Valószínűségszámítás és statisztika

Statisztika témakör jegyzet

Készült Zempléni András előadásai
és Kovács Ágnes gyakorlatai alapján

Sárközi Gergő, 2021-22-2. félév
Nincsen lektorálva!

Tartalomjegyzék

1. Előadás 7: Statisztika bevezetés

- Két fő ág: leíró (újságokba), matematikai (becsléselmélet, hipotizéselmélet)
- Lényeges, hogy válaszainkat értelmezzük, mondatban válaszoljunk: laikusuk is értsék meg az eredményt.

1.1. Leíró statisztika alapfogalmak

- Statisztikai egység: vizsgálat tárgyát képező egység
- Statisztikai sokaság (populáció): egységek összessége, halmaza
  - Lehet hipotetikus is: gyár által jelenleg tervezett gyártandó termékek
- Statisztikai adat: sokaságra vonatkozó számszerű jellemző, mérési eredmény
- Statisztikai ismérv: sokaság egyedeit jellemző tulajdonság
- Ismérvváltozatok: ismérvek lehetséges kimenetelei
- Minta: sokaság véges számosságú részhalmaza
- Statisztikai következtetés: teljes sokaságot nem ismerjük, de a minta alapján következtetünk valamit a teljes sokaságról

1.1.1. Ismérvek típusai

- Első kategória
  - Minőségi: számszerűen nem mérhető
  - Mennyiségi: számszerűen mérhető, lehet diszkrét vagy folytonos
  - Időbeli
  - Területi
- Második kategória
  - Közös: sokaság egyedei között egyformák
  - Megkülönböztető: sokaság egyedei között eltérőek

1.1.2. Mérési skálák (mérési szintek)

- Névleges (nominális): hozzárendelt számok csak megkülönböztetnek, műveleteket végezni rajtuk értelmetlen (pl. személy neme)
- Sorrendi (ordinális): valamilyen tulajdonság alapján sorba rendezünk, egyedek tulajdonsága közötti különbséget nem tudjuk mérni (pl. érdemjegy)
- Intervallumskála: skálaértekek különbségei is valós infót adnak, a nullpont meghatározása a skálán önkényes (pl. hőmérséklet C-ben)
- Aranyskála: skálának van valódi nullpontja és minden matematikai művelet végezhető (pl. személyek magassága)
- Metrikus skála (ritkán használt): intervallum és aranyskála közös neve

1.1.3. Statisztikai tábla

- Statisztikai sorok összefüggő rendszere
- Egyszerű tábla: nincsenek csoportok, összegző sorok
- Csoportosító tábla:
  - Egyetlen csoportosító szempont
  - Gyakoriság van csak benne: hányan esnek a csoportba
- Kombinációs tábla, kontingenciatábla, kereszttábla:
  - Legalább két csoportosító szempont
  - Gyakoriságok vannak csak benne

1.2. Statisztikai elemzés lépései

- Tervezés: mit vizsgálunk, hogyan gyűjtünk adatot, előzetes sejtések/hipotézis
- Adatgyűjtés
- Adatbevitel
- Adatok validálása: nyilván rossz értékek kiszűrése (pl. negatív életkor)
- Adatelemezés, adatellenőrzés (leíró statisztika, grafikonok)
- Hibás adatok kijavítása vagy kihagyása
  - Lehetőleg ki kell javítani, nem pedig kidobni (nehéz feladat)
- Adatelemzés, statisztikai következtetések levonása (matematikai statisztika)
- Eredmények értelmezése, visszacsatolás

1.3. Mennyiségi sorok elemzése

- Ismérv diszkrét $\implies$ gyakorisági sort készítünk
- Ismérv folytonos vagy sok van belőle: osztályközös gyakorisági sor
  - Összevonunk több ismérvet: A_i és B_i közötti gyakoriságot számolunk
- Gyakori jelölések: n a minta mérete, k az ismérvértékek (sorok) száma, f_i a gyakoriság és x_i (vagy $x_{i,a} - x_{i,f}$ ha nem diszkrét) az ismérvérték
  - Osztályközös esetén x_i az osztályközepet jelöli: $x_i = \frac{x_{i,a}+x_{i,f}}{2}$

1.4. Középértékek számolása

- Mintaátlag, közvetlen adatból: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Mintaátlag, osztályközös gyakorisági sorokból: $\overline{x} = \frac{\sum_{i=1}^{n} f_i * x_i}{n}$
- Módusz: legtöbbször előforduló ismérvérték
- Medián: sorba rendezés után középső elem (rendezett minta: X^*)
  - $Me = \begin{cases} x^*[\frac{n+1}{2}] & \text{ha } n \text{ páratlan} \\ \frac{1}{2} * (x^*[\frac{n}{2}] + x^*[\frac{n}{2}+1]) & \text{ha } n \text{ páros} \end{cases}$

1.5. Kvantilisek

- y-kvantilis: $q_y = \inf\{x \mid F(x) \geq y\}$
  - Ha F invertálható: $q_y = F^{-1}(y)$
- Tapasztalati y-kvanatilis: ismérvérték; mintaelemek y-ad része $\leq$ nála
  - Sokféleképpen számolható, interpolációs módszer az egyik:
  - Sorszám megállapítása (e egészrész, t törtrész): $(n+1)y = e + t$
  - Kvantilis kiszámolása: $q_z = X^*_e + t(X^*_{e+1} - X^*_e)$
- Jelölje q_y a tapasztalati y-kvantilist
- Tercilisek: $T_1 = q_{1/3}$, $T_2 = q_{2/3}$
- Kvartilisek: $Q_1 = q_{1/4}$ (alsó), $Q_2 = Me = q_{2/4}$, $Q_3 = q_{3/4}$ (felső)
- Percentilisek: $P_i = q_{i/100}$ ahol $i = 1, 2, ..., 99$

1.6. Tapasztalati eloszlás

- Minden megfigyeléshez $\frac{1}{n}$ súlyt rendelünk $\implies$ diszkrét eloszlás
- $\overline{X} = E(X)$
- Tapasztalati eloszlásfüggvény: $F_n(x) = \frac{1}{n} * \sum_{i=1}^{n} I(x_i < x)$
  - I az indikátor, értéke 0 (ha $X_i < x$) vagy 1 (ha $X_i \geq x$)
  - Ábrázolva: ahol szakadás van, azt az értéket tudja felvenni (az ugrás mértéke a valószínűség)
- k-adik tapasztalati momentum: $m_k = \frac{1}{n} * \sum_{i=1}^{n} X_i^k$
- Glivenko-Cantelli tétel: tapasztalati eloszlásfüggvény ($F_n(x)$) és az elméleti eloszlásfüggvény ($F(x)$) közötti eltérés maximuma 1 valószínűséggel 0-hoz konvergál
  - Következmény: $F_n(x)$ közelít $F(x)$-hez ($\forall x$ esetén) a minta növésével
  - Elég nagy mintával tetszőleges közelséget el lehet érni

1.7. Szóródási mutatók számolása

- Terjedelem (range): $R = x_n^* - x_1^*$
- Interkvartilis terjedelem: $IQR = Q_3 - Q_1$
- Tapasztalati szórás:
  - Átlagtól való átlagos négyzetes eltérés négyzetgyöke
  - Közvetlenül: $S_n = \sqrt{\frac{1}{n} * \sum_{i=1}^{n}(x_i - \overline{x})^2}$
  - Osztályközös gyakoriságból: $S_n = \sqrt{\frac{1}{n} * \sum_{i=1}^{k} f_i(x_i - \overline{x})^2}$
- Korrigált tapasztalati szórás:
  - Átlagtól való korrigált átlagos négyzetes eltérés négyzetgyöke
  - Ez az alapértelmezett általában
  - Számítás: ugyan az, csak n helyett $n-1$-gyel osztunk (Jele: S_n^*)
  - Kis s-sel jelölés jelentése: nem a valószínűségi változóról, hanem a konkrét értékről van szó (nem nagyon számít, csak így helyesebb)
- Szórási együttható, relatív szórás: $V = S_n/\overline{X}$ $(*100\%)$

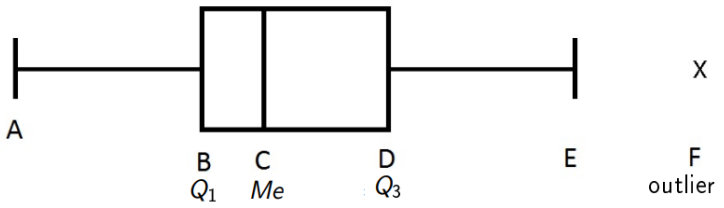
1.8. Grafikus megjelenítés

- Kördiagram rossz

1.8.1. Hisztogram

- Osztályok gyakoriságát ábrázolja (y az f_i gyakoriság, x az ismérv)
- Osztályok száma k, hosszuk (ha azonos): $h = \frac{x_n^* - x_1^*}{k}$
- Sűrűséghisztogram: $g_i = \frac{f_i}{n * h_i}$
  - Relatív gyakoriság / intervallumhossz értéket ábrázoljuk
  - Területarányos, összterület=1

1.8.2. Boxplot ábra (Box & Whiskers diagram)



- A = $max\{x_1^*; Q_1 - 1.5 * IQR\}$
- E = $min\{x_n^*; Q_3 + 1.5 * IQR\}$

2. Előadás 8: Matematikai statisztika, becsléselmélet

2.1. Matematikai statisztika

- Minta alapján teljes populáció tulajdonságaira következtetés
- Paramétertér: Θ (1 vagy több dimenziós, akár végtelen) ($\vartheta \in \Theta$)
- Minta: $X = (X_1, X_2, ..., X_n)$ i.i.d. valószínűségi változók sorozata
  - Minta realizációja $(x_1, ..., x_n)$: konkrét értékeket kap
- Mintatér: $\mathcal{X} : \mathbb{R}^n$, ide eshetnek a mintaelemek
- Mintaelemek eloszlása (F) ismeretlen, de paraméterezhető: F_ϑ
- Gyakori feladat: minta alapján adott eloszlás paraméterjének megállapítása

2.2. Becsléselmélet

2.2.1. Bevezetés, motiváció

- Legyen X egy minta
- Illeszkedésvizsgálat: milyen eloszlású lehet X?
- Pontbecslés: ismert eloszlás esetén mi az eloszlás paramétere?
  - Mintából számoljuk, így valamennyi hiba van benne
  - A kapott eredményben nincs benne, hogy mennyire biztos a becslés
  - Példák erre: Maximum Likelihood, Momentum-módszer
- Intervallumbecslés: milyen intervalban lesz nagy valószínűséggel ϑ?
  - Csak egyetlen szám helyett egy intervallum az eredmény
  - Intervallum hosszából következtethető, hogy mennyi biztos a becslés
  - Példa erre: konfidenciaintervallum (következő EA)

2.2.2. Alapdefiníciók

- Legyen $X = (X_1, ..., X_n)$ i.i.d. minta egy $\vartheta \in \mathbb{R}$ paraméterű eloszláscsaládból
- $T : \mathcal{X} \to \mathbb{R}$ becslés ϑ-ra
  - Torzítatlan $\Leftrightarrow E_\vartheta(T(X)) = \vartheta \quad (\forall \vartheta \in \Theta)$
  - Aszimptotikusan torzítat $\Leftrightarrow E_\vartheta(T_n(X)) \to \vartheta$ ha $n \to \infty$ $(\forall \vartheta \in \Theta)$
  - Konzisztens $\Leftrightarrow T_n(X) \to \vartheta$ sztochasztikusan ha $n \to \infty$ $(\forall \vartheta \in \Theta)$
    - Elégséges, ha T_n aszimptotikusan torzítatlan és $D^2(T_n) \to 0$

Mit becsülünk? $g(\vartheta)$	Mivel becsüljük? $T_n(X)$	Torzítatlan?	Aszimptotikusan torzítatlan?	Gyengén/erősen konzisztens?
$E_\vartheta X_1$	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	igen	igen	igen
$D^2_\vartheta X_1$	$S_n^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$	nem	igen	igen
$D^2_\vartheta X_1$	$(S_n^*)^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$	igen	igen	igen
$F_\vartheta(x)$	$F_n(x) = \frac{\sum_{i=1}^{n} I(X_i < x)}{n}$	igen	igen	igen
$E_\vartheta h(X_1)$	$\frac{\sum_{i=1}^{n} h(X_i)}{n}$	igen	igen	igen

2.2.3. Likelihood függvények

- Likelihood függvény: $L(\vartheta; x)$
  - Folytonos eloszlás: $L(\vartheta; x) = f_\vartheta(x) = \Pi_{i=1}^n f_\vartheta(x_i)$
  - Diszkrét eloszlás: $L(\vartheta; x) = P_\vartheta(X = x) = \Pi_{i=1}^n P_\vartheta(X_i = x_i)$
    - $\implies$ x_i gyakorisága g, akkor $P_\vartheta(X_i = x_i)^g$-ként jelenik meg
- Log-likelihood függvény (latex: ell): $\ell(\vartheta; x) = ln(L(\vartheta; x))$
- $\sum_{i=1}^n x_i$ felírható $n * \overline{x}$ alakban (szebb)
- $f_\vartheta(x)$-ban van elágazás és x függ ϑ-tól (pl. $0 \le x \le \vartheta$), akkor indikátor fv-t vegyünk be f_ϑ-ba: $...*I(0 \le x_i \le \vartheta) \implies ...*I(0 \le x_1^*)*I(x_n^* \le \vartheta)$
  - Maximumot keresünk $\implies$ indikátor értéke legyen 1

2.3. Maximum likelihood becslés (ML-módszer) (pontbecslés)

- $L(\vartheta; x)$ maximumát keressük (ugyan ott van, mint $\ell(\vartheta; x)$ maximuma)
- Maximum keresése deriválttal: $\partial_\vartheta \ell(\vartheta; x) = 0$
  - Nem szükséges további ellenőrzés: ahol 0, ott a max
  - Több dimenziós ϑ esetén: $\partial_{\vartheta_i} \ell(\vartheta; x) = 0$
- ML-becslés invariánsa: ϑ ML becslése $\hat{\vartheta} \implies g(\vartheta)$ ML-becslése $g(\hat{\vartheta})$

2.4. Nevezetes diszkrét eloszlások ML-becslése

- Egyes vizsgálatokhoz szükségünk lesz eloszlások paraméterének becslésére
- Binomiális: $p = \frac{\overline{X}}{m} = \frac{0*db_1+...+m*db_m}{m*\sum_{i=1}^{m} db_i}$ ahol m a másik paraméter
- Poisson: $\lambda = \overline{X} = \frac{1}{n} * \sum_{i=1}^{n} k_i$ ahol $k_i \geq 0$ az érték (nem gyakoriság)
- Geometriai, Pascal: $\frac{1}{\overline{X}} = \frac{n}{\sum_{i=1}^{n} k_i}$ ahol $k_i \geq 1$ az érték (nem gyakoriság)
- Negatív binomiális, hipergeometriaia: nem találtam

2.5. Momentum módszer (pontbecslés)

- Tapasztalati és elméleti momentumokat egyenlővé tesszük
  - Tapasztalati momentum (mintából származik): $m_i = \frac{1}{n} \sum_{j=1}^{n} x_j^i$
  - Elméleti momentum: $M_i(\vartheta) = E_\vartheta(X^i)$
- i értékei: $1, ..., p$ ahol p a ϑ dimenzióinak száma (egyenletrendszer lesz)
  - Egy dimenziós ϑ esetén: $\overline{x} = m_1 = M_1 = E_\vartheta(X)$
  - Két dimhez segítség: $E(X^2) = D^2(X) + E^2(X)$ $(D^2(X) = ...\text{-ból})$

2.6. Becslés hibája, standard hiba

- Becslés standard hibája a becslés szórása
- $s.e.(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ (azaz 0-hoz tart, ahogy az elemszám nő)
- σ ismeretlen $\implies$ becsüljük: $\widehat{s.e.}(\overline{X}) = \frac{\hat{\sigma}}{\sqrt{n}}$
  - Nem torzítatlan, csak aszimptotikusan

2.7. Konfidenciaintervallum (intervallumbecslés)

- Intervallum, ami legalább $1-\alpha$ valószínűséggel tartalmazza a paramétert minden ϑ értékre
  - Azaz a valódi m vagy σ ekkora valószínűséggel van az intervallumban
  - 100 szimulációból kb. $(1-\alpha)*100$-szor lesz az intervallumban
- Mi elsősorban normál eloszlással fogunk csak dolgozni
- Emlékeztető: $\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$
- u_x jelentése: x-hez tartozó $N(0;1)$ eloszlás kvantilis ($\Phi(u_x) = x$)
- t_x jelentése: x-hez tartozó $n-1$ szabadsági fokú t eloszlás kvantilis
- Intervallum hossza csökken, ha n nő és ha σ csökken
- Intervallum hossző nő, ha α csökken

2.7.1. Kétoldali $1-\alpha$ megbízhatóságú konfidenciaintervallum

- m-re, ha σ ismert: $\overline{X} \pm u_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$
  - ```
    ci_also <- mean(sample) - qnorm(1 - alpha / 2) * sigma / sqrt(n)
    ci_felso <- mean(sample) + qnorm(1 - alpha / 2) * sigma / sqrt(n)
    ```
  - Szükséges elemszám adott intervallum hosszhoz:

    ```
    hossz <- 8
    legalabbn <- (2 * qnorm(1 - alpha / 2) * sigma / hossz)^2
    ```
- m-re, ha σ ismeretlen: $\overline{X} \pm t_{n-1;1-\frac{\alpha}{2}} * \frac{S_n^*}{\sqrt{n}}$
  - ```
    sd <- sd(sample)
    ci_also <- mean(sample) - qt(1 - alpha / 2, df=n-1) * sd / sqrt(n)
    ci_felso <- mean(sample) + qt(1 - alpha / 2, df=n-1) * sd / sqrt(n)
    ```
- σ^2-re: $[\frac{(n-1)*(S_n^*)^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}; \frac{(n-1)*(S_n^*)^2}{\chi^2_{n-1;\frac{\alpha}{2}}}]$

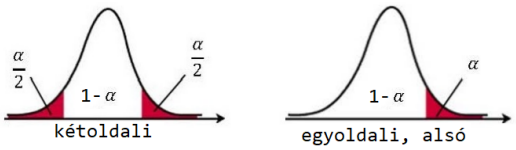
2.7.2. Egyoldali, alsó, $1-\alpha$ megbízhatóságú konfidenciaintervallum

- m-re, ha σ ismert: $(-\infty; \overline{X} + u_{1-\alpha} * \frac{\sigma}{\sqrt{n}})$
  - `ci_felso <- mean(sample) + qnorm(1 - alpha) * sigma / sqrt(n)`
  - Szükséges elemszám adott intervallum hosszhoz:

    ```
    hossz <- 8
    legalabbn <- (2 * qnorm(1 - alpha) * sigma / hossz)^2
    ```

- m-re, ha σ ismeretlen: $(-\infty; \overline{X} + t_{n-1;1-\alpha} * \frac{S_n^*}{\sqrt{n}})$
  - ```
    sd <- sd(sample)
    ci_felso <- mean(sample) + qt(1 - alpha, df=n-1) * sd / sqrt(n)
    ```

- σ^2-re: $(0; \frac{(n-1)*(S_n^*)^2}{\chi^2_{n-1;\alpha}})$


    ```

3. Előadás 9: Hipotézisvizsgálat, próbák

3.1. Hipotézisvizsgálat

- Hipotézis: állítás aminek igazságát vizsgálni szeretnénk: elfogadjuk/elutasítjuk
- Paraméterteret diszjunkt részekre bontjuk: $\Theta = \Theta_0 \cup^* \Theta_1$
- Nullhipotézis: $H_0 : \vartheta \in \Theta_0$
  Ellenhipotézis, alternatív hipotézis: $H_1 : \vartheta \in \Theta_1$
- Nullhipotézist nem "elfogadjuk", hanem "nem tudjuk elvetni".
  Viszont elutasítani el tudjuk.
- Nullhipotézis megválasztása: sok éves tapasztalatnak feleljen meg, reméljük teljesülését, aminek elutasítása negatív következménnyel jár (pl. bírság)
  - Az ellenhipotézis a lényeg, arról fogunk dönteni.
  - Egyenlőségjel mindig a nullhipotézisbe kerül.
- Próba: segítségével döntés hozás a hipotézisről
  - Statisztikai próba vagy próba: minta alapján hozunk döntést
  - Paraméteres próba: eloszlás típusa ismert, a nullhipotézis az eloszlás paraméterére (vagy annak egy függvényére) vonatkozik
    * Továbbiakban ezzel fogunk általában foglalkozni
    * Továbbiakban legyen $\Theta \subset \mathbb{R}$, azaz a paraméter valós
- Mintateret diszjunkt részekre bontjuk: $\chi = \chi_e \cup^* \chi_k$
  - χ_k, kritikus tartomány: megfigyelések, amikre elutasítjuk H_0-t
  - χ_e, elfogadási tartomány: megfigyelések, amikre elfogadjuk H_0-t
- Döntési mátrix hipotézisvizsgálat esetén:

$\downarrow$ valóság | döntés $\rightarrow$	elfogadjuk (χ_e)	elutasítjuk (χ_k)
H_0 teljesül (Θ_0)	helyes döntés	elsőfajú hiba
H_0 nem teljesül (Θ_1)	másodfajú hiba	helyes döntés

3.2. Hiba valószínűségek, erőfüggvény, terjedelem

- Elsőfajú hiba valószínűsége:
  - Egyszerű H_0, $|\Theta_0| = 1$: $\alpha(\vartheta) = P_\vartheta(X \in \chi_k) = P_0(\chi_k) \quad (\vartheta \in \Theta_0)$
  - Összetett H_0, $|\Theta_0| > 1$: $\alpha \geq P_\vartheta(X \in \chi_k) \quad (\forall \vartheta \in \Theta_0)$
- Másodfajú hiba valószínűsége:
  $\beta(\vartheta) = P_\vartheta(X \in \chi_e) = P_1(\chi_e) = 1 - P_\vartheta(\chi_k) \quad (\vartheta \in \Theta_1)$
- Erőfüggvény: $\psi(\vartheta) = 1 - P_\vartheta(\chi_e) = P_\vartheta(\chi_k)$ ahol $\vartheta \in \Theta_1$
  - Jelentése: valószínűsége H_0 elvetésének, amikor az hamis
  - Valószínűsége annak, hogy egy adott különbséget egy adott mintanagyság és terjedelem mellett ki egy statisztikai próba kimutat
- Terjedelem, pontos terjedelem, szignifikanciaszint: $\alpha = \sup_{\vartheta \in \Theta_0} \alpha(\vartheta)$
  - Általában feladat elejekor 5%-on (vagy 1% és 10% között) rögzített
  - Megbízhatósági szint, konfidenciaszint: $1 - \alpha \quad (*100\%)$
    - Valószínűsége, hogy H_0-t elfogadjuk, amikor az igaz
  - Másképp: elsőfajú hiba valószínűsége α lesz

3.3. Próbák bevezetés

- Kétoldali próba: $H_0 : \vartheta = \vartheta_0$ és $H_1 : \vartheta \neq \vartheta_0$
- Egyoldali próba: $H_0 : \vartheta = \vartheta_0$ és $H_1 : \vartheta < \vartheta_0$ (vagy $>$)
- Próbastatisztika: alkalmas T statisztika, amivel a χ_k-t meghatározzuk
  - Kétoldali próbához: $\chi_k = \{x \in \chi \ : \ |T(X)| > c\}$
  - Egyoldali próbához: $\chi_k = \{x \in \chi \ : \ T(X) \lessgtr c\}$
  - c neve: kritikus érték
    - Jellemzően függ a próba terjedelmétől $\implies$ c_α-val jelöljük
    - c_α jelölés jelentése: c_α a $T(X)$ val. változó α-kvantilise
  - Próba meghatározása: előre rögzített α terjedelemhez keressük azt a c_α értéket, amire a próba pontos terjedelme éppen α
    - $\sup_{\vartheta \in \Theta_0} P_\vartheta(T(X) > c_\alpha) = \alpha$

3.4. Hipotézisvizsgálat menete

- Terjedelem (α) lefixálása, általában 5%-on (megbízhatóság: $1 - \alpha$)
- Nullhipotézis: sokévi tapasztalatnak megfelelő paramétertartomány
  - Az egyenlőség (pl. $\leq$) mindig ide kerül
- Alternatív hipotézis: feladat kérdéséhez megfelelő paramétertartomány
  - Erről be tudjuk látni, hogy igaz (H_0-ról csak "nem tudjuk elvetni")
  - Ezért a cél H_1 igazolása, azaz H_0 elvetése
- Problémához alkalmas próba/próbák választása (egy/két oldali, stb.)
- Próbastatisztika kiszámítása

3.4.1. Döntés minta és tartományok alapján

- Kritikus érték kiszámítása, kritikus tartomány megállapítása
  - Számolása általában: eloszlás kvantilis függvény "meghívása" α-ra
- $x \in \chi_k \quad \Leftrightarrow \quad H_1$-et elfogadjuk
- Probléma: nem derül ki, hogy mennyire voltunk közel az elfogadáshoz

3.4.2. Döntés p-érték segítségével

- p-érték kiszámolása (számítógépes számolás esetén lehetőség)
  - Számolása általában: eloszlásfüggvény "meghívása" próbastatisztikával
  - Kétoldali próba esetén bonyolultabb (általában: $2 * pDist(-|T|)$)
- p-érték $< \alpha \quad \Leftrightarrow \quad x \in \chi_k \quad \Leftrightarrow \quad H_1$-et elfogadjuk
- p-érték jelentése: terjedelem, amire a kritikus érték megegyezik a próbastatisztikával
  - Máshogy: legkisebb α, amire az adott minta esetén elvetjük H_0-t
  - Máshogy: igaz H_0 mellett annak a valószínűsége, hogy a tapasztalt eltérést, vagy annál nagyobb eltérést kapunk

3.4.3. Elsőfajú, másodfajú hiba csökkentése

- α csökkentése β növekedésével jár (ha minden más marad)
- Mindkét hiba valószínűségének csökkentése: mintaelemszám növelése

4. Előadás 10: Próbák normális eloszlás paramétereire

4.1. Használt jelölések, emlékeztetők

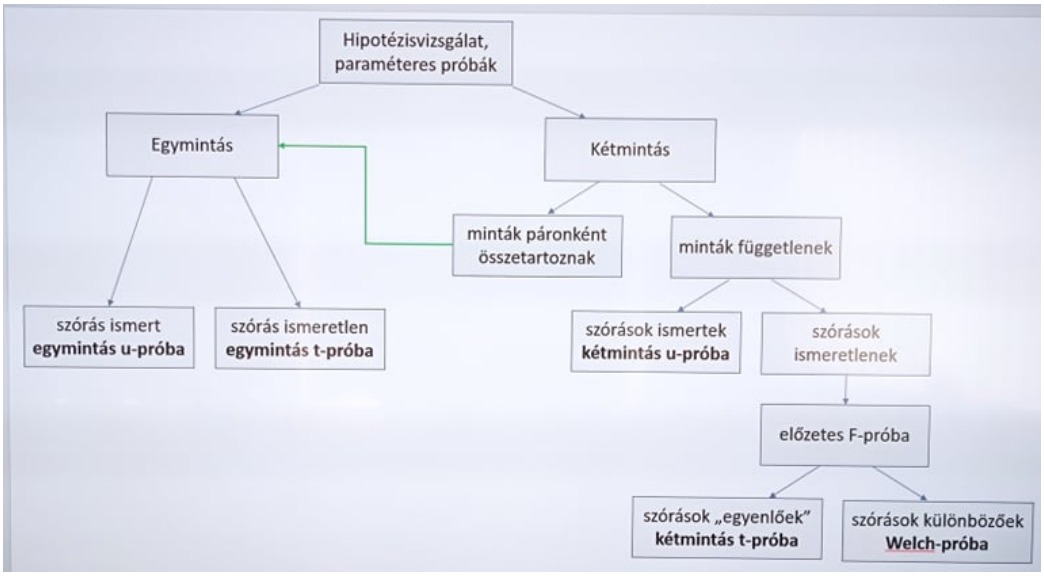
- Emlékeztető: $\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n} * \frac{\overline{X}-m}{\sigma} \sim N(0;1)$
- u_x jelentése: x-hez tartozó $N(0;1)$ eloszlás kvantilis ($\Phi(u_x) = x$)
- t_x jelentése: x-hez tartozó $n-1$ szabadsági fokú t eloszlás kvantilis
- S_n^* jelentése: korrigált tapasztalati szórás, $S_n^* = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{n}(x_i - \overline{x})^2}$

4.2. Próbákról tudnivalók

- Kétmintás próba: két (összefüggő vagy független) mintánk van
- Összefüggő (párosított) minták: vettünk egy mintát, valami megváltozott (pl. gépen valamit állítottunk) és veszünk még egy mintát ugyan onnan
  - Cél: változtatás hatásának vizsgálata (pl. működik-e a gyógyszer)
- Kétoldali próba: egyenlőséget ellenőrzünk (pl. m tényleg az-e)
- Egyoldali próba: gyanúnk, hogy pl. m valaminél kisebb/nagyobb

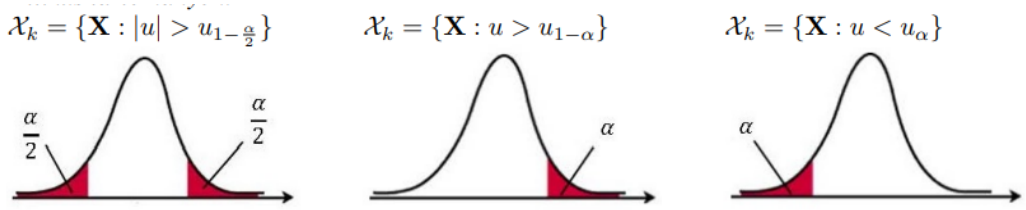
4.3. Próbák normális eloszlás várható értékére (m)

- Egymintás próba
  - Szórás ismert: egymintás u-próba
  - Szórás ismeretlen: egymintás t-próba
- Kétmintás próba, két minta független
  - Szórások ismertek: kétmintás u-próba
  - Szórások ismeretlenek: előzetes F-próba szükséges
    - Szórások megegyeznek: kétmintás t-próba
    - Szórások eltérnek: Welch-próba
- Kétmintás próba, két minta párosított (összefüggő)
  - Szórások ismertek: egymintás u-próba a különbségekre
  - Szórások ismeretlenek: egymintás t-próba a különbségekre



4.3.1. Egymintás u-próba (z-test)

- $X_1, ..., X_n \sim N(m, \sigma^2)$ ahol σ ismert és $m = ?$
- Próbastatisztika: $T(X) = u = \sqrt{n} * \frac{\overline{X} - m_0}{\sigma}$ (H_0 esetén $u \sim N(0,1)$)
  - `u <- sqrt(n) * (mean(sample) - mu0) / sigma`
- Kétoldali: $H_0 : m = m_0$ és $H_1 : m \neq m_0$ és $\chi_k = \{x \ : \ |u| > u_{1-\alpha/2}\}$
  - `pertek <- 2*pnorm(-abs(u))`
  - `krit <- qnorm(c(alpha/2, 1-alpha/2)) #1.alatt és 2.felett`
- Egyoldali: $H_0 : m = m_0$ és
  - $H_1 : m < m_0$ és $\chi_k = \{x \ : \ u < u_\alpha\}$
    - `pertek <- pnorm(u)`
    - `krit <- qnorm(alpha)`
  - $H_1 : m > m_0$ és $\chi_k = \{x \ : \ u > u_{1-\alpha}\}$
    - `pertek <- pnorm(-u)`
    - `krit <- qnorm(1-alpha)`
- Kapcsolat konfidenciaintervallummal:
  - $|u| > u_{1-\alpha/2} \quad \Leftrightarrow \quad m_0 \notin (\overline{X} - u_{1-\alpha/2} * \frac{\sigma}{\sqrt{n}}, \ \overline{X} + u_{1-\alpha/2} * \frac{\sigma}{\sqrt{n}})$
  - Máshogy: H_0-t pontosan akkor utasítjuk el, ha az $1-\alpha$ konfidenciaintervallum nem tartalmazza m_0-t
- ```
  library(TeachingDemos)
  z.test(x, alternative = c("two.sided/less/greater"),
         mu = mu0, stdev = sigma, conf.level = 1 - alpha)
  ```


  ```

4.3.2. Egymintás t-próba (Student's t-test)

- $X_1, ..., X_n \sim N(m, \sigma^2)$ ahol σ és m ismeretlen; $m =?$
- Próbastatisztika: $T(X) = t = \sqrt{n} * \frac{\overline{X} - m_0}{S_n^*}$ (H_0 esetén $t \sim t_{n-1}$)
  - `t <- sqrt(n) * (mean(sample) - mu0) / sd(sample)`
- Kétoldali: $H_0 : m = m_0$ és $H_1 : m \neq m_0$ és $\chi_k = \{x \ : \ |t| > t_{n-1,1-\alpha/2}\}$
  - `pertek <- 2*pt(-abs(t), df=n-1)`
  - `krit <- qt(c(alpha/2, 1-alpha/2), df=n-1) #1.alatt és 2.felett`
- Egyoldali: $H_0 : m = m_0$ és
  - $H_1 : m < m_0$ és $\chi_k = \{x \ : \ t < t_{n-1,\alpha}\}$
    - `pertek <- pt(t, df=n-1)`
    - `krit <- qt(alpha, df=n-1)`
  - $H_1 : m > m_0$ és $\chi_k = \{x \ : \ t > t_{n-1,1-\alpha}\}$
    - `pertek <- pt(-t, df=n-1)`
    - `krit <- qt(1-alpha, df=n-1)`
- `t.test(x, alternative = "two.sided/less/greater",`
  `mu = mu0, conf.level = 1 - alpha)`

4.3.3. Kétmintás u-próba

- Független minták: $X_{1..n} \sim N(m_1, \sigma_1^2)$ és $Y_{1..m} \sim N(m_2, \sigma_2^2)$
  - σ_1, σ_2 ismert és m_1, m_2 ismeretlen (relációjuk a kérdés)
- Próbastatisztika: $T(X,Y) = u = (\overline{X} - \overline{Y}) / \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$
  - H_0 esetén $u \sim N(0,1)$
  - `u <- (mean(sampleA) - mean(sampleB)) / sqrt(sigmaA^2/n + sigmaB^2/m)`
- Kétoldali: $H_0 : m_1 = m_2$; $H_1 : m_1 \neq m_2$ és $\chi_k = \{(x,y) : |u| > u_{1-\alpha/2}\}$
  - `pertek <- 2*pnorm(-abs(u))`
  - `krit <- qnorm(c(alpha/2, 1-alpha/2)) #1.alatt és 2.felett`
- Egyoldali: $H_0 : m_1 = m_2$ és
  - $H_1 : m_1 < m_2$ és $\chi_k = \{(x,y) \ : \ u < u_\alpha\}$
    - `pertek <- pnorm(u)`
    - `krit <- qnorm(alpha)`
  - $H_1 : m_1 > m_2$ és $\chi_k = \{(x,y) \ : \ u > u_{1-\alpha}\}$
    - `pertek <- pnorm(-u)`
    - `krit <- qnorm(1-alpha)`
- ```
  two_sample_u_test <- function(sampleA, sampleB, sigmaA, sigmaB,
                                alternative, conf_level) {
    alpha <- 1 - conf_level; n <- length(sampleA); m <- length(sampleB)
    u <- (mean(sampleA) - mean(sampleB)) / sqrt(sigmaA^2 / n + sigmaB^2 / m)
    if (alternative == "t" || alternative == "two.sided") {
      krit <- qnorm(c(alpha/2, 1 - alpha/2)); pertek <- 2 * pnorm(-abs(u))
    } else if (alternative == "l" || alternative == "less") {
      krit <- qnorm(alpha); pertek <- pnorm(u)
    } else if (alternative == "g" || alternative == "greater") {
      krit <- qnorm(1 - alpha); pertek <- pnorm(-u)
    } else { cat("Invalid alternative: " + alternative); return() }
    cat("Próbastatisztika:", u,
        "\nKritikus tartomány:", krit[1], "alatt és", krit[2], "felett",
        "\nP-érték:", pertek, "\nDöntés:", if (pertek < alpha)
           { "H0 elutasítva" } else { "H0-t nem sikerült elvetni" }, "\n")
  } %sampleA, sampleB sorrendje: lásd jegyzet PDF, kétmintás t-próba
  ```
  ```

4.3.4. Kétmintás t-próba

- Független minták: $X_{1..n} \sim N(m_1, \sigma_1^2)$ és $Y_{1..m} \sim N(m_2, \sigma_2^2)$
  - $\sigma_1 = \sigma_2$ ismeretlen és m_1, m_2 ismeretlen (relációjuk a kérdés)
- Próbastatisztika: $T(X,Y) = t = \sqrt{\frac{n*m}{n+m}} * (\overline{X} - \overline{Y}) / \sqrt{\frac{(n-1)(S_1^*)^2 + (m-1)(S_2^*)^2}{n+m-2}}$
  - H_0 esetén $t \sim t_{n+m-2}$
  - `t <- sqrt((n*m)/(n+m)) * (mean(sampleA) - mean(sampleB)) / sqrt( ((n-1)*sd(sampleA)^2 + (m-1)*sd(sampleB)^2) / (n+m-2) )`
- Kétoldali: $H_0 : m_1 = m_2$ és $H_1 : m_1 \neq m_2$
  és $\chi_k = \{(x,y) : |t| > t_{n+m-2,1-\alpha/2}\}$
- Egyoldali: $H_0 : m_1 = m_2$ és
  - $H_1 : m_1 < m_2$ és $\chi_k = \{(x,y) \ : \ t < t_{n+m-2,\alpha}\}$
    - `pertek <- pt(t)`
    - `krit <- qt(alpha)`
  - $H_1 : m_1 > m_2$ és $\chi_k = \{(x,y) \ : \ t > t_{n+m-2,1-\alpha}\}$
    - `pertek <- pt(-t)`
    - `krit <- qt(1-alpha)`
- `t.test(sampleA, sampleB, alternative="two.sided/less/greater", paired=FALSE, var.equal=TRUE)`
  - A minta sorrend és a reláció egyezzen meg az ellenhipotézissel
  - Ez a kettő ekvivalens: sampleA,sampleB,less $\Leftrightarrow$ sampleB,sampleA,greater

4.3.5. Welch-próba

- Független minták: $X_{1..n} \sim N(m_1, \sigma_1^2)$ és $Y_{1..m} \sim N(m_2, \sigma_2^2)$
  - $\sigma_1 \neq \sigma_2$ ismeretlen és m_1, m_2 ismeretlen (relációjuk a kérdés)
- Próbastatisztika: $T(X,Y) = t' = (\overline{X} - \overline{Y})/\sqrt{\frac{(S_1^*)^2}{n} + \frac{(S_2^*)^2}{m}}$
  - H_0 esetén $t' \sim t_f$ ahol $\frac{1}{f} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$
  - $S_1^* > S_2^*$ (így válasszuk) $\implies c = (\frac{(S_1^*)^2}{n})/(\frac{(S_1^*)^2}{n} + \frac{(S_2^*)^2}{m})$
- Kétoldali: $H_0 : m_1 = m_2$ és $H_1 : m_1 \neq m_2$
  és $\chi_k = \{(x,y) : |t'| > t_{f,\alpha/2}\}$
- Egyoldali: $H_0 : m_1 = m_2$ és
  - $H_1 : m_1 < m_2$ és $\chi_k = \{(x,y) \ : \ t < -t_{f,\alpha}\}$
  - $H_1 : m_1 > m_2$ és $\chi_k = \{(x,y) \ : \ t > t_{f,\alpha}\}$
- `t.test(sampleA, sampleB, alternative="two.sided/less/greater", paired=FALSE, var.equal=FALSE)`
  - Két minta (sampleA, sampleB) sorrendje: lásd kétmintás t-próba

4.4. Próbák normális eloszlás szórásnégyzetére (σ^2)

- Gyakorlaton nem foglalkozunk ilyennel, kivéve az (előzetes) F-próba
- Egymintás próba: χ^2-próba
- Kétmintás próba: F-próba

4.4.1. F-próba

- Független minták: $X_{1..n} \sim N(m_1, \sigma_1^2)$ és $Y_{1..m} \sim N(m_2, \sigma_2^2)$
  - m_1, m_2 ismeretlen és σ_1, σ_2 ismeretlen (relációjuk a kérdés)
- Próbastatisztika: $T(X,Y) = F = \frac{(S_2^*)^2}{(S_1^*)^2}$ ha $S_1^* < S_2^*$
  - H_0 esetén $F \sim F_{n-1,m-1}$
- Kétoldali: $H_0 : \sigma_1 = \sigma_2$ és $H_1 : \sigma_1 \neq \sigma_2$
  - vagy $\chi_k = \{(x,y) : F < F_{n-1,m-1,\alpha/2}\}$
  - vagy $\chi_k = \{(x,y) : F > F_{n-1,m-1,1-\alpha/2}\}$
  - Attól függ, hogy S_1^* vagy S_2^* a nagyobb
- Egyoldali: $H_0 : \sigma_1 = \sigma_2$ és
  - $H_1 : \sigma_1 < \sigma_2$ és $\chi_k = \{(x,y) \ : \ F < F_{n-1,m-1,\alpha}\}$
  - $H_1 : \sigma_1 > \sigma_2$ és $\chi_k = \{(x,y) \ : \ F > F_{n-1,m-1,1-\alpha}\}$
- Előzetes F-próba:
  - Mindig kétoldali
  - Nem számít a minták sorrendje (p-érték nem változik)
  - p-érték nagy $\implies$ nincs bizonyíték, hogy különböznek a szórások
    - Ha nem tudunk dönteni, inkább tekintsük a szórásokat egyenlőnek
  - `var.test(mintaA, mintaB, alternative="two.sided")`
  - ```
    f <- sd(mintaA)^2 / sd(mintaB)^2
    pertek <- 2 * pf(f, df1 = length(mintaA) - 1,
                        df2 = length(mintaB) - 1)
    ```
    ```

4.4.2. χ^2-próba

- $X_1, ..., X_n \sim N(m, \sigma^2)$ ahol σ és m ismeretlen; σ =?
- Próbastatisztika: $T(X) = h = \frac{(n-1)(S_n^*)^2}{\sigma_0^2}$ (H_0 esetén $h \sim \chi^2_{n-1}$)
- Kétoldali: $H_0 : \sigma = \sigma_0$ és $H_1 : \sigma \neq \sigma_0$
  - vagy $\chi_k = \{x \ : \ h < \chi^2_{n-1,\alpha/2}\}$
  - vagy $\chi_k = \{x \ : \ h > \chi^2_{n-1,1-\alpha/2}\}$
  - Attól függ, hogy S_1^* vagy S_2^* a nagyobb
- Egyoldali: $H_0 : m = m_0$ és
  - $H_1 : \sigma < \sigma_0$ és $\chi_k = \{x \ : \ h < \chi^2_{n-1,\alpha}\}$
  - $H_1 : \sigma > \sigma_0$ és $\chi_k = \{x \ : \ h > \chi^2_{n-1,1-\alpha}\}$
- Nem tételezünk fel normális eloszlást (a mintáról)
- TODO teljes eseményrendszerről a dolgok
- Alkalmazások: TODO
- Nincs minden osztályban elég mennyiség: R adhat warning-ot
  - Ökölszabály: min. 5db minden osztályban ($n * p$ szorzás után)
  - Ha nincs elég, akkor vonjunk össze osztályokat:
    - `chisq.test(c(gyakorisag[1:3], sum(gyakorisag[4:5])), p = c(p[1:3], sum(p[4:5]))) #utolsó 2 összevonva`
    - `tbl2 <- cbind(tbl1[,"Left"] + tbl1[,"Neither"], tbl1[,"Right"])`
      `colnames(tbl2) <- c("Left+Neither", "Right")`

5. Előadás 11: illeszkedés-, homogenitás- és függetlenségvizsgálat; regresszióelemzés

5.1. Diszkrét illeszkedésvizsgálat (χ^2-próba)

- H_0: minta egy adott eloszlásból származik (valószínűségek egyeznek)
- H_1: minta nem ilyen eloszlású (min 1x: várt, tap. valószínűségek $\neq$)
  - Próbastat.: $T_n(X) = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \to \chi^2_{r-1}$ (ha H_0 és $n \to \infty$)
  - Kritikus tartomány: $\chi_k = \{x \mid T_n(x) > \chi^2_{r-1,1-\alpha}\}$
- Tiszta illeszkedésvizsgálat: feltételezett eloszlás ismert
- Becsléses illeszkedésvizsgálat: eloszlás paramétere ismeretlen
  - ML-módszerrel s darab paramétert meg kell becsülni
  - Próbastatisztika ekkor H_0 esetén χ^2_{r-1-s}-be tart

Osztály	1	...	r	Összesen
Gyakoriság	v_1	...	v_r	n
Valószínűség	p_1	...	p_r	1

5.1.1. Diszkrét illeszkedésvizsgálat R-ben egyszerűen

- Legyen `gyakorisag` és `fejek_szama` egy-egy vektor
- Példa p-re: `p <- dbinom(fejek_szama, size = 4, p = 0.25)`
- `chisq.test(gyakorisag, p = p)`

5.1.2. Diszkrét illeszkedésvizsgálat R-ben manuálisan

```
s <- 0 #Becsült paraméterek száma
probastat <- sum((gyakorisag - p * sum(gyakorisag))^2 / (p * sum(gyakorisag)))
pertek <- 1 - pchisq(probastat, length(gyakorisag) - 1 - s)
cat('Próbastatisztika:', probastat,
    '\nKritikus érték', qchisq(1 - alpha, length(gyakorisag) - 1 - s),
    '\nP-érték:', pertek,
    '\nDöntés:', if (pertek < alpha) { 'H0 elutasítva' }
    else { 'H0-t nem sikerült elvetni' }, '\n')
```

5.2. Folytonos illeszkedésvizsgálat Kolmogorov-Szmirnov próbával

- TODO 11. előadás 5. oldal
- Gyakorlaton nem vettük, ZH-ban benne volt

5.3. Homogenitásvizsgálat

- Két független minta, 1 közös szemponttal r osztályba soroljuk őket
- H_0: két eloszlás megegyezik ($p_i = q_i$)
- H_1: két eloszlás nem egyezik meg (legalább egy helyen)
- Próbastatisztika: $nm \sum_{i=1}^{r} \frac{(N_i/n - M_i/m)^2}{N_i + M_i} \to \chi^2_{r-1}$ (ha H_0 és $n \to \infty$)
- Kritikus tartomány: $\chi_k = \{(X, Y) \ : \ T_{n,m}(X, Y) > \chi^2_{r-1,1-\alpha}\}$

Osztály	1	...	r	Összesen
1. minta: gyakoriság	N_1	...	N_r	n
1. minta: valószínűség	p_1	...	p_r	1
2. minta: gyakoriság	μ_1	...	μ_r	m
2. minta: valószínűség	M_1	...	M_r	1

5.3.1. Homogenitásvizsgálat R-ben

- `chisq.test(matrix(c(15, 10, 10, 10), ncol=2, byrow=TRUE))`
  - Értelmes táblázat/mátrix kezelés: lásd R jegyzet (lejjebb)

5.3.2. Homogenitásvizsgálat R-ben manuálisan

```
n <- sum(x); m <- sum(y)
probastat <- n * m * sum((x / n - y / m)^2 / (x + y))
pertek <- 1 - pchisq(probastat, length(x) - 1)
cat('Próbastatisztika:', probastat,
    '\nKritikus érték', qchisq(1 - alpha, length(x) - 1),
    '\nP-érték:', pertek,
    '\nDöntés:', if (pertek < alpha) { 'H0 elutasítva' }
    else { 'H0-t nem sikerült elvetni' }, '\n')
```

5.4. Függetlenségvizsgálat

- Egy mintát két szempont alapján osztályokba sorolunk
  - Táblázat: osztály-osztály metszet gyakoriság van benne
- H_0: két szempont független egymástól ($p_{i,j} = p_{i\circ} * p_{\circ j}$)
- H_1: két szempont nem független (nincs egyenlőség legalább egy helyen)
- Próbastatisztika: $\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{i,j} - N_{i\circ}N_{\circ j}/n)^2}{N_{i\circ}N_{\circ j}/n} \to \chi^2_{(r-1)(s-1)}$
  (H_0 és $n \to \infty$ esetén)
- Kritikus tartomány: $\chi_k = \{(X,Y) \ : \ T_n(X,Y) > \chi^2_{(r-1)(s-1),1-\alpha}\}$
- $r = s = 2$ esetén
  - Próbastatisztika: $T_n = n\frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\circ}N_{2\circ}N_{\circ 1}N_{\circ 2}}$
  - Szabadsági foka χ^2-nek pedig 1

		2. szempont					
		1	...	j	...	s	Összesen
	1	N_{11}	...	N_{1j}	...	N_{1s}	$N_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
1. szempont	i	N_{i1}	...	N_{ij}	...	N_{is}	$N_{i\bullet}$
	⋮	⋮		⋮		⋮	⋮
	r	N_{r1}	...	N_{rj}	...	N_{rs}	$N_{r\bullet}$
Összesen		$N_{\bullet 1}$	...	$N_{\bullet j}$	...	$N_{\bullet s}$	n

5.4.1. Függetlenségvizsgálat R-ben

- `chisq.test(matrix(c(15, 10, 10, 10), ncol=2, byrow=TRUE))`
  - Értelmes táblázat/mátrix kezelés: lásd R jegyzet (lejjebb)

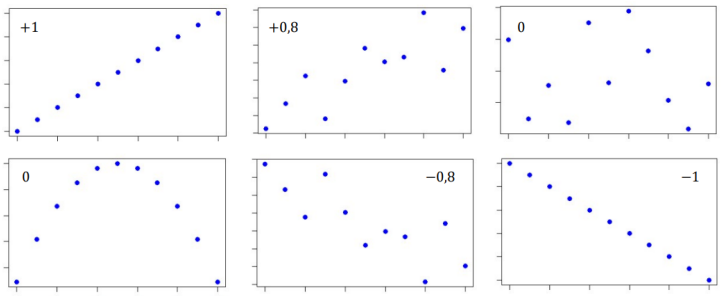
5.4.2. Függetlenségvizsgálat R-ben manuálisan

```
rows <- nrow(tablazat); cols <- ncol(tablazat)
probastat <- 0; for (i in 1:rows) for (j in 1:cols) {
  tmp <- sum(tablazat[i,]) * sum(tablazat[, j]) / sum(tablazat)
  probastat <- probastat + (tablazat[i, j] - tmp)^2 / tmp }
pertek <- 1 - pchisq(probastat, (rows - 1) * (cols - 1))
cat('Próbastatisztika:', probastat,
    '\nKritikus érték', qchisq(1 - alpha, (rows - 1) * (cols - 1)),
    '\nP-érték:', pertek,
    '\nDöntés:', if (pertek < alpha) { 'H0 elutasítva' }
    else { 'H0-t nem sikerült elvetni' }, '\n')
```

5.5. Korreláció- és regresszióelemzés

5.5.1. Korreláció

- Korreláció: szimmetrikus, két változó lineáris kapcsolatának erőssége
- Értéke: -1 és 1 között, ahol -1 az erős negatív kapcsolat
  - Függetlenség esetén az együttható 0 (visszafelé nem igaz)
- Elméleti korrelációs együttható: $R(X,Y) = \frac{cov(X,Y)}{D(X)D(Y)} = \frac{E((X-E(X))(Y-E(Y)))}{D(X)D(Y)}$
- Pearson tapasztalati korreláció: $r_{X,Y} = \frac{\sum_1^n (X_i-\overline{X})(Y_i-\overline{Y})}{(n-1)S_X S_Y} = \frac{\sum_1^n (X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum (x_i-\overline{X})^2 * \sum (y_i-\overline{Y})^2}}$
  - R-ben: `cor(x,y)`
    vagy `sum((x-mean(x))*(y-mean(y))) / ((length(x)-1)*sd(x)*sd(y))`



5.5.2. Regresszió

- Regresszió: két vagy több változó között fennálló kapcsolat modellezése
  - Egyszerű lineáris regresszió: két változó irányított lineáris kapcsolata
- $y_i = a + bx_i + \epsilon_i$ ahol y függő/eredmény és x magyarázó
  - $E(\epsilon) = 0$ és $D^2(\epsilon) = \sigma^2 < \infty$ (normál eloszlás)
  - X legyen hiba nélküli (vagy elhanyagolható hibájú)
- Bármely jelölésen kalap: konkrét értéket jelent (pl. $\widehat{\epsilon_i} = y_i - \widehat{y_i}$)
- Legkisebb négyzetek módszer: $\min \sum_{i=1}^n (y_i - (a + bx_i))^2$

- Reziduális: becsült és valós y függőleges távolsága adott x esetén
- Hiba szórásnégyzet becslés: $\widehat{\sigma}^2 = \frac{\sum(y_i - \widehat{y_i})^2}{n-2}$ (residual standard error)
- Determinációs együttható:
  - $0 \leq R^2 \leq 1$, minél nagyobb, annál jobb a modell
  - Megadja, hogy Y változásainak hány százalék magyarázza a modell
  - Sima: $R^2 = 1 - \frac{\sum(y_i - \widehat{y_i})^2}{\sum(y_i - \overline{y})^2}$
    - Nő, ha több magyarázó változót használunk, tehát nem ideális
  - Korrigált: $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p}$ ahol p a változók száma (itt: 2)
- R-ben: `summary(lm(függő ~ magyarázó))` (két vektor)
  - `Residuals`: kevés adat esetén az értékek, sok esetén összesítés
  - `Intercept` sor: a értéke (függőleges eltolás)
  - `Estimate` oszlop: a, b értéke (szorzó)
  - `Pr(>|t|)` oszlop: mennyire fontos ez a változó a modellbenó
    - H_0: lehetne 0, el lehetne hagyni a változót
    - H_1: nem 0, azaz fontos a változó
    - `Pr(>|t|)` $< \alpha = 0.05 \implies$ fontos a változó, H_0 elvetve
  - `Residual standard error`: reziduális szórás becslése, $\sqrt{\widehat{\sigma}^2}$
  - `Multiple/Adjusted R-squared`: (korrigált) determinációs együttható
- Hasznos R függvények:
  - Legyen `reg <- lm(y ~ x)`
  - Ábrázolás: `plot(x,y); lines(x, reg$fitted.values)`
  - Kiszámolás adott x-re: `reg$coefficients[1] + adottX * reg$coefficients[2]`
- Manuális "megoldása" $y = a + bx$-nek
  - $\hat{b} = \frac{cov(X,Y)}{D^2(X)}$ (R-ben: `cov(x,y) / sd(x)^2`)
  - $D^2(\hat{b}) = \frac{\sigma^2}{\sum(x_i - \overline{X})^2}$
  - $\hat{a} = EY - \hat{b}EX$ (R-ben: `mean(y) -` $\hat{b}$`*mean(x)`)
  - $D^2(\hat{a}) = \sigma^2(\frac{1}{n} + \frac{\overline{X}^2}{\sum(x_i - \overline{X})^2})$

6. Előadás 12: lineáris modell, logisztikus regresszió, vegyes kapcsolat

- TODO 12. előadás
- Remélhetőleg nem lesz benne a ZH-ban: gyakorlaton se néztük

7. R jegyzet

7.1. Hasznos R függvények

- Indexek, ahol `TRUE` van: `which(x == max(x))`

7.2. Grafikonok, plot-ok

- `plot( c(...) )`
- `plot(x, y, type = "l", ...)`
- `plot(seq(from, to, 0.01), sapply(..., f), ...)`
- `barplot(f(0:100), names.arg=0:100, ...)`
- `boxplot(wt ~ cyl, data = mtcars, ...)`
  - $out: outlier values
- `hist(x, breaks=5)`
  - $counts: gyakoriságok osztályokban

7.3. Matematikai függvények

- sum(x), sort(x), min(x), max(x), round(x, 4)
- Mintaátlag, $\overline{X}$: `mean(x)`
- Korrigált tapasztalati szórás, S_n^*: `sd(x)`
- Korrigált szórásnégyzet, $(S_n^*)^2$: `var(x)`
- Tapasztalati k-adik momentum, m_k: `mean(x^2)`
- Statisztikák (min, max, átlag, kvartilisek): `summary(x)`
- Kvartilis: `quantile(x, probs = c(1/4, 1/2, 3/4), type = 6)`
- Tapasztalati eloszlásfüggvény: `plot(ecdf(x), ...)`

7.4. Adathalmaz

- *mtcars* egy adathalmaz, aminek van *cyl* és *wt* oszlopa
- `subset(mtcars, cyl == 4)$wt`
- `mtcars[mtcars$cyl == 4,]$wt`
- Érték-gyakoriság táblázat: `table(vektor)`
  - Oszlop, ahol <valami> igaz: `names(tábla)[tábla==max(tábla)]`

7.5. Táblázat, mátrix

```
ido <- matrix(c(15, 10, 5, 10, 10, 20, 5, 20, 5), ncol=3, byrow=TRUE)
colnames(ido) <- c("kevés", "átlagos", "sok")
rownames(ido) <- c("hüvös", "átlagos", "meleg")
ido <- as.table(ido)
```