Προγραμματισμός με τη γλώσσα python

Alexandros Kanterakis kantale@ics.forth.gr (kantale@ics.forth.gr)

Διάλεξη 12η, Τρίτη 21 Ιανουαρίου 2020

Ας δούμε αυτό το tweet: https://twitter.com/meganinlisbon/status/1101870079858409478 (https://twitter.com/meganinlisbon/status/1101870079858409478)

```
I presented the math for this at the #cosyne19 diversity lunch today.

Success rates for first authors with known gender:
Female: 83/264 accepted = 31.4%
Male: 255/677 accepted = 37.7%

37.7/31.4 = a 20% higher success rate for men
```

Άραγε υπήρχε όντως κάποια προκατάληψη ενάντια στις γυναίκες κατά τη κρίση των δημοσιεύσεων σε αυτό το συνέδριο;

Για να το λύσουμε αυτό πρέπει να κάνουμε τον εξής συλλογισμό:

Είτε υπάρχει όντως προκατάληψη είτε το αποτέλεσμα αυτό βγήκε κατά τύχη! Οπότε πρέπει να υπολογίσουμε τη πιθανότητα αυτή η ανισορροπία στα ποσοστά να "βγήκε" κατά τύχη. Εδώ όμως υπάρχει ένα πρόβλημα: Το να βρούμε τη πιθανότητα δεν είναι αρκετό. Για παράδειγμα ας υποθέσουμε ότι βρίσκουμε ότι η πιθανότητα αυτή είναι 10%. Υπάρχει προκατάληψη ή δεν υπάρχει; Πρέπει λοιπόν να ορίσουμε ένα όριο και να πούμε ότι αν η πιθανότητα που θα βρούμε είναι μικρότερη από αυτό το όριο τότε δεν δεχόμαστε ότι αυτή η ανισορροπία βγήκε κατά τύχη. Αυτό το όριο το ονομάζουμε significance level και συνήθως συμβολίζεται με α.

Μερικά πράγματα για το significance level:

- Παραδοσιακά στις ιατρικές επιστήμες χρησιμοποιούμε το α=0.05
- Μεθοδολογικά, το α πρέπει να οριστεί κατα τον σχηματισμό της υπόθεσης την οποία ερευνούμε (π.χ. υπάρχει προκατάληψη;). Αυτός/αυτή που θα ορίσει το α δεν πρέπει να έχει "δει" τα δεδομένα πριν (αλλά και κανένα είδος ανάλυσης επί των δεδομένων)!
- Υπάρχει μια τεράστια συζήτηση τα τελευταία χρόνια για το αν πρέπει το α να είναι μικρότερο (έχει προταθεί το α=0.005) ή να μην ορίζεται μια σταθερά για όλους, αλλά μια σταθερά για κάθε είδος έρευνας.
- Διαβάστε: https://en.wikipedia.org/wiki/Statistical_significance (https://en.wikipedia.org/wiki/Statistical_significance) και κυρίως το κεφάλαιο "Challenges (https://en.wikipedia.org/w/index.php?title=Statistical_significance&action=edit&section=8)".

Τι σημαίνει όμως αυτό το α=0.05; Ας το αναλύσουμε λίγο περισσότερο. Ακούμε στη καθημερινότητά μας προτάσεις όπως "κάθε φορά που βγαίνω χωρίς ομπρέλα βρέχει!", "το λεωφορείο έρχεται μόλις ανάβω τσιγάρο!", "αυτός ο διατητής μας αδικεί. Κάθε φορά που έχουμε αυτόν/αυτήν χάνουμε!", "κλασσικά τη πρωτοχρονιά έχουμε πολλά συνεχόμενα jack-pot στο joker". Όλα αυτά τα λέμε από παρατηρήσεις που οφείλονται είτε σε πραγματικά φαινόμενα είτε στη τυχαιότητα. Διαβάστε σχετικά και το βιβλίο Fooled by Randomness (https://en.wikipedia.org/wiki/Fooled_by_Randomness). Για να κάνουμε λοιπόν αυτή τη διάκριση (μεταξύ τυχαίου και πραγματικού) λέμε: "αν η πιθανότητα το φαινόμενο που έχεις παρατηρήσει, να συμβαίνει στη τύχη είναι μικρότερη από 5%, τότε το φαινόμενο είναι πραγματικό!".

Πάμε λοιπόν να μελετήσουμε το δεδομένα από το tweet.

Για αρχή ας υπολογίσουμε τα ποσοστά των accepted papers για τους άνδρες και τις γυναίκες

In [1]:
```
import random
```

In [204]:
```
significance_threshold = 0.05

f_pass = 83
f_total = 264
f_rejected = f_total-f_pass
m_pass  = 255
m_total = 677
m_rejected = m_total-m_pass



pass_total = f_pass + m_pass
rejected_total = f_rejected + m_rejected
total = f_total + m_total

print ('Total:', total)
print ('pass_total:', pass_total)
w_conf = f_pass/f_total
m_conf = m_pass/m_total
print ('w_conf:', w_conf)
print ('m_conf:', m_conf)
found_difference = m_conf-w_conf
print ('Percentage difference:', found_difference)
```

```
Total: 941
pass_total: 338
w_conf: 0.3143939393939394
m_conf: 0.3766617429837518
Percentage difference: 0.06226780358981243
```

Τώρα ας υποθέσουμε ότι δεν υπάρχει καμία προκατάληψη. Φτιάχνουμε έναν πίνακα με όλα τα papers. Όπου θα βάλουμε ότι πέρασαν pass_total και ότι απορίφθηκαν rejected_total:

In [199]:
```
papers = [True]*pass_total +  [False]*rejected_total
```

Ας ανακατέψουμε τυχαία αυτόν τον πίνακα!

In [200]:
```
random.shuffle(papers)
```

Ας υποθέσουμε τώρα ότι τα πρώτα 264 τα έστειλαν οι γυναίκες και τα υπόλοιπα 677 οι άντρες. Αφού ο πίνακας papers έχει ανακατευτεί δεν έχει νόημα ποιον θα βάλουμε πρώτο ή δεύτερο.

In [201]:
```
women_send = papers[:f_total]
male_send = papers[f_total:]
```

Ας μετρήσουμε τώρα τα ποσοστά επιτυχίας των ανδρών και των γυναικών:

In [202]:
```
perc_women = sum(women_send)/len(women_send)
perc_men = sum(male_send)/len(male_send)
print ('perc_women:', perc_women)
print ('perc_men:', perc_men)
```

```
perc_women: 0.36363636363636365
perc_men: 0.35745937961595275
```

Ποια είναι η διαφορά στα ποσοστά που βρήκαμε;

In [205]:
```
difference = perc_men-perc_women
difference
```

Out[205]:
```
-0.006176984020410892
```

Θυμηθείτε ότι η διαφορά που παρατηρήσαμε εμείς είναι:

In [206]:
```
found_difference
```

Out[206]:
```
0.06226780358981243
```

Ποια είναι η πιθανότητα το difference να είναι μεγαλύτερο από το found_difference; Δηλαδή ποια είναι η πιθανότητα να βρούμε αυτή τη διαφορά στα ποσοστά κατά τύχη. Ας τρέξουμε το παραπάνω πείραμε 100.000 φορές και ας το μετρήσουμε!

In [211]:
```
papers = [True]*pass_total +  [False]*rejected_total
def f():
    random.shuffle(papers)
    women_send = papers[:f_total]
    male_send = papers[f_total:]

    perc_women = sum(women_send)/len(women_send)
    perc_men = sum(male_send)/len(male_send)

    difference = perc_men-perc_women

    return difference >= found_difference

STEPS=100000
sum(1 for x in range(STEPS) if f())/STEPS
```

Out[211]:
```
0.04288
```

Ας αναλύσουμε λίγο τι είναι το 0.04288. Αν διοργανώσουμε 100.000 συνέδρια τα οποία δεν έχουν κανένα gender bias, όπου και στα 100.000 συνέδρια στείλουμε 941 papers όπου τα 338 γίνονται δεκτά. Και επίσης, και στα 100.000 συνέδρια συμμετέχουν 264 γυναίκες και 677 άντρες. Τότε, από τα 100.000 συνέδρια, στα 4.288, το ποσοστό των accepted στους άνδρες είναι μεγαλύτερο κατά 0.06226780358981243 από ότι το ποσοστό των accepted στις γυναίκες.

Μπορούμε να θέσουμε αυτό το ποσοστό (4.3%) ως τη πιθανότητα ένα gender unbiased συνέδριο να εμφανίσει αυτό τη διαφορά στα ποσοστά των accepted μεταξύ αντρών και γυναικών.

Αφού αυτή η πιθανότητα είναι κάτω από το 5% (significance_threshold = 0.05) λέμε ότι το συνέδριο ΔΕΝ είναι αμερόληπτο και ότι δείχνει μία προτίμηση στο να κάνει accept papers από άντρες.

Παρατηρούμε ότι αυτή η διαδικασία είναι χρονοβόρα υπολογιστικά. Υπάρχει τρόπος να το υπολογίσουμε αυτό με αναλυτικό τρόπο; Ναι! Στη πραγματικότηταα αυτό που κάναμε ήταν ένα Fisher's exact test (https://en.wikipedia.org/wiki/Fisher%27s_exact_test):

In [217]:
```
import numpy as np
from scipy import stats

import scipy.stats as stats
obs = np.array([[m_pass, f_pass], [m_rejected, f_rejected]])
oddsratio, pvalue = stats.fisher_exact(obs, alternative='greater')
pvalue
```

Out[217]:
```
0.04272336653184806
```

Γιατί κάναμε τον αλγοριθμικό τρόπο πρώτα και όχι τον αναλυτικό που στο κάτω-κάτω είναι και πολύ πιο γρήγορος;

Να γιατί: https://www.youtube.com/watch?v=5Dnw46eC-0o (https://www.youtube.com/watch?v=5Dnw46eC-0o)

Ένα άλλο παράδειγμα:

Παίρνουμε 220 ανθρώπους και κοιτάμε αν έχουν μία μετάλλαξη και αν έχουν μία ασθένεια.

ΠΛΗΘΟΣ	Μετάλλαξη=ΝΑΙ	Μετάλλαξη=ΟΧΙ
ΠΑΘΗΣΗ=ΝΑΙ	10	100
ΠΑΘΗΣΗ=ΟΧΙ	40	120

Το ερώτημα είναι.. συσχετίζεται αυτή η μετάλλαξη με αυτή την ασθένεια;

Για αρχή ας μετατρέψουμε τους παραπάνα αριθμούς σε ποσοστά:

Ας βάλουμε τα δεδομένα:

In [329]:
```
disease_mut = 60
disease_nomut = 100
healthy_mut = 40
healthy_nomut = 120
disease = disease_mut + disease_nomut
healthy = healthy_mut + healthy_nomut
mutation = disease_mut + healthy_mut
no_mutation = disease_nomut + healthy_nomut
total = disease + healthy

disease_mut_perc = disease_mut/total
disease_nomut_perc = disease_nomut/total
healthy_mut_perc = healthy_mut/total
healthy_nomut_perc = healthy_nomut/total

print ('disease_mut_perc:', disease_mut_perc)
print ('disease_nomut:', disease_nomut_perc)
print ('healthy_mut_perc:', healthy_mut_perc)
print ('healthy_nomut_perc:', healthy_nomut_perc)
```

```
disease_mut_perc: 0.1875
disease_nomut: 0.3125
healthy_mut_perc: 0.125
healthy_nomut_perc: 0.375
```

ΠΛΗΘΟΣ	Μετάλλαξη=ΝΑΙ	Μετάλλαξη=ΟΧΙ
ΠΑΘΗΣΗ=ΝΑΙ	18.75%	31.25%
ΠΑΘΗΣΗ=ΟΧΙ	12.5%	37.5

Ας υποθέσουμε ότι η μετάλλαξη δεν συσχετίζεται καθόλου με την ασθένεια.

Αν ισχύει αυτό, τότε θα πρέπει η πιθανότητα να έχουν την ασθένεια αυτοί που έχουν τη μετάλλαξη να είναι ίδια με τη πιθανότητα να έχουν την ασθένεια αυτοί που δεν έχουν τη μετάλλαξη. Ποια είναι αυτή η πιθανότητα όμως; ΜΑ.. η πιθανότητα να έχει κάποιος την ασθένεια! η οποία είναι:

In [330]:
```
p_disease = disease / total
print ('p_disease:', p_disease)
```

```
p_disease: 0.5
```

Αν ισχύει λοιπόν ότι η μετάλλαξη δεν ευθύνεται καθόλου με την ασθένεια, τότε θα έπρεπε το πλήθος αυτών που έχουν τη μετάλλαξη και την ασθένεια να είναι `expected_disease_mut = p_disease * mutation` ομοίως για τους υπόλοιπους:

In [331]:
```
expected_disease_mut = p_disease * mutation
expected_disease_nomut = p_disease * no_mutation
expected_healthy_mut = (1-p_disease) * mutation
expected_healthy_nomut = (1-p_disease) * no_mutation

print ('====EXPECTED VALUES====')
print ('expected_disease_mut:', expected_disease_mut)
print ('expected_disease_nomut:', expected_disease_nomut)
print ('expected_healthy_mut:', expected_healthy_mut)
print ('expected_healthy_nomut:', expected_healthy_nomut)

expected_disease_mut_perc = expected_disease_mut / total
expected_disease_nomut_perc = expected_disease_nomut / total
expected_healthy_mut = expected_healthy_mut / total
expected_healthy_nomut = expected_healthy_nomut / total

print ('=====EXPECTED PERCENTAGES=====')
print ('expected_disease_mut_perc:', expected_disease_mut_perc)
print ('expected_disease_nomut_perc:', expected_disease_nomut_perc)
print ('expected_healthy_mut:', expected_healthy_mut)
print ('expected_healthy_nomut:', expected_healthy_nomut)
```

```
====EXPECTED VALUES====
expected_disease_mut: 50.0
expected_disease_nomut: 110.0
expected_healthy_mut: 50.0
expected_healthy_nomut: 110.0
=====EXPECTED PERCENTAGES=====
expected_disease_mut_perc: 0.15625
expected_disease_nomut_perc: 0.34375
expected_healthy_mut: 0.15625
expected_healthy_nomut: 0.34375
```

Παρατηρούμε δηλαδή ότι αν ΔΕΝ σχετίζεται η μετάλλαξη με την ασθένεια, ο πίνακας θα περιμέναμε (expected..) να είναι:

ΠΛΗΘΟΣ	Μετάλλαξη=ΝΑΙ	Μετάλλαξη=ΟΧΙ
ΠΑΘΗΣΗ=ΝΑΙ	15.625%	34.375%
ΠΑΘΗΣΗ=ΟΧΙ	15.625%	34.375%

Παρατηρήστε στον πίνακα ότι είτε κάποιος έχει την μετάλλαξη είτε όχι, η πιθανότητα να έχει τη πάθηση είναι ίδια.

Τώρα όμως είμαστε σκεπτικοί...

Εμείς παρατηρήσαμε (observe) αυτόν τον πίνακα:

ΠΛΗΘΟΣ	Μετάλλαξη=ΝΑΙ	Μετάλλαξη=ΟΧΙ
ΠΑΘΗΣΗ=ΝΑΙ	18.75%	31.25%
ΠΑΘΗΣΗ=ΟΧΙ	12.5%	37.5

Ο οποίος είναι διαφορετικός από αυτόν που θα περιμέναμε αν δεν υπήρχε συσχέτιση. Πόσο διαφορετικός είναι όμως; Ένας τρόπος για να το βρούμε αυτό είναι να υπολογίσουμε το παρακάτω άθροισμα (https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test):

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

In [341]:
```
def chi_square(A,B):

    return sum(((x-y)**2)/y for x,y in zip(A,B))
```

In [343]:
```
difference_from_expected = chi_square(
    [disease_mut_perc, disease_nomut_perc, healthy_mut_perc, healthy_nomut_per
c],
    [expected_disease_mut_perc, expected_disease_nomut_perc, expected_healthy_m
ut, expected_healthy_nomut],
)
difference_from_expected
```

Out[343]:
```
0.01818181818181818
```

Ωραία. Βλέπουμε τώρα ότι αυτό που παρατηρήσαμε διαφέρει από αυτό που περιμέναμε (υπό τη προϋπόθεση της μη συσχέτσισης) κατά 0.018 . Πάλι δεν μπορούμε να καταλήξουμε αν η συσχέτιση υπάρχει ή όχι. Για να το βρούμε αυτό πρέπει να βρούμε ποια είναι η πιθανότητα το 0.018 να προήλθε κατά τύχη!

Οπότε ας φτιάξουμε έναν πίνακα με τυχαίους ανθρώπους υγιείς και ασθενείς:

In [255]:
```
health_status = [True] * disease + [False] * healthy
```

Ομοίως ας κάνουμε το ίδιο και με τις μεταλλάξεις:

In [256]:
```
mutation_status = [True] * mutation + [False] * no_mutation
```

Ας ανακατέψουμε αυτούς του πίνακες και ας τους ταιριάξαμε! Προσοχή! Εδώ κρύβεται "η καρδιά" όλης της μεθοδολογίας. Το "ταίριασμα" μεταξύ ασθένεια και μετάλλαξης γίνεται τυχαία, οπότε γνωρίζουμε από πριν ότι δεν υπάρχει κάποια συσχέτιση

In [286]:
```
random.shuffle(health_status)
random.shuffle(mutation_status)

test_people = list(zip(health_status, mutation_status))
```

Τώρα που τους ανακατέψαμε, θα υπολογίσουμε το πλήθος των στοιχείων του πίνακα:

In [287]:
```
test_dis_mut = 0
test_dis_nomut = 0
test_healthy_mut = 0
test_healthy_nomut = 0

for d,m in test_people:
    if d:
        if m:
            test_dis_mut += 1
        else:
            test_dis_nomut += 1
    else:
        if m:
            test_healthy_mut += 1
        else:
            test_healthy_nomut += 1


test_dis_mut_perc = test_dis_mut/total
test_dis_nomut_perc = test_dis_nomut/total
test_healthy_mut_perc = test_healthy_mut/total
test_healthy_nomut_perc = test_healthy_nomut/total

print ('test_dis_mut_perc:', test_dis_mut_perc)
print ('test_dis_nomut_perc:', test_dis_nomut_perc)
print ('test_healthy_mut_perc:', test_healthy_mut_perc)
print ('test_healthy_nomut_perc:', test_healthy_nomut_perc)
```

```
test_dis_mut_perc: 0.1625
test_dis_nomut_perc: 0.3375
test_healthy_mut_perc: 0.15
test_healthy_nomut_perc: 0.35
```

Ας ξαναφτιάξουμε τον πίνακα:

ΠΛΗΘΟΣ	Μετάλλαξη=ΝΑΙ	Μετάλλαξη=ΟΧΙ
ΠΑΘΗΣΗ=ΝΑΙ	16.25%	33.37%
ΠΑΘΗΣΗ=ΟΧΙ	15%	35

Παρατηρούμε ότι τα ποσοστά είναι λίγο διαφορετικά από αυτά που είχε ο πίνακας που περιμέναμε (expected). Πόσο διαφορετικά όμως;

In [344]:
```
test_difference = chi_square(
    [test_dis_mut_perc, test_dis_nomut_perc, test_healthy_mut_perc, test_health
y_nomut_perc],
    [expected_disease_mut_perc, expected_disease_nomut_perc, expected_healthy_m
ut, expected_healthy_nomut],
)
test_difference
```

Out[344]:
```
0.0007272727272727266
```

Είδαμε ότι η δική μας παρατήρηση (observed) διαφέρει από αυτό που περιμέναμε (με τη πρϋπόθεση της μη-συσχέτισης) κατα 0.018. Ενώ αν πάρουε έναν τυχαίο πίνακα (με την ίδια προϋπόθεση) αυτός διαφέρει πολύ λιγότερο (0.00072). Ας βρούμε τώρα τη πιθανότητα ο τυχαίος πίνακας να διαφέρει περισσότερο από αυτό που παρατηρήσαμε:

In [345]:
```
def f():
    random.shuffle(health_status)
    random.shuffle(mutation_status)

    test_people = list(zip(health_status, mutation_status))

    test_dis_mut = 0
    test_dis_nomut = 0
    test_healthy_mut = 0
    test_healthy_nomut = 0

    for d,m in test_people:
        if d:
            if m:
                test_dis_mut += 1
            else:
                test_dis_nomut += 1
        else:
            if m:
                test_healthy_mut += 1
            else:
                test_healthy_nomut += 1


    test_dis_mut_perc = test_dis_mut/total
    test_dis_nomut_perc = test_dis_nomut/total
    test_healthy_mut_perc = test_healthy_mut/total
    test_healthy_nomut_perc = test_healthy_nomut/total

#    test_difference = (expected_disease_mut_perc-test_dis_mut_perc)**2 + \
#              (expected_disease_nomut_perc-test_dis_nomut_perc)**2 + \
#              (expected_healthy_mut-test_healthy_mut_perc)**2 + \
#              (expected_healthy_nomut-test_healthy_nomut_perc)**2

    test_difference = chi_square(
        [test_dis_mut_perc, test_dis_nomut_perc, test_healthy_mut_perc, test_he
althy_nomut_perc],
        [expected_disease_mut_perc, expected_disease_nomut_perc, expected_healt
hy_mut, expected_healthy_nomut],
    )
    return test_difference >= difference_from_expected

STEPS = 100000
sum(1 for x in range(STEPS) if f())/STEPS
```

Out[345]: 0.02158

Η πιθανότητα αυτή είναι: 0.02158 ή 2.2%

Αυτό είναι λιγότερο από 5% οπότε μπορούμε να πούμε ότι όντως υπάρχει κάποια συσχέτιση.

Αυτό που μόλις κάναμε ονομάζεται chi-squared test (https://en.wikipedia.org/wiki/Chi-squared_test) και μπορούμε να το τρέξουμε απλά με τη παρακάτω εντολή της scipy:

In [326]:
```
obs = np.array([[disease_mut, healthy_mut], [disease_nomut, healthy_nomut]])
chi2, p, dof, expected = chi2_contingency(obs)
p
```

Out[326]: 0.021935308031646828

Βασική ορολογία:

- Όλη αυτή η διαδικασία που παρουσιάστηκε ονομάζεται statistical hypothesis testing (https://en.wikipedia.org/wiki/Statistical_hypothesis_testing)
- Όπως έχουμε πει το α=0.05 είναι το significance_threshold (ή το significance level)
- Οι δύο υποθέσεις που κάναμε:

    1. "Δεν υπάρχει μεροληψία ενάντια στις γυναίκες στο συνέδριο"
    2. "Δεν υπάρχει συσχέτιση της μετάλλαξης με την ασθένεια"

    Ονομάζονται null hypothesis (https://en.wikipedia.org/wiki/Null_hypothesis). Σκοπός της όλης διαδικασίας είναι να καταρρίψουμε (ή όχι) το null hypothesis.
- Οι τιμές που έχουμε υπολογίσει στο τέλος (0.04288 για το πρώτο και 0.02158 για το δεύτερο) ονομάζονται p-values (https://en.wikipedia.org/wiki/P-value).

p-value είναι η πιθανότητα να πάρω τα δεδομένα που πήρα, δεδομένου ότι το null hypothesis ισχύει

In [ ]: