Στατιστική των Επιχειρήσεων Ι

Διάλεξη 3η

Γιώργος Τσιρογιάννης

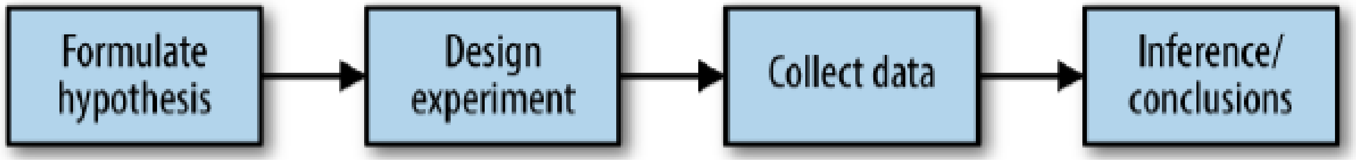
Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων,

Πανεπιστήμιο Πατρών

Ακαδημαϊκό έτος: 2020-2021

Σχεδιασμός στατιστικού πειράματος

- Στόχος: σχεδιασμός των βημάτων/ενεργειών που θα πρέπει να ακολουθηθούν ώστε να αποδεχθούμε ή να απορρίψουμε μια υπόθεση
- Γενική εικόνα της στατιστικής συμπερασματολογίας:



A/B έλεγχος (A/B testing)

- Αφορά δύο ομάδες (A, B)
- Για τις ομάδες αυτές επιθυμούμε να βρούμε ποια θεραπεία (treatment) ή προϊόν ή διαδικασία είναι καλύτερη
- Η «καθιερωμένη» θεραπεία καλείται ελέγχου (control)
- Μια τυπική υπόθεση είναι αν η νέα θεραπεία ή προϊόν ή διαδικασία είναι καλύτερη από την ελέγχου.

Ορολογία

- Θεραπεία (Treatment)
    - Το αντικείμενο μελέτης (θεραπεία (treatment) ή προϊόν ή διαδικασία) που το υποκείμενο υπόκειται
- Ομάδα θεραπείας (Treatment group)
    - Η ομάδα υποκειμένων που υπόκειται σε θεραπεία
- Ομάδα ελέγχου (Control group)
    - Η ομάδα υποκειμένων που υπόκειται στην καθιερωμένη θεραπεία (ή καθόλου θεραπεία)
- Τυχαιοποίηση (Randomization)
    - Η διαδικασία της τυχαίας ανάθεσης στις ομάδες ελέγχου ή θεραπείας
- Υποκείμενα (Subjects)
    - Τα άτομα που υπόκεινται στη θεραπεία ή αγοράζουν προϊόντα ή επισκέπτονται ιστοσελίδες
- Στατιστικό έλεγχο (Test statistic)
    - Το μέτρο σύγκρισης για επιλογή της βέλτιστης θεραπείας

Ο ρόλος της τυχαιοποίησης

- Στόχος: να ελαχιστοποιήσουμε την μεροληψία του δείγματος (sample bias)
- Κάθε διαφοροποίηση στις ομάδες οφείλεται σε:
  - Στην διαφορά των θεραπειών (effect of treatment)
  - Στον παράγοντα τύχη της κλήρωσης

Γιατί control group;

- Εύκολο να το παραλείψουμε! Γιατί δεν το κάνουμε;
- Βεβαιωνόμαστε ότι «όλες οι συνθήκες» σύγκρισης είναι ίδιες
- Και συνεπώς οι πιθανές διαφορές οφείλονται στην θεραπεία ή τον παράγοντα τύχη
- Μια σύγκριση με το baseline είναι πιθανή, αλλά όχι το ίδιο αξιόπιστη

Έλεγχος υποθέσεων (Hypothesis Tests, significance tests)

- Πολύ διαδομένα στην στατιστική
- Null hypothesis
  - Τρέχουσα κατάσταση: η υπόθεση που θέλουμε να «γκρεμίσουμε»
- Alternative hypothesis (εναλλακτική υπόθεση)
  - Η υπόθεση που «ελπίζουμε» να είναι καλύτερη
- Μονής κατεύθυνσης ή διπλής κατεύθυνσης

Γιατί έλεγχος υποθέσεων

- Τείνουμε να ερμηνεύουμε την τυχαιότητα/σύμπτωση ως πραγματικό συστηματικό μηχανισμό και αναγνωρίζουμε πρότυπα
- Γενικά προστατεύουν το ερευνητή ώστε να μην ξεγελαστεί από τα παιχνίδια της τύχης

Πείραμα


- Σκεφτείτε και γράψτε στο chat το αποτέλεσμα την ρίψης ενός νομίσματος 50 φορές.
- H: για τη κεφαλή και T: για τα γραμματα
- Πχ HHTTHHT……

Έλεγχος υποθέσεων: Λογική

- Δεδομένου του ανθρώπινου τρόπου σκέψης να αντιδρά στην «παράξενη» αλλά τυχαία συμπεριφορά με το να προσπαθεί να την ερμηνεύσει ως συστηματική, στην επιστημονική θεώρηση ένας τρόπος είναι να απαιτείται απόδειξη της διαφοράς των δύο ομάδων μεγαλύτερη από εκείνη που οφείλεται σε καθαρή τύχη.
- Η βασική υπόθεση είναι ότι αποτέλεσμα της θεραπείας και στις δύο ομάδες είναι ίδιο (null hypothesis)
- Η ελπίδα μας είναι να δείξουμε κάτω από τις ανωτέρω συνθήκες ότι δεν είναι

Alternative hypothesis (εναλλακτική υπόθεση)

- Null = "καμία διαφορά στους μέσους των ομάδων Α και Β", alternative ="ο μέσος του Α είναι διαφορετικός του Β" (μπορεί μικρότερος ή μεγαλύτερος)
- Null = "A ≤ B", alternative = "A > B"
- Null = "ο μέσος του Β δεν είναι Χ% μεγαλύτερος του Α", alternative = "ο μέσος του Β είναι Χ% μεγαλύτερος του Α"

Απλής/διπλής κατεύθυνσης

- Συχνά στο Α/Β έλεγχο υποθέσεων ελέγχουν μια νέα θεραπεία (Β) με την υπάρχουσα Α. Η (σιωπηλή) υπόθεση είναι ότι θα παραμένουμε στην Α, εκτός και αν η Β (νέα) αποδειχθεί καλύτερη. Ζητάμε από το test να μας προστατέψει από την κατεύθυνση που ίσως ο παράγοντας τύχη ευνοήσει το Β. Επειδή δεν μας ενδιαφέρει αν ευνοήσει η τύχη το Α, ονομάζουμε το test απλής κατεύθυνσης (one way)
- Αν το ζητούμενο είναι το test να μας προστατέψει και προς τις δύο κατευθύνσεις, τότε ονομάζεται διπλής κατεύθυνσης (two way)

Σχόλια

- Η null hypothesis είναι ένα λογικό κατασκεύασμα που εμπεριέχει την ιδέα: «τίποτα σημαντικό δεν συμβαίνει»
- Το test υποθέτει ότι η null hypothesis είναι αλήθεια και δημιουργεί ένα στατιστικό μοντέλο και ελέγχει αν τα δεδομένα είναι ένα λογικά αναμενόμενο αποτέλεσμα.

Στατιστικά σημαντικό αποτέλεσμα

- Όταν το αποτέλεσμα ενός πειράματος είναι πέρα του πεδίου της τύχης, τότε είναι στατιστικά σημαντικό.
- p-value
  - Δεδομένου ενός μοντέλου τύχης που εμπεριέχει την null hypothesis, η p-value είναι η πιθανότητα να λάβουμε τα ασυνήθη/ακραία αποτελέσματα που παρατηρούμε
- Alpha
  - Το κατώφλι της πιθανότητας που πρέπει να ξεπεράσουμε ώστε ο παράγοντας τύχη να είναι κύριος και το αποτέλεσμα να είναι στατιστικά σημαντικό

Τύποι σφάλαματος

- Type 1 error
  - Όταν λανθασμένα καταλήγουμε ότι μια θεραπεία είναι πραγματική (όταν οδηγούμαστε εκεί λόγω τύχης)
- Type 2 error
  - Όταν λανθασμένα μια θεραπεία είναι αποτέλεσμα τύχης (ενώ είναι πραγματική)

Παράδειγμα

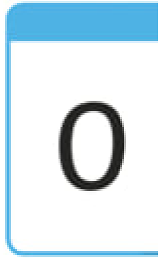
- Ecommerce
- ~46000 σημεία/συναλλαγές συγκεντρωτικά

Outcome	Price A	Price B
Conversion	200	182
No conversion	23,539	22,406

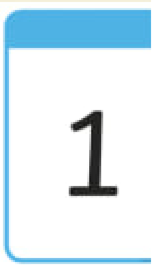
- Μετατροπή (conversion) τιμής A: 0.8425% = 200/(23539+200)*100
- Μετατροπή τιμής B: 0.8057% = 182/(22406+182)*100
- Είναι η διαφορά 0.0368% στατιστικά σημαντική;

Παράδειγμα

- Βήμα 1: βάζουμε κάρτες με 1 και 0 σε μια κάλπη:
  - Με το τρόπο αυτό θέλουμε να αναπαραστήσουμε το κοινό conversion rate των 382 (1) και των 45945 (0): 0.8246%



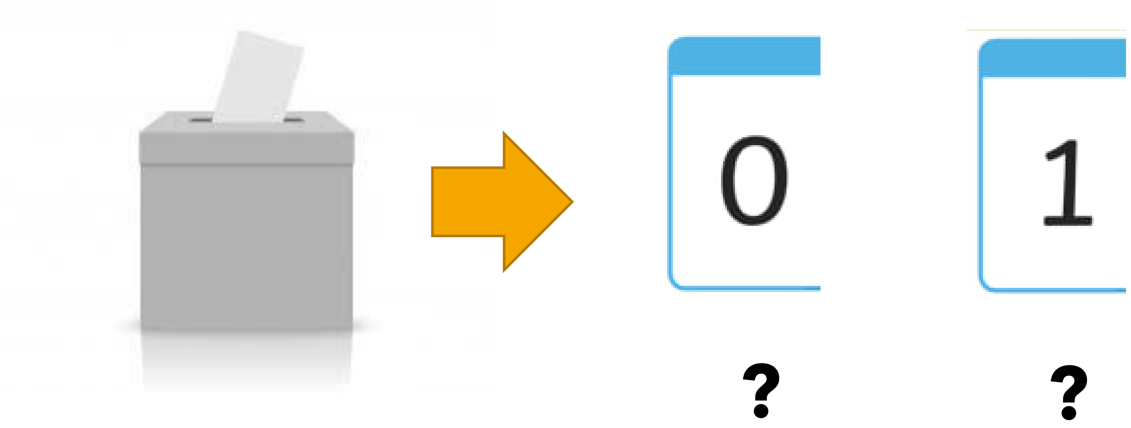
Μη-μετατροπή



μετατροπή

Παράδειγμα

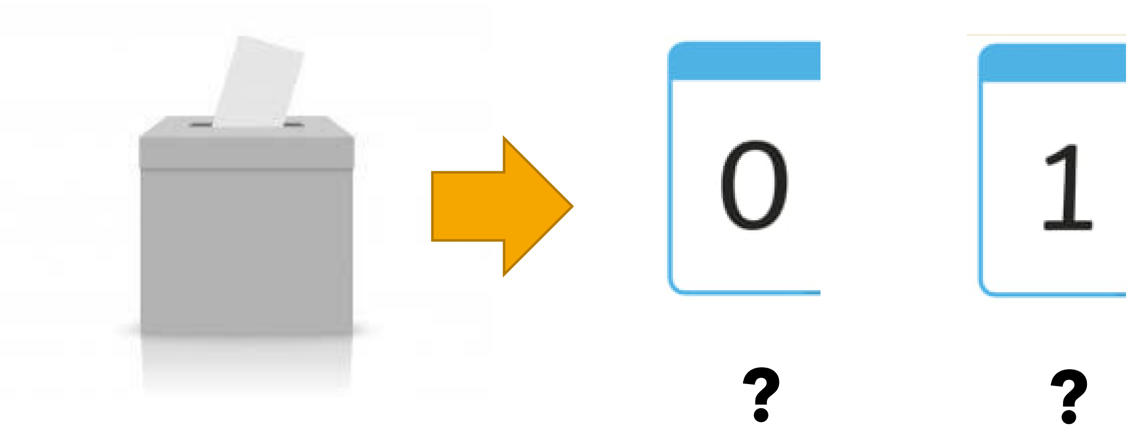
- Βήμα 2: Ανακατεύουμε και επιλέγουμε τυχαία δείγμα 23739 καρτών, και μετράμε των αριθμό των (1)



Outcome	Price A	Price B
Conversion	200	182
No conversion	23,539	22,406

Παράδειγμα

- Βήμα 3: Ανακατεύουμε και επιλέγουμε τυχαία δείγμα 22588 καρτών, και μετράμε των αριθμό των (1)



Outcome	Price A	Price B
Conversion	200	182
No conversion	23,539	22,406

Παράδειγμα

- Βήμα 4: Καταγράφουμε τις διαφορές των δύο δειγμάτων

| αριθμός άσσων βήματος 2 | - | αριθμός άσσων βήματος 3 |

Παράδειγμα

- Βήμα 5: επαναλαμβάνουμε τα βήματα 2 έως 4

Παράδειγμα

- Βήμα 6: πόσο συχνά εμφανίζεται η διαφορά 0.0368%

Παράδειγμα

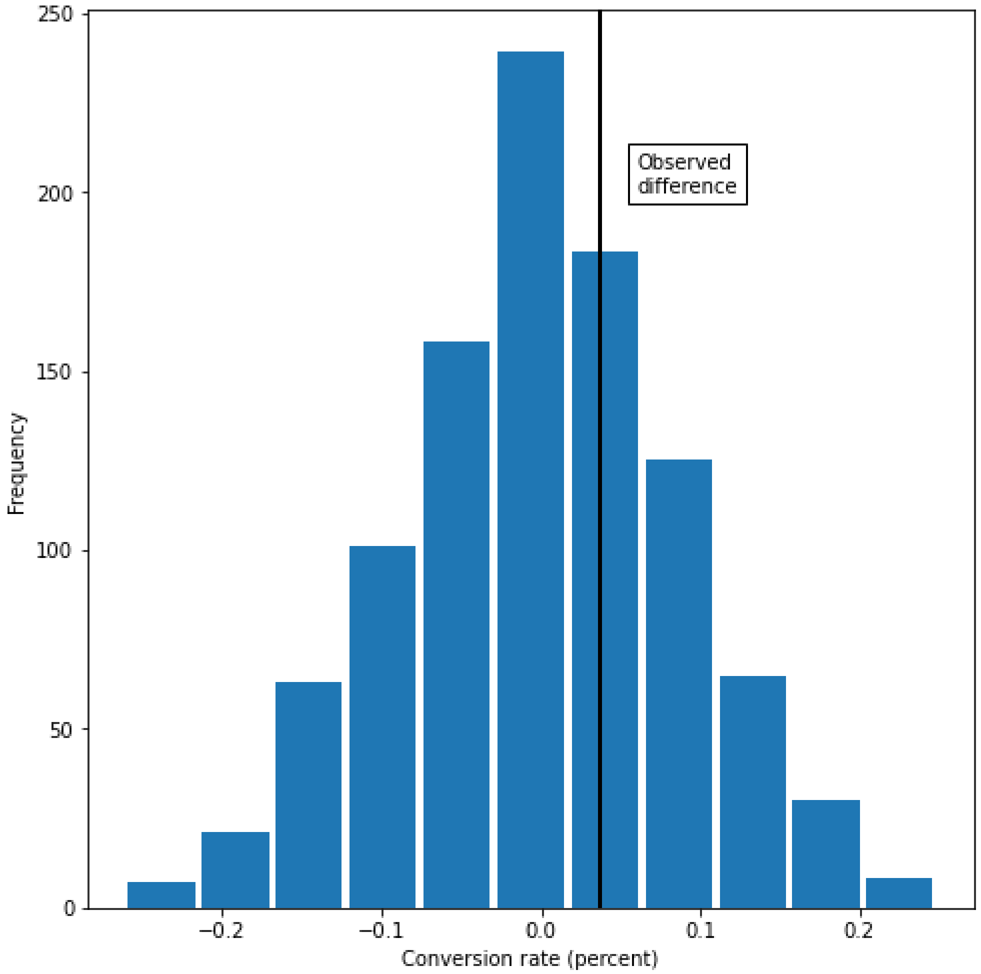
Παράδειγμα Ecommerce

[2]:
```
%matplotlib inline

import random

import pandas as pd
import numpy as np

from scipy import stats
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.stats import power

import matplotlib.pylab as plt
```

Παράδειγμα

[6]:
```
def perm_fun(x, nA, nB):
    n = nA + nB
    idx_B = set(random.sample(range(n), nB))
    idx_A = set(range(n)) - idx_B
    return x.loc[idx_B].mean() - x.loc[idx_A].mean()

random.seed(1)
obs_pct_diff = 100 * (200 / 23739 - 182 / 22588)
print(f'Observed difference: {obs_pct_diff:.4f}%')
conversion = [0] * 45945
conversion.extend([1] * 382)
conversion = pd.Series(conversion)

perm_diffs = [100 * perm_fun(conversion, 23739, 22588)
              for _ in range(1000)]

fig, ax = plt.subplots(figsize=(7, 7))
ax.hist(perm_diffs, bins=11, rwidth=0.9)
ax.axvline(x=obs_pct_diff, color='black', lw=2)
ax.text(0.06, 200, 'Observed\ndifference', bbox={'facecolor':'white'})
ax.set_xlabel('Conversion rate (percent)')
ax.set_ylabel('Frequency')

plt.tight_layout()
plt.show()
```

Παράδειγμα



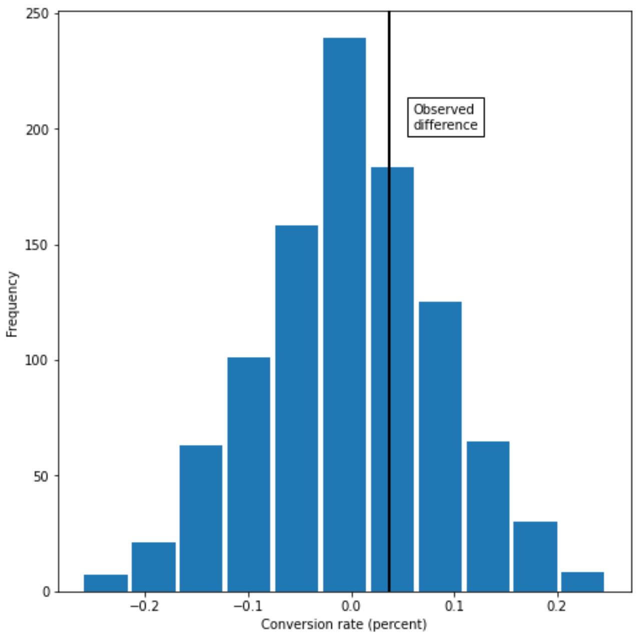
Η διαφορά του 0.0368% είναι εντός της περιοχής της καθαρής τύχης!!!
Αρά μη στατιστικά σημαντική

Είναι το διάγραμμα συχνότητας του μοντέλου τύχης της null hypothesis.
Δηλαδή η συχνότητα να δούμε «ακραία δεδομένα

Παράδειγμα

```
survivors = np.array([[200, 23739 - 200], [182, 22588 - 182]])
chi2, p_value, df, _ = stats.chi2_contingency(survivors)

print(f'p-value for single sided test: {p_value / 2:.4f}')
```

```
p-value for single sided test: 0.3498
```




```
print(np.mean([diff > obs_pct_diff for diff in perm_diffs]))
```

```
0.332
```

Τυπικές τιμές του Alpha

- Εξαρτάται από το πεδίο εφαρμογής:
  - 1%
  - 5%
- Η επιλογή είναι υποκειμενική
- Σχόλιο: η p-value δεν απαντά στο ερώτημα «Ποια είναι η πιθανότητα να είναι αποτέλεσμα τύχης», αλλά το «Δεδομένου ενός μοντέλου τύχης, ποια η πιθανότητα να είναι τόσο ακραία»

Σχόλια


https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.X-DvZNgzaUk
The online home for the publications of the American Statistical Association
The American Statistician
Volume 70, 2016 - Issue 2
Submit an article
Journal homepage
Enter keywords, authors, DOI etc.
This Journal
459,472
Views
1,970
CrossRef citations to date
2,187
Altmetric
Listen
Free access
Editorial
The ASA Statement on p-Values: Context, Process, and Purpose
Ronald L. Wasserstein & Nicole A. Lazar
Pages 129-133 | Accepted author version posted online: 07 Mar 2016, Published online:09 Jun 2016
Download citation
https://doi.org/10.1080/00031305.2016.1154108
Check for updates
Full Article
Figures & data
References
Supplemental
Citations
Metrics
Reprints & Permissions
PDF

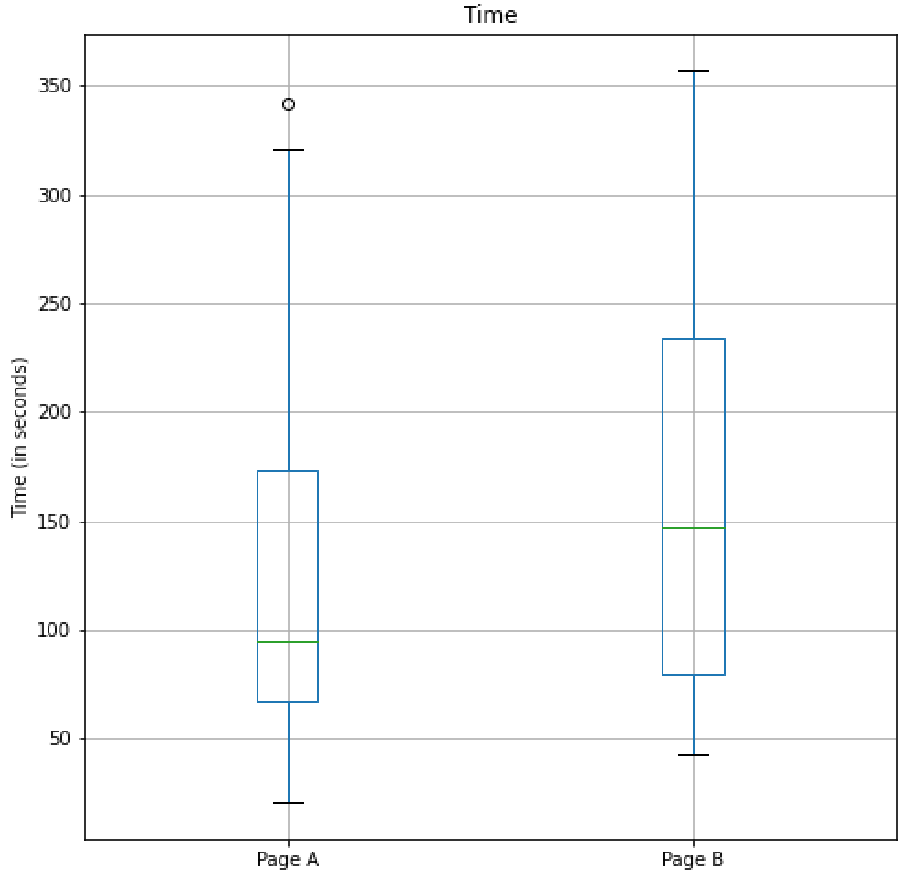
Σχόλια

https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.X-DvZNgzaUk

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Παράδειγμα

- Συγκρίνουμε τους χρόνους που ο χρήστης παραμένει σε δύο ιστοσελίδες A και B
- Υπάρχει στατιστικά σημαντική διαφορά



Παράδειγμα

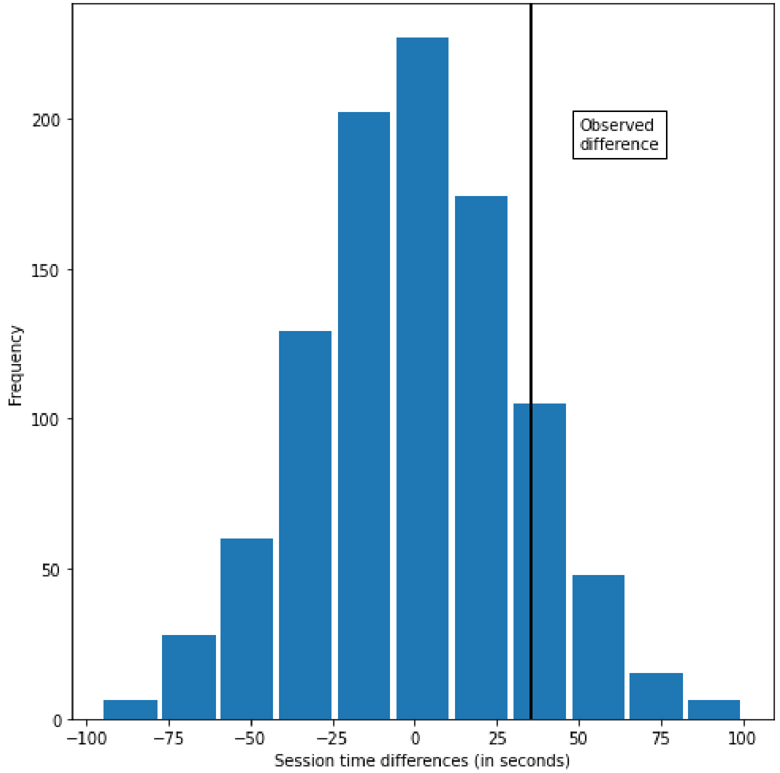
- Προσέγγιση μέσω της τυχαίας δειγματοληψίας

```
mean_a = session_times[session_times.Page == 'Page A'].Time.mean()
mean_b = session_times[session_times.Page == 'Page B'].Time.mean()
print(mean_b - mean_a)
```

```
nA = session_times[session_times.Page == 'Page A'].shape[0]
nB = session_times[session_times.Page == 'Page B'].shape[0]
random.seed(1)
perm_diffs = [perm_fun(session_times.Time, nA, nB) for _ in range(1000)]

fig, ax = plt.subplots(figsize=(7, 7))
ax.hist(perm_diffs, bins=11, rwidth=0.9)
ax.axvline(x = mean_b - mean_a, color='black', lw=2)
ax.text(50, 190, 'Observed\ndifference', bbox={'facecolor':'white'})
ax.set_xlabel('Session time differences (in seconds)')
ax.set_ylabel('Frequency')

plt.tight_layout()
plt.show()
```

Παράδειγμα

- Προσέγγιση μέσω της τυχαίας δειγματοληψίας



Η διαφορά του 35.66 είναι εντός της περιοχής της καθαρής τύχης!!!
Αρά μη στατιστικά σημαντική

Είναι το διάγραμμα συχνότητας του μοντέλου τύχης της null hypothesis.
Δηλαδή η συχνότητα να δούμε «ακραία δεδομένα

Παράδειγμα

- P-values

t-test 1

[28]:
```
res = stats.ttest_ind(session_times[session_times.Page == 'Page A'].Time,
                      session_times[session_times.Page == 'Page B'].Time,
                      equal_var=False)
print(f'p-value for single sided test: {res.pvalue / 2:.4f}')
```

```
p-value for single sided test: 0.1408
```

```
print(np.mean(perm_diffs > mean_b - mean_a))
```

```
0.121
```

t-test 2

[29]:
```
tstat, pvalue, df = sm.stats.ttest_ind(
    session_times[session_times.Page == 'Page A'].Time,
    session_times[session_times.Page == 'Page B'].Time,
    usevar='unequal', alternative='smaller')
print(f'p-value: {pvalue:.4f}')
```

```
p-value: 0.1408
```

Πολύ κοντά στην τιμή του πειράματος τυχαιότητας

Η διαφορά δεν είναι στατιστικά σημαντική και δεν μπορούμε να απορρίψουμε το null hypotheis

Torture the data long enough, and it will confess.

- Αν πάρουμε αρκετές (>20) τυχαίες μετρήσεις (πιθανούς predictors) ενός στόχου (target), με υψηλή πιθανότητα θα υπάρχει στατιστικά σημαντική συσχέτιση με κάποιον για alpha=0.05
- Αυτό σημαίνει πως ασχολούμαστε με τον «ταιριάξουμε» (fit) θόρυβο στον στόχο μας
- Το πρόβλημα μεγαλώνει όταν κάνουμε «ζευγαρωτές» συγκρίσεις πολλών μεταβλητών (πχ. ΑΒ, ΑΓ, ΓΔ, ΑΔ….)
- Πρακτική λύση: διαίρεση του alpha με το πλήθος των tests (Bonferroni adjustment)

Ανάλυση της διακύμανσης (ANOVA)

- Αφορά σε πολλαπλές ομάδες πχ A/B/C/D
- Αφορά σε αριθμητικά δεδομένα
- Ορολογία:
  - Σύγκριση ανά ζεύγη (Pairwise comparison)
    - Έλεγχος υποθέσεων (πχ των μέσων) μεταξύ δύο ομάδων από πολλαπλές ομάδες
  - Συλλογικός έλεγχος (Omnibus test)
    - Μοναδικός έλεγχος υποθέσεων της συλλογικής διακύμανσης μεταξύ πολλαπλών μέσων των ομάδων
  - F-statistic
    - Τυποποιημένο στατιστικό που μετρά τον βαθμό στον οποίο οι διαφορές μεταξύ των μέσων (των ομάδων) ξεπερνάει την αναμενόμενη τιμή του μοντέλου τύχης

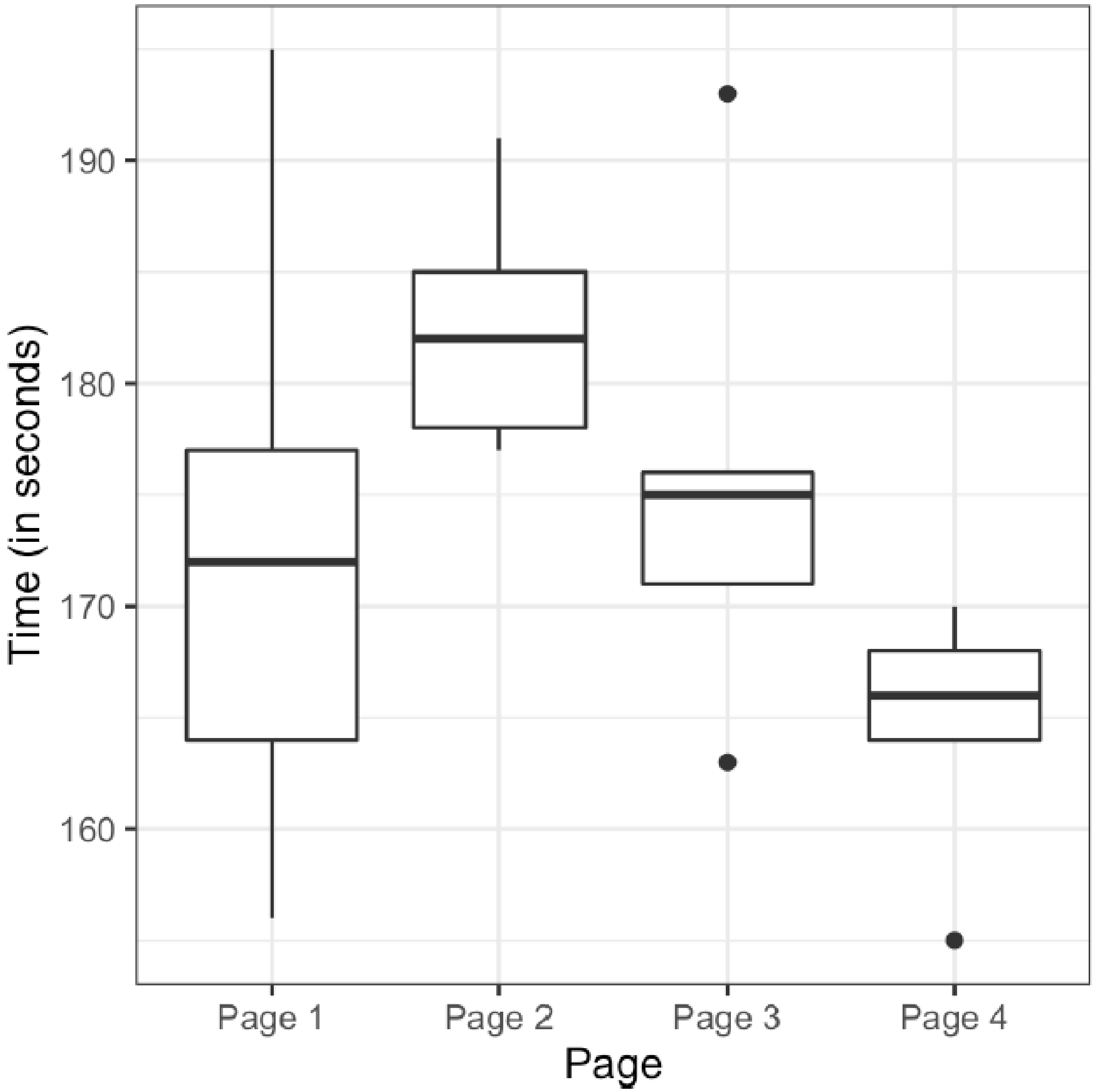
Παράδειγμα

- 4 διαφορετικές σελίδες (web)
- Ύστερα από τυχαιοποίηση της σειράς μετράμε τους χρόνους παραμονής 5 ανθρώπων

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

Παράδειγμα

Ποια σελίδα επιτυγχάνει την μεγαλύτερη παραμονή κατά μέση τιμη;



Παράδειγμα

- 6 συνδυασμοί μέσων
  - 1-2
  - 1-3
  - 1-4
  - 2-3
  - 2-4
  - 3-4

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

Όσο περισσότερα είναι τα πιθανά ζεύγη, δυνητικά τόσο μεγαλύτερη η πιθανότητα τυχαίας υπεροχής

Παράδειγμα

- Υπάρχει η περίπτωση όλες οι σελίδες να έχουν τον ίδιο χρόνο αναμονής για τον πληθυσμό των χρηστών, και οι διαφοροποιήσεις οφείλονται στην τυχαιότητα;

ANOVA

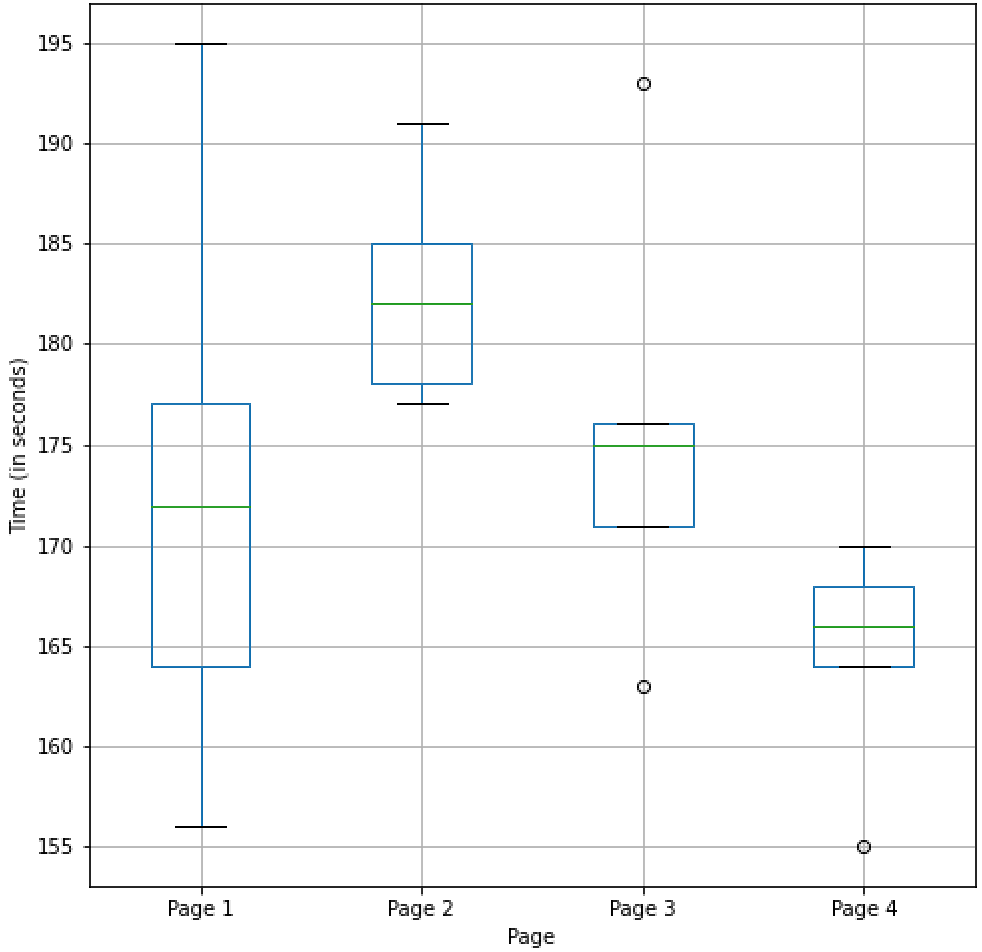
Παράδειγμα

- Προσέγγιση μέσω επαναδειγματοληψίας (resampling)
- 1. Συνδυασμός όλων των δεδομένων σε ένα dataset
- 2. Ανακατανέμουμε την σειρά των σημείων και επιλέγουμε 4 δείγματα με 5 σημεία (χωρίς επανάθεση).
- 3. Υπολογίζουμε τον μέσο της κάθε ομάδας/δείγματος
- 4. Υπολογίζουμε την διακύμανση των 4 ομάδων/δειγμάτων
- 5. Επαναλαμβάνουμε τα βήματα 2-4 πολλές φορές (πχ 1000)

Παράδειγμα

Boxplot

[4]:
```
four_sessions = pd.read_csv('../data/four_sessions.csv')

ax = four_sessions.boxplot(by='Page', column='Time',
                           figsize=(7, 7))
ax.set_xlabel('Page')
ax.set_ylabel('Time (in seconds)')
plt.suptitle('')
plt.title('')

plt.tight_layout()
plt.show()
```

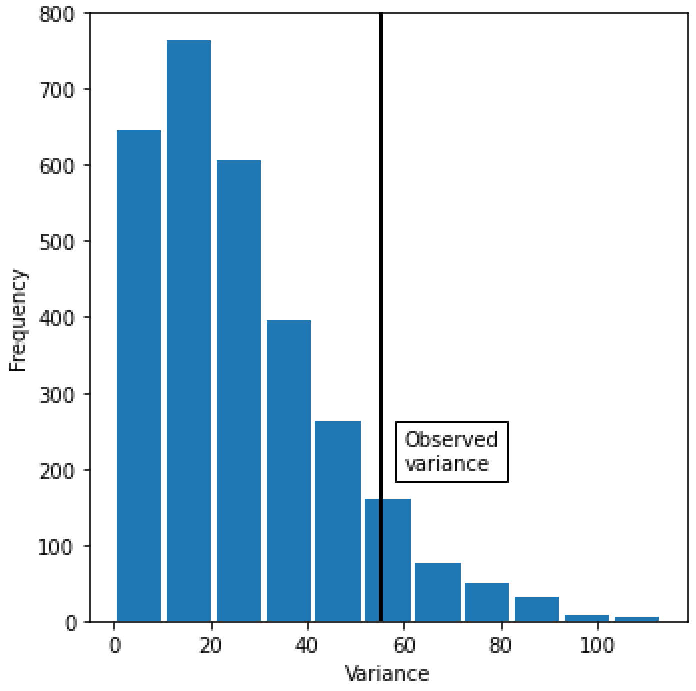


Παράδειγμα

```
observed_variance = four_sessions.groupby('Page').mean().var()[0]
print('Observed means:', four_sessions.groupby('Page').mean().values.ravel())
print('Variance:', observed_variance)
# Permutation test example with stickiness
def perm_test(df):
    df = df.copy()
    df['Time'] = np.random.permutation(df['Time'].values)
    return df.groupby('Page').mean().var()[0]
```

```
Observed means: [172.8 182.6 175.6 164.6]
Variance: 55.426666666666655
```

```
[6]: random.seed(1)
perm_variance = [perm_test(four_sessions) for _ in range(3000)]
print('Pr(Prob)', np.mean([var > observed_variance for var in perm_variance]))

fig, ax = plt.subplots(figsize=(5, 5))
ax.hist(perm_variance, bins=11, rwidth=0.9)
ax.axvline(x = observed_variance, color='black', lw=2)
ax.text(60, 200, 'Observed\nvariance', bbox={'facecolor':'white'})
ax.set_xlabel('Variance')
ax.set_ylabel('Frequency')

plt.tight_layout()
plt.show()
```



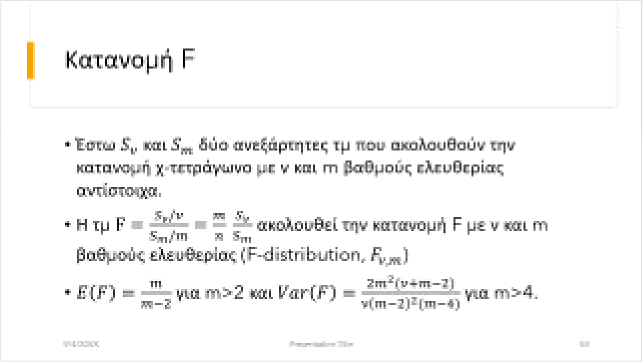

```
Pr(Prob) 0.08733333333333333
```

Πιθανότητα η διακύμανση 55.4 να είναι προϊόν τύχης

Παράδειγμα

- Υπολογισμός μέσω του F-statistic και της F-distribution
- Το F-statistic βασίζεται στο πηλίκο της διακύμανσης των μέσων διηρημένο με την διακύμανση που οφείλεται στο πλεονάζον σφάλμα.
- Όσο υψηλότερο, τόσο πιο στατιστικά σημαντικό το αποτέλεσμα.
- Όταν τα δεδομένα ακολουθούν την κανονική κατανομή, τότε το F-statistic ακολουθεί την F-distribution και η p-value μπορεί να υπολογισθεί

Παράδειγμα


Κατανομή F

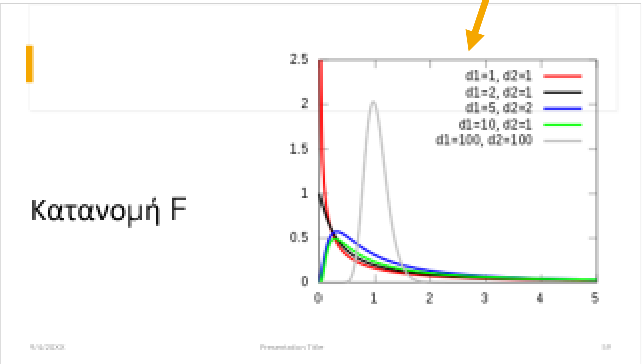
- Έστω S_v και S_m δύο ανεξάρτητες τμ που ακολουθούν την κατανομή χ-τετράγωνα με v και m βαθμούς ελευθερίας αντίστοιχα.
- Η τμ $F = \frac{S_v/v}{S_m/m} = \frac{m}{n}\frac{S_v}{S_m}$ ακολουθεί την κατανομή F με v και m βαθμούς ελευθερίας (F-distribution, $F_{v,m}$)
- $E(F) = \frac{m}{m-2}$ για m>2 και $Var(F) = \frac{2m^2(v+m-2)}{v(m-2)^2(m-4)}$ για m>4.

F-statistic

```
model = smf.ols('Time ~ Page', data=four_sessions).fit()

aov_table = sm.stats.anova_lm(model)
print(aov_table)
```

	df	sum_sq	mean_sq	F	PR(>F)
Page	3.0	831.4	277.133333	2.739825	0.077586
Residual	16.0	1618.4	101.150000	NaN	NaN

Αν alpha=0.05 καταλήγουμε στο ίδιο στατιστικό συμπέρασμα με την επαναδειγματοληψία



F-statistic

p-value

Παράδειγμα

- Εξετάζουμε την αποτελεσματικότητά 3 διαφορετικών headlines (A,B και C)
- Αξιολογούνται από 1000 χρήστες

	Headline A	Headline B	Headline C
Click	14	8	12
No-click	986	992	988

Παράδειγμα

- Null hypothesis: και οι 3 headlines έχουν το ίδιο click-rate

	Headline A	Headline B	Headline C
Click	11.33	11.33	11.33
No-click	988.67	988.67	988.67

chi-square statistic

Pearson residual

$$R = \frac{\text{Observed}-\text{Expected}}{\sqrt{\text{Expected}}}$$

	Headline A	Headline B	Headline C
Click	0.792	−0.990	0.198
No-click	−0.085	0.106	−0.021

$$X = \sum_{i}^{r} \sum_{j}^{c} R^2$$

1.666

Είναι η τιμή αυτή μεγαλύτερη εκείνης που οφείλεται στην τυχαιότητα;

Παράδειγμα

Δεδομένα

```
click_rate = pd.read_csv('../data/click_rates.csv')
clicks = click_rate.pivot(index='Click', columns='Headline', values='Rate')
row_average = clicks.mean(axis=1)
pd.DataFrame({
    'Headline A': row_average,
    'Headline B': row_average,
    'Headline C': row_average,
})
```

	Headline A	Headline B	Headline C
Click			
Click	11.333333	11.333333	11.333333
No-click	988.666667	988.666667	988.666667

Παράδειγμα

Επαναδειγματοληψία (resampling)

```
box = [1] * 34
box.extend([0] * 2966)
random.shuffle(box)

def chi2(observed, expected):
    pearson_residuals = []
    for row, expect in zip(observed, expected):
        pearson_residuals.append([(observe - expect) ** 2 / expect
                                  for observe in row])
    # return sum of squares
    return np.sum(pearson_residuals)

expected_clicks = 34 / 3
expected_noclicks = 1000 - expected_clicks
expected = [34 / 3, 1000 - 34 / 3]
chi2observed = chi2(clicks.values, expected)
```

```
expected_clicks = 34 / 3
expected_noclicks = 1000 - expected_clicks
expected = [34 / 3, 1000 - 34 / 3]
chi2observed = chi2(clicks.values, expected)

def perm_fun(box):
    sample_clicks = [sum(random.sample(box, 1000)),
                     sum(random.sample(box, 1000)),
                     sum(random.sample(box, 1000))]
    sample_noclicks = [1000 - n for n in sample_clicks]
    return chi2([sample_clicks, sample_noclicks], expected)

perm_chi2 = [perm_fun(box) for _ in range(2000)]

resampled_p_value = sum(perm_chi2 > chi2observed) / len(perm_chi2)
print(f'Observed chi2: {chi2observed:.4f}')
print(f'Resampled p-value: {resampled_p_value:.4f}')
```

```
Observed chi2: 1.6659
Resampled p-value: 0.4685
```

Υπολογίσαμε υψηλή πιθανότητα οι διαφορές να οφείλονται στον Παράγοντα τύχη

Παράδειγμα

Chi^2 test

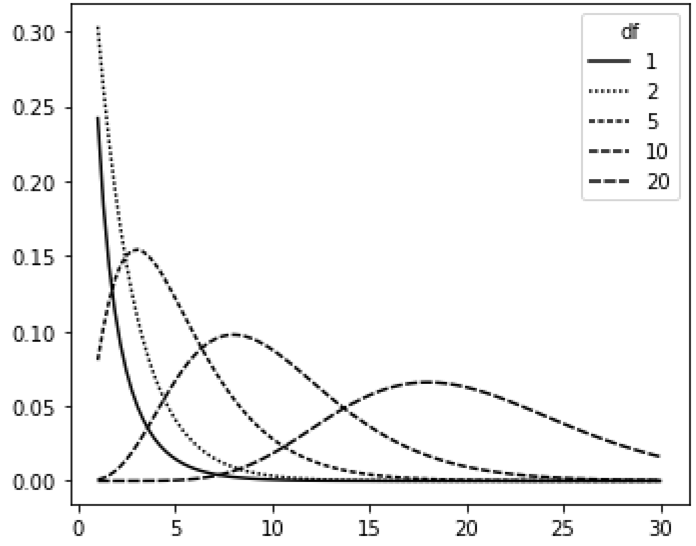
```
chisq, pvalue, df, expected = stats.chi2_contingency(clicks)
print(f'Observed chi2: {chi2observed:.4f}')
print(f'p-value: {pvalue:.4f}')
```

```
Observed chi2: 1.6659
p-value: 0.4348
```

Διαφορά οφείλεται στην προσέγγιση μέσω της κατανομής χ^2

Resampling

```
Observed chi2: 1.6659
Resampled p-value: 0.4685
```



Απαλλακτική εργασία

- Η άσκηση είναι ατομική
- Θα προετοιμάσετε μια παρουσίαση στο θέμα που περιγράφεται στις ακόλουθες διαφάνειες
- Η παρουσίαση σας θα γίνει σε όλη την ομάδα του MBA
- Χρόνος παρουσίασης 10 λεπτά
- Η σειρά παρουσίασης είναι τυχαία και θα ανακοινώνεται μόλις ολοκληρωθεί η προηγουμένη παρουσίαση

Απαλλακτική εργασία

- Πιθανότερη ημερομηνία 10/02/2021 στις 17:00 έως ~22:00
- Μπορείτε να χρησιμοποιήσετε όποιο πρόγραμμα παρουσιάσεων σας βολεύει (πχ powerpoint) και θα πραγματοποιείται με το να κάνετε διαμοίραση της οθόνη σας στο zoom

Απαλλακτική εργασία

- Στόχος:
  - Να μάθατε ένα νέο και βασικό θέμα της στατιστικής επιστήμης
  - Να συνηθίσετε την αυτόνομη αναζήτηση και μελέτη νέων αντικειμένων και εργαλείων χρήσιμων για την δουλειά σας
  - Εξοικείωση στις παρουσιάσεις σε μεγάλα ακροατήρια σε θέματα που οι συμμετέχοντες έχουν παρόμοια γνώση με εσάς

Απαλλακτική εργασία

- Σημεία προσοχής:
  - Προσπαθήστε να συνδυάστε υλικό από περισσότερες πηγές
  - Στις διαφάνειές σας να καλύψετε μεγάλο εύρος του θέματος: πχ θεωρητικό υπόβαθρο, εφαρμογές, παράδειγμα, …
  - Κάντε πρόβα για τον χρόνο που θα σας πάρει (την ημέρα της παρουσίασης θα είναι αυστηρά 10 λεπτά το πολύ)
  - Μην ξεχάσετε να αναφέρετε την βιβλιογραφία και τις άλλες πηγές σας

Απαλλακτική εργασία

- Παραδοτέα:
  - Οι διαφάνειες της παρουσίασης και τυχόν άλλο υλικό (πχ notebook)
  - Θα πρέπει να αποστείλετε με email τα αρχεία αυτά μέχρι και την προηγουμένη μέρα της παρουσίασης (στο:gtsirogianni@upatras.gr με θέμα: Παρουσίαση μεταπτυχιακής εργασίας)

Απαλλακτική εργασία

Θέμα:

Ανάλυση σε κύριες συνιστώσες

(Principal Component Analysis)

Απαλλακτική εργασία

- Ενδεικτική βιβλιογραφία (είναι διαθέσιμη να την κατεβάσετε δωρεάν):

  Αγγλική
  - https://web.stanford.edu/~hastie/Papers/ESLII.pdf, ενότητα 14.5
  - https://statlearning.com/ISLR%20Seventh%20Printing.pdf, ενότητα 10.2
  - http://www.dsc.ufcg.edu.br/~hmg/disciplinas/posgraduacao/rn-copin-2014.3/material/SignalProcPCA.pdf
  - https://arxiv.org/pdf/1404.1100.pdf

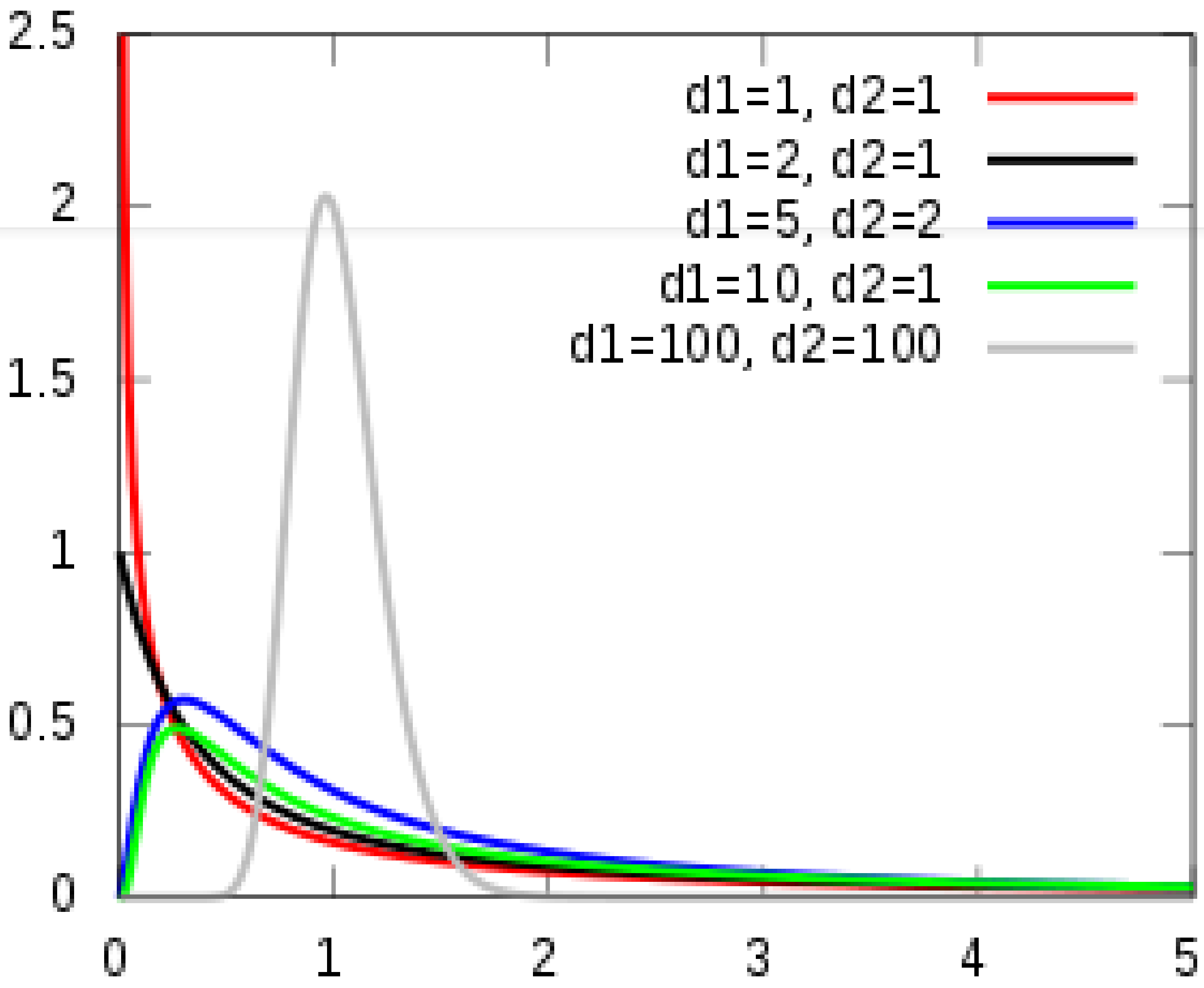
  Ελληνική
  - http://www.mas.ucy.ac.cy/~fokianos/GreekRbook/pca&discriminant.pdf
  - https://repository.kallipos.gr/bitstream/11419/2129/1/05_chapter04.pdf
  - https://gsfakianaki.github.io/docs/bachelor-thesis-sfakianaki.pdf
- Video: https://www.youtube.com/watch?v=FgakZw6K1QQ

Backup

Κατανομή F

- Έστω S_ν και S_m δύο ανεξάρτητες τμ που ακολουθούν την κατανομή χ-τετράγωνο με ν και m βαθμούς ελευθερίας αντίστοιχα.
- Η τμ $\mathrm{F} = \frac{S_\nu/\nu}{S_m/m} = \frac{m}{n}\frac{S_\nu}{S_m}$ ακολουθεί την κατανομή F με ν και m βαθμούς ελευθερίας (F-distribution, $F_{\nu,m}$)
- $E(F) = \frac{m}{m-2}$ για m>2 και $Var(F) = \frac{2m^2(\nu+m-2)}{\nu(m-2)^2(m-4)}$ για m>4.

Κατανομή F



2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=10, d2=1
d1=100, d2=100