Mathematical statistics

A **random variable** (stochastic variable) assigns a probability (probability density) to a possible discrete (continuous) event from a certain discrete (continuous) set of events.

- Discrete example: dice, $p_i = 1/6$ for $i \in \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$
- Continuous example: time of nucleus decay, $p(t) = k\mathrm{e}^{-kt}$

A continuous random variable in 1D ($\boldsymbol{x} \in \mathbb{R}$) is described by a **distribution function**, density of probability, (continuous) probability distribution,... $p(x)$:

$$p(x)\mathrm{d}x = \text{probability that event } \boldsymbol{x} \in [x, x+\mathrm{d}x) \text{ occurs}$$

In 2D, $p(x, y)$ is defined so that event $x \in [x + \mathrm{d}x)$ and $y \in [y + \mathrm{d}y)$ happens with probability $p(x, y)\mathrm{d}x\mathrm{d}y$.

Normalization:

$$\sum_i p_i = 1 \quad \text{or} \quad \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$$

Cumulative (integral) distribution function = probability that $\boldsymbol{x} \le x$:

$$P(x) = \int_{-\infty}^{x} p(x')\mathrm{d}x'$$

Probability distribution

Warning. In physics etc., symbol $\boldsymbol{x}$ (random variable) and x (a value, e.g., in integration) are not distinguished.

Mean value, expectation value (not averaged value = arithmetic average of a sample):

$$\mathrm{E}\,(\boldsymbol{x}) \equiv \langle \boldsymbol{x} \rangle \equiv \langle x \rangle_{\boldsymbol{x}} \overset{\text{loosely}}{=} \langle x \rangle = \int x p(x) \mathrm{d}x \quad \text{or} \quad \sum_i x_i p_i$$

Example. It holds $p(x) = \mathrm{e}^{-x}$ (exponential distribution). Calculate $\langle x \rangle$.

1

Variance, fluctuation, dispersion, mean square deviation (MSD)

$$\mathrm{Var}\,(\boldsymbol{x}) \overset{\text{loosely}}{=} \mathrm{Var}\,x = \langle (x - \langle x \rangle)^2 \rangle = \langle \Delta x^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2, \quad \text{where } \Delta x = x - \langle x \rangle$$

Standard deviation = $\sqrt{\mathrm{Var}\,(\boldsymbol{x})}$, denoted as: $\sigma(\boldsymbol{x})$, $\sigma(x)$, δx

Example. Let distribution $\boldsymbol{u}$ be uniform in interval $[0, 1)$. Calculate the expectation and the variance.

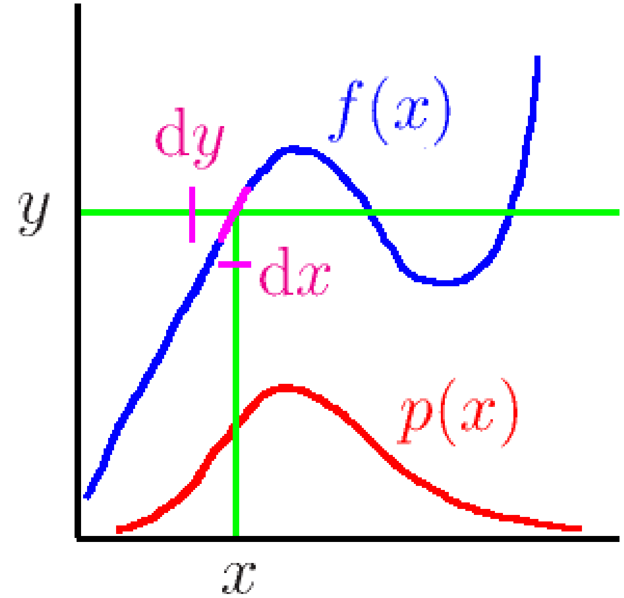
$\langle \boldsymbol{u} \rangle = 1/2$, $\mathrm{Var}(\boldsymbol{u}) = 1/12$; cf. mmpc5.mw:function of random variable)

Function of random variable

Let $\boldsymbol{x}$ be a real random variable with distribution $p(x)$, and $f(x)$ be a real function. A quantity (observable) $f(\boldsymbol{x})$ has the distribution

$$p_f(y) = \sum_{x:f(x)=y} \frac{p(x)}{|f'(x)|}$$

where the sum is over all roots.



Example. Let $\boldsymbol{u}$ be uniform in $u \in [0,1)$. Calculate the distribution function of $t = -\ln u$.

$\exp(-t)$: e.g., time of atom decay for $k = 1$

```
> restart:
> with(Statistics):
> rectf := t->piecewise(t<0,0, t<1,1, 0);
> Rect := Distribution(PDF=(rectf));
> X := RandomVariable(Rect);
> Mean(X); StandardDeviation(X);
> PDF(-log(X),x);
```

Example: Gini coefficient (index)

Measure of income inequality. Income x with probability density $p(x)$, $x \geq 0$.

$$G = \frac{1}{2\langle x \rangle} \int_0^\infty p(x)\mathrm{d}x \int_0^\infty p(y)\mathrm{d}y\, |x-y|, \quad G \in [0,1]$$

Example. Calculate the Gini coefficient for

a) Dirac delta-distribution (all have the same income);

b) exponential distribution of incomes.

a) 0; b) 1/2

```
> restart:
> Gini:=p->int(p(x)*int(p(y)*abs(x-y),y=0..infinity),x=0..infinity)
        /2/int(p(x)*x,x=0..infinity)
> assume(a>0);
> p:=x->Dirac(x-a);
> int(p(x),x=0..infinity);
> Gini(p);
> p:=x->a*exp(-x*a);
> int(p(x),x=0..infinity);
> Gini(p);
```

Function of random variable: mean value

Mean value of quantity f:

$$\langle f\rangle = \int f(x)p(x)\mathrm{d}x \tag{1}$$

Or based on new random variable $\boldsymbol{f} = f(\boldsymbol{x})$:

$$\langle f\rangle = \int y p_f(y)\mathrm{d}y \tag{2}$$

Both mean values are the same:

$$\langle f\rangle = \int f(x)p(x)\mathrm{d}x \overset{\text{subst. } y=f(x)}{=} \int \frac{yp(x)}{f'(x)}\mathrm{d}y = \int y p_f(y)\mathrm{d}y$$

where in the 2nd $\int$, x = root of equation $f(x) = y$, for simplicity we assume: there is only one root, function f is increasing.

Note. Unified and more general description is based on the probability measure μ on a space – so far we have used $\mathbb{R}$, $\mathbb{R}^2$, and a discrete space. We write, e.g., $\langle f\rangle_\mu = \int f(x)\mathrm{d}\mu(x)$ instead of (1) or (2).

Covariance

- Covariance of a 2D distribution:

$$\mathrm{Cov}\,(\boldsymbol{x},\boldsymbol{y}) = \langle \Delta x \Delta y \rangle = \int \Delta x \Delta y p(x,y) \mathrm{d}x \mathrm{d}y$$

- Covariance of two quantities $f(x)$ a $g(x)$ (similarly for a 2D or discrete variable)

$$\mathrm{Cov}\,(f,g) = \langle \Delta f \Delta g \rangle = \int \Delta f \Delta g\, p(x) \mathrm{d}x$$

Independent random variables

Random variables $\boldsymbol{x}$ (with distribution $p_1(x)$) and $\boldsymbol{y}$ (with $p_2(y)$):

$$p(x,y) = p_1(x) p_2(y) \tag{3}$$

In the discrete case (throw a dice twice, $p_{ij} = 1/36$):

$$p_{ij} = p_{1,i} p_{2,j}$$

Covariance of two independent random variables is zero

$$\mathrm{Cov}\,(\boldsymbol{x},\boldsymbol{y}) = \langle \Delta x \Delta y \rangle_{\boldsymbol{x}+\boldsymbol{y}} = \int \mathrm{d}x \int \mathrm{d}y\, \Delta x p_1(x) \Delta y p_2(y) = \langle \Delta x \rangle_{\boldsymbol{x}} \langle \Delta y \rangle_{\boldsymbol{y}} = 0$$

Correlation coefficient

$$r(x,y) = \frac{\mathrm{Cov}\,(x,y)}{\sqrt{\mathrm{Var}\,(x)\mathrm{Var}\,(y)}}$$

Example. Let $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ be two independent random variables with uniform distribution in [0,1]. Calculate:

a) $r(u_1, -u_1)$

b) $r(u_1^2, u_1^2)$

c) $r(u_1, u_2 + u_1)$ (see Maple)

a) -1, b) 1, c) $1/\sqrt{2}$

```
tab 1 100000 | tabproc "rnd(0)" "rnd(0)" | tabproc A A+B | lr
```

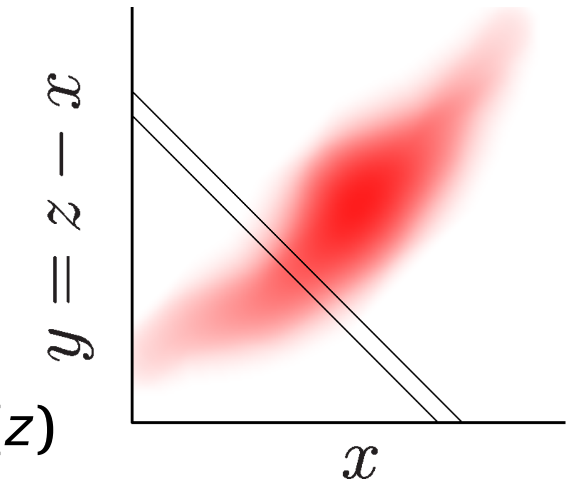
Sum of random variables

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two continuous random variables with distribution $p(x,y)$. The distribution of $\boldsymbol{x}+\boldsymbol{y}$ is

$$p_{\boldsymbol{x}+\boldsymbol{y}}(z)\mathrm{d}z = \int\int_{x+y\in(z,z+\mathrm{d}z)} p(x,y)\mathrm{d}x\mathrm{d}y \overset{y:=z-x}{=} \int p(x,z-x)\mathrm{d}x\mathrm{d}z$$

$\Rightarrow$

$$p_{\boldsymbol{x}+\boldsymbol{y}}(z) = \int p(x,z-x)\mathrm{d}x$$



Now, let $p(x,y) = p_1(x)p_2(y)$. Then

$$p_{\boldsymbol{x}+\boldsymbol{y}}(z) = \int p_1(x)p_2(z-x)\mathrm{d}x \equiv (p_1 * p_2)(z)$$

$p_1 * p_2$ is called the **convolution**.

Discrete example: Let's roll two dice. What is the distribution of the sum of points?

$p(2) = 1/36$, $p(3) = 2/36$, … $p(7) = 6/36$, … $p(12) = 1/36$

Example. Calculate the distribution of $\boldsymbol{u}_1 - \boldsymbol{u}_2$

0 for $|x| > 1$, $1 - |x|$ otherwise

```
tab 1 100000 | tabproc "rnd(0)-rnd(0)" | histogr -1.5 1.5 .1 | plot -
```

Sum of independent random variables

Mean value and variance of independent random variables are additive.

Directly using (3):

$$\mathrm{E}\,(\boldsymbol{x}+\boldsymbol{y}) = \int p_1(x)p_2(y)(x+y)\mathrm{d}x\mathrm{d}y$$

$$= \int p_1(x)p_2(y)x\mathrm{d}x\mathrm{d}y + \int p_1(x)p_2(y)y\mathrm{d}x\mathrm{d}y = \int p_1(x)x\mathrm{d}x + \int p_2(y)y\mathrm{d}y = \mathrm{E}\,(\boldsymbol{x}) + \mathrm{E}\,(\boldsymbol{y})$$

Using the convolution of the distributions:

$$\mathrm{E}\,(\boldsymbol{x}+\boldsymbol{y}) = \int z p_{\boldsymbol{x}+\boldsymbol{y}}(z)\mathrm{d}z = \int z p_1(x)p_2(z-x)\mathrm{d}x\mathrm{d}z$$

$$\overset{y:=z-z}{=} \int (x+y)p_1(x)p_2(y)\mathrm{d}x\mathrm{d}y = \langle x\rangle_1 + \langle y\rangle_2 = \mathrm{E}\,(\boldsymbol{x}) + \mathrm{E}\,(\boldsymbol{y})$$

And the variance:

$$\mathrm{Var}\,(\boldsymbol{x}+\boldsymbol{y}) = \langle(\Delta x+\Delta y)^2\rangle_{\boldsymbol{x}+\boldsymbol{y}}$$

$$= \langle(\Delta x)^2\rangle_{\boldsymbol{x}+\boldsymbol{y}} + 2\langle\Delta x\Delta y\rangle_{\boldsymbol{x}+\boldsymbol{y}} + \langle(\Delta y)^2\rangle_{\boldsymbol{x}+\boldsymbol{y}} = \mathrm{Var}\,(\boldsymbol{x}) + \mathrm{Var}\,(\boldsymbol{y})$$

Central limit theorem

The sum of n equal independent distributions with a finite mean value and variance limits for $n \to \infty$ to the Gaussian distribution (aka normal distribution) with the mean value $n\langle x\rangle$ and variance $n\,\mathrm{Var}\,x$.

Example. Let us consider a discrete distribution $\boldsymbol{b}$: $p(-1/2) = p(1/2) = 1/2$. Let us approximate the sum of n such distributions:

$$
\begin{aligned}
n &= 1 \qquad p(-1/2) = 1/2,\ p(1/2) = 1/2,\ \mathrm{Var}\,\boldsymbol{b} = 1/4 \\
n &= 2 \qquad p(-1) = 1/4,\ p(0) = 1/2,\ p(1) = 1/4,\ \mathrm{Var}\,\boldsymbol{b}^2 = 2/4 \\
n &= 3 \qquad p(\pm 3/2) = 1/8,\ p(\pm 1/2) = 3/8,\ \mathrm{Var}\,\boldsymbol{b}^3 = 3/4
\end{aligned}
$$

Let n be even (for simplicity). Then for $k = -n/2..n/2$:

$$
p(k) = \binom{n}{n/2+k} 2^{-n} \approx \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \sigma^2 = \mathrm{Var}\,(\boldsymbol{b}^n) = \frac{n}{4}
$$

where we have used the Stirling formula $n! \approx n^n \mathrm{e}^{-n} \sqrt{2\pi n}$

See Maple for numerical verification using convolution of rectangular distributions

Gauss' distribution and Chebyshev's inequality

For random variable $\boldsymbol{x}$ with the Gauss' (normal) distrubution it holds:

$$\text{prob}\left(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge t\sigma(\boldsymbol{x})\right) = 2\int_t^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}} = \text{erfc}(k/\sqrt{2})$$

e.g., $\text{prob}\left(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge 2\sigma(\boldsymbol{x})\right) = 0.0455 \approx 5\,\%$

t	normal	general
	$\text{prob}\left(\lvert\boldsymbol{x}-\langle\boldsymbol{x}\rangle\rvert \le t\sigma(\boldsymbol{x})\right)$	
1	68.27 %	≥ 100 %
2	95.45 %	≥ 75 %
3	99.73 %	≥ 88.89 %

Chebyshev's inequality: For a general random variable $\boldsymbol{x}$ with finite mean and variance it holds:

$$\text{prob}\left(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge t\sigma(\boldsymbol{x})\right) \le \frac{1}{t^2}$$

e.g., $\text{prob}\left(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge 2\sigma(\boldsymbol{x})\right) = 25\,\%$

Proof. Let's define (as in C/C++): $(x \le 1) = 1$ for $x \le 1$ and $(x \le 1) = 0$ otherwise.

$$\text{prob}\left(|\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge t\sigma(\boldsymbol{x})\right) = \left\langle |\boldsymbol{x}-\langle\boldsymbol{x}\rangle| \ge t\sigma(\boldsymbol{x}) \right\rangle$$

$$= \left\langle \left(\frac{\boldsymbol{x}-\langle\boldsymbol{x}\rangle}{t\sigma(\boldsymbol{x})}\right)^2 \ge 1 \right\rangle \le \left\langle \left(\frac{\boldsymbol{x}-\langle\boldsymbol{x}\rangle}{t\sigma(\boldsymbol{x})}\right)^2 \right\rangle = \frac{1}{t^2}$$

equality for: $X = \begin{cases} -1, & p = \frac{1}{2t^2} \\ 0, & p = 1-\frac{1}{t^2} \\ +1, & p = \frac{1}{2t^2} \end{cases}$

Mathematical statistics and metrology

The terminology is field-dependent...

Statistic, estimator, “statistical algorithm”, (narrower) “statistical functional”, in metrology “measurement function”, is a formula/algorithm by which a result is calculated from (a sample of) random variables (from data in metrology). A statistic is a random variable, too.

Examples: arithmetic average, parameters of a model in fitting by the least-square method

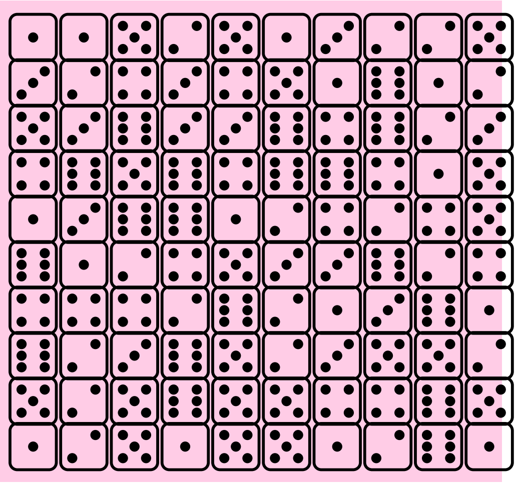
Standard error of a statistic = standard deviation (square root of variance) of the distribution function of the statistic.

Uncertainty (in metrology) includes critical assessment of systematic, random, discretization etc. errors. Similarly as above: “standard uncertainty”.

Distinguish:

- statistic = estimator
- statistics = field of mathematics

Arithmetic average as an example of statistic



Let us have a **sample** of a random variable.
Examples:

- shoe sizes of 1000 people
- 100× rolled dice
- pressure during a simulation

Arithmetic average (sample average, sample mean):

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

for simplicity, I write $\langle x \rangle$ instead of $\langle \boldsymbol{x} \rangle$

It is an **unbiased** estimate of $\langle x \rangle$ because

$$\langle \overline{x}_n \rangle = \langle x \rangle$$

$\sigma(x) \equiv \sqrt{\mathrm{Var}\, x}$

Let's calculate the variance of $\overline{x}_n$:

$$\mathrm{Var}\,(\overline{x}_n) = \langle (\overline{x}_n - \langle x \rangle)^2 \rangle = \left\langle \left(\frac{1}{n}\sum_{i=1}^{n} \Delta x_i \right)^2 \right\rangle = \frac{\mathrm{Var}\, x}{n} \equiv \frac{\sigma(x)^2}{n}, \quad \Delta x_i = x_i - \langle x \rangle$$

We assumed that x_i's are independent, $\langle \Delta x_i \Delta x_j \rangle = 0$ for $i \neq j$.

Standard deviation as an example of statistic

How to estimate $\sigma(x)^2$? We do not know $\langle x\rangle$, but only its estimate, $\overline{x}_n$.

$$\sigma^2(x) = \langle (x - \langle x\rangle)^2 \rangle \approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2$$

$$= \frac{1}{n} n\left[(1-\frac{1}{n})x_1 - \frac{1}{n}x_2 + \cdots\right]^2 = \frac{n-1}{n}\sigma(x)^2$$

Hence for the **corrected sample variance**

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

(1 = number of degrees of freedom) it holds

$$\left\langle \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \right\rangle = \sigma^2(x)$$

so it is an **unbiased** estimate of $\sigma^2(x)$.

But it's square root is a biased estimate of $\sigma(x)$.

Similarly, the **corrected sample variance of the arithmetic average** is

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

The "uncorrected" sample variances do not contain term -1.

The correction comes from Friedrich Wilhelm Bessel.

Summary

For processing of uncorrelated data by the arithmetic average with equal weights, it holds:

- Standard deviation of random variable x = standard error (uncertainty) of one measurement:

$$\sigma(x) = \sqrt{\langle (x - \langle x \rangle)^2 \rangle}$$

is approximated by

$$s_n(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2}$$

- standard error (standard uncertainty) of the arithmetic average $\overline{x}_n$ = uncertainty, with which $\overline{x}_n$ approximates $\langle x \rangle$:

$$\sigma(\overline{x}_n) = \sigma(x)/\sqrt{n}$$

and we calculate (approximate) it by

$$s_n(\overline{x}_n) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2}$$

Habits

We write the result of statistical processing as

$$\underline{\text{quantity}} = \underline{\text{estimate of quantity}} \pm \underline{\text{estimate of error}^\dagger}$$

Physics: $\underline{\text{estimate of error}^\dagger}$ = σ = estimated (standard) error†; loosely (estimated) error†; standard deviation (assumed of the arithmetic average or other statistic).

Common notation: $123.4 \pm 0.5 \equiv 123.4(5) \equiv 123.4_5$

In case of Gaussian distribution, the data are with 68 % probability within the bounds.

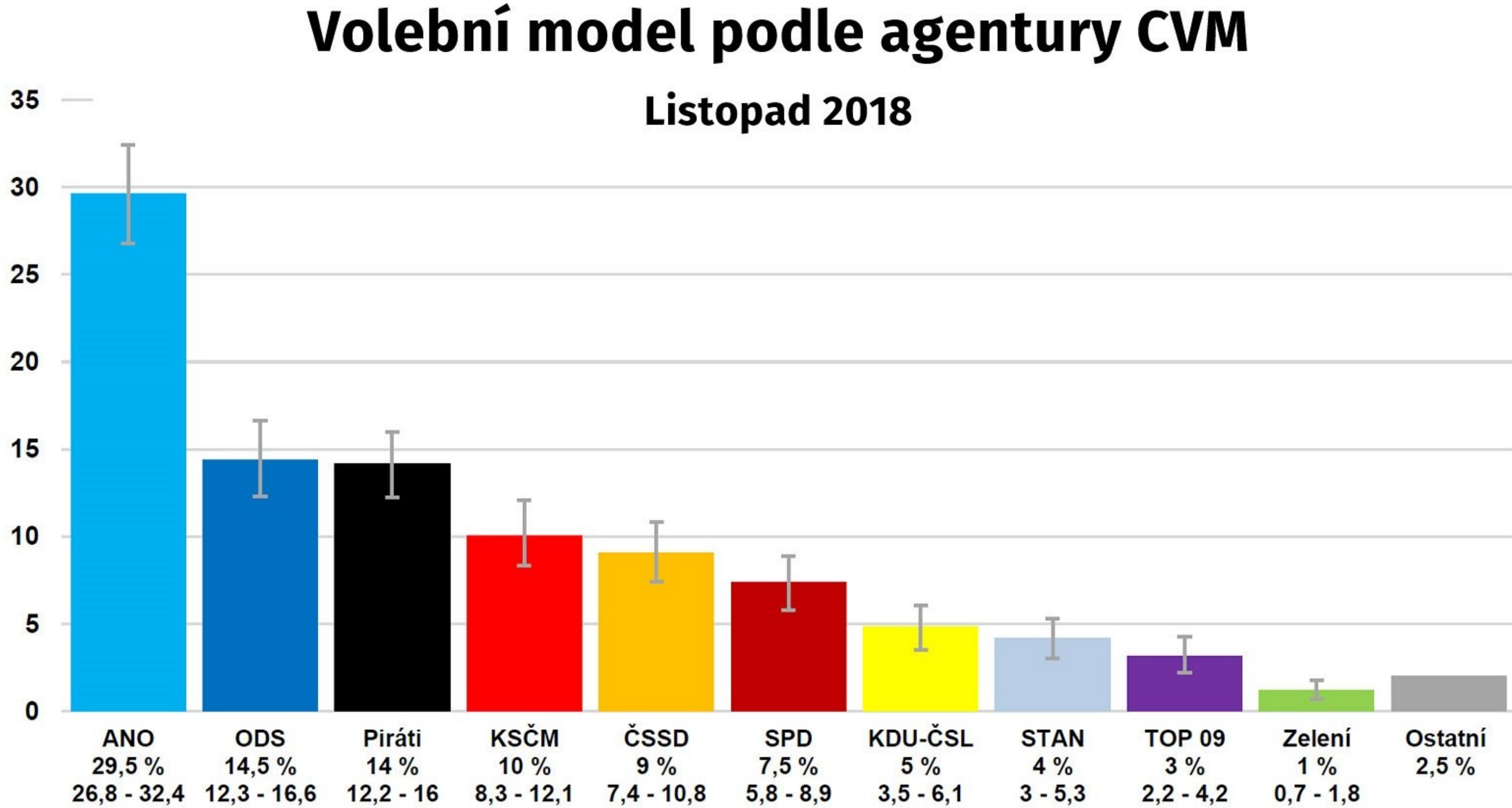
Biology, economy, engineering: Confidence level of 95 % is common (data are with 95 % probability within the bounds); recently, it has been criticized as insufficient. In case of Gaussian distribution:

$$\underline{\text{estimate of error}^\dagger} = 2 \times (\text{estimated standard error})$$

Chemistry: often ignored or nobody knows if σ or 2σ...

The type of the error must be specified!

†or uncertainty



In the opinion poll, 1080 people were asked about their preferences. Determine the confidence level of the error bars shown.

Hint: calculate first the variance of random variable yielding 1 with probability p and 0 otherwise.

$p(1-p)$; 95 %

Testing a hypothesis

Null hypothesis: The hypothesis that a feature (as a particular quantity value, a difference, etc.) derived from the data sample is due to sampling or experimental error and it is not significant.

Example: Students measure their pulse rates (PR). Is the mean pulse rate for college age women equal to 72 (a long-held standard for average pulse rate)?

- Null hypothesis (H_0): $\langle \mathrm{PR} \rangle = 72$
- Alternate (alternative) hypothesis (H_a): $\langle \mathrm{PR} \rangle \neq 72$

From $n = 300$ measurements, we got: $\overline{\mathrm{PR}}_n = 73.23(55)$; i.e., $s_n(\overline{\mathrm{PR}}_n) = 0.55$

For $n = 300$, we can assume that the distribution of $\overline{\mathrm{PR}}_n$ is normal and $s_n(\overline{\mathrm{PR}}_n)$ is accurate enough.

$$t = \frac{\overline{\mathrm{PR}}_n - \langle \mathrm{PR} \rangle_\mathrm{null}}{s_n(\overline{\mathrm{PR}}_n)} = \frac{73.23 - 72}{0.55} = 2.24 \quad (\text{“}2.24\sigma\text{”})$$

$$p = 2\int_t^\infty \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}} = \mathrm{erfc}(k/\sqrt{2}) = 0.025$$

The null hypothesis can be rejected at the 95 % confidence level.

See mmpc5.mw “Normal distribution example”

Student's t-distribution

If x is normal-distributed, random variable $\overline{x}_n$ has the Gauss' distribution with mean value $\langle \overline{x}_n \rangle = \langle x \rangle$ and standard deviation $\sigma(\overline{x}_n) = \sqrt{\mathrm{Var}\, x/n}$. But we have their estimates only – we cannot generally say that $\overline{x}_n$ is within ± estimated $\sigma(\overline{x}_n)$ with probability 68 %.

Let us define the Student's t-distribution with parameter ν (number of degrees of freedom) as the distribution of

$$\frac{\overline{x}_{\nu+1} - \langle x \rangle}{\sigma(\overline{x}_{\nu+1})}$$

$$\Gamma(x) = \int_0^\infty x^{n+1} \mathrm{e}^{-x} \mathrm{d}x,$$
$$\Gamma(n) = (n-1)!,$$
$$\Gamma(n+\tfrac{1}{2}) = \sqrt{\pi}\cdot\tfrac{1}{2}\cdot\tfrac{3}{2}\cdots(n-\tfrac{1}{2})$$

The distribution function is

$$t_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

The large-sample limit is the normalized Gauss' distribution

$$\lim_{\nu\to\infty} t_\nu(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$

Warning: $t_1(x)$ has infinite variance and (strictly) undefined mean value.

See mmpc5.mw “Gauss' (normal) and Student's t-distribution”

Pulse rates again

We have measured 8 persons only: PR = [69,84,67,82,71,81,73,71,76,86], $\overline{\mathrm{PR}}_n = 76$

- Null hypothesis: $\langle \mathrm{PR} \rangle < 72$
- Alternate hypothesis: $\langle \mathrm{PR} \rangle \geq 72$ (one tail)

$$t = \frac{\overline{\mathrm{PR}}_n - \langle \mathrm{PR} \rangle_{\mathrm{null}}}{s_n(\overline{\mathrm{PR}}_n)} = 1.865, \qquad p = \int_t^\infty t_{n-1}(x)\mathrm{d}x = 0.0475 < 0.05$$

The null hypothesis is rejected at the 95 % confidence level, $\langle \mathrm{PR} \rangle < 72$ is improbable.

We may be wrong, this is the “type I error” or “false positive” because we incorrectly accept “our” alternate hypothesis.

- Null hypothesis: $\langle \mathrm{PR} \rangle = 72$
- Alternate hypothesis: $\langle \mathrm{PR} \rangle \neq 72$ (two tails)

$$t = \frac{\overline{\mathrm{PR}}_n - \langle \mathrm{PR} \rangle_{\mathrm{null}}}{s_n(\overline{\mathrm{PR}}_n)} = 1.865, \qquad p = 2\int_t^\infty t_{n-1}(x)\mathrm{d}x = 0.095 > 0.05$$

Not enough evidence to reject the hypothesis, $\langle \mathrm{PR} \rangle = 72$ is quite likely.

We may be wrong, this is the “type II error” or “false negative” because we incorrectly reject “our” alternate hypothesis.

Comparison of two samples

Let us compare two samples (n and m pieces of data, denoted as x_i and y_i) drawn from the same distribution.

Random variable

$$t = \frac{\overline{x}_n - \overline{x}_m}{s\sqrt{1/n + 1/m}}, \quad \text{where } s^2 = \frac{(n-1)[s_n(x)]^2 + (m-1)[s_m(y)]}{n+m-2}$$

has the Student's t-distribution.

- σ_n is the corrected standard deviation of the data (not average)
- For $n = m$, it holds $s^2 = [s_n(\overline{x}_n)]^2 + [s_m(\overline{y}_m)]^2$
- Typical task: We have two sets of measurements obtained in such a way that the expected variances are the same.
  Null hypothesis: Do both means match?

Useful applets:

- `https://stattrek.com/online-calculator/t-distribution.aspx`
- `https://surfstat.anu.edu.au/surfstat-home/tables/t.php`

Excel, LibreOffice: function `T.TEST(array1,array2,tails,type)`

Example (see mmpc5.mw)

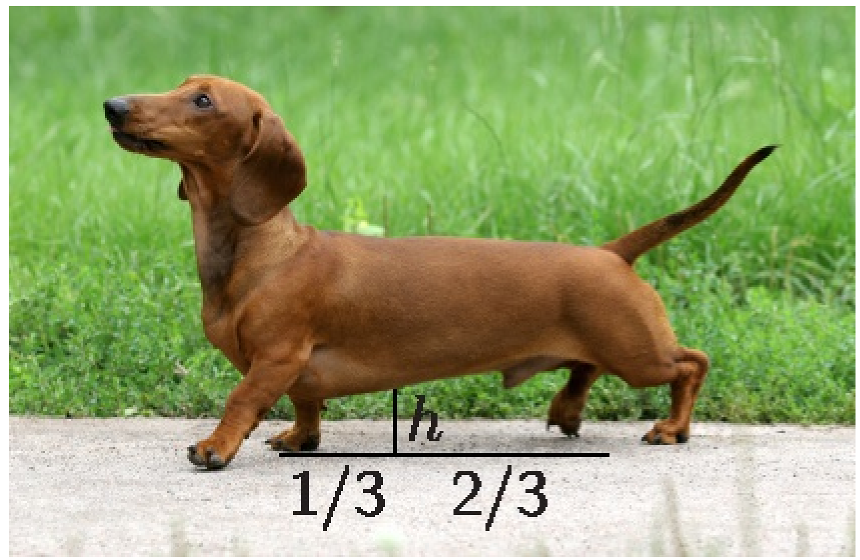
A company produces supports for too long dachshunds. The necessary measurements were outsourced to two companies which measured (in cm):

Company SmileyDog: $x = [12.1, 20, 15.1, 20.8, 19.7]$ cm

Company HappyDog: $y = [18.9, 10.1, 12.1, 9.2, 12.4, 16.7, 12.7]$ cm

a) Are the results in agreement (at the 95 % confidence level)?

b) What is the best estimate of the support height?



a) assuming the same variances: $t = 2.08$, $p = 0.064$ ⇒ both measurements likely do agree

b) 15.0(12) cm

Weights

σ = std.err.

A weighted average (mean):

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

Let us know x_i (independent random variables) with standard errors σ_i.
Which weights are the best?

We will derive the result for two quantities; $w_1 = w$, $w_2 = 1 - w$ (normalized)

$$\overline{x} = w x_1 + (1 - w) x_2$$

$$\sigma^2(\overline{x}) = \langle(\overline{x} - \langle x \rangle)^2\rangle = \langle(w \Delta x_1 + (1 - w)\Delta x_2)^2\rangle = w^2 \sigma_1^2 + (1 - w)^2 \sigma_2^2$$

The minimum is for

$$w = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}, \quad 1 - w = w_2 = \frac{1/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}$$

Consequently (can be generalized to more variables)

$$w_i = \frac{1}{\sigma_i^2}$$

But one must be careful if σ_i are known with a low precision.

Averaging of independent measurements

1. Known weights of data. E.g. (unnormalized) $w_i \approx n_i \gg 1$ (each x_i is a result of processing of many independent measurements), $w_i \approx$ time in simulation,. . .) and σ_i. Then

$$\overline{x} = \frac{\sum_{i=1}^{m} w_i x_i}{\sum_{i=1}^{m} w_i}, \quad \sigma = \frac{\sqrt{\sum_{i=1}^{m} w_i^2 \sigma_i^2}}{\sum_{i=1}^{m} w_i}$$

If available, better use information on w_i rather than $w_i \propto 1/\sigma_i^2$!

Unknown weights of data. Then $w_i = 1/\sigma_i^2$ (assuming that σ_i are accurate enough) and using the above formula

$$\overline{x} = \frac{\sum_{i=1}^{m} x_i/\sigma_i^2}{\sum_{i=1}^{m} 1/\sigma_i^2}, \quad \sigma = \frac{1}{\sqrt{\sum_{i=1}^{m} 1/\sigma_i^2}}$$

3. Few data. Data are samples n_i measurements, where x_i are averages and σ_i are the respective standard error estimates. Then

$$\overline{x} = \frac{\sum_{i=1}^m n_i x_i}{\sum_{i=1}^m n_i}, \quad \sigma = \sqrt{\frac{\sum_{i=1}^m n_i(n_i-1)\sigma_i^2 + \sum_{i=1}^m n_i(x_i-\overline{x})^2}{\left(\sum_{i=1}^m n_i - 1\right)\sum_{i=1}^m n_i}}$$

are the same as if all data are merged.

Example. Accordning to the dachshunds data:

$$x = [12.1, 20, 15.1, 20.8, 19.7] \quad : \quad \overline{x}_5 = 17.54 \pm 1.68$$

$$y = [18.9, 10.1, 12.1, 9.2, 12.4, 16.7, 12.7] \quad : \quad \overline{y}_7 = 13.16 \pm 1.31$$

Calculate the best estimate of the support height by all three methods.

1. $w_i = n_i$: 14.983 ± 1.040
2. $w_i = 1/\sigma_i^2$: 14.812 ± 1.036
3. 14.983 ± 1.185 (the same as for merged data)