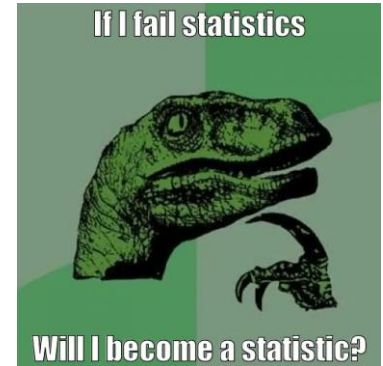


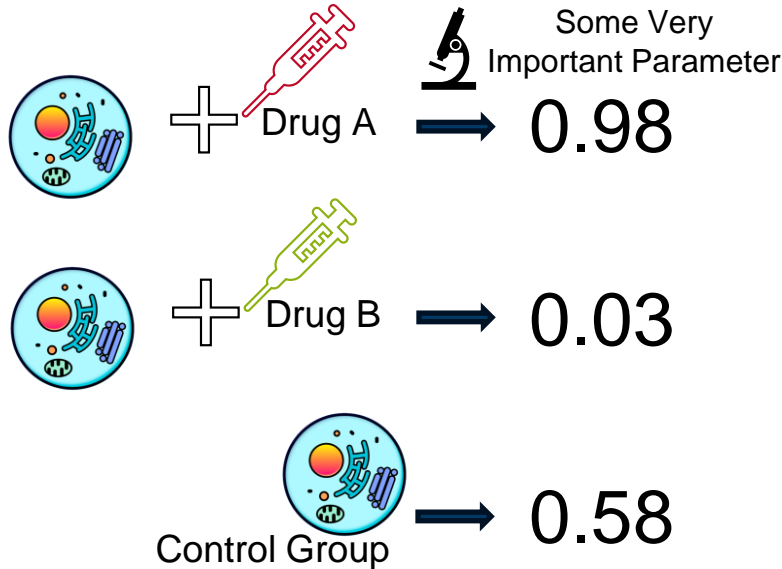
06.01.2025

FUN with ~~Flags~~ STATS

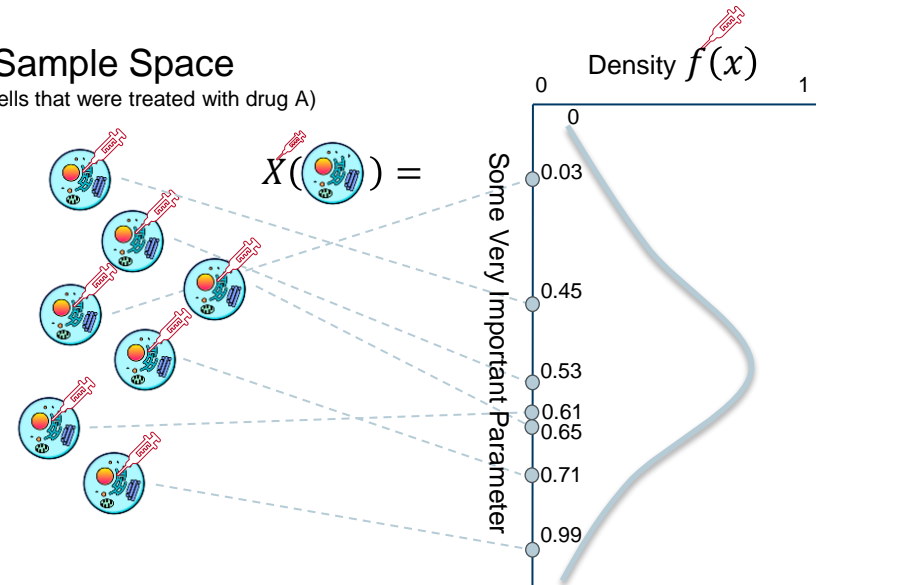
Jan van Grimbergen, Jonathan Bobak



Random Variables and Density Functions



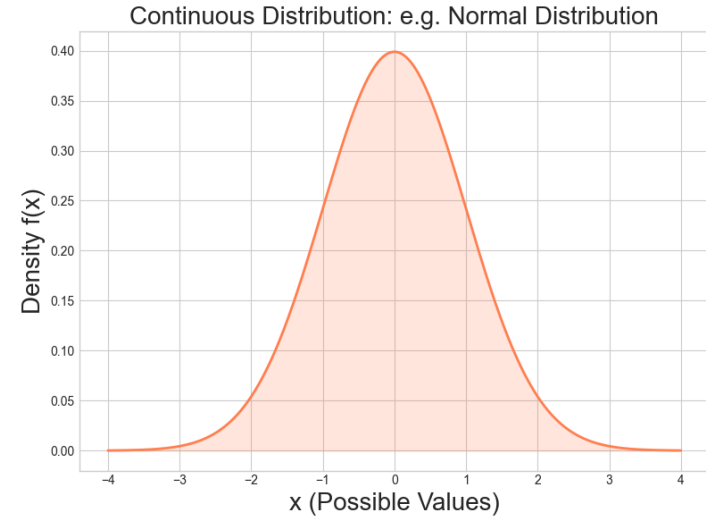
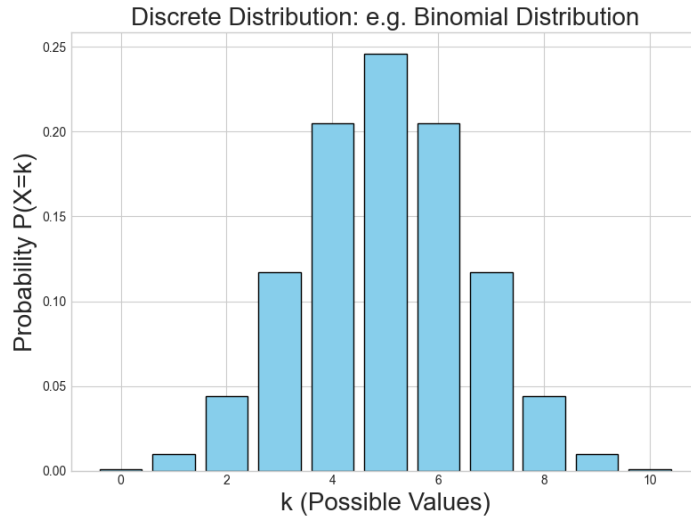
Sample Space
(e.g. all cells that were treated with drug A)



\rightarrow Our experiment can be described with three random variables X, Y and Z

$$\rightarrow P(a < X < b) = \int_a^b f(x) dx$$

- Discrete Distributions
- Continuous Distributions



Discrete

$$\mu = E[X] = \sum_i x_i p_i$$
$$\sigma^2 = Var[X] = \sum_i (x_i - E[X])^2 p_i$$

Continuous

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$
$$\sigma^2 = Var[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$$

Approx Expectation Value and Variance From Measurements

Random variable X with measurements $y_1, y_2, y_3, \dots, y_m$

$$E[X] \approx \hat{\mu} = \frac{1}{m} \sum_{i=1}^m y_i$$
$$Var[X] \approx \hat{\sigma}^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \hat{\mu})^2$$

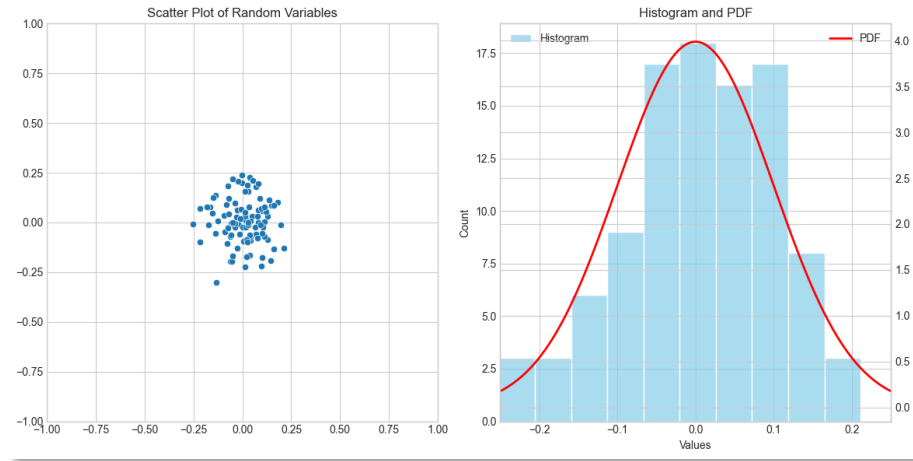
Law of Large Numbers

$$\hat{\mu} \xrightarrow{m \rightarrow \infty} \mu$$




$$\hat{\sigma} \xrightarrow{m \rightarrow \infty} \sigma$$





Short Trip Back To Python



Some Very Important Parameter



	0.75	0.38	...	0.63
	0.15	0.89	...	0.58
	0.3	0.90	...	0.87

Does  or  have a significant effect on  ?

- Want to test if the expected value of cells treated with  or  differ significantly from the control group.
- Question: Is the measured difference bigger than you would expect under randomness?
- Assume that X , Y and Z are normally distributed.

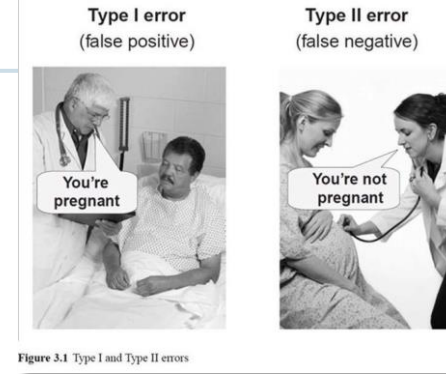
A common assumption with real-world data that comes from the [Central Limit Theorem](#)

Testing - Errors

Does  have a significant effect on  ?
Can we reject the Null Hypothesis ?

Null Hypothesis: $\mu_Y = \mu_Z$

Alternative hypothesis: $\mu_Y \neq \mu_Z$



	Null Hypothesis: True	Alternative Hypothesis: True
Can not reject Null Hypothesis	Correct Decision	Type 2 Error
Have to reject Null Hypothesis	Type 1 Error	Correct Decision

- ➡ The Type 1 Error is controlled to occur with a probability of at most $100 * \alpha$ (α is called the **significance level** of the test)
- ➡ We reject the Null Hypothesis if the probability p to observe the measured differences or a more extreme one in μ_Y and μ_Z is less than α . p is called the **p-value**. It is common to use $\alpha = 0.05$.
- ➡ How to calculate the **p-value**?

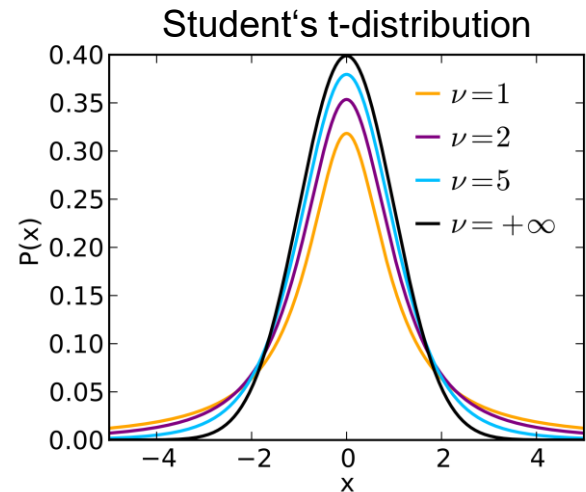
Testing - Welch's t-Test

➡ How to calculate the p-value?

- The test statistic transforms the observed data into a space that follows a known distribution.
- This allows us to calculate the probability of obtaining the observed value or a more extreme one under the null hypothesis (p-value 🎲)

$$t = \frac{\widehat{\mu}_Y - \widehat{\mu}_Z}{\sqrt{\widehat{\sigma}_Y^2 + \widehat{\sigma}_Z^2}} \quad v \approx \frac{\left(\frac{\widehat{\sigma}_Y^2}{m_Y} + \frac{\widehat{\sigma}_Z^2}{m_Z}\right)^2}{\frac{\left(\frac{\widehat{\sigma}_Y^2}{m_Y}\right)^2}{m_Y - 1} + \frac{\left(\frac{\widehat{\sigma}_Z^2}{m_Z}\right)^2}{m_Z - 1}}$$

This test statistic t follows a student's t-distribution with v degrees of freedom



➡ Let us try this in python! 🐍

■ The correct test depends on your data and your knowledge

Test Name	Parametric/Non-Parametric	Distributional Assumptions	Data Level/Requirements	Application
Student's t-test (independent samples)	Parametric	Data in both groups are normally distributed, equal variances	Continuous data, homogeneous variances	Comparing the means of two independent groups
Welch's t-test (independent samples)	Parametric	Data in both groups are normally distributed, unequal variances allowed	Continuous data, no homogeneity of variance required	Comparing the means of two independent groups when variances differ
Mann-Whitney U-test (Wilcoxon rank-sum)	Non-Parametric	No normality assumption, similar shape of distributions recommended	Ordinal or continuous data not necessarily normally distributed	Comparing central tendencies of two independent groups
Wilcoxon signed-rank test (paired)	Non-Parametric	No normality assumption, data are paired	Ordinal or continuous data, paired measurements	Comparing two measurements from the same individuals (e.g., before and after)

Let us assume we do 100 independent tests with $\alpha = 0.05$ how big is the probability of observing at least one type 1 error ?

$$P = 1 - (1 - \alpha)^{100} = 1 - 0.95^{100} = 0.994 = 99\%$$

One approach: Benjamin Hochberg

Let assume those are the ordered p-Values of the tests

$$p_1 \leq p_2 \leq p_3 \leq \dots \leq p_{m-1} \leq p_m$$

Search last i such that $p_i \leq \frac{i}{m} \alpha$. Reject only hypothesis with $j \leq i$

BH Controls the FDR :

$$E[FDR] = \frac{m_0 \alpha}{m} \leq \alpha \text{ with } m_0 \text{ is the count of real true null hypothesis.}$$

Correlation $p_{X,Y}$ measures the **strength** and **direction** of the **linear relationship** between two random Variables X and Y .

$$p_{X,Y} = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \quad \left\{ \begin{array}{l} p_{X,Y} = 1 \Rightarrow \text{Perfect positive linear relationship} \\ p_{X,Y} = 0 \Rightarrow \text{No linear relationship} \\ p_{X,Y} = -1 \Rightarrow \text{Perfect negative linear relationship} \end{array} \right.$$

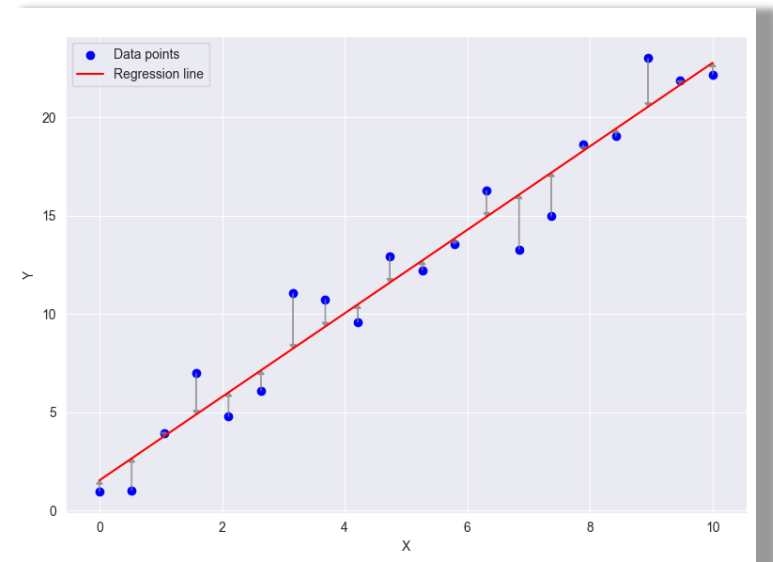
!Correlation does not prove that X causes Y !

Simple Linear Regression

A statistical method to explore linear relationships between a random variable Y and an other random variables X_1, X_2, \dots, X_m .

Find β_i such that: $Y = \beta_0 + \sum_{i=1}^m \beta_i X_i + \varepsilon$

- From measured data, a model is derived that minimizes the deviations (residuals) between predictions and observations.
- Minimization is achieved using the least squares method, which reduces the sum of squared residuals.



Testing the significance of β_i

(t-test for coefficients)

$$Y = \beta_0 + \sum_{i=1}^m \beta_i X_i + \varepsilon$$

Null Hypothesis: $\beta_i = 0$

Alternative Hypothesis: $\beta_i \neq 0$

$\hat{\beta}_i$ is the estimated value from the model.

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \quad SE(\hat{\beta}_i) = \sqrt{\sigma^2 (X^T X)^{-1}_{jj}} \quad \sigma^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - m - 1} \quad \hat{y}_i = \beta_0 + \sum_{i=1}^m \beta_i x_i$$

t follows a Students t-distribution with $n - m - 1$ degrees of freedom

- R^2 is the proportion of the variation in the dependent variable that is predictable from the independent variables

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$R^2 = 1$: Perfect Explanation By the Model

$R^2 = 0$: Model just uses the mean

