Notgepäck Genauigkeit

Beat Hulliger
Dienst Statistische Methoden, Bundesamt für Statistik

20.4.2006

1 Was ist Genauigkeit genau?

Um zu beschreiben, was Genauigkeit in der Statistik ist, müssen wir untersuchen, wo sie auftritt: Wir möchten eine uns unbekannte Grösse θ kennen, z.B. die tatsächliche Erwerbslosenquote in der Schweiz oder den Zuckeranteil bei der Produktion eines Hustensirups. Wir nennen θ einen **Parameter**. Wir können aber θ nicht direkt messen, sondern müssen eine **Stichprobe** nehmen, z.B. eine Stichprobe von Personen oder eine Stichprobe von Flaschen mit Hustensirup, und verschiedene Eigenschaften dieser Einheiten messen. Die Messungen an den Einheiten der Stichprobe sind unsere Daten. Aus den Daten können wir einen **Schätzwert** für unseren Parameter berechnen (Z.B. den Mittelwert der Messungen). Dabei müssen wir uns bewusst sein, dass der Schätzwert (oder die Schätzung) auch anders herausgekommen sein könnte: Wenn wir die Messungen an einer anderen Stichprobe von Einheiten (Personen oder Hustensirup-Flaschen) wiederholen würden, dann würden andere Daten entstehen und der daraus berechnete Schätzwert könnte anders ausfallen[1]. Wir müssen also annehmen, dass unsere Schätzung vom wahren Wert θ abweicht und dass diese Abweichung bei jeder Stichprobe anders sein kann. Die Frage ist: Mit wieviel Abweichung vom Parameter θ müssen wir rechnen? Wie gross ist unsere Unsicherheit, wenn wir aus der Schätzung Aussagen über den wahren Parameter ableiten? Oder eben: Wie genau ist unsere Schätzung?

Wir müssen gut unterscheiden zwischen dem Verfahren, mit dem wir aus den Daten einer Stichprobe einen Schätzwert berechnen, und dem Schätzwert selbst. Das Verfahren nennen wir einen **Schätzer**. Wir bezeichnen den Schätzer mit T. Den Schätzwert, das Ergebnis der uns vorliegenden Anwendung des Schätzers, bezeichnen wir mit t [2]. Man nennt allgemein ein Verfahren, um aus Daten eine Aussage zu gewinnen, eine **Statistik**, und das Ergebnis der Statistik mit konkreten Daten eine **Realisation**. Ein Schätzwert ist also eine Realisation eines Schätzers. Zur Illustration: Wenn der Schätzer ein Kochrezept ist, dann ist der Schätzwert eine mit Hilfe des Rezepts gekochte Mahlzeit.

Wenn wir nun von Genauigkeit sprechen, dann meinen wir die Genauigkeit des Schätzers und nicht des Schätzwerts. Der Schätzwert ist ja eigentlich einfach eine Zahl und kann gar nicht ungenau sein. Wohl aber der Schätzer, dessen Realisationen vom tatsächlichen Wert abweichen und jedesmal wieder anders ausfallen können. Für ein Problem wie die

[1] Wir gehen hier nicht darauf ein, dass auch die Messung selbst einen Fehler haben kann.

[2] Es ist üblich, Statistiken mit grossen Buchstaben und Realisationen mit kleinen Buchstaben zu bezeichnen.

Erwerbslosenquote suchen wir also einen Schätzer, dessen Schätzwerte (die Realisationen) möglichst nahe beim uns unbekannten Parameter θ sind. Um dieses vage "möglichst nahe" präziser zu formulieren, müssen wir jetzt einige Begriffe einführen.

Wenn die Stichprobe **zufällig** ausgewählt worden ist, dann wird sich mit jeder Stichprobe zufällig ein anderer Schätzwert ergeben. Oft kann man in diesen zufälligen Variationen eine Gesetzmässigkeit entdecken: Wir können zwar nicht voraussagen, welcher Schätzwert bei der nächsten Anwendung des Schätzers realisiert werden wird, aber wir können sagen, welche Werte mit welcher Häufigkeit auftreten. Die Häufigkeiten, mit der bestimmte Werte auftreten, nennt man **Verteilung**.

Der **Erwartungswert** eines Schätzers T ist derjenige Wert, der dem Mittelwert von sehr vielen Realisationen der Statistik entspricht. Wir bezeichnen den Erwartungswert mit $E(T)$. Wenn man die Verteilung eines Schätzers kennt, dann kann man den Erwartungswert als eine mit den Häufigkeiten gewichtete Summe der möglichen Werte des Schätzers berechnen. Der Erwartungswert ist ein theoretische Grösse, die normalerweise bei einem praktischen Problem unbekannt ist. Mit anderen Worten kennt man den Erwartungswert meistens nicht, wenn man nur einen Schätzwert vor sich hat. Der Erwartungswert beschreibt aber eine interessante Eigenschaft des Schätzers, nämlich was im Mittel über viele Wiederholungen zu erwarten ist.

Der **Bias** oder **systematische Fehler** ist die Abweichung des Erwartungswerts eines Schätzers vom zu schätzenden Parameter:

$$B(T) = E(T) - \theta.$$

Der Bias beschreibt eine Abweichung, die bei jeder Schätzung in gleicher Art vorhanden ist. Z.B. wird ein Bias sichtbar, wenn man die Treffer eines Schützen mit einem Gewehr mit verzogenem Lauf untersucht: Die Treffer streuen um einen anderen Punkt als die Scheibenmitte.

Die Abweichung der Schätzwerte von θ enthält aber auch eine Komponente, die bei jeder Realisation wieder anders ist. Die Grösse dieser Abweichungen des Schätzers T wird durch die **Varianz** beschrieben. Die Varianz ist der Erwartungswert der quadratischen Abweichungen der Schätzwerte von ihrem Erwartungswert. Die Varianz ist also

$$\mathrm{Var}(T) = E\left([T - E(T)]^2\right)$$

Die Varianz eines Schätzers ist wie der Erwartungswert meistens unbekannt. Die mathematische Statistik hat nun aber gezeigt, dass die Varianz eines Schätzers oft aus der selben Stichprobe wie der Schätzer geschätzt werden kann. Wir gehen hier nicht näher darauf ein, warum und wie das gemacht werden kann. Im nachfolgenden Beispiel werden wir aber eine Schätzung für die Varianz berechnen. Der Einfachheit halber wird bei der Präsentation von Varianzen von Schätzern oft nicht erwähnt, dass es sich selbst wiederum um Schätzungen handelt.

In der Varianz tritt also die Abweichung der Schätzwerte von ihrem Erwartungswert $E(T)$ auf. Wenn der Schätzer einen Bias hat, interessiert auch die Abweichung der Schätzungen vom zu schätzenden Parameter θ. Dies misst der **mittlere quadratische Fehler**, der sich als Summe der Varianz und des quadrierten Bias ergibt. Im Allgemeinen wird die **Genauigkeit** durch den mittleren quadratischen Fehler beschrieben.

Oft gehen wir aber von Schätzern aus, die keinen Bias haben, also so genannt **erwartungstreu** sind. Die Genauigkeit eines erwartungstreuen Schätzers wird durch seine Varianz beschrieben: **Ein erwartungstreuer Schätzer ist umso genauer, je kleiner seine Varianz ist.**

Kann man einen Schätzer genauer machen? In gewissem Sinn ja. Die Varianz hängt nämlich direkt von der Anzahl Einheiten in der Stichprobe, der Stichprobengrösse, ab. Je grösser die Stichprobe ist, desto kleiner ist die Varianz. Darum gilt auch: Je grösser die Stichprobe, desto genauer der Schätzer. Leider gilt dies nicht für den Bias. Systematische Fehler bleiben gleich gross, unabhängig davon, wie viele Einheiten in der Stichprobe sind.

2 Wie kann man die Genauigkeit präsentieren?

Im folgenden nehmen wir an, dass unsere Schätzer erwartungstreu sind.

Die einfachste Beschreibung der Genauigkeit eines Schätzers ist die Standardabweichung. Die **Standardabweichung** $S(T)$ ist die quadratische Wurzel aus der Varianz, also $S(T) = \sqrt{\text{Var}(T)}$.

Wie die Varianz bezieht sich auch die Standardabweichung auf das Schätzverfahren, nicht auf die Schätzung. Sie beschreibt, um wieviel im Mittel über viele Realisationen die Schätzwerte abweichen würden. Aber wir haben natürlich meistens nur eine einzige Realisation anhand der konkreten Stichprobe, die uns vorliegt. Die Standardabweichung sagt uns also übersetzt: "Wir haben jetzt aus unserer Stichprobe eine Schätzung erhalten. Aber selbst wenn wir damit den wahren Wert genau getroffen hätten, müssen wir damit rechnen, dass weitere Schätzungen um eine Standardabweichung verschieden sein werden."

Ein Vorteil der Standardabweichung ist, dass sie die gleiche Masseinheit hat wie der Schätzer selbst. Die Angabe der Standardabweichung nimmt auch wenig Platz ein. Oft wird sie einfach in Klammern hinter der Schätzung angegeben (Z.B. Schätzung der Erwerbslosen in der Schweiz: $134'000(6'200)$.)

Manchmal wird auch die Standardabweichung mit dem Schätzwert verglichen, indem der so genannte **Variationskoeffizient** gebildet wird: $CV(T) = S(T)/T$. Der Variationskoeffizient ist also die Standardabweichung dividiert durch den Schätzer. Er wird üblicherweise in Prozent ausgedrückt. Sein Vorteil ist, dass die Genauigkeit von Merkmalen mit unterschiedlichen Grössenordnungen besser verglichen werden können. Der Variationskoeffizienten wird vor allem bei positiven, absoluten Zahlen verwendet. Bei der Schätzung von Anteilen, also Parametern die zwischen 0 und 1 liegen, ist die Standardabweichung eine bessere Beschreibung der Genauigkeit von Schätzern. Bei Anteilen kann neben der Standardabweichung $S(P)$ die normalisierte Standardabweichung $S'(P) = S(P)/\sqrt{P(1-P)}$ nützlich sein.

Um eine anschauliche Beschreibung der Genauigkeit eines Schätzers zu erhalten, wird oft ein **Vertrauensintervall** herangezogen. Das Vertrauensintervall ist ein Intervall, welches aus den Daten berechnet wird und welches mit grosser Wahrscheinlichkeit den Parameter θ enthält. Oder, um es mit Häufigkeiten auszudrücken, ein Vertrauensintervall enthält in fast allen Realisationen den Parameter θ. Das Vertrauensintervall ist also selbst eine Statistik. Das "Vertrauen" ist die Häufigkeit, mit beim Ziehen von vielen verschiedenen Stichproben die Realisationen des Vertrauensintervalls den Parameter θ enthielten.

Nun kann man unter bestimmten Voraussetzungen ein Vertrauensintervall so bauen, dass von einem Schätzer T ausgegangen wird und die Umgebung des Schätzers das Vertrauensintervall bildet. Dabei werden die Grenzen der Umgebung so gewählt, dass sie proportional zur Standardabweichung des Schätzers sind. Ein Vertrauensintervall ist dann also ein Bereich

$$[T - k\ S(T), T + k\ S(T)].$$

Die Konstante k muss vom Statistiker gewählt werden. Oft setzt man $k = 2$, damit man ziemlich hohes Vertrauen, nämlich 95%, in das Vertrauensintervall haben kann. Mit anderen Worten sollte ein Vertrauensintervall mit $k = 2$ bei vielen Realisationen etwa in 19 von 20 Fällen den Parameter θ enthalten. Solche Vertrauensintervalle werden oft mit der Angabe von $\pm kS(T)$ in Klammern nach dem Schätzwert angegeben, z.B. $134'000(\pm 12'400)$.

Dank der eingebauten Standardabweichung gibt so ein Vertrauensintervall auch eine Beschreibung der Genauigkeit des Schätzers ab: Je breiter das Vertrauensintervall ist, desto grösser ist die Standardabweichung und folglich desto ungenauer der Schätzer. Das Vertrauensintervall ist zudem intuitiv ansprechend, weil direkt ein Bereich bezeichnet wird, in dem der wahre Parameter liegen sollte. Man darf aber nicht vergessen, dass ein konkretes Vertrauensintervall nur eine bildliche Darstellung der Unsicherheit im Schätzverfahren ist. Für eine konkrete Realisation eines Vertrauensintervalls gilt nämlich schlicht, dass es entweder den wahren Parameter enthält oder eben nicht. Da gibt es keine Wahrscheinlichkeit mehr.

3 Ein fiktives Beispiel

Wir spielen hier die Berechnung einer Genauigkeit am bekannten Beispiel der Noten in einer Schulklasse durch. Wir nehmen an, eine Schulklasse habe eine Probe in Biologie geschrieben. Der Lehrer wählt zufällig 5 aus den 25 Probearbeiten der Klasse aus und bewertet sie mit den Noten 4, 4.5, 5, 4.5 und 2. Die Messungen sind die Bewertungen des Lehrers. Wir bezeichnen also die Noten mit $x_1 = 4, x_2 = 4.5, x_3 = 5, x_4 = 4.5, x_5 = 2$ und die Anzahl der Messungen ist $n = 5$.

Der Lehrer ist neugierig und möchte den Klassen-Durchschnitt θ abschätzen, ohne alle Probenarbeiten zu korrigieren. Er verwendet als Schätzer den Mittelwert der Stichprobe $\bar{X} = (X_1 + X_2 + X_3 + X_4 + X_5)/5 = (\sum_{i=1}^{5} X_i)/n$. Der Schätzwert ist dann also $\bar{x} = (4 + 4.5 + 5 + 4.5 + 2)/5 = 4$

Ein Schätzer für die Varianz des Mittelwerts Var($\bar{X}$)ist

$$S^2(\bar{X}) = \frac{\sum_{i=1}^{5}(X_i - \bar{X})^2}{n(n-1)}.$$

Die Realisierung in unserem Fall ist $s^2 = (0^2 + 0.5^2 + 1^2 + 0.5^2 + 2^2)/(5 \times 4) = 0.275$. Die Schätzung für die Standardabweichung ist die Wurzel aus s^2, d.h. $s = \sqrt{0.275} \approx 0.52$. Wenn der Lehrer also 5 andere Probearbeiten in seiner Stichprobe aufnehmen würde, dann müsste er damit rechnen, dass der Mittelwert um etwas mehr als eine halbe Note abweichen würde.

Wir können auch eine Realisation eines Vertrauensintervalls berechnen:

$$[\bar{x} - 2s, \bar{x} + 2s] = [2.96; 5.04].$$

Es wird etwas gross, beschreibt aber die Unsicherheit des Lehrers gut: Wenn er bei vielen vergleichbaren Proben jeweils nur auf Grund von 5 Probearbeiten den Klassendurchschnitt vorhersagen müsste, dann würde er sich nur selten irren, wenn er sich jeweils nur so weit festlegt wie das obige Vertrauensintervall $4(\pm 1.04)$.

4 Genauigkeit in der öffentlichen Statistik

Die Genauigkeit einer statistischen Information wird durch verschiedene Arten von möglichen Fehlern beeinflusst. Alle Erhebungen, auch Registererhebungen, sind z.B. Messfehlern unterworfen. Ergebnisse aus Stichprobenerhebungen haben zusätzlich einen Stichprobenfehler.

Je höher die Anforderungen an die Genauigkeit einer Statistik, desto höher fallen die Kosten aus. Dies gilt für den Stichprobenfehler, der umso kleiner ist, je grösser die Stichprobe ausfällt. Es gilt aber auch für Messfehler, deren Reduktion schwierig und aufwändig ist. Zum Beispiel ist bei der Einkommens- und Verbrauchserhebung der Aufwand, der bei der Bereinigung der von den Haushalten notierten Ausgaben anfällt, beträchtlich.

Angaben über die Genauigkeit von statistischen Informationen der öffentlichen Statistik finden sich in den Methodenkapiteln und Methodenberichten über die entsprechende statistische Aktivität. Das Bundesamt für Statistik ist bestrebt, über die Qualität der statistischen Informationen benutzergerecht zu informieren.

Die Genauigkeit einer statistischen Information ist grundlegender Bestandteil der Diskussion über eine statistische Aktivität in den Experten- und Begleitgruppen der öffentlichen Statistik. Die Benutzer müssen die Anforderungen an die Genauigkeit so ansetzen, dass das Risiko von Fehl-Entscheidungen aufgrund der statistischen Informationen, vertretbar ist. Durch den Dialog mit den wichtigen Benutzern wird sichergestellt, dass die Genauigkeit der veröffentlichten Informationen einer optimalen Abwägung der Bedürfnisse der Benutzer, der Risiken und der Kosten entspricht.

Immer häufiger wird die Anforderung an die Genauigkeit von statistischen Ergebnissen auch in Reglementen der Eropäischen Union festgehalten (z.B. wurden für die Arbeitskräfteerhebung verschiedene Kriterien für die Genauigkeit der Resultate festgelegt).