Histograms Revisited: When are histograms the best approximation method for aggregates over joins?

Alin Dobra
Department of Computer & Information Science & Engineering
University of Florida
Gainesville, FL 32611, USA
adobra@cise.ufl.edu

ABSTRACT

The traditional statistical assumption for interpreting histograms and justifying approximate query processing methods based on them is that all elements in a bucket have the same frequency – the so called uniform distribution assumption. In this paper we show that a significantly less restrictive statistical assumption – the elements within a bucket are randomly arranged even though they might have different frequencies – leads to identical formulae for approximating aggregate queries using histograms. This observation allows us to identify scenarios in which histograms are well suited as approximation methods – in fact we show that in these situations sampling and sketching are significantly worse – and provide tight error guarantees for the quality of approximations. At the same time we show that, on average, histograms are rather poor approximators outside these scenarios.

1. INTRODUCTION

Histograms are among the most widely used and extensively studied approximation technique for aggregate queries [6, 5, 8]. The traditional interpretation of histograms – irrespective of the type – is that the frequencies of items in a bucket are approximated by the average frequency of the bucket and will be used instead of the original frequencies in any computation, for example to estimate the size of the join of two relations F and G. While this interpretation is intuitive and provides simple recipes for performing operations with histograms, it suggests that the histogram approximation of the frequency distribution will work well in the approximation process only if the initial frequency distribution is *smooth* and can be locally approximated using the uniform distribution assumption of histograms. This, as suggested by the following example and apparent from the histogram literature, turns out not to be strictly necessary; histograms might work well even when the average frequency in a bucket is a very rough approximation of the actual frequency.

PODS 2005 June 13-15, 2005, Baltimore, Maryland.

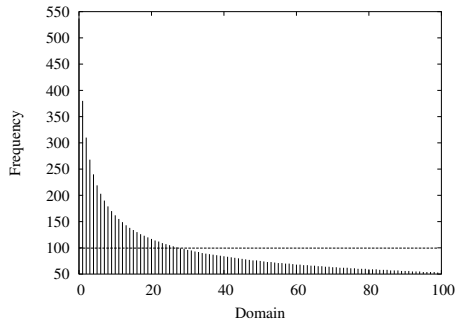

Figure 1: Frequency graph of relation F.

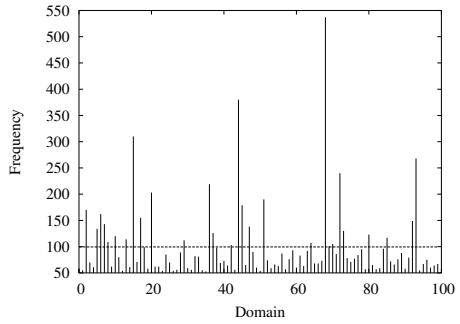

Figure 2: Frequency graph of relation G.

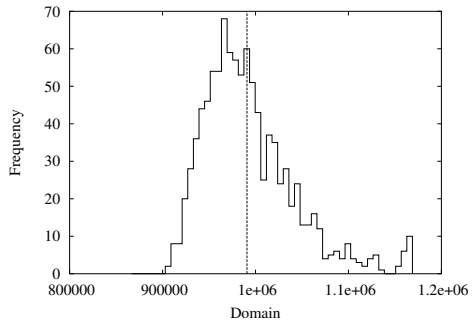

Figure 3: Histogram of the size of the join result

EXAMPLE 1. *Let F and G be two relations, each with a single attribute A with domain $1 \ldots 100$. We generate both F and G to have Zipf distributions with Zipf coefficient 0.5 and with average frequency 99.54 (as close to 100 but allowing all frequencies to be integers). In relation F the frequency is decreasing (see Figure 1); in relation G the frequencies are randomly shuffled (see Figure 2). Observe from these figures that the one-bucket histogram approximation of the frequency, the line at 99.54, is a very poor estimate of the frequencies, thus we would expect poor performance when we use one-bucket histograms to estimate the size of the equi-join of F and G. In the scenario described above, the one bucket histogram prediction is always 990821 irrespective of the shuffling of the domain of G. In the particular case depicted in the figure, the true size of the join is 981565, a mere 1% smaller than the prediction; to make sure this is not happening by chance, we picked 1000 random shufflings of G and plotted the distribution of $|F \bowtie_A G|$ in Figure 3. Notice that the sizes of the joins are compactly distributed (within a 10% relative error) around the prediction using the one-bucket histograms.*

As the previous example suggests, even though within a bucket the uniform frequency approximation is rather crude, the result of approximating the size of the join is surprisingly good and statistically stable. This prompts the question: Why is this happening in spite of the uniform approximation not holding? As we will show in this paper, the answer is that a more general statistical hypothesis holds, namely that the placement of the frequencies in the two relations is *uncorrelated*. This observation is the starting point for the current work and allows us to make the following contributions with this paper:

- We formulate a **new statistical assumption**, random shuffling of frequencies within a bucket, that is more general, thus more likely to hold, than the uniform frequency assumption. As we will show, this new assumption does not change the way the histograms are used for approximating results of queries – thus it is consistent with all the previous work on histograms – but, important from a practitioner's point of view, explains why and when histograms behave well as approximators. Statistically, random shuffling assumption holds when there is no correlation between the frequencies in the two relations being joined, so it is likely to hold in practice quite often.

- We provide **tight error guarantees** for unidimensional histograms when the random shuffling assumption holds. [6] is the only other work that provides tight error guarantees for estimation using histograms, in this case worst case guarantees. The errors we derive allow us to provide theoretical proof that, when the random shuffling assumption holds, histograms are strictly superior to *sampling* and *sketching*. At the same time, we provide compelling theoretical evidence that, when the random shuffling assumption does not hold, histograms are, on average, poor approximators when compared to sampling and sketching.

In the rest of the paper, we first formalize the problem we are solving in Section 2, followed by the formalization of the random shuffling assumption in Section 3. We analyze the behavior of histograms under the random shuffling assumption in Section 4 and compare with the behavior of sampling and sketching in Section 5. In Section 6 we analyze the behavior of histograms when the random shuffling assumption does not hold, then comment on related work in Section 7 and conclude in Section 8.

2. PROBLEM FORMULATION

The general problem we are trying to solve is *approximating aggregates over joins*. As we will show in this section, both the selectivity estimation and the general SUM-like aggregates over join problems can be rephrased as *size of join estimation* problems. Thus all the results developed for the latter can be extended straightforwardly to the other two problems. In this paper we focus on unidimensional problems, postponing the generalization to multiple dimensions for later work.

Symbol(s)	Meaning
F, G	Relations
A	Join Attribute
I	Domain of join attribute A
N	Size of domain I
i, j, i', j'	Indices going over domain I
f_i, g_i	Frequencies of value i in relations F, G
$\overline{f}, \overline{g}$	Average frequencies in relations F, G
$\mathrm{SJ}(F)$	$\sum_{i \in I} f_i^2$, the self join size of F
$\mathrm{SqErr}(F)$	$\sum_{i \in I} (f_i - \overline{f})^2$, the squared error of F
I_l	The lth bucket of domain I
$\mathcal{I}(C)$	Identity function: 1 when C true, 0 otherwise
σ, γ	Uniform random permutations
X	Random variable modeling the query result
X_l	Random variable modeling the contribution to the result of bucket l
n	Number of buckets
N'	Size of the sample
$P[p]$	Probability that predicate p holds
$E[X]$	Expected value of random variable X
$\mathrm{Var}(X)$	Variance of random variable X, $\mathrm{Var}(X) = E\left[X^2\right] - E[X]^2$
$\mathrm{Cov}(X, Y)$	Covariance of random variables X and Y $\mathrm{Cov}(X, Y) = E[XY] - E[X]\,E[Y]$

Table 1: Notation used in the paper.

2.1 Size of Join Problem

Let F and G be two relations, each with a single attribute A with domain I. Furthermore, let f_i and g_i be the frequency of the value i in F and G, respectively. With this, the *size of join problem* is to estimate the quantity:

$$|F \bowtie_A G| = \sum_{i \in I} f_i g_i \tag{1}$$

given synopses of relations F and G (if full information is available, we can simply compute the sum to get the exact answer).

2.2 Selectivity Estimation Problem

Let G be a relation with a single attribute A that has the domain I. Given I', a subset of I, the selectivity estimation problem is to estimate the quantity:

$$|\sigma_{I'}(G)| = \sum_{i \in I'} g_i \tag{2}$$

where g_i is the frequency of value i in G.

With $\mathcal{I}(C)$ the identity function, that takes value 1 if condition C is true and value 0 otherwise, by simply setting the relation F so that $f_i = \mathcal{I}(i \in I')$ we have:

$$\begin{aligned} |F \bowtie_A G| &= \sum_{i \in I} f_i g_i \\ &= \sum_{i \in I} \mathcal{I}(i \in I') g_i \\ &= \sum_{i \in I'} g_i \\ &= |\sigma_{I'}(G)| \end{aligned}$$

Thus, by choosing appropriate relation F we can estimate the selectivity of relation G. These ideas straightforwardly extend to arbitrary selection predicates on G that depend on multiple attributes. The actual generalization is omitted here due to lack of space.

2.3 Aggregates over Joins Problem

Let F and G be two relations that contain a join attribute A and possibly other attributes. Let us first look at aggregates of the from

$$\text{SUM}_{\mathcal{F}_F \mathcal{F}_G}(F \bowtie_A G) = \sum_{t \in F \bowtie_A G} \mathcal{F}_F(t \perp F) \mathcal{F}_G(t \perp G)$$

with $t \perp F$ the part of the tuple in the join that comes from relation F (similarly for G) and $\mathcal{F}_F$ and $\mathcal{F}_G$ arbitrary functions. The only requirement for this to work is to be able to rewrite the expression summed up over the join as the product of expressions depending on attributes of the two relations. To evaluate such expressions we observe that:

$$\begin{aligned} &\text{SUM}_{\mathcal{F}_F \mathcal{F}_G}(F \bowtie_A G) \\ &\quad = \sum_{t \in F \bowtie_A G} \mathcal{F}_F(t \perp F) \mathcal{F}_G(t \perp G) \\ &\quad = \sum_{i \in I} \sum_{t \in F \bowtie_A G, t.A = i} \mathcal{F}_F(t \perp F) \mathcal{F}_G(t \perp G) \\ &\quad = \sum_{i \in I} \left(\sum_{t \in F, t.A = i} \mathcal{F}_F(t) \right) \left(\sum_{t \in G, t.A = i} \mathcal{F}_G(t) \right) \\ &\quad = \sum_{i \in I} \tilde{f}_i \tilde{g}_i \end{aligned} \tag{3}$$

where $\tilde{f}_i$ and $\tilde{g}_i$ are just compact notation for expressions $\sum_{t \in F, t.A = i} \mathcal{F}_F(t)$ and $\sum_{t \in G, t.A = i} \mathcal{F}_G(t)$, respectively. The important observation is that we can use any method designed for size of join estimation to estimate this aggregate as well by simply replacing f_i by $\tilde{f}_i$ and g_i by $\tilde{g}_i$ since then the expression in Equation 1 is identical to the last expression in Equation 3. Thus, computing such aggregates is as easy as computing sizes of joins; the complexity is in the join, and not the expression being summed up.

With the ability to compute estimates of aggregates of the form $\text{SUM}_{\mathcal{F}_F \mathcal{F}_G}(F \bowtie_A G)$, we can immediately compute aggregates of the form AVG and STD as well. For example, to estimate $\text{AVG}_B(F(A, B) \bowtie_A G(A))$ we can estimate $\text{SUM}_B(F(A, B) \bowtie_A G_A)$ and $|F \bowtie_A G|$ and simply take their ratio.

Since both the selectivity estimation and COUNT, SUM, AVG and STD aggregate estimation problems can be reduced to size of join problems, for the rest of the paper we will focus only on the size of join problem.

2.4 Comments on Obtaining Error Guarantees from Expected Value and Variance Estimates

The standard techniques [11, 2] to obtain error guarantees, i.e. confidence intervals, for an estimate is to compute the expected value and variance and then to use either distribution independent bounds given by Chernoff's and Chebyshev's inequalities, or to use distribution dependent bounds. In the latter case, usually the Central Limit Theorem or one of its generalizations is used to argue that the distribution of the estimate is close to normal and then error bounds based on normal distributions with the same expected value and variance are produced. For all the estimates in this paper, either distribution independent bounds could be used to obtain strict characterization of the results or the normal distribution based bounds since all estimates can be expressed as weighted sums of independent identically distributed (iid) random variables so the Central Limit Theorem applies.

In view of the above discussion, in order to simplify the exposition and the comparison, throughout the paper we will just provide results in the form of expected values and variances or squared errors – the variance is equal to the squared error if the random variable is unbiased. Actual error guarantees can be obtained straightforwardly using the above mentioned techniques.

3. RANDOM SHUFFLING ASSUMPTION

Example 1 suggests that the behavior of histograms is interesting when the placement of items in relations F and G are uncorrelated – more precisely, the value of the frequency of the item is independent of the position of the item in the bucket. In order to make any kind of theoretical analysis, we have to build a statistical model that will have exactly this property. As opposed to samples[9] and sketches[2], histograms are not randomized methods. In order to model the lack of correlation between join attributes, we have to set and analyze a probabilistic space over problems and keep the solution fixed. In order to do this, we start with the uniformly random space of permutations σ over the domain of join attributes I, i.e. any of the $N!$ permutations of I is equally probable, where $N = |I|$ is the size of I. Denoting by $\sigma(i)$ the value in position i of the σ permutation of I, the *random shuffling assumption* for a relation F states that the frequency of item i is $f_{\sigma(i)}$, a random variable. We will denote probabilities in this space by P_σ and expectations by E_σ.

The following properties of the probability space set up

above are important:

$$\forall i, i' \in I \quad P_\sigma\left[\sigma(i) = i'\right] = \frac{1}{N} \tag{4}$$

$$\forall i, j, i' \in I, i \neq j \quad P_\sigma\left[\sigma(i) = i' \wedge \sigma(j) = i'\right] = 0 \tag{5}$$

$$\forall i, j, i', j' \in I, i \neq j, i' \neq j', \quad P_\sigma\left[\sigma(i) = i' \wedge \sigma(j) = j'\right] = \frac{1}{N}\frac{1}{N-1} \tag{6}$$

4. HISTOGRAMS UNDER RANDOM SHUFFLING ASSUMPTION

In this section we start the analysis of histograms by looking at their performance when the random shuffling assumption holds. We start with the simplest case, histograms with a single bucket, and work our way up to the fully general case. Throughout this section we use the notation in Table 1.

4.1 One Bucket Histograms

In this section we provide theoretical analysis of the unidimensional size of join problem, as stated in Section 2.1, under random shuffling assumption. As pointed out in Section 2, these results readily extend to selectivity estimation and other aggregates over joins.

PROPOSITION 1. *Let σ be a uniformly random permutation of domain I of size $N = |I|$. Then, for any $i \in I$ and any frequencies of items in relation G, $\{g_j | j \in I\}$ we have:*

$$E_\sigma\left[g_{\sigma(i)}\right] = \overline{g}$$

$$Var_\sigma(g_{\sigma(i)}) = Cov_\sigma(g_{\sigma(i)}, g_{\sigma(i)}) = \frac{SqErr(G)}{N}$$

$$i \neq i', \; Cov_\sigma(g_{\sigma(i)}, g_{\sigma(i')}) = -\frac{SqErr(G)}{N(N-1)}$$

PROOF.

$$\begin{aligned} E_\sigma\left[g_{\sigma(i)}\right] &= \sum_{i' \in I} g_{i'} P_\sigma\left[\sigma(i) = i'\right] \\ &= \frac{1}{N}\sum_{i' \in I} g_{i'} \\ &= \overline{g} \end{aligned}$$

where we used the definition of expectation and Equation 4. Similarly, we have:

$$\begin{aligned} E_\sigma\left[g^2_{\sigma(i)}\right] &= \sum_{i' \in I} g^2_{i'} P_\sigma\left[\sigma(i) = i'\right] \\ &= \frac{1}{N}\sum_{i' \in I} g^2_{i'} \\ &= \mathrm{SJ}(G)\frac{1}{N} \end{aligned}$$

and with this,

$$\begin{aligned} \mathrm{Var}_\sigma(g_{\sigma(i)}) &= E_\sigma\left[g^2_{\sigma(i)}\right] - E_\sigma\left[g_{\sigma(i)}\right]^2 \\ &= \frac{\mathrm{SJ}(G)}{N} - \overline{g}^2 \\ &= \frac{\mathrm{SqErr}(G)}{N} \end{aligned}$$

where we used the fact that $\mathrm{SqErr}(G) = \mathrm{SJ}(G) - N\overline{g}^2$

For $i \neq i'$ we have:

$$\begin{aligned} E_\sigma\left[g_{\sigma(i)} g_{\sigma(i')}\right] &= \sum_{j \in I}\sum_{k \in I} g_j g_k P_\sigma\left[\sigma(i) = j \wedge \sigma(i') = k\right] \\ &= \frac{1}{N(N-1)}\sum_{j \in I}\sum_{k \in I, k \neq j} g_j g_k \\ &= \frac{1}{N(N-1)}\left[\sum_{j \in I} g_j \sum_{k \in I} g_k - \sum_{j \in I} g_j^2\right] \\ &= \overline{g}^2\frac{N}{N-1} - \mathrm{SJ}(G)\frac{1}{N(N-1)} \end{aligned}$$

where we used Equations 5 and 6 to get the second line and the fact that $N\overline{g} = \sum_{i \in I} g_i$ and definition of $\mathrm{SJ}(G)$. With this we have:

$$\begin{aligned} \mathrm{Cov}_\sigma(g_{\sigma(i)}, g_{\sigma(i')}) &= E_\sigma\left[g_{\sigma(i)} g_{\sigma(i')}\right] - E_\sigma\left[g_{\sigma(i)}\right] E_\sigma\left[g_{\sigma(i')}\right] \\ &= \overline{g}^2\frac{N}{N-1} - \frac{\mathrm{SJ}(G)}{N(N-1)} - \overline{g}^2 \\ &= \frac{N\overline{g}^2 - \mathrm{SJ}(G)}{N(N-1)} \\ &= -\frac{\mathrm{SqErr}(G)}{N(N-1)} \end{aligned}$$

□

With this result we obtain the following:

LEMMA 1. *Let σ be a uniformly random permutation of domain I of size $N = |I|$. Let F and G be two relations with a common attribute A with domain I and frequencies f_i and g_i, respectively. Define the random variable over the space of permutations as $X = \sum_{i \in I} f_i g_{\sigma(i)}$, the size of join of F and $\sigma(G)$. Then:*

$$E_\sigma[X] = N\overline{f}\,\overline{g}$$

$$Var_\sigma(X) = \frac{1}{N-1} SqErr(F) SqErr(G)$$

PROOF.

$$\begin{aligned} E_\sigma[X] &= \sum_{i \in I} f_i E_\sigma\left[g_{\sigma(i)}\right] \\ &= \overline{g}\sum_{i \in I} f_i \\ &= N\overline{f}\,\overline{g} \end{aligned}$$

where we used the linearity of expectation and first result in Proposition 1.

$$\begin{aligned}\mathrm{Var}_\sigma(X) &= \mathrm{Var}_\sigma(\sum_{i\in I} f_i g_{\sigma(i)}) \\ &= \sum_{i\in I}\sum_{i'\in I} f_i f_{i'} \mathrm{Cov}_\sigma(g_{\sigma(i)}, g_{\sigma(i')}) \\ &= \sum_{i\in I} f_i^2 \frac{\mathrm{SqErr}(G)}{N} - \sum_{i\in I}\sum_{i'\in I, i'\neq i} f_i f_{i'} \frac{\mathrm{SqErr}(G)}{N(N-1)} \\ &= \frac{\mathrm{SqErr}(G)}{N}\mathrm{SJ}(F) - (N^2\overline{f}^2 - \mathrm{SJ}(F))\frac{\mathrm{SqErr}(G)}{N(N-1)} \\ &= \frac{\mathrm{SqErr}(G)}{N-1}\left(\frac{N-1}{N}\mathrm{SJ}(F) - N\overline{f}^2 + \frac{1}{N}\mathrm{SJ}(F)\right) \\ &= \frac{\mathrm{SqErr}(F)\mathrm{SqErr}(G)}{N-1}\end{aligned}$$

□

We can now prove the following:

THEOREM 1. *Let F and G be two relations. The estimate $N\overline{f}\overline{g}$ of $|F \bowtie_A G|$ using unidimensional histograms for both F and G under the random shuffling assumption is an* unbiased estimator *and has squared error*

$$\frac{1}{N-1} SqErr(F) SqErr(G)$$

PROOF. The proof is direct using Lemma 1 by observing that random variable X models the size of the join under the random shuffling assumption. □

What this theorem tells us is that the way we use one-bucket histograms is consistent with the random shuffling hypothesis and that we should expect the estimate to be very good if the hypothesis holds since the variance is inversely proportional to the size of the domain minus one.

For the scenario in Example 1, the variance of the prediction using the formula in Theorem 1 is $2.53 \cdot 10^9$ and the standard deviation is 50309 – remember that the true result is 985392. Given the fact that most of the mass of a normal distribution is within two standard deviations, using the theory we just developed we would expect the error, in most of the cases, to be at most $2 \cdot 50308/985392 \approx 10\%$ which coincides with the empirical observations based on Figure 3 we made in Section 1.

4.2 Histograms with Aligned Buckets

Usually, we can afford to build histograms with hundreds of buckets, not only one bucket. The natural question to ask is how can we extend the analysis in the previous section to multi-bucket histograms.

LEMMA 2. *Let $I_1, \ldots, I_n$ be an arbitrary bucketization of domain I into n buckets, i.e. $I_1, \ldots, I_n$ form an arbitrary partition of I. If the random shuffling assumption holds independently within each of the buckets for relation G then, by defining random variable $X = \sum_{i\in I} f_i g_{\sigma(i)}$, we have:*

$$E_\sigma[X] = \sum_{l=1}^{n} |I_l| \overline{f}_l \overline{g}_l$$

$$Var_\sigma(X) = \sum_{l=1}^{n} \frac{SqErr(F_l) SqErr(G_l)}{|I_l| - 1}$$

where $\overline{f}_l$ and $\overline{g}_l$ are the average frequencies in bucket I_l of F and G, respectively, and $SqErr(F_l)$ and $SqErr(G_l)$ are the squared errors of F and G, respectively, in bucket I_l.

PROOF. We observe that X can be rewritten as $X = \sum_{l=1}^{n} X_l$ where $X_l = \sum_{i\in I_l} f_i g_{\sigma(i)}$. Since the random shuffling assumption holds independently in each bucket, random variables X_l are independent and the probability space over σ is simply the product of probability spaces, one for each bucket. Since each random variable X_l behaves now exactly like the setup in Lemma 1 (i.e. as in the one bucket histogram case), we have:

$$\begin{aligned}E_\sigma[X] &= \sum_{l=1}^{n} E_\sigma[X_l] \\ &= \sum_{l=1}^{n} |I_l| \overline{f}_l \overline{g}_l\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}_\sigma(X) &= \mathrm{Var}_\sigma(\sum_{l=1}^{n} X_l) \\ &= \sum_{l=1}^{n} \mathrm{Var}_\sigma(X_l)\end{aligned}$$

These results, together with the second result in Lemma 1 gives the second result here. □

THEOREM 2. *Let F and G be two relations. The histogram estimate $\sum_{l=1}^{n} |I_l| \overline{f}_l \overline{g}_l$ of $|F \bowtie_A G|$ using histograms with the same bucketization $I_1, \ldots, I_n$ for both F and G under the random shuffling assumption within each bucket is an* unbiased estimator *and has squared error*

$$\sum_{l=1}^{n} \frac{1}{|I_l| - 1} SqErr(F_l) SqErr(G_l)$$

PROOF. The proof follows directly from Lemma 2 □

The same observations that we made for one-bucket histograms hold here as well: (a) the way the histogram is used is consistent with the random shuffling assumption, and (b) histograms are likely to have a small error if the hypothesis holds. Note that we do not have to make any assumption about the distribution inside a bucket, just the fact that the arrangement for one of the relations is random.

It is possible to generalize Theorem 2 for the case when the buckets are not aligned (see Appendix A) if the random shuffling assumption holds for both relations. This result is unlikely to be useful in practice since, as noted by Ioannidis and Christodulakis [5], for maximal error reduction the same histogram has to be used for both relations.

5. COMPARISON WITH SAMPLING AND SKETCHES

In this section we compare histograms with two other approximate query processing techniques: sampling and sketches. We provide here only the comparison for the case when the random shuffling assumption holds; the comparison for the case when the random shuffling assumption does not hold is deferred to Section 6.

5.1 Sampling

Multiple variations of the sampling method have been proposed in the literature. The most important ones, exemplified on the problem of computing $|F \bowtie_A G|$, are:

- *Sampling from the base relation*[9]: In this type of sampling, random subsets of relations F and G are selected, the query is evaluated on the samples and scaled up to account for the difference in size.
- *Sample Counts:* Select a subsample I' of the domain I of the join attribute A and maintain exact counts f_i and g_i for $i \in I'$. If we do not take samples at exactly the same points in the domain for the two relations, we are wasting samples since we can use only the samples in the intersection. This sampling technique is a small modification of the sampling scheme in [2]. The only difference is that it starts the counting of item i from the beginning not from a random point so the counts are accurate.
- *Sampling from the join.* By producing iid samples from the join, we can simply scale up the number of such samples to get an estimate for the size of the join. The problem is that it is hard to produce iid samples out of join results.[3]

In the theoretical developments in this section we use sample counts since they are easy to produce and are better approximators than samples from the base relation. We assume that the choice of the subset I' is uniformly random; we do not attempt to pick I' so that the performance is improved. If sample counts are maintained for F and G, we know the precise values of f_i and g_i for $i \in I'$ and we do not know anything about the other frequencies. With these observations, the estimate based on the sample is the random variable $X = \frac{N}{N'} \sum_{i \in I'} f_i g_i$. This random variable, thus the estimation using sample counts, is characterized by the following two results:

PROPOSITION 2. *If random shuffling assumption holds for relation G, the sample count estimator for $|F \bowtie_A G|$ that maintains N' samples has average squared error bounded by:*

$$E_\sigma\left[Err(X)\right] \geq \frac{(N-N')(N-2)}{N'(N-1)^2} SqErr(F) SqErr(G)$$

PROOF. See Appendix B □

PROPOSITION 3. *The sample count estimator for $|F \bowtie_A F|$, the self-join size of F, is:*

$$Err(X) = \frac{N-N'}{N'(N-1)} \left[N \sum_{i \in I} f_i^4 - \left(\sum_{i \in I} f_i^2 \right)^2 \right]$$

PROOF. See Appendix C □

5.2 Sketches

A full introduction to AMS sketches [2, 1] is beyond the scope of the paper. The only thing we need here is the following result:

LEMMA 3. *The sketch approximation obtained by averaging n elementary sketch estimates has squared error:*

$$\frac{1}{n}\left[SJ(F)SJ(G) + \left(\sum_{i \in I} f_i g_i \right)^2 - 2 \sum_{i \in I} f_i^2 g_i^2 \right]$$

In particular, when $F = G$, the squared error is upper-bounded by $\frac{1}{n} SJ(F)^2$

PROOF. For a proof see [1]. □

PROPOSITION 4. *If random shuffling assumption holds, then the error averaging n elementary sketches is lower-bounded by*

$$\frac{1}{n} \frac{N-2}{N} SqErr(F) SqErr(G)$$

PROOF. The result follows directly from Lemma 3, the fact that $\mathrm{SqErr}(F) \leq \mathrm{SJ}(F)$ and the observation that

$$\begin{aligned} E_\sigma\left[\sum_{i \in I} f_i^2 g_{\sigma(i)}^2 \right] &= \sum_{i \in I} f_i^2 \frac{1}{N} \sum_{j \in I} g_j^2 \\ &= \frac{1}{N} \mathrm{SJ}(F)\mathrm{SJ}(G) \end{aligned}$$

□

5.3 Comments on Comparison

We postpone the discussion on how histograms compare with sampling and sketching for the self join size problem for Section 6.2. We look here only at the comparison in the case when the random shuffling assumption holds.

From the result in Theorem 1, we notice that the squared error of one-bucket histograms is proportional with the product of the squared errors of relations F and G but it is inversely proportional with the size of the domain of the join attribute minus one. The lower bounds on the squared error for samples, Proposition 2, and sketches, Proposition 4, is also proportional with the product of the squared errors of the two relations but it is roughly proportional with the inverse of the size of the synopsis, N' for the samples and n for the sketches. This means that both samples and sketches need space comparable to the size of the domain of the join attribute in order to compete with one-bucket histograms. Given this, we can conclude that, when the random shuffling assumption holds, histograms are the best approximation technique by a large margin.

6. HISTOGRAMS WHEN RANDOM SHUFFLING ASSUMPTION DOES NOT HOLD

In order to obtain a characterization of how histograms perform on general problems, not only in the case when the frequencies in the two relations are not correlated, we look at the average behavior of two classes of histograms: histograms with buckets of fixed size and general histograms. In Section 7 we survey other work that provided theoretical characterizations of histograms.

6.1 Random Histograms on Arbitrary Problems

THEOREM 3. *Let F and G be two relations. Let $I_1, \ldots, I_n$ be a random partitioning of I, the domain of join attribute A, with the property that $|I_i| = N_i$, a fixed and given number. Then, $X = \sum_{i=1}^{n} N_i \frac{\sum_{j \in I_i} f_i}{N_i} \frac{\sum_{j \in I_i} g_i}{N_i}$ is the histogram estimate for $|F \bowtie G|$ where the buckets are given by $I_1, \ldots, I_n$. Furthermore,*

$$E[X] = \frac{N-n}{N(N-1)} \sum_{j \in I} f_j \sum_{j \in I} g_j + \frac{n-1}{N-1} \sum_{j \in I} f_j g_j$$

where expectation is taken over the probability space in which all partitionings are equiprobable.

PROOF. See Appendix D. □

An immediate consequence of the above result is:

COROLLARY 1. *Under the same conditions as in Theorem 3 but removing the fixed size for each part in the partition, $E[X]$ remains the same.*

PROOF. Since the expected value of X does not depend on the particular values of N_i, i.e. the average behavior is the same irrespective of the bucket sizes, the expectation remains the same irrespective of the particular probability space over the values N_i, thus it is the same when we choose the probability space such that the probability of any of the partitionings is the same. □

Interestingly, on average, the result estimation using histograms is a linear combination between the estimator of the one-bucket histogram and the true size of the join. Unfortunately, the weight of the correct term is $\frac{n-1}{N-1}$, thus in general, histograms will provide accurate estimates only when $n \to N$ unless the one-bucket histogram is accurate. This only happens when random shuffling assumption holds.

6.2 Random Histograms for Self-join Size Computation

Using the result in Theorem 3 and its Corollary 1, when the relation F is joined with itself, on average, the error due to bias of histograms is:

$$\begin{aligned} \text{bias}(X) &= \left(\sum_{i \in I} f_i g_i - E[X] \right)^2 \\ &= \left(\frac{N-n}{N-1} \right)^2 (\text{SJ}(F) - N\bar{f}^2)^2 \\ &\approx \text{SqErr}(F)^2 \end{aligned}$$

Since standard error is the sum of the bias and the variance, when comparing the performance of histograms with sampling (Proposition 3) and sketches (Lemma 3), we notice that it is not inversely proportional with the size of the synopsis – it has a slow linear decrease. This means that, on average, for the self join size problem histograms behave fundamentally worse than sampling and sketching.

6.3 Comments on End-biased histograms

As we have seen in the previous sections, histograms have small error when random shuffling assumption holds and large error, on average, otherwise – for example when self-join sizes are computed. Interestingly, end-biased histograms [8] mostly avoid the poor performance when correlations are present. To see why, we observe that, when the random shuffling hypothesis holds, end-biased histograms behave well like all other histograms. On the other hand, when frequencies are correlated, the result is going to be dominated by the high frequencies so the estimation is again precise since these frequencies are captured accurately.

7. RELATED WORK

The amount of theoretical work on characterizing histograms as approximation methods for database queries is surprisingly small. Piatetsky-Shapiro and Connell[10] provided the first theoretical characterization of histograms, by deriving worst case and average case error guarantees for equi-width and equi-depth histograms when used for selectivity estimation. The only other theoretical characterizations of histograms, the only one applicable to estimation of aggregates over joins, can be found in the work of Ioannidis and his collaborators,[5, 7, 6, 8]. Most of this work is concerned with optimality for histograms, [5, 7, 8], for which, interestingly enough, the issue of computing the error of histograms can be cleverly avoided – the technical mean to do this is to rely on majorization theory instead of a direct optimization. Most of [6] and small parts of the other papers we mentioned are concerned with actual characterizing the error of histograms but most of the results apply only to one-bucket histograms – the only exception is worst case error estimation which results in large bounds. Interestingly, the fact that all histogram estimates are unbiased under the random shuffling assumption was proved in [5] (Theorem 6.1), but the result serves there a completely different purpose.

The idea that size of join algorithms can be extended to other aggregates over joins appeared first in a much simplified form in [4].

8. CONCLUSIONS

In this paper we showed that the random shuffling assumption, a significantly less restrictive statistical assumption than uniform distribution within a bucket, is consistent with

the way histograms are used for approximate query processing. Using this observation we derived tight error guarantees for the case when the assumption holds and showed that, in such cases, histograms are significantly better approximators than sampling and sketches. On the reverse, we showed that, when the assumption does not hold, on average, histograms are inferior to sampling and sketches. These theoretical developments suggest that none of the methods dominates the other throughout the problem spectrum; investigating hybrid approaches might be the key to design approximation methods that are optimal for all problems. The work we presented here is just a first step in the right direction; significant more research is required to obtain a complete theoretical characterization of histograms.

9. ACKNOWLEDGEMENTS

We would like to thank Florin Rusu for making useful comments on a draft of this paper.

10. REFERENCES

[1] N. Alon, P. B. Gibbons, Y. Matias, and M. Szegedy. "Tracking Join and Self-Join Sizes in Limited Storage". In *Proceedings of the Eighteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Philadeplphia, Pennsylvania, May 1999.

[2] N. Alon, Y. Matias, and M. Szegedy. "The Space Complexity of Approximating the Frequency Moments". In *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing*, pages 20–29, Philadelphia, Pennsylvania, May 1996.

[3] S. Chaudhuri, R. Motwani, and V. Narasayya. On random sampling over joins. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, Philadelphia, Pennsylvania, 1999.

[4] A. Dobra, M. Garofalakis, J. Gehrke, and R. Rastogi. "Processing Complex Aggregate Queries over Data Streams". In *Proc. of the 2002 ACM SIGMOD Intl. Conference on Management of Data*, pages 61–72, Madison, Wisconsin, June 2002.

[5] Y. Ioannidis. Universality of serial histograms. In *Proceedings of the 19th International Conference on Very Large Data Bases*, Dublin, Ireland, 1993. IEEE.

[6] Y. Ioannidis and S. Christodulaki. On the propagation of errors in the size of join result. In *Proceedings of the 1991 ACM SIGMOD International Conference on Management of Data*, Denver, Colorado, 1991.

[7] Y. Ioannidis and S. Christodulakis. Optimal histograms for limiting worst-case error propagation in the size of join results. *ACM Transactions on Database Systems*, 18(4):709–748, December 1993.

[8] Y. Ioannidis and V. Poosala. Balancing histogram optimality and practicality for query result size estimation. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, San Jose,CA, 1995.

[9] F. Olken and D. Rotem. Random sampling from databases - a survey, 1995.

[10] G. Piatetsky-Shapiro and C. Connell. Accurate estimation of the number of tuples satisfying a condition. In *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*, Boston, MA, June 1984.

[11] J. Shao. *Mathematical Statistics*. Springer-Verlag, 1999.

APPENDIX

A. GENERAL UNIDIMENSIONAL HISTOGRAMS

In this section we generalize the result in Theorem 2 to the general case in which the buckets might not be aligned. As we showed in Section 4.2, if buckets are aligned, the random shuffling assumption has to hold only for one of the relations. If buckets are not aligned, it is necessary for the assumption to hold for both relations. To obtain the desired result we first prove:

LEMMA 4. *Let $I_1, \ldots, I_n$ be an arbitrary bucketization of domain I. Let $J_1, \ldots, J_m$ be another such bucketization. Let γ and σ be independent random permutations of domain I and define $X = \sum_{i \in I} f_{\gamma(i)} g_{\sigma(i)}$. Then, random variable X has to properties:*

$$E_{\gamma,\sigma}[X] = \sum_{l=1}^{n}\sum_{k=1}^{m} \overline{f}_l \overline{g}_k |I_l \cap J_k|$$

$$Var_{\gamma,\sigma}(X) = \sum_{l=1}^{n}\sum_{k=1}^{m} SqErr(F_l) SqErr(G_k) \frac{|I_l \cap J_k|}{|I_l||J_k|} \left(1 + \frac{|I_l \cap J_k| - 1}{(|I_l|-1)(|J_k|-1)}\right)$$

PROOF. We first have:

$$E_{\gamma,\sigma}[X] = \sum_{i \in I} E_{\gamma,\sigma}\left[f_{\gamma(i)} g_{\sigma(i)}\right] = \sum_{i \in I} E_{\gamma}\left[f_{\gamma(i)}\right] E_{\sigma}\left[g_{\sigma(i)}\right]$$

where we used the linearity of expectation and independence of γ and σ. If $i \in I_l$ for some l, then $E_{\gamma}\left[f_{\gamma(i)}\right] = \overline{f}_l$. To incorporate this if condition in the equations we use the identity function to get $E_{\gamma}\left[f_{\gamma(i)}\right] = \sum_{l=1}^{n} \mathcal{I}(i \in I_l)\overline{f}_l$. Indeed, only for l such that $i \in I_l$ the contribution to the sum is nonzero. Similarly, we have $E_{\sigma}\left[g_{\sigma(i)}\right] = \sum_{k=1}^{m} \mathcal{I}(i \in J_k)\overline{g}_k$. With this:

$$E_{\gamma,\sigma}[X] = \sum_{i \in I}\sum_{l=1}^{n} \mathcal{I}(i \in I_l)\overline{f}_l \sum_{k=1}^{m} \mathcal{I}(i \in J_k)\overline{g}_k = \sum_{l=1}^{n}\sum_{k=1}^{m} \overline{f}_l \overline{g}_k \sum_{i \in I} \mathcal{I}(i \in I_l \wedge i \in J_k) = \sum_{l=1}^{n}\sum_{k=1}^{m} \overline{f}_l \overline{g}_k |I_l \cap J_k|$$

where we used the fact that $\sum_{i \in I} \mathcal{I}(C)$ is the number of elements of I satisfying the condition, the number of elements both in I_l and J_k here.

$$\begin{aligned}\mathrm{Var}_{\gamma,\sigma}(X) &= \mathrm{Var}_{\gamma,\sigma}(\sum_{i\in I} f_{\gamma(i)}g_{\sigma(i)}) \\ &= \sum_{i\in I}\sum_{i'\in I}\mathrm{Cov}_\gamma(f_{\gamma(i)}, f_{\gamma(i')})\mathrm{Cov}_\sigma(g_{\sigma(i)}, g_{\sigma(i')})\end{aligned}$$

Using the fact that, for i and i' within the bucket I_l, we can write

$$\mathrm{Cov}_\gamma(f_{\gamma(i)}, f_{\gamma(i')}) = \frac{\mathrm{SqErr}(F_l)}{|I_l|-1}(\delta_{ii'} - \frac{1}{|I_l|})$$

with $\delta_{ii'}$ the Kronecker symbol that takes value 1 if $i = i'$ and 0 otherwise – this is a simple way to combine Equations 5 and 6. Now, using the same technique as before to avoid guessing the bucket in which i and i' have to be, we can write

$$\begin{aligned}&\mathrm{Cov}_\gamma(f_{\gamma(i)}, f_{\gamma(i')}) \\ &\quad= \sum_{l=1}^{n}\mathcal{I}(i\in I_l)\mathcal{I}(i'\in I_l)\frac{\mathrm{SqErr}(F_l)}{|I_l|-1}(\delta_{ii'} - \frac{1}{|I_l|})\end{aligned}$$

Similarly we have

$$\begin{aligned}&\mathrm{Cov}_\sigma(g_{\sigma(i)}, g_{\sigma(i')}) \\ &\quad= \sum_{k=1}^{m}\mathcal{I}(i\in J_k)\mathcal{I}(i'\in J_k)\frac{\mathrm{SqErr}(G_k)}{|J_k|-1}(\delta_{ii'} - \frac{1}{|J_k|})\end{aligned}$$

Using the fact that

$$\begin{aligned}&\left(\delta_{ii'} - \frac{1}{|I_l|}\right)\left(\delta_{ii'} - \frac{1}{|J_k|}\right) \\ &\quad= \left(1 - \frac{1}{|I_l|} - \frac{1}{|J_k|}\right)\delta_{ii'} + \frac{1}{|I_l||J_k|}\end{aligned}$$

and substituting back in the expression for $\mathrm{Var}_{\gamma,\sigma}(X)$, we get

$$\begin{aligned}&\mathrm{Var}_{\gamma,\sigma}(X) \\ &= \sum_{l=1}^{n}\sum_{k=1}^{m}\frac{\mathrm{SqErr}(F_l)}{|I_l|-1}\frac{\mathrm{SqErr}(G_k)}{|J_k|-1} \\ &\qquad\left[\left(1 - \frac{1}{|I_l|} - \frac{1}{|J_k|}\right)\sum_{i\in I}\mathcal{I}(i\in I_l)^2\mathcal{I}(i\in J_k)^2 + \right. \\ &\qquad\left.\frac{1}{|I_l||J_k|}\sum_{i\in I}\sum_{i'\in I}\mathcal{I}(i\in I_l)\mathcal{I}(i'\in I_l)\mathcal{I}(i\in J_k)\mathcal{I}(i'\in J_k)\right] \\ &= \sum_{l=1}^{n}\sum_{k=1}^{m}\mathrm{SqErr}(F_l)\mathrm{SqErr}(G_k)\frac{|I_l\cap J_k|}{|I_l||J_k|} \\ &\qquad\left(1 + \frac{|I_l\cap J_k| - 1}{(|I_l|-1)(|J_k|-1)}\right)\end{aligned}$$

where we used the fact that $\mathcal{I}(C)^2 = \mathcal{I}(C)$ irrespective of the condition C and that $\sum_{i\in I}\mathcal{I}(i\in I_l)\mathcal{I}(i\in J_k) = |I_l\cap J_k|$. □

We now have the main result:

THEOREM 4. *Let F and G be two relations. The histogram estimate $\sum_{l=1}^{n}\sum_{k=1}^{m}\overline{f}_l\overline{g}_k|I_l\cap J_k|$ of $|F\bowtie_A G|$ using histograms with buckets $I_1,\ldots,I_n$ for F and $J_1,\ldots,J_m$ for G when the random shuffling assumption holds for both relations within each bucket is an* unbiased estimator *and has squared error:*

$$\begin{aligned}&\sum_{l=1}^{n}\sum_{k=1}^{m}SqErr(F_l)SqErr(G_k)\frac{|I_l\cap J_k|}{|I_l||J_k|} \\ &\qquad\left(1 + \frac{|I_l\cap J_k| - 1}{(|I_l|-1)(|J_k|-1)}\right)\end{aligned}$$

PROOF. Follows directly from proof of Lemma 4 □

B. PROOF OF PROPOSITION 2

LEMMA 5. *Let F and G be arbitrary relations with frequencies f_i and g_i for $i\in I$, where I is the domain of the join attribute A. By taking $N = |I|$ and $N' = |I'|$ and setting random variable X to be $X = \frac{N}{N'}\sum_{i\in I'} f_i g_i$,the sample estimate of the size of the join, we have:*

$$E[X] = \sum_{i\in I} f_i g_i$$

$$Var(X) = \frac{N-N'}{N'(N-1)}\left[N\sum_{i\in I} f_i^2 g_i^2 - \left(\sum_{i\in I} f_i g_i\right)^2\right]$$

PROOF. In order to simplify the analysis, we introduce the random variables Y_i that take value 1 if $i\in I'$ and 0 otherwise. If I' is a random subset of I and i, i' are arbitrary elements of I then,

$$E[Y_i] = P[i\in I'] = \frac{N'}{N} \tag{7}$$

$$E[Y_i^2] = P[i\in I'] = \frac{N'}{N} \tag{8}$$

$$\begin{aligned}\mathrm{Var}(Y_i) &= E[Y_i^2] - E[Y_i]^2 \\ &= \frac{N'}{N}\frac{N-N'}{N}\end{aligned} \tag{9}$$

$$\begin{aligned}i\neq i',\quad E[Y_iY_{i'}] &= P[i\in I'\wedge i'\in I'] \\ &= P[i\in I']\,P[i'\in I'|i\in I'] \\ &= \frac{N'}{N}\frac{N'-1}{N-1}\end{aligned} \tag{10}$$

$$\begin{aligned}i\neq i', \mathrm{Cov}(Y_i, Y_{i'}) &= E[Y_iY_{i'}] - E[Y_i]E[Y_{i'}] \\ &= -\frac{N'(N-N')}{N^2(N-1)}\end{aligned} \tag{11}$$

$$\mathrm{Cov}(Y_i, Y_{i'}) = \frac{N'(N-N')}{N(N-1)}\delta_{ii'} - \frac{N'(N-N')}{N^2(N-1)} \tag{12}$$

Using random variables Y_i we can rewrite X as:

$$X = \frac{N}{N'}\sum_{i\in I} Y_i f_i g_i$$

With this, we have

$$E[X] = \frac{N}{N'}\sum_{i\in I} E[Y_i]\, f_i g_i$$
$$= \sum_{i\in I} f_i g_i$$

$$\begin{aligned}\mathrm{Var}(X) &= \frac{N^2}{N'^2}\sum_{i\in I}\sum_{i'\in I} f_i g_i f_{i'} g_{i'} \mathrm{Cov}(Y_i, Y_{i'}) \\ &= \frac{N^2}{N'^2}\left[\frac{N'(N-N')}{N(N-1)}\sum_{i\in I} f_i^2 g_i^2 - \right. \\ &\qquad \left. \frac{N'(N-N')}{N^2(N-1)}\sum_{i\in I}\sum_{i'\in I} f_i g_i f_{i'} g_{i'}\right] \\ &= \frac{N-N'}{N'(N-1)}\left[N\sum_{i\in I} f_i^2 g_i^2 - \left(\sum_{i\in I} f_i g_i\right)^2\right]\end{aligned}$$

□

With this result we can now estimate the error sample counting makes when the random shuffling assumption holds.

Since, according to Lemma 5, sample counting method is unbiased irrespective of the problem, we have $\mathrm{Err}(X) = \mathrm{Var}(X)$.

$$\begin{aligned}&E_\sigma[\mathrm{Err}(X)] \\ &= \frac{N-N'}{N'(N-1)}\left[N\sum_{i\in I} f_i^2 E_\sigma\left[g_{\sigma(i)}^2\right] - \right. \\ &\qquad \left. E_\sigma\left[\left(\sum_{i\in I} f_i g_{\sigma(i)}\right)^2\right]\right] \\ &= \frac{N-N'}{N'(N-1)}\left[\mathrm{SJ}(F)\mathrm{SJ}(G) - \right. \\ &\qquad \left. \mathrm{Var}_\sigma(\sum_{i\in I} f_i g_{\sigma(i)}) - E_\sigma\left[\sum_{i\in I} f_i g_{\sigma(i)}\right]^2\right] \\ &= \frac{N-N'}{N'(N-1)}\left[\mathrm{SJ}(F)\mathrm{SJ}(G) - \right. \\ &\qquad \left. \frac{1}{N-1}\mathrm{SqErr}(F)\mathrm{SqErr}(G) - N^2\overline{f}^2\overline{g}^2\right]\end{aligned}$$

where we used the fact that $\mathrm{Var}(X) = E\left[X^2\right] - E[X]^2$ and the result in Lemma 1. To get the required inequality we observe that:

$$\begin{aligned}&\mathrm{SJ}(F)\mathrm{SJ}(G) - N^2\overline{f}^2\overline{g}^2 \\ &= (\mathrm{SJ}(F) - N\overline{f}^2)(\mathrm{SJ}(G) - N\overline{g}^2) \\ &\quad + N\overline{f}^2(\mathrm{SJ}(G) - N\overline{g}^2) + N\overline{g}^2\mathrm{SJ}(F) - N\overline{f}^2) \\ &= \mathrm{SqErr}(F)\mathrm{SqErr}(G) + N\overline{f}^2\mathrm{SqErr}(G) + N\overline{g}^2\mathrm{SqErr}(F) \\ &\geq \mathrm{SqErr}(F)\mathrm{SqErr}(G)\end{aligned}$$

C. PROOF OF PROPOSITION 3

The proof follows directly from the unbiasedness of the estimator and by substituting g_i by f_i in Lemma 5.

D. PROOF OF THEOREM 3

Within bucket I_i, the average frequencies of relations F and G are $\frac{\sum_{j\in I_i} f_j}{N_i}$ and $\frac{\sum_{j\in I_i} g_j}{N_i}$, respectively. Since the histogram estimate is sum over all buckets over the product of the size of the bucket and the average frequencies, indeed X is the histogram estimate.

In order to analyze X, we introduce random variables Y_{ij} that take value 1 if $j \in I_i$ and 0 otherwise. Y_{ij} has the following two properties:

$$E[Y_{ij}] = E\left[Y_{ij}^2\right] = P[j \in I_i] = \frac{N_i}{N} \tag{13}$$

$$\begin{aligned}j \neq j',\ E[Y_{ij}Y_{ij'}] &= P\left[j \in I_i \wedge j' \in I_i\right] \\ &= P[j \in I_i]\, P\left[j' \in I_i | j \in I_i\right] \\ &= \frac{N_i}{N}\frac{N_i - 1}{N-1}\end{aligned} \tag{14}$$

Using random variables Y_{ij} we can rewrite X as:

$$X = \sum_{i=1}^{n} N_i \frac{\sum_{j\in I} Y_{ij} f_i}{N_i}\frac{\sum_{j\in I} Y_{ij} g_i}{N_i}$$

With this, we have:

$$\begin{aligned}&E[X] \\ &= \sum_{i=1}^{n}\frac{1}{N_i}\sum_{j\in I}\sum_{j'\in I} f_j g_{j'} E[Y_{ij}Y_{ij'}] \\ &= \sum_{i=1}^{n}\frac{1}{N_i}\left[\sum_{j\in I}\sum_{j'\in I, j'\neq j} f_j g_{j'}\frac{N_i}{N}\frac{N_i-1}{N-1} + \sum_{j\in I} f_j g_j \frac{N_i}{N}\right] \\ &= \sum_{i=1}^{n}\frac{1}{N}\left[\sum_{j\in I}\sum_{j'\in I} f_j g_{j'}\frac{N_i-1}{N-1}\right. \\ &\qquad \left. + \sum_{j\in I} f_j g_j\left(1 - \frac{N_i-1}{N-1}\right)\right] \\ &= \frac{N-n}{N(N-1)}\sum_{j\in I} f_j \sum_{j\in I} g_j + \frac{n-1}{N-1}\sum_{j\in I} f_j g_j\end{aligned} \tag{15}$$