Exploratory Data Analysis

Exploratory vs Confirmatory, Stem and Leaf Plots, Histograms and Non-Parametric Kernel Density Estimators

David B King, Ph.D.

August 26, 2015

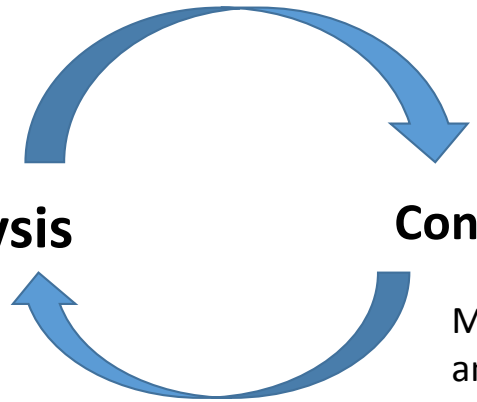
Texts used in this course are

- D. Hoaglin, F. Mosteller, J. Tukey, *Understanding Robust and Exploratory Data Analysis (UREDA)*, Wiley, New York, 1983.
- D. Hoaglin, F. Mosteller, J. Tukey, *Exploring Data Tables, Trends, and Shapes (EDTTS)*, Wiley, New York, 1985.
- W. Cleveland, *Visualizing Data (VD)*, Hobart Press, New Jersey, 1993.
- L. Wilcinson, *The Grammer of Graphics: Second Edition (GG)*, Springer, New York, 2005.

Exploratory vs Confirmatory

Exploratory vs Confirmatory Data Analysis

Broad Phases of Data Analysis



Experimental Data Analysis

Procedures remain flexible. Kids playing in the sandbox - no pressure to give final conclusions.

"Isolates features of the data and reveals those to the analyst"

Comes before any firm choice of models.

Uncovers unexpected departures from the familiar models.

Confirmatory Data Analysis

Methods here address reproducibility and prediction quality of results.

Hypothesis Testing

Confidence Intervals

Give statements of significance under a specific model.

Is your model choice validated by collecting new, heretofore unseen evidence?

Exploratory vs Confirmatory

- EDA: tools for exploring, investigating data
- CDA: tools for validating hypothesis
- Data Analysis: back-and-forth between the two approaches:
  1. Formulate 1-2 specific questions or hypotheses (scientist)
  2. Conduct appropriate hypothesis test, p-value, model-based uncertainty (statistician: CDA)
  3. Explore data, note unexpected patterns (statistician: EDA)
  4. Design future investigation to confirm patterns found through EDA (statistician + scientist communication)

Reasons for Exploratory Data Analysis (EDA)

- detection of mistakes
- checking model assumptions
- preliminary selection of appropriate models
- determining relationships among the explanatory variables, and
- assessing the direction and rough size of relationships between explanatory and outcome variables.

Five R's to EDA

Resistance: is the insensitivity of a procedure to localized "misbehavior" of the data. A **resistant** method produces results that change only slightly when a small part of the data is replaced with new numbers. Resistant methods pay attention to the main body of the data and little to outliers.
Example the median is resistant statistic whereas the mean is not.

$$x = c(1, 2, 3, 4, 5); median(x) = 3; mean(x) = 3$$

If we change $x_5 = 5 \longrightarrow x_5 = 100$ the median is still 3 whereas mean changes a lot.

- "large" change in "small" fraction of data ($5 \rightarrow 100$).
- "small" change in "large" fraction of data (e.g. rounding all numbers).
- robustness: insensitivity to departures from assumptions surrounding an underlying probabilistic model.

Five R's to EDA

Residuals: are departures from fitted model (or fit), where residual = data − fit

- properly analyzed residuals can warn curvature, nonadditivity and nonconstancy of variability.
- examine residuals for additional structures — if so, fit the structure

Five R's to EDA

Re-expression (or Transformation) Two philosophies for fitting data:

1. data $= g$(fit, residual)
   g is a nonlinear combination of fit and residual; residual has some distribution; model parameters are fit via nonlinear least squares or computational maximum likelihood.
2. transform(data) = simple fit + residual
   transform data so the fitting process is simpler, e.g., transformed data are approximate Gaussian, linear additive, ...

EDA tends to choose (b) as we know much more about

- Effects of departures from normality
- Methods that are resistant to Gaussian assumption
- Methods to diagnose nonlinearity
- Methods to detect outliers ("1-10 % errors in data")
- Methods for characterizing the uncertainty in estimates obtained on transformed data (propagation of uncertainty formulae, jackknife, bootstrap)

Five R's to EDA

Revelation: reveal/display data, fit, diagnostic measures, residuals (or patterns, trends, outliers)

- varied tools (graphs, displays, pictures)
- The greatest value of a picture is when it forces us to notice what we never expected to see

Five R's to EDA

Re-iteration:

- Sequential fitting:

$$\begin{aligned} \text{data} &= \text{fit}_1 + \text{rough}_1 \\ \text{rough}_1 &= \text{fit}_2 + \text{rough}_2 \\ \text{rough}_2 &= \text{fit}_3 + \text{rough}_3 \\ \cdots &= \cdots + \cdots \\ \text{data} &= \text{fit}_1 + \text{fit}_2 + \cdots + \text{final rough} \end{aligned}$$

- Many resistant procedures require iteration (resistance line (ch5), median polish (ch6), biweight M-estimators of location in ch11).
- Computers are good at repetition.

Nonparametric vs Parametric

Nonparametric Statistics is a branch of statistics whose methods are not dependent on parametric distributional assumptions most commonly normality.

Advantage of making distributional assumptions (like normality):

Increased Power (Type II error rate is smaller)
Increased Significance Level (Lower Type I error rate)
Smaller Sample Size Requirements

Disadvantages of making parametric distributional assumptions:

Results depend on adherence to assumptions
Results may not be **ROBUST to violations of assumptions**

Nonparametric Statistics:
Few Modeling Assumptions

Parametric Statistics:
More Modeling Assumptions

Less Power
Wider Confidence Intervals
Less Significance Level

Increased Robustness

Greater Power
Smaller Confidence Intervals
Increased Significance Level

Less Robustness

Confirmatory Data Analysis

- Formulate desired model before seeing data
- Analyze the data
- Assess "significance" based on desired model

Dangers of this approach:

- What if assumptions of model don't fit the data?
- Shoe horn the data into model rather than letting the data suggest the model.

Advantages of this approach:

- Well defined hypotheses can be readily tested under model
- Approach avoids data dredging, where hypotheses are formulated post-hoc and significant results found without regard to multiplicity.

Example

Take two pairs of outcomes X and Y measured on the same set of subjects. Let us say we gave a pre-test and a post-test to 15 students

Type this into R:

```
pre=c(17,26,16,28,23,35,41,18,30,29,45,8,38,31,36)
post=c(21,26,19,26,30,40,43,15,29,31,46,7,43,31,37)
student=1:15
diff=post-pre
sign=(diff>0)*1+(diff<0)*-1
Dat=cbind(student=student,pre=pre,post=post,diff=diff,sign=sign)
Dat # look at the data
```

Non-parametric Example

The Sign Test

Want to know if post score is better than pre score.

H0: $\tilde{\mu}_X = \tilde{\mu}_Y \Rightarrow \tilde{\mu}_{X-Y} = 0$ Medians are same

HA: $\tilde{\mu}_X > \tilde{\mu}_Y \Rightarrow \tilde{\mu}_{X-Y} > 0$ Median for X > Median for Y

IDEA: Regardless of the actual distributions of X and Y, the number of times, C, that X >Y is Bin(n , ½) under H0.

C = # of times that X > Y.

Under H0:
C ~ Bin(n, ½)

In data set C=9, n = 15 so exact p-value is sum of exact binomial probabilities that are more extreme than the one observed, where more extreme is in the direction of the alternative

```
> binom.test(9,15)
> binom.test(9,15,alternative="greater")
```

student	pre	post	diff	sign
[1,]	1	17	21	4	1
[2,]	2	26	26	0	0
[3,]	3	16	19	3	1
[4,]	4	28	26	-2	-1
[5,]	5	23	30	7	1
[6,]	6	35	40	5	1
[7,]	7	41	43	2	1
[8,]	8	18	15	-3	-1
[9,]	9	30	29	-1	-1
[10,]	10	29	31	2	1
[11,]	11	45	46	1	1
[12,]	12	8	7	-1	-1
[13,]	13	38	43	5	1
[14,]	14	31	31	0	0
[15,]	15	36	37	1	1

Non-parametric Example

The Sign Test

The sign test makes no assumptions about distributions of X or Y, and results are valid regardless of the distributions of X or Y. But the results are less powerful than t.test(X,Y,alternative="greater").

$$Pval = \sum_{k=C}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^{k} \left(\frac{1}{2}\right)^{n-k} = \left(\frac{1}{2}\right)^{n} \sum_{k=C}^{n} \binom{n}{k}$$

Using the asymptotic large sample approximation to the Binomial Distribution

$$Pval = \left[1 - \Phi\left(\frac{C - \frac{n}{2} - 0.5}{\sqrt{n/4}}\right)\right]$$

Non-parametric Example

The Wilcoxon Signed Rank Test

Example:

The differences were:

```
> diff
 [1]  4  0  3 -2  7  5  2 -3 -1  2  1 -1  5  0  1
```

1. Rank the absolute value of the differences from smallest to largest
2. Calculate the signed rank = sign(X-Y) * rank
3. Sum the signed ranks which are positive

- absdiff=abs(diff)
- rank=rank(absdif)
- signedrank=sign(diff)*rank
- D=cbind(patient=patient,diff=diff,absdiff=absdiff ,rank=rank,signedrank=signedrank)

	student	diff	absdiff	rank	signedrank
[1,]	1	4	4	12.0	12.0
[2,]	2	0	0	1.5	0.0
[3,]	3	3	3	10.5	10.5
[4,]	4	-2	2	8.0	-8.0
[5,]	5	7	7	15.0	15.0
[6,]	6	5	5	13.5	13.5
[7,]	7	2	2	8.0	8.0
[8,]	8	-3	3	10.5	-10.5
[9,]	9	-1	1	4.5	-4.5
[10,]	10	2	2	8.0	8.0
[11,]	11	1	1	4.5	4.5
[12,]	12	-1	1	4.5	-4.5
[13,]	13	5	5	13.5	13.5
[14,]	14	0	0	1.5	0.0

Non-parametric Example

Wilcoxon's Non-parametric Idea

If we replace the actual X and Y values with ranks then the data is akin to drawing marbles Labeled 1, 2, 3, ….., n from an urn containing n marbles

Ranks for | X – Y | in the experiment

Color the marbles red if (X-Y) > 0 and blue if (X-Y) < 0

If we sum the labels on each of the red marbles then this is Wilcoxon's W statistic.

Sum of both blue and red marbles = n (n+1)/2

IF (UNDER H0) the sum of the red marbles = sum of the blue marbles = n(n+1)/4.

Sum of positive signed ranks is the Wilcoxon W statistic

Non-parametric Example

Wilcoxon's Signed Rank Statistic

H0: $\tilde{\mu}_X = \tilde{\mu}_Y \Rightarrow \tilde{\mu}_{X-Y} = 0$ Medians are same

HA: $\tilde{\mu}_X > \tilde{\mu}_Y \Rightarrow \tilde{\mu}_{X-Y} > 0$ Median for X > Median for Y

$$T = \left[\frac{\left| W - \frac{n(n+1)}{4} \right| - \frac{1}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \right]$$

Asymptotically,

T~N(0,1)

Reject H0 whenever,

T > qnorm(1- alpha/2)

Non-parametric Example

Mann-Whitney U Test and Wicoxson Signed Rank Test in R

```
# independent 2-group Mann-Whitney U Test
wilcox.test(y~A)
# where y is numeric and A is A binary factor
```

```
# independent 2-group Mann-Whitney U Test
wilcox.test(y,x) # where y and x are numeric
```

```
# dependent 2-group Wilcoxon Signed Rank Test
wilcox.test(y1,y2,paired=TRUE) # where y1 and y2 are numeric
```

Non-parametric Example

The Kruskal-Wallis Test

Non-Parametric Version of One-Way ANOVA
Generalizes the Wilcoxon Man Whitney Approach for more than one Group.

```
## Hollander & Wolfe (1973), 116.
## Mucociliary efficiency from the rate of removal of dust in normal
##  subjects, subjects with obstructive airway disease, and subjects
##  with asbestosis.
x <- c(2.9, 3.0, 2.5, 2.6, 3.2) # normal subjects
y <- c(3.8, 2.7, 4.0, 2.4)      # with obstructive airway disease
z <- c(2.8, 3.4, 3.7, 2.2, 2.0) # with asbestosis
kruskal.test(list(x, y, z))
## Equivalently,
x <- c(x, y, z)
g <- factor(rep(1:3, c(5, 4, 5)),
        labels = c("Normal subjects",
                "Subjects with obstructive airway disease",
                "Subjects with asbestosis"))
kruskal.test(x, g)
```

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2},$$

Non-parametric vs Parametric Philosophy

- EDA Philosophy: Should adherence to a procedure drive data analysis or should the data drive the chosen procedure?
- Many non-parametric procedures **make no distributional assumptions** so they are less powerful. However they are based upon ranks and percentiles of data which are resistant to outliers and robust to distributional assumptions.
- Big Tradeoff: Is it worthwhile to trade a little statistical power for increased robustness?
- Many methods in EDA use robust and resistant techniques since "we are just playing in the data sandbox." As long as we are in the mode of exploration rather than confirmation, we want to be sure of results, not just sure under some assumptions.

Nonparametric Methods – Ordered Data

Many non-parametric procedures are based upon ranks and percentiles of data. Examples and notation. Suppose we draw n independent samples from some distribution F having mean μ and standard deviation σ, i.e. $X_1, X_2, X_3, \ldots, X_n \sim F_X(x; \mu, \sigma)$. If we sort the n draws from smallest to largest then the **ordered** statistics are denoted

$$\text{smallest} = X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} = \text{largest}.$$

Example: Say your data vector is
$X = c(2.3, 4.6, 1.2, 8.6, 5.6, 7.3, 10.2)$ then

$$X_1 = 2.3, X_2 = 4.6, X_3 = 1.2, \ldots, X_7 = X_n = 10.2$$

The ordered data is

$$X_{(1)} = 1.2, X_{(2)} = 2.3, X_{(3)} = 4.6, \ldots, X_{(7)} = 10.2$$

Distribution of Ordered Data

PDF of Ordered Data

Suppose that $X_1, X_2, \ldots, X_n$ denotes a random sample of size n from a continuous pdf $f(x)$, where $f(x) > 0$ for $a < x < b$. The the pdf of the k^{th} order statistic $X_{(k)}$ is given by

$$g(x_{(k)}) = \frac{n!}{(k-1)!(n-k)!}[F(x_{(k)})]^{k-1}[1 - F(x_{(k)})]^{n-k} f(x_{(k)})$$

if $a < x_{(k)} < b$, and zero otherwise.

Proof: To have $X_k = x_{(k)}$ one must have $(k-1)$ observations less than $x_{(k)}$, one at $x_{(k)}$, and $(n-k)$ observations greater than $x_{(k)}$, where $P(X_k \leq x_{(k)}) = F(x_{(k)})$, $P(X_k \geq x_{(k)}) = 1 - F(x_{(k)})$ and the likelihood of an observation at $x_{(k)}$ is $f(x_{(k)})$. There are $\frac{n!}{(k-1)!(1!)(n-k)!}$ possible orderings of the n independent observations and so the multinomial expression above is the pdf of the ordered statistic.

Sample Ranks

The **sample rank** of a data vector is defined to be the position or index of the vector when that vector is sorted from smallest to largest. The ranks of a vector are denoted $r_1, \ldots r_n$ and are integers (unless there are ties).
Example in R:

$$X = c(2.3, 4.6, 1.2, 8.6, 5.6, 7.3, 10.2)$$

$$rank(X) = c(2, 3, 1, 6, 4, 5, 7)$$

So we would write that

$$r_1 = 2, r_2 = 3, \ldots, r_7 = 7$$

If there are ties in the data we typically average the ranks corresponding to any tied entries. For example,

$$X = c(2.3, 4.6, 4.6, 1.2, 8.6, 5.6, 7.3, 7.3, 10.2)$$

$$rank(X) = c(2, 3.5, 3.5, 1, 8, 5, 6.5, 6.5, 9)$$

Empirical Cumulative Distribution Function (ECDF)

Empirical CDF (ECDF)

```
x = rnorm(12)  # generate 12 N(0,1) random numbers
# look at the numbers
> sort(x)
[1] -0.8542520 -0.7358823  0.2094514  0.4244283  0.7420552  0.8186678
[7]  0.8234618  0.8745359  0.8843882  1.2587840  1.3898083  1.4795486
```

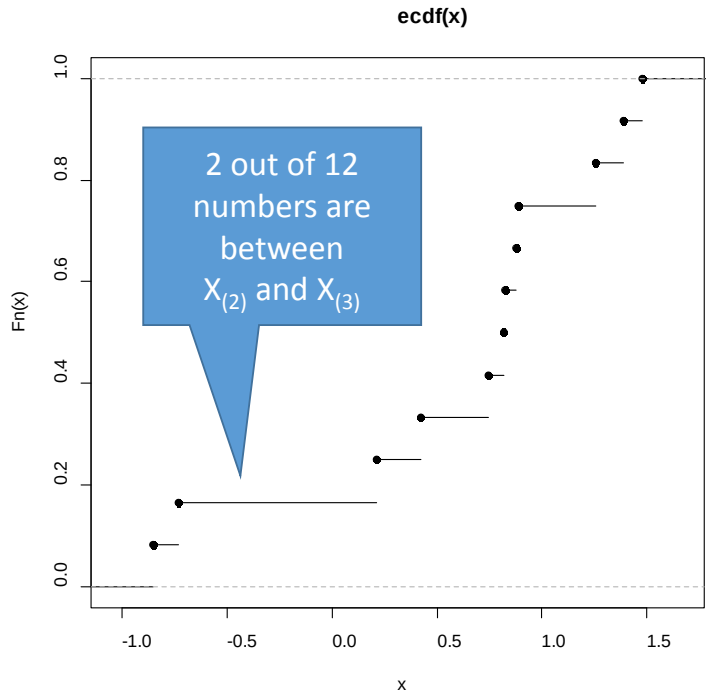
Fn <- ecdf(x) # construct the empirical CDF function in R

Mathematically, the empirical CDF of some data is given by

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I(X_i < x)}{n} = \frac{\#\text{ of data points} < \text{x}}{\text{n}}$$

As $\text{n} \to \infty$ it is well known that $\hat{\text{F}}(\text{x}) \to \text{F}(\text{x})$

in the sense that $|\hat{\text{F}}(\text{x}) - \text{F}(\text{x})| \to 0$ almost surely



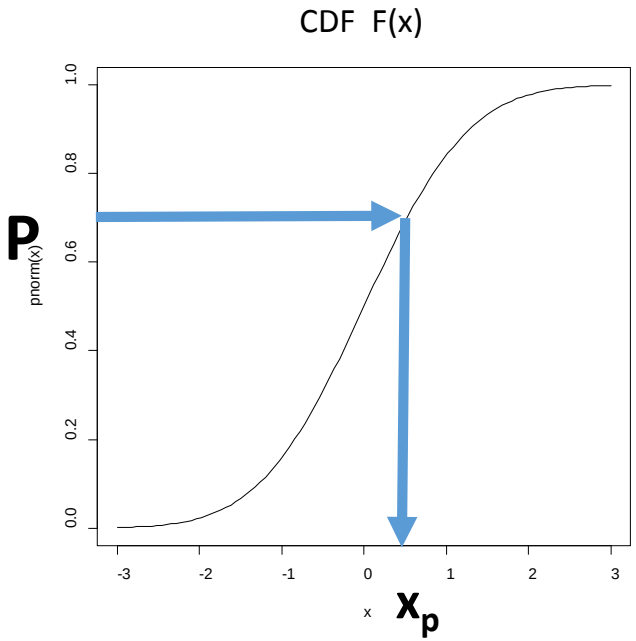
Large Sample Asymptotics of ECDF

Since $\mathrm{E}[I(X < x)] = F(x)$, $\mathrm{Var}[I(X < x)] = F(x)(1 - F(x))$ and $\hat{F}(x) = \frac{\sum_{i=1}^{n} I(X_i < x)}{n}$ the Central Limit Theorem ensures that

$$\boxed{\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x)))}$$

This result has implications in many areas of Statistics, like bootstrapping, quantiles etc.

Percentiles or Quantiles



The p^{th} quantile x_p of a distribution F(x) is the value On the x-axis which satisfies

$$p = F(x_p)$$

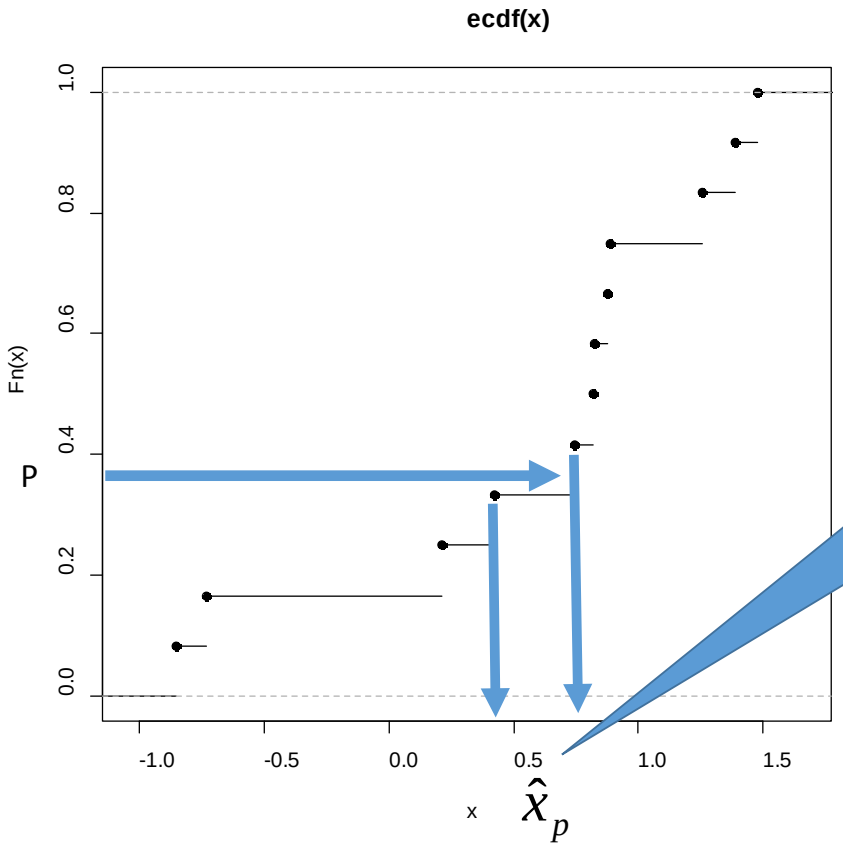
The Inverse function of the CDF is called the Quantile Function and this satisfies the equation

$$Q(p) = F^{-1}(p) = x_p$$

In R, the CDF of x is produced by putting a "P" in front of the distribution name, for example $pnorm(x) = F_{normal}(x) = \Phi(x)$

The quantile function is produced by putting a "q" in front of the distribution name, for example $qnorm(p) = F^{-1}_{normal}(p) = \Phi^{-1}(p) = x_p$

Percentiles or Quantiles

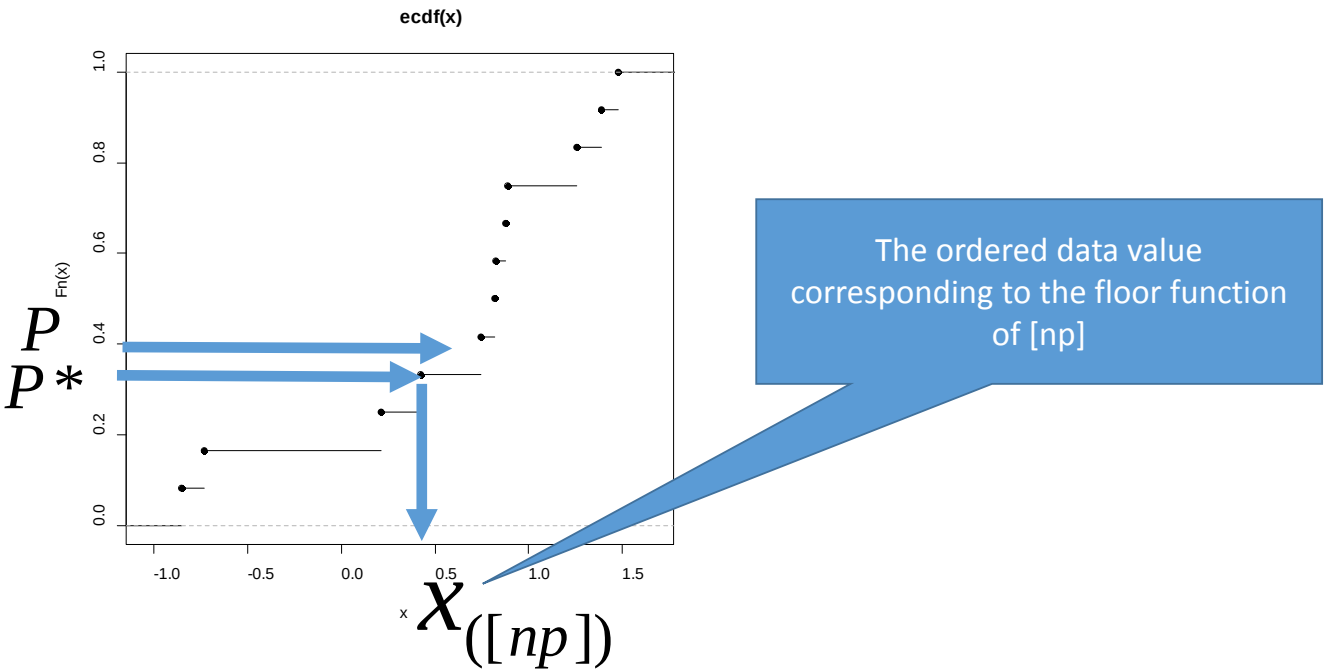


Estimated Quantile Is somewhere in this interval. There are 9 different commonly used definitions however.

Percentiles or Quantiles

Method 1 of finding quantile. Compute [np] where [*] is the floor or truncation function.

Example: Say n = 12 and p=0.4 then [np]= [4.8] = 4 (set the fractional part to zero). Then pictorially, this method finds the closest value on the y-axis P* which corresponds to some data entry and then states the value on the x-axis X_{p*} which is closest to that value

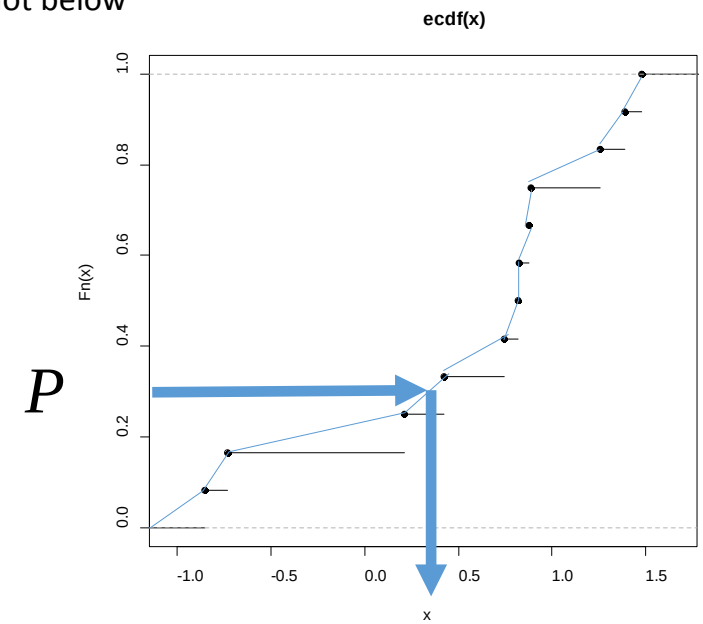


Percentiles or Quantiles

Method 4 of finding a quantile = **Linear Interpolation Technique**.

Draw straight lines which connect the dots in between the discontinuities of the ECDF plot below

Draw



P

$\hat{x}_p$

To compute the quantile draw horizontal line from P until it connects with one of the line segments connecting the dots and then read off the corresponding X-value.

Nine Different Defns of Quantile

There are 9 different definitions of the quantile in R. Type "help(quantile)" in R for more info.

Table: Definitions of Quantile

Quantile Method	Definition
Type 1	Returns smallest quantile value $x_{(\lfloor np \rfloor)}$
Type 2	Returns the average between discontinuities $\frac{x_{(\lfloor np \rfloor)}+x_{(\lceil np \rceil)}}{2}$
Type 3	Returns the nearest order statistic $x_{(\text{round}(np))}$ (SAS defn)
Type 4	Linear interpolation between the points $(p(k), x_{np(k)})$ where $p(k) = \frac{k}{n}$.
Type 5	Linear interpolation between the points $(p(k), x_{np(k)})$ where $p(k) = \frac{k-0.5}{n}$.
Type 6	Linear interpolation between the points $(p(k), x_{np(k)})$ where $p(k) = \frac{k}{n+1}$ (Minitab and SPSS).
Type 7	Linear interpolation between the points $(p(k), x_{np(k)})$ where $p(k) = \frac{k-1}{n-1}$ (Used in S).
Type 8	Linear interpolation between the points $(p(k), x_{np(k)})$ where $p(k) = \frac{k-1/3}{n+1/3}$.
Type 9	Linear interpolation between the points $(p(k), x_{np(k)})$ where $p(k) = \frac{k-3/8}{n+1/4}$.

Large Sample Asymptotics of Quantiles

As far as the large sample asymptotics is concerned, all nine definitions operate the same. Most people regard the differences between definitions 1-9 like "splitting hairs". Since all the asymptotics are the same consider $\hat{x}_p = x_{(\lfloor np \rfloor)}$.

Want asymptotic distribution of $\sqrt{n}(\hat{x}_p - x_p)$ as $n \to \infty$.

The key to this is to recognize that $\hat{x}_p \approx F_X^{-1}(\hat{F}_X(x_p))$ and by the continuity of $F(\cdot)$, $\hat{x}_p \approx F_X^{-1}(\hat{F}_X(x_p)) \to F_X^{-1}(F_X(x_p)) = x_p$. Now if $g(t) = F_X^{-1}(t)$ then

$$\frac{d(g(t))}{dt} = \frac{d(F_X^{-1}(t))}{dt} = \frac{1}{f_X(F_X^{-1}(t))}$$

since

$$y = F_X^{-1}(t) \iff F_X(y) = t \iff f_X(y)dy = dt \iff \frac{dy}{dt} = \frac{1}{f_X(F_X^{-1}(t))}.$$

Hence by the Delta Theorem

$$\begin{aligned}\sqrt{n}(\hat{x}_p - x_p) &= \sqrt{n}(F_X^{-1}(\hat{F}_X(x_p)) - F_X^{-1}(F_X(x_p))) \\ &\xrightarrow{d} N\left(0, \frac{F_X(x_p)[1 - F_X(x_p)]}{\{f(x_p)\}^2}\right) = N\left(0, \frac{p(1-p)}{\{f(x_p)\}^2}\right).\end{aligned}$$

Large Sample Asymptotics of Quantiles

To summarize the asymptotic distribution of the p^{th} quantile is given by

$$\boxed{\sqrt{n}(\hat{x}_p - x_p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{\{f(x_p)\}^2}\right)}$$

This implies that the asymptotic distribution of the sample median is

$$\boxed{\sqrt{n}(\tilde{x} - x_{0.5}) \xrightarrow{d} N\left(0, \frac{1}{4\{f(x_{0.5})\}^2}\right)}$$

Large Sample Asymptotics of Quantiles

Let's assume $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$ then because $x_{0.5} = \text{median} = \text{mean} = \mu$ and

$$f_X(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(t-\mu)^2}{2\sigma^2}\right)$$

then

$$\sqrt{n}(\tilde{x} - x_{0.5}) \xrightarrow{d} N\left(0, \frac{1}{4\{f(\mu)\}^2}\right) = N\left(0, \frac{\pi\sigma^2}{2}\right)$$

so

$$\text{Var}(\tilde{x}) \to \frac{\pi\sigma^2}{2n} \approx 1.57\frac{\sigma^2}{n}.$$

This implies that the asymptotic relative efficiency of the median compared with the mean is

$$\text{ARE}(\bar{x}, \tilde{x}) = \frac{\text{Var}(\bar{x})}{\text{Var}(\tilde{x})} = \frac{\frac{\sigma^2}{n}}{\frac{\pi\sigma^2}{2n}} = \frac{2}{\pi} = 0.6366$$

the median has 64% the efficiency of the mean (not bad!!).

IQR and Psuedovariance

Let's assume $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$ what is the asymptotic distribution of the IQR?
Since $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ we know that

$$\begin{aligned} P(X < x_{0.75}) &= P\left(\frac{X-\mu}{\sigma} < \frac{x_{0.75}-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x_{0.75}-\mu}{\sigma}\right) = 0.75 \\ &\Longrightarrow x_{0.75} = \mu + \Phi^{-1}(0.75)\sigma \text{ and due to symmetry,} \\ &\Longrightarrow x_{0.25} = \mu - \Phi^{-1}(0.75)\sigma. \end{aligned}$$

This implies that if the data are normally distributed

$$IQR = x_{0.75} - x_{0.25} = 2 * \Phi^{-1}(0.75)\sigma \approx 1.35\sigma$$

Hence a rough approximation for the variance is

$$\hat{\sigma} = \text{Pseudo-variance} = \frac{IQR}{1.35} = 0.337 * IQR$$

Mean vs Median Tradeoff

The above shows that **if we assume normality** the mean is a more efficient estimator of μ than the median.
However, the median is more resistant to outliers. How do we know what distribution the data come from?

The nice normal distribution pillow:

The evil Cauchy distribution pillow:

Stem and Leaf Display

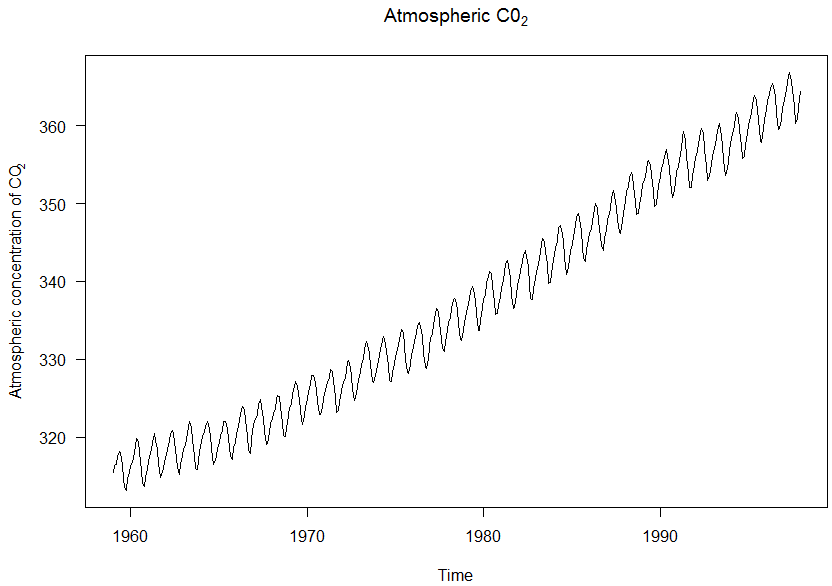
An Early Device to Look at Data: The Stem-and-Leaf Display

Stem and Leaf Display

Data Example: CO2 Concentration in the Atmosphere

> co2

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1959	315.42	316.31	316.50	317.56	318.13	318.00	316.39	314.65	313.68	313.18	314.66	315.43
1960	316.27	316.81	317.42	318.87	319.87	319.43	318.01	315.74	314.00	313.68	314.84	316.03
1961	316.73	317.54	318.38	319.31	320.42	319.61	318.42	316.63	314.83	315.16	315.94	316.85
1962	317.78	318.40	319.53	320.42	320.85	320.45	319.45	317.25	316.11	315.27	316.53	317.53
....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....



Stem and Leaf Display

- Objective of stem and leaf diagram: organize the numbers graphically to notice features of data such as
  1. Symmetry
  2. Approximate center & spread of the data
  3. Values far from the rest
  4. Clusters, gaps
- Widely used basic exploratory technique: compare batches, examine residuals

Stem and Leaf Display

The basic R function for a stem and leaf display is stem() However, the function stem.leaf() in the aplpack package has more options.

```
install.packages(aplpack)
library(aplpack)
stem(co2)
```

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

 31 | 3444
 31 | 55555556666666667777777777778888888888888888999999999999999
 32 | 00000000000000001111111111122222222222222233333333444444444444
 32 | 55555555555566666667777777777777888888888889999999999
 33 | 00000011111111112222222222333333333444444444
 33 | 55555556666666677777778888888888889999999
 34 | 00000001111111111222223333333333444444444
 34 | 55555555566666667777778888888899999999
 35 | 00000011111122222222333333344444444444
 35 | 55555555555666666667777777888888888999999
 36 | 00000000001111112222222333334444444
 36 | 55555667
```

Stem and Leaf Display

The parameter scale roughly adjusts the number of lines in the stem and leaf plot. The default is scale=1, so scale = 2 gives double the number of lines.

```
> stem(co2,scale=2)

 The decimal point is at the |

312 | 277
314 | 077882344789
316 | 011334555677899134555667 89
318 | 00124445566779912333445566679
320 | 11223344556778991223446778
322 | 000112222344779990245567899 9
324 | 0033567899002345789
326 | 0223555680022234455669 9
328 | 002344458891222344699
330 | 112667800333456799
332 | 12344566899123477888
334 | 24557888127899
336 | 1145566679466778888
338 | 022371113447789
340 | 33366900222349
342 | 1234567889912458
344 | 00013578811245678
346 | 123467893367889
348 | 12367782344689
350 | 033880112557
352 | 1112467899245566789
354 | 0011667990233355689
356 | 0011668900155678
358 | 113359112245566667
360 | 00227889035779
362 | 1456222578
```

Mechanics of Stem and Leaf Display

Mechanics behind creating stem and leaf display:

1. Separate number into its stem (the part that need not be repeated each time) and the leaf (extra digit of detail)

$$\text{data value}(22.9) \rightarrow \text{split}(22|9) \rightarrow \text{stem}(22)\text{and leaf}(9)$$

2. Single line per stem: stem.leaf($data, m = 1$)

2	0123456789

3. Two lines per stem: stem.leaf($data, m = 2$)

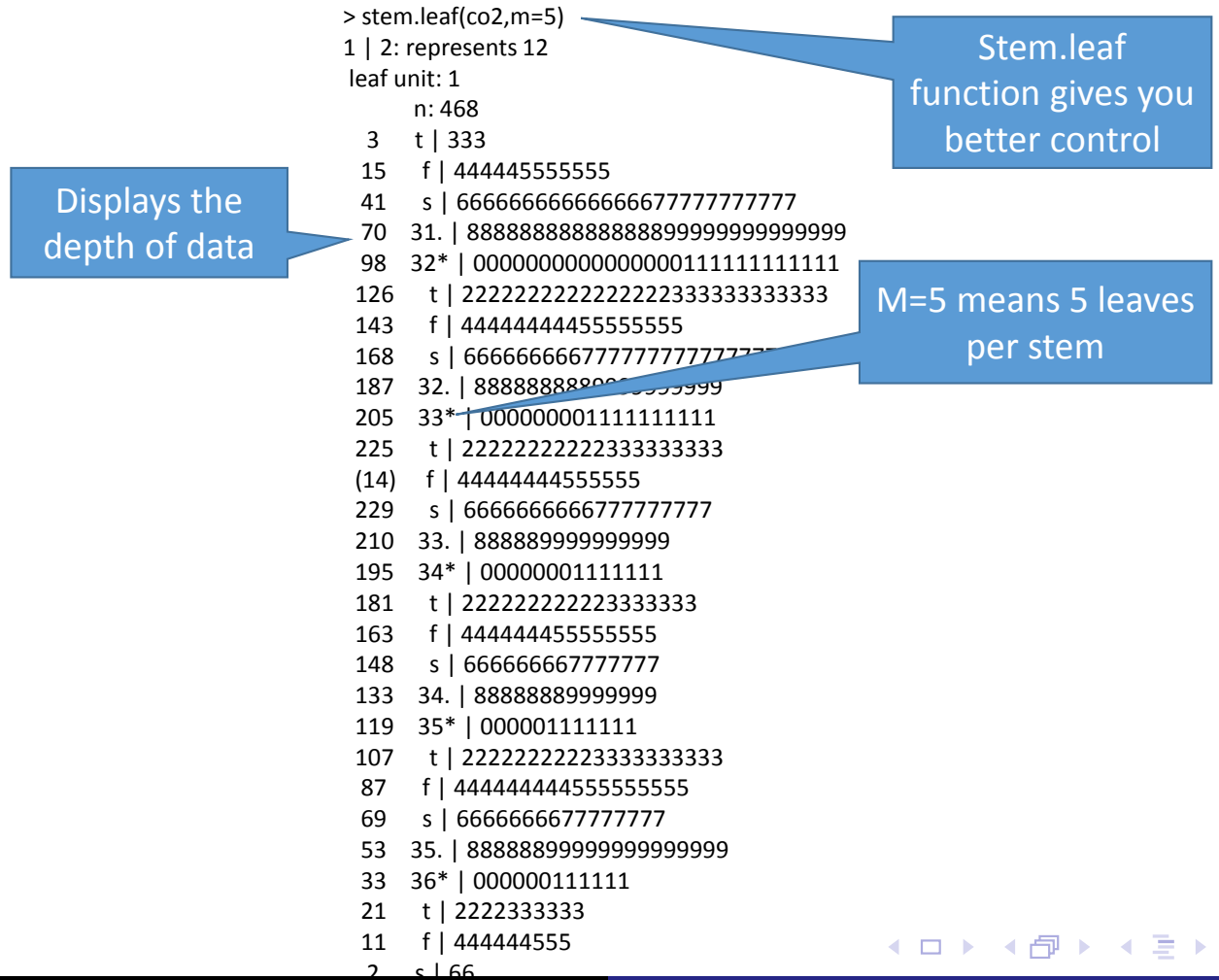
2*	01234
2+	56789

4. Five lines per stem: $*(0,1)t(2,3)f(4,5)s(6,7).(8,9)$

2*	01
T	23
F	45
S	67
.	89

Stem and Leaf Display

The stem.leaf() function in R gives you more information and better control of your stem and leaf diagram.



Stem and Leaf Display

One can compare two different distributions back to back by using the stem.leaf.backback() function under the aplpack package

```
> stem.leaf.backback(co2[1:234],co2[235:468])
______________________________________________________________________________
  1 | 2: represents 12, leaf unit: 1
                                              co2[1:234]       co2[235:468]
______________________________________________________________________________
   8                                            44444333| 31* |
  70     99999999999999888888888888888877777777777666666666666666655555555| 31. |
 (65)  44444444433333333333332222222222222222211111111111100000000000000| 32* |
  99                    9999999999988888888887777777777777777766666666665555555| 32. |
  47                               444444333332222222222111111111000000000| 33* |2233344                         7
   8                                            77766655| 33. |55556666667777788888999999999999    39
                                                        | 34* |000000011111122222222222233333334444444    78
                                                        | 34. |5555555566666666777777888888889999999    115
                                                        | 35* |00000111111222222222222333333333344444444    (41)
                                                        | 35. |555555555666666666777777778888889999999999999999    78
                                                        | 36* |0000001111112222333333444444    33
                                                        | 36. |55566                            5
                                                        | 37* |
______________________________________________________________________________
n:                                                  234       234
______________________________________________________________________________
```

The Depth Display

Record the forward and backward rank by counting from each end of the ordered batch.

- Depth = smaller of the two ranks
- Present depths with the display $\rightarrow$ some summary values can easily be defined in terms of their depths.
- Middle line: includes median. Depth display displays the Number of leaves instead of the depth in parentheses.
- Other lines: maximum depth on that line

Stem and Leaf vs Histogram

The stem and leaf plot is similar to the histogram but has some advantages

Advantages include:

- Retains the most significant digits of the data
- Easier to construct by hand
- Readily find the median and the other summaries
- See the distribution of data values
- Go more easily from a value in the display to the datum that produced it

How many Lines in Stem and Leaf?

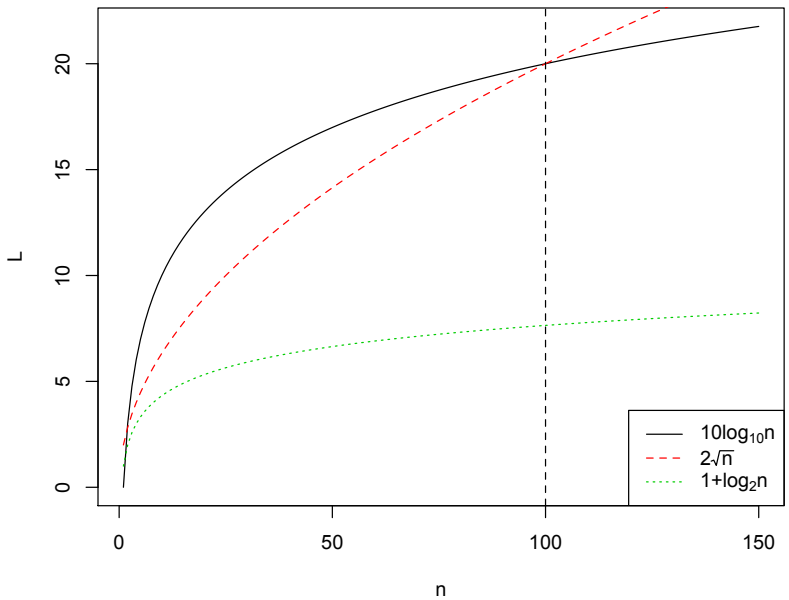
Sturges argued that if n were a power of 2 then "the proper frequency distribution follows the sequence of binomial coefficients." So if $n = 2^4 = 16$, sturges would argue that the line frequencies should be $1, 4, 6, 4, 1$. This is equivalent to saying that $2^{L-1} = n$ thus

$$\text{Sturges line rule } = L = 1 + \log_2(n)$$

Sturges rule based upon "aesthetically pleasing" argument.

Rules for the number of lines

- $L = 10 \cdot \log_{10} n$ as an upper limit (Dixon and Kronmal, 1965)
- $L = 2\sqrt{n}$: fewer lines when n is small, but grows too fast $n > 100$ (Velleman, 1976)
- Sturges' rule: $L = 1 + \log_2 n$, way too small; many data on one line



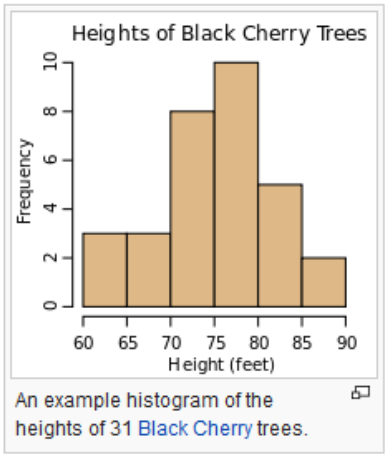
How many Lines in Stem and Leaf?

Connection between stem-and-leaf and histogram

- L =: # of lines for a stem-and-leaf display;
  h = interval width/histogram bin size
- desirable interval width for a histogram brings it close to an assumed density function for the data
- For stem-and-leaf, interval width is 2, 5, or 10 times a power of 10.

Histograms

A histogram records the count (or frequency) of data points which fall into arbitrary bins.



An example histogram of the heights of 31 Black Cherry trees.

Bin	Count
60	3
65	3
70	8
75	10
80	5
85	2

Steps:

1. Compute $\min(x) = x_{(1)}$, $\max(x) = x_{(n)}$ and $\mathrm{Range}(x) = x_{(n)} - x_{(1)}$
2. Decide either # of bins $= k = \left\lceil \frac{(x_{(n)} - x_{(1)})}{h} \right\rceil$ with h the bin width or upon bin width $h = \frac{(x_{(n)} - x_{(1)})}{k}$.
3. Record frequency within the j^{th} bin n_j for $j = 1, \ldots, k$.

Optimum Bin Size

Suppose h is the bin size, x_0 is the position of the first bin and $B_j = (x_0 + (j-1)h, x_0 + jh)$ is the j^{th} bin of the histogram. Then the histogram value at $x \in B_j$ is defined as

$$\hat{f}_h(x) = \frac{n_j}{nh} = \frac{1}{nh} \sum_{i=1}^{n} I(x_i \in B_j)$$

Since the bin count is a Binomial random variable, i.e. $n_j \sim \text{Bin}(n, p_j)$ with $p_j = \int_{B_j} f(x)dx$ then $\text{E}[n_j] = np_j$ and $\text{Var}[n_j] = np_j(1-p_j)$. Thus

$$\text{Bias}[\hat{f}_h(x)] = \text{E}[\hat{f}_h(x)] - f(x) = p_j/h - f(x) = \frac{1}{h}\int_{B_j} f(t)dt - f(x) \text{ and,}$$

$$\text{Var}[\hat{f}_h(x)] = \frac{1}{(nh)^2}\text{Var}[n_j] = \frac{1}{nh^2}p_j(1-p_j)$$

Optimum Bin Size

Now if we utilize the first order Taylor series approximation $f(t) = f(x) + f'(x)(t-x) + O(h^2)$ then

$$\begin{aligned}\text{Bias}[\hat{f}_h(x)] &= \frac{1}{h}\int_{B_j}(f(t) - f(x))dt \\ &= \frac{1}{h}\int_{B_j} f'(x)(t-x)dt + o(h) = f'(x)h + o(h)\end{aligned}$$

Moreover since

$$p_j = \int_{B_j} f(t)dt = \int_{j-1h}^{jh}(f(x)+f'(x)(t-x)+O(h^2))dt = f(x)h+O(h^2)$$

the variance is

$$\text{Var}[\hat{f}_h(x)] = \frac{1}{nh^2}p_j(1-p_j) = \frac{f(x)}{nh} + \text{higher order terms}$$

Optimum Bin Size

Optimum bin width to optimize $MSE(x) = E[\hat{f}_n(x) - f(x)]^2$.
Want to minimize

$$MSE(x) = \text{Bias}^2[\hat{f}_h(x)] + \text{Var}[\hat{f}_h(x)] = \frac{f(x)}{nh} + f'^2(x)h^2$$

Taking $\frac{\partial MSE(x)}{\partial h} = 0 \Longrightarrow$ the MSE is minimized when h is chosen such that

$$h_{opt} = \left[\frac{f(x)}{2(f'(x))^2 n}\right]^{1/3}$$

Rules for interval width (D. W. Scott, 1979)

Objective: choose the interval width, h_n, to minimize IMSE

1. pdf: f, estimate $\hat{f}$, $MSE(x) = E[\hat{f}_n(x) - f(x)]^2$,
2. $IMSE = \int E[\hat{f}_n(x) - f(x)]^2 \, dx$
3. Minimize IMSE as a function of bin width used in $\hat{f}$.

$$h_n = \left[\frac{6}{n \int_{-\infty}^{\infty} [f'(x)]^2 \, dx} \right]^{1/3} \propto n^{-1/3}$$

4. Problem: Need to know f' or f.
5. If $f = \Phi$, Gaussian, $h_n \approx 3.5s/n^{1/3}$, where s is an estimate related to standard deviation.
6. For example, $s = 3/4IQR \rightarrow 2.6IQR/n^{1/3}$.

Rules for interval width (Freedman and Diaconis)

- $$h_{opt} = argmin_h D(h) = argmin_h \{\max_x \|\hat{f}_{h_n}(x) - f(x)\|\}$$

$$\Rightarrow h_n = c(f)(\log_e n/n)^{1/3}$$

- If $f = \Phi$ (Gaussian), $h_n \approx 2(IQR)/n^{1/3}$.

The Quantile-Quantile Plot

QQ-plot

We note already that:

- Eye has trouble detecting departures from curves
- Bell-shaped curve is especially curvy
- Much easier to detect departures from . . . ?

QQ-plot

QQ plots: look at the closeness of the distribution of data to some ideal distribution, and help us to answer the question:

Are the data Gaussian? (exponential, Weibull, chi-squared, . . .)

- Often consider specifically the questions concerning Gaussian data, which is the most commonly assumed distribution, and for which our familiar tests — Student's t, F-tests, CIs for variance — are valid
- Order data, and plot the ith ordered data from the batch of size n against the corresponding Gaussian quantile
- Useful plot to examine residuals

Example: data from normal distribution

$n = 50$ observations from N(0,1)

```
-2.64 -2.00 -1.39 -1.39 -1.33 -1.25
-1.23 -1.14 -0.92 -0.88 -0.83 -0.83
-0.82 -0.77 -0.72 -0.50 -0.48 -0.46
-0.38 -0.29 -0.23 -0.15 -0.05  0.08
 0.13  0.16  0.17  0.18  0.18  0.21
 0.24  0.27  0.28  0.31  0.42  0.52
 0.53  0.61  0.66  0.73  0.80  0.92
 0.97  1.06  1.06  1.70  1.76  1.85
 1.95  3.21
```

Example: stem-and-leaf

$n = 50$ observations from N(0,1); stem():

```
The decimal point is at the |

  -2 | 60
  -1 | 443321
  -0 | 998888755543220
   0 | 11222222333455677 89
   1 | 011788
   2 | 0
   3 | 2
```

Example:

- Around what value do we expect the median to fall?
- The fourths ("quartiles")?
- The extremes?
- Distribution in general?

Example: plot

- y-axis: sorted observations $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$
- x-axis: expected value, $\Phi^{-1}(p_i)$, where
  - $\Phi(x) =$ cumulative distribution function of N(0,1)
  - $p_i =$ expected "area" up to $y_{(i)}$
- What should p_i be?

Example: p_i

What should p_i be?

- i/n?
- $(i-1)/n$?
- $i/(n+1)$? (suggested by Weibull)
- $(i-a)/(n+b)$?
- $a = b = 1/3$:

$$p_i = \frac{i - 1/3}{n + 1/3} = \frac{3i - 1}{3n + 1}$$

- Reason: Blom (1958) (UREDA page 44-46)

$$\text{median of } Y_{(i)} \approx \Phi^{(-1)}(p_i)$$

The above is true for any continuous distribution, not just Φ:

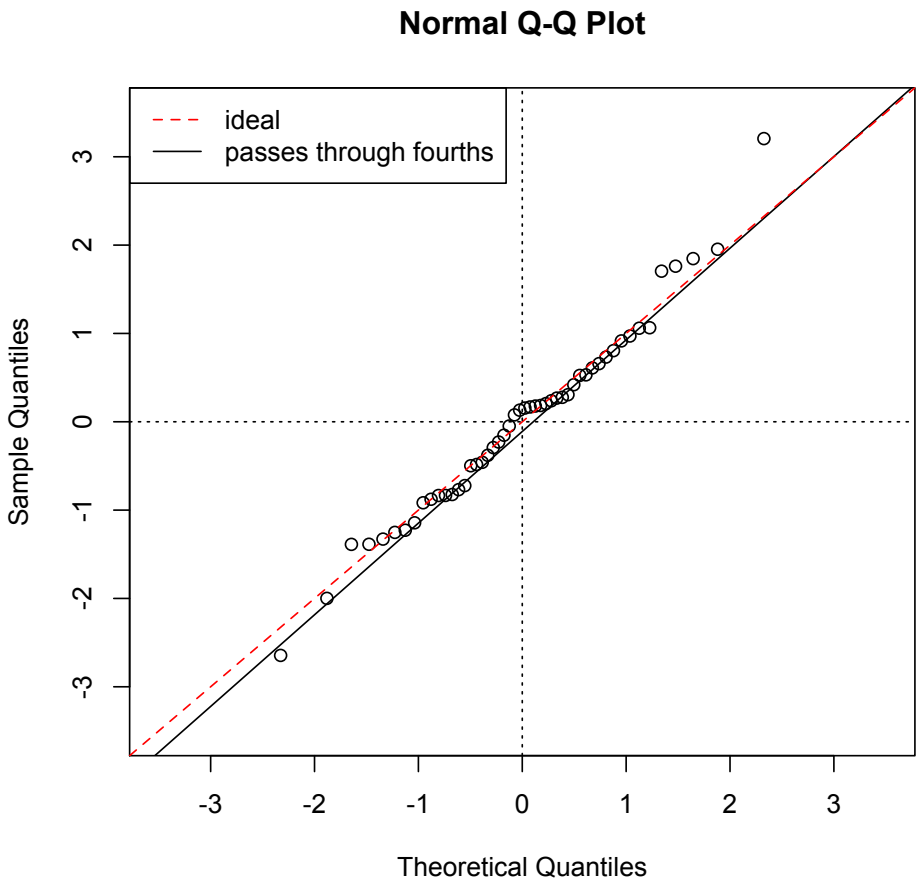
- Exponential: $F(y) = 1 - e^{-\lambda y}$,
  $F^{(-1)}(p_i) = (-1/\lambda) \cdot \log_e(1 - p_i)$
- Weibull: $F(y) = 1 - e^{-\lambda y^{\beta}}$

QQ plots in R

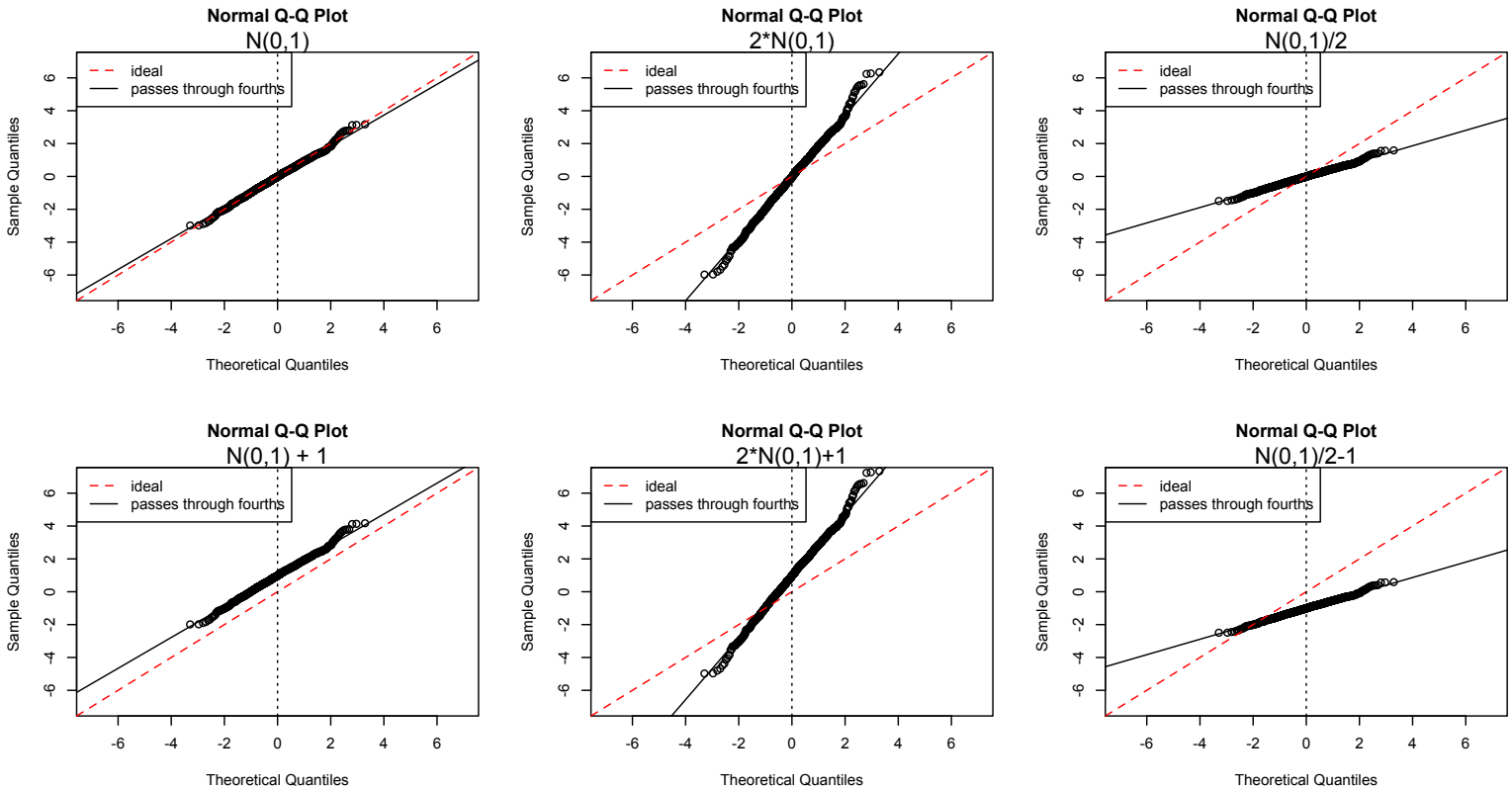
- qqnorm(y) is a generic function the default method of which produces a normal QQ plot of the values in y.
- qqline(y) adds a line to a normal quantile-quantile plot which passes through the first and third quartiles.
- qqplot(x, y) produces a QQ plot of two datasets.

Example: QQ-plot

Data: 50 data from N(0,1)

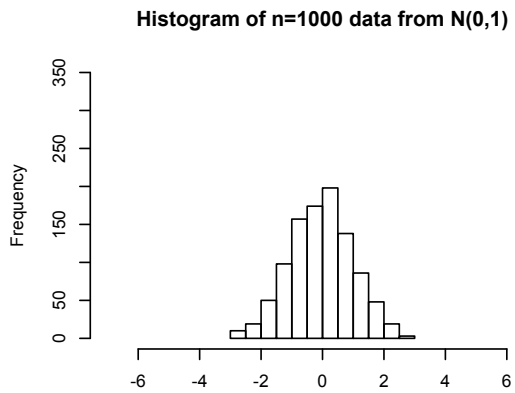


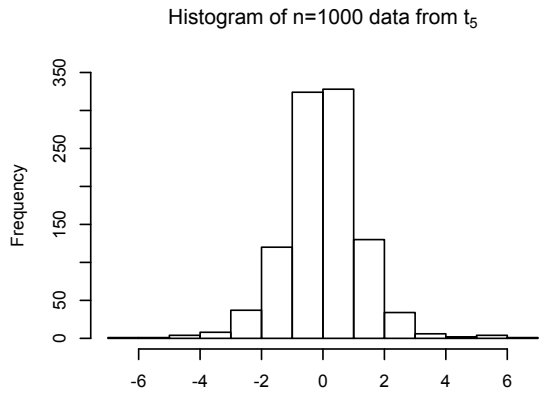
Interpret QQ-plot (1000 data from normal distributions)

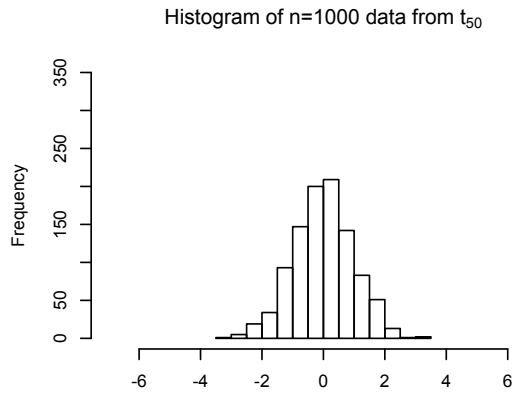


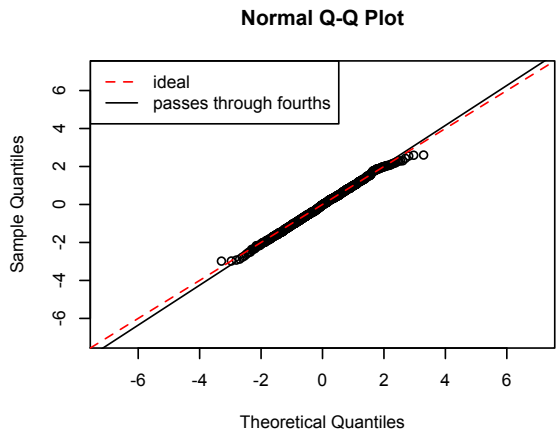
- QQ plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

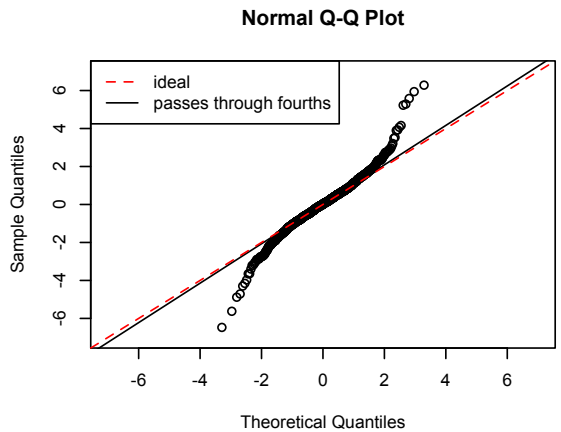
More examples on QQ-plot (heavy-tailed distribution)

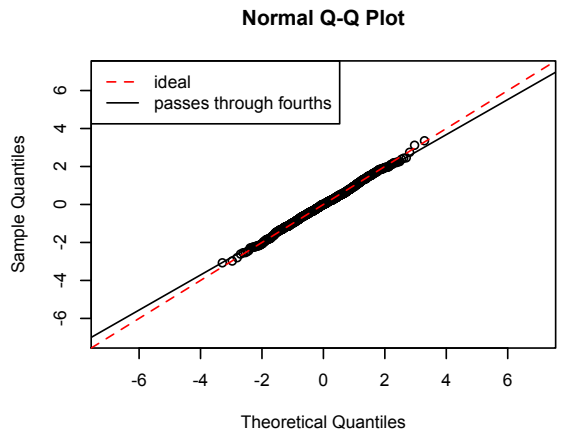




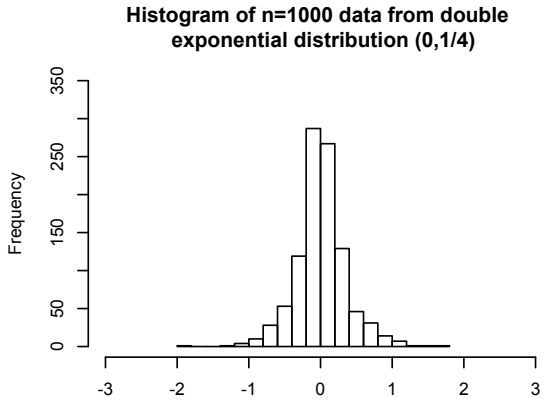


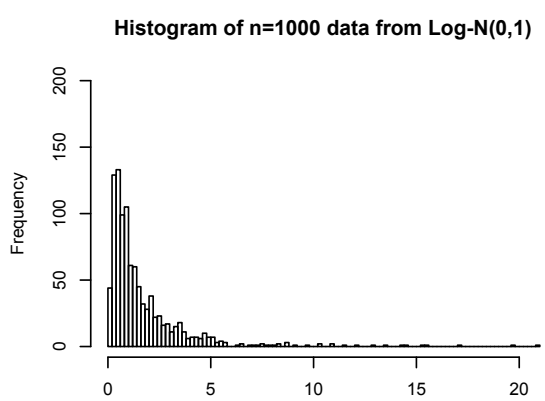


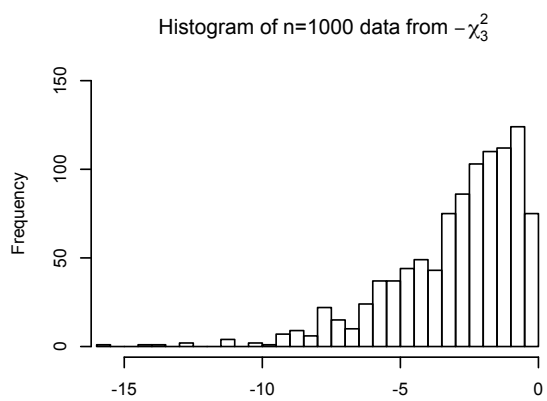


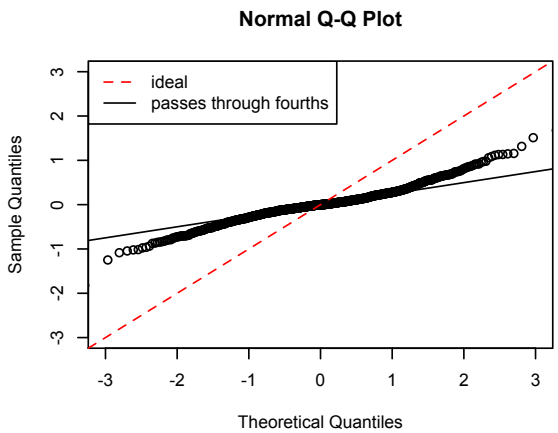


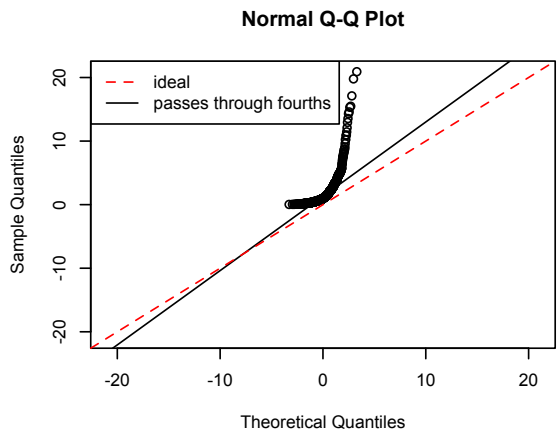
More examples on QQ-plot (light-tailed and skewed distributions)

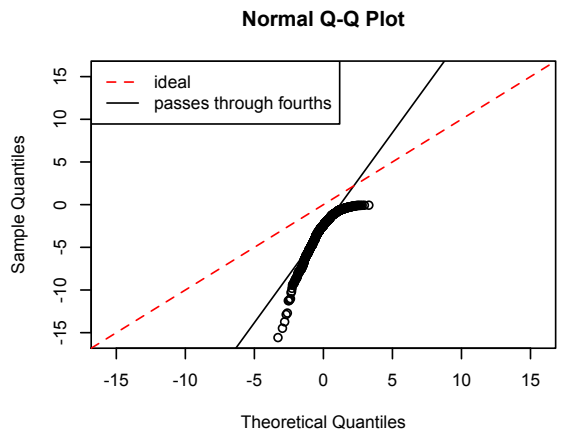




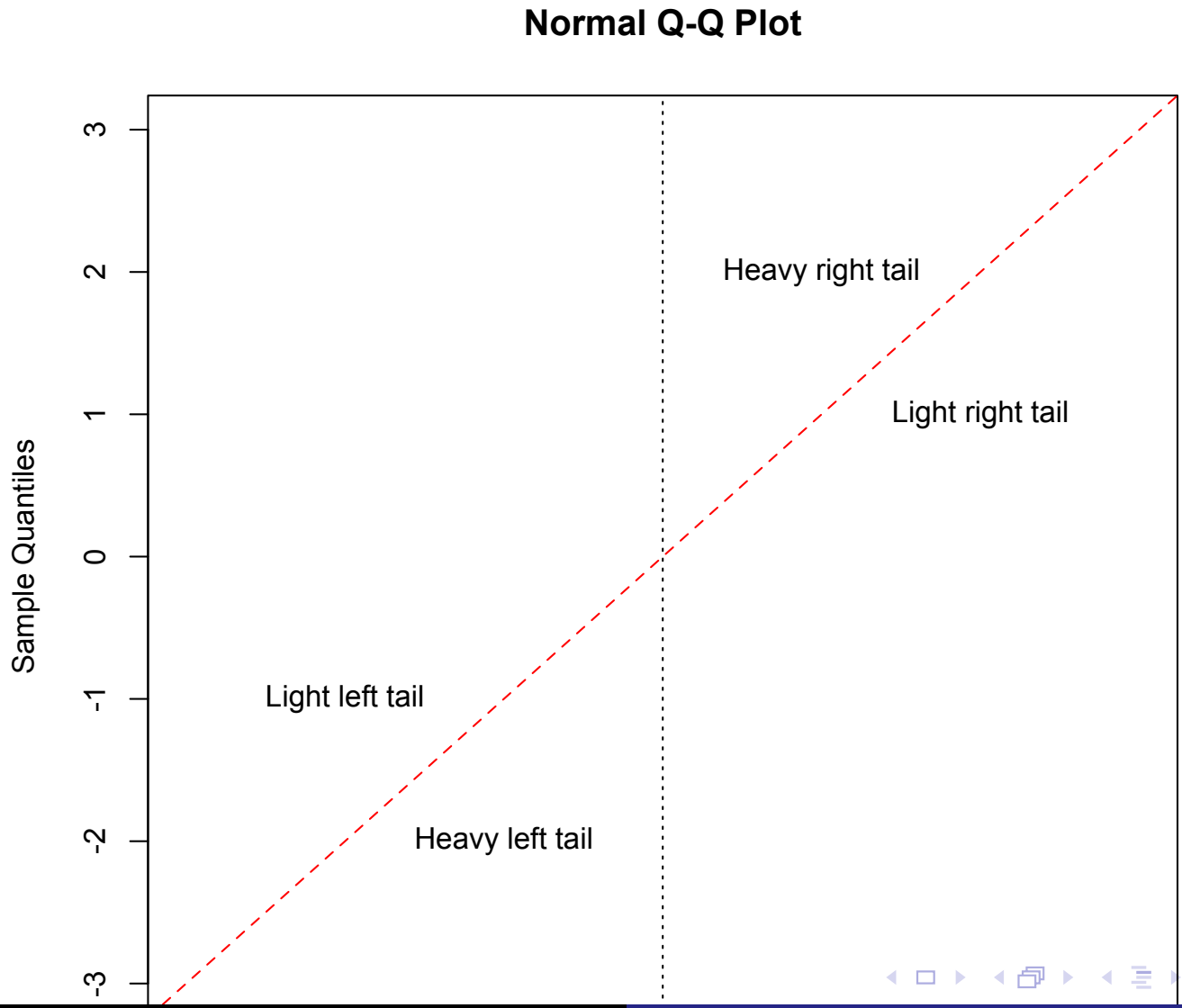




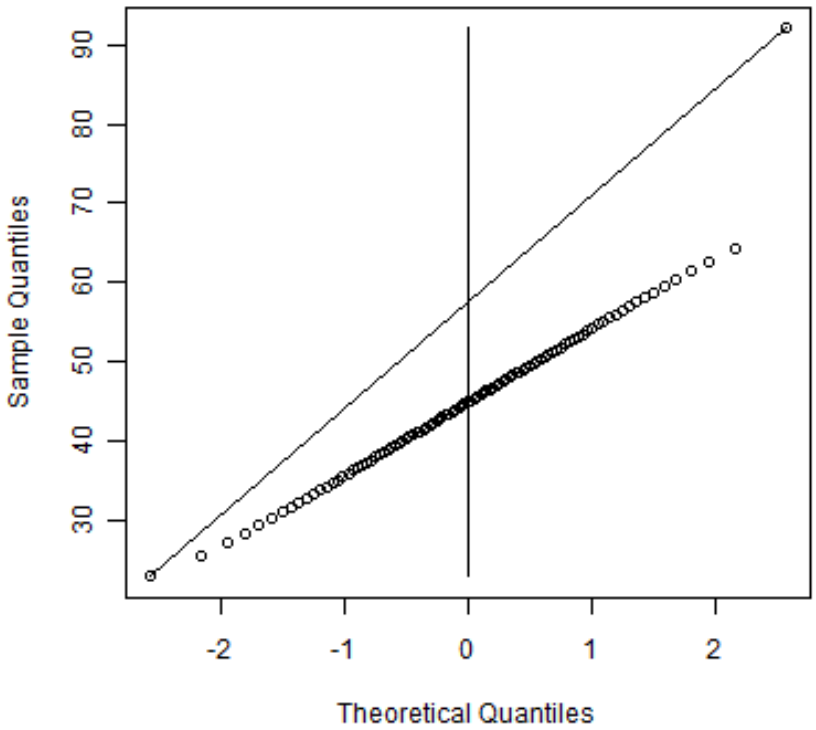


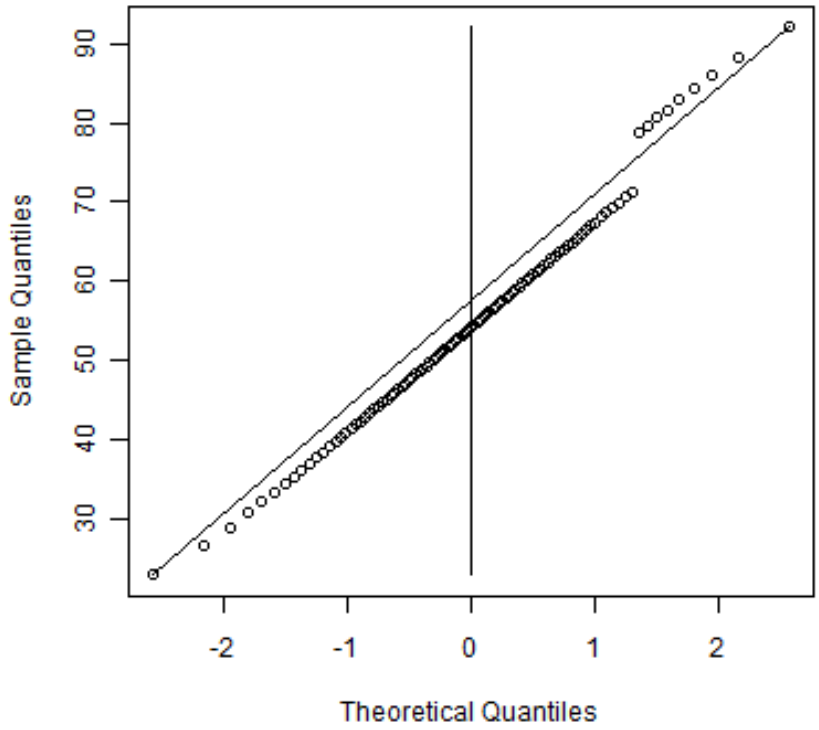


More examples on QQ-plot



More examples on QQ-plot (outlier and multimodality)





Kernel Density Estimators

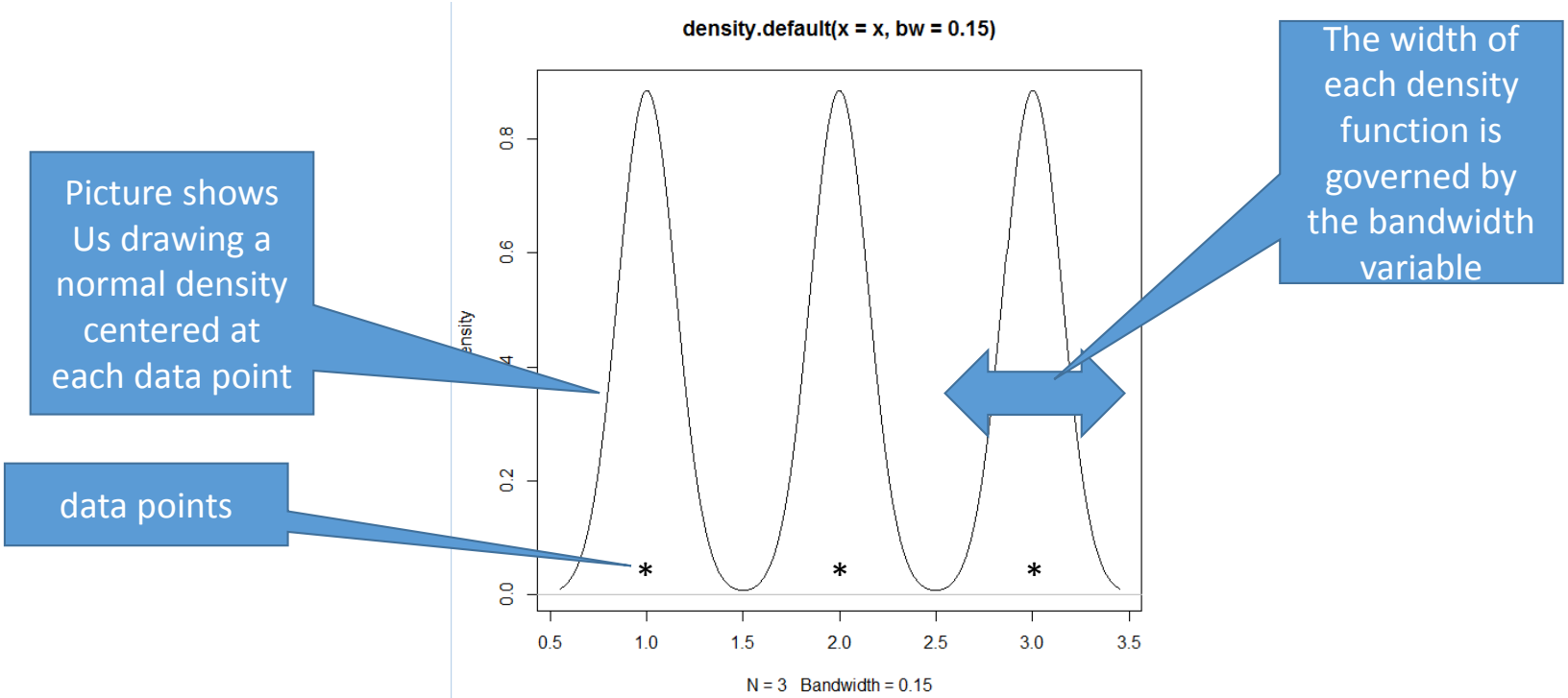
How NonParametric Kernel Density Works

Mechanics of the Non-Parametric Density Plot

Data: X = c(1,2,3) #Just for fun put in 1, 2, 3 into R

Now let's draw a density function called a Kernel K(t) which is centered at each data point

plot(density(x,bw=0.15))



How NonParametric Kernel Density Works

Each density we add to the picture is called a "Kernel" function or "Kernel density" and mathematically the function has the form

$$K(x) = \frac{1}{h} k\left(\frac{x - x_i}{h}\right)$$

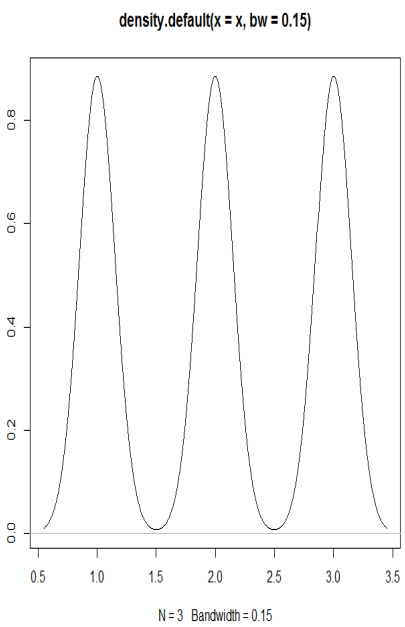
where h is called the **bandwidth and governs the width of the density**.
The kernel density estimator is given by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$$

Each kernel function has area under the curve of $1/n$ so that when you add them all up, the total area is 1.

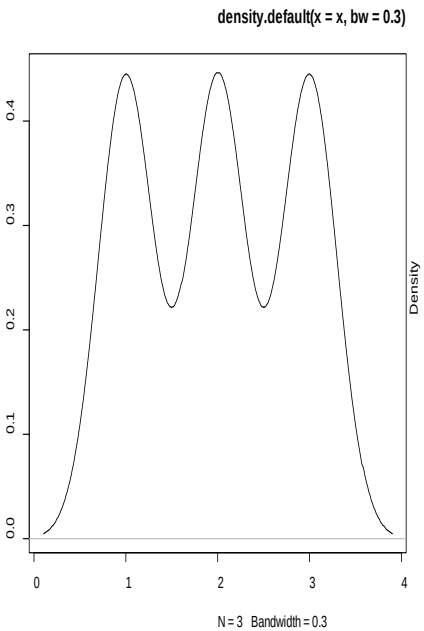
How NonParametric Kernel Density Works

h = 0.15



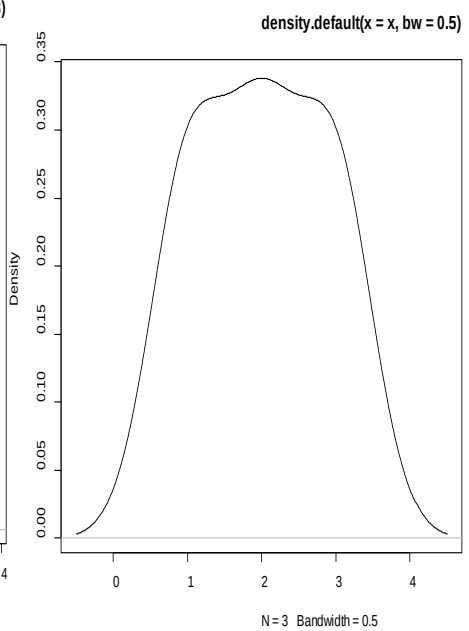
Really Waavy

h = 0.3



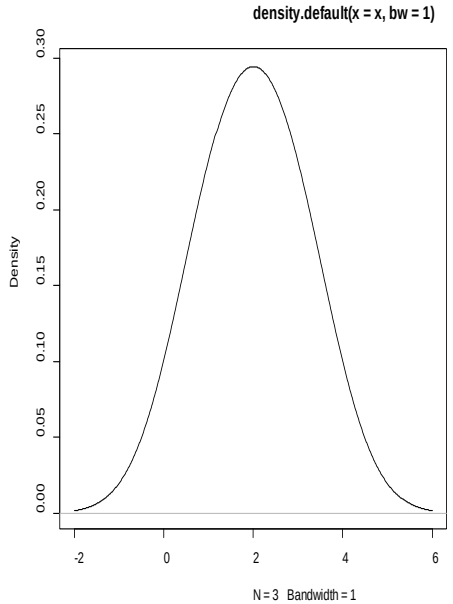
Still Pretty Waavy

h = 0.5



Still Kinda Waavy but not too bad

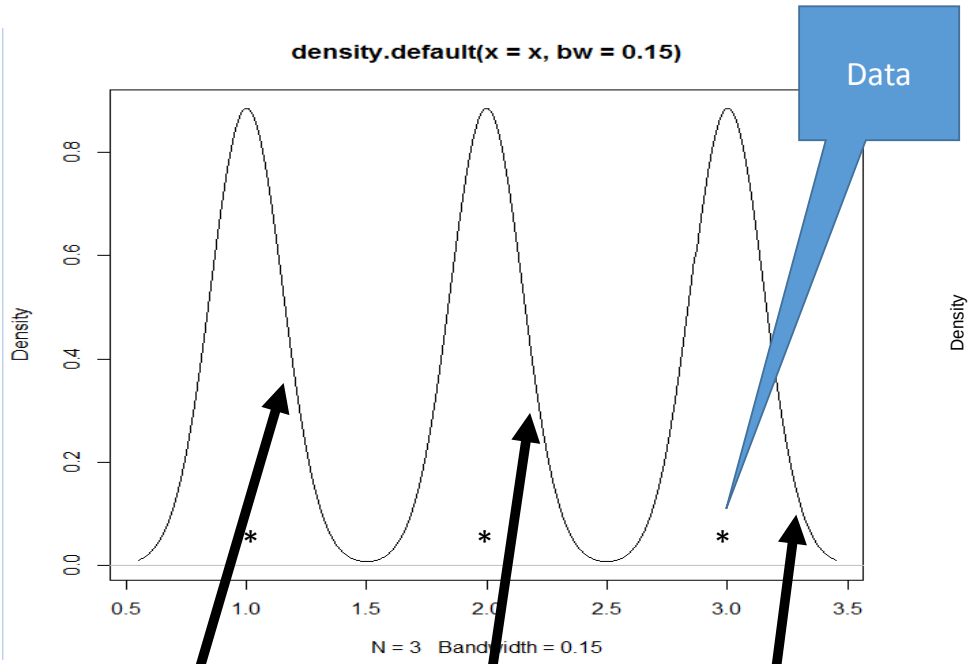
h = 1.0

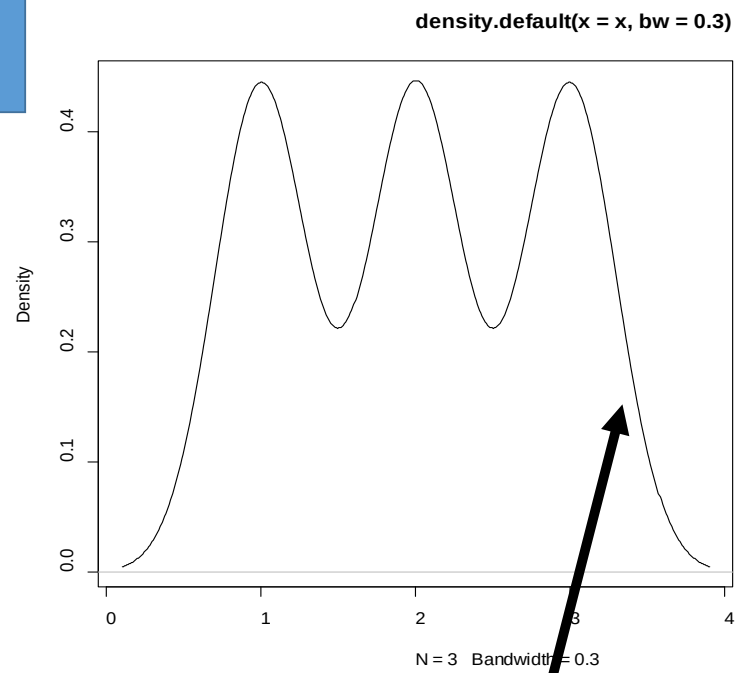


Not Waavy at all, if anything it's probably too smooth.

Increasing band width gives the picture less "Waavyness"

How NonParametric Kernel Density Works





Each Kernel Density is a function of the form

$$\frac{1}{h}k\left(\frac{x-x_1}{h}\right) \quad \frac{1}{h}k\left(\frac{x-x_2}{h}\right) \quad \frac{1}{h}k\left(\frac{x-x_3}{h}\right)$$

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right)$$

Where do Non-Parametric Density Estimates Come From?

Let X be a random variable with continuous distribution $F(x)$ and density $f(x) = \frac{d}{dx}F(x)$. The goal is to estimate $f(x)$ from a random sample $\{X_1, \ldots, X_n\}$.
We now we can estimate the CDF with the ECDF

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$$

Now consider a discrete derivative of $\hat{F}(x)$. For some small $h > 0$, let

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}.$$

Where do Non-Parametric Density Estimates Come From?

- We can write this as

$$\hat{f}(x) = \frac{1}{2nh}\sum_{i=1}^{n} 1(x-h < X_i \leq x+h) = \frac{1}{2nh}\sum_{i=1}^{n} 1\left(\frac{X_i - x}{h} \leq 1\right)$$
$$= \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)$$

  where

$$k(u) = \begin{cases} 1/2 & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

  is the uniform density on $[-1, 1]$.

- Estimator counts the percentage of observations that are within h of x. (Similar to Histogram!!)

Non-Parametric Kernel Density Estimator

The general kernel density estimator can be written as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)$$

where $k(u)$ is the kernel function. The properties of kernel functions are:

- **non-negative** $k(u) \geq 0 \; \forall u$
- $\int_{-\infty}^{\infty} k(u)du = 1$
- **moments** are $\kappa_j(k) = \int_{-\infty}^{\infty} u^j k(u)du$
- **symmetric** $k(-u) = k(u)$
- The **order** of a kernel, ν is the integer corresponding to the first non-zero moment.

Some Common Second-Order Kernels

$$\kappa_1(k) = 0, \quad \kappa_2(k) \neq 0$$

Kernel	Equation	$R(k)$	$\kappa_2(k)$	$\text{eff}(k)$
Uniform	$k_0(u) = \frac{1}{2}1(\lvert u\rvert \leq 1)$	$1/2$	$1/3$	1.0758
Epanechnikov	$k_1(u) = \frac{3}{4}(1-u^2)1(\lvert u\rvert \leq 1)$	$3/5$	$1/5$	1.0000
Biweight	$k_2(u) = \frac{15}{16}(1-u^2)^2 1(\lvert u\rvert \leq 1)$	$5/7$	$1/7$	1.0061
Triweight	$k_3(u) = \frac{35}{32}(1-u^2)^3 1(\lvert u\rvert \leq 1)$	$350/429$	$1/9$	1.0135
Gaussian	$k_\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$	$1/2\sqrt{\pi}$	1	1.0513

Notice that kernels in table are special cases of polynomial family

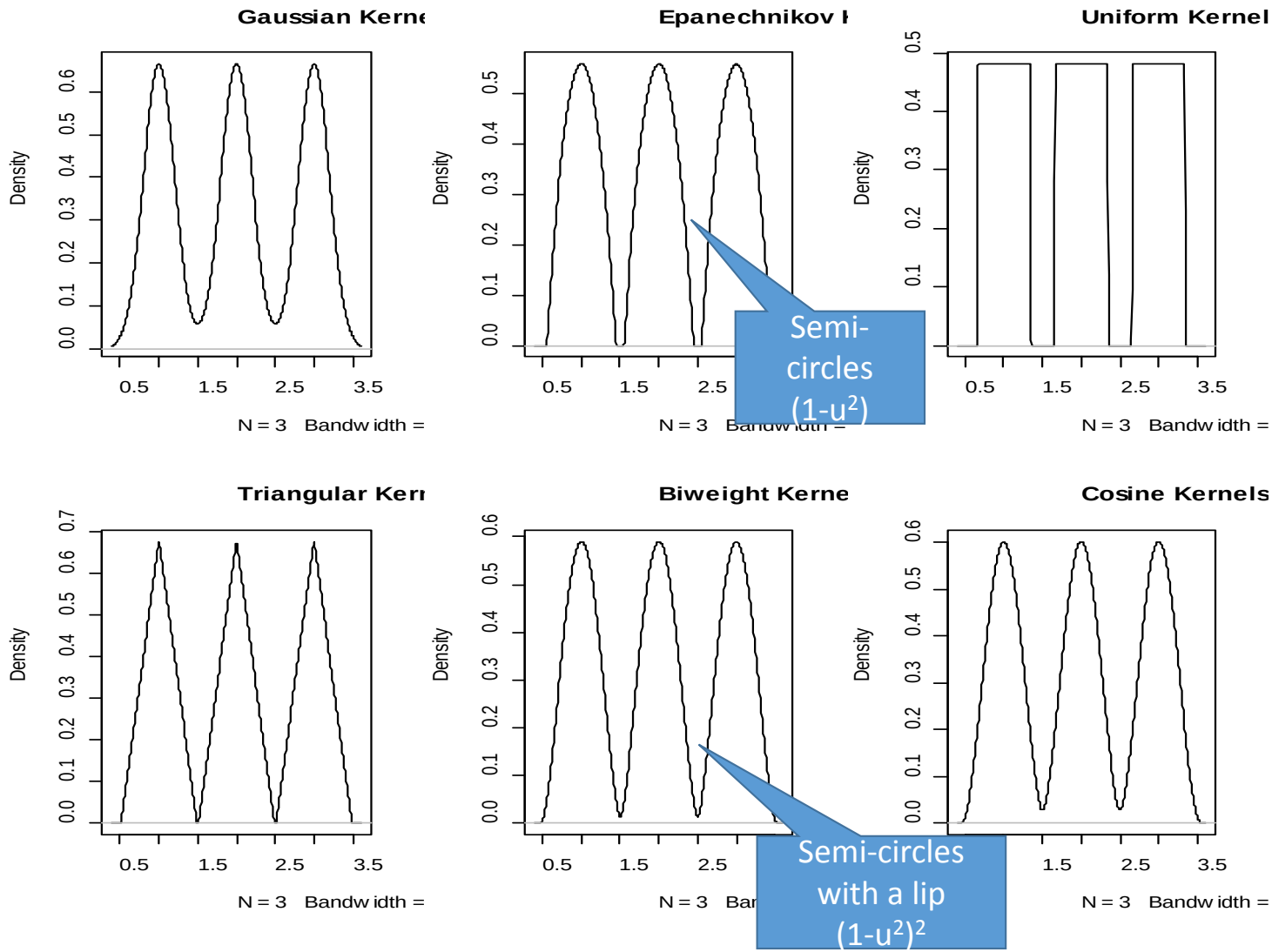
$$k_s(u) = \frac{(2s+1)!!}{2^{s+1}s!}(1-u^2)^s 1(\lvert u\rvert \leq 1)$$

so the Gaussian kernel can be thought of as $\lim_{s\to\infty} k_s(u)$ after rescaling!

R Code for Kernel Density Plots

```
x=c(1,2,3) # data
?density
par(mfrow=c(2,3))
plot(density(x,bw=.2),main="Gaussian Kernel")
plot(density(x,bw=.2,kernel="epanechnikov"),main="Epanechnikov K
plot(density(x,bw=.2,kernel="rectangular"),main="Uniform Kernels
plot(density(x,bw=.2,kernel="triangular"),main="Triangular Kerne
plot(density(x,bw=.2,kernel="biweight"),main="Biweight Kernels")
plot(density(x,bw=.2,kernel="cosine"),main="Cosine Kernels")
```

Shapes of Kernels



Fourth-Order Kernels

Higher-order kernels are obtained by multiplying the second order kernels by a $(\nu/2-1)$ degree polynomial in u^2.

Kernel	Equation	$R(k)$	$\kappa_4(k)$	$\mathrm{eff}(k)$
Epanechnikov	$k_{4,1}(u) = \frac{15}{8}(1-\frac{7}{3}u^2)k_1(u)$	5/4	-1/21	1.0000
Biweight	$k_{4,2}(u) = \frac{7}{4}(1-3u^2)k_2(u)$	805/572	-1/33	1.0056
Triweight	$k_{4,3}(u) = \frac{27}{16}(1-\frac{11}{3}u^2)k_3(u)$	3780/2431	-3/143	1.0134
Gaussian	$k_{4,\phi}(u) = \frac{1}{2}(3-u^2)k_\phi(u)$	$27/32\sqrt{\pi}$	-3	1.0729

$$\kappa_1(k) = \kappa_2(k) = \kappa_2(k) = 0, \quad \kappa_4(k) \neq 0$$

Sixth-Order Kernels

Kernel	Equation	$R(k)$	$\kappa_6(k)$	$\text{eff}(k)$
Epanechnikov	$k_{6,1}(u) = \frac{175}{64}(1 - 6u^2 + \frac{33}{5}u^4)k_1(u)$	1575/832	5/429	1.0000
Biweight	$k_{6,2}(u) = \frac{315}{128}(1 - \frac{22}{3}u^2 + \frac{143}{15}u^4)k_2(u)$	29295/14144	1/143	1.0048
Triweight	$k_{6,3}(u) = \frac{297}{128}(1 - \frac{26}{3}u^2 + 13u^4)k_3(u)$	301455/134468	1/221	1.0122
Gaussian	$k_{6,\phi}(u) = \frac{1}{8}(15 - 10u^2 + u^4)k_\phi(u)$	$2265/2048\sqrt{\pi}$	15	1.0871

$$\kappa_1(k) = \cdots = \kappa_5(k) = 0, \quad \kappa_6(k) \neq 0$$

The **roughness of a kernel** is defined to be

$$R(k) = \int_{-\infty}^{\infty} k^2(u)du$$

Moments of Kernel Densities

We can calculate the moments of the kernel density estimator if we use the transform

$$u = (X_i - x)/h \Longrightarrow x = X_i + uh \Longrightarrow du = hdx.$$

The mean $\mathrm{E}[\hat{f}(x)]$ is given by

$$\begin{aligned}
\int_{=\infty}^{\infty} x\hat{f}(x)dx &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} xk\left(\frac{X_i - x}{h}\right)dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(X_i + uh)k(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i\int_{-\infty}^{\infty}k(u)du + \frac{h}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}uk(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i = \bar{X}.
\end{aligned}$$

Moments of Kernel Densities

The second moment of the kernel density estimator is given by

$$\int_{=\infty}^{\infty} x^2 \hat{f}(x) dx = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} x^2 k\left(\frac{X_i - x}{h}\right) dx$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} (X_i + uh)^2 k(u) du$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i^2 \int_{-\infty}^{\infty} k(u) du + \frac{2h}{n} \sum_{i=1}^{n} X_i \int_{-\infty}^{\infty} u k(u) du + \frac{h^2}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} u^2 k(u) du$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i^2 + h^2 \kappa_2(k).$$

Moments of Kernel Densities

It follows that the variance is given by

$$\int_{=\infty}^{\infty} x^2 \hat{f}(x)dx - \left(\int_{=\infty}^{\infty} x\hat{f}(x)dx\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\kappa_2(k) - (\bar{X})^2$$
$$= \hat{\sigma}^2 + h^2\kappa_2(k).$$

This says that the variance of the density $\hat{f}(x)$ is inflated by a factor of $O(h^2)$ so that

$$\boxed{\mathrm{Var}_{\hat{f}}[X] = \hat{\sigma}^2 + h^2\kappa_2(k)}$$

Estimation Bias

To examine the estimation bias let us look at $\mathrm{E}[\hat{f}(x)]$

$$\mathrm{E}[\hat{f}(x)] = \frac{1}{nh}\sum_{i=1}^{n}\mathrm{E}\left[k\left(\frac{X_i - x}{h}\right)\right] = \int_{=\infty}^{\infty} k(u)f(x+hu)du.$$

Now let's use a local taylor series expansion around x

$$f(x+hu) = f(x) + f^{(1)}(x)hu + \cdots + \frac{1}{\nu^2}f^{(\nu)}(x)h^{\nu}u^{\nu} + o(h^{\nu}).$$

If we now integrate term by term using

$$\int_{-\infty}^{\infty} k(u)du = 1 \text{ and } \int_{-\infty}^{\infty} u^{\nu}k(u)du = \kappa_{\nu}(k)$$

Estimation Bias

we obtain that

$$\mathrm{E}[\hat{f}(x)] = \int_{=\infty}^{\infty} k(u) f(x+hu) du$$

$$= f(x) + hf^{(1)}(x) \int_{=\infty}^{\infty} uk(u)du + \cdots + \frac{1}{\nu!} h^{\nu} f^{(\nu)}(x) \int_{=\infty}^{\infty} u^{\nu} k(u)du + o(h^{\nu}).$$

so if ν is the first non-zero moment of $k(\cdot)$ or **order** of k then

$$\mathrm{E}[\hat{f}(x)] = f(x) + \frac{1}{\nu!} h^{\nu} f^{(\nu)}(x) \kappa_{\nu}(k) + o(h^{\nu})$$

This says that the estimation bias is given by

$$\boxed{\mathrm{Bias}[\hat{f}(x)] = \mathrm{E}[\hat{f}(x)] - f(x) = \frac{1}{\nu!} h^{\nu} f^{(\nu)}(x) \kappa_{\nu}(k) + o(h^{\nu})}$$

Variance

Since the X_i are i.i.d. we see that

$$\begin{aligned}\mathrm{Var}[\hat{f}(x)] &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathrm{Var}\left(k\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{nh^2}\mathrm{E}\left[k^2\left(\frac{X_i - x}{h}\right)\right] - \frac{1}{n}\left(\frac{1}{h}\mathrm{E}\left[k\left(\frac{X_i - x}{h}\right)\right]\right)^2.\end{aligned}$$

Now from the analysis of Bias we found that

$$\mathrm{E}\left[k\left(\frac{X_i - x}{h}\right)\right] = f(x) + O(h^\nu)$$

so the second term is $O(\frac{1}{n})$

Variance

For the first term we see that

$$\begin{aligned}\frac{1}{nh^2}\mathrm{E}\left[k^2\left(\frac{X_i - x}{h}\right)\right] &= \frac{1}{nh^2}\int_{-\infty}^{\infty} k^2\left(\frac{z-x}{h}\right) f(z)dz \\ &= \frac{1}{nh}\int_{-\infty}^{\infty} k^2(u) f(x+hu)du \\ &= \frac{1}{nh}\int_{-\infty}^{\infty} k^2(u)(f(x)+O(h))du \\ &= \frac{f(x)R(k)}{nh} + O(\frac{1}{n})\end{aligned}$$

where $R(k) = \int_{-\infty}^{\infty} k^2(u)du$ is the roughness of the kernel.
Adding the first and second terms together we see that

$$\boxed{\mathrm{Var}\left(\hat{f}(x)\right) = \frac{f(x)R(k)}{nh} + O\left(\frac{1}{n}\right)}$$

Asymptotic Mean Squared Error

Just like the Histogram, we measure the estimation precision by

$$\begin{aligned}
MSE(\hat{f}(x)) &= \mathrm{E}\left(\hat{f}(x) - f(x)\right)^2 \\
&= \mathrm{E}\left(\hat{f}(x) - \mathrm{E}[\hat{f}(x)] + \mathrm{E}[\hat{f}(x)] - f(x)\right)^2 \\
&= \mathrm{E}\left(\hat{f}(x) - \mathrm{E}[\hat{f}(x)]\right)^2 + \left(\mathrm{E}[\hat{f}(x)] - f(x)\right)^2 \\
&= \mathrm{Var}\left(\hat{f}(x)\right) \mathrm{Bias}(\hat{f}(x))^2 \\
&\approx \frac{f(x)R(k)}{nh} + \left(\frac{1}{\nu!} f^{(\nu)}(x) h^{\nu} \kappa_{\nu}(k)\right)^2 \\
&= \mathrm{AMSE}(\hat{f}(x))
\end{aligned}$$

with $\mathrm{AMSE}(\hat{f}(x))$ denoting the **asymptotic mean square error**.

Asymptotic Mean Squared Error

Just like the Histogram, we measure the estimation precision by

$$\begin{aligned}
MSE(\hat{f}(x)) &= \mathrm{E}\left(\hat{f}(x) - f(x)\right)^2 \\
&= \mathrm{E}\left(\hat{f}(x) - \mathrm{E}[\hat{f}(x)] + \mathrm{E}[\hat{f}(x)] - f(x)\right)^2 \\
&= \mathrm{E}\left(\hat{f}(x) - \mathrm{E}[\hat{f}(x)]\right)^2 + \left(\mathrm{E}[\hat{f}(x)] - f(x)\right)^2 \\
&= \mathrm{Var}\left(\hat{f}(x)\right) \mathrm{Bias}(\hat{f}(x))^2 \\
&\approx \frac{f(x)R(k)}{nh} + \left(\frac{1}{\nu!} f^{(\nu)}(x) h^{\nu} \kappa_{\nu}(k)\right)^2 \\
&= \mathrm{AMSE}(\hat{f}(x))
\end{aligned}$$

with $\mathrm{AMSE}(\hat{f}(x))$ denoting the **asymptotic mean square error**.

Asymptotic Integrated Mean Squared Error

Notice that the first term $\propto h^{2\nu}$ and the second term $\propto 1/nh$ so as $n \to \infty$ we want

$$h \to 0 \text{ and } nh \to \infty$$

Notice that $\mathrm{AMSE}(\hat{f}(x))$ is a function of x. A more global measure of of goodness of fit is **Asymptotic Integrated Mean Squared Error (AIMSE)**

$$\begin{aligned} AIMSE &= \int_{-\infty}^{\infty} AMSE(x)dx \\ &= \frac{R(k)}{nh} + \frac{h^{2\nu}}{(\nu!)^2} R(f^{(\nu)})\kappa_\nu^2(k) \end{aligned}$$

where $R(f^{(\nu)}) = \int_{-\infty}^{\infty} (f^{(\nu)})^2(x)dx$ is the roughness of $f^{(\nu)}(x)$.

Optimal Bandwidth

To find the optimal bandwidth we set

$$\frac{d}{dh}(AIMSE) = -\frac{R(k)}{nh^2} + (2\nu)\frac{h^{2\nu-1}}{(\nu!)^2}R(f^{(\nu)})\kappa_\nu^2(k) = 0$$

and solving for h we find that

$$h_{opt} = \left(\frac{(\nu!)^2 R(k)}{2\nu\kappa_{nu}^2(k)R(f^{(\nu)})}\right)^{1/(2\nu+1)} n^{-1/(2\nu+1)}$$

So for second order kernels

$$h \propto n^{-1/5}$$

Optimal Bandwidth

If we plug in the optimal bandwidth into AIMSE we find that, when h is optimal

$$\text{AIMSE}_{opt}(k) = (1+2\nu)\left(\frac{R(f^{(\nu)})\kappa_\nu^2(k)R^{2\nu}(k)}{(\nu!)^2(2\nu)^{2\nu}}\right)^{1/(2\nu+1)} n^{-2\nu/(2\nu+1)}$$

So for second-order kernels

$$\text{AIMSE}_{opt}(k) \propto n^{-4/5}$$

that's a rate just under n^{-1}. That's pretty good for non-parametric techniques. Notice that for $\nu = 4$ or fourth order kernels we have

$$\text{AIMSE}_{opt}(k) \propto n^{-8/9}.$$

With parametric techniques we typically have variance $\propto n^{-1}$

Optimal Bandwidth

Seems a bit magical, if ν get's large rate approaches parametric rate, so why not choose high order kernels every time? Catch is that higher-order kernels need existence of higher order derivatives. As the density gets increasingly smooth the problem gets closer to parametric problem. The other catch is that there is some evidence that the benefits of higher order kernels only kick in when the sample size is large. Suggestion:

- For small n use 2^{nd} order kernels.
- For moderate n use 4^{th} order kernels.
- For large n use 6^{th} order kernels.

Histograms vs Non-Parametric Kernel Density Estimators

Recall for a histogram we had

$$AMSE(x) = \frac{f(x)}{nh} + f'^2(x)h^2$$

thus

$$AIMSE = \int_{-\infty}^{\infty} AMSE(x)dx = \frac{1}{nh} + R(f^{(1)})h^2.$$

Now the optimal bin width was

$$h_{opt} = \left(\frac{1}{2R(f^{(1)})}\right)^{1/3} n^{-1/3}$$

thus

$$AIMSE_{opt} = (\frac{3}{2})(2R(f^{(1)}))^{1/3} n^{-2/3} \propto n^{-2/3}$$

So Histograms converge slower than non-parametric kernel density estimators!! Histograms are probably better when n is small.

Which Kernel Function is Best?

- This is a **Calculus of Variations** Problem. Mathematically the problem can be stated as: Choose the kernel $k(\cdot)$ among the class of all possible kernel functions which minimizes $AIMSE(k)$ where k is subject to certain moment conditions.
- The problem amounts to minimizing $R(k)$ subject to constraints $\int_{-\infty}^{\infty} k(u) = 1$ and $\kappa_\nu(k) = \int_{-\infty}^{\infty} u^\nu k(u) = 1$. (See Muller, Annals of Statistics, 1984).
- The result is the Epinechikov kernel $k_{\nu,1}$!!
- The asymptotic relative efficiency of other kernel functions are defined to be

$$\begin{aligned} \text{eff}(k) &= \left(\frac{AMISE_{opt}(k)}{AMISE_{opt}(k_{\nu,1})} \right)^{(1+2\nu)/2\nu} \\ &= \frac{(\kappa_\nu(k))^{1/2\nu} R(k)}{(\kappa_\nu(k_{\nu,1}))^{1/2\nu} R(k_{\nu,1})} \end{aligned}$$