Homework

Условия

Task 1

$$\Delta^2 = \frac{\mu_1 - \mu_2}{\sigma} = \frac{0.487 - 0.457}{0.0034} = 8.824$$
$$n = \frac{2\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{\Delta^2} = \frac{2(1.96 + 0.84)^2}{8.824} = 1.777 \approx 2$$

Task 2

Скачаем данные отсюда

Установим все необходимые библиотеки

Code

```
install.packages("BiocManager")
BiocManager::install("ShortRead")
BiocManager::install("Biostrings")
BiocManager::install("phyloseq")
```

Возьмем $N = 100$ и подготовим файлы для BLAST:

Code

```
N <- 100

fastq_file1 <- "animal1.01.16S.R1.fastq"
fastq_file2 <- "animal1.02.16S.R1.fastq"

fq1 <- readFastq(fastq_file1)
fq2 <- readFastq(fastq_file2)

cat("Total reads in file 1:", length(fq1), "\n")
cat("Total reads in file 2:", length(fq2), "\n")

set.seed(123)
sampled_fq1 <- fq1[sample(seq_along(fq1), min(N, length(fq1)))]
sampled_fq2 <- fq2[sample(seq_along(fq2), min(N, length(fq2)))]

sampled_reads <- c(sread(sampled_fq1), sread(sampled_fq2))
names(sampled_reads) <- c(as.character(id(sampled_fq1)), as.character(id(sampled_fq2)))

fasta_filename <- "sampled_sequences_2.fasta"
writeXStringSet(sampled_reads, filepath = fasta_filename)
cat("Sampled sequences written to", fasta_filename, "\n")
```

Получим следующий результат

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Uncultured bacterium clone 19-7 16S ribosomal RNA gene, partial sequence	uncultured bacterium	248	248	100%	1e-61	100.00%	1409	KT029396.1
Uncultured bacterium clone 19-24 16S ribosomal RNA gene, partial sequence	uncultured bacterium	248	248	100%	1e-61	100.00%	1409	KT029388.1
Uncultured Rhizobiales bacterium clone SHAI009 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1446	HM108368.1
Uncultured Rhizobiales bacterium clone SHAJ045 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1446	HM108476.1
Uncultured Rhizobiales bacterium clone SHAJ066 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1452	HM108495.1
Bartonella sp. M0280 16S ribosomal RNA gene, partial sequence	Bartonella apihabitans	248	248	100%	1e-61	100.00%	1393	ON614189.1
Bartonella apis strain B2 16S ribosomal RNA gene, partial sequence	Bartonella apis	248	248	100%	1e-61	100.00%	1049	OP359047.1
Uncultured Rhizobiales bacterium clone SHAJ016 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1449	HM108453.1
Uncultured Rhizobiales bacterium clone SHAJ071 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1450	HM108499.1
Uncultured Rhizobiales bacterium clone SHAJ020 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1454	HM108457.1
Bartonella apis strain B10834G6 16S ribosomal RNA gene, partial sequence	Bartonella apis	248	248	100%	1e-61	100.00%	1394	ON564884.1
Uncultured alpha proteobacterium clone HBG_A4R5-2 16S ribosomal RNA gene, partial seq...	uncultured Alphaproteobacteria ...	248	248	100%	1e-61	100.00%	1295	DQ837624.1
Uncultured Rhizobiales bacterium clone SHAI058 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1448	HM108414.1
Uncultured Rhizobiales bacterium clone SHAI019 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1445	HM108378.1
Rhizobiales bacterium PEB0181 16S ribosomal RNA gene, partial sequence	Rhizobiales bacterium PEB0181	248	248	100%	1e-61	100.00%	1015	JQ673258.1
Rhizobiales bacterium PEB0161 16S ribosomal RNA gene, partial sequence	Rhizobiales bacterium PEB0161	248	248	100%	1e-61	100.00%	1037	JQ673232.1
Uncultured Rhizobiales bacterium clone SHAI017 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1449	HM108376.1
Bartonella sp. W8097 16S ribosomal RNA gene, partial sequence	Bartonella apihabitans	248	248	100%	1e-61	100.00%	1396	OK032115.1
Uncultured Rhizobiales bacterium clone SHAJ078 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1430	HM108504.1
Uncultured Rhizobiales bacterium clone SHAI075 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1425	HM108430.1
Uncultured Rhizobiales bacterium clone SHAJ009 16S ribosomal RNA gene, partial sequence	uncultured Hyphomicrobiales b...	248	248	100%	1e-61	100.00%	1445	HM108446.1
Rhizobiales bacterium d3-7 16S ribosomal RNA gene, partial sequence	Rhizobiales bacterium d3-7	248	248	100%	1e-61	100.00%	1041	KM454403.1
Bartonella apis strain Acm18 16S ribosomal RNA gene, partial sequence	Bartonella apis	248	248	100%	1e-61	100.00%	1337	PQ136526.1
Uncultured bacterium clone 19-15 16S ribosomal RNA gene, partial sequence	uncultured bacterium	248	248	100%	1e-61	100.00%	1409	KT029379.1
Uncultured bacterium clone 19-5 16S ribosomal RNA gene, partial sequence	uncultured bacterium	248	248	100%	1e-61	100.00%	1409	KT029394.1
Uncultured Bifidobacterium sp. clone Seq4 16S ribosomal RNA gene, partial sequence	uncultured Bifidobacterium sp.	248	248	100%	1e-61	100.00%	404	OQ809349.1
Bartonella apis strain BBC0108 16S ribosomal RNA gene, partial sequence	Bartonella apis	248	248	100%	1e-61	100.00%	1376	KP987883.1
Rhizobiales bacterium PEB0172 16S ribosomal RNA gene, partial sequence	Rhizobiales bacterium PEB0172	248	248	100%	1e-61	100.00%	1042	JQ673247.1

Результаты BLAST показывают, что многие из наиболее совпадающих последовательностей помечены такими названиями, как Bartonella apis (и вариантами, такими как Bartonella apihabitans) и несколькими некультивируемыми бактериями Rhizobiales.

Bartonella apis — это бактерия, известная как часть микробиоты кишечника медоносной пчелы. Аналогично, многие из некультивируемых клонов Rhizobiales, часто появляются в исследованиях микробиома пчел. Таким образом, из файлов FASTQ (animal1.01 и animal1.02) попадания BLAST четко указывают на медоносную пчелу как на организм-хозяина.

Посмотрим на animal2.rds

Code

```
> animal2_data <- readRDS("animal2.rds")
> print(animal2_data)
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 1605 taxa and 66 samples ]
sample_data() Sample Data:       [ 66 samples by 8 sample variables ]
tax_table()   Taxonomy Table:    [ 1605 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 1605 tips and 1602 internal nodes ]
> animal2_data@tax_table
Taxonomy Table:     [1605 taxa by 7 taxonomic ranks]:
```

Gracilibacteria bacterium canine oral taxon 394, и другие таксоны, а также SR1 bacterium

Таблица 1: Taxonomic Classification

ID	Domain	Phylum	Class	Order
6c3b47bbac4c7af75368b9aa77620a3a	Bacteria	Proteobacteria	Phylum_Proteobacteria	Phylum_Proteobacteria
e9351dc418b0f459ed95e8a2e8a62e6b	Bacteria	Patescibacteria	ABY1	Class_ABY1
e705fdb7baa2313231f39c7b006fe37a	Bacteria	Dependentiae	Babeliae	Babeliales
dd935cf52df4bb39465baf1dcd5213ab	Bacteria	Dependentiae	Babeliae	Babeliales
1f2a4e9732fdce6657fefbb0cc33a65d	Bacteria	Dependentiae	Babeliae	Babeliales
a361007be7cd7ff8604a1199c813406b	Bacteria	Dependentiae	Babeliae	Babeliales
4af2a3b45dc62bbe3d96a1c0c9500190	Bacteria	Dependentiae	Babeliae	Babeliales
50c2209b3c3737fa82aad044728dd787	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales
209d2a7b32e4fd9500896c1910760c01	Bacteria	Acidobacteria	Acidobacteriia	Solibacterales
c4a45ece61cfd37714f3fa0cb99a08e2	Bacteria	Acidobacteria	Acidobacteriia	Solibacterales
a49f96fc5533b4f6050b9457c2ba00b2	Bacteria	Acidobacteria	Acidobacteriia	Solibacterales
a633f75d17e763b5e3063f169d35092b	Bacteria	Acidobacteria	Acidobacteriia	Solibacterales
4820c403cbb92d1f4cf2036b524965d0	Bacteria	Patescibacteria	ABY1	Candidatus Falkowbacteria
ab90adba6e6d0d459f42f5a363e62f9b	Bacteria	Patescibacteria	Parcubacteria	Class_Parcubacteria
6e03661027b3ad3d45bcb252b8002317	Bacteria	Patescibacteria	Parcubacteria	Class_Parcubacteria
0355e1414645c8f8e4312f796e1a5e8c	Bacteria	Patescibacteria	Gracilibacteria	Gracilibacteria bacterium oral taxon 873
3c90991156b7aeeeac6053c02cd95991	Bacteria	Patescibacteria	Gracilibacteria	Gracilibacteria bacterium canine oral taxo
c3049f7dcdde87039ee2a533790f34a2	Bacteria	Patescibacteria	Gracilibacteria	Gracilibacteria bacterium canine oral taxo
0bd202c297ac9312c3391422f172eb47	Bacteria	Patescibacteria	Gracilibacteria	JGI 0000069-P22
695a500552c94eca85fa2e887f8a40b4	Bacteria	Patescibacteria	Gracilibacteria	JGI 0000069-P22
aa41cbc5cc7db4ac206c34c7aa776d82	Bacteria	Tenericutes	Mollicutes	Mycoplasmatales
9e26d3b75fa2e1b0b1b081f748597307	Bacteria	Patescibacteria	CPR2	uncultured Firmicutes bacterium

canine oral, Prevotella sp. canine oral встречаются в полости рта собак, также часть встречается и у кошек. Следовательно это собака или кошка.

По запросам на мультик с пчелой и собакой/кошкой выдается Bee and PuppyCat

Task 3

Скачаем данные отсюда

Установим все:

Code

```
import pandas as pd
import numpy as np
from sklearn.model_selection import StratifiedKFold
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
```

Подготовим данные

Code

```
df = pd.read_csv('task_3.csv', index_col=0)

samples_df = df.iloc[:, :59].T
labels = samples_df.index.str.split('_').str[0]  # NG, NZ, XZ
```

Инициализируем K-fold

Code

```
n_splits = 5
skf = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=42)

accuracy_scores = []
class_reports = []
```

Обучение

Code

```
for train_idx, test_idx in skf.split(samples_df, labels):
    X_train, X_test = samples_df.iloc[train_idx], samples_df.iloc[test_idx]
    y_train, y_test = labels[train_idx], labels[test_idx]

    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)

    model = RandomForestClassifier(n_estimators=100, random_state=42)
    model.fit(X_train_scaled, y_train)

    y_pred = model.predict(X_test_scaled)
    accuracy = accuracy_score(y_test, y_pred)
    report = classification_report(y_test, y_pred, output_dict=True)

    accuracy_scores.append(accuracy)
    class_reports.append(report)
```

Теперь можно посмотреть на точность

Code

```
mean_accuracy = np.mean(accuracy_scores)
std_accuracy = np.std(accuracy_scores)
print(f"Mean Accuracy: {mean_accuracy:.3f} ± {std_accuracy:.3f}\n")
```

Получим следующее: Mean Accuracy: 0.983 ± 0.033

Code

```
avg_precision = np.mean([r['macro avg']['precision'] for r in class_reports])
avg_recall = np.mean([r['macro avg']['recall'] for r in class_reports])
avg_f1 = np.mean([r['macro avg']['f1-score'] for r in class_reports])
print(f"Macro-Averaged Metrics:")
print(f"Precision: {avg_precision:.3f}, Recall: {avg_recall:.3f}, F1-Score: {avg_f1:.3f}")
print("\nClassification Report (Last Fold):")
print(classification_report(y_test, y_pred))
```

Таблица 2: Classification Metrics

Metric	Precision	Recall	F1-Score	Support
Macro-Averaged	0.989	0.967	0.972	-
Classification Report (Last Fold)				
Class	Precision	Recall	F1-Score	Support
NG	1.00	1.00	1.00	1
NZ	1.00	1.00	1.00	5
XZ	1.00	1.00	1.00	5
Accuracy	-	-	-	1.00
Macro Average	1.00	1.00	1.00	11
Weighted Average	1.00	1.00	1.00	11

весь код также можно посмотреть в Google Colab