REVIEW OF STATISTICAL CONCEPTS
**Problem**: Differences in genome sizes among vertebrates. Overall, the chicken genome is $\sim 40\%$ of the size of the human genome. Does the size ratio vary across orthologous regions? Can this be explained by insertion of repeats? Source: Int’l Chicken Genome Seq. Consortium, *Nature* 12/04.

**Population**: 6727 non-overlapping windows (length btw 100 and 150Kb), which cover all alignments of chicken to human genome.

**Variables considered**:  
$Y = \log(hL/cL)$: ratio between human and chicken length, including repeat bases.  
$X = \log(hMS/cMS)$: ratio between the fraction of each window occupied by repeat bases in human and in chicken (proxi for insertion ratio).  
$y$ and $x$ are **quantitative** variables with **continuous** values.

**Sample**: $n=100$ randomly selected windows, and correspondingly $(y_i, x_i), i=1\ldots n$

<table>
<thead>
<tr>
<th>$i$</th>
<th>$y$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.31387</td>
<td>0.48023</td>
</tr>
<tr>
<td>2</td>
<td>0.52634</td>
<td>1.65562</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>0.35218</td>
<td>1.45637</td>
</tr>
<tr>
<td>100</td>
<td>0.22531</td>
<td>0.44261</td>
</tr>
</tbody>
</table>
Center of a quantitative variable (e.g. log of the length ratio) on the sample data.

Mean or arithmetic average:
\[
\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 0.384
\]

A robust alternative evaluation of the center is the Median:

value splitting the data in half:
- \([(n+1)/2]\)th sorted data value, if \(n\) odd
- mid-point btw \((n/2)\)th and \([(n/2)+1]\)th sorted data values, if \(n\) even;

\[
0.358885
\]

Median is lower than mean (in fact there is a slight excess of high values – the data is slightly asymmetric about its center, see graphs below).

However we measure the center though, it seems quite clear that it is > 0 (on log scale; i.e. 1 for length ratios). Based on the sample data, human length exceeds orthologous chicken length on average...
VARIABILITY of a quantitative variable on the sample data.

About the mean: variance and standard deviation (same scale as variable):

\[
s^2(y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = 0.039
\]

\[
sd(y) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} = 0.196
\]

averaging square deviations from the mean; can use n or (n-1) at the denominator (n-1 = degrees of freedom of the sum of squares).

Min-max Range:

\[
\begin{align*}
\text{min} & \quad \text{max} \\
y & \quad -0.07572 \quad 1.07482
\end{align*}
\]

Inter-quartile range:

\[
\begin{align*}
\text{q25} & \quad \text{q75} \\
y & \quad 0.2699500 \quad 0.481065
\end{align*}
\]

Based on the sample data, ratio btw human and chicken length is highly variable across orthologous loci...
Useful summary **statistics** = functions computed on the sample data.

A statistic such as the sample mean or sd can be meant as **descriptor** of the data, but also as an **estimate** of the corresponding feature of the population (**parameter**) from which the sample was drawn.
\[ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \]

**Sample mean** as a **(point) estimate** of the population mean \( \mu_y \).

Its value will change depending on the sample. On average over all possible samples of size \( n \):

\[ E(\bar{y}) = \mu_y \quad \text{(unbiasedness)} \]

\[ MSE(\bar{y}) = \frac{1}{n} \sigma_y^2 \]

Thus we evaluate the **standard error** of the sample mean in estimating the population mean as:

\[ se(\bar{y}) = \frac{1}{\sqrt{n}} sd(y) \]
Inference: CONFIDENCE INTERVAL (CI) for a population parameter.

Use the se to create a CI, with the statistic as “pivot”. For the population mean:

\[ CI(\alpha) = \bar{y} \pm a(\alpha) \cdot se(\bar{y}) \]

multiplier guarantees or approximates a certain coverage or confidence level, (probability that the random interval, over all possible \(n\)-samples contains the population mean):

\[ \Pr\{CI(\alpha) \ni \mu_y\} = 1 - 2\alpha \]

selected using a reference distribution. If the sample size is large, asymptotic arguments (Central Limit Theorem) tell us that:

\[ \frac{\bar{y} - \mu_y}{se(\bar{y})} \sim N(0,1) \text{ \, \, (approx)} \]

roughly this means \( \bar{y} \sim N(\mu_y, se(\bar{y})) \text{ \, \, (approx)} \)

We can use multipliers from a N(0,1) (quantiles).
If the sample size is small, but we can assume the population $y$ to be normal, we have that:

$$\frac{\bar{y} - \mu}{se(\bar{y})} \sim T_{n-1}$$

and we can use Student’s $T$ multipliers instead.

This works approximately also when the population $y$ does not depart too extremely from a Gaussian shape.

A $T$-distribution is symmetric about 0 and bell shaped as a $N(0,1)$, but has heavier tails, so that the multiplier is larger, and the CI broader, for the same coverage.

When the dofs increase, it converges to a $N(0,1)$.

The approximation is already good for $n > 20$. So in practice standard normal and $T$ multipliers are indistinguishable for large samples.
CI with coverage 95% for the population mean of the log-length ratio.
Recall $n = 100$, sample mean $= 0.384$, sample sd $= 0.196$.

- 2.5% quantile of the N(0,1) distribution: $-1.9599$
- 2.5% quantile of the Student’s T-distrib with $n-1 = 99$ dof: $-1.9842$

can get these from tables or statistics software packages.

note quantiles are very close; $n$ large.

$$CI(\alpha) = \bar{y} \pm a(\alpha) \, se(\bar{y}) = 0.384 \pm 1.96 \frac{0.196}{\sqrt{100}} = [0.346, 0.423]$$
Inference: TEST OF HYPOTHESIS for a population parameter.

**Null** hypothesis on a feature of the population, for instance

\[ H_0 : \mu_y = 0 \]

*This is what we would like to refute.* Using our sample data, we investigate the null in comparison to an **alternative**, for instance

\[ H_a : \mu_y \neq 0 \] (two-sided)

or

\[ H_a : \mu_y > 0 \] (one-sided, right – could be left)

*This is what we would like to show is supported by evidence in the data.* Note in this case the null specifies one value, while the alternative specifies a range (these are the most common specifications).

Right-sided: assess if we have enough evidence to conclude, based on our sample of \( n=100 \) observations, that the log length ratio for human vs chicken has a positive mean in the population.
We need to use a **test statistic**, i.e. a function of the data, whose distribution under Ho (**null distribution**) is known and can thus be used as reference. We know that:

\[
\frac{\bar{y} - \mu}{se(\bar{y})} \sim N(0,1) \quad \text{approx} \quad \text{if } n \text{ is large, regardless of the distribution of } y \text{ in the population.}
\]

\[
\sim T_{n-1} \quad \text{if } y \text{ in the population is approximately normal, for any } n.
\]

Thus under Ho (say we use the first result, \(n=100\)):

\[
u = \frac{\bar{y}}{se(\bar{y})} \sim N(0,1)
\]

\[
p(u_{obs}) = Pr(u \geq u_{obs} | \mu_y = 0)
\]

the **p-value**, or achieved significance level, associated with the observed \(u\) is the probability that, **under the null**, the statistic would take the observed value, or a value even more extreme in the direction (here, right) defined by the alternative.
For the two-sided alternative, the p-value is:

\[ p(u_{obs}) = \Pr(u \leq -|u_{obs}| \text{ or } u \geq |u_{obs}| \mid \mu_y = 0) = 2 \Pr(u \geq |u_{obs}| \mid \mu_y = 0) \]

because of symmetry of the null distribution.

**Basic idea**: we can reject Ho in favour of Ha if the observed value of the test statistic is very extreme with respect to what one would expect under the null distribution; that is, if the corresponding p-value is small. The smaller the p-value, the stronger the evidence against Ho provided by the data.

Testing whether the population mean of the log-length ratio is 0 vs positive. Recall \( n = 100 \), sample mean = 0.384, sample sd = 0.196.

\[ u = \frac{\bar{y}}{se(\bar{y})} = \frac{0.384}{0.196/\sqrt{100}} = 19.575 \]

The area on the right of 19.575 under a standard normal is

\[ p(u_{obs}) = 1.260955e-85 \]

0 by all practical means! Very strong evidence that the population mean log length ratio is positive.
**Rejection rule**: reject Ho if the p-value is ≤ a threshold a, say \( \alpha = 0.05 \) (5%).

This is called the **level**, or (target) significance. With this rule, we ensure that

\[
\text{Pr}(\text{rejecting Ho} \mid \text{Ho}) \leq \alpha
\]

i.e. we control the probability of a **false positive**, or so called **type-I error**.

The other error we can make is to fail to reject Ho when Ha holds:

\[
\text{Pr}(\text{not rejecting Ho} \mid \text{Ha})
\]

This is the probability of a **false negative**, or so called **type-II error**.

1- such probability is called the **power** of the test, and the function expressing it for each point in the alternative (in our instance Ha is a range) is called the power function.

Type-I and II error probabilities are in trade-off; test statistics are evaluated based on their power function, once the level is fixed.
GRAPHICAL REPRESENTATIONS of a quantitative variable on the sample data.

These plots show that for almost all windows in the sample the log length ratio is pos (human length larger than chicken), and that it varies substantially across windows – with a slight excess of high values.
ASSOCIATION btw two quantitative variables on the sample data.

Measure of (linear) association, Pearson's correlation coefficient:

$$
\text{cor}(y, x) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}} = 0.396 \in [-1, 1]
$$

+/- 1: maximal direct or inverse linear association (y and x values lie on a line). 0: lack of linear association, which does NOT necessarily mean lack of association in general!

On our data, length ratio and insertion ratio (both on the log scale) present a sizeable positive correlation.

Statistical, rather than exact functional association (unless cor=+/-1!).

Also cor can be meant as descriptor of the sample data, but also as an estimate of the corresponding population feature (parameter).
GRAPHICAL REPRESENTATION of bi-variate sample data.

Scatter plot

Showing a positive statistical association btw length ratio and insertion ratio (both on log scale)
This course is about **REGRESSION ANALYSIS**:

- Constructing quantitative descriptions of the statistical association between $y$ (response variable) and $x$ (predictor, or explanatory variable) on the sample data.
- Introducing models, to interpret estimates and inferences on the parameters of these descriptions in relation to the underlying population.

MULTIPLE regression, when we consider more than one predictor variable.