

Аналитичка и инструментална хемија
Предавања

Статистика

Студии по биохемија и физиологија
и молекуларна биологија

СТАТИСТИКА

Во секојдневниот живот:

Масовни појави:

Голем број единки

Голем број променливи

Хемиски мерења:

Анализирани проби (примероци)

Голем број на карактеристики (обично мерливи физички величини)

Статистиката се занимава со квантитативно изучување на појавите врз основа на голем број набљудувања.

Предмет на проучување се множества од голем број податоци.

Кога се развива и валидира една аналитичка метода, обично се прават **три одделни разгледувања** поврзани со експерименталната грешка:

1. **Пред** да се започне со анализата, се проценуваат грешките поврзани со секое мерење, за да се утврди дали нивниот кумулативен ефект може да ја ограничи применливоста на анализата. Потоа, препознаените грешки за кои се смета дека може да влијаат врз резултатите, се сведуваат на минимум.
2. **Во текот** на анализата, процесот на мерење се следи за да се обезбеди негова контрола.
3. **На крајот** од анализата, квалитетот на мерењата и резултатите се проценуваат и споредуваат според првично поставените критериуми.

Мерки за централна тенденција

Претпоставка: податоците се расејани околу некоја **централна вредност**.

Централната тенденција може да се опише со:

- средна вредност;
- медијана.

Средна вредност

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- сума од сите добиени вредности поделена со вкупниот број извршени мерења
- начин за процена на централната тенденција;
- не е независна величина (влијание од вредности кои отстапуваат.

Пример 1: Анализа на концентрација на железо во крв кај заболени пациент:

$$30,80 + 30,94 + 31,07 + 30,56 + 31,12 + 31,74 + 31,98 = 218,21 \mu\text{g/dL}$$

$$\bar{X} = \frac{218,21 \mu\text{g/dL}}{7} = 31,17 \mu\text{g/dL}$$

Проба бр.	$\gamma(\text{Fe}) / \mu\text{g/dL}$
1	30,80
2	30,94
3	31,07
4	30,56
5	31,12
6	31,74
7	31,98

Медијана, X_{med}

Вредност која се наоѓа на средината од низата, кога податоците се подредени по вредност, од најмал кон најголем.

- Кога во податоците има **непарен** број мерења, тогаш медијаната претставува вредност што се наоѓа на средината од низата;
- За **парен** број на мерења, медијанта претставува средна вредност помеѓу мерењата $n/2$ и $(n/2) + 1$, каде n претставува вкупниот број мерења.

Од пример 1:

30,56 30,80 30,94 31,07 31,12 31,74 31,98

X_{med}

Пример 2 (парен број мерења):

30,56 30,80 30,94 31,07 31,13 31,74 31,98 32,11

$n/2$ $n/2 + 1$

$$X_{\text{med}} = \frac{31,07 + 31,13}{2} = 31,10$$

Мерки за расејување

Отстапување на индивидуалните мерења во однос на централната вредност.

Мерките за расејување се дефинираат во однос на мерките за централната тенденција, но се независни од нив:

- ранг;
- стандардна девијација;
- варијанца.

Ранг

разлика помеѓу најголемата и најмалата вредност во едно множество податоци

$$\gamma = X_{\max} - X_{\min}$$

Од пример 1: $\gamma = (31,99 - 30,56) \mu\text{g/dL} = 1,42 \mu\text{g/dL}$

Стандардна девијација

Апсолутна стандардна девијација, опишува расеаност на податоците на индивидуалните мерења околу средната вредност:

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

Релативна стандардна девијација:

$$s_r = \frac{s}{\bar{X}}$$

Од пример 1:

$$\begin{aligned}(30,80 - 31,17)^2 &= 0,00137 \\(30,94 - 31,17)^2 &= 0,00053 \\(31,07 - 31,17)^2 &= 0,00010 \\(30,56 - 31,17)^2 &= 0,00372 \\(31,12 - 31,17)^2 &= 0,00003 \\(31,74 - 31,17)^2 &= 0,00325 \\(31,98 - 31,17)^2 &= 0,00656\end{aligned}$$

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{0,01556}{7-1}} = 0,051$$

$$s_r = \frac{0,051}{31,17} = 0,0016$$

Варијанца, s^2

Квадрат од стандардната девијација

Од пример 1: $s^2 = (0,051)^2 = 0,0026$

Експериментални грешки

Точност на анализа: грешки поврзани со централната тенденција

Прецизност на анализа: грешки поврзани со расејаноста

Точност

мерка за тоа колку е блиска вредноста за централната тенденција со вистинската или очекуваната вредност μ .

- Апсолутна грешка: $E = \bar{X} - \mu$ (X или X_{med})
- Релативна грешка: $E_r = \frac{\bar{X} - \mu}{\mu}$ (X или X_{med})

Прецизност

мерка за степенот на расејаноста на податоците околу централната вредност

- Повторливост (сите мерења од 1 аналитичар при исти услови);
- Репродуцибилност (мерења од различни аналитичари или со различни методи).

Определени грешки

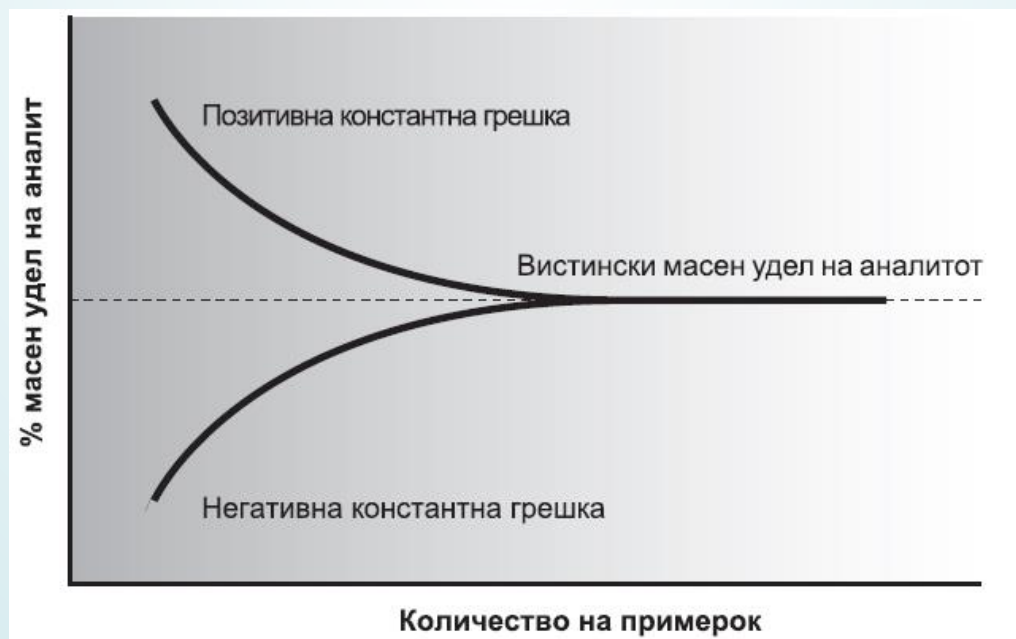
Секоја систематска грешка која доведува до добивање на преголеми или премали резултати при мерењето.

Потеклото може да се идентификува:

- Грешки при земање проба (хетерогени примероци);
- Грешки на методата (осетливост, треба да се стандардизира пробата, слепа проба);
- Грешки при мерење (ограничување на опремата и инструментите - толеранција);
- Лични грешки (субјективни, вградени ставови на аналитичарот).

Идентификација на определените грешки

- Константна определена грешка (**не зависи** од количеството земен примерок)
- Пропорционална определена грешка (**зависи** од количеството земен примерок)



Неопределени грешки (случајни)

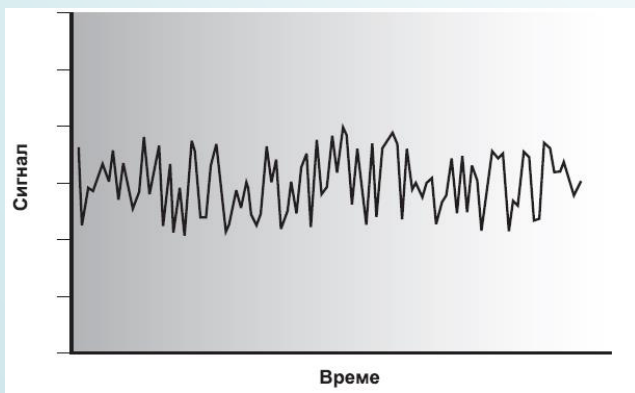
- Не се точно познати
- Флукутираат на случаен начин
- Еднакво веројатни се и позитивните и негативните

Настануваат при:

- Земање проба
- Манипулација со примерокот
- Изведувањето на анализата

Не е можно да се елиминираат, може да се намалат!

Се покоруваат на законот за веројатност



Елиминирање на неопределените грешки кои настануваат како резултат на шумот на инструментот.

Грешка и несигурност

- Грешка – разликата меѓу некое мерење или резултат и вистинската вредност.
- Несигурност – опсег на можни вредности од мерењето или резултатот.

Пренесување на несигурност

Мерење волумен со пипета (пипетирање)

Популација и примерок

- **Популација** – сите членови од еден систем (множество).
- **Примерок** (репрезентативен примерок) – членови од популацијата кои се анализирани.
- **Веројатност** за појава на некоја вредност $P(V)$:
$$P(V) = M/N$$
- **Распределба на веројатноста** – колку пати секој член од популацијата се појавува (крива на распределба):
 - биномна распределба и
 - нормална распределба.

Биномна распределба

Распределба на веројатноста, која ја прикажува можноста за добивање на еден од два можни резултати во текот на точно определен број повторувања. Дискретни вредности.

$$P(X, N) = \frac{N!}{X! (N - X)!} \times p^X \times (1 - p)^{N-X}$$

$$\mu = Np$$

$$\sigma^2 = Np(1 - p)$$

Пример 4.10, Харви

Нормална распределба

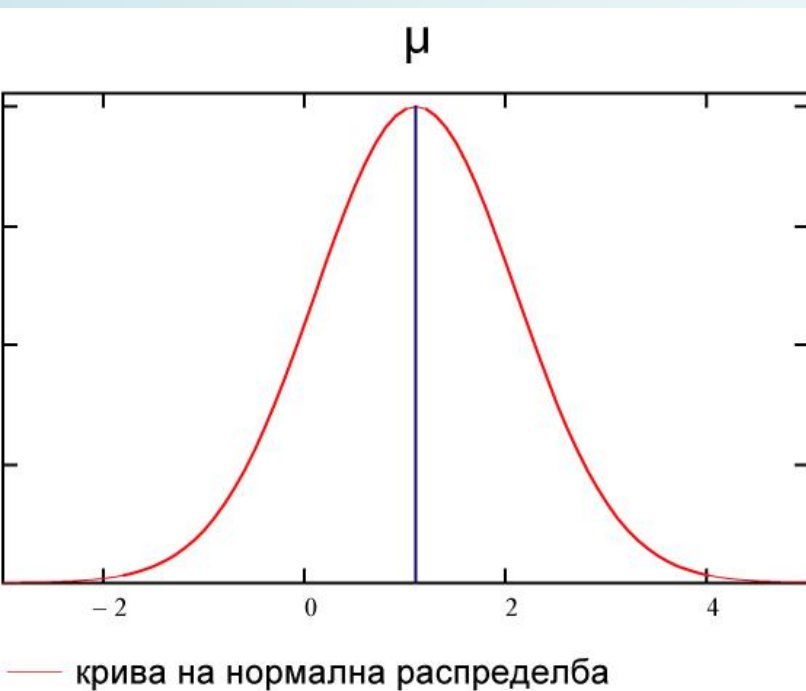
Распределба која укажува на последиците од случајната грешка кај експерименталните мерења. Има облик на своно. Континуирани вредности

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(X-\mu)^2}{2\sigma^2} \right]$$

$$\mu = \frac{\sum_{i=1}^n X_i}{n}$$

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}$$

Крива на нормална распределба, својства

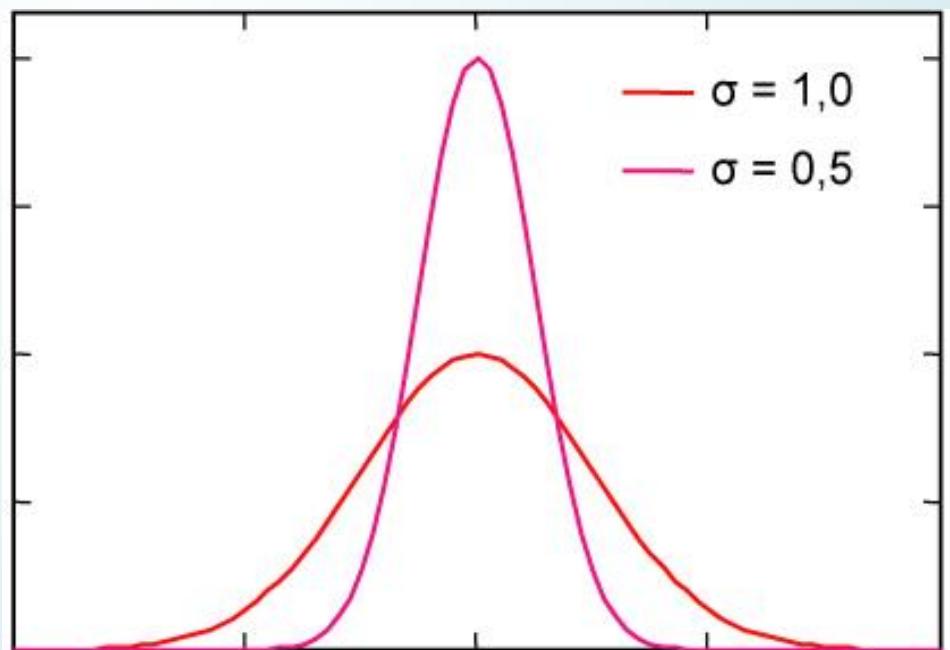


$$\mu = \frac{\sum_{i=1}^n X_i}{n}$$

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}$$

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(X-\mu)^2}{2\sigma^2} \right]$$

68,26 % од членовите на популацијата се во опсегот на вредности од $\mu \pm \sigma$



Интервали на доверба

Отстапување или ниво на доверба z

$$z = \frac{X - \mu}{\sigma}$$

X – гранична вредност

σ – стандардна девијација

μ – вистинска вредност

Која е најверојатната вредност на еден член од популацијата?

Одговор може да даде распределбата на веројатности!

Интервали на доверба

$z = 1,00$; доверба = 68,26 %

$z = 1,96$; доверба = 95,00 %

Пример 4.12, Харви

Интервали на доверба

$$\mu = \bar{X} \pm \frac{z\sigma}{\sqrt{n}}$$

z – ниво на доверба

σ – стандардна девијација (вистинска)

μ – вистинска вредност

$$\mu = \bar{X} \pm \frac{ts}{\sqrt{n}}$$

t – за сите нивоа на доверба ($t \geq z$) студентов коеф.

s – стандардна девијација (експериментална)

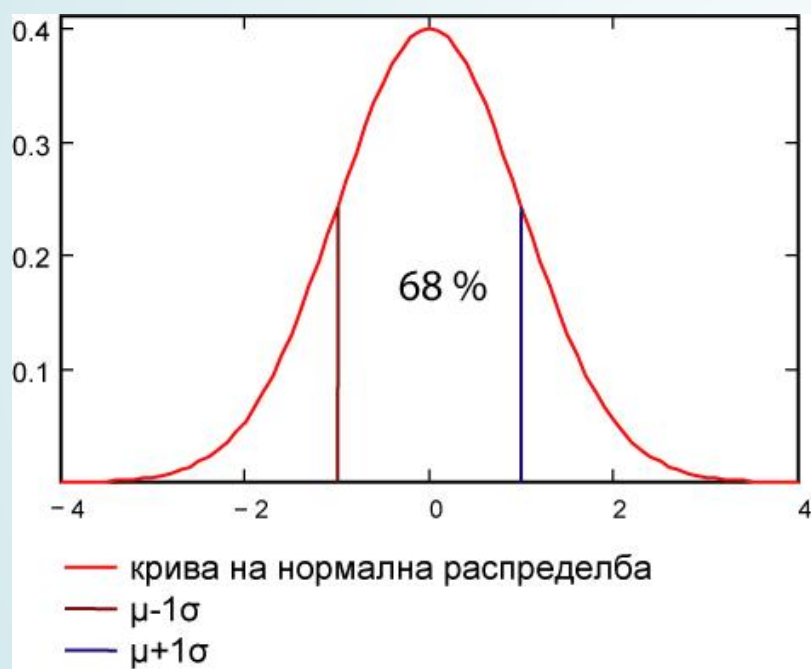
s^2 за случаен примерок, може значително
да се разликува од σ^2

Табела 4.14

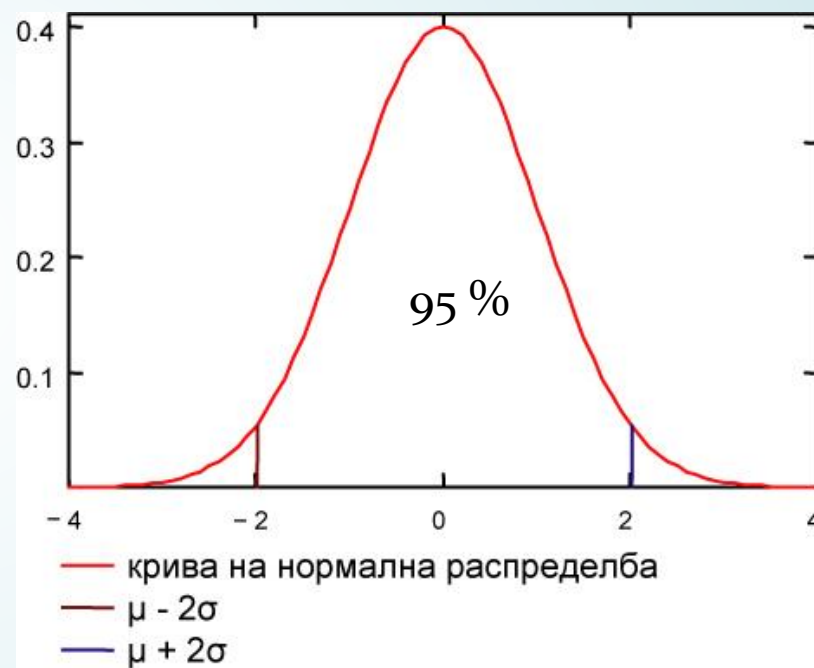
Вредности за t за
ниво на доверба од 95%

Степени на слобода	t
1	12,71
2	4,30
3	3,18
4	2,78
5	2,57
6	2,45
7	2,36
8	2,31
9	2,26
10	2,23
12	2,18
14	2,14
16	2,12
18	2,10
20	2,09
30	2,04
50	2,01
∞	1,96

Интервали на доверба



(од $\mu - 1\sigma$ до $\mu + 1\sigma$)

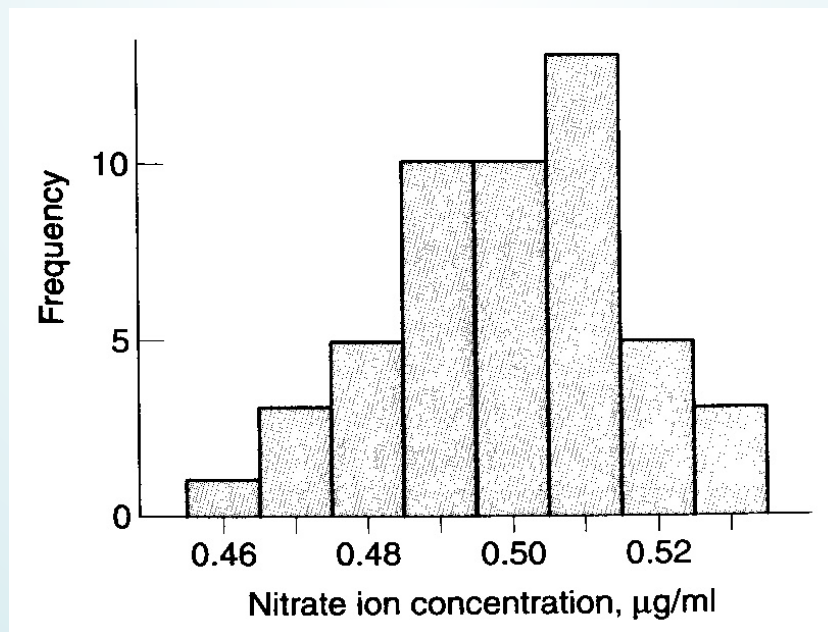


(од $\mu - 2\sigma$ до $\mu + 2\sigma$)

Случајни променливи

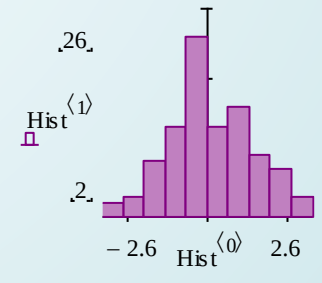
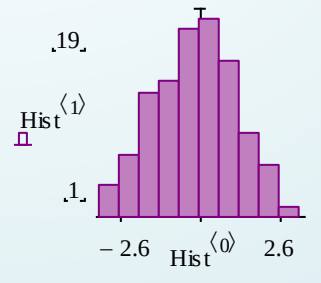
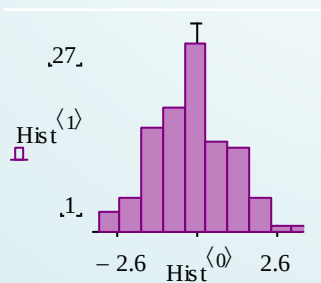
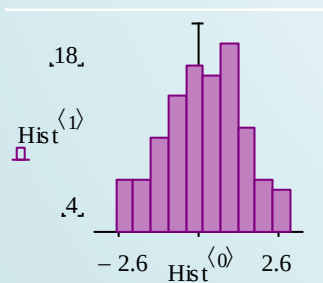
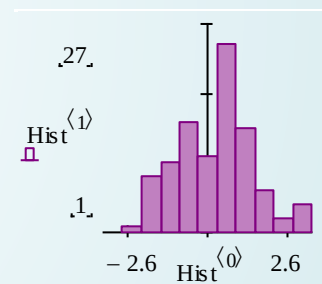
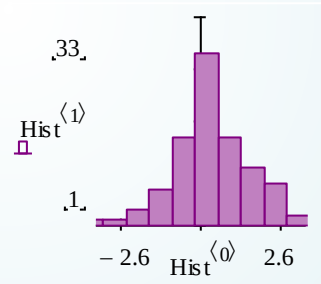
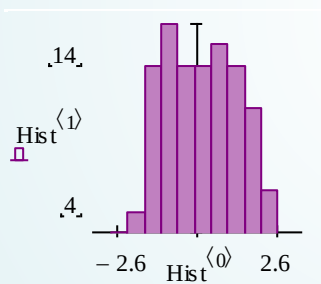
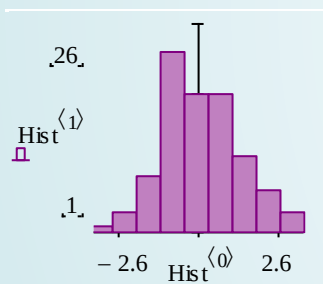
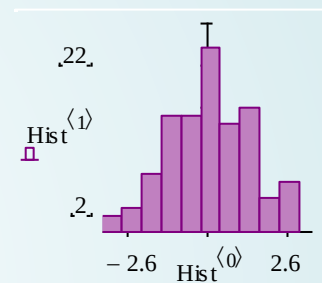
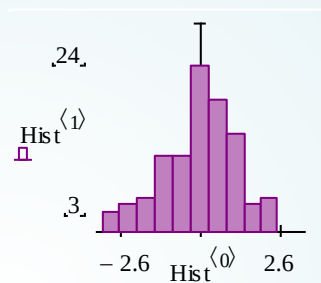
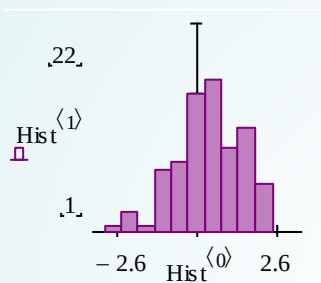
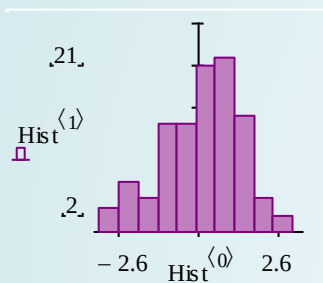
X – случајна променлива од дискретен тип

Хистограм – График кој дава информација за фреквенцијата на појава на дадена вредност како функција од набљудуваните интервали

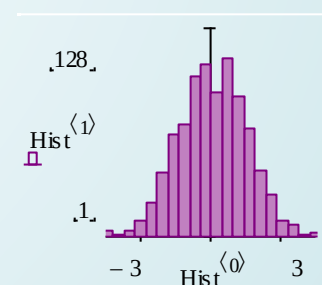
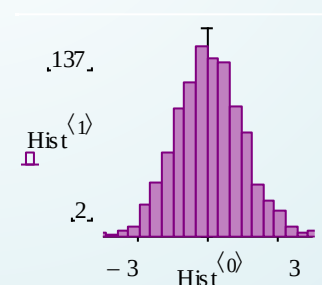
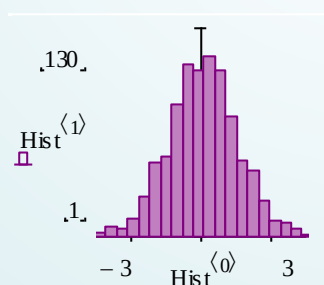
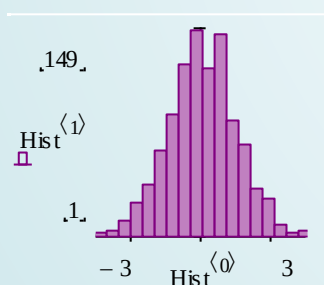
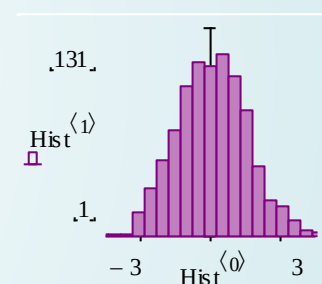
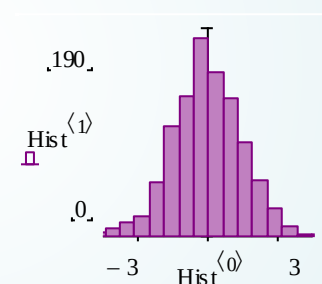
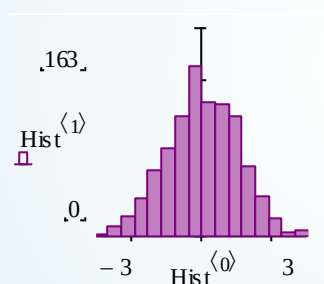
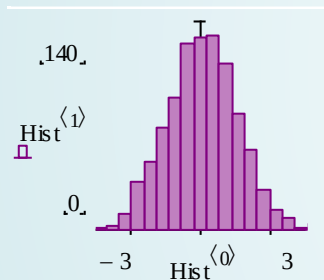
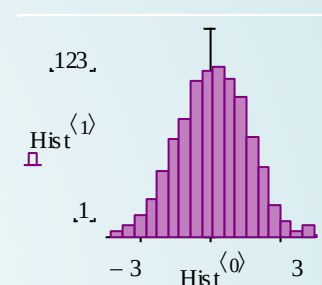
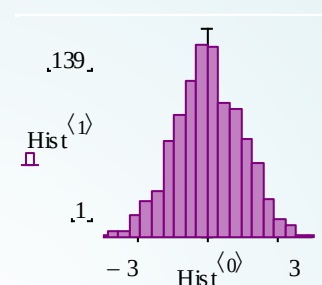
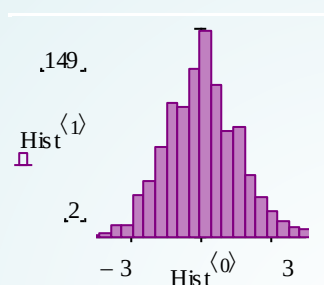
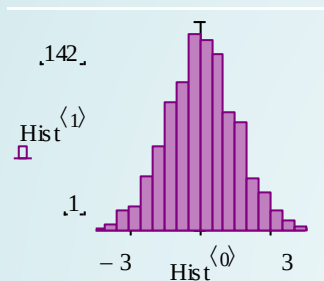


Нормална распределба

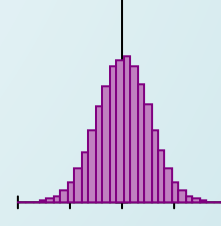
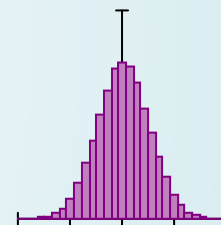
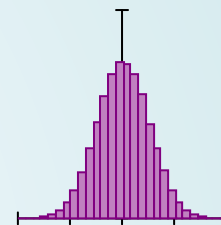
100 пати повторени експерименти



1000 пати повторени експерименти



100.000 пати повторени експерименти



Стандардизација

Со стандардизација, ова:

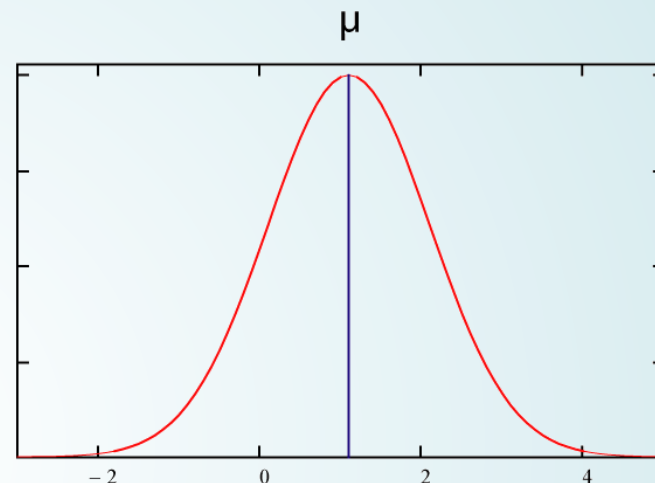
$$\sigma \neq 1$$

$$\mu \neq 0$$

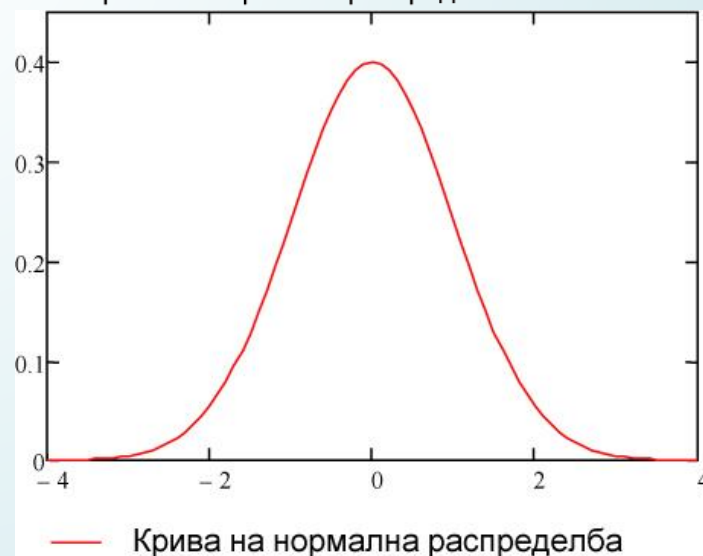
Поминува во ова:

$$\sigma = 1$$

$$\mu = 0$$



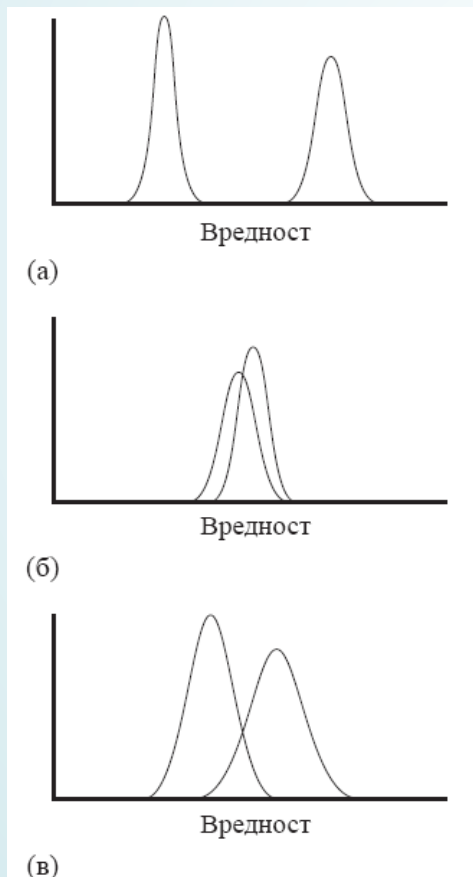
— крива на нормална распределба



— Крива на нормална распределба

Тестови на значајност

За извлекување на квантитативни заклучоци за начинот на кој експериментите се изведуваат



Ниво на шеќер во крвта. Две проби,
пред земање на лек, по земање на лек

Поставување хипотеза

- Нулта хипотеза, H_0 : разликата меѓу вредностите е резултат на неопределени грешки.
- Алтернативна хипотеза, H_A : разликата меѓу вредностите е голема за да се припише на случајни грешки

Вистинитоста на H_0 се испитува со тест на значајност.

Дефинирање ниво на значајност α

Ниво на доверба за задржување на H_0 , т.е. веројатноста дека H_0 ќе биде погрешно отфрлена.

$$\alpha = 1 - (\text{интервал на доверба})/100$$

Студентов t - тест

Статистички тест за споредба на две средни вредности, за да се определи дали разликата во добиените вредности е преголема за да се припише на неопределени грешки.

$H_0: \mu = \bar{X}$; H_A : разликата меѓу μ и $\bar{X}$ е голема

$$\mu = \bar{X} \pm \frac{t_{\text{експ.}} \cdot s}{\sqrt{n}}, \quad n = \frac{n_1 \cdot n_2}{n_1 + n_2}$$

$$t_{\text{експ.}} = \frac{|\mu - \bar{X}| \cdot \sqrt{n}}{s}$$

Од стандардизирани табели, се определува дали $t_{\text{експ.}}$ е во согласност со вредноста за t дадена во табелата. Ако тоа не е случај, хипотезата се отфрла. Заклучокот е дека постојат и определени грешки во методата и интервалот на доверба треба да се зголеми.

F - тест

Статистички тест за споредба на две варијации. Се применува за да се определи дали разликата во варијациите е голема за да се припише на неопределени грешки.

$$H_0: s^2 = \sigma^2; \quad H_A: s^2 \neq \sigma^2$$

$$F_{\text{експ.}} = \frac{s^2}{\sigma^2}, \quad \text{ако } s^2 > \sigma^2$$
$$\text{или } F_{\text{експ.}} = \frac{\sigma^2}{s^2} \quad \text{ако } \sigma^2 > s^2$$

Следува, $F_{\text{експ.}} \geq 1$

Пресметаната вредност за $F_{\text{експ.}}$ се споредува со таблични вредности и за одбраното ниво на доверба.

Ако $F_{\text{експ.}} > F_{\text{табела}}$, постои значајна разлика во прецизността.

Q - тест

Претставува статистички тест за детектирање на екстремни вредности. Се користи за отфрлање на сомнителните вредности.

- Еден резултат многу отстапува од другите.

$$Q = \frac{|\text{сомнителна вредност} - \text{најблиска вредност}|}{\text{најголема вредност} - \text{најмала вредност}}$$

Добиената вредност се споредува со табличните вредности за даден степен на доверба и соодветен број на мерења.

Ако $Q_{\text{експ.}} > Q_{\text{табела}}$ – резултатот се отфрла. Најверојатно е оптеретен со груби грешки.

Регресиона анализа

Ако меѓу величините x и y постои линеарна зависност, тогаш тие може да се поврзат со права:

$$y = a + bx$$

$$a = \frac{\sum y_i - b \sum x_i}{m}; \quad b = \frac{m \sum x_i \cdot y_i - \sum x_i \cdot \sum y}{m \sum x_i^2 - (\sum x_i)^2}$$

Коефициент на корелација – во колкава мера точките отстапуваат од правата:

$r = \pm 1$ одлична корелација

$r < 0,8$ слаба

$r \approx 0$ нема

