Πιθανότητες και Στατιστική

Έλεγχος Υποθέσεων

Κώστας Μανές

Τμήμα Πληροφορικής, Πανεπιστήμιο Πειραιώς

2021-2022

Έλεγχος Υποθέσεων

Ένα διάστημα εμπιστοσύνης που κατασκευάσαμε με βάση τα προηγούμενα, μπορεί να περιέχει την πραγματική τιμή μιας άγνωστης παραμέτρου με μεγάλη πιθανότητα. Όμως, πολύ συχνά θέλουμε να πάρουμε μια απόφαση, δηλαδή να απαντήσουμε με ναι ή όχι σε μια ερώτηση που εμπεριέχει τυχαιότητα, εξασφαλίζοντας ταυτόχρονα ότι η απόφασή μας είναι η σωστή με μεγάλη πιθανότητα.
Τα στατιστικά εργαλεία με τα οποία κατασκευάσαμε διαστήματα εμπιστοσύνης, μπορούν να χρησιμοποιηθούν και για τον σκοπό αυτό με παρόμοιο τρόπο.
Η διαφορά είναι ότι τώρα δεν προσπαθούμε να εκτιμήσουμε την άγνωστη παράμετρο, αλλά να ελέγξουμε πόσο στατιστικά ισχυρή είναι η υπόθεση στην οποία βασίζεται η απόφασή μας.

Έλεγχος Υποθέσεων

Για παράδειγμα, ας υποθέσουμε ότι μια εταιρεία που κατασκευάζει μπαταρίες, ισχυρίζεται ότι έχουν μέση διάρκεια ζωής (τουλάχιστον) 240 ώρες. Για απλότητα, θεωρούμε ότι η διάρκεια ζωής είναι μια ΤΜ $X \sim N(\mu, \sigma^2)$. Έχουμε ένα τυχαίο δείγμα μεγέθους $n = 20$, για το οποίο ο δειγματικός μέσος $\overline{X}$ είναι ίσος με $\overline{x} = 220$. Πόσο σίγουροι είμαστε για τον ισχυρισμό της εταιρείας; Συμβολίζουμε τον ισχυρισμό αυτό ως

$$H_0 : \mu \geq \mu_0,$$

όπου $\mu_0 = 240$. Ο ισχυρισμός αυτός ονομάζεται **υπόθεση 0** ή **μηδενική υπόθεση** (null hypothesis). Ο αντίθετος ισχυρισμός ονομάζεται **υπόθεση 1** ή **εναλλακτική υπόθεση** (alternative hypothesis) και συμβολίζεται με H_1. Εδώ είναι

$$H_1 : \mu < \mu_0.$$

Προκειμένου να απορρίψουμε την H_0, απαιτούμε η πιθανότητα $p = P(\overline{X} \leq \overline{x}|H_0)$ να παρατηρηθεί μια τιμή της $\overline{X}$ τόσο ‘‘μακριά’’ από την υπόθεση H_0, δεδομένου ότι η H_0 είναι σωστή, να είναι πολύ μικρή, μικρότερη από κάποια τιμή $a \in (0, 1)$, η οποία ονομάζεται **επίπεδο σημαντικότητας** του ελέγχου. Αν συμβεί το αντίθετο, τότε αυτό θεωρείται ότι δεν αποτελεί επαρκή ένδειξη για την απόρριψη της H_0, οπότε την αποδεχόμαστε, χωρίς όμως αυτό να σημαίνει ότι είναι σωστή.

Έλεγχος Υποθέσεων

Η αναλογία με τα διαστήματα εμπιστοσύνης είναι άμεση: Ένα (αριστερόπλευρο) $(1-a)100\%$ δ.ε. για την μ είναι το $(-\infty, \overline{X} + z_a\sigma/\sqrt{n}]$. Πράγματι, είναι

$$P(\mu \in (-\infty, \overline{X} + z_a\sigma/\sqrt{n}]) = P(\overline{X} \geq \mu - z_a\sigma/\sqrt{n}) = P(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \geq -z_a)$$
$$= 1 - \Phi(-z_a) = 1 - a.$$

Επομένως, αν

$$\mu_0 \notin (-\infty, \overline{X}+z_a\sigma/\sqrt{n}] \Leftrightarrow \mu_0 > \overline{X}+z_a\sigma/\sqrt{n} \Leftrightarrow \overline{X} < \mu_0 - z_a\frac{\sigma}{\sqrt{n}} \Leftrightarrow Z < -z_a,$$

όπου $Z = \dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ η στατιστική συνάρτηση του ελέγχου, τότε απορρίπτουμε την H_0, με εμπιστοσύνη $(1-a)100\%$. Η τιμή $x_c = \mu_0 - z_a\dfrac{\sigma}{\sqrt{n}}$ ονομάζεται **κρίσιμη τιμή** του ελέγχου.

Προφανώς, η απόρριψη εξαρτάται τόσο από την τιμή της στατιστικής συνάρτησης Z, όσο και από την τιμή του a. Καθώς μειώνεται το a, μειώνεται και το $-z_a$, οπότε γίνεται πιο πιθανό να ισχύει $Z \geq -z_a$, οπότε η H_0 δεν απορρίπτεται.
Η μέγιστη τιμή του a για την οποία δεν απορρίπτεται η H_0 ονομάζεται **p-τιμή (p-value)** του ελέγχου. Στο συγκεκριμένο παράδειγμα, αυτό συμβαίνει όταν

$$\overline{X} = x_c \Leftrightarrow Z = -z_a \Leftrightarrow \Phi(Z) = a,$$

οπότε είναι p-value $= \Phi(Z)$. Σημειώνεται ότι η p-value είναι μια ΤΜ, αφού εξαρτάται από την Z. Μάλιστα, αποτελεί ένα άνω φράγμα της πιθανότητας p, αφού είναι

$$p = P(\overline{X} \leq \overline{x}|H_0) \overset{*}{\leq} P(\overline{X} \leq \overline{x}|\mu = \mu_0) = P(Z \leq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}|\mu = \mu_0)$$

$$\overset{**}{=} \Phi(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}) = \text{p-value}.$$

Η ισότητα (**) ισχύει διότι όταν είναι $\mu = \mu_0$, τότε $Z \sim N(0,1)$. Η ανισότητα (*) δείχνει ότι η p μεγιστοποιείται όταν $\mu = \mu_0$.
Επομένως, κρατώντας συντηρητική στάση, μπορούμε να θεωρούμε ότι σε κάθε περίπτωση η μηδενική υπόθεση είναι η $H_0 : \mu = \mu_0$ και ότι αυτή που αλλάζει είναι η H_1. Το πλεονέκτημα είναι ότι η $H_0 : \mu = \mu_0$ καθορίζει πλήρως την κατανομή της στατιστικής συνάρτησης και για το λόγο αυτό ονομάζεται απλή (αλλιώς ονομάζεται σύνθετη). Στη συγκεκριμένη περίπτωση, υπό την απλή υπόθεση H_0, είναι $Z \sim N(0,1)$. Για τον λόγο αυτόν, στα επόμενα θα θεωρούμε ότι η H_0 είναι απλή (ισότητα) και θα διακρίνουμε περιπτώσεις ανάλογα με την H_1.

Τελικά, απορρίπτουμε την H_0, όταν

$$\text{p-value} < a \Leftrightarrow \Phi(Z) < a \Leftrightarrow Z < -z_a \Leftrightarrow \overline{X} < \mu_0 - z_a \frac{\sigma}{\sqrt{n}},$$

και την αποδεχόμαστε όταν p-value $\geq a$.
Το σύνολο $R = (-\infty, -z_a) \subseteq S_Z$ ονομάζεται **χωρίο απόρριψης** (ή κρίσιμη περιοχή), και συμβολίζεται για συντομία ως $R = \{Z < -z_a\}$.
Επειδή δεν είναι κάτω φραγμένο, ο έλεγχος αυτός ονομάζεται **αριστερόπλευρος έλεγχος**. Αντίστοιχα προκύπτουν ο δεξιόπλευρος και ο δίπλευρος έλεγχος, ανάλογα με τη μορφή της H_1. Οι τρεις περιπτώσεις συνοψίζονται στον επόμενο πίνακα:

Έλεγχος για την μέση τιμή μ κανονικού πληθυσμού με διακύμανση σ^2 γνωστή

H_1	Χωρίο απόρριψης	Κρίσιμη τιμή	p-value
(αρ. έλεγχος) $\mu < \mu_0$	$R = \{Z < -z_a\}$	$x_c = \mu_0 - z_a \frac{\sigma}{\sqrt{n}}$	$\Phi(Z)$
(δεξ. έλεγχος) $\mu > \mu_0$	$R = \{Z > z_a\}$	$x_c = \mu_0 + z_a \frac{\sigma}{\sqrt{n}}$	$\Phi(-Z)$
(δίπλευρος έλεγχος) $\mu \neq \mu_0$	$R = \{\|Z\| > z_{a/2}\}$	$x_c = \mu_0 \pm z_{a/2} \frac{\sigma}{\sqrt{n}}$	$2\Phi(-\|Z\|)$

Στατιστική συνάρτηση ελέγχου: $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$.

Η απόρριψη μιας σωστής μηδενικής υπόθεσης ονομάζεται **σφάλμα τύπου Ι**. Ο παραπάνω έλεγχος ορίσθηκε με τέτοιο τρόπο ώστε να ελαχιστοποιεί την πιθανότητα σφάλματος τύπου Ι: Η πιθανότητα αυτή με βάση τα παραπάνω είναι το πολύ ίση με a. Πράγματι, είναι

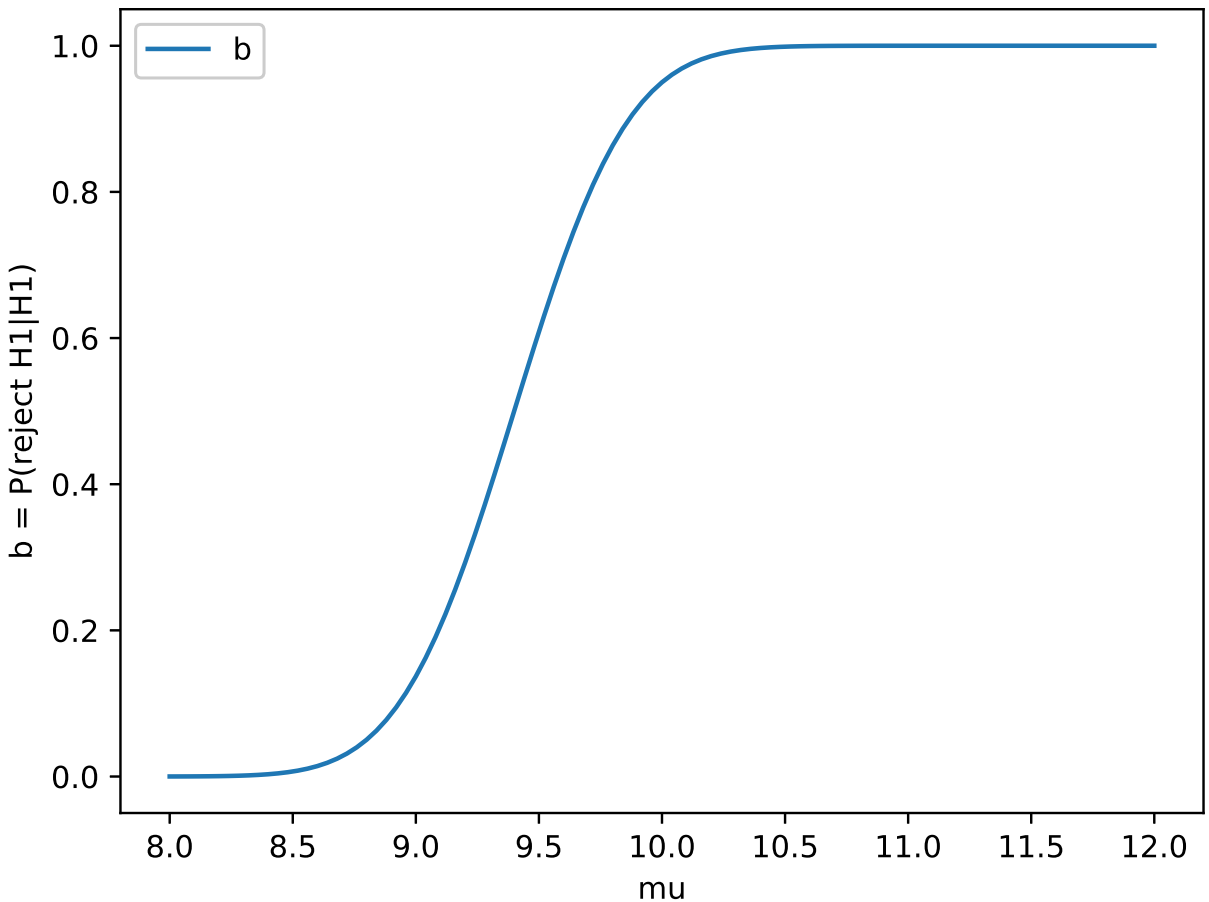
$$P(\overline{X} < \mu_0 - z_a\sigma/\sqrt{n} | \mu \geq \mu_0) \leq P(\overline{X} < \mu_0 - z_a\sigma/\sqrt{n} | \mu = \mu_0) = \Phi(-z_a) = a.$$

Επομένως, επιλέγοντας μικρότερο a, μειώνουμε αυτή την πιθανότητα. Όμως, υπάρχει και το ενδεχόμενο αποδοχής μιας λανθασμένης μηδενικής υπόθεσης, το οποίο ονομάζεται **σφάλμα τύπου ΙΙ**, με πιθανότητα που συμβολίζεται με β. Στη συγκεκριμένη περίπτωση, είναι

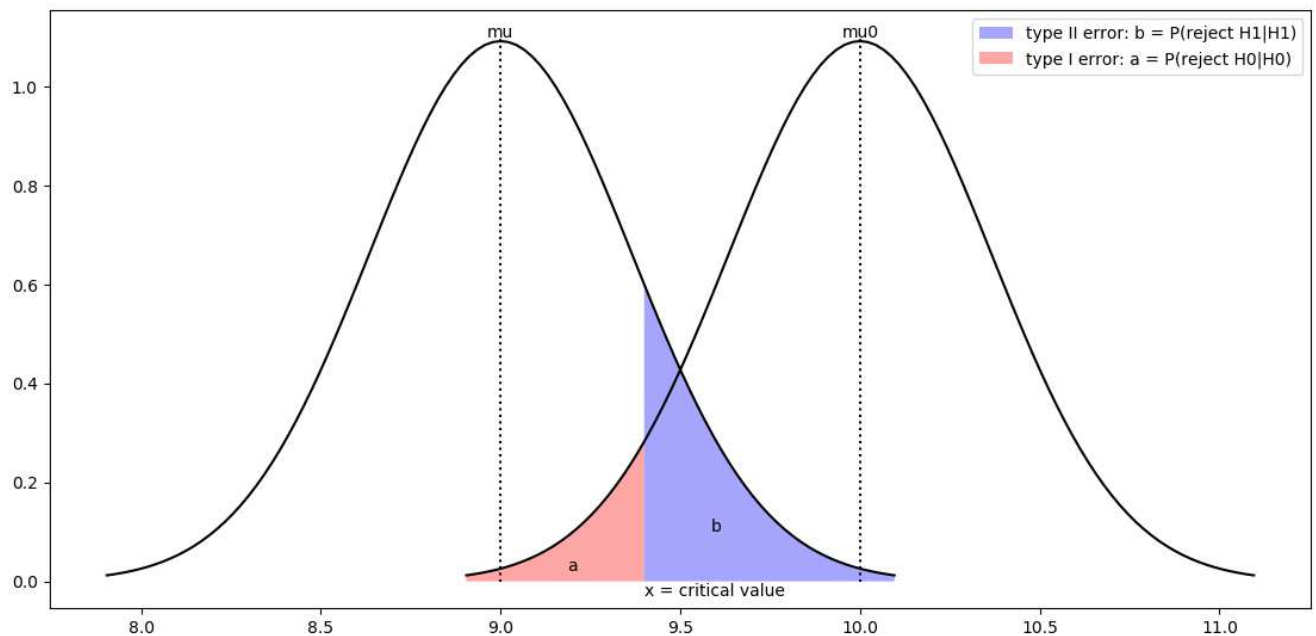
$$\beta = P(\overline{X} \geq \mu_0 - z_a\frac{\sigma}{\sqrt{n}} | H_1) = P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_a | \mu < \mu_0)$$

$$= \Phi(z_a - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}) \leq \Phi(z_a) = 1 - a$$

Στην επόμενη εικόνα φαίνεται η γραφική παράσταση του $\beta = \Phi(z_a - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}})$, ως συνάρτηση του μ, για $\mu_0 = 10$, $n = 30$, $a = 0.05$, $\sigma = 2$:



Στην επόμενη εικόνα, για τις ίδες τιμές και για $\mu = 9$, φαίνονται οι τιμές των a και β ως τα εμβαδά του κόκκινου και του μπλε χωρίου αντίστοιχα. Η αριστερή καμπύλη αντιστοιχεί στην κατανομή της ΤΜ $\overline{X} \sim N(\mu, \sigma^2/n)$, όταν $\mu = 9$. Η δεξιά αντιστοιχεί στην κατανομή της $\overline{X}$ όταν ισχύει η $H_0 : \mu = \mu_0 = 10$.



Η πιθανότητα $1-\beta$ απόρριψης μιας λανθασμένης μηδενικής υπόθεσης ονομάζεται **ισχύς του ελέγχου**. Προφανώς, το β είναι συνάρτηση των μ, a, n. Λύνοντας ως προς n, έχουμε ότι

$$\Phi(z_a - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}) = \beta = \Phi(-z_\beta) \Rightarrow z_a - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} = -z_\beta$$
$$\Rightarrow \frac{\mu_0 - \mu}{\sigma}\sqrt{n} = z_a + z_\beta$$
$$\Rightarrow \sqrt{n} = \frac{(z_a + z_\beta)\sigma}{\mu_0 - \mu},$$

δηλαδή, αν δίνονται τα μ, a, μ_0, μπορούμε να υπολογίσουμε το κατάλληλο μέγεθος δείγματος n ώστε να εξασφαλίσουμε ένα συγκεκριμένο μικρό άνω φράγμα για το β.

Παράδειγμα

Κανονικός πληθυσμός έχει διακύμανση $\sigma^2 = 1$. Θέλουμε να ελέγξουμε την υπόθεση ότι η μέση τιμή μ του πληθυσμού είναι $\mu_0 = 9$, έναντι της $H_1 : \mu \neq 9$. Λαμβάνουμε ένα τυχαίο δείγμα μεγέθους $n = 16$, στο οποίο ο δειγματικός μέσος ισούται με $\overline{X} = 9.4$. Να ελεγχθεί η υπόθεσή μας σε επίπεδο σημαντικότητας $a = 0.05$.

Λύση

Είναι $H_1 : \mu \neq \mu_0$, οπότε θα πραγματοποιηθεί δίπλευρος έλεγχος. Για $a = 0.05$, είναι $z_{a/2} = 1.96$.

Η στατιστική συνάρτηση ελέγχου είναι η $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$, με

$$Z = \frac{9.4 - 9}{1/4} = 1.6 < 1.96 = z_{a/2},$$

οπότε η H_0 δεν απορρίπτεται.

Παρατηρήσεις

Στο ίδιο συμπέρασμα καταλήγουμε, υπολογίζοντας

$$\text{p-value} = 2\Phi(-1.6) = 0.1096 > 0.05 = a.$$

Μπορούμε να υπολογίσουμε το μέγεθος n του δείγματος, για το οποίο θα απορρίπταμε την H_0:

$$Z > z_{a/2} \Leftrightarrow \sqrt{n}(\overline{X} - \mu_0) > 1.96 \Leftrightarrow \sqrt{n} > 1.96/0.4 = 4.9 \Leftrightarrow n \geq 25.$$

Για την ισχύ του ελέγχου, έχουμε

$$\begin{aligned} 1 - \beta &= 1 - P(-z_{a/2} \leq \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_{a/2} | H_1) \\ &= 1 - P(-z_{a/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \leq z_{a/2} | H_1) \\ &= 1 - \Phi(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{a/2}) + \Phi(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{a/2}) \end{aligned}$$

Έλεγχος Υποθέσεων

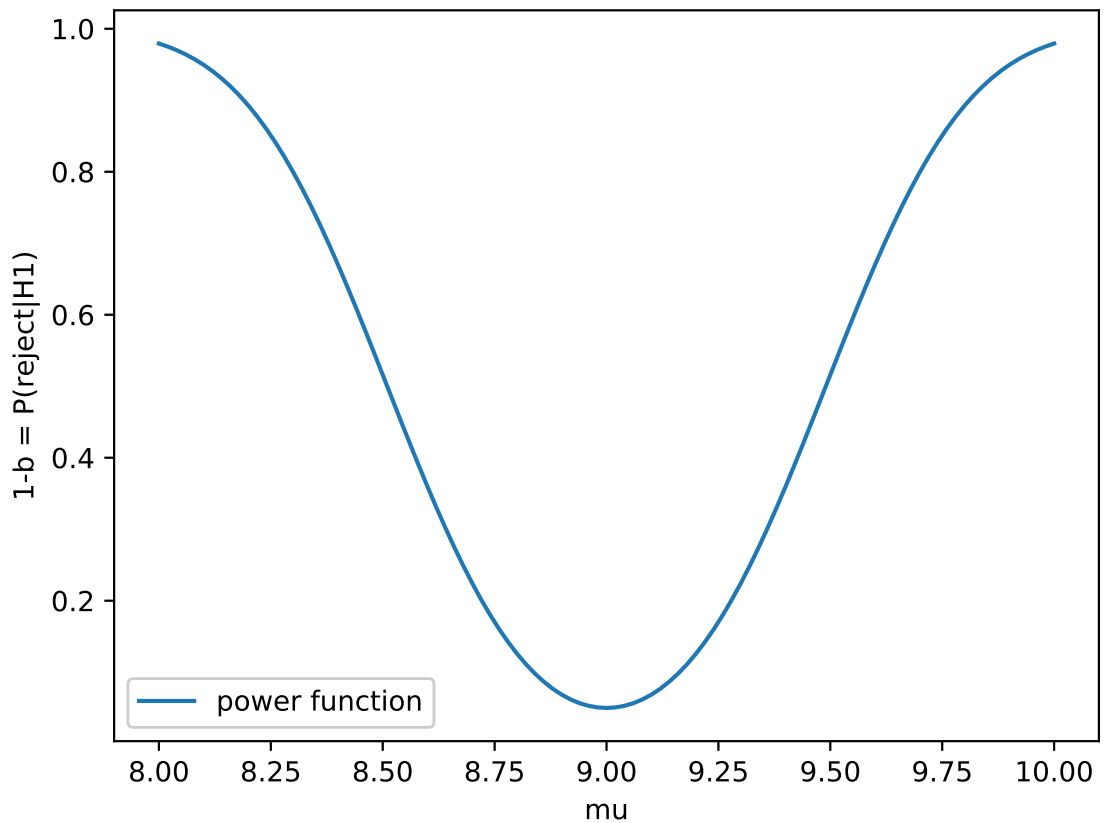
Παρακάτω δίνεται ο κώδικας για τη λύση του παραδείγματος 14.

```
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt
xbar, mu0, n, a, sigma = 9.4, 9, 16, 0.05, 1 #9.1.1
Z = np.sqrt(n)*(xbar-mu0)/sigma
pval = 2*st.norm.cdf(-np.abs(Z))
zpp = st.norm.ppf(1-a/2)
if(pval < a): print("H0 is rejected")
else: print("H0 is accepted")
print("Z=%s, p-value = %s, z_{a/2} = %s"%(Z, pval, zpp))

mu = np.linspace(8,10, 101)
err = (mu0-mu)*np.sqrt(n)
p = 1-st.norm.cdf(err+zpp) + st.norm.cdf(err-zpp)
fig, ax = plt.subplots(1, 1)
ax.plot(mu, p, label = 'power function')
ax.legend()
ax.set_xlabel("mu")
ax.set_ylabel("1-b = P(reject|H1)")
plt.show()
```

Output:

```
H0 is accepted
Z=1.6000000000000014, p-value = 0.10959858339911567, z_{a/2}
     = 1.95996398454054
```



Παράδειγμα

Εικάζουμε ότι κανονικός πληθυσμός έχει μέση τιμή $\mu_0 = 20$ και γνωστή διακύμανση $\sigma^2 = 7$. Ένα δείγμα μεγέθους $n = 25$ από τον πληθυσμό έδωσε $\overline{x} = 21.5$. Να ελεχθεί, σε επίπεδο σημαντικότητας $a = 0.04$, η υπόθεση $H_0 : \mu = 20$, έναντι της: *i*) $H_1 : \mu \neq 20$, *ii*) $H_1 : \mu > 20$.

Λύση

i) Είναι $H_1 : \mu \neq \mu_0$, οπότε θα πραγματοποιηθεί δίπλευρος έλεγχος. Για $a = 0.04$, είναι $z_{a/2} = 2.054$. Η στατιστική συνάρτηση ελέγχου είναι η

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}, \text{ με } Z = \frac{21.5 - 20}{\sqrt{7/25}} = 2.8347 > 2.054 = z_{a/2},$$

οπότε η H_0 απορρίπτεται. Στο ίδιο συμπέρασμα καταλήγουμε αν υπολογίσουμε την p-value του ελέγχου:

$$\text{p-value} = 2\Phi(-|Z|) = 2\Phi(-2.8347) = 0.0046 < 0.04 = a.$$

Λύση (συνέχεια)

ii) Στην περίπτωση αυτή, είναι p-value $= \Phi(-Z) = 0.0023 < a$, οπότε η H_0 απορρίπτεται.

Παρατηρήστε ότι η p-value είναι διπλάσια στην περίπτωση δίπλευρου ελέγχου, διότι στην πιθανότητα μιας τιμής $\overline{x}$ τόσο μεγαλύτερης από την μ_0 προστίθεται και η πιθανότητα μια τιμής εξίσου μικρότερης.

Έλεγχος για την μέση τιμή μ κανονικού πληθυσμού με διακύμανση σ^2 άγνωστη (one sample t-test)
Η περίπτωση αυτή αντιμετωπίζεται ομοίως με την προηγούμενη, με τη διαφορά ότι τώρα η στατιστική συνάρτηση του ελέγχου είναι η

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

Υπό την υπόθεση $H_0 : \mu = \mu_0$, η T ως γνωστό ακολουθεί την κατανομή t με $n-1$ βαθμούς ελευθερίας, οπότε, αντίστοιχα με πριν, προκύπτει ο ακόλουθος πίνακας:

Έλεγχος για την μέση τιμή μ κανονικού πληθυσμού με διακύμανση σ^2 άγνωστη (one sample t-test)

H_1	Χωρίο απόρριψης	Κρίσιμη τιμή	p-value
$\mu < \mu_0$	$R = \{T < -t_{n-1,a}\}$	$x_c = \mu_0 - t_{n-1,a}\frac{S}{\sqrt{n}}$	$F_{n-1}(T)$
$\mu > \mu_0$	$R = \{T > t_{n-1,a}\}$	$x_c = \mu_0 + t_{n-1,a}\frac{S}{\sqrt{n}}$	$F_{n-1}(-T)$
$\mu \neq \mu_0$	$R = \{\lvert T\rvert > t_{n-1,a/2}\}$	$x_c = \mu_0 \pm t_{n-1,a/2}\frac{\sigma}{\sqrt{n}}$	$2F_{n-1}(-\lvert T\rvert)$

όπου F_{n-1} η αθροιστική συνάρτηση της κατανομής t με $n-1$ βαθμούς ελευθερίας.

Παράδειγμα

Δείγμα μεγέθους $n = 36$ έδωσε δειγματικό μέσο $\overline{X} = 71$ και δειγματική διακύμανση $S^2 = 15$. Να ελεχθεί, σε επίπεδο σημαντικότητας $a = 0.02$, η υπόθεση $H_0 : \mu = 70$, έναντι της *i*) $H_1 : \mu \neq 70$, *ii*) $H_1 : \mu > 70$.

Λύση

i) Είναι $H_1 : \mu \neq \mu_0$, όπου $\mu_0 = 70$, οπότε θα πραγματοποιηθεί δίπλευρος έλεγχος. Για $a = 0.02$, είναι $t_{n-1,a/2} = 2.4377$.
Η στατιστική συνάρτηση ελέγχου είναι η

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}, \text{ με } T = \frac{71 - 70}{\sqrt{15/36}} = 1.55 < 2.4377 = t_{n-1,a/2},$$

οπότε η H_0 δεν απορρίπτεται. Στο ίδιο συμπέρασμα καταλήγουμε αν υπολογίσουμε την p-value του ελέγχου:

$$\text{p-value} = 2F_{n-1}(-1.55) = 0.13 > 0.02 = a.$$

Λύση (συνέχεια)

ii) Είναι $H_1 : \mu > \mu_0$, όπου $\mu_0 = 70$, οπότε θα πραγματοποιηθεί δεξιόπλευρος έλεγχος. Είναι $T = 1.55 < 2.133 = t_{n-1,a}$ και p-value $= F_{n-1}(-T) = 0.065 > a$, οπότε πάλι η H_0 δεν απορρίπτεται.

```
import numpy as np
import scipy.stats as st

xbar, mu0, n, a, s = 71, 70, 36, 0.02, np.sqrt(15)
T = np.sqrt(n)*(xbar-mu0)/s
pval = 2*st.t.cdf(-np.abs(T), n-1)
tpp = st.t.ppf(1-a/2, n-1)
print("i) H1: mu != %s, T=%s, p-value = %s, t_{a/2,n-1} = %s
    "%(mu0, T, pval, tpp))
```

Output:

```
i) H1: mu != 70, T=1.5491933384829668, p-value =
    0.1303312488722083 4, t_{a/2,n-1} = 2.437722547143737
```

Έλεγχος για το ποσοστό p πληθυσμού

Έστω τυχαίο δείγμα $(X_1, \ldots, X_n)$, με $X_i \sim \text{Bernoulli}(p)$ και έστω η μηδενική υπόθεση $H_0 : p = p_0$, για το ποσοστό p του πληθυσμού. Χρησιμοποιούμε τον συμβολισμό

$$F(x; n, p) := \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}.$$

για την αθροιστική συνάρτηση της κατανομής $\text{Binom}(n, p)$.
Ως γνωστό, η ΤΜ $X = \sum_{i=1}^{n} X_i$ ακολουθεί διωνυμική κατανομή με παραμέτρους n, p. Υπό την υπόθεση H_0, έχουμε ότι $X \sim \text{Binom}(n, p_0)$, με μέση τιμή $\mu_0 = np_0$. Έστω ότι η X έχει πάρει την τιμή k στο συγκεκριμένο δείγμα. Προκειμένου να απορρίψουμε την H_0, θα πρέπει η πιθανότητα η X να πάρει μια τιμή τόσο μακριά από την μ_0 όσο η παρατηρηθείσα τιμή k, δεδομένου ότι η H_0 είναι σωστή, να είναι μικρότερη από ένα επίπεδο εμπιστοσύνης a.

Έλεγχος για το ποσοστό p πληθυσμού

Στην περίπτωση αριστερόπλευρου ελέγχου, δηλαδή όταν $H_1 : p < p_0$, η πιθανότητα αυτή είναι

$$\text{p-value} = P(X \leq k|H_0) = P(X \leq k|p = p_0) = F(k; n, p_0),$$

Στην περίπτωση δεξιόπλευρου ελέγχου, δηλαδή όταν $H_1 : p > p_0$, η πιθανότητα αυτή είναι

$$\text{p-value} = P(X \geq k|H_0) = 1 - P(X < k|p = p_0) = 1 - F(k-1; n, p_0).$$

Στην περίπτωση δίπλευρου ελέγχου, δηλαδή όταν $H_1 : p \neq p_0$, η H_0 απορρίπτεται ανν ισχύουν οι ισοδύναμες συνθήκες

$$\begin{aligned} & P(X \leq k|H_0) < a/2 \text{ ή } P(X \geq k|H_0) < a/2 \\ \Leftrightarrow & 2F(k; n, p_0) < a \text{ ή } 2(1 - F(k-1; n, p_0)) < a \\ \Leftrightarrow & 2\min\{F(k; n, p_0), 1 - F(k-1; n, p_0)\} < a. \end{aligned}$$

Έλεγχος για το ποσοστό p πληθυσμού

Εναλλακτικά, όταν το n είναι μεγάλο, βάσει του ΚΟΘ, μπορεί να χρησιμοποιηθεί προσεγγιστικά ο z-έλεγχος που παρουσιάστηκε προηγουμένως. Πράγματι, ως γνωστό, όταν ισχύει η $H_0 : p = p_0$, τότε είναι

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)/n}} \to N(0,1)$$

Ο τρόπος αυτός προτιμάται όταν το n είναι μεγάλο, αφού τότε οι υπολογισμοί μέσω της διωνυμικής κατανομής είναι υπολογιστικά απαιτητικοί. Εισάγοντας και τη διόρθωση συνέχειας, ανάλογα με τη μορφή της H_1, η προσέγγιση μπορεί να βελτιωθεί σημαντικά, όταν το n δεν είναι πολύ μεγάλο. Συγκεκριμένα, όταν έχουμε αριστερόπλευρο (αντ. δεξιόπλευρο) έλεγχο, είναι $P(X \le k) = P(X \le k + 1/2)$ (αντ. $P(X \ge k) = P(X \ge k - 1/2)$), οπότε χρησιμοποιούμε αντίστοιχα τις στατιστικές συναρτήσεις

$$Z_+ = \frac{X - np_0 + 1/2}{\sqrt{np_0(1-p_0)}}, \qquad \text{αντ.} \quad Z_- = \frac{X - np_0 - 1/2}{\sqrt{np_0(1-p_0)}}$$

Έλεγχος για το ποσοστό p πληθυσμού

Κατόπιν τούτων, προκύπτουν οι παρακάτω περιπτώσεις για τον έλεγχο:

H_1	p-value (binomial test)	p-value (z-test)
$p < p_0$	$F(X; n, p_0)$	$\Phi(Z_+)$
$p > p_0$	$1 - F(X-1; n, p_0)$	$\Phi(-Z_-)$
$p \neq p_0$	$2\min\{F(X; n, p_0), 1-F(X-1; n, p_0)\}$	$2\min\{\Phi(Z_+), \Phi(-Z_-)\}$

όπου

$$Z_+ = \frac{X - np_0 + 1/2}{\sqrt{np_0(1-p_0)}}, \qquad \text{και} \quad Z_- = \frac{X - np_0 - 1/2}{\sqrt{np_0(1-p_0)}}$$

Έλεγχος για το ποσοστό p πληθυσμού

Στον κώδικα που ακολουθεί, εκτελείται αρχικά αριστερόπλευρος, δεξιόπλευρος και δίπλευρος διωνυμικός έλεγχος με μηδενική υπόθεση $H_0 : p = p_0 = 0.5$, όπου το X παράγεται τυχαία από την κατανομή $\mathrm{Binom(n, p)}$, όπου $n = 30, p = 0.6$. Στη συνέχεια, εκτελείται z-έλεγχος πρώτα χωρίς και έπειτα με διόρθωση συνέχειας. Ο τελευταίος δίνει καλύτερη προσέγγιση, όπως φαίνεται και από το αποτέλεσμα.

Έλεγχος για το ποσοστό p πληθυσμού

```
import numpy as np
import scipy.stats as st

n, p, p0 = 30, 0.6, 0.5
x = np.random.binomial(n, p)
H1 = {'greater':'p>p0', 'less':'p<p0', 'two-sided':'p!=p0'}
print("X = %s, X/n = %s"%(x, x/n))

print("\nexact binomial test:")
theory_pval = {'greater':1-st.binom.cdf(x-1,n,p0),
               'less':st.binom.cdf(x,n,p0),
               'two-sided':2*np.min([1-st.binom.cdf(x-1,n,p0
    ),st.binom.cdf(x,n,p0)])}
for k, v in H1.items():
    result = st.binom_test(x, n, p0, alternative = k)
    print("H1:%s, p-value = %s, theory p-value = %s"%(v,
    result, theory_pval[k]))
```

Έλεγχος για το ποσοστό p πληθυσμού

```
print("\nz-test:")
se = np.sqrt(n*p0*(1 - p0))
z = (x - n*p0)/se
theory_pval = {'greater':st.norm.cdf(-z),
               'less':st.norm.cdf(z),
               'two-sided':2*st.norm.cdf(-np.abs(z))}

for i in H1.keys():
    print("H1:%s, theory p-value = %s"%(H1[i], theory_pval[i
    ]))

print("\nz-test with continuity correction:")
zm = (x - n*p0 - 1/2)/se
zp = (x - n*p0 + 1/2)/se
theory_pval = {'greater':st.norm.cdf(-zm),
               'less':st.norm.cdf(zp),
               'two-sided':2*np.min([st.norm.cdf(-zm),st.
    norm.cdf(zp)])}
for i in H1.keys():
    print("H1:%s, theory p-value = %s"%(H1[i], theory_pval[i
    ]))
```

Έλεγχος για το ποσοστό p πληθυσμού

Output:

```
X = 19, X/n = 0.6333333333333333

exact binomial test:
H1:p>p0, p-value = 0.10024421103298661, theory p-value =
    0.10024421103298664
H1:p<p0, p-value = 0.9506314266473055, theory p-value =
    0.9506314266473055
H1:p!=p0, p-value = 0.20048842206597323, theory p-value =
    0.20048842206597328

z-test:
H1:p>p0, theory p-value = 0.07206351740800766
H1:p<p0, theory p-value = 0.9279364825919924
H1:p!=p0, theory p-value = 0.1441270348160153 3

z-test with continuity correction:
H1:p>p0, theory p-value = 0.1006213104788619 7
H1:p<p0, theory p-value = 0.9498258767688547
H1:p!=p0, theory p-value = 0.2012426209577239 4
```

Έλεγχος για το ποσοστό p πληθυσμού

Παράδειγμα

Σε μια πόλη, 1 στους 10 καταναλωτές προτιμά την μάρκα Α. Μετά από μια έντονη διαφημιστική καμπάνια, εξετάστηκε ένα τυχαίο δείγμα 200 ατόμων, προκειμένου να διαπιστωθεί η αποτελεσματικότητα της καμπάνιας. Οι μετρήσεις έδειξαν ότι 26 άτομα προτιμούν την Α. Να ελεχθεί σε επίπεδο σημαντικότητας $a = 0.05$ η αποτελεσματικότητα της καμπάνιας.

Λύση

Η μηδενική υπόθεση είναι η $H_0 : p \leq p_0$, όπου $p_0 = 0.1$ και $H_1 : p > p_0$. Η τελευταία δηλώνει ότι το ποσοστό αυξήθηκε μετά την καμπάνια. Έχουμε $X = 26 > 20$, οπότε εφαρμόζοντας δεξιόπλευρο διωνυμικό έλεγχο, βρίσκουμε ότι

$$\text{p-value} = 1 - F(X - 1; n, p_0) = 1 - F(25; 200, 0.1) = 0.1$$

οπότε η H_0 δεν μπορεί να απορριφθεί.

Έλεγχος για το ποσοστό p πληθυσμού

```
import numpy as np
import scipy.stats as st
n, a, p0, x, H1 = 200, 0.05, 0.1, 26, 'greater' #9.1.4
#binomial test
pval = st.binom_test(x, n, p0, alternative = H1)
print("binomial test:\nx =", x, ", p-value =", pval, ",
    Upper a-percent point:", st.binom.ppf(1 - a,n,p0),"\n")
#print("p-value = ", 1-st.binom.cdf(x-1,n,p0))
#z-test
z = (x - n*p0 - 1/2)/np.sqrt(p0*(1 - p0)*n)
print("z-test:\nz = %f, z_{a} = %f, p-value = %f\n"
      %(z, st.norm.ppf(1-a), st.norm.cdf(-z)))
```

Output:

```
binomial test:
x = 26 , p-value = 0.10045728651286649 , Upper a-percent
    point: 27.0

z-test:
z = 1.296362, z_{a} = 1.644854, p-value = 0.097425
```

Έλεγχος για το ποσοστό p πληθυσμού

Παράδειγμα

Εικάζεται ότι το ποσοστό των χορτοφάγων σε μια πόλη είναι 22%. Επιλέγουμε τυχαία 260 άτομα και βρίσκουμε ανάμεσά τους 62 χορτοφάγους. Σε επίπεδο σημαντικότητας $a = 0.02$, είναι σωστή αυτή η εικασία;

Λύση

Θέτουμε $H_0 : p = p_0$ και $H_1 : p \neq p_0$, όπου $p_0 = 0.22$. Εφαρμόζοντας δίπλευρο z-έλεγχο (χωρίς διόρθωση συνέχειας), βρίσκουμε

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = 0.718614, \quad z_{a/2} = 2.326348, \quad \text{p-value} = 0.472379$$

οπότε η H_0 δεν απορρίπτεται. Σημειώνεται ότι το δειγματικό ποσοστό $62/260 = 0.23846$ είναι πολύ κοντά στο 0.22. Αν εισάγουμε διόρθωση συνέχειας, τότε προκύπτει p-value $= 0.51973$.

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 γνωστές

Αν $(X_1, \ldots, X_{n_1}) \sim N(\mu_1, \sigma_1^2)$ και $(Y_1, \ldots, Y_{n_2}) \sim N(\mu_2, \sigma_2^2)$ είναι ανεξάρτητα τυχαία δείγματα από 2 πληθυσμούς και $\overline{X}$, $\overline{Y}$ οι αντίστοιχοι δειγματικοί μέσοι, τότε ως γνωστό είναι

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \sigma^2), \qquad \text{όπου } \sigma^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2.$$

Αν θέλουμε να ελέγξουμε την $H_0 : \mu_1 = \mu_2$ έναντι π.χ. της $H_1 : \mu_1 \neq \mu_2$, τότε θεωρούμε την στατιστική συνάρτηση

$$Z = (\overline{X} - \overline{Y})/\sigma = (\overline{X} - \overline{Y})/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

η οποία ακολουθεί την $N(0,1)$, όταν ισχύει η H_0. Αυτό σημαίνει ότι η πιθανότητα να παρατηρηθεί μια ακραία τιμή $c > 0$ της $|\overline{X} - \overline{Y}|$, όταν ισχύει η H_0, είναι

$$P(|\overline{X} - \overline{Y}| \geq c | H_0) = P(|Z| \geq c/\sigma | H_0) = 2\Phi(-c/\sigma).$$

Επομένως, η p-value του ελέγχου είναι η $2\Phi(-|Z|)$ και η H_0 απορρίπτεται σε επίπεδο σημαντικότητας a, όταν p-value $< a \Leftrightarrow 2\Phi(-|Z|) < a \Leftrightarrow -|Z| < -z_{a/2} \Leftrightarrow |Z| > z_{a/2} \Leftrightarrow |\overline{X} - \overline{Y}| > z_{a/2}\sigma$.

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 γνωστές

Ομοίως προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον επόμενο πίνακα:

H_1	Χωρίο απόρριψης	p-value
(αριστερόπλευρος έλεγχος) $\mu_1 - \mu_2 < 0$	$R = \{Z < -z_a\}$	$\Phi(Z)$
(δεξιόπλευρος έλεγχος) $\mu_1 - \mu_2 > 0$	$R = \{Z > z_a\}$	$\Phi(-Z)$
(δίπλευρος έλεγχος) $\mu_1 - \mu_2 \neq 0$	$R = \{\|Z\| > z_{a/2}\}$	$2\Phi(-\|Z\|)$

όπου

$$Z = (\overline{X} - \overline{Y}) / \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 γνωστές

Παράδειγμα

Δοκιμάζουμε ελαστικά αυτοκινήτου μιας μάρκας σε μια τοποθεσία Α και μιας άλλης μάρκας σε μια άλλη τοποθεσία Β και παίρνουμε τα ακόλουθα δείγματα για τη διάρκεια ζωής τους (σε 100km):

$$A : 61.1, 58.2, 62.3, 64, 59.7, 66.2, 57.8, 61.4, 62.2, 63.6$$

$$B : 62.2, 56.6, 66.4, 56.2, 56.2, 57.4, 58.4, 57.6, 65.4$$

Από την εμπειρία γνωρίζουμε ότι η τυπική απόκλιση της διάρκειας ζωής εξαρτάται από την τοποθεσία, με $\sigma_A = 3$ και $\sigma_B = 4$, καθώς και ότι η διάρκεια ζωής ακολουθεί κανονική τιμή με μέσο που εξαρτάται μόνο από την κατασκευή. Να ελεχθεί η υπόθεση $H_0 : \mu_A = \mu_B$, έναντι της υπόθεσης $H_1 : \mu_A \neq \mu_B$.

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 γνωστές

Λύση

Υπολογίζοντας όπως παρακάτω ότι p-value $= 0.21$, δεν απορρίπτουμε την H_0.

```
import numpy as np
import scipy.stats as st
A = [61.1, 58.2, 62.3, 64, 59.7, 66.2, 57.8, 61.4, 62.2,
    63.6]
B = [62.2, 56.6, 66.4, 56.2, 56.2, 57.4, 58.4, 57.6, 65.4]
sA, sB = 3, 4
xbar, ybar, sigma = st.tmean(A), st.tmean(B), np.sqrt(sA**2/
    len(A) + sB**2/len(B))
z = (xbar-ybar)/sigma
pval = 2*st.norm.cdf(-np.abs(z))
print("Z = %s, sigma = %s, p-value = %s"%(z, sigma, pval))
```

Output:

```
Z = 1.252756297229828 7, sigma = 1.636391694844772, p-value =
    0.2102944108000274 2
```

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 άγνωστες, αλλά ίσες

Θεωρούμε 2 ανεξάρτητα δείγματα $(X_1, \ldots, X_{n_1}) \sim N(\mu_1, \sigma_1^2)$ και $(Y_1, \ldots, Y_{n_2}) \sim N(\mu_2, \sigma_2^2)$ όπως πριν, αλλά τώρα είναι $\sigma_1 = \sigma_2 = \sigma$, αλλά σ άγνωστη. Αν θέλουμε να ελέγξουμε π.χ. την $H_0 : \mu_1 = \mu_2$, έναντι της $H_1 : \mu_1 \neq \mu_2$, τότε θεωρούμε την στατιστική συνάρτηση

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{1/n_1 + 1/n_2}}, \qquad \text{όπου } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

η οποία ως γνωστό ακολουθεί την $t_{n_1+n_2-2}$, όταν ισχύει η H_0. Αυτό σημαίνει ότι η πιθανότητα να παρατηρηθεί μια ακραία τιμή $c > 0$ της $|\overline{X} - \overline{Y}|$, όταν ισχύει η H_0, είναι

$$P(|\overline{X} - \overline{Y}| \geq c|H_0) = P(|T| \geq c/D|H_0) = 2F_{n_1+n_2-2}(-c/D),$$

όπου $F_{n_1+n_2-2}$ η αθροιστική συνάρτηση της κατανομής $t_{n_1+n_2-2}$ και D ο παρονομαστής της T. Επομένως, η p-value του ελέγχου είναι η $2F_{n_1+n_2-2}(-|T|)$ και η H_0 απορρίπτεται σε επίπεδο σημαντικότητας a, όταν

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 άγνωστες, αλλά ίσες

$$\text{p-value} < a \Leftrightarrow 2_{n_1+n_2-2}(-|T|) < a \Leftrightarrow -|T| < -t_{a/2,n_1+n_2-2}$$

$$\Leftrightarrow |T| > t_{a/2,n_1+n_2-2} \Leftrightarrow |\overline{X} - \overline{Y}| > t_{a/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Ομοίως προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον επόμενο πίνακα:

H_1	Χωρίο απόρριψης	p-value
$\mu_1 - \mu_2 < 0$	$R = \{T < -t_{a,n_1+n_2-2}\}$	$F_{n_1+n_2-2}(T)$
$\mu_1 - \mu_2 > 0$	$R = \{T > t_{a,n_1+n_2-2}\}$	$F_{n_1+n_2-2}(-T)$
$\mu_1 - \mu_2 \neq 0$	$R = \{\|T\| > t_{a/2,n_1+n_2-2}\}$	$2F_{n_1+n_2-2}(-\|T\|)$

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 άγνωστες και άνισες

Σε αυτή την περίπτωση, χρησιμοποιείται η στατιστική

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \to t_d, \quad d = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \left(\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}\right)^{-1},$$

της οποίας η κατανομή αποδεικνύεται ότι συγκλίνει στην κατανομή t με d βαθμούς ελευθερίας.
Ο έλεγχος αυτός είναι γνωστός ως t-έλεγχος Welch.
Αν τα n_1, n_2 είναι αρκετά μεγάλα, τότε για απλότητα χρησιμοποιείται απλός z-έλεγχος, αφού $T \to N(0,1)$.
Σημειώνεται ότι και οι δύο έλεγχοι είναι προσεγγιστικοί.

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 άγνωστες και άνισες

H_1	Χωρίο απόρριψης (t-test)	p-value (t-test)
$\mu_1 - \mu_2 < 0$	$R = \{T < -t_{a,d}\}$	$F_d(T)$
$\mu_1 - \mu_2 > 0$	$R = \{T > t_{a,d}\}$	$F_d(-T)$
$\mu_1 - \mu_2 \neq 0$	$R = \{\|T\| > t_{a/2,d}\}$	$2F_d(-\|T\|)$

Αν δίνονται οι λίστες A, B τιμών των δύο δειγμάτων, οι παραπάνω t-έλεγχοι εκτελούνται στην Python με τη συνάρτηση (equal_var = True, για ίσες διακυμάνσεις):

```
scipy.stats.ttest_ind(A, B, equal_var=False, alternative=s)
```

όπου το s παίρνει αντίστοιχα τιμή "less", "greater", "two-sided".

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 άγνωστες και άνισες

Παράδειγμα

Από τυχαίο δείγμα 40 εργαζομένων μιας εταιρείας A, προέκυψε ότι το μέσο ετήσιο εισόδημα είναι $\overline{X} = 54000$, με τυπική απόκλιση $S_1 = 6000$. Επίσης, από τυχαίο δείγμα 50 εργαζομένων μιας εταιρείας B, προέκυψε ότι το μέσο ετήσιο εισόδημα είναι $\overline{Y} = 57000$, με τυπική απόκλιση $S_2 = 8000$. Μπορούμε να δεχθούμε ότι οι εργαζόμενοι στις δύο εταιρείες έχουν τον ίδιο μέσο μισθό σε επίπεδο σημαντικότητας $a = 0.01$; Να γίνει έλεγχος και για τις δύο περιπτώσεις που για τις άγνωστες διακυμάνσεις είναι $\sigma_A = \sigma_B$ και $\sigma_A \neq \sigma_B$.

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 άγνωστες και άνισες

Λύση

Θεωρούμε $H_0 : \mu_1 = \mu_2$. Ο κώδικας που ακολουθεί εκτελεί αριστερόπλευρο και δίπλευρο έλεγχο. Η στατιστική συνάρτηση ελέγχου είναι η

$$T = \frac{\overline{X} - \overline{Y}}{s},$$

όπου

$$s^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \text{ όταν } \sigma_A = \sigma_B,$$

$$s^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}, \text{ όταν } \sigma_A \neq \sigma_B,$$

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 άγνωστες και άνισες

```
import numpy as np
import scipy.stats as st
n1, n2, xbar, ybar, s1, s2, a = 40, 50, 54000, 57000, 6000, 8000, 0.01 #9.1.7
H1 = ['mu1 < mu2', 'mu1 != mu2']
s = np.sqrt(((n1-1)*s1**2+(n2-1)*s2**2)/(n1+n2-2)*(1/n1+1/n2))

T = (xbar-ybar)/s #equal variances
tp, pval = st.t.ppf(1-a, n1+n2-2), st.t.cdf(T, n1+n2-2) #less
tp2, pval2 = st.t.ppf(1-a/2, n1+n2-2), 2*st.t.cdf(-np.abs(T), n1+n2-2) #two-sided
print("t-test, equal variances: a =", a)
print("\tH1: %s: T = %.6f, -t_{a} = %.6f, p-value = %.6f"%(H1[0], T, -tp, pval))
print("\tH1: %s: T = %.6f, t_{a/2} = %.6f, p-value = %.6f"%(H1[1], T,tp2,pval2))

T = (xbar-ybar)/(np.sqrt(s1**2/n1 + s2**2/n2)) #unequal variances
zp, zpval = st.norm.ppf(1-a), st.norm.cdf(T) #less
zp2, zpval2 = st.norm.ppf(1-a/2), 2*st.norm.cdf(-np.abs(T)) #two-sided
print("z-test, unequal variances: a =", a)
print("\tH1: %s: Z = %.6f, -z_{a} = %.6f, p-value = %.6f"%(H1[0], T, -zp, zpval))
print("\tH1: %s: Z = %.6f, z_{a/2} = %.6f, p-value = %.6f"%(H1[1], T,zp2,zpval2))

df = np.floor((s1**2/n1+s2**2/n2)**2/(s1**4/(n1*n1*(n1-1))+s2**4/(n2*n2*(n2-1))))
tp, tpval = st.t.ppf(1-a, df), st.t.cdf(T, df) #less
tp2, tpval2 = st.t.ppf(1-a/2, df), 2*st.t.cdf(-np.abs(T), df) #two-sided
print("t-test, unequal variances: a =", a)
print("\tH1: %s: T = %.6f, -t_{a} = %.6f, p-value = %.6f"%(H1[0], T, -tp, tpval))
print("\tH1: %s: T = %.6f, t_{a/2} = %.6f, p-value = %.6f"%(H1[1], T,tp2,tpval2))
```

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών όταν σ_1, σ_2 άγνωστες και άνισες

Output:

```
t-test, equal variances: a = 0.01
  H1: mu1 < mu2: T = -1.968922, -t_{a} = -2.369472, p-value
    = 0.026054
  H1: mu1 != mu2: T = -1.968922, t_{a/2} = 2.632858, p-value
    = 0.052109
z-test, unequal variances: a = 0.01
  H1: mu1 < mu2: Z = -2.031856, -z_{a} = -2.326348, p-value
    = 0.021084
  H1: mu1 != mu2: Z = -2.031856, z_{a/2} = 2.575829, p-value
    = 0.042168
t-test, unequal variances: a = 0.01
  H1: mu1 < mu2: T = -1.968922, -t_{a} = -2.369977, p-value
    = 0.022609
  H1: mu1 != mu2: T = -1.968922, t_{a/2} = 2.633527, p-value
    = 0.052145
```

Σε κάθε περίπτωση, δεν απορρίπτουμε την H_0. Η περίπτωση $\sigma_A \neq \sigma_B$ έχει μεγαλύτερη p-value, διότι η παρατηρηθείσα διαφορά των μέσων μπορεί να οφείλεται σε κάποιο βαθμό στη διαφορά των διακυμάνσεων.

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών, όχι ανεξάρτητων (paired t-test)

Θεωρούμε 2 τυχαία δείγματα

$$(X_1, \ldots, X_n) \sim N(\mu_1, \sigma_1^2) \text{ και } (Y_1, \ldots, Y_n) \sim N(\mu_2, \sigma_2^2),$$

αλλά τώρα η X_i εξαρτάται από την Y_i. Για παράδειγμα, X_i και Y_i είναι η κατανάλωση βενζίνης του ίδιου αυτοκινήτου i χρησιμοποιώντας την βενζίνη Α και Β αντίστοιχα. Αν θέλουμε να ελέγξουμε π.χ. την $H_0 : \mu_1 = \mu_2$, έναντι της $H_1 : \mu_1 \neq \mu_2$, τότε θεωρούμε την ΤΜ $W_i = X_i - Y_i$, η οποία είναι κανονική $N(\mu_1 - \mu_2, \sigma^2)$, όπου $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\mathrm{Cov}(X_i, Y_i)$, και, αν ισχύει η H_0, έχει μέση τιμή 0. Το διάνυσμα $(W_1, \ldots, W_n)$ είναι ένα τυχαίο δείγμα με δειγματική διακύμανση, έστω S^2. Επομένως, η στατιστική συνάρτηση

$$T = \frac{\overline{X} - \overline{Y}}{S/\sqrt{n}}$$

ακολουθεί την t_{n-1}, όταν ισχύει η H_0. Μπορούμε λοιπόν και εδώ να εφαρμόσουμε t-έλεγχο με τη στατιστική συνάρτηση T.

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών, όχι ανεξάρτητων (paired t-test)

Αν δίνονται οι λίστες A, B τιμών των δύο δειγμάτων, ο παραπάνω t-έλεγχος εκτελείται στην Python με τη συνάρτηση:

```
scipy.stats.ttest_rel(A, B, alternative='two-sided')
```

Παράδειγμα

Σε ένα δείγμα $n = 28$ ανθρώπων, ο μέσος χρόνος απόκρισης σε ένα ερέθισμα είναι $\overline{X} = 1$ sec. Μετά τη λήψη $100ml$ αλκοόλ, ο μέσος χρόνος απόκρισης αυξήθηκε σε $\overline{Y} = 1.2$ sec, με δειγματική διακύμανση $S^2 = 0.25$. Σε ποιο επίπεδο εμπιστοσύνης μπορούμε να συμπεράνουμε ότι το αλκοόλ αυξάνει το χρόνο απόκρισης;

Ο κώδικας που ακολουθεί εκτελεί αριστερόπλευρο και δίπλευρο έλεγχο. Από τον δίπλευρο έλεγχο και για $a = 0.05$, συμπεραίνουμε με εμπιστοσύνη 95% ότι το αλκοόλ επηρεάζει τον χρόνο απόκρισης και από τον αριστερόπλευρο έλεγχο και για $a = 0.03$, συμπεραίνουμε με εμπιστοσύνη 97% ότι το αλκοόλ αυξάνει τον χρόνο απόκρισης.

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο κανονικών πληθυσμών, όχι ανεξάρτητων (paired t-test)

```
import numpy as np
import scipy.stats as st
n, xbar, ybar, s, a = 28, 1, 1.2, 0.5, 0.05 #9.1.9
H1 = ['mu1 < mu2', 'mu1 != mu2']
T = np.sqrt(n)*(xbar-ybar)/s
tp, pval = st.t.ppf(1-a,n-1), st.t.cdf(T, n-1) #less
tp2, pval2 = st.t.ppf(1-a/2,n-1), 2*st.t.cdf(-np.abs(T), n
    -1) #two-sided
print("t-test, a =", a)
print("H1: %s: T = %.6f, -t_{a} = %.6f, p-value = %.6f"%(H1
    [0], T, -tp, pval))
print("H1: %s: T = %.6f, t_{a/2} = %.6f, p-value = %.6f"%(H1
    [1], T, tp2, pval2))
```

Output:

```
t-test, a = 0.05
H1: mu1 < mu2: T = -2.116601, -t_{a} = -1.703288, p-value =
    0.021827
H1: mu1 != mu2: T = -2.116601, t_{a/2} = 2.051831, p-value =
     0.043655
```

Έλεγχος για την διαφορά $p_1 - p_2$ των ποσοστών δύο πληθυσμών

Θεωρούμε 2 ανεξάρτητα τυχαία δείγματα

$$(X_1, \ldots, X_{n_1}) \sim \text{Bernoulli}(p_1) \kappa a i (Y_1, \ldots, Y_{n_2}) \sim \text{Bernoulli}(p_2).$$

Ως γνωστό, είναι $X = \sum_{i=1}^{n_1} X_i \sim \text{Binom}(n_1, p_1)$ και $Y = \sum_{i=1}^{n_2} Y_i \sim \text{Binom}(n_2, p_2)$. Επίσης, ως γνωστό, από το ΚΟΘ, είναι

$$\overline{X} = X/n_1 \to N(p_1, p_1(1-p_1)/n_1), \qquad \overline{Y} = Y/n_2 \to N(p_2, p_2(1-p_2)/n_2),$$

και αφού $\overline{X}, \overline{Y}$ ανεξάρτητες, έχουμε ότι

$$\overline{X} - \overline{Y} \to N(p_1 - p_2, p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2).$$

Επομένως, υπό την μηδενική υπόθεση $H_0 : p_1 = p_2$, έχουμε ότι

$$\frac{\overline{X} - \overline{Y}}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \to N(0,1).$$

Έλεγχος για την διαφορά $p_1 - p_2$ των ποσοστών δύο πληθυσμών

Επειδή τα p_1, p_2 είναι άγνωστα, ο παρονομαστής αντικαθίσταται από την εκτιμήτρια

$$S_p = \sqrt{\overline{P}(1-\overline{P})(1/n_1 + 1/n_2)}, \qquad \text{όπου } \overline{P} = \frac{X+Y}{n_1+n_2}$$

και μπορούμε να εφαρμόσουμε (προσεγγιστικό) z-έλεγχο με στατιστική συνάρτηση την

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\overline{P}(1-\overline{P})(1/n_1 + 1/n_2)}} \to N(0,1), \qquad \overline{P} = \frac{X+Y}{n_1+n_2}.$$

Έλεγχος για την διαφορά $p_1 - p_2$ των ποσοστών δύο πληθυσμών

Αν θέλουμε να εφαρμόσουμε ακριβή έλεγχο, τότε, αν παρατηρήσαμε $X = k_1$ και $Y = k_2$, είναι

$$P(X = k_1 | X + Y = k_1 + k_2, H_0) = \binom{n_1}{k_1}\binom{n_2}{k_2} / \binom{n_1 + n_2}{k_1 + k_2} = f(k_1; M, n, N),$$

όπου $M = n_1 + n_2, n = n_1, N = k_1 + k_2$ και $f(x; M, n, N)$ η PMF της $\mathrm{HGeom(M, n, N)}$.
Στην περίπτωση δίπλευρου ελέγχου, η H_0 απορρίπτεται όταν ισχύει μια από τις ισοδύναμες συνθήκες

$$P(X \leq k_1 | X + Y = k_1 + k_2, H_0) \leq \frac{a}{2} \text{ ή } P(X \geq k_1 | X + Y = k_1 + k_2, H_0) \leq \frac{a}{2}$$

$$\Leftrightarrow 2 \min\{F(k_1; M, n, N), 1 - F(k_1 - 1; M, n, N)\} < a$$

Επομένως, είναι p-value $= 2 \min\{F(k_1; M, n, N), 1 - F(k_1 - 1; M, n, N)\}$.

Έλεγχος για την διαφορά $p_1 - p_2$ των ποσοστών δύο πληθυσμών

Ανάλογα, προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον επόμενο πίνακα:

H_1	p-value (exact test)	p-value (z-test)
$p_1 < p_2$	$F(X; M, n, N)$	$\Phi(Z)$
$p_1 > p_2$	$1 - F(X-1; M, n, N)$	$\Phi(-Z)$
$p_1 \neq p_2$	$2\min\{F(X; M, n, N), 1 - F(X-1; M, n, N)\}$	$2\Phi(-\lvert Z\rvert)$

όπου

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\overline{P}(1-\overline{P})(1/n_1 + 1/n_2)}}, \qquad \overline{P} = \frac{X+Y}{n_1+n_2},$$

και $F(x; M, n, N)$ η CDF της $\mathrm{HGeom(M, n, N)}$.

Έλεγχος για την διαφορά $p_1 - p_2$ των ποσοστών δύο πληθυσμών

Παράδειγμα

Δύο απορρυπαντικά δοκιμάστηκαν για την ικανότητά τους να καθαρίζουν λεκέδες. Το πρώτο ήταν επιτυχές σε 63 από 91 δοκιμές, ενώ το δεύτερο σε 49 από 79. Κάποιοι πιστεύουν ότι το πρώτο είναι καλύτερο. Να ελεχθεί η υπόθεση αυτή σε επίπεδο σημαντικότητας $a = 0.05$.

Ο κώδικας που ακολουθεί εκτελεί δεξιόπλευρο έλεγχο, με $H_0 : p_1 = p_2$ και $H_1 : p_1 > p_2$.

```
import numpy as np
import scipy.stats as st
x,n1, y, n2, a = 63, 91, 49, 79, 0.05
xbar, ybar = x/n1, y/n2
pbar = (x + y)/(n1 + n2)
sp = np.sqrt(pbar*(1 - pbar)*(1/n1 + 1/n2))
sp2 = np.sqrt(xbar*(1-xbar)/n1 + ybar*(1-ybar)/n2)
Z = (xbar - ybar)/sp
print(xbar, ybar, pbar, sp, sp2)
```

Έλεγχος για την διαφορά $p_1 - p_2$ των ποσοστών δύο πληθυσμών

```
print("\nz-test: a =", a)
za = st.norm.ppf(1-a)
pval = st.norm.cdf(-Z) #greater
print("Z = %.6f, z_{a} = %.6f, p-value = %.6f"%(Z,za,pval))

print("\nexact test: a =", a)
M,n,N = n1+n2, n1, x+y
pval2 = st.hypergeom.sf(x,M,n,N)
print("p-value = %.6f"%(pval2))
```

Output:

```
0.6923076923076923 0.620253164556962 0.6588235294117647
    0.07290617364270878 0.07295452675511863

z-test: a = 0.05
Z = 0.988319, z_{a} = 1.644854, p-value = 0.161498

exact test: a = 0.05
p-value = 0.125001
```

Έλεγχος για την διακύμανση σ^2 κανονικού πληθυσμού

Θεωρούμε τυχαίο δείγμα $(X_1, \ldots, X_{n_1}) \sim N(\mu, \sigma^2)$, με δειγματικό μέσο $\overline{X}$ και δειγματική διακύμανση S^2. Υπό τη μηδενική υπόθεση $H_0 : \sigma = \sigma_0$, η στατιστική συνάρτηση

$$X = \frac{(n-1)S^2}{\sigma_0^2}$$

ακολουθεί κατανομή χ^2_{n-1} με $n-1$ βαθμούς ελευθερίας και για $a \in (0,1)$ είναι

$$P(\chi^2_{n-1,1-a/2} \leq X \leq \chi^2_{n-1,a/2}) = 1 - a$$

Επομένως, για έναν δίπλευρο έλεγχο με επίπεδο εμπιστοσύνης a, αποδεχόμαστε την H_0 όταν $X \in [\chi^2_{n-1,1-a/2}, \chi^2_{n-1,a/2}]$, αλλιώς την απορρίπτουμε.

Έλεγχος για την διακύμανση σ^2 κανονικού πληθυσμού

Ανάλογα προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον ακόλουθο πίνακα:

H_1	χωρίο απόρριψης	p-value
$\sigma < \sigma_0$	$R = \{X < \chi^2_{1-a,n-1}\}$	$F_{n-1}(X)$
$\sigma > \sigma_0$	$R = \{X > \chi^2_{a,n-1}\}$	$1 - F_{n-1}(X)$
$\sigma \neq \sigma_0$	$R = \{X < \chi^2_{1-a/2,n-1}\}$ $\cup \{X > \chi^2_{a/2,n-1}\}$	$2\min\{F_{n-1}(X), 1 - F_{n-1}(X)\}$

Έλεγχος για την διακύμανση σ^2 κανονικού πληθυσμού

Παράδειγμα

Αν δείγμα μεγέθους $n = 20$ έχει δειγματική τυπική απόκλιση $S = 0.12$, να ελεγχθεί η υπόθεση $H_0 : \sigma \leq 0.1$ έναντι της $H_1 : \sigma > 0.1$.

```
import numpy as np
import scipy.stats as st
n, sigma0, S, a = 20, 0.1, 0.12, 0.05
H1 = ['sigma > sigma0', 'sigma != sigma0']
X = (n-1)*S**2/sigma0**2
pval = 1-st.chi2.cdf(X, n-1) #greater
pp = st.chi2.ppf(1-a, n-1)
print("H1:%s, a = %s, X = %f, x_{a} = %f, p-value = %f"%(H1
    [0], a, X, pp, pval))
```

Output:

```
H1:sigma > sigma0, a = 0.05, X = 27.360000, x_{a} =
    30.143527, p-value = 0.096543
```

Συμπεραίνουμε ότι η $H_0 : \sigma \leq 0.1$ μπορεί να απορριφθεί με επίπεδο εμπιστοσύνης 90% (όχι όμως και όταν ο έλεγχος είναι δίπλευρος).

Έλεγχος για τον λόγο σ_1/σ_2 των διακυμάνσεων δύο κανονικών πληθυσμών

Θεωρούμε 2 ανεξάρτητα τυχαία δείγματα $(X_1, \ldots, X_{n_1}) \sim N(\mu_1, \sigma_1^2)$ και $(Y_1, \ldots, Y_{n_2}) \sim N(\mu_2, \sigma_2^2)$. Αν S_1^2 και S_2^2 είναι οι αντίστοιχες δειγματικές διακυμάνσεις, υπό τη μηδενική υπόθεση $H_0 : \sigma_1 = \sigma_2$, η στατιστική συνάρτηση

$$X = \frac{S_1^2}{S_2^2}$$

ακολουθεί ως γνωστό την κατανομή Fisher F_{n_1-1,n_2-1}, επομένως, για $a \in (0,1)$, είναι

$$P(F_{1-a/2,n_1-1,n_2-1} \leq X \leq F_{a/2,n_1-1,n_2-1}) = 1 - a$$

Επομένως, για έναν δίπλευρο έλεγχο με επίπεδο εμπιστοσύνης a, αποδεχόμαστε την H_0 όταν $F \in [F_{1-a/2,n_1-1,n_2-1}, F_{a/2,n_1-1,n_2-1}]$, αλλιώς την απορρίπτουμε.

Έλεγχος για τον λόγο σ_1/σ_2 των διακυμάνσεων δύο κανονικών πληθυσμών

Ανάλογα προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον ακόλουθο πίνακα:

H_1	χωρίο απόρριψης	p-value
$\sigma_1 < \sigma_2$	$R = \{X < F_{1-a}\}$	$F(X)$
$\sigma_1 > \sigma_2$	$R = \{X > F_a\}$	$1 - F(X)$
$\sigma_1 \neq \sigma_2$	$R = \{X < F_{1-a/2}\} \cup \{X > F_{a/2}\}$	$2\min\{F(X), 1 - F(X)\}$

όπου $X = \dfrac{S_1^2}{S_2^2}$, F η CDF της F_{n_1-1,n_2-1} και F_a το άνω a-ποσοστιαίο σημείο αυτής.

Έλεγχος για τον λόγο σ_1/σ_2 των διακυμάνσεων δύο κανονικών πληθυσμών

Παράδειγμα

Αν δύο ανεξάρτητα δείγματα μεγέθους $n_1 = 10$ και $n_2 = 12$ έχουν δειγματική διακύμανση $S_1^2 = 0.14$ και $S_2^2 = 0.28$ αντίστοιχα, να ελεγχθεί η υπόθεση $H_0 : \sigma_1 = \sigma_2$ έναντι της $H_1 : \sigma_1 \neq \sigma_2$, σε επίπεδο σημαντικότητας $a = 0.05$.

Λύση

Εφαρμόζουμε δίπλευρο έλεγχο. Για $a = 0.05$, βρίσκουμε $\ell = F_{1-a/2,n_1-1,n_2-1} = 0.2556$ και $u = F_{a/2,n_1-1,n_2-1} = 3.5879$. Επειδή $X = S_1^2/S_2^2 = 0.5 \in [\ell, u]$, η H_0 δεν μπορεί να απορριφθεί.

Ο κώδικας που ακολουθεί, υπολογίζει τα παραπάνω μεγέθη, καθώς και την p-value του ελέγχου.

Έλεγχος για τον λόγο σ_1/σ_2 των διακυμάνσεων δύο κανονικών πληθυσμών

```
import numpy as np
import scipy.stats as st

n1, n2, Svar1, Svar2, a = 10, 12, 0.14, 0.28, 0.05
H1 = 'sigma1 != sigma2'

F = Svar1/Svar2
pval = 2*np.min([st.f.cdf(F, n1-1,n2-1), st.f.sf(F, n1-1,n2
    -1)]) #two-sided
lpp = st.f.ppf(a/2, n1-1,n2-1)
upp = st.f.ppf(1-a/2, n1-1,n2-1)
print("H1:%s, a = %s, F = %f, lpp = %f, upp = %f, p-value =
    %f"%(H1, a, F, lpp, upp, pval))
```

Output:

```
H1:sigma1 != sigma2, a = 0.05, F = 0.500000, lpp = 0.255619,
     upp = 3.587899, p-value = 0.307519
```

Έλεγχος καλής προσαρμογής (goodness of fit test)

Έστω τυχαίο δείγμα $(X_1, \ldots, X_n)$ από άγνωστη διακριτή κατανομή, και έστω ότι τα X_i παίρνουν τιμές στο $[k]$. Κάποιος ισχυρίζεται ότι οι τιμές προέρχονται από μια συγκεκριμένη διακριτή κατανομή με $P(X = i) = p_i$, για δεδομένα p_i. Θέλουμε να ελέγξουμε την υπόθεση

$$H_0 : \forall i \in [k], P(X = i) = p_i, \qquad \text{έναντι της } H_1 : \exists i \in [k], P(X = i) \neq p_i,$$

όπου το X αντιπροσωπεύει οποιοδήποτε από τα X_i και $(p_1, \ldots, p_k)$ η υποτιθέμενη συνάρτηση πιθανότητας της X.
Θέτουμε $N_i = |\{j \in [n] : X_j = i\}|$ το πλήθος των εμφανίσεων της τιμής i στο δείγμα.
Υπό την H_0, η ΤΜ N_i ακολουθεί διωνυμική κατανομή με παραμέτρους n και p_i, οπότε $E(N_i) = np_i$.

Έλεγχος καλής προσαρμογής (goodness of fit test)

Θεωρούμε τη στατιστική συνάρτηση

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{N_i^2}{np_i} - n$$

η οποία αντιπροσωπεύει την ''απόσταση'' των τιμών που παρατηρήθηκαν από τις αναμενόμενες σύμφωνα με την H_0. Δηλαδή, όσο μεγαλύτερη τιμή πάρει η T, τόσο ισχυρότερη ένδειξη έχουμε ότι η H_0 δεν ισχύει. Αποδεικνύεται ότι $T \to \chi^2_{k-1}$ καθώς $n \to \infty$, οπότε ένας έλεγχος με επίπεδο σημαντικότητας a είναι να απορρίψουμε την H_0 αν $T > \chi^2_{a,k-1}$, αλλιώς την αποδεχόμαστε. Η p-value του ελέγχου είναι η

$$\text{p-value} = 1 - F_{k-1}(T)$$

όπου F_{k-1} η αθροιστική συνάρτηση κατανομής της χ^2_{k-1}.
Η προσέγγιση θεωρείται ικανοποιητική όταν $N_i \geq 5$, για κάθε i.

Έλεγχος καλής προσαρμογής (goodness of fit test)

Παράδειγμα

Ένας κατασκευαστής ισχυρίζεται ότι τα προϊόντα του είναι ποιότητας 1, 2, 3, 4 ή 5 (μεγαλύτερος αριθμός = χαμηλότερη ποιότητα) με πιθανότητες $(p_1, p_2, p_3, p_4, p_5) = (0.15, 0.25, 0.35, 0.20, 0.05)$. Αν σε 30 προϊόντα μετρήσαμε αντίστοιχα πλήθη εμφανίσεων $(N_1, N_2, N_3, N_4, N_5) = (3, 6, 9, 7, 5)$, μπορούμε να απορρίψουμε τον ισχυρισμό σε επίπεδο εμπιστοσύνης 95%;

Λύση

Έχουμε $k = 5$ διαφορετικές ποιότητες - δυνατές τιμές της τυχαίας μεταβλητής. Υπολογίζουμε την στατιστική συνάρτηση

$$T = \sum_{i=1}^{5} \frac{(N_i - np_i)^2}{np_i} = \sum_{i=1}^{5} \frac{N_i^2}{np_i} - n = 9.35$$

και στη συνέχεια την p-value $= 1 - F_4(T) = 0.053$, όπου F_4 η συνάρτηση κατανομής της χ^2_4. Επομένως, δεν μπορούμε να απορρίψουμε την H_0 (οριακά).

Έλεγχος καλής προσαρμογής (goodness of fit test)

```
import numpy as np
import scipy.stats as st

a = 0.05
f_obs = np.array([3,6,9,7,5])
n, k = f_obs.sum(), len(f_obs)
f_exp = n*np.array([0.15, 0.25, 0.35, 0.20, 0.05])
T, pval = st.chisquare(f_obs, f_exp) #use built-in function
print("T = %s, p-value = %s"%(T, pval))

T = np.sum((f_obs - f_exp)**2/ f_exp) #use formulas from
    theory
print("T = %s, chi2_{a,k-1} = %s, p-value = %s"%(T,st.chi2.
    isf(a,k-1),st.chi2.sf(T,k-1)))
```

Output:

```
T = 9.347619047619046, p-value = 0.052974280436963686
T = 9.347619047619046, chi2_{a,k-1} = 9.487729036781158, p-
    value = 0.052974280436963686
```

Έλεγχος καλής προσαρμογής (goodness of fit test)

Στην περίπτωση που η ΤΜ παίρνει άπειρες τιμές, μπορούμε να διαμερίσουμε το σύνολο τιμών σε k υποσύνολα S_i, $i \in [k]$, οπότε N_i είναι το πλήθος των εμφανίσεων τιμών του S_i στο δείγμα. Αν η H_0 δεν καθορίζει πλήρως τις παραμέτρους της κατανομής, τότε χρησιμοποιούμε στη θέση τους τις εκτιμήσεις αυτών από το δείγμα.

Παράδειγμα

Έστω ότι ο αριθμός ατυχημάτων ανά ημέρα για 30 ημέρες από τον ακόλουθο πίνακα:

$$8, 0, 0, 1, 3, 4, 0, 2, 12, 5, 1, 8, 0, 2, 0, 1, 9, 3, 4, 5, 3, 3, 4, 7, 4, 0, 1, 2, 1, 2$$

Να ελεγχθεί η υπόθεση ότι η κατανομή των ατυχημάτων είναι Poisson.

Λύση

Εκτιμάμε την παράμετρο της Poisson: $\hat{\lambda} = 3.167$ (πλήθος ατυχημάτων/πλήθος ημερών). Έστω ΤΜ $X \sim P(\hat{\lambda})$. Διαμερίζουμε το σύνολο τιμών S_X σε 5 υποσύνολα:

$$p_0 = P(X = 0), \quad p_1 = P(X = 1), \quad p_2 = P(X = 2) + P(X = 3),$$

$$p_3 = P(X = 4) + P(X = 5), \quad p_4 = P(X > 5)$$

Υπολογίζουμε τα N_i, $0 \leq i \leq 4$. και στη συνέχεια τη στατιστική συνάρτηση $T = \sum_{i=0}^{4} \frac{(N_i - np_i)^2}{np_i}$ και την p-value $= 0.000311$. Τελικά απορρίπτουμε την H_0 με μεγάλη εμπιστοσύνη.

Έλεγχος καλής προσαρμογής (goodness of fit test)

```
import numpy as np
import scipy.stats as st
a = 0.05
obs=np.array([8,0,0,1,3,4,0,2,12,5,1,8,0,2,0,
              1,9,3,4,5,3,3,4,7,4,0,1,2,1,2])
lbar = np.mean(obs) #estimate lambda
f_obs = np.bincount(obs) #count occurrences
n, k = np.sum(f_obs), len(f_obs)
f_obs2 = [f_obs[0], f_obs[1], f_obs[2]+f_obs[3], f_obs[4]+
    f_obs[5], np.sum(f_obs[6:-1])]
f_exp = n*st.poisson.pmf(np.arange(0,k), lbar)
f_exp2 = [f_exp[0], f_exp[1], f_exp[2]+f_exp[3], f_exp[4]+
    f_exp[5], np.sum(f_exp[6:-1])]
print("N_i's: %s\nlbar = %f"%(f_obs,lbar))
T, pval = st.chisquare(f_obs2, f_exp2)
print("T = %s, p-value = %s"%(T, pval))
```

Output:

```
N_i's: [6 5 4 4 4 2 0 1 2 1 0 0 1]
lbar = 3.166667
T = 21.040126868212, p-value = 0.0003109205102098774
```

Έλεγχος ανεξαρτησίας

Θεωρούμε ότι κάθε άτομο ενός πληθυσμού χαρακτηρίζεται από ένα ζεύγος χαρακτηριστικών (ΤΜ) (X, Y) με τιμές στο $[r] \times [c]$. Κάθε άτομο έχει μια τιμή σε κάθε χαρακτηριστικό με κάποια άγνωστη πιθανότητα. Συμβολίζουμε με

$$P_{i,j} = P(X = i, Y = j), \qquad p_i = P(X = i), \qquad q_j = P(Y = j)$$

τις αντίστοιχες πιθανότητες. Η μηδενική υπόθεση είναι ότι τα χαρακτηριστικά X, Y είναι ανεξάρτητα, δηλαδή

$$H_0 : \forall (i,j), P_{i,j} = p_i q_j, \qquad H_1 : \exists (i,j), P_{i,j} \neq p_i q_j.$$

Σε ένα τυχαίο δείγμα μεγέθους n βρέθηκαν $N_{i,j}$ σε πλήθος άτομα με $(X, Y) = (i, j)$. Επομένως

$$E(N_{i,j}) = nP_{i,j} = np_i q_j.$$

Η τελευταία ισότητα ισχύει δεδομένης της H_0.

Έλεγχος ανεξαρτησίας

Θέτουμε

$$N_i = \sum_{j=1}^{c} N_{i,j}, \qquad M_j = \sum_{i=1}^{r} N_{i,j}$$

οπότε προκύπτουν οι εκτιμήτριες των p_i, q_j

$$\hat{p}_i = \frac{N_i}{n}, \qquad \hat{q}_j = \frac{M_j}{n}.$$

Κατόπιν τούτων, προκύπτει ότι η στατιστική συνάρτηση

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{i,j} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{i,j}^2}{n\hat{p}_i\hat{q}_j} - n$$

για την οποία αποδεικνύεται ότι $T \to \chi^2_{(r-1)(c-1)}$, όταν $n \to \infty$.
Επομένως, η p-value του ελέγχου είναι η p-value $= 1 - F(T)$, όπου F η συνάρτηση κατανομής της χ^2 με $(r-1)(c-1)$ βαθμούς ελευθερίας.

Παράδειγμα

Δίνεται ο επόμενος πίνακας για την προτίμηση πολιτικού κόμματος ενός δείγματος 300 ψηφοφόρων.

$i \backslash j$	1	2	3	Σύνολο (N_i)
Γυναίκες	68	56	32	156
Άνδρες	52	72	20	144
Σύνολο (M_j)	120	128	52	$n = 300$

Να ελεγχθεί η υπόθεση ότι η επιλογή κόμματος είναι ανεξάρτητη του φύλου.

Λύση

Θέτουμε X το φύλο του ατόμου και Y το κόμμα που υποστηρίζει. Υπολογίζουμε p-value $= 0.04$, σύμφωνα με τον επόμενο κώδικα, και τελικά απορρίπτουμε την H_0 με επίπεδο εμπιστοσύνης 95%.

Έλεγχος ανεξαρτησίας

```
import numpy as np
import scipy.stats as st
obs = [[68, 56, 32],
       [52, 72, 20]]
T, pval, df, e = st.chi2_contingency(obs)
print("Expected:\n", e)
print("\nT = %s, p-value = %s, df = %s"%(T, pval, df))
```

Output:

```
Expected:
 [[62.4  66.56 27.04]
 [57.6  61.44 24.96]]

T = 6.432856673241291, p-value = 0.04009801943167609, df = 2
```

Έλεγχος ανεξαρτησίας

Στα παραπάνω, μπορούμε να θεωρήσουμε ότι μια από τις 2 ΤΜ, π.χ. η Y, αντιπροσωπεύει τον αριθμό δείγματος. Δηλαδή στην περίπτωση που έχουμε c δείγματα από c πληθυσμούς, μπορούμε να ελέγξουμε όπως παραπάνω την υπόθεση ότι ένα χαρακτηριστικό X ακολουθεί την ίδια κατανομή σε κάθε πληθυσμό, δηλαδή είναι ανεξάρτητο επιλογής πληθυσμού: $H_0 : \forall j, P_1(X = j) = P_2(X = j) = \cdots = P_c(X = j)$.

Παράδειγμα

Σε 4 χώρες επιλέχθηκε δείγμα 500 γυναικών και ρωτήθηκαν αν έχουν υποστεί σεξουαλική παρενόχληση στον χώρο εργασίας.

$i \backslash j$	AU	DE	JP	US	Σύνολο (N_i)
Ναι	28	30	58	55	171
Όχι	472	470	442	445	1829
Σύνολο (M_j)	500	500	500	500	$n = 2000$

Να ελεγχθεί η υπόθεση ότι τα ποσοστά παρενόχλησης είναι ίδια σε κάθε χώρα.

Έλεγχος ανεξαρτησίας

```
import numpy as np
import scipy.stats as st
obs = [[28, 30, 58, 55],
       [472, 470, 442, 445]]
T, pval, df, e = st.chi2_contingency(obs)

print("Observed:\n", obs)
print("Expected:\n", e)
print("\nT = %s, p-value = %s, df = %s"%(T, pval, df))
```

Output:

```
Expected:
 [[ 42.75  42.75  42.75  42.75]
 [457.25 457.25 457.25 457.25]]

T = 19.510229921441113, p-value = 0.00021440518558965625, df
    = 3
```

Απορρίπτουμε την H_0 με επίπεδο εμπιστοσύνης 99%.