An Introduction to the Simple Biostatistics Program (SBP)

Stanley B. Pounds

7/11/2022

What is SBP

- SBP is the **S**imple **B**iostatistics **P**rogram.
- SBP was develped by Dr. Stanley B. Pounds, a faculty member of the Department of Biostatistics and the Graduate School of Biomedical Sciences at St. Jude Children's Research Hospital.
- SBP is an extension of the *R* statistical computing software that simplifies introductory biostatistics for students.
- SBP defines a few simple functions that perform all the computational tasks for an introductory biostatistics course.
- SBP minimizes the technicalities of computational tasks so students can focus on concepts and interpretation.

Setting Up SBP

```
# source the SBP.setup.file
SBP.setup.file="https://raw.githubusercontent.com/stan-pounds/Simple-Biostats-Program/main/setup-SBP.R"
source(SBP.setup.file)
```

- The above commands will need to be performed during *each* R session.
- You will not be able to use the SBP functions until after the above commands are executed in R.

Reading Data with SBP

```
data.set=read.data()
```

- The above command will open a window for the user to interactively navigate through folders to the data file.
- It can read data in the *csv*, tab-delimited *txt*, *xlsx*, and *Rdata* formats.
- For *xlsx* files, it will also prompt the user to choose the sheet to be read.
- It will then read the data and open a viewer to see it.
- The data from the file will be stored under the name `data.set` in R.

Example

- Demonstrate in R studio

Get an R Package

```
get.package("penalized")
```

- R packages are R add-ons that define useful functions to perform specific tasks.
- Some R packages include example data sets.
- The above code downloads the `penalized` R package and makes it available for use in the R session.

Get an R Package data set

```
get.package("penalized")  # make the penalized package available for use
data("nki70")             # make the nki70 data set available for use
help("nki70")             # open a help page about the nki70 data set
View(nki70)               # open the nki70 data set in a data viewer
```

Data Analysis Functions

Function	Actions
`describe("x",data.set)`	Compute descriptive stats & graphs for the `data.set` column named `x` using `data.set`
`estimate("x",data.set)`	Estimate the population value for the `x` column variable using `data.set`
`compare(y~grp,data.set)`	Compare the variable `y` across the `grp` groups using `data.set`
`correlate(y~x,data.set)`	Correlate the numeric variables `y` and `x` using `data.set`
`model(y~x+grp,data.set)`	Model `y` as a function of `x` and `grp` using `data.set`

Example with `nki70` data

```r
get.package("penalized")
data("nki70")
head(nki70[,1:10])
```

```
        time event  Diam   N       ER        Grade Age      TSPYL5
125 7.748118     0 <=2cm 1-3 Positive Intermediate  50 -0.18752814
127 4.662560     1 <=2cm 1-3 Positive    Well diff  42  0.15099047
128 8.739220     0  >2cm 1-3 Positive    Well diff  50  0.11695046
129 7.567420     0 <=2cm 1-3 Positive Intermediate  43  0.10493318
130 7.296372     0 <=2cm 1-3 Negative  Poorly diff  47  0.30821656
132 6.718686     0 <=2cm 1-3 Positive Intermediate  47 -0.09643536
    Contig63649_RC      DIAPH3
125    -0.15304662 -0.29514052
127    -0.21005843  0.03355057
128    -0.25813878  0.07791767
129    -0.13687348 -0.01984126
130     0.03544526  0.15589646
132    -0.03772432 -0.05882551
```

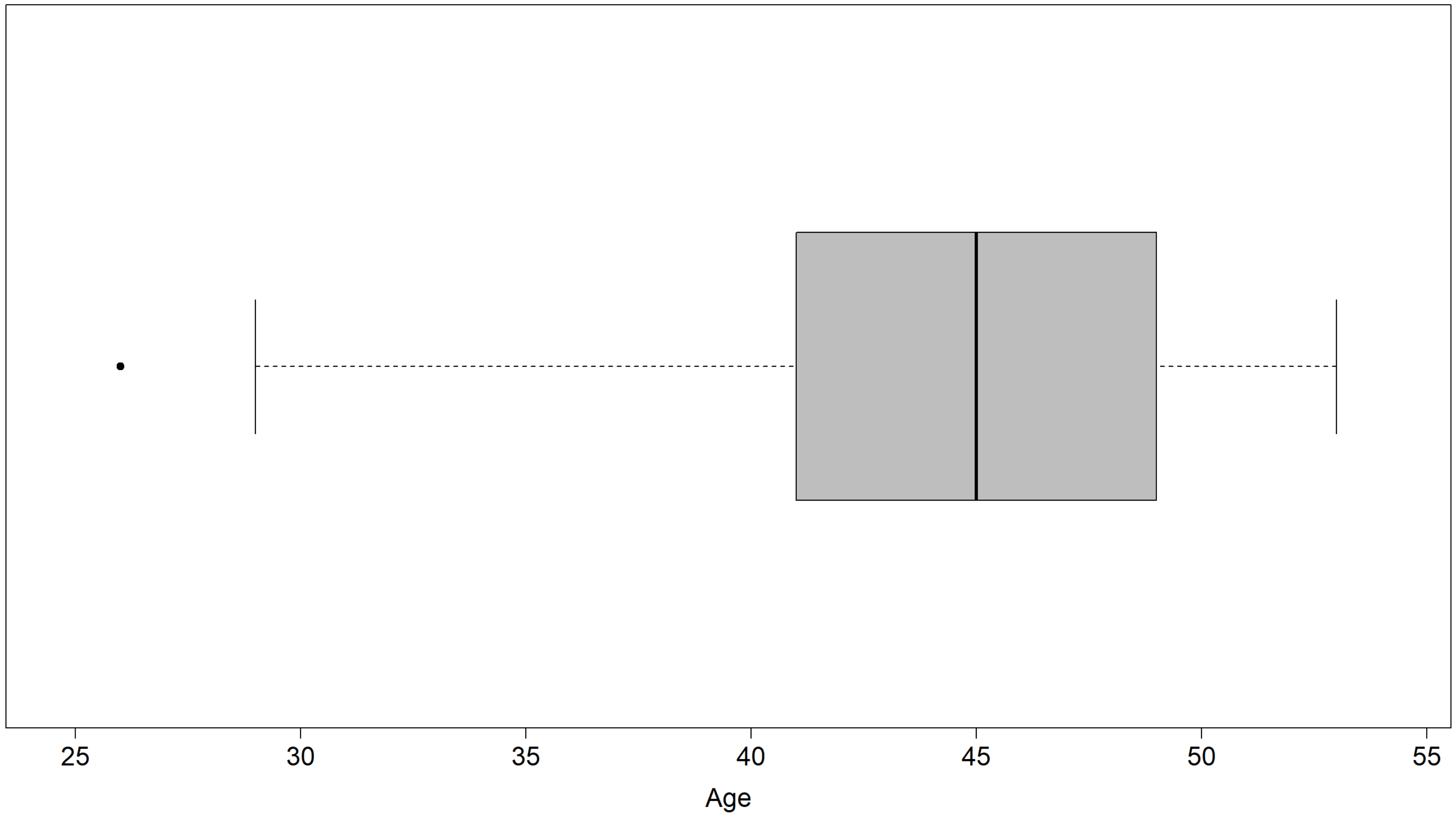
Column Names of a data set

```
colnames(nki70)
```

```
 [1] "time"           "event"          "Diam"           "N"
 [5] "ER"             "Grade"          "Age"            "TSPYL5"
 [9] "Contig63649_RC" "DIAPH3"         "NUSAP1"         "AA555029_RC"
[13] "ALDH4A1"        "QSCN6L1"        "FGF18"          "DIAPH3.1"
[17] "Contig32125_RC" "BBC3"           "DIAPH3.2"       "RP5.860F19.3"
[21] "C16orf61"       "SCUBE2"         "EXT1"           "FLT1"
[25] "GNAZ"           "OXCT1"          "MMP9"           "RUNDC1"
[29] "Contig35251_RC" "ECT2"           "GMPS"           "KNTC2"
[33] "WISP1"          "CDC42BPA"       "SERF1A"         "AYTL2"
[37] "GSTM3"          "GPR180"         "RAB6B"          "ZNF533"
[41] "RTN4RL1"        "UCHL5"          "PECI"           "MTDH"
[45] "Contig40831_RC" "TGFB3"          "MELK"           "COL4A2"
[49] "DTL"            "STK32B"         "DCK"            "FBXO31"
[53] "GPR126"         "SLC2A3"         "PECI.1"         "ORC6L"
[57] "RFC4"           "CDCA7"          "LOC643008"      "MS4A7"
[61] "MCM6"           "AP2B1"          "C9orf30"        "IGFBP5"
[65] "HRASLS"         "PITRM1"         "IGFBP5.1"       "NMU"
[69] "PALM2.AKAP2"    "LGP2"           "PRC1"           "Contig20217_RC"
[73] "CENPA"          "EGLN1"          "NM_004702"      "ESM1"
[77] "C20orf46"
```

Describe Age in nki70 data

```
age.result=describe("Age",nki70)
```



Describe Age in nki70 data

```
age.result
```

```
 **TABLES**

|               |        Age|
|:--------------|----------:|
|n.total        | 144.000000|
|n.missing      |   0.000000|
|n.available    | 144.000000|
|mean           |  44.305556|
|stdev          |   5.339230|
|median         |  45.000000|
|lower.quartile |  41.000000|
|upper.quartile |  49.000000|
|minimum        |  26.000000|
|maximum        |  53.000000|
|shapiro.pvalue |   0.000182|



 **RESULTS**

The variable Age has 144 observations (144 available; 0 missing)  with mean 44.3, standard deviation 5.3, median 45, lower quartile 41, upper
quartile 49, minimum 26, and maximum 53.
```

Describe Age in nki70 data

```
 **METHODS**

The Shapiro-Wilk (1965) test was used to evaluate the normality of the distribution of Age.


 **REFERENCES**

Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3-4): 591-611.
doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384.
```

Describe Age in `nki70` data

TABLES

	Age
n.total	144.000000
n.missing	0.000000
n.available	144.000000
mean	44.305556
stdev	5.339230
median	45.000000
lower.quartile	41.000000
upper.quartile	49.000000
minimum	26.000000
maximum	53.000000
shapiro.pvalue	0.000182

RESULTS

The variable Age has 144 observations (144 available; 0 missing) with mean 44.3, standard deviation 5.3, median 45, lower quartile 41, upper quartile 49, minimum 26, and maximum 53.

Describe Age in `nki70` data

METHODS

The Shapiro-Wilk (1965) test was used to evaluate the normality of the distribution of Age.

REFERENCES

Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3-4): 591-611. doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384.

Common Options for Data Analysis Functions

Option	Purpose
`txt=number`	produce no narrative (`txt=0`), basic narrative (`txt=1`), or detailed narrative (`txt=2`)
`fig=number`	produce no figures (`fig=0`), basic figures (`fig=1`), or more tables (`fig=2`,`fig=3`,etc)
`tbl=number`	produce no tables (`tbl=0`), basic tables (`tbl=1`), or more tables (`tbl=2`,`tbl=3`,etc)
`clr="color.name"`	use the color `color.name` in the figures
`clr=c("name1","name2")`	Use the color(s) `name1` and `name2` in the figures
`clr="palette.name"`	use the palette `palette.name` to define colors for the figures

Set fig=0 to Suppress Figures

```
describe("Age",nki70,fig=0)
```

```
 **TABLES**

|               |        Age|
|:--------------|----------:|
|n.total        | 144.000000|
|n.missing      |   0.000000|
|n.available    | 144.000000|
|mean           |  44.305556|
|stdev          |   5.339230|
|median         |  45.000000|
|lower.quartile |  41.000000|
|upper.quartile |  49.000000|
|minimum        |  26.000000|
|maximum        |  53.000000|
|shapiro.pvalue |   0.000182|



 **RESULTS**

The variable Age has 144 observations (144 available; 0 missing)  with mean 44.3, standard deviation 5.3, median 45, lower quartile 41, upper
quartile 49, minimum 26, and maximum 53.


 **METHODS**

The Shapiro-Wilk (1965) test was used to evaluate the normality of the distribution of Age.


 **REFERENCES**

Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3-4): 591-611.
doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384.
```

Set `tbl=0` to Suppress Tables

```
describe("Age",nki70,tbl=0,fig=0)
```

```
 **RESULTS**

The variable Age has 144 observations (144 available; 0 missing)  with mean 44.3, standard deviation 5.3, median 45, lower quartile 41, upper
quartile 49, minimum 26, and maximum 53.


 **METHODS**

The Shapiro-Wilk (1965) test was used to evaluate the normality of the distribution of Age.


 **REFERENCES**

Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3-4): 591-611.
doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384.
```

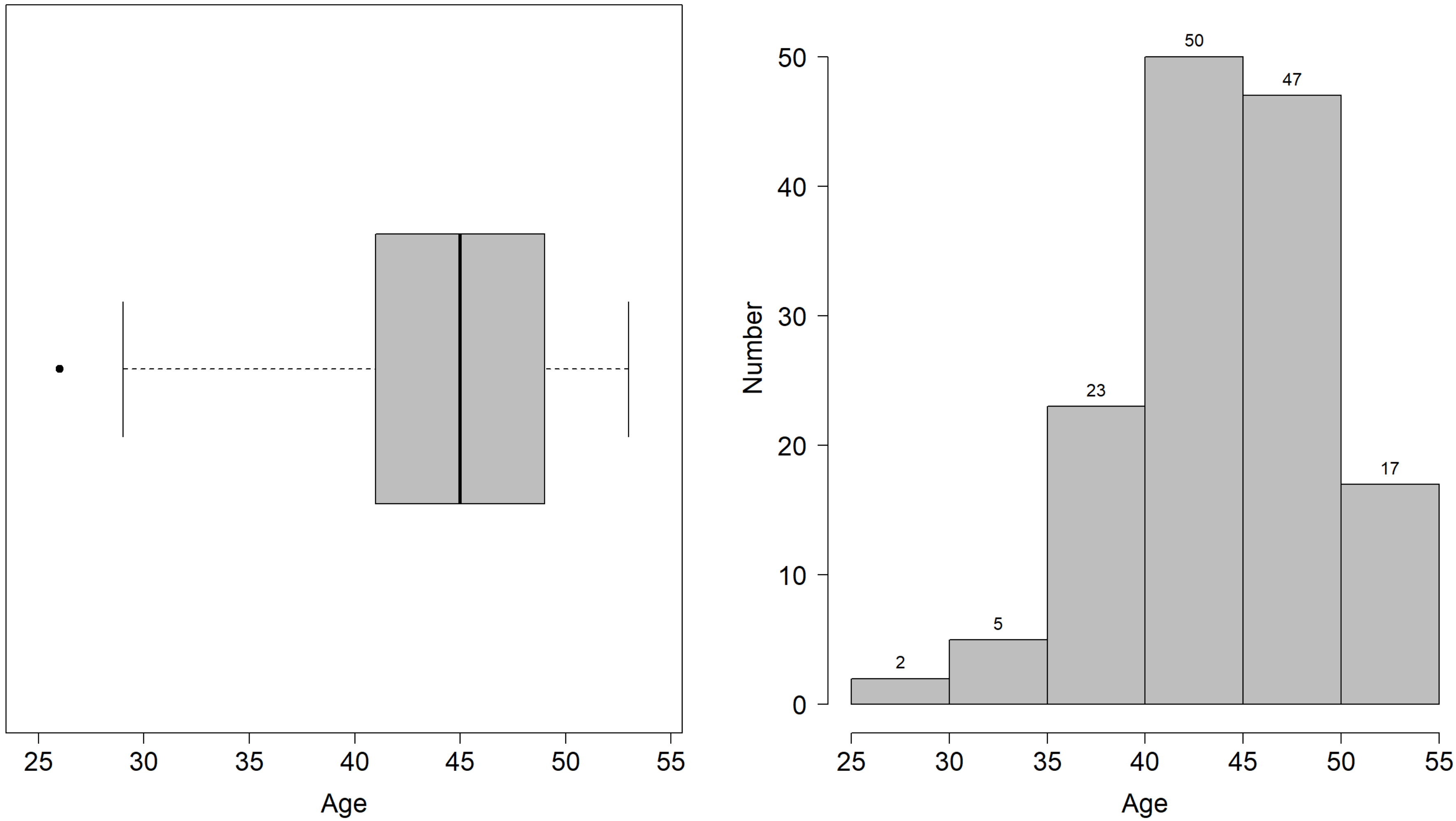
Set `txt=0` to Suppress Text

```
describe("Age",nki70,txt=0,fig=0)
```

```
 **TABLES**

|               |        Age|
|:--------------|----------:|
|n.total        | 144.000000|
|n.missing      |   0.000000|
|n.available    | 144.000000|
|mean           |  44.305556|
|stdev          |   5.339230|
|median         |  45.000000|
|lower.quartile |  41.000000|
|upper.quartile |  49.000000|
|minimum        |  26.000000|
|maximum        |  53.000000|
|shapiro.pvalue |   0.000182|



 **METHODS**

The Shapiro-Wilk (1965) test was used to evaluate the normality of the distribution of Age.


 **REFERENCES**

Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3-4): 591-611.
doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384.
```

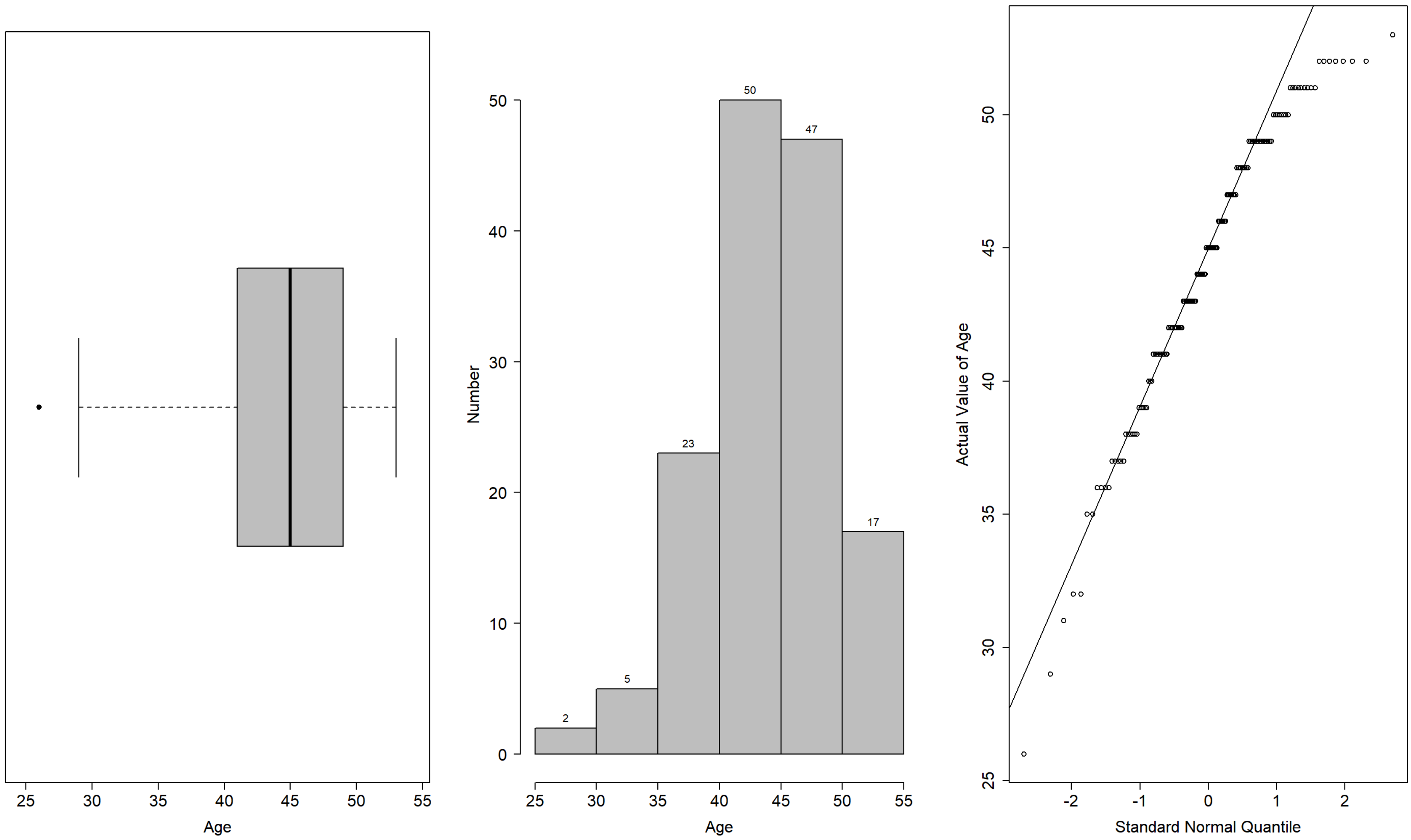
Set fig=2 to Get More Figures

```
describe("Age",nki70,fig=2)
```



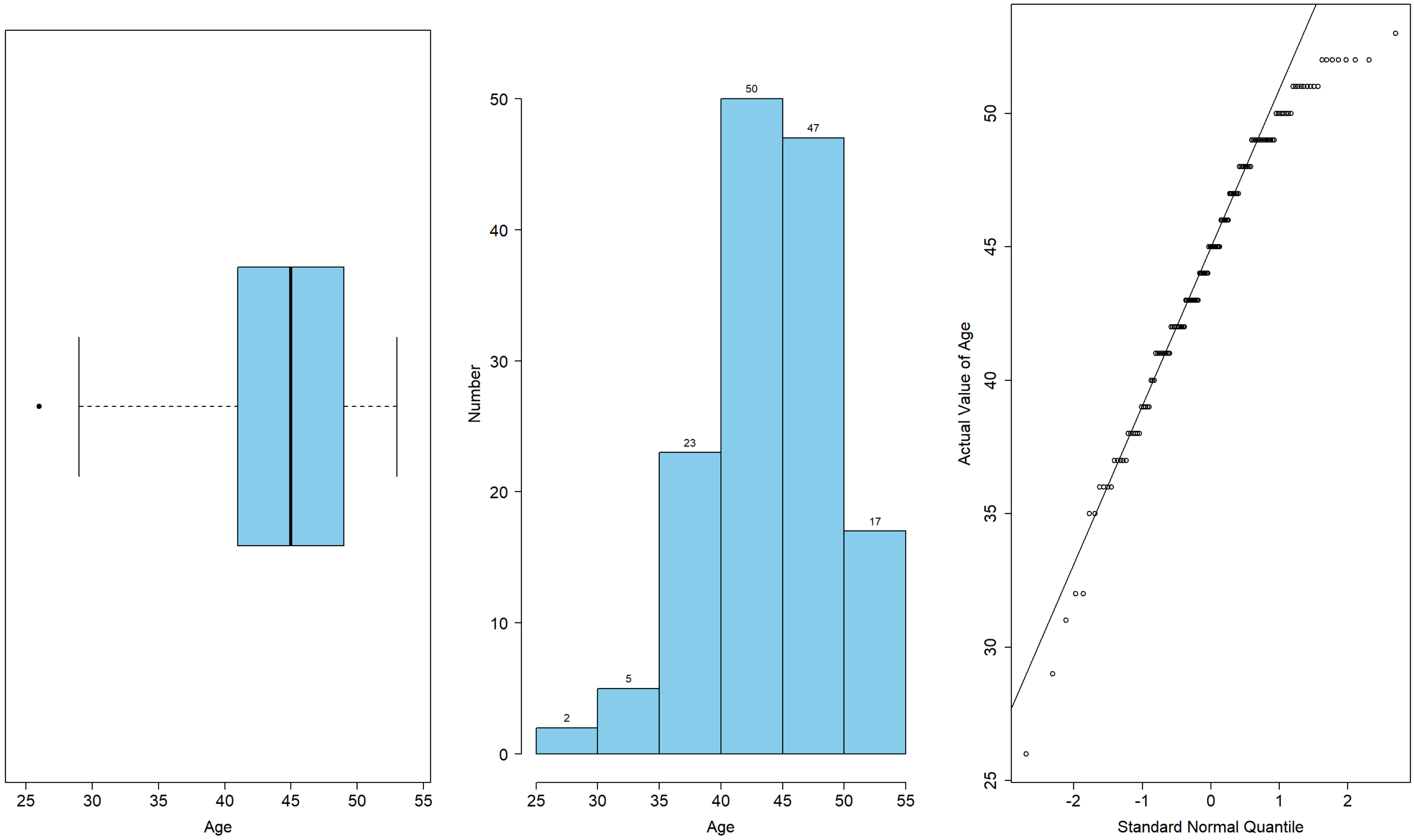
Set fig=3 to Get Even More Figures

```
describe("Age",nki70,fig=3)
```



Set `clr="skyblue"` to Get Sky Blue Figures

```
describe("Age",nki70,fig=3,clr="skyblue")
```

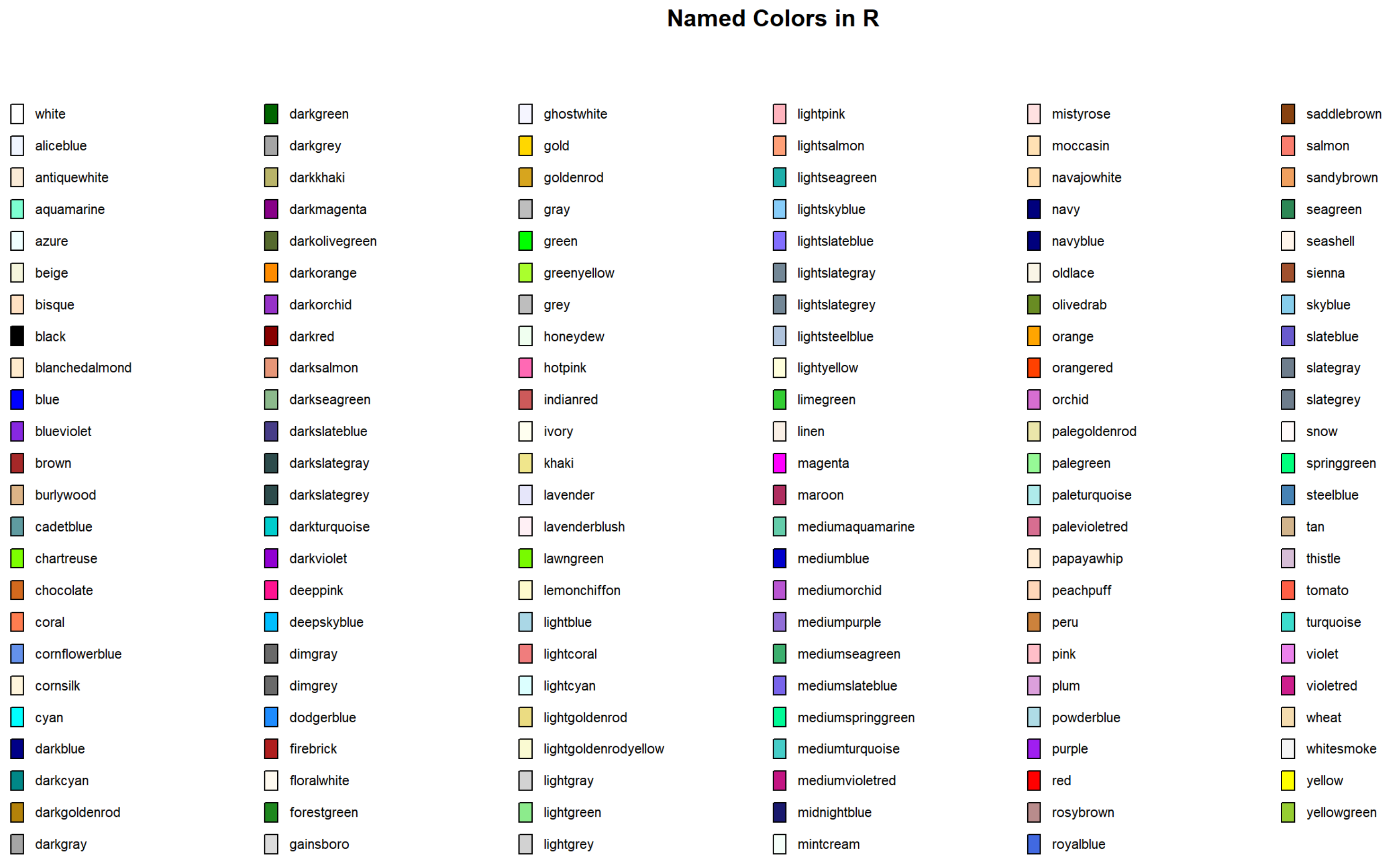


Colors in SBP

- Use the `clr` option to specify colors for figures.
- One may specify the name of one color, names of multiple colors, or the name of a color palette.
- Use the function `show.colors()` to see the colors and their names.
- Use `show.palettes(n)` to see palettes of `n` colors.

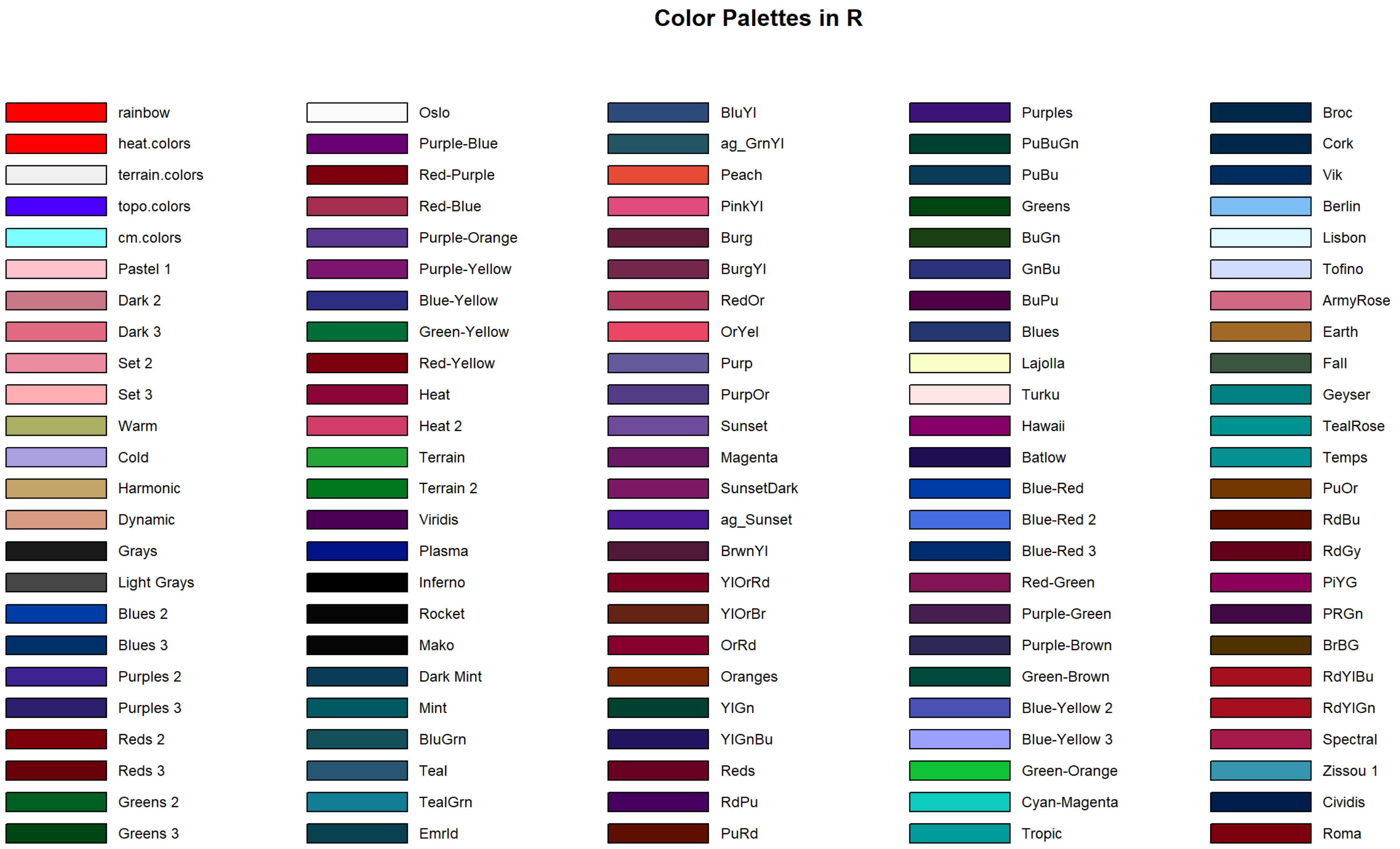
Colors in SBP

```
show.colors()
```



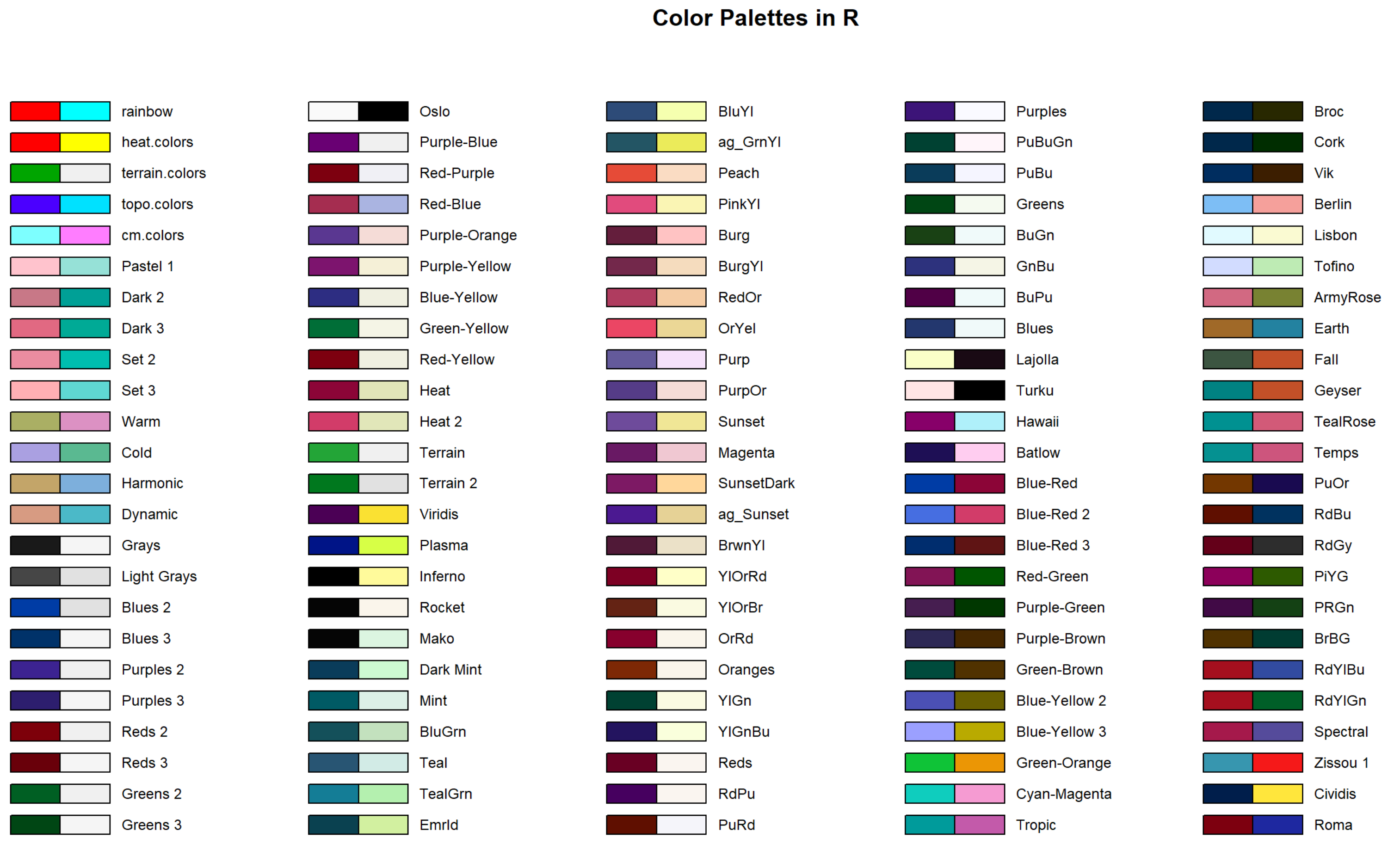
One-Color Palettes in SBP

```
show.palettes(1)
```



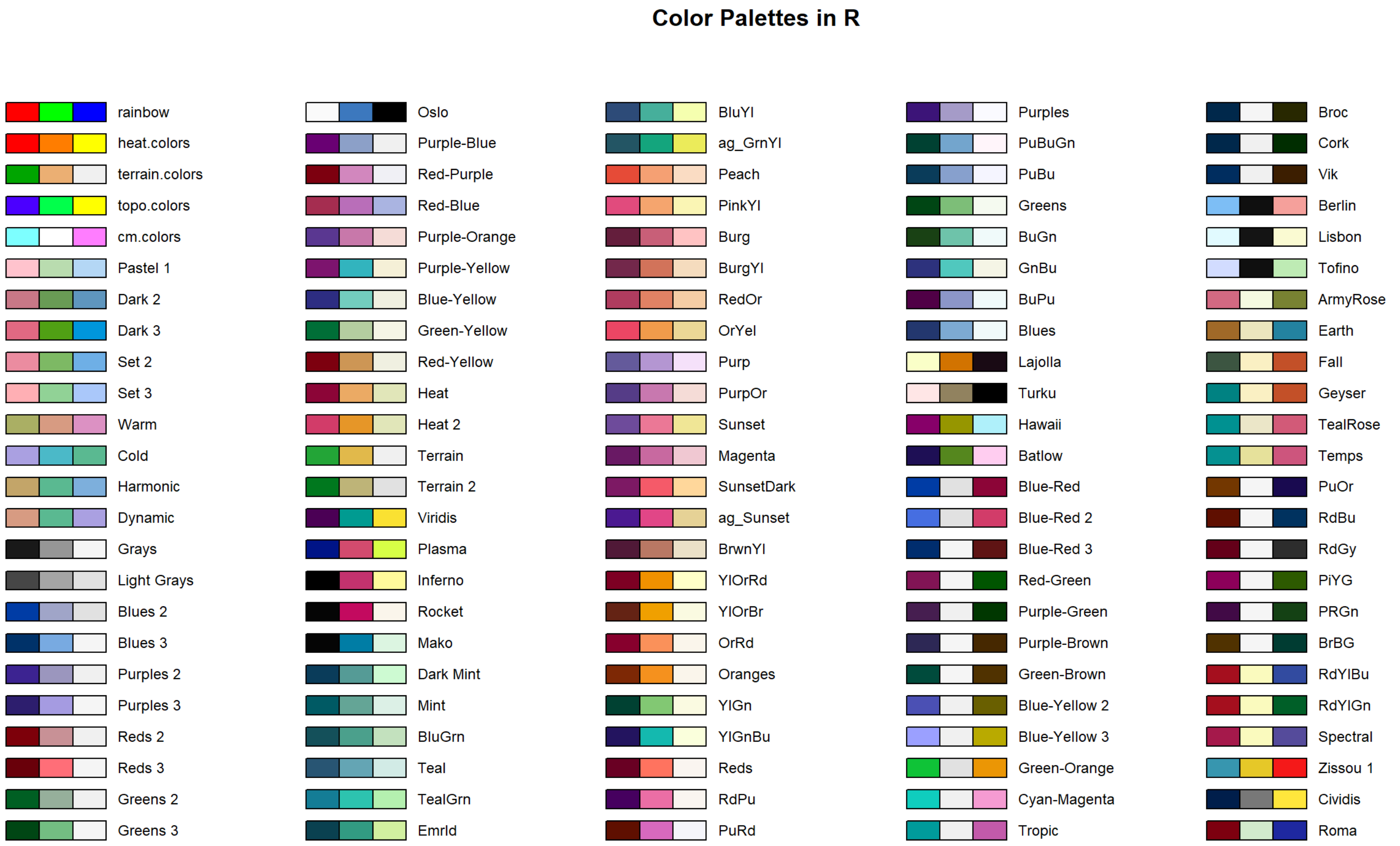
Two-Color Palettes in SBP

```
show.palettes(2)
```



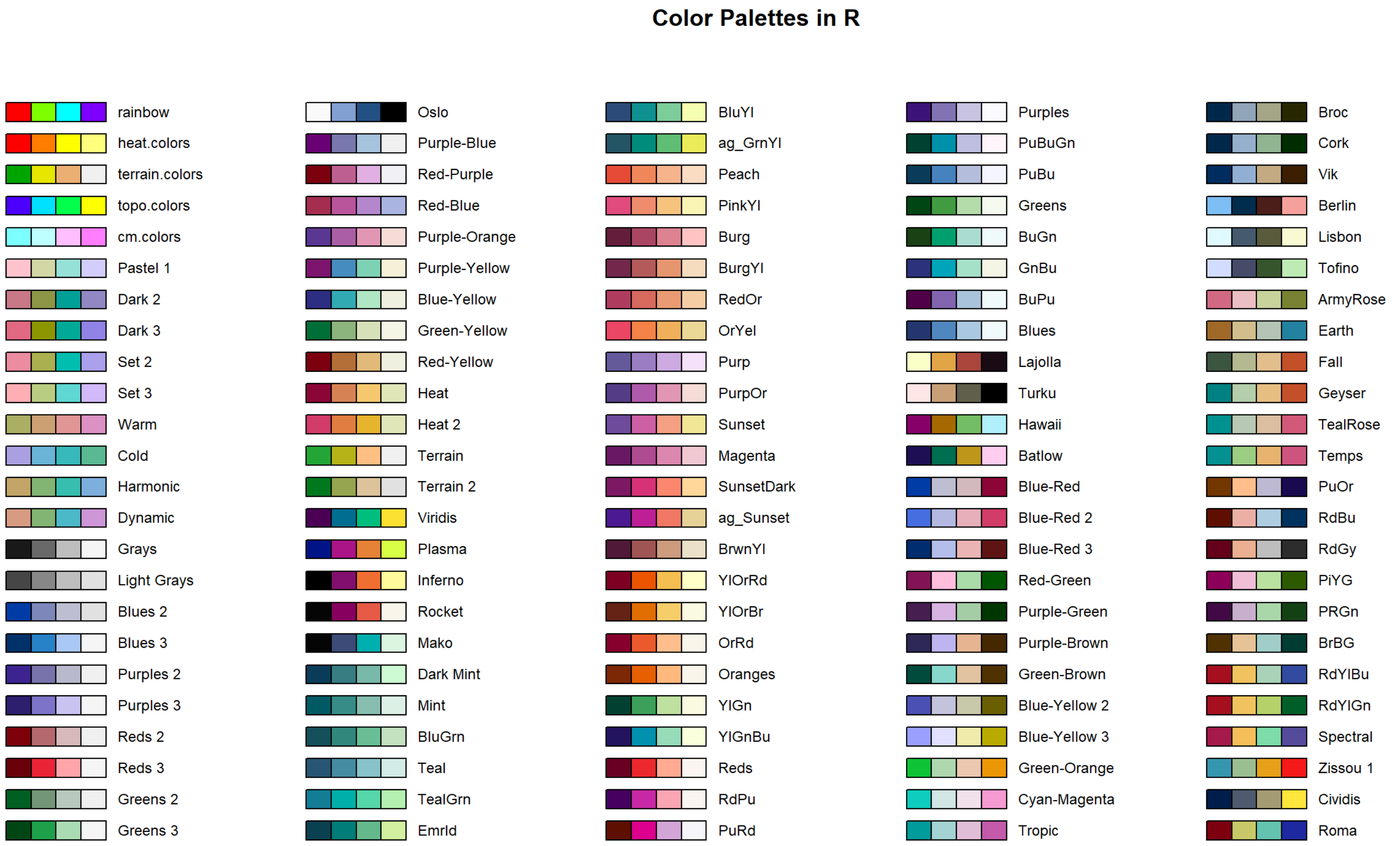
Three-Color Palettes in SBP

```r
show.palettes(3)
```



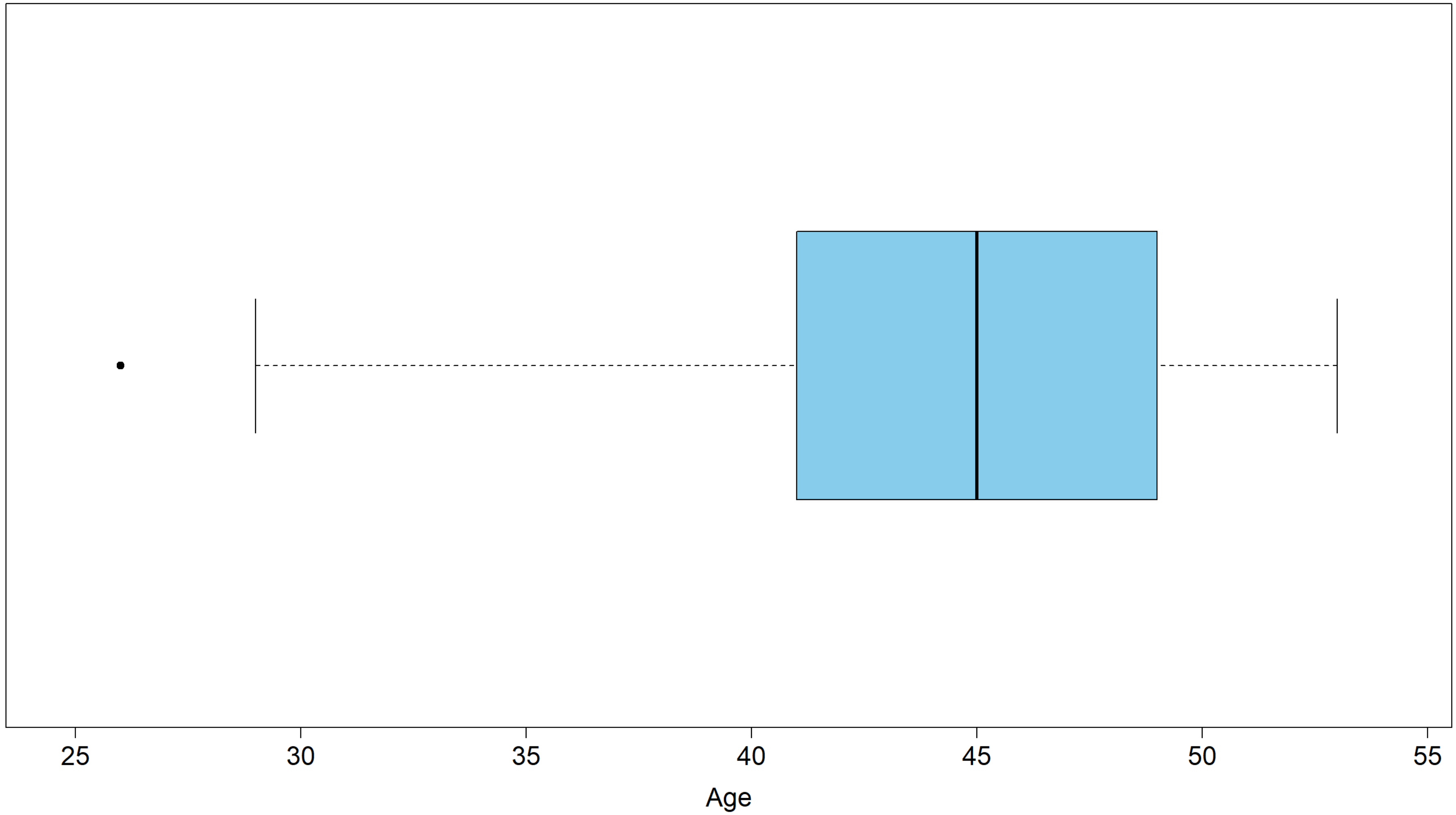
Four-Color Palettes in SBP

```r
show.palettes(4)
```



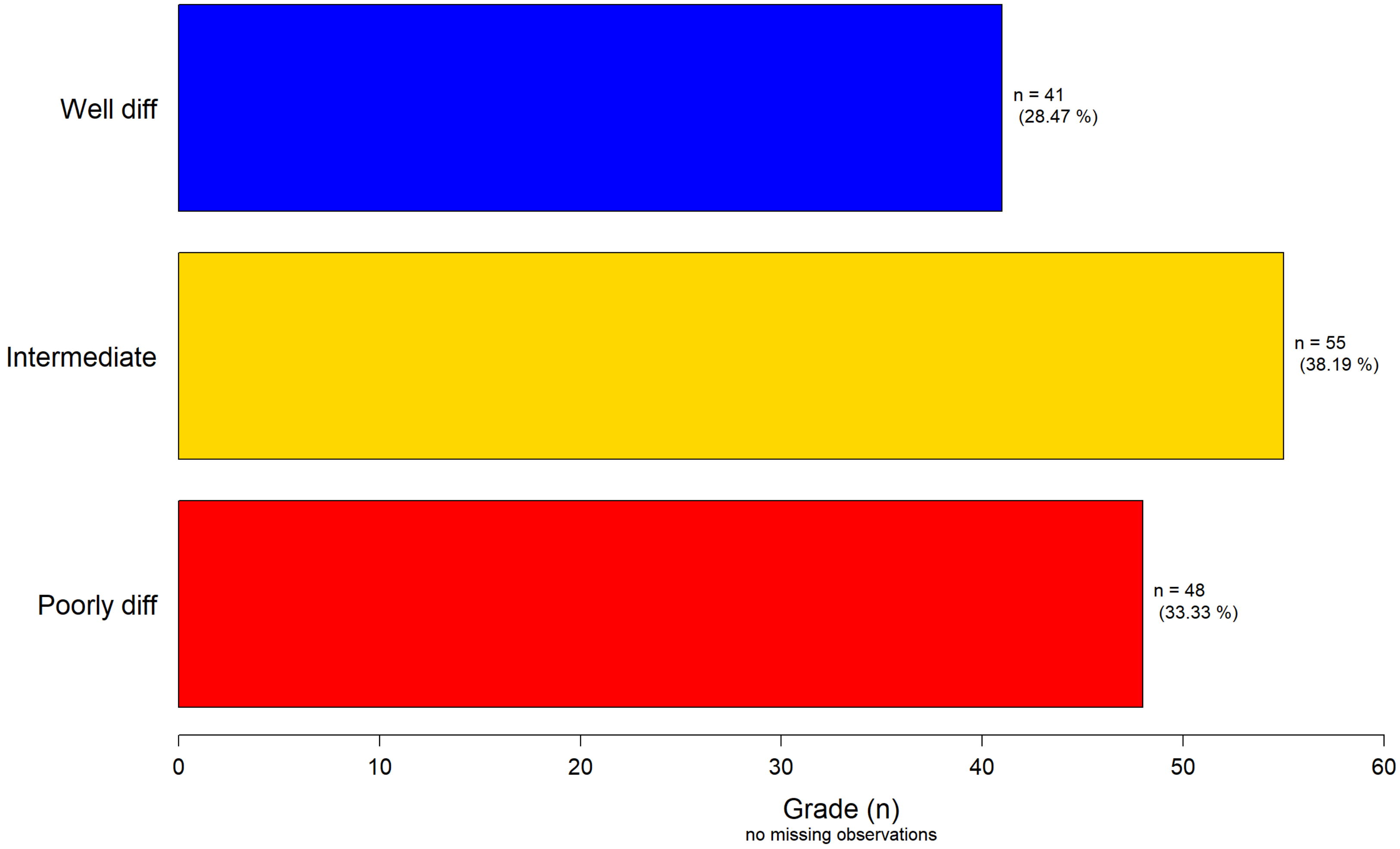
Produce Sky Blue Figures

```
describe("Age",nki70,clr="skyblue")
```



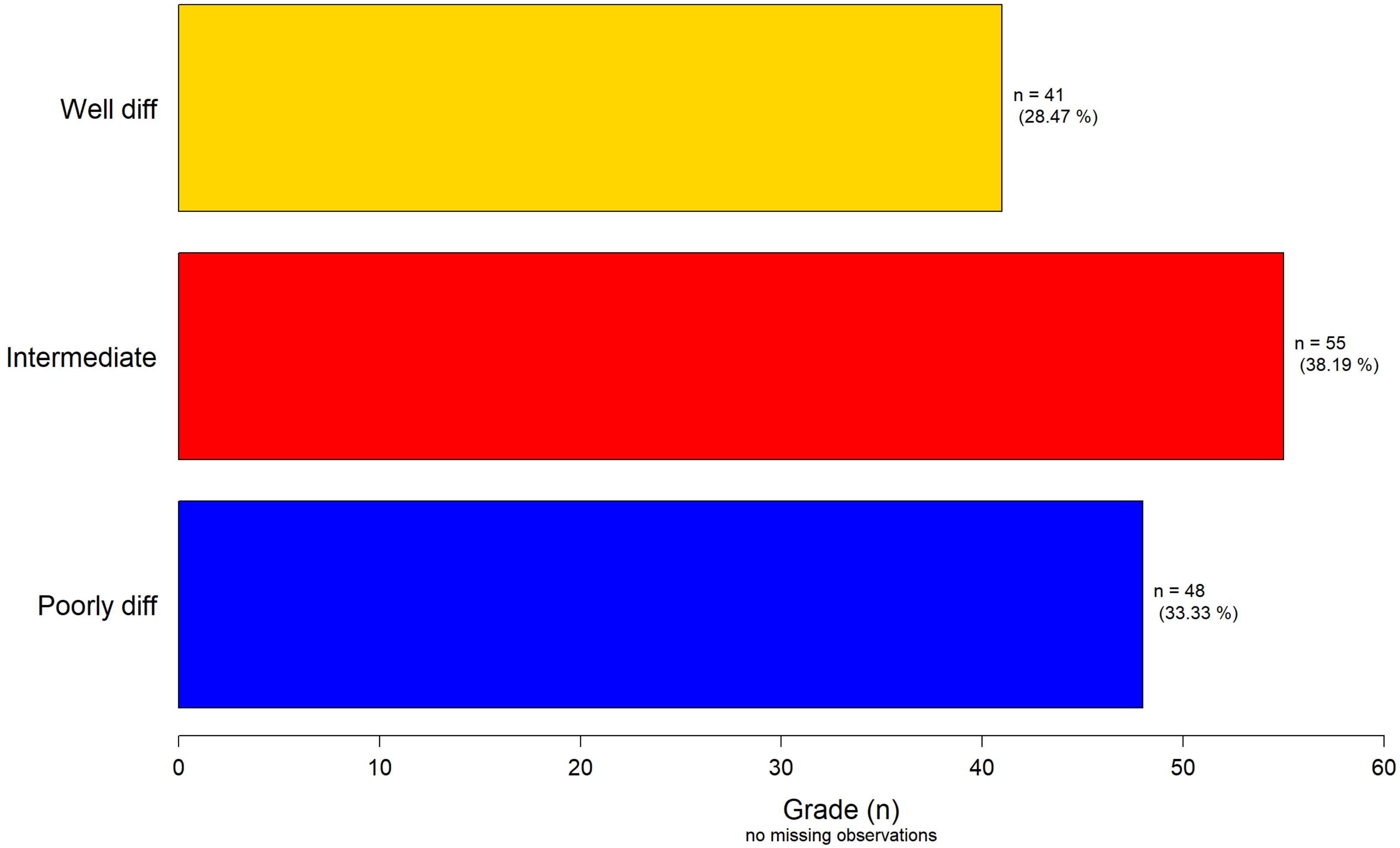
Specify Multiple Color Names

```
describe("Grade",nki70,clr=c("red","gold","blue"))
```



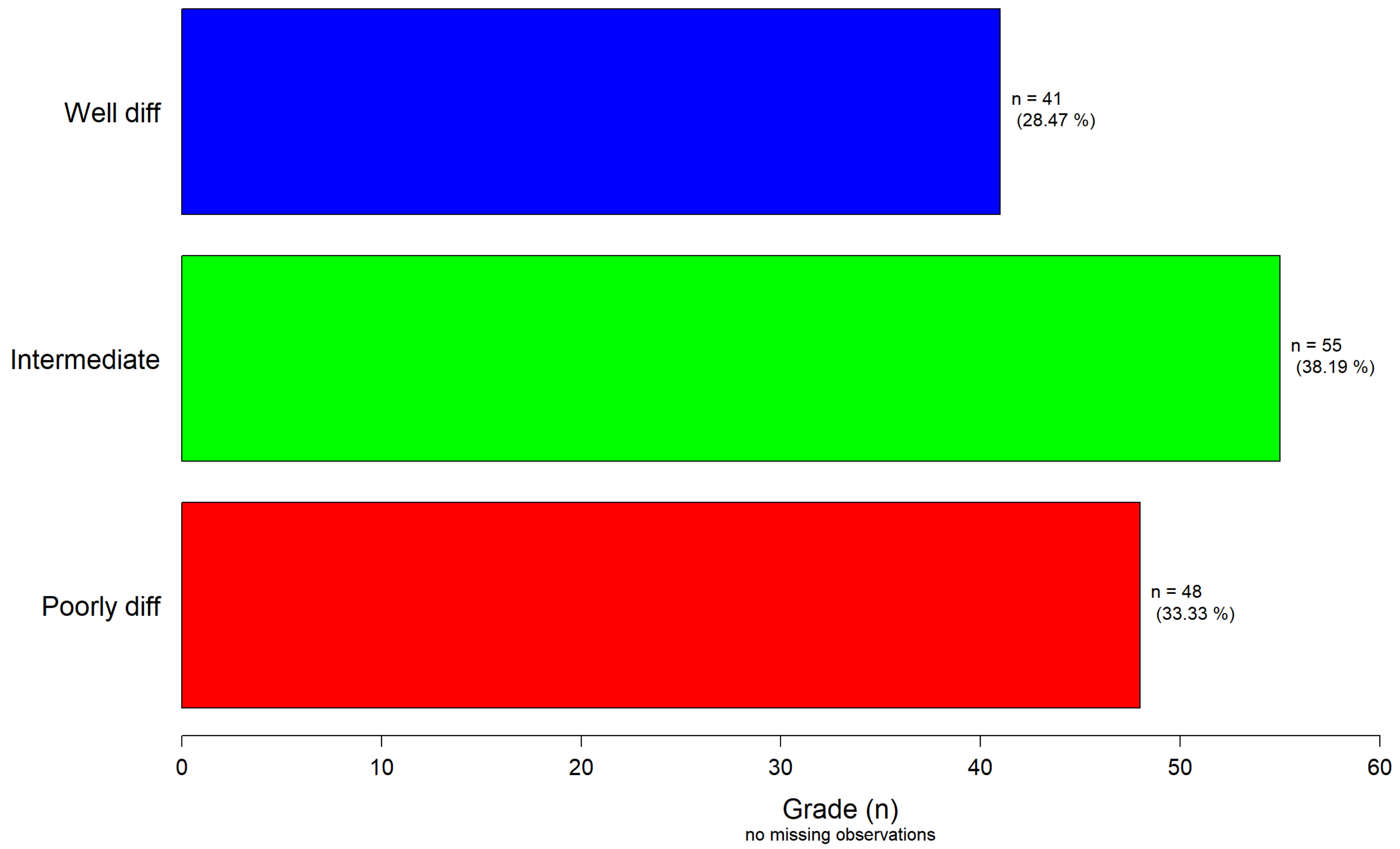
Specify Multiple Color Names

```
describe("Grade",nki70,clr=c("blue","red","gold"))
```



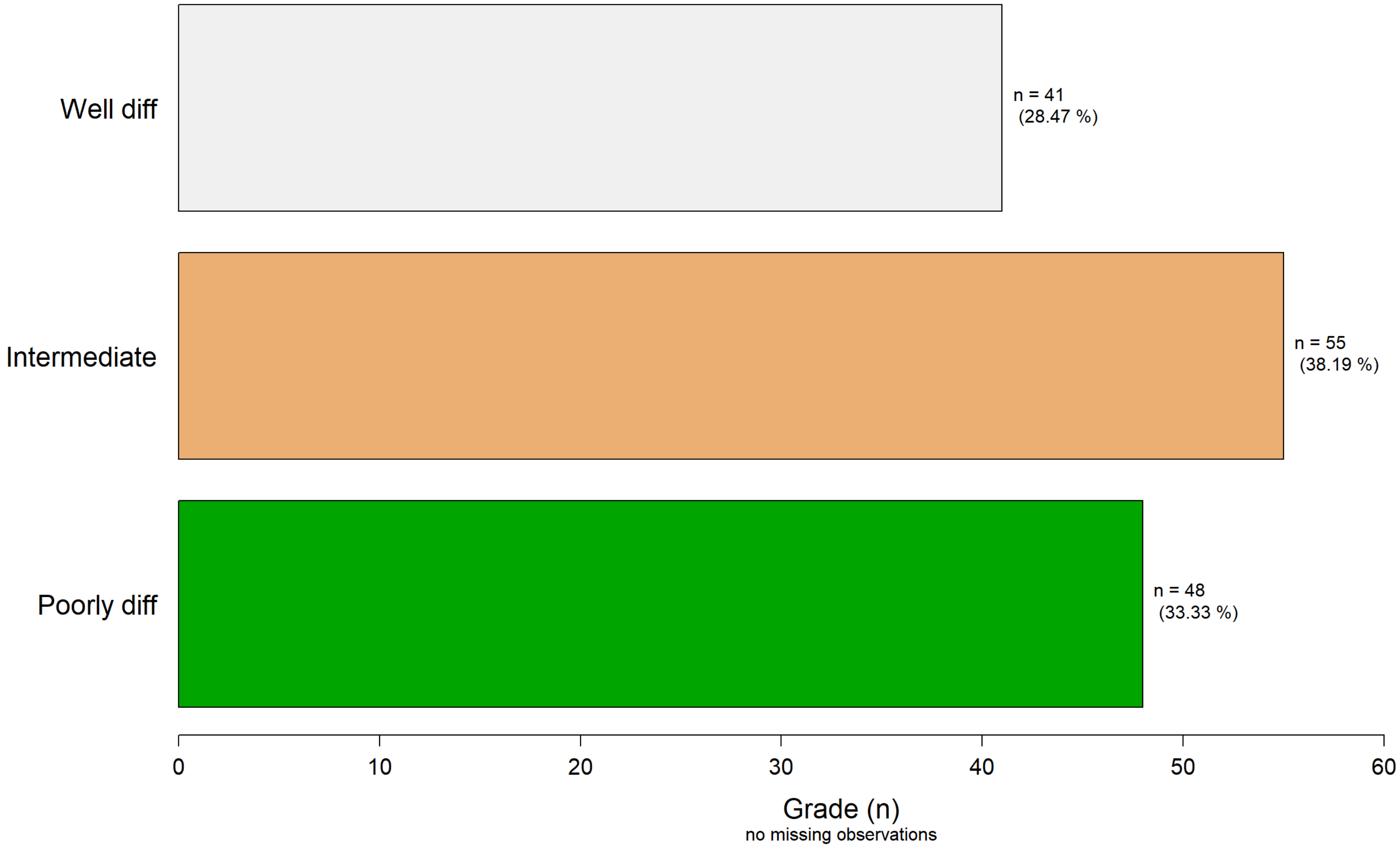
Specify the Rainbow Color Palette

```
describe("Grade",nki70,clr="rainbow")
```



Specify the Terrain Color Palette

```
describe("Grade",nki70,clr="terrain.colors")
```



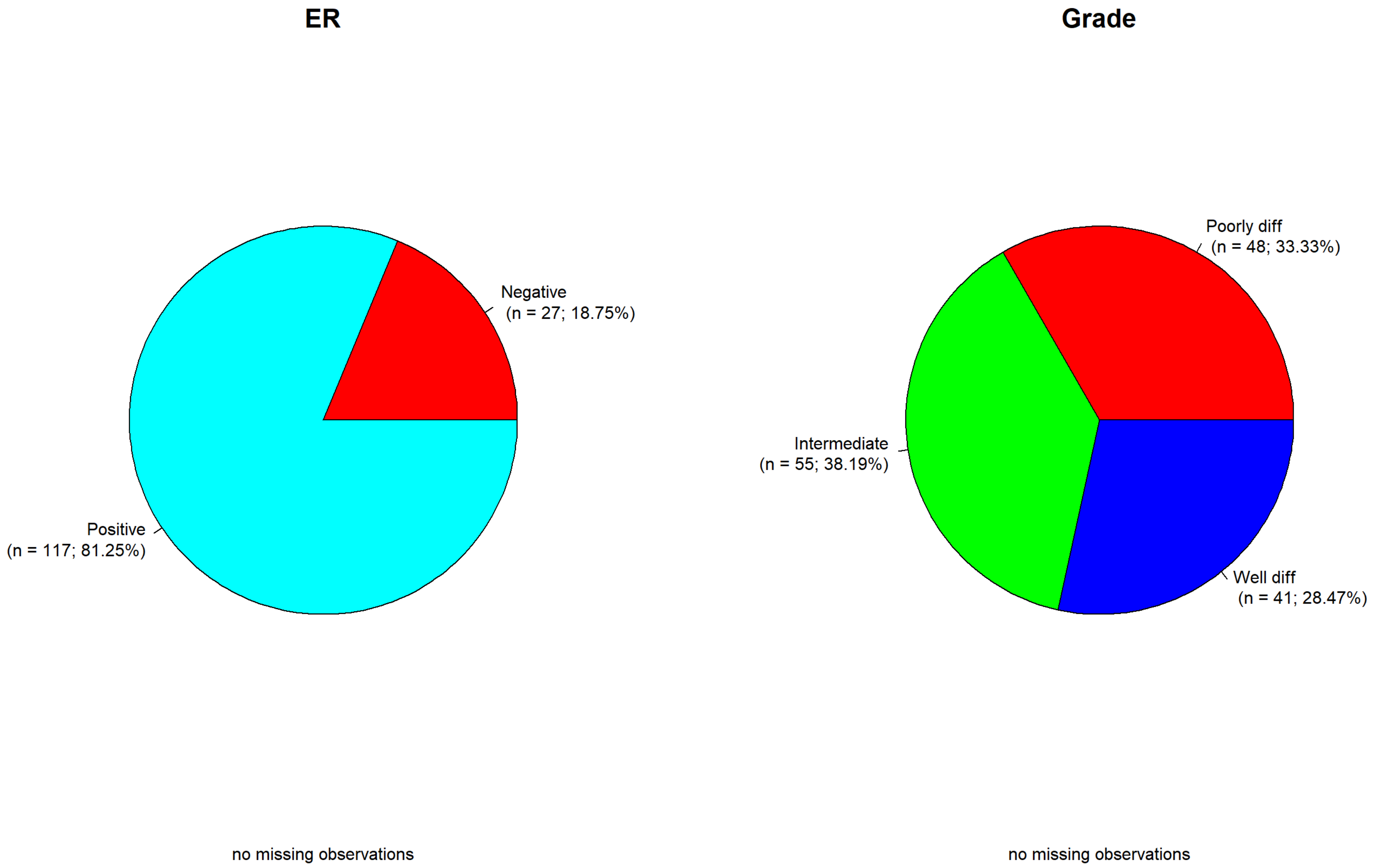
Specific Options for Data Analysis Functions

Graphics Functions

Function	Purpose
`pie.plot("y",data.set)`	Produce a pie chart of the categorical data column y of `data.set`
`bar.plot("y",data.set)`	Produce a bar plot or histogram of the y column of `data.set`
`box.plot("y",data.set)`	Produce a box plot of the numeric data column y of `data.set`
`box.plot(y~grp,data.set)`	Produce side-by-side boxplots of the numeric y by the group `grp` of `data.set`
`nqq.plot("y",data.set)`	Produce a normal quantile-quantile plot of the numeric y column of `data.set`
`mosaic.plot(y~x,data.set)`	Produce a mosaic plot for the categorical data columns y and x of `data.set`
`scatter.plot(y~x,data.set)`	Produce a scatter plot of y versus x for `data.set`
`event.plot("evnt",data.set)`	Plot the survival or cumulative incidence of the `evnt` column of `data.set`
`event.plot(evnt~grp,data.set)`	Plot the survival or cumulative incidence of the `evnt` by `grp` groups

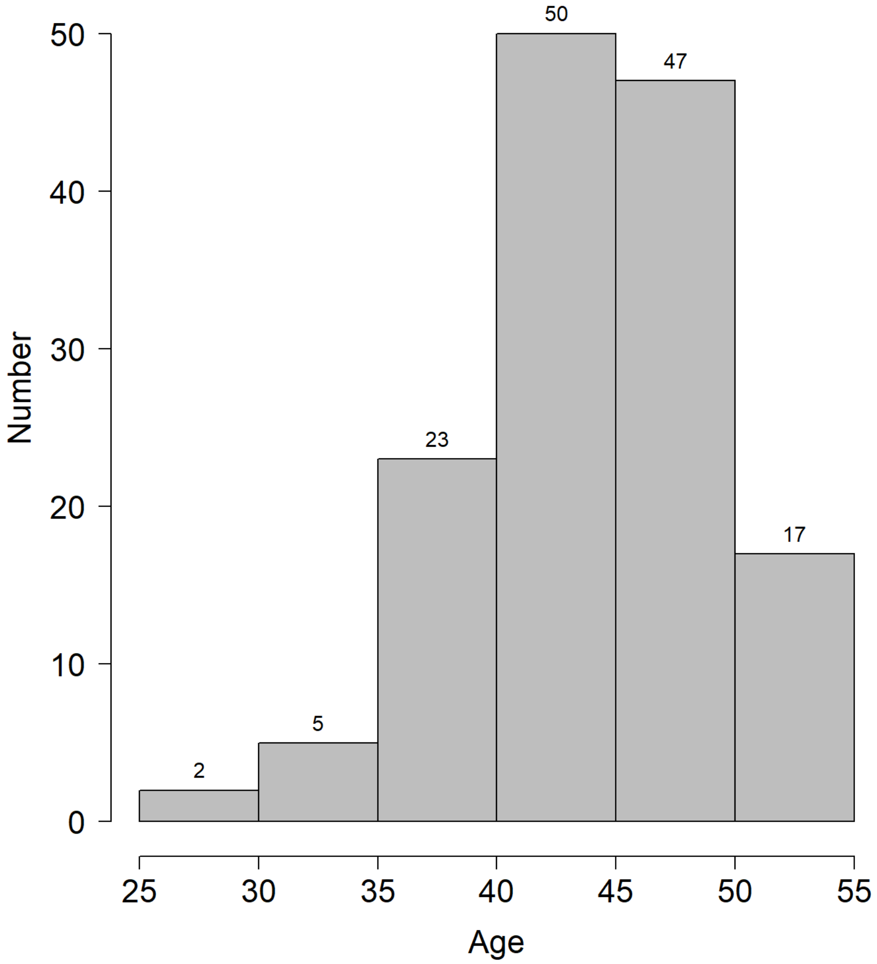
Pie Plot Examples

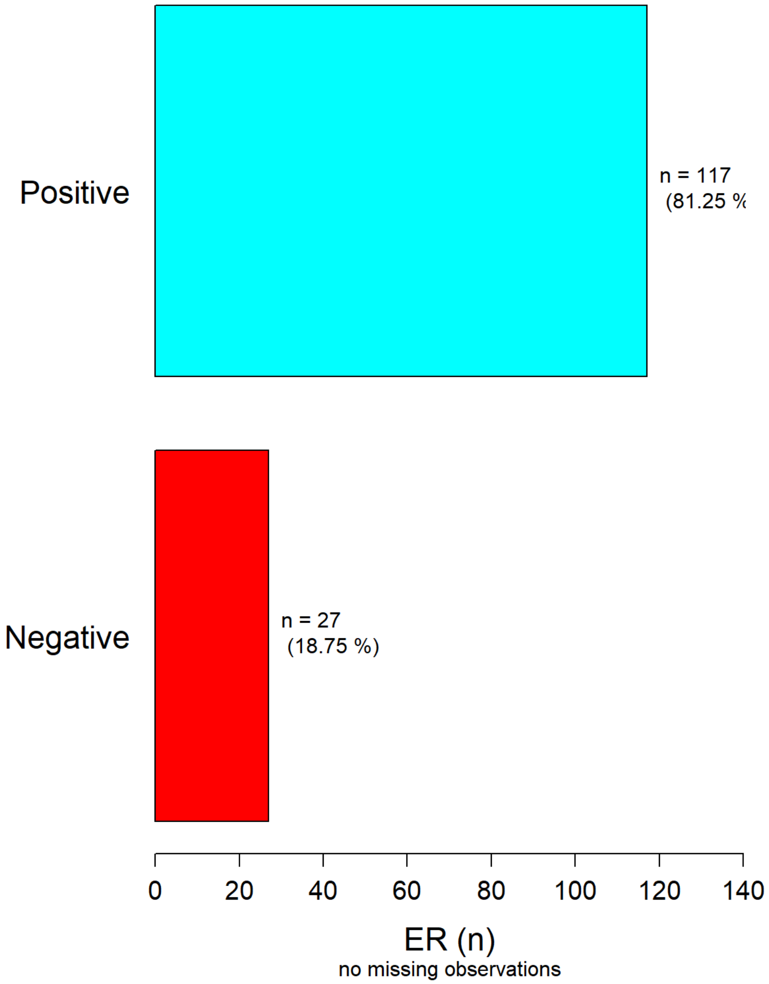
```
pie.plot("ER",nki70)     # pie chart for ER status
pie.plot("Grade",nki70)  # pie chart of Tumor Grade
```



Bar Plot Examples

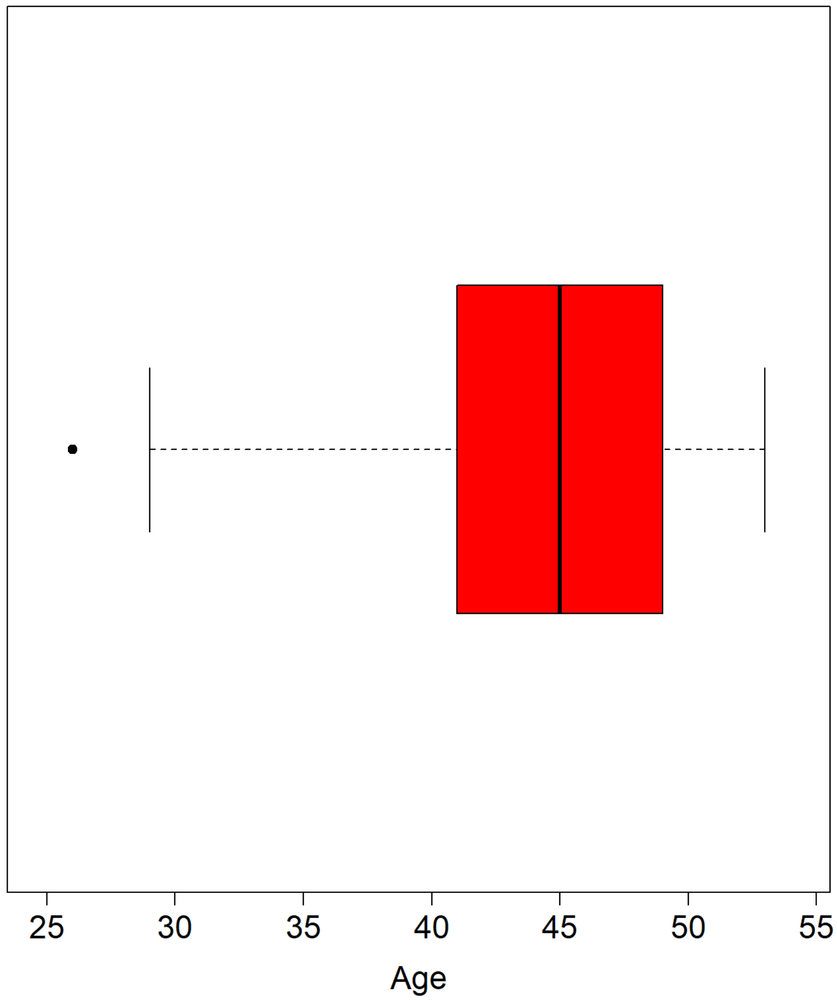
```
bar.plot("Age",nki70) # histogram for a numeric variable
bar.plot("ER",nki70)  # bar plot for a categorical variable
```

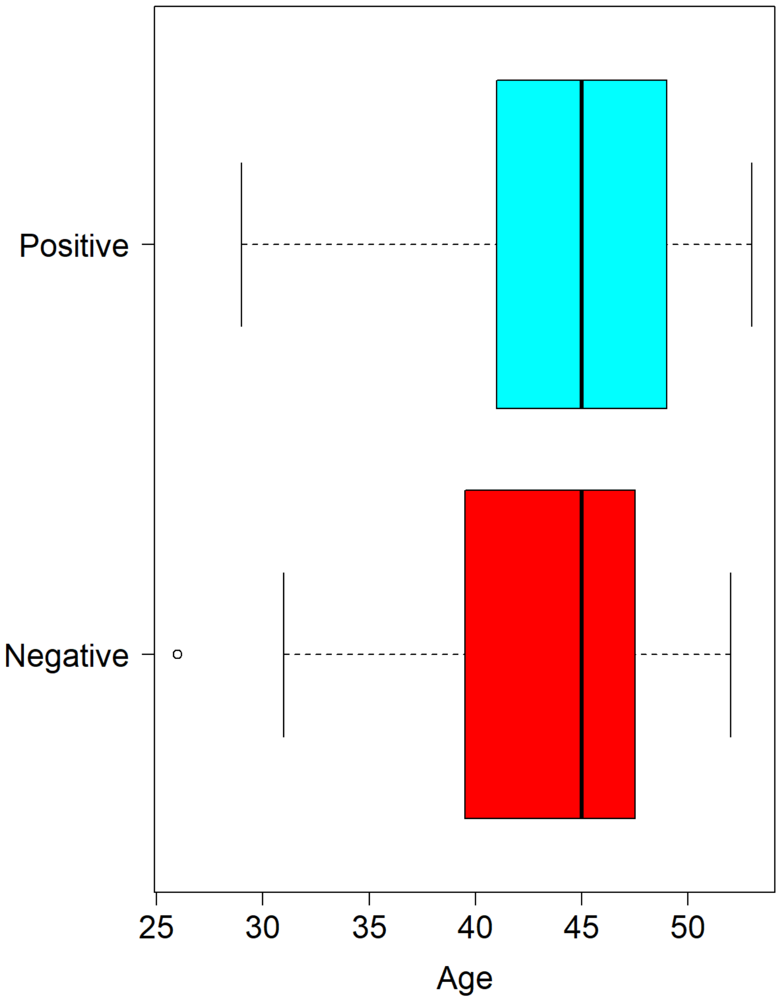




Box Plot Examples

```
box.plot("Age",nki70)   # box plot of Age for all data
box.plot(Age~ER,nki70)  # side-by-side boxplots of Age by ER status
```



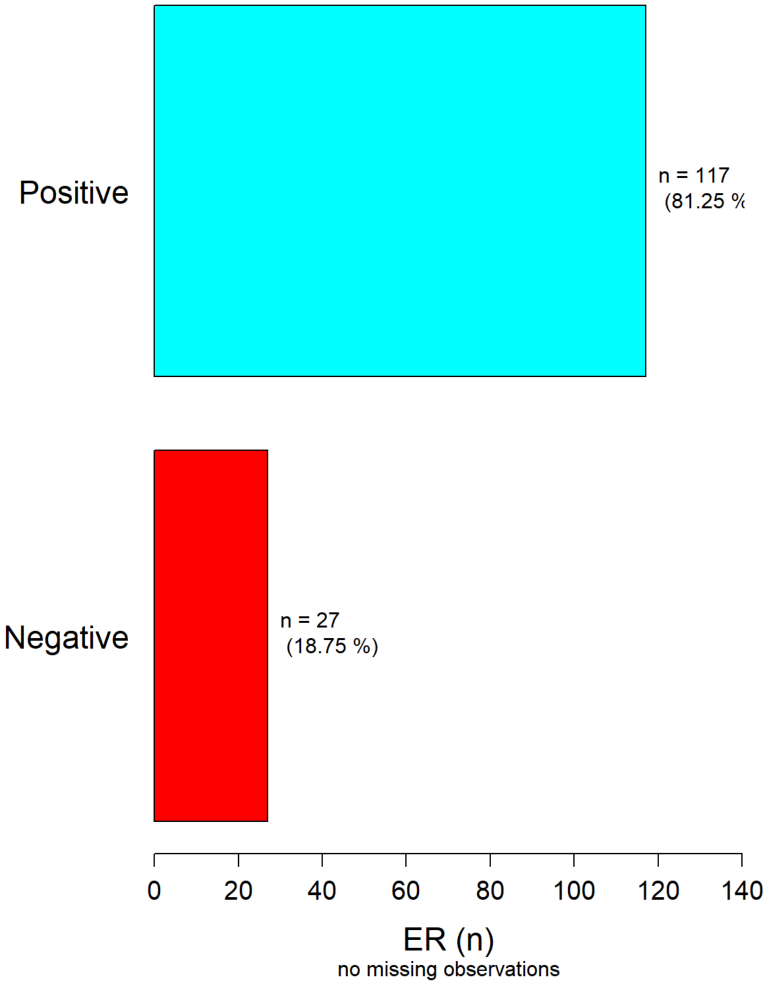


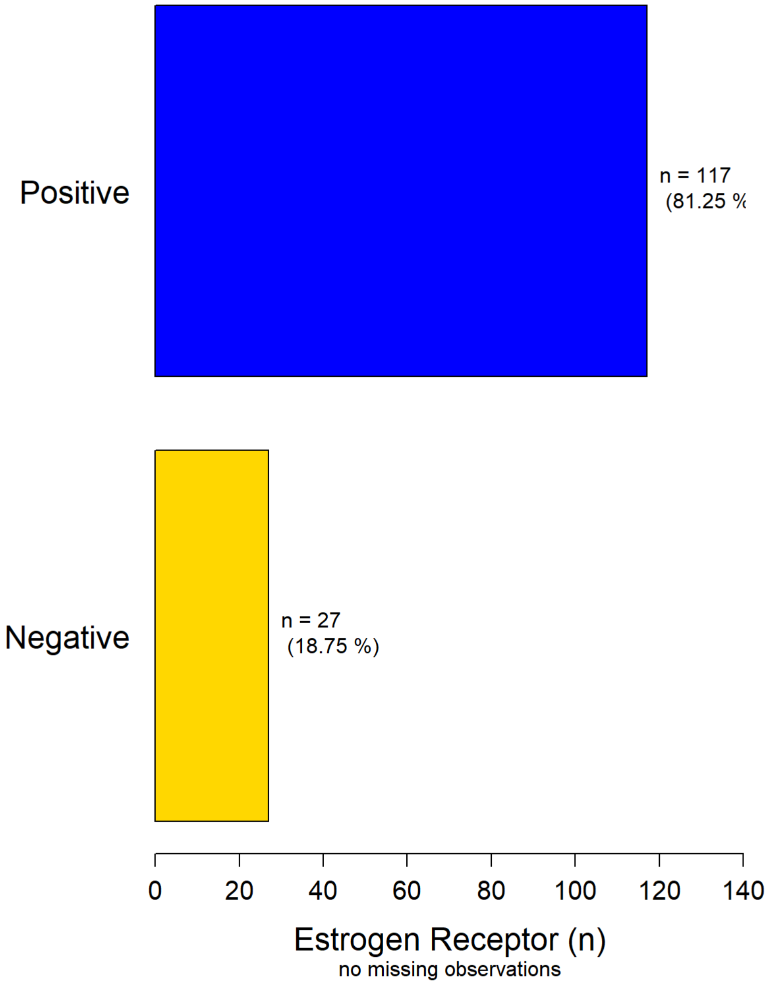
Common Options for Graphics Functions

Option	Action
`clr=color.name`	Specify colors or palette for figure
`y.name="name"`	Use `"name"` to label the y variable in the figure

Example

```
bar.plot("ER",nki70)
bar.plot("ER",nki70,y.name="Estrogen Receptor",clr=c("gold","blue"))
```



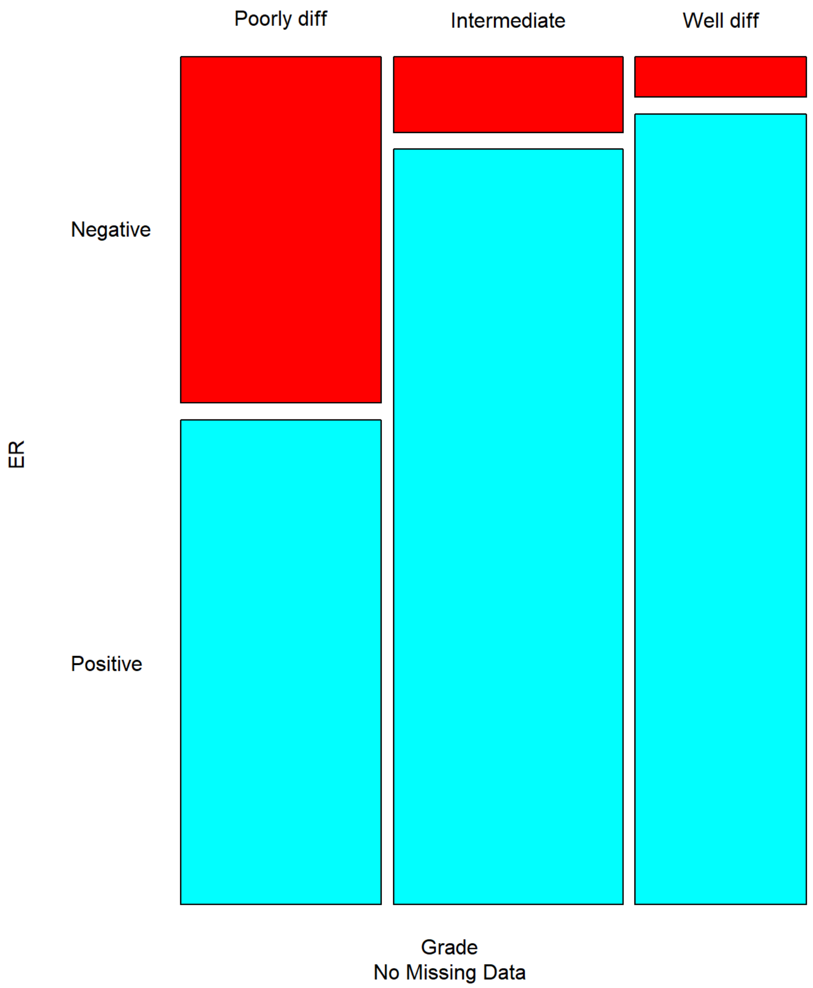


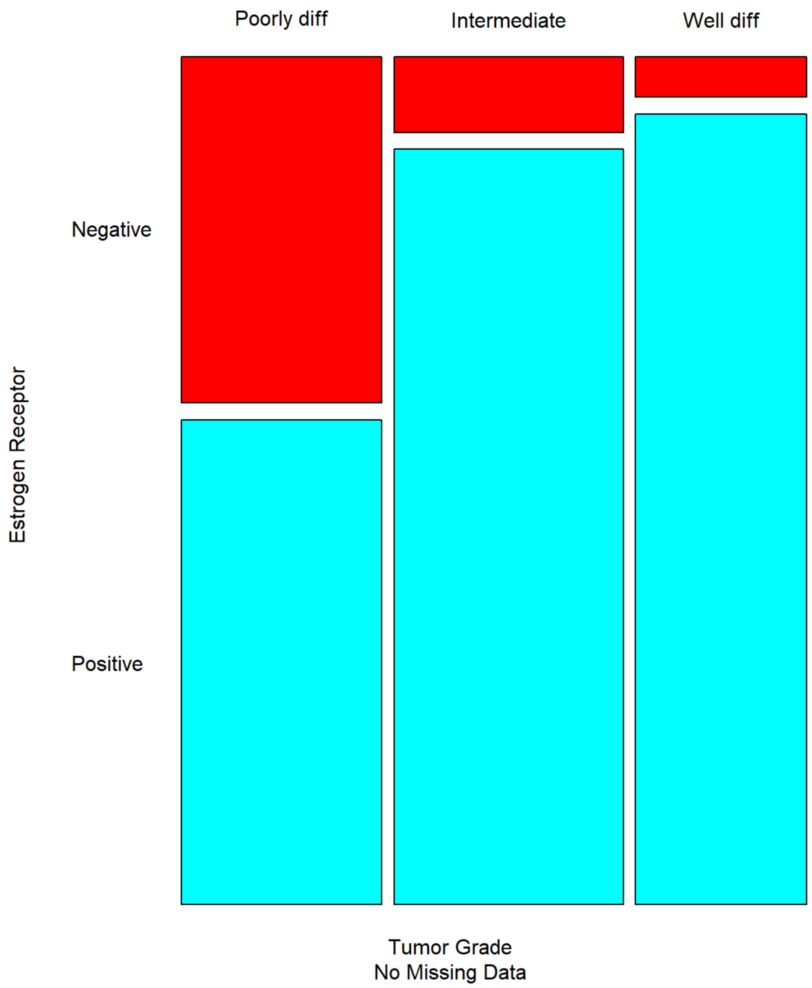
Specific Options for Graphics Functions

Function(s)	Option	Action
`mosiac.plot`	`grp.name="name"`	Use `"name"` to label the group variable in the figure
`scatter.plot`	`x.name="name"`	Use `"name"` to label the x-axis in the figure
`bar.plot`, `pie.plot`	`all=FALSE` or `all=TRUE`	Indicates whether to use **all** data by including missing data as a distinct category
`scatter.plot`	`line=0,1,`, or `NA`	Add a flat line (`line=0`), a fitted line (`line=1`) or no line (`line=NA`)

Example

```
mosaic.plot(ER~Grade,nki70)
mosaic.plot(ER~Grade,nki70,y.name="Estrogen Receptor",grp.name="Tumor Grade")
```





Including Tables in Word

```
age.result=describe("Age",nki70,fig=0)
word.table(age.result)
```

```
Age
n.total,144
n.missing,0
n.available,144
mean,44.3055555555556
stdev,5.3392304227652
median,45
lower.quartile,41
upper.quartile,49
minimum,26
maximum,53
shapiro.pvalue,0.000181978293787236



**INSTRUCTIONS**
1. Copy the output into Word.
2. Highlight the output in Word.
3. Go to Insert>Table>Convert Text to Table.
```

Including Figures in Word

- In R Studio, click on the *Plots* panel.
- Use the left and right arrows to navigate to the figure of interest.
- Click the *Export* button.
- Choose *Copy to Clipboard*.
- Click "Copy Plot"
- Paste the plot into Word.

Summary

- The Simple Biostatistics Program (SBP) provides a simple *R* interface to produce tables, figures, and narratives for the statistical procedures covered in an introductory biostatistics class.
- Use the function `` `read.data `` to read data into R.
- Use the function `get.package` to make an R package available for use in an R session.
- Data Analysis Functions: `describe`, `estimate`, `compare`, `correlate`, `model`.
- Graphics Functions: `bar.plot`, `pie.plot`, `box.plot`, `nqq.plot`, `mosaic.plot`, `scatter.plot`, `event.plot`
- Use `word.table` to generate tabular output to copy and paste into Word. Then, use *Insert>Table>Convert Text to Table* to represent the output as a table in Word.
- Use the *Export* button in R studio to copy a figure to the clipboard. Then paste the figure into Word.

Practice Exercise (Not Homework)

- Describe the `Diam`, `FGF18`, `GSTM3` columns of the `nki70` data. Tinker with the `txt`, `fig`, `tbl`, and `clr` options.
- Use the graphics functions to create the figures without generating the narrative and tabular output.
- Copy your narrative, tabular, and graphical results into Word.