Aufgabe 9) Angenommen Sie vergleichen Ihren neu entwickelten Spam-Klassifizierer *Spammy* mit einem Baseline-Spam-Klassifizierer und erhalten folgende Ergebnisse:

Goldstandard	Baseline	Spammy	Häufigkeit
Spam	Spam	Spam	57
Spam	Spam	NoSpam	5
Spam	NoSpam	Spam	7
Spam	NoSpam	NoSpam	3
NoSpam	Spam	Spam	2
NoSpam	Spam	NoSpam	6
NoSpam	NoSpam	Spam	2
NoSpam	NoSpam	NoSpam	154

Hier gab es bspw. 57 Emails, die sowohl vom Baseline-Tagger als auch von Spammy korrekt als Spam klassifiziert wurden.

Sagen Sie so genau wie möglich, wie Sie hier mit dem **Vorzeichentest** berechnen, ob Spammy signifikant besser als der Baseline-Klassifikator ist. (3 Punkte)

Vorlesung-SS20-WH.pdf

Goldstandard	Baseline	Spammy	Häufigkeit
Spam	Spam	Spam	57
Spam	Spam	NoSpam	5
Spam	NoSpam	Spam	7
Spam	NoSpam	NoSpam	3
NoSpam	Spam	Spam	2
NoSpam	Spam	NoSpam	6
NoSpam	NoSpam	Spam	2
NoSpam	NoSpam	NoSpam	154

hier kan man auch
b(>=13,0.5,20) schreiben

Antwort:

Man zählt zunächst, wieviele Emails Spammy richtig klassifiziert und der andere Tagger falsch. Das sind 13. Dann zählt man, wieviele Emails Spammy falsch klassifiziert und der andere Tagger richtig. Das sind 7. Nur diese 20 Beispiele sind für den Vorzeichentest relevant. Die Nullhypothese besagt, dass Spammy nicht besser als der andere Tagger ist. Die Wahrscheinlichkeit, dass Spammy ein beliebiges der 20 Beispiele korrekt klassifiziert hat, ist daher unter Annahme der Nullhypothese maximal 0.5. Die Wahrscheinlichkeit, dass man bei Gültigkeit der Nullhypothese das beobachtete Ergebnis (Spammy 13 Mal korrekt) oder ein noch unwahrscheinlicheres Ergebnis (≥ 13) bekommt, ist durch die Summe $\sum_{i=13}^{20} b(i, 0.5, 20)$ gegeben, wobei $b(r, p, n)$ die Binomialverteilung mit Wahrscheinlichkeit p und Stichprobengröße n ist. Wenn diese Summe kleiner als 0.05 ist, kann die Nullhypothese zurückgewiesen werden. Man sagt dann: Spammy hat eine signifikant höhere Genauigkeit.

Aufgabe 9) Angenommen Sie vergleichen Ihren neu entwickelten Spam-Klassifizierer *Spammy* mit einem Baseline-Spam-Klassifizierer und erhalten folgende Ergebnisse:

Goldstandard	Baseline	Spammy	Häufigkeit
Spam	Spam	Spam	57
Spam	Spam	NoSpam	5
Spam	NoSpam	Spam	7
Spam	NoSpam	NoSpam	3
NoSpam	Spam	Spam	2
NoSpam	Spam	NoSpam	6
NoSpam	NoSpam	Spam	2
NoSpam	NoSpam	NoSpam	154

Hier gab es bspw. 57 Emails, die sowohl vom Baseline-Tagger als auch von Spammy korrekt als Spam klassifiziert wurden.

Sagen Sie so genau wie möglich, wie Sie hier mit dem **Vorzeichentest** berechnen, ob Spammy signifikant besser als der Baseline-Klassifikator ist. (3 Punkte)

Baseline richtig getaggt(und Spammy falsch getaggt) = 5 +2 = 7

Spammy richtig getaggt (und Baseline falsch getaggt) = 7 + 6 = 13

insgesamt = 17+3 = 20 Samples

Nullhypothese: Spammy ist nicht besser als Baseline

Berechne
b(>= 13, 0.5, 20) = …

Falls der Wert <= 0.05, dann heißt das, dass Spammy signitikant besser als Baseline ist.

Signifikanztest

Beim Vergleich zweier Systeme ist es wichtig zu wissen, ob der Unterschied zwischen ihren Evaluierungsergebnissen **signifikant** (also bedeutsam) ist oder ob er Zufall sein könnte.

Dazu wenden wir einen **Signifikanztest** an. Wir werden den Vorzeichentest nehmen.

Vorzeichentest zur Taggerevaluierung

Beim Vorzeichentest interessieren nur die Wörter, die **genau einer** der Tagger falsch annotiert hat. Alle anderen werden ignoriert. Angenommen es gibt **60** Wörter in den Testdaten, die NewTagger richtig taggt und OldTagger falsch, und **40** Wörter, die OldTagger richtig taggt und NewTagger falsch.

Nullhypothese: NewTagger ist nicht besser als OldTagger.

⇒ Bei jedem der 100 Wörter ist die Wahrscheinlichkeit, dass NewTagger richtig taggte, maximal 0.5.

Dies ist ein einseitiger Binomialtest, weil uns nur interessiert, ob NewTagger signifikant besser ist, nicht aber, ob er signifikant schlechter ist.

Wir summieren die Werte der Binomialfunktion:

$$b(\geq 60, 0.5, 100) \approx 0.03$$

⇒ Der Unterschied ist signifikant mit einer Fehlerwk. von etwa 3 %.

Wir haben die Evaluationsergebnisse von beiden Taggers gesammelt und wollen wissen, ob man aus diesen Ergebnisse sagen kann, dass der neue Tagger besser ist als der alte.

Wir benutzen dafür den Binomialtest. Der test kann uns sagen, ob das Ergebnis vom neuen Tagger(60 richtig getaggte Wörter) signifikant besser ist oder es nicht genug ist, so zu betimmen.

Wir stellen zuerst eine Nullhypothese. Wir nehmen an, dass der neue Tagger nicht besser ist als der alte. D.h. die Wk dass der neue Tagger ein Wort richtig taggt ist maxmimal 0.5 (p=0.5).
Dann berechnen wir die Binomial-Wk b(>=60, 0.5,100) unter der Nullhypothese.
Diese Wk sagt uns, wie wahrscheinlich ist es, dass der neue Tagger 60 Wörter oder mehr richtig taggt, wenn p = 0.5.
Wenn 60 ist normal(zu erwarten/hoch wahrscheinlich) bei p=0.5, dann erwarten wir, dass b(>=60, 0.5,100) eine hohe Wk ergibt. Wenn aber es unwahrscheinlich ist, dann können wir sagen, es ist höher als erwartet. (p ist dann wohl nicht 0.5 sonder höher)

Wir berechnet nicht nur X = 60 (X=richtig taggt)sondern X>= 60, um sicher zu gehen, dass obwohl wir viele Binomial-Wk (X=60,X=61,…,X=100) summieren, bekommen wir immer noch eine niedrige Wk.

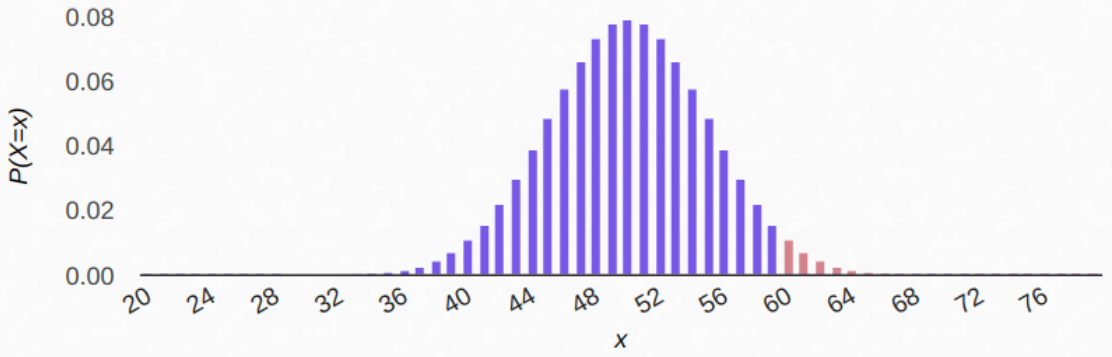
Wenn diese Wk ist kleiner als 0.05, dann können wir die Nullhypothese verwerfen und sagen, das Ergebnis (60 richtig getaggte Wörter) ist ein sigifikant besseres ergebnis. Also, neue Tagger ist signifikant besser als der alte.

Binomial Distribution

$X \sim Bin(n, p)$

$n =$ 100 $p =$ 0.5

$x =$ 60 P(X ≥ x) = 0.02844



$\mu = E(X) = 50$ $\sigma = SD(X) = 5$ $\sigma^2 = Var(X) = 25$

Behandlung unbekannter Wörter

Die Apriori-Wahrscheinlichkeit $p(t) = f(t)/N$ (N = Korpusgröße) der Tags kann leicht aus dem Trainingskorpus geschätzt werden.

Die bedingte Wahrscheinlichkeit $p(t|w)$ kann anhand des Wortsuffixes (z.B. der Länge l) geschätzt werden:

$$p(t|w) = p(t|a_1 a_2 ... a_n) \approx p_{suff}(t|a_{n-l+1} ... a_n)$$

Why with +1?

$p(\text{NN}|\text{schwärzlich}) \approx p_{suff}(\text{NN}|\text{zlich})$ falls $l = 5$

Beispiel:
a1,a2,..an
w = “house”
l=3
n=5
suffix should start at a3 (“use” = a3,a4,a5)
n-l+1 = 5-3+1 = 3

20.01.21

Könntest du bitte die **“Uebung-WS17-18.pdf”** und “**Uebung-WS16-17-WH.pdf”** die Aufgabe 6. aus den Altklausuren lösen? BITTE!!

Uebung-WS17-18.pdf

Lösungvorschlag:
https://colab.research.google.com/drive/1Wqw9InSPfpb_D7hXHyJ6GnL3GiShSR2i?usp=sharing