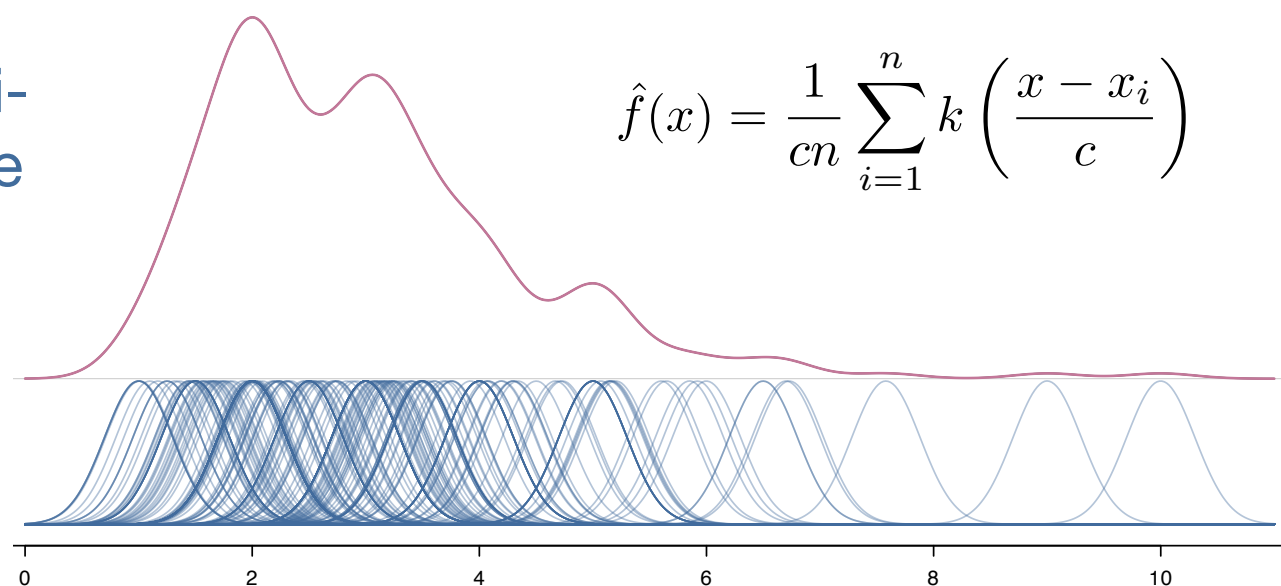


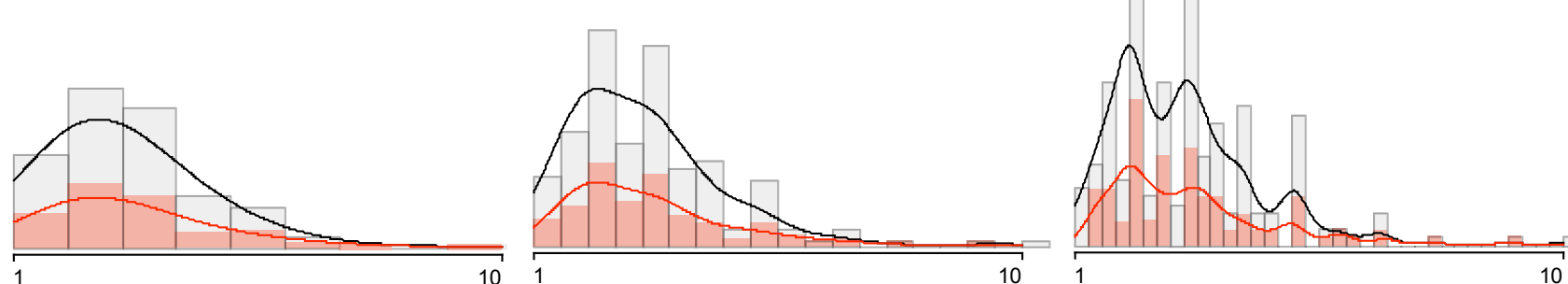
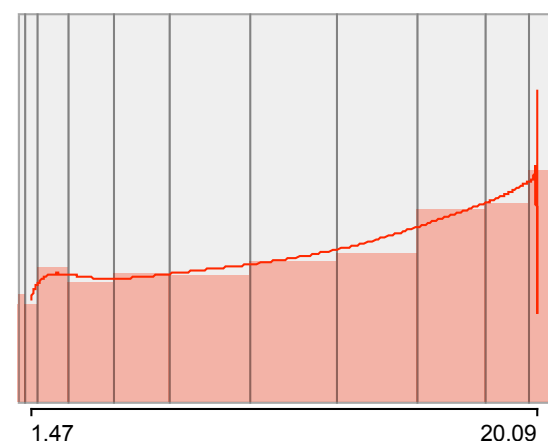
## Density Estimators

- Density estimators try to approximate the continuous distribution function of a sample as close as possible. They are computationally far more intensive than histograms
- In that respect they are far superior to histograms, but fail to identify gaps and accumulations as (interactive) histograms can do
- The idea of a kernel density estimator is to sum up the contribution of each single point to the overall distribution
- The interactive control of the bandwidth is crucial



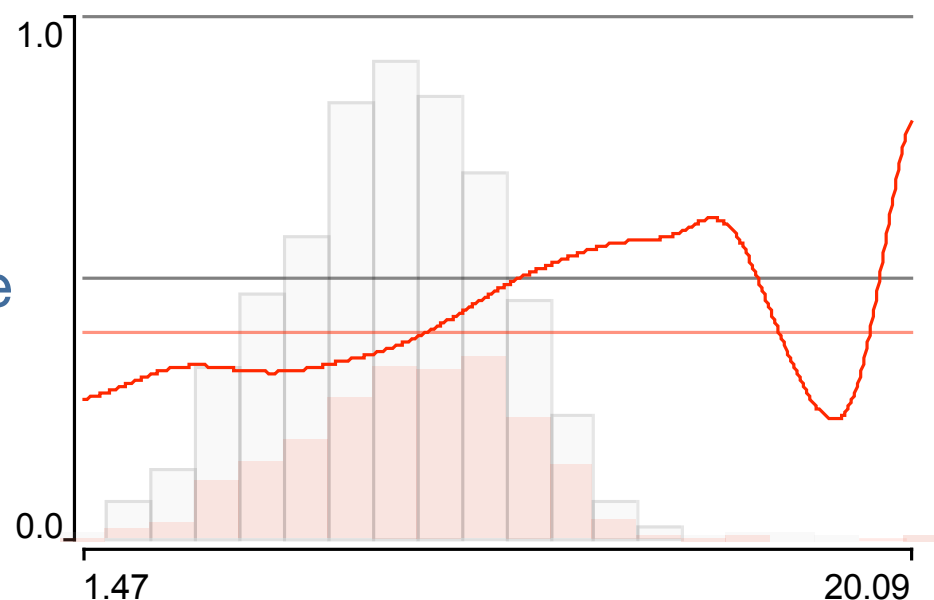
## Densities in Histograms and Spinograms

- We can benefit from both plots at a time, when we superpose densities onto histograms.
- Control of the bin width (histogram) and bandwidth (kernel density estimator) should go in line.
- Switching to a spinogram, we get a continuous approximation of the conditional density
- But, we need to keep in mind that the x-axis is still non-linear and thus harder to interpret



## CD-Plot

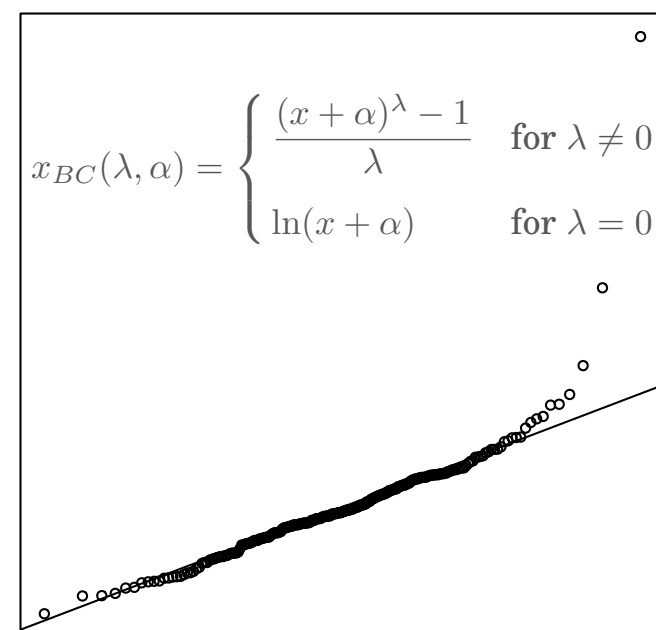
- The CD-plot (Conditional Distribution plot) overcomes the problem of the non-linear x-axis.
- For any point on the x-axis, it shows the proportion highlighted, i.e., the conditional distribution of this selected subset
- Thus, it can be plotted along with the standard histogram
- Although the linear x-axis is easier to interpret, it also has a risk
- Whereas intervals with fewer points and thus higher variance are squished together in a spinogram, we need to keep the variance of the CD-plot estimate in mind



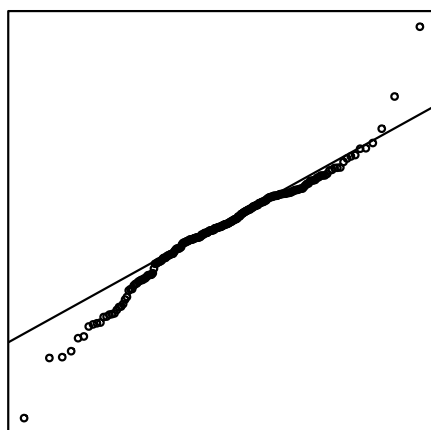
## Transforming Data: Continuous Data

- Many statistical procedures assume normality or at least work better if data are not too far from being gaussian
- Transformations can help, but are also questionable as long as we cannot give a content related justification
- Sometimes we only need to remove or explain outliers

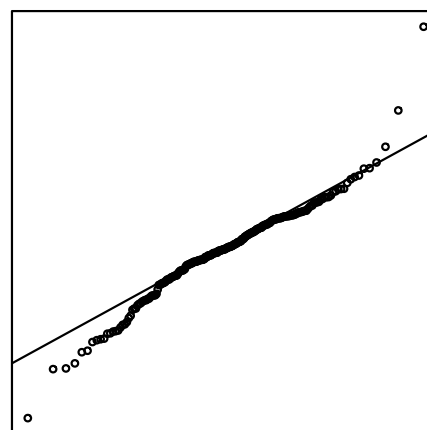
to make a  
distribution  
“normal”



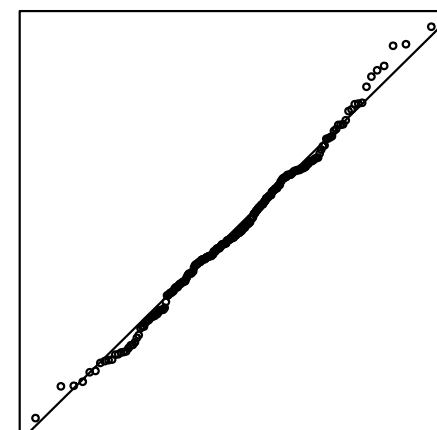
Log Transformation



Box-Cox Transformation (lambda=0.15)

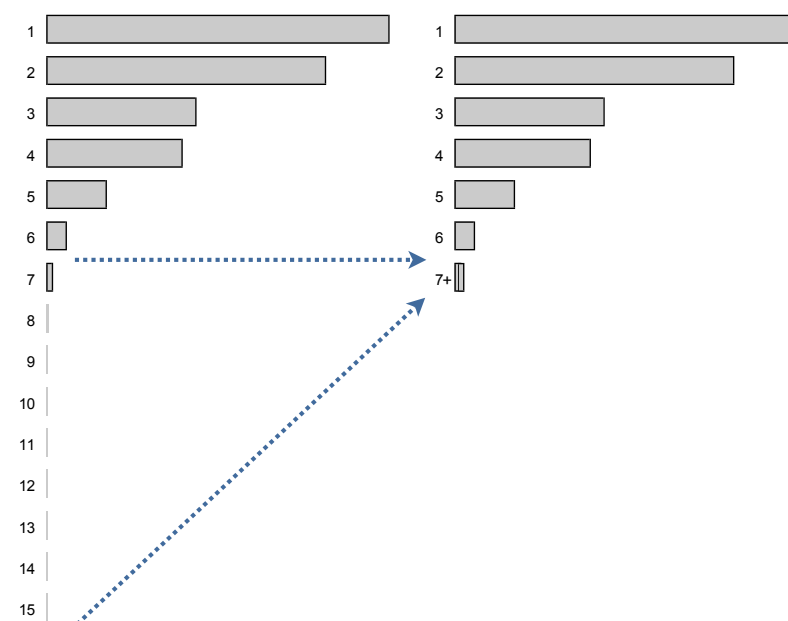
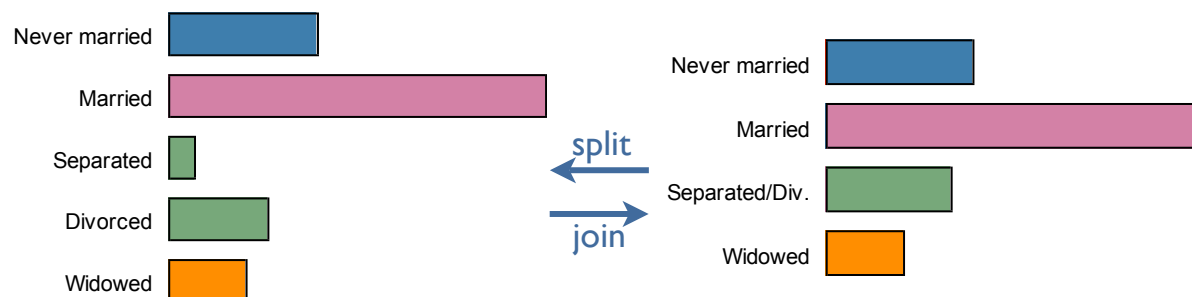


Outlier removed



## Transforming Data: Categorical Data

- Transformations are not restricted to continuous data and can be necessary for categorical data as well
- With very large datasets we often have the problem of having a lot of very small categories, which can be joined to one group
- In other cases it might be sensible to join or split categories in order to reflect their influence on other variables more properly



## Weighted Data/Plots: Simple

- Especially with categorical data it is usually much more efficient to store data only in aggregated form
- Statistical graphic tools should be able to handle such data
- If data is of different importance this can be expressed by sample weights – no matter if the data is continuous or categorical
- For area based plots the generalization for weighted data is usually straight forward

Class	Age	Gender	Survived
First	Adult	Male	Yes
First	Adult	Male	Yes
First	Adult	Male	Yes
First	Adult	Male	Yes
First	Adult	Male	Yes
First	Adult	Male	Yes

...

Class	Age	Gender	Survived	Count
First	Adult	Female	No	4
First	Adult	Female	Yes	140
First	Adult	Male	No	118
First	Adult	Male	Yes	57
First	Child	Female	Yes	1
First	Child	Female	No	0

...

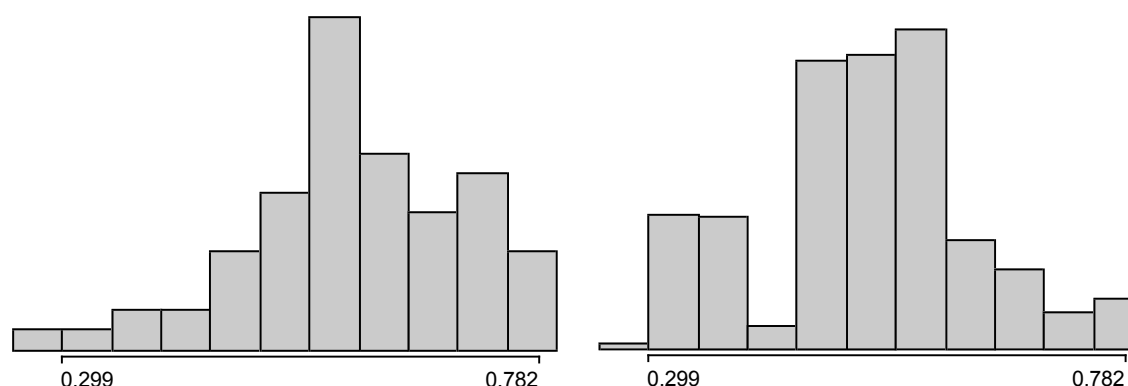
## Weighted Plots: Advanced

- **Example:** Florida Election (Case Study I)

We look at the percentage votes for G.W. Bush in the 65 counties in Florida. In the left plot, each county gets the same weight, i.e., we look at the distribution of counties, in the right plot, each county is weighted by its number of voters, we look at the distribution of voters

- **Interpretation:**

Bush's support was stronger in the less populated counties since high percentages are downweighted and low percentages are upweighted



Each county gets the same weight

... is weighted by the number of voters