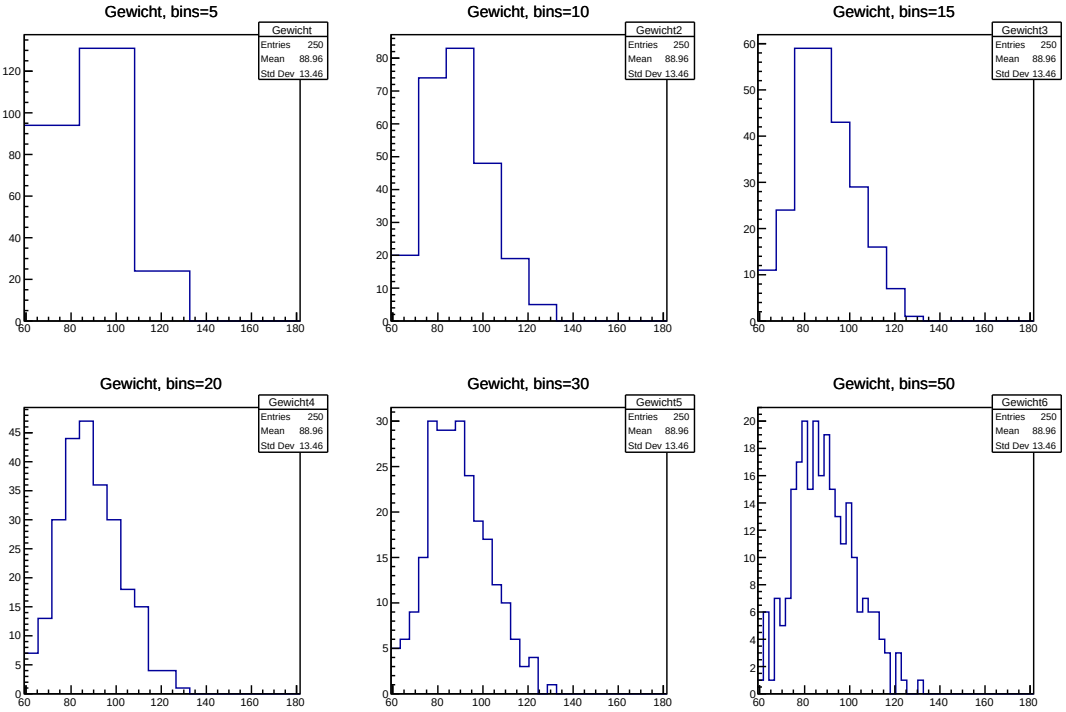
Blatt1

Aufgabe 2



a)

In der obigen Abbildung sind Histogramme resultierend aus den Daten der *Groesse_Gewicht.txt* in Abhängigkeit der Bins abgebildet. Erkennbar ist, dass für eine große Binbreite die Werte sehr grob dargestellt werden. Dies ist vorallem im ersten Plot zu sehen, der nicht eindeutig einem Trend einer Verteilungsfunktion zugeordnet werden kann. Bei zu kleiner Binbreite (letzer Plot) werden die Daten zu genau dargestellt; Die individuellen Ausreisser, welche bei einer Messanzahl von nur 250 unvermeidbar sind (keine aussagekräftige Anzahl), machen sich zu stark bemerkbar. Dies ist nicht zuletzt an den zwei bis mehreren Peaks zu erkennen. Das selbe Argument trifft auf die "Löcher" zu, da diese Löcher in der Realität nicht festgestellt werden können: Es ist sinnlos, dass es keine Menschen gäbe, die ein bestimmtes Gewicht haben, aber es Menschen gibt, deren Gewicht um jenes liegt.

Am besten bietet sich hier ein Mittelweg an, sprich Plot 3 oder 4. Zusätzlich kann hier noch die *Regel von Scott* hinzugezogen werden,

$$h = \frac{3.49 \cdot \sigma}{\sqrt[3]{n}},$$

wobei h der Binbreite entspricht. Für das Gewicht ergibt die Formel ca. 16 Bins, was Plot 3 entspricht.

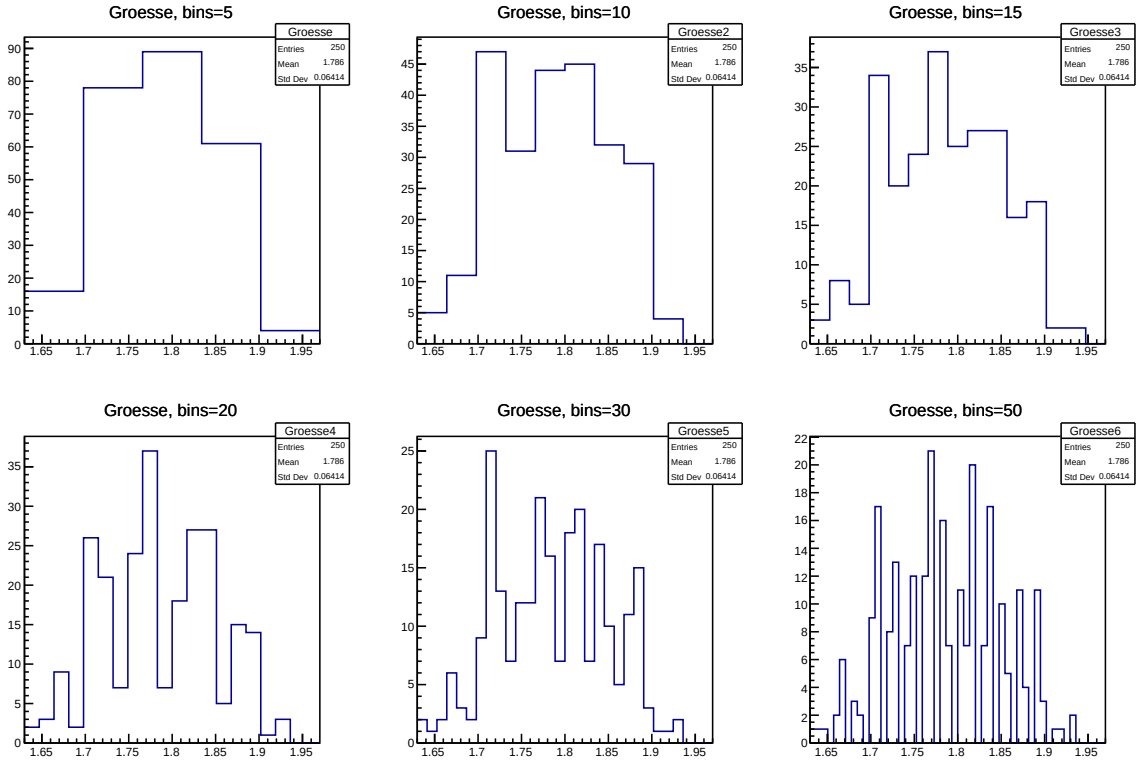


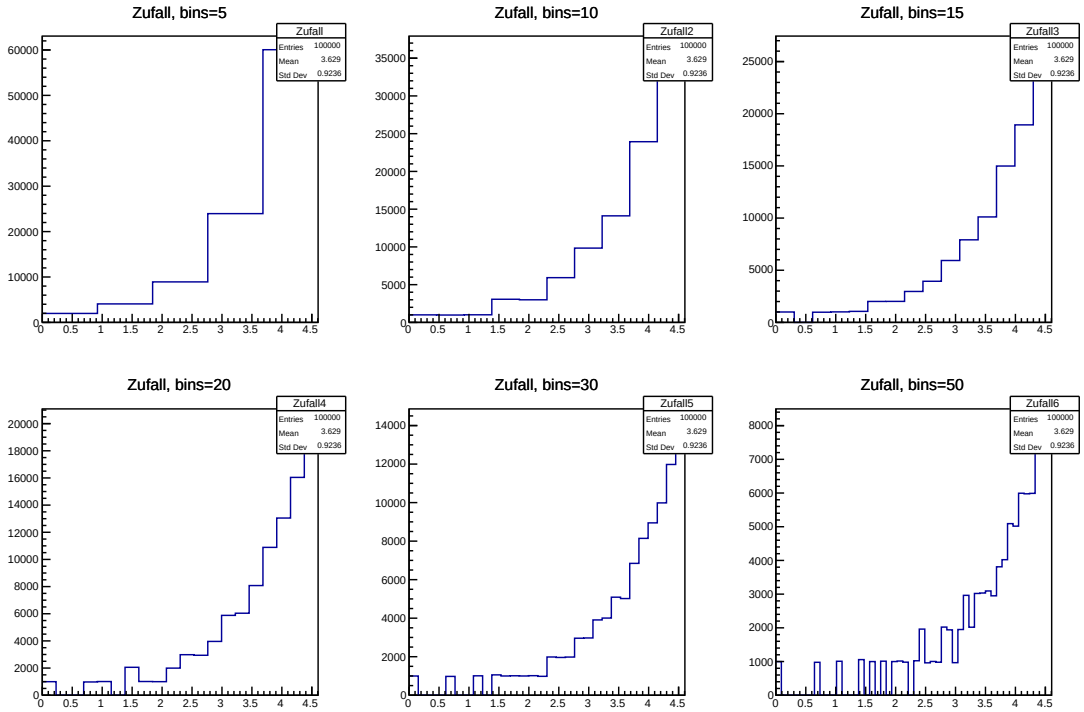
Abbildung 1: Histogramme in Abhängigkeit der Bins (Größe)

Die gleichen Argumente treffen auf die Größe zu, wobei der zweite Plot das Auge am meisten anspricht, da das Histogramm relativ gleichverteilt aussieht. Die Regel von Scott ergibt eine Binzahl von 9.5, was dieser Wahl bepflichtet.

b)

Generell ist es immer sinnvoll möglichst viele Daten zu verwenden, da sich die Verteilung dementsprechend immer mehr der realen annähert. Ist die Anzahl der möglichen Messergebnisse bzw das Interval, auf dem sich diese verteilen, sehr groß, ist es auch besser mehr Messdaten einzuholen, da es sonst zu diesen "Löchern" aus Aufgabenteil a) kommt. Folglich können bei weniger möglichen Ergebnissen weniger Bins verwendet werden und andersherum.

Die Anzahl der Bins muss auf jeden Fall kleiner sein als die Anzahl der Messwerte. Diese Frage ist jedoch situationsabhängig. Angenommen es gäbe ein Problem mit vielen möglichen Ergebnissen, die jedoch einen scharfen Häufungspunkt in der Verteilung besitzen. Dann ist es auch bei wenigen Ergebnissen sinnvoll eine kleine Binbreite (bzw. viele Bins) zu verwenden. Daher auch das Zitat *"There is no right or wrong answer as to how wide a bin should be"*.



c)

Bei den hier angegebenen Histogrammen fällt bei einer hohen Binzahl auf, dass bestimmte Binbereiche nicht bestetzt sind, obwohl die danebenliegenden Bins noch stark besetzt sind. Dies liegt daran, dass lediglich ganze Zahlen von 1 bis 100 gewürfelt werden. Bereits im dritten Plot ergibt sich somit eine Lücke zwischen 0.3 und 0.6, da keine logarithmierte ganze Zahl in diesem Bereich liegt. Eine Vergrößerung der Binanzahl macht also, zumindest im linken Bereich des Histogrammes, keinen Sinn mehr. Zudem fällt auf, dass das Histogram im rechten Bereich mehr Werte aufweist. Dies liegt daran, dass der Logaritmus für große Werte langsamer ansteigt, so dass viele große Zahlen auf ein kleineres Intervall projeziert werden.