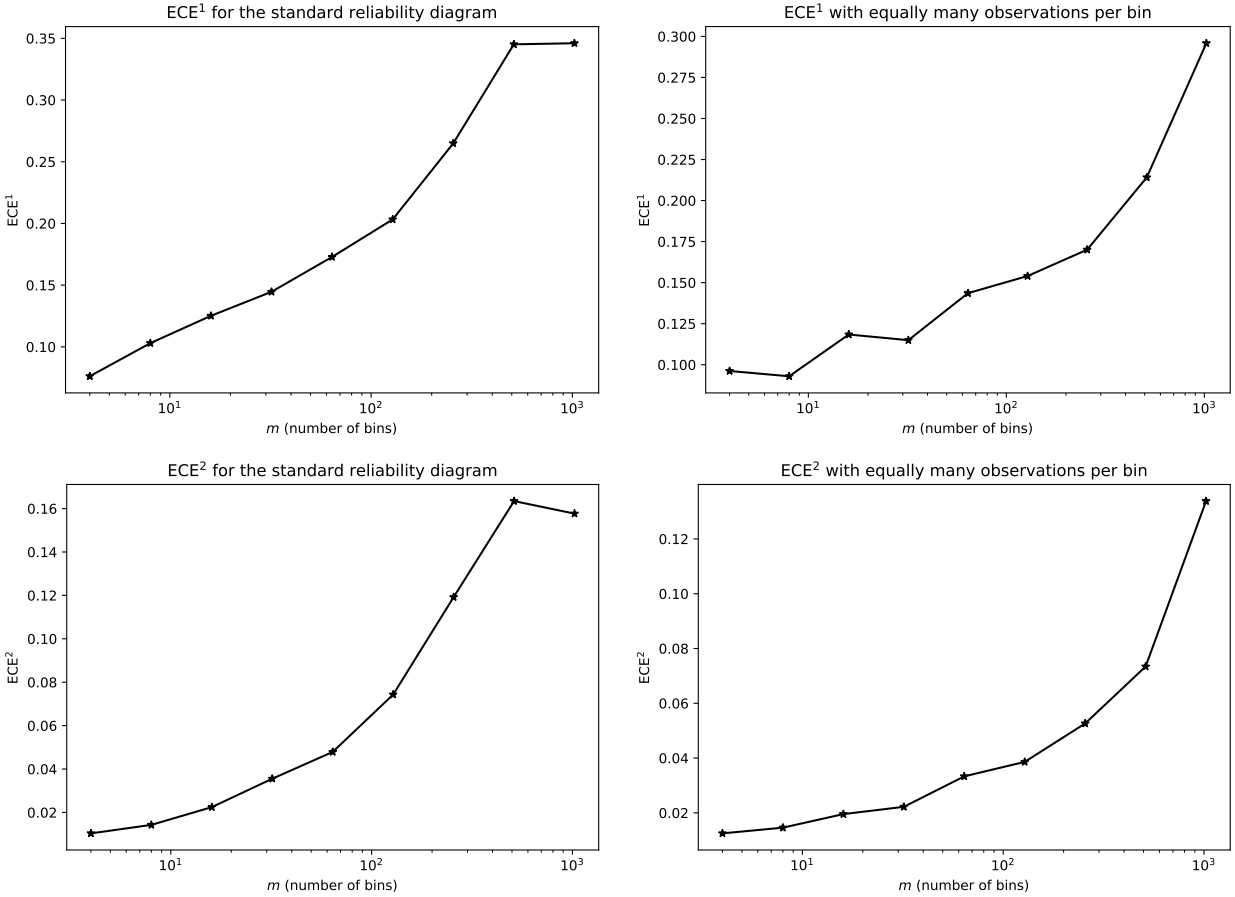
$m = 50$




$n$=1300

# Wild boars' empirical calibration errors




What is the right choice of bins? Which is the right value for $m$?

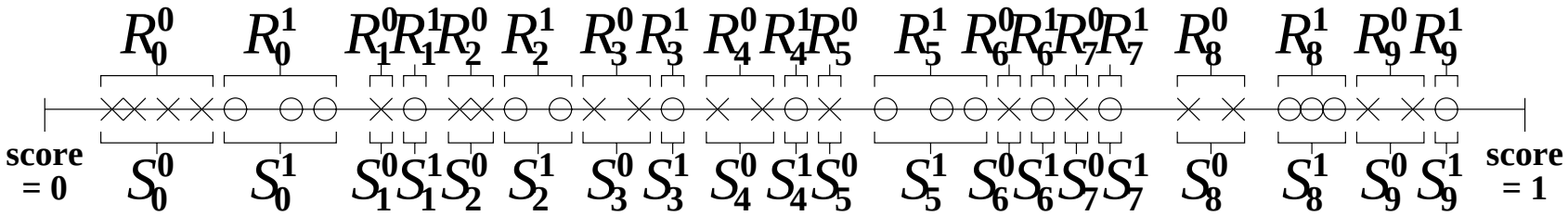
# Deviation between two disjoint subpopulations

Deviation between two disjoint subpopulations

# Caveats

1. Comparing a subpopulation to the full population while controlling for scores is always legitimate — every member of the subpopulation matches up with at least one member of the full population, namely, that very same member (that is the very definition of "*sub*-population"!).
2. Comparing different subpopulations directly might be absurd. Indeed, the scores for one subpopulation might even all be less than all the scores for another subpopulation; there is no way to match up members according to score in such a case.
3. Nevertheless, it can make sense to compare subpopulations directly (rather than via the individual subpopulations' deviations from the full population). An example is comparing the subpopulation of individuals who received a certain medical treatment to the untreated subpopulation.

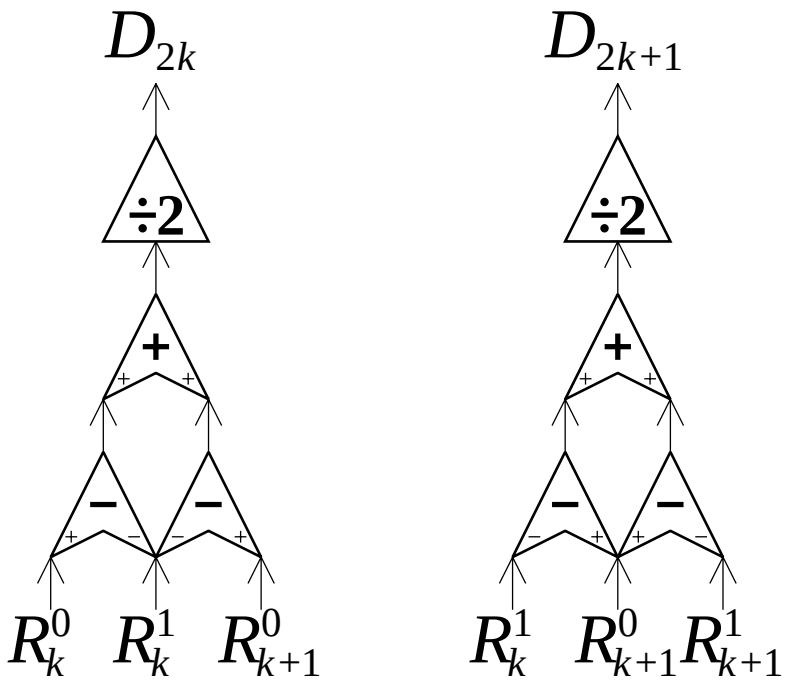
# Finest-possible binning




1. The crosses ("x") indicate the scores for subpopulation 0 while the circles ("o") indicate the scores for subpopulation 1.
2. The averages of the scores for subpopulation $k$ corresponding to the indicated blocks of observed scores are $S_0^k$, $S_1^k$, ..., $S_9^k$, for $k = 0, 1$.
3. The averages of the responses for subpop. $k$ corresponding to the indicated blocks of observed scores are $R_0^k$, $R_1^k$, ..., $R_9^k$, for $k = 0, 1$.

# Average forward and backward differences

Define $D_{2k}$ to be the average of the even indexed forward and backward differences and $D_{2k+1}$ to be the average of the odd indexed forward and backward differences:



# Accumulation of averaged differences

Define the cumulative differences

$$C_j = \frac{1}{n} \sum_{k=0}^{j-1} D_k \tag{13}$$

for $j = 1, 2, \ldots, n$.

The graph of $C_k$ versus $k/n$ and the metrics $C_{\mathrm{MAD}}$ and $C_{\mathrm{range}}$ — defined via the same formulas (3) and (4) as on slide 13 above — admit interpretations analogous to those discussed on earlier slides for the related settings treated there.

# Controlling for multiple covariates

## Controlling for multiple covariates

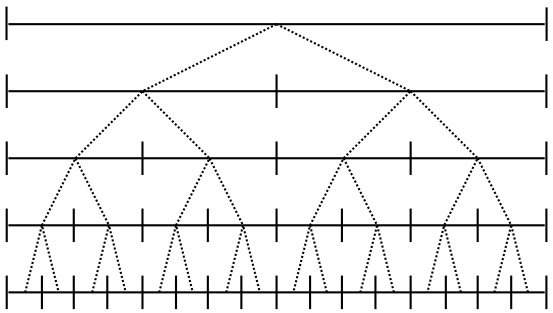
# Multiple covariates require multiple dimensions

1. "Covariates" is yet another name for the "input variables," "independent variables," or "control variables" in a regression.
2. Earlier slides considered only a single real-valued covariate; the scores specified the values of the covariate.
3. When there are multiple real-valued covariates, their values fill a multidimensional vector space; denoting by $d$ the number of scalar covariates, the vector space of all their possible values will be $d$-dimensional.
4. There is a natural notion of locality in $d$ dimensions that looks over all possible length scales, hierarchically organizing $d$-dimensional space into the canonical "dyadic tree" defined on the following slide.

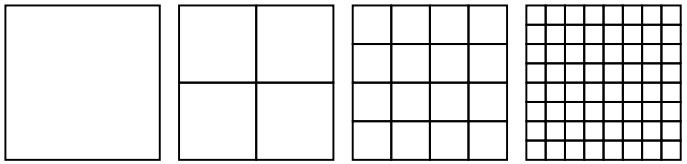
# Dyadic trees

1. The dyadic tree in a $d$-dimensional vector space is the partition of space obtained by recursively subdividing equally along every coordinate axis during every split.
2. In one dimension, the canonical dyadic tree is the balanced binary tree. In two dimensions, the canonical dyadic tree is the canonical quad tree. In three dimensions, the canonical dyadic tree is the canonical oct tree.

five levels of the canonical binary tree:

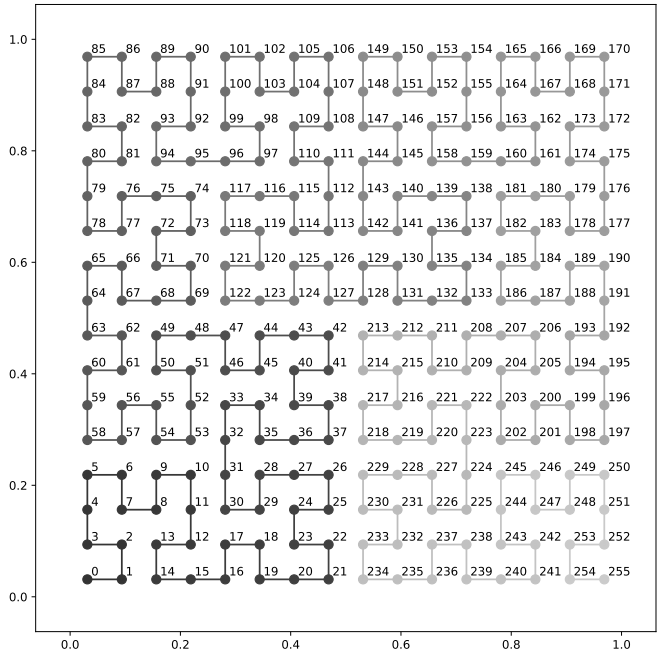


four levels of the canonical quad tree:



# Hilbert space-filling curves

The Hilbert curve puts a total order on the unit hypercube $[0, 1]^d$ via the depth-first traversal of the canonical $2^d$-ary (dyadic) tree. With $d = 2$, an approximation with 255 line segments is

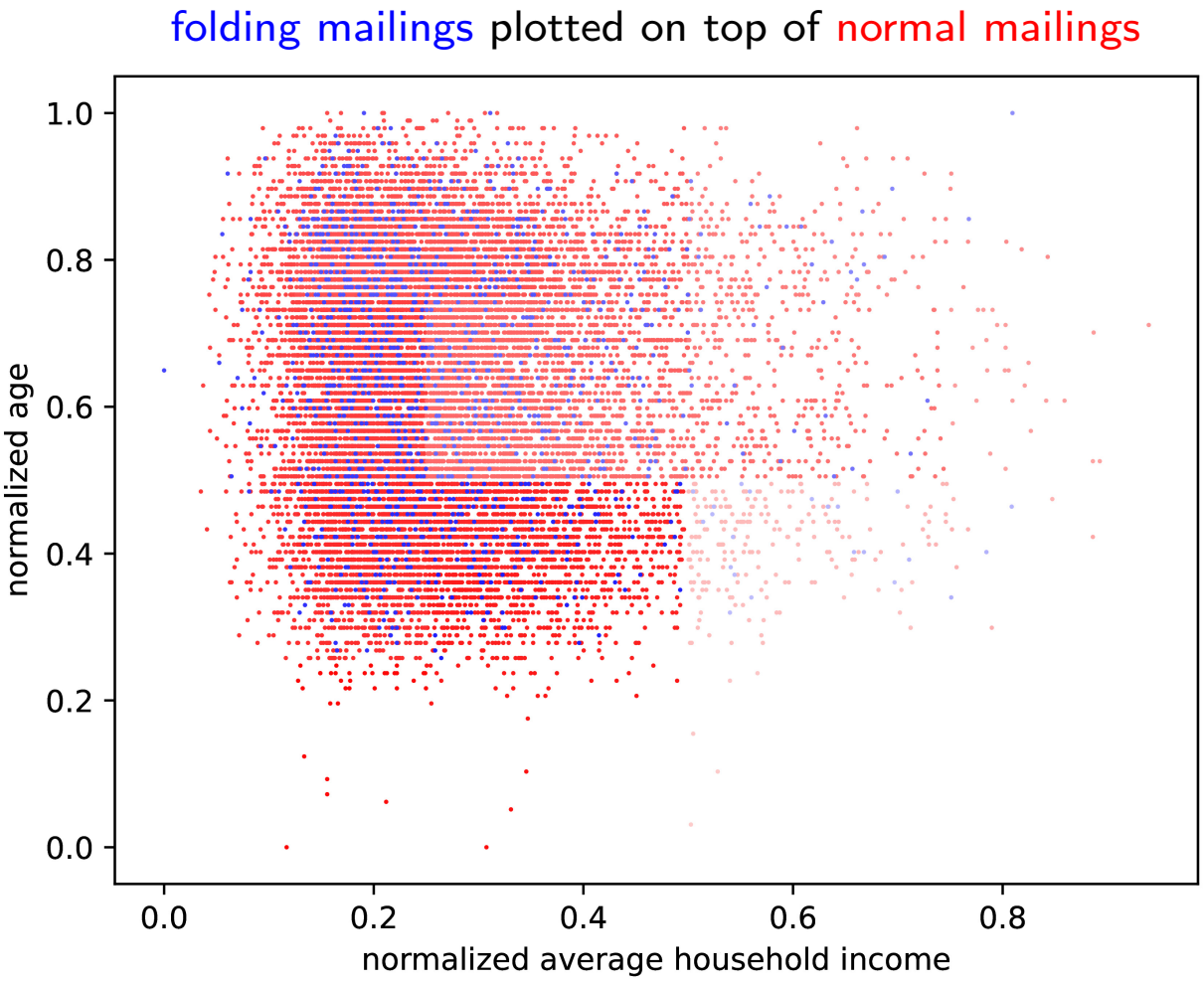


# Reduction to one dimension

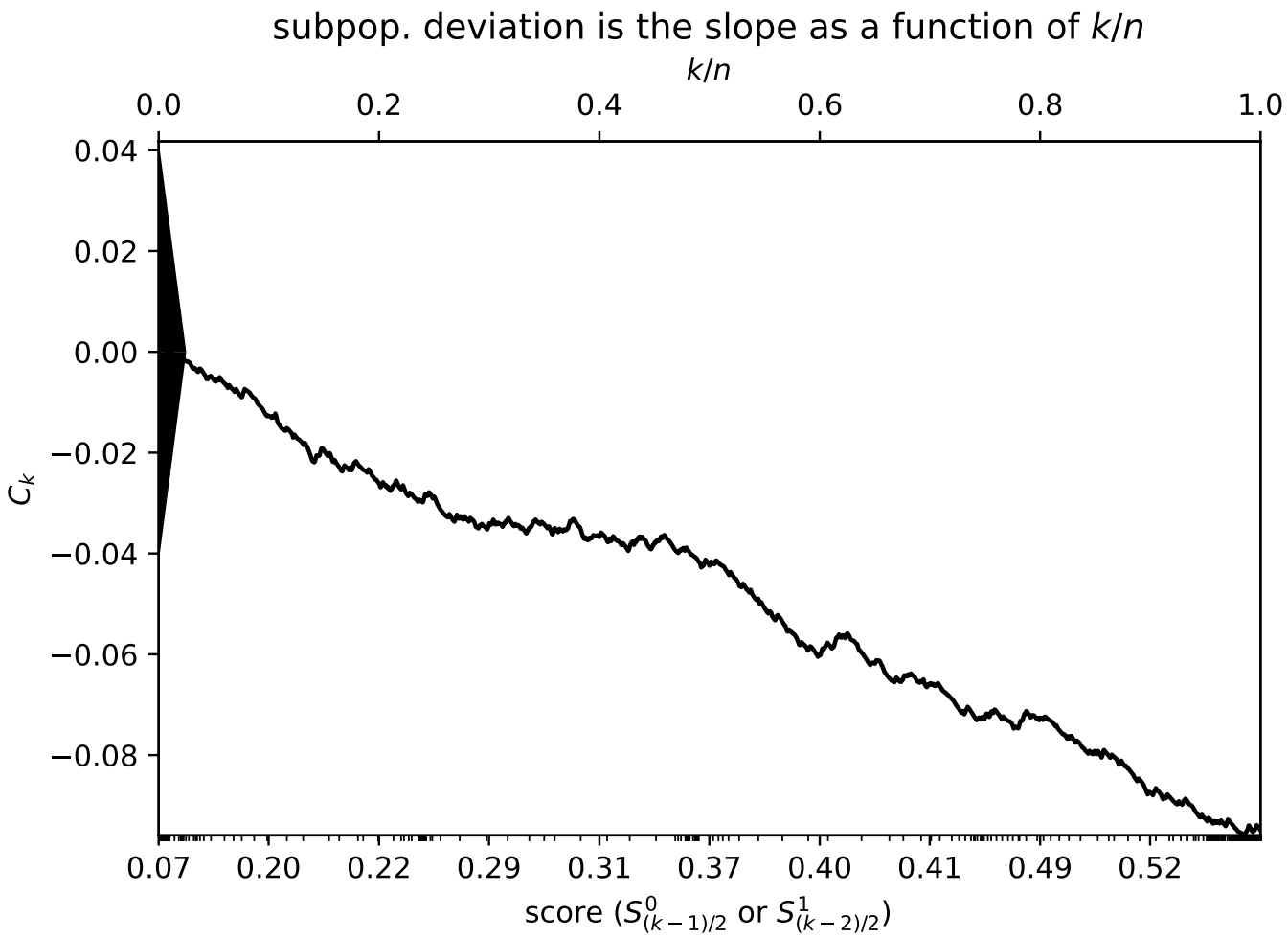
A space-filling curve $h$ (e.g., Hilbert's) maps continuously from the unit interval $[0, 1]$ onto (surjectively) the unit hypercube $[0, 1]^d$. This ensures that, given a function $f$ on the unit hypercube $[0, 1]^d$, local averages of the composition $f \circ h$ are also local averages of $f$; as usual, the definition of the composition is $(f \circ h)(t) = f(h(t))$ for all $t$ in the unit interval $[0, 1]$.

Given points in the unit hypercube $[0, 1]^d$, use the Hilbert curve to map each of the points to a score in the unit interval $[0, 1]$, then use the methods of the earlier slides with these scalar scores.

# Hilbert-curve score for folding or normal as brightness



# Cumulative diffs. between folding and normal mailings



# Description of the data (skip these details)

1. In 1994, a national veterans organization mailed folding cards to $n_0 = 1{,}236$ prospective donors and mailed normal cards to $n_1 = 15{,}866$.
2. The responses (indicating whether the recipients responded to the solicitation or not), together with covariates such as age and average household income in the associated Census block, formed the core of the data for the 1998 KDD Cup.
3. The combined sample size derived from the diagram on slide 41 is $n = 2{,}279$ (rather than the $n = 20$ on slide 41).

# Extensions

# Extensions

# Weighted sampling

## Weighted sampling

# Graphs for weighted samples (skip if low on time)

All methods generalize to weighted sampling; in weighted samples each pair of score $S_k$ and response $R_k$ comes with a positive real number $W_k$. The ordinates (vertical coordinates) become

$$C_j = \frac{\sum_{k=1}^{j} (R_{i_k} - \tilde{R}_{i_k})\, W_{i_k}}{\sum_{k=1}^{n} W_{i_k}} \tag{14}$$

(with $\tilde{R}_{i_k} := S_{i_k}$ and $i_k := k$ when assessing calibration) and the abscissae (horizontal coordinates) become

$$A_j = \frac{\sum_{k=1}^{j} W_{i_k}}{\sum_{k=1}^{n} W_{i_k}} \tag{15}$$

for $j = 1, 2, \ldots, n$.

# P-values for weighted samples (skip if low on time)

1. The interpretation of slope in the graph of $C_k$ versus $A_k$ is the same as in the case of unweighted sampling.
2. P-values for the maximum absolute value of $C_1, C_2, \ldots, C_n$ and for the range of $C_0, C_1, \ldots, C_n$ arise similarly, courtesy of convergence in distribution to the maximum absolute value or range of the standard Brownian motion over $[0, 1]$, as before.
3. For example, formula (6) on slide 16 above, giving the standard deviation under the null hypothesis of perfect calibration (so that $\tilde{R}_{i_k} := S_{i_k}$ and $i_k := k$), becomes

$$\sigma = \frac{\sqrt{\sum_{k=1}^{n} S_k(1 - S_k)\,(W_k)^2}}{\sum_{k=1}^{n} W_k}. \tag{16}$$

# Avoiding randomization

# Avoiding randomization

# Slight random perturbations (skip if low on time)

1. If any of the inequalities $S_1 \leq S_2 \leq \cdots \leq S_n$ were not strict, one possibility would be to perturb the scores at random such that the scores become all distinct from each other, so that $S_1 < S_2 < \cdots < S_n$. Slide 7 suggested this approach.
2. To describe another possibility, we can keep the notation $S_1 < S_2 < \cdots < S_n$ while denoting the responses and weights for $n_k$ repetitions of score $S_k$ by $R_k^{(1)}, R_k^{(2)}, \ldots, R_k^{(n_k)}$ and $W_k^{(1)}, W_k^{(2)}, \ldots, W_k^{(n_k)}$ (so then the total number of observations is $\sum_{k=1}^{n} n_k$).

# A new data set (skip if low on time)

We construct a new data set, with the response being the weighted average

$$R_k = \frac{\sum_{j=1}^{n_k} R_k^{(j)} W_k^{(j)}}{\sum_{j=1}^{n_k} W_k^{(j)}} \tag{17}$$

and the weight being the sum

$$W_k = \sum_{j=1}^{n_k} W_k^{(j)} \tag{18}$$

for $k = 1, 2, \ldots, n$.

# Cumulative differences stay the same (skip if low on time)

1. The cumulative differences for the new data set are

$$B_\ell = \frac{\sum_{k=1}^{\ell} (R_k - r(S_k))\, W_k}{\sum_{k=1}^{n} W_k} \tag{19}$$

for $\ell = 1, 2, \ldots, n$, with $r(S_k) = S_k$ for gauging calibration.

2. The cumulative differences for the original data set with the scores perturbed infinitesimally at random are

$$C_\ell = \frac{\sum_{k=1}^{\ell} \sum_{j=1}^{n_k} \left(R_k^{(j)} - r(S_k)\right) W_k^{(j)}}{\sum_{k=1}^{n} \sum_{j=1}^{n_k} W_k^{(j)}} \tag{20}$$

for $\ell = 1, 2, \ldots, n$, where the responses $R_k^{(1)}, R_k^{(2)}, \ldots, R_k^{(n_k)}$ & weights paired with them are randomly permuted for each $k$.

3. So $B_\ell = C_\ell$ for all $\ell = 1, 2, \ldots, n$, due to the previous slide.

# Ties in scores can be treated as weighted samples

The aggregated abscissae (horizontal coordinates) are

$$A_\ell = \frac{\sum_{k=1}^{\ell} W_k}{\sum_{k=1}^{n} W_k} = \frac{\sum_{k=1}^{\ell} \sum_{j=1}^{n_k} W_k^{(j)}}{\sum_{k=1}^{n} \sum_{j=1}^{n_k} W_k^{(j)}} \tag{21}$$

for $\ell = 1, 2, \ldots, n$.

Thus, the horizontal coordinates in the cumulative graphs for the original and new data sets are the same, as are the vertical coords. (as shown on the previous slide), at least when the graph consists of straight line segments connecting $(A_{k-1}, B_{k-1}) = (A_{k-1}, C_{k-1})$ to $(A_k, B_k) = (A_k, C_k)$ for $k = 1, 2, \ldots, n$, with $A_0 = B_0 = C_0 = 0$. The original data set's cumulative graph which interpolates linearly from every perturbed score to the next greatest perturbed score will be almost the same, aside from having slightly larger random excursions between scores that were the same prior to perturbation.

# Advantages and disadvantages (skip if low on time)

| | randomized | non-randomized |
|---|---|---|
| pros | displays every single member of the original data set | horizontal axis has no randomization |
| cons | horizontal axis is randomized to some extent (the ordering for repeated scores must be randomized, obviously) | displays only averaged responses for repeated (degenerate) scores |

# Interactive plots

Interactive plots

# Interactive traversal of the Hilbert curve

Census Bureau's weighted sampling of counties in California: the county is the subpopulation of the full state pop.; responses are given by "Variable".

(Click to play the movie.)

# Concluding thoughts

# Concluding thoughts

# Ideal limit of the empirical Kuiper metric

Given 3 real-valued random variables, $X$, $Y$, $Z$, the Kuiper metric is the following summary stat of differences between the regression $\mathbb{E}[Y|X]$ of $Y$ on $X$ and the regression $\mathbb{E}[Z|X]$ of $Z$ on $X$; this stat compares $Y$ and $Z$ while controlling for $X$ (say $X$ is age or income):

$$\max_{-\infty \leq a \leq b \leq \infty} \Big| \mathbb{E}_{a \leq X \leq b} \big[ \mathbb{E}[Y|X] - \mathbb{E}[Z|X] \big] \Big|, \tag{22}$$

where $\mathbb{E}_{a \leq X \leq b} \big[ \mathbb{E}[Y|X] - \mathbb{E}[Z|X] \big]$

$$= \mathbb{E}\Big[ \big( \mathbb{E}[Y|X] - \mathbb{E}[Z|X] \big) \cdot \mathbb{1}\{a \leq X \leq b\} \Big], \tag{23}$$

$$\mathbb{1}\{a \leq X \leq b\} = 1 \text{ when } a \leq X \leq b, \text{ and} \tag{24}$$

$$\mathbb{1}\{a \leq X \leq b\} = 0 \text{ when } X < a \text{ or } X > b. \tag{25}$$

Calibration is the special case with $Z = X$ (making $\mathbb{E}[Z|X] = X$).

# Matched-pair analysis

The simplest, most common paired samples consist of observations from two subpopulations, with each observed response from one subpopulation corresponding to an observed response from the other subpopulation at the same value(s) of the covariate(s). Such a pair of observed responses (one from each subpopulation) at the same value(s) of the covariate(s) is known as a "matched pair," with the matching based on the value(s) of the covariate(s). Generalization of all of this tutorial's cumulative methodologies to the analysis of paired samples should be obvious; full details are in the paper available at https://arxiv.org/abs/2305.11323

# Open-access articles and open-source software

Open-access references are available at the top of
http://tygert.com/research.html

Open-source software is available at the bottom of
http://tygert.com/software.html

# Randomized controlled trials and A/B tests

The cumulative approach is relevant for analyzing the treatment effect in any randomized controlled trial. Conditioning on covariates traditionally can lead to issues with Simpson's Paradox, due to the need to choose the sizes of the bins for binning ("binning" is also known as "bucketing," "segmenting," or "stratifying"). Cumulative stats avoid Simpson's Paradox entirely.

A/B tests are a form of randomized controlled trial that is especially popular in the data-driven industries.

# Observational studies

Observational studies can also benefit from cumulative statistics. "Adjusting" for covariates is common in such studies. Adjustment often involves conditioning on (controlling for) covariates. Conditioning on confounding covariates facilitates causal or counterfactual analysis.

# Conclusion

1. Kolmogorov, Kuiper, Smirnov, and Wiener had good reason to introduce cumulative statistics, and Hilbert and Peano had good reason to introduce their space-filling curves.
2. Classical non-cumulative statistics have an explicit trade-off between statistical confidence and resolving power (resolution of variations as a function of score). The classical metrics and reliability diagrams are typically misleading (or manipulated) due to the strong dependence on the arbitrary choice of bins.
3. Meta's Fairness Flow promotes the cumulative metrics as the preferred method for detecting differences while controlling for variates, thanks to the hard work of the Responsible AI team.
4. There is plenty more to come in this space — stay tuned!