RANDOM VARIABLE and its CHARACTERISTICS
suppose we have 3 000 numerical values. All these data follow a certain distribution – behave in a specific manner. In order to depict this behaviour we may construct a

Histogram
the horizontal axis are the intervals (‘bins’) our values belong to. The range is (practically) from 2 to 9.5. And this range has been divided into 15 bins:

<table>
<thead>
<tr>
<th>Interval</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-2.5</td>
<td>5</td>
</tr>
<tr>
<td>2.5-3</td>
<td>15</td>
</tr>
<tr>
<td>3-3.5</td>
<td>60</td>
</tr>
<tr>
<td>3.5-4</td>
<td>132</td>
</tr>
<tr>
<td>4-4.5</td>
<td>287</td>
</tr>
<tr>
<td>4.5-5</td>
<td>455</td>
</tr>
<tr>
<td>5-5.5</td>
<td>583</td>
</tr>
<tr>
<td>5.5-6</td>
<td>545</td>
</tr>
<tr>
<td>6-6.5</td>
<td>435</td>
</tr>
<tr>
<td>6.5-7</td>
<td>275</td>
</tr>
<tr>
<td>7-7.5</td>
<td>141</td>
</tr>
<tr>
<td>7.5-8</td>
<td>45</td>
</tr>
<tr>
<td>8-8.5</td>
<td>20</td>
</tr>
<tr>
<td>8.5-9</td>
<td>2</td>
</tr>
<tr>
<td>9-9.5</td>
<td>2</td>
</tr>
</tbody>
</table>

The second row shows how many values belong to a given interval: e.g. the third entry 60 shows that sixty values are greater than 3.0 and equal to or less than 3.5. In the given bin we have thus 60 out of 3000.

The **probability that our RV** $X$ **has the values**: $3.0 < x \leq 3.5$ is $60/3000 = 0.002$. The height of the vertical bar is the measure of this probability.

But for practical reasons we have to depict distributions with numbers rather than graphs. ☺️
A FUNCTION OF A RANDOM VARIABLE: 

\[ Y = H(X) \ldots \]

is also a random variable — so it also has — \( F(y) \) a cumulative distribution with some \textit{PARAMETERS} (which may be known from an experiment)

\[ \text{MATHEMATICAL EXPECTATION or the MEAN VALUE OF A RANDOM VARIABLE} \]

\[ E(X) = \hat{x} = \begin{cases} \sum_{k=0}^{n} x_k \mathcal{P}(X = x_k) = \sum_{k=0}^{n} p_k x_k & \text{for a discrete RV} \\ \int_{-\infty}^{\infty} x f(x) \, dx & \text{for a continuous RV} \end{cases} \]
...every mathematical expectation is a number, so $E(X)$ is no longer something which may be called ‘random’.

We use various conventions of notation: $E(X)$, $\hat{x}$, $\mu$ (the ‘true’ mean value for the given RV $X$) and $m$ (the estimated mean value for the given $X$).

For a physicist (well, not only) $E(X)$ may be perceived as a “centre-of-mass” of the $X$, or ...

**weighted mean:** $E(X) = \sum w_i x_i / \sum w_i$.

The weights are:

$p_i$’s for discrete RV — $E(X) = \sum p_i x_i$

and $f(x) \, dx = \mathcal{P}(X \in [x, x + dx])$ for continuous RV

In the second case the sum is of course replaced by an integral.

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) \, dx$$
Let our function of the random variable \(V\) be:

\[
H(X) = (X - c)^l
\]

(\(c - \text{any number}\)); its mathematical expectation

\[
E\{(X - c)^l\}
\]

is called the \(l\)-th moment of the random variable \(X\) with respect to \(c\).

\[
\alpha_l \overset{\text{def}}{=} E\{(X - c)^l\}
\]

It is logical to put \(c = E(X)(\hat{x})\) — in this manner we obtain the so-called CENTRAL MOMENTS:

\[
\mu_l = E\{(X - \hat{x})^l\}
\]
Let's consider the case of a continuous variable:

\[
\mu_0 = \int_{-\infty}^{\infty} (x - \hat{x})^0 f(x) \, dx = 1
\]

\[
\mu_1 = \int_{-\infty}^{\infty} (x - \hat{x})^1 f(x) \, dx = 0
\]

\[
\mu_2 = \int_{-\infty}^{\infty} (x - \hat{x})^2 f(x) \, dx \overset{\text{def}}{=} \text{VAR}(X) = \sigma^2(X) = \text{VARIANCE}
\]

\[
\mu_3 = \int_{-\infty}^{\infty} (x - \hat{x})^3 f(x) \, dx = \text{SKEWNESS}
\]

\[
\mu_4 = \int_{-\infty}^{\infty} (x - \hat{x})^4 f(x) \, dx = \text{KURTOSIS}
\]
what is the meaning of those moments?

- **VARIANCE** — a measure of the spread (dispersion) (always > 0)
- **SKEWNESS** — a measure of asymmetry
- **KURTOSIS** — a measure of the spread as compared with a special type of distribution — normal distribution

\[ \sigma = \sqrt{VAR(X)} = \sigma(X) = \sigma_x \]

— STANDARD MEAN DEVIATION OF A RANDOM VARIABLE X — N.B. it is expressed in the same UNITS as X!

\[ \sigma(X) \ldots \]

\ldots may be regarded as a *natural unit* for measuring our Random Variable.
a short-cut formula for calculating the variance:

\[
[X - E(X)]^2 = X^2 - 2E(X)X + [E(X)]^2
\]

but we have \((X - a \text{ R.V.}; a, b - \text{constants})\)

\[
E(aX + b) = aE(X) + b.
\]

proof (for a discrete-type R.V)

\[
\sum_i p_i x_i = \sum_i p_i (ax_i + b) = a \sum_i p_i x_i + b \sum_i p_i = aE(X) + b.
\]

(Repeat this proof for the case of a continuous RV.)
a short-cut formula for calculating the variance:

\[
[X - E(X)]^2 = X^2 - 2E(X)X + [E(X)]^2
\]

but we have ($X$ – a R.V.; $a, b$ – constants)

\[
E(aX + b) = aE(X) + b.
\]

proof (for a discrete-type R.V)

\[
\sum_i p_i x_i = \sum_i p_i(ax_i + b) = a \sum_i p_i x_i + b \sum_i p_i = aE(X) + b.
\]

(Repeat this proof for the case of a continuous RV.)

Applying the $E$ operator to the right member of the equation $\star$

\[
E(X^2) - 2E(X)E(X) + E\{[E(X)]^2\} = E(X^2) - [E(X)]^2.
\]
One may prefer the so-called standardised parameters

\[ \gamma_3 = \frac{\mu_3}{\sigma^3} \quad (= \gamma) \]

\[ \gamma_4 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_4}{\mu_2^2} - 3 \]

\[ \gamma_3 > 0 \rightarrow E(X) - Mo > 0 \]

\[ \gamma_4 > 0 \rightarrow \text{the distribution is ,,slimmer” than the Normal distribution} \]
standardised RANDOM VARIABLE

\[ Z = \frac{X - \hat{x}}{\sigma} \]

\( \hat{x} \) — is a ”natural” zero (origin)
\( \sigma \) — is a ”natural” unit

Let \( X \) be a RV with \( E(X) = \hat{x} \) and \( VAR(X) = \sigma^2 \). Then, for \( d \) being a number:

\[ \mathcal{P}(|X - \hat{x}| \geq d) \leq \frac{\sigma^2}{d^2}, \text{ or} \]

putting: \( d = k \cdot \sigma \) we get

\[ \mathcal{P}(|X - \hat{x}| \geq k \cdot \sigma) \leq \frac{1}{k^2}. \]

This is CHEBYSHEV INEQUALITY – a rather crude estimate of the dispersion of our \( X \) around \( E(X) \).
DESCRIPTIVE STATISTICS

- **quantile:**
  A **quantile** $q(f)$ or $x_f$, is a value of $x$ for which a specified fraction, $f$, of the $X$ values is less than or equal to $x_f$:

  $F(x_f) = P(X \leq x_f) \geq f$

  $1 - F(x_f) = P(X > x_f) \leq 1 - f$

  (for a continuous RV we have the " $\geq$" or " $\leq$" sign)

  **QUANTILE** for $f = 0.5$ (50%) is called **median**; for $f = 0.25$ (25%) we have the first (lower) quartile, and for $f = 0.75$ (75%) we have the fourth (upper) quartile

- **MODE** (modal value — $Mo(X)$)
  is a value $x$, for which: $\frac{df}{dx} = 0$ and $\frac{d^2f}{dx^2} < 0$
  — (local maximum of the probability density function)

- **RANGE:** $x_{max} - x_{min}$
Box-and-whiskers plot

- maximum value
- upper quartile
- expected value
- median
- lower quartile
- minimum value

after Jacek Tarasiuk "Wyklady ze statystyki inżynierskiej"
WFiIS 2013
DESCRIPTIVE STATISTICS

RANDOM VARIABLE and its CHARACTERISTICS