Proseminar SS 99
Prof. Dr.-Ing . Gerhard Weikum

”Datenreduktionstechniken”

Thema 1 :	**Sampling Techniken**
Bearbeitung :	**Vanessa WALTER**
Betreuer.:	**Dipl.-Inform. Arnd Christian König**

1. EINLEITUNG

Zu den ” Datenreduktionstechniken” liegt das Thema "Sampling Techniken" vor.

Als Vorlage für das Seminar werden nachstehende Litaraturhinweise angeführt:

- Yibei Ling, Wei Sun, An Evalation of Sampling–Based Size Estimation Methods for Selections in Database Systems, International Conference on Data Engineering, 1995

- Peter J. Haas, Arun N. Swami, Sequential Sampling Procedures for Query Size Estimation, ACM SIGMOD Conference, 1992

- Richard J. Lipton, Jeffrey F. Naughton, Donovan A. Schneider, Practical Selectivity Estimation through Adaptive Sampling

Das Wort ”Sample” bedeutet Stichprobe. Wobei wir unter ”Sampling Techniken ” nicht die Anwendungen im Zusammenhang zu Sampling beim Digitalisieren von Analogsignalen (Abtastrate/Sampling-Rate, Sampling-Tiefe z.B. beim Digitalisieren von Musik) betrachten.

Thema sind die Sampling-Techniken zu den Stichprobenverfahren für die Anwendungen:

- Politik : Wahlforschung und Analysen
- Wirtschaft : Marketing, Qualitätsforschung, Personal-, Produktionsplanungen
- Soziologie : Verhaltensforschung
- Medizin : Arzneimittelforschung
- Medien : Konsumentenforschung, Programmqualität
- Staat/Verwaltung : Amtliche Statistiken
- Informatik : Anforderungsanalysen bei Softwareentwicklungen, Technikbewertungen

2. STICHPROBENVERFAHREN

Mit den Sichprobenverfahren werden aus einenem geringen Prozentsatz gesampelter Daten Rückschlüsse und Aussagen auf die Gesamtmenge ermittelt. Es werden z.B. nur einige hundert Bürger befragt, um die Tendenz der Gesamtheit der Bürger zu ermitteln. Es besteht ein Zusammenhang zwischen der Genauigkeit und der Anzahl der Proben (Sampels). Im Normalfall verursachen mehr Sampels höhere Kosten, dafür ist das Ergebnis genauer.

Vorteile und Gründe, warum Stichprobenverfahren gegenüber einer kompletten Zählung oder Auswertung bevorzugt oder erforderlich werden:

- Geringe Kosten
- Eine grobe Schätzung kann ausreichend sein
- Durchführung von Untersuchungen, bei denen der Untersuchungsgegenstand zerstört wird (Materialprüfung, Qualitätskontrolle)
- Schnelle Durchführung, somit größere Aktualität
- Endlose dynamische Vorgänge können bei einer kompletten Zählung nicht erfasst werden
- Eine unvollständige Zählung kann einen nonsampling Fehler hervorrufen, der größer ist als bei einem Stichprobenverfahren

Nachteile der Stichprobenverfahren

- Fehler bei der Übertragung von Ergebnissen aus der Stichprobe auf die Gesamtmenge (z.B. durch das Auswahlverfahren)
- Probleme mit der Repräsentativität der Stichprobe bei einer stark heterogenen Gesamtmenge

Zunehmende Bedeutung erhalten Sampling-Selektionen in Verbindung mit Data Management Systems (DBMS's). Für Anfragen in Datenbanksystemen werden verschiedene Sampling Methoden / Techniken angewandt. Zur Optimierung von Zeit und Kosten werden Anfrage-Opimierer (Query Optimizer) eingesetzt. Ziel ist es den kostengünstigsten Ablaufplan zu finden.

Mit Sampling–basierenden Methoden kann die Größe eines Anfrageergebnisses (Anzahl der Tupel, die die Anfragebedingungen erfüllen) ermittelt werden, die erforderlich sind für:

- Effizient anzufragen in Relationalen Datenbank-Systemen
- Planungen zu Speicherbedarf und Systemressourcen der Datenbanken
- Audits/Überprüfungen und statistische Studien

Die Sample-Based Methoden selbst:

- Liefern ziemlich genaue Ergebnisse
- Können einfach implementiert werden
- Benötigen keinen Offset für Information über die zu bearbeitende Datenmenge
- Können für eine Vielzahl von Anfragen eingesetzt werden
- Benötigen weniger Zeit, um die Anzahl der Tupel einer Anfrage abzuschätzen, als für die Berechnung des eigentlichen Wertes.nötig ist

Im Vergleich zu den herkömmlichen Methoden (Parametric Methods, Histogramm oder Table Based Nonparametric Methods) werden nicht gebraucht:

- Speicherung und Wartung von detaillierten Gesamtstatistiken über die Daten
- Zusätzliches Wissen / Annahmen über die zugrundeliegende Datenverteilung

3. KLASSIFIKATION DER SAMPLING-METHODEN

Sampling-Techniken werden unterteilt in **Random Sampling (Zufallsauswahl)** und **Systematic Sampling**.

3.1 Random Sampling:

Aus einer Gesamtmenge werden zufällig einzelne Stichproben entnommen, die für eine Auswertung / Studie weiterverwendet werden. Bei der Auswahl sind alle möglichen Proben gleichmäßig wahrscheinlich.

Random Sampling wird weiterhin unterteilt in:

(a) Random Sampling ohne Zurücklegen:
Eine einmal gezogene Zufallszahl (Tupel) steht beim nächsten Experiment nicht mehr zur Verfügung, um zu vermeiden, daß dieses Tupel wieder gezogen wird.

(b) Random Sampling mit Zurücklegen:
Eine bereits gewähltes Tupel muß wieder in die Gesamtmenge zurückgelegt werden für den weiteren Prozessablauf, d.h. jede Ziehung erfolgt auf der Basis der ursprünglichen Gesamtmenge. Dieser Fall ist der einfachere, da sich die Verhältnisse der Gesamtmenge während des Ziehungsprozesses nicht verändern.

Die Zufallsauswahl mit echten Zufallszahlen läßt sich am besten am Urnenmodell verdeutlichen, bei dem die Zufallszahlen der Stichprobe aus einer Urne gezogen werden, die alle zur Auswahl stehenden Zahlen enthält.
zu (a): In einer Urne befinden sich 4 gleichartige Kugeln mit den Nummern 1,2,3,4. Eine Kugel wird gezogen, ihre Nummer festgestellt. Die Kugel wird nicht wieder zurückgelegt. Der Versuch wird wiederholt.
zu (b): In einer Urne befinden sich 4 gleichartige Kugeln mit den Nummern 1,2,3,4. Eine Kugel wird gezogen, ihre Nummer festgestellt und anschließend wieder zurückgelegt. Der Versuch wird wiederholt.

3.1 Systematic Sampling:

Systematic sampling verwendet anstatt eines einzelnen Tupels Datenblöcke / Datenseiten als Sampeleinheit. Ausgehend von einem zufällig ausgewählten Startelement werden alle übrigen Stichprobenelemente systematisch ermittelt. Die Ergebnisse hängen stark von der Datenverteilung ab. Insbesondere wenn die Daten gebündelt (clustered) sind, können die gesampelten Daten unzuverlässig sein. Vorteil dieser Sampling Methode ist die hohe Effizienz bei der Datenauswertung.

Beispiel: Auswahl einer Bevölkerungsstichprobe von 1000 Personen.
Aus einer Datei wird immer die k-te Adresse ausgewählt, wobei sich der Abstand zwischen den Adressen durch k = N / n (N = Größe der Gesamtmenge, n = Größe der Stichprobe) ergibt. Die Startadresse wird dabei zufällig aus den ersten k-Adressen gezogen.

Im weiteren Verlauf wird zu der Auswahl der Proben (Sampels) ausschließlich auf das Auswahlverfahren "Zufallsauswahl (Random Sampling) " Bezug genommen. Merkmal dieses Verfahrens ist die Angabe einer Auswahlregel, mit der die Wahlscheinlichkeit für die Auswahl eines Samples aus der Gesamtmenge berechnet werden kann. Eine Auswahl aufs Geradewohl scheidet aus, da nur die Zufallsauswahl die Anwendungen inferenz-statistischer Techniken erlaubt.

4. BESTIMMUNG DES STICHPROBENUMFANGS

Die Vorlagen von Haas und Swami, Lipton, Naughton und Schneider, und Ling und Sun beschreiben Sequential Sampling Methoden. Der Unterschied zwischen den Methoden liegt hauptsächlich in den verschiedenen Möglichkeiten zur Bestimmung der Anzahl der Tupel , die bezüglich einer vorgegebenen Genauigkeit und Konfidenz genommen werden müssen.

Verwendet werden für die folgenden Notationen:

m	Gesamtmenge an Tupeln
n	Stichprobengröße
$1-\alpha$	Vorgegebene Konfidenz
$\bar{Y}$	Mittelwert aller Tupel
$\bar{y}$	Mittelwert der Stichprobe
V	Varianz der Gesamtmenge
e	erforderliche relative Genauigkeit der Abschätzung
$t_{1-\alpha}$	Abszisse der Normalverteilung an der Stelle 1-α
S_n	Anzahl der gesampelten Tupel, welche die Auswahlbedingung erfüllen
Y_n	die geschätzte Größe

f ( )   charakteristische Funktion einer Auswahl

$x_i Tupel$ (Zufallsvariable)

$$y_i = f(x_i) = \left\{ \begin{array}{ll} 1 & falls\ x_i\ die\ Auswahlbedingung\ erfüllt \\ 0 & sonst \end{array} \right\}$$

Bei Stichproben stimmen die ermittelten Ergebnisse (z.B. Mittelwert, Varianz) nicht mit den eigentlichen Unbekannten in der Gesamtmenge überein. Zufallsauswahlen ermöglichen es aber Aussagen darüber zu treffen mit welcher Irrtumswahrscheinlichkeit der unbekannte Erwartungswert der Gesamtmenge in einem bestimmten vorgegebenen Intervall $[\bar{y}-d, \bar{y}+d]$ liegt. Der wahre Wert der Gesamtmenge läßt sich also exakt eingrenzen.

Nun hängt d und damit die Maßzahl für die Größe dieses Konfidenzintervalls auch von der Größe der Stichprobe ab. Es ist erkennbar, daß mit Erhöhung des Stichprobenumfangs auch eine Erhöhung der Genauigkeit des Mittelwertes $\bar{y}$ in Bezug auf den Mittelwert der

Gesamtmenge $\bar{Y}$ und damit eine Verkleinerung des Konfidenzintervalls verbunden ist. Durch die Abhängigkeit der Intervallgröße von der Stichprobengröße kann festgestellt werden, wie groß der Stichprobenumfang einer Untersuchung sein muß, damit bei einer verlangten Genauigkeit eine bestimmte Irrtumswahrscheinlichkeit nicht überschritten wird. Bei gegebenen Genauigkeitsanforderungen läßt sich der benötigte Stichprobenumfang also bereits im voraus exakt bestimmen.

Hierzu folgende Ableitungen:

4.1 Mittelwert- und Varianzschätzung

Ausgangsfrage: Welche Beziehung besteht zwischen dem Mittelwert einer Stichprobe

$$\bar{y} = \sum_{i=1}^{n} y_{k_i} \ / \ n = \sum_{i=1}^{n} f(x_{k_i}) \ / \ n$$

und dem Mittelwert über allen Tupeln

$$\bar{Y} = \sum_{i=1}^{m} f(x_i) \ / \ m \quad ?$$

(a) Der **Stichprobenmittelwert** ist in reinen Zufallsproben eine erwartungstreue (d.h. unverzerrte) Schätzung des Mittelwertes der Gesamtmenge, d.h. die aus den einzelnen Stichproben berechneten Mittelwerte verteilen sich “gleichmäßig” um den wahren Mittelwert.

Herleitung:

$$\begin{aligned}
E(\bar{y}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n} E\left(\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \bar{Y} \\
&= \frac{1}{n} n \ \bar{Y} \\
&= \bar{Y}
\end{aligned}$$

Die Häufigkeitsverteilung von Stichproben aus zufälligen, voneinander unabhängigen Stichproben nähert sich mit steigender Stichprobengröße immer stärker der Normalverteilung an (Zentraler Grenzwertsatz).

Neben dem Erwartungswert für den Mittelwert als Kennwert zur Beschreibung der “zentralen Tendenz” ist ebenso die Varianz für den Stichprobenmittelwert zur Beschreibung seiner “Streuung” von Bedeutung.

(b) Für identisch und unabhängig verteilte Zufallsvaribalen mit dem Erwartungswert $\bar{Y}$ und der **Varianz** V **der Gesamtmenge mit**

$$V = \frac{\sum_{i=1}^{m} (f(x_i) - \bar{Y})^2}{m-1}$$

gilt für die **Varianz des Stichprobenmittelwertes**:

$$V(\bar{y}) = \frac{V}{n}$$

(c) Daraus folgt für die **Standardabweichung**, d.h. für die durchschnittliche Abweichung eines Stichprobenmittelwertes vom wahren Mittelwert:

$$SE(\bar{y}) = \frac{\sqrt{V}}{\sqrt{n}}$$

4.2 Das Konfidenzintervall für den Erwartungswert

Nach dem zentralen Grenzwertsatz verteilen sich die Stichprobenmittelwerte bei unabhängigen Stichproben gemäß der Normalverteilung. Durch Standardisierung der Zufallsvariablen $\bar{y}$ mit dem wahren Mittelwert $\bar{Y}$ und der Standardabweichung $SE(\bar{y})$ erhalten wir die **standardnormalverteilte Zufallsvariable**

$$\frac{\bar{y} - \bar{Y}}{\frac{\sqrt{V}}{\sqrt{n}}}$$

mit dem Mittelwert 0 und der Varianz 1. Für eine vordefinierte relative Einschätzgenauigkeit e und Konfidenz

$$1 - \alpha$$

kann der Betrag der Samples, die notwendig sind, um die erforderliche Genauigkeit e und die Konfidenz $1 - \alpha$ einzuhalten, durch folgende Formel berechnet werden:

$$Pr\left(\frac{|\bar{y} - Y|}{\sqrt{V} / \sqrt{n}} \leq \frac{e * \bar{Y}}{\sqrt{V} / \sqrt{n}}\right) = 1 - \alpha \quad (1)$$

Oder: Mit Intervallgrenzen k, -k läßt sich aufgrund der bekannten Eigenschaften der Standardnormalverteilung folgende Beziehung für obige Varibale aufstellen:

$$Pr\left(-k \le \frac{\bar{y}-\bar{Y}}{\frac{\sqrt{V}}{\sqrt{n}}} \le k\right) = 1-\alpha \quad (1')$$

Die Wahrscheinlichkeit, daß der Wert der Zufallsvariablen sich zwischen den durch k und - k definierten Intervallgrenzen befindet ist $1-\alpha$.

Ist die Sampelgröße groß genug, so daß der Zentrale Grenzwertsatz angewendet werden kann, so gilt

$$n = \left(\frac{t_{1-\alpha} \cdot \sqrt{V}}{e \cdot \bar{Y}}\right)^2 \quad (2)$$

5. SEQUENTIAL SAMPLING METHODEN

Auf den vorgenannten Herleitungen bauen die S**equential Sampling** Methoden auf. Ein Sequential Sampling nimmt ein Sample nach dem anderen aus der Gesamtmenge heraus (Randomauswahl), berechnet die Anzahl der Samples, die den vorgegebenen Bedingungen bezüglich der Genauigkeit genügen, und überprüft, ob die Abbruchbedingung für den Sampelvorgang erfüllt ist.
Nachstehende Sequential Sampling Methoden werden angeführt, wobei hier unterschiedliche Bestimmungen der Abbruchbedingungen definiert werden.

5.1 Adaptive Sampling nach Lipton, Naughton und Schneider

Zum Adaptive Sampling und seinen Varianten gibt es eine Studie von Lipton, Naughton und Schneider. Die Hauptidee dieses sogenannten LNS-Algorithmus ist das Partitionieren der Anfrage. Um die Größe der Anfrage zu bestimmen, wird zuerst die Antwort zu dieser Anfrage in eine Anzahl von disjunkten Untermengen aufgeteilt, so daß es möglich ist, zufällig eine dieser Untermengen auszuwählen und ihre Größe zu berechnen. Der Sampling-Algorithmus arbeitet, durch wiederholte zufällige Auswahl einer dieser Untermengen, Berechnung der Größe dieser Untermenge und anschließender Einschätzung der Größe des Anfragergebnisses, die sich auf diesen Sampels stützen. Die Abbruchbedingung wird in Form der Summe der Sampels S_n und des Betrages n der Stichprobengröße ausgedrückt. Dies verleiht dem Algorithmus einen adaptiven Charakter: Wenn die Sampels groß sind, werden weniger genommen, sind die Sampels dagegen klein, so werden mehr genommen.

Das Sampling wird fortgesetzt, bis die folgende Abbruchbedingung erfüllt ist:

$$S_n > k_1 \cdot b \cdot (\frac{1}{e} + 1) \cdot \frac{1}{e} \quad (3)$$

wobei $b > V / \bar{Y}$ (b=1 ist eine obere Grenze für $V / \bar{Y}$) und k_1 durch das Konfidenzintervall folgerdermaßen bestimmt wird:

(1) Falls die Sampeldaten ziemlich wilkürlich sind und der Zentrale Grenzwertsatz nicht anwendbar ist, wird

$$k_1 = \frac{1}{1 - \sqrt{(1-\alpha)}}$$

verwendet. Dies ist im wesentlichen eine obere Schranke für k_1 .
(2) Falls die Sampeldaten normalverteilt sind und der Zentrale Grenzwertatz angewendet werden kann, dann kann eine strengere Grenze verwendet werden:

$$k_1 > \left[\Phi^{-1} \left(\frac{1 + \sqrt{(1-\alpha)}}{2} \right) \right]^2$$

Die eingeschätzte Größe kann folgendernaßen bestimmt werden:

$$Y_n = m \cdot \frac{S_n}{n} \quad (4)$$

Die Verwendung einer oberen Schranke b = 1 könnte ineffizient sein, was zu einem übergroßen Sampling führen würde. Deshalb wird bei der verbesserten Methode von Lipton, Naughton und Schneider (LNS-Methode) zuerst ein Mustersampel (Pilot Sample) genommen, um Vorinformationen über V und $\bar{Y}$ zu erhalten und somit eine präzisere Einschätzung von $V / \bar{Y}$ für die Variable b in Gleichung (3) zu erhalten. Als Folge hiervon hat der verbesserte LNS-Algorithmus eine bessere Performance.
Benutzt man die Abbruchbedingung nur für die Summe der Samples, so könnte dies ein Problem darstellen, wenn die Selektivität der Auswahlbedingung klein ist. Der Anzahl der Sampels könnte sogar größer werden als die Größe der Gesamtheit (Over-Sampling) . Um diese ungewöhnliche Situation zu vermeiden, haben Lipton und Naughton die Verwendung des sogenannten "Sanity Bound" (s. Abschnitt 6) eingeführt. Der Samplingprozess terminiert, wenn der Betrag der bereits genommenen Sampels (n) folgende Einschränkung erfüllt:

$$n > k_2 \cdot h^2 \quad (5)$$

wobei h so gewählt ist, so dass die eingeschätzte Größe innerhalb $\frac{100}{h}$ % der Worse-Case Obergrenze der Selektionsgröße liegt, und k_2 von der gewünschten Konfidenz abhängig ist:

(1)Falls die gesampelten Daten ziemlich willkürlich sind und der Zentrale Grenzwertsatz nicht anwendbar ist, wird $k_2 = \frac{1}{\alpha}$ verwendet, was im wesentlichen eine obere Grenze für k_2 darstellt.

(2)Falls die gesampelten Daten normalverteilt sind und der Zentrale Grenzwertsatz anwendbar ist, kann eine strengere Grenze verwendet werden:

$$k_2 > \left[\Phi^{-1}\left(\frac{2-\alpha}{2}\right)\right]^2$$

Adaptive Sampling kann effizient eingesetzt werden, da es möglich ist bei kostenaufwendigen Proben, also wenn sich herausstellt, daß die Anzahl der Sampels zu groß wird, früh abzubrechen.

5.2 Double Sampling nach Hou, Ozsoyoglu und Dogdu

Double Sampling gehört zu den mehrphasigen Auswahlverfahren. Diese zeichnen sich dadurch aus, daß zunächst eine relativ große, einfache Zufallsstichprobe aus der Gesamtheit entnommen wird, deren Informationen für die nächste Phase der Ziehung, bei der eine Unterstichprobe genommen wird, verwertet werden. Der Vorteil hiervon liegt vorallem darin, daß sich die Kenntnisse über die Grobstruktur der Gesamtheit, die mit der ersten Stichprobe ermittelt wurden, durch weitere detaillierte Erhebungen an Unterstichproben verfeinern lassen.

Beispiel:
1.Phase: Erhebung einer kleinen Anzahl von einfach zu ermittelnden Merkmalen in einer großen Stichprobe
2.Phase: Erhebung einer umfangreichen Menge von Merkmalen an einer Stichprobe aus der Ausgangsstichprobe

Bei der nachstehenden Methode nach **Hou**, **Ozsoyoglu** und **Dogdu** wird das Sampling in 2 Stufen (Zweiphasenauswahl) unterteilt. In der ersten Stufe werden m_1 Tupel gesampelt, um die Varianz Vund den Mittelwert der Gesamtmenge $\bar{Y}$ herzuleiten und um den Betrag m_2 der Samples zu berechnen, die insgesamt benötigt werden, wobei man die erforderliche Einschätzgenauigkeit e, Konfidenz $1-\alpha$, und den geschätzten Mittelwert der Gesamtmenge $\bar{Y}$ und die Varianz V aus der ersten Stufe zu verwendet. In der zweiten Stufe Stufe werden falls
(a) $m_2 > m_1$ zusätzliche $m_2 - m_1$ Sampels genommen, um die erforderliche Einschätzgenauigkeit und Konfidenzintervall zu erhalten

(b) $m_2 \leq m_1$ keine weiteren Samples benötigt

Die Anzahl der Sampels, die in dieser zweiten Stufe benötigt werden, wird folgendermaßen berechnet:

$$m_2 = \frac{\left(\frac{t_1 - \alpha}{e}\right)^2 (1 - p_1)}{p_1} + \frac{3}{p_1 \cdot (1 - p_1)} + \frac{t^2_{1-\alpha}}{e^2 \cdot p_1 \cdot m_1} \quad (6)$$

Die Variable p_1 ist der Prozentsatz der Sampels, die der gegebenen Selektion aus der ersten Stufe genügen.

Die Größe der Schätzung kann schließlich berechnet werden durch folgende Formel:

$$Y_n = m \cdot \left(S_n / n - \left(\frac{e}{t_{1-\alpha}}\right)^2 \cdot \frac{S_n}{n - S_n} \right) \quad (7)$$

5.3 Sequential Sampling nach Haas und Swami

Um die Größe eines Anfrageergebnisses zu bestimmen, schlagen Haas und Swami eine weitere Sequential Sampling Methode vor, die im Gegensatz zu den beiden bereits beschriebenen Sequential Methoden (Adaptive Sampling und LNS-Algorithmus) asymptotisch effizient ist. Eine sequentielle Einschätzungsprozedur ist asymptotisch effizient, wenn die Samplingkosten der Prozedur sich den minimalen Samplingkosten nähern, wenn die erforderliche Genauigkeit zunehmend strenger wird. Im Gegensatz zu den vorherigen beschriebenen Methoden wird bei der Sampling Methode von Haas und Swami kein initiales Mustersampel benötigt.

Bei dieser Methode wird nach jedem Sampelschritt die Entscheidung getroffen, ob das Sampling fortgesetzt oder gestoppt werden soll (sequential procedure) in Abhängigkeit von der vorgegebenen Genauigkeit und der Konfidenz.

Der Betrag der Sampels n wird ermittelt durch:

$$n = n(t) = inf\{n \geq 1 : max(S_n, n \cdot d) > t\} \quad (8),$$

wobei t ein fester Parameter ist, der durch die gewünschte Konfidenz und den relativen Fehler folgendermaßen berechnet wird:

$$t = \frac{t^2_{1-\alpha} \cdot V}{e^2 \cdot max(\bar{Y}, d)} \quad (9),$$

Der Parameter d (Sanity Bound) wird verwendet, wenn die Selektivität klein ist (d=0 bedeutet, daß kein Sanity Bound verwendet wird).

Die Abbruchbedingung wird also berechnet durch:

$$n = inf\{n \geq 1 : V_n > 1 , e \cdot max(S_{n,} n \cdot d) \geq t_{1-\alpha}(n \cdot V_n)^{1/2}\} \quad (10)$$

Die Größe der Einschätzung kann mit Hilfe der Formel(4) berechnet werden.
Im Vergleich zu den vorherigen beschriebenen Methoden wird beim Sequential Sampling kein fester Betrag von Sampels verwendet, um die totale mittlere Varianz abzuschätzen sondern dieser dynamisch während des Samplings bestimmt.

Der **Algorithmus des Sequential Samplings** ist folgender:

1. (Initialisierung) Wähle zufällig ein Tupel x und setze n = 1, s = x und w = 0;

2. if w > 0 und $e \cdot max(s,nd) \geq t_{1-\alpha,n}(n \cdot w / (n-1))^{1/2}$, dann stop und Schätzung $y(e) = m \cdot s / n$ zurückgegeben;

3. Wähle zufällig ein weiteres Tupel aus;

4. w_{new} Update mit $\frac{n-1}{n}{}_{wold} + \frac{(s_{old} - n \cdot x)^2}{(n+1)n}$, $s_{new} = s_{old} + x$, inkrementiere n um 1 und return zu Schritt 2;

6. VERGLEICH DER VERSCHIEDENEN METHODEN

Bei einer Studie von Yibei Ling und Wei Sun zum Vergleich der verschiedenen Sampling-Methoden wurden betrachtet:

- Durchschnittliche Anzahl der gesampelten Tupel
- Durchschnittlicher relativer Schätzfehler
- Abdeckung (Coverage = echte Wahrscheinlichkeit, daß die Einschätzgenauigkeit einer Samplingprozedur innerhalb der vordefinierten erforderlichen Genauigkeit liegt)

Berücksichtigt wurden Abläufe mit und ohne Sanity Bound (= Bedingung, die den Fall abfängt, daß die Abbruchbedingung (Stopping Condition) aufgrund von geringer Selektivität keine gute Abschätzung der zu sampelnden Werte liefert).
Gemessen wurden die durchschnittliche Anzahl der Tupel, die benötigt wurden, um die vorspezifizierte Konfidenz und den relativen Schätzfehler bei verschiedenen Selektivitäten, sowie der aktuelle durchschnittliche Abdeckungsbereich.

6.1 **Ohne Sanity Bound**

Bei dem Vergleich wurden angewandt:

- Adaptive Sampling (verbesserte LNS-Methode) ; die Abbruchbedingung basiert auf der Anzahl der Sampels
- Sequential Sampling mit $max(S_{n,} n \cdot d) = S_n$ und d = 0
- Double Sampling

Zusätzliche wurden beim Sequential Sampling die folgenden Varianten bezüglich der Abbruchbedingung verwendet:

- $F_{t,n}^{-1}((2-\alpha) / 2)$
- l = 2
- l = 3

(Die Abbruchbedingung muß l Mal erfüllt sein, bevor das Sampling beendet wird).

Für das Adaptive Sampling und die erste Stufe des Double Sampling wurden 100 Mustersampel genommen.

Nachstehende Beobachtungen:

Anzahl der gesampelten Tupel:

Adaptive Sampling benötigt die größte Anzahl an gesampelten Tupeln und Sequential Sampling die geringste Anzahl, um die erforderliche Genauigkeit und Konfidenz einzuhalten.

Genauigkeit:

Adaptive Sampling scheint die höchste Einschätzgenauigkeit zu erreichen, wahrscheinlich aufgrund seiner relativ hohen Samplegröße.
Die Genauigkeit beim Sequentiel Sampling unter Verwendung der Abbruchbedingung $F_{t,n}^{-1}((2-\alpha) / 2)$ beträgt 10%.

Die Genauigkeit des Sequentiel Sampling bei Verwendung von l=2 ist dagegen nicht gut, da das Sampling möglicherweise zu früh abgebrochen wurde.
Die Verwendung von l=3 zeigt eine Verbesserung gegenüber l=2.
Vorschlag: Verwendung von l=3 bei dieser Methode, um die Genauigkeit zu verbessern.

Covergage:

Die durchschnittliche Abdeckung bei Anwendung des Sequential Sampling beträgt 80%, unabhängig von l. Adaptive Sampling und Double Sampling ohne Sanity Bound dagegen erreichen sogar eine Abdeckung von 100%, aufgrund der relativ hohen Sampelgröße.

<u>Größenskalierbarkeit</u> (Size Scalability):

Mit Hilfe der Skalierbarkeit wird festgestellt, wie sich ein Samplingalgorithmus verhält (Effizienz), wenn die Datenmenge zunimmt.
Die Größenskalierbarkeit ist bei allen Methoden gut und die Beziehungen zwischen der Anzahl der Sampels/relativer Fehler/durchschnittliche Erfassung vs. Selektivität sind fast identisch trotz verschiedener Bezugsgrößen.

6.2 Mit Sanity Bound

Bei geringer Selektivität wird die Anzahl der Sampels bei den Methoden des Adaptive Sampling und Sequential Sampling sehr groß. Um dieses Problem des Over-Sampling Problem zu vermeiden, wurde die Verwendung des Sanity Bound vorgeschlagen. Da für Double Sampling kein Sanity Bound verwendet wird, wird diese Methode im folgenden nicht mehr erwähnt.

Bei den Methoden Adaptive und Senquential Sampling ist die Performance nicht gut, wenn die Selektivität klein ist. Sie nähert sich den Methoden ohne Sanity Bound, bei Zunahme der Selektivität.

<u>Coverage:</u>

Die Abdeckunng der Methoden mit Sanity Bound nähert sich der vorspezifizierten Konfidenz bei Zunahme der Selektivität. Beim Sequential Sampling jedoch ist weicht die Abdeckung weit von der vordefinierten Konfidenz ab, sogar wenn die Selektivität groß ist.
Der Abdeckung bei geringer Selektivität kann jedoch verbessert werden durch Reduzierung des Wertes von d.
Beim Adaptive Sampling kann die Abdeckung bei geringer Selektivität durch Erhöhung des Wertes von h verbessert werden.
Zusammenfassung: Die Einführung des Sanity Bound könnte den ungewöhnlich hohen Betrag der Sampels wenn die Selektivität hoch ist vermeiden, was jedoch zu einer Verschlechterung der Abdeckung und der relativen Einschätzgenauigkeit führt. Die Skalierbarkeit bei Adaptive Sampling und Sequential Sampling mit Sanity Bound ist ausgezeichnet. Die Anzahl der Sampels ist hier fast unabhängig von der Anzahl der Tupel in der gesampelten Relation.

6.3 **Schräge Verteilung der Daten** (Skewed Data Distributions)

Mit “skewed data” bezeichnet man Daten, bei denen die Häufigkeit einzelner Werte unregelmäßig verteilt ist.
Die Performance einer Samplingprozedur ist unabhängig von der Verteilung der Daten, wenn die Selektivität konstant ist. Jedoch haben verschiedene Datenverteilungen eine bedeutende Auswirkung auf die Performance. Für eine schräge Datenverteilung sind Anfragen mit einer geringen Selektivität wahrscheinlicher, was aber zu einem relativ hohen Einschätzfehler und einer geringen Abdeckung führen könnte, ganz gleich welche Samplingmethode angewendet

wird. Bei schräger Datenverteilung sind geringe Selektivitäten wahrscheinlicher, wenn Anfragebereiche randomverteilt sind.

7. ZUSAMMENFASSUNG

Sequential Sampling hat eine bessere Performance aufgrund seiner sparsamen Anzahl der Sampels, obwohl diese Methode die schlechteste Abdeckung hat.
In der Praxis ist die Verwendung des Sanity Bound für die Methoden des Adaptive und Sequential Sampling erforderlich, um ein Over-Sampling zu vermeiden. Die Methoden mit Sanity Bound können die Anzahl der Sampels bei geringer Selektivität reduzieren, was aber eine schlechte Abdeckung und einen größeren relativen Einschätzfehler zur Folge hat.
Eine schräge Datenverteilung scheint die Performance aller Samplig-Methoden zu verschlechtern, da Anfragen mit geringer Selektivität wahrscheinlicher zu erzeugen sind. Alle Sampling-Methoden skalieren extrem gut mit Bezugsgrößen.