Robust topics

- Median
- MAD
- Spearman
- Wilcoxon rank test
- Weighted least squares
- Cook's distance
- M-estimators

Robust topics

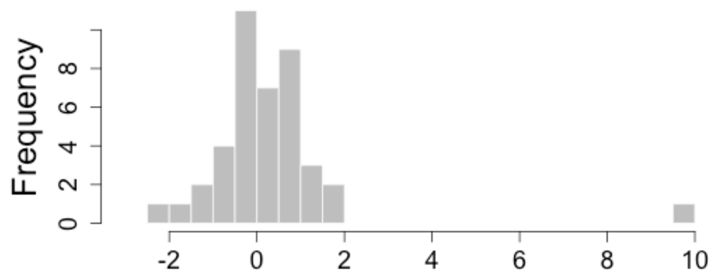
- Median => middle
- MAD => spread
- Spearman => association
- Wilcoxon rank test => group diffs
- Weighted least squares
- Cook's distance => observation influence
- M-estimators => framework for estimation

What do we mean by robust?

- “robust to outliers”
- “robust to misspecification of the model”

- Low variance (“precise”), low bias (“accurate”)
- Accuracy (TP+TN/total), precision (1-FDR), sensitivity (TPR), specificity (1-FPR)

What do we mean by outlier?

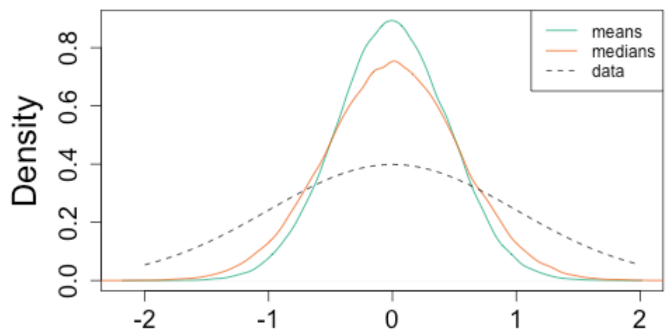


- Technical error?
- Data entry error?
- Unaccounted for tail of data distribution?

How do most statistics work

Median

```
dat <- matrix(rnorm(5*1e5),ncol=5)
means <- rowMeans(dat)
medians <- apply(dat,1,median)
```



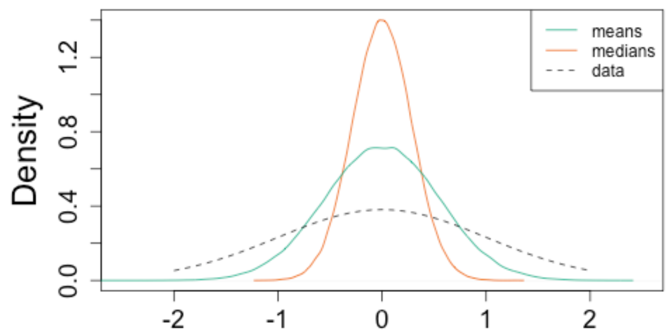
Efficiency

```
sd(medians)/sd(means)
```

```
[1] 1.197183
```

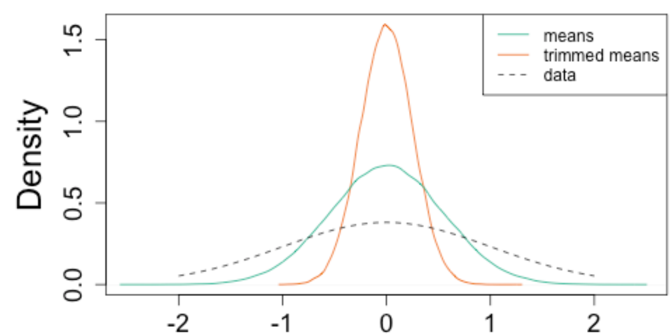
Median for non-normal data

```
dat <- cbind(matrix(rnorm(19*1e5,sd=1),ncol=19),
             matrix(rnorm(1*1e5,sd=10),ncol=1))
means <- rowMeans(dat)
medians <- apply(dat,1,median)
```

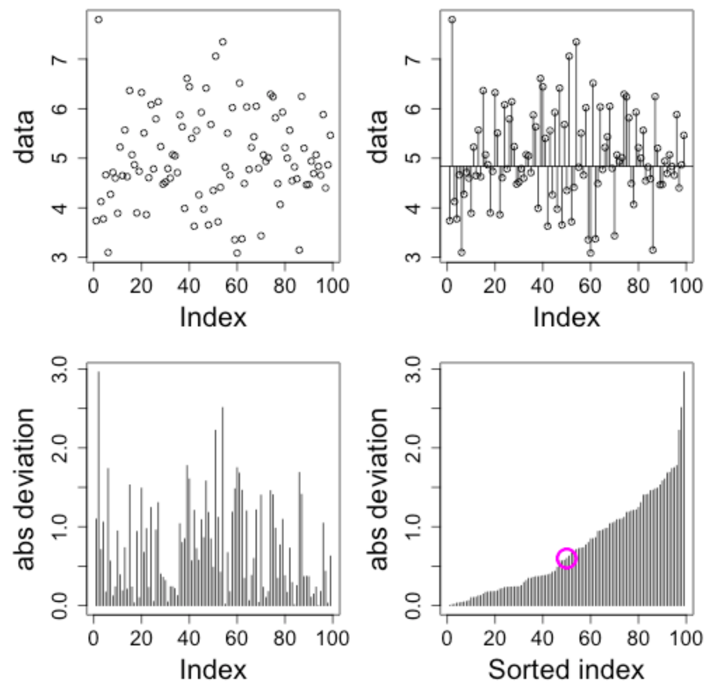


Trimmed mean for non-normal data

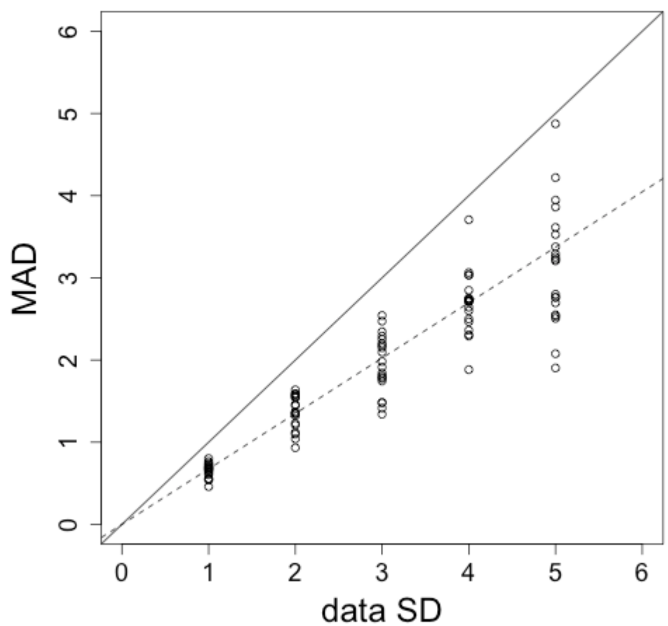
```
dat <- cbind(matrix(rnorm(19*1e5,sd=1),ncol=19),
             matrix(rnorm(1*1e5,sd=10),ncol=1))
means <- rowMeans(dat)
# trim 5% from each end = 10% of data
tmeans <- apply(dat,1,mean,trim=.05)
```



MAD: median absolute deviation

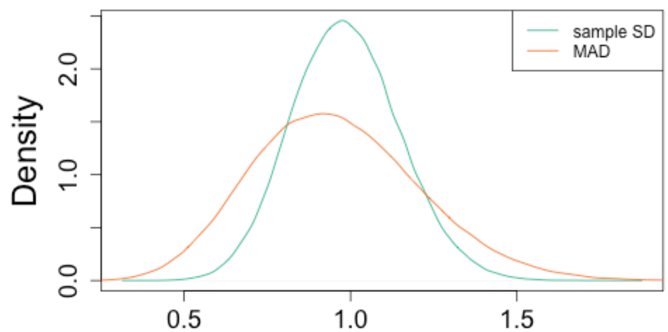


Median absolute deviation & SD



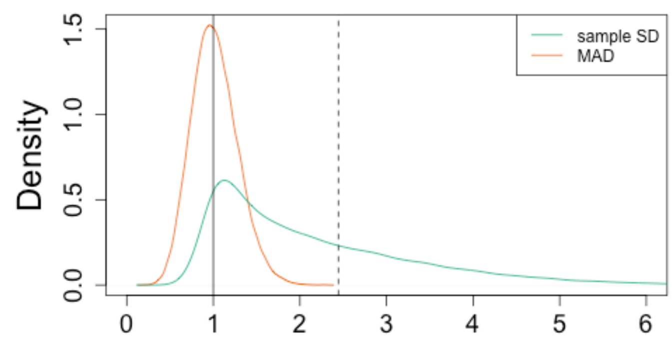
Efficiency of MAD

```
dat <- matrix(rnorm(20*1e5),ncol=20)
sds <- apply(dat,1,sd)
mads <- apply(dat,1,mad)
```



MAD with outliers

```
dat <- cbind(matrix(rnorm(19*1e5,sd=1),ncol=19),
             matrix(rnorm(1*1e5,sd=10),ncol=1))
sds <- apply(dat,1,sd)
mads <- apply(dat,1,mad)
```

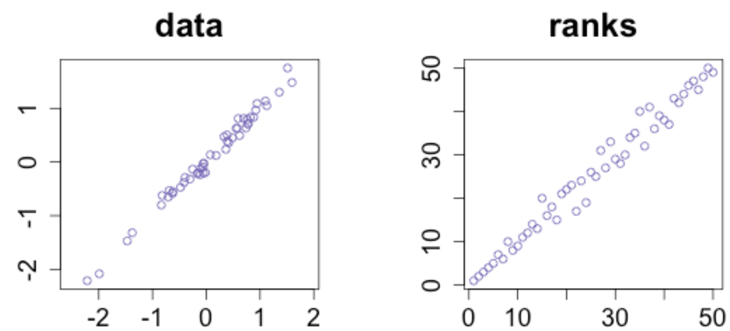


Spearman correlation

+ less sensitive to outliers

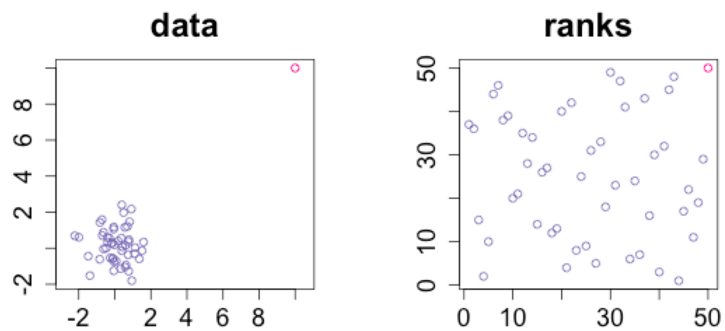
- all subregions of range count equally

Spearman correlation




```
  pearson  spearman
0.9932279 0.9879952
```

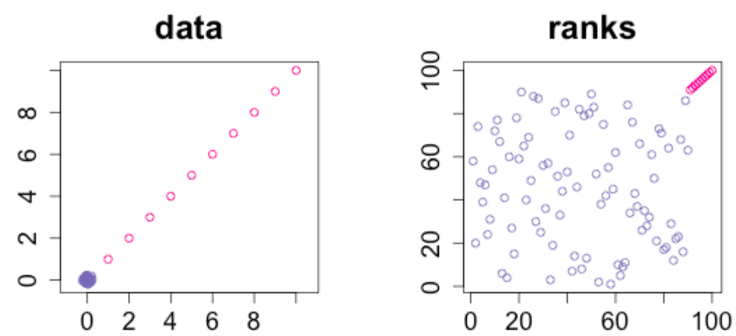
Spearman correlation




```
   pearson    spearman
0.69943398 -0.03404562
```

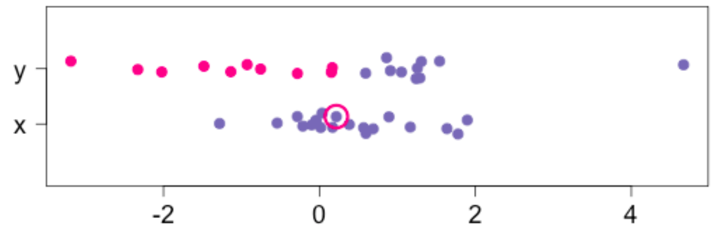
Spearman correlation

- Drug resistance in cell lines
- Gene expression




```
  pearson  spearman
0.9977304 0.1939154
```

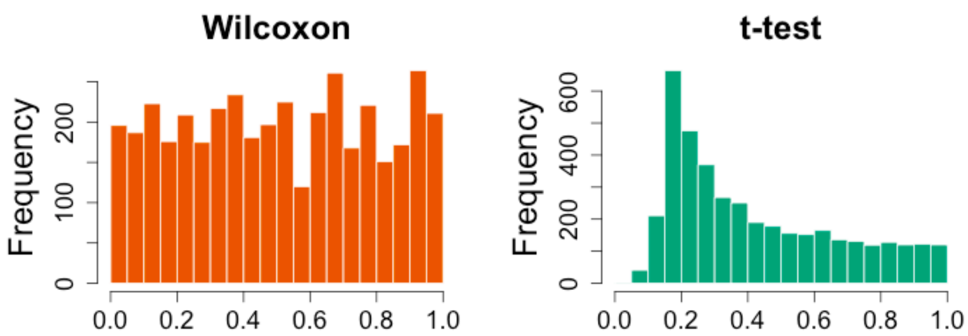
Wilcoxon / Mann-Whitney rank test



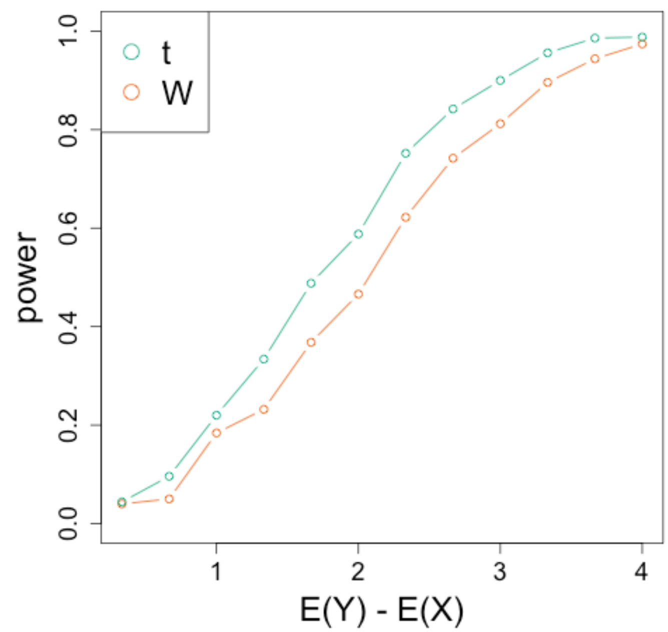

```r
W <- 0
for (i in seq_along(x)) {
  W <- W + sum(y <= x[i])
}
print(W)
```

```
[1] 211
```

Wilcoxon vs t-test p distribution (n=20)



Wilcoxon vs t-test sensitivity (SD=1, n=4)

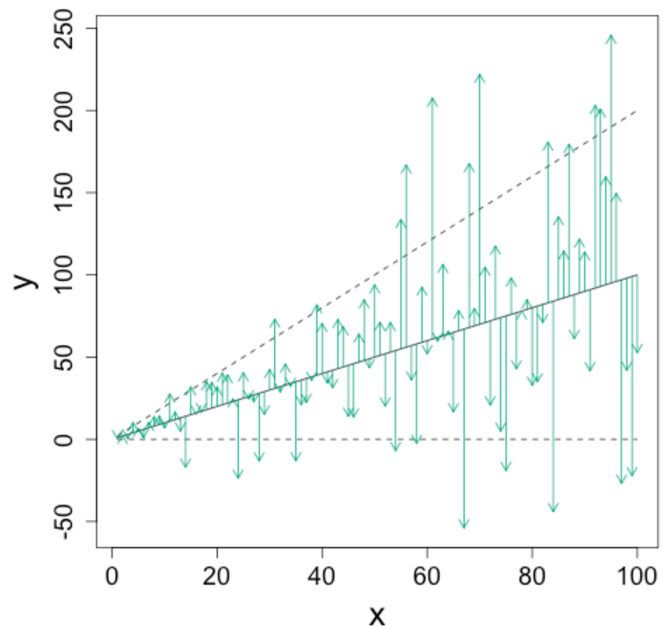


Wilcoxon for small sample size

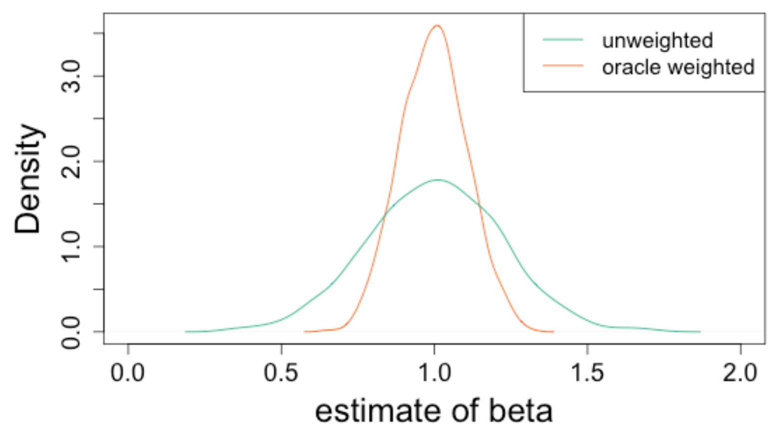
```
wilcox.test(x=101:103, y=1:3)
```

```

	Wilcoxon rank sum test

data:  101:103 and 1:3
W = 9, p-value = 0.1
alternative hypothesis: true location shift is
not equal to 0
```

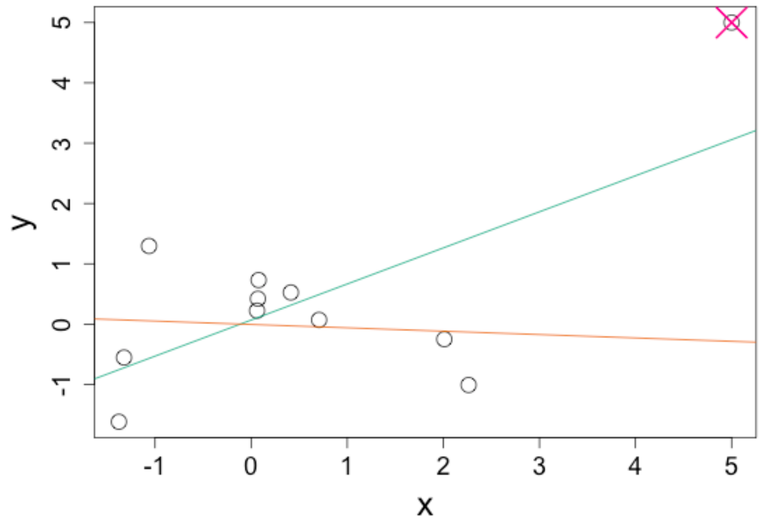
Weighted least squares



Weighted least squares



Cook's distance




```
                   fit1         fit2
(Intercept) 0.06755603 -0.006076753
x           0.59821827 -0.056409684
```

Cook's distance

```
dfbeta(fit1)[1,"x"]
```

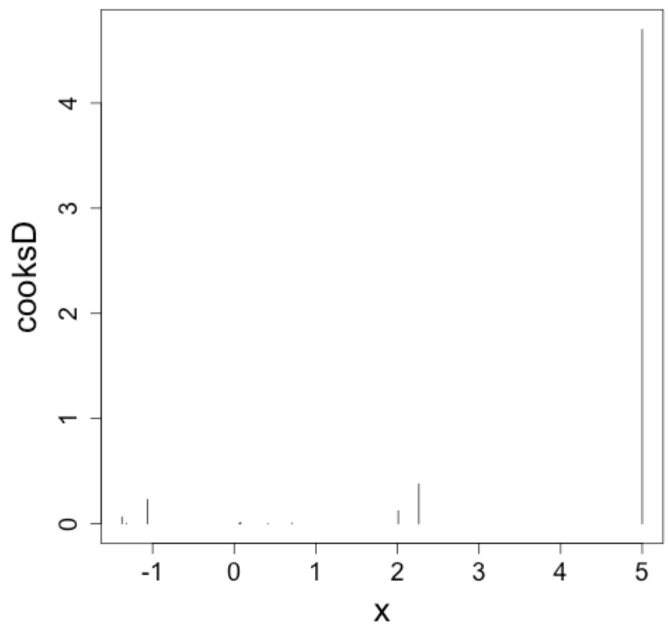
```
[1] 0.654628
```

```
coef(fit1)[2] - coef(fit2)[2]
```

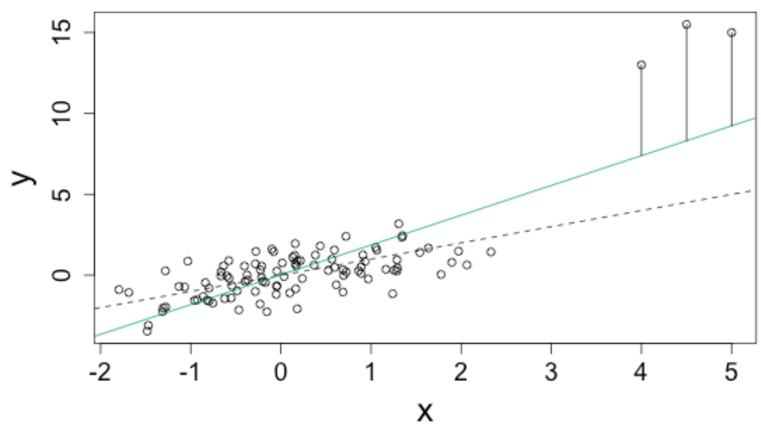
```
       x
0.654628
```

Cook's distance

```
cooksD <- cooks.distance(fit1)
```



M-estimators

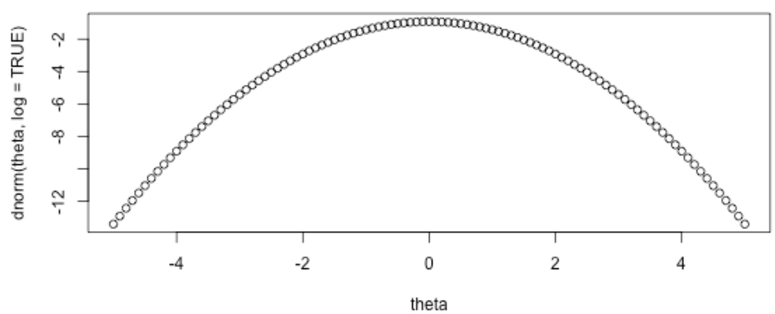


M-estimators

- M-estimators are a generalized framework for estimation
- M for *Maximum likelihood-type* estimation
- Least squares is a maximum likelihood estimate for data with normally-distributed error.

MLE reminder

```
theta <- seq(-5,5,.1)
plot(theta, dnorm(theta,log=TRUE))
```



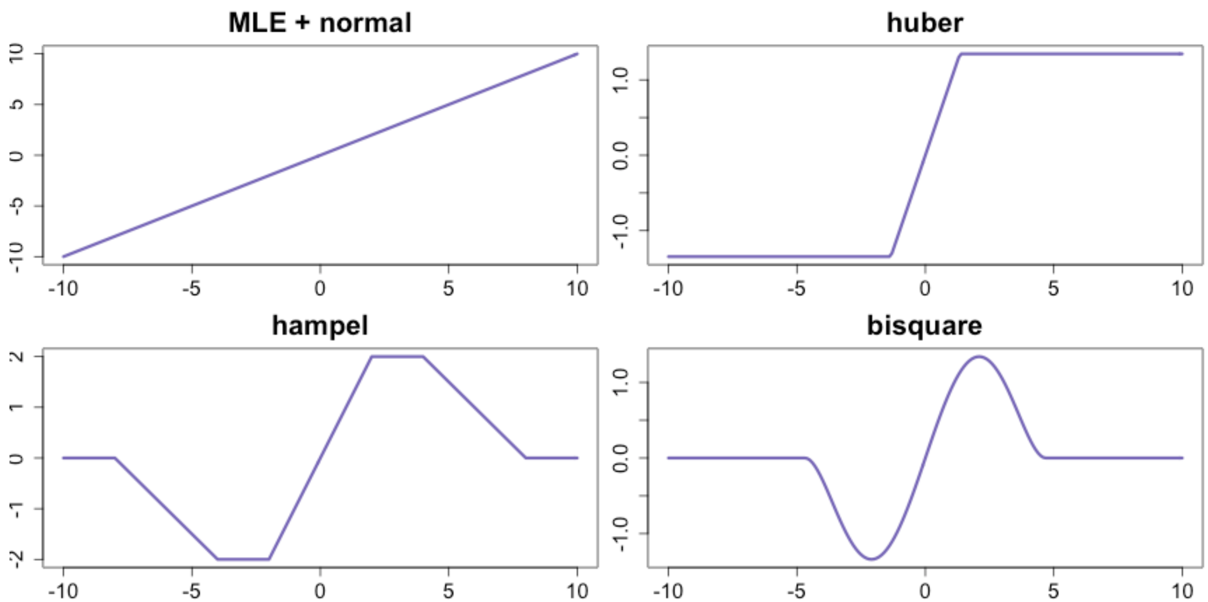
Theory of estimation

It is interesting to look back to the very origin of the theory of estimation, namely to Gauss and his theory of least squares. Gauss was fully aware that his main reason for assuming an underlying normal distribution and a quadratic loss function was mathematical, i.e., **computational, convenience**. In later times, this was often forgotten, partly because of the central limit theorem.

Theory of estimation

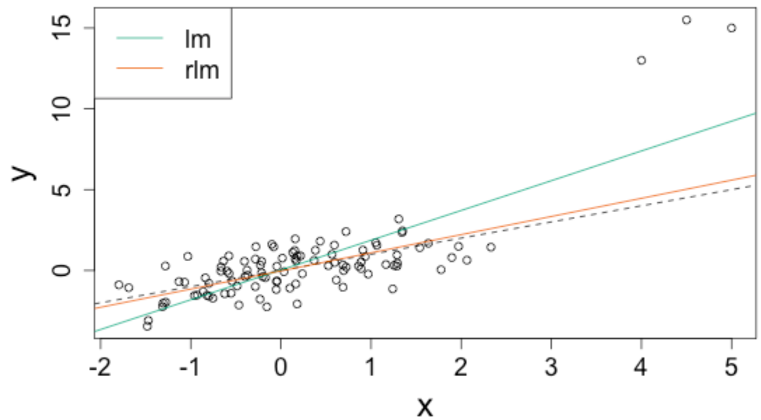
However, if one wants to be honest, the central limit theorem can at most explain why many distributions occurring in practice are approximately normal. The stress is on the word "approximately." This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): **What happens if the true distribution deviates slightly from the assumed normal one**?

M-estimators



M-estimators

```
library(MASS)
rob.fit <- rlm(y ~ x)
```



Links on robust statistics in genomics

- **SAMseq**'s implementation of rank test
  - sequencing depth
  - noise of low counts
  - false discovery rate
- **voom** weighted linear model
- **edgeR** and **limma-voom** sample quality weights
- **DESeq2** use of Cook's distance