University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Exam#1 (PRACTICE TEST)

October 2nd, 2023 Name: _____

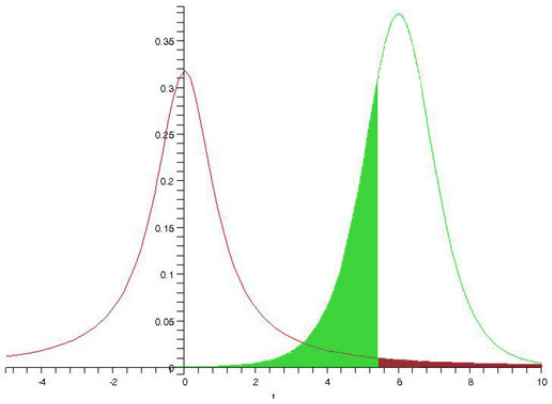
25 Questions, 4 points each

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: _____

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
(B) The one on the right

The red shaded area represents what?

(A) The False-Negative Rate
(B) The p-value
(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:

Question 3.

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

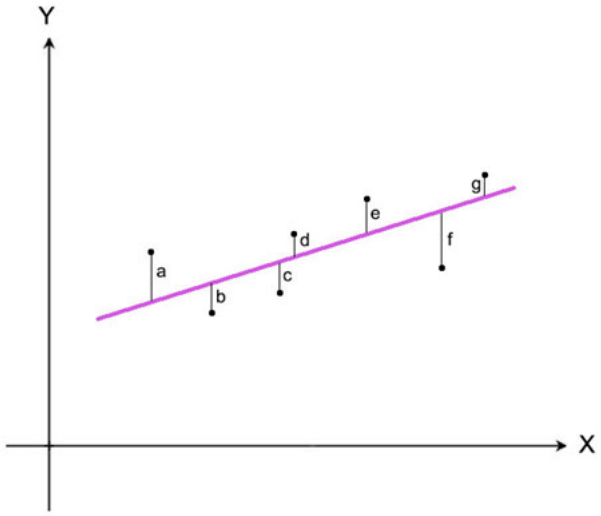
(A) Observations are normally distributed.
(B) The means of the two groups are not equal.
(C) Variance is equal in both conditions.
(D) The null hypotheses is false.

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

(A) It is linear in the independent variable
(B) It is linear in the coefficients
(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

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: _____

Question 6.

Consider the linear regression model:

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

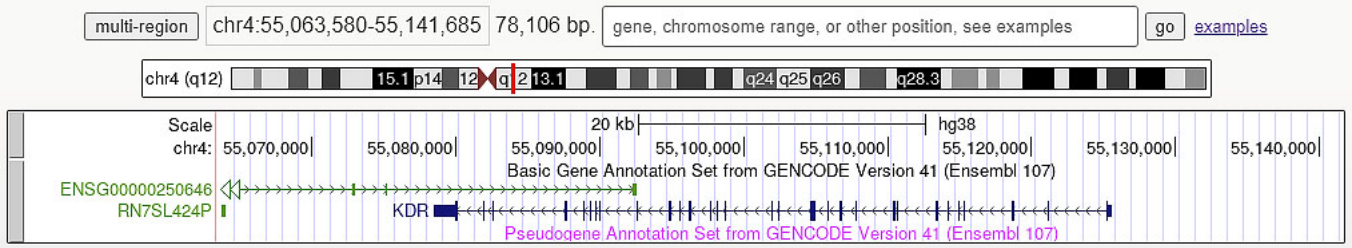
Assuming $\beta_2 > 0$, what is the shape of the regression curve? (The name of the type of curve.)

Answer: _____

What is the mean of ϵ?

Answer: _____

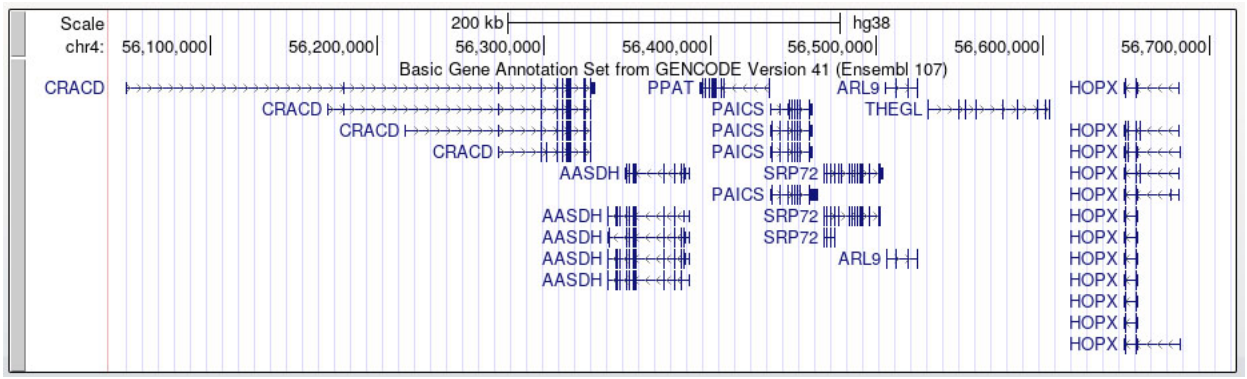
Question 7.



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

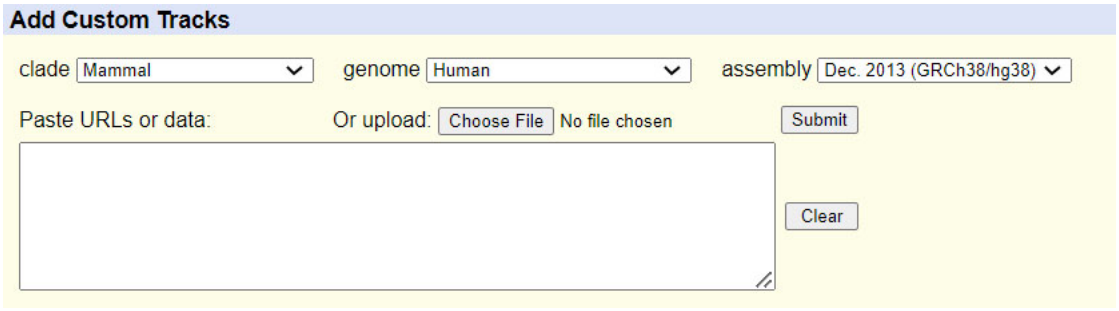
Question 8.



How many different *genes* are show here?

Answer: _____

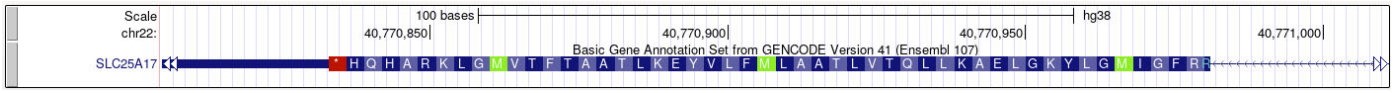
Question 9.



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: _____

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

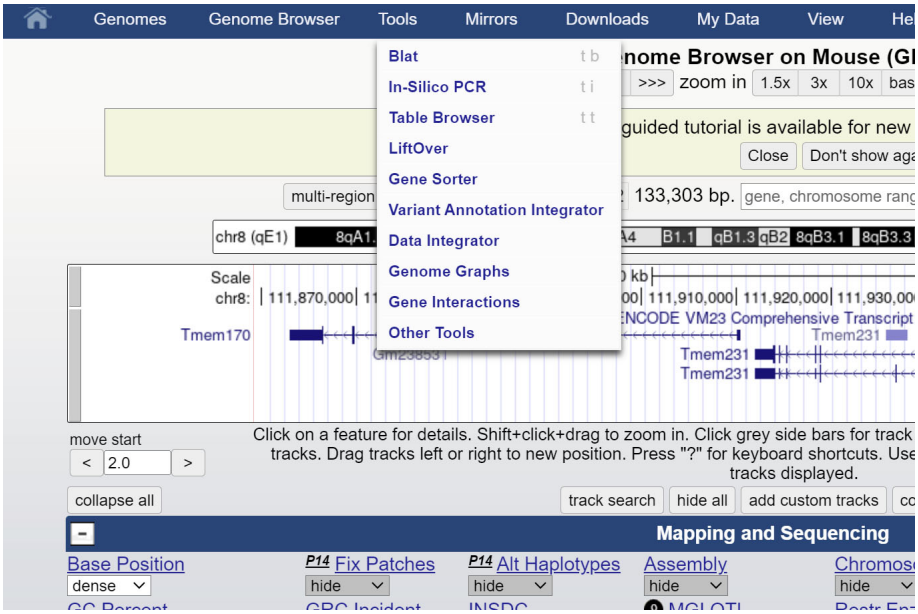


Answer: _____

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.

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: _____

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: _____

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: _____

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: _____

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: _____

Question 18.

Explain the difference between the pipe operator and redirection.
Answer:

Question 19. What does the following grep command do?

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

Answer:

Question 20. Write down the alignment inferred by the traceback path and give its score.

		SEQUENCE 1								
			A	C	G	T	T	G	C	A
	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

SEQUENCE 2

Answer:

Question 21. Fill in the square marked *X* in the Needleman-Wunch 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	X		
A	**-5**				

Answer: _____

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

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

Answer: _____

Question 25.

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

Answer: _____