5LN445 Mathematics for Language Technologists

Q & A, Exam prep

Exam on Wednesday

This level of fancy is enough.
No need for trigonometry!

You bring:
Calculator
Pen
Any handwritten notes you think you need

You will be given:
~~Table of formulas~~
Table for normal CDF
Table for χ^2

The exam will cover statistics and linear algebra. Max 40 points, just like last time. Points are distributed approximately proportional to the number of lectures per major theme.

L8: Sampling distributions
L9: Confidence intervals
L10: Significance for a proportion
L11: Significance for several proportions
L12: Linear regression
L13: Vectors (basic definitions)
L14: Geometry
L15: Applications / Exercises
L16: Matrices
L17: Applications in python / Exercises

Learning goals

Statistics: discuss and apply elementary concepts in statistics such as distribution, sample, estimation, and hypothesis testing. Readings: OpenIntro Statistics.

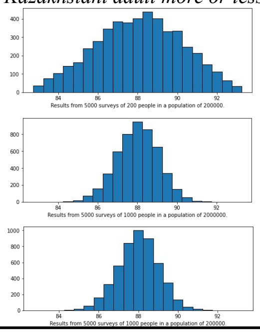
- Some Special Distributions
  - Binomial distribution
  - Normal distribution
  - *Uniform distribution*
  - Categorical distribution
- Modelling
  - Hypothesis testing
  - Statistical significance
  - Confidence interval
  - Significance interval
  - Regression

Linear Algebra: discuss and apply elementary concepts relating to vector spaces. Readings: A Concise Introduction to Linear Algebra

- Vector spaces
  - Vectors
  - Axis
  - Vector operations: addition, subtraction and scalar multiplication
  - Vector length, distance metrics
  - Dot product
- Geometry
  - Shapes in higher dimensions
  - Intersecting objects
  - *Classification*
  - Projection

Sampling distribution

"Suppose the proportion of Kazakhstani adults who support the expansion of solar energy is 88%, which is our parameter of interest. Is a randomly selected Kazakhstani adult more or less likely to support the expansion of solar energy?"



Sample proportions will be nearly normally distributed as:

$$\hat{p} \sim N\left(\mu = p, \sigma = \sqrt{\frac{p\cdot(1-p)}{n}}\right)$$

$$p \pm z^* \cdot \sigma$$

Sampled observations must be **independent**. For a low **sample size**, 95% of the spread of the sampling distribution might cover a large area.

Multiple samples (yes/no) → Binomial distr. → Normal distr.

ST Example 5.11-13

In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a *"mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient"*. This poll included responses of 1,042 New York adults between Oct 26th and 28th, 2014.

What is the point estimate in this case, and is it reasonable to use a normal distribution to model that point estimate?

$$p = 0.82$$

Estimate the standard error of p from the Ebola survey.

$$SE = \sqrt{\frac{p\cdot(1-p)}{n}} = \sqrt{\frac{0.82\cdot(1-0.82)}{1042}} \approx 0.012$$

Construct a 95% confidence interval for p, the proportion of New York adults who supported a quarantine for anyone who has come into contact with an Ebola patient.

$$p \pm z^* \cdot SE \rightarrow 0.82 \pm 1.96 \cdot 0.012 \rightarrow (0.796, 0.844)$$

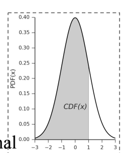
Normal distribution

Standard normal (or "z-points")

$$z_i = \frac{x_i - x^*}{\sigma}$$

$$Z \sim \mathcal{N}(0,1)$$

Normal CDF



$\Phi(z)$ = probability mass up to z

$\Phi^{-1}(x)$ = position in a standard normal distribution given a probability mass.

Approximating the binomial distribution

$$B \sim Binom(n,p)$$

$$P(a \le B \le b) \approx \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

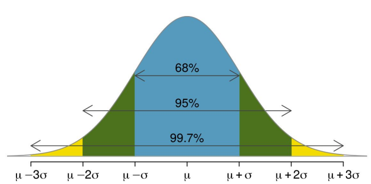
Proportion $\frac{B}{n} \rightarrow X \sim \mathcal{N}\left(\mu = p, \sigma = \sqrt{\frac{p(1-p)}{n}}\right)$

Adding/subtracting normal *random variables*:

$aX + c \quad \Rightarrow \mu = a\mu + c, \quad \sigma^2 = (a\sigma)^2$

$aX + bY + c \quad \Rightarrow \mu = a\mu_X + b\mu_Y + c, \ \sigma^2 = (a\sigma_X)^2 + (b\sigma_Y)^2$

$a,b,c \in \mathbb{R}$, $a,b \neq 0$, X and Y are independent normal r.v.



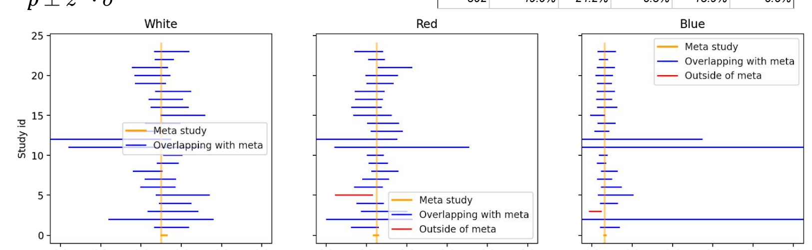
Exercise

Choose an ailment (colour) and calculate the parameters of the the sampling distribution for a handful of studies. Calculate the 95% confidence intervals.

$$\hat{p} \sim N\left(\mu = p, \sigma = \sqrt{\frac{p\cdot(1-p)}{n}}\right)$$

$$p \pm z^* \cdot \sigma$$

N	P(white)	P(red)	P(blue)	P(green)	P(yellow)
237	51.9%	26.6%	4.6%	10.1%	6.8%
111	55.9%	26.1%	3.6%	6.3%	8.1%
302	49.0%	24.2%	6.3%	13.9%	6.6%



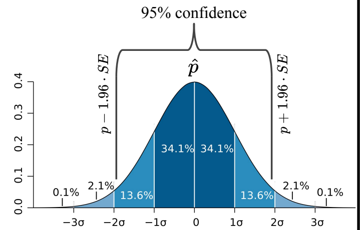
Can Facebook categorize user interests?

$p = 0.67, n = 850$ We also know that: $\hat{p} \sim N\left(\mu = p, \sigma = \sqrt{\frac{p\cdot(1-p)}{n}}\right)$

The 95% confidence interval is defined as: "point estimate" ± 1.96 · "standard error"

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.67 \cdot 0.33}{850}} \approx 1.6\%$$

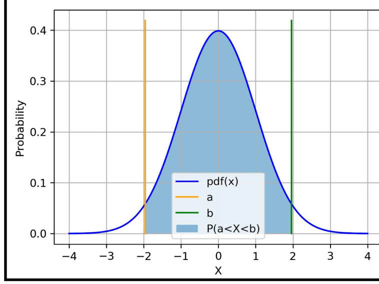
$$p \pm 1.96 \cdot SE = 0.67 \pm 1.96 \cdot 0.016$$
$$\approx (0.67 - 0.03, 0.67 + 0.3)$$
$$= (0.64, 0.70)$$



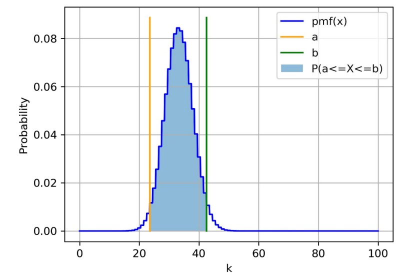
"Standard error" is another name for the standard deviation of the sampling distribution.

Confidence interval for non-normal distributions

X ~ Normal(0, 1)
95.0% confidence is the range [-1.96, 1.96]
P(X > a)=0.975, P(X < b)=0.975, P(a < X < b)=0.950



X ~ Binomial(100, 0.33)
95.0% confidence is the range [24, 42]
P(X >= a)=0.984, P(X <= b)=0.972, P(a <= X <= b)=0.956



Choosing a sample size

Given a poll where 85% out of 500 people answered yes to something interesting, how wide was the margin of error (ME) for a 95% confidence level?

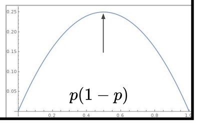
$$ME = z^* \cdot SE = 1.96 \cdot \sqrt{\frac{0.85 \cdot 0.15}{500}} \approx 0.031$$

How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%?

$$n \geq \left(\frac{z^*}{\text{ME}}\right)^2 p(1-p) \Rightarrow n \geq \left(\frac{1.96}{.01}\right)^2 \cdot 0.85 \cdot 0.15 \Rightarrow n \geq 4898.04$$

What about if we don't have any information about p?

$$n \geq \left(\frac{1.96}{.01}\right)^2 \cdot 0.50 \cdot 0.50 \Rightarrow n \geq 9604$$



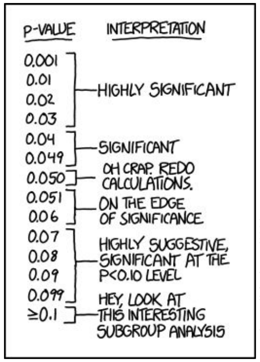
p-values

We then use this test statistic to calculate the p-value, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

$P(\hat{p} > .50 \mid H_0) = ?$

If the **p-value is low** (lower than a significance level α), we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence **reject H_0**.

If the **p-value is high** (higher than a significance level α) we say that it is likely to observe the data even if the null hypothesis were true, and hence do **keep H_0**.



https://xkcd.com/882/

Results from the GSS

The General Social Survey (GSS) by the University of Chicago asks the same question, below are the distributions of responses from the 2010 GSS as well as from a group of introductory statistics students at Duke University:

	GSS	Duke
A great deal	454	69
Some	124	30
A little	52	4
Not at all	50	2
Total	680	105

Parameter and point estimate

Parameter of interest: Difference between the proportions of *all* Duke students and *all* Americans who would be bothered a great deal by the northern ice cap completely melting.

$p^{\star}_{Duke} - p^{\star}_{USA}$ ← Actual proportions in the populations

Point estimate: Difference between the proportions of *sampled* Duke students and *sampled* Americans who would be bothered a great deal by the northern ice cap completely melting.

$p_{Duke} - p_{USA}$ ← Estimated proportions from samples

From before:

CI: *point estimate* ± *margin of error*

HT: Use *Z = (point estimate - null value) / SE* to find appropriate p-value.

New standard error: $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ ← How did this happen?

Sample proportions are also nearly normally distributed

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap (p_{Duke} - p_{US}).

Data	Duke	USA
A great deal	69	454
Not a great deal	36	226
Total	105	680
p	0.657	0.668

$$(p_{Duke} - p_{USA}) \pm z^{\star} \cdot \sqrt{\frac{p_{Duke}(1-p_{Duke})}{n_{Duke}} + \frac{p_{USA}(1-p_{USA})}{n_{USA}}}$$

$$= (0.657 - 0.668) \pm 1.96 \cdot \sqrt{\frac{0.657 \cdot 0.343}{105} + \frac{0.668 \cdot 0.332}{680}}$$

$$= -0.011 \pm 1.96 \cdot 0.0497 = -0.011 \pm 0.097 = (-0.108, 0.086)$$

CI vs. HT for proportions

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Duke	USA
A great deal	69	454
Not a great deal	36	226
Total	105	680
p	0.657	0.668

Null hypothesis:

$$Y \sim \mathcal{N}(0, \sigma) \quad \sigma = \sqrt{\frac{p_{pooled}(1-p_{pooled})}{n_{Duke}} + \frac{p_{pooled}(1-p_{pooled})}{n_{USA}}}$$

$$= \sqrt{\frac{0.666 \cdot 0.334}{105} + \frac{0.666 \cdot 0.334}{680}}$$

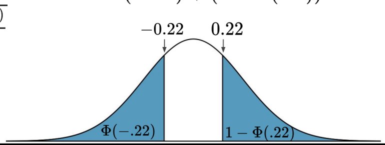
$$\approx 0.0495$$

p-value (two sided):

$$\mathrm{P}(|Y| > |p_{Duke} - p_{USA}|) = \mathrm{P}(|Y| > |0.657 - 0.668|)$$

$$= \mathrm{P}(Y > \tfrac{0.11}{0.0495}) - \mathrm{P}(Y < -\tfrac{0.11}{0.0495})$$

$$\approx \Phi(-.22) + (1 - \Phi(.22)) \approx 0.82$$



2009 Iranian presidential election

In the 2009 Iran election, there were accusations of election fraud. We'll compare the data from a poll conducted before the election (observed data) to the reported votes in the election to see if the two follow the same distribution.

Candidate	Observed # of voters in poll	Reported % of votes in election
(1) Ahmedinajad	338	63.29%
(2) Mousavi	136	34.10%
(3) Minor candidates	30	2.61%
Total	504	100%
	↓ observed	↓ expected distribution

2009 Iranian presidential election

Hypotheses for testing if the distributions of reported and polled votes are different:

H_0: The observed counts from the poll **follow** the same distribution as the reported votes.

H_A: The observed counts from the poll do **not follow** the same distribution as the reported votes.

Mahmoud Ahmadinejad

Mir-Hossein Mousavi

Candidate	Observed # of voters in poll	Reported % of votes in election
(1) Ahmedinajad	338	63.29%
(2) Mousavi	136	34.10%
(3) Minor candidates	30	2.61%
Total	504	100%

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df=k-1}$$

2009 Iranian presidential election

Candidate	Observed # of voters in poll	Reported % of votes in election	Expected # of votes in poll
(1) Ahmedinajad	338	63.29%	504 × 0.6329 = 319
(2) Mousavi	136	34.10%	504 × 0.3410 = 172
(3) Minor candidates	30	2.61%	504 × 0.0261 = 13
Total	504	100%	504

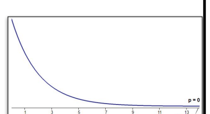
$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df=k-1}$$

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(30 - 13)^2}{13} = 22.23$$

$$\chi^2_{df=3-1=2} = 30.89$$



http://www.statdistributions.com/chisquare/

The least squares line

Form of the model: $y = \beta_0 + \beta_1 \cdot x + \epsilon$

What we try to learn/fit: $\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$

Predicted y — Intercept — Slope — Explanatory variable

Notation:

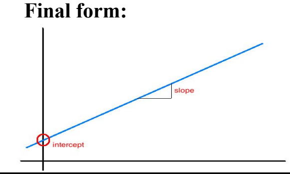
Intercept parameter: β_0

Slope parameter: β_1

For a good fit:

- Linearity
- Nearly normal residuals
- Constant variability
- No extreme outliers

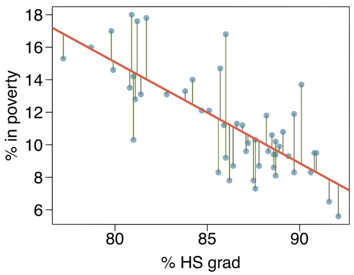
Final form:



Residuals

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

$$residual = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \ldots + e_n^2 \qquad e_i = y_i - \hat{y}_i$$



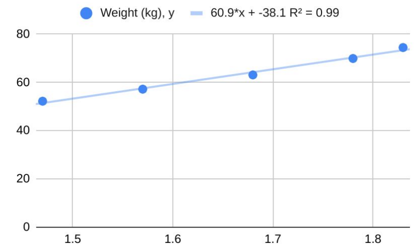
Residuals are the leftovers from the model fit:
Data = Fit + Residual

Given a model f(x):

$$residual = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - f(x))^2$$

A smaller example

Height (m), x	Weight (kg), y
1.47	52.21
1.57	57.2
1.68	63.11
1.78	69.92
1.83	74.46



$$C = \begin{bmatrix} 0.018 & 1.068 \\ 1.068 & 65.714 \end{bmatrix}$$

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

$$\beta_1 = R \cdot \frac{s_y}{s_x}$$

$$C = \begin{pmatrix} s_x^2 & R s_x s_y \\ R s_y s_x & s_y^2 \end{pmatrix} = \begin{pmatrix} Var(x,x) & Var(x,y) \\ Var(y,x) & Var(y,y) \end{pmatrix}$$

$$Var(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

R: 0.99
s_x: 0.13
s_y: 8.11
R^2: 0.99
slope: 60.89
intercept: -38.07
mean x: 1.666
mean y: 63.38

Linear algebra

A **vector** is an ordered sequence of numbers that describe a position in some space.

Centre
Axis
Position
Length / Distance
Dimensionality

$\mathbf{u} = (1, 2, 3, 4)^T$, $\mathbf{v} = (-2, 3, -4, 1)^T$
$\mathbf{u}, \mathbf{v} \in \mathbb{R}^4$

An equation like x+y=2 might have many solutions (x, y). In some cases, it can be of interest to find a solution (x, y) that several equations share. Ex:

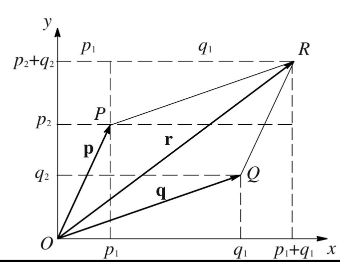
$$\begin{cases} 2x + y &= 1 \\ x + y &= 0 \end{cases}$$

Here we will focus on solving such a **system of equations** by using the **substitution method**.

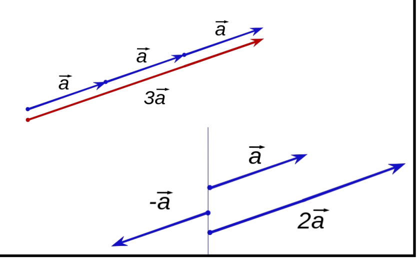
1. We know from equation 2 that x+y=0 then y=-x.
2. Substituting into equation 1 gives 2x+(-x)=1 ⇒ x=1.
3. Since y=-x , then y=-1

Addition & scalar multiplication

p + q = r Definition: $p_i + q_i = r_i$
Vector addition is performed elementwise



ap + bq = r Definition: $ap_i + bq_i = r_i$
Scalar multiplication is performed elementwise



Dot product

f: $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

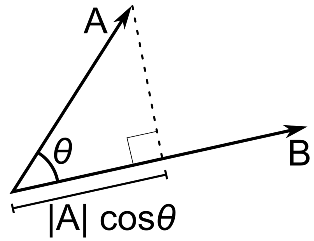
$\bar{u} \cdot \bar{v} = \sum_{i=1}^{n} u_i \cdot v_i$

$\bar{u} \cdot \bar{v} = \bar{v} \cdot \bar{u} = |\bar{u}||\bar{v}|cos(\theta)$

$\mathbf{u} \cdot \mathbf{v} = \mathbf{0} \Rightarrow$ Orthagonality

$cos\ \theta$ tells us:

1. if **u** and **v** are pointing in similar directions,
2. the angle between the two using $cos^{-1}\ \theta$



Ex:

$\mathbf{p} = (1, 2)^T$

$\mathbf{q} = (-1, 3)^T$

$\mathbf{p} \cdot \mathbf{q} = 1 \cdot -1 + 2 \cdot 3 = 5$

$\mathbf{p} = (2, 4)^T$

$\mathbf{q} = (-1, 2.5)^T$

$\mathbf{p} \cdot \mathbf{q} =$

You probably have inverse sin and cos functions on your calculator.

Distance & Lines

Distance function

d: $\mathbb{R}^n \to \mathbb{R}$

A **unit vector** has length 1

$\hat{\mathbf{u}} = \frac{\mathbf{u}}{||\mathbf{u}||}$

Euclidean distance (l_2 norm)

$d_{euclidean}(\mathbf{u}, \mathbf{v}) = \sqrt{\sum (u_i - v_i)^2}$

$\|\mathbf{p}\| = \sqrt{p_1^2 + p_2^2 + \cdots + p_n^2} = \sqrt{\mathbf{p} \cdot \mathbf{p}}$

Definition 1

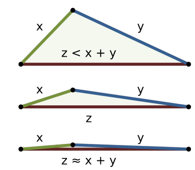
$\bar{r} = \bar{p} + t \cdot \bar{v} \qquad r_i = p_i + t \cdot v_i$

Definition 2

$$\begin{cases} x &= x_0 + t \cdot v_1 \\ y &= y_0 + t \cdot v_2 \\ z &= z_0 + t \cdot v_3 \end{cases} \qquad \frac{x - x_0}{v_1} = \frac{y - y_0}{v_2} = \frac{z - z_0}{v_3}$$

Triangle inequality

$\|\mathbf{v} + \mathbf{u}\| \le \|\mathbf{v}\| + \|\mathbf{u}\|$



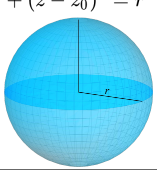
Spheres & Planes

Definition 1: $\|\mathbf{u} - \mathbf{q}\|$ = radius
, where: **u, q**$\in \mathbb{R}^m$

Definition 2: $x^2 + y^2 = r^2$ *(for m=2)*
$x^2 + y^2 + z^2 = r^2$ *(for m=3)*

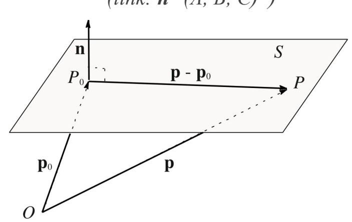
Origin at $(x_0, y_0, z_0)^T$:

$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = r^2$



Definition 1: $\mathbf{n}(\mathbf{p} - \mathbf{q}) = 0$
, where **n, p, q**$\in \mathbb{R}^m$

Definition 2: Ax + By + Cz + D = 0
*(link: **n**=(A, B, C)^T)*



Matrices

A matrix is basically a set of stacked vectors.

$$\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times m}$$

n number of rows

m number of columns,

$\bar{v}^{(1)} = (1, 2, 3, 4)^\top$

$\bar{v}^{(2)} = (4, \pi, 9, 4)^\top$

$\bar{v}^{(3)} = (5, 7, 3, -2)^\top$

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & \pi & 9 & 4 \\ 5 & 7 & 3 & -2 \end{pmatrix}$$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = (a_{ij}) \in \mathbb{R}^{m \times n}.$$

(The matrix' dimensions are ordered as in numpy.)

Operators

Addition / Subtraction

$A + B = C \Rightarrow a_{ij} + b_{ij} = c_{ij}$ $A, B, C \in \mathbb{R}^{n\times m}$

$A - B = C \Rightarrow a_{ij} - b_{ij} = c_{ij}$

Scalar multiplication

$\lambda \cdot A = C \Rightarrow \lambda \cdot a_{ij} = c_{ij}$ $A, C \in \mathbb{R}^{n\times m}$
$\lambda \in \mathbb{R}$

Scalar division

$\frac{A}{\lambda} = C \Rightarrow \frac{1}{\lambda} \cdot a_{ij} = c_{ij}$ $A, C \in \mathbb{R}^{n\times m}$
$\lambda \in \mathbb{R}$

(Note that these definitions are the same as those for vectors. Matrices work like stacked vectors.)

Matrix multiplication

$A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times o}, C \in \mathbb{R}^{m\times o}$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1o} \\ b_{21} & b_{22} & \cdots & b_{2o} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{no} \end{pmatrix} = \begin{pmatrix} c_{11} & a_{12} & \cdots & c_{1o} \\ c_{21} & c_{22} & \cdots & c_{2o} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mo} \end{pmatrix}$$

$c_{ij} = \sum_{k=1}^{n} a_{ik} \cdot b_{kj}$

Each c_{ij} is a dot product of row vector from A and column vectors from B. Note that AB is not the same as BA.

A way of setting up the multiplication:

$$\begin{array}{cc} & \begin{pmatrix} b_{11} & b_{12} \\ \vdots & \vdots \\ b_{n1} & b_{n2} \end{pmatrix} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} & \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ \vdots & \vdots \\ c_{m1} & c_{m2} \end{pmatrix} \end{array}$$