Umgang mit Zahlen und Daten
Montag 28. Februar

■ **Datensätze:** Die Ergebnisse von Beobachtungen und Messungen liegen in der Regel in Form von *Datensätzen* vor.

Ein (numerischer) **Datensatz** ist so etwas wie eine zweispaltige Liste oder "Wertetabelle". Dabei ist stark zu unterscheiden, ob die Eintragungen der ersten Spalte (unabhängige Variable) eine inhaltliche Bedeutung haben - etwa Zeitpunkt der Temperaturmessung - oder ob sie reine Numerierungsfunktion für die Zahlwerte der zweiten Spalte haben! Wir sagen im ersten Fall "Auszählungstyp" und im zweiten "Zeitreihentyp" .

Wir betrachten hier vornehmlich den **Auszählungstyp**:

♦ Zerlege den Wertebereich (mögliche Werte) disjunkt in angemessene Teile.

◊ Im diskreten Fall kann das vorgegeben sein durch die überhaupt möglichen Werte, sonst ist künstlich zu diskretisieren.

♦ Die Teile geeignet benennen

♦ Zähle aus, wieviele Daten in jeden dieser Teile fallen

♦ Das gibt die "absoluten Häufigkeiten". Etwa N_i = Zahl der Daten, die in den i-ten Bereich fallen.

♦ Daraus kann man die "relativen Häufigkeiten" bilden: Ist N die Gesamtzahl aller Fälle, und N_i die absolute Häufigkeit, dann ist $n_i = \frac{N_i}{N}$ die "relative Häufigkeit" für den i-ten Bereich.

♦ Veranschaulicht ("graphisch dargestellt") werden beide Typen von Häufigkeiten in der Regel als Histogramme.

◊ Unterschied: Läßt man die Gesamtzahl N wachsen, dann wächst das Digramm für die ansoluten Häufigkeiten mit, wogegen die Form der relativen Häufigkeiten sich stabilisieren sollte!

◊ Aber es gibt noch zahlreiche andere nützliche Veranschaulichungsmethoden. Beispielsweise den Typ "Punktschwarm in der Ebene". (Sowohl Zeitreihe wie auch Auszählungstyp mit Zahlpaar als Wert.

■ **Das Codierungsproblem der beschreibenden Statistik**

Wir betrachten einen numerischen Datensatz vom Auszählungstyp. Er enthält sehr viel Information, wobei in der Regel jedoch viel Unwichtiges mit wenig und schwer erkennbarem Wichtigem vermischt ist.

Ziel: Beschreibe und charakterisiere den Datensatz durch wenige (in der Regel 2) Zahlangaben, die möglichst viel Information liefern sollen:

> Man nimmt dazu "Mittelwert und Streuung". Die erste Zahl beschreibt die Lage, den "Ort" der Daten auf der Zahlachse in Form eines Ersatzpunktes. Die zweite Zahl gibt ein Maß dafür, wie weit die Daten um diesen Ersatzpunkt streuen.

Die Berechnungsformeln:

$$\begin{aligned} a &= (a_1, a_2, ..., a_N) \quad \text{numerischer Datensatz} \\ \bar{a} &= \mu(a) = \frac{1}{N}(a_1 + .. + a_i + ... + a_N) = \frac{1}{N}\sum_{i=1}^{N} a_i \qquad \text{arithm. Mittel} \\ Var(a) &= \frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2 = ... \qquad \sigma(a) = \sqrt{Var(a)} \quad \text{Streuung und Varianz} \end{aligned}$$

♦ Wir wählen ā als neuen Ursprung oder Aufpunkt auf der Zahlengeraden und σ als neue Einheit.

Jetzt können wir die Daten neu codieren, indem wir schreiben

$$\boxed{a_i = \bar{a} + \alpha_i \sigma}$$

Gehe (auf der Zahlgeraden) zum neuen Ursprung $\bar{a}$, und von dort um α_i (in Einheiten von σ) weiter. Entscheidend ist: Die neuen Koordinaten α_i sind Zahlen in der Größenordnung von 1. Völlig unabhängig davon, wie groß die a_i sind.

Überschlägig kann man sagen: Für etwa 50% der Daten hat man $|\alpha_i| < 1$ und für etwa 90% hat man $|\alpha_1| < 2$.

Das ist die gewünschte Umcodierung des Datensatzes!

■ **Wahrer Wert, Schätzwert und Fehler.**
(Systematischer und statistischer Fehler)
Bei Messergebnissen hat man es häufig mit folgender Situation zu tun:

Man mißt eine Größe, von der man animmt, dass sie nach ausreichender Idealisierung einen "wahren Wert" besitzt. Das sei a_{Wahr}. Aber jede Messung ist ungenau. Man macht einen Fehler. Das Messergebnis der i-ten Messung sei a_i. Dann haben wir $a_i = a_{Wahr} + f_i$, wenn f_i der Fehler ist. Man hofft, dass man irgendwie zu a_{Wahr} gelangt, wenn man nur häufig genug misst.

Das ist vielfach ein schwieriges Problem, dass sorgfältiger fallspezifischer Analyse bedarf. (Systematische Fehler). In der Regel kann man wie folgt vorgehen:

♦ Man habe einen Datensatz von n Messungen a_i der gesuchten Größe

♦ Bilde das Datenmittel $\bar{a}$ des Datensatzes. Das ist ein **Schätzwert** für a_{Wahr}.

♦ Jetzt braucht man noch einen Schätzwert für den Fehler. Das ist **fast** die Datensatzstreuung. diese selbst sollte man nich nehmen, sondern stattdessen ("1/(N-1)-Regel!)

$Var_S(a) = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - \bar{a})^2 = ...$	$\sigma_S(a) = \sqrt{Var_S(a)}$	Datenschätzung der Streuung

Erneut sieht die Situation dann wir folgt aus:

$$a_i = \bar{a} + \varepsilon_i \sigma_S(a)$$

a_i Das i-te Messergebnis
$\bar{a}$ bekannt, Schätzwert
$\sigma_S(a)$ bekannt
ε_i Unbekannt, aber Größenordnung 1

Und wie steht es mit a_{wahr}? Es sei N die Gesamtzahl der Messungen des Datensatzes:

$$a_{Wahr} = \bar{a} + \beta \Delta a \qquad \Delta a = \frac{\sigma_A(A)}{\sqrt{N}} \qquad !!!!!!!$$

β Größenordnung 1

Dieser Sachverhalt wird in der Regel in folgender Form verkürzt angegeben/dargestellt:

$$a_{Schätz} = \underbrace{3.71}_{\text{für } \bar{a}} \pm \underbrace{0.05}_{\text{Für } \Delta a}$$

Man gibt also einen Bereich an, in dem, bzw. in dessen Nähe der wahre Wert liegen sollte.

☐ Gegeben ein numerischer Datensatz a von insgesamt N Messungen. Der Messbereich werde in K Teile $I_1, I_2, ..., I_I$ zerlegt. Weiter sei N_i die Zahl der Beobachtungen, di in den i-ten Bereci fallen.

a) Was ist dann $N_1 + N_2 + ... + N_K = \sum_{i=1}^{K} N_i$? (Der Wert kann angeben werden)

b) Weiter sei $h_i = \frac{1}{N} N_i$ die zugehörige relative Häufigkeit. Was ist dann $\sum_{i=1}^{K} h_i$?

☐ Das sind die ersten 100 Stelle der Zahl π. Interpretieren Sie das all Datensatz wobei die i-te Ziffer der i-te Wert a_i ist. also $a_1 = 3$, $a_2 = 1$ usw. Nehmen Sie eine Auszählung der Ziffern vor und bestimmen Sie deren relative Häufigkeiten. Tragen Sie das Resultat als Diagramm auf .

$\pi = 3.\,141\,592\,653\,589\,793\,238\,462\,643\,383\,279\,502\,884\,197\,169\,399\,375\,105\,820\,974\,944\,592\,307\,816\,406\,286\,208\,998\,628\,034\,825\,342\,117\,068$

▼

–	0	1	2	3	4	5	6	7	8	9
N_i	8	8	12	12	10	8	9	7	13	13

$h_i = \frac{N_i}{100}$ bedeutet hier nur Kommaverschiebung. ▲

☐ Gegeben die folgende Datenliste:

$$-66 \quad \mathbf{-136} \quad 128 \quad 27 \quad \mathbf{157} \quad 115 \quad 49 \quad -72 \quad 19 \quad -118$$

Bestimmen Sie Mittelwert ā und Datenstreuung σ. Stellen Sie die Daten in der Form $a_i = \bar{a} + \varepsilon_i \sigma$ dar.

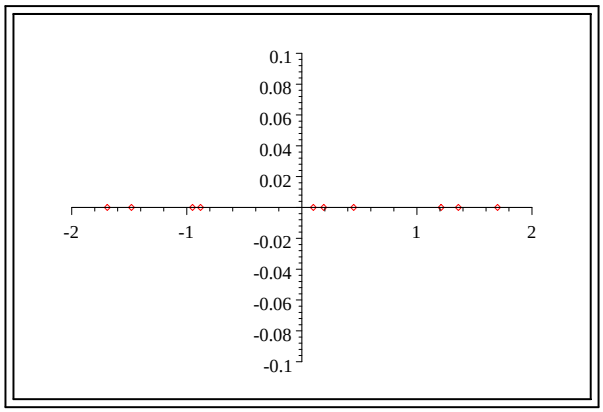
▼ Man findet (per Taschenrechner) ā=10.3 und $\sigma = 86.6$. D. h. als Zweizahlbeschreibung

$$a_{Cod} = 10.3 \pm 86.6$$

Für die datensatzbezogenen Beschreibung $a_i = \bar{a} + \varepsilon_i \sigma$ erhält man folgende Liste der ε_i :

$$-0.88 \quad \mathbf{-1.69} \quad 1.36 \quad 0.19 \quad \mathbf{1.70} \quad 1.21 \quad 0.45 \quad -0.95 \quad 0.10 \quad -1.48$$

In diesem Fall liegen alle Werte zwischen -2 und +2.



Ein Strahl von Teilchen bewegt sich durch eine Flüssigkeit. Bei Eintritt sind es 1000 Stück (x=0). In der Flüssigkeit werden die Teilchen zunehmend absorbiert. Hat der Strahl die Strecke x der Flüssigkeit durchflogen, sind nur noch N(x) Teilchen vorhanden. In bestimmten Abständen sind die Teilchen gezählt. Die Ergebnisse finden Sie in folgender Tabelle:

x	0	.1	.2	.3	.4	.5	.6	.7	.8	.9
N(x)	1000	945	880	855	786	730	659	651	610	582

Bestimmen Sie näherungsweise über diese Tabelle **die relative Absorbtionsrate**. (Genau überlegen, was das bedeutet! Was hat das mit dem Absorbtionskoeffizienten zu tun?) Tragen Sie dann Log(N) gegen x auf! Wie kann man dann dieselbe Größe erneut erhalten? Natürlich sind die Daten mit Fehlern und Unsicherheiten behaftet.

(Das ist ein Datensatz vom Zeitreihentyp. Wieso?)

▼ ▲

□ Für die Varianz eines Datensatzes gilt folgende Formel:

$$Var(a) = \frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2 = \frac{1}{N}\sum_{i=1}^{N} a_i^2 - \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N} a_i\right)^2}_{\text{Das ist } \bar{a}^2}$$

Verstehen Sie zunächst den Rechenweg, der von der rechten Seite beschrieben wird und wie man dort von den Daten zum Resultat kommt. Danach: Wie erhält man die rechten Seite? Indem links $(a_i - \bar{a})^2$ mit dem Binomialsatz ausgerechnet wird. Versuchen Sie diese Rechnung nachzuvollziehen!

Schließlich: Wozu ist die Umformung nutze? Wenn man N erhöht, und dabei verfolgen will, wie sich die Varainz dabei ändert, ist die zweite Form besser handhabbar!

▼ ▲

Beispiel eines Datensatzes vom *Auszählungstyp:*

♦ Nehme eine Buchseite. Numeriere die Zeilen durch. Zähle wieviel Worte in der i-ten Zeile stehen. Bezeichnung W_i. Dann ist $(W_1, ..., W_N)$ ein typischer Datensatz vom Auszählungstyp. Jetzt wird diese Liste nach der Anzahl ausgezählt. n_k bezeichnet dann die Anzahl der Zeilen auf der Seite mit k Worten.

Ein Beispiel mit nur 4 Zeilen ergibt (mathematische Ausdrücke zählen als ein Wort):

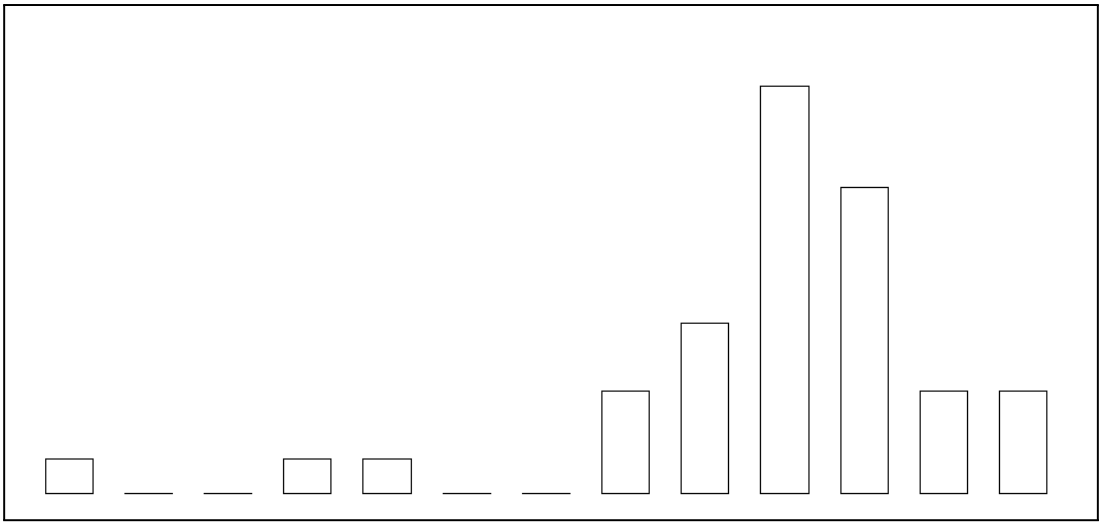
Nehme eine Buchseite. Numeriere die Zeilen durch. Zähle	$W_1 = 8$
wieviel Worte in der i-ten Zeile. Bezeichnung W_i. Dann	$W_2 = 9$
ist $(W_1, ..., W_N)$ ein typischer Datensatz vom Auszählungstyp	$W_3 = 7$
Jetzt wird diese Liste nach der Anzahl ausgezählt. n_k	W_4=9

Auszählung gibt $n_7 = n_8 = n_{10} = 1$ und $n_9 = 2$. Und alle übrigen $n_k = 0$!

Eine Auszählung der Zeilen einer ganzen Seite ergab folgendes Resultat:

$n_1 = n_4 = n_5 = 1$. Und $n_8 = n_{12} = n_{13} = 3$. Und $n_9 = 5$, $n_{10} = 12$, $n_{11} = 9$. Alle weiteren $n_k = 0$.

Wie sieht das als Histogramm aus? (Sie sollten verstehen, was horizontal, was vertikal aufgetragen ist. Erklären Sie das Auftreten der drei Werte für n_1, n_4 und n_5.)



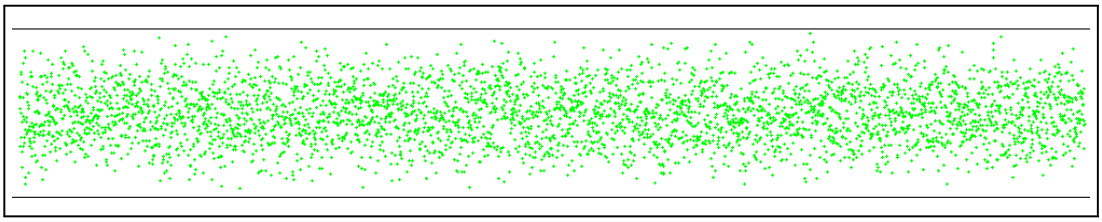
Das ist eine Verteilung mit einem Mittelwert bei etwa 10. Genauer hat man 9.82±2.33.

☐ Zeichnen Sie den Mittelwert an der "richtigen" (geeigenten) Stelle vertikal ein. Dann die Steuung.horizontal nach beiden Seiten vom Mittelwert ausgehend.

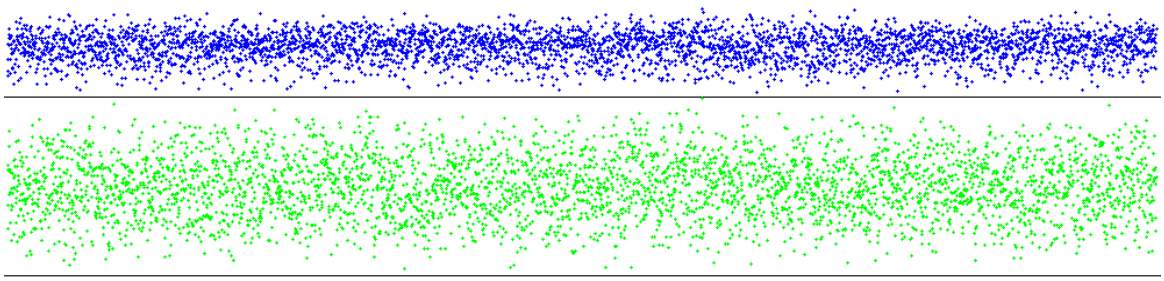
Ergänzung und Erläuterung zur "1/(N-1)-Regel" zur Schätzung der Varianz.

Wir wählen drei Zahlen zufällig und gleichverteilt zwischen 0 und 1 aus und bilden das arithmetische Mittel dieser drei Zahlen. (Etwa $a_1 = 0.25$, $a_2 = 0.83$ und $a_3 = 0.61$ mit $\bar{a}$=0.56)

Das tun wir 1000 Mal. alle diese Mittel liegen zwischen 0 und 1. Die erste Figur zeigt diese 1000 Mittelwerte (horiontal die Nr. der Ziehung, vertikal die Größe des jeweiligen arithm. Mittelwertes zwischen 0 und 1).

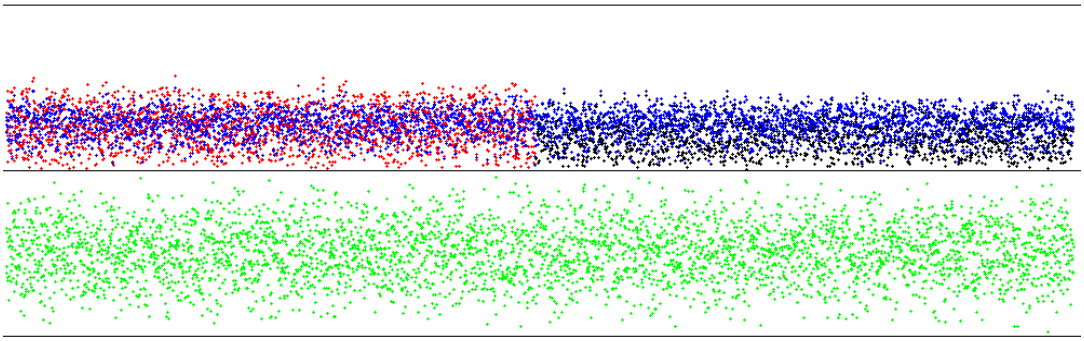


Der "wahre Mittelwert" ist hier natürlich 0.5. Denn wenn man viele (einige hundert) solcher Zahlen zufällig zieht, dann wird deren Mittel gegen diesen Wert streben. Wir bilden jetzt den Unterschied zwischen arithmetischem Mittel der drei Zahlen und dem wahren Mittel 0.5. Das nächste Bild zeigt, wie das in unserem Experiment aussieht:



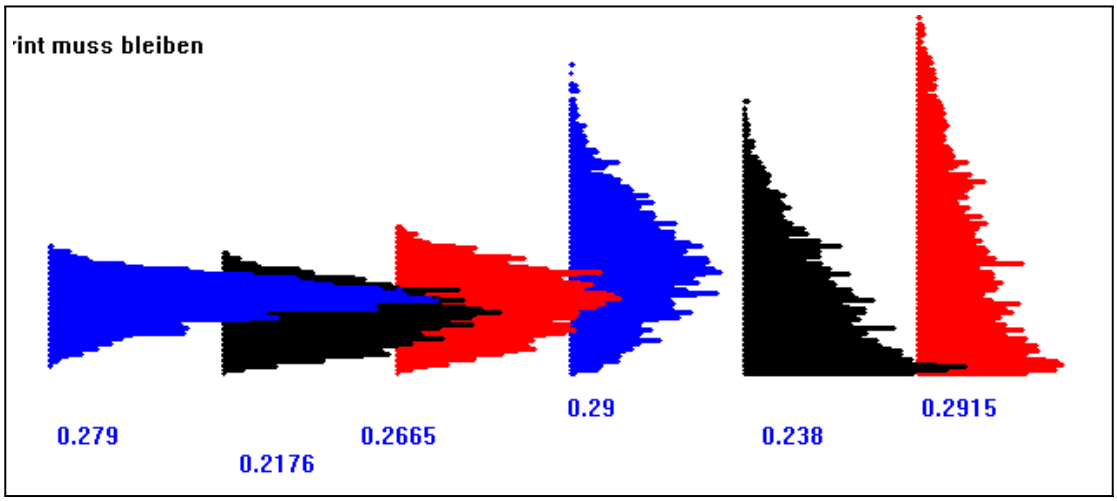
Oben ist blau diese Differenz aufgetragen mit Wert zwischen 0 (unterer Strich) und 1(oberer Strich. Man sieht, dass sich so relativ kleine Werte (gegen 1) ergeben. Das Bild zeigt, wie die arithmetischen Mittel aus den drei Zahlen (Also unser Schätzwert für den wahren Mittelwert von diesem wahren Mittel abweichen, wi sie streuen

Kann man nun die Größe dieser Abweichung mit Hilfe eines Dreierdatensatzes allein schätzen? Das behauptete die 1/(N-1) Regel. Darum berechnen als nächstes einamal die Streuung der Daten mit 1/N und einmal mit 1/(N-1).



Links rot ist für die ersten 500 Fälle die 1/(N-1)-Schaätzung für σ aufgetragen und man sieht, dass blau und rot etwa dieselbe typische Größe haben. Rechts ist dagegen schwarz die nach der 1/N-Regel bestimmte Schätzung aufgetragen und die ist offensichtlich systematisch zu klein! Die schwarzen Punkte liegen zu tief, würden eine zu kleine Schätzung ergeben.

Jetzt nehmen wir noch Auszählungen all dieser Größen vor (Datensatz ist die beschriebene 1000-fache Wahl von drei Zufallszahlen). Die auszählungen werden als Histogramme dargestellt, wobei allerdings die Bin-Werte vertikal, die Anzahlen horizontal aufgetragen werden.



Von links nach rechts: Blau die Abweichungen vom wahren Mittel 0.5. Der Mittelwert ist 0.279.. Daneben schwarz die nach der 1/N - Regel geschätzte Streuung mit der miitleren Vorhersage (über die 1000 Fälle) von 0.218. Dieser Wert ist zu klein. Rot daneben die mit der 1/(N-1)-Regel gefundene Schätzung mit dem viele besseren Wert von 0.267. Weiter rechts daneben noch die Verteilungen für die Quadrate dieser drei Größen!

Fassen wir zusammen: Kennt man nur eine einzige Messreihe von drei Werten (a_1, a_2, a_3), dann ergibt deren arithmetisches Mittel $\bar{a}$ den besten Schätzwert für den wahren Wert. Wie weit aber der (dann unbekannte) wahre Wert und der Schätzwert voneinander entfernt sind, sollte man mit der 1/(N-1)-Regel schätzen, in unserem Fall also über die Formel

$$\sigma_{Schätz} = \sqrt{Var_{Schätz}} \qquad \mathrm{V}_{Schätz} = \frac{1}{2}\sqrt{(a_1-\bar{a})^2+(a_2-\bar{a})^2+(a_3-\bar{a})^2}$$

Ist N größer, sagen wir 10, dann ergeben beide Formeln kaum einen Unterschied. Der Unterschied wird nur für ganz kleine N beachtenswert.

Näherung für den (absoluten) Unterschied:

$$\frac{1}{\sqrt{N-1}} - \frac{1}{\sqrt{N}} = \frac{1}{\sqrt{N}}\left(\left(1-\frac{1}{N}\right)^{-\frac{1}{2}} - 1\right) \approx \frac{1}{\sqrt{N}} \cdot \frac{1}{2}\frac{1}{N}$$

.