COURSE NOTES: INFERENTIAL STATISTICS
**Definition**

In statistics, when we talk about distributions we usually mean probability distributions.

Definition (informal): A distribution is a function that shows the possible values for a variable and how often they occur.

Definition (Wikipedia): In probability theory and statistics, a probability distribution is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment.

**Examples:** Normal distribution, Student’s T distribution, Poisson distribution, Uniform distribution, Binomial distribution

---

**Graphical representation**

It is a common mistake to believe that the distribution is the graph. In fact the distribution is the ‘rule’ that determines how values are positioned in relation to each other.

Very often, we use a graph to visualize the data. Since different distributions have a particular graphical representation, statisticians like to plot them.

**Examples:**
- Uniform distribution
- Binomial distribution
- Normal distribution
- Student’s T distribution
The Normal distribution is also known as Gaussian distribution or the Bell curve. It is one of the most common distributions due to the following reasons:

- It approximates a wide variety of random variables
- Distributions of sample means with large enough samples sizes could be approximated to normal
- All computable statistics are elegant
- Heavily used in regression analysis
- Good track record

Examples:

- Biology. Most biological measures are normally distributed, such as: height; length of arms, legs, nails; blood pressure; thickness of tree barks, etc.
- IQ tests
- Stock market information

\[ N \sim (\mu, \sigma^2) \]

\( N \) stands for normal; 
\( \sim \) stands for a distribution; 
\( \mu \) is the mean; 
\( \sigma^2 \) is the variance.
Controlling for the standard deviation

Keeping the standard deviation constant, the graph of a normal distribution with:

- a smaller mean would look in the same way, but be situated to the left (in gray)
- a larger mean would look in the same way, but be situated to the right (in red)
The Normal Distribution

Controlling for the mean

Keeping the mean constant, a normal distribution with:
  • a smaller standard deviation would be situated in the same spot, but have a higher peak and thinner tails (in red)
  • a larger standard deviation would be situated in the same spot, but have a lower peak and fatter tails (in gray)
The Standard Normal distribution is a particular case of the Normal distribution. It has a mean of 0 and a standard deviation of 1.

Every Normal distribution can be 'standardized' using the standardization formula:

\[ z = \frac{x - \mu}{\sigma} \]

A variable following the Standard Normal distribution is denoted with the letter z.

Why standardize?
Standardization allows us to:
• compare different normally distributed datasets
• detect normality
• detect outliers
• create confidence intervals
• test hypotheses
• perform regression analysis

Rationale of the formula for standardization:
We want to transform a random variable from \( N \sim (\mu, \sigma^2) \) to \( N \sim (0,1) \). Subtracting the mean from all observations would cause a transformation from \( N \sim (\mu, \sigma^2) \) to \( N \sim (0,\sigma^2) \), moving the graph to the origin. Subsequently, dividing all observations by the standard deviation would cause a transformation from \( N \sim (0,\sigma^2) \) to \( N \sim (0,1) \), standardizing the peak and the tails of the graph.
The Central Limit Theorem (CLT) is one of the greatest statistical insights. It states that no matter the underlying distribution of the dataset, the sampling distribution of the means would approximate a normal distribution. Moreover, the mean of the sampling distribution would be equal to the mean of the original distribution and the variance would be \( n \) times smaller, where \( n \) is the size of the samples. The CLT applies whenever we have a sum or an average of many variables (e.g. sum of rolled numbers when rolling dice).

<table>
<thead>
<tr>
<th>The theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>➢ No matter the distribution</td>
</tr>
<tr>
<td>➢ The distribution of ( \bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \ldots, \bar{x}_k ) would tend to ( N(\mu, \sigma^2/n) )</td>
</tr>
<tr>
<td>➢ The more samples, the closer to Normal ( ( k \to \infty ) )</td>
</tr>
<tr>
<td>➢ The bigger the samples, the closer to Normal ( ( n \to \infty ) )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Why is it useful?</th>
</tr>
</thead>
<tbody>
<tr>
<td>The CLT allows us to assume normality for many different variables. That is very useful for confidence intervals, hypothesis testing, and regression analysis. In fact, the Normal distribution is so predominantly observed around us due to the fact that following the CLT, many variables converge to Normal.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Where can we see it?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Since many concepts and events are a sum or an average of different effects, CLT applies and we observe normality all the time. For example, in regression analysis, the dependent variable is explained through the sum of error terms.</td>
</tr>
</tbody>
</table>

Click here for a CLT simulator.
Estimators and Estimates

Broadly, an estimator is a mathematical function that approximates a population parameter depending only on sample information.

Examples of estimators and the corresponding parameters:

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimator</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\bar{x}$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Variance</td>
<td>$s^2$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>Correlation</td>
<td>$r$</td>
<td>$\rho$</td>
</tr>
</tbody>
</table>

Estimators have two important properties:

- **Bias**
  The expected value of an unbiased estimator is the population parameter. The bias in this case is 0. If the expected value of an estimator is (parameter + b), then the bias is b.
- **Efficiency**
  The most efficient estimator is the one with the smallest variance.

An estimate is the output that you get from the estimator (when you apply the formula). There are two types of estimates: point estimates and confidence interval estimates.

**Point estimates**
A single value.
Examples:
- 1
- 5
- 122.67
- 0.32

**Confidence intervals**
An interval.
Examples:
- (1, 5)
- (12, 33)
- (221.78, 745.66)
- (-0.71, 0.11)

Confidence intervals are much more precise than point estimates. That is why they are preferred when making inferences.
Definition: A confidence interval is an interval within which we are confident (with a certain percentage of confidence) the population parameter will fall.

We build the confidence interval around the point estimate.

(1-\(\alpha\)) is the level of confidence. We are \((1-\alpha)\times100\%\) confident that the population parameter will fall in the specified interval. Common alphas are: 0.01, 0.05, 0.1.

General formula:

\[
\left[ \bar{x} - ME, \bar{x} + ME \right] , \text{ where } ME \text{ is the margin of error.}
\]

\[
ME = \text{reliability factor} \times \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}
\]

<table>
<thead>
<tr>
<th>Term</th>
<th>Effect on width of CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>((1-\alpha))</td>
<td>↑</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>↑</td>
</tr>
<tr>
<td>(n)</td>
<td>↓</td>
</tr>
</tbody>
</table>
The Student’s T distribution is used predominantly for creating confidence intervals and testing hypotheses with normally distributed populations when the sample sizes are small. It is particularly useful when we don’t have enough information or it is too costly to obtain it.

A random variable following the t-distribution is denoted $t_{\nu,\alpha}$, where $\nu$ are the degrees of freedom.

We can obtain the student’s T distribution for a variable with a Normally distributed population using the formula:

$$t_{\nu,\alpha} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

All else equal, the Student’s T distribution has fatter tails than the Normal distribution and a lower peak. This is to reflect the higher level of uncertainty, caused by the small sample size.
## Formulas for Confidence Intervals

<table>
<thead>
<tr>
<th># populations</th>
<th>Population variance</th>
<th>Samples</th>
<th>Statistic</th>
<th>Variance</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>known</td>
<td>-</td>
<td>z</td>
<td>$\sigma^2$</td>
<td>$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$</td>
</tr>
<tr>
<td>One</td>
<td>unknown</td>
<td>-</td>
<td>t</td>
<td>$s^2$</td>
<td>$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$</td>
</tr>
<tr>
<td>Two</td>
<td>-</td>
<td>dependent</td>
<td>t</td>
<td>$s^2_{\text{difference}}$</td>
<td>$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$</td>
</tr>
<tr>
<td>Two</td>
<td>known</td>
<td>independent</td>
<td>z</td>
<td>$\sigma_x^2, \sigma_y^2$</td>
<td>$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$</td>
</tr>
<tr>
<td>Two</td>
<td>unknown, assumed equal</td>
<td>independent</td>
<td>t</td>
<td>$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$</td>
<td>$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$</td>
</tr>
<tr>
<td>Two</td>
<td>unknown, assumed different</td>
<td>independent</td>
<td>t</td>
<td>$s_x^2, s_y^2$</td>
<td>$(\bar{x} - \bar{y}) \pm t_{\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$</td>
</tr>
</tbody>
</table>