

Statistic 1

-Distribution

Bernoulli Distribution

Binomial Distribution

Multinomial Distribution

Poisson Distribution

Normal Distribution

Gamma Distribution

Exponential Distribution

χ^2 distribution

-Hypothesis testing

T-test

Wilcoxon

ANOVA, MANOVA

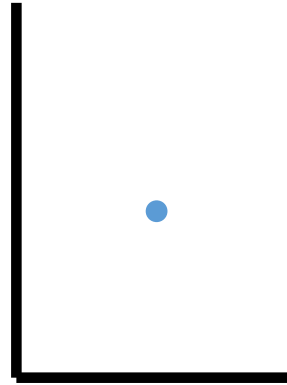
Kruskal-wallis

Fisher's exact test

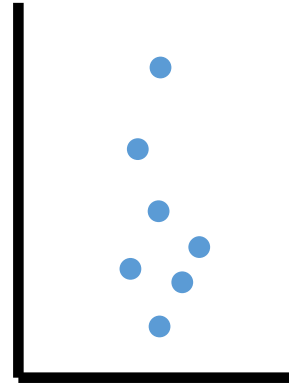
Post-hoc analysis

Distribution

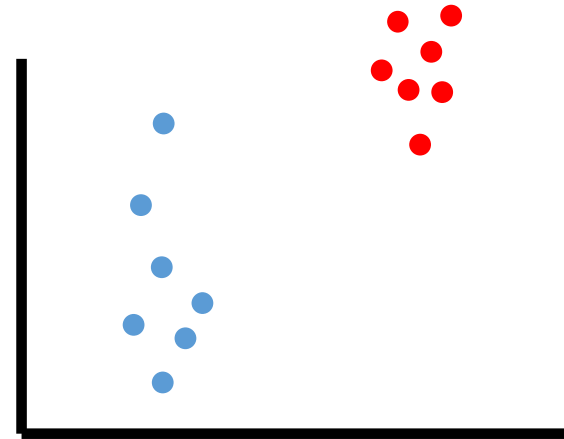
Handling biological data



Handling biological data



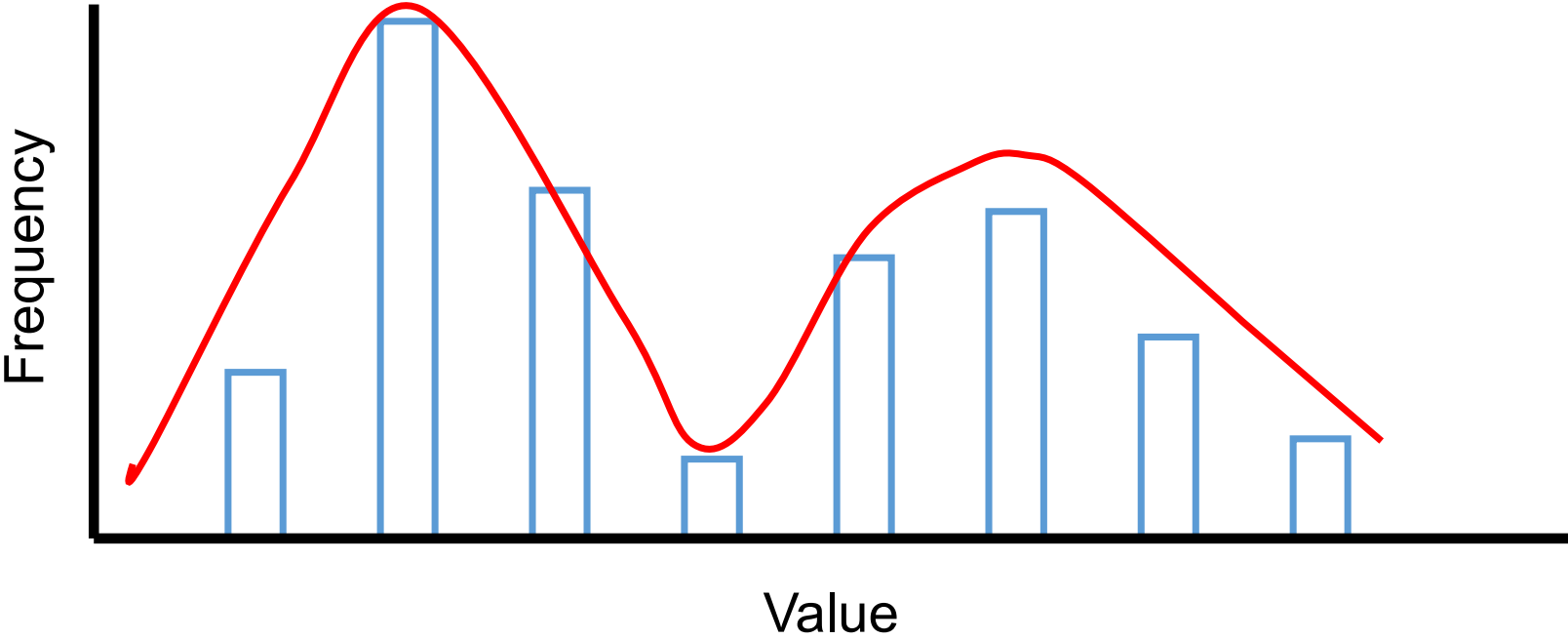
Handling biological data



Is it high?

Statistics!

Data distribution



Real data → Modeling

Bernoulli Distribution

- Probability of 1: p
 - Probability of 0: $1-p$
- Only two cases (0, 1)

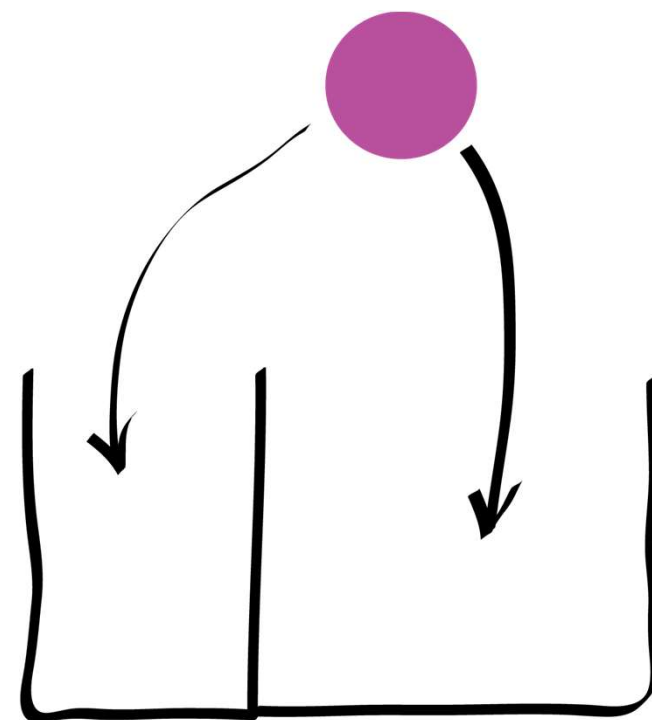
$$F(X=x|p) = p^x(1-p)^{1-x}$$

- $x=0$: $1-p$
- $x=1$: p

Expected value: p $E(x) = \sum_x xP(X=x) = (1)p + (0)(1-p) = p$

Variance: $p(1-p)$

$F(X=x|p)$: Probability Mass Function



Binomial Distribution

- n * Bernoulli trial

$$B(x; n, p) = {}_n C_x p^x (1 - p)^{n-x}$$

$$\text{※} {}_n C_x = \frac{n!}{x!(n-x)!}$$

- Expected value: np

- Variance: npq

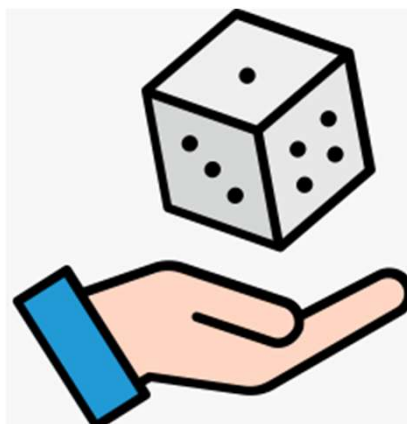
(q: 1-p)

Example

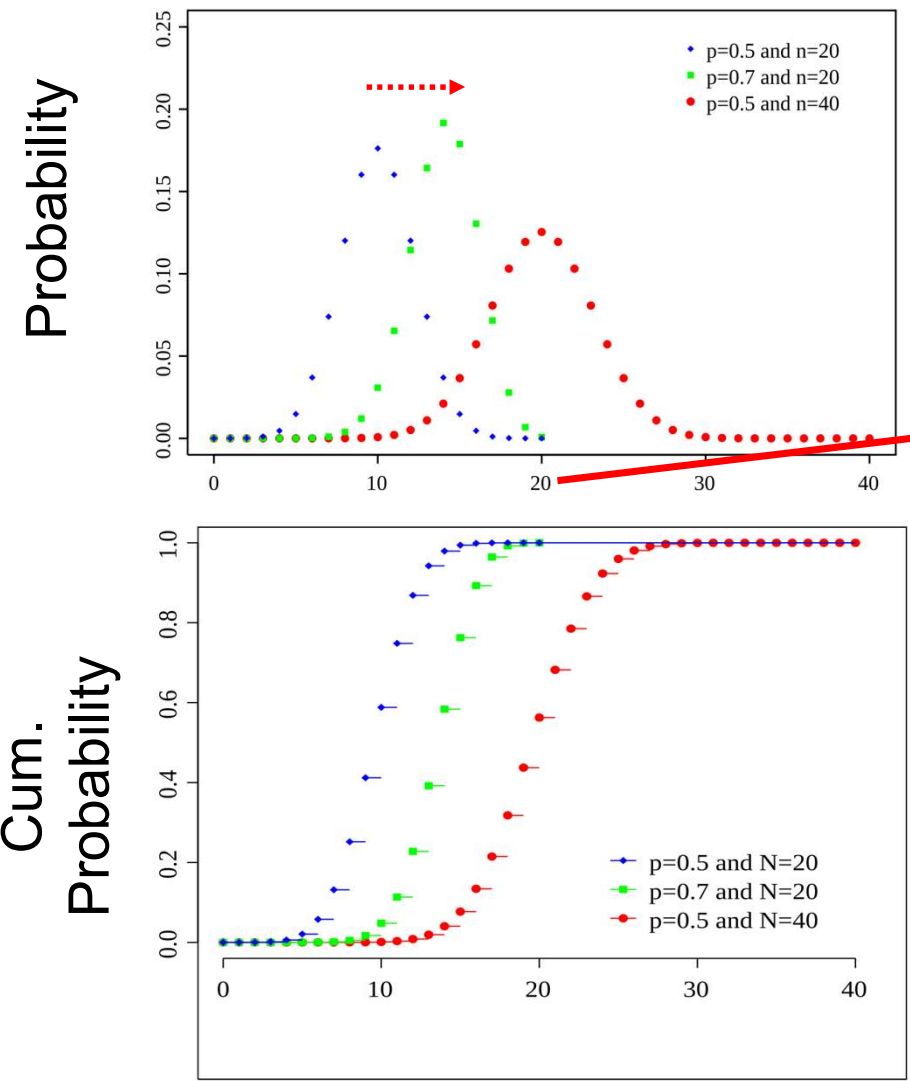
Dice: 1/6 for each number

10 times

How many times will “1” show



Binomial Distribution



Higher $p \rightarrow$ higher expected value

chance for x-hit

Binomial Distribution

*risk SNP: $n \rightarrow$ significant overlap ?

... ATATTGCGGCATGCGCGAT ...

Detectable genomic region: N bp

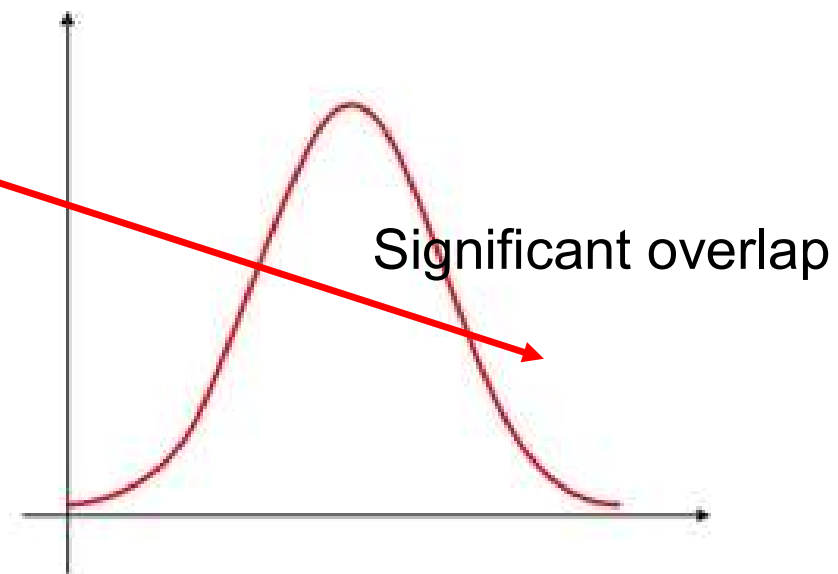
Annotated region: k bp

(ex: functional annotation, enhancer, methylation)

$\rightarrow$ binomial distribution: $B(n,p)$ vs my overlap

n : # risk SNP

p : k/N



Multinomial Distribution

- Binomial: 2
- Multinomial: $m > 2$

$$P(x_1, x_2, \dots, x_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$$
$$= \binom{n}{x_1, x_2, \dots, x_m} p_1^{x_1} p_2^{x_2} \dots p_m^{x_m}.$$

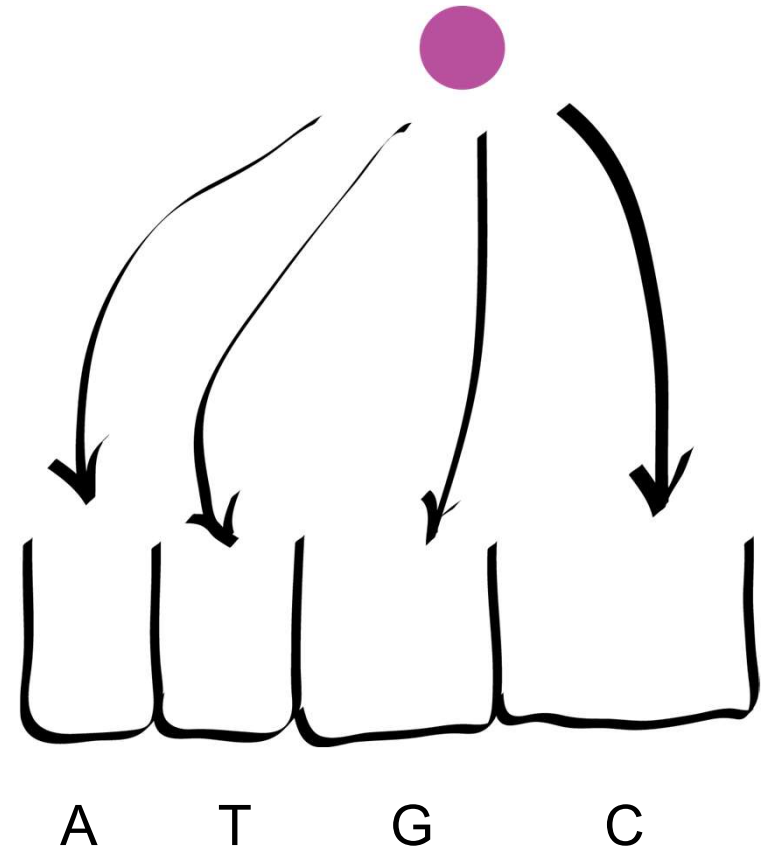
*Probability

A: 0.2, B: 0.3, C: 0.5

P for A=1, B=2, C=3 events during 6 votes

Events combination

$$\Pr(A=1, B=2, C=3) = \frac{6!}{1!2!3!} (0.2^1) (0.3^2) (0.5^3) = 0.135$$



Poisson Distribution

*Trying to describe a distribution for how many occurrence of a certain event per “unit time”

*Assumption

- Constant mean: same occurrence per time / space (or **constant p** for the event)
+ p is very small
- Independency: independent for each period / interval
- Only one occurrence: No multiple occurrence

Ex)

- the number of cars passed during 1 hr
- the number of errors in one page
- how many people visited the website in one day

Poisson Distribution

$$f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

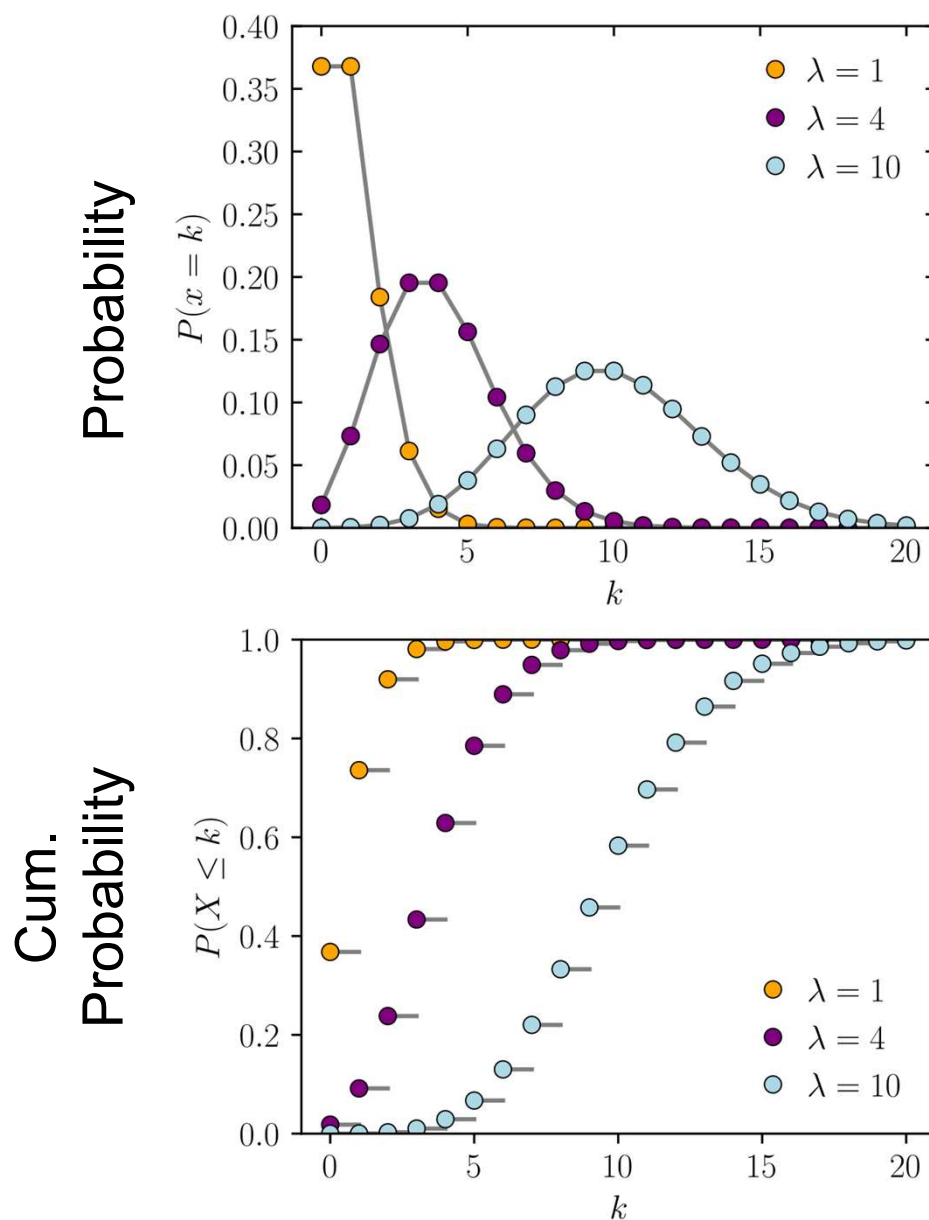
λ : expected mean

k : number of the event

$f(k; \lambda)$: probability for k events

Expected mean: λ

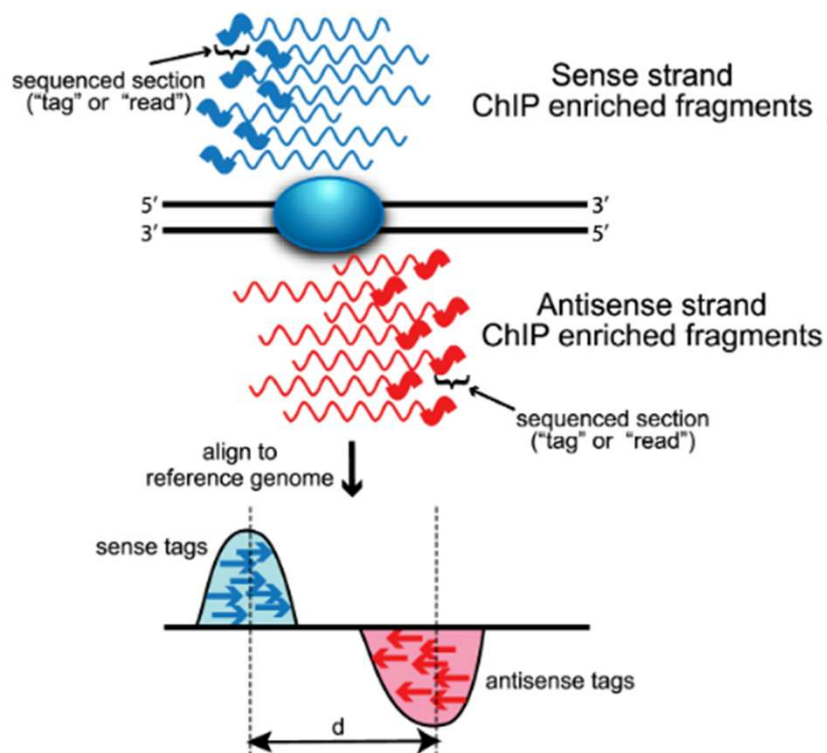
Variance: λ



Poisson Distribution

*MACS2 peak calling

- Number of read for a given (genomic) window

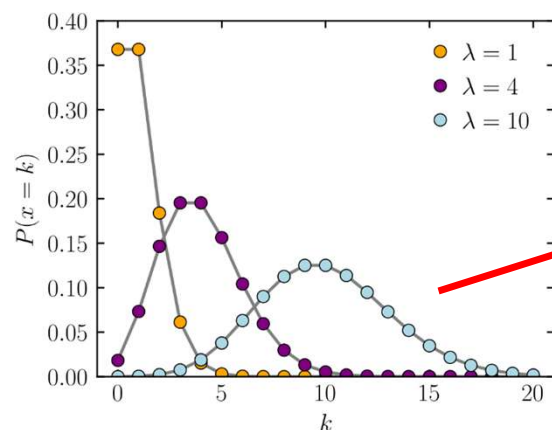


λ = mean = expected value = variance

$$\lambda = \frac{\text{total number of events (k)}}{\text{number of units (n) in the data}}$$

$$= \frac{\text{Read length (nt)} * \text{Total read number}}{\text{Effective genome length (nt)}}$$

Effective genome: mappable genome size



Significant peak

Normal Distribution

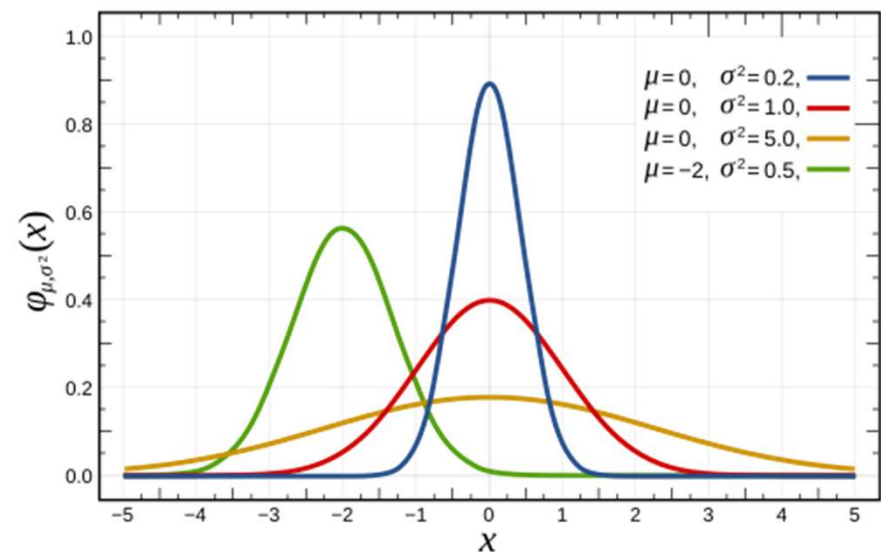
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean: μ

Standard deviation: σ

Symmetric by mean

→ Nature might be under normal distribution
(due to noise ...)



Gamma Distribution

*Probability of the duration for k-times event

$$f(x; k, \theta) = x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)} \text{ for } x > 0$$

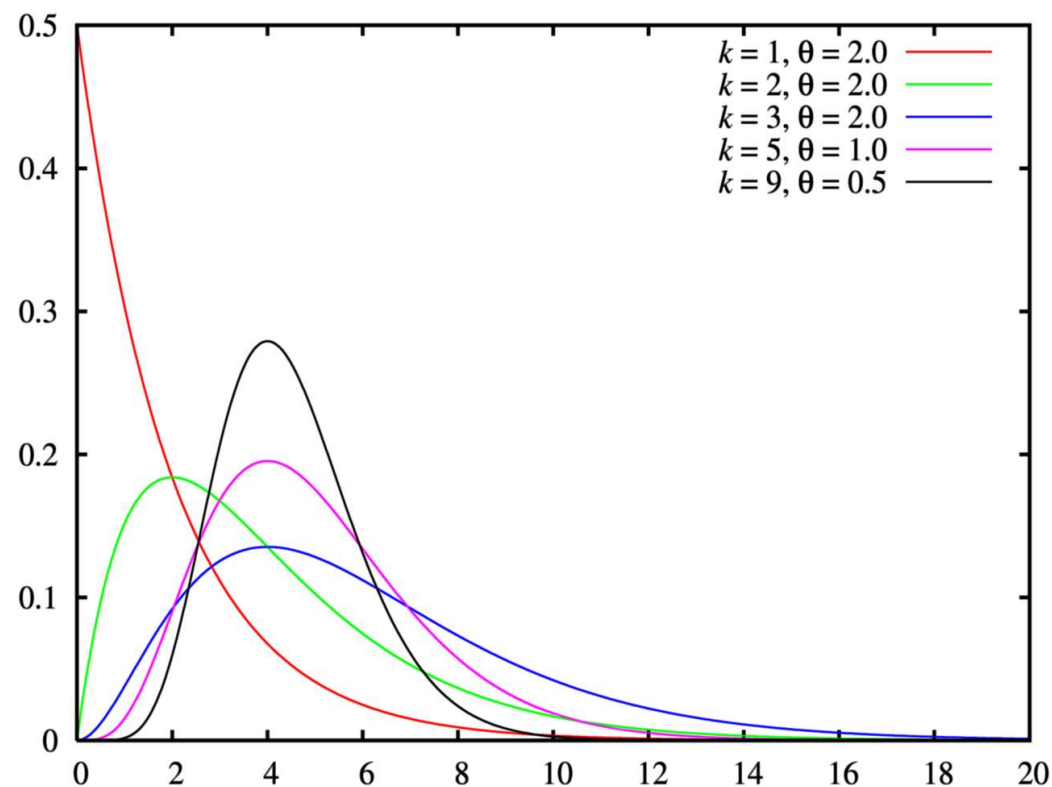
k: k-times event $> 0 \rightarrow$ sharp shape

θ : $> 0 \rightarrow$ spread shape

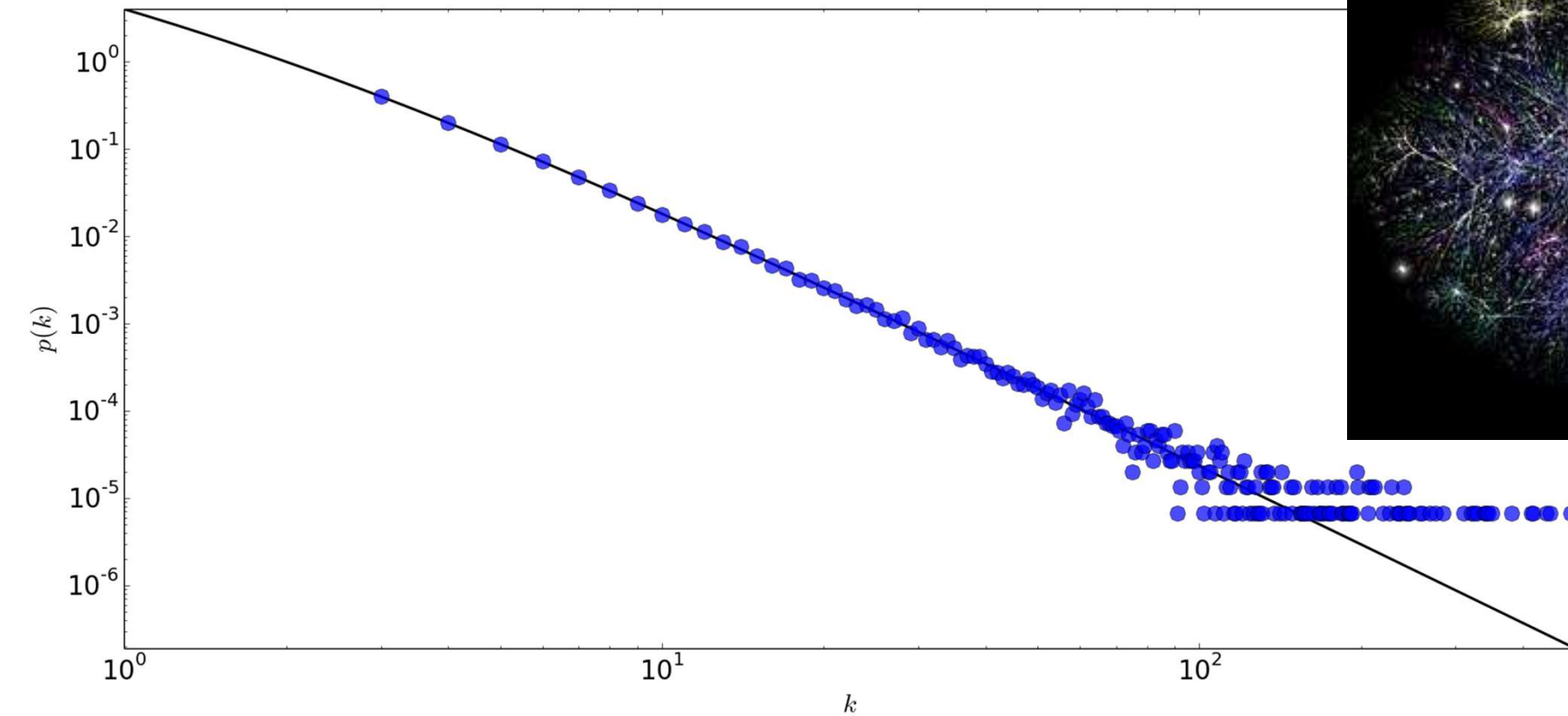
Mean: $k \theta$

Variance: $k \theta^2$

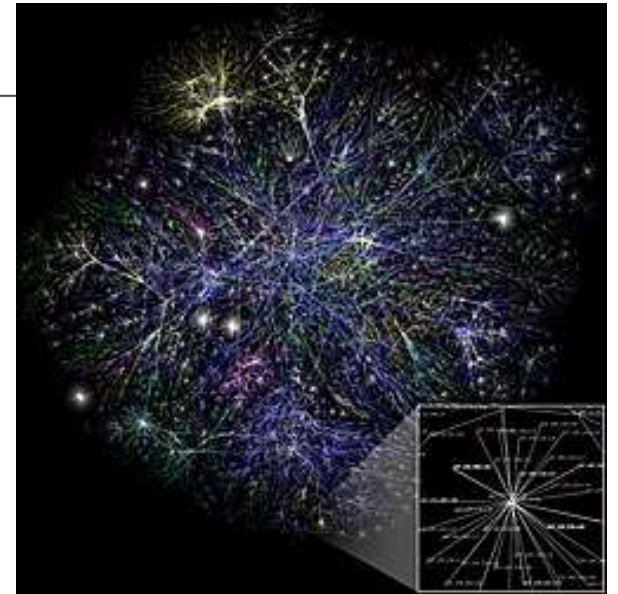
*explains when it is skewed



Scale-free network ($\sim$ gamma function)



Node degree in "gene-network"

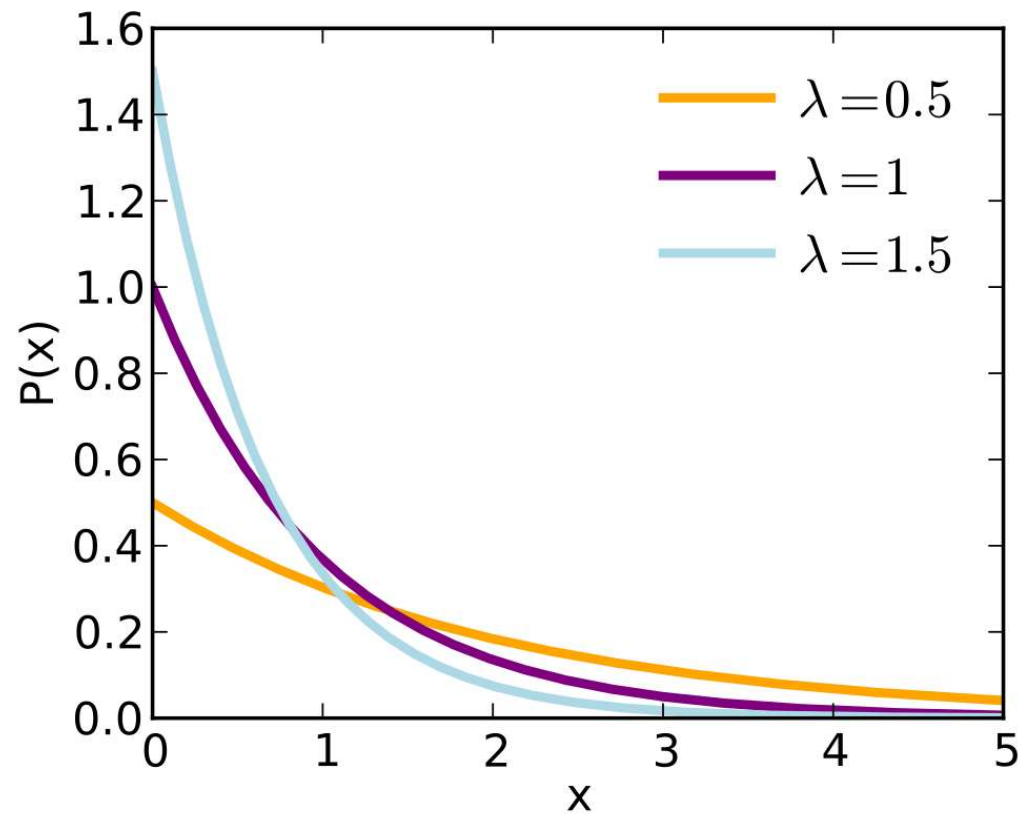


Exponential Distribution

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$t_{1/2} = \frac{\ln(2)}{\lambda}$$

-just for half-life $\rightarrow$ constant
 $\sim \lambda$



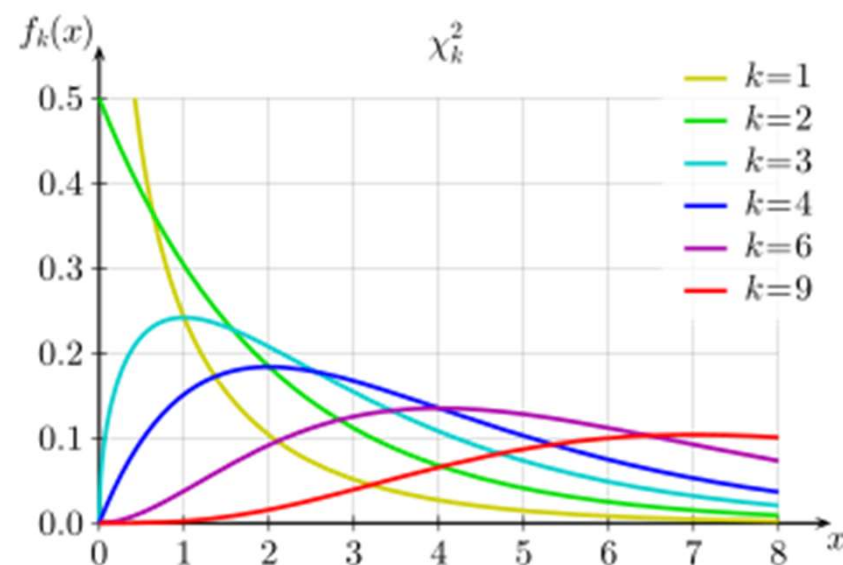
χ^2 distribution

If $Z_1, \dots, Z_k$ are independent, standard normal random variables, then the sum of their squares,

$$X = \sum_{i=1}^k Z_i^2,$$

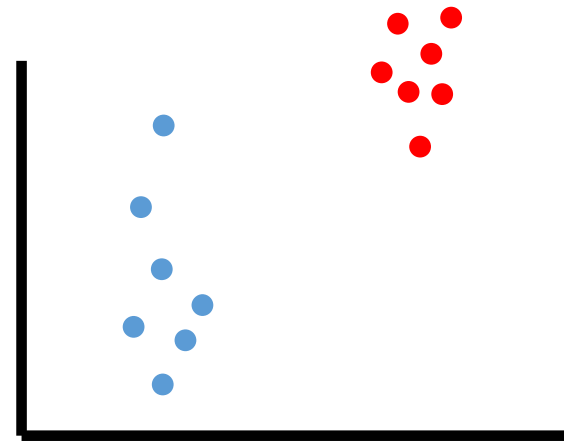
is distributed according to the chi-squared distribution with k degrees of freedom. This is usually denoted as

$$X \sim \chi^2(k) \text{ or } X \sim \chi_k^2.$$



Hypothesis testing

Hypothesis testing



Is it high?

Is it significant ??

T test

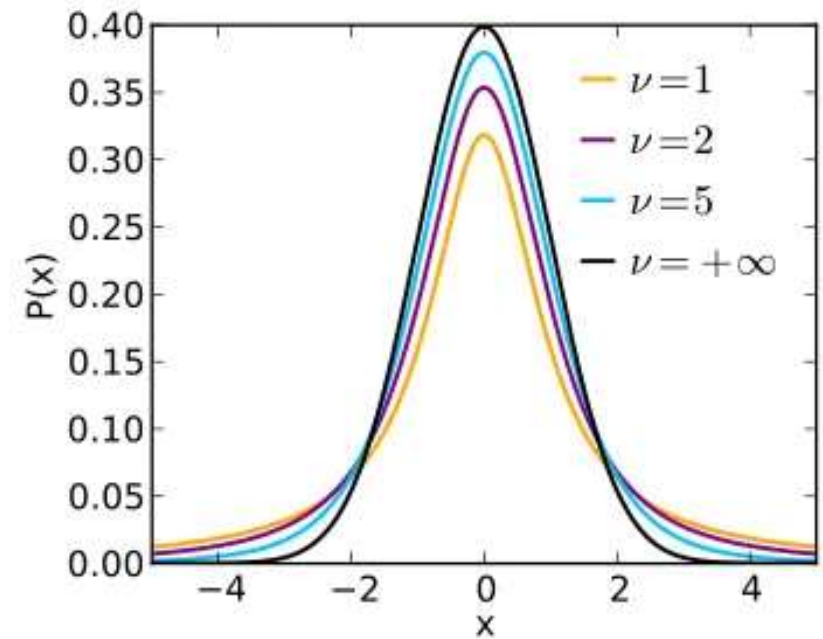
$$T = \frac{Z}{\sqrt{U/v}} \sim t_v$$

T distribution

Z: follow Normal(0,1)

U: follow X^2

V: degree of freedom

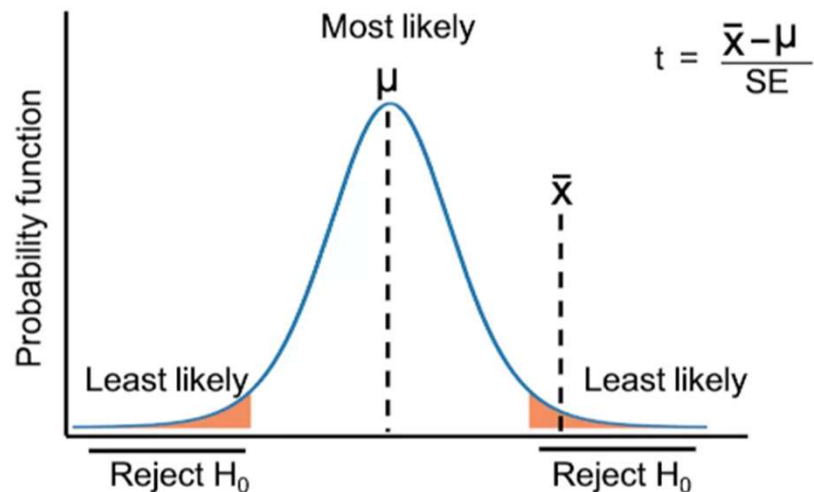


T test

- Is it different ??
- H₀: null hypothesis → it is the same

T-test: follow normal distribution (parametric), independent
T-score $\sim \Delta \text{mean} / \text{s.d.}$

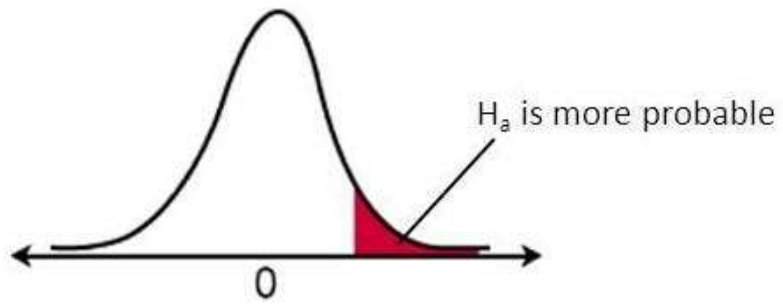
- one-sample test: is it different from the “distribution”?



$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

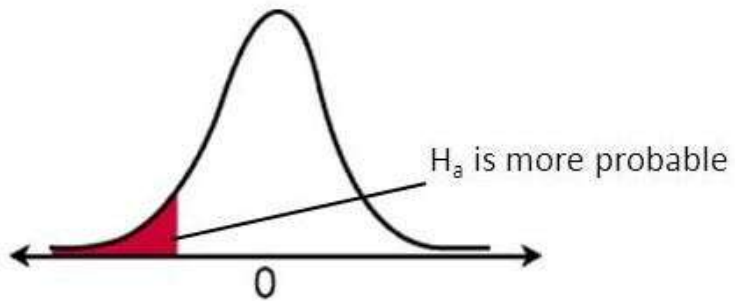
- Sample mean
- estimated s.d. of the population

T test



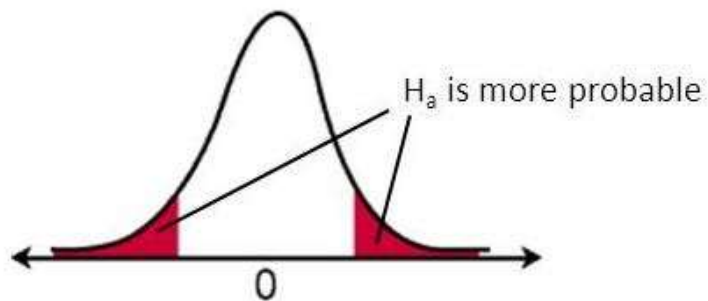
Right-tail test

$$H_a: \mu > \text{value}$$



Left-tail test

$$H_a: \mu < \text{value}$$

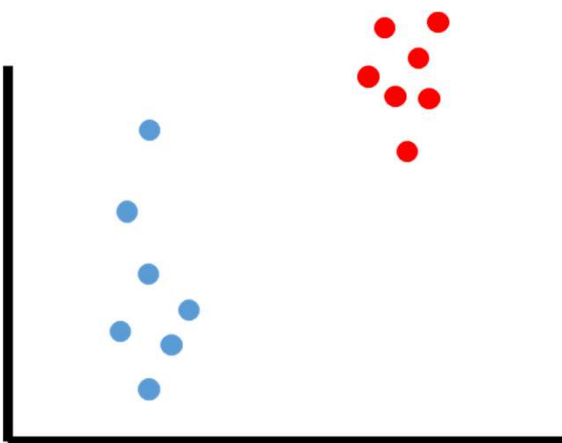
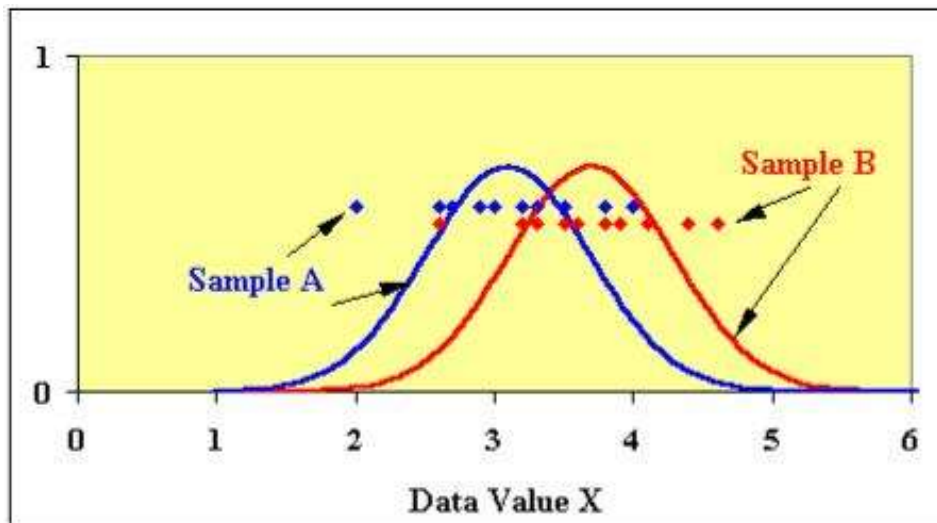


Two-tail test

$$H_a: \mu \neq \text{value}$$

T test

-two-sample test: Are two distributions different
(unpaired, paired)



Is it high?

$$\bar{X}_A = \frac{\sum_{i=1}^{n_1} X_{A,i}}{n_1}, \quad S_A = \sqrt{\frac{\sum_{i=1}^{n_1} (X_{A,i} - \bar{X}_A)^2}{n_1 - 1}}$$

Two-Sample T-Test

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{X}_1$ = observed mean of 1st sample

$\bar{X}_2$ = observed mean of 2nd sample

s_1 = standard deviation of 1st sample

s_2 = standard deviation of 2nd sample

n_1 = sample size of 1st sample

n_2 = sample size of 2nd sample

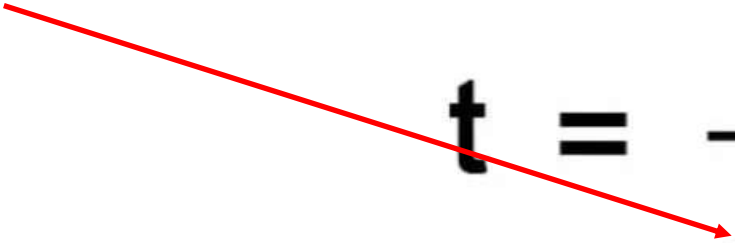
T test

*two-sample test:

Same variance: student's t test

Different variance: welch's t test

Two-Sample T-Test


$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{X}_1$ = observed mean of 1st sample

$\bar{X}_2$ = observed mean of 2nd sample

s_1 = standard deviation of 1st sample

s_2 = standard deviation of 2nd sample

n_1 = sample size of 1st sample

n_2 = sample size of 2nd sample

T test

-two-sample test: Are two distributions different
(unpaired, **paired**)

$$t = \frac{\bar{x}_d - \mu_0}{s_d / \sqrt{n}}$$

-d: difference

Mean of difference & s.d. of difference

Wilcoxon rank-sum test

When the data does not follow (or guarantee) the normal distribution (non-parametric)

- Two-sample test (paired, unpaired)
- one,two-tailed

→ Rank based
($>20 \rightarrow$ U distribution follows Norm)

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

with

R_1, R_2 being the sums of the ranks in groups 1 and 2, after ranking all samples from both groups such that the smallest value obtains rank 1 and the largest rank $n_1 + n_2$.^[7]

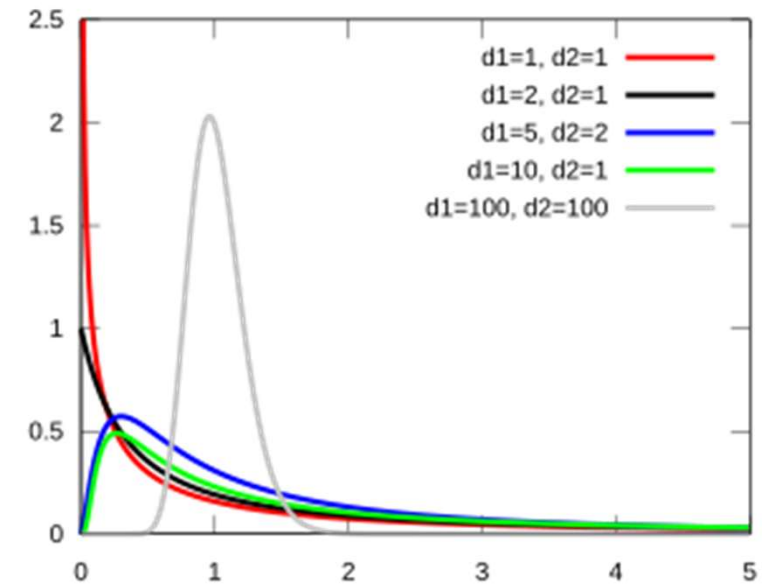
n_1 : $X_1 \sim X_n$, n_2 : $Y_1 \sim Y_n$ (number of sample $\rightarrow$ highest rank)

ANOVA (Analysis of variance)

-F distribution

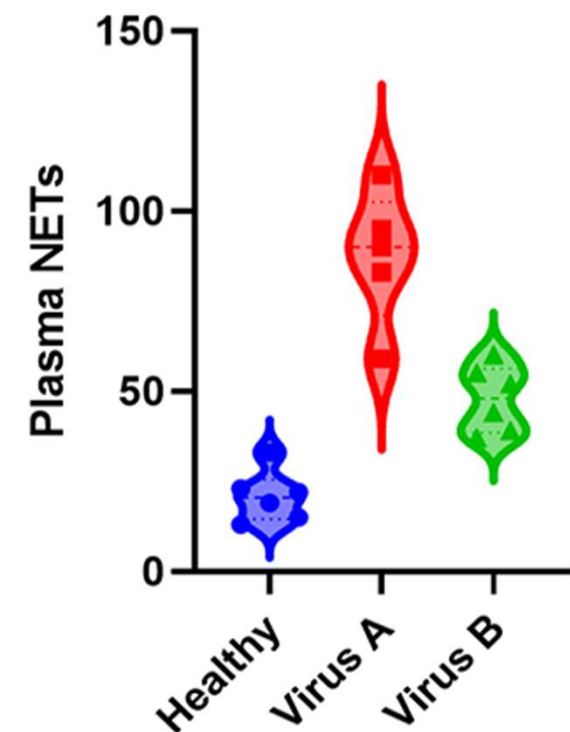
$$X = \frac{U_1/d_1}{U_2/d_2}$$

where U_1 and U_2 are independent random variables with chi-square distributions with respective degrees of freedom d_1 and d_2 .

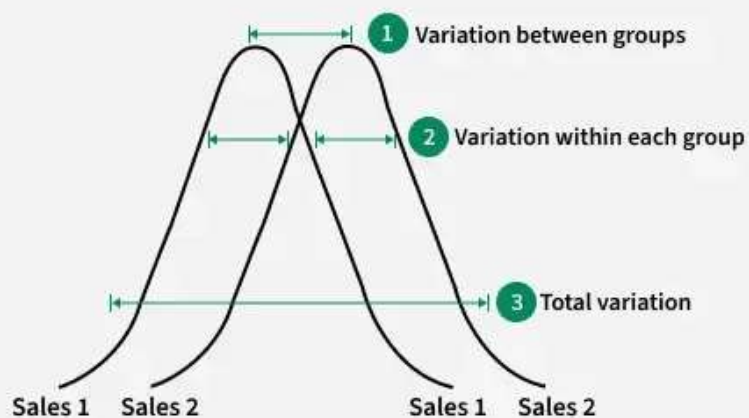


ANOVA

Viral Infection NET Levels



ANOVA testing



Sales 1	Sales 2
150	170
150	162
157	177
145	192
130	184
170	169
165	155

Two sample groups of sales data

- More than 2 groups (follow normal distribution)
- Compare by variance (rather than the mean value)
- Σ Variance within group vs Σ Variance across groups

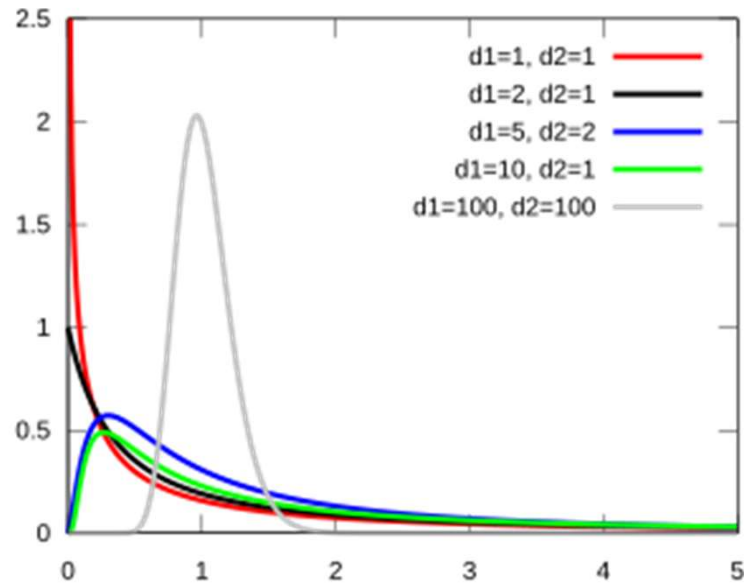
$$F = \frac{MS_b}{MS_w}$$

* Σ Variance across groups
Diff btw total mean

ANOVA

$$F = \frac{MS_b}{MS_w}$$

-F distribution
→ P value



ANOVA

-One-way ANOVA

Factors

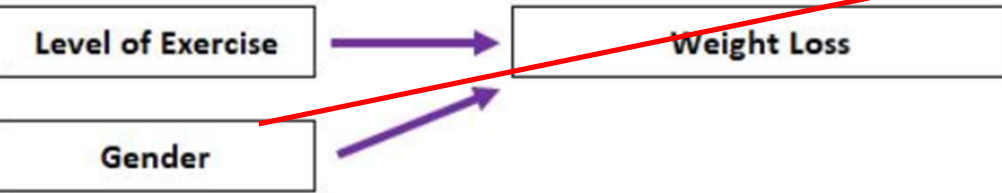
Response Variable



-two-way ANOVA

Factors

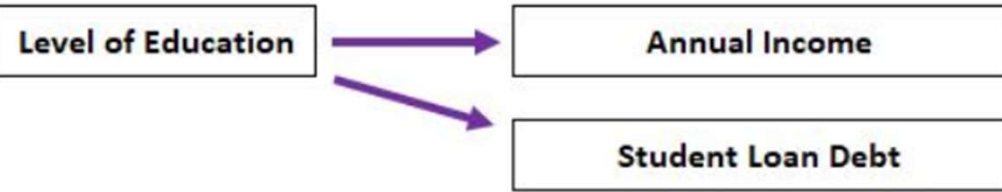
Response Variable



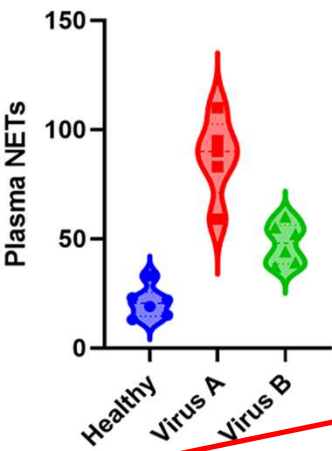
MANOVA (multivariate anova)

Factors

Response Variables



Viral Infection NET Levels

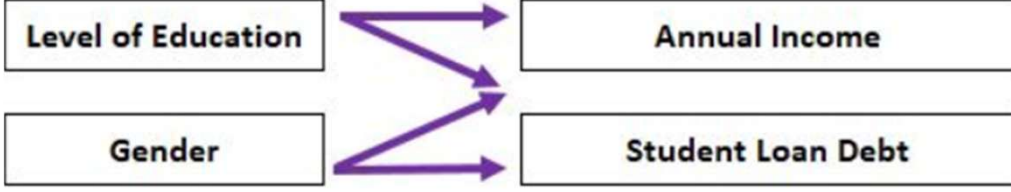


+ Gender

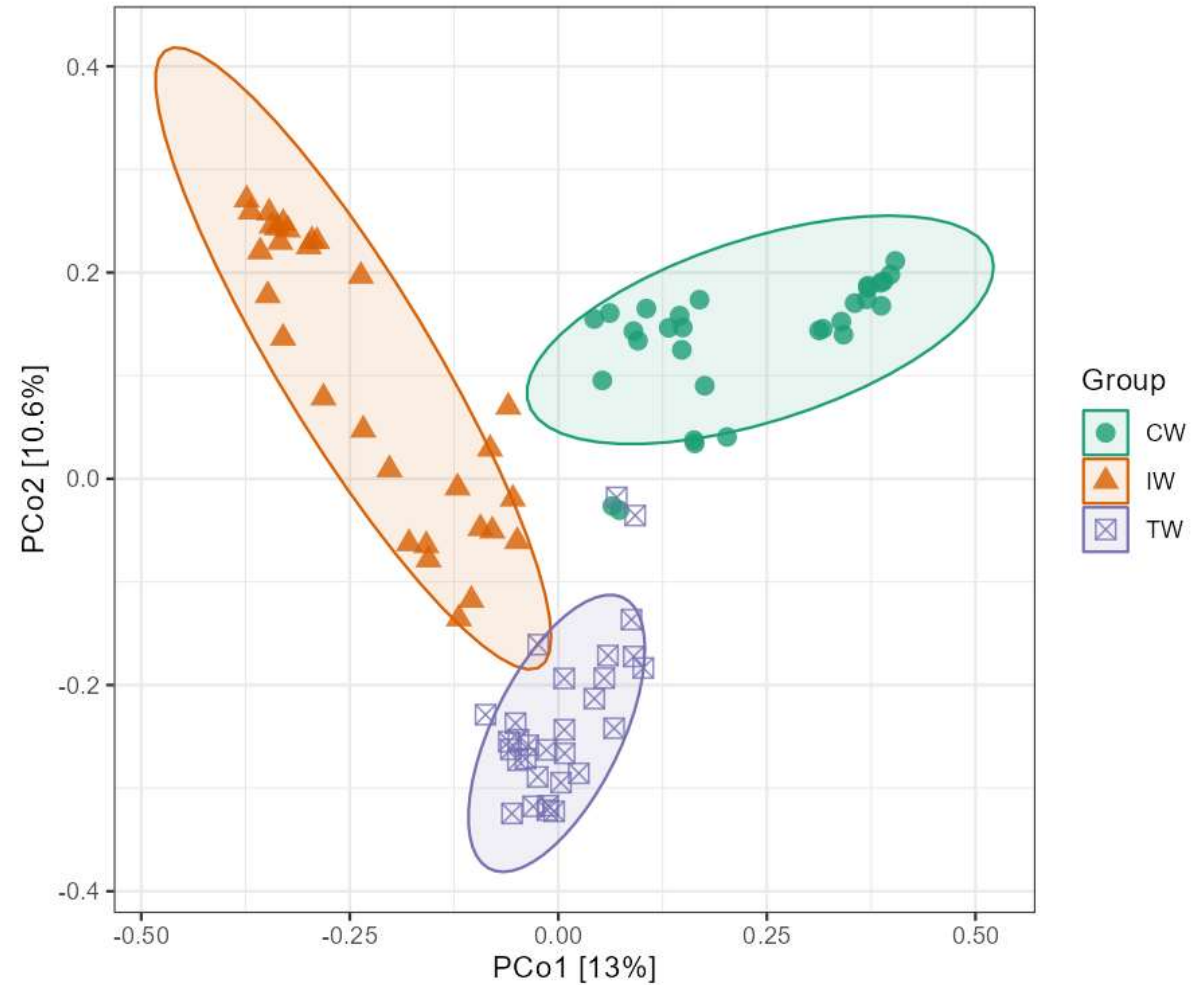
MANOVA (two-way)

Factors

Response Variables



MANOVA



Metagenome

Whether metagenomes from 3 groups are overlapping or not

Different group → different abundance of each taxa

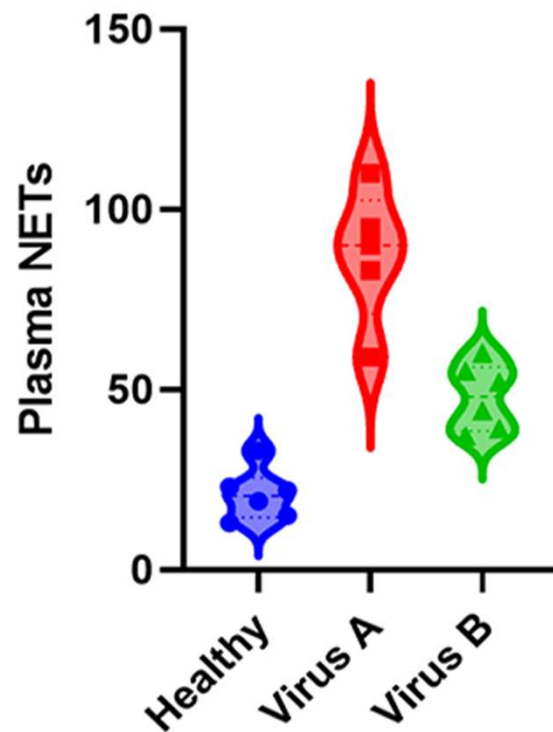
Post-hoc T test

After ANOVA is significant → which groups are different?

→ Pairwise t-test (healthy vs A, healthy vs B, A vs B)

→ Multiple hypothesis correction

Viral Infection NET Levels



Kruskal-wallis test

-I want to use ANOVA

→ non-parametric (normal distribution X)

→ Rank-based

(Then, post-hoc should be Wilcoxon rank-sum test)

Fisher's exact test

	Class I	Class II	Row Total
Blue	a	b	$a + b$
Red	c	d	$c + d$
Column Total	$a + c$	$b + d$	$a + b + c + d (=n)$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

- Whether two groups have a bias
- Compare the expected ratio vs real ratio

*probability

$$\frac{\binom{a+c}{a} \binom{b+d}{b} (a+b)!(c+d)!}{n!} = \frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{n}{a+b}}$$

Fisher's exact test

	Men	Women	Row Total
Studying	0	10	10
Non-studying	12	2	14
Column Total	12	12	24

→ Contingency table

For this table (with extremely unequal studying proportions) the probability is $p = \frac{\binom{10}{0} \binom{14}{12}}{\binom{24}{12}} \approx 0.000033652$.

-Men : Women → 12 : 12

Studying: 0 : 10 or 12 : 2 → deviated from 12 : 12

Two-tail: whether it is associated

One-tail: enriched or depleted

Ex) drug-treatment → effective or defective

Fisher's exact test

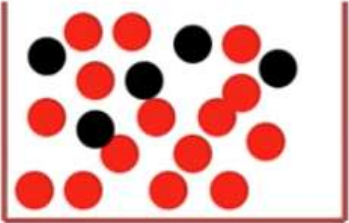
- ❖ **Hypergeometric test** (also known as **Fisher's exact test**)
 - **Null hypothesis:** Observed list is a random sample from population.
 - **Alternative hypothesis:** More black genes than expected in my list.

*My genelist
→ Which pathway (geneset) is enriched?

2x2 contingency table for Fisher's Exact Test

Gene list	In gene list	Not in gene list	
	In pathway	Not in pathway	
● RRP6	x = 4	496	m = 500
● MRD1	k - x = 1	4499	t - m = 4500
● RRP7	k = 5	4995	t = 5000
● RRP43			
● RRP42			

$$P(X = x > q) = \sum_{x=q}^m \frac{\binom{m}{x} \binom{t-m}{k-x}}{\binom{t}{k}}$$

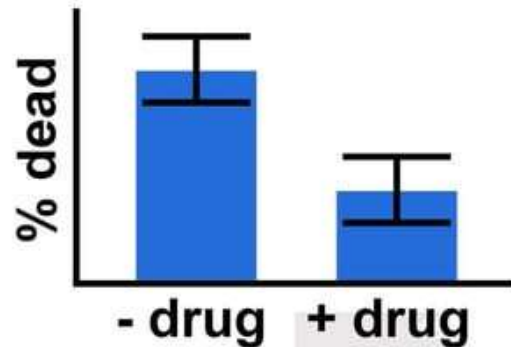


Background population:
500 black genes,
4500 red genes

Statistical Significance Tests

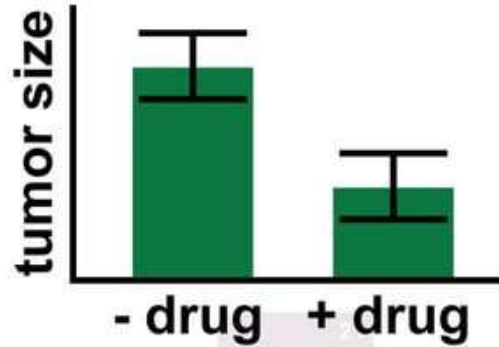
Fisher's Exact Test

x-axis: 2 categories (- or + drug)
y-axis: 2 categories (dead or alive)



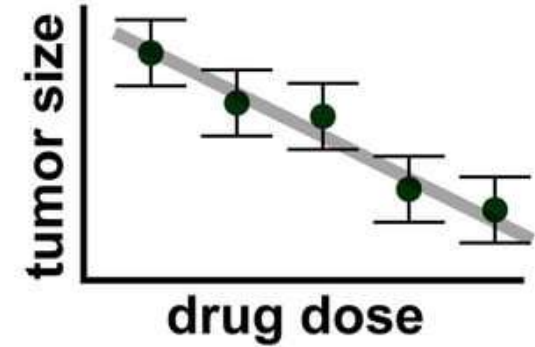
Students T-Test

x-axis: 2 categories (- or + drug)
y-axis: continuous (tumor size)



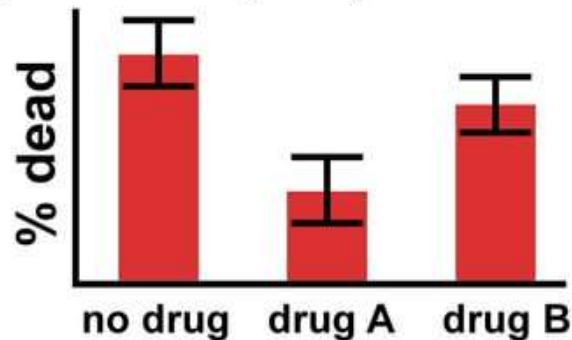
Regression Analysis

x-axis: continuous (drug dose)
y-axis: continuous (tumor size)



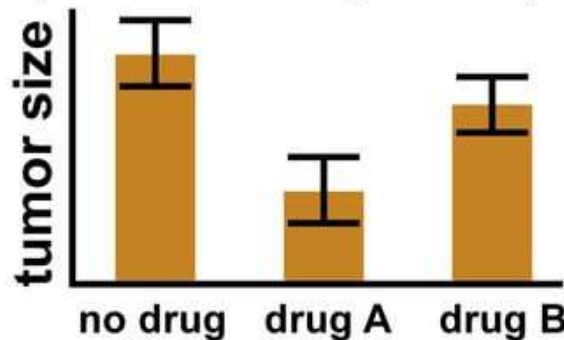
Chi-Square Test

x-axis: >2 categories (drug A, B, etc.)
y-axis: >2 categories (dead or alive)



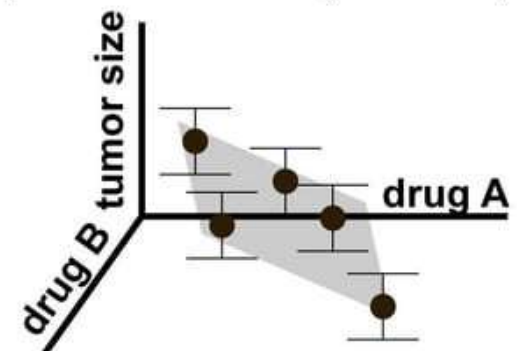
ANOVA

x-axis: >2 categories (drug A, B, etc.)
y-axis: continuous (tumor size)



Multiple Regression

x/z-axis: >2 continuous (drug dose)
y-axis: continuous (tumor size)



*Degree of freedom
Number of sample without decisive sample

평균 알면 n-1 만 샘플 정보 알아도 나머지 값은 자동으로 결정됨

핵심 정리	
상황	자유도
표본 평균을 계산한 뒤 분산 계산	$n - 1$
한 표본 t 검정	$n - 1$
두 집단 평균 비교(t 검정)	$n_1 + n_2 - 2$
카이제곱 검정	범주 수 - 1
회귀분석	관측치 수 - 추정 모수 수