Einführung in:
Schätzwert, Streuung, Standardfehler, Konfidenzintervall, p-Wert und Signifikanz

Theo ist ein interessierter Bienenzüchter der wissen will, wie schnell seine Bienen fliegen können. Bisher hat er die Geschwindigkeit von drei Bienen gemessen. Damit kann Theo bereits eine durchschnittliche Fluggeschwindigkeit ausrechnen. Wenn er in Zukunft *abschätzen* will, wie schnell eine Biene fliegen kann (wie schnell ist z.B. Biene Maja?), dann nimmt er den *Mittelwert* seiner drei gemessenen Bienen als *Schätzung*:

> Eine Biene fliegt "höchst"-wahrscheinlich mit der durchschnittlichen "Bienenfluggeschwindigkeit". Das leuchtet Theo ein. Aber ist das ein zuverlässiger *Schätzwert*, fragt er sich.

Stellen wir uns vor, dass die eine Biene sehr schnell ist, eine andere dagegen sehr langsam. Wenn Theo nun just drei besonders schnelle Bienen gemessen hätte, dann läge sein *Schätzwert* ja völlig daneben. Theo erkennt, dass für eine *zuverlässige Schätzung* eine möglichst grosse und *repräsentative* Bienen-*Stichprobe* notwendig ist: Je mehr Bienen Theo misst, umso zuverlässiger kennt er die wahre durchschnittliche Bienenfluggeschwindigkeit. Damit Theo sicher sein kann, dass seine Stichprobe gut ausgewählt ist, muss er die Auswahl einer höheren Instanz überlassen, dem sogenannten *Zufall*.

Mit der durchschnittlichen Bienenfluggeschwindigkeit wissen wir noch immer wenig darüber, wie schnell nun Biene Maja fliegen kann. Denn wir wissen nicht, ob alle Bienen in etwa gleich schnell fliegen, oder ob es ganz schnelle und ganz langsame Bienen gibt. Nur wenn alle Bienen gleich schnell fliegen könnten wir zuverlässig *schätzen*, wie schnell Biene Maja fliegt. Ansonsten brauchen wir für eine interpretierbare Schätzung ein Mass, welches uns angibt, wie unterschiedlich schnell die einzelnen Bienen sind:

> Wenn alle Bienen in etwa gleich schnell fliegen, dann ist die *Streuung* klein - wir haben dann eine kleine *Standardabweichung SD* (≅ kleine mittlere Abweichung).

Theo erkennt nun den Zusammenhang, dass eine kleine *Streuung* günstig ist für eine gute *Schätzung*. Wenn alle Bienen etwa gleich schnell, dann fliegt auch Biene Maja mit grosser Wahrscheinlichkeit in etwa mit dieser Geschwindigkeit. Die Genauigkeit des Mittelwerts als *Schätzwert* hängt von der *Streuung* ab - sie muss immer mit berücksichtigt werden.

Theo ist ein stolzer Bienenzüchter; er ist überzeugt, dass seine Bienen besonders schnell fliegen können. Er fordert deshalb seinen Imkernachbarn Bruno zum Bienenschnellflug-Duell auf. Weil Theo von seinen Bienen überzeugt ist, stellt er die Behauptung *(=Hypothese)* auf, dass seine Bienen schneller sind als die von Bruno. Weil sich diese generelle Aussage auf alle Bienen von Theo bezieht, müssten alle Bienen von Theo mit denen von Bruno, also jeweils die vollständigen *Populationen* verglichen werden. Beide merken rasch, dass dieser Aufwand nicht zu bewältigen ist. Wie kann nun dieses Duell trotzdem ausgetragen werden?

Wenn alle Bienen gemessen werden, dann kennt man zuverlässig und genau die wahre Durchschnitts-Bienenfluggeschwindigkeit. Misst man nur einen Teil der Bienen, dann kann man die wahre Durchschnittsgeschwindigkeit *schätzen* und die *Schätzwerte* (= Mittelwerte) vergleichen. Theo prüft eifrig einige seiner Bienen und vergleicht das Resultat mit einigen von Bruno's Bienen. Tatsächlich, Theo's Bienen sind schneller! Bruno meint, dass dies reiner Zufall sei; bei einer erneuten Messung könnten gerade so gut seine Bienen schneller sein. Wer hat nun Recht?

Hätte Theo alle seine Bienen gemessen und auch alle von Bruno, dann gäbe es keine Zweifel am Resultat, die ganze *Population* wäre erfasst und der *Zufall* (dass Theo etwa zufälligerweise von seinen Bienen nur besonders schnelle gemessen hat) wäre aus dem Spiel. Nun sind aber nur *Stichproben* gemessen worden. Die durchschnittliche Bienenfluggeschwindigkeit ist daher nicht der wahre/wirkliche/echte (Populations-)Wert, sondern eine *Schätzung* des wahren Wertes, die auch vom Zufall beeinflusst ist, und damit einen (*Zufalls-*)*Fehler* aufweist.

Wie lassen sich denn überhaupt Vergleiche anstellen, wenn doch beide *Schätzwerte* (Theo's & Bruno's mittlere Bienengeschwindigkeit) vom *Zufall* beeinflusst sind. Zum Glück verhält sich der *Zufall* nicht ganz "zufällig", sondern gemäss einer Gesetzmässigkeit, die übrigens Bernoulli als einer der ersten beschrieben hat (u.a. Gesetz der grossen Zahl). Letztendlich nur dank diesen Gesetzmässigkeiten lässt sich der (*Zufalls-*)*Fehler* aus den Daten erschliessen*:*

> *Je kleiner die Streuung, umso höher ist die Aussagekraft des Mittelwertes*, das hat Theo bereits begriffen. *Je mehr Messungen, umso höher die Zuverlässigkeit* - auch das hat Theo begriffen. Nun lassen sich beide Argumente vereinen:
>
> *Die Streuung wird mit der Anzahl Messungen verrechnet! (SD / Wurzel n)*
>
> Das Resultat ist der *Standardfehler (SE). Dieser gibt den* (*Zufalls-*)*Fehler* der *Schätzwerte an*: Ein kleiner *Fehler* bedeutet, dass der *wahre* (Populations-)Wert, also die echte durchschnittliche Bienenfluggeschwindigkeit, mit grosser Wahrscheinlichkeit in der Nähe des *Schätzwertes* liegt; ein grosser *Fehler* gibt an, dass der wahre Wert auch weiter vom *Schätzwert* entfernt sein könnte.

Der tatsächliche (Populations-)Wert liegt auf jeden Fall mit grosser Wahrscheinlichkeit in einem Intervall um den *Schätzwert* herum. Dieses *Konfidenzintervall* ist bei kleinem *Fehler* ebenfalls klein, d.h. umfasst einen kleinen Bereich. Ein kleines *Konfidenzintervall* bedeutet, dass der *Schätzwert* zuverlässig bzw. genau ist. Würde man die ganze *Population* messen, dann wäre dieser *Fehler* null. Der *Schätzwert* würde exakt dem wahren (Populations-)Wert entsprechen.

Nun können die beiden Schätzungen von Theo's und Bruno's Bienen miteinander verglichen werden. Ist die *Differenz* der beiden Mittelwerte (= *Effekt*) gross und der *Fehler* klein, dann können wir ziemlich sicher sein, dass die Bienen unterschiedlich schnell fliegen. Ist hingegen der *Effekt* gering und der *Fehler* gross, dann sind wir nicht sicher, ob die Bienen tatsächlich unterschiedlich schnell sind und der *Effekt* nicht rein *zufällig* zustande gekommen ist.

Bruno fällt auf, dass man nie hundertprozentig sicher sein kann - ausser man würde wirklich alle Bienen messen. Das ist korrekt, aber wir brauchen nicht hundertprozentige Sicherheit - es reichen auch 95 Prozent. Wir können zum Beispiel die Aussage machen, dass die Bienenvölker unterschiedlich schnell fliegen. Das Risiko, dass diese Aussage falsch ist, lässt sich anhand der *Mittelwerte* und *Standardfehlern* mittels einem *Signifikanztest* berechnen.

Ein *Signifikanztest* gibt an, wie gross die Wahrscheinlichkeit eines Irrtums ist, wenn wir behaupten, dass ein *Effekt* besteht. Je kleiner diese *Irrtumswahrscheinlichkeit* (= *p-Wert*) ist, desto sicherer bzw. *signifikanter* ist das Ergebnis. *P-Werte* kleiner als das übliche *alpha-Fehlernivea* von 5% werden als *signifikant* bezeichnet, d.h. es besteht dann ein "überzufälliger" und damit wahrer, nicht mit dem Zufall erklärbaren, d.h. wirklich bestehender *Effekt*!! Eine Aussage/*Hypothese* ist dadurch *empirisch* bestätigt (oder falsifiziert). Diese *inferenzstatistische* Methode bietet eine elegante und effiziente Möglichkeit, *allgemeingültige Aussagen* (die ganze *Population* betreffend) zu machen, ohne dass alle Bienen gemessen werden müssen. Replikationen erhöhen die *Bewährung* bzw. Sicherheit der Aussage.