

University of Pennsylvania
BIOL4536 Fall 2023
Professor: Gregory R. Grant
Final Exam Practice Problems SOLUTIONS

PATHWAY ENRICHMENT ANALYSIS

Question 1. True or False. Suppose we are doing a pathway enrichment analysis. For a set of 200 DE genes, we calculate one enrichment p -value for each gene.

ANSWER: False. We calculate one for each gene set (pathway)

Question 2. True or False. You may have to convert gene identifiers to use a particular pathway enrichment tool.

ANSWER: True. They tend to accept some ID's and not others.

Question 3. True or False. Pathways are specific to protein coding genes.

ANSWER: False. Any type of gene can be involved in a pathway.

Question 4. True or False. In gene set enrichment analysis, there's one p -value for each of the categories of gene sets: Biological Process, Molecular Function and Cellular Component.

ANSWER: False. Those are categories of gene sets and there's a p -value for each gene set.

Question 5. Explain why you might want to upload your own "background" genes in a pathway enrichment analysis.

ANSWER: Because including genes in the background that weren't measured exaggerates p -values.

Question 6. What is one alternative algorithm to using the hypergeometric test?

ANSWER: GSEA, Gene Set Enrichment Analysis. We only talked about those two.

Question 7. True or False. Gene Set Enrichment Analysis (GSEA) utilizes a random walk to define its score.

ANSWER: True.

Question 8. The hypergeometric distribution is

- (A) Discrete
- (B) Continuous

ANSWER: Discrete.

Question 9. Saying a step taken while doing a hypothesis test is "anti-conservative" means it tends to make the p -value:

- (A) Larger
- (B) Smaller ← **THIS ONE**

(C) It doesn't mean either of those things

REASONING: Smaller. Thereby making it more likely to call something significant when it's not.

Question 10. True or False. Pathway enrichment analysis based on the hypergeometric test can be used for any list of genes, not just ones from a DE analysis.

ANSWER: True.

DIMENSIONALITY REDUCTION

Question 1. Suppose we have a spreadsheet of gene expression data with 27,132 genes (rows) from 10 subjects (columns). Circle all that are true

- (A) Each gene is a point in 27132-dimensional space.
- (B) Each subject is a point in 27132-dimensional space. ← **THIS ONE**
- (C) Each gene is a point in 10-dimensional space. ← **THIS ONE**
- (D) Each subject is a point in 10-dimensional space.

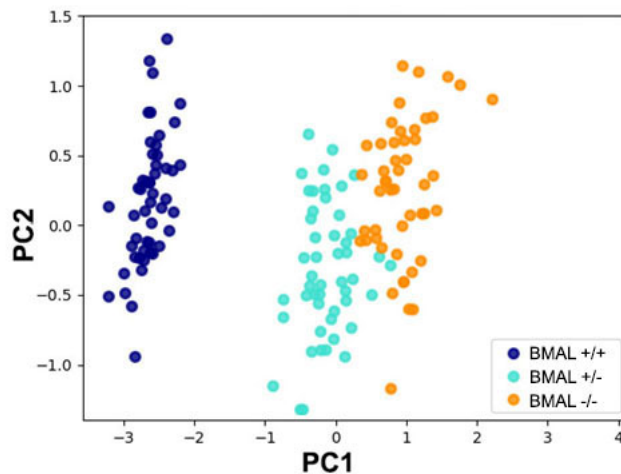
Question 2. Suppose there are 30,000 genes and we represent subjects' gene expression at these genes as points in 30,000-dimensional space. Explain why we care about the euclidean distances of these points in space.

ANSWER: Because the closer they are in space, the most similar their gene expression is. And that can reveal relationships and differences between the samples.

Question 3. True or False. It is possible for a latent variable to combine the information from all genes (dimensions) into one variable (dimension).

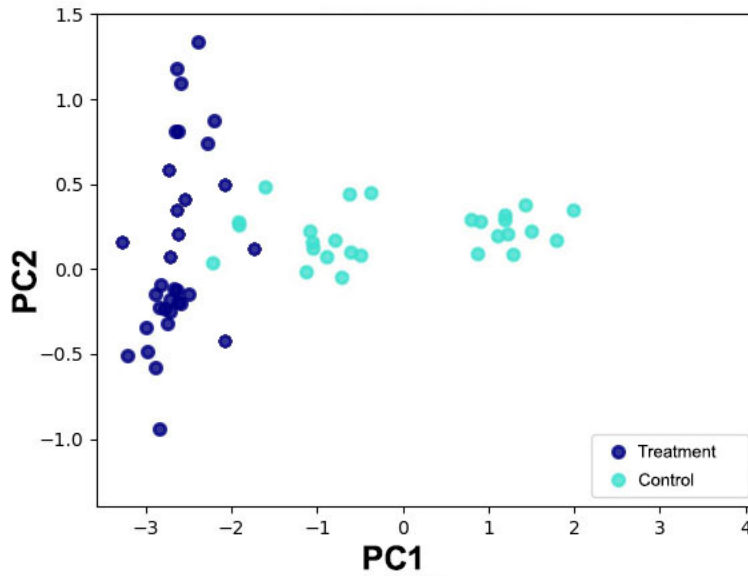
ANSWER: True. It could be that no loadings are zero.

Question 4. Interpret PC1 and PC2 in the following plot PCA plot



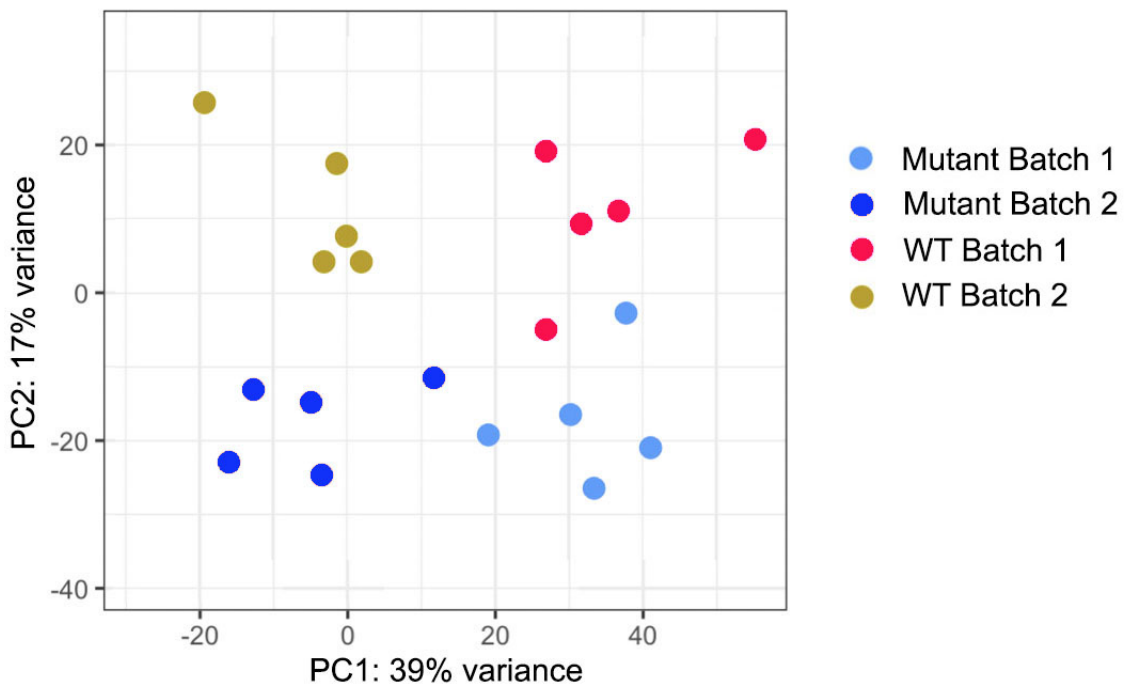
ANSWER: PC1 is driven by the difference *between* the three genotypes and PC2 is driven by the biological variation *within* the conditions.

Question 5. Suppose you have RNA-Seq data from Treatment and Control and you get the following PCA plot. Suppose the loadings for PC1 are positive only in pathway P_1 and the loadings for PC2 are positive only in pathway P_2 . Which pathway P_1 or P_2 would you expect to have differential means between treatment and control?



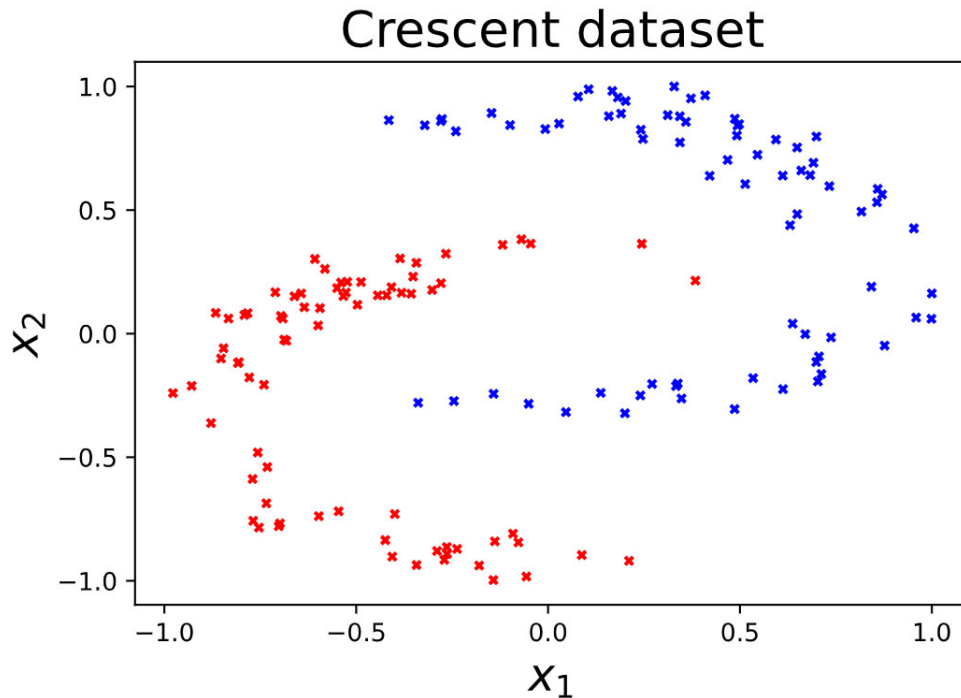
ANSWER: We see separation between treatment and control when we project onto PC1 while PC2 mixes the two groups together. So PC1 is about differential means between the two groups. Therefore, it's pathway P_1 .

Question 6. Consider the following PCA plot. Interpret the latent variable given by PC2.



ANSWER: PC2 is driven by the difference between WT and Mutant. So that latent variable captures the differential effect.

Question 7. True or False. The data in the figure below calls for non-linear methods.



ANSWER: True.

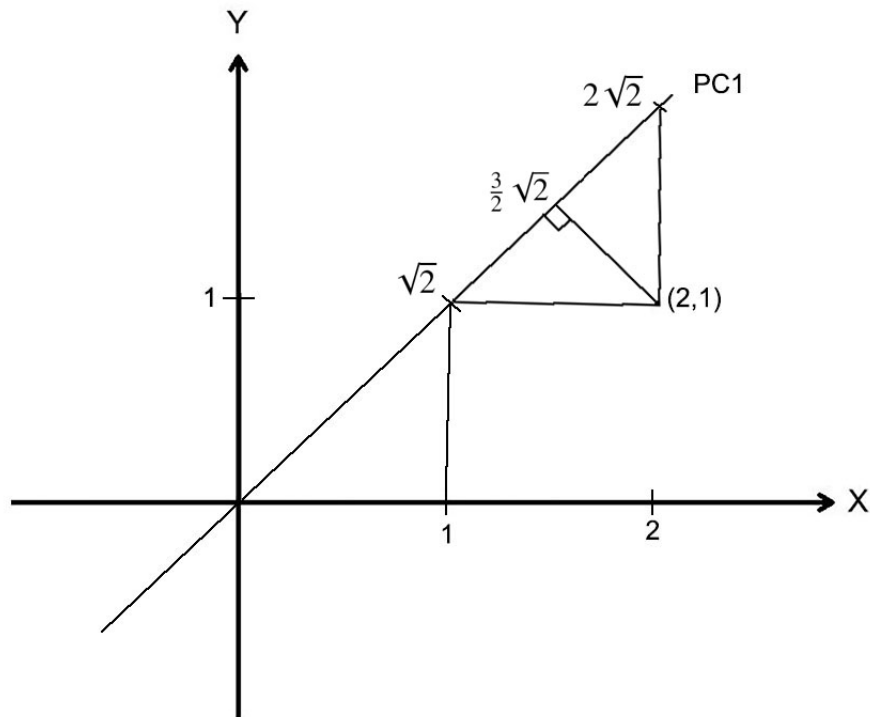
Question 8. Suppose a set of genes are differentially expressed between two experimental conditions C_1 and C_2 . Suppose the variation within each of the two conditions is greater than the variation between the two conditions. Which principle component PC1 or PC2 is more likely to capture the biological difference of interest between the conditions?

ANSWER: PC2 is more likely. Variation within condition is greater than between conditions means that PC1 (probably) captures variation within conditions and PC2 (probably) captures variance between conditions.

Question 9. Suppose we compare control mice to mice that have undergone a drug treatment. Give an example of a possible source of biological variation that is not of primary interest.

ANSWER: Gene expression could vary because of gender, or any other factor not of primary interest such as weight, age, mood, basically anything that can vary. But not measurement error or anything like that which would be technical variations not biological.

Question 10. In the diagram below, project the point (2, 1) onto the PC1 and give its value as a latent variable.



ANSWER: We did this on the board at some point. $\frac{3}{2}\sqrt{2}$. Draw in the relevant triangles, then find the $\sqrt{2}$ from pythagoras and the projection is then half way on the line between $\sqrt{2}$ and $2\sqrt{2}$.

NON-PARAMETRIC METHODS

Question 1. Suppose you have a set of observed (random) ratios and you want to test if the mean of the distribution of the ratios is one. What would be a sensible permutation of the data?


ANSWER: For each ratio flip a coin, if it's tails replace the ratio with its reciprocal.

Question 2. True or False. In a permutation test, replacing the statistic S with cS where c is a constant, does not change the p -values.

ANSWER: True. Because multiplying two numbers by a constant does not change which one is bigger.

Question 3. Consider the repeated measures data on the left spreadsheet. Does the spreadsheet on the right constitute a valid permutation for a repeated measures permutation test?

	Measurement 1	Measurement 2
Subject 1	11.3	32.4
Subject 2	1.8	2.8
Subject 3	0.2	0.7
Subject 4	2.2	7.2
Subject 5	42.4	116.3
Subject 6	1.7	5.4
Subject 7	0.8	2.1
Subject 8	23.3	52.1
Subject 9	3.4	11.2
Subject 10	1.4	4.4



	Measurement 1	Measurement 2
Subject 1	2.8	32.4
Subject 2	1.8	11.3
Subject 3	0.2	0.7
Subject 4	2.2	7.2
Subject 5	42.4	116.3
Subject 6	1.7	5.4
Subject 7	0.8	2.1
Subject 8	23.3	52.1
Subject 9	3.4	11.2
Subject 10	1.4	4.4

ANSWER: No, because Measurement 1 of Subject 1 was swapped with Measurement 2 of Subject 2. To be a valid permutation for paired data the subject has to be the same.

Question 4. True or False. In a permutation test, significance can only be obtained if the value of the statistic on the unpermuted data is greater or equal to the value of the statistic on every permutation.

ANSWER: False. It has to be greater on 95% of them (or 99% if using a 0.01 significance cutoff). But not necessarily all.

Question 5. True or False. You can apply a paired test to unpaired data, but not the other way around.

ANSWER: False. You can't do it either way.

Question 6. Suppose you have two sequences S_1 and S_2 and you align them with Smith-Waterman and the optimal local alignment score is R . Suppose you want to test if the two sequences are actually related, so in other words you want to know if R is significantly large. You want to design a permutation test. What should you permute?

ANSWER: False. Take two letters in a sequence and swap them and do that a bunch of times for both sequences. If they're related that should lower the optimal score between them leading to a significant p -value.

Question 7. Suppose we want to compare the expression of a gene between two different times of day. Give one reason why we might prefer a repeated measures design and give one reason why a repeated measures design might not be possible.

ANSWER: If there's any considerable subject-to-subject variability then we should do repeated measures. But if we have to sacrifice the organism to make the measurement then it is not possible.

Question 8. In a Mann-Whitney for four versus four replicates, recall R is the sum of the ranks in one group. What is the largest and smallest R could be?

ANSWER: The smallest is $R = 1 + 2 + 3 + 4 = 10$ and the largest is $R = 5 + 6 + 7 + 8 = 26$.

Question 9. In regression we assume the random error term ϵ is normally distributed. True or False, this assumption about ϵ constitutes a "parametric" assumption.

ANSWER: True.

Question 10. True or False. It is possible to design a test with false-positive rate equal to zero.

ANSWER: True. Just always reject nothing. Not a very useful test but it does have FP rate equal to zero.

GWAS

Question 1. True or False. In GWAS if a variant causes a disease 100% of the time, then it has low penetrance.

ANSWER: False, the opposite is true.

Question 2. True or False. GWAS is strictly for associating SNPs with pathological conditions and diseases.

ANSWER: False, it's also for associating phenotypes in healthy individuals, just as eye color.

Question 3. Consider a short contiguous stretch of genome on chromosome 10 of length 1Mb. Explain why there's at least a 75% chance that stretch of genome will not be passed on to a particular offspring in the next generation.

ANSWER: Because the offspring might get the other copy of chromosome 10 or because of crossovers get the wrong half of the same copy.

Question 4. Suppose a chromosome is 100Mb long and two SNPs are 1000 bases apart. Suppose there's one crossover on that chromosome in one generation that occurs with equal probability anywhere on the chromosome. What's the probability that the crossover happens between the two SNPs?

ANSWER: The crossover has to happen on the 1000 base stretch between the SNPs so $1000/100000000 = 0.00001$.

Question 5. True or False. The term "cis" means proximal and the term "trans" means distant.

ANSWER: True.

Question 6. Height is associated with (circle one)

- (A) One gene
- (B) Approximately 10 genes
- (C) Approximately 100 genes
- (D) Approximately 1000 genes
- (E) At least 1/3 of all genes ← **THIS ONE**

Question 7. True or False. GWAS is strictly for associating traits with non-synonymous SNP's in protein coding genes.

ANSWER: False, it works for all types of SNPs.

Question 8. True or False. Determining the causative SNPs is called the "Fine Mapping Problem".

ANSWER: True.

Question 9. True or False. GWAS associates SNPs with phenotypes thereby revealing their mechanism of action.

ANSWER: False. It reveals only associations, possibly even causative associations, but that's a long way from uncovering their mechanism of action.

Question 10. For multiple testing corrections in GWAS, both FWER and FDR approaches are taken in practice.

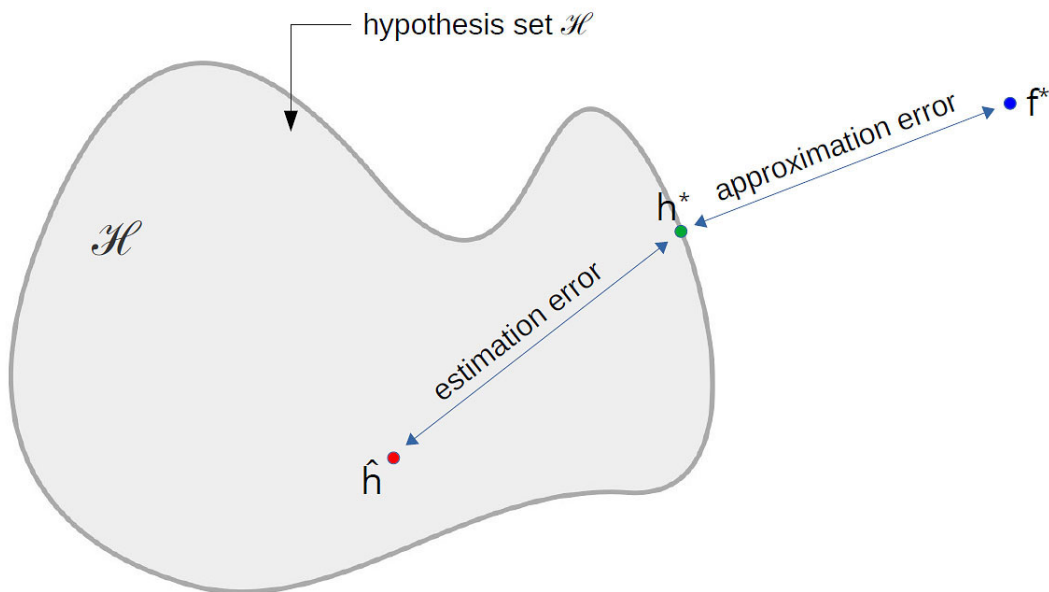
ANSWER: True.

MACHINE LEARNING

Question 1. True or False. We need training data where we know the truth to do both supervised and unsupervised learning.

ANSWER: False. Training data is for supervised learning.

Question 2. True or False. If the learning algorithm finds h^* then $\hat{h} = h^*$.

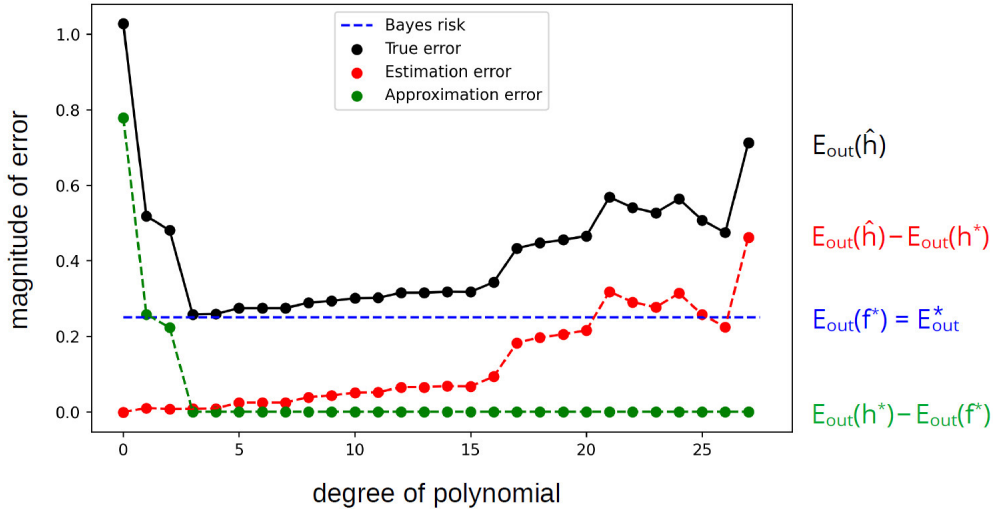
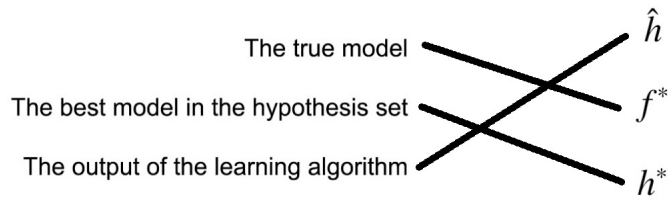


ANSWER: True. \hat{h} is the function returned by the learning algorithm.

Question 3. True or False. If the Bayes' Risk of a model is zero, then there's a deterministic relationship between the independent and dependent variables.

ANSWER: True.

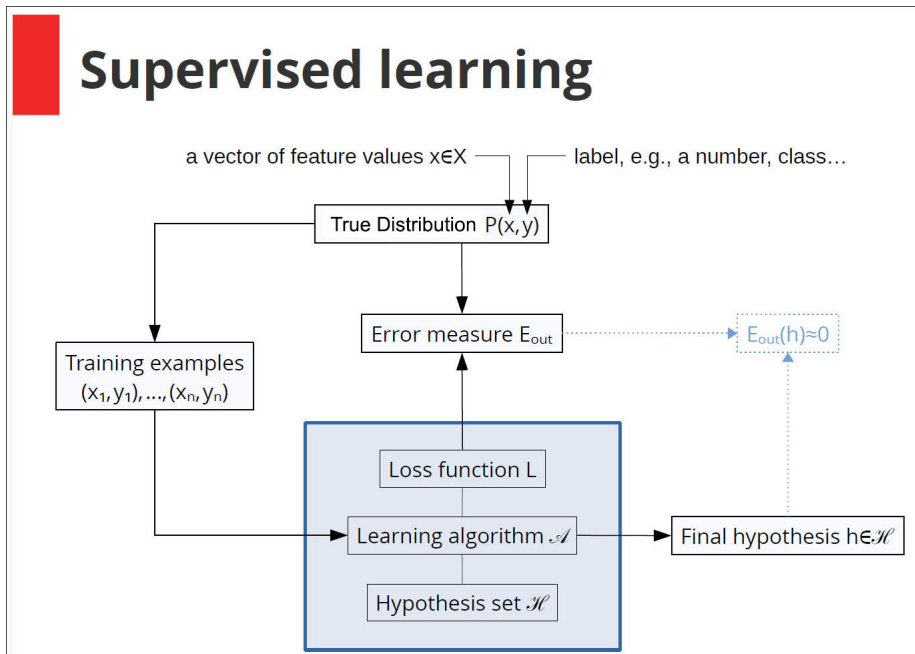
Question 4. Referring to the figure below, connect the things on the left to the corresponding things on the right.



Question 5. True or False. The 0-1 loss function is for a dependent variable which is categorical and quadratic loss is for a dependent variable which is continuous.

ANSWER: True.

Question 6. Refer to the diagram below. True or False, the Learning algorithm always finds the best (most accurate) function in \mathcal{H} .



ANSWER: False. It tries to find it but cannot guarantee it.

Question 7. Consider the linear model with n independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

To train the model we need at least

- (A) 2 data points.
- (B) $n - 1$ data points.
- (C) $n + 1$ data points. ← **THIS ONE**
- (D) n^2 data points.

Question 8. True or False. In supervised learning if, after evaluation using the test data, the model is further refined, then the test data needs to be replaced by the training data to do final evaluation of the model.

ANSWER: False. The test data needs to be replaced by entirely new data.

Question 9. In supervised learning, let f^* be the true model of the data (as usual). True or False, $E_{\text{in}}(f^*)$ necessarily equals $E_{\text{out}}(f^*)$.

ANSWER: False. $E_{\text{in}}(f^*)$ is calculated from the training data and changes if we get new data, while $E_{\text{out}}(f^*)$ is the true error, it does not depend on any data.

Question 10. True or False. Overfitting is only a problem in supervised learning, not unsupervised.

ANSWER: False. It happens in both.

Question 11. In k -means clustering, the iterative part is to:

- (A) Determine k .
- (B) Determine the clusters. ← **THIS ONE**

R ANALYSIS

Question 1. What is the name of the R function used to download and install packages from CRAN?

ANSWER: `install.packages()`

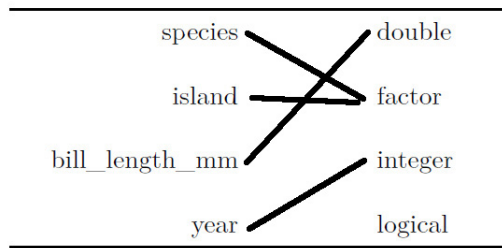
Question 2. Circle the R 'less than or equals' operator:

- (A) `<=` ← **THIS ONE**
- (B) `<`
- (C) `< -`
- (D) `==`

Question 3. Here are the first six lines from a text file:

```
species,island,bill_length_mm,year
Adelie,Torgersen,39.1,2007
Adelie,Torgersen,39.5,2007
Adelie,Torgersen,40.3,2007
Adelie,Torgersen,NA,2007
Adelie,Torgersen,36.7,2007
```

For each column name, draw a line connecting it with the most appropriate R data type to represent the values contained in that column. *Note: You may not need to use all data types, and you may need to draw lines connecting multiple columns to the same data type.*



Question 4. Which of the following lines of code will return the 'species', 'island' and 'year' columns from the penguins data frame? (circle all that apply)

- (A) `select(species, island, year, penguins)`
- (B) `select(penguins, species, island, year) ← THIS ONE`
- (C) `select(species, island, year, .data=penguins) ← THIS ONE`
- (D) `select(penguins=.data, species, island, year)`

Question 5. Consider the following table of differential expression results:

```
# A tibble: 6 x 4
  gene_id      baseMean log2FoldChange  pvalue
  <chr>          <dbl>          <dbl>    <dbl>
1 ENSMUSG00000026822  23019.          9.87  0
2 ENSMUSG00000040026  22475.          9.41  0
3 ENSMUSG00000021091  50363.          4.39  1.25e-285
4 ENSMUSG00000057465  50544.          8.50  1.44e-240
5 ENSMUSG00000016024   3589.          2.47  1.45e-193
6 ENSMUSG00000051439   934.           6.89  2.99e-188
```

Fill in the blanks to make this code perform a Benjamini-Hochberg multiple-testing correction on the data in the 'pvalue' column and store the result in a new column named 'padj'

```
il1b_de_results |>
  mutate(padj = p.adjust(pvalue, method = "BH"))
```

Question 6. Consider the following *ggplot2* code:

```
mtcars |>
  ggplot(aes(x = mpg,
             y = qsec)) +
  geom_point(color = "red")
```

Is this code an example of **mapping** or **setting** the 'color' aesthetic?

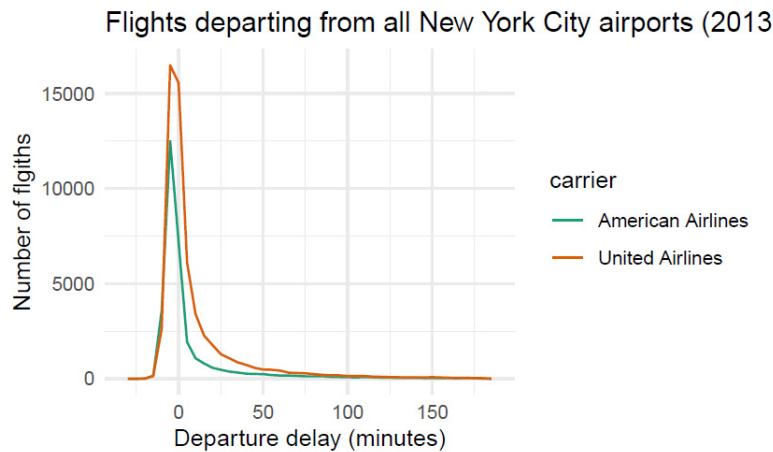
ANSWER: setting

Question 7. Which `geom_` function is designed to plot distributions of categorical variables?

- (A) `geom_histogram()`
- (B) `geom_freqpoly()`
- (C) `geom_density()`
- (D) `geom_bar()` ← **THIS ONE**

Question 8. Consider this table and graph:

```
# A tibble: 6 x 6
  flight carrier      sched_dep_time dep_delay sched_arr_time arr_delay
  <int> <chr>          <int>      <dbl>      <int>      <dbl>
1  1545 United Airlines      515         2         819         11
2  1714 United Airlines      529         4         830         20
3  1141 American Airlines     540         2         850         33
4  1696 United Airlines     558        -4         728         12
5   301 American Airlines     600        -2         745          8
6   194 United Airlines     600        -2         917          7
```



Which column(s) from the input table would you need to create the graph? (circle all that apply)

- (A) `flight`
- (B) `carrier` ← **THIS ONE**
- (C) `sched_dep_time`
- (D) `dep_delay` ← **THIS ONE**
- (E) `sched_arr_time`
- (F) `arr_delay`

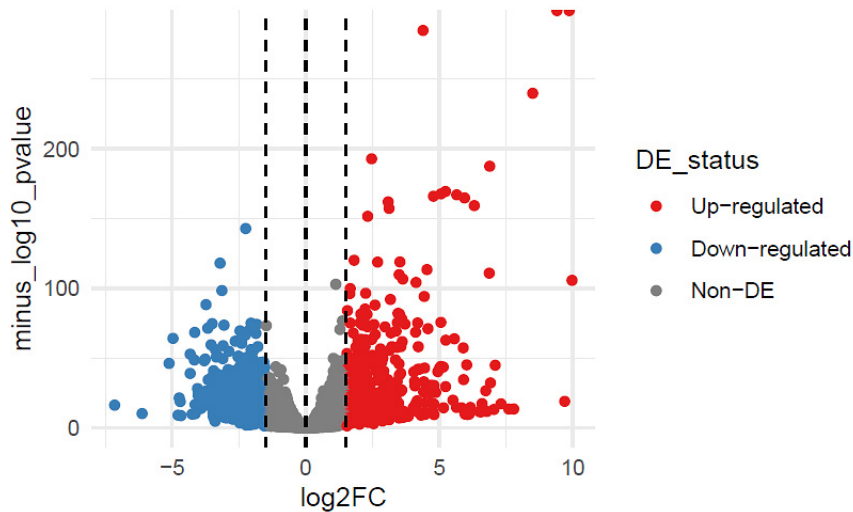
Question 9. Consider the same table and graph from question 8. Which filter function code could we use to limit the input data for this graph to just United Airlines flights?

- (A) `filter(carrier==United Airlines)`
- (B) `filter(carrier='United Airlines')`
- (C) `filter('United Airlines')`
- (D) `filter(carrier=='United Airlines')` ← **THIS ONE**

Question 10. Consider the 'de_results' data frame of differential expression results:

```
# A tibble: 6 x 6
  gene_id      log2FC  pvalue  padj  minus_log10_pvalue DE_status
  <chr>      <dbl>  <dbl>  <dbl>  <dbl> <fct>
1 ENSMUSG00000058006  1.12  1.90e-11 1.63e-10      10.7 Non-DE
2 ENSMUSG00000021336 -1.32  1.34e- 8 7.94e- 8       7.87 Non-DE
3 ENSMUSG00000011158 -0.199 1.30e- 1 1.93e- 1       0.885 Non-DE
4 ENSMUSG00000032085  0.246 1.47e- 1 2.14e- 1       0.833 Non-DE
5 ENSMUSG00000004364  0.0278 8.19e- 1 8.64e- 1       0.0866 Non-DE
6 ENSMUSG00000113428  0.0823 8.22e- 1 8.66e- 1       0.0853 Non-DE
```

Here's a volcano plot made from the 'de_results' data frame:



Fill in the aesthetic mappings you would need to create this volcano plot:

```
aes(x      = log2FC,
     y      = minus_log10_pvalue,
     color  = DE_status)
```

Question 11. Which R function do you use to access the 'samples' component of a SummarizedExperiment object?

- (A) assay()
- (B) colData() ← **THIS ONE**
- (C) rowData()
- (D) metadata()

Question 12. The following code generates an error:

```
filter(penguins, species = "Gentoo")
```

```
Error in `filter()`:  
! We detected a named input.  
i This usually means that you've used `=` instead of `==`.  
i Did you mean `species == "Gentoo"`?
```

Based on the error message, re-write this code so it runs correctly and returns all rows of from Gentoo penguins.

ANSWER: `filter(penguins, species == "Gentoo")`

Answer note: For the expression in the filter function, we need to compare the values in the 'species' column to the string "Gentoo". For comparisons, we use relational operators, like "==". The error was caused because the code originally used a single equal sign ("="; used to assign values to arguments), instead of a double equal sign ("==" ; a relational operator returning TRUE whenever values on its left and right side match).

Question 13. The following code generates an error:

```
ggplot(covid_testing, aes(x = age)) + Geom_histogram()
```

```
Error in Geom_histogram(): could not find function "Geom_histogram"
```

Based on the error message, re-write this code so it runs correctly and generates a histogram showing the age distribution for patients taking COVID-19 tests.

ANSWER: `ggplot(covid_testing, aes(x = age)) + geom_histogram()`

Answer note: R is a case-sensitive language. The error in the original code happened because it was trying to call the "Geom_histogram" function, beginning with a capital "G". This function does not exist in the ggplot2 package, which is why the error message states it cannot find it. The correct function is "geom_histogram", beginning with a lower-case "g".

Question 14. Which R package contains functions for reading data from raw text files?

- (A) readxl
- (B) readr ← **THIS ONE**
- (C) dplyr
- (D) tidyr

Question 15. Draw a line connecting the *ggplot2* function to its role in creating a figure.

Function	Role in figure creation
labs()	Create the canvas on which everything else is painted
ggplot()	Set the main figure title and axis titles
aes()	Paint data on the figure as a scatterplot
geom_point()	Map columns from the input data to visual properties of the figure

Question 16. Consider the following function definition:

```
get_de_genes <- function(de_results,
                          de_method,
                          q_value_cutoff = 0.05,
                          log2_fc_cutoff = 1) {
  # Body of the function
}
```

If we call the function with this code:

```
input_data |>
  get_de_genes(log2_fc_cutoff = 2,
              "limma")
```

What value is assigned to the 'de_method' argument?

- (A) contents of 'input_data'
- (B) 2
- (C) 0.05
- (D) 'limma' ← **THIS ONE**
- (E) 1

Question 17. Consider the two tibbles:

Tibble A:

```
# A tibble: 24 x 3
  gene_name sample_id  read_counts
  <chr>      <chr>          <int>
1 Lcn2      Saline_9574      63
2 Lcn2      Saline_9575      41
3 Lcn2      IL1B_9577       39976
4 Lcn2      IL1B_9578      44056
5 Ido2      Saline_9574     1734
6 Ido2      Saline_9575     1129
# i 18 more rows
```

Tibble B:

```
# A tibble: 6 x 5
  gene_name Saline_9574 Saline_9575 IL1B_9577 IL1B_9578
  <chr>          <int>      <int>      <int>      <int>
1 Lcn2             63         41       39976     44056
2 Ido2            1734       1129        280        230
3 Fam83a           6           5          94         210
# i 3 more rows
```

List the number of rows present in each tibble:

Tibble A: _____

Tibble B: _____

ANSWER: Tibble A: 24, Tibble B: 6

Question 18. There is a blank in the following filter expression:

```
filter(penguins,
       species == "Gentoo" ___ species == "Chinstrap")
```

Which operator would go in the blank to make this expression return all rows with data from “Gentoo” or “Chinstrap” penguins?

- (A) &
- (B) | >
- (C) | ← **THIS ONE**
- (D) +