



3000788 Intro to Comp Molec Biol

Lecture 2: Probability and statistics

August 21, 2023



Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Probability is the basis of statistics



- **P-value** = probability of observing the same or more extreme result, given that the null hypothesis is true
 - Probability of seeing >2-fold up-regulation of gene A in drug treated patient by chance, given that the drug does not affect gene A
- **Likelihood ratio** = $\frac{P(\text{observed data} \mid \text{model 1})}{P(\text{observed data} \mid \text{model 2})}$
 - If LR is high, reject model 2. If LR is low, reject model 1
 - $\frac{P(\text{observed monkey pox genome diversity} \mid \text{mutation rate}=0.3)}{P(\text{observed monkey pox genome diversity} \mid \text{mutation rate}=0.01)}$

Probability is about what can happen



- In a human genome with 20,000 genes, there are 518 kinases. What is the probability that a randomly selected gene is a kinase?
- Gene A has two alleles, A and *a*. The frequency of A in Thai population is 0.8. What is the probability that your genotype is AA? What about Aa?
- If the rate of death due to pancreatic cancer is 15% per year, what is the probability that a pancreatic cancer patient survives for at least 5 years?

Probability is about what can happen



- In the human genome with 3 billion base pairs, what is the probability of observing the pattern TAATTA by chance?
- Given a 50-bp DNA sequencing read, what is the expected number of locations on a 3 billion base pairs genome that this read will match to by chance?

More advanced examples



- In a human genome with 20,000 genes, there are 518 kinases. From a comparison of gene expression between control and drug-treated cells, there are 300 differentially expressed genes (DEGs). What is the expected number of kinases among these DEGs?
- Gene A has two alleles, A and *a*. The frequency of A in Thai population is 0.8, and the allele *aa* is embryonic lethal. What is the probability that your genotype is AA? What about A*a*?

Permutation and combination



- There are $N! = N(N-1)(N-2)\dots\times 1$ ways to permute N objects (order them)
 - $O_1 O_2 \dots O_N$
 - There are N choices for the first position
 - There are $N-1$ choices for the second position, anything but the first object
 - And so on
- There are $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ ways to select k objects from a pool of N objects
 - Do you know how we get this expression?

Permutation and combination



- There are $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ ways to select k objects from a pool of N objects
 - Order all N objects and choose the first k objects
 - There are $N!$ ways. But some of them result in the same selection
 - $[O_1 O_2 \dots O_k] O_{k+1} O_{k+2} \dots O_N$
 - $[O_2 O_1 \dots O_k] O_{k+2} O_{k+1} \dots O_N$
 - $[O_k O_2 \dots O_1] O_N O_{k+2} \dots O_{k+1}$
 - There are $k!$ ways to order the first k objects and get the same selection
 - There are $(N-k)!$ ways to order the last $N-k$ objects and get the same selection
- In probability, division usually means you overcount and then compensate by dividing by the number of duplicates

Discrete probability distributions



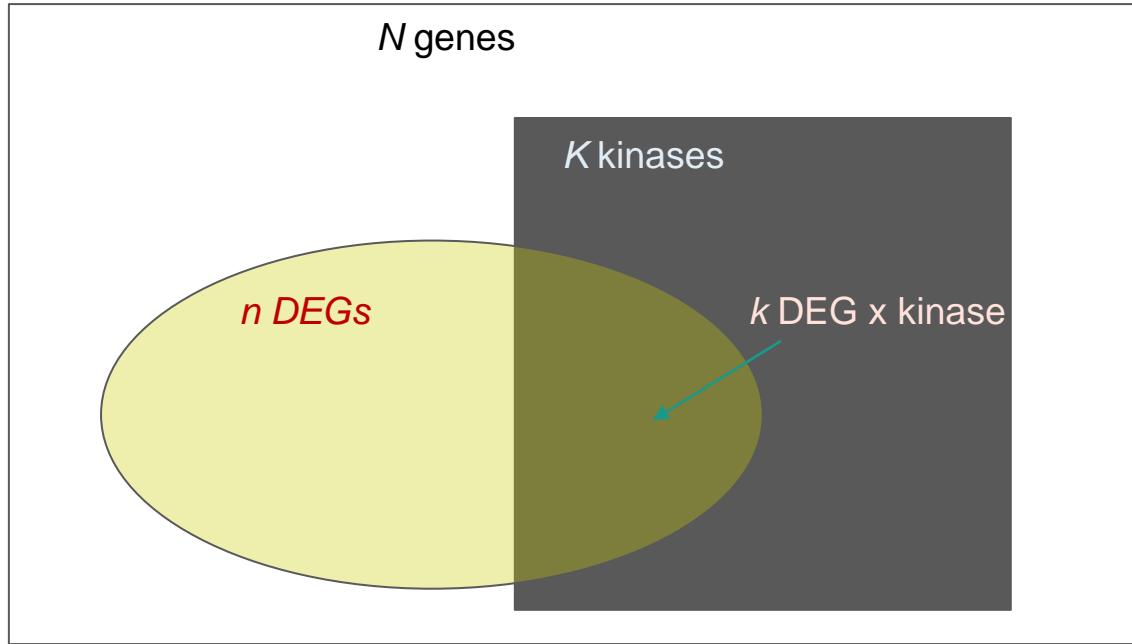
- **Binomial distribution**
 - Independent trial with two mutually exclusive outcomes: Win and Lose
 - $P(\text{Win}) = p$, $P(\text{Lose}) = 1 - p$
 - Probability of getting k Wins out of N trials = $\binom{N}{k} p^k (1 - p)^{N-k}$
 - Expected number of Win = pN
- **Poisson distribution**
 - Independent events occurring with expected count = λ
 - Probability of observing exactly k events = $\frac{\lambda^k e^{-\lambda}}{k!}$
 - Expected number of events = λ

Derivation of Poisson distribution



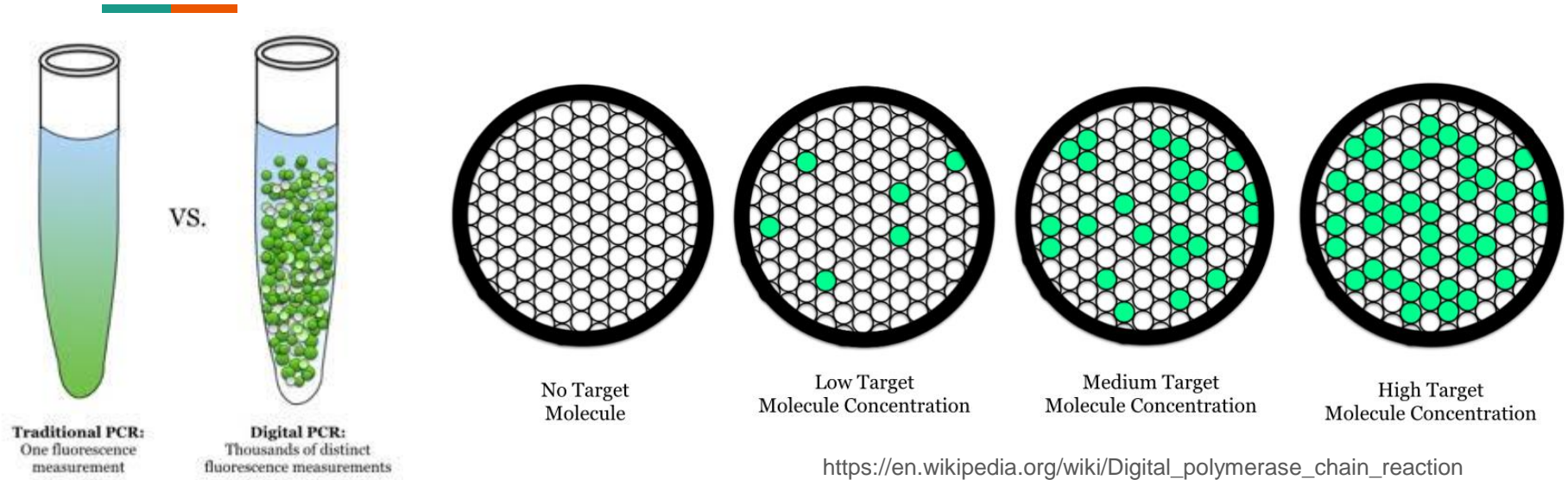
- Binomial model
 - N independent discrete trials with probability of success p
- Poisson = Binomial with $N \rightarrow \text{infinity}$
 - Events have the probability to occur continuously
 - Probability of success $p = \frac{\lambda}{N}$
 - $$\lim_{N \rightarrow \infty} \binom{N}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k} = \lim_{N \rightarrow \infty} \frac{N!}{N^k (N-k)!} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}\right)^N \left(1 - \frac{\lambda}{N}\right)^{-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

Hypergeometric distribution



- Why does the probability = $\frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$?

Example 1: Digital Droplet PCR



- M total DNA molecules: M_a of allele a and M_A of allele A
- N droplets, k are positive for allele a
- Likelihood ratio test of $\frac{P(k \text{ positive droplets} \mid \text{patient genotype } AA)}{P(k \text{ positive droplets} \mid \text{patient genotype } Aa)}$

Digital Droplet PCR

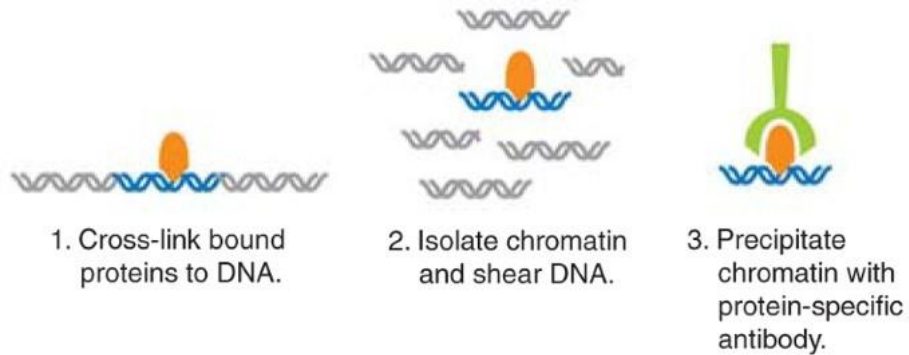


- Fluorescent signal is a “success” of PCR probe binding to a molecule of allele a
- M total DNA molecules: M_a of allele a and M_A of allele A
- N droplets, k are positive for allele a

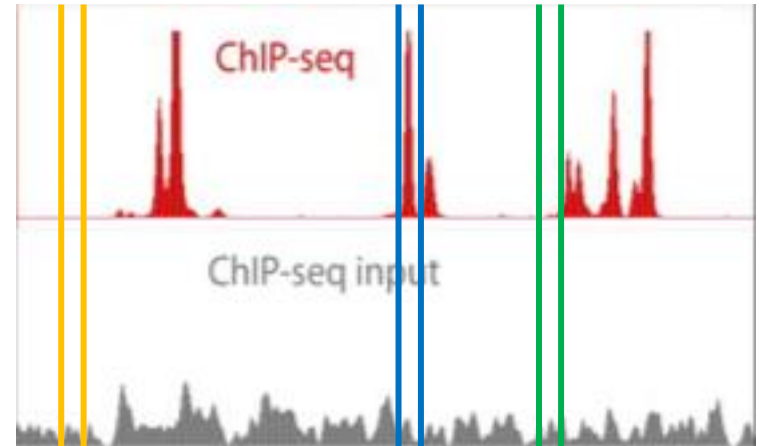
- Q1: Which distribution captures the number of “success” in each droplet?
- Q2: What is the expected number of “successes” in each droplet?
- Q3: What is the probability of observing a positive droplet (≥ 1 success)?

- Q4: Which distribution captures the number of positive droplets?
- Q5: What is the probability of observing k positive droplets in this experiment?

Example 2: ChIP-seq peak assessment



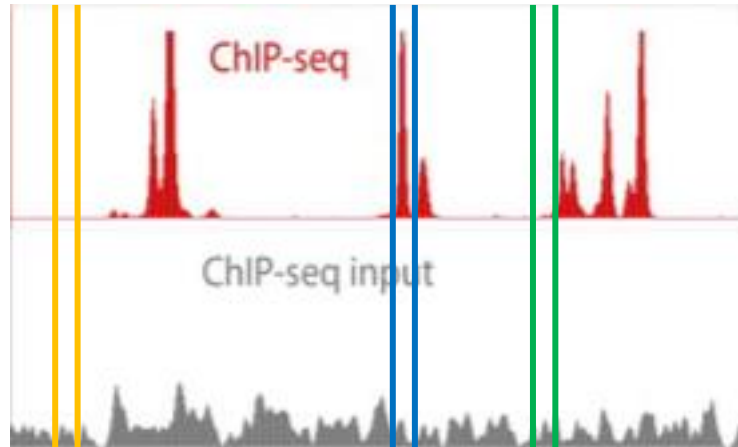
Shah, A. Nature Methods 6:i-ii (2009)



Park et al. Nat Rev Genet 10:669-680 (2009)

- ChIP-seq peak = enrichment of DNA read at certain genomic position
 - Imply protein binding or histone modification
- Does the peak arise from bias in DNA sequencing?

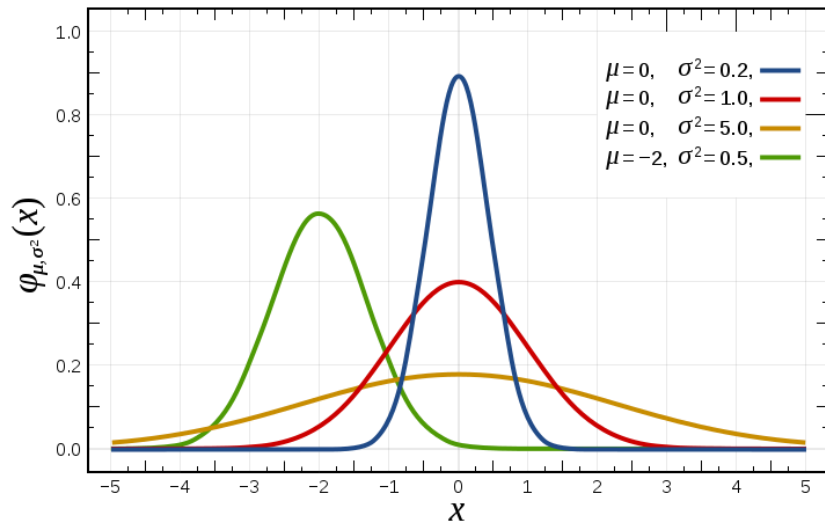
ChIP-seq peak assessment



Park et al. Nat Rev Genet 10:669-680 (2009)

- Estimate expected number λ_g of reads at a genomic position g from a control
 - Also called ChIP-seq input, without immunoprecipitation
- Probability of observing k reads at a position g = $\text{Poisson}(\lambda_g)$

Continuous distributions

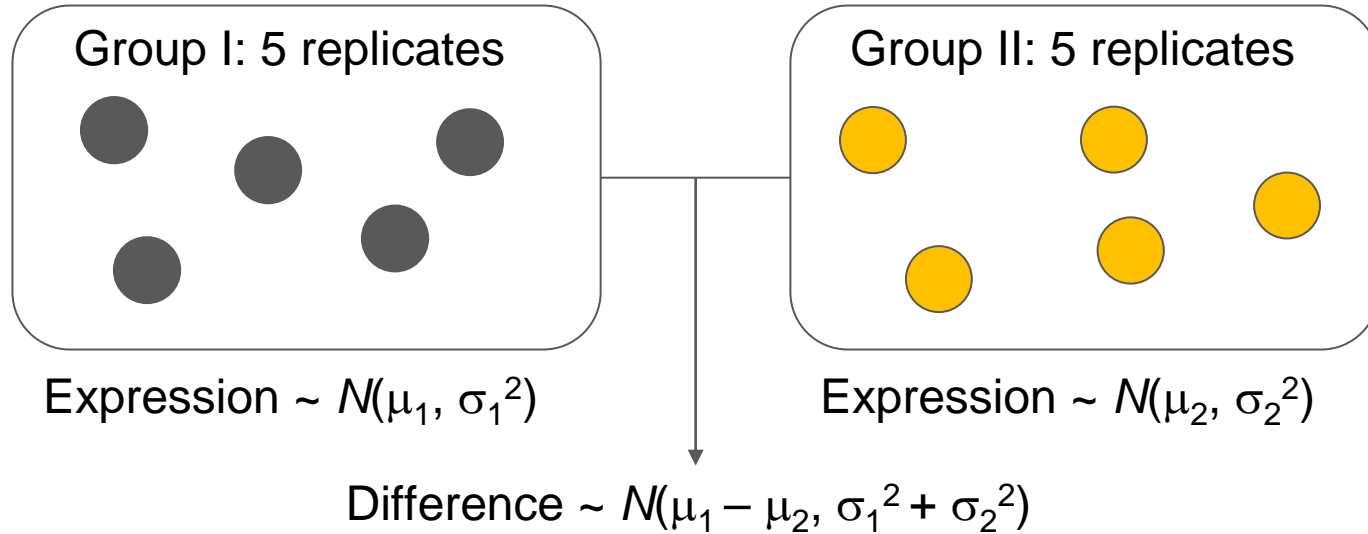


Images from https://en.wikipedia.org/wiki/Normal_distribution

- Normal or Gaussian distribution
- Defined by location (mean) and spread (variance)

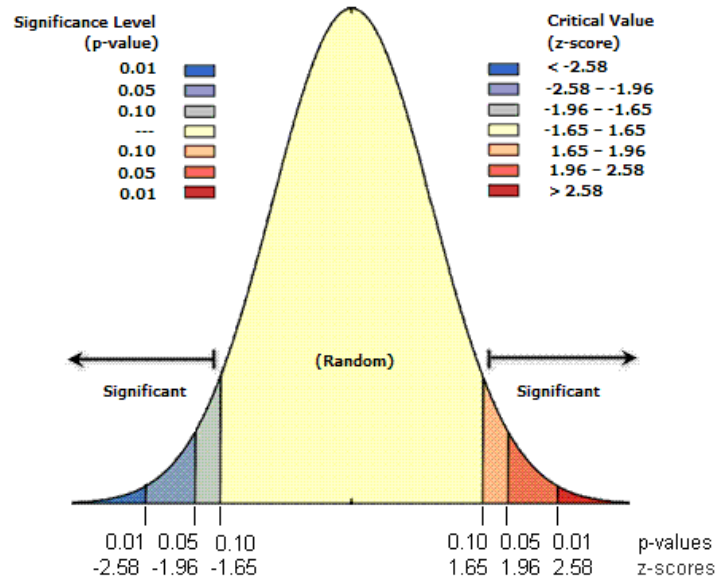
$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

Properties of normal distribution



- $N(\mu, \sigma^2) + c = N(\mu + c, \sigma^2)$
- $N(\mu, \sigma^2) \times c = N(c\mu, (c\sigma)^2)$

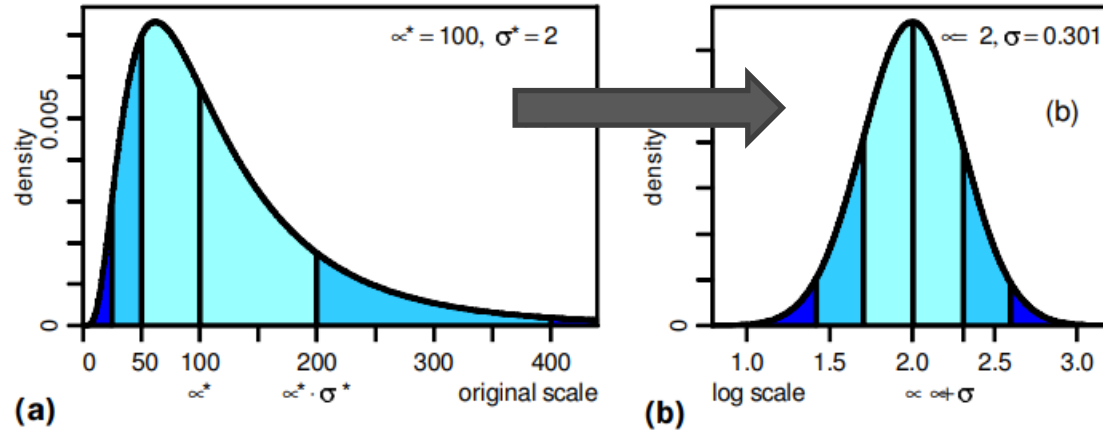
Standard normal distribution and Z-score



<https://desktop.arcgis.com/en/arcmap/10.4/tools/spatial-statistics-toolbox/what-is-a-z-score-what-is-a-p-value.htm>

- $$Z\text{-score} = \frac{x - \text{mean}}{\text{standard deviation}}$$
- If the data came from a normal distribution, $N(\mu, \sigma^2)$, Z-score is the transformation to standard normal distribution, $N(0, 1)$
- Z-score can be converted to p-value

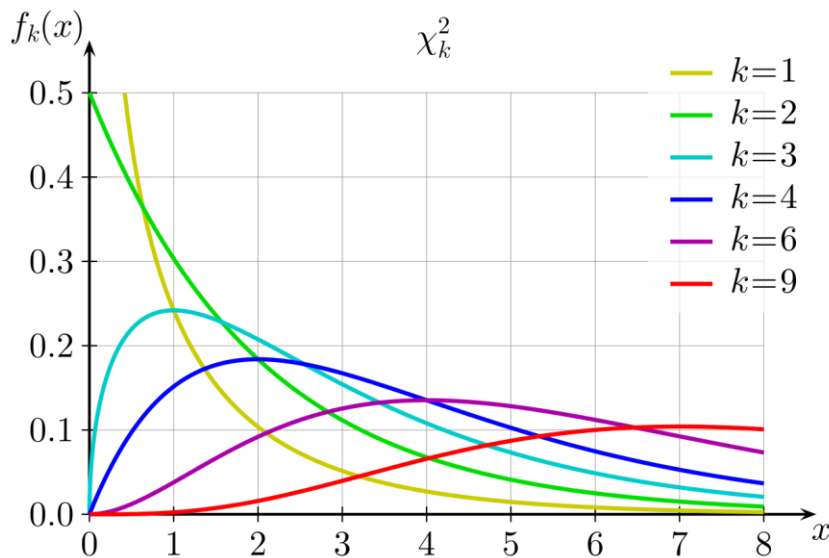
Log-normal distribution



Limpert, Stahel, and Abbt. BioScience 2001.

- Some data, especially intensity, are not normally distributed but their log are
 - Fluorescence intensity
 - Gene expression from microarray
 - Peptide abundance from mass spectrometry

Chi-squared distribution



Images from https://en.wikipedia.org/wiki/Chi-squared_distribution

- Equal to $\sum_{i=1}^k Z_i^2$ where Z_i are standard normal (this may seem unnatural but is useful in many statistical tests)

Statistical tests related to Chi-squared



Genotype	<i>A/A</i>	<i>A/a</i>	<i>a/a</i>
Expected frequency	200	120	70
Observed frequency	90	210	80

- $\sum_i \frac{(O_i - E_i)^2}{E_i}$ follows Chi-squared distribution
- Test of nested models, such as multiple mutation rate models
 - Likelihood ratios = $\frac{P(\text{data} \mid \text{complex model})}{P(\text{data} \mid \text{simple model})}$
 - $-2 \times \text{Log Likelihood ratio}$ follows Chi-squared distribution

Statistics explains things that already happened



- Gene A has two alleles, A and a . A study of 1,000 Thai individuals found 700 with genotype AA, 200 with genotype Aa, and 100 with genotype aa . What is the estimated allele frequency of a ?
- In a study of 5 pancreatic cancer patients, they survived for 1, 5, 3, 4, and 5 years. What is the estimated yearly survival rate?

Maximum likelihood principle

- Likelihood = $P(\text{data} \mid \text{model})$
- Find **model** that maximize likelihood
- Gene A has two alleles, A and a . A study of 1,000 Thai individuals found 700 with genotype AA, 200 with genotype Aa, and 100 with genotype aa . What is the estimated allele frequency of a ?
 - Let's set the allele frequencies $f_A = p$ and $f_a = 1 - p$
 - $P(AA) = p^2$, $P(Aa) = 2p(1 - p)$, and $P(aa) = (1 - p)^2$
 - Likelihood = $P(AA)^{700} P(Aa)^{200} P(aa)^{100} = p^{1400} 2^{200} p^{200} (1 - p)^{200} (1 - p)^{200}$
 $= 2^{200} p^{1600} (1 - p)^{400}$
 - Which p maximize the likelihood?
 - Solve the equation $\frac{d\text{Likelihood}}{dp} = 0 \rightarrow p_{\text{MLE}} = 0.8$

Maximum likelihood principle



- In a study of 5 pancreatic cancer patients, they passed away after 1, 5, 3, 4, and 5 years, respectively. What is the estimated yearly survival rate?
 - Let's set the yearly survival rate = r
 - $P(\text{survive exactly } k \text{ years}) = r^k(1 - r)$
 - $P(\text{data} \mid r) = r^1(1 - r) r^5(1 - r) r^3(1 - r) r^4(1 - r) r^5(1 - r) = r^{18}(1 - r)^5$
 - Which r maximize the likelihood?
 - Solve the equation $\frac{d\text{Likelihood}}{dr} = 0 \rightarrow r_{\text{MLE}} = 18/23$

Likelihood ratio test



- Likelihood ratio = $\frac{P(\text{data} \mid \text{model 1})}{P(\text{data} \mid \text{model 2})}$
 - If LR is high, reject model 2. If LR is low, reject model 1
- Example:
 - Test for viral spreading rate: $\frac{P(\text{COVID-19 infection} \mid \text{spreading rate} > 1.5)}{P(\text{COVID-19 infection} \mid \text{spreading rate} < 1.5)}$
 - Test for impact of treatment: $\frac{P(\text{gene expression} \mid \text{control and treatment differ})}{P(\text{gene expression} \mid \text{all samples are the same})}$
- This is theoretically the most powerful test (Neyman-Pearson Lemma)

Nested model testing



- **Simple model:** Omicron and Delta have the same spreading rate
 - One parameter
- **Complex model:** Omicron and Delta have different spreading rates
 - Two parameters
- Complex model always achieve higher likelihood
 - Find two spreading rates that fit the data better than a single rate
- But is the additional complexity worth it?
 - $\frac{P(\text{data} \mid \text{complex model})}{P(\text{data} \mid \text{simple model})}$ must be much greater than 1 to reject the simple model
 - Akaike information criterion (AIC): $2 \times \# \text{ parameters} - \log \text{ likelihood}$

Null hypothesis-based testing



- **Alternative hypothesis:** Omicron BA.5 spreads more easily than other strains
- **Null hypothesis:** Omicron BA.5 spreads at the same rate as other strains
- Data = Rise of COVID-19 cases with frequency of BA.5 in population
- Likelihood under alternative hypothesis = $P(\text{Data} \mid \text{spread rates for all strains})$
 - We want to show that this is more likely than null hypothesis
 - But difficult to calculate!
- Likelihood under null hypothesis = $P(\text{Data} \mid \text{same spread rate})$
 - Easier to calculate
 - We will try to show that this is unlikely instead

P-value



- Probability of observing **the same or more extreme result**, given that the null hypothesis is true
- How to quantify **the same or more extreme**?
 - If BA.5 has the same spread rate as other strains, rise in BA.5 frequency shouldn't increase the number of new daily infections
 - **What if we measure the rate of increase in daily infections?**

P-value example



- Before BA.5, a study estimated the rate of daily increase in COVID-19 infections with a normal distribution $N(1.3, 0.01)$
- After BA.5, data show that the rate of daily increase in COVID-19 infections is 1.5
- P-value = $P(\text{daily rate} \geq 1.5 \mid \text{BA.5 has the same spread rate as prior strains})$
= $P(\text{getting value} \geq 1.5 \text{ from } N(1.3, 0.01))$
= P-value of Z-score of 2 = 0.02275
- Reject null hypothesis

P-value caution



- Before BA.5, a study estimated the rate of daily increase in COVID-19 infections with a normal distribution $N(1.3, 0.01)$
- After BA.5, data show that the rate of daily increase in COVID-19 infections is 1.4
- P-value = $P(\text{daily rate} \geq 1.4 \mid \text{BA.5 has the same spread rate as prior strains})$
= $P(\text{getting value} \geq 1.4 \text{ from } N(1.3, 0.01))$
= P-value of Z-score of 1 = 0.158655
- Do we accept null hypothesis?

Test statistics



- A measure, or score, of **the same or more extreme result**
- In one-sample t -test of whether the mean \bar{x} of data $\{x_1, x_2, \dots, x_n\}$ is equal to β , the test statistics is $t = \frac{\bar{x} - \beta}{\frac{SD}{\sqrt{n}}}$
 - Measure how close is \bar{x} to β , subject to the variability of the data
 - Correspond to the null hypothesis that the mean of the data is β
 - **The more t deviates from zero, the more extreme the result**
- P-value = $P(t \geq t_{\text{observed}} \mid \text{the data is normally distributed with mean } \beta)$
 - t follows $N(0, 1)$ by Central Limit Theorem

Test statistics behind popular tests

- Mann-Whitney U test: $U = \sum_{i=1}^n \sum_{j=1}^m S(X_i, Y_j), \quad S(X, Y) = \begin{cases} 1, & \text{if } X > Y, \\ \frac{1}{2}, & \text{if } Y = X, \\ 0, & \text{if } X < Y. \end{cases}$
- Wilcoxon rank-sum test:
 1. Compute $|X_1|, \dots, |X_n|$.
 2. Sort $|X_1|, \dots, |X_n|$, and use this sorted list to assign ranks R_1, \dots, R_n
$$T = \sum_{i=1}^N \text{sgn}(X_i) R_i.$$
- Sign test: Assume that each observation is equally likely to be positive or negative
 - Probability of k positive values out of N observations = $\text{Binomial}(N, k, p = 0.5)$

Impact of null hypothesis choices



Cell	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
C1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
C2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

- Are gene expressions up-regulated in cell C1?
 - **Unpaired** Student's t-test p-value = 0.5687
 - Mann-Whitney U test p-value = 0.6101
 - **Paired** Student's t-test p-value = 0.0137
 - Wilcoxon signed rank test p-value = 0.0156
 - Sign test p-value = 0.00815

Null hypothesis-based testing framework



- Propose null hypothesis
 - The data are normally distributed with mean = β
- Design the test statistic $t = \frac{\bar{x} - \beta}{\frac{SD}{\sqrt{n}}}$
- Derive the distribution of test statistic under the null hypothesis
 - This is where probability knowledge comes in
- Specify the significance level α to reject null hypothesis (e.g., 0.05)
- Calculate p-value: $P(t \geq t_{\text{observed}} \mid \text{null hypothesis})$
- By following this framework, new tests can be created!

Correlation



Cell	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
C1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
C2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

- Correlation of gene expression between C1 and C2 = 0.9746
- How significant is this?
 - We have only one dataset and no information about the underlying distribution
- Null hypothesis
 - Gene expression between C1 and C2 are uncorrelated
 - C2 data can be shuffled and still give the same correlation score

Permutation test



- **Alternative hypothesis:** The **observed property** of the data, such as high correlation, is due to **some structure**, such as the pairing of genes, in the data
- **Null hypothesis:** **That structure** in the data does not contribute to the **property of interest**
- P-value = Probability that **the shuffled** data has the **same or more extreme property** than the original data
- Shuffle data in such a way that **the structure of interest** is disrupted
- Calculate the **property of interest** and compared to the original score

Permutation test



Cell	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
C1	0.701	0.503	0.991	0.827	0.623	0.728	0.596
C2	0.691	0.478	0.905	0.739	0.589	0.719	0.508

Correlation = 0.97



1,000 times

Correlation = 0.24

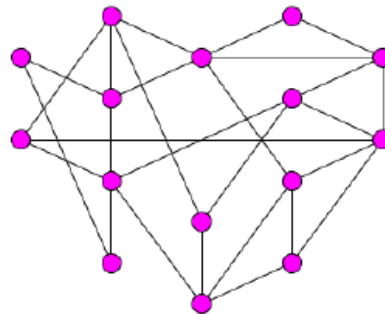
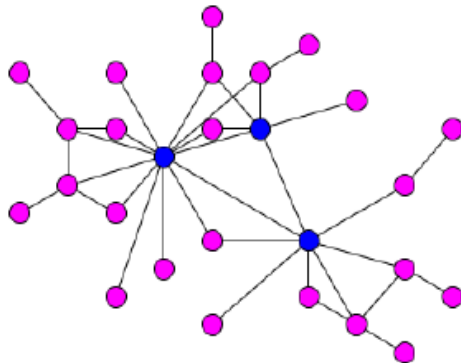
Correlation = -0.30

C2	0.719	0.691	0.739	0.589	0.508	0.905	0.478
----	-------	-------	-------	-------	-------	-------	-------

Correlation = 0.32

Permutation test for network data

Biological network
has hubs that serve
as shortcut between
other nodes



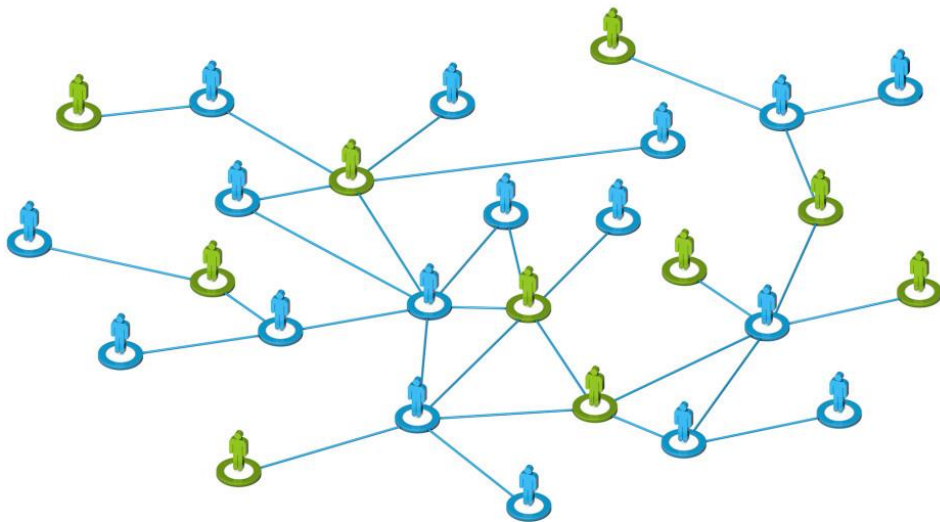
Random network is
not well-connected

Source: Segura-Cabrera *et al.* Analysis of Protein Interaction Networks to Prioritize Drug Targets of Neglected-Diseases Pathogens

- **Null hypothesis:** The **small diameter** of biological network can be achieved by chance in networks with **the same number of nodes and edges**
- **Permutation test:** Generate 1,000 **random networks with the same number of nodes and edges** and **compute the diameter of these networks**

Permutation test for network data

Same-gender FB friendship
occur more easily than
different-gender ones



- **Null hypothesis:** The **high number of same-gender edge** of Facebook friendship network can be achieved by chance in random networks with **the same number of nodes and edges**

Correction for multiple testings



- P-value cutoff of 0.05 means that under the null hypothesis, there is only 5% chance of observing the same or more extreme result
- Applying the same test 1,000 times will result in 50 tests on average with smaller p-value than 0.05 just by chance
 - Differential expression analysis tests thousands of gene at once
- This is unacceptable if a conclusion relies on multiple tests
 - Functional enrichment analysis assumes that all input DEGs are true

Bonferroni method



- Divide the p-value cutoff by the number of test
- Adjusted p-value cutoff = $0.05 / 1000 = 0.00005$
- Applying the same test 1,000 times will result in 0.05 tests on average with smaller p-value than 0.00005 just by chance
- Easy but lose power

False discovery rate (FDR)



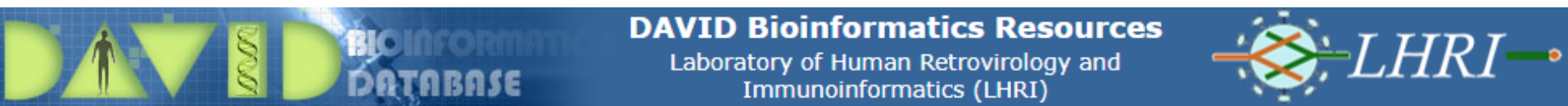
- P-value operates under the null hypothesis
- But in practice, we want to control the number of errors in the output
 - The number of DEGs that were incorrectly proposed
- FDR = Probability of getting a false positive
= # false positive / # all predicted positives
- But FDR involves alternative hypothesis, which is difficult to calculate
- We can control FDR somewhat through p-value!

Benjamini-Hochberg procedure



- Valid under broad assumption (independent tests, etc.)
- Given a series of tests with p-values, p_1, p_2, \dots, p_n
- To control FDR to be within 0.05
 - Sort p-values from low to high, p'_1, p'_2, \dots, p'_n
 - Find largest k such that $p'_k \leq 0.05 \times k / n$
 - For the smallest p-value, this is equivalent to Bonferroni
 - For other p-values, this technique gradually loosens the cutoff
 - Reject null hypothesis for tests corresponding to p'_1, p'_2, \dots, p'_k

Example of functional enrichment report



Functional Annotation Clustering

[Help and Manual](#)

Current Gene List: demolist1

Current Background: Homo sapiens

145 DAVID IDs

☒ Options Classification Stringency

39 Cluster(s)

 [Download File](#)

Annotation Cluster 1		Enrichment Score: 4.64		Count	P_Value	Bonferroni	Benjamini	FDR
<input type="checkbox"/>	GOTERM_CC_DIRECT	extracellular space	RT	38	2.8E-9	6.8E-7	6.8E-7	6.6E-7

Any question?



- See you on August 24th