Motivation for Hamilton Chapter 1

- ***Describing the variability*** and the observed distribution of data is the ***required*** first step of any data analysis.
- The ***shape*** of an univariate distribution can have ***substantial impact*** on the outcome of statistical procedures.
  E.g.: ***Outliers*** or ***heavy tails*** may detrimentally influence the outcome of model calibrations and estimations.
- Not accounting for the distribution of variables can force a researcher to redo their data analysis at a later state.
- Most methods assume ***symmetric*** or preferably ***normally*** distributed variables.
- Remember:
  - Data tell a story about the phenomena under investigation.
  - Always handle data and analysis results with a critical attitude and use common sense.
  - Always ask yourself: Do the data or the generated analysis results make sense?
- Transformations to symmetry are discussed in Chapter 1. Note, statisticians use many more transformations under in particular circumstances.
  E.g., we will encounter later the logit-transformation.

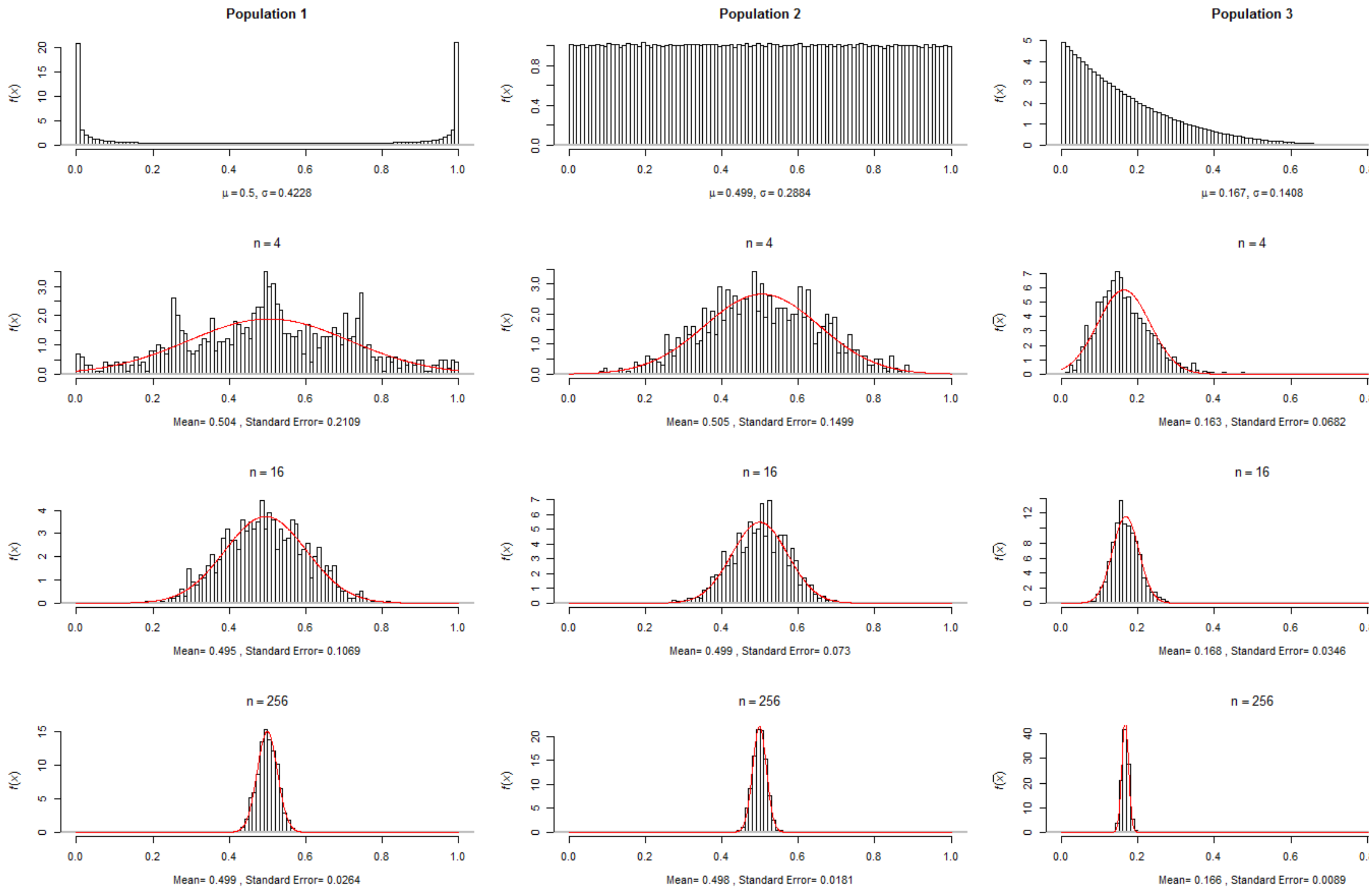
Central Limit Theorem (skipped)

- Def. Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be a ***random independent*** sample of size *n* drawn from an ***arbitrarily distributed*** population with expectation μ and standard deviation σ.

  Then for large enough sample sizes *n*, the sampling distribution of $\bar{X}$ is [a] asymptotically (i.e., as $n \to \infty$) normal distributed [b] with $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.

  Proof for independent sample objects: $Var\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \underbrace{Var(X_i)}_{=\sigma^2} = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$

- <u>Example:</u> Central limit theorem with the R-script **CENTRALLIMIT.R**:



Review: The Shape of Distributions (skipped)

- Distributions can be distinguished with regards the ***balance*** of their left and right tails:
  - ***Symmetric*** distributions. Tails are balanced into either direction from a central value.
  - ***Negatively*** skewed distributions (long tail into the negative direction)
  - ***Positively*** skewed distributions (long tail into the positive direction). These distributions frequently emerge for variables with a binding lower origin (like zero income).
  - Extreme skewness may hint at ***outliers*** that do not match the rest of the observed data.
- The number of meaningful clusters of observations is described by the term modality:
  - Uni-modality refers to just one peak
  - Bi-modality refers to two outstanding peaks
  - Multimodality refers to more than two outstanding peaks.
- Multimodality may hint at a heterogeneous underlying data generating process.

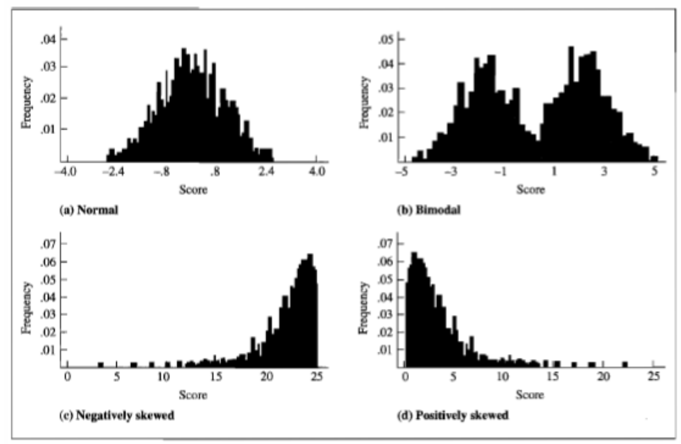


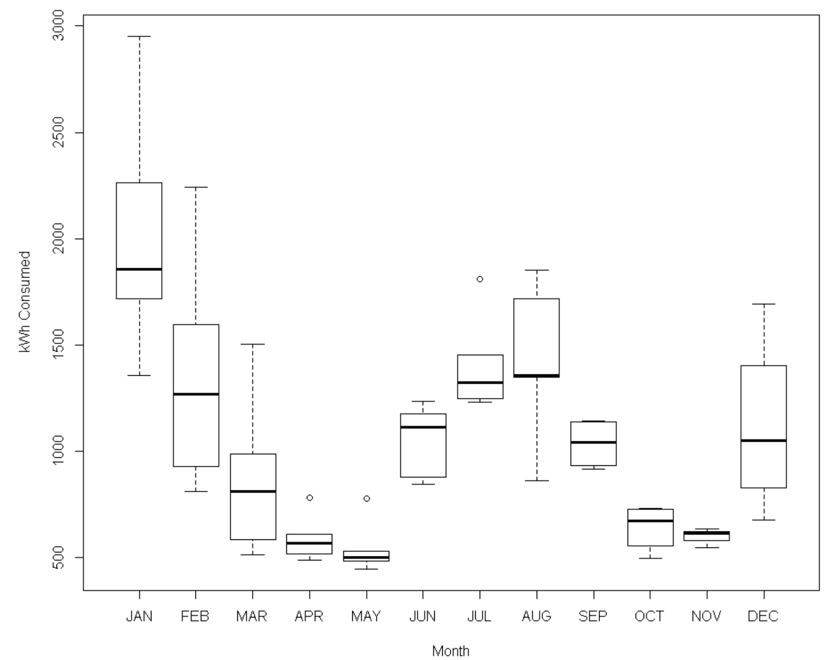
Figure 3.9
Shapes of frequency distributions: (a) Normal; (b) Bimodal; (c) Negatively skewed; (d) Positively skewed

Quantiles and Percentiles (skipped)

- Technically, quantiles and percentiles are generated from a ***sorted list*** of the original data points $x_{[1]} \leq x_{[2]} \leq x_{[3]} \leq \cdots \leq x_{[n-1]} \leq x_{[n]}$ where each observations has an assigned rank $i \in \{1, 2, \ldots, n\}$, with $i = 1$ for the smallest observation and $i = n$ for the largest observation.
- For a give data value $\boldsymbol{x}_{[i]}$ the ***percentile*** approximates the proportion of sample observations less or equal to $\boldsymbol{x}_{[i]}$, that is, $p_{[i]} = \frac{i-\frac{1}{2}}{n} \approx \Pr\left(X \leq x_{[i]}\right) = \int_0^{x_{[i]}} f(x) \cdot dx$.
- A ***quantile*** is that observed data value of a distribution, which is associated with a particular percentile point.
- Important quantiles are:
  - 0.25 quantile also called Q_1 quartile (25 % of the observations are smaller or equal to this quantile value)
  - 0.50 quantile also called the median (50 % of the observations are smaller or larger than the given quantile value)
  - 0.75 quantile also called Q_3 quartile (75 % of the observations are smaller or equal to this quantile value and 25 % of the observations are larger than this value)
  - A measure of spread is the inter-quartile range: $IQR = Q_3 - Q_1$

Box-Plots (skipped)

- Construction of the box-plot
  - Draw a ***box*** from Q_1 to Q_3. Mark the ***median*** $\boldsymbol{Q_2}$ in the center of the box with a line.
  - Definition of ***adjacent values*** $x_{low}^{adj} = \min\left(x_{[i]} \in \left(Q_1, Q_1 - 1.5 \cdot IQR\right) \text{ plus } x_{[i]} \text{ in dataset}\right)$ and $x_{high}^{adj} = \max\left(x_{[i]} \in \left(Q_3, Q_3 + 1.5 \cdot IQR\right) \text{ plus } x_{[i]} \text{ in dataset}\right)$.
    The term $x \in (a,b)$ means, all x-values in the interval between a and b.
    Draw the "fences" so they just include the smallest and largest data values x_{low}^{adj} and x_{high}^{adj}, respectively.
  - ***Outliers*** are in the interval $\left[1.5 \cdot IQR, 3.0 \cdot IQR\right]$ starting from Q_1 below or Q_3 above, respectively.
    Severe outliers are beyond that range $(> 3.0 \cdot IQR)$



- Use of box-plots:
  - Easy visual description of the distribution of a variable and potential outliers

- Comparison of distributions for several variables side-by-side.

Quantile-Normal Plot

- Calculate the theoretical quantiles of a normally distributed random variable $X_{[i]}$ (assuming the mean μ and the variance σ^2 were estimated from the sample data) based on the given percentiles $p_{[i]}$ of the observed variable $Y_{[i]}$.
- ***Quantile-Normal Plot***: Plot the theoretical normal distribution quantiles $X_{[i]}$ on the abscissa (X-axis) against their matching empirical distribution of $Y_{[i]}$ on the ordinate (*Y*-axis).
  Interpretation:
    - Diagonal with slope 1 => equal distributions;
    - Not a straight-line => different shapes

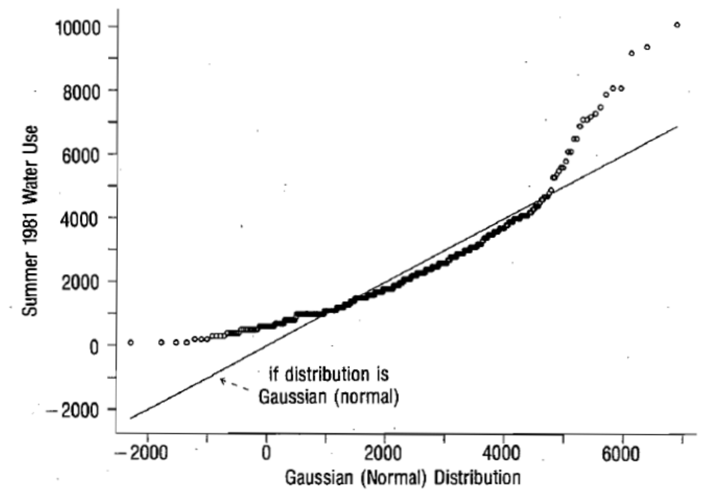


Figure 1.9 Quantile-normal plot of household water use (positively skewed).

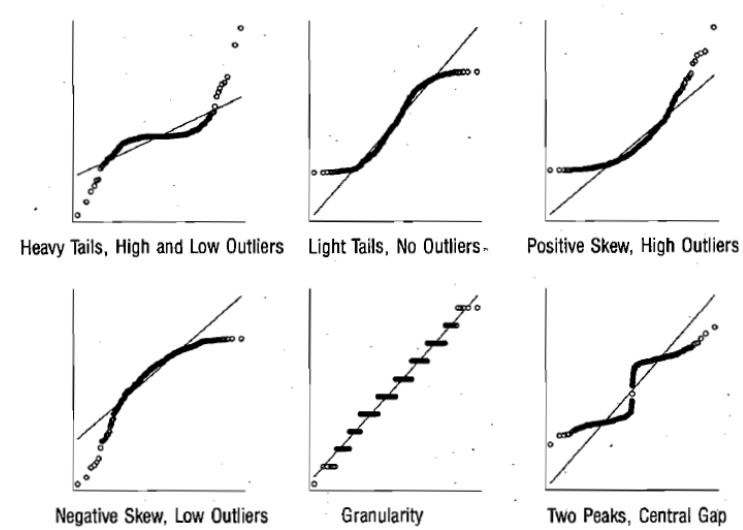


Figure 1.10 Quantile-normal plots reflect distribution shape.

Properties of Arithmetic Mean (skipped)

- Implications of the ***zero-sum*** property $\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right) = 0$: Assuming the mean is known, then *n-1* observation can vary freely, whereas we can predict the last observation with certainty.

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right) = 0$$
$$\Rightarrow \sum_{i=1}^{n} Y_i = n \cdot \bar{Y}$$
$$\Rightarrow Y_n = n \cdot \bar{Y} - \sum_{i=1}^{n-1} Y_i$$

That implies that we loose on degree of freedom.

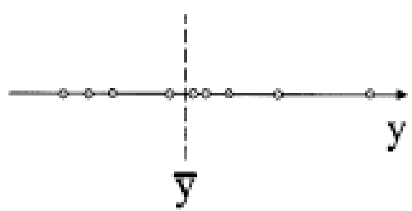
- Implication of the ***least squares property*** $\min_{\theta} \sum_{i=1}^{n}(Y_i - \theta)^2 \Rightarrow \theta = \bar{Y}$.

  Large deviations have a strong impact on the estimated mean, variance etc. because the large deviations are squared

  $\Rightarrow$ Thus, large deviations pull the mean into their direction.

  $\Rightarrow$ Standard deviations are drastically inflated.

- Lacking any other information, the arithmetic mean will become best ***predictor*** for the variable under question.
- The deviations from the mean are the ***unexplained*** part or the ***residuals*** of the observations, i.e., $y_i = \bar{y} + \varepsilon_i$.



- Definition of total sum of squares: $TSS = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$ or $TSS = \sum_{i=1}^{n} Y_i^2 - n \cdot \bar{Y}^2$.

variance

- Why is the population variance estimated $(n-1)$ in the denominator, that is, by $s^2 = TSS/(n-1)$:
  Explanation 1: If we calculate the mean from the sample then there are only $n-1$ "degrees of freedom" left because of the ***zero sum property*** of the mean.
  Explanation 2: The mean is calculated to minimize the *TSS*.
  Thus the sample mean always fits the observed sample data better than any ***unobserved but true*** population expectation μ.
  For the true expectation μ, the TSS would be slightly larger. That is why the sample TSS needs to be inflated by dividing it by a slight smaller value than n, that is, $n-1$.
- ***Standard deviation*** measures the variation in ***original units*** rather than in squared units.

Review: Skewness (skipped)

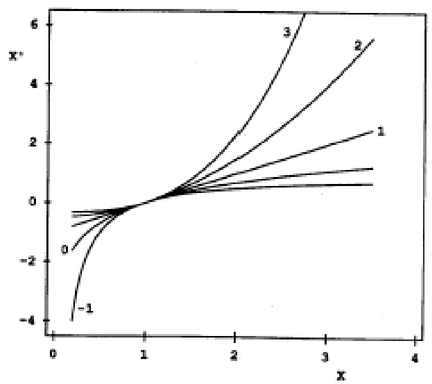
- Why does the distribution of the water consumption in the Concord dataset deviate from the normal distribution?
  Reason: Fixed lower bound (negative consumption impossible).
- ***Skewness*** and bounded/truncated distributions: For skewed distributions the notion of the center of the distribution (mean) becomes ambiguous and the ***median*** may be a better representation of the central tendency in the data.
- The ***skewness*** is defined by $skew(X) \equiv \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n \cdot s_X^3}$
- The normal distribution has a skewness of 0.

Box-Cox Transformation

- This lecture focuses on the more general *Box-Cox* transformation rather than the slightly simpler *power*-transformation, which is discussed in Hamilton.
  For both transformations the general interpretation of the parameter λ does not change.
- Causes for ***extreme observations***: [a] skewed distributions, [b] measurement or recoding errors, [c] extreme but feasible events (perhaps not belonging to the population under investigation).
- The *power*-transformation presented in book and the Box-Cox transformation only work for **X's** ***larger than zero***.
- Thus the standard Box-Cox transformation will not work for data values X that are zero or negative.
  - In order to avoid this problem, a constant such as $\min(X)$ or, say, 5% quantile, needs to be added to X with negative values, in order to make it positive.
  - However, if the constant is too small leading to positive but close to zero values, outliers may be introduced.
  - On the other hand, choosing the constant too large, may make the transformation to normality ineffective.
- The Box-Cox transformation is a generalization of the power transformation: $Y = \dfrac{X^{\lambda} - 1}{\lambda}$ and for $\lambda = 0 \text{ we get } y = \ln(x)$. See also note 11 on page 28 in Hamilton
- $\lambda > 1$ reduce negative skewness, whereas $\lambda < 1$ reduce positive skewness.
  Remember: Positive skewness is very common for variables with a natural bound of zero.

- If power $\lambda < 0$ then all values are multiplied by a negative number to preserve the order of observations.
  This explains the value λ in the denominator of the Box-Cox transformation

FOX Fig 4.1



X	X^{-1}	$\frac{X^{-1}}{-1}$
1	1	-1
2	1/2	$-1/2$
3	1/3	$-1/3$
4	1/4	$-1/4$

Figure 4.1. The family of power transformations X' of X. The curve labeled p is the transformation $X^{(p)}$, that is, $(X^p - 1)/p$; $X^{(0)}$ is $\log_e X$.

- Note: R’s function **`car::powerTransform( )`** is performing several statistical tests whether a variable either needs to be transformed or whether a *log*-transformation is sufficient by using the likelihood ratio test (LR) principle:
  - The first LR tests the null hypotheses $H_0: \lambda^{optimal} = 0$. If we cannot reject the null hypothesis then we should tentatively work with a *log*-transformation to achieve normality/symmetry.
  - The second LR tests the null hypotheses $H_0: \lambda^{optimal} = 1$. If we cannot reject the null hypothesis then we should tentatively should work with an untransformed variable because it is approximately symmetric.

  - The Wald confidence interval provides the 95% probability range within which the true population transformation parameter λ lies.

Handling Transformations with Negative Data Values

- In case some data values are negative or zero a small constant γ can be added to the data X (see the **`?car::bcPower( )`** and Fox & Weissberg pp 161-162 for the **`bcnPower`** transformation family)
- A more informed way avoiding some of the problems by just adding a constant is to first transform the data by:

$$z(X,\gamma) = \frac{\left(X + \sqrt{X^2 + \gamma^2}\right)}{2} \text{ with}$$

  - The transformation $z(X,\gamma)$ is monotonic (i.e., if $x_1 < x_2$.then $z(x_1,\gamma) < z(x_2,\gamma)$)
  - For large positive X relative to γ ($X \gg \gamma$) the transformation is approximately linear with $z(X,\gamma) \approx X$.
  - If $\gamma = 0$ then $z(X,\gamma) = X$ for $X > 0$ and $z(X,\gamma) = 0$ for $X \leq 0$.
- Subsequently, once the γ-parameter is determined a standard Box-Cox transformation is applied to $z(X,\gamma)$.

LOESS Smoother of Y~X Relationships

- Many of R’s scatterplot functions not only show a linear regression fit through the data cloud but also show a locally smoothed loess-curve:
    - In essence the sliding window moves over the value range of X.
    - In each window a local regression line is estimated.
    - These local regression lines are “splined” together into the smooth loess curve over the whole value range of X

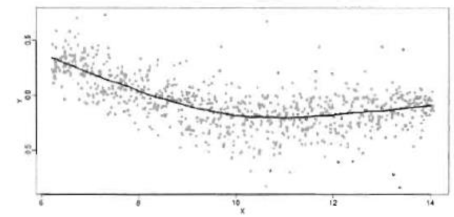


FIGURE 9.17
MA-plot with curve obtained with loess.

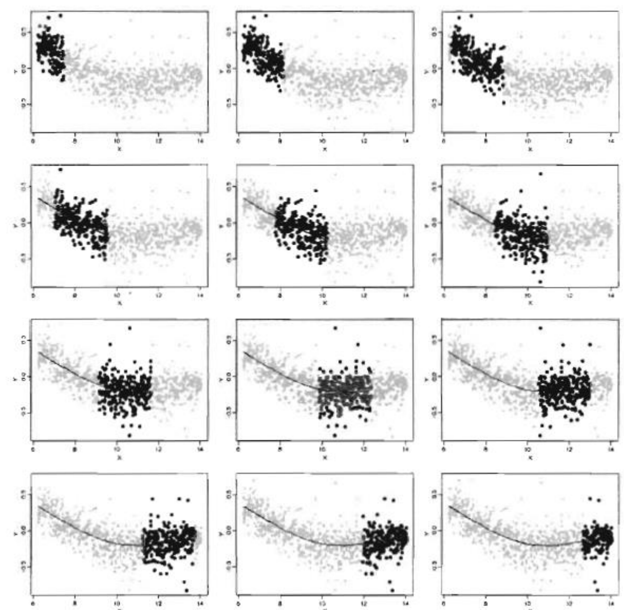


FIGURE 9.16
Illustration of how loess estimates a curve. Showing 12 steps of the process.