Methods Lecture

Data Science and AI for Neuroscience Summer School

Tara Chari and Lior Pachter

July 11, 2022

Overview of Topics

Fundamental Methods and Metrics

- Matrix Definitions
- Linear and Logistic Regression
- Correlation and Partial Correlation

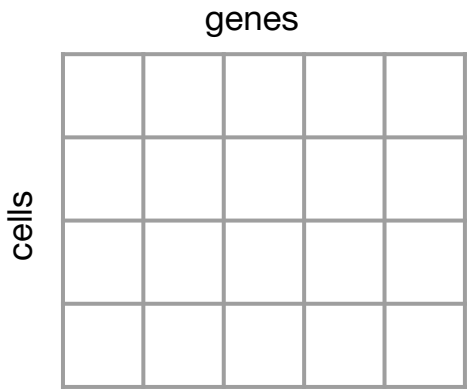
Applications to Biological Count Data

- Count Distributions
- Normalization/Stabilization of Counts

Single-cell RNA-seq

- Single-cell RNA-seq is neither single, cell, nor RNA.
  So what is it?

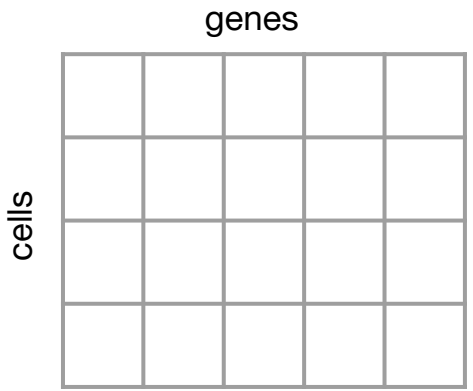
- Single-cell RNA-seq refers to a group of (constantly improving) technologies and analysis tools that
  - start with an **INPUT** of cells,
  - **OUTPUT** a (proxy for a) gene expression matrix.



Single-cell RNA-seq

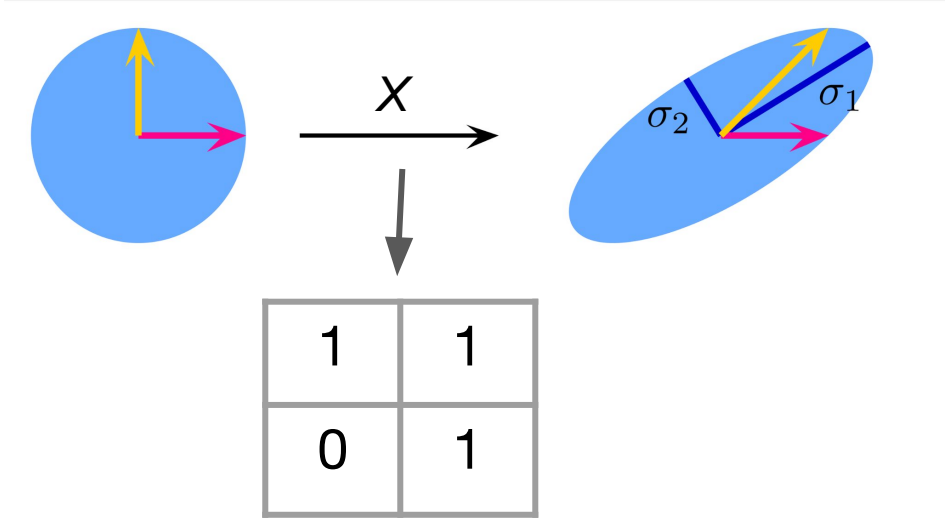
- Single-cell RNA-seq is neither single, cell, nor RNA.
  So what is it?

- Single-cell RNA-seq refers to a group of (constantly improving) technologies and analysis tools that
  - start with an **INPUT** of cells,
  - **OUTPUT** a (proxy for a) gene expression matrix.

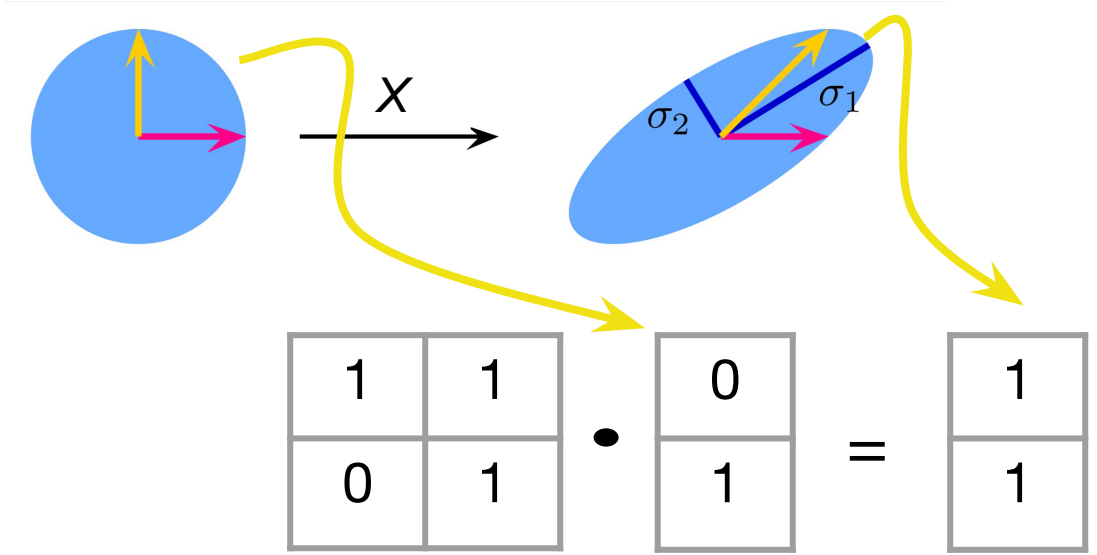


What is a gene expression matrix?

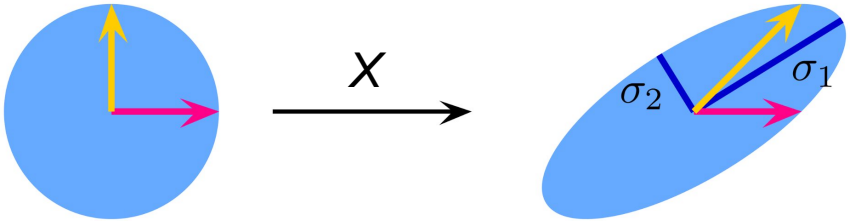
A matrix is code for a (linear) function



How a matrix describes (is code for) a function



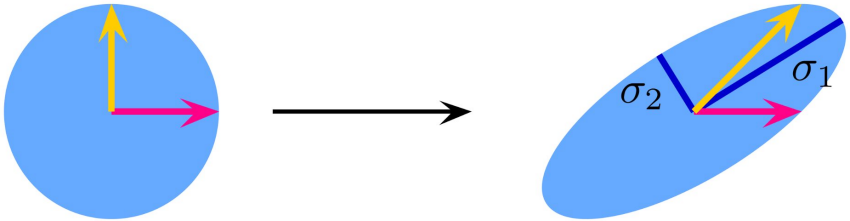
How a matrix describes (is code for) a function



$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}$$

$$X \left(\quad b \quad \right)$$

The rank of a matrix is...



$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}$$

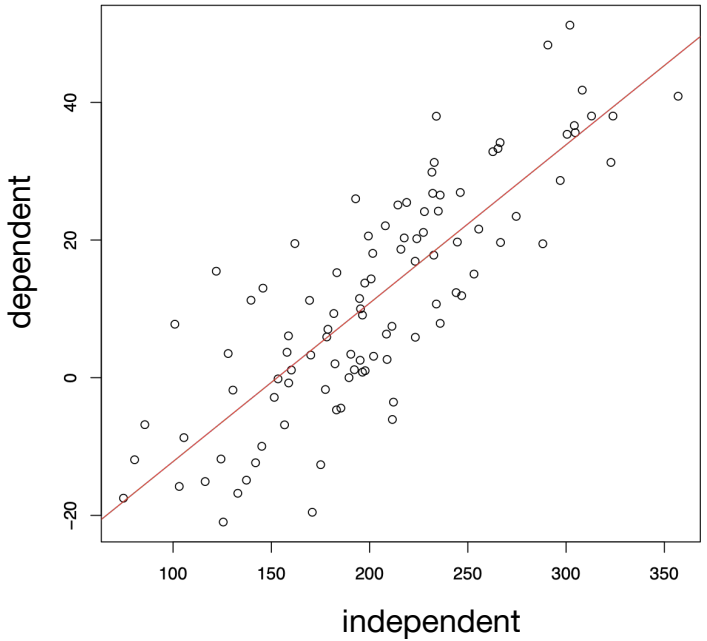
$$X \left(\quad b \quad \right)$$

… the dimension of the image.

Linear regression

- “Regression analysis” refers to the problem of estimating relationships between a dependent variable, and one or more independent variables.

- The simplest example of this is linear regression, where the relationship to is assumed to be linear, and regression analysis then refers to finding a linear combination of the independent variables that provides the *best fit* to the dependent variable.

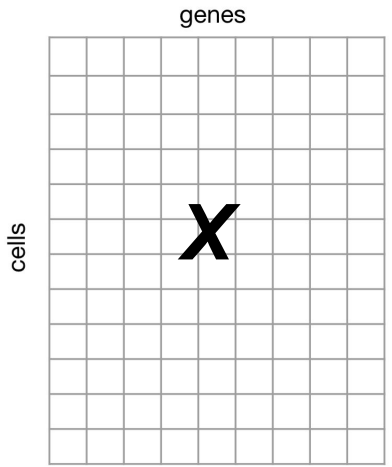


A (statistics) convention for tables

- The convention for representing data tables in ***statistics*** is to use the *rows* for observations, and the *columns* for features.
- One reason for this convention is the form of regression models, which describe observations as linear combinations of explanatory variables with some added noise using the form:

$$y = X\beta + \epsilon.$$

  - With this matrix notation, *X*, which is also known as the design matrix

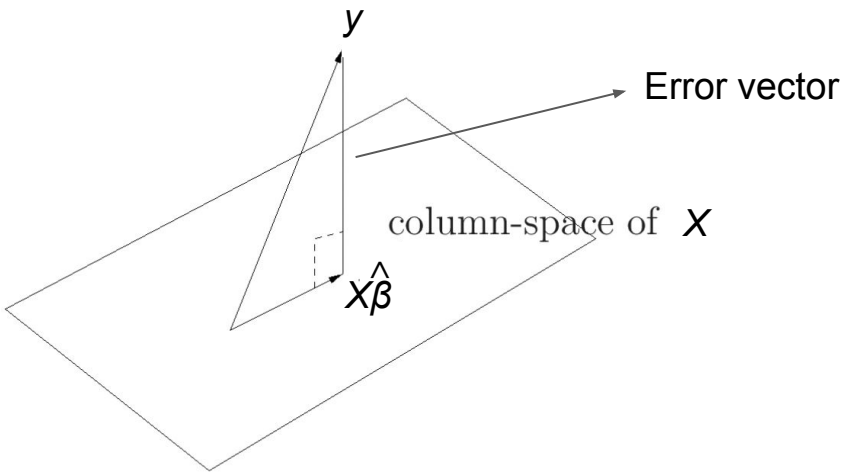


Solving the least squares problem (geometrically)

- Least squares optimization = finding nearest point in col(X) i.e. find the value β that minimizes $(||X\beta - Y||_2)^2$; the minimal β is denoted $\hat{\beta}$.

standard notation for an estimator

- As we will investigate, the solution emerges naturally as $(X^TX)^{-1}X^TY$ …



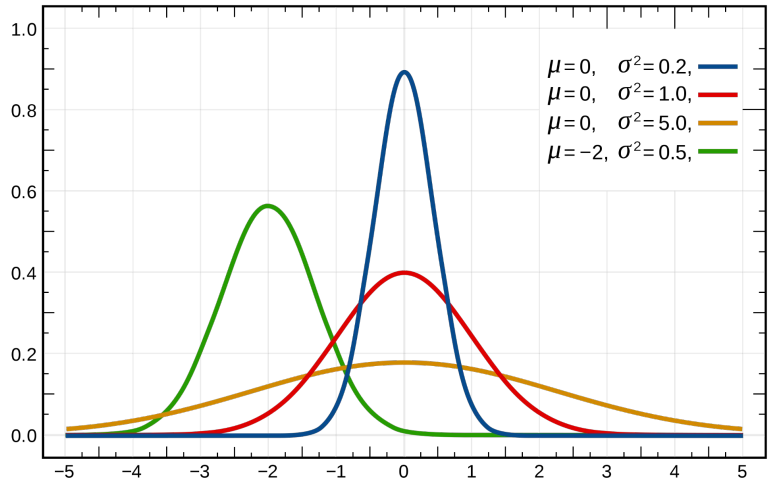
Gaussian (normal) distributions for error ϵ

$$y = X\beta + \epsilon.$$

- The normal distribution is given by

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

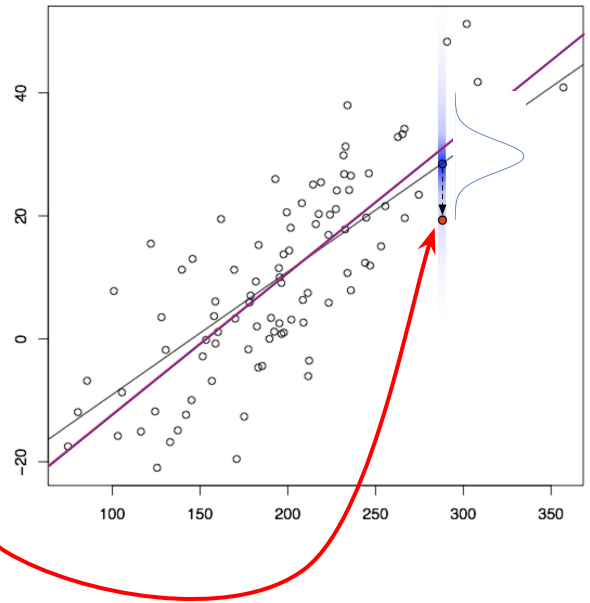
- The distribution has two parameters: μ is the mean and σ^2 is the variance.
- The normal distribution is also called the "Gaussian".
- The *standard normal* has $\mu = 0$, $\sigma^2 = 1^2$.



Gaussian error model for linear regression

- Assuming the points were initially on the line $y=mx+b$, the probability of the perturbation of point i, which was initially at height mx_i+b but after vertical perturbation is at y_i, is

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(mx_i+b))^2}{2\sigma^2}}.$$



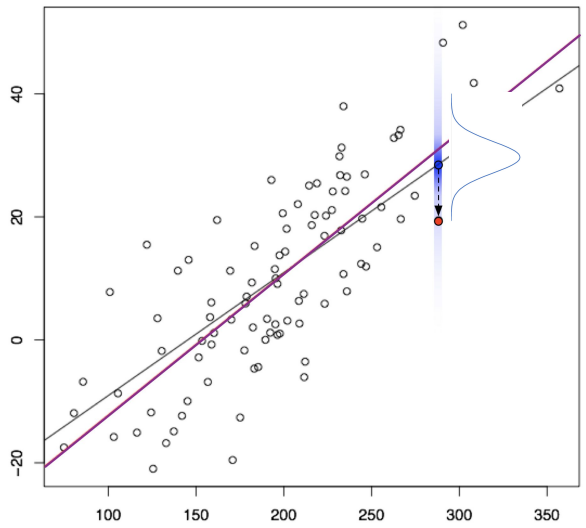
Gaussian error model for linear regression

- Assuming the points were initially on the line $y=mx+b$, the probability of the perturbation of point i, which was initially at height mx_i+b but after vertical perturbation is at y_i, is

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(mx_i+b))^2}{2\sigma^2}}.$$

- The likelihood of observing data $(x_1,\ldots,x_n)$ and $(y_1,\ldots y_n)$ is therefore the product of these probabilities, namely

$$\mathcal{L}(m,b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(mx_i+b))^2}{2\sigma^2}}.$$



Least squares solution from Gaussian error model

- Finding the original line from which the data is generated, is mathematically the problem of finding the parameters *m* and *b* that maximize the likelihood function. That is, $\text{argmax}_{m,b}\mathcal{L}(m,b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(mx_i+b))^2}{2\sigma^2}}$.

  Since the logarithm is a monotonically increasing function, one can instead find the parameters maximizing the logarithm of the likelihood function:

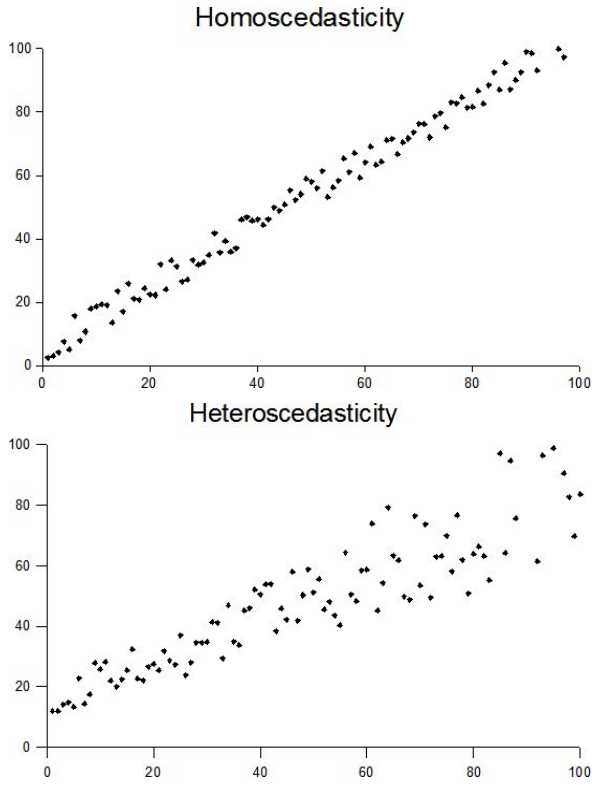
$$\text{argmax}_{m,b} log(\mathcal{L}(m,b)) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(mx_i+b))^2$$

$$= \text{argmin}_{m,b}\sum_{i=1}^{n}(y_i-(mx_i+b))^2.$$ ← least squares

Homoscedasticity and heteroscedasticity

- One of the assumptions in the Gaussian latent model for linear regression, is that of **homoscedasticity**. This means that the **variance σ^2 is constant** across all observations, and does not depend on the value of the explanatory variables. The assumption is necessary for the *Gauss-Markov* theorem to hold and for least squares to yield the **B**est **L**inear **U**nbiased **E**stimator

- The opposite of homoscedasticity is heteroscedasticity.



Example application in single-cell RNA-seq

- From the preprint (Kamimoto et al., 2020):

  “Here, **we present CellOracle, a machine learning-based tool** to infer [gene regulatory networks, i.e.] GRNs via the integration of different single-cell data modalities. CellOracle overcomes current challenges in GRN inference by using single-cell transcriptomic and chromatin accessibility profiles, integrating prior biological knowledge via regulatory sequence analysis to infer transcription factor (TF)-target gene interactions.”

- From the Methods section:

learning model. CellOracle builds a model that predicts a target gene expression based on the gene expression of regulatory candidate genes:

$$y = \sum_i \beta_i x_i + \alpha$$

Where x_i is gene expression value of the regulatory candidate gene, y is target gene expression, β_i is a coefficient value of the linear model, and α is an intercept for this model. Here, we use the list of potential regulatory genes generated in the previous step. In CellOracle, we use the coefficient β as a network edge strength between a TF and its target gene. For example, β will

Logistic Regression Model

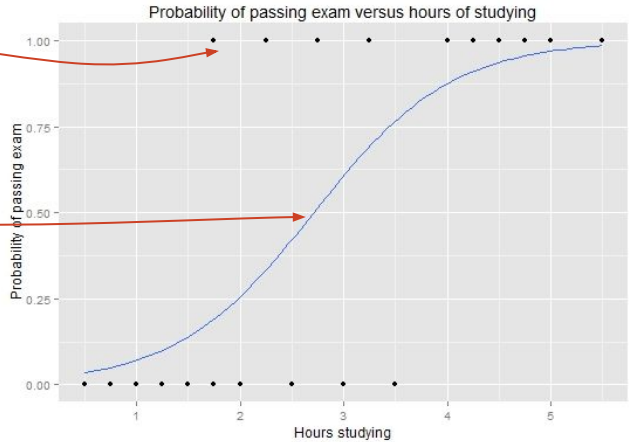
- A model for binary dependent variables, i.e. *y = 0* or *y = 1.*
- Model <u>probability</u> of event using the logistic functions

Max

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Steepness

Midpoint



$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Logistic regression for classification

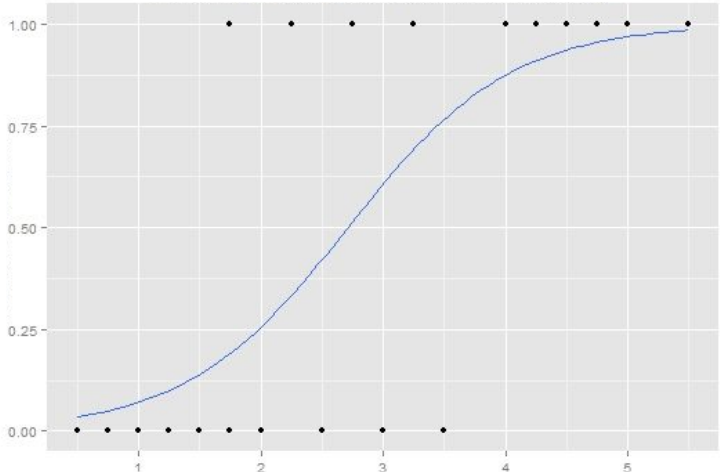
- The likelihood function is the product of the probability for each point:

$$
\begin{aligned}
L(\theta \mid y; x) &= \Pr(Y \mid X; \theta) \\
&= \prod_i \Pr(y_i \mid x_i; \theta) \\
&= \prod_i h_\theta(x_i)^{y_i} (1 - h_\theta(x_i))^{(1-y_i)}
\end{aligned}
$$

where

$$
h_\theta(X) = \frac{1}{1 + e^{-\theta^T X}} = \Pr(Y = 1 \mid X; \theta)
$$

$p(x)$

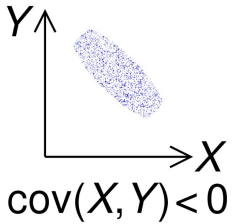


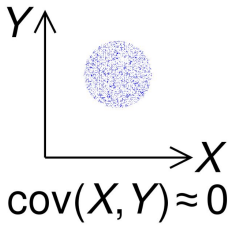
Covariance of random variables

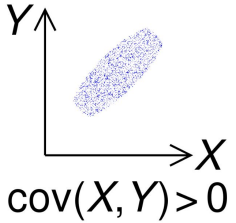
- The covariance of two random variables X and Y is
  $cov[X,Y] = E[\,(X-E[X])(Y-E[Y])\,]$
  $= E[XY]-E[X]E[Y]$.

- $cov[X,X] = var[X]$.

- If X and Y are independent random variables then the covariance is zero:
  Proof: Independence means that $E[XY]=E[X]E[Y]$.
  The converse is not true.







Correlation

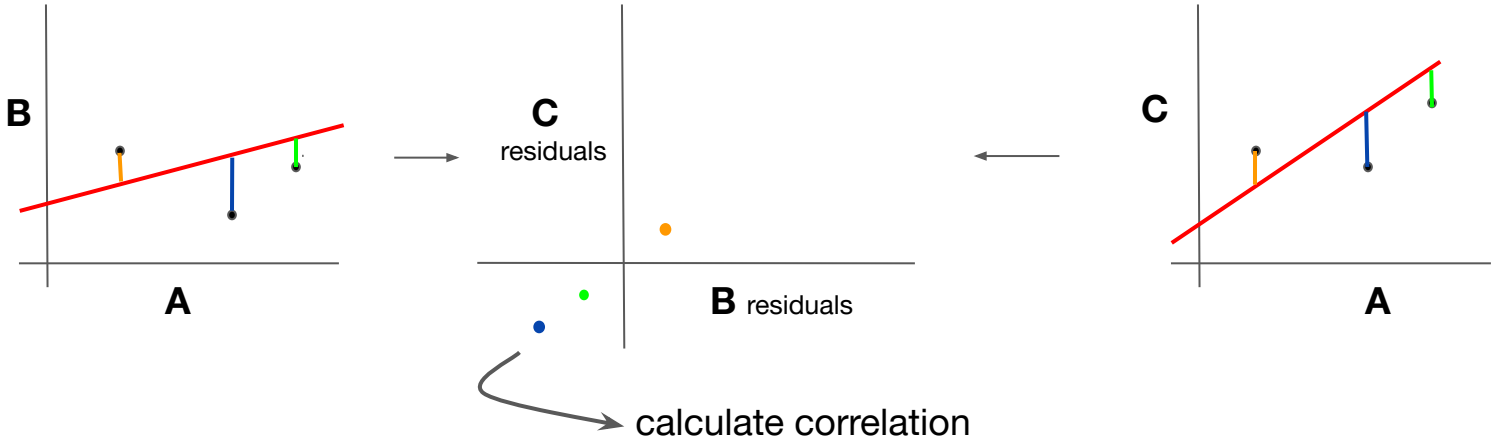
- The covariance of two random variables *X* and *Y* is in units that are a product of those of *X* and *Y*. To obtain a dimensionless number, the covariance can be divided by the product of the standard deviation of *X* and the standard deviation of *Y*. This is called the ***correlation coefficient***:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

  Other names include Pearson's product-moment correlation coefficient, Pearson's coefficient, or Pearson's correlation.

Partial correlation

- In order to better assess the “direct” correlation between **B** and **C**, instead of computing the correlation between them, the correlation between the residuals after *regressing* **B** against **A**, and **C** against **A**, are computed. This is known as *partial correlation*.



Computing partial correlation

- The partial correlation between **B** and **C** given **A**, which is denoted by $\rho_{BC \cdot A}$, can be shown to be simplify to

$$\rho_{BC \cdot A} = \frac{\rho_{BC} - \rho_{BA}\rho_{AC}}{\sqrt{1 - \rho_{BA}^2}\sqrt{1 - \rho_{AC}^2}}.$$

- Partial correlation between two random variables can be extended to the case where residuals that are computed by regressing against not one, but multiple other random variables, are correlated with each other.

Overview of Topics

Fundamental Methods and Metrics

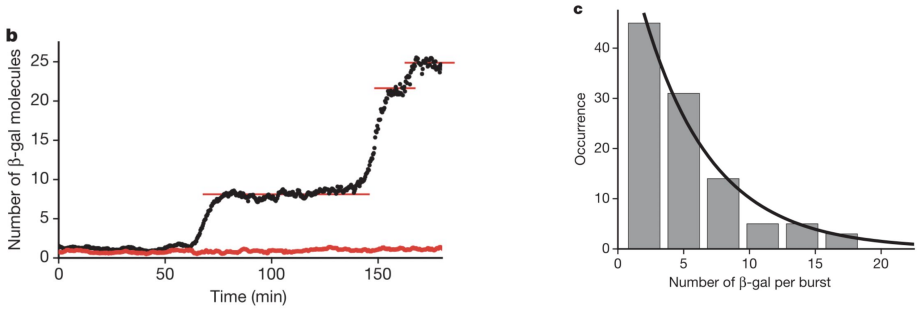
- Matrix Definitions
- Linear and Logistic Regression
- Correlation and Partial Correlation

Modeling and Analyzing Single-cell Data

- Count Distributions (Modeling Data)
- Normalization/Stabilization of Counts (Preprocessing)

Single-cell data is sparse, discrete, and noisy

- Lots of zero counts
- Dealing with approximate counts of molecules (UMIs)
- Sources of intrinsic and extrinsic noise, from cells and sampling during sequencing



Cai, Friedman, and Xie, 2006

Describing gene-count matrices as count data

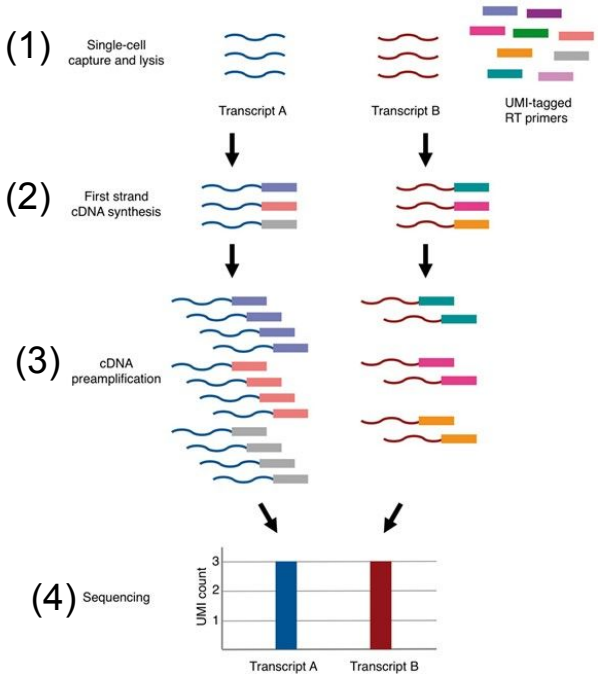
- Data where the observations result from counting some objects and are in the set of non-negative integers *{0,1,2,...}*.

- Entries in single-cell gene expression matrices are counts of molecules derived from read alignment followed by counting of UMIs.

- Count data are usually represented with
  - Binomial distribution.
  - Poisson distribution.
  - Negative binomial distribution.

Counting molecules with UMIs

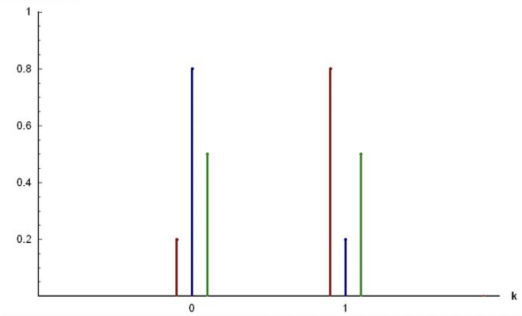
- **U**nique **M**olecular **I**dentifiers (UMIs) allow for counting the number of distinct RNA molecules represented in a cDNA library, rather than the number of distinct DNA molecules comprising the cDNA library.
- Unlike "**UMI Counts"**, “**Read counts**” means the DNA (not RNA) molecules comprising the cDNA library
- The process of deriving UMI counts (4) from read counts (3) is called *UMI collapsing*.

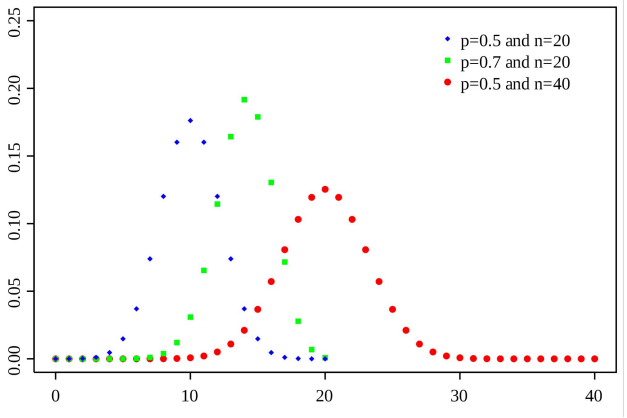


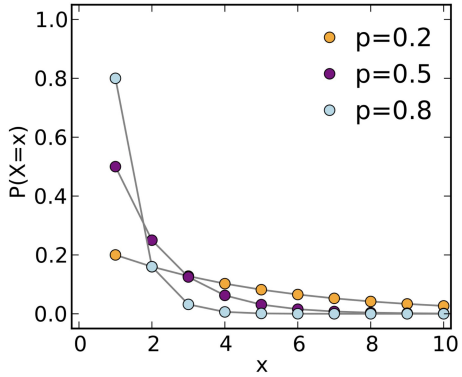
Kolodziejczyk et al., 2017

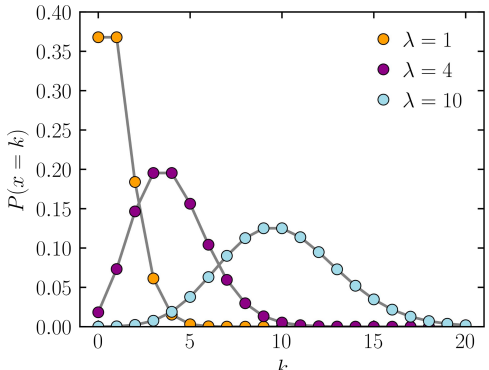
Six essential distributions for count data:

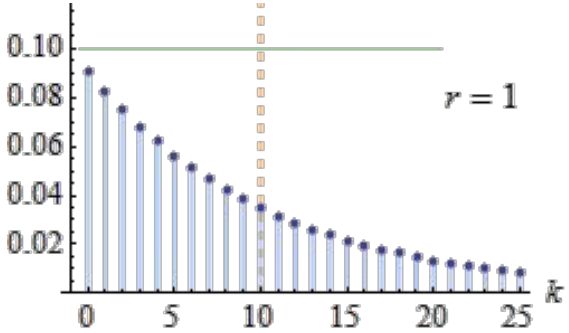
Bernoulli, Binomial, Poisson, Geometric, Negative Binomial, Gamma

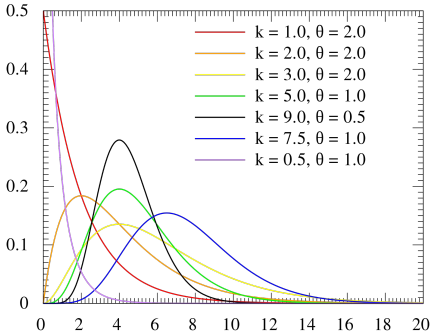






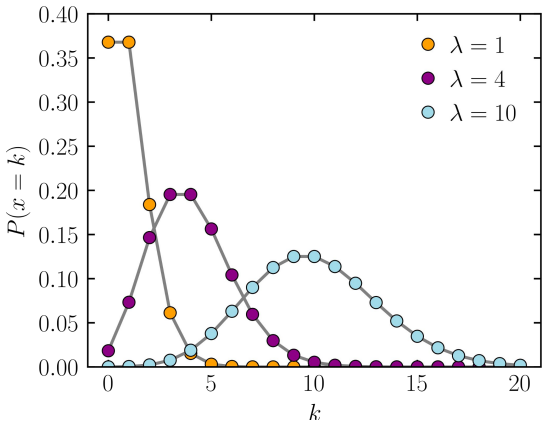






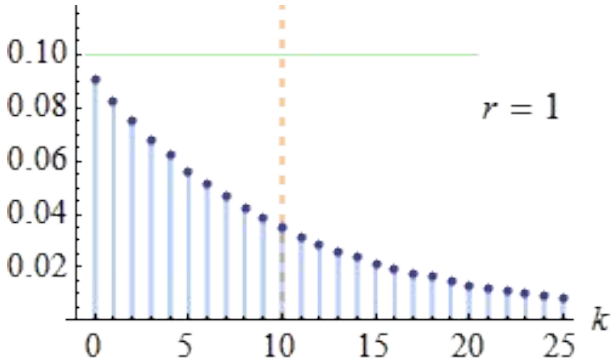
Poisson distribution

Parameters	$\lambda \in (0, \infty)$ (rate)
Support	$k \in \mathbb{N}_0$ (Natural numbers starting from 0)
PMF	$\frac{\lambda^k e^{-\lambda}}{k!}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)



Negative binomial distribution I

Parameters	$r > 0$ — number of failures until the experiment is stopped (integer, but the definition can also be extended to reals) $p \in [0,1]$ — success probability in each experiment (real)
Support	$k \in \{0, 1, 2, 3, \ldots\}$ — number of successes
PMF	$k \mapsto \binom{k+r-1}{k} \cdot (1-p)^r p^k,$ involving a binomial coefficient
CDF	$k \mapsto 1 - I_p(k+1, r)$, the regularized incomplete beta function



Negative binomial distribution II

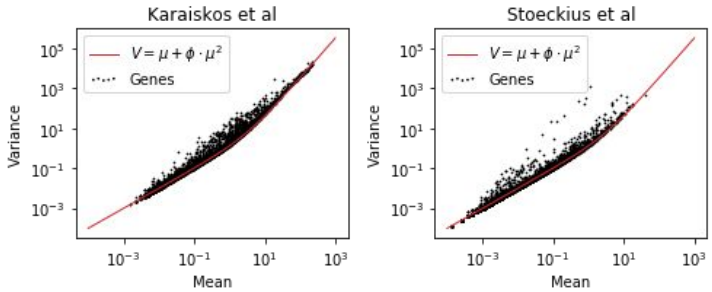
- The negative binomial distribution arises as a mixture of Poisson distributions via a hierarchical model where the Poisson parameter is a *Gamma*-distributed random variable:

$$X|\lambda \sim \text{Pois}(\lambda)$$
$$\lambda \sim \text{Gamma}\,(\alpha, \beta)\,.$$

- This interpretation is fundamental to understanding the negative binomial distribution, as it shows that the negative binomial can be seen as a Poisson random variable where there is “noise” in the rate parameter.

The gene mean-variance relationship in single-cell

- The negative binomial distributions provide good fits for gene counts. This results from a particular **mean-variance** relationship that is evident in numerous datasets:



Svensson, 2017

- In particular, the negative binomial distribution yields a variance that is quadratic in the mean, namely $V = \mu + \phi \cdot \mu^2$, where μ is the mean and Φ is a parameter called the *dispersion* or *shape* parameter.

Processing Count Data for Analysis

- What assumptions do analysis methods make?

- How to pre-process data to use such methods?

Normalization

- Analysis methods such as Least-Squares and PCA are based on models that make strong assumptions about variance e.g. **equal variance** in each coordinate (gene).
- Similarly, clustering methods that rely on computation of distances between cells from genes that have unequal variance, will be disproportionately affected by highly expressed genes.

- For these reasons, it is desirable to transform gene expression data so that genes are placed on an equal (variance) footing. This process is called **variance stabilization.**

Variance stabilizing transformations for negative binomial data

- Anscombe also showed that a good approximation to the transformation

$$y = 2sinh^{-1}\sqrt{\frac{x+c}{k+d}},$$

is given by

$$y = ln\left(x + \frac{k}{2}\right).$$

- The term *k/2* is called a *pseudocount*.
- This is why it makes sense to log transform count data (e.g. log1p).

Scaling and size factors

- Single-cell RNA-seq data is inherently *compositional*, i.e. the counts are not absolute measures of molecule counts but rather samples from a cDNA library. Therefore, the gene counts are only *relatively* meaningful.

- Some cells may be sampled more than others, i.e. the total number of reads, or read depth, across cells is not equal.

Adjusting for read depth

- The process of adjusting counts in a gene expression matrix to account for read depth is called size factor normalization.

- The simplest scaling procedure for transformation of gene expression matrices to account for their compositional nature, or for performing size factor normalization, is to divide the counts x_{cg} for gene g in cell c by the sum of all counts for that cell, i.e. $\tilde{x}_{cg} = \frac{x_{cg}}{\sum_c x_{cg}}$.

Overview of Topics

Fundamental Methods and Metrics

- Matrix Definitions
- Linear and Logistic Regression
- Correlation and Partial Correlation

Modeling and Analyzing Single-cell Data

- Count Distributions
- Normalization/Stabilization of Counts