Hoeffding's Inequality for Sampling Without Replacement

Ruban Vishnu Pandian

1 Introduction

Hoeffding's inequality is a very famous result in concentration theory. It is a classic result which upper bounds the probability that the sum of bounded, independent random variables differ from the sum of their expectations beyond an error parameter. The formal mathematical statement is:

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - E[X_i]) \geq t\right) \leq \exp\left(\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{1}$$

where $X_i's$ are the independent RVs, a_i, b_i are the lower and upper bounds respectively for the RV X_i, i.e., $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$ and t is a positive error parameter.

The above result can be applied to a specialized setting where $X_i's$ are IID RVs, i.e., all have the same marginal distribution and hence, same expectation and bounds. Let the bounds be a, b and let $E[X_i] = \mu$. In this case, the theorem simplifies as:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}X_i - \mu \geq t\right) \leq \exp\left(\frac{2nt^2}{(b-a)^2}\right)$$

This is a very useful result in concentration theory and statistical learning theory since it helps us provide sample complexity based guarantees for the PAC learning framework. This result can also be applied when the $X_i's$ are sampled with replacement from a finite dataset:

$$D = \{a_1, a_2, ..., a_N\} \;\; (n \leq N)$$

Since when the RVs are sampled with replacement, they do not influence each other and have the same marginal distributions and hence are IID. But what happens if they are sampled without replacement? Is the above theorem still applicable? The answer turns out to be **Yes!** and Hoeffding even proved it in

his original paper. The formal statement is:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq t\right) \leq \exp\left(\frac{2nt^2}{(b-a)^2}\right) \tag{2}$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} a_i$ is the population average of the dataset. But how is this theorem applicable when the samples are obtained without replacement? Clearly the original theorem cannot be directly applied since sampling without replacement introduces dependency between the samples. So some other method must be applied.

This is where Hoeffding proposed a clever idea of relating the two cases using an auxillary function. This whole article is just my attempt at rigorously going through Hoeffding's original proof of extending the result for samples obtained without replacement. In particular, to explore the combinatorial nuances present in this proof and how it cleanly connects to a very useful combinatorial object: **Sterling numbers of Second kind**.

2 Notations:

Important notations used in the following sections are defined here:

1. $[N]$: Set of natural numbers from 1 to N.
2. $[N]^n$: Cartesian product of $[N]$ with itself n times.
3. N^n : N raised to the power n which is also the cardinality of the set $[N]^n$.
4. $\{N\}^n$: Set of all ordered n-tuples of natural numbers from 1 to N with no repetition, i.e., no number appears twice in the tuple.
5. $(N)^n$: The falling factorial $N(N-1)...(N-n+1)$. Turns out it is also the cardinality of $\{N\}^n$.
6. $\{\{N\}\}^n$: Set of all unordered n-tuples of natural numbers from 1 to N with no repetition, i.e., no number appears twice in the tuple and no two tuples in the set are permutations of each other.
7. $\binom{N}{n}$: The standard binomial coefficient which also turns out to be the cardinality of $\{\{N\}\}^n$.
8. $< N >^n = \{(r_1, r_2, ..., r_n) : r_1, r_2, ...r_n \in \mathbb{N}, r_1 + r_2 + ... + r_n = N\}$: The set of ordered n-tuples of natural numbers adding to N.
9. $<< N >>^n$: The set of unordered n-tuples of natural numbers adding to N, i.e., no two tuples in this set are permutations of each other.
10. $\pi(B)$: The set of all unique permutations of the tuple B.

11. $\binom{N}{r_1\ r_2\ r_3\ ...\ r_n}$: The multinomial coefficient where $r_1 + r_2 + ... + r_n = N$.
12. $D = \{a_1, a_2, ..., a_N\}$: Finite dataset from which samples will be obtained.
13. $X_1, X_2, ...X_n$: IID samples obtained by sampling D uniformly with replacement.
14. $Y_1, Y_2, ..., Y_n$: Samples obtained by sampling D uniformly without replacement.
15. $S(n, k)$: Sterling numbers of Second kind with arguments n and k. They count how many distinct, unlabeled partitions of size k can be made from a set of n distinct objects.

3 Expectation of Convex function applied on Sample average with and without replacement

An important step in this endeavor involves the Chernoff bound in upper bounding the probability. Chernoff bound is another useful result in concentration theory which applies the Markov inequality to an exponential version of the original event. Formally, we have:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq t\right) = \mathbb{P}\left(\exp\left(\lambda \sum_{i=1}^{n} \tilde{X}_i\right) \geq \exp(\lambda n t)\right)$$

$$\leq \frac{E\left[\exp\left(\lambda \sum_{i=1}^{n} \tilde{X}_i\right)\right]}{\exp(\lambda n t)} \tag{3}$$

for all $\lambda > 0$. Here $\tilde{X_i}'s$ denote the mean-corrected versions of $X_i's$ i.e., $\tilde{X}_i = X_i - \mu$. Now if the $X_i's$ are independent then the expectation in the RHS factorizes into product of individual expectations which can be upper bounded using the Hoeffding's Lemma. This is the whole idea behind proof of Hoeffding's inequality. But this idea is not directly applicable when the $X_i's$ are obtained by sampling without replacement.

Note that $f(\sum_{i=1}^{n} X_i) = \exp(\lambda \sum_{i=1}^{n} X_i)$ is a convex function it is argument. An important result relating expectations of convex functions when its input argument is sample average of samples obtained with and without replacement turns out to be the bridge in extending the Hoeffding's inequality for samples obtained without replacement. It is the following:

$$E\left[f\left(\sum_{i=1}^{n} Y_i\right)\right] \leq E\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \tag{4}$$

Proving this result will be the main body of this article.

4 Proof:

4.1 Construction of the auxillary function:

We first construct an auxillary function $g(\cdot)$ as follows:

$$g(t_1, t_2, ..., t_n) = \sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k, R_k, J_k) f(J_k \otimes R_k)$$

where:

$$R_k = (r_1, r_2, ..., r_k)$$
$$J_k = (j_1, j_2, ..., j_k)$$
$$J_k \otimes R_k = r_1 t_{j_1} + r_2 t_{j_2} + ... + r_k t_{j_k}$$

$P(k, R_k, J_k)$ are the coefficients in the auxillary function definition and $J_k \otimes R_k$ is simply the weighted sum of sample values $t'_{j_i}s$ using counts $r'_i s$.

Basically, the auxillary function contains contributions from all possible n-tuples that can be formed by sampling $\{t_1, t_2, ..., t_n\}$ with replacement. A key point to note is $\{\{n\}\}^k$ contains tuples such that no two of them are permutations of each other. Hence, without loss of generality, we can assume any $J \in \{\{n\}\}^k$ is always ordered, i.e., any $(j_1, j_2, ..., j_k) \in \{\{n\}\}^k$ satisfies $j_1 < j_2 < ... < j_k$. This also ensures every possible combination of indices contributes exactly once with a given $(r_1, r_2, ..., r_k)$.

Key point to note: It is important that only distinct combinations of $(j_1, j_2, ..., j_k)$ must be considered, i.e., no two $J_k \in \{\{n\}\}^k$ can be permutations of each other. This is because the permutations of counts are handled in the notation $R_k \in < n >^k$, i.e., two tuples in $< n >^k$ can be permutations of each other. Consider the following example where $n = 5, k = 3, r_1 = 2, r_2 = 2, r_3 = 1$. We have two cases:

- *Permutation allowed in indices but not in counts:* In this case, we'll have the following possibilities:
  1. $(t_{j_1}, t_{j_1}, t_{j_2}, t_{j_2}, t_{j_3})$
  2. $(t_{j_1}, t_{j_1}, t_{j_3}, t_{j_3}, t_{j_2})$
  3. $(t_{j_2}, t_{j_2}, t_{j_1}, t_{j_1}, t_{j_3})$
  4. $(t_{j_2}, t_{j_2}, t_{j_3}, t_{j_3}, t_{j_1})$
  5. $(t_{j_3}, t_{j_3}, t_{j_1}, t_{j_1}, t_{j_2})$
  6. $(t_{j_3}, t_{j_3}, t_{j_2}, t_{j_2}, t_{j_1})$
- *Permutation allowed in counts but not in indices:* In this case, we'll have the following possibilities:
  1. $(t_{j_1}, t_{j_1}, t_{j_2}, t_{j_2}, t_{j_3})$

2. $(t_{j_1}, t_{j_1}, t_{j_2}, t_{j_3}, t_{j_3})$
3. $(t_{j_1}, t_{j_2}, t_{j_2}, t_{j_3}, t_{j_3})$

In the first case, double counting happens. For example $(t_{j_1}, t_{j_1}, t_{j_2}, t_{j_2}, t_{j_3})$ and $(t_{j_2}, t_{j_2}, t_{j_1}, t_{j_1}, t_{j_3})$ correspond to the same case but are counted twice. This happens when some of the $r_i's$ are equal. But when permutation is allowed only in the counts r_i, it already handles these kinds of cases and avoids double counting. Clearly, permutations cannot be allowed in both counts and indices and also at least one of them must have permutations. Else, if both had no permutations then the only possible tuple would have been $(t_{j_1}, t_{j_1}, t_{j_2}, t_{j_2}, t_{j_3})$ and distinct tuples like $(t_{j_2}, t_{j_2}, t_{j_3}, t_{j_3}, t_{j_1})$ would have been missed.

This is one of the important combinatorial nuances which needs to be addressed in this proof and also turns out to be useful later. In short, allow permutations in counts r_i but not in indices j_i.

4.2 Defining the coefficients $P(k, R_k, J_k)$:

To fully define $g(\cdot)$, $P(k, R_k, J_k)$ needs to be defined. To do so, we use the next defining property of $g(\cdot)$:

$$E[g(Y_1, Y_2, ..., Y_n)] = E\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \tag{5}$$

This is the key defining property which would help us derive the needed result. Also note that $Y_i's$ and $X_i's$ can be equivalently written as:

$$X_i = a_{x_i}; \;\; Y_i = a_{y_i}$$

where $x_1, x_2, ..., x_n$ are indices obtained by sampling $[N]$ uniformly with replacement and $y_1, y_2, ..., y_n$ are obtained by sampling $[N]$ without replacement. Now let us write down the explicit expressions for both the expectations using the equivalent notation defined above:

$$E[f(X_1, X_2, ..., X_n)] = \frac{1}{N^n} \sum_{(x_1, x_2, ..., x_n) \in [N]^n} f(a_{x_1}, a_{x_2}, ..., a_{x_n})$$

$$E[g(Y_1, Y_2, ..., Y_n)] = \frac{1}{(N)^n} \sum_{(y_1, y_2, ..., y_n) \in \{N\}^n} g(a_{y_1}, a_{y_2}, ..., a_{y_n})$$

Hence, we have:

$$\frac{1}{N^n} \sum_{(x_1, x_2, ..., x_n) \in [N]^n} f(a_{x_1}, a_{x_2}, ..., a_{x_n}) = \frac{1}{(N)^n} \sum_{(y_1, y_2, ..., y_n) \in \{N\}^n} g(a_{y_1}, a_{y_2}, ..., a_{y_n})$$

Now to define $P(k, R_k, J_k)$, we'll do coefficient matching. For a given $R_k = (r_1, r_2, ..., r_k)$ and J_k, there will be exactly $\binom{n}{r1\ r2\ r3\ ...\ r_k}$ terms in the LHS corresponding to that. This is because we want k unique elements with counts

$r_1, r_2, ..., r_k$ to be sampled. In other words, rearrange k distinct elements with counts $r_1, r_2, ..., r_k$ in n slots which can be done in $\binom{n}{r_1\ r_2\ r_3\ ...\ r_k}$ ways.

In the RHS, the contribution from J_k would be present in those terms whose set of indices contain J_k as a subset, i.e., the term corresponding to $(y_1, y_2, ..., y_n)$ will contain the contribution from J_k if $J_k \in (y_1, y_2, ..., y_n)$. This is simply based on how we have defined $g(\cdot)$. And in our construction, $P(k, R_k, J_k)$ only depends on the counts and indices, not on the specific set $(y_1, y_2, ...y_n)$ which serves as the input index set to $g(\cdot)$. Hence, $P(k, R_k, J_k)$ will be same in all the index sets $(y_1, y_2, ..., y_n)$ containing J_k. This means to find the coefficient on the RHS, we simply need to find the total number of index sets containing J_k.

To do so, we first need to choose k slots among $(y_1, y_2, ..., y_n)$ for the elements of J_k, which can be done in $\binom{n}{k}$ ways and then rearrange them since order of $(y_1, y_2, ..., y_n)$ matters. Hence, elements of J_k can be placed in $\binom{n}{k} \times k! = (n)^k$ ways. The remaining $n-k$ slots must be filled using the remaining $N-k$ indices with order being distinct, which can be done in $(N-k)^{n-k}$ ways. Hence, in the RHS, $(n)^k(N-k)^{n-k}$ terms would contain J_k. Hence, we finally have:

$$\frac{\binom{n}{r_1\ r_2\ r_3\ ...\ r_k}}{N^n} = \frac{(n)^k(N-k)^{n-k}}{(N)^n} P(k, R_k, J_k)$$

$$\implies P(k, R_k, J_k) = \binom{n}{r_1\ r_2\ r_3\ ...\ r_k} \frac{(N)^n}{N^n} \frac{1}{(n)^k(N-k)^{n-k}} \tag{6}$$

Note that $P(k, R_k, J_k)$ does not depend on J_k at all. Hence, we can simply re-parametrize it as $P(k, R_k)$. With $P(k, R_k)$ fully defined, $g(\cdot)$ is also properly defined. The important property of $g(\cdot)$ is 5.

4.3 The coefficients add to 1:

In this section we'll prove the following statement:

$$\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k, R_k) = 1$$

To do so, first note that the summation over index sets can be simplified since the coefficients do not depend on J_k. Hence, we simply need to find how many $J_k's$ are present in $\{\{n\}\}^k$. In other words, the cardinality of $\{\{n\}\}^k$ which is simply $\binom{n}{k}$. Hence, we have:

$$\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k, R_k) = \sum_{k=1}^{n} \sum_{R_k \in <n>^k} \binom{n}{k} P(k, R_k)$$

Writing out the expression we derived for $P(k, R_k)$, we have:

$$\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \binom{n}{k} P(k, R_k) = \sum_{k=1}^{n} \sum_{R_k \in <n>^k} \binom{n}{k} \binom{n}{r_1\ r_2\ r_3\ ...\ r_k} \frac{(N)^n}{N^n} \frac{1}{(n)^k(N-k)^{n-k}}$$

Expanding out the relevant numbers in their factorial form, we get:

$$\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \binom{n}{r1\ r2\ r3\ ...\ r_k} \frac{n!}{k!(n-k)!} \frac{N!}{(N-n)!N^n} \frac{(n-k)!(N-n)!}{(n!)(N-k)!}$$

$$= \sum_{k=1}^{n} \sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ r_3\ ...\ r_k} \frac{N!}{k!N^n} \frac{1}{(N-k)!} = \sum_{k=1}^{n} \frac{1}{N^n} \binom{N}{k} \sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ r_3\ ...\ r_k}$$

The inner sum is simply the sum of multinomial coefficients of order n and groups k with the added constraint that none of the counts are zero. This sum is simply $k!$ times the Sterling number of second kind $S(n,k)$, i.e., we have:

$$\sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ r_3\ ...\ r_k} = k!S(n,k) \tag{7}$$

Hence, our original sum becomes:

$$\sum_{k=1}^{n} \frac{1}{N^n} \binom{N}{k} \sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ r_3\ ...\ r_k} = \sum_{k=1}^{n} \frac{1}{N^n} \binom{N}{k} k!S(n,k)$$

$$= \sum_{k=1}^{n} \frac{(N)^k}{N^n} S(n,k)$$

There is a polynomial identity involving Sterling numbers of second kind and falling factorials. It states:

$$\sum_{k=1}^{n} (N)^k S(n,k) = N^n \implies \sum_{k=1}^{n} \frac{(N)^k}{N^n} S(n,k) = 1 \tag{8}$$

But this was exactly what our original sum, i.e., sum of coefficients $P(k, R_k)$ was which we have shown to be 1 and hence, proven. Details about Sterling numbers and their properties which were used above are all proved in detail in the Appendix 7.

4.4 The weighted sum of $P(k, R_k)$ and $J_k \otimes R_k$ adds to $\sum_{i=1}^{n} t_i$:

In this section we'll prove the following statement:

$$\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k, R_k)(J_k \otimes R_k) = \sum_{i=1}^{n} t_i$$

Note that $J_k \otimes R_k$ is defined as:

$$J_k \otimes R_k = r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k}$$

After substituting this expression and rearranging the two inner summations, we get:

$$\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k, R_k)(J_k \otimes R_k)$$

$$= \sum_{k=1}^{n} \sum_{J_k \in \{\{n\}\}^k} \sum_{R_k \in <n>^k} P(k, R_k)(r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k}) \tag{9}$$

Let us focus only on the inner most sum of a fixed J_k. Note that it can be rewritten in the following manner:

$$\sum_{R_k \in <n>^k} P(k, R_k)(r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k})$$

$$= \sum_{B \in <<n>>^k} \sum_{R_k \in \pi(B)} P(k, R_k)(r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k})$$

We are simply decomposing the sum over all ordered tuples into a double summation over all unordered tuples and all permutations of an unordered tuple. Also note that $P(k, R_k)$ is invariant under permutations of R_k, which is evident from its expression in 6. Hence, $P(k, R_k)$ will be same for all $R_k \in \pi(B)$. Hence, we can further simplify the summation as:

$$\sum_{B \in <<n>>^k} \sum_{R_k \in \pi(B)} P(k, R_k)(r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k})$$

$$= \sum_{B \in <<n>>^k} P(k, B) \sum_{R_k \in \pi(B)} (r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k}) \tag{10}$$

Now let us focus only on the inner sum. It can be simplified as:

$$\sum_{R_k \in \pi(B)} (r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k})$$

$$= \left(\sum_{R_k \in \pi(B)} r_1 \right) t_{j_1} + \left(\sum_{R_k \in \pi(B)} r_2 \right) t_{j_2} + ... + \left(\sum_{R_k \in \pi(B)} r_k \right) t_{j_k}$$

Note $\pi(B)$ is exactly the set of all permutations of the tuple B. Hence, by symmetry, the summation $\sum_{R_k \in \pi(B)} r_i$ will be the same for all $i \in [K]$. Hence, we have:

$$\sum_{R_k \in \pi(B)} (r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k}) = \left(\sum_{R_k \in \pi(B)} r_1 \right) (t_{j_1} + t_{j_2} + ... + t_{j_k}) \tag{11}$$

Once again, it is enough to just focus on the first summation. Let $B = (b_1, b_2, ..., b_k)$. Moreoever, let $C = \{c_1, c_2, ..., c_l\}$ be the set of unique values

present in the tuple B (clearly $1 \leq l \leq k$). Let their counts be $w_1, w_2, ..., w_l$ i.e. $w_1, w_2, ...w_l \in \mathbb{N}$ and $w_1 + w_2 + ... + w_l = k$. So the distinct values r_1 can take in the summation are exactly contained in the set C.

Let's assume it takes value c_1. This means the tuple R_k has its first value as c_1. We need to find how many such tuples are present in $\pi(B)$. But that is easy, the remaining $k-1$ slots must be filled by distinct object $c_1, c_2, ..., c_l$ with counts $w_1 - 1, w_2, ..., w_l$. This can be done in $\binom{k-1}{w_1-1\ w_2\ w_3\ ...\ w_l}$ ways. A similar analysis can be done when r_1 takes the other c_i values too. Based on this logic, we have:

$$\sum_{R_k \in \pi(B)} r_1$$

$$= c_1 \binom{k-1}{w_1 - 1\ w_2\ w_3\ ...\ w_l} + c_2 \binom{k-1}{w_1\ w_2 - 1\ w_3\ ...\ w_l} + ... + c_l \binom{k-1}{w_1\ w_2\ w_3\ ...\ w_l - 1} \tag{12}$$

This is another key combinatorial nuance that must be properly understood. Now note that we have:

$$\binom{k-1}{w_1\ w_2\ ...w_i - 1 ...\ w_l} = \frac{(k-1)!}{(w_1!)(w_2!)...(w_i - 1)!...(w_k!)}$$

$$= \frac{w_i}{k} \frac{k!}{(w_1!)(w_2!)...(w_i!)...(w_k!)} = \frac{w_i}{k} \binom{k}{w_1\ w_2\ w_3\ ...\ w_l}$$

Substituting this expression in 12, we get:

$$\sum_{R_k \in \pi(B)} r_1 = \frac{1}{k} \binom{k}{w_1\ w_2\ w_3\ ...\ w_l} (c_1 w_1 + c_2 w_2 + ... + c_l w_l)$$

But note that:

$$c_1 w_1 + c_2 w_2 + ... + c_l w_l = r_1 + r_2 + ... + r_k = n$$

Hence, putting all together, 11 becomes:

$$\sum_{R_k \in \pi(B)} (r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k}) = \frac{n}{k} \binom{k}{w_1\ w_2\ w_3\ ...\ w_l} (t_{j_1} + t_{j_2} + ... + t_{j_k})$$

This means 10 becomes:

$$\sum_{B \in <<n>>^k} P(k, B) \sum_{R_k \in \pi(B)} (r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k})$$

$$= \sum_{B \in <<n>>^k} P(k, B) \frac{n}{k} \binom{k}{w_1\ w_2\ w_3\ ...\ w_l} (t_{j_1} + t_{j_2} + ... + t_{j_k}) \tag{13}$$

Note that the sum $t_{j_1} + t_{j_2} + ... + t_{j_k}$ and the constant $\frac{n}{k}$ are independent of B. Hence, we can only focus on:

$$\sum_{B \in <<n>>^k} P(k,B) \binom{k}{w_1\ w_2\ w_3\ ...\ w_l}$$

$$= \sum_{B \in <<n>>^k} \binom{n}{b_1\ b_2\ b_3\ ...\ b_k} \frac{(N)^n}{N^n} \frac{1}{(n)^k (N-k)^{n-k}} \binom{k}{w_1\ w_2\ w_3\ ...\ w_l}$$

$$= \frac{(N)^n}{N^n} \frac{1}{(n)^k (N-k)^{n-k}} \sum_{B \in <<n>>^k} \binom{n}{b_1\ b_2\ b_3\ ...\ b_k} \binom{k}{w_1\ w_2\ w_3\ ...\ w_l}$$

The inner sum is precisely $k!S(n,k)$ from its definition. Hence, we have:

$$\sum_{B \in <<n>>^k} P(k,B) \binom{k}{w_1\ w_2\ w_3\ ...\ w_l} = \frac{(N)^n}{N^n} \frac{k!S(n,k)}{(n)^k (N-k)^{n-k}}$$

Hence, 13 which is referring to 10 becomes:

$$\sum_{B \in <<n>>^k} P(k,B) \sum_{R_k \in \pi(B)} (r_1 t_{j_1} + r_2 t_{j_2} + ...r_k t_{j_k})$$

$$= \frac{n}{k} (t_{j_1} + t_{j_2} + ... + t_{j_k}) \frac{(N)^n}{N^n} \frac{k!S(n,k)}{(n)^k (N-k)^{n-k}}$$

This was precisely the innermost sum in 9. Hence it becomes:

$$\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k,R_k)(J_k \otimes R_k)$$

$$= \sum_{k=1}^{n} \sum_{J_k \in \{\{n\}\}^k} \frac{n}{k} (t_{j_1} + t_{j_2} + ... + t_{j_k}) \frac{(N)^n}{N^n} \frac{k!S(n,k)}{(n)^k (N-k)^{n-k}} \tag{14}$$

Let us do some simplifications using factorial forms. We have:

$$\frac{n}{k} \frac{(N)^n}{N^n} \frac{1}{(n)^k (N-k)^{n-k}} = \frac{n}{k} \frac{N!}{N^n (N-n)!} \frac{(N-n)!}{(n)^k (N-k)!}$$

$$= \frac{n}{k} \frac{(N)^k}{N^n} \frac{1}{(n)^k}$$

Hence, 14 becomes:

$$\sum_{k=1}^{n} \sum_{J_k \in \{\{n\}\}^k} \frac{n}{k} (t_{j_1} + t_{j_2} + ... + t_{j_k}) \frac{(N)^n}{N^n} \frac{k!S(n,k)}{(n)^k (N-k)^{n-k}}$$

$$= \sum_{k=1}^{n} \sum_{J_k \in \{\{n\}\}^k} (t_{j_1} + t_{j_2} + ... + t_{j_k}) S(n,k) \frac{n}{k} \frac{(N)^k}{N^n} \frac{k!}{(n)^k}$$

$$\sum_{k=1}^{n} S(n,k)\frac{n}{k}\frac{(N)^k}{N^n}\frac{1}{(n)^k}\left(k!\sum_{J_k\in\{\{n\}\}^k}(t_{j_1}+t_{j_2}+...+t_{j_k})\right) \tag{15}$$

Note $t_{j_1}+t_{j_2}+...+t_{j_k}$ will be same for any permutation of J_k. And for every J_k, there are precisely $k!$ permutations equivalent to it. Hence, the inner sum can be written as:

$$k!\sum_{J_k\in\{\{n\}\}^k}(t_{j_1}+t_{j_2}+...+t_{j_k})=\sum_{J_k\in\{n\}^k}(t_{j_1}+t_{j_2}+...+t_{j_k})$$

This step provides another powerful way in which the summation can be simplified. The summation is now over all ordered k-tuples. Hence using symmetry, we can say that $\sum_{J_k\in\{n\}^k} t_{j_i}$ will be same for all $i\in[K]$. Hence, the inner sum becomes:

$$\sum_{J_k\in\{n\}^k}(t_{j_1}+t_{j_2}+...+t_{j_k})=k\sum_{J_k\in\{n\}^k}t_{j_1}$$

t_{j_1} takes values over $t_1,t_2,...,t_n$. Let us say it is t_1 i.e., $j_1=1$. How many such ordered tuples J_k are in $\{n\}^k$? Precisely $(n-1)^{k-1}$ since the remaining $k-1$ positions can be filled by the remaining $n-1$ indices with their order being mattered. Similar logic can be applied when j_1 takes other values too. Hence, the inner sum becomes:

$$\sum_{J_k\in\{n\}^k}(t_{j_1}+t_{j_2}+...+t_{j_k})=k(n-1)^{k-1}(t_1+t_2+...t_n)$$

This means 15 which refers 14 becomes:

$$\sum_{k=1}^{n}\sum_{J_k\in\{\{n\}\}^k}\frac{n}{k}(t_{j_1}+t_{j_2}+...+t_{j_k})\frac{(N)^n}{N^n}\frac{k!S(n,k)}{(n)^k(N-k)^{n-k}}$$

$$=\sum_{k=1}^{n}S(n,k)\frac{n}{k}\frac{(N)^k}{N^n}\frac{1}{(n)^k}k(n-1)^{k-1}(t_1+t_2+...t_n)$$

$$\sum_{k=1}^{n}S(n,k)\frac{n(N)^k}{N^n}\frac{1}{(n)^k}(n-1)^{k-1}(t_1+t_2+...t_n) \tag{16}$$

The term $t_1+t_2+...+t_n$ is independent of k and hence, can be taken outside the summation. We then have:

$$\sum_{k=1}^{n}S(n,k)\frac{n(N)^k}{N^n}\frac{1}{(n)^k}(n-1)^{k-1}=\sum_{k=1}^{n}S(n,k)\frac{n(N)^k}{N^n}\frac{(n-k)!(n-1)!}{n!(n-k)!}$$

$$=\sum_{k=1}^{n}S(n,k)\frac{(N)^k}{N^n}=1$$

We already saw that the above sum is equal to 1 in the previous section. Hence, 16 which actually refers to 9 becomes:

$$\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k, R_k)(J_k \otimes R_k) = \sum_{i=1}^{n} t_i$$

Hence the original statement we intended to prove is derived.

4.5 Convexity property:

The previous two statements are very important since they let us apply the property of convexity in a very useful form. Once again, $g(\cdot)$ is defined as:

$$g(t_1, t_2, ..., t_n) = \sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k, R_k, J_k) f(J_k \otimes R_k)$$

We have shown that $P(k, R_k)$ add to 1. Clearly, they are non-negative quantities too, evident from their expression. And since $f(\cdot)$ is a convex function, we have:

$$g(t_1, t_2, ..., t_n) \geq f\left(\sum_{k=1}^{n} \sum_{R_k \in <n>^k} \sum_{J_k \in \{\{n\}\}^k} P(k, R_k, J_k)(J_k \otimes R_k)\right) = f\left(\sum_{i=1}^{n} t_i\right)$$

Hence, this implies:

$$g(a_{y_1}, a_{y_2}, ..., a_{y_n}) \geq f\left(\sum_{i=1}^{n} a_{y_i}\right)$$

Or in other words:

$$g(Y_1, Y_2, ..., Y_n) \geq f\left(\sum_{i=1}^{n} Y_i\right)$$

Since expectation preserves comparison order, substituting this back in 5 we get:

$$E[g(Y_1, Y_2, ..., Y_n)] = E\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \geq E\left[f\left(\sum_{i=1}^{n} Y_i\right)\right]$$

which is exactly the main result 4 we wanted to prove in this article.

5 How it extends Hoeffding's inequality to Samples obtained without replacement:

Note that in 3, the upper bound is of the form $\frac{E\left[\exp\left(\lambda \sum_{i=1}^{n} \tilde{X_i}\right)\right]}{\exp(\lambda n t)}$. Now instead of IID samples X_i, if we had Y_i, using the main result we proved, we would have had:

$$\frac{E\left[\exp\left(\lambda \sum_{i=1}^{n} \tilde{Y_i}\right)\right]}{\exp(\lambda n t)} \leq \frac{E\left[\exp\left(\lambda \sum_{i=1}^{n} \tilde{X_i}\right)\right]}{\exp(\lambda n t)}$$

This works since $f(x) = \exp(\lambda x)$ is a convex function and hence, the main result applies. Once we have the Chernoff bound involving IID RVs, it can be factorized into product of individual expectations and the steps following it based on Hoeffding's lemma can be applied.

6 Conclusion:

This article is not a new contribution or anything like that. This is just my attempt at rigorously deriving the Hoeffding's lemma for sampling without replacement by proving the intermediate main result involving expectations of convex function applied on sample averages. In particular, this study helped my understand the combinatorial nuances present in this problem and entirely opened me up to the concept of Sterling numbers. Rigorously writing this all down helped me understand the intricate details one should understand when working on a problem like this.

7 Appendix:

7.1 Sterling numbers of Second Kind:

Sterling numbers of Second kind $S(n,k)$ is a useful combinatorial quantity that measures how many distinct, unlabeled partitions of size k can be made from a set with n distinct elements. For example, if $n = 3, k = 2, S = \{a, b, c\}$, then the possible partitions are:

1. $\{a\}, \{b, c\}$
2. $\{b\}, \{a, c\}$
3. $\{c\}, \{a, b\}$

Thus $S(3,2) = 3$. By convention, $S(n,0) = 0$ since for any $n > 0$, it is impossible to have partitions of size 0. It is possible to write down a concrete formula for $S(n,k)$ using multinomial coefficients. The construction can be done as follows:

1. First fixate on a count tuple $R_k = (r_1, r_2, ..., r_k) \in << n >>^k$.
2. For such a tuple, find how many unique partitions can be made.

For a given tuple R_k, this is extremely simple. We are trying to place n distinct objects into boxes of sizes $r_1, r_2, ..., r_k$ and the order does not matter, i.e., the boxes are unlabeled. Hence, all we need to do is choose what objects go into box 1, what objects go into box 2 from the remaining and so on. This is simply:

$$\binom{n}{r_1}\binom{n-r_1}{r_2}...\binom{r_k}{r_k} = \binom{n}{r_1\ r_2\ ...\ r_k}$$

There is a slight caveat in this calculation. What if some counts are equal? For example, assume $n = 5, k = 3, r_1 = 2, r_2 = 2, r_3 = 1, S = \{a, b, c, d, e\}$. When we are placing the objects into different boxes based on the formula above, we are counting the following two cases as distinct cases:

- $\{a, b\}, \{c, d\}, \{e\}$
- $\{c, d\}, \{a, b\}, \{e\}$

Since the boxes are unlabeled, the two cases actually correspond to the same event. This double counting happens only when some of the counts are equal. To correct for this double counting, let $B = \{b_1, b_2, ..., b_l\}$ denote the set of unique elements present in the tuple R_k and let $c_1, c_2, ..., c_l$ be their counts. That is we have: $1 \leq l \leq k$, $c_1, c_2, ..., c_l \in \mathbb{N}$ and $c_1 + c_2 + ... + c_l = k$. Take boxes with count b_1. There are c_1 such boxes. Any rearrangement of those boxes correspond to the same partition and there are $c_1!$ such rearrangements. Using similar logic, we can say the corrected term will be:

$$\binom{n}{r_1}\binom{n-r_1}{r_2}...\binom{r_k}{r_k}\frac{1}{(c_1)!(c_2)!...(c_l)!} = \frac{\binom{n}{r_1\ r_2\ ...\ r_k}}{(c_1)!(c_2)!...(c_l)!}$$

Hence, for a given $R_k \in << n >>^k$, we know how many unique partitions are possible. Now we simply need to add this term over all R_k. Hence, we finally have:

$$S(n,k) = \sum_{R_k \in <<n>>^k} \frac{\binom{n}{r_1\ r_2\ ...\ r_k}}{(c_1)!(c_2)!...(c_l)!} \tag{17}$$

7.2 Relation to Sum of Multinomial Coefficients with no Zeros:

Let us understand how $S(n,k)$ relates to sum of multinomial coefficients with no zeros. In particular, we are interested in the sum:

$$\sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ ...\ r_k}$$

Note that the above summation can be written as:

$$\sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ ...\ r_k} = \sum_{M \in <<n>>^k} \sum_{R_k \in \pi(M)} \binom{n}{r_1\ r_2\ ...\ r_k}$$

The multinomial coefficient will be the same for all $R_k \in \pi(M)$. Hence, the inner sum simply becomes:

$$\sum_{M \in <<n>>^k} \sum_{R_k \in \pi(M)} \binom{n}{r_1\ r_2\ ...\ r_k} = \sum_{M \in <<n>>^k} \binom{n}{m_1\ m_2\ ...\ m_k} |\pi(M)|$$

where $M = (m_1, m_2, ..., m_k)$. The cardinality of $\pi(M)$ is simply the total number of unique permutations of a count tuple $M \in << n >>^k$. Assuming the unique elements of M are captured in set $B = \{b_1, b_2, ..., b_l\}$ with counts $c_1, c_2, ...c_l$, the total number of unique permutations of M is simply $\binom{k}{c_1\ c_2\ ...\ c_l}$. This is because we have l distinct objects with counts $c_1, c_2, ..., c_l$ which we want to rearrange over k slots. Hence, we have:

$$\sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ ...\ r_k} = \sum_{M \in <<n>>^k} \binom{n}{r_1\ r_2\ ...\ r_k} \binom{k}{c_1\ c_2\ ...\ c_l}$$

Expanding the second term, we get:

$$\sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ ...\ r_k} = \sum_{M \in <<n>>^k} \binom{n}{m_1\ m_2\ ...\ m_k} \frac{k!}{(c_1)!(c_2)!...(c_l)!}$$

$$= k! \left(\sum_{M \in <<n>>^k} \binom{n}{m_1\ m_2\ ...\ m_k} \frac{1}{(c_1)!(c_2)!...(c_l)!} \right)$$

From the expression we have for $S(n,k)$ in 17, the above expression simply becomes:

$$\sum_{R_k \in <n>^k} \binom{n}{r_1\ r_2\ ...\ r_k} = k!S(n,k) \tag{18}$$

Hence, the sum of multinomial coefficients with no zeros cleanly connects to $S(n,k)$ in a very simple final expression.

7.3 Polynomial Identity involving Falling Factorials:

$S(n,k)$ also connects to falling factorial terms $(x)^k = x(x-1)...(x-k+1)$ to provide a very useful polynomial identity. In particular, we have the following statement:

$$\sum_{k=1}^{n} (x)^k S(n,k) = x^n$$

for any natural number $x \geq n$. To prove this, we will count the size of the same object in two different ways. Consider a domain set A of size n and a co-domain set B of size x. The total number of functions that can be defined from A to B is x^n. Now let us count the same number in a different way:

1. First fixate on how many distinct images we want in the range, i.e., k
2. Next we will count how many partitions of that size k of the domain set A are possible.
3. Next we will count for each such partition, how many unique output assignments can be done.

For the first step, it is very easy to see the possible values for k are $1 \leq k \leq n$. Now for a given k, the total number of unique partitions of size k on the domain set A are simply $S(n,k)$ from its definition. And now we need to assign output values/labels. We'll need to select k values from B, i.e., x objects and also which object is assigned to which partition element matters, i.e., rearrangements of the picked output values are treated as distinct cases. This rearrangement can be done in $k!$ ways. In other words, for every such partition, there are $\binom{x}{k}k! = (x)^k$ output labellings.

Hence, the total number of functions counted in this way is:

$$\sum_{k=1}^{n} (x)^k S(n,k)$$

This must be equal to x^n and thus, the result we wanted to prove is derived.