Eddig: többváltozós térben bonyolult kérdések voltak:

- adatok mintázata → klaszterek
- csoportok-osztályok-klassifikáció
- folytonos változó modellezése

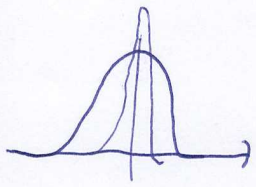
Valóság egyszerű kérdés egyváltozós térben:

Mit mérünk?

nyers adatok - scatter plot

túl sok adat: átlag, szórás, min-max, kvartilisek, hisztogram, boxplot, $\bar{y}(s)$ forma

Mennyi az értéke - átlagra való kérdések:



levezethető: $\sigma^2_{\bar{y}_N} = \frac{\sigma_y^2}{N} \rightarrow \frac{1}{\sqrt{N}}$ a szórása

átlagnak $\frac{1}{\sqrt{N}}$-re csökken a szórása

t-eloszlás: $\bar{y}$ átlagokból μ valós sokaságbeli átlagra

$$\frac{\bar{y} - \mu}{s/\sqrt{N}}$$ mintából

$\nu = N-1$ szabadsági fok

t-eloszlást követ

Gosset 1908 (student)

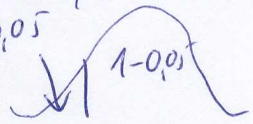
intervallumot adjunk meg ahová pl. 95% esik

$1 - 0{,}95 = 0{,}05 = \alpha$ szignifikancia szint mellett

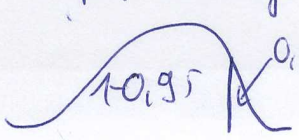
$$\bar{y} \pm \frac{s \cdot t^{-1}\left(\frac{\alpha}{2}, \nu = N-1\right)}{\sqrt{N}}$$

kétoldali eset ↑

egyoldali, pl. hatóanyag



pl. szennyező



lehet 0,01; 0,005; 0,001 ...

Statisztikai tesztek (próbák, hipotézis vizsgálat...)

döntési célra használjuk: állításról mondjuk meg, hogy igaz-e? NEM! Többségre: Csak annyit mondunk, hogy nem nemigaz valószínűleg....

- állítás, amire hivatkozni vagyok (populációra! nem a mintára)
- elméleti modell felállítása
- elméleti modellhez paraméterek (mérésből, máshonnan)
- adott eset $\Leftrightarrow$ elméleti modell valószínűségi összehasonlítása

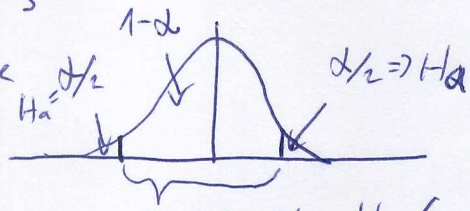
Példa: **z-próba**: $n>30$ minta y_i, E_0

$H_0: E=E_0$ $H_a: E\neq E_0$ (H_0 állításának ellentett eseménye)

mintából: $\bar{y}$ és s kiszámolása, majd

próba statisztika $\rightarrow$ $z_0=\frac{E_0-\bar{y}}{s}$ kiszámolása, E_0 hova esik, ha std. normál eloszlás lenne $z_0=\frac{\bar{y}-E_0}{\sigma/\sqrt{n}}$

σ ismert, ha $N>30$



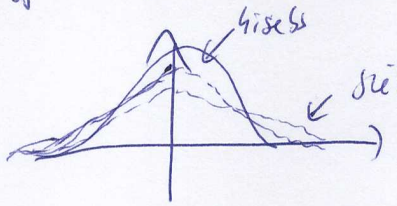
területet alulról integrálva

$0,025 \leq z_0 \leq 0,975$ H_0 ✓

tössi H_a

egymintás t-próba

ugyanez, csak a fv. a t-eloszlás $\gamma=N-1$-gyel



$t_0=\frac{\bar{y}-E_0}{s/\sqrt{n}}$

kétmintás t-próba

két mérési sor $H_0: E_1=E_2$ $H_a: E_1\neq E_2$

$d=\bar{y}_1-\bar{y}_2 \approx 0$

$t_0=\frac{d-E(d)}{s_d}$ vizsgálat

$\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{n}}$

két populáció átlaga egyezik-e? (pontosabban...)

ún. p-értéket itt többségre $0,05\leq p \Rightarrow H_0$

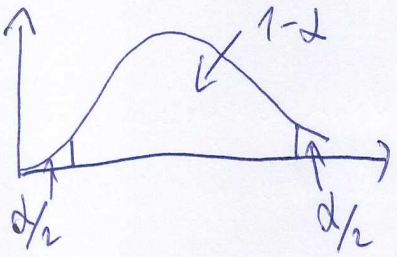
$> \Rightarrow H_a$

F-próba

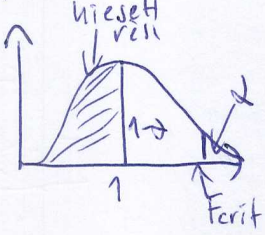
a két populáció varianciája egyezik-e (pl. két gép sorrája, laborán...)

$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$ $F_0 = \frac{s_1^2}{s_2^2}$

F-eloszlás $\nu_1 = N_1 - 1$ és $\nu_2 = N_2 - 1$ szabadsági fokok mellett



ha úgy kiptem, hogy $s_1^2/s_2^2 \geq 1$ legyen elővalogatással



$F_0 > F_{crit} \Rightarrow H_a$

régi könyvekben, most pl. $i H$

← integrál erre

p-value ← pl. $\alpha = 0.05$-nél $0,05 \leq p$ H_0

$>$ H_a

számít-e, hogy $p =$ pl. $0,2$; $0,5$; $0,9$?

Nem! Nem! p érték $\neq H_0$ valószínűsége!

Ha dönteni akarunk, legyen α előre megadva és csak $\alpha <> p$-t nézzük!

I. és II. fajú hiba:

döntés / Valóság	H_0 elfogadva	H_0 elutasítva
H_0 igaz	✓	1. fajú hiba ←
H_0 hamis	2. fajú hiba	✓

α-val tudjuk hol, de hogy romlik a 2. fajú hiba...

ANOVA - egyfaktor (analysis of variance ⇐ átlag miatt) ANOVA

alapkérdés: ha több csoportban kaptunk eredményeket, függ-e az átlagok szempontjából, melyikben

additív modell $y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ ← egyéni bizonytalanság

↑ mert; ↑ teljes várható érték; ↑ csoport várható értékének eltérése

$\mu + \alpha_j = \mu_j$

H_0: $\mu_1 = \mu_2 = \mu_3 = \dots$ vagy $\alpha_1 = \alpha_2 = \alpha_3 = \dots = 0$

H_1: létezik $\neq$

$\bar{y}$ ← teljes átlag $\quad \bar{y}_j$ ← csoport átlag $\quad$ n elem egy-egy csoportban, g csoport

$$SS_T = \sum_1^n \sum_j^g (y_{ij} - \bar{y})^2 = SS_{csop.belül} + SS_{csop.között} =$$

$$= \sum_1^n \sum_j^g (y_{ij} - \bar{y}_j)^2 + \sum_j^g n(\bar{y}_j - \bar{y})^2$$

mert ~~négyzetösszeg~~ négyzetösszeg/ν ⇓ variancia

szabadsági fokok: SS_T $n \cdot g - 1$ $\quad SS_{csb}$ $g \cdot (n-1)$ $\quad SS_{cs.k}$ $g-1$

⇓

$s_T^2 = SS_T/(n \cdot g - 1)$ ~~SS_{csb}~~/ $s_{csb}^2 = \frac{SS_{csb}}{g(n-1)}$ $\quad s_{csk}^2 = \frac{SS_{csk}}{g-1}$

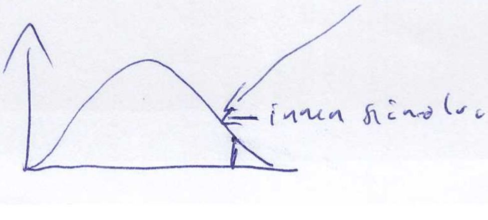
mindegyik F eloszlással írható le páronként osztva (normál eloszl.....)

$$F = \frac{s_{csk}^2}{s_{csb}^2} \Rightarrow F\text{-el. } \nu_1 = (g-1),\ \nu_2 = g(n-1) \leftarrow \text{kell között } F_{\nu_1, \nu_2}$$

táblázatban:

forrás	négyzetösszeg	ν	var	F	P(H_0) p érték
csop. között	$SS_{cs.k}$	$g-1$	s_{csk}^2	s_{csk}^2/s_{csb}^2	
csop. belül	$SS_{cs.b}$	$g(n-1)$	s_{csb}^2		
teljes	SS_T	$g \cdot n - 1$			

↑ pontosabban viszont! $p(X > F)$



~~további régi trükk~~: gyanús a $p > 0{,}9\dots$ is! miért?

Rengeteg teszt közelíthető a χ^2 eloszláshoz

$\chi^2 = \sum_{i=1}^{\nu} x_i^2$ $\quad x_i \in N(0,1)$ $\quad \nu$ = degrees of freedom = N

$E(\chi^2) = \nu$ $\quad \max \chi^2 \to \nu - 2$, ha $\nu > 2$

$\sigma^2(\chi^2) = 2\nu$

pl: Illeszkedés jósága egy adott eloszláshoz: (k darab)

H_0: minta A-ból származik

H_a: nem

A felosztása A_i részekre, hisztogram készítése

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(y_i - e_i)^2}{e_i}$$

y_i : mintából A_i-be esők száma

$e_i : N \cdot P(A_i)$ $\quad N = \sum_{i=1}^{k} y_i$

elvetés, ha $\chi_0^2 \geq \chi_{\alpha,\nu}^2$ $\quad \nu = k - 1 - r$

r = paraméterek száma az eloszlásban

Kontingencia táblázat

mért:

	dadog	pösze	
férfi	32	28	→60
nő	18	22	→40
	↓ 50	↓ 50	↓ 100

y_{ij}

függetlenség esetén várt

	d	p	
f	30	30	→60
n	20	20	→40
	↓ 50	↓ 50	

e_{ij}

H_0: A populációk ugyanazok a kategóriákban

nem

H_a:

vagy: H_0: Sor és Oszlop kategóriák függetlenek egymástól

H_a:

$$\chi_0^2 = \sum_{i} \sum_{j} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{\sum_k y_{ik} \cdot \sum_l y_{lj}}{\sum_k \sum_l y_{kl}}$$

függnek

elvetés, ha $\chi_0^2 \geq \chi_{\alpha,\nu}^2$ $\quad \nu = (\text{sorok} - 1)(\text{oszlopok} - 1)$

Másik gyakori: t-eloszlás

párosított adatokra, pl.: meredekségre

$y = ax + b$ modellnél ($y = ax + b + \varepsilon$)
↑ normál elosz.

$S_{yy} = \sum y_i y_i - \bar{y}\bar{y}$

H_0: $a = a_0$

H_a: $\neq$, vagy $>$, vagy $<$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - \bar{x}\bar{y}}{\sum x_i x_i - \bar{x}\bar{x}}$$

$$t_0 = \frac{a - a_0}{s_{xy}/\sqrt{S_{xx}}} \qquad s_{xy}^2 = (S_{yy} - a S_{xy})/(N-2)$$

elvetés:

$|t_0| \geq t_{\alpha/2, n-2}$, vagy $t_0 > t_{\alpha/2, n-2}$, vagy $t < -t_{\alpha/2, n-2}$

korrelációs e.h-ra

H_0: $r = 0$

$$t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

H_a: $r \neq 0$ vagy $r > 0$ vagy $r < 0$

elvetés: $|t_0| \geq t_{\alpha/2, n-2}$, $t \geq t_{\alpha, n-2}$, $t \leq -t$

létezik $r = r_0$ és $r_1 = r_2$ változatis, máshogy

tengelymetszetre

$$t_0 = \frac{b - b_0}{RMSE \cdot \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}}$$

$\nu = n - 2$ t-vel összevetve

Ha nem normál eloszlás:

páros t-teszt ⟷ Mann-Whitney u-test

anova ⟷ Kruskal-Wallis test

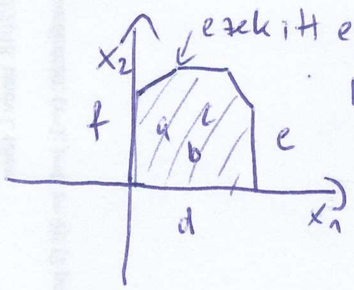
egymintás t ⟷ Wilcoxon előjeles rang test

Több kritériumú döntések (Multi Criteria Decision Making) MCDM/1

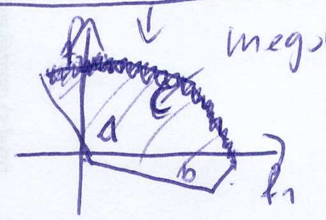
Optimális döntés keresése, ha több szempont van.

$x_1 \dots x_m$ kritérium $f_1(\underline{x}) \dots f_n(\underline{x})$ döntési függvény (min. v. max)

elfogadható tere definiálható a kritériumok terében, pl. 2db kritérium ← döntési fv.-ek terében



pl. $x_2 < 10$ (költség)
$x_1 < 40$ ⇒



megoldásokat itt keressük

lásd: Multicriteria decision analysis wikipedia

Hol vannak a pontok ⇒ a, b, c van belül b, c - értéke a legmesszebb az origótól, ha pl. f_1, f_2 max a kérdés, b vagy c

non-dominated pont (nem leuralható): ha nincs olyan, ami mindenben jobb:
pl. ~~nem~~ ~~kell~~ az ábra ↗ ≈ pareto hatékonyság = pontok, ahol nem javítható valamelyik tovább a jóléte, anélkül, hogy a többieknek romlana. „Pareto optimális pontok"
Pareto efficient/optimális pontok

Derringer módszer MCDM-re:

$x_i \in$ i-dik kritériuma egy adott esetnek (nagy legyen a jó)

x_{max}: legjobb x_{min}: legrosszabb $d_i = \dfrac{x_i - x_{min}}{x_{max} - x_{min}}$ vagy

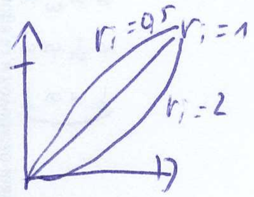
lehet úgy is, hogy x_{min}, x_{max} az értelmes határ $x_j < x_{min}$ $d_j = 0$
és akkor ha $x_j < x_{min} \rightarrow d_j = 0$ $x_j > x_{max} \rightarrow d_j = 1$

udott esetre

$$D = \sqrt[m]{\prod_{i=1}^{m} d_i^{r_i}}$$ számolása

$r_i = 1$ lineáris

$r_i = 0{,}5$, $r_i = 2$



max D a legjobb döntés

Topsis algoritmus

- $X^{n \times m}$ mátrix, n eset, m kritérium (X_{ij} nagy lehet rosszabb $\exists_-$, jó is $\exists_+$)
- normálás:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{k=1}^{n} x_{kj}^2}} \quad i = 1, 2 \dots n$$

- W_j eredeti oszlopsúlyból $w_j = \frac{W_j}{\sum W_k}$ normált oszlopsúlyok
- legrosszabb és legjobb elvi pont összeszedése

  A_w ← hipotetikus pont, összes kritériumból a legrosszabb az övé

  A_b ← hipotetikus pont, összes kritériumból a legjobb az övé
- n esetre távolság kiszámítása A_w-től és A_b-től

  pl: $d_{iw} = \sqrt{\sum_{j=1}^{m} (t_{ij} - t_{wj})^2}$ $\quad d_{ib} = \dots$
- $s_{iw} = \frac{d_{iw}}{d_{iw} + d_{ib}}$ ← hasonlóság a legrosszabbhoz

  s_{iw} rangsorba rakható, vagy a legjobb döntés

  $s_{iw} = 1$ elvi legjobb

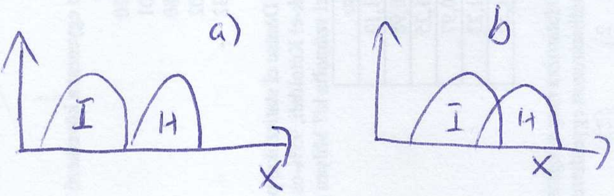
  $s_{iw} = 0$ elvi legrosszabb

Egy döntési fv létrehozása (lehetőleg előre!)

$f(\underline{w}, \underline{x}_i)$ alapján pl. maxra

ROC görbe (Receiver-Operating Characteristic)

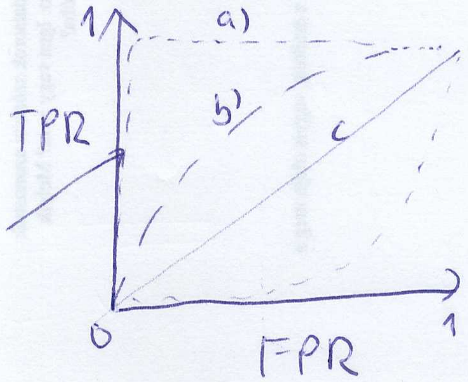
Klasszifikáció 1 változó szerint:



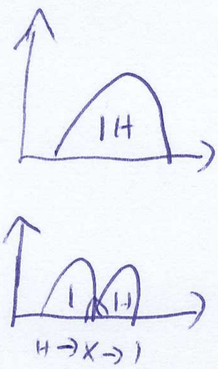
ne x-legyen a változónk egy görbén, hanem két függ x-től:

valós\döntés	igaz	hamis
igaz	TP	FN
hamis	FP	TN

$TPR = \frac{TP}{TP+FN}$ $FPR = \frac{FP}{TN+FP}$



a háttérben a döntési határ változik



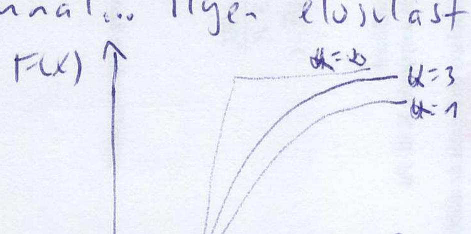
AUC = görbe alatti terület ⇒ 1 tökéletes a klassz.

Pareto-eloszlás

szerinte: 80%-a a javaknak 20%-nál a társadalomnak... Ilyen eloszlást sok minden követ.

Pareto index ↓

$$F(x) = \begin{cases} 1-\left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m \\ 0 & x < x_m \end{cases}$$



kumulatív!

Lehet pl. faktorokat is így kategorizálni →
(bemenő változók)

mit javítsunk...

Box-Cox transzformáció:

Szeretjük, ha a hiba/bizonytalanság homoszkedasztikus, vagyis nem függ pl. y értékétől. Átalakítjuk az eredetit:

$$\sigma_y = y^{\alpha} \Rightarrow y^* = y^{\lambda} \Rightarrow \mathrm{var}(y^*) = \text{konstans}$$

α	λ	$y^* =$
2	-1	$1/y$
1,5	-0,5	$1/\sqrt{y}$
1	0	$\ln y$
0,5	0,5	$\sqrt{y}$
0	1	y (nincs transzformáció)