Korrelationen im Länderquartett

$p(😀|🎲) = 1$

Ein Unterrichtsentwurf aus der Reihe **Probably Fun – Games to teach Statistics**

von Dr. Kristian Rother (www.academis.eu/probably_fun/)

Unterrichtsziel

Die Teilnehmer zeichnen eine Ausgleichsgerade und berechnen einen Korrelationskoeffizienten.

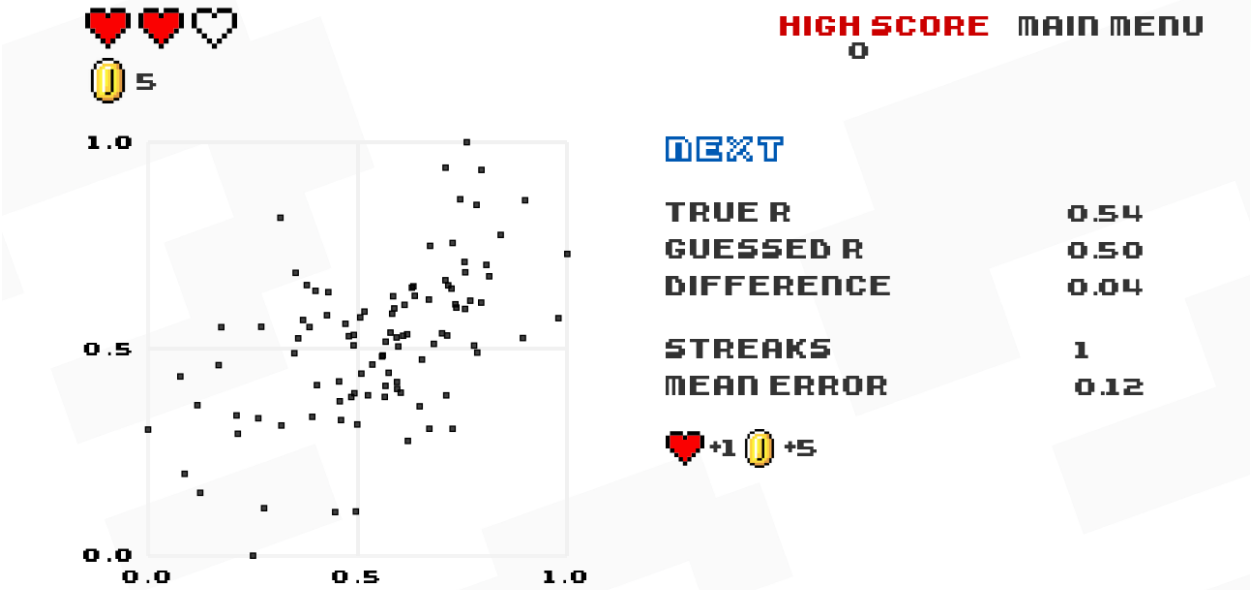
Zeit

90 Minuten

Begriffe

- Streudiagramm
- logarithmische Skala
- Lineare Regression
- Geradengleichung
- Mean Squared Error
- Korrelationskoeffizient

Das Spiel: Guess the Correlation



Hättest du diesen Korrelationskoeffizienten korrekt geschätzt?

Guesss the Correlation (www.guessthecorrelation.com/) von Omar Wagih ist ein kleines online-Spiel im Retro-Stil, bei dem die Spieler eine Wolke von Datenpunkten gezeigt bekommen und den Korrelationskoeffizienten erraten müssen. Je genauer der Wert geschätzt wird, desto mehr Punkte gibt es. Kein Gesellschaftsspiel im herkömmlichen Sinne, aber es passt trotzdem gut in diese Sammlung.

Unterrichtsablauf

Diese Lektion ist die logische Fortsetzung der Unterrichtseinheiten mit den Quartettkarten. Dieses Mal widmen wir uns dem **Streudiagramm** und daraus abgeleiteten statistischen Werkzeugen. Hier werden die Materialien weiter verwendet. Die Begriffe bauen aber aufeinander auf, und ich hielt diese Lektion für wichtig.

Der Unterricht zum Thema ist zweigeteilt: Im ersten Teil geht es um Streudiagramme von **Fläche** über **Bevölkerung** in den Ländern im Quartett. Der Effekt der linksschiefen Verteilung (viele kleine Länder) wird mathematisch durch die **doppelt logarithmische Darstellung** ausgeglicheen. Mit einem guten Werkzeug läßt sich bei dieser Gelegenheit auch das Histogramm einer Variablen vor und nach dem Logarithmieren darstellen. Am Ende sollt auf jeden Fall eine **lineare Regression** in Form einer **Ausgleichsgerade** stehen.

$p(😀 \mid 🎲) = 1$

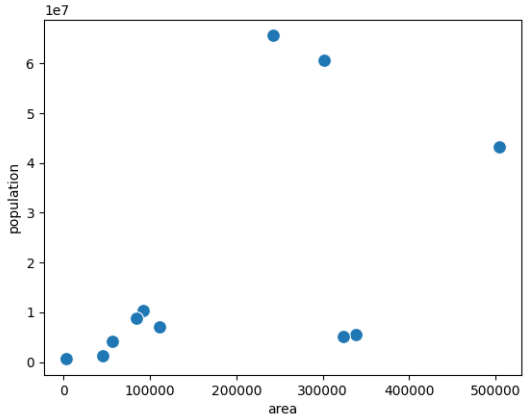
Im zweiten Teil wird der Begriff der **Korrelation** eingeführt. Die gebräuchliche Formel für den **Korrelationkoeffizienten** ist wegen der Kovarianzen leider für die meisten Gelegenheitsstatistiker etwas schwer verdaulich. Hat man aber soeben die lineare Regression vollendet, läßt sich der Korrelationskoeffizient r bequem aus der Steigung und den Standardabweichungen beider Variablen berechnen:

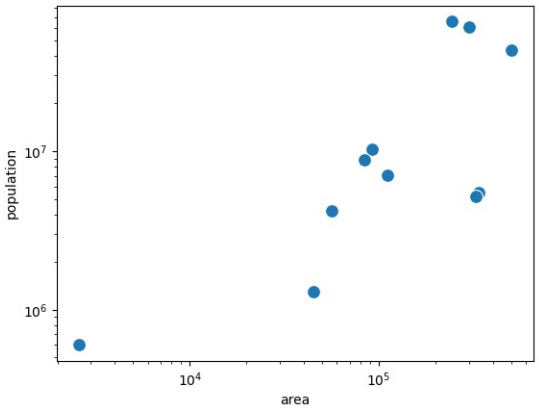
$$r = \frac{m \cdot std(x)}{std(y)}$$

Schritt	Aktivität	Zeit
1.	Teile je 16 Quartettkarten mit Ländern an Gruppen von Teilnehmern aus	1’
2.	Lass die Teilnehmer ein Streudiagramm von **Bevölkerung** über **Fläche** zeichnen (auf Papier oder mit einem elektronischen Werkzeug)	15’
3.	Betrachtet und besprecht gemeinsam die Ergebnisse (es sollte auffallen, dass die Punkte sich einseitig ballen, aber grundsätzlich größere Länder auch mehr Leute haben)	10’
4.	Zeichnet die gleichen Daten gemeinsam auf einer einfach logarithmischen und einer doppelt logarithmischen Skala. Bespreche genau was ein Logarithmus tut, zur Veranschaulichung kann das Histogramm der Variablen helfen.	10’
5.	Zeichnet eine Ausgleichsgerade. Idealerweise zunächst von Hand an der Tafel. Besprecht die Geradengleichung und schätzt Steigung und Achsenabschnitt. Nehmt den fit danach automatisiert vor und prüft. Berechnet gegebenenfalls die quadratische Abweichung.	20’
6.	Spielt einige Runden www.guessthecorrelation.com	10’
7.	Frage was hohe und niedrige Werte für r sind. Stelle danach die Gleichung für den Korrelationskoeffizienten vor. Erwähne negative Korrelation (kommt im Spiel nicht vor!)	10’
8.	Berechnet den Korrelationskoeffizienten für ein Beispiel manuell	10’
9.	Erörtert einige der Reflexionsfragen	10’

Ergebnisse

Einige mit Python erstellte Diagramme für eine Stichprobe von Länderkarten. Links ein gewöhnliches Streudiagramm, rechts die doppelt logarithmische Darstellung. Der Ausreißer links unten (Luxemburg) ist links nicht ohne weiteres erkennbar.





$p(😀 \mid 🎲) = 1$

Fortsetzung

Natürlich läßt sich eine Menge mehr zu linearer Regression sagen. Für den Erstkontakt genügt eine vollautomatische 2D-Ausgleichsgerade wie sie von Tabellenkalkulationen angeboten wird. Sollte der Kurs sich aber in Richtung maschinelles Lernen entwickeln, könnten folgende Themen angbaut werden:

- multiple lineare Regression
- bei mehreren Variablen: Korrelationsmatrix berechnen
- partielle Ableitungen
- analytische Lösung (Normal Equation)
- Gradientenmethode
  (diese läßt sich übrigens prima über das Partyspiel “Topfschlagen” veranschaulichen)
- Regularisierung (Ridge, Lasso)
- Annahmen linearer Modelle und das Gauß-Markov-Theorem

Da wir uns hier aber vom Einsatz von Spielen entfernen, werden diese Themen an anderer Stelle ausgeführt.

Reflexionsfragen für den Unterricht

- Nenne ein Beispiel für hohe und niedrige Korrelation aus dem Alltag
- Wenn eine hohe Korrelation vorliegt, bedeutet dies auch einen kausalen Zusammenhang?
- Was bedeutet der Begriff “statistisch unabhängig”?

Links

- Spurious Correlations
  tylervigen.com/spurious-correlations
- Maschinelles Lernen auf Academis
  www.academis.eu/machine_learning/