stackfuel

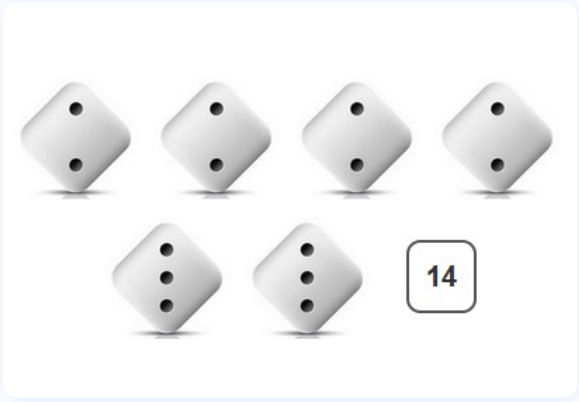
Statistische Beschreibung von Daten

Im Folgenden wollen wir kurz die Bereiche Stochastik, Statistik und Wahrscheinlichkeitsrechnung besprechen. Hierbei legen wir den Schwerpunkt auf die Aspekte dieser Bereiche, die für die Beschreibung und Auswertung von Daten besonders wichtig sind.

STOCHASTIK

Stochastik bedeutet übersetzt so etwas wie “die Kunst des Ratens”. In der Mathematik ist sie die übergeordnete Disziplin zur Wahrscheinlichkeitstheorie und Statistik. Gemein ist allen, dass sie sich mit der Beschreibung von Zufallsprozessen beschäftigen.
Wir fassen hierbei den Begriff Zufall sehr weit. Im Prinzip bedeutet es für uns einfach, dass Aspekte des Prozesses außerhalb unserer Kontrolle liegen bzw. nicht von uns vollständig quantifiziert werden können. Das lässt sich de facto auf jeden realen Fall anwenden. Ziel von Wahrscheinlichkeitsrechnung und Statistik ist es dann, solche Prozesse zu beschreiben.
Zentrales Element der Wahrscheinlichkeitsrechnung bzw. -lehre sind die **Wahrscheinlichkeitsverteilungen**. Wie der Name schon nahelegt, ordnen diese allen möglichen Ausgängen von Zufallsprozessen eine Wahrscheinlichkeit zu. Nehmen wir einen fairen Würfel: jede der 6 Seiten kommt gleich häufig vor. Die Wahrscheinlichkeitsverteilung ist somit 1/6 für jeden der 6 Ausgänge. Wenn wir den Würfel jedoch 6-mal würfeln, was wird passieren?
Man mag geneigt sein zu denken, dass jede beliebige Seite mindestens 1-mal auftritt. Wenn du Lust und einen Würfel zur Hand hast, kannst du es mal ausprobieren.

Ansonsten gibt es auch Seiten wie https://rolladie.net/roll-6-dice, wo man den Wurf simulieren kann. Ich habe zum Beispiel folgendes Ergebnis erhalten:



4-mal die 2 und 2-mal die 3. Die Verteilung ist also 4/6 für die 2 und 2/6 für die 3.
Wieso weicht diese Verteilung von den vermuteten 1/6 für alle Ergebnisse ab? Bei den geworfenen Würfeln handelt es sich um eine sogenannte Stichprobe. Dies ist die Verknüpfung zwischen Wahrscheinlichkeitstheorie und Statistik. Wahrscheinlichkeitstheorie hat als zentrales Element die Wahrscheinlichkeitsverteilung, während die Statistik sich auf die Beschreibung solcher Stichproben fokussiert. Diese Begrifflichkeit ist nicht scharf umrissen: häufig werden gleiche oder ähnliche Methoden und Formalismen in beiden Bereichen angewendet.

In der Praxis haben wir selten Zugriff auf die “wahre” Verteilung. Wir behelfen uns mit der Annahme, dass sich die Stichprobenverteilung der “wahren” Verteilung annähert.
Unser Fokus liegt hier darauf, einige wichtige Grundlagen aus dem Bereich Stochastik einzuführen. An den Stellen, an denen es sich anbietet, führen wir aber auch relevante Power BI Befehle ein. Wenn du Lust hast, kannst du dann diese Befehle auf die vorher erzeugten Zufallszahlen oder andere Datensätze anwenden. So wird die Theorie für dich vielleicht etwas lebendiger. Weitestgehend versuchen wir aber auch konkrete, anschauliche Beispiele zu geben.

BESCHREIBENDE STATISTIK

Wie der Name schon impliziert, geht es hierbei vor allem darum einen ist-Zustand abzubilden und zu beschreiben. Eine Datengrundlage soll dargestellt und gegebenenfalls zusammengefasst werden. Ein wichtiges Mittel hierfür sind statistische Kenngrößen.

Im Wesentlichen unterscheiden wir hierbei zwei Klassen von Maßen:
- Lage-Maße
- Dispersions-Maße

Lage-Maße

Beginnen wir mit dem **Mittelwert**. Du wirst ihn vermutlich schon kennen.
Mathematisch ist er definiert als:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Hierbei wird eine Zufallszahl x betrachtet, die von einer Stichprobe der Größe N genommen wurde. Wenden wir das auf unser Beispiel mit den Würfeln an. Die Zufallszahl ist das Ergebnis des Würfelwurfs, also die Augenzahl, die oben liegt. Wir haben 6-mal gewürfelt, entsprechend ist N=6.

Die x_i sind hierbei die einzelnen Ergebnisse der Würfe. Im obigen Beispiel wären also x_1,… ,x_4=2 und x_5,x_6=3. Der Mittelwert wäre somit:

$$\bar{x} = \frac{1}{6}(4 \cdot 2 + 2 \cdot 3) = \frac{14}{6} = 2.33$$

Wählen wir N groß genug, sollte sich dieser Mittelwert einer Stichprobe dem “Mittelwert” der Zufallsverteilung annähern. Diesen nennen wir **Erwartungswert**. Es ist der Wert, den wir “erwarten” können, wenn wir eine einzelne Stichprobe entnehmen. Aber hier kann der Name etwas täuschen, denn der Erwartungswert muss kein häufiger bzw. typischer Wert sein. Da hier der Wert des Ergebnisses mit einfließt, können seltene, aber betragsmäßig große Elemente der Verteilung einen überproportionalen Einfluss auf den Mittelwert nehmen.

Als Test kannst du mal selbst überprüfen, wie sich der Mittelwert verändern würde, wenn die Seite mit der 6 stattdessen einen Wert von 100 oder gar 1000 hätte.

Besonders bei ganzzahligen oder gar kategorischen Größen kann es aber auch sinnvoll sein, das häufigste Element einer Verteilung betrachten. Man spricht hier von Modus. Dieser Modus kann auch gut das typische Verhalten einer Zufallszahl abbilden. Speziell bei Fließkommazahlen passt Modus oft jedoch nicht so gut. Es kann sein, dass sich überhaupt keine Zahlen wiederholen, oder dass diese Wiederholungen rein zufällig sind ohne, dass in diesem Bereich generell viele Datenpunkte liegen.

Daher gibt es noch ein weiteres Lage-Maß, um einen “mittleres” Verhalten abzubilden. Dies ist der Median. Beim Median wird die Stichprobe aufsteigend geordnet. Dann wird der “mittelste” Wert ausgewählt. Haben wir eine ungerade Anzahl an Elementen in der Stichprobe, ist es genau der Wert in der Mitte. Stellen wir uns vor, die Stichprobe wäre geordnet 1,2,3,4,5. Dann wäre der Wert in der Mitte genau die 3.

Was aber, wenn die Stichprobe außerdem 6 enthalten wurde? Dann ist es Konvention, den Mittelwert der zwei Werte neben der Mitte zu wählen. Bei 1,2,3,4,5,6 wären das genau die 3 und die 4. Der Mittelwert von 3 und 4 wäre 3.5. Somit wären in diesem Fall der Median und der Mittelwert der Stichprobe identisch. Das ist aber in der Regel nicht der Fall. Nehmen wir das Beispiel 1,2,3,4,5,6 von eben. Wir könnten den Wert 6 durch einen beliebig großen Wert ersetzen, ohne den Median zu beeinflussen. Der Mittelwert wird dadurch aber stark beeinflusst.

Häufig wird der Median deshalb in Bereichen eingesetzt, in denen man den Einfluss von seltenen, aber betragsmäßig großen Ergebnissen minimieren möchte. Beispiele hierfür sind statistische Erhebungen zu Einkommen, Reichtum oder Mietpreisen. In einigen Ländern verdient der Großteil der Bevölkerung zum Beispiel im globalen Vergleich sehr wenig, aber es gibt eine Handvoll sehr wohlhabender Personen, die den Mittelwert des Einkommens erhöhen. Hier vermittelt dann der Median ein besseres Bild der Einkommenssituation der allgemeinen Bevölkerung.

Quantile

Der Median ist ein Spezialfall einer ganzen Reihe von Maßen. Häufig teilt man den Werte-Bereich in wohldefinierte, gleichgroße Teile ein. Im einfachsten Fall halbieren wir den Wertebereich. Dann ist der Median genau der Wert, der die beiden Intervalle trennt. Andere gängige Einteilungen sind Viertel (Quartile) und Hundertstel (Perzentile).

Die Perzentile sind besonders im englischsprachigen Raum verbreitet. Dort werden häufig bei Tests bei der Auswertung der Performance die Perzentile angegeben. Befindet man sich im 97ten Perzentil, gibt es nur 3 Prozent, die besser oder gleichgut abgeschnitten haben. Aber es wäre etwas unpraktisch die Position aller 100 Perzentile anzugeben, um einen Eindruck von der Verteilung zu erhalten.

In Dax gibt es für die Perzentile zwei Funktionen:

- https://docs.microsoft.com/en-us/dax/percentile-inc-function-dax
- https://docs.microsoft.com/en-us/dax/percentilex-exc-function-dax

Wenn du möchtest, kannst du diese Funktionen mal in einem der Datensätze ausprobieren.

Um einen guten Eindruck für die Verteilung zu bekommen, eignen sich eher die **Quartile**. Dabei handelt es sich um die 25, 50 und 75er Perzentile. Zusammen mit Minimum und Maximum (0er und 100er Perzentile) erhält man so einen guten Überblick darüber, wie sich die Daten über den Wertebereich verteilen. Man spricht hier auch von der **5-Werte Zusammenfassung**.

Eine ähnliche Aufgabe haben die Dispersionsmaße.

Dispersionsmaße

Im weitesten Sinne sollen Dispersionsmaße abbilden, wie stark die Daten sich über den Wertebereich verteilen oder streuen. Der Mittelwert soll abdecken, welchen Wert wir bei einer Zufallszahl ungefähr erwarten sollen. Eine naheliegende, weiterführende Frage wäre, welche **Abweichung** wir von diesem Mittelwert erwarten können. Also den Mittelwert der Differenz des Ergebnisses zum Mittelwert.

$$\overline{(x-\bar{x})} = \frac{1}{N}\sum_{i=1}^{N} x_i - \bar{x}$$

Leider fällt einem schnell auf, dass dieser Wert immer Null ergibt. Das liegt daran, dass sich Abweichungen nach unten und nach oben gegenseitig aufheben. Das können wir umgehen, indem wir die Quadrate der Abweichungen berechnen.

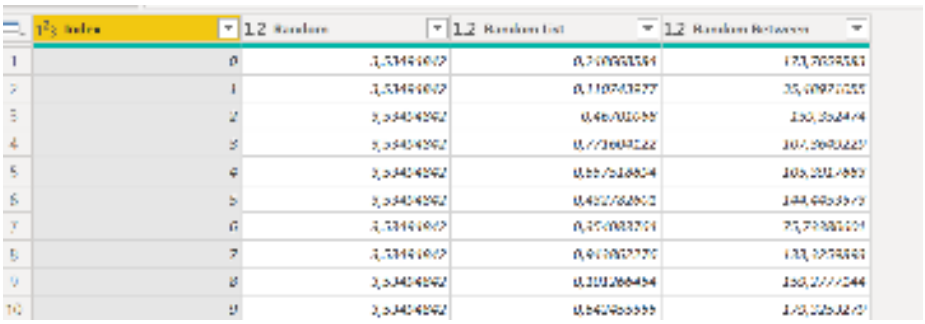
$$\overline{(x-\bar{x})^2} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2$$

Es handelt sich also um die quadrierte mittlere Abweichung vom Mittelwert. Diese Größe nennen wir **Varianz**. In vielen Fällen betrachtet man die Quadratwurzel dieser Größe. Dies nennt man die **Standardabweichung** (eng. standard deviation).

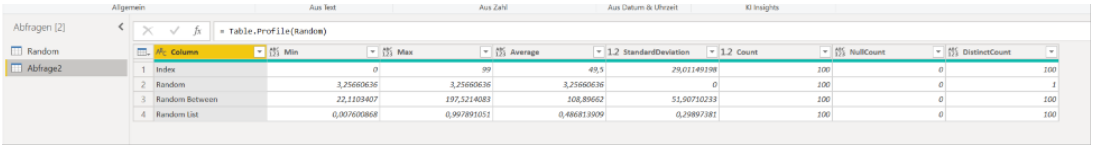
Ein weiteres wichtiges Dispersionsmaß basiert auf den Quartilen. Man bildet hierbei die Differenz aus dem 75er und 25er Quartil. Diese Größe nennen wir den **Interquartilsabstand**.

Manchmal reichen aber auch Lage- und Dispersions-Maße im Detailgrad nicht aus, um eine Verteilung zu beschreiben. Dazu gibt es noch Größen, die die Gestalt bzw. Form einer Verteilung beschreiben. Bevor wir uns diese Größen ansehen, wollen wir aber noch kurz ein paar Verteilungen ansehen.

Der Power Query Editor bietet dir mit Table.Profile() eine Möglichkeit um schnell einen Überblick über einen Datensatz zu erlagen. Das geht zum Beispiel so. Zunächst brauchen wir eine Table. Ich benutze zum Beispiel diese Table mit den erzeugten Zufallszahlen aus der vorherigen Übung:



Wenn ich nun wieder eine neue leere Abfrage erstelle, kann ich dort Table.Profile() in die Eingabe schreiben. In die Klammern füge ich jetzt noch den Namen der Tabelle ein, für die ich das Profil erstellen möchte:



Wird bei dir noch in einigen Zeilen *null* angegeben, kann es sein, dass du den Datentyp in der Tabelle anpassen musst.

(ZUFALLS-)VERTEILUNGEN.

Betrachten wir noch mal den Münzwurf. Wir hatten im Beispiel 4-mal die 2 und 2-mal die 3 erhalten. Die 4 und die 2 sind hierbei die **absoluten Häufigkeiten** der Wurfergebnisse. Teilen wir diese durch die Stichprobengröße N, erhalten wir die **relativen Häufigkeiten**.

Wir könnten jeden der geworfenen Würfel unter ein Hütchen setzen. Wenn wir nun eins dieser 6 Hütchen zufällig auswählen würden, hätten wir in 2/3 der Fälle eine 2 aufgedeckt und in 1/3 der Fälle eine 3. Wir können diese relativen Häufigkeiten also als Wahrscheinlichkeiten interpretieren.

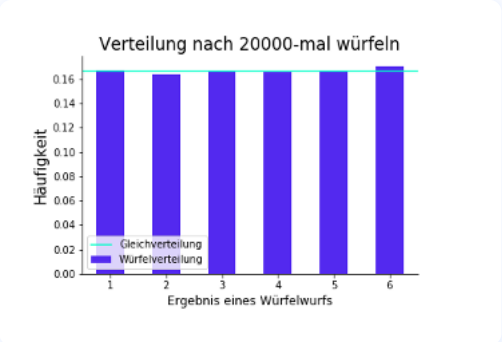
Weiterhin können wir annehmen, dass sich diese Wahrscheinlichkeiten den “wahren” Wahrscheinlichkeiten annähern, wenn die Stichprobengröße N groß genug ist. Formal spricht man davon, N gegen unendlich laufen zu lassen. In der Praxis nähern sich diese Wahrscheinlichkeiten schon bei endlichen Stichprobengrößen den wahren Größen gut an.

Wir wollen uns nun die Wahrscheinlichkeitsverteilungen einiger wichtiger Verteilungen ansehen.

Gleichverteilung

Ein Beispiel für eine “idealisierte” oder “wahre” Gleichverteilung haben wir schon kennengelernt.

Werfen wir einen fairen Würfel, erwarten wir für jede der 6 Seiten die gleiche Wahrscheinlichkeit nach einem Wurf oben zu liegen, nämlich eine Chance von 1 zu 6. Im Folgenden haben wir das Ergebnis von 20.000 Würfelwürfen dargestellt. Im Vergleich dazu haben wir die Gleichverteilung über die 6 möglichen Ergebnisse als Referenz Linie dargestellt. Die tatsächliche Verteilung ähnelt der Gleichverteilung schon sehr, es gibt aber auch kleine Abweichungen.



Das bring uns direkt zur Erweiterung des Konzepts der Verteilung für Zufallszahlen, die kontinuierliche Werte, zum Beispiel dargestellt durch Fließkommazahlen, annehmen können.

In diesem Fall ist es nicht mehr direkt möglich eine Wahrscheinlichkeit für jedes einzelne Element aus dem Wertebereich anzugeben, denn wir können nicht mehr über alle Elemente summieren. Denn Summen können wir nur über Abzählbar unendliche Mengen bilden. Ein Intervall auf den reellen Zahlen enthält aber überabzählbar viele Elemente.

Stattdessen können wir dann nur Wahrscheinlichkeiten für Wertebereiche definieren. Stellen wir uns vor wir betrachten den Regen, der pro Tag fällt. Dieser wird häufig in ml pro Quadratmeter gemessen. Diese Größe kann beliebige positiv reelle Werte annehmen. Entsprechend würden wir hier keine Wahrscheinlichkeiten für konkrete Niederschlagsmengen angeben, sondern eher für Wertebereiche. Das können Intervalle sein, zum Beispiel von 10 bis 20 ml/qm, oder auch für Halbintervalle. Um Beispiel die Wahrscheinlichkeit mehr als 100 ml/qm Niederschlag zu beobachten.

Stellen wir uns vor, dass wir den Bereich der Werte, die angenommen werden können, sehr fein in Bereiche der Länge dx aufteilen. Wir definieren die Wahrscheinlichkeit, dass ein Wert in einem bestimmten Bereich zwischen x und x+dx angenommen wird, als p(x)dx. Hierbei bezeichnen wir p(x) als die Wahrscheinlichkeitsdichte. Für eine Gleichverteilung erwarten wir, dass zu gleichgroßen Intervallen bzw. Bereichen eine gleichgroße Wahrscheinlichkeit gehört. In diesem Fall wäre die Wahrscheinlichkeitsdichte einfach eine Konstante. Im diskreten Fall, wie im Beispiel der Würfel, ist die Konstante einfach 1 geteilt durch die Anzahl der möglichen Ergebnisse.
Im kontinuierlichen Fall können wir die möglichen Ergebnisse nicht mehr abzählen.

Hier nehmen wir stattdessen die Länge des Intervalls. Das liegt daran, dass wir hier integrieren, statt zu addieren.

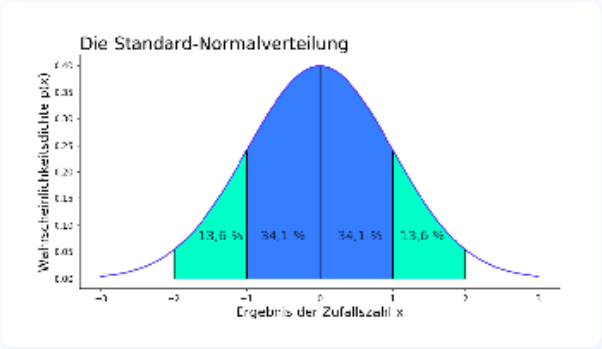
Haben wir also eine kontinuierliche Zufallszahl, die Werte zwischen a und b annehmen kann, und in diesem Bereich gleich verteilt ist, so ist die Zufallsdichte in diesem Fall 1/(b-a).

Normalverteilung

Eine weitere wichtige Klasse von kontinuierlichen Zufallszahlen bilden die Zufallszahlen, deren Wahrscheinlichkeitsdichte eine sog. Normalverteilung bildet. Dir ist die entsprechende Funktion vielleicht als Gauß'sche Glockenkurve oder unter einem ähnlichen Namen schon mal begegnet. Die mathematische Formel hierzu lautet:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

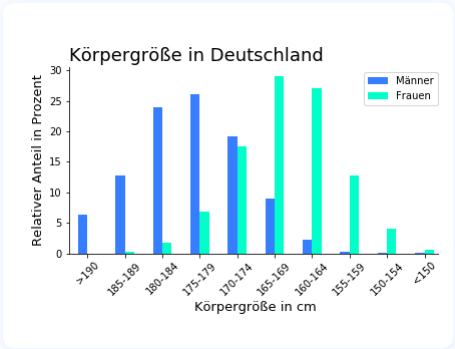
Hierbei bezeichnet σ die Standardabweichung und somit σ^2 die Varianz. Der Mittelwert, bzw. der Erwartungswert, wird mit μ bezeichnet. Der Spezialfall μ=0, σ=1 wird als Standard-Normal-verteilung bezeichnet.



In dieser Grafik siehst du die charakteristische Glockenform dieser Verteilung. Bei der Standard-Normalverteilung liegen 68,2 Prozent der Werte im Intervall [-1,1]. Das Intervall umspannt 2 Standardabweichungen und enthält den Mittelwert in der Mitte.
Die Normalverteilungen spielen eine besondere Rolle in der Stochastik, weil eine sehr breite Menge an Zufallsprozessen in solche normalverteilten Zufallsgrößen übergehen. Diese Beobachtung wird im Zentralen Grenzwertsatz der Statistik konkretisiert. Dieser geht über den aktuellen Kurs hinaus, wenn du möchtest, kannst du aber hier bei Wikipedia mehr erfahren:

https://de.wikipedia.org/wiki/Zentraler_Grenzwertsatz

Wir wollen nun kurz ein anschauliches Beispiel geben. Ein reales Beispiel für eine Größe, die annähernd Normalverteilt ist, ist die Körpergröße. Diese haben wir hier exemplarisch dargestellt. Wir sehen, dass die Verteilung für Männer und Frauen jeweils stark an die charakteristische Glockenform der Normalverteilung erinnert.



Wir sehen hier auch noch einen interessanten Effekt: Obwohl Männer im Schnitt ungefähr 10 Zentimeter größer sind als Frauen, gibt es einen kleinen Anteil Männer die kleiner als die meisten Frauen sind. Umgekehrt gibt es auch eine Reihe von Frauen, die mit ihrer Körpergröße buchstäblich eine Vielzahl von Männern in den Schatten stellen. Das illustriert das wir nicht ohne weiteres von der Gesamtheit auf Individuen schließen können. Hier können wir noch einen wichtigen weiteren Punkt ansprechen. Wir hatten gesagt, dass wir generell darauf hoffen, dass sich die Stichprobenverteilung der Gesamtverteilung annähert, wenn wir die Stichprobe vergrößern. Das gilt aber nur, wenn wir keine Verzerrungen in der Datenerhebung haben. Solche Verzerrungen sind in der Praxis nicht immer leicht zu entdecken, weil wir die Gesamtverteilung in der Regel nicht kennen, sondern durch Datenerhebung ermitteln wollen.

Ein kleines anschauliches Beispiel kann dies noch etwas klarer machen. Stellen wir uns vor, ein Außerirdischer landet auf der Erde und möchte statistisch die Größe der Menschen ermitteln. Durch Zufall landet er in Holland. Nun sind Holländer im Schnitt das größte Volk der Welt. Egal wie viele Stichproben in Holland unser Außerirdischer sammelt, er wird diese Verzerrung nicht auflösen können. Und da er ein Außerirdischer ist, weiß er nicht, dass es Regionale Unterschiede bei der Größe gibt. In vielen Fällen geht es uns wie den Außerirdischen. Weil wir die Absolute Verteilung nicht kennen, können wir nicht immer wissen, ob wir Verzerrungen im Datensatz haben.

Generell können wir uns folgendes merken: statistische Analysen profitieren davon, wenn die zugrunde liegende Stichprobe möglichst groß ist. Reine Größe der Stichprobe allein reicht jedoch nicht. Wir sollten auch Verzerrungen im Datensatz vermeiden. Es ist nicht immer leicht, diese Verzerrungen aufzudecken. Aber wenn wir Vergleiche zwischen Teilgruppen eines Datensatzes machen, sollten wir insbesondere dann vorsichtig sein, wenn diese Teilgruppen, die wir vergleichen, aus deutlich verschieden großen Mengen bestehen.

Pareto-Verteilung

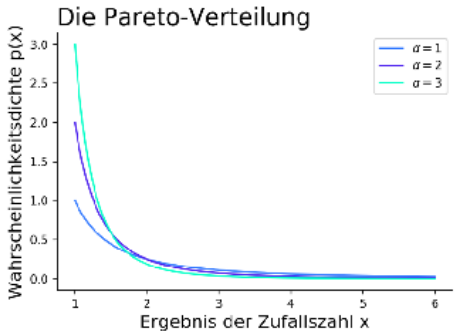
Eine weitere Klasse von Verteilungen wollen wir noch besprechen, sind die **Pareto-Verteilungen**. Pareto hat unter anderem die Wohlstandsverhältnisse während des 19. Jahrhunderts untersucht. Dabei fiel ihm auf, dass ein geringer Teil der Bevölkerung einen signifikanten Teil des Wohlstands auf sich vereinigte.

Heute bezeichnet das Pareto Prinzip allgemein, dass ein geringer Teil der Ursachen für einen signifikanten Anteil der Ergebnisse verantwortlich ist. Umgangssprachlich findet man dabei die 80-20 Regel, nach der man 80 Prozent der Ergebnisse mit 20 Prozent des Aufwandes entfällt.

Solche Beobachtungen schlagen sich oft in sogenannten Pareto-Verteilungen nieder. Diese sind sog. Potenzgesetze (engl. Power Law Distribution). Das bedeutet, dass ihre Wahrscheinlichkeits-dichtefunktion die folgende Form haben

$$p(x) = \frac{x_{min} \cdot \alpha}{x^{1+\alpha}} \; f\ddot{u}r \; x > x \min$$

Hier ist $p(x)$ proportional zu $x^{-(1+\alpha)}$.



Hier wurde die Pareto-Verteilung für $x_{min} = 1$ für verschiedene Werte von α dargestellt. Wir sehen, dass der Minimalwert gleichzeitig die höchste Wahrscheinlichkeitsdichte aufweist. Entsprechend sind so geringe Ergebnisse von x sehr wahrscheinlich. Für große Werte von x fällt die Verteilung zunächst rapide ab. Aber nach einem gewissen Wert scheinen die Verteilungen abzuflachen.

Es bildet sich etwas, dass uns vielleicht an einen “Rattenschwanz” erinnert. Die Verteilung ist noch in einem breiten Bereich von Null verschieden. Außerdem ist sie etwas schief in dem Sinne, dass sie nicht so spiegelsymmetrisch wie die Gauß-Verteilung ist. Um diese zwei Ein-drücke zu quantifizieren, Führt man die Schiefe (skewness) und die Wölbung (Kurtosis) ein.

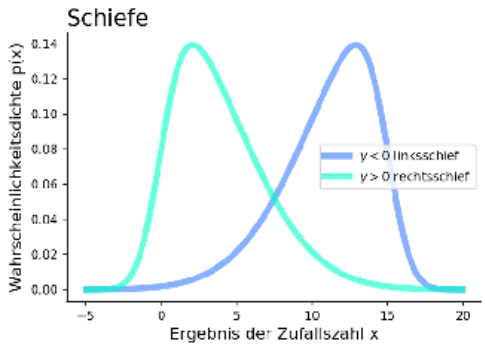
SCHIEFE UND WÖLBUNG

Bei der Standardabweichung haben wir uns schon mit den mittleren Abweichungen beschäftigt. Dabei hatten wir gesehen, dass die mittlere Abweichung vom Mittelwert verschwindet. Die mit-tlere quadrierte Abweichung hatte uns auf die Varianz bzw. Die Standardabweichung geführt. Was passiert, wenn wir höhere Potenzen betrachten?

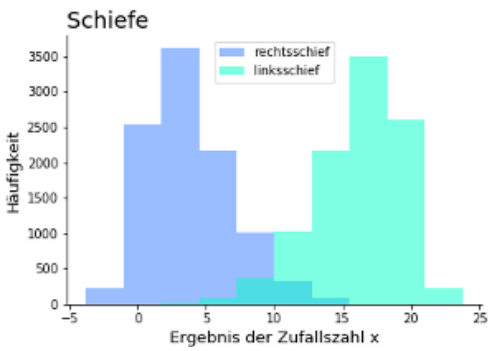
Schiefe (skewness)

$$\gamma_m = \frac{\overline{(x-\mu)^3}}{\sigma^3}$$

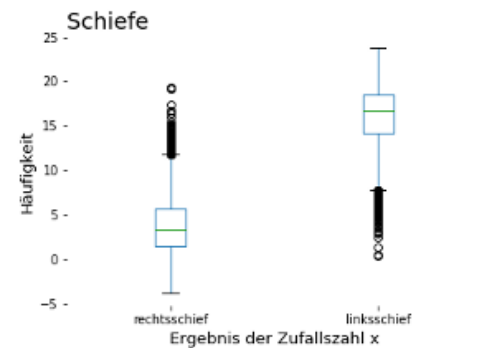
Dies benutzt man zur Definition der **Schiefe** oder auf Englisch *skewness*.
Wir wollen nur kurz motivieren, warum es sich hier um eine Größe handelt, die die Schiefe misst. Im einfachen Fall in dem der Mittel- bzw. Erwartungswert verwindet, mitteln wir die Zufallsgröße zur dritten Potenz. Bei ungeraden Potenzen erhalten wir, anders als bei geraden Potenzen, negative Beiträge beim Mitteln. Ist die Verteilung symmetrisch, wie zum Beispiel die Normalverteilung, heben sich die positiven und negativen Beiträge genau gegenseitig auf. Aber wenn die Verteilung sich etwas nach links oder rechts neigt, also schief ist, wird diese Symmetrie gebrochen.



Wir können die Schiefe auf andere Weise erkennen. Oft kann man sie qualitativ aus einem Histogramm ablesen. Dies könnte zum Beispiel so aussehen:



Auch in einem Boxplot kann man die Schiefe gut erkennen:



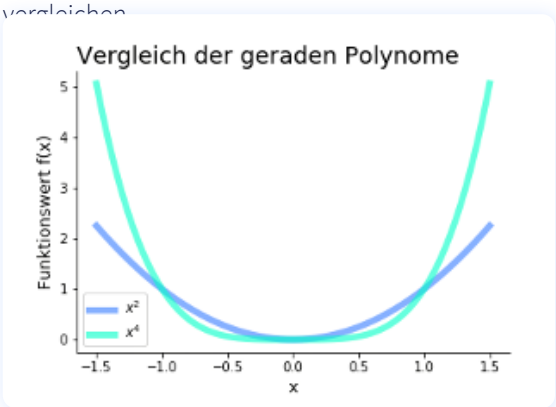
Man kann es auch aus statistischen Größen ablesen:

- Rechtsschief: Modus<Median<Mittelwert
- Linksschief: Modus>Median>Mittelwert
- Symmetrisch: Modus = Median = Mittelwert

Die Normalverteilung ist symmetrisch und hat somit keine Schiefe. Die Pareto-Verteilungen sind hingegen nicht symmetrisch und besitzen, zumindest für bestimmte Parameter, eine Schiefe.

Wölbung (Kurtosis)

Die nächsthöhere Potenz führt uns auf die Wölbung. Die Interpretation basierend auf der Formel ist hier noch etwas schwerer als bei der Schiefe. Da es sich aber hier wieder um eine gerade Potenz handelt, heben sich Beiträge nicht so einfach gegenseitig auf, ähnlich wie bei der Varianz. Noch mehr Einsicht erhalten wir, wenn wir mal die x zur zweiten Potenz mit x zur vierten Potenz vergleichen.



Wir sehen, dass hier im Bereich zwischen –1 und 1 die zweite Potenz größer ist. Außerhalb dieses Bereichs dominiert aber die vierte Potenz. Wenn wir x zur vierten Potenz mitteln, erhalten wir wenig Beitrage in der Nähe des Mittelwertes. Genau umgekehrt verhält es sich mit Punkten, die weiter vom Mittelwert entfernt sind. Wir messen also, wie stark sich die Werte um den Mittelpunkt zentrieren.

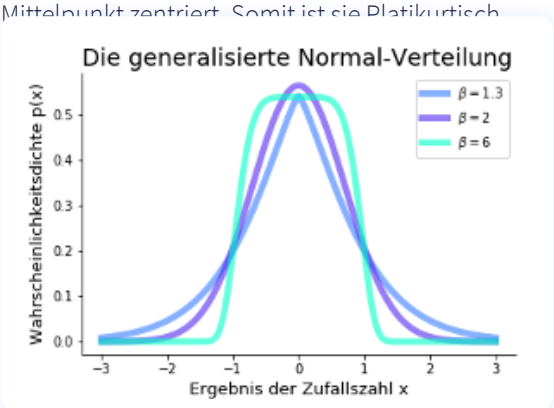
Speziell geht es um die Existenz dieser “Rattenschwänze”, die wir vorhin bei der Pareto-Verteilung gesehen haben. Die Normalverteilung hat immer eine Wölbung von 3. Basierend hierauf kann man eine **Exzess Wölbung** definieren, die sich als Differenz zur Wölbung der Normalverteilung ergibt.

Exzess Kurtosis = Kurtosis –3

Hierbei unterscheidet man 3 Fälle:

1. Exzess Kurtosis =0: Mesokurtisch. Die Verteilung ist genauso gewölbt wie die Normalverteilung
2. Exzess Kurtosis >1: leptokurtisch oder steilgipflig. >spitzere Verteilung mit mehr “Rattenschwanz” (englisch tail)

3. Exzess Kurtosis <1: Platikurtisch oder flachgipflig. > flachere Verteilung
Hier siehst du die generalisierte Normalverteilung. Diese hat einen Parameter β der die Form der Verteilung bestimmt. Für β =2 ist es die uns schon bekannte Normalverteilung. Entsprechend ist die Verteilung mesokurtisch. für β=1.3 ist sie etwas spitzer aber hat dafür einen ausgeprägteren "Rattenschwanz". Sie ist somit leptokurtisch. Für β = 6 ist sie etwas platter und stärker um den Mittelpunkt zentriert. Somit ist sie Platikurtisch.



FORM VON VERTEILUNGEN: ZUSAMMENFASSUNG

Fassen wir nun nochmal zusammen: Es gibt zwei wichtige Parameter, die die Form einer Verteilung beschreiben. Dabei geht es um die Schiefe und die Wölbung. Spiegel symmetrische Verteilungen wie die Normalverteilung, besitzen keine Schiefe. Nicht symmetrische Verteilungen wie die Pareto-Verteilung, können eine Schiefe besitzen. Im übertragenen Sinne kann man sich die schiefen Verteilungen so vorstellen, als ob es sich um "betrunkene" Normalverteilungen handelt, die etwas "Schlagseite" haben.

Die Wölbung beschreibt, wie flach oder Spitz eine Verteilung ist und somit, wie stark die "Rattenschwänze" der Verteilung ausgeprägt sind.

Neben der Beschreibenden Statistik, gibt es noch die schließende und die explorative Statistik. Die schließende Statistik leitet aus den Daten zusammenhänge ab. Klassische Beispiele hierfür sind die Korrelationsanalyse und die Regression. Einen Einblick in diese Bereiche wirst du in den nächsten Lektionen erhalten. Die Explorative Statistik kombiniert Methoden aus den beiden vorhergehenden Bereichen der beschreibenden und schließenden Statistik. Hierbei geht es meist darum tiefergehende Analysen vorzubereiten.

Herzlichen Glückwunsch! Du hast nun dein Wissen aus den Bereichen Statistik erweitert und bist nun gut gewappnet, wenn dir im Alltag oder im Job statistische Konzepte begegnen! Wir wünschen dir viel Vergnügen mit der nächsten Lektion!