Στατιστική Επαγωγή

Στα πλαίσια του μαθήματος αυτού μας ενδιαφέρει η στατιστική επαγωγή για τον μέσο ενός πληθυσμού. Δηλαδή, έχοντας ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από μια τυχαία μεταβλητή X με μέσο $E(X) = \mu$ μας ενδιαφέρει να βγάλουμε συμπεράσματα για το άγνωστο μ. Η στατιστική επαγωγή για το μ μπορεί να πάρει τις ακόλουθες μορφές:

- Σημειακή Εκτίμηση του μ,
- Εκτίμηση Διαστήματος για το μ,
- Έλεγχος Υποθέσεων για το μ.

Το πρώτο βήμα για την στατιστική επαγωγή είναι η εύρεση κατάλληλου εκτιμητή (με τις επιθυμητές ιδιότητες) για το μ. Δεδομένου ενός παρατηρούμενου δείγματος που στην πράξη θα έχουμε, ο εκτιμητής θα μας δώσει όπως έχουμε δει την (σημειακή) εκτίμηση για το μ. Επομένως, την πρώτη μορφή της στατιστικής επαγωγής την έχουμε ήδη εξετάσει χρησιμοποιώντας τον δειγματικό μέσο $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ως εκτιμητή του μ.

Ακολούθως, αν μας ενδιαφέρει να αποκτήσουμε καλύτερη εικόνα για το πόσο κοντά είναι η εκτίμηση στην παράμετρο που εκτιμάει, πρέπει να εξάγουμε την κατανομή του εκτιμητή. Η κατανομή του εκτιμητή θα εξαρτάται από άγνωστες παραμέτρους, μεταξύ αυτών και το μ. Στα πλαίσια του μαθήματος θα μας ενδιαφέρει μόνο το μ οπότε οποιεσδήποτε άλλες παράμετροι είτε θα μας είναι γνωστές ή θα πρέπει να τις εκτιμήσουμε και αυτές. Συγκεκριμένα έχουμε εξετάσει τις ακόλουθες περιπτώσεις όπου η κατανομή του εκτιμητή, μετά από απλούς μετασχηματισμούς, θα ακολουθεί κάποια γνωστή κατανομή:

- Κανονικός πληθυσμός και γνωστή διακύμανση: $Z := \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- Κανονικός πληθυσμός και άγνωστη διακύμανση: $T := \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, όπου $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ η δειγματική διακύμανση.
- (Μη κανονικός Πληθυσμός) Άγνωστη κατανομή δειγματικού μέσου αλλά μεγάλο δείγμα: προσεγγιστικά $Z := \frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} \sim N(0,1)$ όπου $\hat{\sigma}$ οποιοσδήποτε συνεπής εκτιμητής του σ, δηλ. $\hat{\sigma} \xrightarrow{p} \sigma$ ή ισοδύναμα $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

Παρατηρήστε ότι σε κάθε περίπτωση στους λόγους που σχηματίζουμε για το Z και το T μόνο το μ θα είναι άγνωστο. Επομένως, μπορούμε να προσδιορίσουμε την πιθανότητα ο δειγματικός μέσος να είναι κοντά στο μ, και κατ' επέκταση διάστημα με κέντρο το μ στο οποίο ο δειγματικός μέσος θα παίρνει τιμές με κάποια πιθανότητα που εμείς επιλέγουμε. Συγκεκριμένα, σε καθεμία από τις παραπάνω περιπτώσεις θα έχουμε:

- $P\left(\mu - z_{a/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{a/2}\frac{\sigma}{\sqrt{n}}\right) = 1-a$, όπου $z_{a/2}$ τέτοιο ώστε $P(Z > z_{a/2}) = \frac{a}{2}$ για $0 \leq a \leq 1$ και $Z \sim N(0,1)$.
- $P\left(\mu - t_{n-1,a/2}\frac{S}{\sqrt{n}} \leq \bar{X} \leq \mu + t_{n-1,a/2}\frac{S}{\sqrt{n}}\right) = 1-a$, όπου $t_{n-1,a/2}$ τέτοια ώστε $P(T > t_{n-1,a/2}) = \frac{a}{2}$ για $0 \leq a \leq 1$ και $T \sim t_{n-1}$.
- $P\left(\mu - z_{a/2}\frac{\hat{\sigma}}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{a/2}\frac{\hat{\sigma}}{\sqrt{n}}\right) \simeq 1-a$, όπου $z_{a/2}$ τέτοιο ώστε $P(Z > z_{a/2}) = \frac{a}{2}$ για $0 \leq a \leq 1$ και $Z \sim N(0,1)$.

Το a λέγεται επίπεδο στατιστικής σημαντικότητας και το $1-a$ επίπεδο εμπιστοσύνης και επιλογή του έχει μεγάλη σημασία για να βγάλουμε χρήσιμα συμπεράσματα σχετικά με το άγνωστο μ. Συγκεκριμένα σημειώστε ότι αν επέλεγα $\alpha = 0$ τότε απλά θα είχα ότι $P\left(-\infty < \bar{X} < \infty\right) = 1$ (καθώς η κανονική

κατανομή και η Student t παίρνουν τιμές σε όλους τους πραγματικούς) το οποίο δεν θα ήταν καθόλου πληροφοριακό σχετικά με το μ. Επίσης, παρατηρήστε ότι όσο πιο μικρό επιλέξω το a τόσο μεγαλύτερες θα είναι οι τιμές $z_{a/2}$ και $t_{n-1,a/2}$ οπότε και τα αντίστοιχα διαστήματα θα είναι μεγαλύτερα.

Οπότε το πρώτο βήμα είναι να επιλέξουμε μια τιμή για το a η οποία να είναι αρκετά μικρή (συνήθως επιλέγουμε $a = 0.01$ ή $a = 0.05$ ή $a = 0.10$) η οποία υποδηλώνει την πιθανότητα να παρατηρήσουμε μια ακραία εκτίμηση για το $\bar{X}$. Ακολούθως, θεωρώντας ότι η εκτίμηση που παρατηρούμε θα βρίσκεται στο διάστημα που έχουμε κατασκευάσει, μπορούμε να βγάλουμε συμπέρασμα για τις λογικές τιμές που μπορεί να πάρει το μ ώστε διαστήματα με κέντρο αυτές τις τιμές να περιλαμβάνουν και την εκτίμηση που πήραμε (για το δεδομένο a).

Διάστημα Εμπιστοσύνης

Συγκεκριμένα δεδομένου ότι έχουμε παρατηρήσει κάποιο δείγμα, κάθε εκτιμητής θα μας δώσει μια εκτίμηση για το μ ή/και το σ^2 όταν αυτό είναι άγνωστο. Δεδομένων των εκτιμήσεων μπορώ να βγάλω συμπέρασμα για το μ κατασκευάζοντας ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης, αφού έχω επιλέξει ένα επίπεδο στατιστικής σημαντικότητας a. Για παράδειγμα για την τελευταία περίπτωση όπου έχω μη κανονικό πληθυσμό αλλά μεγάλο δείγμα αν λύσω ως προς μ (αντί για $\bar{X}$ που έλυνα πριν)

$$\begin{aligned} -z_{a/2} &\leq \frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{n}} \leq z_{a/2} \\ -z_{a/2}\hat{\sigma}/\sqrt{n} &\leq \bar{x}-\mu \leq z_{a/2}\hat{\sigma}/\sqrt{n} \\ -\bar{x}-z_{a/2}\hat{\sigma}/\sqrt{n} &\leq -\mu \leq -\bar{x}+z_{a/2}\hat{\sigma}/\sqrt{n} \\ \bar{x}-z_{a/2}\hat{\sigma}/\sqrt{n} &\leq \mu \leq \bar{x}+z_{a/2}\hat{\sigma}/\sqrt{n} \end{aligned}$$

όπου $\bar{x}$ και $\hat{\sigma}$ θα παίρνουν κάποιες αριθμητικές τιμές. Άρα το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το μ θα είναι το $[\bar{x}-z_{a/2}\hat{\sigma}/\sqrt{n}, \bar{x}+z_{a/2}\hat{\sigma}/\sqrt{n}]$.

Συνοψίζοντας στις 3 περιπτώσεις που με ενδιαφέρουν, το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το μ θα είναι

- Κανονικός πληθυσμός και γνωστή διακύμανση: $\left[\bar{x}-z_{a/2}\frac{\sigma}{\sqrt{n}}, \bar{x}+z_{a/2}\frac{\sigma}{\sqrt{n}}\right]$
- Κανονικός πληθυσμός και άγνωστη διακύμανση: $\left[\bar{x}-t_{n-1,a/2}\frac{s}{\sqrt{n}}, \bar{x}+z_{a/2}\frac{s}{\sqrt{n}}\right]$
- (Μη κανονικός Πληθυσμός) Άγνωστη κατανομή δειγματικού μέσου αλλά μεγάλο δείγμα: $\left[\bar{x}-z_{a/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x}+z_{a/2}\frac{\hat{\sigma}}{\sqrt{n}}\right]$

Εφαρμογή: Bernoulli

Αν έχω τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από την τυχαία μεταβλητή $X \sim Bernoulli(p)$, όπου το p είναι άγνωστο. Όπως έχουμε δει ισχύει $E(X) = p$ και $Var(X) = p(1-p)$. Ένας εκτιμητής για το p είναι ο εκτιμητής των ροπών:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Η κατανομή του εκτιμητή $\hat{p}$ δεν είναι κάποια γνωστή κατανομή[1] αλλά αν έχω μεγάλο δείγμα θα έχω προσεγγιστικά από το Κεντρικό Οριακό Θεώρημα ότι

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$$

[1] Ξέρω ότι $\sum_{i=1}^{n} X_i \sim Bin(n,p)$) αλλά εδώ πολλαπλασιάζω με $\frac{1}{n}$.

όπου $\hat{p}(1-\hat{p})/n$ είναι ένας εκτιμητής για την διακύμανση του $\hat{p}$ η οποία είναι $Var(\hat{p}) = p(1-p)/n$.

Δηλαδή, ισοδύναμα, θα έχω ότι προσεγγιστικά

$$\hat{p} \sim N\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

Επομένως, για να φτιάξω το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το p θα λύσω την ακόλουθη ανισότητα ως προς p:

$$-z_{a/2} \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{a/2}$$

οπότε το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το p θα είναι

$$\left[\hat{p} - z_{a/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{a/2}\sqrt{\hat{p}(1-\hat{p})/n}\right]$$

Για παράδειγμα αν έχω $n = 50$ και $\hat{p} = 0.54$ τότε

- Το 99% διάστημα εμπιστοσύνης για το p θα είναι:

$$\begin{aligned}[0.54 - 2.57\sqrt{0.54(1-0.54)/50}, 0.54 + 2.57\sqrt{0.54(1-0.54)/50}] \\ \simeq [0.36, 0.72]\end{aligned}$$

  καθώς $P(Z < -2.57) = P(Z > 2.57) = 0.01/2 = 0.005$, για $Z \sim N(0,1)$.

- Το 95% διάστημα εμπιστοσύνης για το p θα είναι:

$$\begin{aligned}[0.54 - 1.96\sqrt{0.54(1-0.54)/50}, 0.54 + 1.96\sqrt{0.54(1-0.54)/50}] \\ \simeq [0.40, 0.68]\end{aligned}$$

  καθώς $P(Z < -1.96) = P(Z > 1.96) = 0.05/2 = 0.025$, για $Z \sim N(0,1)$.

- Το 90% διάστημα εμπιστοσύνης για το p θα είναι:

$$\begin{aligned}[0.54 - 1.65\sqrt{0.54(1-0.54)/50}, 0.54 + 1.65\sqrt{0.54(1-0.54)/50}] \\ \simeq [0.42, 0.66]\end{aligned}$$

  καθώς $P(Z < -1.65) = P(Z > 1.65) = 0.10/2 = 0.05$, για $Z \sim N(0,1)$.

Έλεγχοι Υποθέσεων

Στους ελέγχους υποθέσεων αντί να βρίσκουμε όλες τις δυνατές εύλογες τιμές που μπορεί να πάρει το άγνωστο μ με βάση την εκτίμηση που παρατηρούμε, ελέγχουμε μια συγκεκριμένη τιμή για το μ. Συγκεκριμένα μας ενδιαφέρει να ελέγξουμε μια συγκεκριμένη υπόθεση έναντι μιας εναλλακτικής υπόθεσης. Ο στατιστικός έλεγχος υπόθεσης μπορεί να έχει μία από τις παρακάτω 3 μορφές (για κάποιο $\mu_0 \in \mathbb{R}$):

1. $H_0 : \mu = \mu_0$
   $H_1 : \mu \neq \mu_0$
2. $H_0 : \mu \geq \mu_0$ ή $H_0 : \mu = \mu_0$
   $H_1 : \mu < \mu_0$
3. $H_0 : \mu \leq \mu_0$ ή $H_0 : \mu = \mu_0$
   $H_1 : \mu > \mu_0$

όπου H_0 ονομάζεται "μηδενική υπόθεση" και H_1 ονομάζεται "εναλλακτική υπόθεση".

Για να κάνουμε έναν έλεγχο υποθέσεων χρειαζόμαστε μια "στατιστική ελέγχου" η οποία δίνει πάντα μια αριθμητική τιμή αν θεωρήσουμε ότι η H_0 ισχύει, δηλαδή αν ισχύει ότι $\mu = \mu_0$. Θα έχουμε πάλι τις εξής 3 περιπτώσεις:

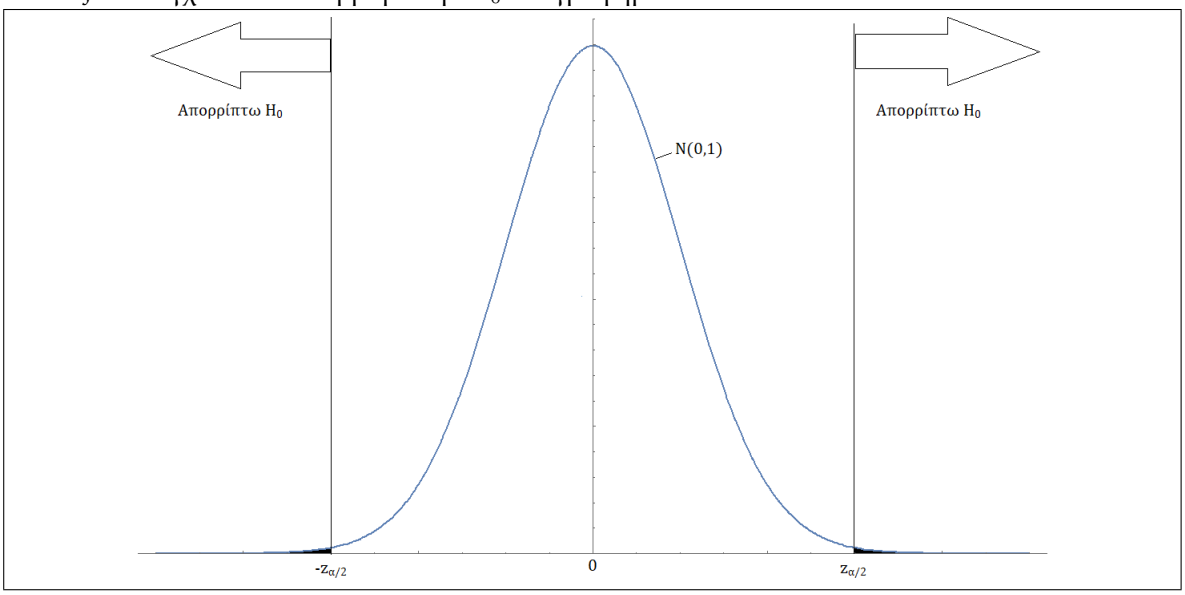
- Κανονικός πληθυσμός και γνωστή διακύμανση: $Z := \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$
- Κανονικός πληθυσμός και άγνωστη διακύμανση: $T := \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$, όπου $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ η δειγματική διακύμανση.
- (Μη κανονικός Πληθυσμός) Άγνωστη κατανομή δειγματικού μέσου αλλά μεγάλο δείγμα: προσεγγιστικά $Z := \frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$ όπου $\hat{\sigma}$ οποιοσδήποτε συνεπής εκτιμητής του σ.

όπου ο συμβολισμός $\overset{H_0}{\sim}$ υποδηλώνει "υπό την μηδενική υπόθεση".

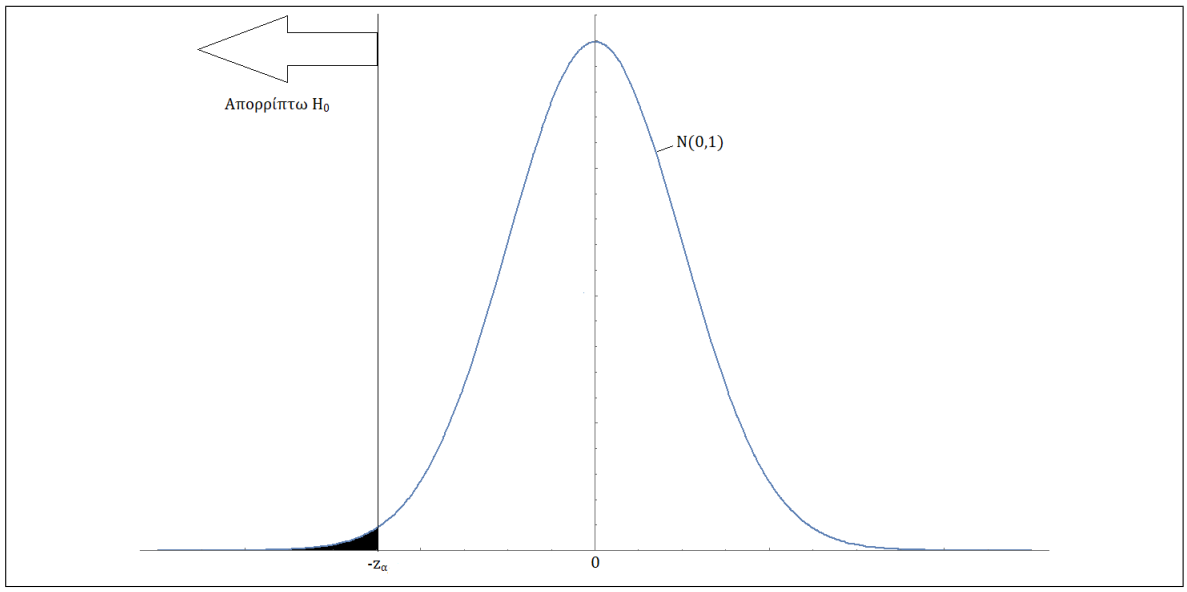
Παρακάτω θα δούμε πώς κάνουμε κάθε ένα στατιστικό έλεγχο υποθέσεων σε κάθε περίπτωση για κάποιο επίπεδο στατιστικής σημαντικότητας a.

Κανονικός πληθυσμός, γνωστή διακύμανση

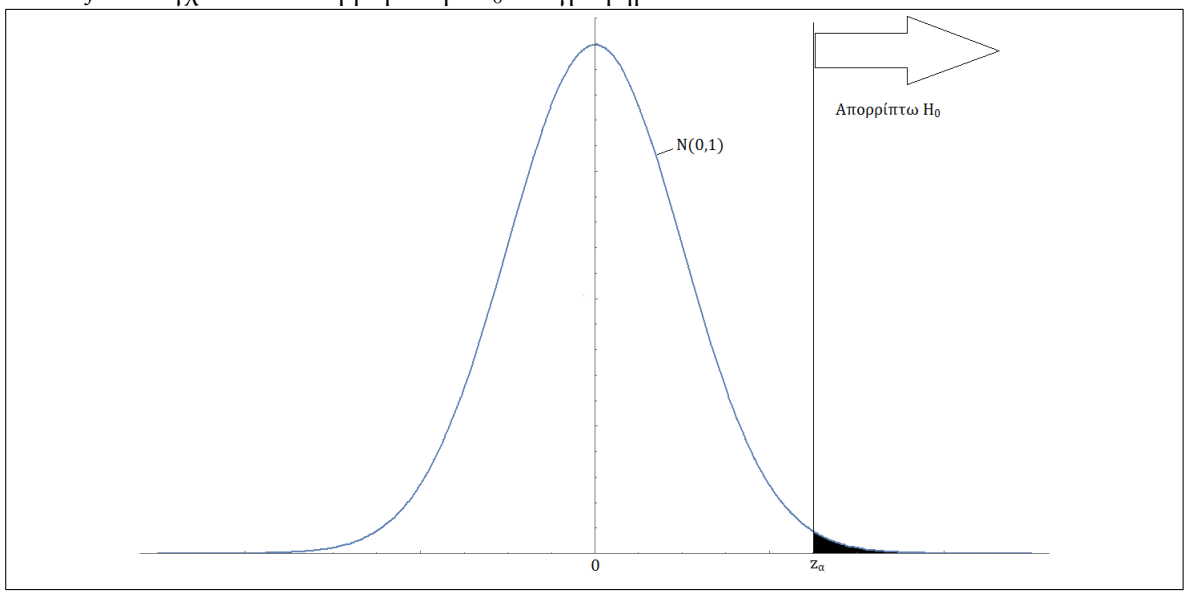
- $H_0 : \mu = \mu_0$
  $H_1 : \mu \neq \mu_0$
  Απορρίπτω την H_0 αν $\left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > z_{a/2} \Leftrightarrow \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > z_{a/2}$ ή $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < -z_{a/2}$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0. Σε γράφημα:



- $H_0 : \mu \geq \mu_0$ ή $H_0 : \mu = \mu_0$
  $H_1 : \mu < \mu_0$
  Απορρίπτω την H_0 αν $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < -z_a$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0. Σε γράφημα:

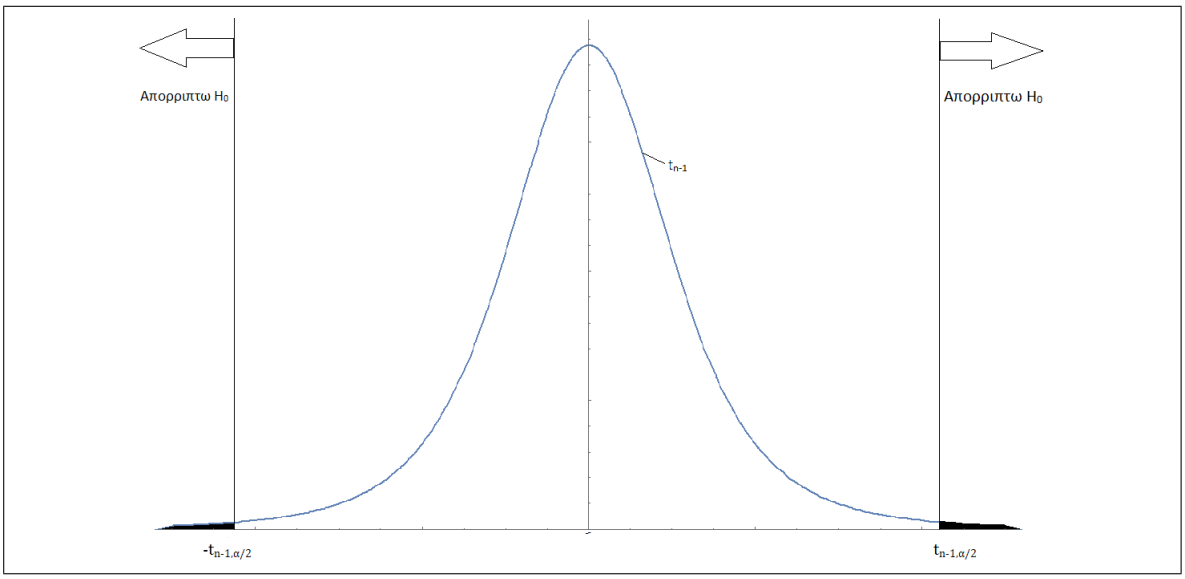


- $H_0 : \mu \leq \mu_0$ ή $H_0 : \mu = \mu_0$
  $H_1 : \mu > \mu_0$
  Απορρίπτω την H_0 αν $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > z_a$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0. Σε γράφημα:



Κανονικός πληθυσμός, άγνωστη διακύμανση

- $H_0 : \mu = \mu_0$
  $H_1 : \mu \neq \mu_0$
  Απορρίπτω την H_0 αν $\left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > t_{n-1,a/2} \Leftrightarrow \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > t_{n-1,a/2}$ ή $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < -t_{n-1,a/2}$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0. Σε γράφημα:



- $H_0 : \mu \geq \mu_0$ ή $H_0 : \mu = \mu_0$
  $H_1 : \mu < \mu_0$
  Απορρίπτω την H_0 αν $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} < -t_{n-1,a}$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0.

- $H_0 : \mu \leq \mu_0$ ή $H_0 : \mu = \mu_0$
  $H_1 : \mu > \mu_0$
  Απορρίπτω την H_0 αν $\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > t_{n-1,a}$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0.

Μη Κανονικός πληθυσμός

- $H_0 : \mu = \mu_0$
  $H_1 : \mu \neq \mu_0$
  Απορρίπτω την H_0 αν $\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > z_{a/2} \Leftrightarrow \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} > z_{a/2}$ ή $\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < -z_{a/2}$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0.

- $H_0 : \mu \geq \mu_0$ ή $H_0 : \mu = \mu_0$
  $H_1 : \mu < \mu_0$
  Απορρίπτω την H_0 αν $\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < -z_a$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0.

- $H_0 : \mu \leq \mu_0$ ή $H_0 : \mu = \mu_0$
  $H_1 : \mu > \mu_0$
  Απορρίπτω την H_0 αν $\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} > z_a$.
  Αλλιώς αποτυγχάνω να απορρίψω την H_0.