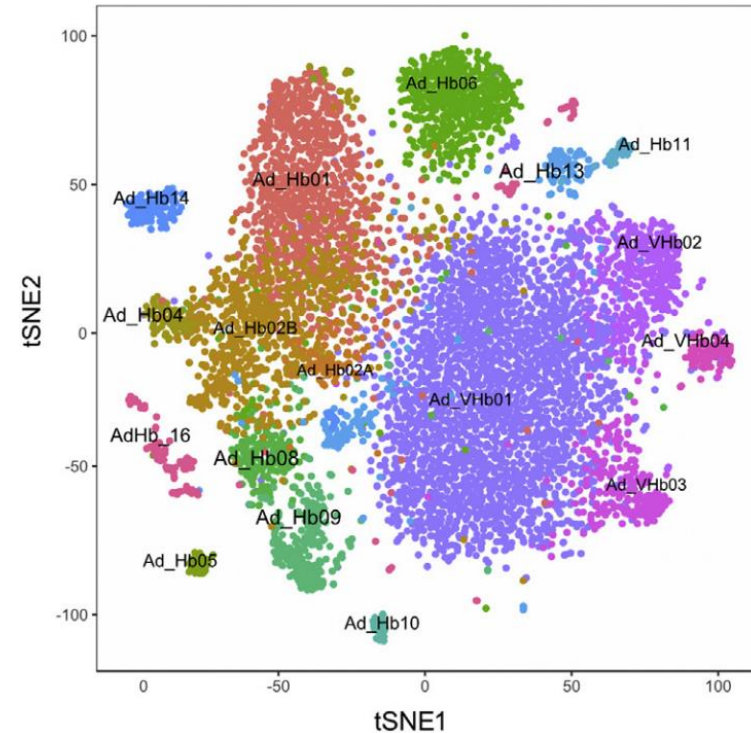


# Statistics and data analysis

Dave Zhao  
Department of Statistics  
University of Illinois Urbana-Champaign



# Is statistics really necessary?



## Official release of Seurat 4.0

We are excited to release Seurat v4.0! This update brings the following new features and functionality:

- **Integrative multimodal analysis.** The ability to make simultaneous measurements of multiple data types from the same cell, known as multimodal analysis, represents a new and exciting frontier for single-cell genomics. In Seurat v4, we introduce weighted nearest neighbor (WNN) analysis, an unsupervised strategy to learn the information content of each modality in each cell, and to define cellular state based on a weighted combination of both modalities. In our new paper, we generate a CITE-seq dataset featuring paired measurements of the transcriptome and 228 surface proteins, and leverage WNN to define a multimodal reference of human PBMC. You can use WNN to analyze multimodal data from a variety of technologies, including CITE-seq, ASAP-seq, 10X Genomics ATAC + RNA, and SHARE-seq.

Seurat 4.0.0

InstallGet startedVignettesExtensionsFAQNewsReferenceArchive

Dataset: Download here

Rapid mapping of query datasets to references.

We introduce Azimuth, a workflow to leverage high-quality reference datasets to rapidly map new scRNA-seq datasets (queries). For example, you can map any scRNA-seq dataset of human PBMC onto our reference, automating the process of visualization, clustering annotation, and differential expression. Azimuth can be run within Seurat, or using a standalone web application that requires no installation or programming experience.

Vignette: Mapping scRNA-seq queries onto reference datasets

Web app: Automated mapping, visualization, and annotation of scRNA-seq datasets from human PBMC

Additional speed and usability updates: We have made minor changes in v4, primarily to improve the performance of Seurat v4 on large datasets. These changes substantially improve the speed and memory requirements, but do not adversely impact downstream results. We provide a detailed description of key changes [here](#). Users who wish to fully reproduce existing results can continue to do so by continuing to install Seurat v3.

We believe that users who are familiar with Seurat v3 should experience a smooth transition to Seurat v4. While we have introduced extensive new functionality, existing workflows, functions, and syntax are largely unchanged in this update. In addition, Seurat objects that have been previously generated in Seurat v3 can be seamlessly loaded into Seurat v4 for further analysis.

## About Seurat

Seurat is an R package designed for QC, analysis, and exploration of single-cell RNA-seq data. Seurat aims to enable users to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements, and to integrate diverse types of single-cell data.

If you use Seurat in your research, please considering citing:

- Hao<sup>1</sup>, Hao<sup>2</sup>, et al., *Cell* 2021 [Seurat V4]
- Stuart<sup>1</sup>, Butler<sup>1</sup>, et al., *Cell* 2019 [Seurat V3]
- Butler<sup>1</sup>, et al., *Nat Biotechnol* 2018 [Seurat V2]
- Satija<sup>1</sup>, Farrell<sup>1</sup>, et al., *Nat Biotechnol* 2015 [Seurat V1]

All methods emphasize clear, attractive, and interpretable visualizations, and were designed to be [easily used](#) by both dry-lab and wet-lab researchers.

Seurat is developed and maintained by the [Satija lab](#) and is released under the GNU Public License (GPL 3.0).

Developed by Paul Hoffman, Satija Lab and Collaborators.



Home » Bioconductor 3.15 » Software Packages » edgeR

## edgeR

platforms allrank 25 / 2140support 5.5 / 6.1in Bioc 13.5 years

build okupdated before releasedependencies 10

DOI: [10.18129/B9.bioc.edgeR](https://doi.org/10.18129/B9.bioc.edgeR)

### Empirical Analysis of Digital Gene Expression Data in R

Bioconductor version: Release (3.15)

Differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. As well as RNA-seq, it be applied to differential signal analysis of other types of genomic data that produce read counts, including ChIP-seq, ATAC-seq, Bisulfite-seq, SAGE and CAGE.

Author: Yunshun Chen, Aaron TL Lun, Davis J McCarthy, Matthew E Ritchie, Belinda Phipson, Yifang Hu, Xiaobei Zhou, Mark D Robinson, Gordon K Smyth

Maintainer: Yunshun Chen <yuchen at wehi.edu.au>, Gordon Smyth <smyth at wehi.edu.au>, Aaron Lun <infinite.monkeys.with.keyboards at gmail.com>, Mark Robinson <mark.robinson at imls.uzh.ch>

Citation (from within R, enter `citation("edgeR")`):

Robinson MD, McCarthy DJ, Smyth GK (2010), "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, 26(1), 139-140. doi: [10.1093/bioinformatics/bto516](https://doi.org/10.1093/bioinformatics/bto516).

McCarthy DJ, Chen Y, Smyth GK (2012), "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic Acids Research*, 40(10), 4288-4297. doi: [10.1093/nar/aks042](https://doi.org/10.1093/nar/aks042).

Chen Y, Lun AAT, Smyth GK (2016), "From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline." *F1000Research*, 5, 1438. doi: [10.12688/f1000research.8997.2](https://doi.org/10.12688/f1000research.8997.2).

### Installation

To install this package, start R (version "4.2") and enter:

```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("edgeR")
```

For older versions of R, please refer to the appropriate [Bioconductor release](#).

### Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("edgeR")
```

<a href="#">PDF</a>	edgeR Vignette
<a href="#">PDF</a>	edgeRUsersGuide.pdf
<a href="#">PDF</a>	Reference Manual
<a href="#">Text</a>	NEWS



Home » Bioconductor 3.15 » Software Packages » phyloseq

## phyloseq

platforms allrank 78 / 2140support 3 / 7in Bioc 10 years

build okupdated before releasedependencies 76

DOI: [10.18129/B9.bioc.phyloseq](https://doi.org/10.18129/B9.bioc.phyloseq)

### Handling and analysis of high-throughput microbiome census data

Bioconductor version: Release (3.15)

phyloseq provides a set of classes and tools to facilitate the import, storage, analysis, and graphical display of microbiome census data.

Author: Paul J. McMurdie <joe711 at gmail.com>, Susan Holmes <susan at stat.stanford.edu>, with contributions from Gregory Jordan and Scott Chamberlain

Maintainer: Paul J. McMurdie <joe711 at gmail.com>

Citation (from within R, enter `citation("phyloseq")`):

McMurdie PJ, Holmes S (2013), "phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data." *PLoS ONE*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>.

### Installation

To install this package, start R (version "4.2") and enter:

```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("phyloseq")
```

For older versions of R, please refer to the appropriate [Bioconductor release](#).

### Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("phyloseq")
```

<a href="#">HTML</a>	<a href="#">R Script</a>	analysis vignette
<a href="#">HTML</a>	<a href="#">R Script</a>	phyloseq and DESeq2 on Colorectal Cancer Data
<a href="#">HTML</a>	<a href="#">R Script</a>	phyloseq basics vignette
<a href="#">HTML</a>	<a href="#">R Script</a>	phyloseq Frequently Asked Questions (FAQ)
<a href="#">PDF</a>		Reference Manual
<a href="#">Text</a>		NEWS

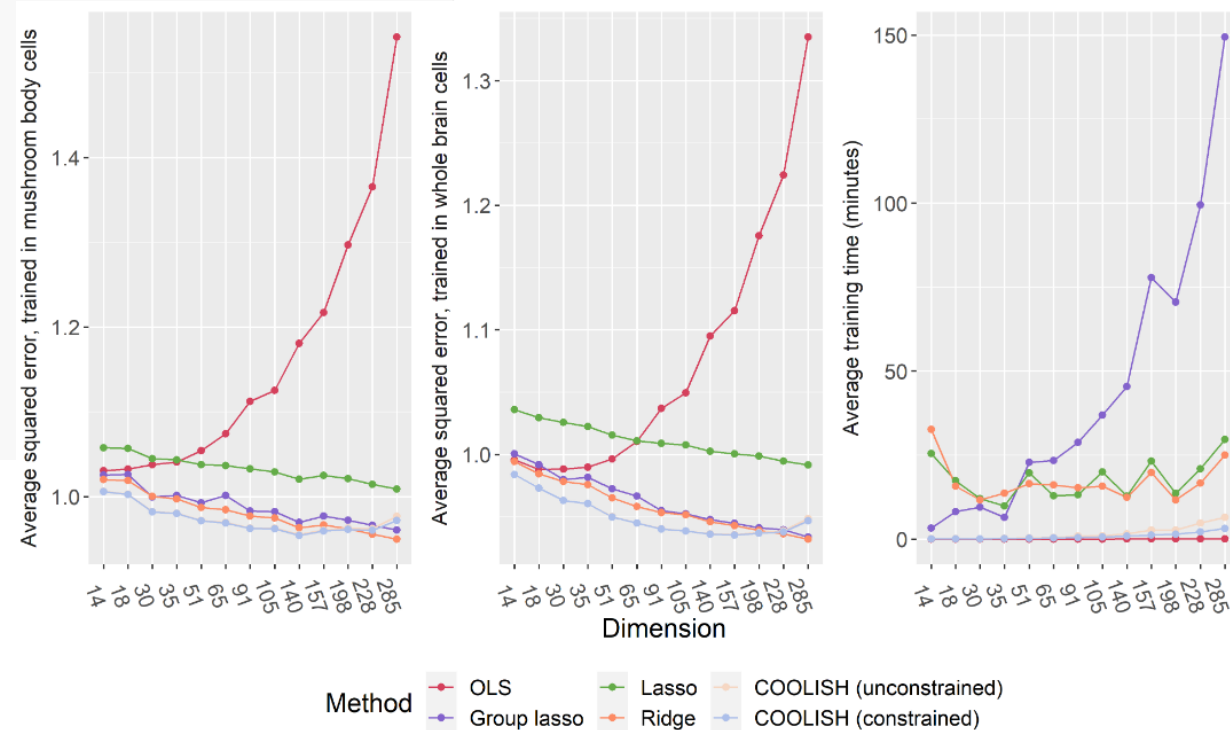
# Statistics is necessary for more complicated analyses.

## Run Principal Component Analysis

Source: `R/generics.R`, `R/dimensional_reduction.R`

Run a PCA dimensionality reduction. For details about stored PCA calculation parameters, see `PrintPCAParams`.

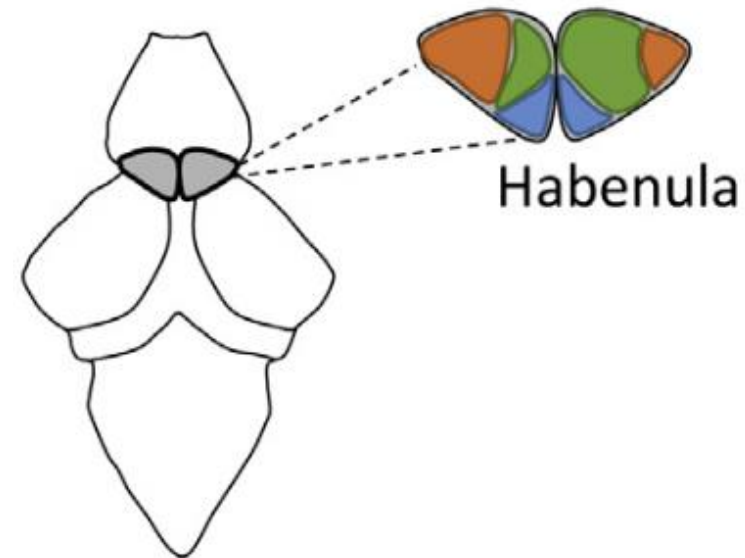
```
RunPCA(object, ...)  
  
# S3 method for default  
RunPCA(  
  object,  
  assay = NULL,  
  npcs = 50,  
  rev.pca = FALSE,  
  weight.by.var = TRUE,  
  verbose = TRUE,  
  ndims.print = 1:5,  
  nfeatures.print = 30,  
  reduction.key = "PC_",  
  seed.use = 42,  
  approx = TRUE,  
  ...  
)
```



# Role of statistics

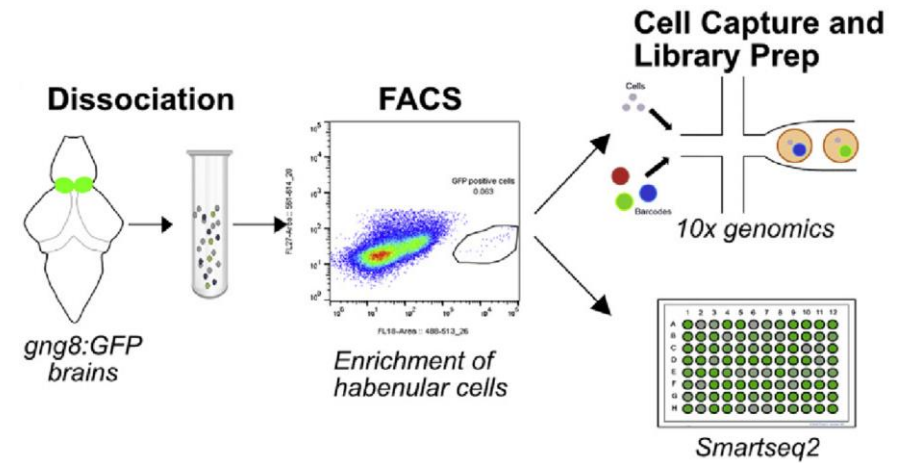
## Scientific question

How do cells in the larval zebrafish habenula coordinate their functions?



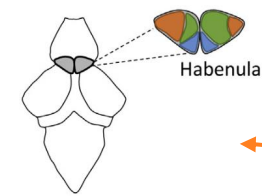
# Experimental data

Single-cell RNA-seq on 10 pooled larval zebrafish habenula.

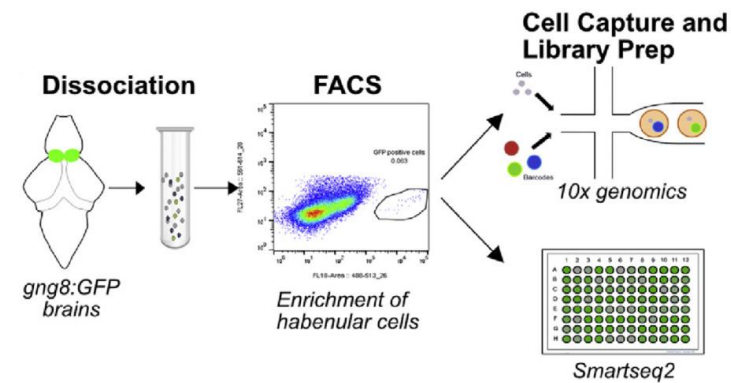


# The problem of induction

How can we make general conclusions from specific examples?

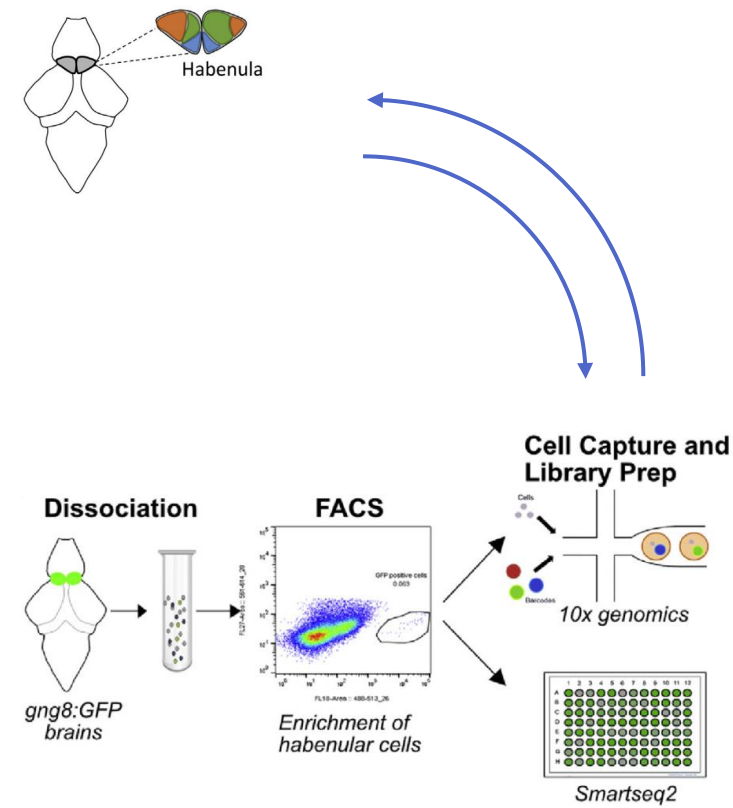


- ③ What exactly is a general conclusion?
- Why is induction justified?
- How to make inferences?
- How accurate are the inferences?



## Statistics

Statistics provides a mathematical theory of induction.





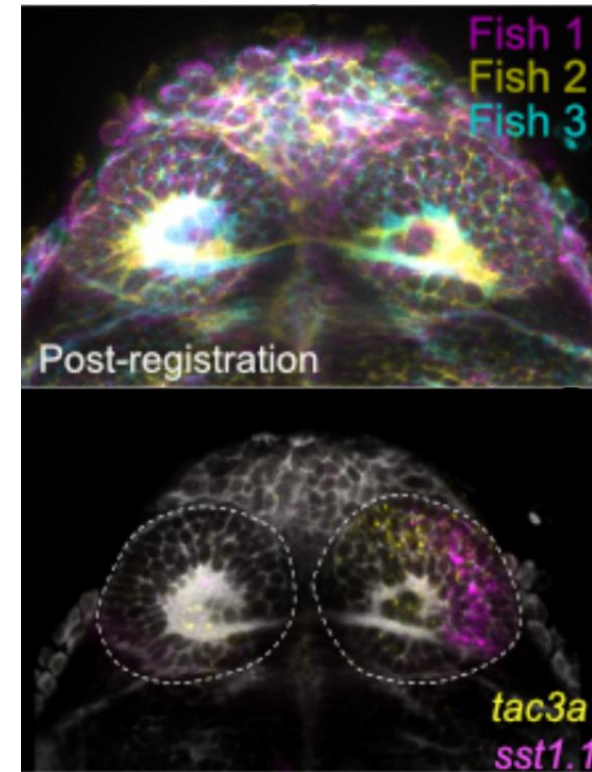
# Statistical concepts

What exactly is a general conclusion?

A general conclusion is viewed as some statement about a **population**.

---

The population is the hypothetical collection of all objects you want to generalize to, e.g., all cells in the adult zebrafish habenula.

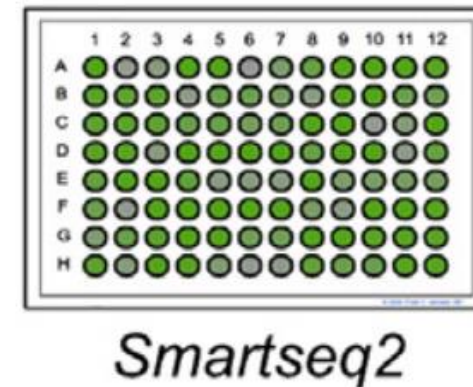
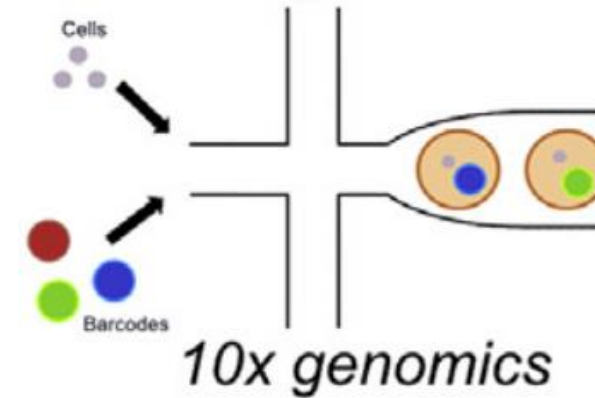


What exactly is a general conclusion?

Each member of the population can be characterized by **numerical variables**.

---

Variables need to be operationally defined and may need to be recoded numerically. Some important variables may not be directly observable.

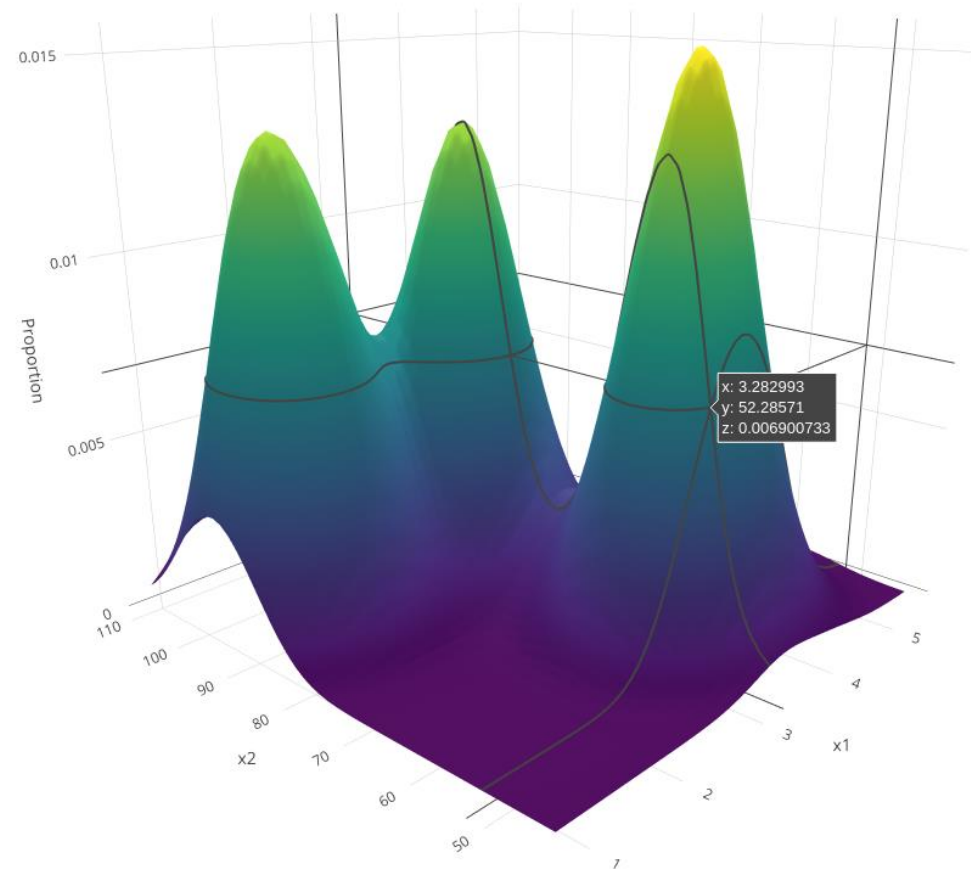


What exactly is a general conclusion?

A general conclusion describes properties of the **population distribution function** of the variables.

---

$P(X_1 = x_1, \dots, X_p = x_p) \approx$  proportion of the population whose first variable has value  $x_1$ , second variable has value  $x_2$ , ..., and  $p$ th variable has value  $x_p$ .



What exactly is a general conclusion?

There are three important  
types of properties:  
1. Define **latent variables**.

---

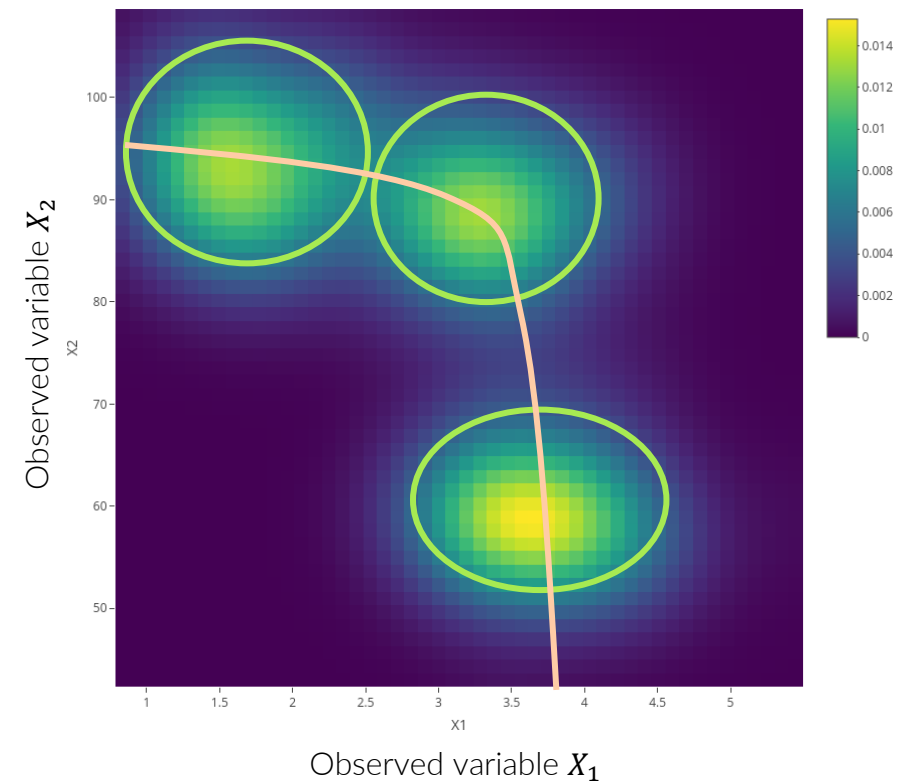
These properties define the “best”  
latent variables, according to some  
investigator-defined metric.

Latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$

Latent (factor/component) variable

$Z_2 = g_2(X_1, \dots, X_p)$

$g_1$  and  $g_2$  depend on pop. dist. function

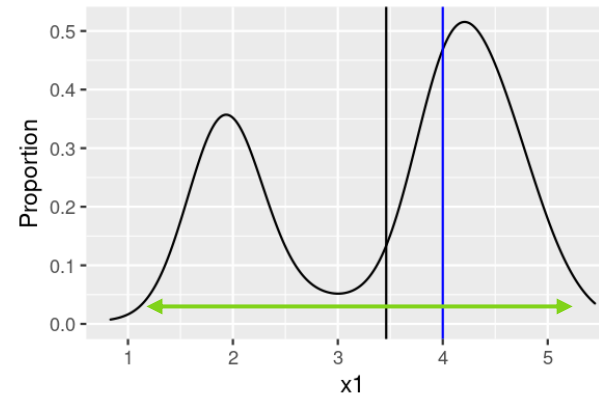
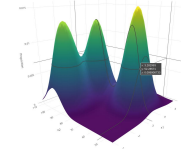


What exactly is a general conclusion?

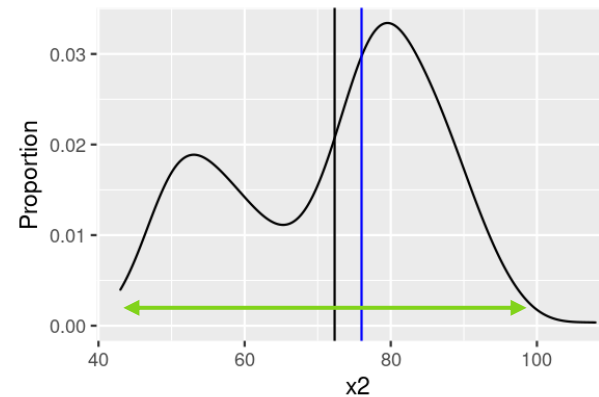
## 2. Describe **univariate** characteristics.

---

These properties include measures of central tendency, variability, etc.



Mean, **median**,  
and **standard deviation** of  $X_1$ .



Mean, **median**,  
and **standard deviation** of  $X_1$ .

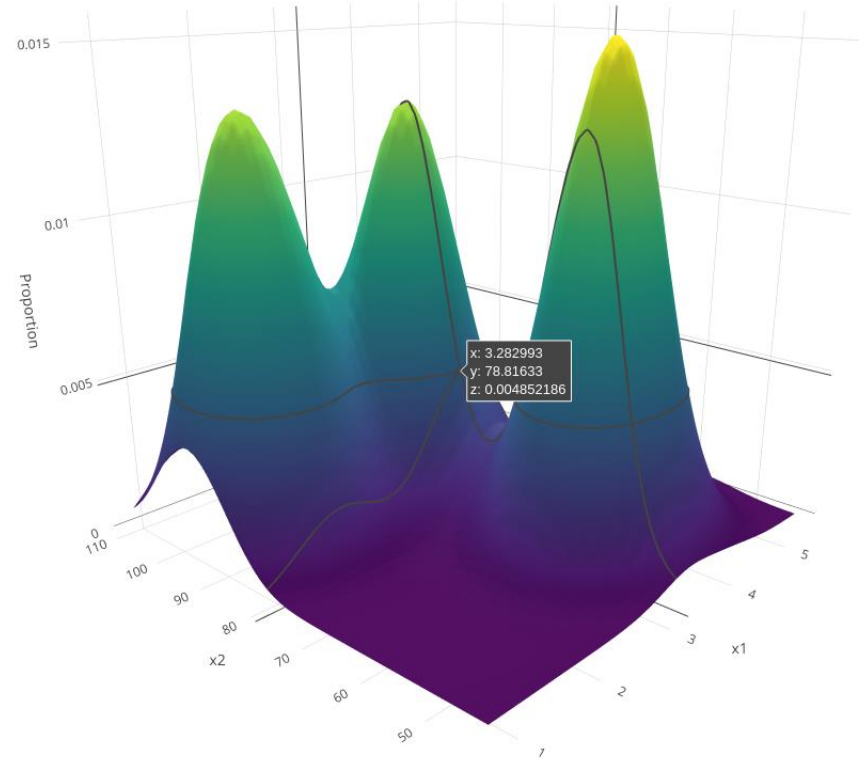
What exactly is a general conclusion?

### 3. Describe **dependence** between variables.

---

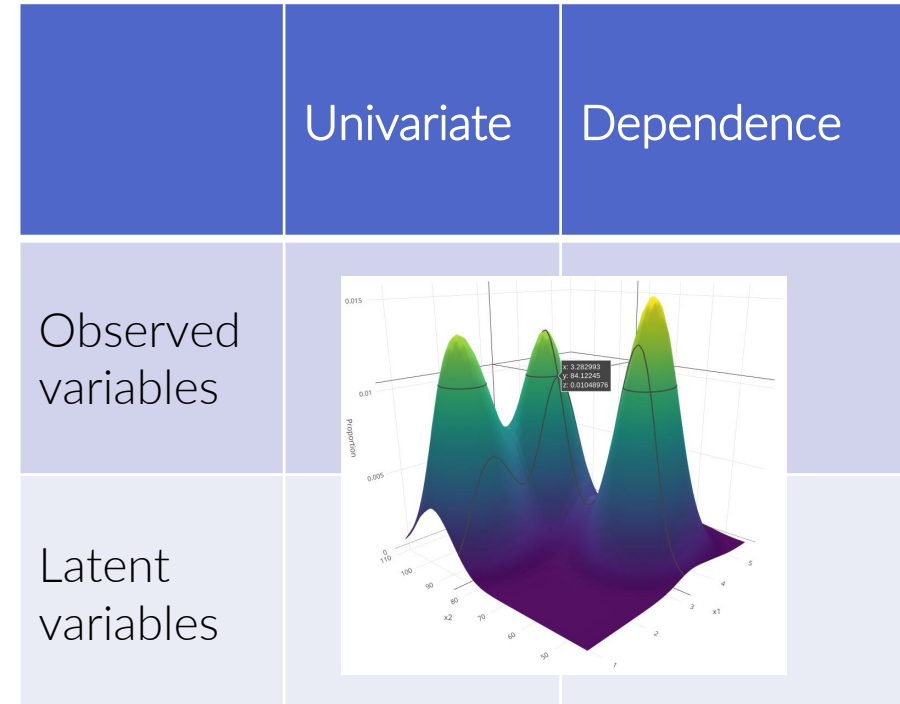
Dependence between independent and dependent variables is the most common type of multivariate characteristic.

$P(X_1 = x_1 | X_2 = x_2) \approx$  the population distribution of  $X_1$  in the subgroup of the population with  $X_2 = x_2$ :



What exactly is a general conclusion?

A general conclusion is a mathematical statement regarding **population parameters** of interest.





What exactly is a general conclusion?

Mathematical models of probability distributions impose assumptions but are easier to understand.

---

Regression modeling is a common type of mathematical model and trades stronger assumptions for less complexity in expressing dependence between variables.

Model  $P(X_1 = x_1 \mid X_2 = x_2)$  using a generalized linear model:

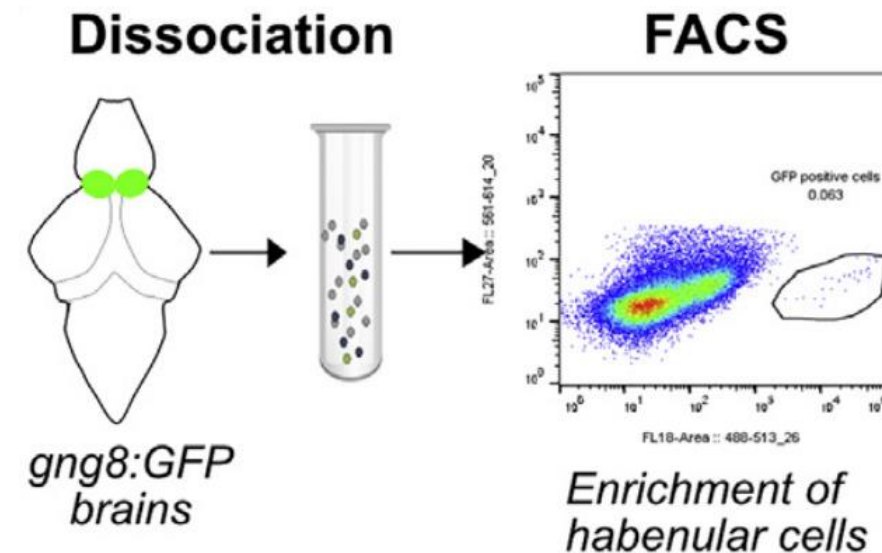
- $X_1 \mid X_2 \sim \text{NegBin}(\mu(X_2), \sigma(X_2))$
- $\log \mu(X_2) = \beta_0 + \beta_1 X_2$
- $\sigma(X_2) = \mu(X_2)(1 + \phi \mu(X_2))$

Why is induction justified?

Specific examples are viewed as being **sampled** from the population.

---

Sampling must be designed by the experimenter and should maximize information, optimize efficiency, and minimize systematic biases.



Why is induction justified?

By the laws of **probability**,  
a properly chosen sample  
will be representative of  
the population.

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| > \epsilon\right) \leq \frac{\text{var}(X_i)}{\epsilon^2 n}$$

How to make inferences?

Data are viewed as  
**numerical variables**  
measured on each sample.

---

The “samples by variables” data table  
is the fundamental unit of statistical  
analysis.

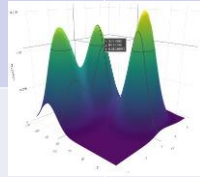
Variables (e.g., genes)

$X_1$	$X_2$		...	$X_p$
		$X_{ij}$		

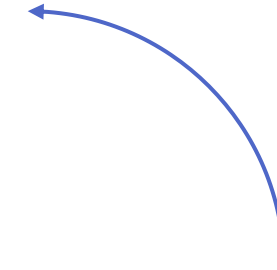
Samples (e.g., cells)

How to make inferences?

Inductive processes  
correspond to **functions** of  
the observed data.

	Univariate	Dependence
Observed variables		
Latent variables		

$$f(X_{11}, \dots, X_{np})$$



$X_1$	$X_2$		...	$X_p$
		$X_{ij}$		

How accurate are the inferences?

By the laws of **probability**,  
it is possible to quantify  
the uncertainty of  
inferences from a properly  
chosen sample.

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\text{var}(X_i)} \xrightarrow{d} N(0,1)$$

# Data analysis concepts

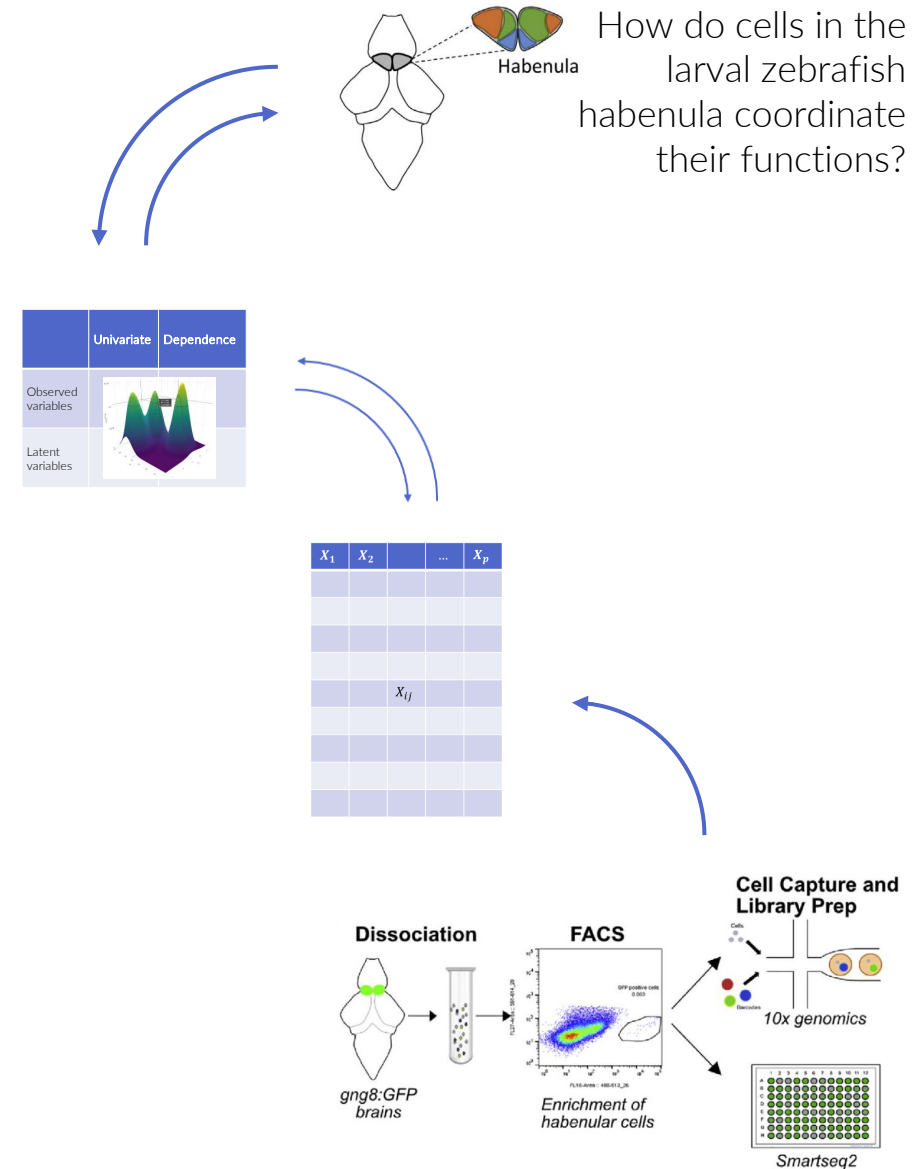
How can the **mathematical** theory of statistics help answer **explanatory** scientific questions using **biological** experimental results?





## Data analysis

Data analysis mathematizes scientific questions and experimental data and interprets statistical results.



How to express explanatory questions as math?

Pose relevant **descriptive**  
questions.

---

The descriptive questions must be  
answerable by statistical methods.

Explanatory question	Descriptive question
How do cells work?	??

How to express explanatory questions as math?

Express descriptive  
question in terms of  
population parameters.

---

Determine what types of variables  
and dependence structures the  
question is asking about.

	Univariate	Dependence
Observed variables		What genes differentiate adult zebrafish habenula cell types?
Latent variables		



Define latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$ .

For which genes  $j$  does  $P(X_j | Z_1 = z)$  change for different clusters  $z$ ?

How to convert biological results to numerical data table?

# Experimental data must be preprocessed into numerical form.

Preprocessing usually include  
quantification, quality control,  
normalization, and additional steps.

## Computational Methods for Data Analysis

### Alignment and quantification

For the 10X droplet data, raw sequencing data was converted to matrices of expression counts using the cellranger software provided by 10X genomics<sup>1</sup>. Briefly raw BCL files from the Illumina NextSeq or HiSeq were demultiplexed into paired-end, gzip-compressed FASTQ files for each channel using “cellranger mkfastq.” Both pairs of FASTQ files were then provided as input to “cellranger count” which partitioned the reads into their cell of origin based on the 16bp cell barcode on the left read. Reads were aligned to a zebrafish reference transcriptome (ENSEMBL Zv10, release 82 reference transcriptome), and transcript counts quantified for each annotated gene within every cell. Here, the 10-base pair unique molecular identifier (UMI) on the left read was used to collapse PCR duplicates, and accurately quantify the number of transcript molecules captured for each gene in every cell. Both cellranger mkfastq and cellranger count were run with default command line options. This resulted in an expression matrix (genes x cells) of UMI counts for each sample.

For SS2 data, raw reads were mapped to a zebrafish transcriptome index (Zv10 Ensembl build) using Bowtie 2 [60], and expression levels of each gene was quantified using RSEM [61]. We also mapped the reads to the Zv10 genome using Tophat2. We only used libraries with genome alignment rate > 90% and transcriptome alignment rate (exonic) > 30%. RSEM yielded an expression matrix (genes x samples) of inferred gene counts, which was converted to TPX (transcripts per 10<sup>4</sup>) values and then log-transformed after the addition of 1, consistent with the normalization of the droplet data.

### Filtering expression matrix and correcting for batch effects

Cells were first filtered to remove those that contain less than 500 genes detected and those in which > 6% of the transcript counts were derived from mitochondrial-encoded genes (a sign of cellular stress and apoptosis). Genes that were detected in less than 30 cells were also removed. Among the remaining cells, the median number of UMIs per cell was 2,279 and the median number of genes was 1,319 for larval data. The same for adult data was 1,614 UMI/cell and 709 genes/ cell, respectively (Figures S1C, S1D, S5A, and S5B).

We used a linear regression model to correct for batch effects in the gene expression matrix using the RegressOut function in the Seurat R package, and used the residual expression values for further analysis. The residual matrix was then scaled, centered and used for the selection of variable genes, PCA and clustering.

```
file = "/home/user/data/stat530_2022/scrna-seq/GSM2818521_larva_counts_matrix.txt"

larval = read.table(file, header = TRUE)
dim(larval)

library(Seurat)
## set random seed for reproducibility
set.seed(1)

obj = CreateSeuratObject(counts = larval,
                          min.cells = 30, min.features = 500)

## #####
## preprocessing
## #####
## -----
## quality control
## -----
obj[["percent.mt"]] = PercentageFeatureSet(obj, pattern = "^MT-")

VlnPlot(obj,
        features = c("nCount_RNA",
                      "nFeature_RNA",
                      "percent.mt"))

obj = subset(obj, percent.mt <= 6)

## -----
## normalization
## -----
obj = NormalizeData(obj)
obj = FindVariableFeatures(obj)
obj = ScaleData(obj, vars.to.regress = "percent.mt")
```

How to interpret statistical results?

# Bioinformatics databases can help annotate results.

Interpreting statements about population distribution functions in their scientific context is challenging.

The screenshot shows the DAVID Functional Annotation Tool interface. At the top, there's a navigation bar with links like Home, Start Analysis, and About DAVID. A welcome message states: '\*\*\* Welcome to DAVID (2021 Update) \*\*\*' and '\*\*\* If you are looking for DAVID 6.8, it is still accessible on this server until retirement on June 1, 2022. \*\*\*'. The main section is titled 'Functional Annotation Tool' and includes a 'Submit your gene list to start the tool!' button. Below this, there are sections for 'Key Concepts' including 'Term/Gene Co-Occurrence Probability', 'Gene Similarity Search', and 'Term Similarity Search'. At the bottom, there's an 'Integrated Solutions' section with a list of features: Functional Annotation, Numerous Data Sources, Co-occurrence Probability, Use Homolog Annotation, Dynamic Pathway Maps, and Disease Associations. The interface also has a sidebar with steps: Step 1: Enter Gene List, Step 2: Select Identifier, Step 3: List Type, and Step 4: Submit List.

Descriptive answer

Genes ... differentiate adult zebrafish habenula cell types.

Explanatory answer

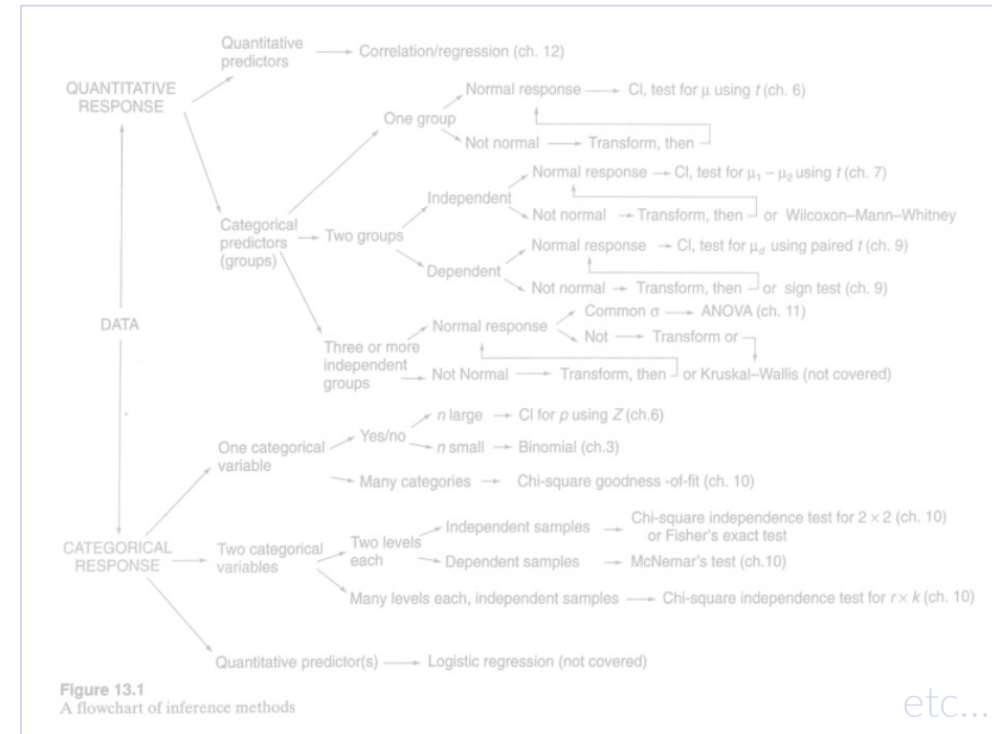
Cells in the larval zebrafish habenula coordinate their functions by...

# Inference methods

Define latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$ .  
 For which genes  $j$  does  $P(X_j | Z_1 = z)$  change for different clusters  $z$ ?

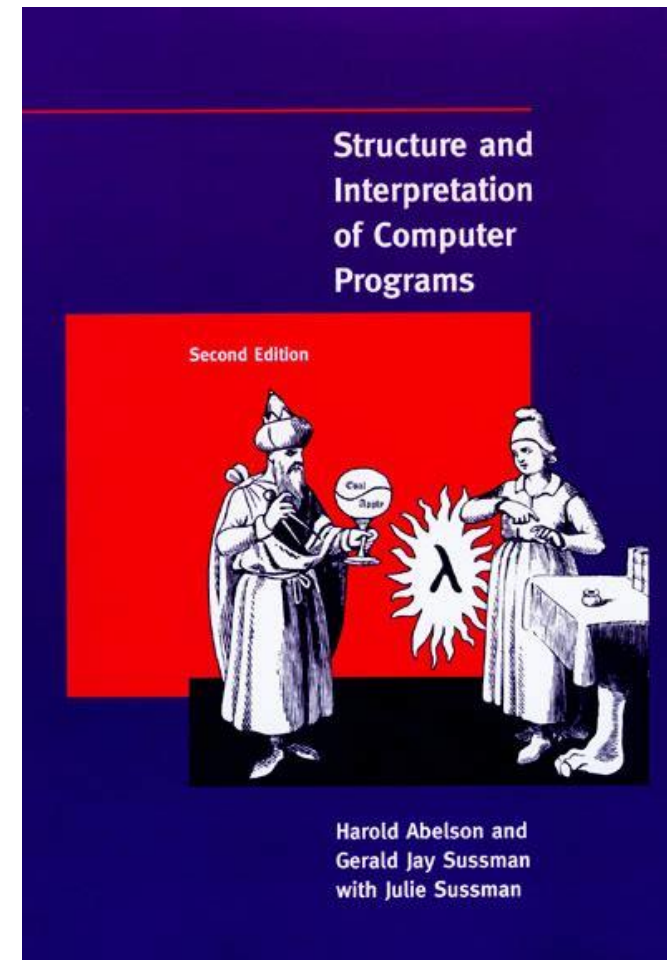
$$f(X_{11}, \dots, X_{np})$$

$X_1$	$X_2$		...	$X_p$
		$X_{ij}$		



“The language in which you'll spend most of your working life hasn't been invented yet, so we can't teach it to you. Instead we have to give you the skills you need to learn new languages as they appear.”

*Why Structure and Interpretation  
of Computer Programs matters*  
(<https://people.eecs.berkeley.edu/~bh/sicp.html>)





	Population parameters of interest						
Question to be answered	Mean	Median	Var	GLMs	...	Clusters	Factors
Testing "Is?"				Regression and classification			
Estimation "How much?"						Clustering	Dimension reduction

	Population parameters of interest						
Question to be answered	Mean	Median	Var	GLMs	...	Clusters	Factors
Testing <i>"Is?"</i>							
Estimation <i>"How much?"</i>							

Other considerations:

1. Data structure

- Variable types
- Missingness
- Censoring
- Etc.

2. Assumptions

- Parametric
- Semiparametric
- Nonparametric

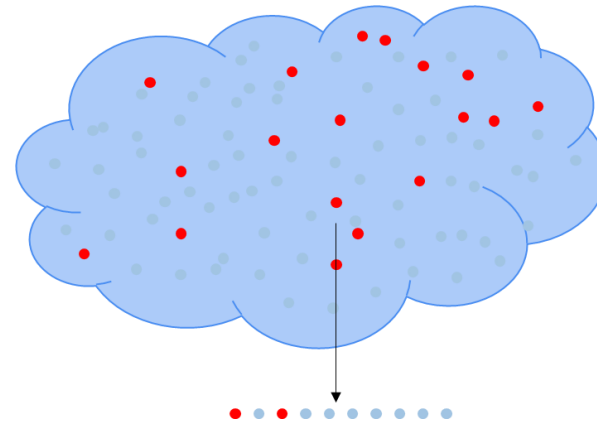
3. Culture

- Best practices
- Trends
- Etc.

Many statistical methods  
come in families indexed  
by **tuning parameters**.

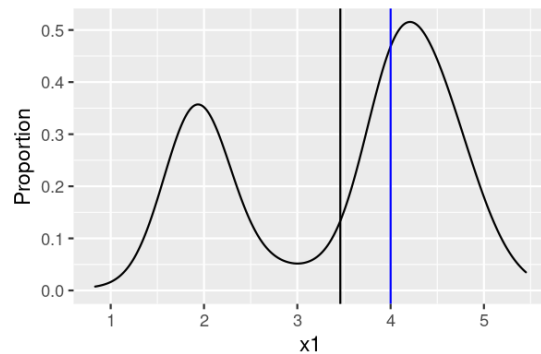
---

Tuning parameters generally trade off  
variability for bias and can be very  
difficult to choose.



$$\hat{p}_{ab} = \frac{2 + a}{10 + a + b}$$

	Population parameters of interest						
Question to be answered	Mean	Median	Var.	Depend.	...	Clusters	Factors
Testing "Is?"	Is $\mu_1 = 0$ ?						
Estimation "How much?"	What is $\mu_1$ ?						



Mean, median,  
and standard  
deviation of  $X_1$ .

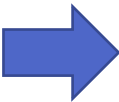
	Population parameters of interest						
Question to be answered	Mean	Median	Var.	Depend.	...	Clusters	Factors
Testing "Is?"				Is $\beta_1 = 0$ ?			
Estimation "How much?"				What is $\beta_1$ ?			

Model  $P(X_1 = x_1 \mid X_2 = x_2)$  as a generalized linear model:

- $X_1 \mid X_2 \sim \text{NegBin}(\mu(X_2), \sigma(X_2))$
- $\log \mu(X_2) = \beta_0 + \beta_1 X_2$
- $\sigma(X_2) = \mu(X_2)(1 + \phi \mu(X_2))$

	Population parameters of interest						
Question to be answered	Mean	Median	Var.	Depend.	...	Clusters	Factors
Testing "Is?"							
Estimation "How much?"							

	Univariate	Dependence
Observed variables		What genes differentiate adult zebrafish habenula cell types?
Latent variables		



Define latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$ .  
 For which genes  $j$  does  $P(X_j | Z_1 = z)$  change for different clusters  $z$ ?

Define latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$ .

## Estimate clusters using shared nearest neighbor clustering.

A cluster is a collection of samples that are more closely related to each other than to samples outside the cluster.

	Population parameters		
Question	Dep	Clusts	Factors
Testing			
Est.			

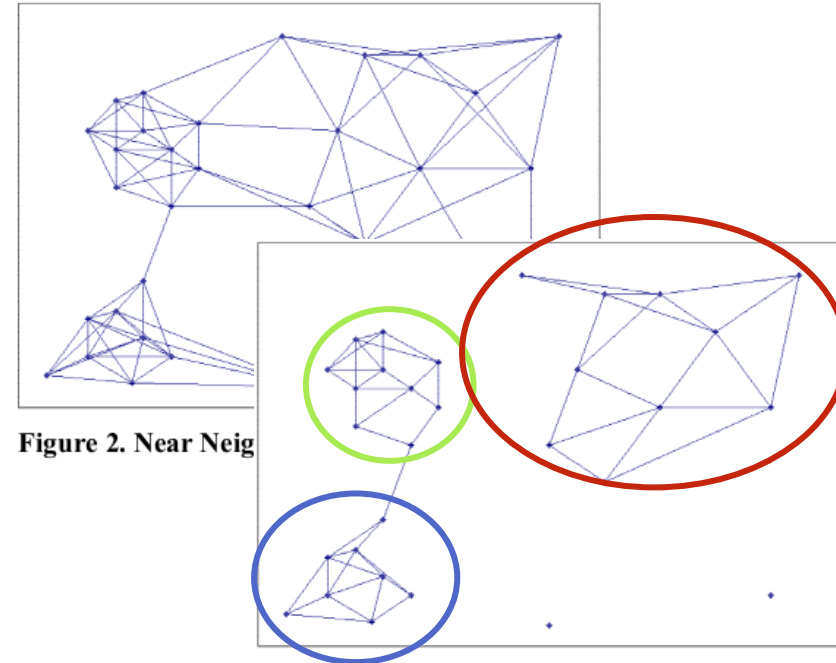


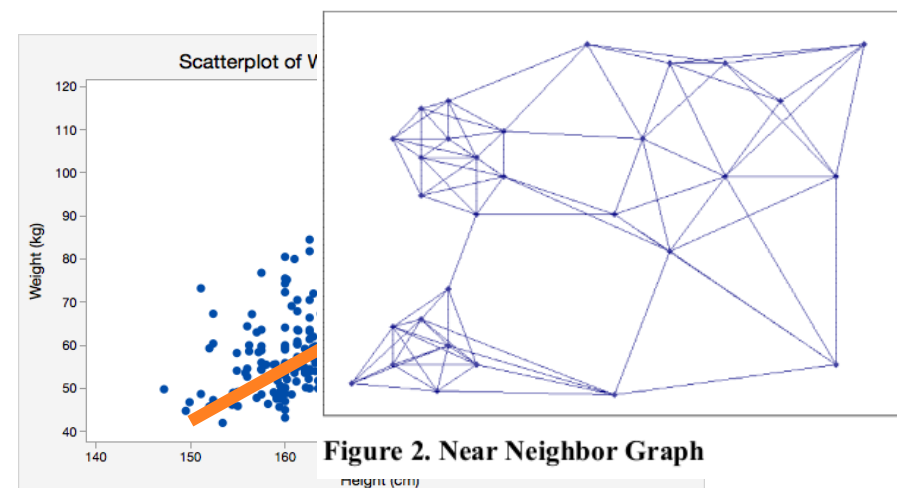
Figure 3. Unweighted Shared Near Neighbor Graph

Define latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$ .

## Construct shared nearest neighbors by estimating principal components.

The  $k$ th principal component is  $Z_k = g_k(X_1, \dots, X_p)$  where  $g_k(x) = \alpha_{k1}x_1 + \dots + \alpha_{kp}x_p$  such that  $\text{var } g_k(x)$  is maximized for  $\|a_k\|_2 = 1$  and the  $Z_k$  are uncorrelated. The number of PCs to use is a tuning parameters.

	Population parameters		
Question	Dep	Clusts	Factors
Testing			
Est.			



$X_1$	...	$X_p$	$Z_1$	...	$Z_{10}$





Define latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$ .

Choose the number of  
clusters by choosing  
resolution.

---

The resolution is a tuning parameter;  
there are no “true” clusters.

	Population parameters		
Question	Dep	Clusts	Factors
Testing			
Est.			

```
## dimension reduction  
obj = RunPCA(obj)
```

```
## clustering  
obj = FindNeighbors(obj)  
obj = FindClusters(obj,  
                    resolution = 0.5)
```

Define latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$ .

	Population parameters		
Question	Dep	Clusts	Factors
Testing			
Est.			

## Visualize clusters by estimating UMAP coordinates.

---

A UMAP coordinate is another type of latent variable  $Z_k = g_k(X_1, \dots, X_p)$  where  $g_k$  is nonlinear. It has many tuning parameters.

```
## dimension reduction  
obj = RunUMAP(obj, dims = 1:20)
```

```
## visualization  
DimPlot(obj)
```

For which genes  $j$  does  $P(X_j | Z_1 = z)$  change for different clusters  $z$ ?

## Test each gene's association with cluster using a Wilcoxon test.

---

The tests are then adjusted for multiple comparisons.

	Population parameters		
Question	Dep	Clusts	Factors
Testing			
Est.			

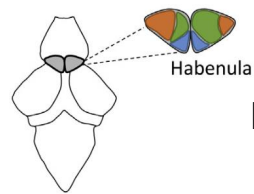
```
markers = FindAllMarkers(obj)

## view top markers for cluster 0
head(markers[markers$cluster == 0,])

## view top markers for cluster 5
head(markers[markers$cluster == 5,])

## visualize markers
FeaturePlot(obj,
            features = c("G0S2",
                        "LRRTM1"))
```

# Summary



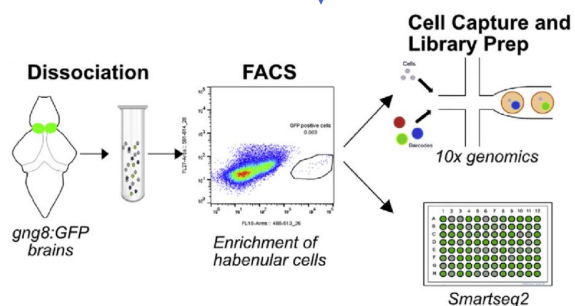
How do cells in the larval zebrafish habenula coordinate their functions?



What genes differentiate adult zebrafish habenula cell types?



Define latent (cluster) variable  $Z_1 = g_1(X_1, \dots, X_p)$ .  
For which genes  $j$  does  $P(X_j | Z_1 = z)$  change for different clusters  $z$ ?



$X_1$	$X_2$	...	$X_p$

cluster	avg_log2FC	pct.1	pct.2	p_val_adj	gene	
1:	0	3.100836	0.946	0.312	0.000000e+00	G0S2
2:	1	2.253668	0.667	0.170	2.114650e-186	TENM3
3:	2	2.110081	0.926	0.226	7.071570e-274	TSPAN18B
4:	3	1.859363	0.524	0.017	0.000000e+00	PTH2R
5:	4	2.233065	0.521	0.014	0.000000e+00	ADRB2A
6:	5	3.356025	0.868	0.029	0.000000e+00	LRRTM1
7:	6	7.112995	1.000	0.060	0.000000e+00	ICN
8:	7	4.561219	0.856	0.061	0.000000e+00	CBLN1
9:	8	3.328361	0.860	0.132	2.878485e-204	TUBB5
10:	9	2.443857	0.781	0.008	0.000000e+00	PPP1R1C
11:	10	1.500814	0.345	0.045	1.549424e-51	MAB21L2
12:	11	3.400962	0.971	0.009	0.000000e+00	GAD2
13:	12	6.452704	0.960	0.010	0.000000e+00	SI:DKEY-117I10.1