

R ve R Studio Kurulumu

R Kurulumu

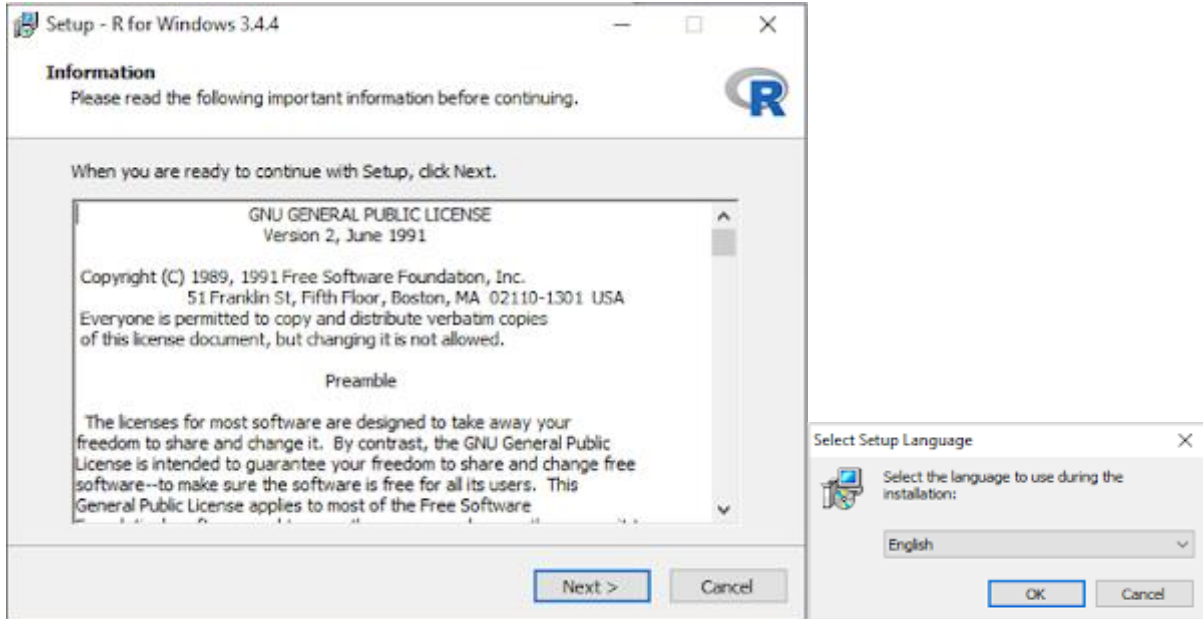
R istatistiksel ve ekonometrik hesaplar gibi birçok özelliğin yanı sıra veri manipülasyonu, programlama ve grafiksel gösterimi bir arada sunan, çok yönlü, entegre bir yazılım ortamıdır. R açık kaynak bir programlama dilidir. Bu nedenle R'nin tüm sürümleri ücretsiz olarak internet sitesinde yer almaktadır.

R'nin internet sitesinden işletim sisteminize uygun programı indirip kurabilirsiniz. Linux, Mac OS ve Windows işletim sistemleri için sürümleri mevcuttur. Windows işletim sistemi için "Download R for Windows" seçeneğini seçeriz. Aşağıdaki linkten indirme yapılabilecek internet sitesine gidebilirsiniz.

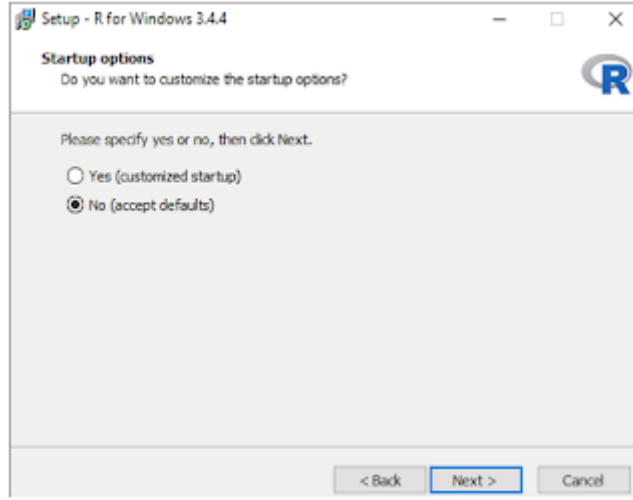
- <https://cran.r-project.org/>



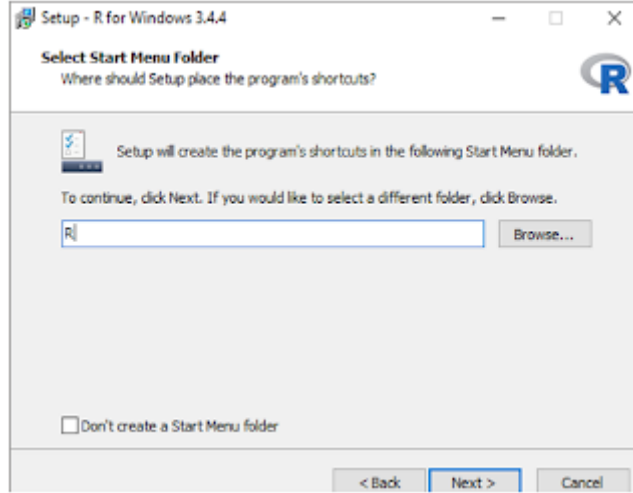
İlk defa yükleme yaptığımız için yani sistemimizde daha önceden R kurulu olmadığı için "install R for the first time" seçeneğini seçeriz. İndirme işlemini başlatırızKullanım Sözleşmesi ekrana gelir.



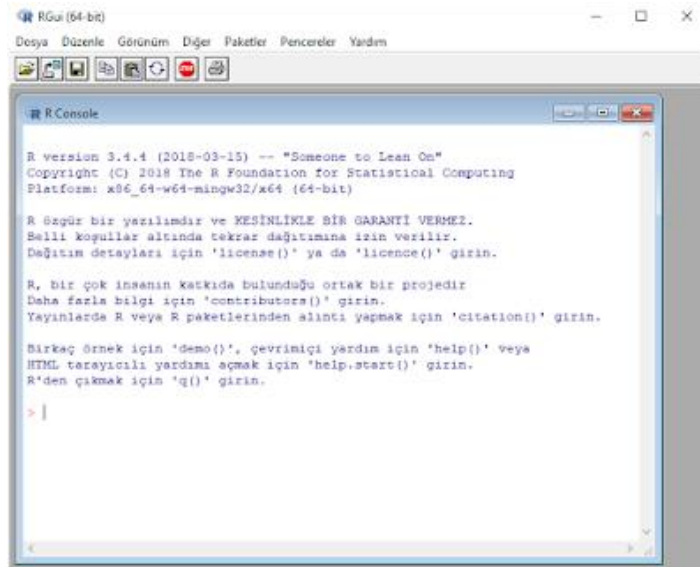
R'ı kuracağımız dosyanın belirtilmesi istenir. Default dosyaya kurulumu yaparız. Başlangıç seçeneklerinde değişiklik yapılmak isteniyorsa "customized startup" seçeneği seçilir.



Yaptığınız çalışmaların kaydedileği klasörü seçeceğiniz ekran gelir. Burada default olarak Belgelerim klasörünün altına R adında klasör oluşturur ve yapılan çalışmaları buraya kaydeder.



Bu aşamalar sonunda kurulum dosyaları yüklenir. Kurulum tamamlandığında R programını çalıştırdığımızda aşağıdaki gibi bir çalışma ortamı açılır.



Kullanıcıların bir kısmı bu çalışma ortamının gelişmiş olmadığı ve grafiksel arayüz olarak zayıf olduğunu düşündüklerinden yardımcı bir programa ihtiyaç duymaktadırlar. Bu noktada ise farklı bir "IDE(Integrated Development Environment)" Türkçe'si "Tümleşik Geliştirme Ortamı" olan bir arayüzden faydalanırlar. Yardımcı program olarak da seçenekler arasında en çok tercih edilen Rstudio ortamını dersimizde kullanacağımız için Rstudio kurulumu ile devam edelim.

R Studio Kurulumu

R editörü grafiksel bir arayüz olmayıp eski tip bir yazılım konsolidur. Ücretsiz olan Rstudio programı daha modern ve kullanışlı bir arayüz sunmaktadır. Rstudio dışında Revolution Analytics, StatET ve ESS gibi editörler de mevcuttur. R Studio programını indirmek için aşağıdaki linke gidilir.

- <https://rstudio.com/products/rstudio/download/>

The screenshot shows the RStudio website's download page. The header includes the RStudio logo and navigation links: Products, Resources, Pricing, About Us, and Blogs. The main heading is "Choose Your Version of RStudio". Below this, a paragraph describes RStudio as a set of integrated tools. To the right, there is an image of a computer monitor displaying RStudio. The main content area displays five download options with their respective prices and buttons:

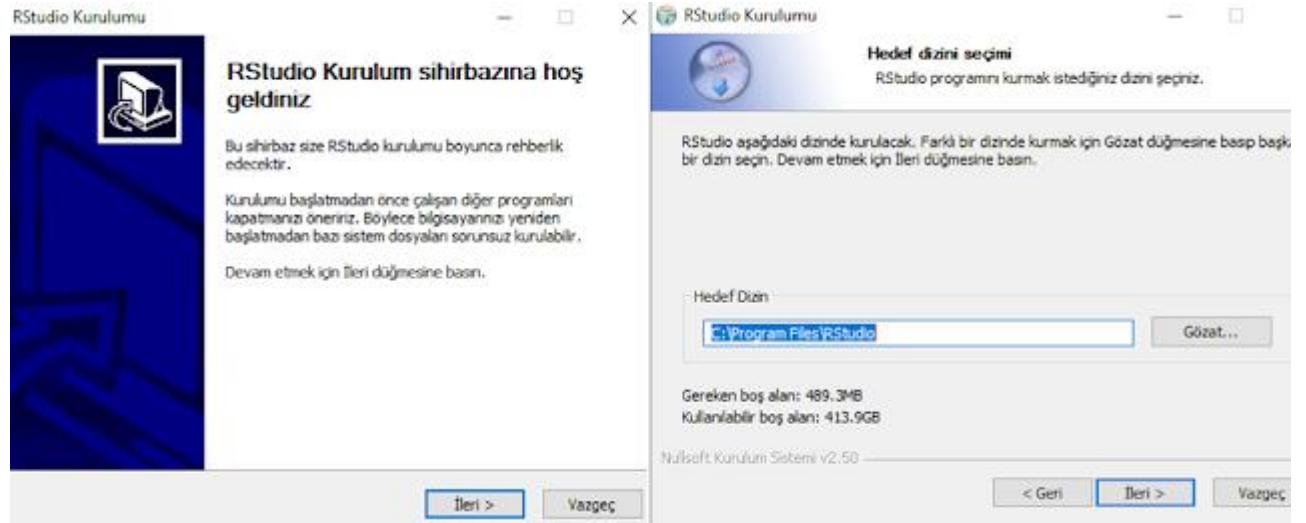
Product	License	Price	Button
RStudio Desktop	Open Source License	FREE	DOWNLOAD
RStudio Desktop	Commercial License	\$995 per year	BUY
RStudio Server	Open Source License	FREE	DOWNLOAD
RStudio Server Pro	Commercial License	\$9,995 per year	DOWNLOAD
RStudio Server Pro + RStudio Connect	Commercial License	\$29,995 per year	TALK

İşletim sistemine uygun ve son yayınlanan sürüm olan RStudio platformunu seçeriz.

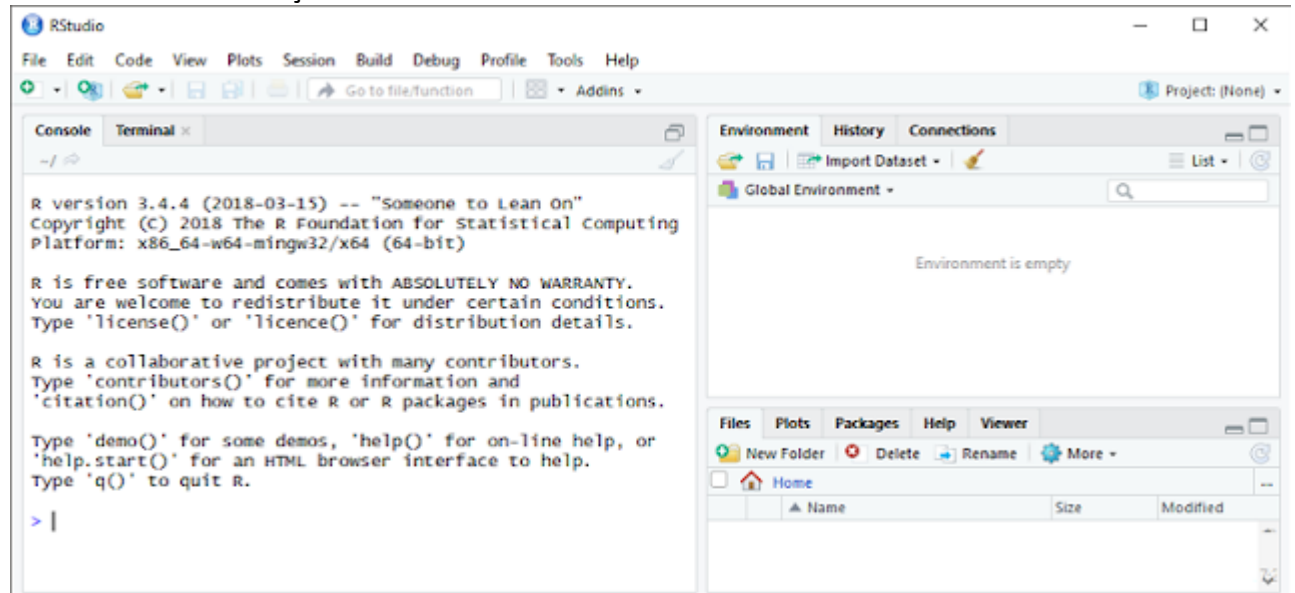
The screenshot shows the RStudio website's download page for RStudio Desktop 1.1.442. The header includes the RStudio logo and navigation links: Products, Resources, Pricing, About Us, and Blogs. The main heading is "RStudio Desktop 1.1.442 — Release Notes". Below this, a paragraph states "RStudio requires R 3.0.1+. If you don't already have R, download it here." The main content area is titled "Installers for Supported Platforms" and contains a table with the following data:

Installer	Size	Date	MD5
RStudio 1.1.442 - Windows Vista/7/8/10	85.8 MB	2019-05-12	25a8eb8ecae4fd71901c977dbcfb104b
RStudio 1.1.442 - Mac OS X 10.8+ (64-bit)	74.5 MB	2019-05-12	89613427803a1e516372075ec2e264b2
RStudio 1.1.442 - Ubuntu 12.04-13.10/Debian 8 (32-bit)	88.3 MB	2019-05-12	090fcb1fec90e3d621bc89e113c8dc28
RStudio 1.1.442 - Ubuntu 12.04-13.10/Debian 8 (64-bit)	87.4 MB	2019-05-12	2c0005a6a8f12b06c7e6b343692288fd
RStudio 1.1.442 - Ubuntu 16.04+/Debian 9+ (64-bit)	63.1 MB	2019-05-12	c9eb372938b10626fbc4d65fa01c7175
RStudio 1.1.442 - Fedora 19+/Redhat 7+/openSUSE 13.1+ (32-bit)	88.1 MB	2019-05-12	77ced16b9ca8d9c636d388b842a60e1c
RStudio 1.1.442 - Fedora 19+/Redhat 7+/openSUSE 13.1+ (64-bit)	90.8 MB	2019-05-12	8e6435aa53fa8ea9878ef9c09b6419f4

İndirme işlemi tamamlandıktan sonra kurulum aşamasına geçeriz. Rstudio programının kurulumunu başlatırız ve kurulum dosyalarının kopyalanacağı yeri seçeriz.

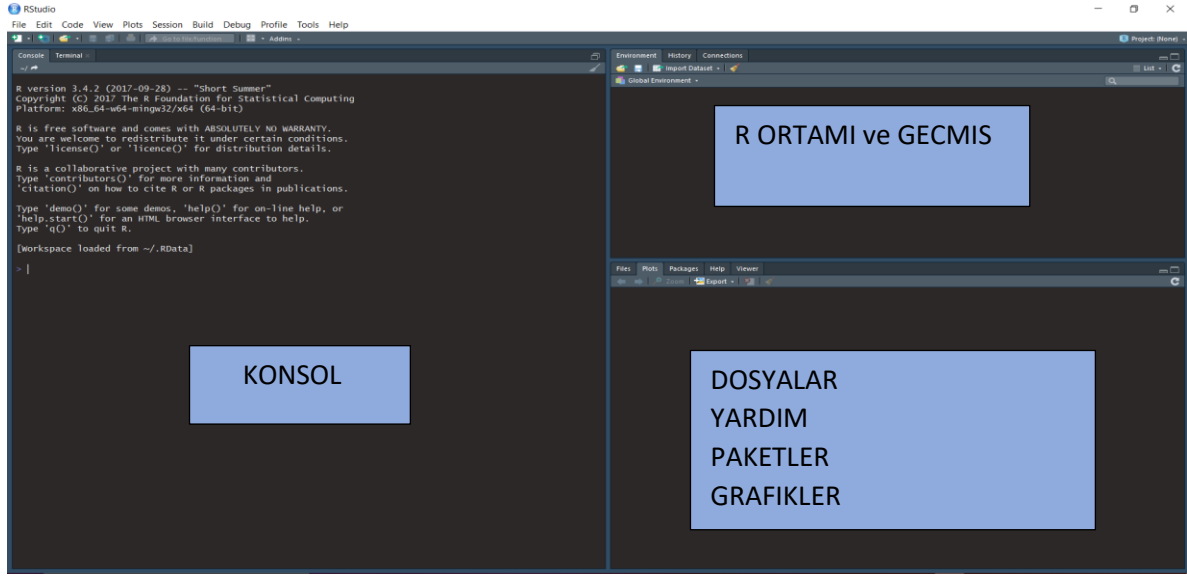


Kurulum işlemini devam ettiririz. Kurulum dosyaları bilgisayara çıkarılır. Bu işlem bittiğinde kurulum tamamlanmış olur.

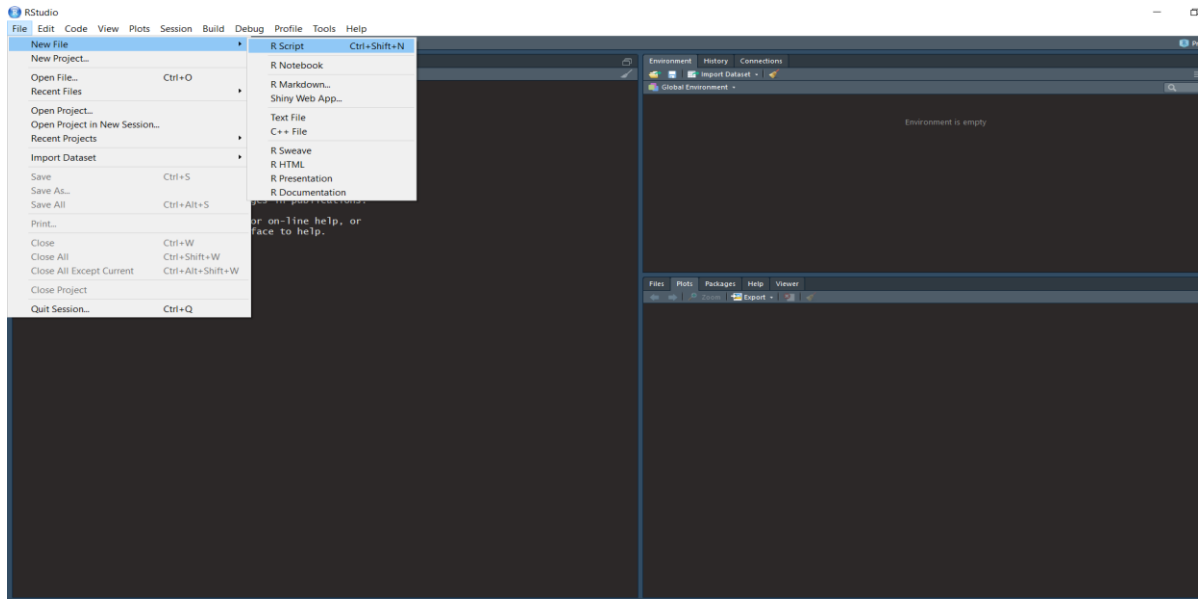


RStudio Ortamına Genel Bakış

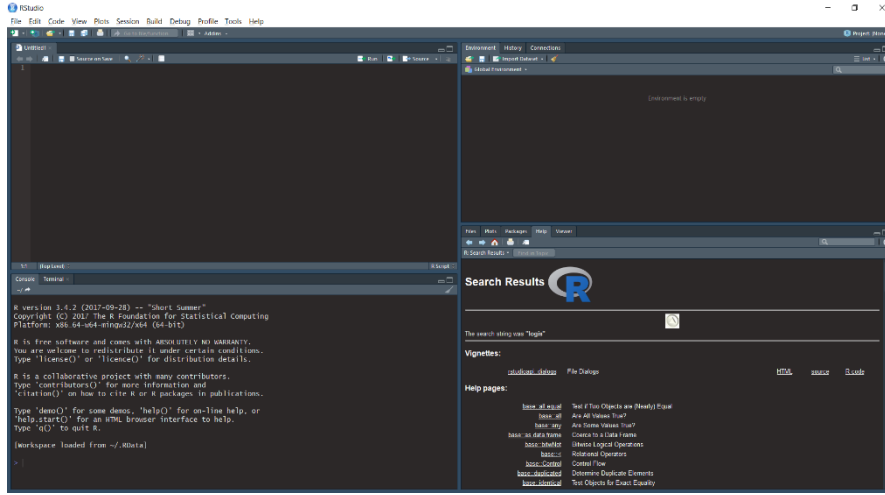
RStudio açıldığında 3 ana ekran olarak görülmektedir. Bunların içeriği aşağıdaki resimde gösterilmiştir.



Açıldığında 3 ana ekrana ayrılmış vaziyette olan R de kod yazımının sağlandığı bölge sol üst kısımdaki konsol yeridir. Bu kısımda kod yazımı satır satır icra edilip kodlar çalıştırıldığından toplu bir şekilde yazılmış olan komutların çalıştırılması için betik(scrypt) ekleme işlemi yapmalıyız. Bunun için sol üst kısımda bulunan **File-> New File-> R Script** seçeneği seçilir.



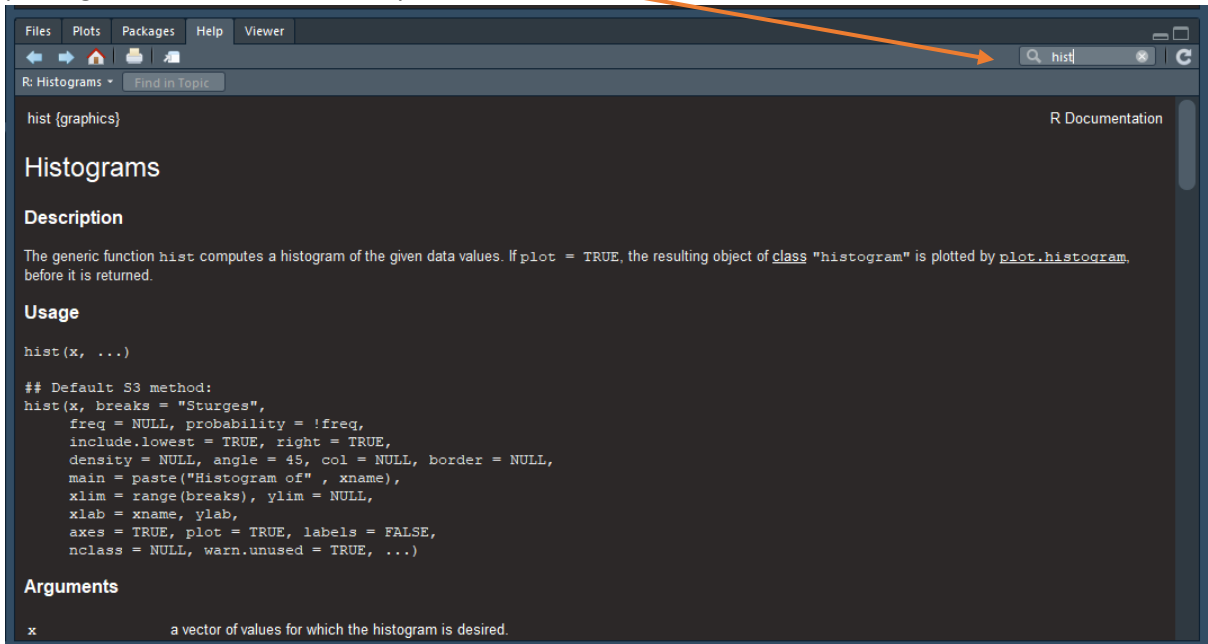
Bu asamadan sonra R kodlarinin yazilabilecegi editor kismina da ulasmis oluyoruz.



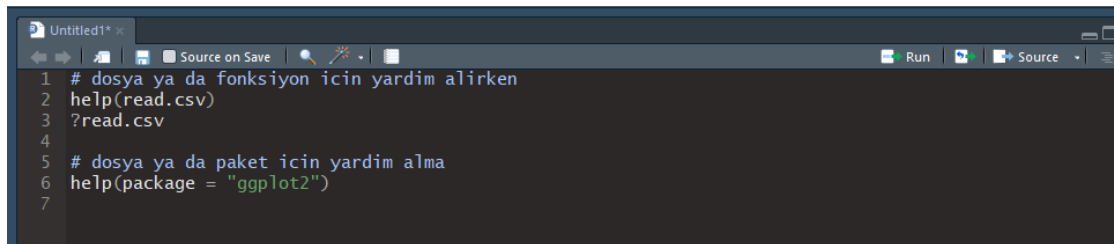
Yardım Alma

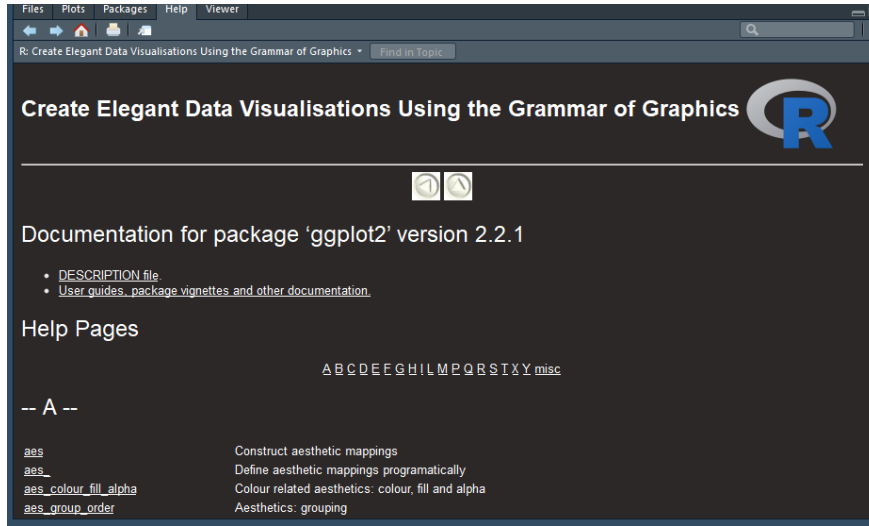
1.Yardim menu su üzerinden yardım

Sağ üst kösedeki arama kısmına yardım almak istedigimiz fonksiyon, paket gibi enstrümanların ismini yazarız.



2.Konsol veya Editor Üzerinden Yardım





Üstteki kodlar yazıldıktan sonra RStudio nun sağ alt kısmındaki help kısmında ilgili fonksiyon, paket, dosya vs. ile ilgili açıklamaları ve bunlarla ilgili örnekleri görebiliriz.

Paketler

R programında paketler, isteğe bağlı olarak yüklenebilen ek işlevlerin koleksiyonudur. Genellikle bu işlevleri göstermek için kullanılacak örnek veriler içerirler. R birçok ortak istatistiksel işlev ve modelle birlikte sunulsa da çalışmalarımızın çoğu ek paketler gerektirir. Bu nedenle bazı durumlarda bu ek paketleri yüklememiz gerekmektedir.

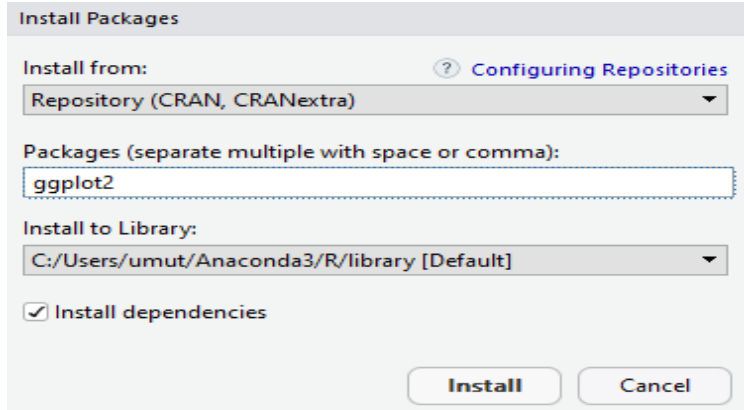
Paket Kurma ve Yükleme

Bir paketi kullanmak için, önce paketi kurmalı ve sonra yüklemelisiniz. Bu adımlar komut satırında veya Paketler Sekmesi kullanılarak yapılabilir. Her iki yaklaşımın örnekleri aşağıda verilmektedir. R paketlerinin yalnızca bir kez kurulması gerekir (R yükseltilene veya yeniden kurulana kadar). Bununla birlikte, her yeni bir R oturumu başlattığınızda, o oturumda kullanmak istediğiniz her paketi yüklemeniz gerekir.

Paketler sekmesinde, bilgisayarınızda yüklü olan tüm paketlerin bir listesini ve “Yükle” veya “Güncelle” etiketli 2 düğmeyi göreceksiniz. Yeni bir paket yüklemek için Yükle düğmesini seçmeniz yeterlidir. Bir seferde bir veya daha fazla paketi, virgülle ayırarak girebilirsiniz. Öncelikle paket kurulumu ile işleme başlayalım.

Files	Plots	Packages	Help	Viewer
Install Update				
Name	Install R packages	Description	Version	
System Library				
<input type="checkbox"/>	assertthat	Easy Pre and Post Assertions	0.2.0	
<input type="checkbox"/>	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.1	
<input type="checkbox"/>	base64enc	Tools for base64 encoding	0.1-3	
<input type="checkbox"/>	BH	Boost C++ Header Files	1.65.0-1	
<input type="checkbox"/>	bindr	Parametrized Active Bindings	0.1	
<input type="checkbox"/>	bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2	
<input type="checkbox"/>	bit	A class for vectors of 1-bit booleans	1.1-12	
<input type="checkbox"/>	bit64	A S3 Class for Vectors of 64bit Integers	0.9-7	
<input type="checkbox"/>	bitops	Bitwise Operations	1.0-6	
<input type="checkbox"/>	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.1.0	
<input type="checkbox"/>	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-20	
<input type="checkbox"/>	broom	Convert Statistical Analysis Objects into Tidy Data Frames	0.4.2	
<input type="checkbox"/>	caret	Classification and Regression Training	6.0-77	
<input type="checkbox"/>	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1	
<input type="checkbox"/>	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0	
<input type="checkbox"/>	class	Functions for Classification	7.3-14	
<input type="checkbox"/>	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.6	
<input type="checkbox"/>	codetools	Code Analysis Tools for R	0.2-15	
<input type="checkbox"/>	colorspace	Color Space Manipulation	1.3-2	
<input type="checkbox"/>	compiler	The R Compiler Package	3.4.2	
<input type="checkbox"/>	config	Manage Environment Specific Configuration Values	0.2	
<input type="checkbox"/>	crayon	Colored Terminal Output	1.3.4	
<input type="checkbox"/>	curl	A Modern and Flexible Web Client for R	3.0	

Install seçildikten sonra ise yükleyeceğimiz paket ismi aşağıdaki resimdeki gibi “Packages” kısmında yazılır ve alt kısımdaki install seçilir.

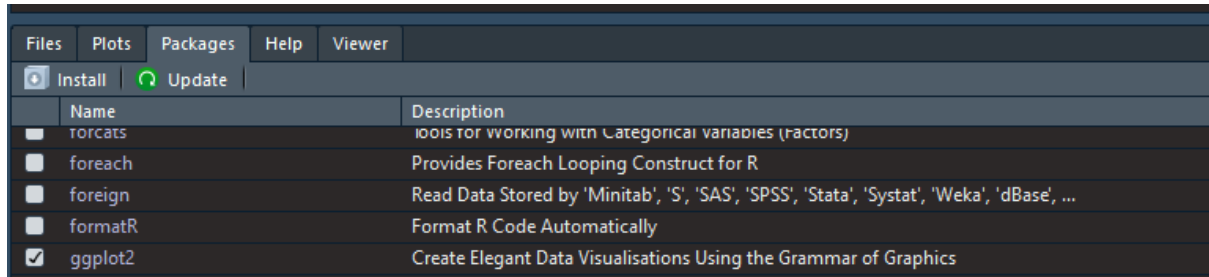


Bu işlemlerden sonra ekranda yükleniyor imgesi çıkar ve aynı zamanda Konsol kısmında yükleme ile ilgili yazılar aktıktan sonra yuklenen paketin hangi konuma kurulduğu ile ilgili bilgi yazisiyla paketin kurulumu tamamlanmis olur.

Aynı kurulum islemini ise editore ya da konsola yazacagimiz kod araciligıyla da yapabiliriz.

```
install.packages("ggplot2")
```

Paket kurulduktan sonra yükleme işlemine geçebiliriz. Bu da iki şekilde gerçekleşir. Birincisi yine paketler kısmında kurulum yaptığımız paketi seçerek yüklemis oluruz. Bu paket yükleme işlemini RStudio programında actığımız her oturumda yapmamız gereklidir.



```
library("ggplot2")
```

Paketi yüklerken ikinci yöntemimiz ise soldaki kodu yazmaktır.

Hangi paketlere sahip bulunduğunu ise aşağıdaki kod yardımıyla öğrenebiliriz.

```
installed.packages()
```

Komut Yazma

RStudio kullanıcılarına 2 şekilde kod-komut yazma olanağı sağlamaktadır. Bunlardan birincisi Konsol üzerinden tek satirlik kodlar yazarak işlem yapmaktır. İkincisi ise açtığımız betik yani script üzerinden komut dosyaları yazmaktır.

Komut Dosyaları Yazma

RStudio'nun Kaynak Sekmeleri (Source) yerleşik bir metin düzenleyici görevi görür. Konsolda komutlar genellikle R işlevlerini çıkarmadan önce yazılır. Komut dosyası oluşturma aslında eseriniz gösterir. Görevi tamamlamak için gereken işlevler sırası, bir görevi belgelemek veya otomatikleştirmek için yazılır. Komut dosyaları ilk başta hantal gibi görünse de uzun vadede, özellikle tekrarlanan görevler için zaman kazandırır. Bazı yararları:

- Bir görevin nasıl gerçekleştirileceği konusundaki talimat hatırlatma görevi görür
- Zamandan tasarruf sağlayan ve değişikliklerin tekrar değerlendirilmesini sağlayan hızlı yinelemeye izin vererek yazılanların güncellenmesini sağlar.
- Kullanıcı hatası olasılığını azaltır.

R Dosyalarını Kaydetme

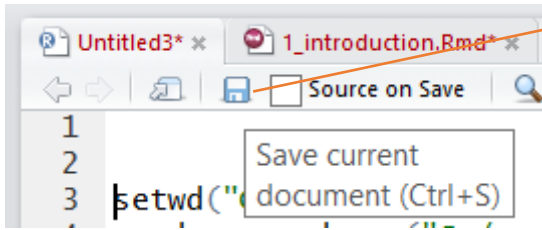
R'de, yaptığınız işi takip etmek için çeşitli dosya türlerini kaydedebilirsiniz. Dosya türleri şunlardır: çalışma alanı, komut dosyası, geçmiş ve grafikler. Sık sık kaydetmek önemlidir, çünkü R, diğer yazılımlar gibi periyodik olarak çökebilir. Bu tür sorunlar özellikle büyük dosyalarla çalışırken mümkündür. Çalışma alanınızı komut satırından veya Dosya menüsünden R'ye kaydedebilirsiniz.

R komut dosyası (.R)

R komut dosyası, yazdığınız R komutlarının metin dosyasıdır.

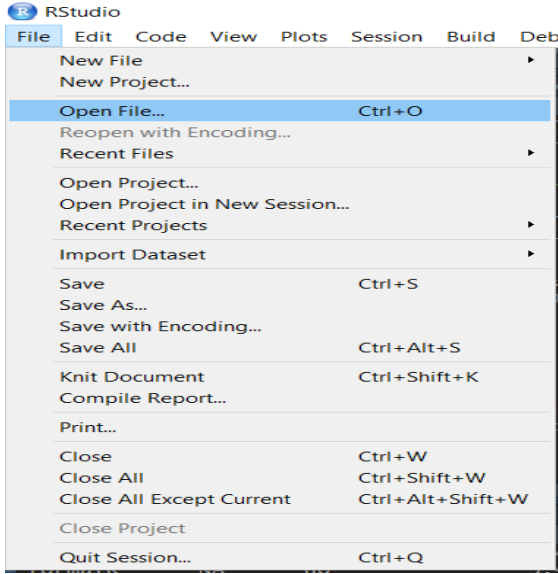
Komut dosyalarınızı (ister R Editör'de isterse Not Defteri gibi başka bir programda yazılmış olsunlar) kaydetmek isteyebilirsiniz, böylece ileride bunlara başvurabilir, gerektiğinde düzenleyebilir ve ne yaptığınızı takip edebilirsiniz.

R komut dosyalarını RStudio'ya kaydetmek için, R komut dosyası sekmenizdeki kaydet düğmesini tıklamanız yeterlidir.



Komut dosyalarını .R uzantısıyla kaydedin. R, komut dosyalarının yalnızca bu uzantıyla kaydedildiğini varsayar. Başka bir metin düzenleyicisi kullanıyorsanız, komut dosyalarınızı R'ye kaydetme konusunda endişelenmenize gerek yoktur. Metin dosyalarını RStudio metin düzenleyicisinde

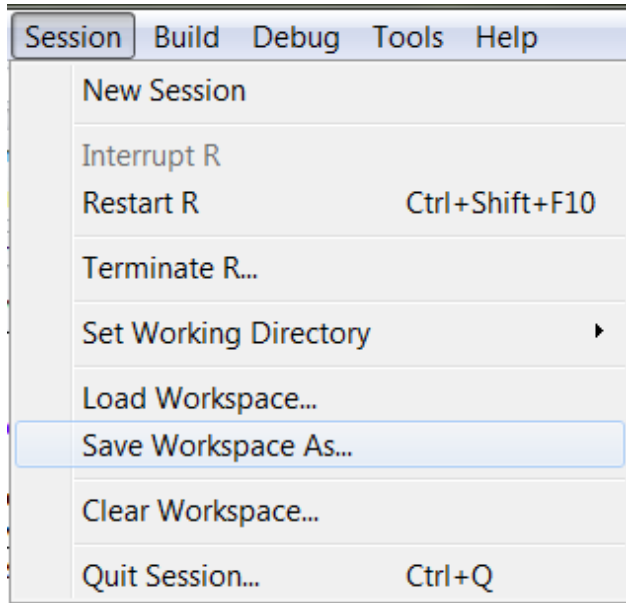
açabilirsiniz, ancak R komut dosyalarını yazmak veya kaydetmek için Microsoft Word'ü kullanmak genellikle kötü bir fikirdir. "" Alıntılar gibi bazı klavye karakterleri Word'de aynı şekilde depolanmaz (örneğin "" ve ""). Fark büyük ölçüde insan gözüyle ayırt edilemez, ancak R'de çalışmaz



Bir R komut dosyasını (R Scripti) açmak için ise aşağıdaki resimdeki adımlarla dosyayı yüklediğiniz konumdan alarak çalıştırabilirsiniz.

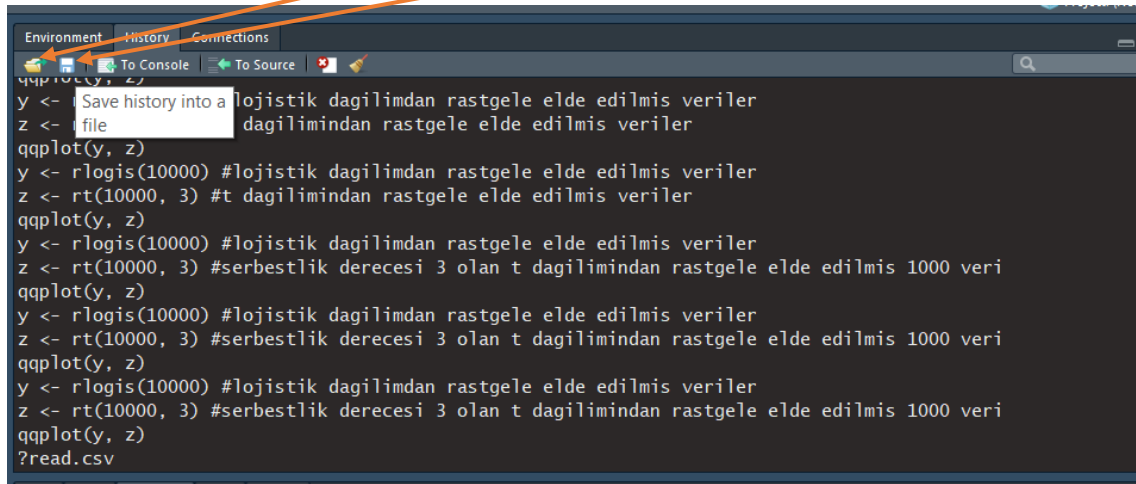
Çalışma Alanı (.Rdata)[Workspace]

R çalışma alanı, R oturumunuz sırasında oluşturduğunuz veya yüklediğiniz tüm veri nesnelerinden oluşur. Q () yazarak veya uygulama penceresinden çıkarak R'den çıktığınızda, R çalışma alanınızı kaydetmenizi ister. Evet'i seçerseniz, R, .Rdata adlı bir dosyayı çalışma dizininize kaydeder. R'yi bir sonraki açışınızda ve .Rdata çalışma alanınızı yeniden yüklediğinizde, tüm veri nesneleriniz R'de kullanılabilir ve yazdığınız tüm komutlara klavyenizdeki yukarı ok ve aşağı ok tuşlarını kullanarak erişilebilir. Ayrıca **Session>Save Workspace As** seçeneğini tıklayarak R oturumunuz sırasında istediğiniz zaman çalışma alanınızı kaydedebilir veya yükleyebilirsiniz.



.Rhistory (R geçmiş)

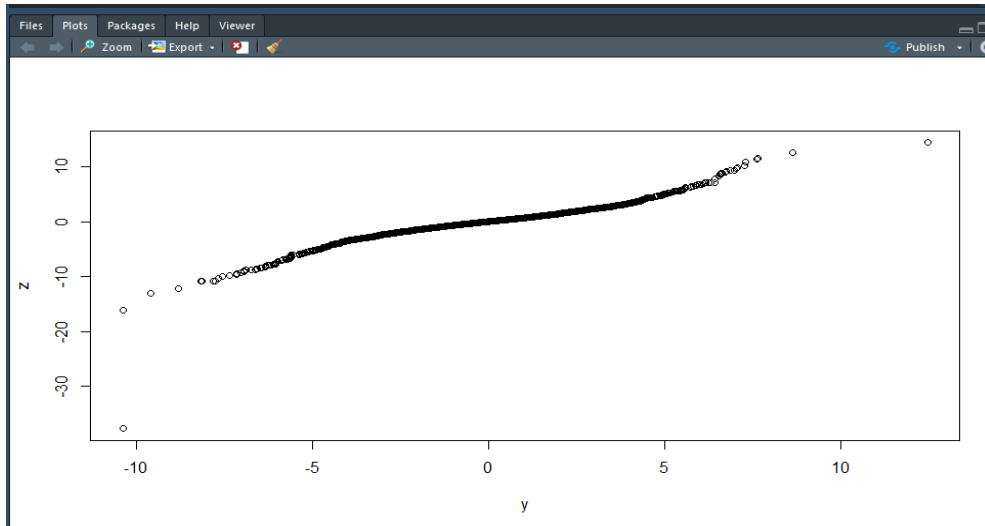
Bir R geçmişi dosyası, yazdığımız tüm komutların bir kopyasıdır. Bir R komut dosyasındaki tüm adımlarınızı belgelemediyseniz faydalı olabilir. R dosyası gibi, Rhistory dosyası da yürüttüğünüz tüm komutları listeleyen bir metin dosyasıdır. Sonuçların kaydını tutmaz. R geçmişinizi geçmiş sekmesinden yüklemek veya kaydetmek için Dosya Aç veya Kaydet düğmesini tıklayın. Bir Rhistory dosyası yüklerseniz, önceki komutlarınız tekrar yukarı ok ve aşağı ok tuşlarıyla kullanılabilir hale gelir.



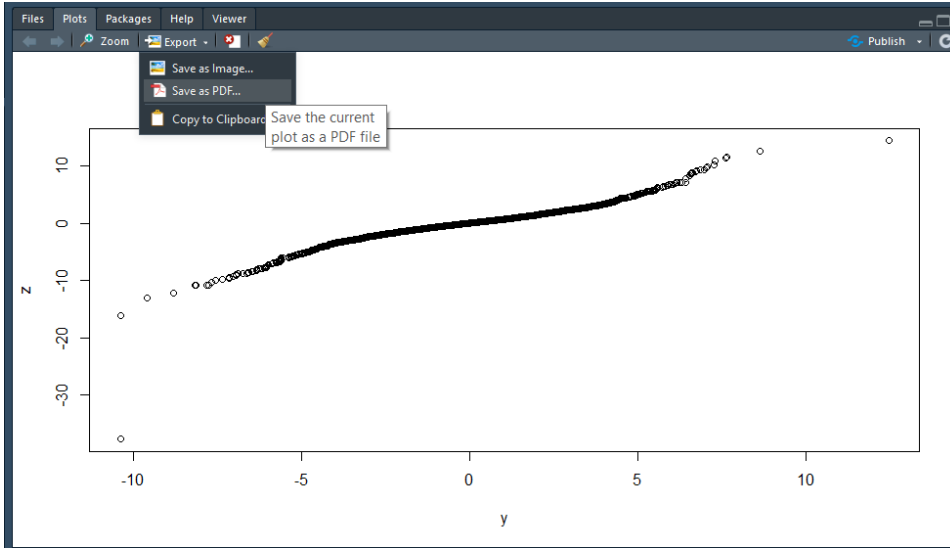
Grafik

Grafik çıkışlar çeşitli formatlarda kaydedilebilir. Bir grafiği kaydetmek için:

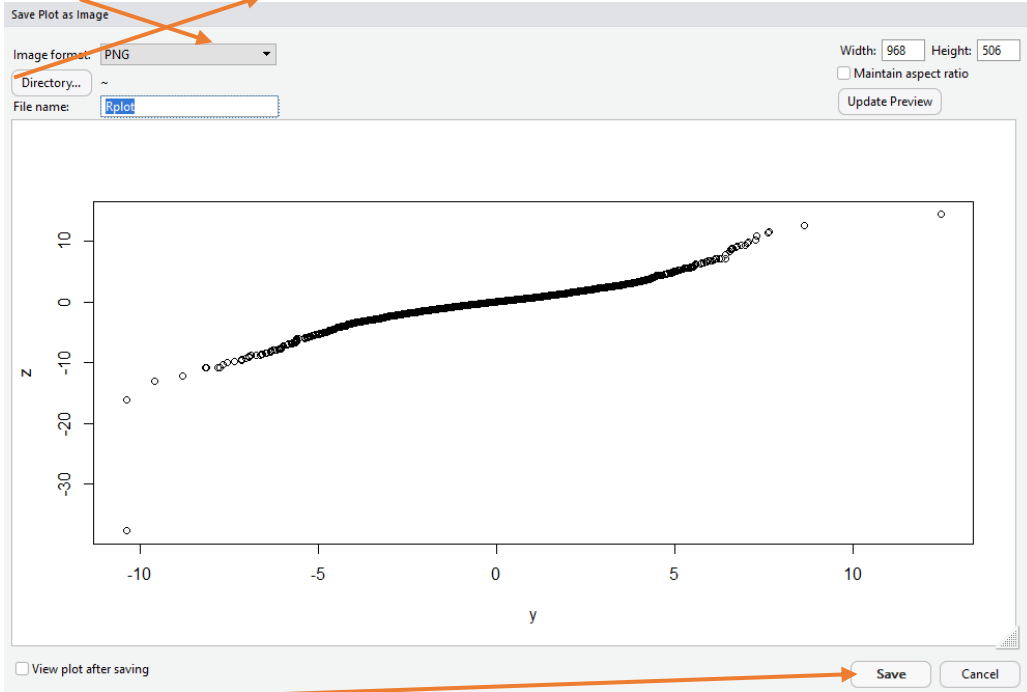
- Plots(Grafikler) Sekmesi penceresine tıklayın,



- Dışa Aktar düğmesine tıklayın ve istediğiniz formatı seçin

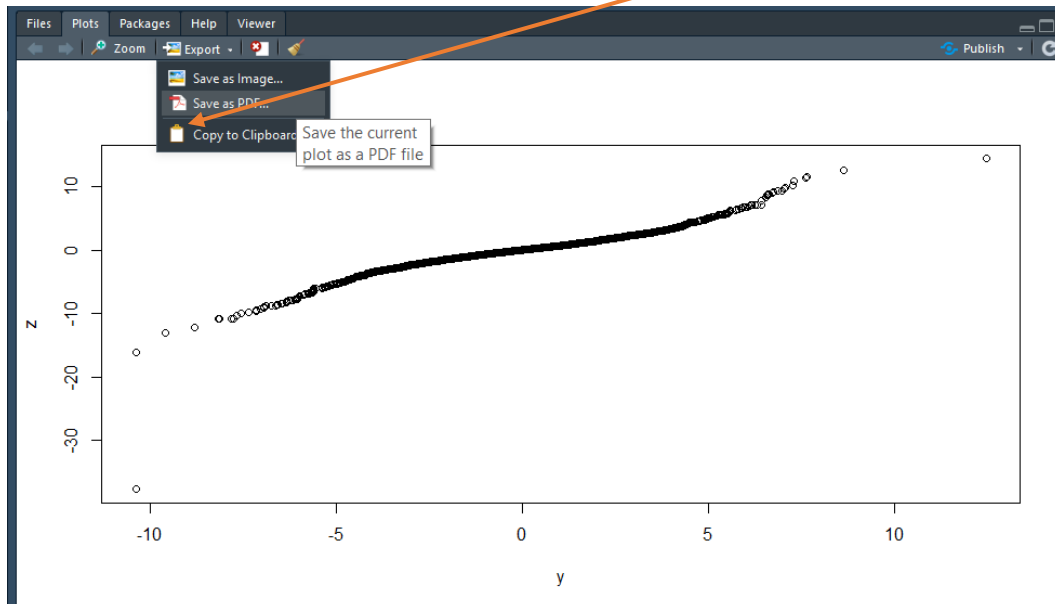


- Dışa aktarma ayarlarını istediğiniz gibi değiştirin

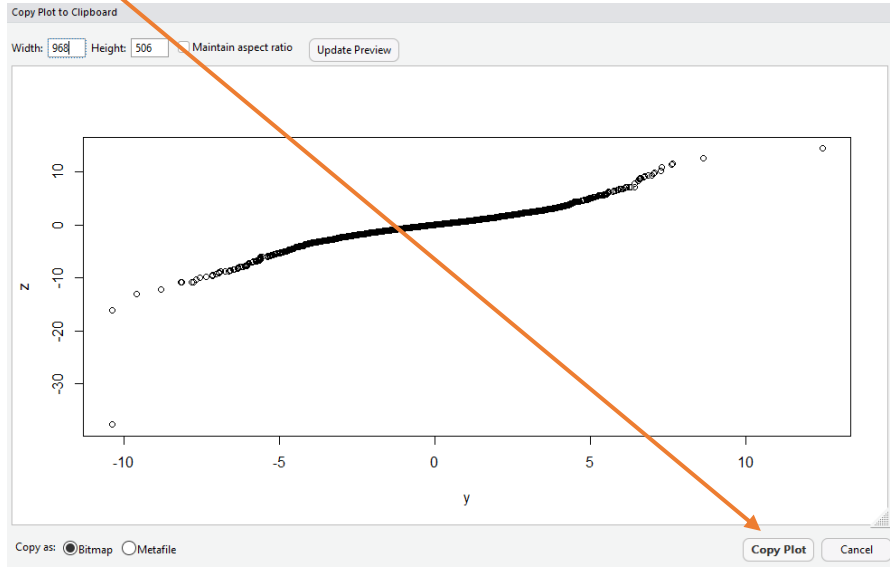


- Kaydet'e tıklayın.

Not: Bu grafik farklı bir dosya içine kopyalanmak istenirse Copy to Clipboard seçilir.



Grafigi Kopyala seçeneđi seçilip kopyalanacak yerde CTRL + V ile yapıştırılarak işlem bitirilir.



IST374

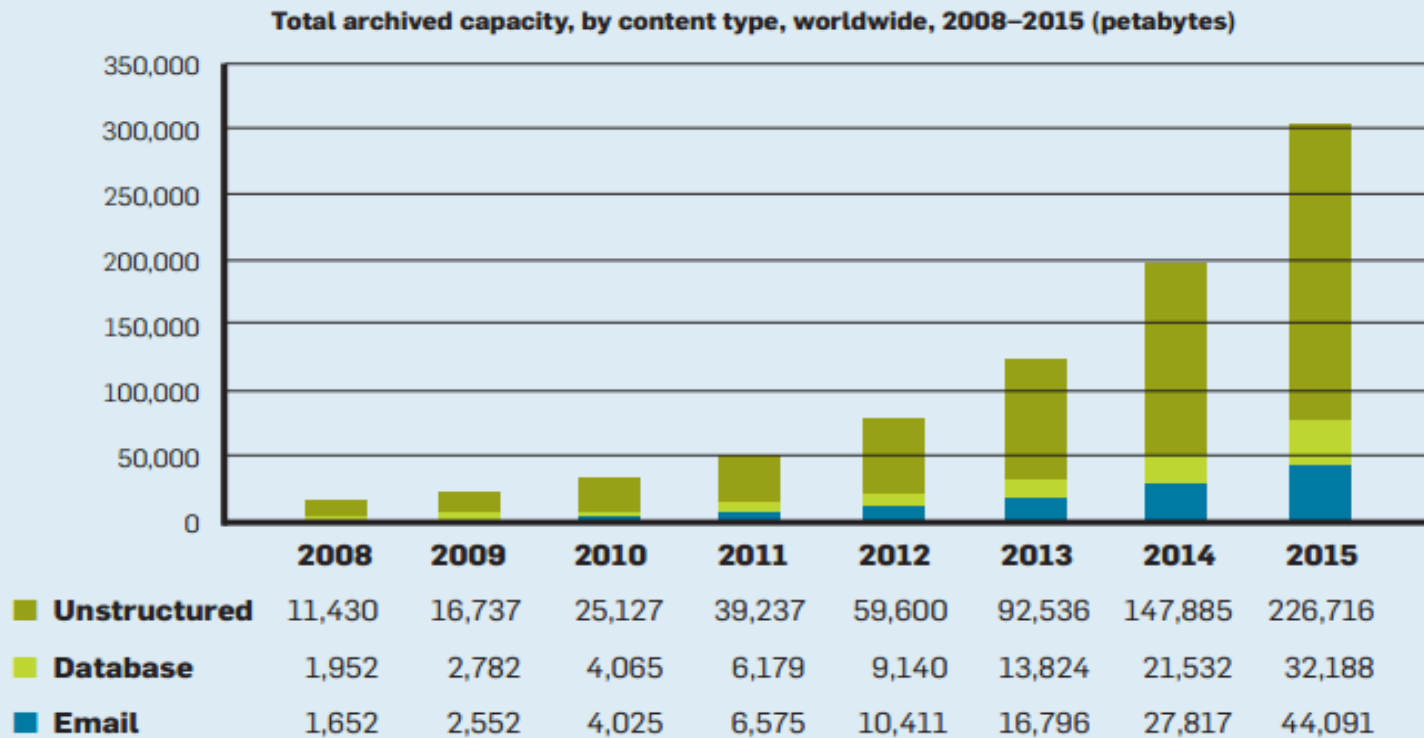
VERİ ANALİZİ DERSİ

GİRİS & TEMEL KAVRAMLAR

Umut YAMAK

Veri?

Figure 1. Projected growth of unstructured and structured data.



Data are everywhere...



Computer Science

- Leibniz – Binary Logic.
- Turing machines
- Information Theory
- Weiner & Cybernetics
- Von Neumann Architecture.
- Babbage, Lovelace
- Boolean Algebra
- Punch cards.
- Sort & Search Algorithms – Dijkstra, Kruskal, Shell Sort, ...
- Heuristics – Simulated Annealing, ...
- Text/ string search
- 1974 Peter Naur "Concise Survey of Computer Methods", **Data Science, Datalogy**
- Knuth – Art of Computer Programming.

Data Technology

- Cartography
- Astronomical Charts.
- William Playfair
- Charles Minard
- Florence Nightingale.
- First IBM Computers
- DBMS.
- Removable Disk drives
- Relational DBMS.
- Graph Algorithms
- Multigrid methods
- Tree based methods.
- Desktop, floppy
- SQL, OOP
- High level languages.
- 1989 First KDD Workshop
- Gregory Piatetsky-Shapiro.
- William Cleveland: Data Science
- Leo Breimann: Statistical Modeling: 2 Cultures.

Visualization

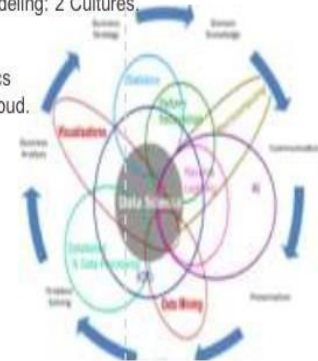
- Calculus
- Logarithms
- Newton-Raphson.
- Optimization Methods
- Fourier and other transforms
- Matrix & Generalizations
- Non-euclidean geometries.
- Applications to Military, manufacturing, Communications.
- Networks
- Assignment Problems
- Automation
- Scheduling
- John Tukey
- Jacques Bertin.
- Edward Tufte.
- Grammar of Graphics
- Word Cloud, Tag Cloud.

Mathematics/ OR

- Probability
- Correlation
- Bayes Theorem.
- Regression, Least Squares
- Time Series.
- Theoretical Foundations of Modern Stats
- Hypothesis, DOE
- Mathematical Statistics.
- 1962 John W. Tukey, Future of Data Analysis
- 1976 – SAS Institute
- 1977 The International Association for Statistical Computing (IASC).
- Decision Science
- Pattern recognition
- Machine learning.

Statistics

- Bayesian Methods
- Time Series Methods (Box Cox, Survival, etc.)
- Stochastic Methods.
- Simulation, Markov
- Computational Statistics.



Pre 1800s

1800-1900

1900-1940

1940-1960

1960

1970

1980

1990

2000

2010

Veri analizcileri uzerine

Analyzing the Analyzers

An Introspective Survey of
Data Scientists and Their Work

Harlan D. Harris, Sean Patrick Murphy
& Marck Vaisman



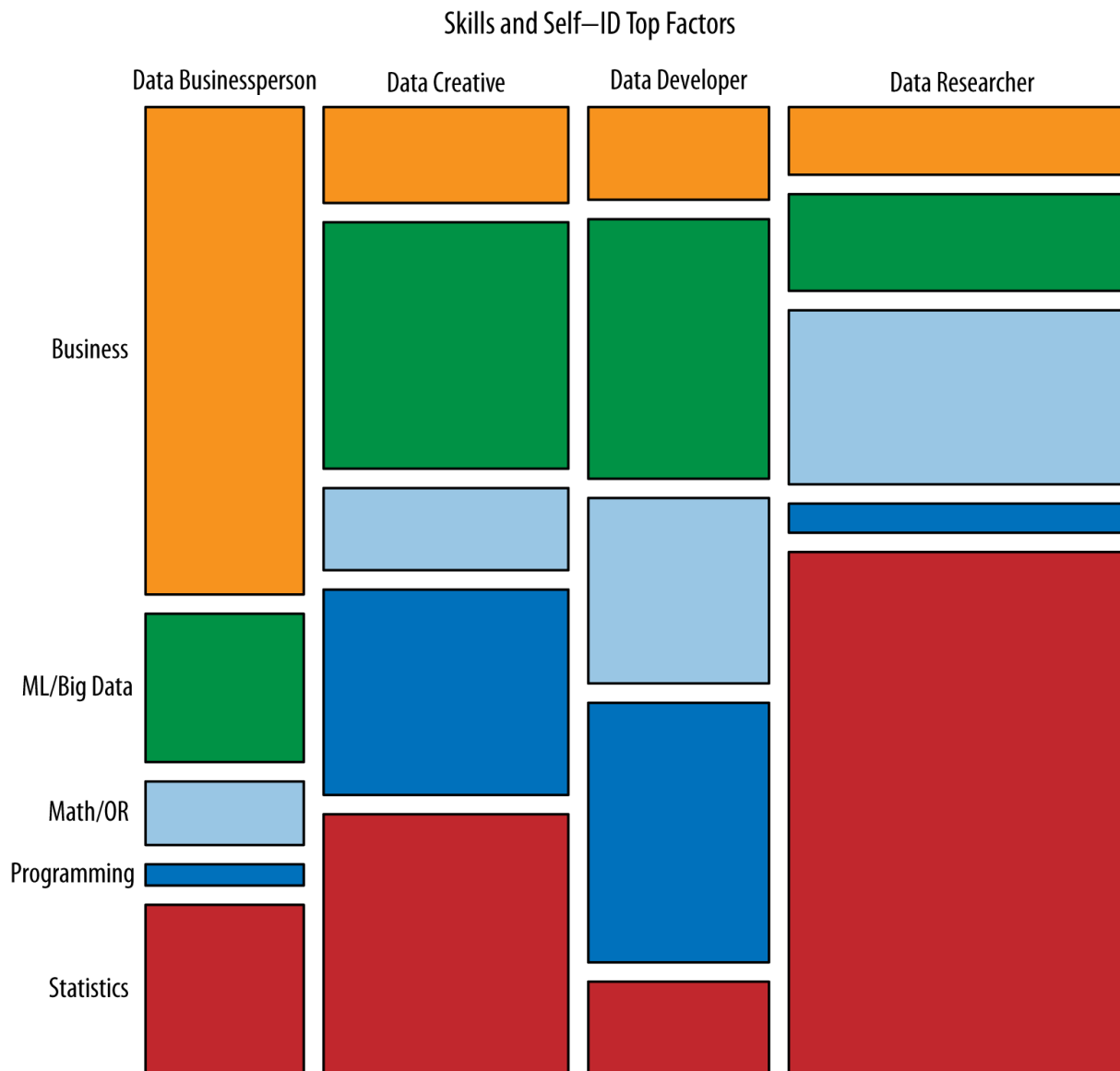
Business	ML/ Big Data	Math / OR	Programming	Statistics
Product Development	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Veri analizcileri uzerine

Analyzing the Analyzers

An Introspective Survey of
Data Scientists and Their Work

Harlan D. Harris, Sean Patrick Murphy
& Marck Vaisman



Istatistik?

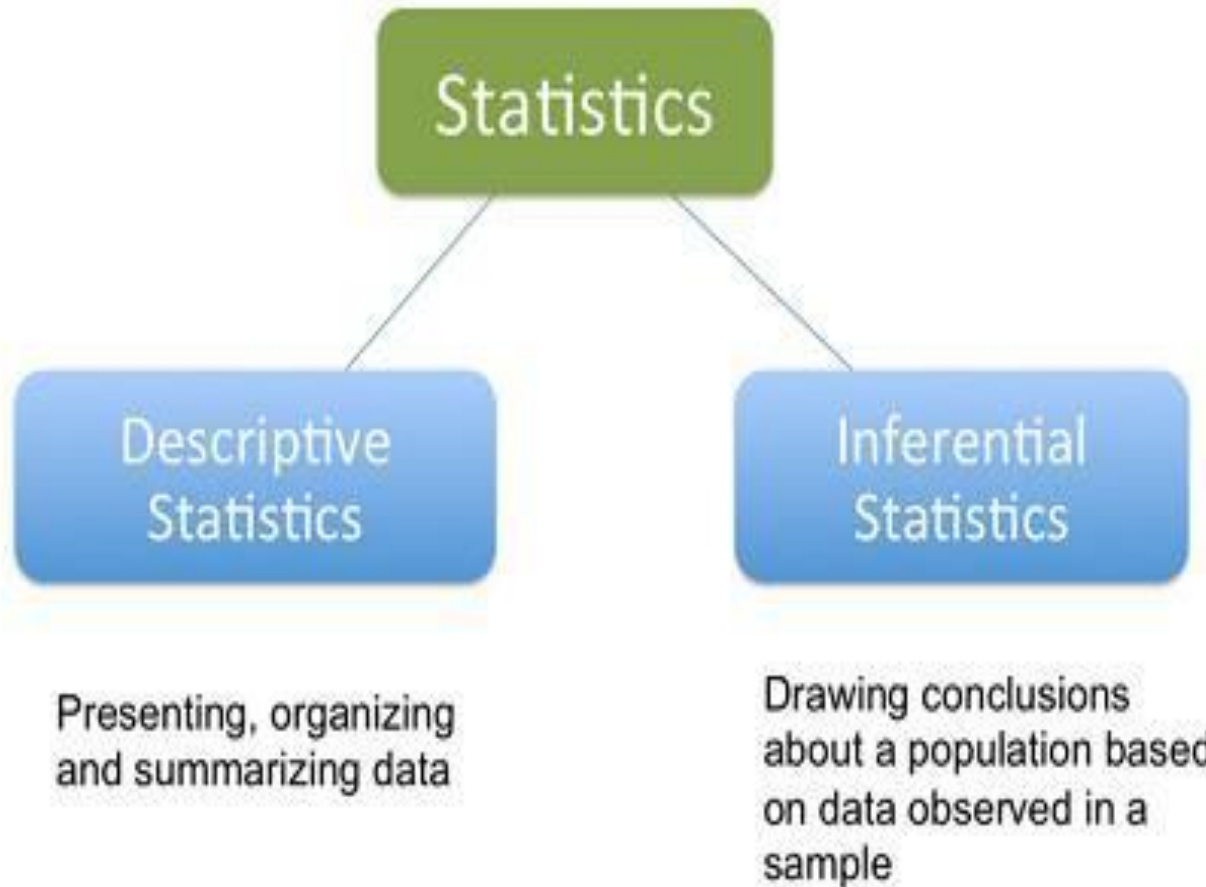
İstatistiğin birinci anlamı, belli konularda toplanan sayısal değerler ile ileri sürülen bir takım figürler, değerlerdir.

İkinci olarak, birçok farklı alanda ortaya çıkan soruları ve/veya sorunları yanıtlamak için yapılan araştırmalarda temel unsur olan verileri doğru bir şekilde toplayarak analiz ederek bazı anlamlı bilgilere ulaşma işlemini “**istatistik**” olarak nitelendirebiliriz. Bu amaca ulaşmak için, istatistikte aşağıdaki adımlardan faydalanılır:

- Deneme desenlerinin planlanması veya tasarım,
- Verilerin toplanması,
- Verilerin özetlenmesi ve analizi,
- Sonuçların yorumlanması ve genelleştirilmesiyle ilgili yöntemler ve prensiplerden yararlanılır.

Istatistik Tipleri

İstatistik yöntemler veya işlemler temel olarak iki kategoriye ayrılarak gruplandırılabilir.



İstatistik Tipleri

Tanımlayıcı İstatistik Yöntemler

Tanımlayıcı istatistik veya açıklayıcı istatistik (descriptive statistics), gözlem değerleri üzerinde yapılan ilk istatistiklerdir. Bir iş yerinde çalışan kişilerin ortalama ücreti, istatistik dersini almakta olan öğrencilerin ortalama notu vb. istatistikler açıklayıcı istatistiklere örnek verilebilir.

Istatistik Tipleri

Yorumsal İstatistik Yöntemler

Yorumsal istatistik (inferential statistics) veya genelleştirme istatistiği (inductive statistics), sayma, sorgulama ve şansa bağlı olarak yapılan örneklemeler aracılığıyla çalışılan populasyon hakkında mümkün olduğunca fazla bilgi edinmeyi konu edinir. Bu aşamada, elde edilen ortalamalar kullanılarak iki ortalamanın testi (t-testi) ve çoklu ortalamaların testi (varyans analizi) gibi yapılır.

“İstatistik dersini alan öğrencilerin notları %95 güvenilirlikle 35 ile 86 arasındadır” örneği yorumsal istatistiğe bir örnek olarak verilebilir.

Yorumsal istatistikte iki temel yöntem kullanılır. Bunlar tahminleme ve hipotez testidir. Tahminlemede, populasyondan çekilen örneklerden populasyona ait parametreleri tahmin etmek ve oluşturulan tahminler hakkında güven aralığı vermek söz konusudur. Hipotez testinin en yaygın kullanımında bir sıfır hipotez ileri sürülür ve verinin bunu red etmek için yeterince güçlü olup olmadığı belirlenir.

Bazi Temel Kavramlar

Farklı deneklerden yada hastalardan farklı değerler alabilen, yani bireyden bireye, gözlemden gözleme değişiklik gösteren özellikler **değişken** olarak adlandırılır.

Değişkenler ifade edilişlerine göre en basit hali ile sayısal ve kategorik olarak sınıflandırılabilirler.

1.Sayısal Değişkenler

Ölçme yapılarak yada sayma yoluyla elde edilen değişkenlerdir.

Örnekler:

- Yaş
- Vücut kitle indeksi
- Ölçekten elde edilen toplam puan
- Hastanede kalış süresi
- Hastanın maaliyeti

Bazi Temel Kavramlar

2.Kategorik değişkenler

Düzeyleri sözel olarak ifade edilenler yada sayısal bir değişkenin sınıflandırılmasından elde edilmiş değişkenlerdir. Değişkenin aldığı değerler belirli gruplar altında toplanabilir.

Örnekler:

- Cinsiyet (kadın-erkek)
- Diyabet (var-yok)-
- HT (var-yok)
- Test sonucu (pozitif-negatif)
- Gelir (düşük-orta-yüksek)
- Hastalığın evresi (I-II-III)
- Eğitim Durumu (ilkokul-ortaokul- lise-universite)

Asagidaki degiskenleri kategorik ya da sayisal olarak siralayiniz.

- Sac Rengi
- Plaka Numarasi
- PinKodu
- Posta Kodu
- Ayakkabi Numarasi
- Tisort Boyutu (S, M, L)
- Boy uzunlugu
- Okuldan eve varis suresi

Olcme Duzeyleri

1.Nominal (Sınıflandırma) Ölçeği

Değişkenler arasında hiçbir sıralama yapılmaksızın, ortak

özellikler yönüyle değişkenlerin kategorilere ayrılmasıdır. Sınıflandırmada kategoriler homojen olmalı, karşılıklı birbirini dışta tutmalı yani bir değişken aynı anda iki kategoriye birden dahil edilmemeli ve kategoriler arasında sıralı ilişkiler hakkında hiçbir ilişki olmamalıdır.

Örneğin doğum yeri, cinsiyet, dini mezhep, politik tercihler gibi değişkenler nominal ölçek örnekleridir.

Olcme Duzeyleri

2.Siralama (Ordinal) Ölçeği

Bu ölçekte veriler arasında bir sıralama yapılması söz konusudur. Burada verilmiş olan puan bir sıra dizisini göstermektedir. 1 puan değişkenin en düşük değerini verirken 2, 3,, puanlar ile daha yüksekdeğerler ifade ederek bir sıralamayapılması söz konusudur. Bu sıralamada 1 ile 2 arasındaki uzaklıkla 2 ile 3 arasındaki uzaklığın aynı olduğunu ya da 2 büyüklüğünün 1 ve 3 ile orantılı olduğu söyleyemeyiz. Dolayısıyla sıralamada birinci sıradaki nesne ile ikinci sıradaki nesne arasında daha büyük, daha küçük, daha iyi, daha kötü, daha canlı, daha cansız vb. ilişkiler söz konusudur. Bu ölçekte kullanılan uygun istatistikler ortanca, yüzdelik, çeyrek sapma, sıra farkları korelasyon katsayısı, işaret testi, Mann-Whitney U –testi vb.dir.

Olcme Duzeyleri

3.Aralık (Mesafeli) Ölçeği

Sıralama ölçeğinden farklı olarak değişkenler arasındaki puanlar açısından

eşit mesafe söz konusudur.

Aralık ölçeğindeki veriler düzen ve uzaklık özelliklerine sahiptir. Ancak aralık ölçeği verileri ile nesneler arasında mutlak büyüklük karşılaştırması yapmak mümkün değildir. Ayrıca aralık ölçeğinde sıfır noktası (mutlak orjin) yoktur. Aralık ölçeği ile hemen hemen bütün istatistiksel analizler kullanılmakla beraber en yaygın kullanımı aritmetik ortalamadır. Aritmetik ortalama dışında standart sapma, pearson korelasyon katsayısı, t-testi, F-testi kullanılabilir.

Olcme Duzeyleri

4.Oran Ölçeği


Yukarıda sayılan bütün ölçeklerin özelliklerini taşıyan oran ölçeğinde mutlak veya doğal bir orjini (sıfır noktası) de vardır.

Oran ölçekleri negatif değerler içermemektedir.

Her türlü matematiksel ve istatistiksel işleme imkan sağlamaktadırlar. Örneğin, pazar payının, satışların, gelirin ölçülmesi vb. oranlı ölçeğe uygundur.

Bu degiskenlerin olcme duzeylerini belirtiniz.

- Cinsiyet
- Sınıf
- IQ
- Göz rengi
- Posta kodu
- Yükseklik
- Sıcaklık (F, C)
- Ağırlık
- Zaman



A large, bold, black question mark is centered on a white background. Above the white area is a solid gray horizontal bar. The question mark is composed of a thick, curved hook and a solid, rounded dot below it.

Veri Analizinde Grafiksel Yöntemler

1.Histogram

Histogram, gruplandırılmış bir veri dağılımının sütun grafiğiyle gösterimidir. Bir histogramın amacı, tek değişkenli bir veri kümesinin dağılımını grafiksel olarak özetlemektir. Histogram grafiksel olarak aşağıdakileri gösterir:

- Verilerin merkezini (yani konumunu).
- Verinin yayılımını
- Verilerin çarpıklığını.
- Aykırı değerlerin varlığı.

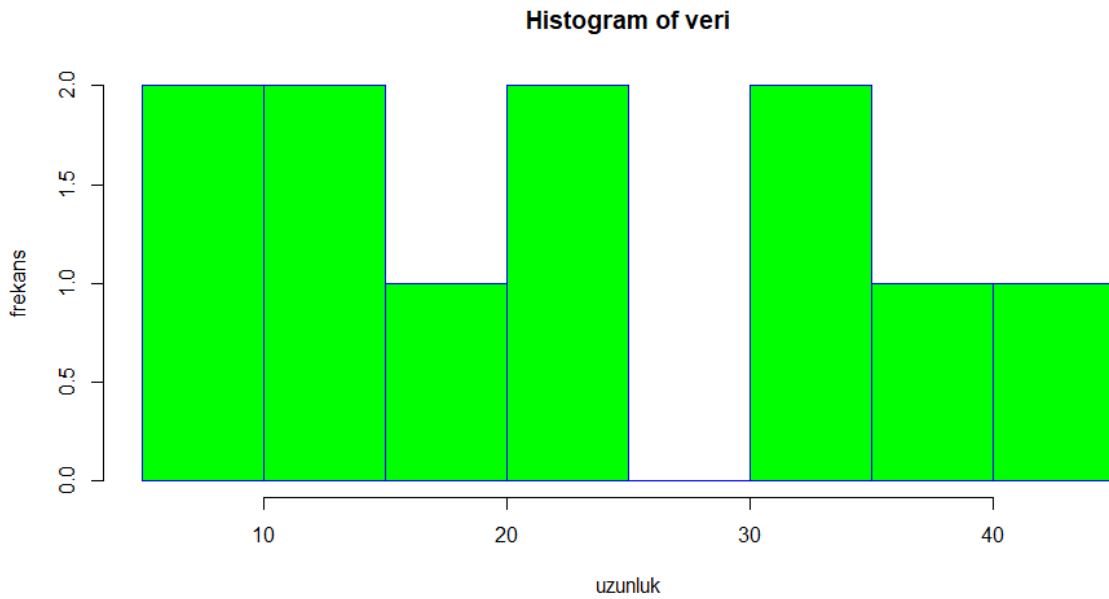
Bu özellikler, verilerin uygun dağılımı için güçlü göstergeleri sağlamaktadır.

Dersimizde kullandığımız R, *hist()* işlevini kullanarak histogram oluşturur. Bu işlev şu şekildedir; *hist(x, ...)* burada x kullandığımız verinin ismiyken noktalarla ifade edilen kısımlar çok sayıda parametreden oluşur. Bunları teker teker incelemek adına help kısmına *hist* yazarak ulaşılabilir.

R üzerinden bir örnek:

```
# grafik için bir veri seti oluşturalım.  
veri <- c(9,13,21,8,36,22,12,41,31,33,19)  
  
# histogram çizdirir.  
hist(veri,xlab = "uzunluk",ylab="frekans",col = "green",border = "blue")
```

Üstteki kodlarla temel olarak aşağıdaki gibi bir histogram çizdirilebilir.



2.Kutu Grafiđi (Box-plot)

Veri kümeleri arasındaki benzerlik ya da farklılıkları görmek için kullanılır. Kutu-grafigi kullanılarak veri kümesinin

- Konumu
- Yayılımı
- Çarpıklığı
- Kuyruk uzunluğu
- Aykırı değeri

tespit edilebilmektedir. Veri setinin yayılımı ve konum değeriyle verilerin dağılımı hakkında ipucu da vermektedir.

Ayrıca, birden fazla veri seti olduğunda kutu grafikleri veri kümeleri arasındaki benzerlik ve farklılıkları görmek için kullanılır. Bu ilerleyen konularda gruplar arası kıyaslama yapılırken bizlere testlerimizi destekleyici argümanlar sunacaktır.

Örnek: Aşağıdaki veri setinde Türkiye’de nüfusu en çok olan 16 şehrin 2018 yılına ait nüfus sayıları 10 binlik olarak aşağıda verilmektedir. Bu verilere göre kutu grafiğini çizmek için gerekli olan noktaları hesaplayınız ve kutu grafiğinden çıkan sonucu yorumlayınız.

No	Şehir	Nüfus(×10000)
1	Samsun	134
2	Kayseri	139
3	Manisa	143
4	Hatay	161
5	Diyarbakır	173
6	Mersin	181
7	Kocaeli	191
8	Gaziantep	203
9	Şanlıurfa	204
10	Konya	221
11	Adana	222
12	Antalya	243
13	Bursa	299
14	İzmir	432
15	Ankara	550
16	İstanbul	1506

Bu veri setindeki toplam gözlem sayısının 16 olduğu görülür.16 elemanımız olduğu için medyanı bulurken veri setinde yapılan sıralamaya göre ortadaki iki elemanın ortalamasını alırız. Bunlar 8. Ve 9. şehirlerdir. Buna göre **medyan yani ortanca değeri 203.5** bulunur.

Q1 değeri bulmak içinse 4. ve 5. değeri ortalaması alınır ve $\frac{161+173}{2} = 167$ elde edilir.

Q3 değeri bulmak içinse 12. ve 13. değeri ortalaması alınır ve $\frac{243+299}{2} = 271$ elde edilir.

Alt sinir yani minimum değeri için şu formül kullanılır,

$$\text{Min. Değer} = Q1 - 1.5 \times (Q3 - Q1)$$

Üst sinir yani Maksimum değeri için şu formül kullanılır,

$$\text{Maks. Değer} = Q3 + 1.5 \times (Q3 - Q1)$$

Bu formüllere göre alt sinir 11, üst sinir 427 bulunur.

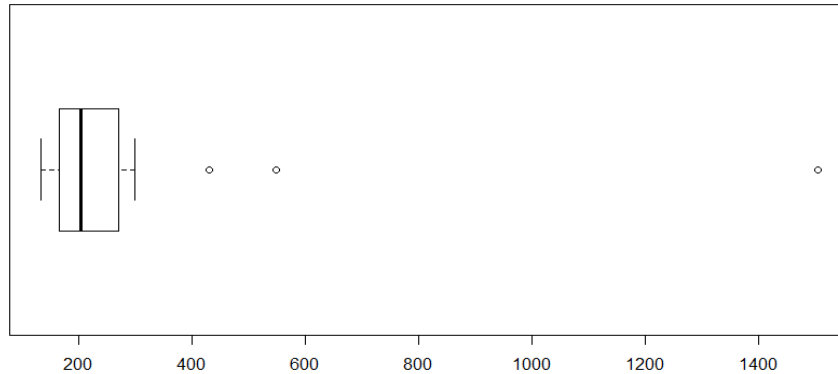
Bu veri setinde nüfusu 110.000'den az olan şehir bulunmamaktadır. Ancak, İstanbul, Ankara ve İzmir şehirlerinin nüfusları 4.270.000'den fazla olduğu için bu şehirler nüfus bakımından aykırı değer olarak belirlenir.

Yukarıdaki veri seti için bu adımlar izlenerek elde edilen kutu-grafiği aşağıdaki gibi elde edilir.

R programında bu kutu grafiğini çizdirmek istersek aşağıdaki kodu yazarız.

```
veri<- c(1506,550,432,299,243,222,221,204,203,191,181,173,161,143,139,134)
boxplot(veri,horizontal = TRUE)
```

ve aşağıdaki kutu grafiği şeklini elde ederiz.



Yorum

- Kutu grafiğine göre 3 tane gözlem noktası aykırı gözlem olarak tespit edilmiştir. Buna göre; İstanbul, Ankara ve İzmir nüfus bakımından aykırı değerdir.
- Ayrıca bu grafik verilerin dağılımı açısından ipucu vermiştir ve verilerin sağa çarpık olduğunu söyleyebiliriz.

3.Q-Q Grafiği (Q-Q Plot)

Kantil kantil (q-q) grafiği, iki veri setinin ortak bir dağılıma sahip popülasyonlardan gelip gelmediğini belirlemek için kullanılan grafiksel bir tekniktir.

Bir q-q grafiği, ikinci veri kümesinin miktarlarına karşı birinci veri kümesinin miktarlarının grafiğidir. 45 derecelik bir referans çizgisi de çizilir. İki veri seti aynı dağılıma sahip bir popülasyondan geliyorsa, noktalar yaklaşık olarak bu referans çizgisi boyunca düşmelidir. Bu

referans çizgisinden ne kadar uzaklaşırsa, iki veri kümesinin farklı dağılımlara sahip popülasyonlardan geldiğine dair kanıtlar o kadar büyük olur.

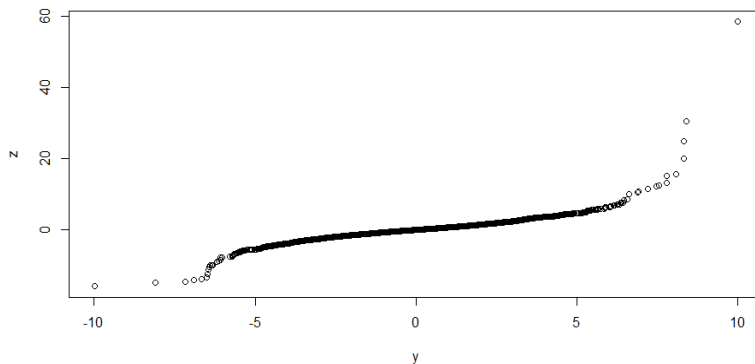
Q-q grafiğinin avantajları:

- Örnek boyutlarının eşit olması gerekmez.
- Dağılımla ilgili olarak bilgi verebilecek birçok argüman aynı anda test edilebilir. Örneğin, konumdaki değişimler, ölçek değişimleri, simetri değişiklikleri ve aykırı değerlerin varlığı bu grafikten tespit edilebilir. Örneğin, iki veri kümesi, dağılımları yalnızca konumdaki bir kayma ile farklılık gösteren popülasyonlardan geliyorsa, noktalar 45 derecelik referans çizgisinden yukarı veya aşağı yer değiştiren düz bir çizgi boyunca uzanmalıdır.
- Ayrıca normallik varsayım testi olarak da kullanılabilen bu grafik oldukça faydalıdır.

R üzerinden bir örnek:

```
y <- rlogis(10000) #lojistik dagilimdan rastgele elde edilmiş veriler
z <- rt(10000, 3) #serbestlik derecesi 3 olan t dagilimindan rastgele elde edilmiş 1000 veri
qqplot(y, z)
```

Burada ilk iki satırda yorum kısmında ifade edildiği gibi lojistik ve t dağılımlarından veri türetildi. Bu verilerin dağılımlarının ortusup ortusmedigine q-q grafiğiyle karar vermek istenirse 3. Satırdaki qqplot() işlevi uygulanır ve aşağıdaki grafik elde edilir.



Bu grafikte y ve z verileri 45 derecelik düz bir çizgi üzerinde buluşmamışlardır. Bu da bizlere verilerin dağılımlarının farklı olduklarını göstermektedir.

4. Dal-Yaprak Grafiği (Stem-and-Leaf Plot)

Nicel bir değişkenin dal ve yaprak grafiği, veri öğelerini en önemli sayısal rakamlarına göre sınıflandıran metinsel bir grafikdir. Buna ek olarak, grafiği okunabilirlik için basitleştirmek amacıyla her bir alternatif satırı bir sonraki satırla birleştiririz.

Kodumuzu aşağıdaki gibi oluşturduktan sonra alttaki şekli elde ederiz.

```
# grafik için bir veri seti oluşturalım.
veri <- c(9,13,21,8,36,22,12,41,31,33,19)
stem(veri)
```

Eldeki verinin özetinin aşağıdaki gibi olduğu gözlenir.

The decimal point is 1 digit(s) to the right of the |

0		89
1		239
2		12
3		136
4		1

Kaynaklar:

- PROF. DR. BİRDAL ŞENOĞLU Ankara Üniversitesi Açık Kaynak Ders Notları.
- https://www.tutorialspoint.com/r/r_boxplots.htm
- Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), Graphical Methods for Data Analysis, Wadsworth.
- Tukey, John (1977), EDA Exploratory Data Analysis, Addison-Wesley.

Normallik Testleri

İstatistiksel yöntemlerin birçoğu normallik varsayımına dayanır. Anova, Regresyon, t testi gibi testler verilerin normal dağılıma sahip olduğu durumlarda kullanılmaktadır. Bu nedenle de analize başlamadan önce veri setinin dağılımının normal olup olmadığının test edilmesi gerekir.

Eğer veri setinin dağılımı normal değilse veri seti aşağıda gösterilen alternatif yaklaşımlardan bir tanesi kullanılarak analiz edilmelidir.

Bu alternatif yaklaşımlar

- Parametrik Olmayan Teknikler
- Dayanıklı (Robust) Teknikler
- Dönüşümler

olarak belirtilebilir.

Literatüre girmiş birçok normallik varsayımı testi vardır. Bunlar şöyle sıralanabilir:

- Anderson – Darling Testi
- Shapiro – Wilk Testi
- Kolmogorov – Smirnov Tek Örnek Testi
- Pearson Ki Kare Uygunluk Testi
- Cramer- Von Mises Testi
- Shapiro – Francia Testi

Bu testler birbirinin türevi olarak şekilde çoğaltılmıştır. Yani bu listeyi daha da genişletmek mümkündür. Yapılan çalışmalarda ise Shapiro-Wilk testi ve Anderson-Darling testi diğerlerine göre daha güçlü bulunmuştur. Bu sebeple kullanılan paket program ya da R ve Python gibi istatistik ve veri bilimi alanında kullanılan yazılım dillerinde farklı tiplerde testler uygulanabilmektedir. Örnek vermek gerekirse SPSS’te Shapiro – Wilk ve Kolmogorov Smirnov testleri Minitab programıysa Anderson-Darling, Ryan Joiner ve Kolmogorov Smirnov testlerini sunmaktadır.

Dersimizde faydalandığımız R programında ise kullanıcıların ekledikleri paketlerle bu metotlar sürekli artış gösterebilmektedir. Temel olarak bulunan Shapiro – Wilk testi herhangi bir paket yüklemeye gerek kalmadan “shapiro.test” komutuyla hemen uygulanmaktadır. Bu testin yanında “nortest” paketinde bulunan diğer testleri de kullanabiliriz. Bu pakette aşağıdaki normallik testleri bulunmaktadır.

- ad.test Anderson-Darling normallik testi
- cvm.test Cramer-von Mises normallik testi
- lillie.test Lilliefors (Kolmogorov-Smirnov) normallik testi
- pearson.test Pearson chi-square normallik testi
- sf.test Shapiro-Francia normallik testi

Burada test edilen hipotezler şu şekildedir.

H_0 : Verilerin geldiği dağılım ile normal dağılım arasında farklılık yoktur.

H_1 : Verilerin geldiği dağılım ile normal dağılım arasında farklılık vardır.

Bu hipotezlerin yaygın olarak yazılımı ise şu şekilde olmaktadır.

H_0 : Veriler normal dağılmıştır.

H_1 : Veriler normal dağılmamıştır.

Yapılacak testin önem değeri(p) eğer testten önce belirlenen alfa değerinden yani 1.tip hatadan küçükse H_0 hipotezi reddedilebilir, değilse reddedilemez. Bu sonuca göre de karar ve yorum yapılarak normallik testi tamamlanmış olur.

Eğer normallik testine göre veriler normal dağılmamışsa; verilere logaritma - karekök gibi matematiksel fonksiyonlar uygulanarak normallik sağlanmaya çalışılır. Eğer bu işlem de sonuç vermezse parametrik olmayan testler tercih edilir. Ya da yazının girişinde de ifade edildiği gibi dayanıklı yani robust metotlardan faydalanılır.

Örnek: R üzerinden yapılacak uygulamada öncelikle, normal dağılım üzerinden veri üretimi yaparak normal olan veriler üzerinden grafiksel enstrümanlarla inceleme yapıp sonrasında R üzerinden normallik testlerini uygulayacağız. Öte yandan ise, normal olmayan bir dağılımdan veri üretimi yapıldığında elde ettiğimiz grafikleri ve test skorlarını da normal dağılan veriye göre elde ettiğimiz sonuçlarla kıyaslayacağız.

1.Asama: Normal Dağılımdan Veri Üretimi

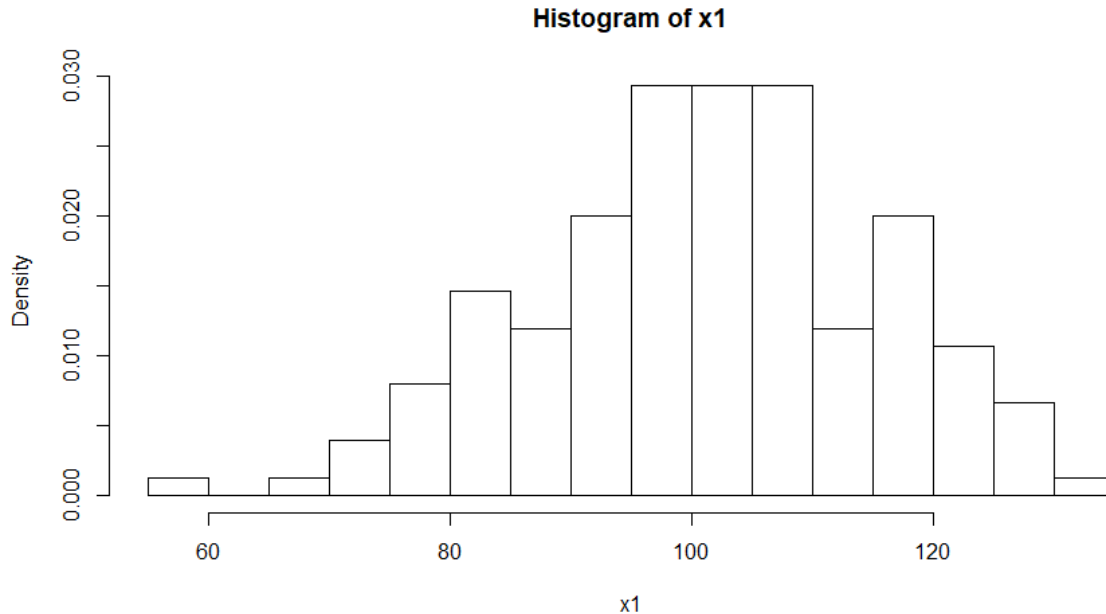
```
x1 <- rnorm(150, mean = 100, sd = 15)
```

Bu komut vasıtasıyla ortalaması 100, St. sapması 15 olup normal dağılan 150 tane veri ürettik.

2.Asama: Histogram

```
hist(x1, freq = FALSE, breaks = 20)
```

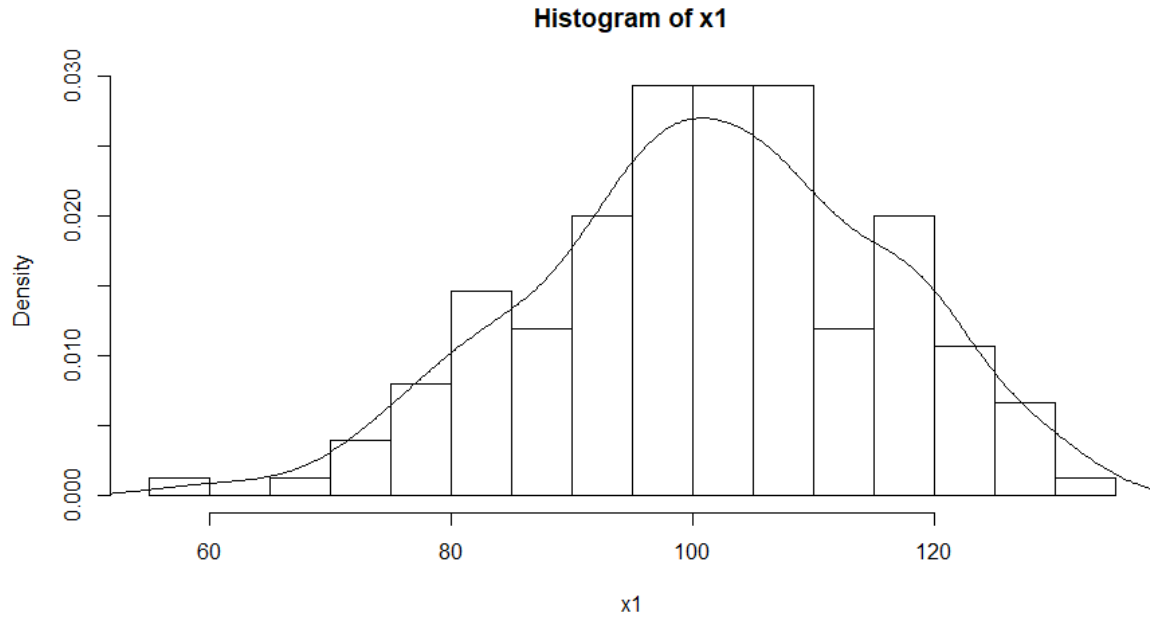
Bu kod yardımıyla ise x1 adıyla tanımladığımız veri setini 20 tane aralığa ayırarak çizdiren bir histogram oluşturduk.



- Buradan dağılım ile ilgili nasıl bir fikir yürütebiliriz?

Bu histograma yoğunluk eğrisini de eklemek istersem aşağıdaki kodu ekleyebilirim.

```
points(density(x1), type = "l") # yoğunluk çizgisini de çizdirmek istersek
```

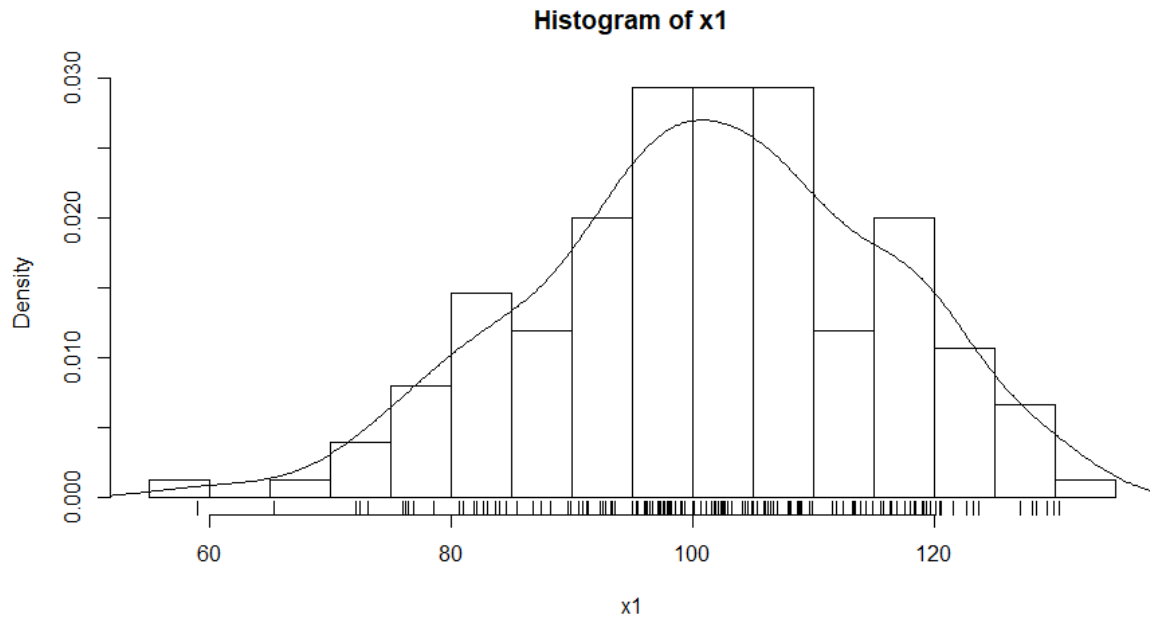


Bu sayede, dağılım ile görsel olarak daha kolay bir yorum yapabiliriz.

Verilerin hangi noktalar olduğunu görmek istersek de “rug” fonksiyonu isimize yarayacaktır.

```
rug(x1) # verileri de alt kismida göstermek icin bu komuttan faydalaniriz.
```

Böylece tam olarak su histogrami elde ederiz.

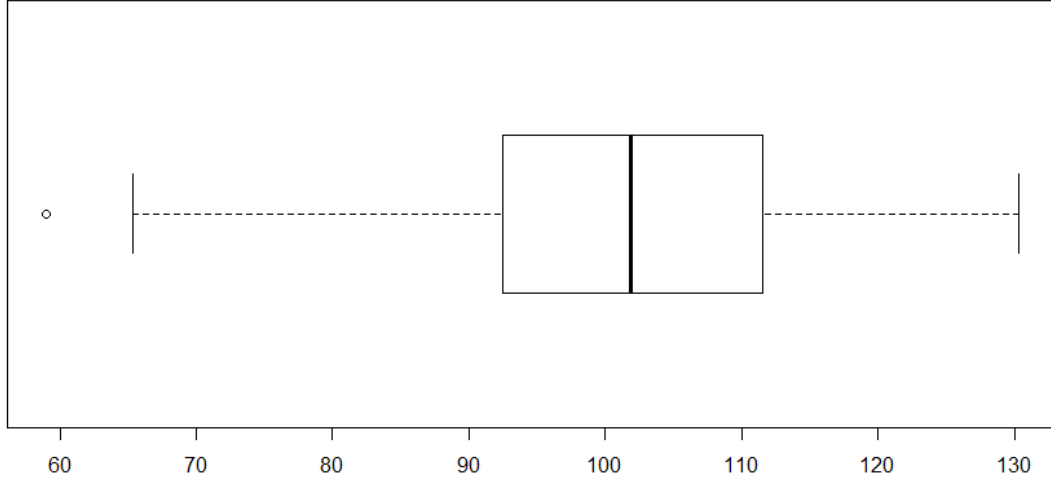


3.Asama: Kutu Grafiği

Bu grafik farklı yöntemlerle yapılabilir ama biz en temel olan fonksiyonu baz alarak grafiği çizdirelim.

```
boxplot(x1, horizontal=TRUE)
```

burada çizdirmek istediğimiz kutu grafiğini yatay eksenle çizdirmek istersek “horizontal=TRUE” argümanını fonksiyona ilave etmeliyiz.

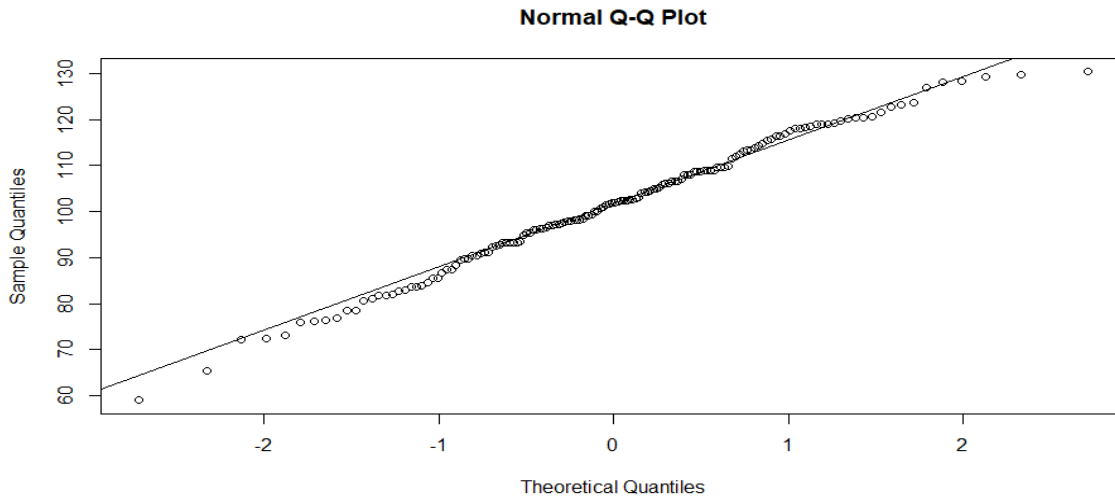


- Buradaki yorumlarınız nelerdir?

4.Asama: Q-Q Grafiği

R üzerinde 2 satır kod ile temel olarak bir Q-Q plot çizdirebiliriz. Birinci satırdaki kod iki tane veri setini eksenlere koyarak çizdirir. İkinci satırdaki kod ise yorum yapacağımız zaman bize lazım olacak olan referans çizgisini çizer. Bu doğru 45 derecelik çizilmektedir yani verilerin dağılımı birbirleriyle ne kadar örtüşüyorsa 45 derecelik bu çizgiden fazla ayrılmazlar.

```
qqnorm(x1) # veriyi normal skorlarına karşılık çizdirir.  
qqline(x1) # referans alacağımız çizgimizi çizdirir.
```



Sonuç olarak elimizdeki veriyi yine normal dağılımdan gelen veriyle karşılaştırdık ve orta çizgimizden uzağa düşen noktalar görmedik bu sebeple veriler normal dağılıma sahiptir diyebiliriz.

5.Asama: Normallik Varsayımı Testleri

R programında temel olarak bulunan Shapiro – Wilk testi herhangi bir paket yüklemeye gerek kalmadan “shapiro.test” komutuyla hemen uygulanmaktadır ve aşağıdaki gibi kodlama yapılır.

```
shapiro.test(x1)
```

Komutun içine sadece veri setimizi argüman olarak yazılır ve direkt olarak testi uygulanır ve sonucu R Studio nun konsol kısmından okuruz.

```
Shapiro-Wilk normality test
data:  x1
W = 0.99054, p-value = 0.4142
```

Karar: Buradan görüldüğü üzere p değeri önceden belirlenen ve genellikle yüzde 5 olarak alınan 1. Tip hata yani $\alpha = 0.05$ değerinden büyük olması sebebiyle yokluk hipotezi reddedilemez.

Yorum: H_0 yokluk hipotezi reddedilemediğine göre, veriler yüzde 95 güven düzeyinde normal dağılıma sahiptir.

Konunun ilk kısmında da ifade edildiği üzere R programında bulunan diğer normallik testlerini “nortest ” paketini yükleyerek ve bu kütüphaneyi çağırarak diğer normallik testlerinin fonksiyonlar aktif hale getirilir.

```
install.packages("nortest") # paketimiz elimizde olmadığı için öncelikle yüklememiz gerekli
library(nortest) # sonrasında ise bu komut ile yüklediğimiz paketi aktif hale getirelim.
```

Böylece aşağıdaki testler kullanılabilir hale getirilmiştir.

- Anderson-Darling normallik testi
- Cramer-von Mises normallik testi
- Lilliefors (Kolmogorov-Smirnov) normallik testi
- Pearson chi-square normallik testi
- Shapiro-Francia normallik testi

Bunlardan bazıları aşağıdaki gibi kodlanır.

```
ad.test(x1) #Anderson-Darling normallik testi
lillie.test(x1) #Lilliefors (Kolmogorov-Smirnov) normallik testi
cvm.test(x1) #Cramer-von Mises normallik testi için
```

Bu testlerin sonuçları ise aşağıdaki gibi sıralanır.

```
Anderson-Darling normality test
data:  x1
A = 0.27076, p-value = 0.6706
```

```
Cramer-von Mises normality test
data:  x1
W = 0.033815, p-value = 0.7889
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
data:  x1
D = 0.034305, p-value = 0.9364
```

Elde edilen skorlara göre karar ve yorumları yapınız.

Bu yapılan uygulama bir de normal olmayan dağılımdan üretilen verilerle yapılırsa neler çıkarılabilir inceleyelim.

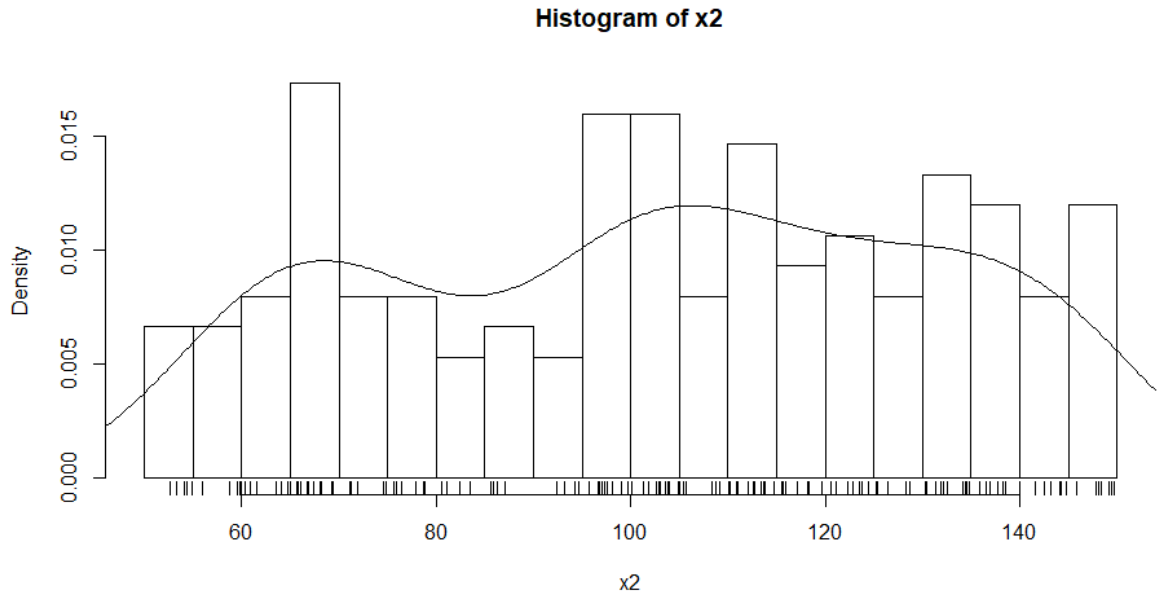
1.Asama: Normal Olmayan Bir Dağılımdan Veri Üretimi

```
x2 <- runif(150, min = 50, max = 150)
```

Bu komuta göre, tekdüze(uniform) dağılımdan 50 ile 150 aralığında 150 tane veri simule edilir.

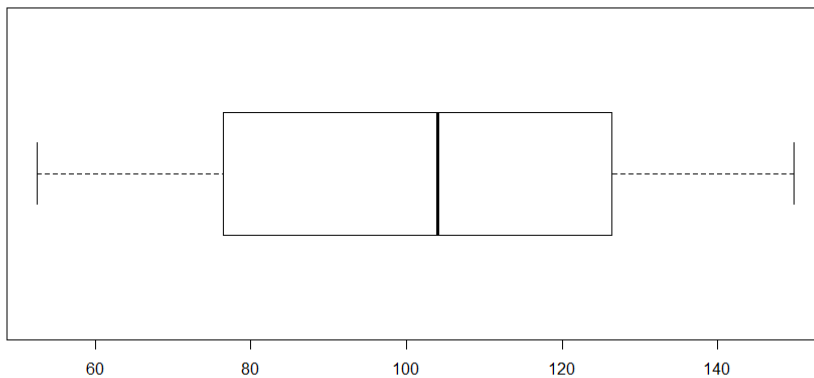
2.Asama: Histogram

```
hist(x2, freq = FALSE, breaks = 20)  
points(density(x2), type = "l")  
rug(x2)
```



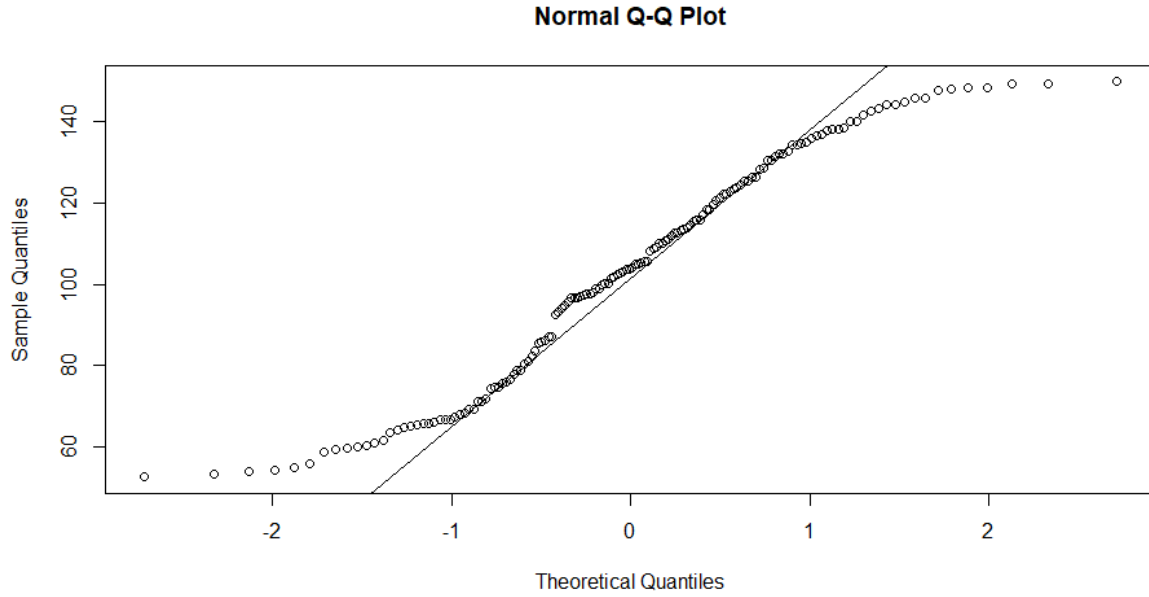
3.Asama: Kutu Grafiği

```
# boxplot  
boxplot(x2, horizontal=TRUE)
```



4.Asama: Q-Q Grafiği

```
qqnorm(x2) # veri setini normal skorlarına karşılık çizdirir.  
qqline(x2) # referans alacağımız çizgimizi çizdirir
```



5.Asama: Normallik Varsayımı Testleri

```
shapiro.test(x2)  
ad.test(x2)  
lillie.test(x2)  
cvm.test(x2)
```

- Bu testleri hangi paketler aracılığıyla kullanıyoruz? Tekrar hatırlatınız. Elde ettiğiniz sonuçları ve skorları kullanarak hipotezlerinizin hakkındaki kararınızı ve yorumunuzu yazınız.

Karar:

Yorum:

Temel Kavramlar ile İlgili Hatırlatma

Araştırmacılar genelde örneklemden elde edilen bilgiler yardımıyla anakütle parametreleri hakkında bir karara varmaya çalışırlar. Bu kararlar verilirken ya bir tahmin yapılır ya da konu ile ilgili belirli bir varsayımda bulunulur. Gerçekleşsin veya gerçekleşmesin ileri sürülen bu tip varsayımlara HİPOTEZ denir.

Örneklem istatistiklerinden yararlanarak bir hipotezin geçerli olup olmadığını ortaya koyma işlemine ise istatistiksel hipotez testi veya hipotez testi denir.

İstatistiksel bir araştırmada iki tür hipotez kurulur. “Eşit, fark yoktur, önemli değildir, en az (fazla) ... kadardır” biçiminde kurulan hipoteze “yokluk (null), boş ya da sıfır hipotezi” denir ve H_0 ile gösterilir. H_0 hipotezine karşı test edilen hipoteze ise “alternatif (alternative), seçenek ya da karşıt hipotez “ denir ve H_1 ile gösterilir. H_0 ve H_1 hipotezleri genelde aşağıdaki gibi kurulur. H_0 : Örneklemden elde edilen değer ile anakütlenin bilinen değeri arasında bir fark yoktur. H_1 : Örneklemden elde edilen değer ile anakütlenin bilinen değeri arasında bir önemli (anamlı) bir fark vardır.

Genellikle araştırmacılar sıfır hipotezinin reddedilmesini ve karşıt hipotezin kabul edilmesini isterler. Hipotez testlerinin aşamalarıyla şu şekildedir.

1. Hipotezlerin oluşturulması
2. Anlam düzeyinin (α) belirlenmesi
3. Örneklem dağılımının belirlenmesi
4. Ret bölgesinin belirlenmesi
5. Test istatistiğinin hesaplanması ve yorum

Bir Anakütle Ortalamasının Hipotez Testi

Bu tür hipotezlerin testinin amacı, karşıt hipotezde ileri sürülen iddianın kabul edilip edilmeyeceğinin ortaya çıkarılmasıdır. Ancak karşıt hipotezi direk test etmek mümkün olmadığından, sıfır hipotezi test edilir ve elde edilen sonuç karşıt hipotez için genellenir.

Tek grup anakütlenin parametreleriyle ilgili parametrik hipotez testlerin varsayımları şunlardır:

1. Örneklem alınıldığı anakütle normal dağılıma sahiptir.
2. Örneklemdeki birimler eşit olasılıkla ve iadeli olarak seçilmiş veya anakütle sonsuz büyüklüktedir. Bu testlerde ileri sürülebilecek karşıt hipotezlere şu şekilde örnek verebiliriz.

Bir firmanın tereyağı paketlerinin ağırlığının 250 gr olması gerektiği halde, firma buna uymamaktadır. Günlük ortalama üretimi 1000 kg olan bir ilaç fabrikasında uygulanan yeni

teknik üretimi artırmıştır. Turistik amaçla yurt dışına giden vatandaşlarımızın ortalama konaklama süresi 20 günden azdır.

Bu örneklere göre sıfır ve karşıt hipotezlerimiz sırasıyla aşağıdaki gibi olacaktır.

$$H_0: \mu = \mu_0 = 250 \text{ gr} \quad H_0: \mu = \mu_0 = 1000 \text{ kg} \quad H_0: \mu = \mu_0 = 20 \text{ gün}$$

$$H_1: \mu \neq \mu_0 = 250 \text{ gr} \quad H_1: \mu > \mu_0 = 1000 \text{ kg} \quad H_1: \mu < \mu_0 = 20 \text{ gün}$$

Peki bu oluşturulan hipotezleri hangi istatistiksel test kullanarak sınavacağımıza karar verirken bir örneklemden oluşan veri için;

- Örnekleme dağılımı normal, varyans biliniyor ve örnekleme sayısı 30'dan büyükse z testi
- Örnekleme dağılımı normal, varyans bilinmiyorsa z testi
- Örnekleme dağılımı normal, varyans biliniyor ve örnekleme sayısı 30'dan büyükse z testi
- Örnekleme dağılımı normal, varyans biliniyor ve örnekleme sayısı 30 dan küçükse t testi

Böylece bir genelleme yapmak gerekirse;

- Anakütle varyansinin bilindiği durumlarda z testi
- Bilinmediği durumlarda ise; örnek hacmi 30'dan küçükse t testini kullanmalıyız.

SORU 1: İlkokul mezunu öğrencilerin Türkçeyi kullanma başarı puanları $X \sim N(65, 144)$ normal dağılıma sahiptir. Rastgele seçilen ilkököl mezunu 45 öğrencinin Türkçeyi kullanma başarı puanları TDKO ölçeği aracılığıyla aşağıdaki gibi belirlenmiştir.

BAŞARI PUANLARI

23	45	65	47	65	48	56	78	90	87
62	68	79	30	48	65	67	65	45	67
66	88	73	52	44	50	35	76	62	59
83	23	43	54	44	66	74	81	88	32
43	61	74	70	75					

Buna göre öğrencilerin başarı puanlarının ortalama değeri 65'e eşit olup olmadığını test ediniz.

ÇÖZÜM:

Veri Girişi:

Veriler için 3 farklı şekilde yükleme işlemi yapılabilir. Bunlardan ikisini uygulayalım.

İlk olarak kodlama yöntemiyle yapalım.

```
veri2<-read.table("C:/Users/umut/Desktop/31 mart 1.soru.txt", header=FALSE)  
colnames(veri2)
```

Buraya yükleyeceğimiz verinin bilgisayarımız içerisindeki konumu yazılır ve sizler kendi bilgisayarınızdaki dosya konumunu yazacaksınız.

İkinci olaraksa elimizle vektör girişi yaparak işlemi uygulayalım.

```
veri <- c(23,45, 65, 47, 65, 48, 56, 78, 90, 87, 62, 68, 79, 30, 48, 65, 67, 65, 45, 67,  
        66, 88, 73, 52, 44, 50, 35, 76, 62, 59, 83, 23, 43, 54, 44, 66, 74, 81,  
        88, 32, 43, 61, 74, 70, 75)
```

Öncelikle, rastgele seçilen verilerin normal dağılıma sahip olup olmadığı hipotezi test edilir.

H_0 : Veriler normal dağılmıştır.

H_1 : Veriler normal dağılmamıştır.

Normallik Testleri

```
shapiro.test(veri2$V1)  
shapiro.test(veri)  
  
library("nortest")  
lillie.test(veri)  
ad.test(veri)  
lillie.test(veri)  
cvm.test(veri)
```

Verimizi iki farklı şekilde R programına yükledikten sonra testler için iki şekilde de uygulama yapabiliriz. Rstudio programında hazır olan Shapiro Wilk testini uygulayabiliriz. Ayrıca geçen haftaki konumuzda işlediğimiz gibi nortest paketini yükleyerek oradaki normallik testlerinden de faydalanabiliriz.

```
Shapiro-Wilk normality test  
  
data: veri  
W = 0.96918, p-value = 0.2703
```

```

> cvm.test(veri)

      Cramer-von Mises normality test

data:  veri
W = 0.066911, p-value = 0.3002

> lillie.test(veri)

      Lilliefors (Kolmogorov-Smirnov) normality test

data:  veri
D = 0.11532, p-value = 0.1399

> ad.test(veri)

      Anderson-Darling normality test

data:  veri
A = 0.39191, p-value = 0.3647

```

Tüm normallik testlerindeki p değeri önceden belirlemiş olduğumuz alfa değerinden büyük olması sebebiyle H_0 hipotezini reddedemeyiz ve verilerimiz normal dağılmaktadır diyebiliriz.

Elimizde rastgele seçilmiş, normal dağılıma sahip, tek grup veriler bulunmaktadır. Ayrıca bu verilerin varyansı bilinmektedir. Bu nedenler sebebiyle tek grup ortalama testlerinden Z testi seçilir.

Z testi

R Studio programında z testini uygulayan bir komut hali hazırda bulunmamaktadır. Bu sebeple bu testin bulunduğu *TeachingDemos* paketi kurulur ve yüklenir.

```

install.packages("TeachingDemos")
library(TeachingDemos)

z.test(veri, alternative="two.sided", mu=65, stdev=12)
#alternative="greater"
#alternative="less"

```

Kodlar girilip z testi uygulandığında alttaki çıktı elde edilmiştir.

One Sample z-test

```
data: veri
z = -2.5963, n = 45.0000, Std. Dev. = 12.0000, Std. Dev. of the sample mean = 1.7889,
p-value = 0.009423
alternative hypothesis: true mean is not equal to 65
95 percent confidence interval:
 56.84947 63.86165
sample estimates:
mean of veri
 60.35556
```

Buna göre, p değeri alfa değerinden düşüktür bu da demektir ki yokluk hipotezi reddedilebilir.

Karar: Yüzde 95 güven düzeyinde verilerin ortalaması 65'ten farklıdır.

SORU 2: Bir okuldaki 5. sınıf öğrencilerinin bir sınavdan aldığı skorların ortalaması 65'tir. Rastgele seçilen 10 öğrencinin sınav sonuçları aşağıdaki gibi belirlenmiştir.

SINAV SONUÇLARI

60	55	75	80	90	40	70	100	60	70
----	----	----	----	----	----	----	-----	----	----

Bu seçilen örnekler tüm okulun bir rastgele örneği midir?

ÇÖZÜM:

Buradaki araştırma sorusu için hipotezlerimizi yazarsak;

$$H_0: \mu = \mu_0 = 65$$

$$H_1: \mu \neq \mu_0 = 65$$

Veri Girişi:

Buradaki veri sayısı az olduğu için verileri elle girebiliriz.

```
# bir örneklem ortalamasıyla ilgili örnekler
# öncelikle verimiz için vektörümüzü oluşturualım
veri <- c(60, 55, 75, 80, 90, 40, 70, 100, 60, 70)
```

Tek grup verilerin ortalama testini yapmadan önce verilerin normallik şartını sağlayıp sağlamadığına bakılır.

Normallik Testi

H_0 : Veriler normal dağılmıştır.

H_1 : Veriler normal dağılmamıştır.

Rstudio programında hazır olan Shapiro Wilk testini uygulayabiliriz.

```
shapiro.test(veri)
```

Buna göre elde edilen sonuç çıktısı aşağıdaki gibidir.

```
Shapiro-Wilk normality test
data:  veri
W = 0.98576, p-value = 0.9885
```

p değeri önceden belirlemiş olduğumuz alfa değerinden büyük olması sebebiyle H_0 hipotezini reddedemeyiz ve veriler normal dağılmaktadır kararı verilir.

Elimizdeki veriler sadece bir örneklemden (örnek, grup, sample) gelmektedir ve bu örneklemin ortalaması ile çekildiği kitledeki ortalamanın kıyası yapılmak istenmektedir. Buradaki uygulamada örneklem dağılımı normal dağılım, örneklem hacmi 30'dan az ve kitlenin varyansı bilinmemektedir. Bu sebepler bizlere bu uygulama için 1 Örneklem T testi (one sample t test) uygulanmasını işaret etmektedir.

Bir örneklem için t testi

RStudio programında z testinin aksine t testini uygulayan bir komut hali hazırda bulunmaktadır ve aşağıdaki şekilde yazılır.

```
t.test(veri,mu=65)
```

Bu komutun çıktısıysa şu şekildedir;

```
One Sample t-test

data:  veri
t = 0.90453, df = 9, p-value = 0.3893
alternative hypothesis: true mean is not equal to 65
95 percent confidence interval:
 57.49546 82.50454
sample estimates:
mean of x
      70
```

Buna göre, p değeri alfa değerinden büyüktür yani H_0 (yokluk) hipotezi reddedilemez.

Karar: Yüzde 95 güven düzeyinde öğrencilerin sınav notu ortalaması 65'ten farklı değildir.

SORU 3: Atakum Denizevleri Mahallesiindeki ev fiyat ortalaması Emlakçılar Birliği tarafından 150 birim olarak belirtilmiştir. Yapılan bir araştırmada 29 tane örnek seçilip onların fiyatları aşağıdaki gibi belirtilmiştir. Bu seçilen örneklemin ortalaması 150 birime eşit midir test ediniz.

EV FİYATLARI

60	50	56	72	80	80	80	99	101	110
110	110	120	140	144	145	150	180	201	210
220	240	290	309	320	325	400	500	507	

ÇÖZÜM:

Buradaki araştırma sorusu için hipotezlerimizi yazarsak;

$$H_0: \mu = \mu_0 = 150$$

$$H_1: \mu \neq \mu_0 = 150$$

Veri Girişi:

```
veri3<-read.table("C:/Users/umut/Desktop/31_mart_3.soru.txt", header=TRUE)
colnames(veri3)
```

Verileri girdik veri3 olarak tanımladığımız veri setini environment kısmından görebiliriz. Burada ayrıca verinin bulunduğu sütun ismini de istedik ve su çıktıyı aldık.

```
[1] "EvFiyatlari"
```

Tek grup verilerin ortalama testini yapmadan önce verilerin normallik şartını sağlayıp sağlamadığına bakılır.

Normallik Testi

Öncelikle hipotezleri kuralım.

H_0 : Veriler normal dağılmıştır.

H_1 : Veriler normal dağılmamıştır.

```
shapiro.test(veri3$EvFiyatlari)
```

Buna göre elde edilen sonuç çıktıysa

aşağıdaki gibidir.

```
Shapiro-wilk normality test
data: veri3$EvFiyatlari
W = 0.85934, p-value = 0.00119
```

p değerinin önceden belirlemiş olduğumuz alfa değerinden küçük olması sebebiyle H_0 hipotezi

reddedilir ve veriler normal dağılmamakta olduğuna karar verilir.

Böylece, 1 örneklem için parametrik olmayan test olan Wilcoxon testi aşağıdaki hipotezler için uygulanır.

H_0 : Ev fiyatlarının ortancası 150 birimdir.

H_1 : Ev fiyatlarının ortancası 150 birim değildir.

Wilcoxon Testi

RStudio programında t testi gibi t testini uygulayan bir komut hali hazırda bulunmaktadır ve aşağıdaki şekilde yazılır.

```
wilcox.test(veri3$EvFiyatlari,mu = 150)
```

Bu komutun çıktısıysa şu şekildedir;

```
wilcoxon signed rank test with continuity correction  
data: veri3$EvFiyatlari  
V = 235, p-value = 0.4729  
alternative hypothesis: true location is not equal to 150
```

Buna göre, p değeri alfa değerinden büyüktür yani H_0 (yokluk) hipotezi reddedilemez.

Karar: Yüzde 95 güven düzeyinde ev fiyatlarının ortancası 150 birimden farklı değildir.

Bağımsız İki Örneklem İçin Ortalama Testleri

İki tane bağımsız örneklem(grup) için normal dağılıma bağımlı bir değişkenin ortalamalarının karşılaştırılmak istendiğinde kullanılır. Verilerin rastgele seçilmesi, normallik varsayımı ve varyansların homojenliği varsayımları altında parametrik olanları uygulanır.

Varsayımları test etmek için normallik testi ve varyans homojenliği testi uygulanır. Bu varsayımlar sağlanmadığında ise parametrik olmayan testlerden Mann Whitney U testi kullanılır.

İki ortalama arasındaki farkın testi yapılırken, kullanılacak test istatistikleri anakütle varyansının bilinmesi ve örnek büyüklüğü dikkate alınarak aşağıdaki şekilde bir sınıflama yapılabilir. Gözlemler Normal dağılışı gösteriyorsa ve

- Popülasyon (anakütle) varyansları (σ_1^2 , σ_2^2) biliniyor ya da popülasyon varyansları bilinmiyorken ancak örneklem sayısı büyükse ($n \geq 30$) → Z testi
- Popülasyon varyansları bilinmiyor fakat eşit kabul edilebiliyorsa ($\sigma_1^2 = \sigma_2^2$) → t testi
- Popülasyon varyansları bilinmiyor fakat eşit kabul edilemiyorsa ($\sigma_1^2 \neq \sigma_2^2$) → t testi

kullanılır.

Buraya kadar yazılan tüm durumlar iki örneklemdeki verilerin de normal dağılıma sahip olmalarını kapsamaktadır.

Eğer örneklemdeki veriler normal dağılışı göstermiyorsa parametrik olmayan istatistiksel yöntemlerden olan Mann Whitney U / Mann Witney Wilcoxon / Wilcoxon Sıralama Toplamı testi uygulanır. Bu test, R/RStudio üzerinde wilcoxon olarak geçmektedir.

Örnek1:

Bir psikolog uykunun hatırlama üzerine etkisini araştırıyor. Rastgele seçilmiş günde 8 saat uyuyan 6 öğrenci ile yine rastgele seçilmiş günde 5 saat uyuyan diğer 6 öğrenciye hatırlama testi uyguluyor. Öğrencilerin aldığı puanlar aşağıdaki şekilde ölçülüyor.

8 saat uyuyanlar: 40 45 52 61 65 75

5 saat uyuyanlar: 30 35 48 52 54 60

Bu iki grup arasında önemli bir fark olup olmadığını test ediniz.

Cözüm:

Veri Girişi:

```
# verileri vektor halinde girisini yapalim
x <- c(40,45,52,61,65,75)
y <- c(30,35,48,52,54,60)
```


Araştırmadaki veriler rastgele seçilmiştir ve iki bağımsız örneklemin karşılaştırılması hakkında uygulama yapılmaktadır. Bu aşamada yapılacak teste geçmeden hipotezleri oluşturalım.

H_0 :İki örneklem(grup) ortalaması arasında fark yoktur

H_1 :İki örneklem(grup) ortalaması arasında fark vardır.

→ $H_0: \mu_1 - \mu_2 = 0$

→ $H_1: \mu_1 - \mu_2 \neq 0$

Bu uygulanacak test için öncelikle parametrik mi parametrik olmayan bir yöntem mi seçileceğine karar verilir. Bu karar da verilerin normal dağılıp dağılmamasına bakılarak ortaya çıkartılır.

Normallik Varsayımı Kontrolü:

Hipotezler:

H_0 : Veriler normal dağılmıştır.

H_1 : Veriler normal dağılmamıştır.

RStudio Programında Normallik Testi:

```
# normallik testi
#her iki grup için de yaparız

shapiro.test(x)
shapiro.test(y)
```

Diğer normallik testleri için olan fonksiyonlar hangi pakette bulunmaktadır?

```
> shapiro.test(x)

Shapiro-Wilk normality test

data:  x
W = 0.97277, p-value = 0.9105

> shapiro.test(y)

Shapiro-Wilk normality test

data:  y
W = 0.92344, p-value = 0.5305
```

İki örneklem için de p değeri 0.05 değerinden büyüktür ($p > \alpha$). Bu nedenle normallik testi için H_0 hipotezi reddedilemez. Yani, yüzde 95 güven düzeyinde veriler normal dağılmaktadır.

Varyans Homojenliği Varsayımı Kontrolü:

Varyans homojenliği varsayımı, bağımsız karşılaştırma t testi ve ANOVA'nın, tüm karşılaştırma gruplarının aynı varyansa sahip olduğunu belirten bir varsayımdır. Varyans homojenliğini test etmek için kullanılabilecek birkaç istatistiksel test vardır. Bu testler şunları içermektedir:

- Hartley'nin Fmax testi.
- F testi.
- Cochran'ın testi.
- Levene'nin testi.
- Bartlett'in testi.

Bu testlerin birçoğunun normalliklere karşı çok hassas olduğu bulunmuştur ve sıklıkla kullanılmamaktadır. Bu testlerden varyans homojenliği için en yaygın değerlendirmenin yapıldığı test Levene'nin testidir.

Hipotezler:

H_0 : Varyanslar homojendir.

$$H_0: \sigma_1^2 = \sigma_2^2$$

H_1 : Varyanslar homojen değildir.

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

RStudio Programında Varyans Homojenliği Testleri:

Yukarıda da ifade edildiği gibi varyans homojenliğini kontrol etmenin farklı metotları bulunmaktadır. Dersimizde RStudio içinde hazır komuta sahip olan F testi ile Levene's Test aracılığı ile yapılan testleri hesaplayalım.

- F testi

```
var.test(x, y, ratio = 1,
         alternative = "two.sided",
         conf.level = 0.95)
```

Argümanlar sırası ile bağımlı değişken, gruplar, oran miktarı, hipotez testinin yönü ve güven düzeyi.

F test to compare two variances

```
data: x and y
F = 1.2721, num df = 5, denom df = 5, p-value = 0.7981
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1780123 9.0912274
sample estimates:
ratio of variances
 1.272144
```

- Levene's Test

```
install.packages("car")
library(car)
library(carData)
#####
##elle veri girerek veri seti olusturma
notlar <- c(x,y)#### oncelikle sinav skorlarini vektorde toplayalim
uyku <- c(rep("5saat",6),rep("8saat",6)) ##### bagimsiz gruplari olusturalim
veri <- data.frame(notlar,uyku) #### veri cercevesi yaratalim
#####levene testi #####
leveneTest(veri$notlar,veri$uyku,center = mean)
```

Levene testi "car" paketi içinde bulunmaktadır. Bu sebeple öncelikle bu paketi kurmalıyız ve yüklemeliyiz. "car" paketi "carData" paketini de ihtiyaç duyabilir bu sebeple bu paketi de yükleyebiliriz.

Ayrıca fonksiyonun argümanları gereği önceki satırlarda x ve y olarak oluşturduğumuz değişkenleri bir veri seti haline getirilmelidir.

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  0.1619 0.6959
      10
```

RStudio çıktısına göre Levene testine göre p değeri yüzde 5'ten büyüktür. Bu nedenle varyanslar homojendir diyebiliriz. Veriler normal dağılıp, gruplar arası varyanslar eşit yani homojen olduğuna göre parametrik bir yöntem olan 2 bağımsız örneklem t testi uygulanır.

Bağımsız 2 örneklem t testi

RStudio üzerinde t testi uygulaması aşağıdaki şekilde yapılır.

```
#####t testi#####
t.test(veri$notlar~veri$uyku, var.equal = TRUE, conf.level = 0.95)
```

Uygulamadan elde edilen sonuçlar şu şekildedir.

Two Sample t-test

```
data: veri$notlar by veri$uyku
t = 1.3748, df = 10, p-value = 0.1992
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.103874 25.770541
sample estimates:
mean in group 5saat mean in group 8saat
      56.33333      46.50000
```

Buna göre p değeri 0.05 değerinden büyüktür ($p > \alpha$). Bu yüzden H_0 hipotezi reddedilemez yani yüzde 95 güven düzeyinde iki grup (örnek, örneklem) ortalaması arasında fark yoktur.

Örnek2: Ayaklarında kırık olan sporcuların tedavi sürecinde çektikleri acı seviyesini ölçen medikal alete göre ağrı kesici skorları aşağıdaki gibi olsun. Aşağıdaki veri setine göre ilaç grubu ile ilaç almayan grubunun ortalama skorları arasında anlamlı bir farklılık var olup olmadığını test ediniz.

İlaç Almayan	7500	8000	4000	1550	1250	1000	2250	6800	3400	6300	9100	9700	1040	1670	1400
İlaç Alan	400	250	800	1400	8000	7400	1020	6000	920	1420	2700	4200	5200	4100	

Cözüm:

Veri Girişi:

```
#####veri girişi #####
```

```
library(readxl)
```

```
sporcu <- read_excel("C:/Users/umut/Desktop/sporcu.xlsx")
```

```
sporcu
```

```
#####
```

```
#####veri girişi2 #####
```

```
b <- c(7500,8000,4000,1550,1250,1000,2250,6800,3400,6300,9100,9700,1040,1670,1400)
```

```
c <- c(400,250,800,1400,8000,7400,1020,6000,920,1420,2700,4200,5200,4100)
```

Kodlama ile veri girişi

Vektör ile veri girişi

Araştırmadaki veriler rastgele seçilmiştir ve iki bağımsız örneklemin karşılaştırılması hakkında uygulama yapılmaktadır. Bu aşamada yapılacak teste geçmeden hipotezleri oluşturalım.

H_0 :İki örneklem(grup) ortalaması arasında fark yoktur

H_1 :İki örneklem(grup) ortalaması arasında fark vardır.

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

Bu uygulanacak test için öncelikle parametrik mi parametrik olmayan bir yöntem mi seçileceğine karar verilir. Bu karar da verilerin normal dağılıp dağılmamasına bakılarak ortaya çıkartılır.

Normallik Varsayımı Kontrolü:

Hipotezler:

H_0 : Veriler normal dağılmıştır.

H_1 : Veriler normal dağılmamıştır.

RStudio Programında Normallik Testi:

```
#####normallik testi#####
```

```
shapiro.test(subset(sporcu, Ilac == 1, select = puan)$puan)
```

```
shapiro.test(subset(sporcu, Ilac == 0, select = puan)$puan)
```

```
#####normallik 2#####
```

```
shapiro.test(b)
```

```
shapiro.test(c)
```

```
#####
```

Diğer normallik testlerini kapsayan paket hangi pakettir ve hangi testler bulunmaktadır?

```
> shapiro.test(subset(sporcu, Ilac == 1, select = puan)$puan)

Shapiro-Wilk normality test

data: subset(sporcu, Ilac == 1, select = puan)$puan
W = 0.85936, p-value = 0.02363

> shapiro.test(subset(sporcu, Ilac == 0, select = puan)$puan)

Shapiro-Wilk normality test

data: subset(sporcu, Ilac == 0, select = puan)$puan
W = 0.88179, p-value = 0.06164
```

İki örneklem için de p değeri 0.05 değerinden büyük değildir. Zira ilaç almayan sporcular için oluşturulmuş örneklem verileri normal dağılmamıştır. Bu nedenle normallik testi için H_0 hipotezi reddedilir. Yani, yüzde 95 güven düzeyinde veriler normal dağılmamaktadır.

İki bağımsız gruba ait veri setini analiz ederken eğer veriler normal dağılış göstermiyorsa, parametrik olmayan testlerden Mann-Whitney U / Wilcoxon testi kullanılır. Bu uygulamadaki veriler de normal dağılıma sahip olmadığı için parametrik olmayan bağımsız iki örneklem testi uygulanır. Burada parametrik testte karşılaştırılan ortalamalar yerine medyanlar kıyaslanarak anlamlı farklılık olup olmaması test edilir.

2 Örneklem için Wilcoxon Testi

RStudio üzerinde 2 örneklem Wilcoxon testi uygulaması aşağıdaki şekilde yapılır.

```
wilcox.test(puan ~ Ilac, data=sporcu)
```

Uygulamadan elde edilen sonuçlar ise şu şekildedir.

```
> wilcox.test(puan ~ Ilac, data=sporcu)

Wilcoxon rank sum test with continuity correction

data: puan by Ilac
W = 75, p-value = 0.1978
alternative hypothesis: true location shift is not equal to 0
```

Buna göre p değeri 0.05 değerinden büyüktür ($p > \alpha$). Bu yüzden H_0 hipotezi reddedilemez yani yüzde 95 güven düzeyinde iki grup (örnek, örneklem) ortancaları (medyanları) arasında fark yoktur.

Örnek3: İki grubun ortalama boy uzunluklarının karşılaştırması istenir. Birinci grup (A) İtalyan uyruklu bireylerden oluşur ; ikinci grupsa(B) Alman uyruklu bireylerden alınmıştır. Veriler aşağıda verilmiştir:

A: 175, 168, 168, 190, 156, 181, 182, 175, 174, 179

B: 150, 180, 145, 168, 130, 190, 140, 185, 142, 188

Bağımlı (Eşli Gözlemler) Örnekler için Ortalama Testleri

Aynı birey/cisim/alet.. üzerinde farklı zamanlarda ölçümler alındığında ve bunların karşılaştırılması söz konusu olduğu durumlarda bağımlı (eşli) grup ortaya çıkar. Eşleştirilmiş fertlerle yapılan testlerde kullanılan test istatistiği daha önceki grup karşılaştırmalarında kullanılanlardan daha farklıdır. Çünkü grup karşılaştırmalarında X1 ile X2 değişkenlerinin birbirinden bağımsız olduğu varsayılmaktaydı. Eşleştirilmiş gözlemlerde ise X1 ve X2 ölçümleri aynı birey üzerinde veya çok benzer bireyler üzerinden yapıldığı için bağımlı olacaktır. Yani $n1 = n2 = n$ (gözlem çifti sayısı) olacaktır.

Aynı fert üzerinde farklı zamanlarda ölçümler alındığında ve bunların karşılaştırılması söz konusu olduğu durumlarda bağımlı (eşli) grup ortaya çıkar.

Özetlemek gerekirse bağımlı iki örneklem testleri, rastgele seçilen n birimlik gruptan iki farklı zamanda ve iki farklı uygulamaya ilişkin elde edilen iki veri setinin farklarının sıfır ortalamalı toplumun rastgele örnekleri olup olmadığını test etmek için uygulanan bir yöntemdir.

Bu şekilde düzenlenmiş uygulamalarda dersimizde 2 yöntem uygulanacaktır.

1. Bağımlı 2 Örnek t Testi
2. Wilcoxon İşaretli Sıra Testi

1.Bağımlı 2 Örnek t Testi (Paired t testi)

Aynı fert üzerinde farklı zamanlarda ölçümler alındığında ve bunların karşılaştırılması söz konusu olduğu durumlarda eğer verilerin normal dağılıma sahipse bu test uygulanmaktadır.

Örnek: Üniversite öğrencilerinin vize ve final notları arasındaki başarı durumunu ölçmek isteyen bir öğretim elemanı 15 kişilik öğrenci grubunun vize ve final notlarında anlamlı bir farklılık olup olmadığını nasıl test edebilir?

Vize	45	67	60	55	48	62	48	63	72	50	77	81	56	45	68
Final	75	73	85	72	56	73	76	80	95	82	92	90	70	60	87

Çözüm

Bu uygulamada öğretim üyesi aynı bireylerin vize ve final sonuçlarını kıyaslamak istemektedir. Yani bağımlı grupların kıyası söz konusudur. Bu sebeple bağımlı 2 örnek için ortalama karşılaştırma testi uygulanacaktır. Bu testlerden hangisini kullanacağımıza ise verilerin normallik varsayımını sağlayıp sağlamamasına göre karar verilecektir.

Veri Girişi:

```
library(readxl)
notlar <- read_excel("C:/Users/umut/Desktop/notlar.xls")
```

Araştırmadaki veriler rastgele seçilmiştir ve iki bağımsız örneklemin karşılaştırılması hakkında uygulama yapılmaktadır. Bu aşamada yapılacak teste geçmeden hipotezleri oluşturalım.

H_0 : Vize ve final ortalamaları arasında fark yoktur
 H_1 : Vize ve final ortalamaları arasında fark vardır.

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Uygulanacak test için öncelikle parametrik mi parametrik olmayan bir yöntem mi seçileceğine karar verilir. Bu karar da verilerin normal dağılıp dağılmamasına bakılarak ortaya çıkartılır.

Normallik Varsayımı Kontrolü:

Hipotezler:

H_0 : Vize sınavı verileri normal dağılmıştır.
 H_1 : Vize sınavı verileri normal dağılmamıştır.

H_0 : Final sınavı verileri normal dağılmıştır.
 H_1 : Final sınavı verileri normal dağılmamıştır.

RStudio Programında Normallik Testi:

```
##### normallik testleri#####
```

```
shapiro.test(notlar$vize)  
shapiro.test(notlar$final)
```

```
> shapiro.test(notlar$vize)
```

Shapiro-wilk normality test

```
data: notlar$vize  
W = 0.94495, p-value = 0.4486
```

```
> shapiro.test(notlar$final)
```

Shapiro-wilk normality test

```
data: notlar$final  
W = 0.9662, p-value = 0.7983
```

Vize ve final sınavı için oluşturulan verilerde p değeri 0.05 değerinden büyüktür ($p > \alpha$). Bu nedenle normallik testi için H_0 hipotezi reddedilemez. Yani, yüzde 95 güven düzeyinde veriler normal dağılmaktadır.

Aynı öğrencilerin farklı zamanlardaki sınav skorları ölçülmüş ve bunların karşılaştırılması söz konusudur ve verilerin normal dağılıma sahiptir bu neden bağımlı gruplar için t testi uygulanır.

Bağımlı 2 örneklem t testi

RStudio üzerinde uygulama aşağıdaki şekilde yapılır.

```
#####bağımlı gruplar için t testi#####  
t.test(notlar$vize,notlar$final, paired = TRUE, alternative = "two.sided" )
```



```
> t.test(notlar$vize,notlar$final, paired = TRUE, alternative = "two.sided" )
```

Paired t-test

```
data: notlar$vize and notlar$final  
t = -8.5451, df = 14, p-value = 6.306e-07  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-22.43455 -13.43212  
sample estimates:  
mean of the differences  
-17.93333
```

Buna göre p değeri 0.05 değerinden küçüktür ($p < \alpha$). Bu yüzden H_0 hipotezi reddedilir yani yüzde 95 güven düzeyinde vize ve final skorları ortalaması arasında istatistiksel olarak anlamlı bir farklılık vardır.

Eğer veriler normal dağılıma sahip olmazsa parametrik olmayan yöntemlerden Wilcoxon işaretli sıra testi uygulanır.

2. Wilcoxon İşaretli Sıra Testi (Wilcoxon signed ranks test)

Veriler normal dağılmadığında bağımlı iki örnek arasındaki farkın önemliliğini test eder. Eşleştirilmiş t testini parametrik olmayan alternatifidir. Ortalama olarak medyan kullanılır. n birimlik örnekten elde edilen iki gözlem seti farkının medyanı sıfır olan toplumdaki çekilmiş rasgele bir örnek olup olmadığını test eder.

Örnek: Yeni bir rejim programına başlayan bireyler üzerinde bu programın etkisi araştırılmak istenmektedir. Rasgele seçilen 8 bireyin bir rejimi uygulamadan önceki ve sonraki puanları aşağıdaki gibidir. Bu sonuçlara göre verileri parametrik olmayan yöntem ile test ederek rejimin denekler üzerinde anlamlı bir etkisi olup olmadığını test ediniz.

Öncesi	53	51	47	67	128	74	108	48
Sonrası	48	53	37	74	78	67	78	57

Çözüm:

Veri Girişi:

```
library(readxl)  
on_son <- read_excel("C:/Users/umut/Desktop/on_son.xlsx")
```

Araştırmadaki veriler rastgele seçilmiştir ve bireylerin rejimin öncesinde ve sonrasındaki puanları kıyas edilerek rejimin etkisi araştırılmaktadır . Bu aşamada yapılacak teste geçmeden hipotezleri oluşturalım.

H_0 : Rejim öncesi ve sonrası puan ortalamaları arasında fark yoktur.

H_1 : Rejim öncesi ve sonrası puan ortalamaları arasında fark vardır.

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Uygulanacak test için öncelikle parametrik mi parametrik olmayan bir yöntem mi seçileceğine karar verilir. Bu karar da verilerin normal dağılıp dağılmamasına bakılarak ortaya çıkartılır.

Normallik Varsayımı Kontrolü:

Hipotezler:

H₀: Rejim öncesi verileri normal dağılmıştır.

H₁: Rejim öncesi verileri normal dağılmamıştır.

H₀: Rejim sonrası verileri normal dağılmıştır.

H₁: Rejim sonrası verileri normal dağılmamıştır.

RStudio Programında Normallik Testi:

```
##### normallik testleri#####  
shapiro.test(on_son$Ön)  
shapiro.test(on_son$Son)
```

```
> ##### normallik testleri#####  
> shapiro.test(on_son$Ön)
```

Shapiro-Wilk normality test

```
data: on_son$Ön  
W = 0.82005, p-value = 0.04671
```

```
> shapiro.test(on_son$Son)
```

Shapiro-Wilk normality test

```
data: on_son$Son  
W = 0.9248, p-value = 0.47
```

Rejim öncesi ve sonrası için kayıt altına alınmış verilerde p değeri rejim sonrası verisi için 0.05 değerinden büyüktür ($p < \alpha$). Fakat rejim öncesi verisi için 0.05 değerinden küçüktür ($p < \alpha$). Bu nedenle normallik testi için H₀ hipotezi reddedilir. Yani, yüzde 95 güven düzeyinde veriler normal dağılmamaktadır.

Veriler normal dağılmadığı için 2 bağımlı örnek için parametrik olmayan Wilcoxon signed rank test uygulanır.

Wilcoxon İşaretli Sıra Testi

```
##### bagimli 2 orneklem icin parametrik olmayan test#####  
wilcox.test(on_son$Ön, on_son$Son, paired = TRUE, alternative = "two.sided")
```

```
> wilcox.test(on_son$Ön, on_son$Son, paired = TRUE, alternative = "two.sided")
```

Wilcoxon signed rank test with continuity correction

```
data: on_son$Ön and on_son$Son  
V = 26.5, p-value = 0.262  
alternative hypothesis: true location shift is not equal to 0
```

Buna göre p değeri 0.05 değerinden büyüktür ($p > \alpha$). Bu yüzden H₀ hipotezi reddedilemez yani yüzde 95 güven düzeyinde vize ve final skorları medyanları (ortancaları) arasında istatistiksel olarak anlamlı bir farklılık yoktur.

Tek Yönlü Varyans Analizi

Varyans Analizi (ANOVA, Analysis Of Variance), üç ya da daha fazla gruba ait ortalamalar arasındaki farkın anlamlı olup olmadığı ile ilgili hipotezleri test etmek için kullanılmaktadır.

İki grubun ortalamaları arasındaki farkın anlamlı olup olmadığı t testi kullanılarak incelenebilirdi. Eğer ikiden fazla grubun ortalamaları karşılaştırılacak ise F Testi diğer bir ismiyle Varyans Analizi uygulanır. Burada grupların 2'li kombinasyonları alınarak gruplar t testleriyle karşılaştırılabilir. Fakat bu işlem hem daha uzun sürecektir hem de göze alınan hata payı miktarı test sayısı arttıkça çoğalacaktır. Bu nedenler sebebiyle birden fazla t testi uygulamak yerine Anova uygulanarak bu dezavantajlardan kaçınılır.

Varyans Analizi, gruplar arasında bir farklılık olup olmadığını gösterir ama hangi grupların birbirinden farklılık gösterdiğinin bilgisini sunmaz. Bu farklılıkları tespit edebilmek için çoklu karşılaştırma testlerinden faydalanılır.

Varyans analizi hangi varsayımlar altında uygulanmalıdır?

- Bağımlı değişken yani nicel olarak oluşturulan değişken için veriler tek yönlü Anovada karşılaştırılan her grup için normal dağılmalıdır. (Teknik olarak, normal olarak dağılması gereken artıklardır, ancak sonuçlar aynı olacaktır). Örneğin, üç grubun (amatör, yarı profesyonel ve profesyonel rugby oyuncuları) bacak güçlerini karşılaştırırsak, bacak gücü değerlerinin (bağımlı değişken) amatör grup, yarı profesyoneller ve profesyonel oyuncular yani 3 grup için de normal dağılmalıdır.
- Varyansların homojen olmalıdır. Bu, her gruptaki nüfus farklılıklarının eşit olduğu anlamına gelir.
- Gözlemlerin bağımsızlığı. Bu çoğunlukla bir çalışma tasarımı konusudur ve bu nedenle, gözlemlerinizin çalışma tasarımınıza (örneğin grup çalışması / aileler / vb.) Dayalı olarak bağımsız olmadığına inanıp inanmadığınızı belirlemeniz gerekecektir.

Verilerim bu varsayımları sağlamakta başarısız olursa ne olur?

- Tek yönlü ANOVA, normallik varsayımına karşı sağlam bir test olarak kabul edilir. Bu, normallik varsayımına yönelik ihlalleri oldukça iyi tolere ettiği anlamına gelir. Fakat yine de gruplardaki verilerin normalliğini sağlayamadığı durumlar olabilir. Burada yapılabilecek iki seçenek vardır:
(1) Verilerinizi çeşitli algoritmalar kullanarak dönüştürün, böylece dağılımlarınızın şekli normal dağılım olur.
(2) Normallik varsayımı gerektirmeyen yani parametrik olmayan Kruskal-Wallis H Testi'ni seçilebilir.
- Varyansların homojenliği varsayımı ihlal edildiğinde uygulanabilecek farklı yaklaşımlar vardır. Bu yaklaşımlardan Welch testini seçerek verilerin normal ama varyansların homojen olmadığı durum çözüme kavuşturulur.
- Olguların bağımsızlığının olmayışı başarısızlığın en ciddi varsayımı olarak ifade edilmiştir. Genellikle, bu soruna iyi bir çözüm sunan yapabileceğiniz çok az şey vardır. Zira bu varsayım verilerin toplanma aşamasında ortaya çıkarılır.

Çoklu Karşılaştırma Testleri

Varyans analizi sonucunda çıkan sonuç grupların birbirinden farklı olduğunu gösterir şekilde ise bu farklılıkları ortaya çıkan ve ortalamaları farklılık gösteren grupları tespit etmek amacıyla oluşan testlere çoklu karşılaştırma testleri denir.

Bu testler; homojen varyans yaklaşımını kullananlar ve heterojen varyans yaklaşımını kullananlar olarak iki temel gruba ayrılırlar.

Homojen varyans yaklaşımından en çok kullanılan testlere TUKEY ve DUNCAN, heterojen varyans yaklaşımından en çok kullanılan testlere ise Tamhane ve Games – Howell testlerini örnek verebiliriz.

Veriler normal dağılıma sahip değilken ise Dunn çoklu karşılaştırma testini kullanabiliriz.

Uygulama

Derste yapılacak örnekler üzerinden konu uygulaması yapılacaktır.