Linearna kombinacija slučajnih varijabli

Neka imamo niz slučajnih varijabli X_1, X_2,...,X_n koje imaju redom očekivanja μ_1, μ_2,…, μ_n i standardne devijacije σ_1, σ_2,…, σ_n.

Def:

Za bilo koji skup slučajnih varijabli X_1, X_2,...,X_n i konstanti a_1, a_2,...,a_n, slučajna varijabla

$$Y= a_1X_{1+} a_2X_2+...+a_nX_n$$

zove se **linearna kombinacija** X_i-ova.

Očekivanje linearne kombinacije

Bez obzira jesu li X_i-ovi nezavisni ili ne, vrijedi:

$$E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$$

dokaz na ploči za n=2.

Varijanca linearne kombinacije

Bez obzira jesu li X_i-ovi nezavisni ili ne, vrijedi:

$$V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n) + \sum_i \sum_{j\neq i} a_i a_j Cov(X_i, X_j)$$

dokaz na ploči za n=2.

Ako su X_i-ovi nezavisni, vrijedi:

$$V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n)$$

Varijanca razlike nezavisnih varijabli

Ako stavimo n=2, a_1=1 i a_2=-1, dobivamo poseban slučaj:

$$V(X_1 - X_2) = V(X_1) + V(X_2) = \sigma_1^2 + \sigma_2^2$$

!!! Varijanca razlike je zbroj varijanci, a ne razlika!!!

Linearna kombinacija nezavisnih normalnih slučajnih varijabli

Propozicija:

Neka su $X_1, X_2,...,X_n$ nezavisne normalne slučajne varijable. Tada njihova linearna kombinacija

$$Y= a_1X_{1+} a_2X_2+...+a_nX_n$$

također ima normalnu raspodjelu s očekivanjem $\mu_Y = \sum_{i=1}^{n} a_i \mu_i$ i varijancom $\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$.

IV.4. SLUČAJNI UZORAK

Na prvom predavanju govorili smo o uzimanju uzorka iz cijele populacije. Elementi n-članog uzorka označeni su s $x_1, x_2,..., x_n$. Da smo uzeli neki drugi uzorak, dobili bismo druge vrijednosti x_i-ova. Općenito, prije uzimanja uzorka ne možemo znati vrijednost i-tog člana uzorka pa je to za nas slučajna varijabla X_i.

U mnogim statističkim problemima vrijednosti u uzorku od n elemenata $\{x_1, x_2,..., x_n\}$ iz neke populacije mogu se zamisliti kao opažene vrijednosti niza od n slučajnih varijabli. Neka su $X_1, X_2,...,X_n$ takve nove slučajne varijable.

Def: Neka su $X_1, X_2,...,X_n$ slučajne varijable. Kažemo da one čine **slučajni uzorak** veličine n, ako su varijable X_i
(a) nezavisne i
(b) imaju iste raspodjele vjerojatnosti.

Kažemo da su X_i **nezavisne i identično raspodijeljene.**

Primjena:

- mjerenje fizikalne veličine (populacija je beskonačna)
- uzimanje uzorka s povratom
- uzimanje uzorka iz velike populacije

nije ispredavano:
Svaka veličina koja se može izračunati iz podataka o uzorku naziva se **statistika**. Prije uzimanja uzorka ne možemo znati vrijednost statistike pa je to onda slučajna varijabla koju označavamo velikim slovom. Primjeri statistike su prosjek uzorka, standardna devijacija uzorka, momenti uzorka, medijan uzorka,…

IV.4.1. Prosjek i total slučajnog uzorka

Def:

Slučajna varijabla $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ naziva se **prosjek** uzorka $X_1, X_2,...,X_n$.

Def:

Slučajna varijabla $T_0 = \sum_{i=1}^{n} X_i$ naziva se **total** uzorka $X_1, X_2,...,X_n$.

Prosjek i total su linearne kombinacije nezavisnih varijabli

Očekivanje prosjeka i totala

$$E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$$

$E(\overline{X}) = \mu_{\overline{X}} = \mu$ $\qquad$ $E(T_0) = \mu_{T_0} = n\mu$

Varijanca prosjeka i totala

Ako su X_i-ovi nezavisni, vrijedi:

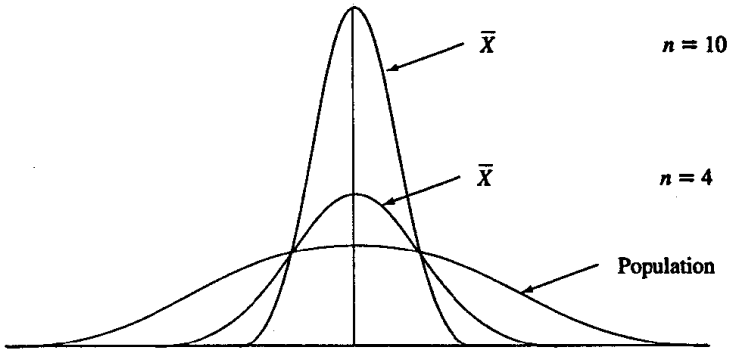
$$V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n)$$

$$V(\overline{X}) = \sigma_{\overline{X}}^2 = \frac{\sigma^2}{n} \qquad V(T_0) = \sigma_{T_0}^2 = n\sigma^2$$

Raspodjela prosjeka i totala

Za slučajne varijable prosjek i total znamo očekivanje i varijancu. Međutim, ne znamo cijelu raspodjelu. U specijalnom slučaju uzorka iz normalne raspodjele, raspodjela je normalna.

Ako imamo n-člani uzorak iz normalne raspodjele koja ima očekivanje μ i varijancu σ^2, onda je slučajna varijabla $\overline{X}$ normalno raspodijeljena s očekivanjem μ i varijancom $\frac{\sigma^2}{n}$.



U općenitom slučaju, raspodjela vjerojatnosti za statistiku slučajnog uzorka može se dobiti primjenom pravila vjerojatnosti, ako su raspodjela vjerojatnosti populacije i funkcija za izračunavanje statistike dovoljno jednostavne. U protivnom, provodi se simulacijski eksperiment pomoću računala. Ovdje navodimo dva jednostavna primjera raspodjele prosjeka i totala za dvočlani slučajni uzorak.

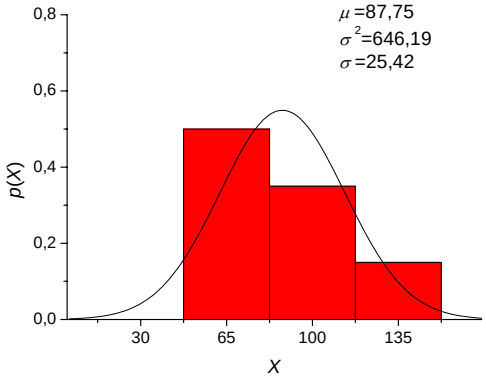
Primjer 1 (diskretna raspodjela):

Prodavač automobila prodaje 50% automobila niže klase po cijeni 65.000 kuna, 35% automobila srednje klase po 100.000 kuna i 15% automobila više klase po 135.000 kuna. Definirajmo slučajnu varijablu

X=prihod od prodaje jednog automobila.(u tisućama kuna)

Ona je diskretno raspodijeljena na slijedeći način:

x	65	100	135
$p(x)$	0,5	0,35	0,15



Određenog dana najavila su se dva kupca. Neka su slučajne varijable:

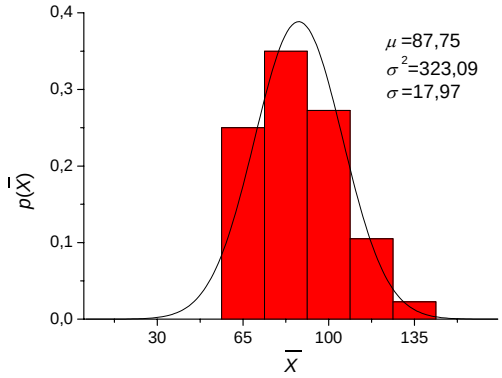
X_1=prihod od prvog kupca; X_2=prihod od drugog kupca

Tablica mogućih prihoda:

x_1	x_2	$p(x_1,x_2)$	x_1+x_2	$\overline{x}$
65	65	0,25	130	65
65	100	0,175	165	82,5
65	135	0,075	200	100
100	65	0,175	165	82,5
100	100	0,1225	200	100
100	135	0,0525	235	117,5
135	65	0,075	200	100
135	100	0,0525	235	117,5
135	135	0,0225	270	135

Slučajna varijabla $\overline{X}$ raspodijeljena je ovako:

$\overline{x}$	$p(\overline{x})$
65	0,25
82,5	0,35
100	0,2725
117,5	0,105
135	0,0225

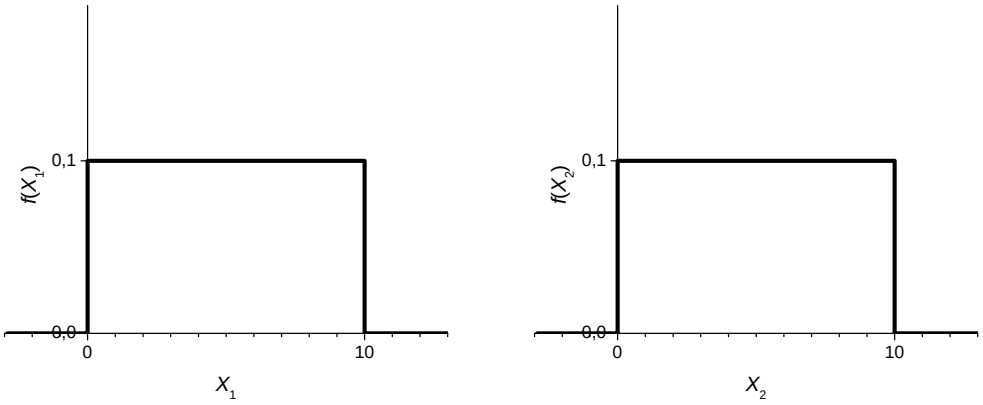


Primjer 2 (kontinuirana raspodjela):

Na putu do posla čekam autobus koji vozi svakih 10 minuta, a zatim tramvaj koji također vozi svakih 10 minuta.
Neka su slučajne varijable:

X_1 = vrijeme čekanja autobusa
X_2 = vrijeme čekanja tramvaja

$$f(x_1)=\begin{cases}1/10 & , \quad 0\le x_1\le 10\\ 0 & , \quad x_1<0 \quad ili \quad x_1>10\end{cases} \qquad f(x_2)=\begin{cases}1/10 & , \quad 0\le x_2\le 10\\ 0 & , \quad x_2<0 \quad ili \quad x_2>10\end{cases}$$



Dodatak (Izračun nije ispredavan)

Definirajmo slučajnu varijablu:
T_0 = ukupno vrijeme čekanja
i nađimo njezinu funkciju gustoće vjerojatnosti!

$f(t)dt$ je vjerojatnost da X_1+X_2 poprimi vrijednost u intervalu $(t,t+dt)$

$$f(t)dt = \int\limits_{\substack{(x_1,x_2)\\ x_1+x_2=t}} f(x_1)f(x_2)dx_1dx_2 =$$

x_1 je varijabla integracije, a supstituiramo $x_2=t-x_1$

$$= dt\int f(x_1)f(t-x_1)dx_1$$

Imamo dva područja u kojima je podintegralna funkcija različita od nule:
i) $0 < t < 10$
ii) $10 < t < 20$

- Najprije rješavamo $t<10$:

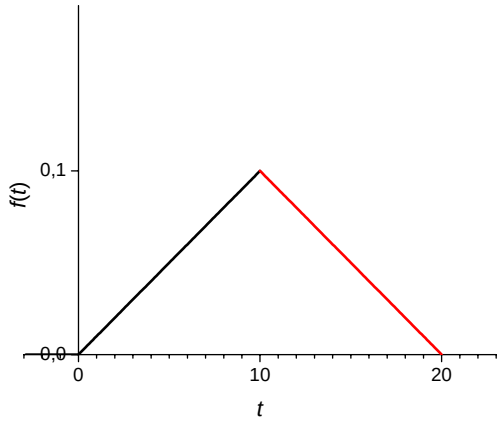
$$f(t) = \int\limits_{x_1=0}^{t} f(x_1)f(t-x_1)dx_1 = \int\limits_{x_1=0}^{t} \frac{1}{100}dx_1 = \frac{t}{100}$$

- Sada riješimo slučaj $t > 10$:

$$f(t) = \int_{x_1=t-10}^{10} f(x_1) f(t-x_1) dx_1 = \int_{x_1=t-10}^{10} \frac{1}{100} dx_1 = \frac{20-t}{100}$$

U cijelom području funkcija gustoće vjerojatnosti je:

$$f(x_1) = \begin{cases} \frac{t}{100} & , \quad 0 \le t \le 10 \\ \frac{20-t}{100} & , \quad 10 \le t \le 20 \end{cases}$$



Vidimo da prosjek dvočlanog uzorka pravokutne raspodjele ima trokutnu raspodjelu.

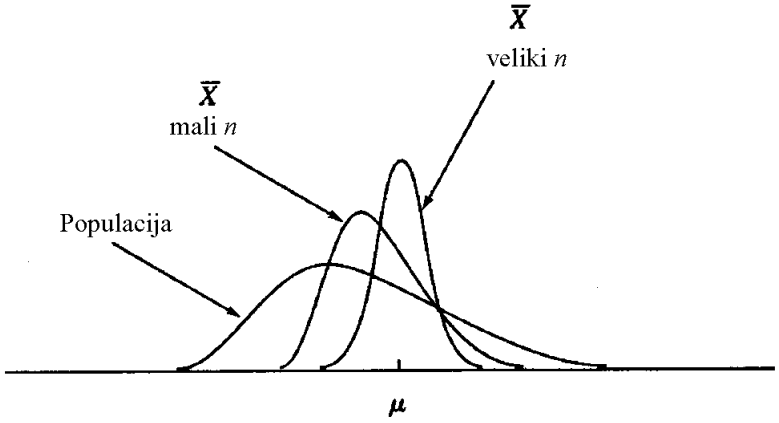
Primjer simulacijskog eksperimenta:
http://www.stat.sc.edu//~west/javahtml/CLT.html

Gornji primjeri zorno pokazuju općenito pravilo: Čak i kada populacija nije normalno raspodijeljena, uzmemo li veliki uzorak, prosjek i total bit će normalno raspodijeljeni. Upravo to nam kaže središnji granični teorem:

IV.4.2. Središnji granični teorem (CLT)

Teorem:

Neka je X_1, X_2,...,X_n slučajni uzorak bilo koje raspodjele s očekivanjem μ i standardnom devijacijom σ. Ako je n dovoljno velik, $\overline{X}$ ima približno normalnu raspodjelu s očekivanjem $\mu_{\overline{X}} = \mu$ i standardnom devijacijom $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$. T_0 ima također približno normalnu raspodjelu s očekivanjem $\mu_{T_0} = n\mu$ i standardnom devijacijom $\sigma_{T_0} = \sqrt{n} \cdot \sigma$.



Praktično pravilo: **CLT se može primijeniti kad je $n \geq 30$.**

V. PROCJENA PARAMETARA POPULACIJE NA OSNOVI UZORKA

Populacija koja nas zanima ima karakteristične parametre μ, σ, σ^2, momenti, proporcije itd. To su brojevi koji nas zanimaju, ali ih nismo u mogućnosti odrediti na cijelom skupu. Mogući razlozi:

1. Osnovni skup je prevelik (npr. svi birači u nekoj državi)
2. Ne možemo odrediti veličinu osnovnog skupa (npr. svi gledatelji neke TV emisije)
3. Osnovni skup je beskonačan (npr. sva moguća mjerenja neke fizikalne veličine)

Stoga uzimamo uzorak i na osnovu njegovih parametara želimo procijeniti tražene parametre osnovnog skupa

Def:

Neka je θ neki parametar raspodjele osnovnog skupa (μ, σ,...). Slučajnu varijablu $\hat{\theta}$ koju upotrebljavamo da bismo procijenili parametar θ nazivamo **procjenjitelj** tog parametra.

Zamislimo da imamo dva mjerna instrumenta. Jedan je dobro baždaren, a drugi sistematski daje manje rezultate. Rezultati mjerenja pomoću prvoga bit će raspodijeljeni oko prave vrijednosti mjerene veličine. Kažemo da je taj instrument nepristran. Drugi instrument je pristran. Slično možemo reći za procjenjitelje.

Def:

Za procjenjitelj $\hat{\theta}$ kažemo da je **nepristrani procjenjitelj** parametra θ ako je njegovo očekivanje

$$E(\hat{\theta}) = \theta$$

Propozicija:

Ako je X_1, X_2, ... X_n slučajni uzorak neke populacije, onda je slučajna varijabla $\overline{X}$ nepristrani procjenjitelj parametra μ te populacije.

(Dokaz: Već smo pokazali da je očekivanje raspodjele prosjeka uzorka jednako očekivanju osnovnog skupa.)

Propozicija:

Ako je X_1, X_2, ... X_n slučajni uzorak neke populacije koja ima očekivanje μ i varijancu σ^2, onda je slučajna varijabla

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

nepristrani procjenjitelj varijance osnovnog skupa σ^2.

Dokaz:

Za bilo koju slučajnu varijablu vrijedi $E(Y^2) = [E(Y)]^2 + V(Y)$, pa tako i za svaki član slučajnog uzorka vrijedi $E(X_i^2) = \mu^2 + \sigma^2$.

Primjenimo li to na $S^2 = \frac{1}{n-1}\left[\sum X_i^2 - (\sum X_i)^2 / n\right]$, dobivamo:

$$E(S^2) = \frac{1}{n-1}\left\{\sum E(X_i^2) - \frac{1}{n}E\left[\left(\sum X_i\right)^2\right]\right\} =$$

$$\frac{1}{n-1}\left\{\sum(\mu^2+\sigma^2) - \frac{1}{n}\left\{V\left[\left(\sum X_i\right)\right] + \left[E\left(\sum X_i\right)\right]^2\right\}\right\} =$$

$$\frac{1}{n-1}\left\{\sum(\mu^2+\sigma^2) - \frac{1}{n}\left\{V[T_0] + [E(T_0)]^2\right\}\right\} =$$

$$\frac{1}{n-1}\left\{\sum(\mu^2+\sigma^2) - \frac{1}{n}\left\{n\sigma^2 + [n\mu]^2\right\}\right\} = \sigma^2$$

čime smo dokazali da se radi o nepristranom procjenjitelju.

Koliko je očekivanje slučajne varijable $\hat{\sigma}_u^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$?

Nju možemo raspisati:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n} = \frac{\sum_{i=1}^{n}\left({X_i}^2 - 2X_i\overline{X} + \overline{X}^2\right)}{n} = \frac{\sum_{i=1}^{n}{X_i}^2}{n} - 2\overline{X}^2 + \overline{X}^2 = \frac{\sum_{i=1}^{n}{X_i}^2}{n} - \overline{X}^2$$

$$E(\hat{\sigma}_u^2) = \frac{1}{n}\sum_{i=1}^{n}E({X_i}^2) - E(\overline{X}^2) = \frac{1}{n}\sum_{i=1}^{n}(\mu^2+\sigma^2) - (\mu^2 + \frac{\sigma^2}{n}) = \sigma^2 - \frac{\sigma^2}{n}$$

Dakle $\hat{\sigma}_u^2$ je pristrani procjenjitelj

$$E(S^2) = \frac{n}{n-1}E(\hat{\sigma}_u^2) = \sigma^2 \quad \Rightarrow \quad S^2 \text{ je nepristrani procjenjitelj}$$

Nije ispredavano:

Da bismo shvatili značenje ove razlike, zamislimo da promatramo sve moguće uzorke iz velike populacije. Pogledajmo vrijednosti procjenjitelja za *k*-ti uzorak:

$\hat{\sigma}_{uk}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ik} - \overline{x}_k\right)^2$ odstupanje od prosjeka *k*-tog uzorka

$s_k^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ik} - \mu\right)^2$ odstupanje od prosjeka populacije.

Raspišimo:

$$s_k^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_{ik} - \overline{x}_k\right) + \left(\overline{x}_k - \mu\right)\right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(x_{ik} - \overline{x}_k\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(x_{ik} - \overline{x}_k\right)\left(\overline{x}_k - \mu\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\overline{x}_k - \mu\right)^2 =$$

$= \sigma_{uk}^2 + 0 + (\bar{x}_k - \mu)^2$

Posljednji član predstavlja odstupanje prosjeka *k*-tog uzorka od prosjeka populacije. Mi ne znamo koliko je to odstupanje, ali znademo da je ono u prosjeku $\frac{\sigma^2}{n}$.

Ovaj rezultat je samo naizgled nelogičan. Naime, ako imamo uzorak od 1 elementa, ne možemo ništa reći o varijanci. Tek za *n*=2 možemo nešto reći.

Odabir najboljeg procjenjitelja $\hat{\theta}$ za parametar populacije θ koji želimo odrediti iz uzorka nije uvijek jednostavan. U dodatku 1 prikazane su dvije metode procjenjivanja.

V.1. INTERVALI POUZDANOSTI i STANDARDNA POGREŠKA

Interval pouzdanosti prosjeka populacije

Procjena parametra daje nam jedan jedini broj koji ne sadrži informaciju o tome koliko je ta procjena vjerodostojna. Ako smo uzeli uzorak od *n* elemenata i izračunali njegov prosjek $\bar{x}$, on nikad nije jednak prosjeku populacije μ, ali zanima nas koliko blizu bi mogao biti. Zanima nas u kojem intervalu je on pouzdano.

Pretpostavimo da varijanca osnovnog skupa iznosi σ.

Prema središnjem graničnom teoremu (CLT), $\bar{X}$ je normalno raspodijeljen:

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

Za normalnu raspodjelu vrijedi:

$$P\left(-1 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1\right) = 68\%$$

$$P\left(-1{,}96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1{,}96\right) = 95\%$$

$$P\left(-2{,}575 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 2{,}575\right) = 99\%$$

To možemo pisati:

Prava vrijednost prosjeka populacije nalazi se u intervalu

$$\boxed{\bar{x} \pm \frac{\sigma}{\sqrt{n}}}$$

s vjerojatnošću 68%.

U statistici su rijetki problemi u kojima poznajemo varijancu, a ne poznajemo očekivanje. Najčešće su nam oba parametra nepoznata:

Ako ne znamo varijancu osnovnog skupa, moramo ju procijeniti:

Nepristrani procjenjitelj varijance je

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Za procjenjitelja standardne devijacije najčešće uzimamo slučajnu varijablu

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

koja je vrlo bliska nepristranom procjenjitelju standardne devijacije.

Interval 68% pouzdanosti je tada $\bar{x} \pm \frac{s}{\sqrt{n}}$.

U fizici se dogovorno svi rezultati mjerenja pišu s navođenjem iznosa 68-postotne nepouzdanosti:

$$\bar{x} \pm M$$

gdje je $M = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}}$ **nepouzdanost prosjeka** izračunata iz n nezavisnih mjerenja. Veličina M još se naziva i **standardna pogreška**.

Definiramo i **relativnu pogrešku** $R = \frac{M}{\bar{x}}$

Često se navodi i veličina $m = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = s$ koja označava **preciznost** svakog pojedinog mjerenja i govori o raspršenju cijele populacije mjerenja (tj. o preciznosti mjernog instrumenta).

Uočimo da preciznost mjerenja m govori o rasipanju mjerenja od srednje vrijednosti. Preciznost određujemo iz prvih nekoliko mjerenja, a svako sljedeće mjerenje ne smanjuje tu vrijednost.

Međutim, nepouzdanost $M = \frac{m}{\sqrt{n}}$ postaje to manja što veći broj mjerenja obavimo. Stoga su uvijek pouzdaniji rezultati dobiveni iz većeg broja mjerenja. U pravilu je potrebno obaviti barem 10 mjerenja. Broj obavljenih mjerenja planira se prije eksperimenta na osnovu željene pouzdanosti i raspoloživog vremena.

U statistici primjenjenoj na medicinu, ekonomiju, sociologiju,... obično se uzimaju intervali pouzdanosti 95% (pouzdano) ili 99% (vrlo pouzdano).

V.2. MJERENJA RAZLIČITIH STATISTIČKIH TEŽINA

Kada neku fizikalnu veličinu mjerimo u više navrata i dobivamo različite rezultate, zanima nas koja je prava, ili barem najvjerojatnija, vrijednost te fizikalne veličine. Mjerenja se obično izvode na različite načine i s različitim pouzdanostima pa ih ne smijemo tretirati ravnopravno. Kažemo da mjerenja imaju različite statističke težine.

Jedna takva serija mjerenja daje rezultat u obliku

$$x = \bar{x} \pm M \ ,$$

gdje M predstavlja interval 68% pouzdanosti za nalaženje prave vrijednosti x_p (standardna pogreška mjerenja).
U jednoj seriji mjerenja veličinu M smo odredili na sljedeći način:

$$M = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}} \ .$$

Nepristrani procjenjitelj prave vrijednosti x_p na osnovi te serije mjerenja je normalno raspodijeljen prema:

$$X_p \sim N(\bar{x}, M^2) \ .$$

<u>Opća srednja vrijednost</u>

Ako istu veličinu x mjerimo u raznim prigodama dobivamo različite rezultate s različitim standardnim pogreškama. One u svakoj seriji mjerenja ovise o preciznosti mjerenja u toj seriji i o broju mjerenja. Neka su rezultati takvih k mjerenja:

$$\begin{aligned} x_1 &= \bar{x}_1 \pm M_1 \\ x_2 &= \bar{x}_2 \pm M_2 \\ &\vdots \\ x_k &= \bar{x}_k \pm M_k \end{aligned}$$

Na osnovi tih k serija mjerenja možemo ponovno tražiti najvjerojatniju vrijednost x_p tražene fizikalne veličine. Funkcija gustoće vjerojatnosti za varijablu x_p iz k serija mjerenja glasi

$$f\left(x_p; \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k\right) = \frac{1}{M_1 M_2 \cdots M_k \left(\sqrt{2\pi}\right)^k} e^{-\sum_{i=1}^{k} \frac{\left(\bar{x}_i - x_p\right)^2}{2M_i^2}}$$

Logaritmiramo i tražimo maksimum funkcije.
Najvjerojatniji je onaj x_p za koji je

$$\sum_{i=1}^{k} \frac{\left(\bar{x}_i - x_p\right)^2}{M_i^2} = \min.$$

Nađimo taj minimum!

$$\frac{d}{dx_{\mathrm{p}}}\left(\sum_{i=1}^{m}\frac{\left(\bar{x}_i - x_p\right)^2}{M_i^2}\right) = -2\sum_{i=1}^{m}\frac{\left(x_i - x_{\mathrm{p}}\right)}{M_i^2}$$

Za najvjerojatniji x_{p}^* vrijedi

$$\sum_{i=1}^{m}\frac{\left(x_i - x_p^*\right)}{M_i^2} = \sum_{i=1}^{m}\frac{\bar{x}_i}{M_i^2} - x_{\mathrm{p}}^*\sum_{i=1}^{m}\frac{1}{M_i^2} = 0$$

$$\boxed{x_{\mathrm{p}}^* = \frac{\sum_{i=1}^{m}\frac{\bar{x}_i}{M_i^2}}{\sum_{i=1}^{m}\frac{1}{M_i^2}}}$$

Kažemo da je x_p^* opća srednja vrijednost svih mjerenja.

Interpretacija ovog rezultata:
Zamislimo da su sva mjerenja bila izvršena s jednakom preciznošću. Tada pouzdanost svake serije mjerenja ovisi samo o broju mjerenja u toj seriji. Neka je broj mjerenja u i-toj seriji (broj elemenata u i-tom uzorku) n_i. Tada se n_i-ovi međusobno odnose kao:

$$n_1 : n_2 : \cdots : n_k = \frac{1}{M_1^2} : \frac{1}{M_2^2} : \cdots : \frac{1}{M_k^2}$$

pa je najvjerojatnija prava vrijednost tražene fizikalne veličine u stvari srednja vrijednost svih mjerenja u svim serijama.

Definirajmo **statističku težinu** w_i pojedine serije:

$$w_i = \frac{\frac{1}{M_i^2}}{\sum_{j=1}^{k}\frac{1}{M_j^2}} \ .$$

Vidimo da su tako definirane statističke težine normirane, tj. vrijedi

$$\sum_{i=1}^{m} w_i = 1 \ .$$

Statističke težine odnose se obrnuto proporcionalno s kvadratima nepouzdanosti:

$$w_1 : w_2 : \cdots : w_k = \frac{1}{M_1^2} : \frac{1}{M_2^2} : \cdots : \frac{1}{M_k^2}$$

Uz ovu notaciju, najvjerojatnija prava vrijednost mjerene veličine je

$$\boxed{x_p^* = \sum_{i=1}^{k} w_i \bar{x}_i} \ .$$

Nepouzdanost opće srednje vrijednosti

Osim što želimo znati najvjerojatniju vrijednost tražene fizikalne veličine, zanima nas i kolika je pouzdanost tog rezultata.

Procjenjitelj prave vrijednosti mjerene veličine jest linearna kombinacija prosjeka iz *k* mjerenja:

$$X_p^* = \sum_{i=1}^{k} w_i \overline{X}_i$$

Varijancu linearne kombinacije znademo izračunati:

$$M^2 = V\left(X_p^*\right) = \sum_{i=1}^{k} w_i^2 V\left(\overline{X}_i\right) = \sum_{i=1}^{k} w_i^2 M_i^2$$

Uvrstimo li definiciju statističkih težina, dobivamo sljedeći rezultat za nepouzdanost opće srednje vrijednosti:

$$\boxed{M^2 = \frac{1}{\sum_{i=1}^{k} \frac{1}{M_i^2}}}$$

V.3. PROPAGACIJA POGREŠKE MJERENJA

Kada želimo posredno odrediti neku fizikalnu veličinu, a poznajemo njezinu funkcijsku zavisnost o neposredno mjerenim veličinama, važno nam je znati kako se s pomoću pogrešaka mjerenja osnovnih veličina određuje pogreška tražene veličine.
Npr.:

1. Mjerimo visinu stola h_1 i visinu stolca h_2, a zanima nas koliko će visok biti stolac ako ga stavimo na stol $h=h_1+h_2$.
2. Mjerimo vrijeme t padanja kuglice s vrha nebodera, a želimo odrediti visinu nebodera $h=gt^2/2$.
3. Mjerimo masu m i volumen V nekog predmeta, a želimo odrediti njegovu gustoću $\rho=m/V$.

Pretpostavljamo da su pogreške mjerenja slučajne (normalno raspodijeljene) i malene u odnosu na vrijednosti fizikalnih veličina, tj. $X_i \sim N(\mu_i, M_i^2)$ i $M_i << \mu_i$.

Razmatrat ćemo tri slučaja:

I. Posredna veličina je linearna kombinacija izravno mjerenih veličina

$$h(X,Y) = aX + bY$$

To je poseban slučaj linearne kombinacije pa odmah možemo pisati

$$E[h(X,Y)] = a\mu_X + b\mu_Y \qquad \text{ili} \qquad \bar{h}(X,Y) = a\bar{x} + b\bar{y}$$

$$\boxed{M_h^2 = a^2 M_X^2 + b^2 M_Y^2}$$

II. Posredna veličina je nelinearna funkcija jedne mjerene veličine $h(X)$.

Budući da smo pretpostavili da je pogreška malena, možemo pisati

$$E[h(X)] \cong h(\mu_X) = h(\bar{x})$$

Pogledajmo sad kako propagira pogreška!

$$M_h^2 = V[h(X)] = E\left[(h(X) - h(\mu_X))^2\right]$$

Razvijmo $h(X)$ u red:

$$h(X) = h(\mu_X) + (X-\mu_X)\left.\frac{dh}{dX}\right|_{\mu_X} + \frac{1}{2}(X-\mu_X)^2\left.\frac{d^2h}{dX^2}\right|_{\mu_X} + \cdots$$

pa približno vrijedi $h(X) - h(\mu_X) \approx (X-\mu_X)\left.\frac{dh}{dX}\right|_{\mu_X}$

$$E\left[(h(X) - h(\mu_X))^2\right] = E\left[\left((X-\mu_X)\left.\frac{dh}{dX}\right|_{\mu_X}\right)^2\right] = \left(\left.\frac{dh}{dX}\right|_{\mu_X}\right)^2 \cdot M_X^2$$

Za male pogreške dakle vrijedi: $\boxed{M_h = \left|\left.\frac{dh}{dx}\right|_{\bar{x}}\right| \cdot M_X}$

primjer: Pogreška potencije

$$h(X) = X^{\alpha}$$

$$M_h = \alpha\, \bar{x}^{\alpha-1} \cdot M_X$$

za relativne pogreške dakle vrijedi:

$$\frac{M_h}{\mu_h} = \alpha \cdot \frac{M_X}{\bar{x}} \qquad \text{ili} \qquad R_h = \alpha \cdot R_X$$

III. Najopćenitiji slučaj: Posredna veličina je nelinearna funkcija dvije ili više mjerenih veličina

$$h(X, Y).$$

Pokazuje se da malena pogreška propagira na sljedeći način:

$$M_h^2 = \left(\frac{\partial h}{\partial X}\right)^2_{\bar{x},\bar{y}} \cdot M_X^2 + \left(\frac{\partial h}{\partial Y}\right)^2_{\bar{x},\bar{y}} \cdot M_Y^2.$$

Općenito za *n* izravno mjerenih veličina imamo:

$$h(X_1, X_2, ..., X_n)$$

$$\Rightarrow \qquad \bar{h}(X_1, X_2, ..., X_n) = h(\bar{x}_1, \bar{x}_2, ..., \bar{x}_n)$$

$$M_h = \sqrt{\sum_i \left(\frac{\partial h}{\partial X_i}\right)^2_{\bar{x}_1,\bar{x}_2,...,\bar{x}_n} M_{X_i}^2}$$

Primjer iz praktikuma (vježba 8):
Modul elastičnosti čelika određuje se mjerenjem dimenzija čelične šipke, savijenosti čelične šipke po jedinici sile i udaljenosti potpornja. Teorijska formula za savijenost šipke je:

$$\lambda = \frac{1}{4E}\frac{L^3}{ab^3}F,$$

gdje je E modul elastičnosti čelika, L je udaljenost potporanja, a je širina šipke, b je debljina šipke, a F je sila teže utega.

Izmjerene su sljedeće veličine:

$$A = \frac{\lambda}{F} = (0{,}76 \pm 0{,}01)\frac{\text{mm}}{\text{N}} = (7{,}6 \pm 0{,}1)\cdot 10^{-4}\frac{\text{m}}{\text{N}} = \bar{A} \pm M_A$$

$$a = (10{,}26 \pm 0{,}05)\,\text{mm} = (1{,}026 \pm 0{,}005)\cdot 10^{-2}\,\text{m} = \bar{a} \pm M_a$$

$$b = (1{,}53 \pm 0{,}03)\,\text{mm} = (1{,}53 \pm 0{,}03)\cdot 10^{-3}\,\text{m} = \bar{b} \pm M_b$$

$$L = (29{,}0 \pm 0{,}1)\,\text{cm} = (2{,}90 \pm 0{,}01)\cdot 10^{-1}\,\text{m} = \bar{L} \pm M_L$$

Najvjerojatnija vrijednost za modul elastičnosti čelika je:

$$\bar{E} = \frac{1}{4}\frac{\bar{L}^3}{\bar{a}\bar{b}^3}\frac{1}{\bar{A}} = 2{,}183226\ \text{N/m}^2$$

Varijanca modula elastičnosti je:

$$M_E^2 = \left(\frac{\partial E}{\partial A}\right)^2 M_A^2 + \left(\frac{\partial E}{\partial a}\right)^2 M_a^2 + \left(\frac{\partial E}{\partial b}\right)^2 M_b^2 + \left(\frac{\partial E}{\partial L}\right)^2 M_L^2 =$$

$$= \left(\frac{\overline{L}^3}{4\overline{a}\overline{b}^3\overline{A}^2} \right)^2 M_A^2 + \left(\frac{\overline{L}^3}{4\overline{a}^2\overline{b}^3\overline{A}} \right)^2 M_a^2 + \left(\frac{3\overline{L}^3}{4\overline{a}\overline{b}^4\overline{A}} \right)^2 M_b^2 + \left(\frac{3\overline{L}^2}{4\overline{a}\overline{b}^3\overline{A}} \right)^2 M_L^2 =$$

$$= \overline{E}^2 \left[\left(\frac{M_A}{\overline{A}} \right)^2 + \left(\frac{M_a}{\overline{a}} \right)^2 + \left(3\frac{M_b}{\overline{b}} \right)^2 + \left(3\frac{M_L}{\overline{L}} \right)^2 \right] =$$

$$= \overline{E}^2 \left[1{,}73 \cdot 10^{-4} + 2{,}37 \cdot 10^{-5} + 3{,}46 \cdot 10^{-3} + 1{,}07 \cdot 10^{-4} \right]$$

Standardna pogreška je:

$$M_E = \overline{E}\sqrt{3{,}76 \cdot 10^{-3}} = 0{,}13 \cdot 10^{11}\ \mathrm{N/m^2}$$

Rezultat za modul elastičnosti pišemo:

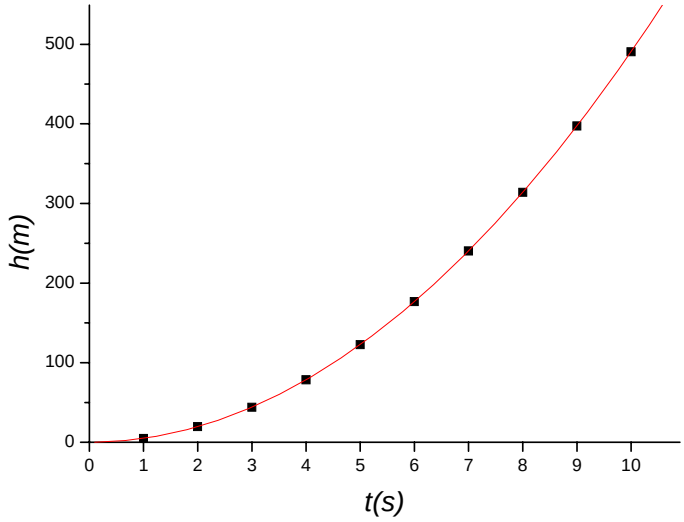
$$E = (2{,}2 \pm 0{,}1) \cdot 10^{11}\ \mathrm{N/m^2}$$

VI. LINEARNA REGRESIJA I KORELACIJA

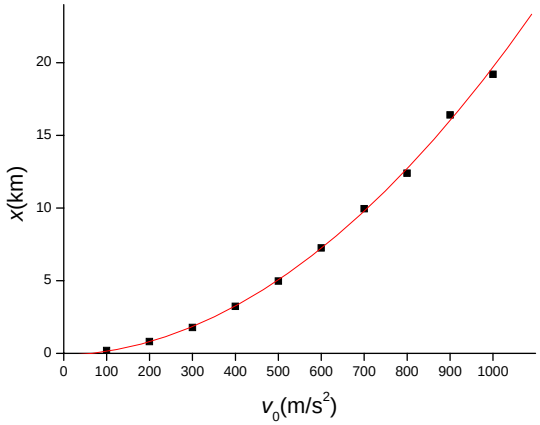
Često proučavamo rezultate mjerenja ili statističkog istraživanja uspoređujući dvije skupine podataka *X* i *Y*. Npr.:

opažanje	**X**	**Y**
elastična sila	masa utega	duljina opruge
slobodni pad	vrijeme	prijeđeni put
topovsko tane	masa eksploziva	Domet
studenti	uspjeh na prijemnom	uspjeh na studiju
Ljudi	visina	masa
Automobili	snaga	potrošnja goriva

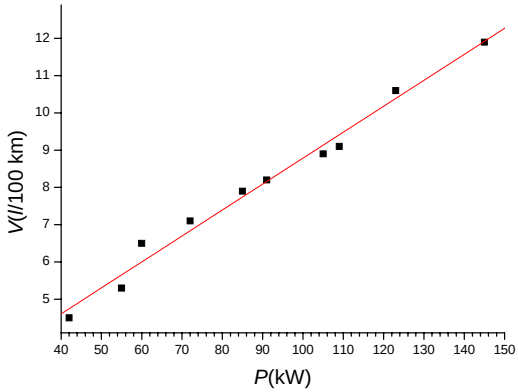
Neki parovi varijabli su deterministički povezani. Npr.: Ako u slobodnom padu znamo vrijeme padanja, točno znamo i prijeđeni put (dubinu poniranja).



Međutim, mnoštvo je parova varijabli *X* i *Y* za koje bismo rekli da su slabije ili jače povezane, ali ne postoji deterministička veza među njima. Npr.: Kada ispaljujemo topovsko tane pod nekim kutom, domet tog taneta ovisit će o početnoj brzini, ali i o nekim drugim utjecajima koje ne možemo kontrolirati (vjetar, vibracije prilikom ispaljivanja, masa taneta, ...).



Ili, drugi primjer: Potrošnja goriva nekog automobila ovisi o snazi motora, ali nikako ne možemo govoriti o determinističkom odnosu:



Regresijska analiza dio je statistike koji se bavi istraživanjem odnosa između dvije ili više varijabli povezanih na nedeterministički način.

Linearan odnos

Najjednostavniji odnos dviju deterministički povezanih varijabli jest linearan:

$$y = ax + b$$

Npr. ako znademo konstantu neke opruge K, onda će njezina duljina kad se o nju objesi uteg mase m biti dana relacijom:

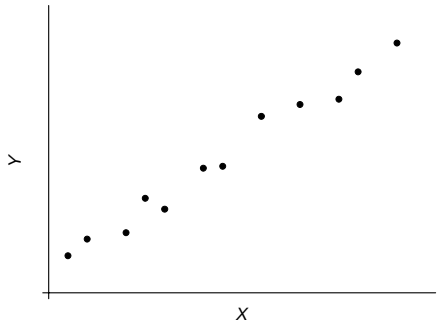
$$l = \frac{g}{K} m + l_0 \ ,$$

gdje je l_0 duljina neopterećene opruge.

Ako dvije varijable nisu deterministički povezane, onda će za fiksiranu vrijednost x_i varijable X, varijabla Y_i biti slučajna. Općenito, varijabla čiju vrijednost određuje izvođač pokusa označava se s X te naziva **nezavisnom varijablom**. Za određeni x_i, slučajna varijabla Y_i naziva se **zavisnom** i poprima neku vrijednost y_i. U praksi se za nezavisnu varijablu uzima uvijek ona veličina koja je točnije mjerena.

Primjer: Utege vješamo na niz opruga koje su izašle iz istog stroja. Očekujemo da su konstante opruga vrlo slične, ali da postoji nekakvo rasipanje. Mase utega su određene prije pokusa i predstavljaju vrijednosti nezavisne varijable. Izmjerene duljine opterećenih opruga predstavljaju vrijednosti zavisne varijable.

Neka su $x_1, x_2, \ldots, x_n$ vrijednosti nezavisne varijable za koje se čine opažanja, a s Y_i i y_i označimo slučajnu varijablu i opaženu vrijednost pridruženu x_i-u. Skup opaženih podataka sastoji se od n parova $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Takve podatke najprije ucrtamo u graf.



Promatranjem grafa procjenjujemo može li se dobivena ovisnost aproksimirati linearnom. Kada bi varijable X i Y bile deterministički linearno povezane, za svaki par (x_i, y_i) vrijedila bi relacija

$$y_i = a\, x_i + b \ ,$$

a parametri a i b bili bi isti za sve parove. Međutim, budući da su Y_i slučajne varijable, ne postoje jedinstveni koeficijenti a i b koji bi zadovoljavali gornju relaciju za svaki par. Stoga nastojimo pronaći takve a i b koji su najvjerojatniji parametri pretpostavljene linearne ovisnosti.

VI.1. Određivanje koeficijenata metodom najmanjih kvadrata

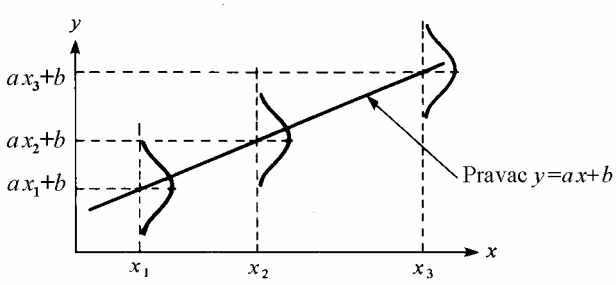
Pretpostavka:

Postoje parametri a i b takvi da za svaku vrijednost x_i nezavisne varijable X, zavisnu varijablu Y_i možemo pisati:

$$Y_i = ax_i + b + \hat{\varepsilon} \ ,$$

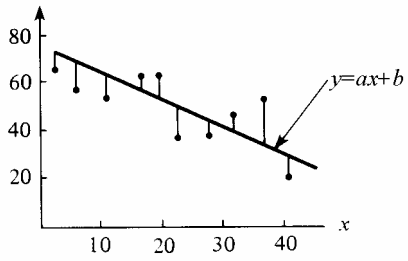
gdje je $\hat{\varepsilon}\,\varepsilon$ normalna slučajna varijabla s očekivanjem $E(\hat{\varepsilon}) = 0$ i varijancom $V(\hat{\varepsilon}) = \sigma^2$. σ^2 je jednaka za sve vrijednosti x.

Što to znači? Pretpostavili smo da postoji jedan pravi (ispravan) pravac regresije koji za poznati x_i određuje očekivanu vrijednost $E(Y_i)$, a da su odstupanja vrijednosti y_i od očekivanja $E(Y_i)$ slučajne pogreške raspodijeljene prema Gaussu i jednako vjerojatne u cijelom području.



Za izmjerene (opažene) parove vrijedi:

$$y_i = ax_i + b + \varepsilon_i \ ,$$



Princip najmanjih kvadrata:

Od svih pravaca $y = ax + b$, najvjerojatniji pravac regresije jest onaj za koji je suma kvadrata odstupanja

$$f(a,b) = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

minimalna.

Suma kvadrata odstupanja je minimalna kada istodobno vrijedi:

$$\frac{\partial f(a,b)}{\partial a} = 0 \qquad \text{i} \qquad \frac{\partial f(a,b)}{\partial b} = 0$$

Uz te uvjete dobivamo **normalne jednadžbe**:

$$\sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb = 0$$

Rješavanjem ovog sustava jednadžbi dobivamo izraze za najvjerojatnije vrijednosti koeficijenata *a* i *b*:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad \text{i} \qquad b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Odsječak *b* na *y*-osi možemo jednostavnije pisati:

$$b = \bar{y} - a\bar{x} \ .$$

To nam olakšava račun kad već imamo izračunati *a*.

Napomene:

- Prije računanja pravca regresije treba u grafu provjeriti ima li smisla linearna regresija i jesu li podaci podjednako raspršeni.
- Rezultate sumiranja ne smije se zaokruživati jer pogreška zaokruživanja bitno utječe na razliku velikih sličnih brojeva

Za nastavak računa bit će nam korisno napisati parametar *a* u obliku:

$$a = \frac{n}{N_z}\sum_{i=1}^{n}(x_i - \bar{x})y_i \ ,$$

gdje je $N_z = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$ nazivnik gornjeg izraza.

Možemo reći da je najvjerojatnija vrijednost parametra *a* u stvari očekivanje slučajne varijable $\hat{a}$ definirane kao:

$$\hat{a} = \frac{n}{N_z}\sum_{i=1}^{n}(x_i - \bar{x})Y_i \ .$$

Pouzdanost parametra *a*

$\hat{a}$ je linearna kombinacija normalnih slučajnih varijabli Y_i pa njezino očekivanje i varijancu možemo izračunati prema pravilima za linearnu kombinaciju.
Za varijancu linearne kombinacije vrijedi (vidi starija predavanja):

$$V(c_1Y_1 + c_2Y_2 + ... + c_nY_n) = c_1^2V(Y_1) + c_2^2V(Y_2) + ... + c_n^2V(Y_n)$$

pa lako možemo odrediti varijancu našeg parametra *a*:

$$\sigma_a^2 = \frac{n^2}{N_z^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 V(Y_i)\,.$$

Budući da smo pretpostavili da je $V(Y_i)=\sigma^2$ jednaka za svaki *i*, možemo pisati:

$$\sigma_a^2 = n^2\sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]^2} = n^2\sigma^2 \frac{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}{\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]^2} = n^2\sigma^2 \frac{N_z/n}{N_z^2} = n\frac{\sigma^2}{N_z}\,.$$

Preostaje nam još da procijenimo σ (=*m*, preciznost jednog mjerenja).

Def:

Predviđena vrijednost slučajne varijable Y_i je

$$y_i^* = ax_i + b\ .$$

Može se pokazati da je nepristrani procjenjitelj za σ^2 dan izrazom:

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n-2} = \frac{\sum_{i=1}^{n}[y_i - (ax_i + b)]^2}{n-2}\ .$$

Nećemo to izvoditi, ali razmotrimo značenje:
Pravac je određen s pomoću dvije točke. Dakle, ako su nam poznata samo dva para (x_i,y_i), nećemo imati nikakvo odstupanje od pravca pa niti procjenu varijance.
Raspisivanjem gornjeg izraza s^2 se može napisati u obliku:

$$s^2 = \frac{\sum_{i=1}^{n}y_i^2 - a\sum_{i=1}^{n}x_iy_i - b\sum_{i=1}^{n}y_i}{n-2}$$

pa uvrštavajući izraz za *b* dobivamo:

$$s^2 = \frac{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2 - a\left(n\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right)}{n(n-2)} = \frac{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2 - a^2N_z}{n(n-2)}$$

I konačno za nepouzdanost parametra *a* dobivamo:

$$M_a^2 = \sigma_a^2 = n\frac{\sigma}{N_z} = \frac{1}{(n-2)}\left[\frac{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2}{N_z} - a^2\right]$$

Parametar b je linearna kombinacija slučajnih varijabli Y_i i $\hat{a}$ pa njezinu nepouzdanost lako odredimo (nećemo pokazivati):

$$M_b^2 = M_a^2 \frac{\sum_{i=1}^{n} x_i^2}{n} .$$

Sažetak rezultata za linearnu regresiju:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$M_a = \sqrt{\frac{1}{(n-2)}\left[\frac{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} - a^2\right]}$$

$$M_b = M_a \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} .$$

VI.2. Nelinearne regresije

U fizici često mjerimo parove varijabli među kojima postoji neka funkcionalna zavisnost, ali ona nije linearna. Najčešće iz teorijskih razmatranja pretpostavljamo oblik funkcionalne ovisnosti, ali nam nisu poznate sve konstante (parametri). Npr.: amplituda titranja y gušenog tjeranog harmoničkog oscilatora ovisi o frekvenciji pobude ω prema relaciji:

$$y = \frac{A}{\sqrt{\left(\omega_0^2 - \omega^2\right)^2 + \left(\omega/\tau\right)^2}} \ .$$

Parametre A, ω_0 i τ želimo odrediti iz serije mjerenja. Princip određivanja tih parametara metodom najmanjih kvadrata jednak je onom kad smo imali linearnu regresiju. Postavljamo uvjet da suma kvadrata odstupanja mjerenih amplituda od očekivanih

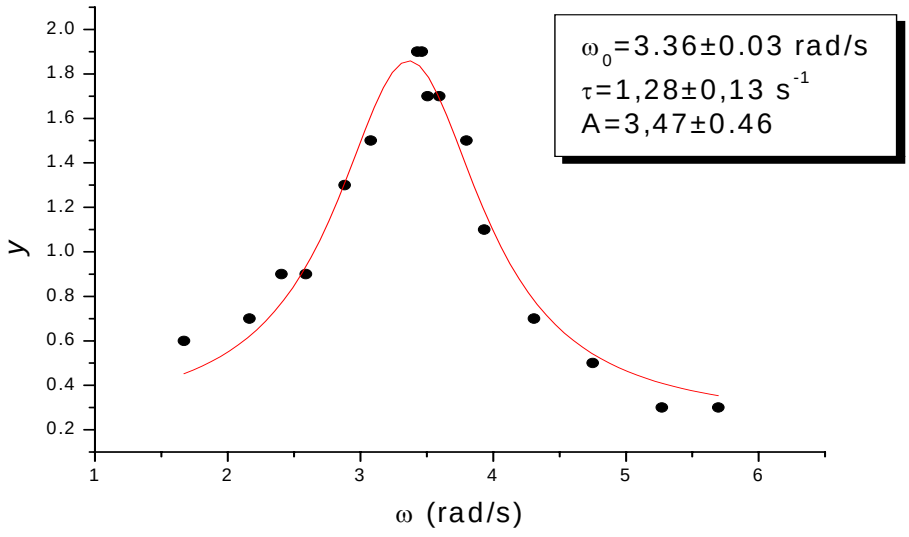
$$f(A,\omega_0,\tau) = \sum_i \left(y_i - \frac{A}{\sqrt{\left(\omega_0^2 - \omega_i^2\right)^2 + \left(\omega/\tau\right)^2}} \right)^2$$

bude minimalna.

Imamo tri nepoznata parametra i postavljamo sustav od tri jednadžbe s tri nepoznanice iz uvjeta:

$$\frac{\partial f(A,\omega_0,\tau)}{\partial A} = 0 \quad ; \quad \frac{\partial f(A,\omega_0,\tau)}{\partial \omega_0} = 0 \quad ; \quad \frac{\partial f(A,\omega_0,\tau)}{\partial \tau} = 0$$

Rješavanje ovakvog sustava vrlo je kompliciranо i najčešće zahtijeva numeričko rješavanje. U današnje doba računala i programa specijaliziranih za obradu podataka, nema potrebe da rješavamo takve specifične probleme. Primjer s praktikuma:



Ipak, jedan lagan i rješiv problem riješit ćete na vježbama kad je funkcionalna ovisnost polinom drugog stupnja.

Regresija s transformiranim varijablama

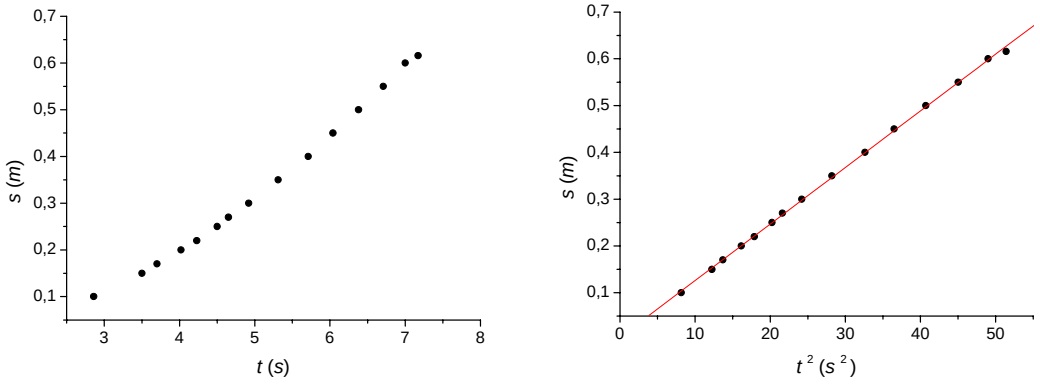
Vrlo često se neka nelinearna ovisnost želi prikazati u linearnom obliku. Dva su razloga za to:

1. Mogu se primjeniti jednadžbe za linearnu regresiju
2. Takav grafički prikaz zorno potvrđuje (ili odbacuje) ispravnost primijenjene teorije.

Primjer iz praktikuma:
Dubina poniranja Maxwellova diska (vježba 6) ovisi o vremenu prema relaciji:

$$s = \frac{1}{2}\frac{mg}{m + \frac{I_z}{r^2}}t^2$$

Prikažemo li rezultate jedne serije mjerenja u *s-t* dijagramu, točke će se grupirati oko parabole i morali bismo primijeniti nelinearnu regresiju. Međutim, uvedemo li novu varijablu $u=t^2$, rezultati će biti grupirani oko pravca i moći ćemo primijeniti linearnu regresiju. Takvom transformacijom varijabli mnogi kompliciranim problemi se pojednostavljuju.



Logaritamski grafovi

Često imamo slučaj da je zavisna veličina proporcionalna potenciji nezavisne veličine, a mi ili ne znamo iznos eksponenta ili želimo provjeriti predviđeni iznos. Tada varijable logaritmiramo pa veza među novim, transformiranim varijablama postaje linearna.

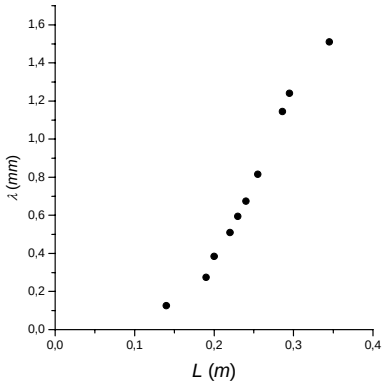
Primjer iz praktikuma (vježba 8):
Modul elastičnosti čelika određuje se mjerenjem dimenzija čelične šipke, savijenosti čelične šipke po jedinici sile i udaljenosti potpornja. Teorijska formula za savijenost šipke je:

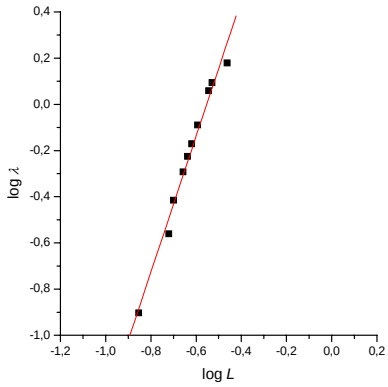
$$\lambda = \frac{1}{4E}\frac{L^3}{ab^3}F,$$

gdje je E modul elastičnosti čelika, L je udaljenost potporanja, a je širina šipke, b je debljina šipke, a F je sila teže utega.
Želimo li provjeriti ovisi li savijenost šipke uistinu o trećoj potenciji udaljenosti među potpornjima, obavit ćemo mjerenja za razne udaljenosti potporanja držeći sve ostale veličine konstantnima. Logaritmirat ćemo gornju jednadžbu:

$$\log\lambda = \log\frac{1}{4E}\frac{F}{ab^3} + 3\log L$$

Uvest ćemo nove varijable $x = \log L$ i $y = \log\lambda$. Rezultate mjerenja izračunate u tim novim varijablama prikazat ćemo u grafu. Ako taj graf možemo opisati linearnom regresijom $y = ax + b$, onda je savijenost uistinu proporcionalna potenciji od L. Ako je parametar a približno jednak 3, onda je eksponent teorijski dobro predviđen.

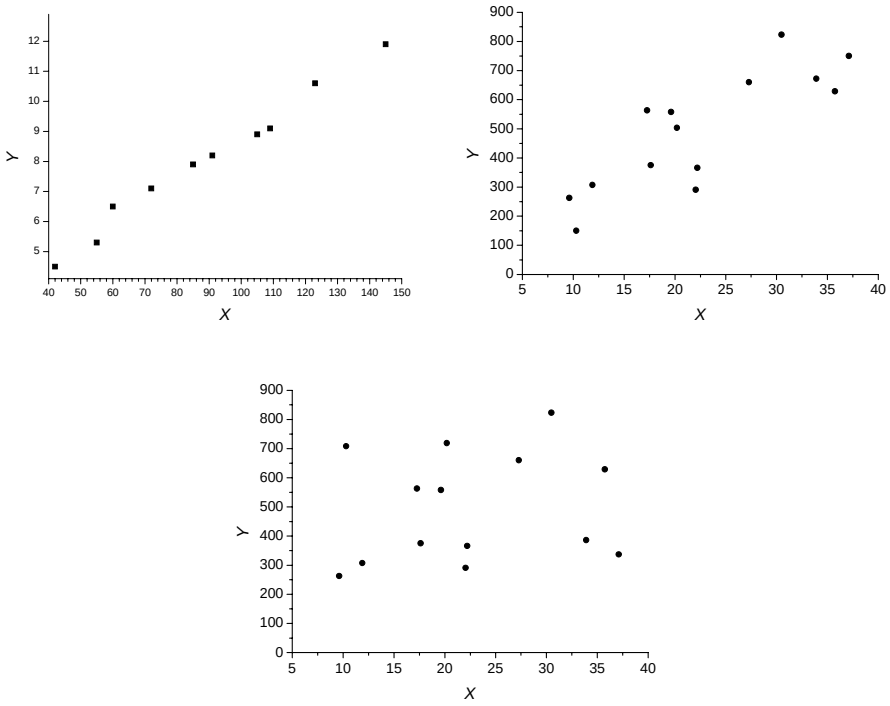




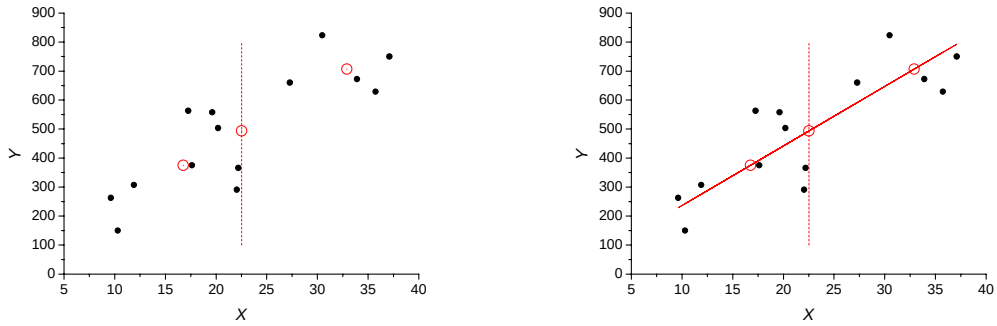
Linearnom regresijom dobili smo koeficijent smjera $a = 2{,}94 \pm 0{,}14$ što znači da je teorijsko predviđanje eksponenta bilo ispravno.

VI.3. Korelacije

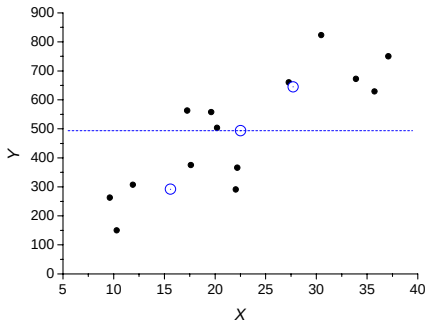
Kada veza između parova varijabli nije deterministički linearna, možemo imati razne stupnjeve rasipanja. Zanima nas koliko su varijable X i Y korelirane. Evo nekoliko primjera nedeterminističkog odnosa:

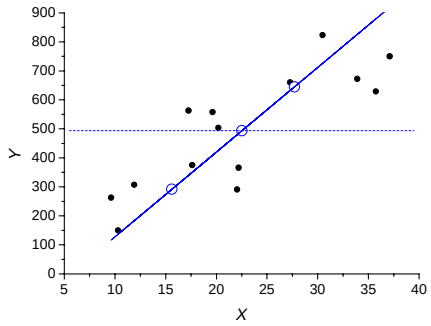


U prvom slučaju možemo prilično uvjerljivo govoriti o linearnom odnosu varijabli X i Y, u drugom slučaju možemo pretpostaviti da postoji pozitivna korelacija među tim varijablama, dok su u trećem slučaju varijable X i Y naizgled nezavisne. Želja nam je ove tvrdnje na neki način kvantizirati. Razmatrat ćemo podatke iz drugog grafa. Ne znamo je li jedna od varijabli točnije mjerena. Probajmo odoka odrediti pravce regresije najprije uz pretpostavku da je varijabla X točnije mjerena:

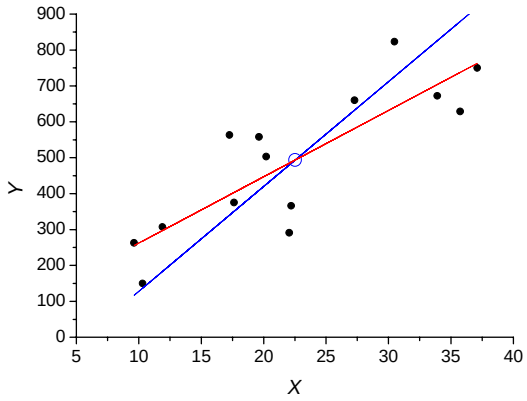


a zatim da je varijabla Y točnije mjerena. U tom slučaju tražimo ovisnost varijable X o Y pa nakon nalaženja središnje točke, graf dijelimo na gornju i donju polovicu, a ne na lijevu i desnu:

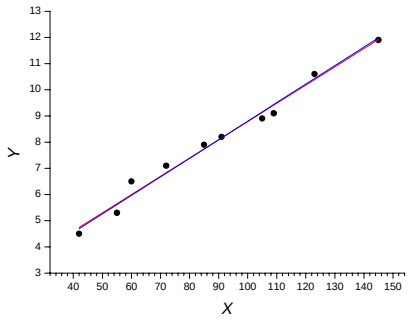


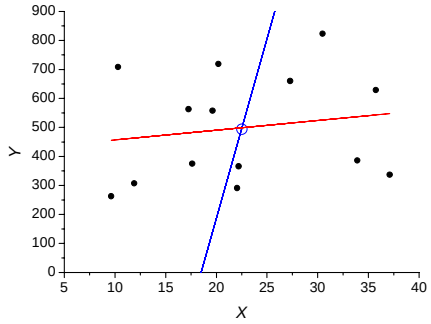


Dobiveni pravci se ne podudaraju, a odstupanje je to veće što su podaci slabije korelirani. Za ovaj skup podataka pravci regresije nacrtani su na istom grafu:



Za prvi i treći skup podataka pravci regresije izgledaju ovako:





U prvom slučaju pravci se gotovo poklapaju dok se u drugom slučaju kut među njima približava π/2.

Pravce regresije točnije dobivamo primjenom metode najmanjih kvadrata. Za ovisnost varijable Y o X postavljamo jednadžbu

$$y = ax + b\,,$$

a vrijednosti koeficijenata a i b su:

$$a = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} \qquad b = \frac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - \bar{x}^2}\,.$$

Za ovisnost varijable X o Y postavljamo jednadžbu

$$x = cy + d\,,$$

a vrijednosti koeficijenata c i d su:

$$c = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{y^2} - \bar{y}^2} \qquad d = \frac{\overline{y^2}\bar{x} - \bar{y}\overline{xy}}{\overline{y^2} - \bar{y}^2} \ .$$

Izračunavanje koeficijenata pravaca regresije pomoću kovarijance

Prisjetimo se momenata dvodimenzionalnih raspodjela!
Vrlo je važan moment M_{11} koji se naziva **kovarijanca**. Za n parova mjerenih veličina, kovarijanca je:

$$M_{11} = \sigma_{XY} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\bar{y} = m_{11} - m_{10}m_{01}$$

Koeficijent smjera a pravca regresije Y o X možemo pisati:

$$a = \frac{m_{11} - m_{10}m_{01}}{m_{20} - m_{10}^2} = \frac{M_{11}}{M_{20}} = \frac{\sigma_{XY}}{\sigma_X^2} \ .$$

Budući da pravac prolazi kroz točku $(\bar{x}, \bar{y})$, jednadžba pravca je

$$(y - \bar{y}) = \frac{\sigma_{XY}}{\sigma_X^2}(x - \bar{x}) \ .$$

Slično za ovisnost X o Y vrijedi:

$$c = \frac{m_{11} - m_{10}m_{01}}{m_{02} - m_{01}^2} = \frac{M_{11}}{M_{02}} = \frac{\sigma_{XY}}{\sigma_Y^2}$$

pa je jednadžba pravca

$$(x - \bar{x}) = \frac{\sigma_{XY}}{\sigma_Y^2}(y - \bar{y}) \ .$$

Koeficijent korelacije

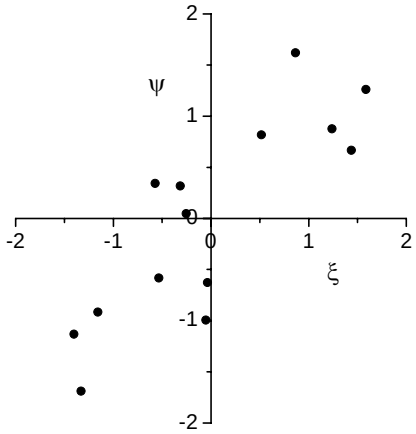
Jednadžbe pravaca možemo "standardizirati":

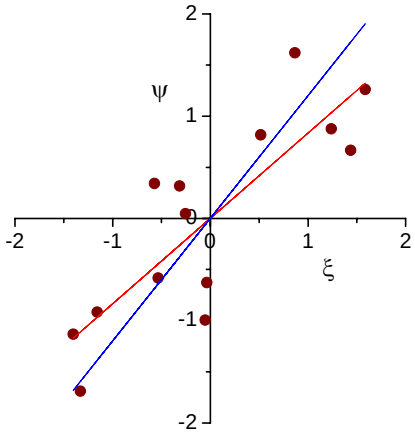
$$\frac{(y - \bar{y})}{\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}\frac{(x - \bar{x})}{\sigma_X} \quad \text{(ovisnost } Y \text{ o } X\text{)}$$

$$\frac{(x - \bar{x})}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}\frac{(y - \bar{y})}{\sigma_Y} \quad \text{(ovisnost } X \text{ o } Y\text{)}$$

Možemo nacrtati novi graf koji ima ishodište u $(\bar{x}, \bar{y})$, vrijednosti na apscisi izražene u jedinicama σ_X, a vrijednosti na ordinati izražene u

jedinicama σ_Y. Nove, bezdimenzionalne varijable su $\xi = \frac{(x-\bar{x})}{\sigma_X}$ i $\psi = \frac{(y-\bar{y})}{\sigma_Y}$.





Jednadžbe pravaca regresije su:

$$\psi = \rho\xi$$

za ovisnost ψ o ξ, i

$$\xi = \rho\psi$$

za ovisnost ξ o ψ.

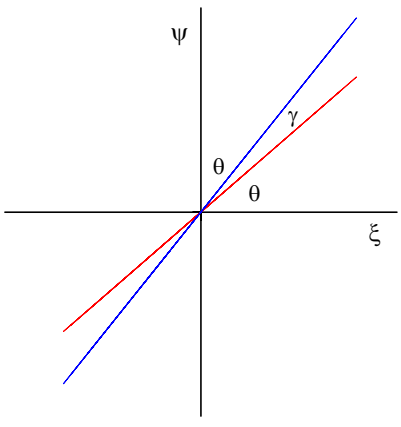
Upravo smo definirali **koeficijent korelacije** (vidi 9. predavanje)

$$\boxed{\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}}.$$

To je bezdimenzionalna veličina koja ne ovisi o jedinicama u kojima su varijable X i Y izražene.

Značenje koeficijenta korelacije

U bezdimenzionalnom grafu ρ određuje tangens kuta između pravca ovisnosti ψ o ξ i apscise, a isto tako određuje tangens kuta između pravca ovisnosti ξ o ψ i ordinate.



Kut između pravaca regresije, γ povezan je s ρ na sljedeći način:

$$\gamma = \frac{\pi}{2} - 2\theta$$

$$\operatorname{tg}\gamma = \operatorname{ctg} 2\theta = \frac{1 - \operatorname{tg}^2 \theta}{2 \operatorname{tg}\theta} = \frac{1 - \rho^2}{2\rho}$$

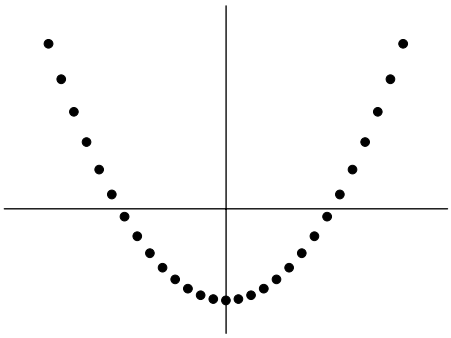
Koeficijent korelacije može poprimiti vrijednosti $-1 \le \rho \le 1$.

Ako su varijable nezavisne, pravci su paralelni s osima koordinatnog sustava, tj. $\gamma = \frac{\pi}{2}$, $\rho = 0$.

Ako postoji linearna deterministička veza između *X* i *Y*, pravci se poklapaju, a ρ = 1 ili ρ = -1 . Dogovorno, kažemo da je korelacija dobra kad je $|\rho|$>0,5 .

Napomena:

Kad su *X* i *Y* nezavisne varijable, koeficijent korelacije težit će nuli ($\rho \approx 0$). Međutim, činjenica da je ρ=0 ne znači nužno da su varijable nezavisne. Primjer:



VII. PROVJERAVANJE (testiranje) HIPOTEZA

Provjera hipoteze vrlo je bitan dio statističkog zaključivanja.
Da bi se takva provjera formulirala, potrebno je postaviti neku teoriju koja se želi dokazati.

Npr.:

- Novi lijek bolji je za liječenje određenih simptoma od starog.
- Igraća kocka ima pomaknuto težište
- U Svijetu se rađa više muškaraca nego žena

U svakom takvom problemu postavljamo dvije tvrdnje (hipoteze) od kojih je točno jedna istinita:

H_0= nul-hipoteza
H_1= alternativna hipoteza

One se ne tretiraju ravnopravno:

U sudskom procesu: H_0= optuženi je nevin
H_1= optuženi je kriv

Hipoteza H_0 smatra se ispravnom dok se ne dokaže H_1.

Za H_0 obično se uzima stara, postojeća teorija:

Npr., za lijek: H_0= stari lijek jednako je dobar kao i novi
za kocku: H_0= kocka je poštena

Ako ne odbacimo H_0, to ne znači da je ona ispravna, nego samo da nemamo dovoljno dokaza da ju odbacimo.

Postavljanje problema, kritično područje

Da bismo donijeli odluku, moramo uzeti uzorak iz populacije koju istražujemo.
Prije uzimanja uzorka moramo skup mogućih ishoda podijeliti u dva područja:

A = područje prihvaćanja H_0
B = područje odbacivanja H_0 = **kritično područje**

Primjer bacanja novčića 6 puta: X = broj grbova u šest bacanja
H_0 = novčić je pošten, tj. $X \sim Bin(6,1/2)$
H_1 = češće pada grb

Mogući ishodi:

X	0	1	2	3	4	5	6
$P(X/H_0)$	0,0156	0,0938	0,2344	0,3125	0,2344	0,0938	0,0156

Odlučimo se: A={0,1,2,3,4}
B={5,6}

Vrste pogrešaka

		odluka	
		odbaci H_0	prihvati H_0
istina	H_0	pogreška I. vrste	ispravan zaključak
	H_1	ispravan zaključak	pogreška II. vrste

Smatra se da je pogreška I. vrste mnogo ozbiljnija od pogreške II. vrste.

Vjerojatnost pogreške I. vrste:

α = vjerojatnost odbacivanja H_0 kada je istinita = $P(\mathrm{B}/H_0)$

Vjerojatnost pogreške II. vrste:

β = vjerojatnost prihvaćanja H_0 kada nije istinita = $P(\mathrm{A}/H_1)$

U našem primjeru je $\alpha = 0{,}1094 \approx 11\%$.
Međutim, čak i da je novčić toliko nepošten da je $p = 2/3$, raspodjela varijable X bi bila

X	0	1	2	3	4	5	6
$P(X/H_1)$	0,0014	0,0165	0,0823	0,2195	0,3292	0,2634	0,0878

pa bi vjerojatnost pogreške II. vrste uz naš odabir kritičnog područja bila $\beta = 0{,}6488$.

Veliki β se događa kada je uzorak premalen ($n = 6$).

Signifikantnost testa (važnost, značajnost)

Def:

Za postupak provjere kažemo da ima **razinu signifikantnosti** (važnosti) α ako je
$P(\text{pogreška I. vrste}) \leq \alpha$.
Kažemo da je to **test razine α.**

Naš primjer je test razine 0,11.

Tradicionalno se kao razine signifikantnosti uzimaju vrijednosti 0,01; 0,05 ili 0,10.
Za $\alpha = 0{,}05$ kažemo da je test signifikantan, a za $\alpha = 0{,}01$ kažemo da je test vrlo signifikantan.

Moć testa

Moć statističkog testa mjeri sposobnost testa da odbaci nul-hipotezu kad je uistinu pogrešna, tj. da učini ispravnu odluku.
P = 1-β

U našem primjeru je P = 0,35. Idealno je P = 1. Za snažniji test morali bismo imati veći uzorak.

Opća pravila odabira testa

Prije uzimanja uzorka:
1) Izreci nul-hipotezu i alternativnu hipotezu.
2) Razmotri odgovarajuću raspodjelu danu nul-hipotezom.
3) Odluči o razini signifikantnosti testa.
4) Odredi kritično područje (odluči o kriteriju odbacivanja nul-hipoteze)

Sada uzmi uzorak!
5) Izračunaj vrijednost statistike testa
6) Učini odluku:
   Ako je vrijednost statistike testa u kritičnom području, odbaci H_0!
   Ako vrijednost statistike testa nije u kritičnom području, nemoj odbaciti H_0!