CSE 5370: Bioinformatics
Homework 1

Yogesh Kalapala, yxk9640, 1001879640

1 Molecular Biology/Genetics Questions [30 points total]

1.1 Mendelian Genetics

heterozygoyus F1 pea (genotype RrYy) with a homozygous (genotype rryy) pea.

	RY	Ry	rY	ry
ry	RrYy	Rryy	rrYy	rryy
ry	rRyY	rRyy	rryY	ryry
ry	rRYy	rRyy	rrYy	rryy
ry	rRyY	rRyy	rryY	rryy

- Round Yellow(R,Y) - 4/16=0.25
- Round Green(R,y) - 4/16=0.25
- Wrinkle Yellow(r,Y) - 4/16=0.25
- Wrinkle Green(r,y) - 4/16=0.25

1.2 Genome Wide Association Studies (GWAS)

$$\text{The Bonferroni-corrected p-value} = \frac{original\, p-value}{no\, of\, hypothesis}$$

n is number of hypothesis and it is mentioned that SNP and height association is independent hypothesis we will have 2.6 million hypothesis = $\frac{0.05}{26000000}$

$= 0.000000001923077$

$= 1.923^{*}\ 10^{-}9$

$n is number of hypothesis and it is mentioned that metabolites will be only 1000$
$= \frac{0.05}{1000}$

$= 0.00005$

$= 5^{*}10^{-}5$

Type –1 error is false positive i.e a test showing a positive result when the disease is not present and Type:II error is showing a negative result even if the disease is present inside. In order to avoid Type: I error bonferroni conservative nature leads to Type II error.

2 Statistics Questions [30 points total]

2.1 Drug Approval

- Null Hypothesis: When the mean of sample measurement of treated population is same as the mean of sample measurement of the untreated population then the samples are said to have null hypothesis and the drug or the treatment had not effect on the population. For instance, assume in the above scenario if the systolic blood pressure of untreated mice is same as the systolic blood pressure of the treated mice then the drug would have null hypothesis.

- Alternate Hypothesis: When the mean of sample measurement of treated population is less than the mean of sample measurement of the untreated population then the samples are said to have null hypothesis and the drug or the treatment had not effect on the population. In the above scenario the systolic blood pressure of untreated mice is 120mmHg and systolic blood pressure of the treated mice is 115mmHg which is less than the untreated mice which tells us that the drug would has some effect on mice.

- For the given scenario Z-test is appropriate because we already know the standard deviation which is SD = 15. Sample Mean = 115mmHg Population Mean = 120mmHg Standard Deviation = 15 Total population = 20
  $Z = \frac{(sample\, mean) - (population\, mean)}{SD} \star \sqrt{n}$
  $Z = \frac{(115)-(120)}{SD} \star \sqrt{20}$

= -0.33333 ⋆20
= −1.490711985

Z-score is -1.490711985

P value is : 0.068112

The p-value for our drug is 0.068112 > 0.05 which tells us that the drug is not significant and cannot be released in the market yet.

- Advantage of choosing Non-parametric tests is they do not need the data to follow a particular distribution and non-parametric tests can be used with small sample sizes. The Real-world data does not fit to neat distribution and would not have a lot of data.
- Parametric tests have good statistical power and the data has to follow a distribution. These factors help to find significant effect.

3 Programming Question [40 points]

3.1

- Colab https://colab.research.google.com/drive/126oLyC6VqOoYHRCdmnvoMUFkoeYL4Y73?usp=shar
  Sumbitted here prokka files,colab .ipynb files.

4 Difficulty Adjustment

- How long did this assignment take you to complete?
- 1 hour for the conceptual part and programming prokka was not working and glimmer was also not working.
- If the assignment took you longer than the designed 3 hours, which parts were overly difficult?
- Coding part installing the prokka,glimmer and implementing them.

5 References

How to code in colab