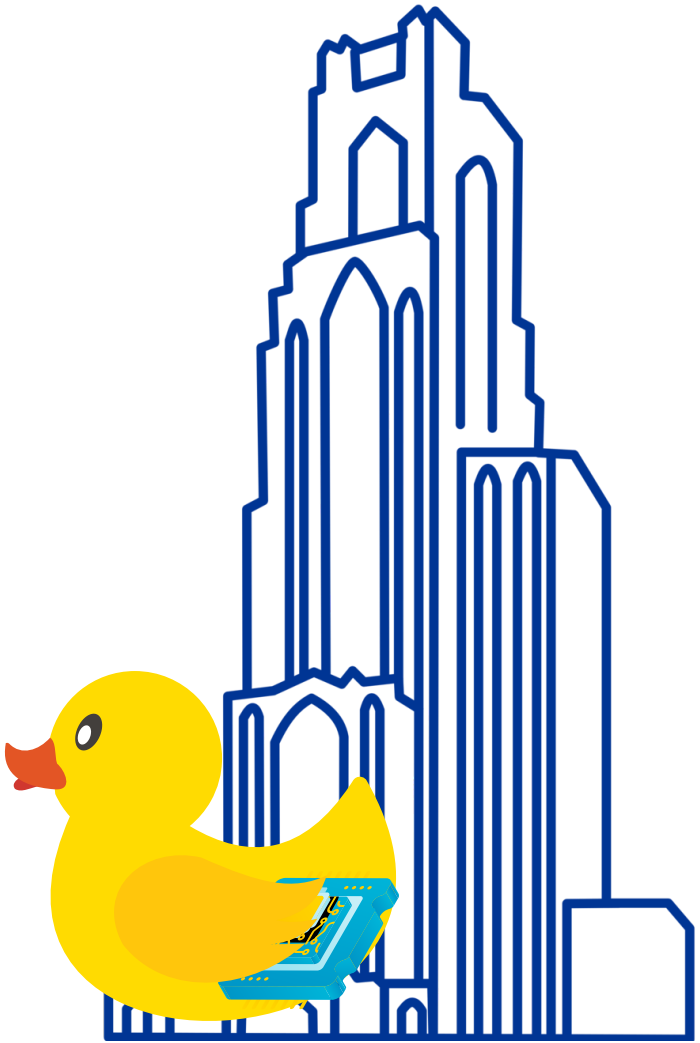


Computational Biology

(BIOSC 1540)

Lecture 08A

Differential gene expression

Foundations

Feb 25, 2025

University of Pittsburgh

Announcements

Assignments

- P02A is due Mar 14
- P02B will be published sometime this week

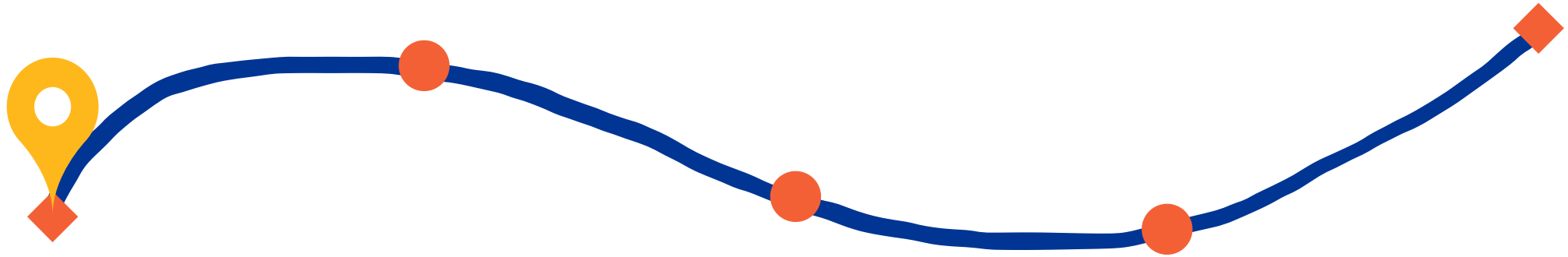
Quizzes

- Quiz 03 is on Mar 18 and will cover L06B to L08B

CBits

- César optional Python recitations are on Fridays from 2 - 3 pm in L1 Clapp Hall
- Please fill out the Canvas discussion for CBit 07

After today, you should have a better understanding of

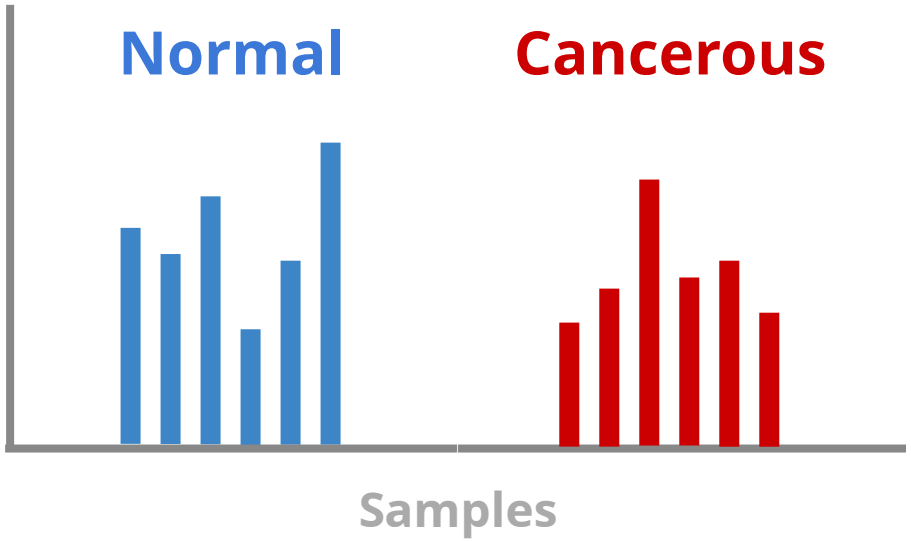


Hypothesis testing for comparing gene expression

Let's remember the big picture: We want to quantify differences in gene expression

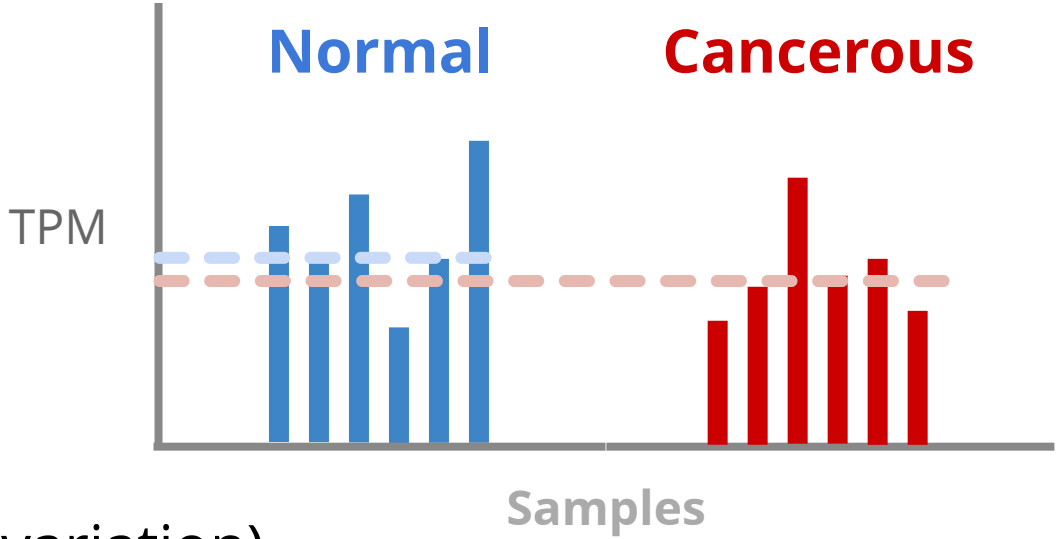
We have been focused on **quantifying gene expression** in quantities like Transcripts Per Million (TPM)

Differential gene expression quantifies changes in gene expression levels between different sample groups or conditions



We cannot rely on simple comparisons when analyzing gene expression

We could technically directly compare means between our different conditions



However, biological data are **inherently noisy**, and observed differences may arise by chance

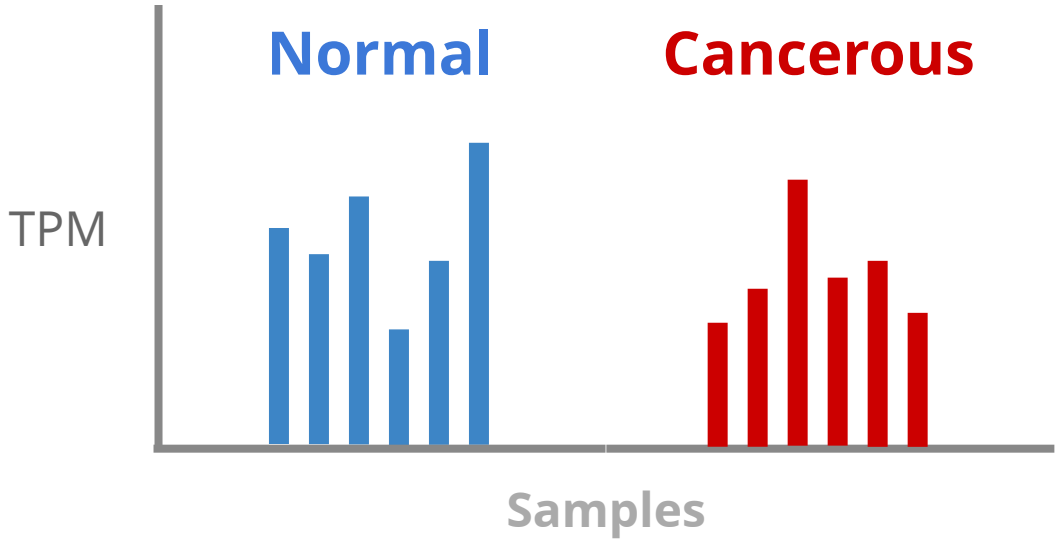
Examples of experimental biases (besides sample variation)

Sequencing depth: Higher depth could appear as higher expression levels simply due to having more data

Batch effects: Processing sampling with different equipment, reagents, times, etc. can show systematic differences

We need approaches that address these sources of variation and noise

Statistical models can account for variability and separate signal from noise



Hypothesis testing between statistical models provides a quantitative way to compare conditions

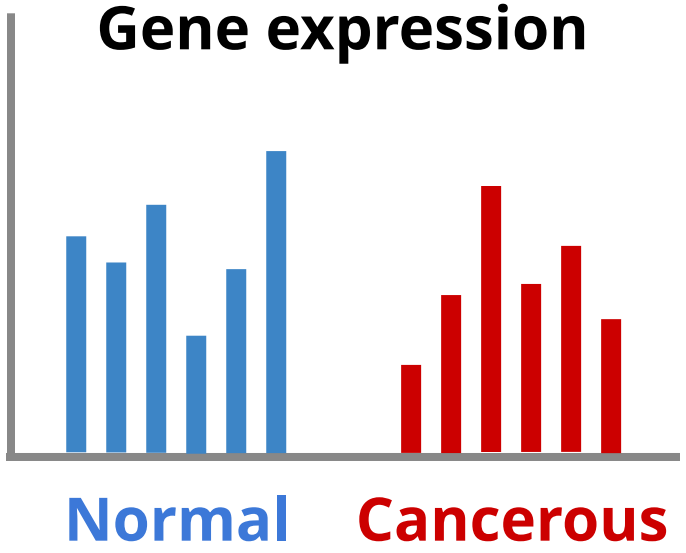
Hypothesis testing in RNA-seq data

After fitting a statistical model, we need to perform **hypothesis testing** to see if the difference in expression between conditions is statistically significant

We have two hypotheses:

Null Hypothesis (H_0): There is **no difference** in gene expression between the two conditions

Alternative Hypothesis (H_1): There is a **significant difference** in gene expression between the conditions



We **reject the null hypothesis** when our statistical test demonstrates that the observed difference, if any, is unlikely to have happened by random chance

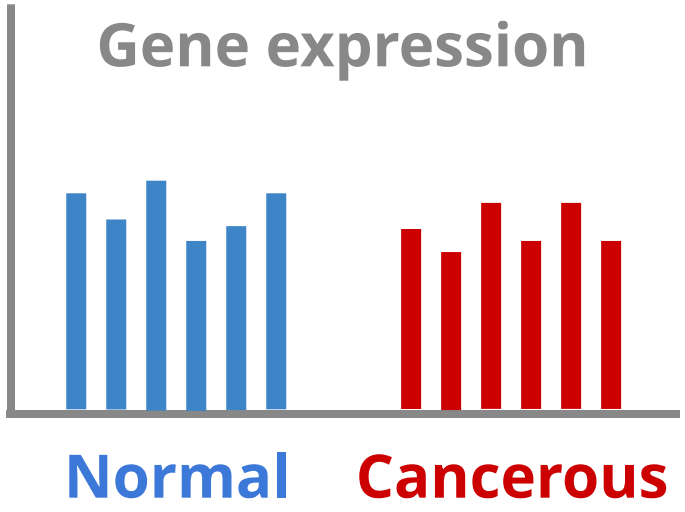
The p-value is the probability of the null hypothesis being true

Probability value (p-value):

What is the probability that any difference is either **(1)** nonexistent or **(2)** due to random chance

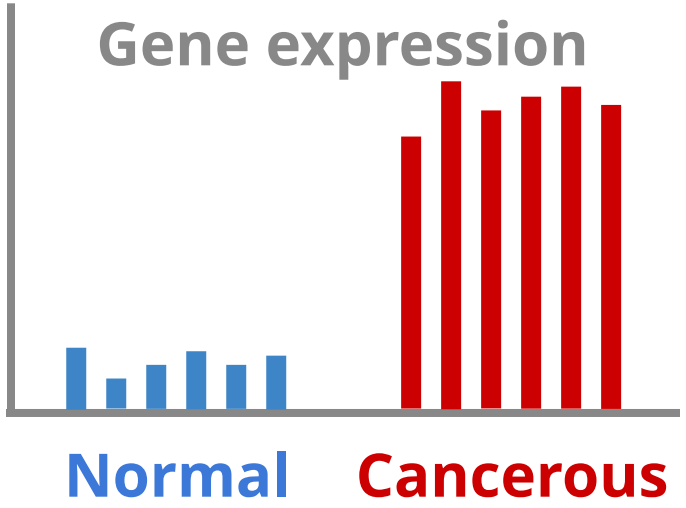
The **higher the p-value**, the more our model **supports the null hypothesis**

$p \approx 0.9999$

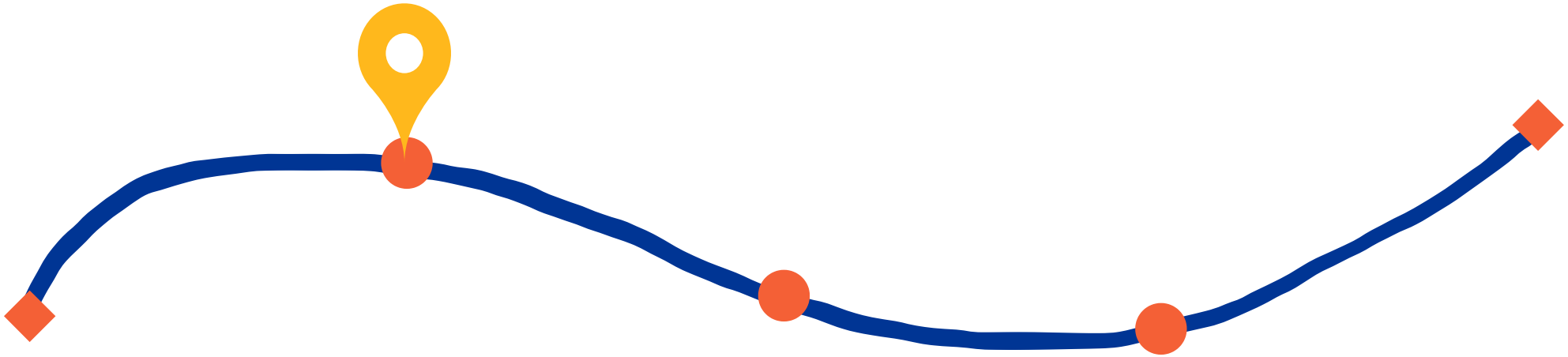


The **lower the p-value**, the more our model **supports the alternative hypothesis**

$p \approx 0.00001$



After today, you should have a better understanding of



Reliable statistical models for gene expression data

Binomial distribution

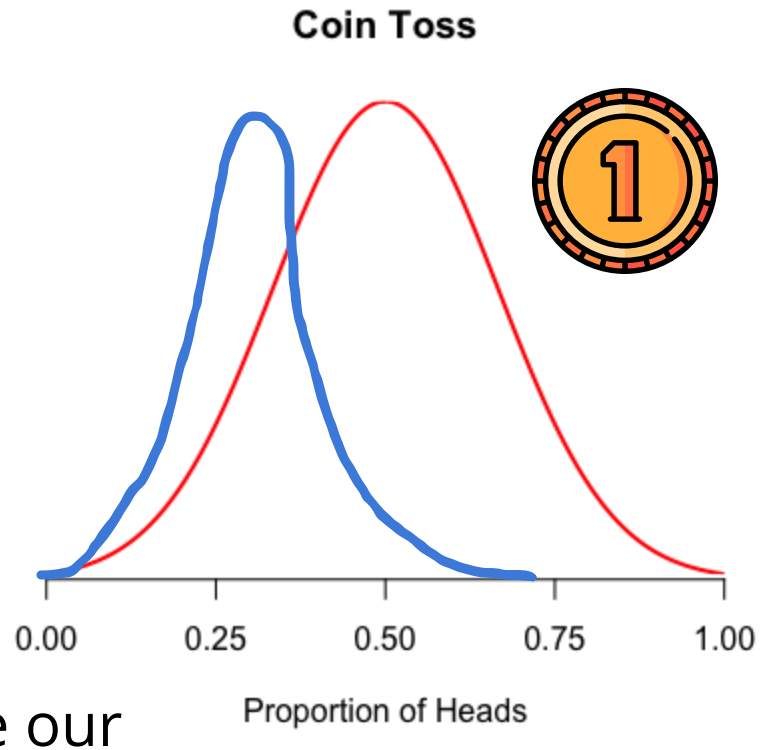
To compute probabilities under H_0, we need a model that describes expected variation

A **statistical model** describes how data is expected to behave if H_0 is true.

For example, a fair coin flip should result in a normal distribution centered on 50% of each side

This is our statistical model that describes our coin flip observations under H_0

If we flip a coin 10 million times and our distribution looks like **this**



We are probably flipping a weighted coin because our observations do not match our H_0 statistical model

Gene expression data have unique challenges that require specific statistical models

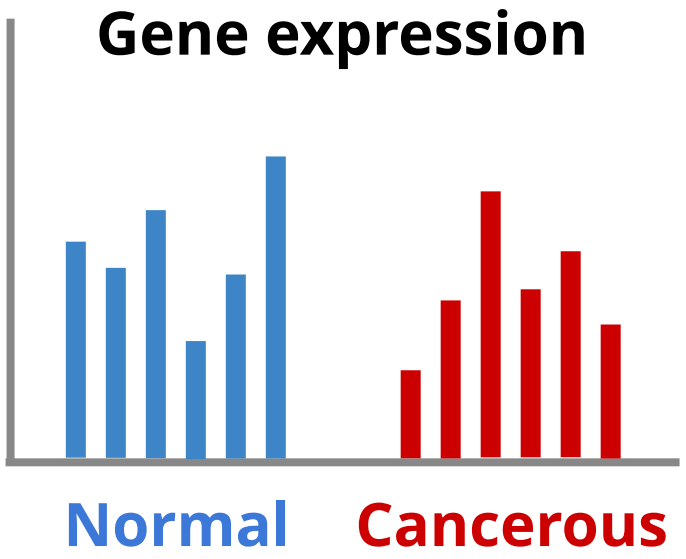
The nature of count data

RNA-seq generates **count data** – the number of RNA fragments that map to each gene

Example: 573,282 TPM

What is discrete data?

- Data that can only take whole numbers
- In RNA-seq, we measure the **number of transcripts**, so the data are **count-based**
- For example, you cannot have "half a transcript"



Discrete data requires us to use **special statistical models**

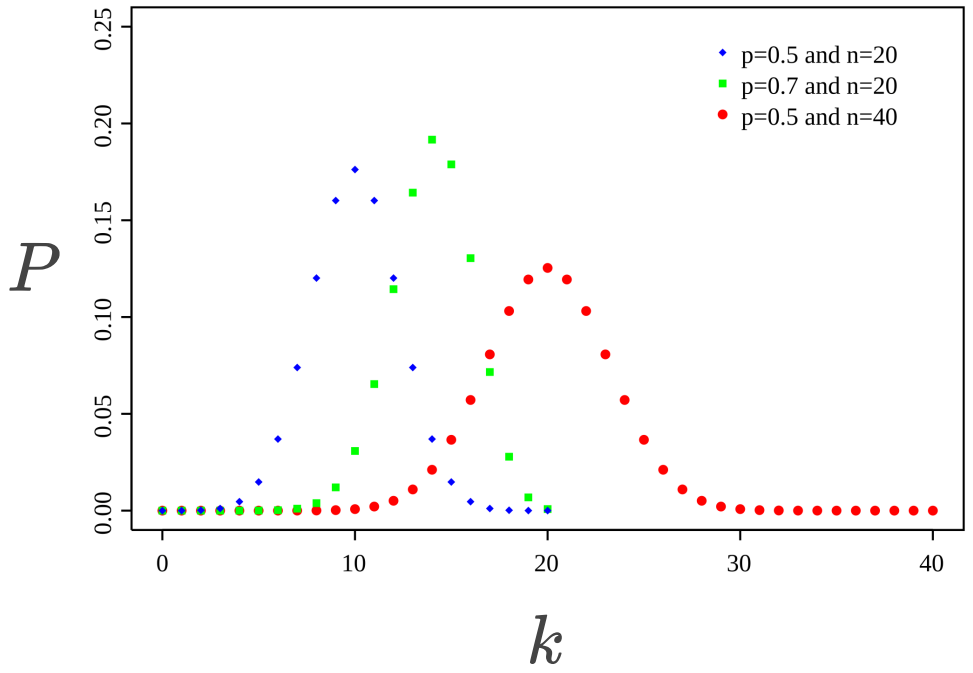
Binomial: A Simple Model for Discrete Counts

The **Binomial distribution** models the number of **successes** in a fixed number of independent trials, where each trial has the same probability of success

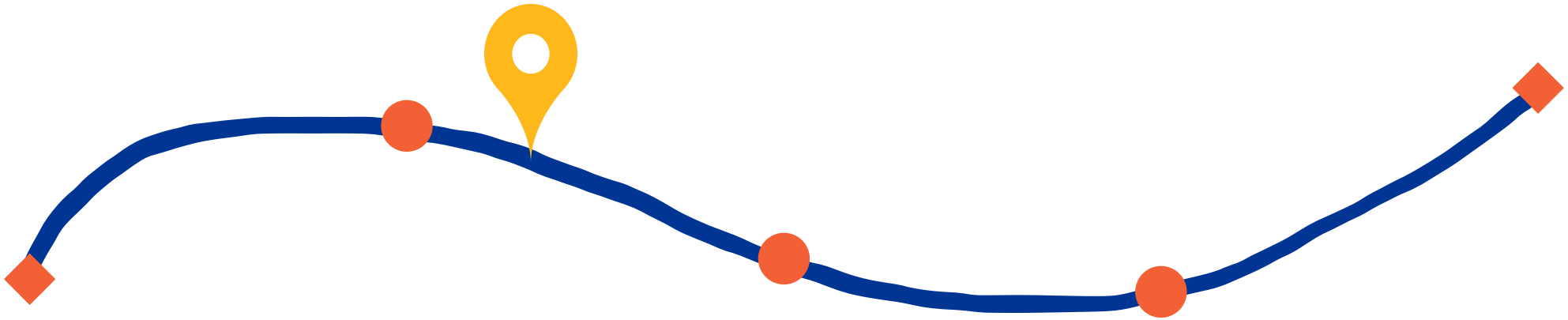
RNA-seq analogy: Each read can be considered a "trial," and the probability that a read maps to a specific gene is the "probability of success."

$$P(X = k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

P	Probability
k	Number of successes
n	Number of trials
p	Probability of success



After today, you should have a better understanding of



Reliable statistical models for gene expression data

Poisson distribution

Challenge #1: The binomial distribution assumes that the probability of success (*p*) is the same for every trial

For example, if I have 10 samples from cancerous cells, the binomial distribution assumes they are perfect replicates with no biases

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

P	Probability
k	Number of successes
n	Number of trials
p	Probability of success

Ignoring sample-to-sample variability can lead to underestimating the true uncertainty in the data

Challenge #2: High sequencing depth results in an extremely large number of trials, posing both computational and modeling challenges

$$P(X = k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

When sequencing depth is high, *n* (the total number of reads) becomes very large

Factorials when *n* is large makes accurate calculations impractical

Challenge #3: For many genes, the probability of expression (*p*) is extremely low, further complicating the use of the binomial distribution

$$P(X = k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

With very low *p*, the expected number of successes (reads mapping to a lowly expressed gene) is minuscule compared to *n*

Calculations with very small probabilities may lead to numerical underflow/imprecise results

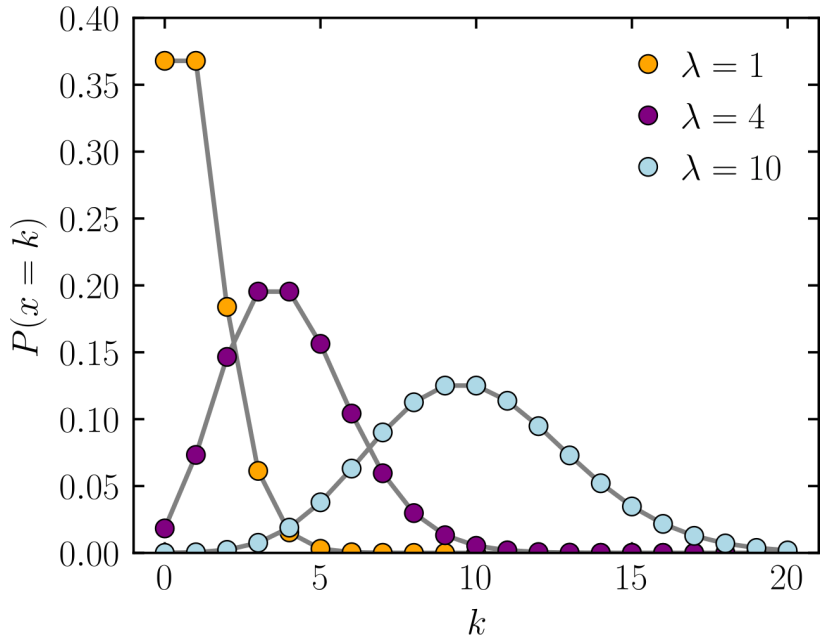
Poisson distribution: A tractable model for large discrete counts

The Poisson distribution is a statistical tool used to model the number of events (i.e., counts) that happen in a fixed period of time or space, where:

- The events are **independent** of each other
- Each event has a **constant average rate** (i.e., allows variation between events)

Assuming the constant average rate of success allows some variation around the mean

I.e., sample variation and batch effects



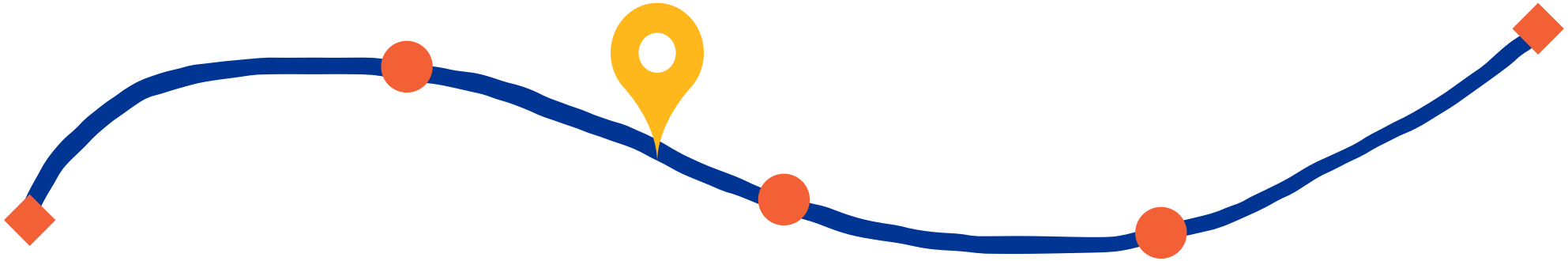
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

P Probability

k Number of events or counts

λ Expected average of X

After today, you should have a better understanding of



Reliable statistical models for gene expression data

Negative binomial distribution

Poisson distribution assumes mean and variance are equal

The expected value (i.e., mean)

$$E[X] = \sum_{k=0}^{\infty} k \cdot P(X = k)$$

$$= \sum_{k=1}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!}$$

When *k = 0*, the term is zero

$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

$$k\frac{\lambda^k}{k!} = \lambda\frac{\lambda^{k-1}}{(k-1)!}$$

$$= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}$$

Use $j = k - 1$

$$= \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

$$\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{\lambda}$$

$$= \lambda$$

You don't need to understand these derivations—just the outcome

Poisson distribution assumes mean and variance are equal

$$E[X^2] = E[X(X-1)] + E[X] \qquad E[X(X-1)] = \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k e^{-\lambda}}{k!}$$

$$Var(X) = E[X^2] - (E[X])^2$$

$$= (\lambda^2 + \lambda) - \lambda^2$$

$$= \lambda$$

$$E[X] = Var(X) = \lambda$$

If our variance is different from our mean, our Poisson model breaks down

When k = 0 or 1, the term is zero

$$= \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k e^{-\lambda}}{k!}$$

$$k(k-1)\frac{\lambda^k}{k!} = \frac{\lambda^k}{(k-2)!}$$

$$= e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!}$$

Use $j = k - 2$

$$= \lambda^2 e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}$$

$$\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{\lambda}$$

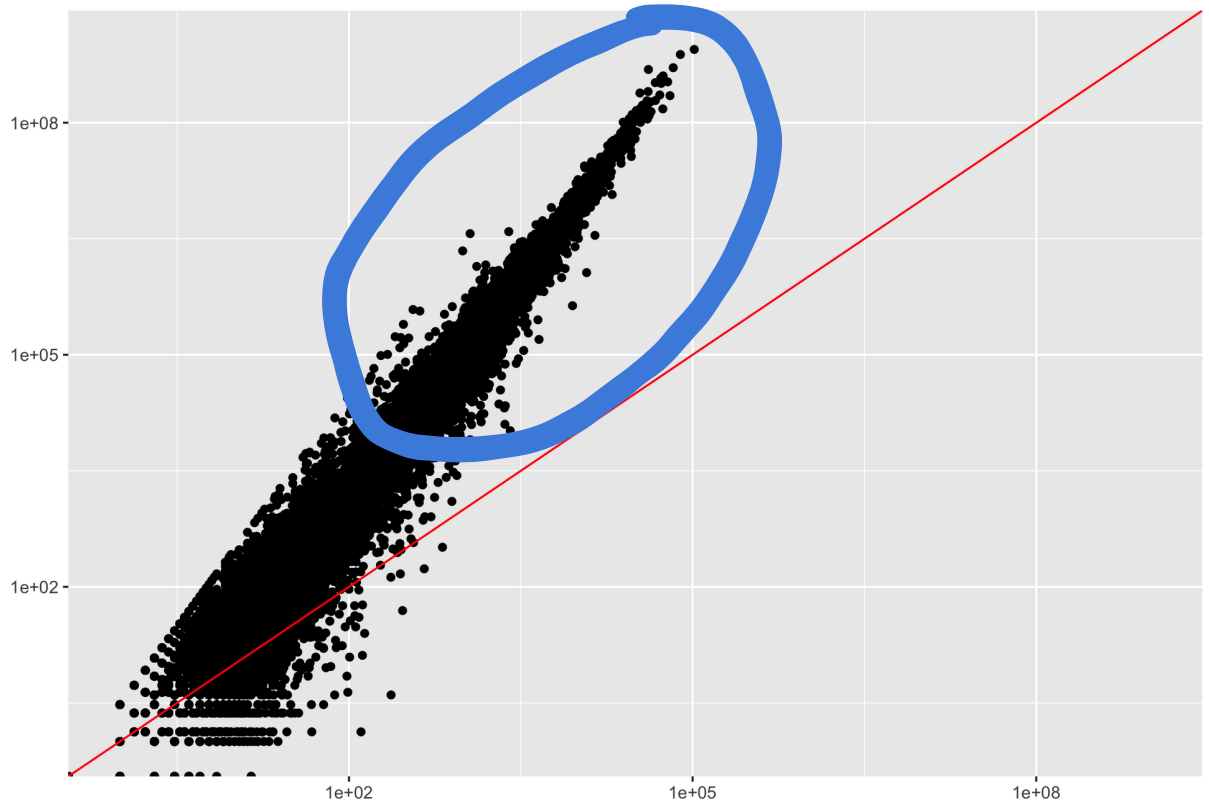
$$= \lambda^2 e^{-\lambda} \cdot e^{\lambda}$$

$$= \lambda^2$$

You don't need to understand these derivations—just the outcome

Parity plots with mean and variance show deviations with Poisson distributions

Mean = variance line



Count variance

Count mean

Higher counts typically have a larger variance

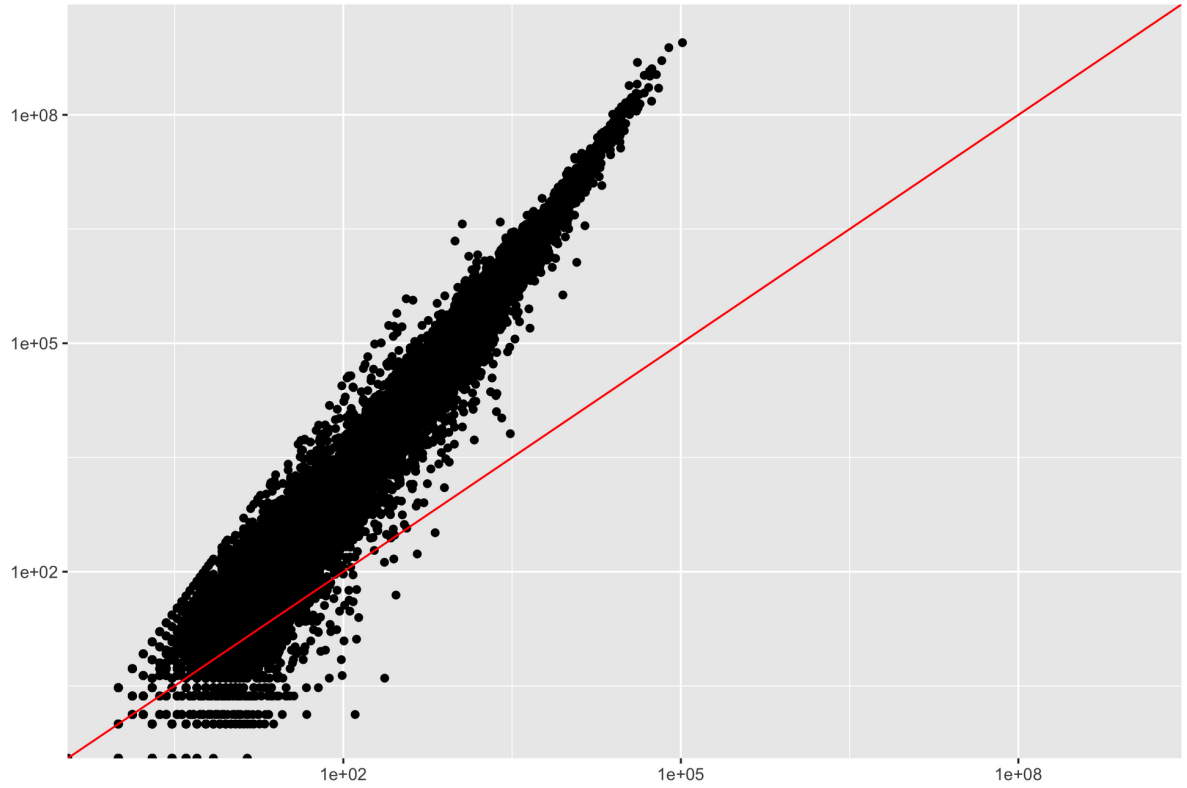
Overdispersion in RNA-Seq

Overdispersion: It happens when the variance in the data is larger than what is predicted by simpler models (e.g., Poisson distribution)

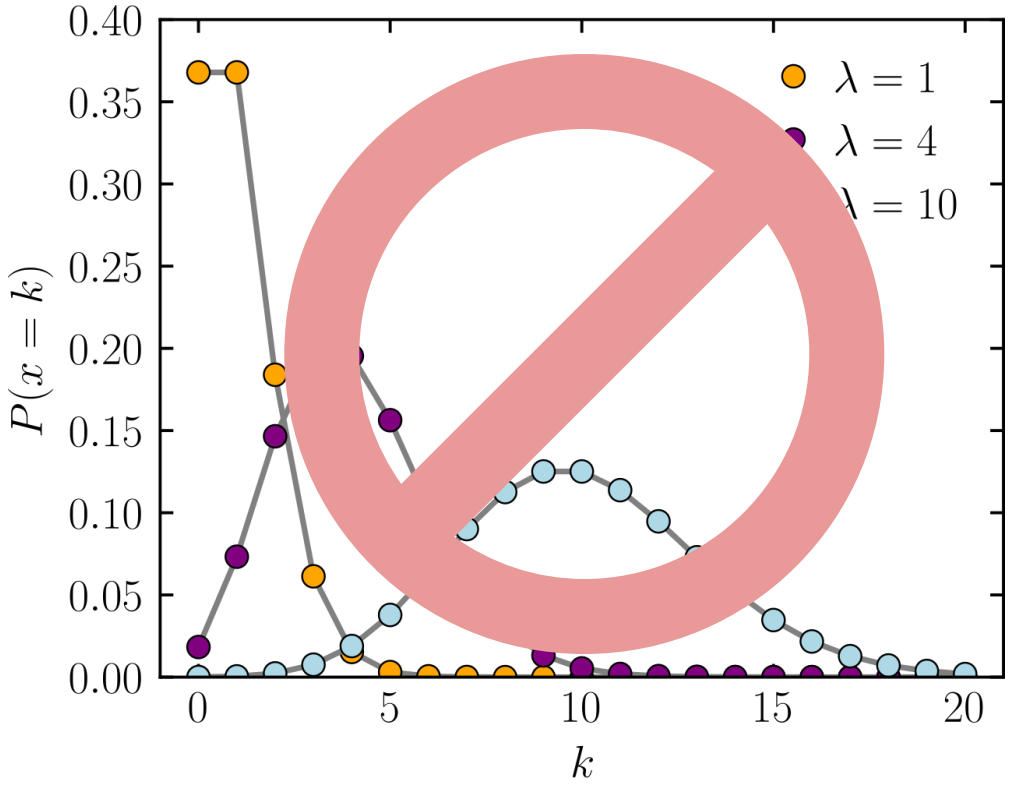
- **Expected variance** for Poisson-distributed data equals the mean: $\text{Variance} = \mu$
- Variance is often larger than the mean for RNA-Seq: $\text{Variance} > \mu$

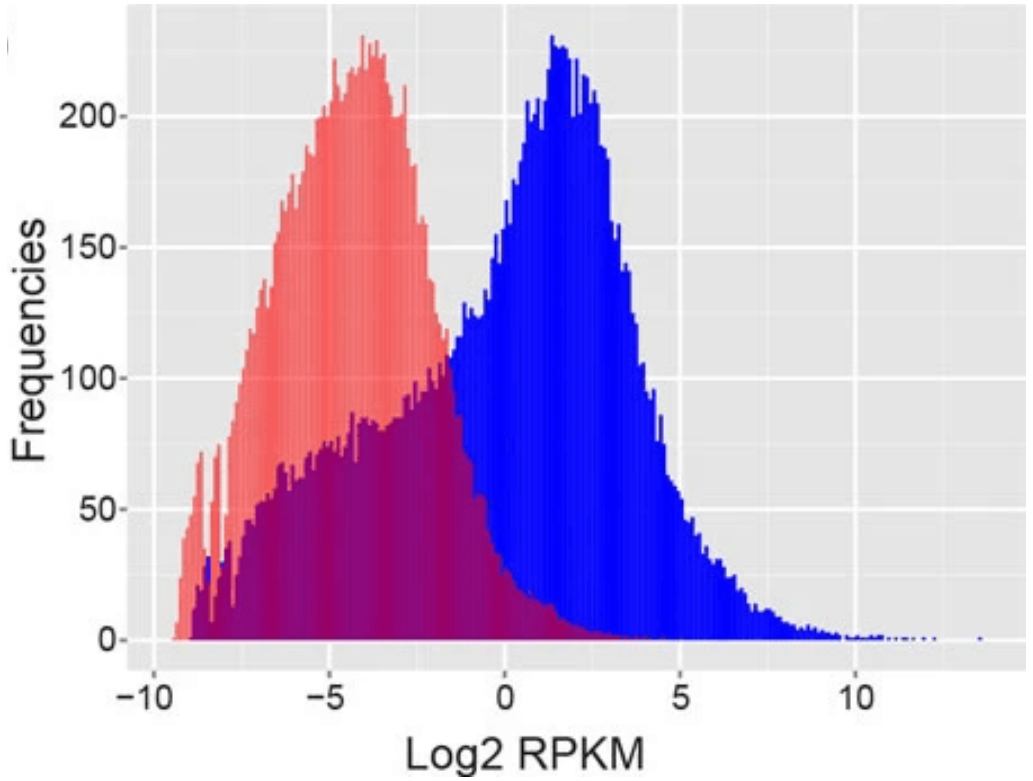
Overdispersion may reflect **biological variability** between samples not captured by the experimental conditions

- Differences in RNA quality
- sequencing depth,
- biological factors like different cell types within the same tissue



Poisson distribution is unsuitable for RNA-seq data because of high noise





Negative Binomial distribution accounts for high dispersion

$$P(X = k) = \frac{\Gamma(k + \frac{1}{\alpha})}{k!\,\Gamma(\frac{1}{\alpha})} \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{k}$$

k	Observed number of counts
μ	Mean or expected value of counts
α	Dispersion parameter, controlling how much the variance exceeds the mean
$\Gamma(\cdot)$	Gamma function, which generalizes the factorial to floats

$$\mathrm{Var}(X) = \mu + \alpha\mu^2$$

If $\alpha=0$, the Negative Binomial distribution reduces to the **Poisson distribution**

The challenge of zeros in RNA-seq data

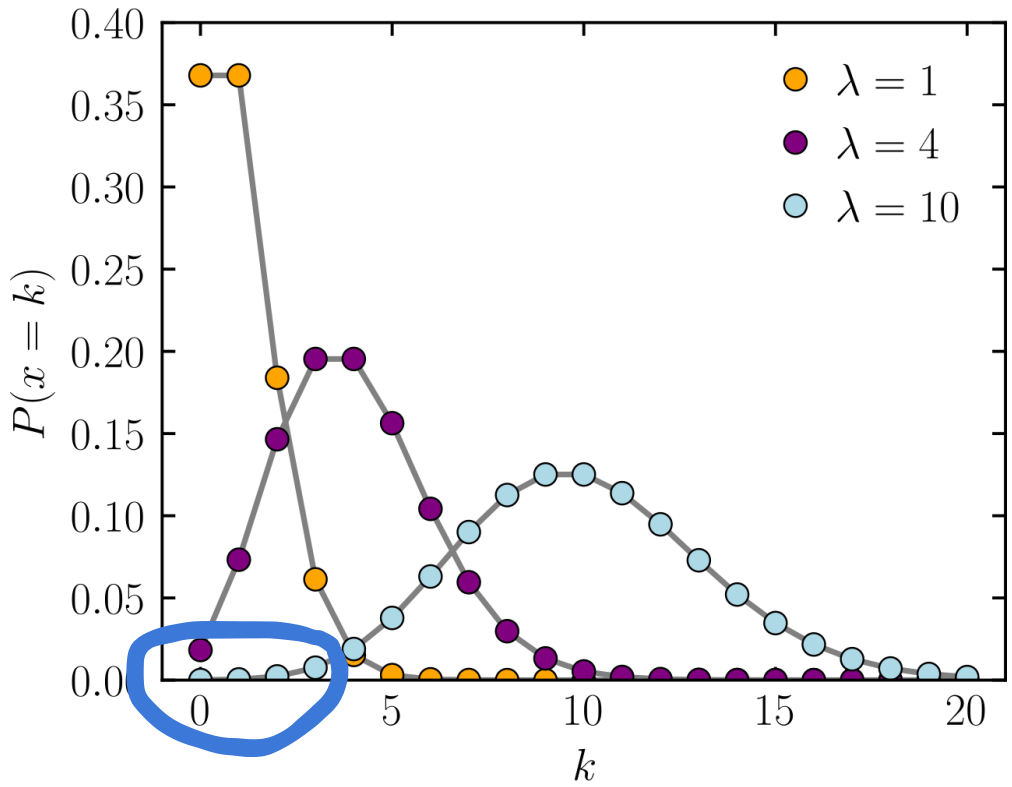
RNA-seq data frequently contains **zero counts for some genes** because not all genes are expressed under all conditions

Most statistical models account for variance, but not that zeros can dominate counts

For example, if we have a high expected mean with Poisson distribution we can still have zeros or very low counts

In these circumstances, we have to use zero-inflated models

We will ignore these for now



After today, you should have a better understanding of

Fitting statistical models

Likelihood quantifies the probability of the observed data given a model

The likelihood of model parameters θ given data $\mathbf{y}$ is defined as

$$L(\theta) = P(\mathbf{y}|\theta)$$

When individual data points $y_1, y_2, \ldots, y_n$ are independent, the joint probability is calculated by multiplying their individual probabilities:

$$P(y_1, y_2, \ldots, y_n|\theta) = \prod_{i=1}^{n} P(y_i|\theta)$$

Multiplying these probabilities aggregates the evidence from each data point, providing a comprehensive measure of how well the model with parameter θ fits all the data

A higher product (or joint likelihood) means the model assigns a higher probability to the observed data, indicating a better fit.

The log transformation simplifies computation and interpretation

Log likelihood

$$\log L(\theta) = \sum_{i=1}^{n} \log P(y_i|\theta)$$

Converts products into sums, reducing computational issues.

Makes differentiation easier for optimization

Maximum Likelihood Estimation (MLE) finds the parameters that maximize the log likelihood

Optimization problem

$$\hat{\theta} = \arg\max_{\theta} \log L(\theta)$$

At the optimum, the model parameters provide the best explanation of the observed data.

Before the next class, you should

Lecture 08A:

Differential gene expression - Foundations

Lecture 08B:

Differential gene expression - Methodology

Today

Thursday

- Work on P02A (due Mar 14)