Scientific Computing – Statistics

Fabian Sinz
Dept. Neuroethology, University Tübingen
Bernstein Center Tübingen

10/20/2014

my expectations to this course

- interest and participation
- motivation to understand and question concepts
- high scientific standard
- intellectual honesty
- sincere cooperation

this week will be ...

... no ~~fun~~ piece of cake

this week will be ...

... no ~~fun~~ piece of cake

this week will be ...

... no lecture (please!)

What you should learn this week

- What makes good plots?
- What is descriptive/inferential statistics?
- What is the general structure of a statistical test?
- What does a p-value mean?
- How can I build my own tests?
- How large should my n be?
- What is maximum likelihood and why is it important?

Day 1 – descriptive statistics and plots

data scales

What data types are distinguished in statistics?

Why are data types important?

data scales

What data types are distinguished in statistics?

Why are data types important?

- selection of statistics
- selection of plots
- selection of correct tests

data scales

nominal/categorial scale

- properties like cell type, experimental group (i.e. treatment 1, treatment 2, control)
- each observation/sample is put into one category
- there is no reasonable order among the categories
- example: [rods, cones] vs. [cones, rods]

data scales

ordinal scale

- like nominal scale, but there is an order
- **but:** there is no reasonable measure of distance between the classes
- examples: ranks, ratings

data scales

interval scale

- quantitative/metric values
- reasonable measure of distance between values but no absolute zero
- examples: temperature in °C

data scales

absolut/ratio scale

- like interval scale but with absolute zero
- example: temperature in °K

data scales

absolut/ratio scale

- like interval scale but with absolute zero
- example: temperature in °K

relationsships between scales

- scales exhibit increasing information content from nominal to absolute
- conversion „downwards" always possible

examples from neuroscience and psychology

- **nominal:**

examples from neuroscience and psychology

- **nominal:**
  - treatment group
  - stimulus class
  - cell type
- **ordinal:**

examples from neuroscience and psychology

- **nominal:**
  - treatment group
  - stimulus class
  - cell type
- **ordinal:**
  - ratings
  - clinical stages of a disease
  - states of an ion channel
- **Absolut-/Ratioskala:**

examples from neuroscience and psychology

- **nominal:**
    - treatment group
    - stimulus class
    - cell type
- **ordinal:**
    - ratings
    - clinical stages of a disease
    - states of an ion channel
- **Absolut-/Ratioskala:**
    - firing rate
    - membrane potential
    - ion concentration

Day 1 – descriptive statistics and plots

What is "a statistic"?

statistic

A statistic (singular) is a single measure of some attribute of a sample (e.g., its arithmetic mean value). It is calculated by applying a function (statistical algorithm) to the values of the items of the sample, which are known together as a set of data.

http://en.wikipedia.org/wiki/Statistic

Beispiele für Teststatistiken

- **nominal:**

Beispiele für Teststatistiken

- **nominal:**
  - count
  - relative frequency/proportion
- **ordinal:**

Beispiele für Teststatistiken

- **nominal:**
  - count
  - relative frequency/proportion
- **ordinal:**
  - median
  - quantile/percentile
  - rank correlation
- **absolute/ratio:**

Beispiele für Teststatistiken

- **nominal:**
  - count
  - relative frequency/proportion
- **ordinal:**
  - median
  - quantile/percentile
  - rank correlation
- **absolute/ratio:**
  - mean
  - variance/ standard deviation
  - Pearson correlation

exercise

Spearman rank correlation

1. Use `randi` to generate two vectors `x,y` with 100 random integers between 0 and 10 each.
2. Find out how to compute the Spearman rank correlation

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

   with Matlab. $d_i = x_i - y_i$ is the difference in the rank between the single data points.
3. Compute ρ between x and y, between x and y^2, between $\log(x+1)$ and y^2.
4. Compute the "standard" (Pearson) correlation coefficient between these values.
5. What can you observe and why does it make sense?

solution

Spearman rank correlation

```
>>> x = randi(10, 100, 1);
>>> y = randi(10, 100, 1);
>>> corr(x,y,'type','Spearman')
ans =
    0.1220
>>> corr(x,y.^2,'type','Spearman')
ans =
    0.1220
>>> corr(x,y,'type','Pearson')
ans =
    0.1074
>>> corr(x,y.^2,'type','Pearson')
ans =
    0.0551
```

The rank correlation does not change under a monotone transformation of the data. Therefore, it can be used for ordinal data. The Pearson correlation coefficient does not have that property.

Day 1 – descriptive statistics and plots

What makes a good plot?

features of a good plot

A good plot

- helps the reader to clearly understand your point.

features of a good plot

A good plot

- helps the reader to clearly understand your point.
- is not misleading and let's the reader judge the information on her own (different y-axis/length scales in two related plots, "squeezing" via log-plots).

features of a good plot

A good plot

- helps the reader to clearly understand your point.
- is not misleading and let's the reader judge the information on her own (different y-axis/length scales in two related plots, "squeezing" via log-plots).
- contains information about the data (a comic might be illustrative, but does not contain information about the data).

features of a good plot

A good plot

- helps the reader to clearly understand your point.
- is not misleading and let's the reader judge the information on her own (different y-axis/length scales in two related plots, "squeezing" via log-plots).
- contains information about the data (a comic might be illustrative, but does not contain information about the data).
- adheres to the principle of ink minimization.

features of a good plot

design/organization

- Is the display consistent with the model or hypothesis being tested?

features of a good plot

design/organization

- Is the display consistent with the model or hypothesis being tested?
- Are there "empty dimensions" in the display that could be removed (A 3D pie chart for 2D categorical data, extraneous colors that do not encode meaningful information)?

features of a good plot

design/organization

- Is the display consistent with the model or hypothesis being tested?
- Are there "empty dimensions" in the display that could be removed (A 3D pie chart for 2D categorical data, extraneous colors that do not encode meaningful information)?
- Does the display provide an honest and transparent portrayal of the data (hiding, smoothing, modifying data points should be avoided or explicitly mentioned)?

Allen et al. 2012, Neuron

features of a good plot

axes

- Are axes scales defined as linear, log, or radial?

features of a good plot

axes

- Are axes scales defined as linear, log, or radial?
- Does each axis label describe the variable and its units (use "a.u." for arbitrary units)?

features of a good plot

axes

- Are axes scales defined as linear, log, or radial?
- Does each axis label describe the variable and its units (use "a.u." for arbitrary units)?
- Are axes limits appropriate for the data (The graphic should not be bounded at zero if the data can take on both positive and negative values.)?

features of a good plot

axes

- Are axes scales defined as linear, log, or radial?
- Does each axis label describe the variable and its units (use "a.u." for arbitrary units)?
- Are axes limits appropriate for the data (The graphic should not be bounded at zero if the data can take on both positive and negative values.)?
- Is the aspect ratio appropriate for the data (When x and y axes contrast the same variable under different conditions the graphic should be square.)?

Allen et al. 2012, Neuron

features of a good plot

color mapping

- Is a color bar provided?

features of a good plot

color mapping

- Is a color bar provided?
- Is the color map sensible for the data type (does the data extend to both $\pm$, does it live in an interval, is it circular)?

features of a good plot

color mapping

- Is a color bar provided?
- Is the color map sensible for the data type (does the data extend to both $\pm$, does it live in an interval, is it circular)?
- Are contrasting colors consistent with a natural interpretation?
- Can features be discriminated when printed in grayscale?
- Has red/green contrast been avoided to accommodate common forms of colorblindness?

Allen et al. 2012, Neuron

features of a good plot

uncertainty

- Does the display indicate the uncertainty of estimated parameters?

features of a good plot

uncertainty

- Does the display indicate the uncertainty of estimated parameters?
- Is the type of error surface appropriate for the data?
  - Use standard deviations to describe variability in the population.

features of a good plot

uncertainty

- Does the display indicate the uncertainty of estimated parameters?
- Is the type of error surface appropriate for the data?
    - Use standard deviations to describe variability in the population.
    - Use standard errors or confidence intervals to make inferences about parameters estimated from a sample.

features of a good plot

uncertainty

- Does the display indicate the uncertainty of estimated parameters?
- Is the type of error surface appropriate for the data?
  - Use standard deviations to describe variability in the population.
  - Use standard errors or confidence intervals to make inferences about parameters estimated from a sample.
  - Parametric confidence intervals should only be used if data meet the assumptions of the underlying model.

features of a good plot

uncertainty

- Does the display indicate the uncertainty of estimated parameters?
- Is the type of error surface appropriate for the data?
  - Use standard deviations to describe variability in the population.
  - Use standard errors or confidence intervals to make inferences about parameters estimated from a sample.
  - Parametric confidence intervals should only be used if data meet the assumptions of the underlying model.
- Are the units of uncertainty defined (is it standard error, is it 95% confidence interval)?

Allen et al. 2012, Neuron

features of a good plot

annotation

- Are all symbols defined, preferably by directly labeling objects?

features of a good plot

annotation

- Are all symbols defined, preferably by directly labeling objects?
- Is the directionality of a contrast between conditions obvious?

features of a good plot

annotation

- Are all symbols defined, preferably by directly labeling objects?
- Is the directionality of a contrast between conditions obvious?
- Is the number of samples or independent experiments indicated?

features of a good plot

annotation

- Are all symbols defined, preferably by directly labeling objects?
- Is the directionality of a contrast between conditions obvious?
- Is the number of samples or independent experiments indicated?
- Are statistical procedures and criteria for significance described?

features of a good plot

annotation

- Are all symbols defined, preferably by directly labeling objects?
- Is the directionality of a contrast between conditions obvious?
- Is the number of samples or independent experiments indicated?
- Are statistical procedures and criteria for significance described?
- Are uncommon abbreviations avoided or clearly defined?

features of a good plot

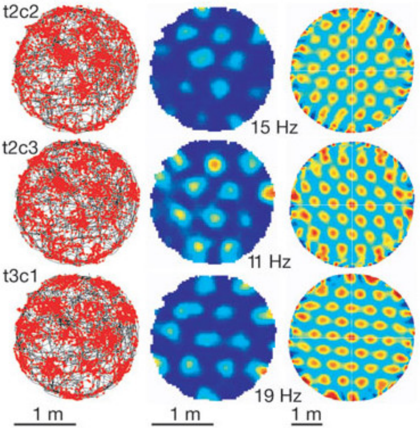
annotation

- Are all symbols defined, preferably by directly labeling objects?
- Is the directionality of a contrast between conditions obvious?
- Is the number of samples or independent experiments indicated?
- Are statistical procedures and criteria for significance described?
- Are uncommon abbreviations avoided or clearly defined?
- Are abbreviations consistent with those used in the text?

Allen et al. 2012, Neuron

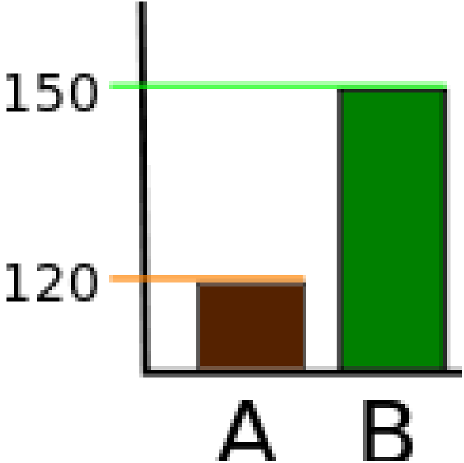
Day 1 – descriptive statistics and plots

suboptimal example



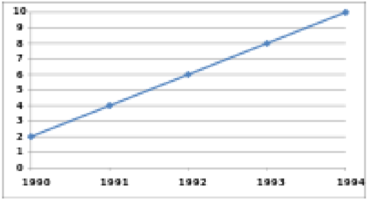
Hafting et al. 2005, nature

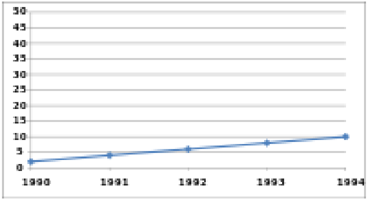
suboptimal example



http://en.wikipedia.org/wiki/Misleading_graph

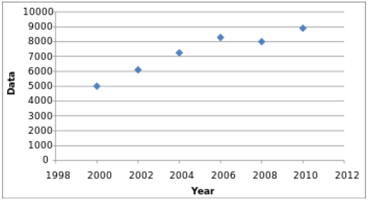
suboptimal example

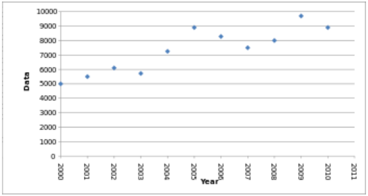




http://en.wikipedia.org/wiki/Misleading_graph

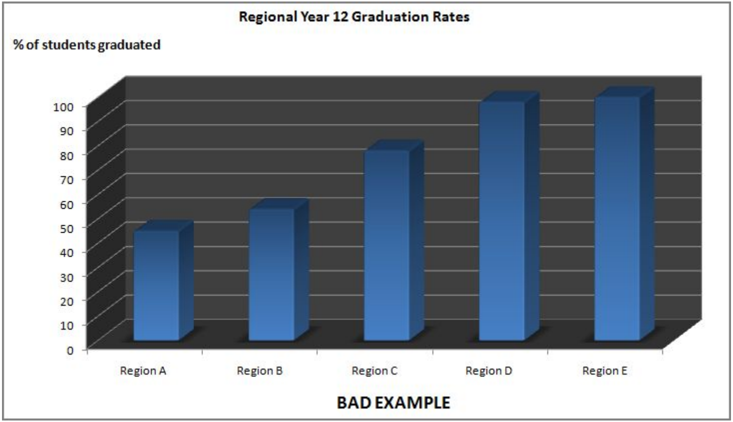
suboptimal example





http://en.wikipedia.org/wiki/Misleading_graph

suboptimal example

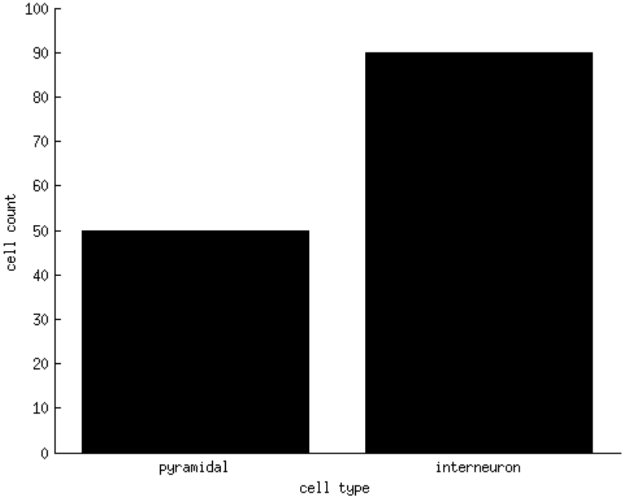


www.enfovis.com

Day 1 – descriptive statistics and plots

plotting nominal data

bar plot for count and relative frequency



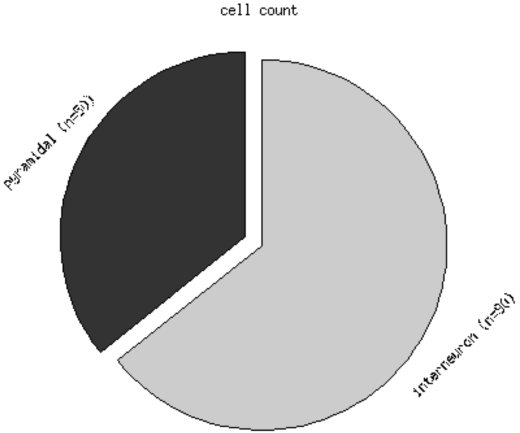
plotting nominal data

bar plot for count and relative frequency

```
% plot
bar([1,2], [50, 90], 'facecolor', 'k')

% labels axes
ylabel('cell count')
xlabel('cell type')

% cosmetics
xlim([0.5,2.5])
ylim([0, 100])
box('off')
set(gca,'XTick',1:2,'XTickLabel',{'pyramidal','interneuron'},'FontSize',20)

% settings for saving the figure
set(gcf, 'PaperUnits', 'centimeters');
set(gcf, 'PaperSize', [11.7 9.0]);
set(gcf, 'PaperPosition',[0.0 0.0 11.7 9.0]);
```

plotting nominal data

pie chart for count and relative frequency



plotting nominal data

exercise

pie chart

Plot the same data ($n_{py} = 50$, $n_{in} = 90$) as a pie chart in Matlab.

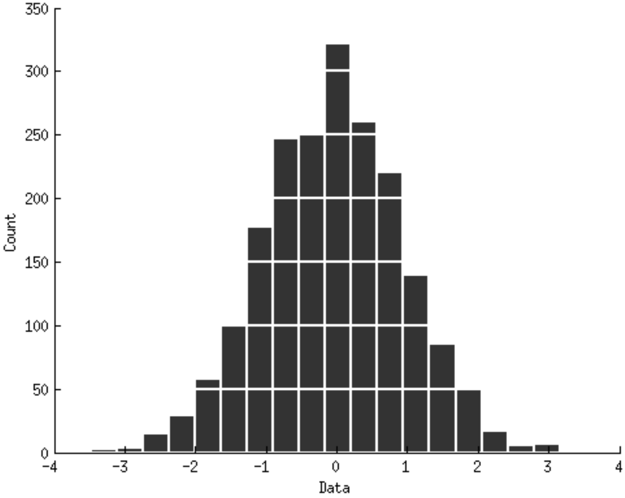
plotting nominal data

pie chart for relative frequency

```
data = [50, 90];
h = pie(data, [1,0], {'pyramidal (n=50)', 'interneuron (n=90)'})
hText = findobj(h,'Type','text') % text object handles

set(h(1), 'FaceColor', [.2,.2,.2]);
set(h(2), 'Rotation', 45);
set(h(3), 'FaceColor', [.8,.8,.8]);
set(h(4), 'Rotation', 45);

title('cell count')
set(gca,'XTick',1:2,'XTickLabel',{'pyramidal', 'interneuron'})
box('off')
set(gcf, 'PaperUnits', 'centimeters');
set(gcf, 'PaperSize', [11.7 9.0]);
set(gcf, 'PaperPosition',[0.0 0.0 11.7 9.0]);
```

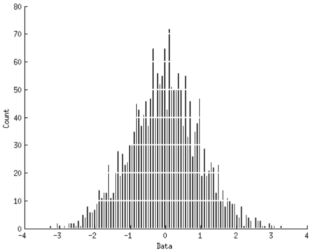
plotting interval/ratio/absolute data

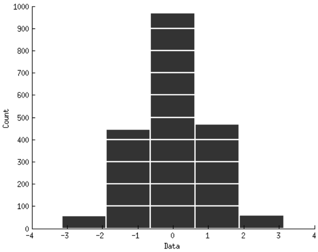
histogram



plotting interval/ratio/absolute data

bad choice of bins





Rule of thumb

Choose the bins $b \approx n/20$.

plotting interval/ratio/absolute data

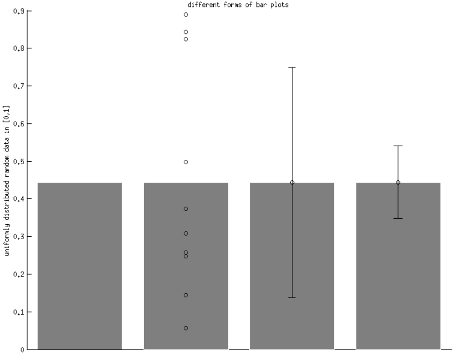
how to do in Matlab

```
x = randn(2000,1); % generate Gaussian data

hist(x, 50); % generate histogram

% set facecolor to gray
h = findobj(gca, 'Type','patch');
set(h(1), 'FaceColor',[.2,.2,.2], 'EdgeColor','w', 'linewidth',2)

% plot a white grid over it
h = gridxy([],get(gca,'ytick'),'color','w','linewidth',2)
uistack(h, 'top')

% cosmetics
box('off');
xlabel('Data')
ylabel('Count')
```

plotting interval/ratio/absolute data

bar plot

There are several ways to plot a sample $x_1, \ldots, x_n$ of interval/ratio/absolute scale with a bar plot



plotting interval/ratio/absolute data

bar plot

```
% bar plot
x = rand(10,1);
gray = [.5,.5,.5];

bar(1, mean(x), 'EdgeColor','w','FaceColor', gray);
hold on

bar(2, mean(x), 'EdgeColor','w','FaceColor', gray);
plot(0*x + 2, x, 'ok');

bar(3, mean(x), 'EdgeColor','w','FaceColor', gray);
errorbar(3, mean(x), std(x), 'ok');

bar(4, mean(x), 'EdgeColor','w','FaceColor', gray);
errorbar(4, mean(x), std(x)/sqrt(length(x)), 'ok');
set(gca, 'xtick',[])
ylabel('uniformly distributed random data in [0,1]')
box('off')
title('different forms of bar plots')
hold off
```

plotting interval/ratio/absolute data

bar plot and measure of central tendency and spread

- A bar plot collapses real data onto a single number and some measure of spread. This number is usually a measure of central tendency, i.e. a typical/central value for the probability distribution of the data.

plotting interval/ratio/absolute data

bar plot and measure of central tendency and spread

- A bar plot collapses real data onto a single number and some measure of spread. This number is usually a measure of central tendency, i.e. a typical/central value for the probability distribution of the data.
- What measures of central tendency can you think of?

plotting interval/ratio/absolute data

bar plot and measure of central tendency and spread

- A bar plot collapses real data onto a single number and some measure of spread. This number is usually a measure of central tendency, i.e. a typical/central value for the probability distribution of the data.
- What measures of central tendency can you think of?
  - mean
  - median
  - geometric mean (the nth root of the product of the data values)
  - weighted mean
  - midrange (mean of the maximum and minimum values of a data set)

plotting interval/ratio/absolute data

bar plot and measure of central tendency and spread

- A bar plot collapses real data onto a single number and some measure of spread. This number is usually a measure of central tendency, i.e. a typical/central value for the probability distribution of the data.
- What measures of central tendency can you think of?
  - mean
  - median
  - geometric mean (the nth root of the product of the data values)
  - weighted mean
  - midrange (mean of the maximum and minimum values of a data set)
- Additionally, the bar plot is equipped with a measure of spread or dispersion. What measure of spread can you think of?

plotting interval/ratio/absolute data

bar plot and measure of central tendency and spread

- A bar plot collapses real data onto a single number and some measure of spread. This number is usually a measure of central tendency, i.e. a typical/central value for the probability distribution of the data.
- What measures of central tendency can you think of?
  - mean
  - median
  - geometric mean (the nth root of the product of the data values)
  - weighted mean
  - midrange (mean of the maximum and minimum values of a data set)
- Additionally, the bar plot is equipped with a measure of spread or dispersion. What measure of spread can you think of?
  - standard deviation
  - range (maximum minus minimum of a dataset)
  - inter-quartile range

plotting interval/ratio/absolute data

measure of central tendency and spread

The part of statistics that summarizes data in a small number of values is called descriptive statistics.

robust statistics

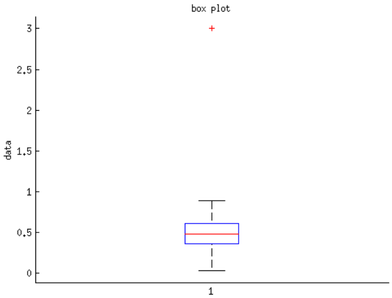
When is statistic called robust (leave-one-out)?

- Generate an array with 20 random numbers using `randn`.
- Compute 20 means: the i^{th} mean is computed from the data set without the i^{th} example.
- Repeat this with the median.
- Make a bar plot that depicts the means of the computed means and medians along with an appropriate measure of dispersion.
- What can you observe? Do you understand why?

plotting interval/ratio/absolute data

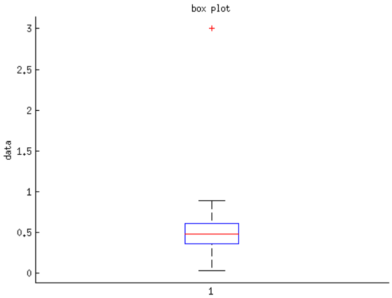
boxplot

Who knows what the elements mean?



plotting interval/ratio/absolute data

boxplot



Who knows what the elements mean?

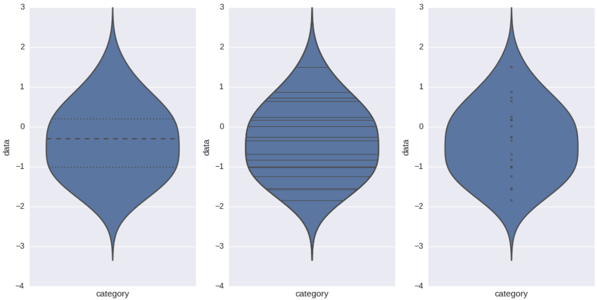
- the box depicts the inter-quartile range
- the line denotes the median
- the whiskers denote the extreme value of the data not considered outliers
- outliers are plotted separately

Outliers

- Find out how an outlier is defined in a matlab boxplot.
- Can you remove an outlier from the dataset?

plotting interval/ratio/absolute data

violinplot



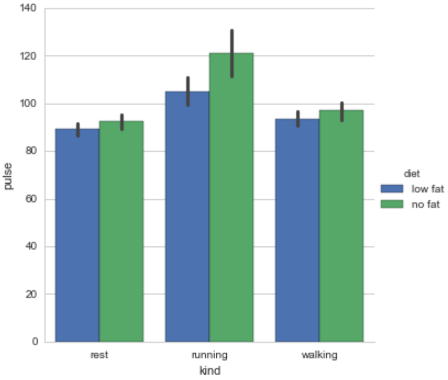
- Violinplots depict the distribution of the data by a smoothed histogram.
- Additional information (data points, median, inter-quartile range) are plotted inside.

plotting combinations of scales

What could we use for a combination of categorial/nominal and interval/ratio/absolute?

plotting combinations of scales

What could we use for a combination of categorial/nominal and interval/ratio/absolute?



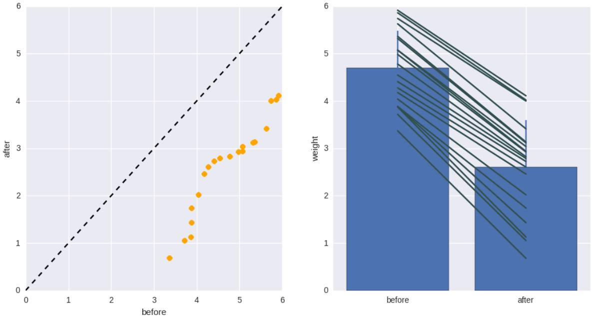
Each category is a single bar.

plotting combinations of scales

What could we use for a combination of interval/ratio/absolute and interval/ratio/absolute, e.g. $(x_1, y_1), \ldots, (x_n, y_n)$?

plotting combinations of scales

What could we use for a combination of interval/ratio/absolute and interval/ratio/absolute, e.g. $(x_1, y_1), \ldots, (x_n, y_n)$?



Scatter plot or paired bar chart. Scatter plot can also be used for ordinal vs. ordinal data (why not the bar chart?).

That's it.