University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

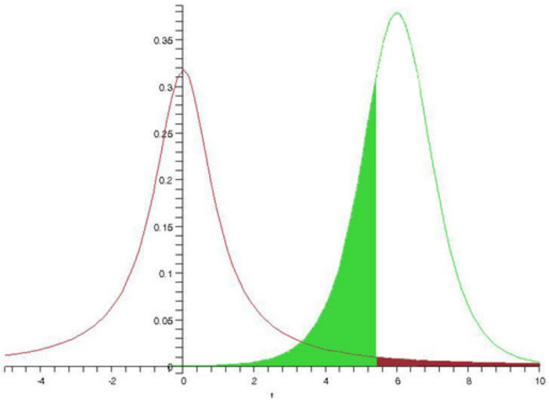
Exam#1 (PRACTICE TEST SOLUTIONS)

Question 1.

Suppose you have a test for a disease that has false positive rate of 0.01 (so 1%). Suppose you test a population of 1000 individuals, 200 of which have the disease and 800 of which do not. How many false positives do you expect?

ANSWER: $800 \cdot 0.01 = 8$. Only a negative can be a false-positive.

Question 2. This graph shows the null and alternate distributions for a T-test.



Which distribution is the null hypothesis distribution?
(A) The one on the left $\longleftarrow$ **THIS ONE**
(B) The one on the right

The red shaded area represents what?
(A) The False-Negative Rate
(B) The p-value $\longleftarrow$ **THIS ONE**
(C) The Power
(D) The probability that we reject the alternative hypothesis

What does the green shaded area represent? Will accept any of the standard terms for it, or you can explain it in your own words.

ANSWER: The "Power" of the test. Would also accept "Type II" error or the "False-Negative Rate"

Question 3.

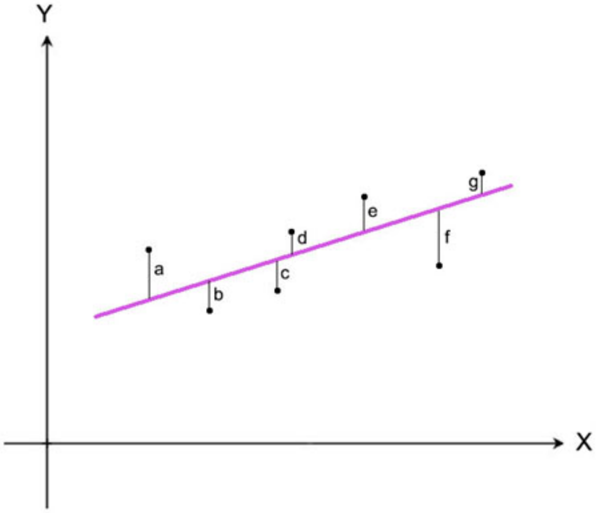
Which of the following are assumptions of a T-test, circle all that apply.

(A) Observations are normally distributed. ⟵ **THIS ONE**
(B) The means of the two groups are not equal.
(C) Variance is equal in both conditions. ⟵ **THIS ONE**
(D) The null hypotheses is false.

Question 4. Linear regression is called "linear" because:

(A) It is linear in the independent variable
(B) It is linear in the coefficients ⟵ **THIS ONE**
(C) The regression curve is a straight line
(D) $\beta_0 \leq \beta_1$
(Circle the correct answer)

Question 5.



This algorithm to estimate the regression line is called:

(A) Minimal Slope Derivation
(B) Maximal Conjunction
(C) Anterior Magnus Ambulation
(D) Least Squares ⟵ **THIS ONE**

Write down the formula of the lengths a, b, c, d, e, f and g that we minimize in order to estimate the regression line.

ANSWER: $a^2 + b^2 + c^2 + d^2 + e^2 + f^2 + g^2$.

Question 6.

Consider the linear regression model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

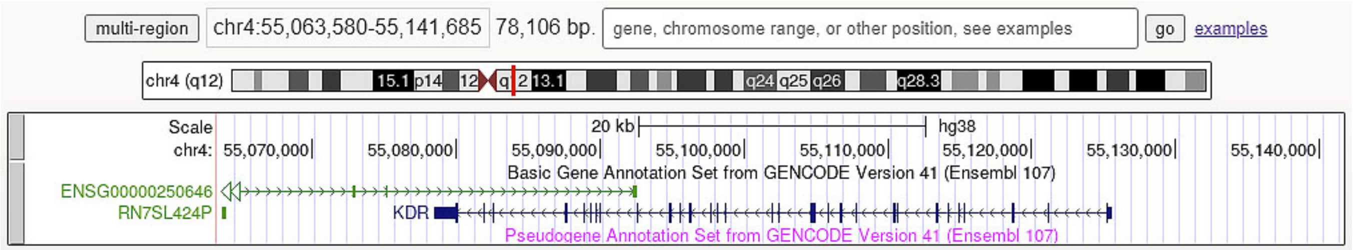
Assuming $\beta_2 > 0$, what is the shape of the regression curve? (The name of the type of curve.)

ANSWER: It's a parabola (specifically $Y = \beta_0 + \beta_1 X + \beta_2 X^2$).

What is the mean of ϵ?

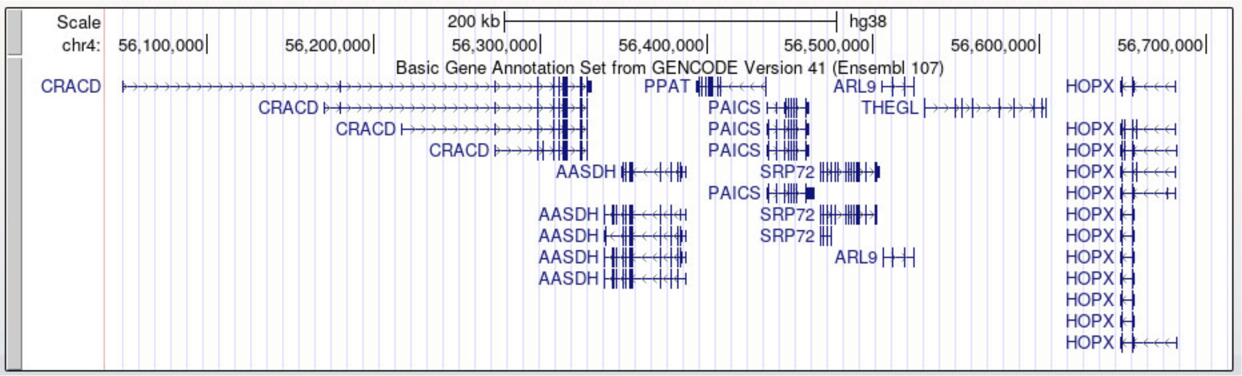
ANSWER: Zero. (That's always an assumption of linear regression.)

Question 7.


multi-region
chr4:55,063,580-55,141,685
78,106 bp.
gene, chromosome range, or other position, see examples
go
examples
chr4 (q12)
15.1
p14
12
q
2
13.1
q24
q25
q26
q28.3
Scale
chr4:
20 kb
hg38
55,070,000|
55,080,000|
55,090,000|
55,100,000|
55,110,000|
55,120,000|
55,130,000|
55,140,000|
Basic Gene Annotation Set from GENCODE Version 41 (Ensembl 107)
ENSG00000250646
RN7SL424P
KDR
Pseudogene Annotation Set from GENCODE Version 41 (Ensembl 107)

The gene shown KDR is on chr4:
(A) Near the telomere of the short arm of the chromosome
(B) Near the telomere of the long arm of the chromosome
(C) Near the centromere on the short arm of chromosome
(D) Near the centromere on the long arm of chromosome ⟵ **THIS ONE**

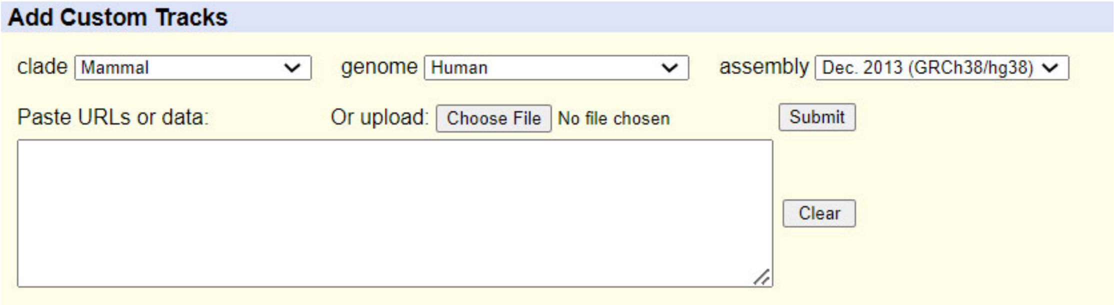
Question 8.



How many different *genes* are show here?

ANSWER: Eight. CRACD, AASDH, PPAT, PAICS, SRP72, ARL9, THEGL and HOPX

Question 9.



Add Custom Tracks

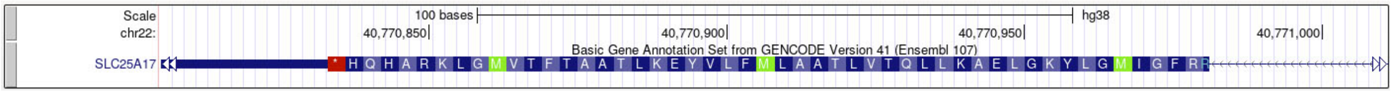
clade Mammal genome Human assembly Dec. 2013 (GRCh38/hg38)

Paste URLs or data: Or upload: Choose File No file chosen

This is the form to enter a custom track in the genome browser. What do you enter in the box to create a span on chromosome 10 from base 10,000,000 to base 20,000,000?

ANSWER: chr10 9999999 20000000

Question 10. What is the red box at the end of the CDS?

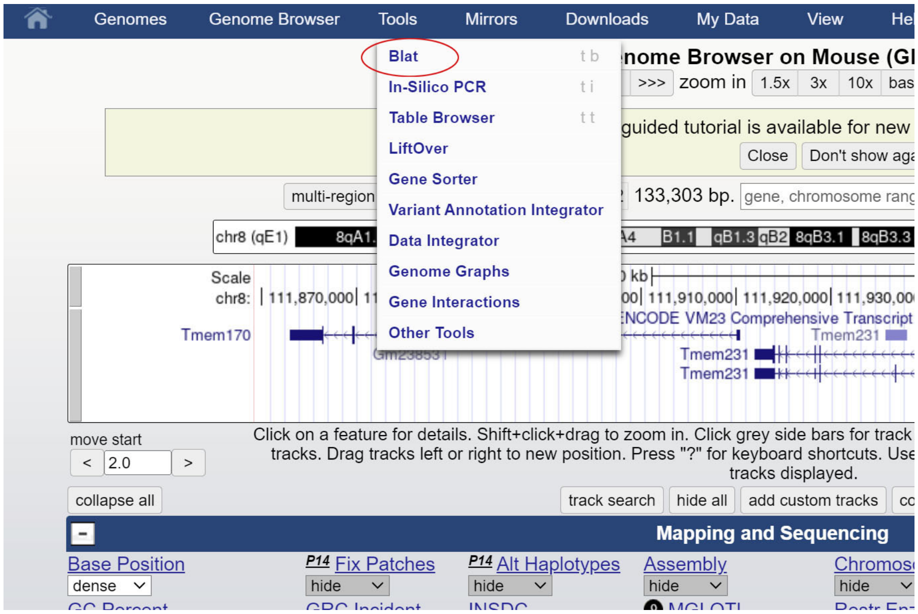

Scale
100 bases
hg38
chr22:
40,770,850
40,770,900
40,770,950
40,771,000
Basic Gene Annotation Set from GENCODE Version 41 (Ensembl 107)
SLC25A17
* H Q H A R K L G M V T F T A A T L K E Y V L F M L A A T L V T Q L L K A E L G K Y L G M I G F R R

ANSWER: That’s the stop codon.

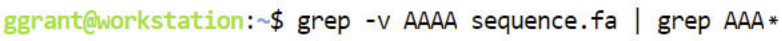
Question 11. Circle all that are necessarily true. A 3’ UTR

(A) is the first exon.
(B) is contained in the first exon.
(C) contains the first exon.
(D) none of the above are necessarily true. ⟵ **THIS ONE**

Question 12. Which of these genome browser Tools would you use to align a high-throughput sequencing read to the genome? (circle the right one)



Question 13. What does the following grep command return?



```
ggrant@workstation:~$ grep -v AAAA sequence.fa | grep AAA*
```

ANSWER: All rows that have at three `A`’s in a row and not four or more `A`’s in a row.

Question 14.

Write down the grep command to search for sequences in a file named `data.txt` that contain the string PCGA, but do not contain the string PCGAQ.

ANSWER: `grep PCGA data.txt | grep -v PCGAQ`

Question 15.

Write down the grep command to search for sequences in a file named `data.txt` that contain the string PCGA exactly once.

ANSWER: `grep PCGA data.txt | grep -v PCGA.*PCGA`

Question 16. Write down a grep command that will find, in the file named data.fa, all lines with and `A` and with a `B` but that do not have an `A` anywhere in the string after the `B`.

ANSWER: `grep A.*B[^A]*$ data.txt`

Question 17.

Write down the grep command to search for sequences in a file named `data.txt` that have an A somewhere before a B, and do not have another A following the B.

ANSWER: Note that 17 specifies `A` comes before `B` while 16 does not. Nonetheless, I believe the same answer works for both.

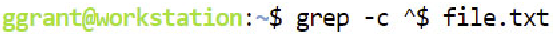
```
grep A.*B[^A]*$ data.txt
```

Question 18.

Explain the difference between the pipe operator and redirection.

ANSWER: The pipe operator `|` is for feeding the output of a command to the input of another. While redirection operator > is for feedint the output of a command to a file.

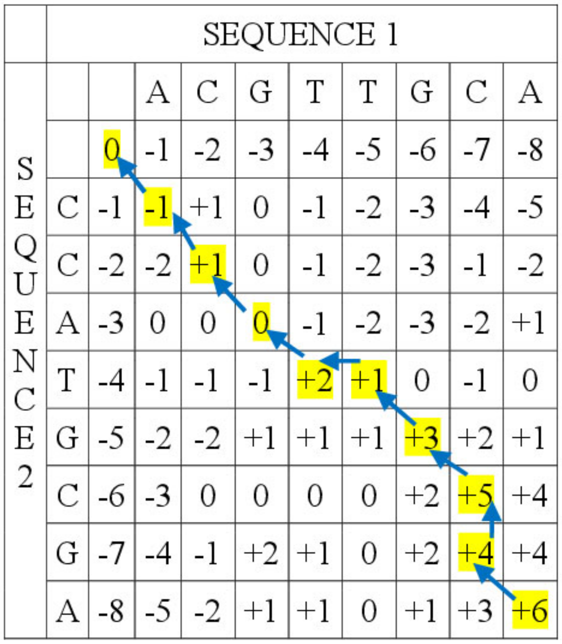
Question 19. What does the following grep command do?



```
ggrant@workstation:~$ grep -c ^$ file.txt
```

ANSWER: It counts the number of blank rows in the file `file.txt`

Question 20. Write down the alignment inferred by the traceback path.



		SEQUENCE 1								
			A	C	G	T	T	G	C	A
SEQUENCE 2		0	-1	-2	-3	-4	-5	-6	-7	-8
	C	-1	-1	+1	0	-1	-2	-3	-4	-5
	C	-2	-2	+1	0	-1	-2	-3	-1	-2
	A	-3	0	0	0	-1	-2	-3	-2	+1
	T	-4	-1	-1	-1	+2	+1	0	-1	0
	G	-5	-2	-2	+1	+1	+1	+3	+2	+1
	C	-6	-3	0	0	0	0	+2	+5	+4
	G	-7	-4	-1	+2	+1	0	+2	+4	+4
	A	-8	-5	-2	+1	+1	0	+1	+3	+6

ANSWER: The score is +6 and the alignment is:

$$\begin{array}{ccccccccc} A & C & G & T & T & G & C & - & A \\ C & C & A & T & - & G & C & G & A \end{array}$$

Question 21. Fill in the square marked X in the Needleman-Wunsch table. Matches score +1, mismatches score -1 and indels score -1. Draw in also the appropriate arrows.

		A	C	A	A
	0	**-1**	**-2**	**-3**	**-4**
A	**-1**	1	0	-1	-2
C	**-2**	0	2		
T	**-3**	-1	1		
G	**-4**	-2	↓ 0		
A	**-5**				

Question 22. Given an alignment scoring scheme where larger is better, what does it mean to have an alignment with "optimal" score?

ANSWER: It means there are no other alignments with higher score.

Question 23. Suppose the score for a match is +1, the score for a mismatch is X and the score for an indel is Y. (*assume X and Y are negative*). Suppose the following alignment has score -4

$$\begin{array}{ccc} A & - & G \\ A & C & C \end{array}$$

And suppose the following alignment has score -7

$$\begin{array}{cccc} A & - & - & G \\ A & C & C & C \end{array}$$

What exactly are the values of X and Y?

ANSWER: The first alignment implies:

$$1 + X + Y = -4$$

and the second implies

$$1 + X + 2Y = -7$$

Subtracting the first equation from the second gives $Y = -3$ and substituting this into either equation gives $X = -2$.

Question 24. How many different DNA sequences of length four are there?

ANSWER: $4^4 = 2^8 = 256$

Question 25.

True or False, $t^3 + \sqrt{t} + 5$ is big O of t^4?

ANSWER: True since

$$\lim_{t\to\infty} \frac{t^3 + \sqrt{t} + 5}{t^4} = \lim_{t\to\infty}\left(\frac{1}{t} + \frac{\sqrt{t}}{t^4} + \frac{5}{t^4}\right) = 0 + 0 + 0 = 0$$