KAPITEL 7: EINIGE GRUNDBEGRIFFE DER BESCHREIBENDEN STATISTIK

Die Grundlagen der Statistik werden in den Lehrveranstaltungen "Forstliche Biometrie" und "Statistische Datenanalyse" genauer behandelt. Hier sollen lediglich die allerwichtigsten Grundbegriffe eingeführt werden, die in der Literatur und in anderen Lehrveranstaltungen sehr häufig auftreten und mit denen man sich daher von Anfang an vertraut machen sollte: Mittelwert, Median, Varianz, Standardabweichung von Stichproben, und der Korrelationskoeffizient zweier Merkmale.

Die zentrale Problemstellung der beschreibenden Statistik sieht in vielen Fällen wie folgt aus: Gegeben sind *n* Objekte (auch: "Probanden", "Gegenstände", "Fälle", "cases" etc.), von denen jedes durch eines oder mehrere Merkmale (auch: "Variablen") charakterisiert wird (z.B. Bäume einer Probefläche mit den Merkmalen "Höhe", "Baumart", "Schadensklasse" usw.). Jedes Merkmal kann verschiedene *Ausprägungen* annehmen: Hierbei kann es sich um Zahlen oder um nicht-numerische Qualitäten handeln. Im Folgenden betrachten wir nur Merkmale, deren Ausprägungen reelle Zahlen sein können (sogenannte *metrisch skalierte* Merkmale).

Das Ziel ist es nun, folgendes zu beschreiben (und zu quantifizieren):
- bei einem einzelnen Merkmal:
  Die *Lage* der *n* Objekte im Bereich aller möglichen Merkmalsausprägungen (sogenannte *Lokalisation*), sowie
  die *Streuung* des Merkmals (*Dispersion*),
  und oft auch die Form der Verteilung der Merkmalsausprägungen,
- bei mehreren Merkmalen (zusätzlich):
  ihren *Zusammenhang* untereinander.

Mittelwert, Median, Standardabweichung

Wir betrachten zunächst den Fall, dass wir es nur mit einem einzigen Merkmal zu tun haben. Die Ausprägungen, die bei den *n* Objekten vorliegen, seien $x_1, x_2, ..., x_i, ..., x_n$. (Das sind also so viele Zahlen, wie es Objekte gibt.)
Das wichtigste Lokalisationsmaß ist das *arithmetische Mittel*:

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

(in Worten: Es wird die Summe der Merkmalsausprägungen aller Objekte gebildet und durch die Anzahl der Objekte geteilt.)
In der Summe, die in dieser Formel vorkommt, können Werte mehrfach auftreten (nämlich immer dann, wenn verschiedene Objekte dieselbe Merkmalsausprägung haben). Mitunter ist es einfacher, die Summe nicht über alle Objekte, sondern über alle möglichen Merkmalsausprägungen laufen zu lassen: Das heißt, wir fassen die in der obigen Summe mehrfach auftretenden Summanden zusammen und gewichten jedes x, was als Ausprägung vorkommen kann, mit seiner *absoluten Häufigkeit* $H(x)$. In der neuen Summe wird jeder Wert x nur einmal betrachtet. Wenn wir den Vorfaktor in die Summe hineinziehen, können wir zu *relativen Häufigkeiten* $h(x) = H(x)/n$ übergehen (diese liegen immer zwischen 0 und 1). Wir erhalten somit als alternative Berechnungsformeln für den Mittelwert:

$$\bar{x} = \frac{1}{n}\sum_{x} H(x)\cdot x = \sum_{x} h(x)\cdot x \ .$$

Ein anderes Lokalisationsmaß ist der *Median* (auch: Halbwert). Um diesen zu bestimmen, werden die Werte der einzelnen Objekte zunächst nach der Größe sortiert. Bei ungerader Anzahl von Objekten ist der Median der mittlere der sortierten Werte.

Wenn wir also z.B. $n = 5$ Fälle haben, und die Messwerte für das betrachtete Merkmal waren

$$7;\ 5;\ 1;\ 4;\ -2,$$

so ist die Folge der sortierten Werte

$$-2;\ 1;\ \mathbf{4};\ 5;\ 7,$$

und der zentrale Wert ist der Median $\tilde{x} = 4$.
Bei gerader Anzahl von Objekten definiert man den Median als das arithmetische Mittel der beiden mittleren Werte in der sortierten Folge: Haben wir z.B. die Messwerte

$$-3;\ 10;\ 4;\ 3;\ 1;\ -3,$$

so ist die sortierte Folge

$$-3;\ -3;\ \mathbf{1};\ \mathbf{3};\ 4;\ 10,$$

und der Median ist (1+3)/2 = 2.

Ein wichtiges Maß für die *Streuung* von Merkmalsausprägungen ist die sogenannte *Stichprobenvarianz* s^2. Sie wird definiert durch die folgende Formel:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ .$$

Es handelt sich also "im wesentlichen" um das mittlere Quadrat der Abweichung vom Mittelwert — mit der kleinen Modifikation, dass durch $n-1$ statt durch n geteilt wird.
Die *Standardabweichung* des Merkmals ist einfach die Quadratwurzel aus der Stichprobenvarianz:

$$s = \sqrt{s^2} \ .$$

Sie wird häufig in Tabellen zusätzlich zum Mittelwert angegeben (oft in Klammern hinter dem Mittelwert, oder in der Form "$\bar{x} \pm s$"), um eine Information darüber zu liefern, wie stark das Merkmal um seinen Mittelwert herum streut.

Beispiel:
$n = 5$ Fälle; Messwerte 7; 5; 1; 4; –2 (vgl. Beispiel oben). Wir erhalten:

$$\bar{x} = \tfrac{1}{5}(7+5+1+4-2) = \tfrac{15}{5} = 3{,}0 \quad \text{und}$$

$$s^2 = \tfrac{1}{4}\left((7-3)^2 + (5-3)^2 + (1-3)^2 + (4-3)^2 + (-2-3)^2\right)$$

$$= \tfrac{1}{4}\left(4^2 + 2^2 + (-2)^2 + 1^2 + (-5)^2\right) = \tfrac{1}{4}\left(16+4+4+1+25\right) = \tfrac{50}{4} = 12{,}5 \ ;$$

$$s = \sqrt{12{,}5} \approx 3{,}536.$$

Wenn wir zu *zwei* Merkmalen übergehen, ist neben der Lokalisation und Streuung der beiden Einzelmerkmale auch ihr eventueller Zusammenhang untereinander von Interesse. Man unterscheidet:

Regressionsrechnung: Hier geht es darum, einen (angenommenen) Zusammenhang zwischen den beiden Merkmalen x und y, der als Funktion verstanden wird, zu quantifizieren: $y = f(x)$. Im einfachsten Fall nimmt man f als lineare Funktion an (Regressionsgerade). Die Anpassung der Funktion an die Datenmenge (Menge aller Punkte (x, y); "scatterplot") erfolgt gewöhnlich mit der "Methode der kleinsten Quadrate", die wir schon in Kapitel 5 kennengelernt haben.

Korrelationsrechnung: Es wird ein Maß dafür bestimmt, wie stark beide Merkmale linear zusammenhängen, bzw. welcher Anteil der Streuung des einen Merkmals durch das andere erklärt werden kann. Hierfür ist der Begriff des Korrelationskoeffizienten von Bedeutung.

Beachte: Das Vorliegen eines statistischen Zusammenhangs (einer Korrelation) zwischen zwei Merkmalen bedeutet noch nicht, dass zwischen diesen Merkmalen ein *kausaler* Zusammenhang besteht. Der im Datensatz manifeste Zusammenhang kann zufällig entstanden sein, oder er kann auf eine gemeinsame Ursache für die Ausprägung beider Merkmale zurückzuführen sein, ohne dass diese sich untereinander direkt beeinflussen.

Korrelation

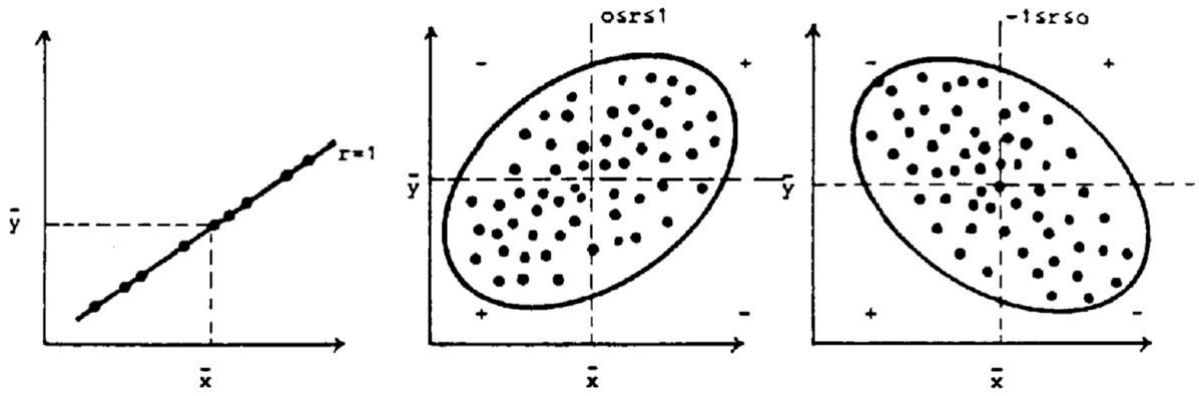
Einfacher Korrelationskoeffizient

Unter Korrelation versteht man den Grad des Zusammenhangs in der Punktwolke, von der die Regressionsgerade ja nur die Richtung angibt. Sie befasst sich also mit der Abhängigkeit der Zufallsgrößen (x, y) voneinander, mit der Streuung der Messwerte um die Regressionsgerade herum und wird quantitativ ausgedrückt durch den ***einfachen Korrelationskoeffizienten***:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad \text{mit} \quad \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad \bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} \; .$$

Der Korrelationskoeffizient r_{xy} kann ***Werte zwischen –1 und 1*** annehmen.
Ist $r_{xy} = 1$ oder $r_{xy} = -1$, gibt es ***keine Streuung*** um die Regressionsgerade, ***alle Punkte liegen genau auf der Geraden*** *(siehe Abb. 137)*. Je nach Vorzeichen hat die Gerade positive oder negative Steigung.
Ist $r_{xy} = 0$, so ist auch die Steigung der Regressionsgeraden Null. x und y heißen dann ***unkorreliert***. Insbesondere sind voneinander unabhängige Zufallsgrößen unkorreliert, die Punktwolke könnte dann z.B. kreisförmig sein. Jedoch müssen unkorrelierte Größen nicht unbedingt unabhängig sein.

Abbildung 137 **Abbildung 138** **Abbildung 139**



Multipler Korrelationskoeffizient

Der Korrelationskoeffizient lässt sich auch zur ***Bewertung der Güte der Anpassung*** bei der Methode der kleinsten Quadrate verwenden. Man ermittelt dann nicht den Zusammenhang zwischen x und y (dieser ändert sich ja nicht, wenn man an verschiedene Funktionen anpasst), sondern ***die Korrelation zwischen den Messwerten y und den Schätzwerten auf der Kurve***. Diese ist abhängig von der Güte der Anpassung:

$$r_{y\hat{y}} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \cdot \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}}_i)^2}}$$

$r_{y\hat{y}}$ kann Werte zwischen 0 und 1 annehmen. Bei $r_{y\hat{y}} = 1$ liegen alle Messwerte genau auf der Regressionsgeraden.

Wichtig:
Der multiple Korrelationskoeffizient ist eine Maßzahl für die Güte der Anpassung, der einfache Korrelationskoeffizient sagt hingegen nichts über die Qualität der Anpassung aus!

Das sollte man nach dem Besuch von Vorlesung und Übung beherrschen:

- Berechnung des arithmetischen Mittels eines Merkmals einer Stichprobe
- Bestimmung des Medians
- Berechnung von Stichprobenvarianz und Standardabweichung
- Was sagen der einfache und der multiple Korrelationskoeffizient aus?