# 6B Hypoteesintestaus

Tämä on viimeinen harjoituspaketti, ja tässä on vain tuntitehtäviä. Tehtäviin kannattaa valmistautua tutustumalla luentoon 6A ja luentomonisteen lukuun 11.

## Tuntitehtävät

**6B1** (Naulatehdas) Tehdas valmistaa nauloja, joiden keskipituuden tulisi olla 10.00 cm. Valmistusprosessin epätarkkuuksien vuoksi naulojen pituudet ovat satunnaisia, normaalijakaumalla jolla on tuntematon odotusarvo $\mu$ ja tuntematon keskihajonta $\sigma$. Laadunvalvontaa varten mitattiin 120 naulaa. Mitattujen naulojen pituuden keskiarvo oli 10.08 cm ja keskihajonta 0.40 cm. Datajoukkoa pidetään "suurena datajoukkona", eli datan keskihajontaa käytetään sellaisenaan datalähteen keskihajonnan $\sigma$ estimaattina eikä siihen liittyvää epävarmuutta huomioida.

(a) Laske tuntemattomalle odotusarvoparametrille $\mu$ luottamusväli luottamustasolla 95%.

(b) Laske tuntemattomalle odotusarvoparametrille $\mu$ luottamusväli luottamustasolla 99%.

(c) Laske p-arvo testille, jossa nollahypoteesina on, että $\mu = 10.00$, ja vastahypoteesina sen komplementti.

(d) Käyttäen c-kohdan p-arvoa, hylätäänkö nollahypoteesi merkitsevyystasolla $\alpha = 0.05$?

(e) Käyttäen c-kohdan p-arvoa, hylätäänkö nollahypoteesi merkitsevyystasolla $\alpha = 0.01$?

(f) (Vapaaehtoinen lisätehtävä.) Tee kohdat a–e uudestaan huomioiden $\sigma$:aan liittyvä epävarmuus, ts. käyttäen standardinormaalijakauman sijasta $t$-jakaumaa parametrilla $n - 1 = 119$. Vertaa numeerisia tuloksia aiempiin.

T-jakauman kertymäfunktion ja kvantiilifunktion voi laskea R-komennoilla `pt` ja `qt`, tai Matlabin ja Octaven funktioilla `tcdf` ja `tinv`. T-jakauman taulukoita on myös netissä esim. `https://en.wikipedia.org/wiki/Student%27s_t-distribution`.

**Ratkaisu.**

(a) 95% luottamusväli on

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} = 10.08 \pm 1.96\,\frac{0.40}{\sqrt{120}} \approx \mathbf{10.08 \pm 0.07} = [\mathbf{10.01}, \mathbf{10.15}],$$

missä $z = 1.96$ on sellainen luku, että standardinormaalijakaumasta on sen yläpuolella osuus 0.025, toisin sanoen $\Phi(1.96) = 0.975$ (ja vastaavasti $-z$:n alapuolella on osuus 0.025, eli $\Phi(-1.96) = 0.025$).

Huomataan, että luottamusväli on kokonaan tavoitearvon 10.00 yläpuolella.

(b) 99% luottamustasoa varten käytämme kerrointa $z = 2.58$, koska $\Phi(2.58) \approx 0.995$ (= vasen ja häntä kumpikin 0.005, yhteensä 0.01). Luottamusväli on

$$m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}} = 10.08 \pm 2.58\,\frac{0.40}{\sqrt{120}} \approx \mathbf{10.08 \pm 0.09} = [\mathbf{9.99}, \mathbf{10.17}].$$

Huomataan, että tämä leveämpi luottamusväli sisältää tavoitearvon 10.00.

(c) Nollahypoteesi on $H_0 : \mu = 10.00$ ja vastahypoteesi $H_1 : \mu \neq 10.00$. Lasketaan aluksi testisuure

$$t(\vec{x}) = \frac{m(\vec{x}) - \mu}{\sigma/\sqrt{n}} = \frac{10.08 - 10.00}{0.40/\sqrt{120}} \approx 2.19.$$

Jos nollahypoteesi on totta, niin testisuureella $t(\vec{X})$ on standardinormaalijakauma. P-arvo on todennäköisyys havaita (nollahypoteesin vallitessa) testisuureelle arvo, joka on *ainakin* niin kaukana odotusarvosta (nollasta) kuin havaitsimme.

Koska testisuureen arvo 2.19 on positiivinen, lasketaan aluksi *oikean* hännän todennäköisyys

$$\mathrm{P}(t(\vec{X}) \geq 2.19) = 1 - \Phi(2.19) \approx 0.0142.$$

Vasemman hännän todennäköisyys $\mathrm{P}(t(\vec{X}) \leq -2.19)$ on (standardinormaalijakauman symmetrian nojalla) sama, joten yhteensä on

$$\mathrm{P}(\,|t(\vec{X})| \;\geq\; 2.19) = 2 \cdot (1 - \Phi(2.19)) \approx \mathbf{0.0285}.$$

Tämä on kysytty p-arvo.

(d) Koska p-arvo on $0.0285 < 0.05$, nollahypoteesi **hylätään**. Toisin sanoen havaittua naulanäytteen keskiarvoa 10.08 pidetään (merkitsevyystasolla $\alpha = 0.05$) liian poikkeavana ollakseen sopusoinnussa nollahypoteesin $\mu = 10.00$ kanssa.

(e) Koska p-arvo $0.0285 \geq 0.01$, nollahypoteesi **hyväksytään** (jää voimaan). Toisin sanoen naulanäytteen keskiarvo 10.08 ei ole niin poikkeava, että olisi (tällä merkitsevyystasolla) saatu riittävästi syytä hylätä nollahypoteesia.

(f) Luottamusvälin laskennassa käytetään nyt t-jakauman (parametrilla $n-1 = 119$) kvantiileja 0.975 ja 0.995. R-komennoilla `qt(.975, 119)` ja `qt(.995, 119)` saadaan kertoimet 1.98 ja 2.62, eli luottamusvälit ovat noin $1 \ldots 2\%$ leveämmät kuin standardinormaalijakaumaa käytettäessä.

Vastaavasti hypoteesitestissä lasketaan nyt p-arvoa varten häntätodennäköisyydet t(119)-jakauman mukaisesti, eli oikean hännän tn on `1-pt(2.19, 119)` $\approx 0.0152$, vasemman hännän tn on `pt(-2.19, 119)` $\approx 0.0152$ (huom. myös t-jakauma on symmetrinen origon suhteen), joten p-arvo on $\approx 2 \cdot 0.0152 \approx 0.0305$. P-arvo on hiukan suurempi kuin c-kohdassa, ts. havaintoa ei nyt pidetä aivan niin poikkeavana, mutta edelleen nollahypoteesi hylätään tasolla $\alpha = 0.05$ ja hyväksytään tasolla $\alpha = 0.01$.

**6B2** (Herneet) Gregor Mendel (1822–1884) oli modernin genetiikan uranuurtajia. Hän tutki muun muassa eräiden ominaisuuksien periytymistä herneissä. Eräiden tutkimusten nojalla hän päätyi seuraavaan tulokseen: Kun hernekasveja risteytetään eräällä tietyllä tavalla, niin kullakin jälkeläisellä on 75% todennäköisyys tuottaa keltaisia herneitä, ja 25% todennäköisyys tuottaa vihreitä herneitä.

Otetaan nollahypoteesiksi Mendelin väite, että kullakin kasvilla on todennäköisyys $\theta = 0.75$ tuottaa keltaisia herneitä. Väitettä tutkittiin tuottamalla 80 risteytettyä kasvia. Havaittiin, että niistä 56 tuotti keltaisia herneitä. Suorita tilastollinen testi luentomonisteen esimerkin 11.1 tapaan ja käytä merkitsevyystasoa $\alpha = 0.05$.

Ohje: Tutki ensin, paljonko havaittu lukumäärä poikkeaa nollahypoteesin mukaisesta odotusarvosta, ja laske tn, että havaitaan ainakin näin suuri poikkeama kumpaan tahansa suuntaan. Älä käytä normaaliapproksimaatiota. Todennäköisyyden komplementtisäännöstä (vastakohdan todennäköisyydestä) voi olla apua laskennassa, jotta ei tarvitse laskea kovin monta termiä.

**Ratkaisu.** Nollahypoteesi ja vastahypoteesi ovat:

$$H_0 : \theta = 0.75, \qquad H_1 : \theta \neq 0.75.$$

Käytetään tunnuslukuna keltaisia herneitä tuottavien kasvien lukumäärää $X$, jonka arvoksi on havaittu 56. Jos nollahypoteesi on tosi, niin $X$ on binomijakautunut parametrein $n = 80$ ja $\theta = 0.75$, jolloin sen odotusarvo on $n\theta = 60$. Merkitään kyseisen binomijakauman tiheysfunktiota

$$b(x) = \binom{80}{x} \theta^x (1-\theta)^{80-x}.$$

Havaittu arvo poikkeaa odotusarvosta 4 yksikköä alaspäin. Tn havaita ainakin näin suuri poikkeama alaspäin on

$$\mathrm{P}(X \leq 60 - 4) = \mathrm{P}(X \leq 56) = \sum_{x=0}^{56} b(x),$$

ja tn havaita ainakin näin suuri poikkeama ylöspäin on

$$\mathrm{P}(X \geq 60 + 4) = \mathrm{P}(X \geq 64) = \sum_{x=64}^{80} b(x).$$

Nämä summat voitaisiin laskea suoraankin (niissä on yhteensä 74 termiä), mutta ainakin käsin laskettaessa kannattaa käyttää vihjeen mukaisesti komplementtisääntöä:

$$\mathrm{P}(X \leq 56 \text{ tai } X \geq 64) = 1 - \mathrm{P}(57 \leq X \leq 63) = 1 - \sum_{x=57}^{63} b(x) \approx 0.366.$$

(Huomaa, että $X$ on kokonaislukuarvoinen, joten lukuarvojen kanssa täytyy olla tarkkana; jos $X$ ei ole $\leq 56$, niin se on $\geq 57$.) Tässä tarvitsee laskea vain 7 termiä. Laskun voi tehdä myös seuraavalla R-komennolla: `1-sum(dbinom(57:63, 80, 0.75))`

Koska p-arvo $0.366 \geq 0.05$, nollahypoteesi **hyväksytään** eli havainnot eivät antaneet aihetta epäillä Mendelin väitettä.