B4M36SAN BE4M36SAN Seminar/Tutorial

Alikhan Anuarbekov

Ali

anuarali@fel.cvut.cz

- For those who was absent last week, the course webpage:

  https://cw.fel.cvut.cz/wiki/courses/b4m36san

- If you want to have all R libraries ready for labs, download **r_setup.zip**

Materials
san_intro.zip, r_setup.zip, pres-1.pdf

Big team project (Final assignment section in courseware**)**

- Start looking for 4 people team **NOW**

- Try to find a question/dataset that is interesting for you

- Look for previous year best projects for inspiration

- Ask question before the deadline for team creation (3.11)

- If no team on 3.11 (midterm week), then I will forcefully assign a team for you

Tips and recommendations:

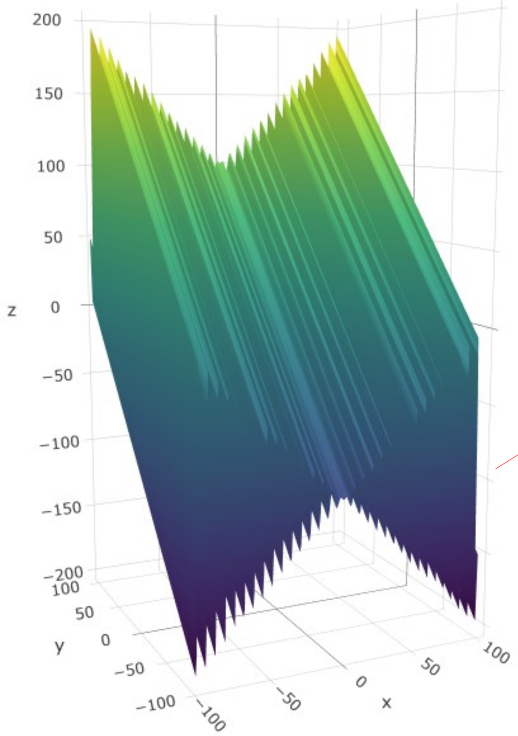
- Do not focus on single question, try to combine two or more questions
  - Example: Not only look for medical question, but try to add socio-economy

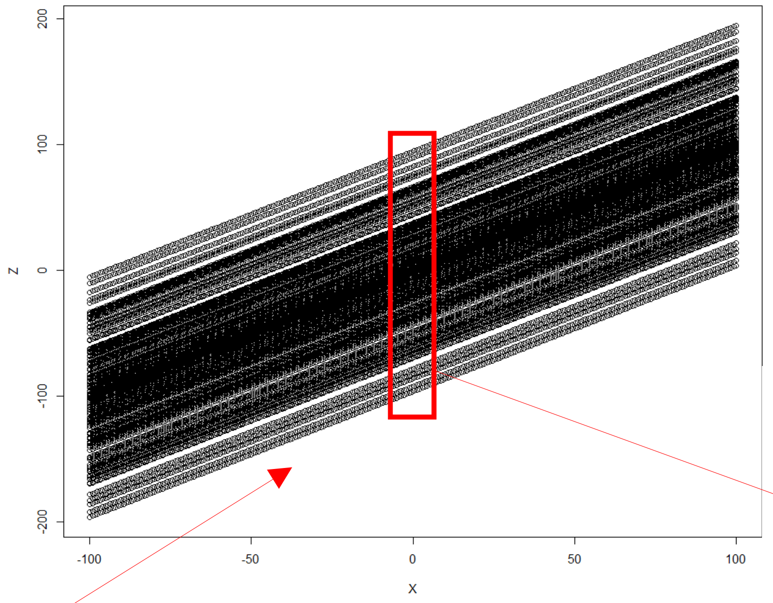
- Try to find more than one dataset
  - Example: UNICEF + World Bank

- Start working early, at least to formulate a question

From deterministic formula to randomness (Last week)

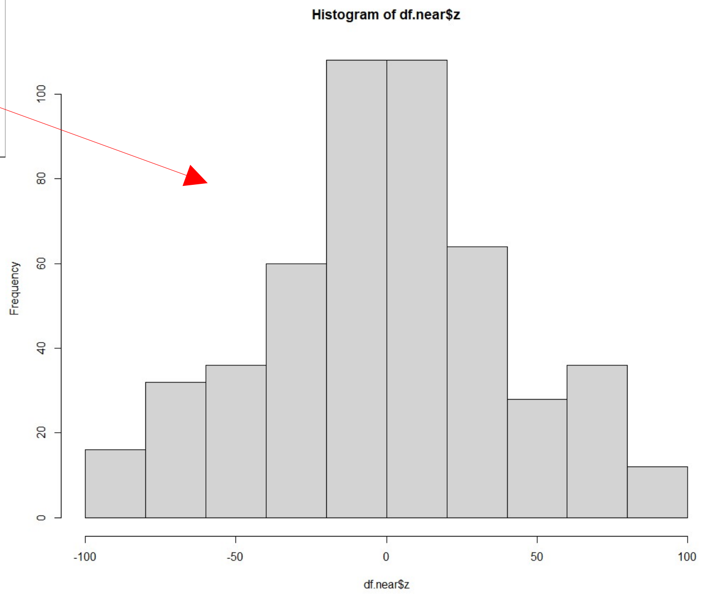
Z = X + Y * sin(Y)





Hide Y axis

Look at some X-Z range:



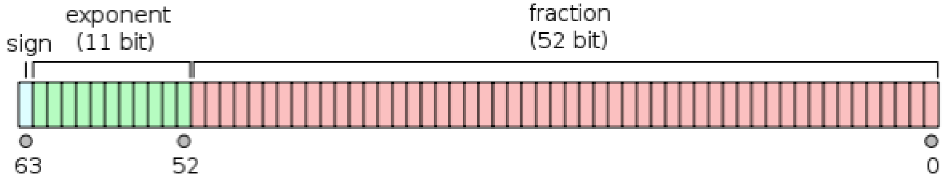
Sometimes deterministic formulas are infinitely complex

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \qquad \text{For small x,} \ \sin x \approx x$$

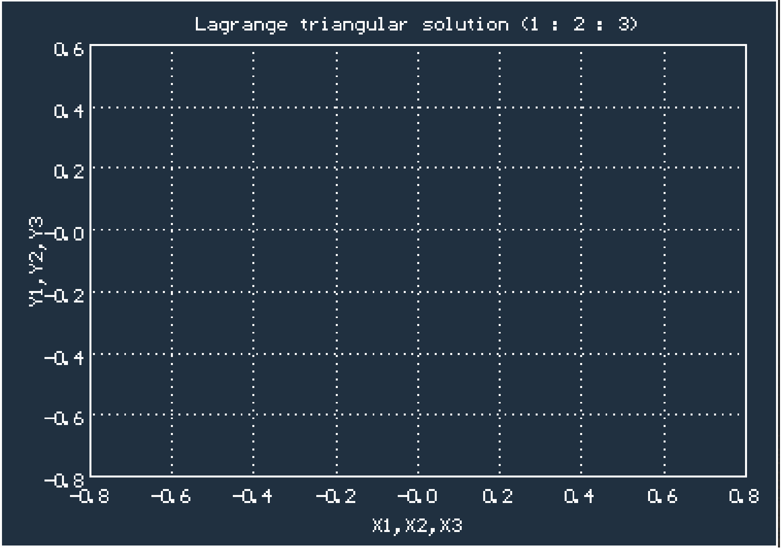
$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \qquad \text{For small x,} \ \cos x \approx 1$$

$$\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \dots \qquad \text{For small x,} \ \tan x \approx x$$

- You could compute every single digit
- But it has ∞ number of digits!
- If you cut it somewhere, you introduce an error!

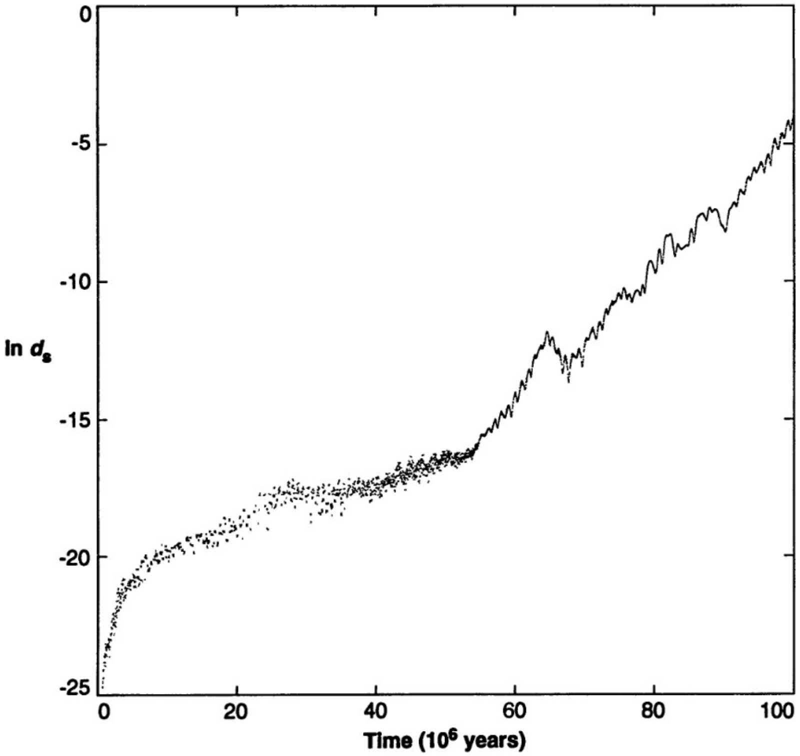


1/3 = 0011 1111 1101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 $+\Delta_{error}$



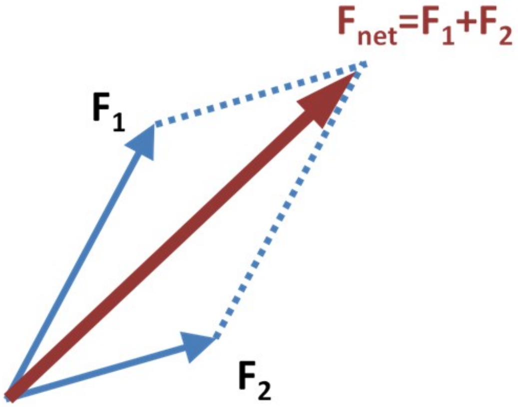
Practical case: Three body problem (Chaotic behavior)

$$
\begin{cases}
\ddot{\boldsymbol{r}}_1 = \displaystyle\sum_{j=1, j\neq i}^{3} m_j \frac{\boldsymbol{r}_j - \boldsymbol{r}_1}{|\boldsymbol{r}_j - \boldsymbol{r}_1|^3} \\[2em]
\ddot{\boldsymbol{r}}_2 = \displaystyle\sum_{j=1, j\neq 2}^{3} m_j \frac{\boldsymbol{r}_j - \boldsymbol{r}_2}{|\boldsymbol{r}_j - \boldsymbol{r}_2|^3} \\[2em]
\ddot{\boldsymbol{r}}_3 = \displaystyle\sum_{j=1, j\neq 3}^{3} m_j \frac{\boldsymbol{r}_j - \boldsymbol{r}_3}{|\boldsymbol{r}_j - \boldsymbol{r}_3|^3}
\end{cases}
$$



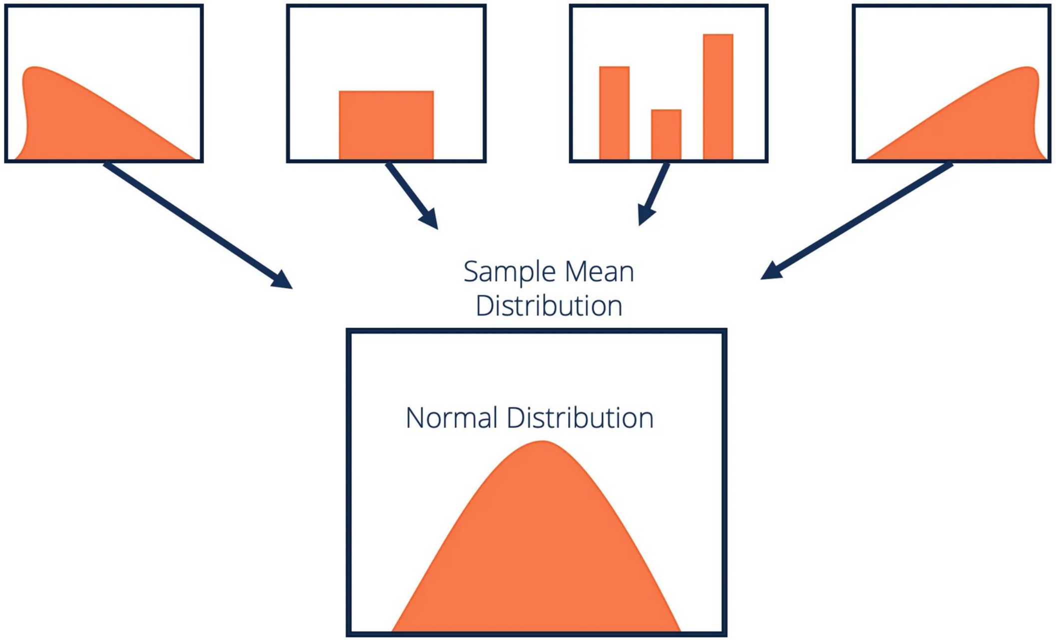
Why Gaussian normal distribution?

- Why we use a Gaussian normal? What if there is another distribution?
- Maybe, but typically when you have multiple factors/causes, they are typically summed



Example: Physics
sum of multiple forces

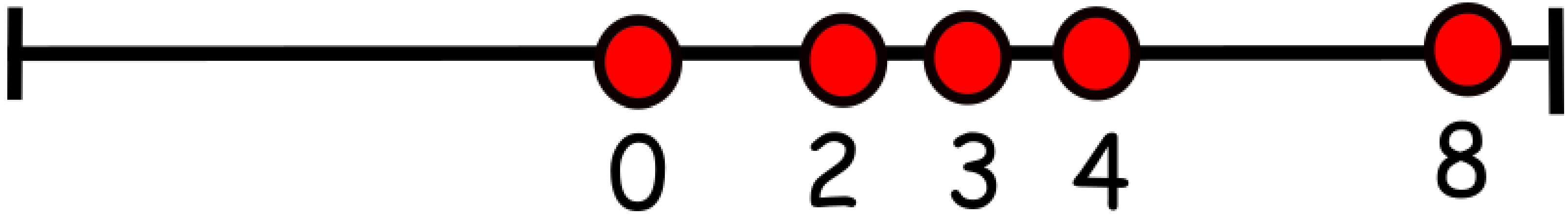
- Then, when we compute their mean (sum / number),
  it is mathematically Gaussian in a limit!
  (Central limit theorem)



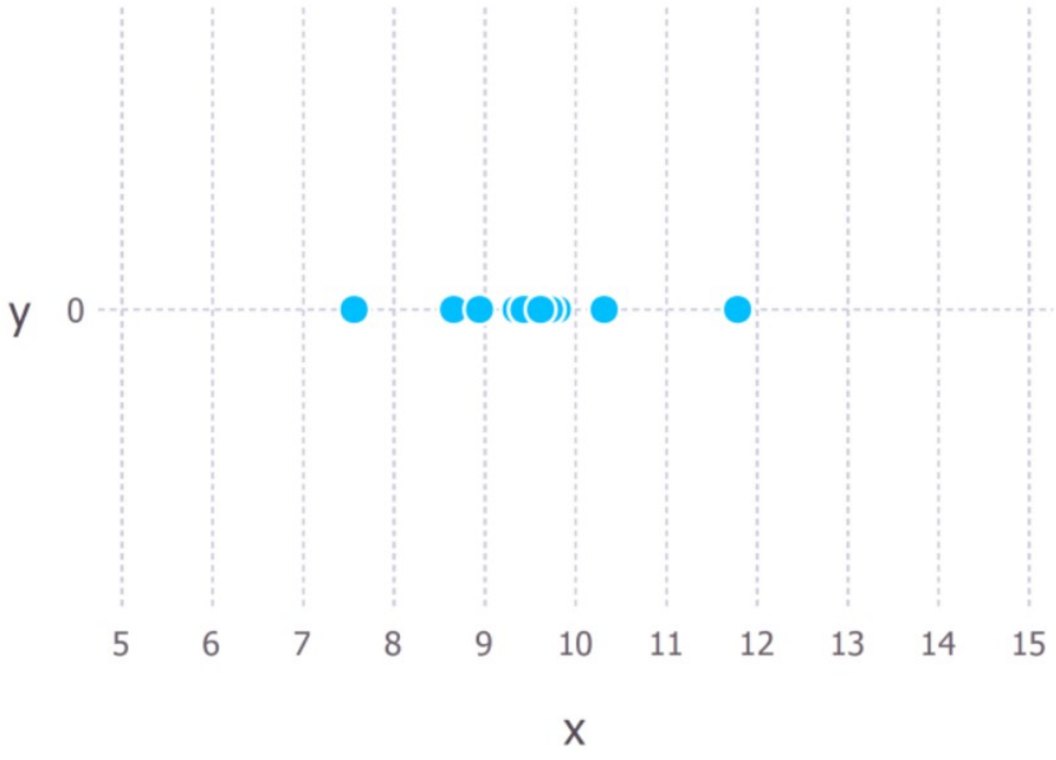
Consider the following problem:

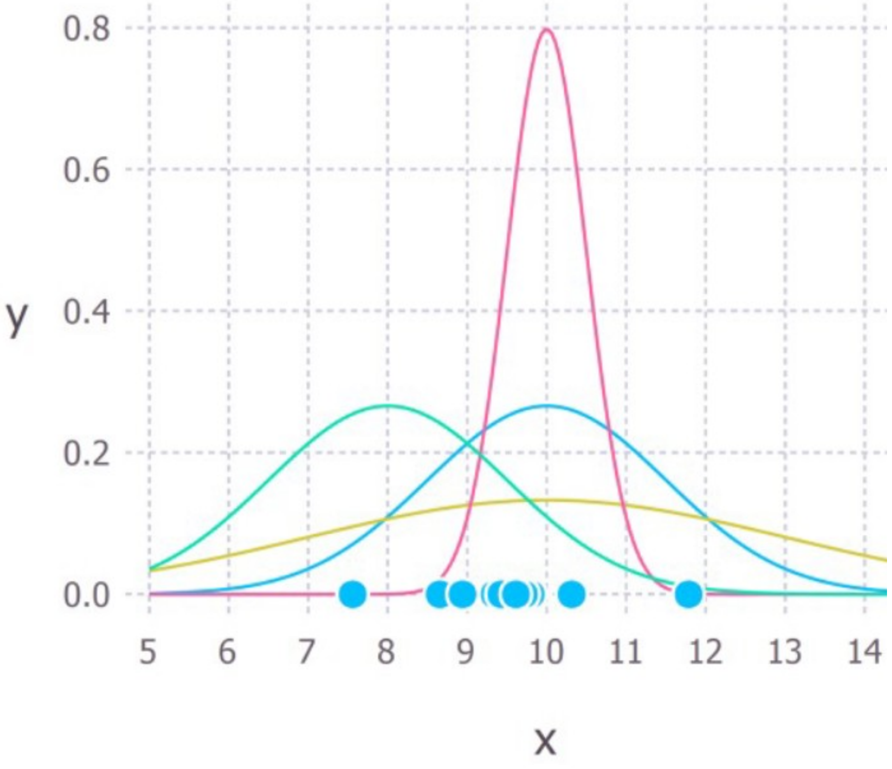
- We get a sample of points
- We know that they come from the Gaussian Normal distribution
- But how do we estimate the parameters?

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

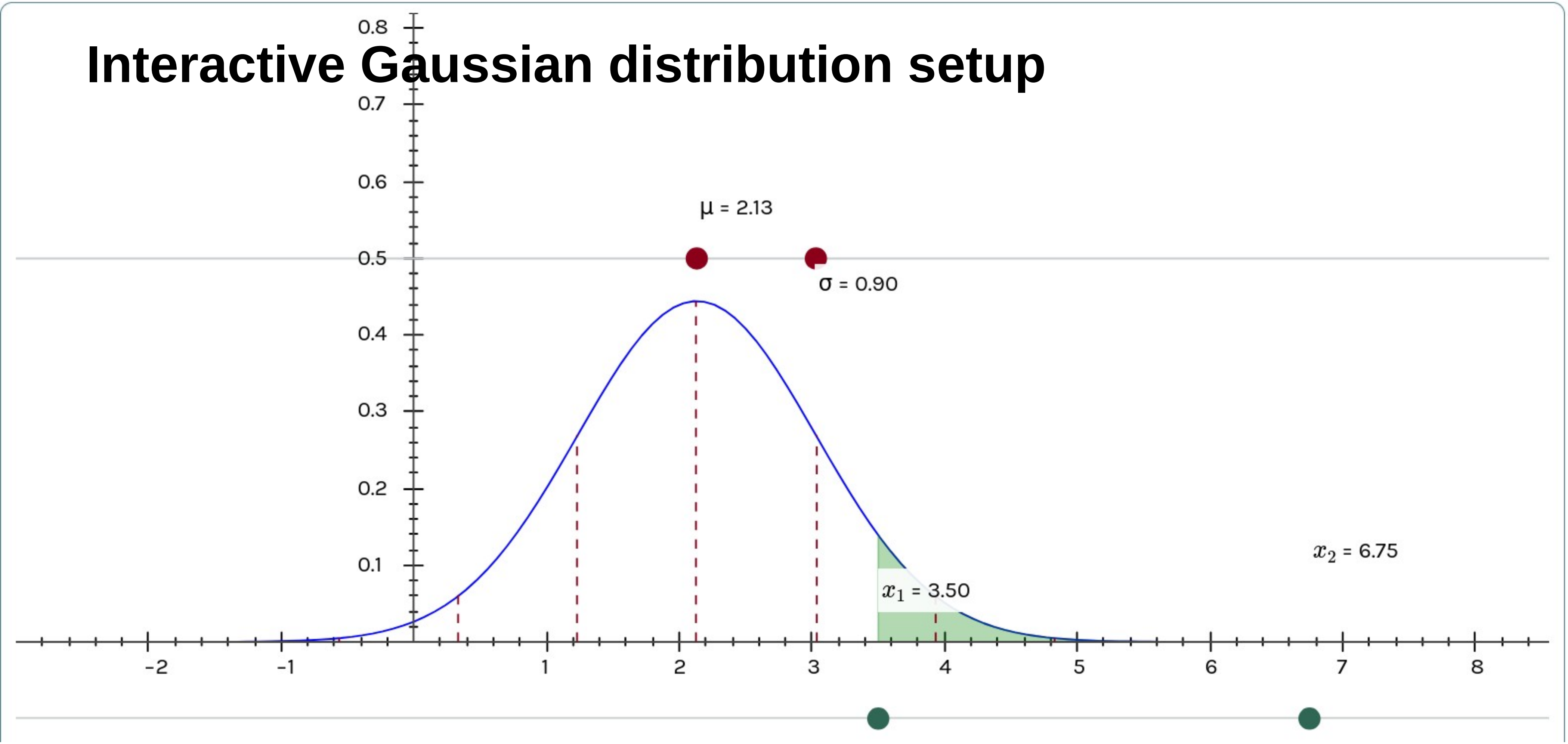


What value to use to measure how good each distribution is?



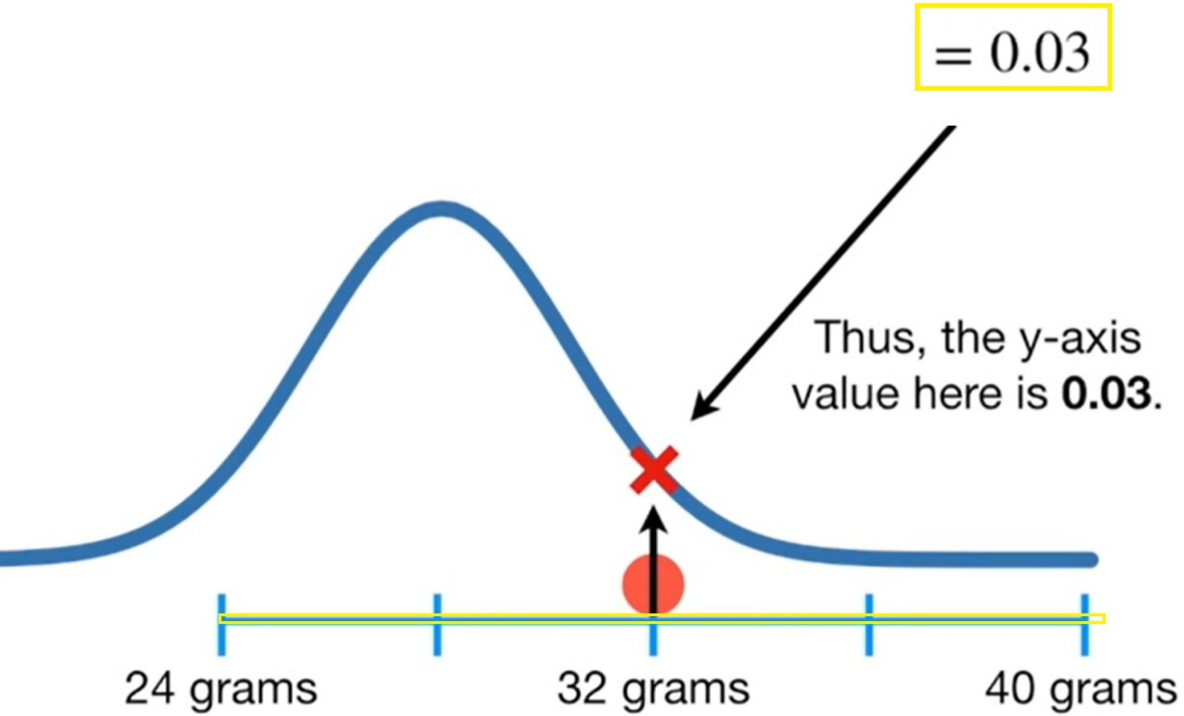


Interactive Gaussian distribution setup



https://www.intmath.com/counting-probability/normal-distribution-graph-interactive.php

Maximum Likelihood estimation (MLE)



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- For every point – compute its probability based on density formula

- Multiply all point values to get the total value

$$\text{score}(\mu, \sigma) = \prod_i f(x_i | \mu, \sigma)$$

See full explanation:
https://didawiki.cli.di.unipi.it/lib/exe/fetch.php/dm/19_dm2_maximum_likelihood_estimation_2022_23.pdf

$$\textit{minimize} \quad \text{score}(\mu, \sigma) \quad \textit{for } \mu, \sigma$$

$$\text{score}(\mu, \sigma) = \prod_i f(x_i | \mu, \sigma)$$

$$x_1, x_2, \ldots, x_m$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Can anybody solve this analytically?

$$\textit{minimize} \quad \text{score}(\mu,\sigma) \quad \textit{for } \mu,\sigma$$

$$\text{score}(\mu,\sigma)=\prod_i f(x_i|\mu,\sigma)$$

$$x_1, x_2, \ldots, x_m$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\log f(x) = -\log\sigma - \log\sqrt{2\pi} - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

Rewrite as:

$$\log \textit{score}(\mu,\sigma)=\sum_i \log f(x_i|\mu,\sigma)=\sum_i -\log\sigma-\log\sqrt{2\pi}-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

Solve with first derivative conditions:

$$\frac{\partial}{\partial\mu}\log \textit{score}(\mu,\sigma)=0$$

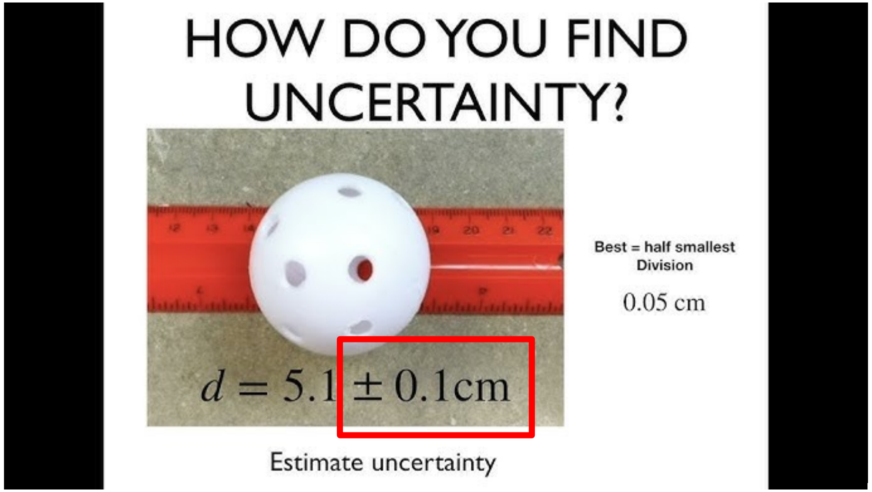
$$\frac{\partial}{\partial\sigma}\log \textit{score}(\mu,\sigma)=0$$

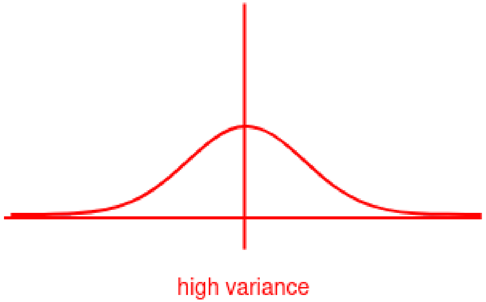
$$\mu=\frac{1}{m}\sum_i x_i$$

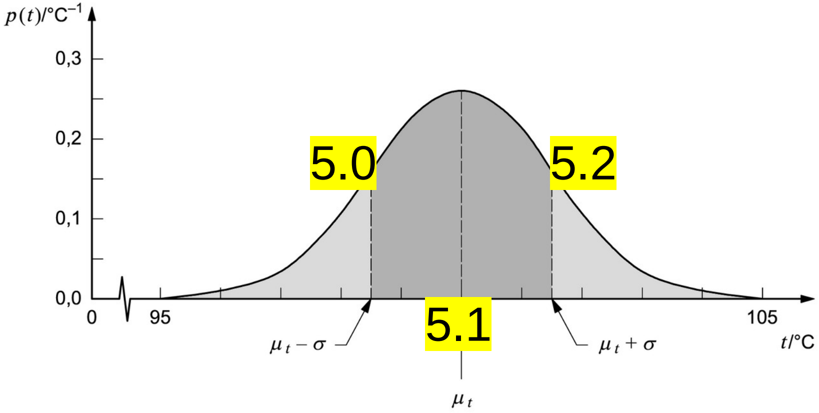
$$\sigma=\sqrt{\frac{1}{m}(x_i-\mu)^2}$$

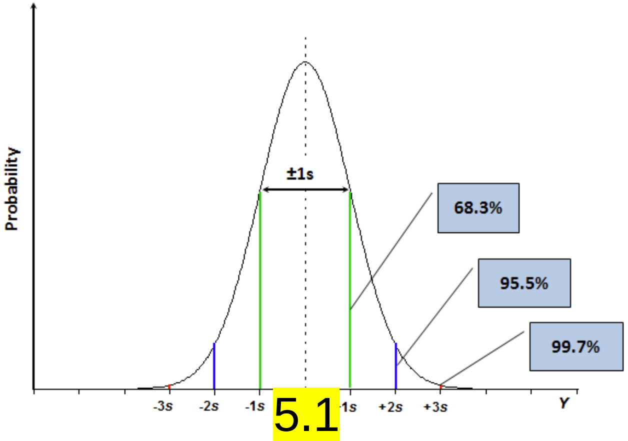
OR we can use something more familiar...

Variance is the quality of distribution for continuous variable









Minimum variance estimation

$$S=\sqrt{\frac{1}{m}(x_i-\mu)^2} \quad \text{OR} \quad S^2=\frac{1}{m}(x_i-\mu)^2$$

$$\text{minimize} \quad \underset{x_1,x_2,\dots,x_m}{S(\mu)} \quad \text{for } \mu$$

Can anybody solve this analytically?

Minimum variance estimation

$$S=\sqrt{\frac{1}{m}(x_i-\mu)^2} \quad \text{OR} \quad S^2=\frac{1}{m}(x_i-\mu)^2$$

$$\text{minimize} \quad \underset{x_1, x_2, \ldots, x_m}{S(\mu)} \quad \text{for } \mu$$

$$\frac{\partial}{\partial\mu}S^2=\frac{1}{m}\sum_i 2(x_i-\mu)(-1)=0$$

$$\sum_i(\mu-x_i)=0$$

$$m\mu-\sum_i x_i=0$$

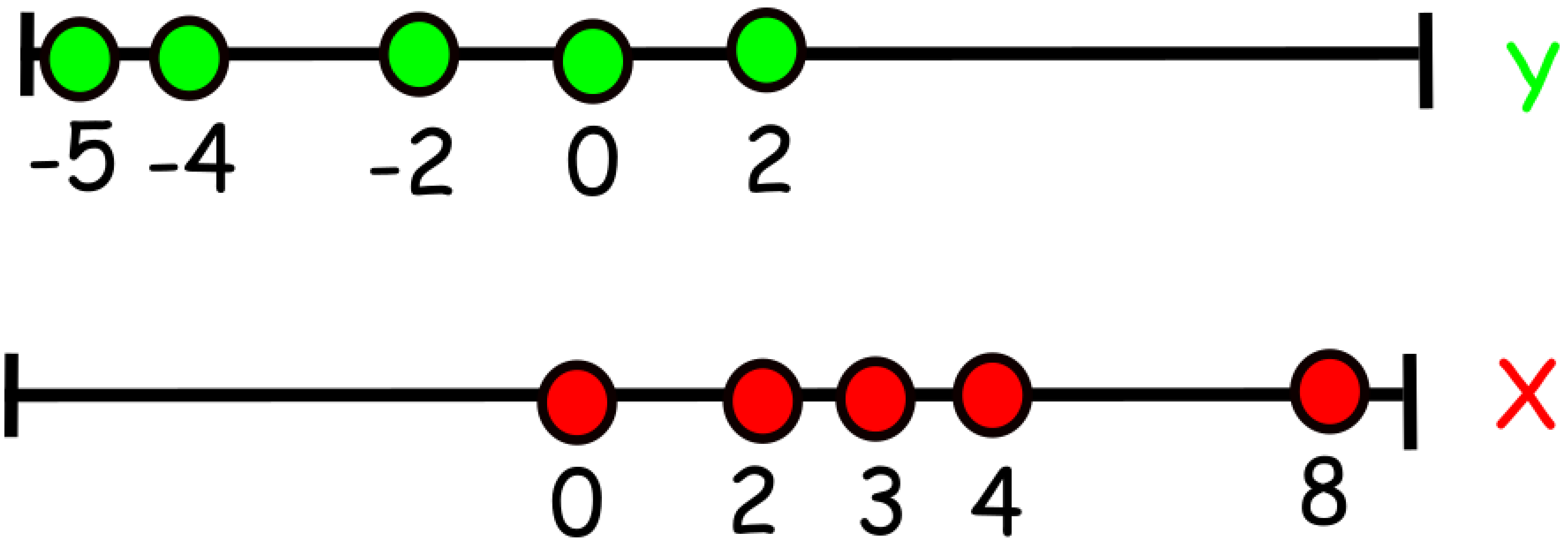
$$\mu=\frac{1}{m}\sum_i x_i$$

Same as in MLE! But only for Gaussian Normal!
For now we will use this formula because of simplicity

Now we switch to multi-variable (multivariate) case

Consider the new problem:

- We get a sample of points for two variables
- We know that they both come from the Gaussian Normal distribution
- If they are independent, just solve every 1D problem as previously
- But what if we have a condition that they come from the same distribution?



Could you formulate the 2-variate optimization problem?

Could you formulate the 2-variate optimization problem?

$$S_X^2 = (x_i - \mu)^2 \quad \text{and} \quad S_Y^2 = (y_i - \mu)^2$$

$$\text{minimize} \quad S_{total}^2(\mu) = S_X^2(\mu + 5) + S_Y^2(\mu) \quad \text{for } \mu$$

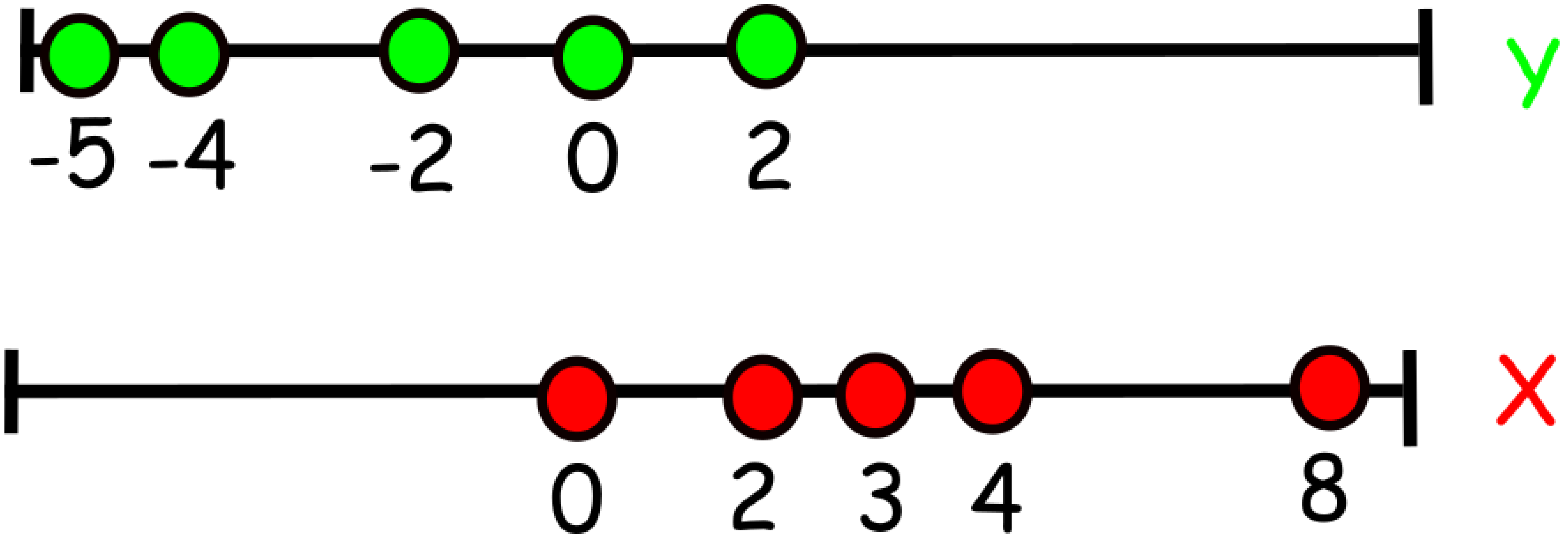
$$x_1, x_2, \ldots, x_m, y_1, \ldots, y_n$$

$$\frac{\partial}{\partial \mu} S^2 = \sum_i 2(x_i - \mu)(-1) + \sum_i 2(y_i - \mu)(-1) = 0$$

……. .

Consider a little modified new problem:

- We get a sample of points for two variables
- We know that they both come from the Gaussian Normal distribution
- If they are independent, just solve every 1D problem as previously
- But what if we have a condition that:
  - **Both share the same variance**
  - **But X has its mean shift by exactly 5 units**



Could you formulate the 2-variate optimization problem?

$$S_X^2 = (x_i - \mu)^2 \quad \text{and} \quad S_Y^2 = (y_i - \mu)^2$$

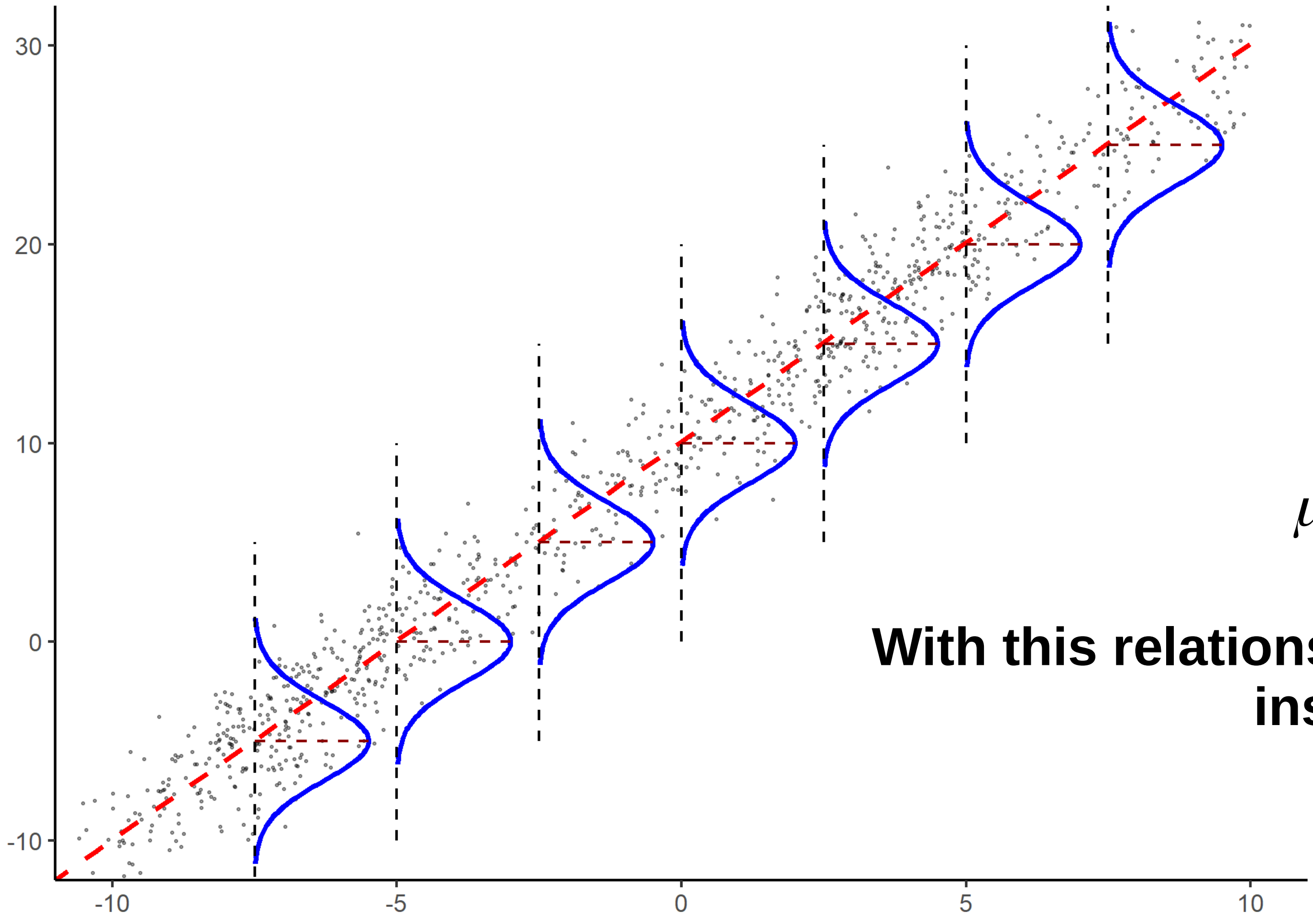
$$\text{minimize} \quad S_{total}^2(\mu) = S_X^2(\mu + 5) + S_Y^2(\mu) \quad \text{for } \mu$$

$$x_1, x_2, \ldots, x_m, y_1, \ldots, y_n$$

$$\frac{\partial}{\partial \mu} S^2 = \sum_i 2(x_i - \mu - 5)(-1) + \sum_i 2(y_i - \mu)(-1) = 0$$

....

Finally, the linear regression is just a set of such tasks



$$\mu_x = \beta_0 + X * \beta_1$$

With this relationship between parameters instead of +5

And this is simply Minimum Variance objective

- we define the **residual sum of squares** (RSS) as

$$RSS = e_1^2 + e_2^2 + \cdots + e_m^2 = \boxed{\sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$$

- the least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = 0 \rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m} (x_i - \bar{x})^2}$$

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = 0 \rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

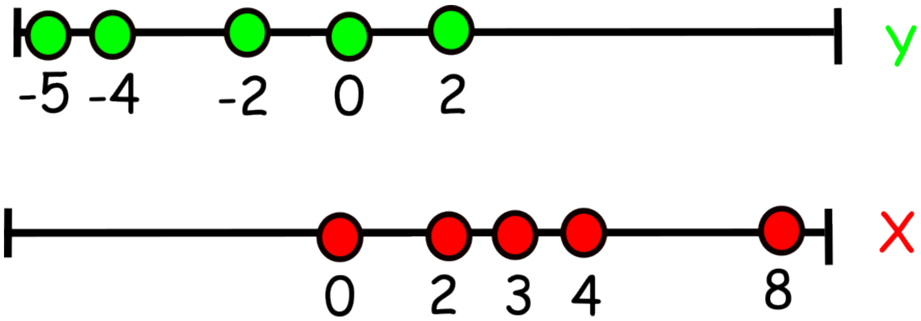
Same derivative computation

What is the p-value of the coefficient?

Lets recall how we defined a problem for 2 distributions:

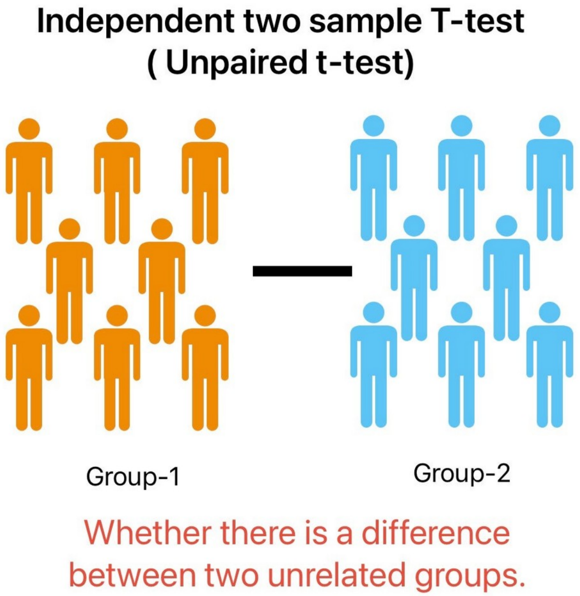
- We get a sample of points for two variables
- We know that they both come from the Gaussian Normal distribution
- If they are independent, just solve every 1D problem as previously
- But what if we have a condition that:
  - **Both share the same variance**
  - **But X has its mean shift by exactly 5 units**

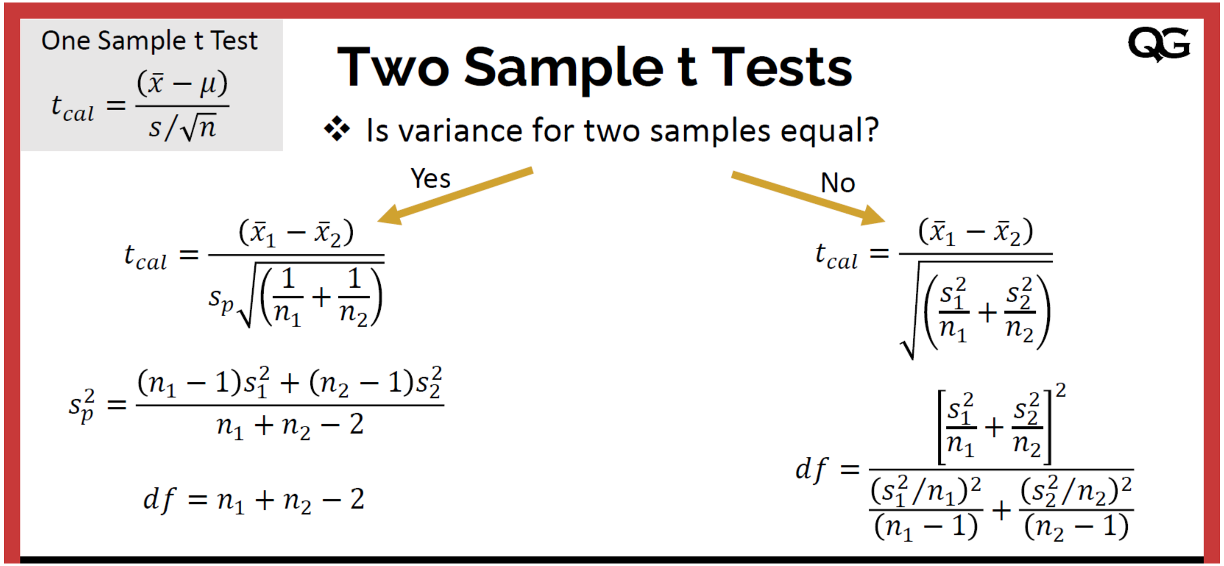
- **But what if the measured variable (whether X or Y) has no impact on the distribution? Is there a formal (hypothesis testing) way to compute a p-value to decide?**



What is the p-value of the coefficient?

This is the same case as two sample t-test from statistics 101!





See: https://www.qualitygurus.com/two-sample-t-test/

See: https://pub.aimind.so/t-test-essentials-understanding-one-sample-independent-two-sample-and-paired-test-methodologies-5dbbecd319b5

What is the p-value of the coefficient?

- The most common hypothesis test
  - H_0: there is no relationship between X and Y,
  - H_A: there is some relationship between X and Y,
- mathematically this corresponds to testing
  - $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$,
- the test stems from the standard error and t-statistic given by

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

  - the statistic will have a t-distribution with m-2 degrees of freedom,
  - the corresponding p-value is the probability of observing any value equal to $|t|$ or larger,
  - both under the H_0 assumption, i.e., assuming $\beta_1 = 0$.

What is R^2 (R-squared)?

- This is another way to measure how good we are in comparison to X-Y independent (baseline) model (beta1 = 0)
- But unlike previous p-value, this is based on metric value
- Recall the optimization problem for Variance minimization:

$$S^2_{x=1} = \sum_i (y_{1i} - \mu)^2 \quad \text{and} \quad S^2_{x=2} = \sum_i (y_{2i} - \mu_{x=2})^2$$

$$\text{minimize} \quad S^2_{total}(\mu) = S^2_{X=1}(\mu_{x=1}) + S^2_{x=2}(\mu_{x=2}) \quad \text{for} \ \mu_x = \beta_0 + \beta_1 X$$

$$x = 1 \rightarrow y_{11}, y_{12}, \ldots$$

$$x = 2 \rightarrow y_{21}, y_{22}, \ldots$$

What is R^2 (R-squared)?

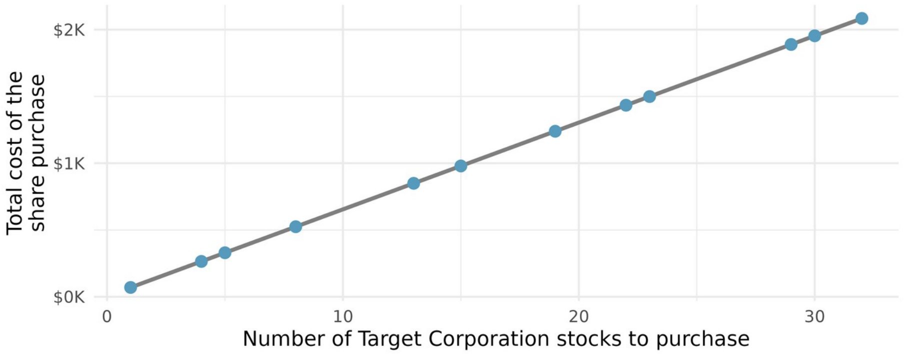
- This is another way to measure how good we are in comparison to X-Y independent (baseline) model (beta1 = 0)
- But unlike previous p-value, this is based on metric value
- We start with baseline model (beta1 = 0), it has a certain value of

$$S^2_{total}(\beta_0, \beta_1 = 0)$$

- We want to improve it for our derivative-based solution:

$$S^2_{total}(\beta_0, \hat{\beta}_1)$$

- And finally, if we were to ideally solve the problem:



$$S^2_{total}(perfect) = 0$$

What is R^2 (R-squared)?

- This is another way to measure how good we are in comparison to X-Y independent (baseline) model (beta1 = 0)
- RSS = our variance
- TSS = variance of baseline model
- Decrease of minimization value in percentage

- **R-squared** gives the fraction of variance explained by the model

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- where $TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2$ stands for the total sum of squares,
- and $RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$ stands for the residual sum of squares,

What is F-statistics and how to evaluate a total model?

- It is R-squared converted to hypothesis format with p-value
- Simply speaking – if p-value < 0.05, you model is useful

- What would you say about the following model?

```
summary(lm(y ~ x,d))

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671   0.02559 *
x             0.4999     0.1178   4.243   0.00216 **

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,  Adjusted R-squared:  0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

>= 1 star is GOOD

> 0.5 is GOOD

< 0.05, GOOD

Questions?

Rstudio/R basic programming

LOGICAL -> INTEGER -> NUMERIC -> CHARACTER

“boolean” “integer” “double” “string”

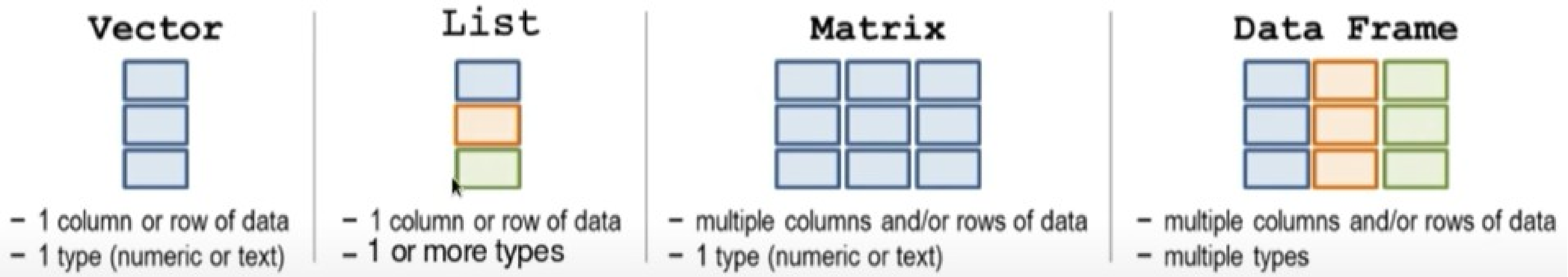
a = 1
class(a) = type of variable
vector = array of same type
….

https://www.w3schools.com/r/r_get_started.asp

https://www.datacamp.com/courses/free-introduction-to-r

Download today’s tutorial script:

2	29.9.	JB, AA, JK	Simple linear regression.	san_lreg.zip, pres_2.pdf



	dma	day	amount_of_deaths	new_deaths	weekly_deaths	gtrends
66	1	2021-05-02	66247	47	491	62
67	1	2021-05-09	66690	41	443	62
68	1	2021-05-16	67030	43	340	55
7481	2	2020-02-02	0	0	0	68
7482	2	2020-02-09	0	0	0	66
7483	2	2020-02-16	0	0	0	68
7484	2	2020-02-23	0	0	0	65
7485	2	2020-03-01	0	0	0	66
7486	2	2020-03-08	0	0	0	66
7487	2	2020-03-15	1	0	1	68
7488	2	2020-03-22	12	3	11	77
7489	2	2020-03-29	56	6	44	79
7490	2	2020-04-05	184	16	128	100
7491	2	2020-04-12	397	38	213	96
7492	2	2020-04-19	777	25	380	77
7493	2	2020-04-26	1170	21	393	77
7494	2	2020-05-03	1559	20	389	73
7495	2	2020-05-10	1945	21	386	69

Showing 66 to 84 of 14,212 entries, 6 total columns

<----- **View**(**myDataFrame**)

myDataFrame$new_deaths = access column as vector

$$H = \beta_0 + \beta_1 \cdot V + \beta_2 \cdot K + \epsilon , \qquad \epsilon \in N(0, \sigma^2)$$

can be denoted shortly as (R notation): H ~ V + K

```
lm.fit <- lm(medv ~ lstat, data = Boston)
```

Also a useful notation: **H ~ .** (H = regression of all other variables)

If you are not sure about something about function/data:

help(lm) **help(Boston)**

?lm **?Boston**

Today’s homework (voluntary) for 1 point

2	29.9.	JB, AA, JK	Simple linear regression.	san_lreg.zip pres_2.pdf

Deadline: 06.10.2025 (next seminar)

Goals:

*) Try to write some simple R code

*) Understand basic concepts regarding the linear regression

*) You could still answer with pseudocode and theoretical answers if you do not want to use R