BEVEZETÉS AZ INFORMATIKÁBA II.
SZÁMÍTÓGÉPES STATISZTIKA

Csendes Tibor
csendes@inf.u-szeged.hu

- 1 kredit
- heti 1 óra előadás + 2 óra gyakorlat (Bánhelyi Balázs)
- Az előadáshoz 3 fokozatú minősítés tartozik (1, 3, 5).
- A félév végén egy 1 órás dolgozat lesz az előadás anyagából.
- A dolgozat javasolt időpontja 2005. május 3., kedd 16-17 óra, itt, az előadás helyszínén.
- Ez alapján megajánlott jegy, amit lehet vizsgával javítani.
- A jegyzet: Csendes Tibor: Bevezetés a Számítógépes Statisztikába. Kapható a Vitéz utcai jegyzetboltban, de van belőle a könyvtárban is. Elérhető még a

  `http://www.inf.u-szeged.hu/~csendes/stat.ps.gz`

  címen (egy kivonat).
- Az előadásra járni nem kötelező, de katalógus alapján + pont jár annak, aki rendszeresen jelen van.
- Aki követi az előadásokat, és a feltett kérdésekre felelni tud, újabb + pontokat kaphat.
- A tárgyhoz kapcsolódik a később hallgatott Statisztika tárgy.
- A tárgy bevezetést ad az SPSS nevű statisztikai program használatába, és az ehhez szükséges alapvető statisztikai fogalmakat ismerteti meg.
- Felmentést az kaphat, aki az SPSS programot a leadott anyagnak megfelelő szinten ismeri.
- Motiváció: ebből is meg lehet élni...

A STATISZTIKA

„van hazugság, nagy hazugság és statisztika"

„elegendő számú adatból statisztikával bármit ki lehet mutatni"

„csak abban a statisztikában hiszek, amit magam hamisítottam"

A statisztikai eljárások nem elég gondos, nem elegendően körültekintő használata esetén megkérdőjelezhetetlennek tűnő hibás eredményeket kaphatunk.

A statisztikai programcsomagok ismertetése során a leggyakoribb hibalehetőségeket is megtárgyaljuk az elkerülésükhöz szükséges lépésekkel.

A statisztika szó jelentései:

- Ez a neve a szélesen értelmezett diszciplinának, tudományágnak, bár ezen belül is van általános statisztika, matematikai statisztika stb.),
- így hívják a statisztikai eljárásokat (... végrehajtani a statisztikát...)
- adatgyűjtést is értenek ez alatt (készítsen statisztikát...).

A mi szóhasználatunkban a *statisztika* olyan eljárásokkal foglalkozik, amelyek mérési adatok, felmérésekre kapott válaszok vagy más véletlen eseményektől függő adatok jellemzőit vagy összefüggésük mértékét és jellegét határozzák meg.

Ide tartozik a kapott eredmények olyan megjelenítése is, amely az adatok értelmezését megkönnyíti.

`http://www.inf.u-szeged.hu/~csendes/statfolia.pdf`

SZÁMÍTÓGÉPES STATISZTIKA

A jegyzet címében a számítógépes jelző arra utal, hogy közvetlenül nem a statisztika fogalmaival, összefüggéseivel foglalkozunk, hanem statisztikai eljárások, próbák, mutatók konkrét adatokra való meghatározásával.

A statisztikai programcsomagok ismertetése során a leggyakoribb hibalehetőségeket is megtárgyaljuk az elkerülésükhöz szükséges lépésekkel.

Néhány statisztikai eljárás más jellegű programban is elérhető, így például gyakran táblázatkezelő programban, vagy általános numerikus programcsomagokban is találunk ilyeneket:

Excel, StarOffice, As-Easy-As, Matlab, Maple, Mathematica, R stb.

Az egyszerűbb statisztikai programok, mint a **SigmaStat** is, csak egyváltozós statisztikákat képesek kiszámolni, cserében viszont könnyen kezelhetők és kisebb kapacitású gépen is futtathatók, olcsóbbak.

A statisztikai eljárások közel teljes körét rendelkezésre bocsátó professzionális programokból sok van, ezeket főleg PC-n vagy munkaállomásokon használhatjuk. Ide tartozik a részletesen tárgyalt **SPSS** mellett például a **StatGraphics**, a **Statistica**, a **BMPD** és az **SAS**.

Ezen osztály által kínált algoritmusok köre nem nagyon tér el, és bár a használatuk nagyon különböző lehet, a céljainkra elegendő ezek közül egyet ismertetni.

Linux operációs rendszerhez is számos programot lehet találni. Egy bő lista van ezekről a `http://chps06.ch.unito.it/linux/A/3` internetes címen (további linkekkel és rövid ismertetéssel minden programról).

STATISZTIKAI ALAPFOGALMAK

A *valószínűség:* egy 0 és 1 közötti szám ($0 \leq p \leq 1$), amely azt jellemzi, hogy egy esemény bekövetkezte milyen eséllyel, gyakorisággal várható.

Az 1 valószínűség csaknem biztos bekövetkezést, a nulla valószínűség csaknem lehetetlen előfordulást jelent. (A köznyelvben itt használhatunk biztos, illetve lehetetlen előfordulást is, a „csaknem" a matematikai pontosság kedvéért áll itt.)

A kísérletezés során tapasztalt *relatív gyakoriság*ok megközelítik az elméleti valószínűséget.

Az adatokat általában egy táblázatban célszerű elrendezni.

Az *eset* az összetartozó statisztikai adatok olyan egysége, amelyek amiatt képeznek egységet, mert egy egyedre, vagy mérési kísérletre vonatkoznak (pl. a kísérletben résztvevő személy, állat, vegyület stb.). Az eseteket általában egy-egy számítógépes rekordban, rendszerint a táblázat soraiban adjuk meg.

A tulajdonságokat, jellemzőket az egyes egyedekre vonatkozóan a *valószínűségi változó*k (röviden változók) tartalmazzák.

Az esetekre vonatkozó változóértékek alkotják a *statisztikai mintá*t, vagy röviden *mintá*t. Sok esetben jellemző az, hogy a teljes sokaságból csak kevés egyedre vonatkozó adat áll rendelkezésre.

PÉLDA: statisztikai mintának tekinthetjük a szavazási hajlandóságot, illetve a választási preferenciákat vizsgáló közvéleménykutatás alapadatát. Az eseteknek ekkor egy-egy megkérdezettre vonatkozó adathalmaz felel meg, míg a feltett kérdésekre kapott válaszok változók értékeit adják. A válaszadók átlagéletkora például egy olyan statisztikai mutató, amit a fenti értelemben statisztikának is szoktak röviden nevezni.

STATISZTIKAI ALAPFOGALMAK 2.

PÉLDA: egy új gyógyszer hatásosságának vizsgálatára gyűjtött adatsor feldolgozása. Ilyenkor két csoportra szokás osztani a pácienseket, az egyik csoport kapja a vizsgálandó kezelést, a másik (az ún. kontroll csoport) hatástalan gyógyszert kap — hogy valóban csak a szer hatását mérjük, ne az egyéb, pl. pszichés következményeket.

A betegenként gyűjtött adatok tartoznak egy esethez, a mért értékek pedig egy-egy változóhoz. Olyan statisztikát szokás vizsgálni, mint a megcélzott mérhető értékek átlagos eltérése a csoportok által reprezentált sokaságok között.

A táblázatkezelő programok az eseteket sorokban, a változókat oszlopokban tárolják. Ezt követik a statisztikai programok is. Másrészt több statisztikai eljárás szempontjából az esetek és a változók szerepe felcserélhető (mint pl. a klaszterezés esetén). A legtöbb statisztikai feldolgozás nyilvánvalóvá teszi, hogy mik lesznek az esetek és mik a változók.

A VALÓSZÍNŰSÉGI VÁLTOZÓK TÍPUSAI

A valószínűségi változók típusa fontos a végrehajtandó eljárás szempontjából, és az előzetes adatkezelést is befolyásolja. Alapvetően két típust különböztetünk meg:

1. *diszkrét valószínűségi változó* által felvehető értékek száma véges (vagy megszámlálhatóan végtelen, mint pl. az egész számok halmaza), vagy
2. *folytonos valószínűségi változó*: amely a valós számok halmazának egy vagy több intervallumán bármely értéket felvehet. Más szóval adott határok között bármely valós értéket felvehet (ilyen például a valós számok halmaza 0 és 1 között).

PÉLDA: Az előbbire a megfelelt – nem felelt meg – kiválóan megfelelt minősítés, illetve a kék – zöld – piros színhármas. Az utóbbi csoportba tartozik a testmagasság, a termésátlag, vagy az autók fogyasztása.

A diszkrét változókon belül van a *bináris* vagy *dichotom változó*k (alternatív ismérvek) csoportja: ezek csak két értéket vehetnek fel (pl. igen – nem, vagy férfi – nő).

ADATTÍPUSOK

Az adatokat először is a jelentésük jellege alapján lehet osztályozni: eszerint az adat *kvalitatív* vagy *kvantitatív* lehet.

A kvalitatív (vagy minőségi) adattípus az objektumok fajtáit adja meg (pl. neme: férfi – nő).

A kvantitatív (vagy mennyiségi) adattípus a számmal kifejezhető jellemzőket mutatja (pl. életkor, jövedelem).

Az adatok ilyen osztályozása általában természetes, könnyen megadható, mégis, ha az adatokat számokkal kódoljuk, akkor ezek a típusok csak a jellemzők eredeti jelentése alapján határozhatók meg. Így egy 100-as adatérték lehet mérési eredmény (tehát kvantitatív típusú), de például színkód is, ami pedig kvalitatív adattípusnak felel meg.

A *mérési skálák* (vagy mérési szintek) részletesebb osztályozást adnak az adatokra. Ezek mondják meg, hogy az adatainkat pontosan hogyan szabad értelmezni, milyen összefüggéseket használhatnak a statisztikai eljárások. Ennek megadása döntően befolyásolhatja az eredményünket, és emiatt ez komoly hibalehetőséget is jelent.

MÉRÉSI SKÁLÁK

Az alkalmazandó mérési skálát a statisztikai program nem tudja maga kiválasztani, mindenképpen a felhasználó, illetve a program kezelője segítségére lesz szükség.

Ezért ennek az osztályozásnak a megfelelő ismerete elengedhetetlen a statisztikai programok megbízható használatához.

Bár erről megkérdezhetjük a végső felhasználót, vagy kideríthetjük a szűkebb szakmában szokásos, elfogadott osztályozást, de ezt magunk is tisztázhatjuk.

Másrészt számos később ismertetendő részletkérdésben mindenképp a szakterület elfogadott módszertanára kell támaszkodnunk, így ebben az esetben törekedni kell az önálló döntésre.

Legyen A és B két objektum, x egy változó, x_A és x_B pedig az x változó értékei A és B esetén. A következő skálatípusokat tárgyaljuk (amelyek ebben a sorrendben tartalmazzák egymást):

1. A *névleges (vagy nominális) skála* minden értéke egy önálló kategóriát jelöl, az objektumok között csak az azonosság vagy különbözőség viszonyát tételezi fel (pl. a nem, szín, születési hely). A-ról és B-ről csak annyit tudunk, hogy $x_A = x_B$ vagy $x_A \neq x_B$. Ez a legkevésbé informatív mérési skála.

   Ennek esetében tehát hiába kódoltuk az adatokat számokkal, azokkal a szokásos műveleteket nincs értelme elvégezni, hiszen az eredeti információ-tartalom azt nem engedi meg (két színnek nincs pl. sorrendje). Ennek megfelelően az adatunkra vonatkozóan mindig a legtöbb információt nyújtó érvényes mérési skálát kell megadni.

MÉRÉSI SKÁLÁK 2.

2. A *sorrendi (vagy ordinális, ill. rang-) skála* esetén az objektumok között az azonosságon kívül nagyságrendi, illetve sorrendi különbséget is megállapíthatunk (például jó – közepes – rossz, magas – alacsony). A-ról és B-ről mondhatjuk, hogy $x_A < x_B$ vagy $x_A = x_B$ vagy $x_A > x_B$.

   A statisztikai programok gyakran támogatják ezt a mérési skálát, és a rá vonatkozó eljárások természetesen eltérnek a többi mérési skálán mért változókra írtaktól.

3. Ha az adatainkat *intervallum (vagy különbségi) skálán* mérhetjük, akkor a különbségek mértékét is értelmezhetjük (például a hőmérséklet, a dátum). Ha $x_A > x_B$, akkor B az A-tól $x_A - x_B$ egységgel különbözik.

   Ez a skálatípus már a legtöbb magasszintű statisztikai eljárást megengedi, ebben az értelemben ennek megléte már nem nagyon korlátozza a végrehajtható algoritmusok körét.

4. Az *arányskálán* az előbbieken túl még értelmezhető kezdőpont is van, tehát két objektum között nemcsak a különbséget, hanem az arányt is megállapíthatjuk (pl. a sorszámok, a fizetés, az életkor). Ha $x_A > x_B$, akkor az A objektum x_A/x_B-szer nagyobb, mint B. Az arányskálát nevezhetjük a legmagasabb mérési szintnek.

   Ismét meg kell jegyezni, hogy a számmal való kódolás miatt természetesen minden esetben van ugyan kezdőpont (hiszen kódolásra használt valós számok arányskálának felelnek meg), de a lényeges kérdés, hogy a mért mennyiségre értelmezhető-e ez, illetve hogy annak kitüntetett szerepe van-e a feldolgozás szempontjából.

Az utóbbi két mérési skálát együttesen *metrikus skálá*nak szokás nevezni. A minőségi ismérvek többnyire névleges skálán mértek (de nem mindig), a mennyiségi ismérvek pedig általában ennél erősebb mérési skálához tartoznak. A statisztikai programok nem minden mérési skála megadását teszik lehetővé, például az SPSS a scale, ordinal és nominal lehetőségeket adja meg (az első az arány- és intervallum skálát is fedi).

A MINTA JELLEMZŐI

A statisztikai feldolgozás adatait más szempontból is lehet jellemezni. A *sokaság vagy populáció* a statisztikai vizsgálat egyedeinek összessége (halmaza). Ennek minden elemre kiterjedő teljeskörű vizsgálatát nem mindig lehet, vagy nem gazdaságos elvégezni. Ilyen statisztikai sokaság például a szavazati joggal rendelkezők köre, vagy egy más feldolgozásban egy gyógyszerkísérletben résztvevő személyek csoportja.

A *statisztikai minta* ezzel szemben a vizsgált sokaságból kiválasztott egyedekhez tartozó megfigyelési adatok halmaza, részsokasága. Mintavételnél fontos szempont a *reprezentativitás* (azaz a kiválasztott mintának jól kell reprezentálnia a vizsgálni kívánt sokaságot az adott vizsgálatok szempontjából), és a *függetlenség* (ugyanazon egyed többszöri mérése nem független adatokat eredményez — a minta elemszámát így nem szabad növelni). *Cenzorált minta* az, amikor az eredeti minta elemeinek csak egy részét használjuk fel a következtetések levonásához.

Az alábbiakban röviden áttekintjük a minta legfontosabb jellemzőit, egyszerű definíciókkal, illetve ahol kell, rövid magyarázattal. A minta egyszerű jellemzői elsősorban a statisztikai feldolgozás első fázisában hasznosak, amikor a feldolgozandó adatok helyességét kell megállapítani. Ehhez nagy segítséget adnak a mért mennyiségek várt mutatói és a ténylegesen feldolgozott számokra adódó mutatók esetleges eltérései. Ez persze inkább nagyobb adatmennyiség esetén jelentős, kevés adatot könnyen össze lehet vetni akár teljes egészében is. Erre nagy adathalmaz esetén nincs reális lehetőség.

A minta eloszlásának (a folytonos változó értékei elhelyezkedésének) megjelenítésére általában *hisztogram*ot használunk. Utóbbi előállításához a legkisebb és a legnagyobb mintaelem közti különbséget valahány (általában 5-nél több) intervallumra osztjuk. Ezután készítünk egy ábrát, amelyben az intervallumokra olyan magas téglalapokat rajzolunk, mint ahány megfigyelés abba az intervallumba esik. Minél több a mintaelem, és minél több az intervallumok száma, a hisztogram annál jobban megközelíti az elméleti eloszlást. Ha ez az elméleti eloszlás a harang-görbe (Gauss-görbe), akkor azt mondjuk, hogy a minta *normális eloszlású* populációból származik.

A MINTA JELLEMZŐI II.

Az *elemszám* a statisztikai mérések száma (az esetek száma). A *hiányzóadat kódok* (missing values) olyan értékek, amik az illető változó lehetséges, értelmes értékei között nem fordulnak elő, de annak legszűkebb ábrázolásába belefférnek. Például a cipőméretek hiányzóadat kódja lehet a 99. A hiányzóadat kód figyelembevételével szokás külön megadni az *érvényes esetek számá*t is (number of valid cases).

Az adatgyűjtés során nyilván üresen maradhat a hiányzó adatok helye, de a számítógépes bevitel, tárolás során nem helyes, ha a véletlenre bízzuk, hogy milyen szám rendelődik a szóköz(ök)höz. Ha a hiányzóadat kódok elkülönített kezelését nem oldjuk meg, akkor olyan hibák adódhatnak, hogy például egy átlagba mondjuk 0 értékkel beleszámítódik a hiányzó érték is, és így irreális, a valós helyzetet nem tükröző eredményt kapunk.

A hiányzó adatok helyes kezelése a statisztikai feldolgozás egyik kritikus eleme, ami a látszólagos lényegtelenség és egyszerű érthetőség miatt is komoly csapdát jelent. Másrészt a korrekt hiányzóadat kód használata számos statisztikai eljárás eredményességét, leíró erejét tudja lényegesen javítani.

Példánkban a tanulmányi átlagra a 9.9 jó hiányzóadat kód, mert a szokásos értékek hosszába belefér, és mégsem értelmes adat, ilyen érték az adatgyűjtés során nem adódhat.

Rossz lenne viszont a tanulmányi átlag hiányzóadat kódjának pl. 3.3, hiszen ez előfordulhat értelmes adatként, és 999.9 is, mert ez pedig túl hosszú, egyes megjelenítések, feldolgozások során gondot okozhat. Hasonlóan helytelen lenne a tanulmányi átlag hiányzóadat kódja számára a ”hiány” szöveg bevitele, mert ennek típusa nem egyezik meg az eredeti változóéval.

KÖZÉPÉRTÉKEK

Egy változó *középértéke* a gyakorisági eloszlás helyzetét tömören, egy számmal kifejező érték, azonos mértékegységű adatok olyan jellemzője, amelytől azt várjuk, hogy közepes helyzetű, könnyen meghatározható és értelmezhető legyen. A középérték mértékegysége megegyezik a jellemzett változóéval. Ide tartoznak a helyzeti középértékek: a módusz és a medián, valamint a számított középértékek vagy átlagok, mint pl. a számtani átlag.

Az *átlag*, vagy *számtani átlag* egy adott mennyiségi, metrikus változó értékei összege osztva az elemszámmal (angolul mean).

A mintaelemeket nagyság szerint rendezve a középső elem (páratlan számú elem esetén), vagy a két középső elem átlaga (páros számú elem esetén) a *medián* (rövidítése Me). Ebben az értelemben ez a minta közepe. Más szóval az a szám, aminél a mintaelemek 50%-a kisebb vagy egyenlő. Angolul egyszerűen median.

A *módusz* a leggyakrabban előforduló érték(ek). A később ismertetendő normális eloszlás esetén az átlag, a módusz és a medián egybeesik. Az angol neve mode.

PÉLDA. Ha egy y változó értékei 5, 2, 3, 4, 4 és 1, akkor az ezekre vonatkozó átlag $3,1\dot{6}$ a módusz 4, a medián pedig 3,5.

FELADAT.

Mutassunk olyan rövid adatsort, amelynek átlaga 2, módusza 3, mediánja pedig 4!

Mutassunk olyan rövid adatsort, amelynek átlaga, módusza és mediánja megegyezik!

AZ ELOSZLÁS JELLEMZŐI

Itt a statisztikai változók, vagy más szóval ismérvek további jellemzőit ismertetjük röviden. A *szórás* (angol rövidítése SD a standard deviationből) a minta szórása, azaz a minta elemeinek az átlagtól való eltérésének négyzetes átlaga.

Normális eloszlás esetén az átlag $\pm$ 2 $*$ SD intervallumban található a mintaelemek 95,45%-a. A szórás (elméleti) és a korrigált tapasztalati szórás:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}, \quad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

ahol x_i az i-edik értéke az x valószínűségi változónak, és $\bar{x}$ a mintaelemek átlaga — inkább csak arra az esetre, ha számítógép nélkül kellene meghatározni.

Bár a szórást a legtöbb kalkulátor közvetlenül is meg tudja adni. Tekintsük az y változót, amelynek értékei 5, 2, 3, 4, 4 és 1. Ennek szórása $\sigma = 1,3437$, illetve $s = 1,4720$. A szórás négyzete, a *szórásnégyzet* is gyakran használt mutató.

Becsléskor a becslőfüggvény szórását az illető becslés *standard hibájának* (SE vagy SEM) nevezzük. Átlag esetén ez az átlag szórása. Ez azt fejezi ki, hogy az adott részminta alapján kapott átlag mennyire jól közelíti a valódi populáció átlagot. Az átlag $\pm$ 2 $*$ SE jelenti azt az intervallumot, amelyben a populáció átlaga kb. 95% valószínűséggel benne van.

A *relatív szórás* = (SD/átlag)$*100$. Megadja százalékos értelemben (mértékegység nélkül), hogy a szórás hányszorosa az átlagnak. Relatív jellege miatt alkalmas a különböző nagyságrendű változók szórásának összehasonlítására.

AZ ELOSZLÁS JELLEMZŐI II.

A *percentil* vagy *percentilis* a mediánhoz hasonló mutató a minta jellemzésére. A P_{25} 25%-os percentil pl. az a szám, aminél a mintaelemek 25%-a kisebb (vagy egyenlő). Az 5, 2, 0, 3, 1, 4, 6, 8, számokra P_{75} értéke 5, mert ennél nem nagyobb számból pont 6 van ($8 \times 0,75$).

Ha az értékkészletet nem száz, hanem 4 részre osztjuk, akkor kvartilisről (Q_i), ha tízre, akkor decilisről (D_i) beszélünk. Az ilyen mutatók összefoglaló neve a *kvantilis*.

A mennyiségi jellegű minta *terjedelme* a legnagyobb és a legkisebb mintaelem közötti különbség. Hasznos lehet a hibásan bevitt adatok kiderítéséhez. Az előző bekezdésben említett minta terjedelme 8 - 0 = 8.

A *ferdeség*, vagy *ferdeségi együttható, aszimmetria* egy mérőszám arra, hogy az eloszlás szimmetrikus-e vagy ferde. Negatív ferdeségi együttható esetén baloldali (negatív) ferdeségről van szó, ekkor az átlagnál nagyobb értékek a gyakoribbak.

A *lapultság* (kurtózitás) is az eloszlás egy alaki tulajdonságát fejezi ki: ha ez a mutató pozitív, az azt jelenti, hogy az eloszlás a normális eloszláshoz képest csúcsosabb, negatív esetben pedig lapultabb. Ennek megfelelően szokásos a *csúcsosság* név is.

ELOSZLÁSOK

A valószínűségi változók *eloszlásfüggvénye* azt mutatja meg, hogy ezek a változók milyen valószínűséggel vesznek fel egy adott számnál kisebb értéket: $F(x) = P(\xi < x)$, ahol P a $\xi < x$ esemény valószínűsége. Az $F(x)$ abszolut folytonos eloszlásfüggvény deriváltja $f(x)$, az ún. *sűrűségfüggvény*.

A sűrűségfüggvénnyel adott változók *várható értéké*t az

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

képlettel (ahol $f(x)$ a megfelelő sűrűségfüggvény), a diszkrét változókét a súlyozott középpel definiáljuk: $\sum_{i=1}^{n} x_i P(\xi = x_i)$. Az E betű az angol expectation szóra utal. A következőben a leggyakrabban előforduló, illetve a statisztikai feldolgozáshoz leginkább használatos eloszlásokat mutatjuk be röviden.

Binomiális eloszlás

Tekintsünk egy olyan kísérletet, amelynek két kimenetele van, A és B, és amelyek valószínűségei p és $q = 1-p$. Ekkor annak a valószínűsége, hogy n számú független kísérletből az A lehetőség pontosan k-szor következik be, $P_k = \binom{n}{k} p^k q^{n-k}$. A P_k valószínűségek n-edrendű p paraméterű *binomiális eloszlást* határoznak meg. A binomiális változó várható értéke np, szórásnégyzete npq.

ELOSZLÁSOK 2.

Poisson-eloszlás

A ξ diszkrét valószínűségi változót λ ($0 < \lambda < \infty$) paraméterű *Poisson-eloszlás*únak nevezzük, ha lehetséges értékei a nemnegatív egész számok, és

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-k}$$

teljesül ($k = 0, 1, 2, \ldots$). Várható értéke és szórásnégyzete is λ.

A binomiális eloszlás határeseteként lehet megkapni a kísérletek számának (n) növelésével és a p csökkentésével úgy, hogy az $np = \lambda$ szorzat állandó maradjon. Pontok térbeli vagy időbeli véletlen elhelyezkedése akkor követ Poisson-eloszlást, ha azok egymástól függetlenül minden térrészben vagy időszakaszban egyforma valószínűséggel oszlanak meg (pl. a vérsejtek száma a mikroszkóp látómezejében, radioaktív anyag adott idő alatt elbomlott atomjainak a száma).

Egyenletes eloszlás

Az egyik leggyakrabban használt eloszlás: lényegében azt fejezi ki, hogy a szóba jöhető alternatívák egyforma valószínűségűek. Diszkrét esetben, amikor a változó csak véges számú értéket vehet fel, ezek mindegyike egyenlő valószínűségű (mint például a kockadobás).

Folytonos esetben akkor beszélünk *egyenletes eloszlásról*, ha a változónak egy adott szakaszra, tartományra esésének a valószínűsége arányos a szakasz hosszával, illetve a tartomány mértékével. Az egyenletes eloszlású ξ diszkrét változó várható értéke $\frac{1}{n}\sum_{i=1}^{n} x_i$, és szórásnégyzete $\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$, amennyiben a felvehető értékei $x_1, x_2, \ldots, x_n$.

ELOSZLÁSOK 3.

Normális eloszlás

Egy valószínűségi változó *normális eloszlás*ú (jelölése $N(\mu, \sigma)$), ha az eloszlásfüggvénye

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt.$$

A binomiális eloszlás határeseteként is előáll a normális eloszlás, ha n növekedése közben p állandó marad. A képletében szereplő két paraméter a várható érték (μ) és a szórás (σ). A μ az eloszlás várható értéke, mediánja és módusza is egyben.

Független valószínűségi változók összegének az eloszlása közelítően normális eloszlású, ez biztosítja gyakori előfordulását. Hasonló okból, ha csak egyenletes eloszlású pszeudovéletlen-szám generátor áll rendelkezésre, akkor pl. n darab ($n > 10$) ilyen véletlen szám összege közelítőleg normális eloszlású véletlen számot ad. A standard normális eloszlás a 0 várható értékű, 1 szórású normális eloszlás ($N(0, 1)$).

Khi-négyzet eloszlás

A $\xi_1, \xi_2, \ldots, \xi_n$ független, standard normális eloszlású változók négyzetei összegének eloszlása n szabadságfokú *khi-négyzet (χ^2) eloszlás*. Ennek a várható értéke n, a szórásnégyzete pedig $2n$. Az előző szakaszban elmondottak miatt nagy n szabadságfok esetén alig tér el a normális eloszlástól.

Paraméter (vagy az eloszlás paramétere) az eloszlásfüggvényt meghatározó képletben szereplő valamely változó. Például a normális eloszlás paraméterei a várható érték (μ) és a szórás (σ).

Paraméteres módszer: olyan matematikai statisztikai módszerek összefoglaló neve, melyek paraméterrel vagy paraméterekkel (véges sok) leírható sokaságokra alkalmazhatók. Ebből adódóan nyilván vannak *nemparaméteres statisztikai eljárás*ok is, amelyek tehát nem a véges sok paraméterrel megadható eloszlásokon alapulnak. Hasonlóan a *paraméteres próba* a hipotézisvizsgálatnál az előírt parametrikus eloszlású sokaság valamelyik paraméterére vonatkozó próba.

Statisztikai becslés: a populáció eloszlásának valamely ismeretlen paraméterét egy alkalmas minta alapján közelítjük. A minta elemeit egy megfelelő formulába helyettesítve közelíthetjük a paraméter igazi értékét (pl. a populáció "elméleti" átlagát a mintaelemekből szokásos módon számolt átlaggal közelítjük).

Egy megfigyelés *szabadságfok*a a magyarázó rendszeren belül önkényesen megválasztható értékek száma, speciális esetben az egymástól független összeadandók száma.

Megbízhatósági intervallum (vagy *konfidencia intervallum, megbízhatósági tartomány):* olyan intervallum, amely (általában) nagy, előre megadott valószínűséggel tartalmazza a becsült paraméter valódi értékét.

STATISZTIKAI PRÓBÁK

Ez a szakasz a statisztikai próba felállításához és az eredmény kiértékeléséhez ad segítséget, összefoglalva a legfontosabb fogalmakat. A szokásos, gyakori hipotézisvizsgálatokat a statisztikai programok közvetlenül támogatják. A *statisztikai próba* olyan eljárás, amely valamilyen hipotézisnek (az alapsokaságra vonatkozó feltevésnek) az ellenőrzését teszi lehetővé a minta adatai és a próbafüggvény alapján.

A nullhipotézis: hipotézisvizsgálatban általában az a feltevés, hogy bizonyos különbségek vagy hatások a populációban adott értékkel egyenlők. Például, hogy két átlag különbsége 0, vagy az, hogy a korrelációs együttható nulla. De lehet az is a kiindulási feltevésünk, hogy pl. a várható érték 10.

Szignifikancia, szignifikáns eltérés: a nullhipotézistől való, adott szintet meghaladó eltérés. A *szignifikancia-szintet* általában valószínűséggel adjuk meg. Ez lehet pl. 5% (azaz $\alpha = 0,05$ annak a hibának a valószínűsége, hogy tévesen állapítottuk meg a különbséget, ha a nullhipotézis igaz). Ha tehát a próba eredménye $p < 0,05$, akkor ez azt jelenti, hogy szignifikáns különbséget vagy hatást állapítottunk meg. Ha százszor megismételnénk a kísérletet, a százból csak kb. 95 esetben kapnánk ugyanezt az eredményt, 5 esetben nem találnánk eltérést (elsőfajú hiba). A szokásos szintek: 5%, 1%, 0,1% (azaz $\alpha = 0,05$, 0,01, 0,001). A *megbízhatósági szint*ek ennek megfelelően 95%, 99% és 99,1%. A szignifikáns eredményt leggyakrabban a p-érték és a szignifikancia-szint (α) összehasonlításával szokás megállapítani. Ez a gyakorlat abból az időből származik, amikor csak táblázatok álltak rendelkezésre. Jelenleg egyre elterjedtebb magának a p értéknek a megadása.

Nem szignifikáns: $p > 0,05$ (p nagyobb, mint 0,05). Az 5%-os szinten nem szignifikáns különbség azt jelenti, hogy nem sikerült a különbséget kimutatni. Ez nem feltétlenül jelenti azt, hogy egyáltalán nincs különbség. Ha az eredmény nem szignifikáns, akkor lényegében semmit sem tudunk mondani a vizsgált jelenségről. Ebben az értelemben végül is elfogadjuk azt a nullhipotézist, hogy nincs eltérés. Az elkövetett hibáról csak annyit tudunk, hogy nagy mintaelemszám esetén elég kicsi, ha a nullhipotézis nem igaz (másodfajú hiba).

A STATISZTIKAI PRÓBÁKKAL KAPCSOLATOS TOVÁBBI ALAPFOGALMAK

Az *elsőfajú hiba* akkor fordul elő, amikor a nullhipotézist elvetjük, bár az igaz. Valószínűsége egyenlő a szignifikancia-szinttel (α).

A *másodfajú hibá*t akkor követjük el, amikor a nullhipotézist elfogadjuk, bár az nem igaz. Valószínűségét (β) nem ismerjük. Ha az elsőfajú hiba valószínűségét csökkentjük, a másodfajú hibáé nő, de $\alpha + \beta \neq 1$. Nagy mintaelemszám esetén általában a másodfajú hiba valószínűsége csökken.

Egyoldali próba amikor a nullhipotézissel szemben felállított *alternatív hipotézis*ben (ellenhipotézisben) csak egyirányú változást tételezünk fel.

Kétoldali próba: ekkor a nullhipotézissel szemben felállított alternatív hipotézisben minden irányú változást figyelembe veszünk.

A következő oldalon megadott szempontok és útmutatások új statisztikai próbák összeállításához és végrehajtásához adnak segítséget. Másrészt a leggyakoribb ilyen teszteket a tárgyalt statisztikai programok közvetlenül is támogatják, vagyis ekkor inkább csak az eredmények helyes értelmezéséhez, vagy a jó paraméterezéshez használhatjuk ezeket az ismereteket.

STATISZTIKAI PRÓBA VÉGREHAJTÁSA

A statisztikai próbák végrahajtásának a következő lépései vannak:

1. Az előzetes ismereteink alapján állítunk valamit, amit statisztikai módszerrel szeretnénk igazolni. Először a kiinduló hipotézist (H_0) kell felállítani, a nullhipotézist megfogalmazni. A nullhipotézisben sok esetben (de nem mindig) azt rögzítjük, hogy nincs változás.

2. Ezután az alternatív hipotézis (H_1) felállítása következik.

3. A következő lépés a próba szignifikancia-szintjének meghatározása ($\alpha = 0,05$, $\alpha = 0,01$, vagy $\alpha = 0,001$). Ezt az értéket az adott szakterület szokásos értékeihez kell igazítani.

4. Határozzuk meg ezután a használt véletlen minta elemszámát. Ezt idő-, illetve pénzkorlátok és előzetes ismereteink is meghatározzák, különben nyilván a nagyobb minta megbízhatóbb eredményt ad. Ezután jön a véletlen minta előállítása, és a próbastatisztika kiszámítása. (Az érintett változó a nullhipotézis fennállása esetén valamely ismert eloszlást követ.)

5. Meghatározzuk a döntési szabályt, és azt a kritikus értéket vagy értékeket (ha kétoldali próbát hajtunk végre), amelynél a mintából kiszámított próbastatisztika csak kis ($< \alpha$) valószínűséggel vesz fel nagyobb értéket.

6. Ha a kiszámított próbastatisztika a kritikus értéknél nagyobb (illetve az elfogadási tartományon kívül esik), akkor elvetjük a nullhipotézist, mivel egy kis valószínűségű esemény következett be (egyúttal elfogadjuk az alternatív hipotézist). Ilyenkor azt mondjuk, hogy az eltérés szignifikáns az α szinten ($p < \alpha$), az alternatív hipotézis teljesül.

7. Ha a kiszámított próbastatisztika a kritikus értéknél kisebb (illetve az elfogadási tartományon belül van), akkor megtartjuk a nullhipotézist és azt mondjuk, hogy az eltérés nem szignifikáns α szinten. Azt is mondhatjuk, hogy nem vetjük el a nullhipotézist, ami egy óvatos megfogalmazás, és arra utal, hogy a szignifikancia-szint függvényében általában nem állíthatjuk, hogy a nullhipotézis igaz.

A *korrelációs* eljárások két valószínűségi változó közötti összefüggés szorosságát mérik, ami aztán a predikció minősége mértékeként is használható. A regressziótól eltérően itt nem szükséges az egyik változó, mint eredményváltozó kijelölése. Az r korrelációs együttható egy -1 és 1 köztött változó szám.

Ha ennek értéke -1, akkor függvényszerű negatív lineáris összefüggés van a változók között, azaz amíg az egyik nő, addig a másik csökken. Ha a korrelációs együttható 1, akkor függvényszerű pozitív lineáris összefüggés van. A nulla korrelációs együttható pedig azt jelenti, hogy nincs lineáris összefüggés a változók között. Más érték esetén óvatos diszkusszió mellett a közelálló említett eseteknek megfelelő következtetést vonhatjuk le. Ha a kapott korreláció abszolútértéke nincs közel 1-hez vagy nullához, akkor nem állapíthatunk meg korrelációt.

A *regressziós* eljárás feltételezi, hogy olyan összefüggés van a magyarázóváltozók és az eredményváltozó között, hogy ha az adatokat térben ábrázoljuk, akkor egyenest, síkot, vagy adott típusú görbét kapunk megközelítőleg. A regresszió azt a paraméterezést keresi meg, amely a legjobb illesztést adja az aktuális adathoz.

A többváltozós lineáris esetben a magyarázóváltozók (nyilván többváltozós) lineáris függvényével modellezzük az eredményváltozó értékét. A regresszió egy paraméteres statisztikai módszer, amely feltételezi, hogy a reziduumok (a becsült és a tényleges eredményváltozó értékek közti eltérések) normális eloszlásúak. Mivel a regressziós együtthatók kiszámításakor a reziduumok négyzetösszegét minimalizáljuk, ezért szokás ezt az eljárást a legkisebb négyzetek módszerének is hívni.

EGY DOLGOZAT FELADATAI

1. Adjon példát mindegyik skálatípusra úgy, hogy épp az legyen a legjobb skálatípus, amely az adott mintára érvényes!
2. Határozza meg az átlag, a medián és a módusz értékét a következő adatsorra: 1, 2, 2, 3, 4, 5!
3. Indokolja a hiányzóadat kód fontosságát!
4. Határozza meg, milyen mérési skála felel meg a színeknek, a hőmérsékleti fokoknak, a fizetés összegének és a {nem felelt meg, megfelelt, kiválóan megfelelt} értékelésnek!
5. Mutassunk egy rövid mérési adatsort, amelyre az átlag, a medián és a módusz három különböző érték!
6. Milyen hiányzóadat kódot használna a cipőméretek megadásakor? Jelölje meg az alábbi öt lehetőség egyikét:

   semmilyent 99-et 42-est tetszőlegest -

Tartalomjegyzék

1. óra 1
2. óra 4
3. óra 6
4. óra 9
5. óra 11
6. óra 13
7. óra 15
8. óra 19
9. óra 22

Tartalomjegyzék **24**