Question 1:

Please provide a short description of the following statistical tests. Make sure to include any key assumptions and what the tests are used for.
1. ANOVA
2. T-test
3. Wilcoxon rank sum test.

Ans:

Testing for the difference between two dependent groups, such as before and after measurements on the same subjects, is typically done by testing for a difference in centers of distribution (means or medians). In this situation, the data are paired; two observations are obtained on each n subject, resulting in one sample of 2n observations. The paired t-test looks for a difference in means, while the non-parametric sign and signed rank tests examine differences in medians. It is assumed that the spreads and shapes of the two populations are the same for all tests.

Independent t-test tests the difference between two independent groups, while the non-parametric Wilcoxon rank sum test examines differences in medians.

T-test:

Assumptions:

- The data are continuous.
- The sample data have been randomly sampled from a population.
- There is homogeneity of variance (i.e., the variability of the data in each group is similar).
- The distribution is approximately normal.

The null hypothesis is H0: $\mu1 = \mu2$, and

the alternative hypothesis is Ha: $\mu1 = \mu2$ vs Ha: $\mu1 < \mu2$ (Left tailed) or Ha: $\mu1 > \mu2$ (Right tailed) or Ha: $\mu1 \neq \mu2$ (Two tailed).

Tests for the Means of Two Populations, Using a Paired Sample

$$t = \frac{\bar{d}}{\left(s_d/\sqrt{n}\right)}$$

where $\bar{d}$ = mean of paired differences, s_d = standard deviation of paired differences and n= number of pairs.

Hypothesis

The null hypothesis is H0: $\mu1 = \mu2$, and

The alternative hypothesis is Ha: $\mu1 = \mu2$ vs Ha: $\mu1 < \mu2$ (Left tailed) or Ha: $\mu1 > \mu2$ (Right tailed) or Ha: $\mu1 \neq \mu2$ (Two tailed).

Independent sample t-test for unequal variances,

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} - \frac{s_2^2}{n_2}}}$$

Where $s^2 = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$

Independent sample t-test for equal variances

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s_p = \sqrt{\frac{(n1-1)*s_1^2 + (n2-1)*s_2^2}{n1+n2-2}}$

where $\overline{x_1}$ and $\overline{x_2}$ are means, s_1 and s_2 are standard deviations and n1 and n2 are sample sizes for group1 and group2 respectively.

Interpretation

If tabulated value is greater than calculated t value then null hypothesis was accepted otherwise, rejected.

If $P \leq \alpha$, reject $H0$; otherwise, do not reject $H0$.

If Ho is accepted, we conclude that both means are equal otherwise means are not equal.

Wilcoxon rank sum test:

Mann–Whitney U test (also called the Wilcoxon rank-sum test or Wilcoxon–Mann–Whitney test).

The Mann–Whitney U test / Wilcoxon rank-sum test is not the same as the Wilcoxon signed-rank test, although both are nonparametric and involve the summation of ranks. The Mann–Whitney U test is applied to independent samples. The Wilcoxon signed-rank test is used to matched or dependent samples.

Assumptions:

1. The two samples are randomly and independently drawn from their respective populations.
2. The variable under study is continuous.
3. The measurement scale is at least ordinal.
4. The distributions of two populations differ only with respect to location parameter.

Hypothesis

The null hypothesis is H0: μ1 = μ2, and

the alternative hypothesis is Ha: μ1 = μ2 vs Ha: μ1 < μ2 (Left tailed) or Ha: μ1 > μ2 (Right tailed) or Ha: μ1 ≠ μ2 (Two tailed).

The test statistic for the Wilcoxon rank-sum test:

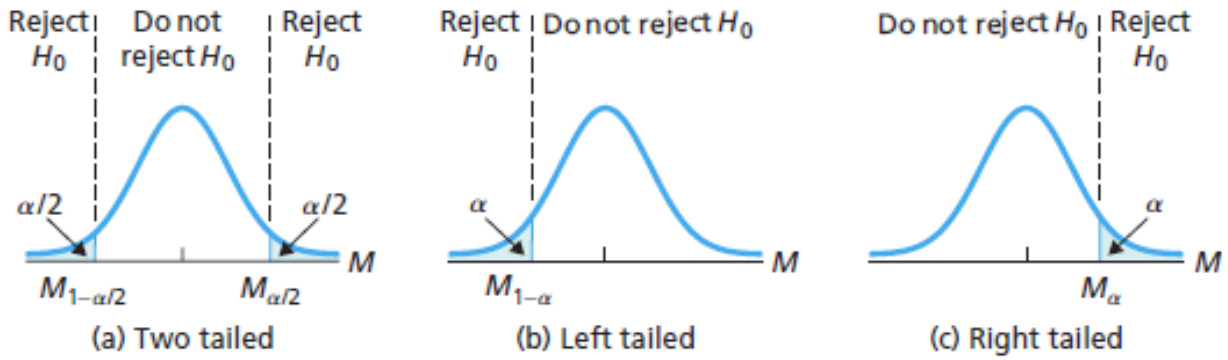
U1= n1*n2 + (n2*(n2+1)/2) - R2

U2= n1*n2 + (n1*(n1+1)/2) – R1

Where $U_1 + U_2 = n_1 * n_2$, $R_1 + R_2 = N*(N + 1)/2$ and N = n1 + n2.

R1 and R2 being the sum of the ranks in groups 1 and 2, after pooling all samples in one set (see below) and where the smallest value obtains rank 1 and so on.

Critical value(s) for a Wilcoxon rank sum test at the significance level α if the test is (a) two tailed, (b) left tailed, or (c) right tailed.



Interpretation:

If the value of the test statistic falls in the rejection region, reject *H*0; otherwise, do not reject *H*0 at α level of significance.

If $P \leq \alpha$, reject *H*0; otherwise, do not reject *H*0.

If H0 is rejected, two medians are not equal otherwise equal.

ANOVA:

Analysis of variance (ANOVA) is a procedure to test comparison between the several means, that is, the means of a single variable for several groups. Compares the means of a variable for groups that result from a classification by one other variable called a One-way ANOVA. The possible values of the factor are referred to as the levels of the factor.

Assumptions:

- Simple random samples: The samples taken from the populations under consideration are simple random samples.
- Independent samples: The samples taken from the populations under consideration are independent of one another.
- Normal populations: For each population, the variable under consideration is normally distributed.
- Equal standard deviations: The standard deviations of the variable under consideration is the same for all the populations.

Hypothesis:

H0: μ1= μ2= μ3 = …… = μk

Verses

H1: μ1≠ μ2 ≠ μ3 ≠ ……. ≠ μk

Source	df	*SS*	$MS = SS/df$	*F*-statistic
Treatment	$k-1$	*SSTR*	$MSTR = \frac{SSTR}{k-1}$	$F = \frac{MSTR}{MSE}$
Error	$n-k$	*SSE*	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	*SST*		

SSTR = $\sum n_i\,(\bar{x}_i - \bar{x})^2$ is between sum of squares or tretment sum of squares

SSE = $\sum(n_i - 1) * s_i^2$ is with in sum of squares or error sum of squares

SST = $\sum(x_i - \bar{x})^2$ is total sum of squares

SST = SSE + SSTR

Where,
n = total number of observations
$\bar{x}$ = mean of all n observations.
and, for i = 1, 2, . . . , k,
nj = i^{th} sample size
$\bar{x}_i$= i^{th} sample mean
s_i = i^{th} sample standard deviation

Interpretation
If calculated f-statistic value is less than the tabulated f value at α level of significance, then accept H0 otherwise reject the H0.
If $P \leq \alpha$, reject $H0$; otherwise, do not reject $H0$.
Hence, we conclude that all group means are not equal if H0 is rejected, otherwise wise all means are equal.

Question 3
In a study, physicians were asked what the odds of breast cancer would be in a woman who was initially thought to have a 1% risk of cancer but who ended up with a positive mammogram result (a mammogram accurately classifies about 70% of cancerous tumors and 80% of benign tumors.) What is the probability of cancer in the woman?

Ans:
P is who mammogram result is positive
B is tumor is malignant
NB is tumor is benign

P(B) = 0.01 who have the initially breast cancer is 1%

P(NB) = 1-0.01 = 0.99 who do not have the breast cancer

probability of mammogram result is positive when tumor is malignant, P(C|B) = 0.7
probability of mammogram result is not positive when tumor is benign tumors, P($\bar{P}$|NB) = 0.8

probability of mammogram result is positive when tumor is benign tumors,

P(P|NB) = 1- P($\bar{P}$|NB) = 1- 0.8 = 0.2

probability of cancer in the woman P(B|P) = $\frac{P(B)*P(P|B)}{P(B)*P(\overline{P}|B)+P(NB)*p(P|NA)}$

$$= \frac{0.01*0.7}{0.01*0.7+0.2*0.99}$$
$$= 0.03414634$$

probability of cancer in the woman is 0.034.

Question 4

Assume A and B are matrices with entries that are from the standard normal distribution. Will entries in the product of the matrices (AB) be normally distributed? If we needed to control the variance in the product, how would we do that?

Ans:
The product of two matrices *A* and *B*, with entries that are from the standard normal distribution will not necessarily result in a matrix with entries that are normally distributed. The distribution of the entries in the product matrix *AB* depends on the specific entries of *A* and *B*, as well as their dimensions and the properties of matrix multiplication.
The entries in the product matrix *AB* are essentially linear combinations of the entries of *A* and *B*, which means their distribution can be quite complex and may not follow a simple distribution like the normal distribution.
However, under certain conditions and in the limit of large matrix dimensions, the entries in the product matrix *AB* can tend to approach a normal distribution due to the Central Limit Theorem. This theorem states that the sum (or average) of a large number of independent random variables, each with finite variance, tends to be normally distributed regardless of the underlying distribution of the individual variables.
To control the variance in the product of matrices *AB*, you can consider various approaches:

1. **Normalize the Entries**: Normalize the entries of matrices *A* and *B* such that they have unit variance. This can help control the overall variance in the product matrix *AB*.
2. **Scale the Matrices**: Multiply one or both of the matrices *A* and *B* by a scalar value to adjust their overall variance. For example, scaling one of the matrices by a factor *c* will scale the variance of the entries in the resulting product matrix by the square of *c*.
3. **Use Matrix Factorization**: Decompose one or both of the matrices *A* and *B* into factors with desired properties (e.g., lower variance) and then perform matrix multiplication with the decomposed matrices.
4. **Regularization**: Apply regularization techniques to the matrices *A* and *B* before multiplication to control their magnitude and variance.
5. **Post-processing**: After computing the product matrix *AB*, you can apply post-processing techniques such as scaling, clipping, or transformation to adjust the variance of the resulting entries as needed.

Overall, controlling the variance in the product of matrices *AB* involves adjusting the properties of the input matrices, using appropriate scaling or normalization techniques, and considering the specific requirements of the application.

Question 5

Explain Bessel's correction in the context of calculating the sample variance as compared to the population variance.

Ans:

This method corrects the bias in the estimation of the population variance. It also partially corrects the bias in estimating the population standard deviation. However, the correction often increases the mean squared error in these estimations.

Bessels' correction refers to the "n-1" found in several formulas, including the sample variance and standard deviation. This correction is made to correct because these sample statistics tend to underestimate the actual parameters found in the population.

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

$$\text{sd} = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

why do we subtract 1 when using these formulas?

When you have an entire population and calculate any parameter (like the population variance or population standard deviation), your results will be accurate. That's because you have all the data about your population. However, when you work with a sample, you've only got a small fraction of the population to work with. Therefore, your answers aren't going to be as accurate as those you would have got, if you had the entire set of data to work with.

In the case of the sample variance & standard deviation, the statistic you are working with is the sample mean ($\bar{x}$) instead of the population mean (μ). Any x-value in your sample is going to be closer to $\bar{x}$ than to μ.

This fact alters the sums of squares (in the numerator of the above formulas). The sum of squares for μ

$$\sum_{i=1}^{N}(x_i - \mu)^2$$

is going to be larger than the sum of squares for $\bar{x}$.

$$\sum_{i=1}^{N}(x_i - \bar{x})^2$$

For small sample sizes, Bessel's correction is going to be quite severe. If you have a small sample, it's highly unlikely that it's going to be a very good estimate of the population mean anyway. If you have a very large sample size, you're going to approach a point when your sample statistics are going to be almost equal to your population parameters. In that case, Bessel's correction simply isn't needed at all.

Question 6

Describe an appropriate distribution of model stock returns.

Ans:

Modeling stock returns often involves selecting a probability distribution that accurately reflects the behaviour of the returns. The **Normal distribution** (the Gaussian distribution) is commonly used for modeling stock returns. However, it's important to note that stock returns often exhibit fat tails and excess kurtosis, meaning extreme events occur more frequently than expected under a normal distribution. In such cases, alternative distributions may provide a better fit. Here are a few distributions commonly used to model stock returns:
This skewness is important in determining which distribution is appropriate for investment decision-making. A further distinction is that the values derived from a lognormal distribution are normally distributed.

Lognormal distribution: This distribution differs from the normal distribution in several ways. A significant difference is in its shape: the normal distribution is symmetrical, whereas the lognormal distribution is not. Because the values in a lognormal distribution are positive, they create a right-skewed curve.

Student's t-distribution: This distribution is similar to the Normal distribution but has heavier tails, making it more suitable for modeling stock returns with fat tails and excess kurtosis. The t-distribution is characterized by a single parameter called the degrees of freedom, which controls the thickness of the tails.

Generalized Extreme Value (GEV) distribution: The GEV distribution is often used to model extreme events, such as large stock market movements. It allows for various tail behaviors, including heavy tails, and can be parameterized to capture different degrees of tail thickness.

GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models: Rather than directly modeling the distribution of stock returns, GARCH models focus on modeling the volatility or variance of returns over time. These models capture the tendency of volatility to cluster and exhibit persistence, which is commonly observed in financial time series data.

Exponential distribution: While less common for modeling stock returns directly, the exponential distribution is sometimes used to model waiting times between stock price movements or the durations of market phases.

Lévy distribution: This distribution is used in the context of Lévy processes, which are stochastic processes that exhibit certain types of jumps or discontinuities. Lévy distributions can capture extreme events and jump in stock returns.

When selecting a distribution to model stock returns, it's important to consider the specific characteristics of the data, such as fat tails, skewness, and autocorrelation. Researchers often use statistical tests and model diagnostics to assess the goodness of fit of different distributions and choose the most appropriate one for their analysis. Additionally, combinations of distributions or more complex models may capture the full range of behaviour observed in stock returns.