Keele University

# Statistics & Experimental Design with R

Barbara Kitchenham
Keele University

# Descriptive Statistics

Part 3

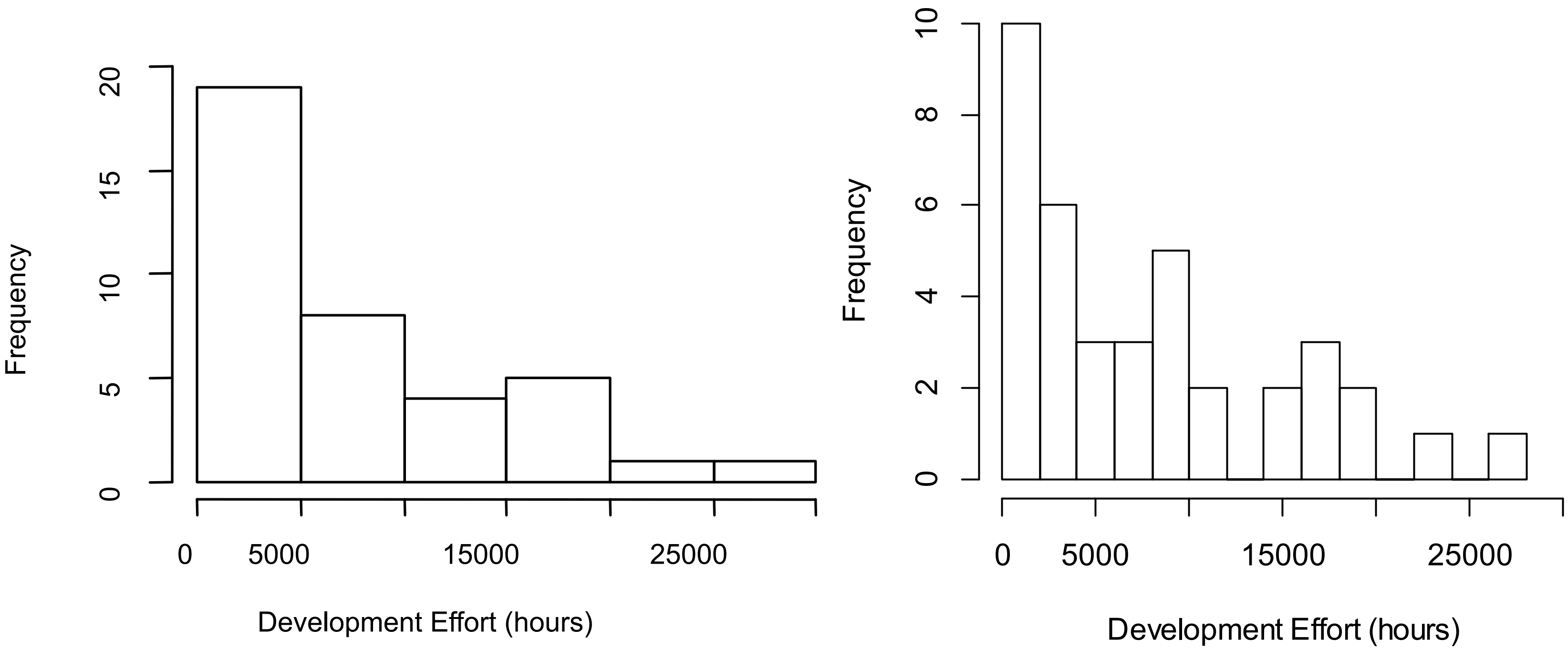
# Aim

- Visualise a data set
  - Understand nature and limitations
- Identify basic descriptive statistics
  - Statistics of Central Tendency/Location
  - Statistics of Dispersion Scale
  - Standard error of Location metrics

# Visualising Distribution of Sample

- Histogram
  - Represents the frequency distribution by "discretising" the sample range into bins
  - Calculating the proportion of sample values in each bin
- Box plots
  - Shows central tendency, dispersion and skewness
- Kernel Density Estimators
  - Smooth histograms to represent continuous frequency function

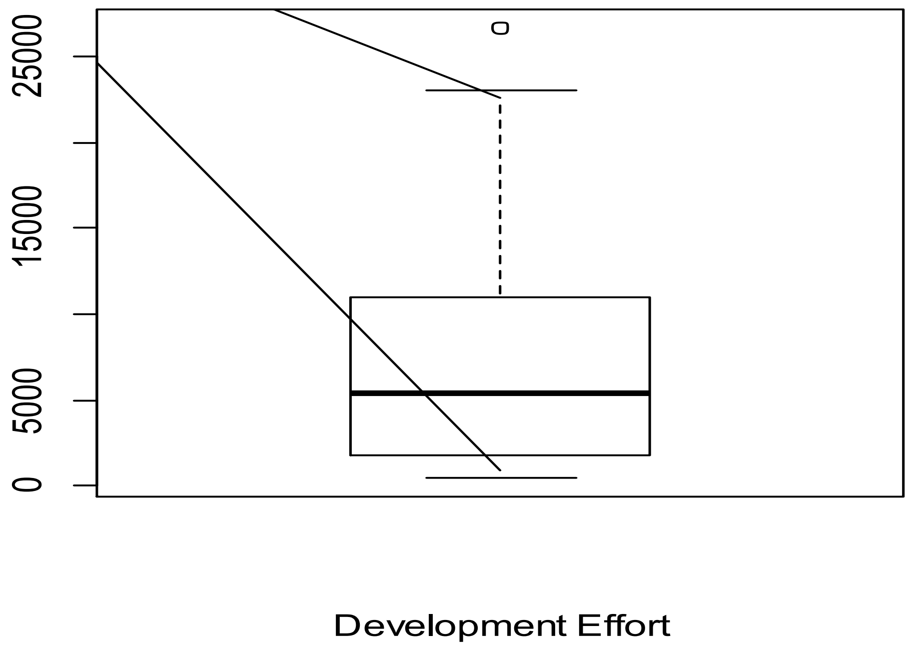
# Example Histograms



# Histograms

- Give an indication of shape of frequency distribution
- Indicate whether data are symmetric or skewed
- Depend on bin width
  - Suggested bin size $2 \times IQR \times n^{-1/3}$
    - Interquantile range ISQ=75%ile-25%ile
- Properties
  - Not smooth
  - Dependent on end point of bins
  - Depend on width of bins

# Box Plot




- Box length = Interquartile length
- Line through box = median
- Upper Tail= 1.5×Box_Lenth + 75%ile rounded down to nearest data point
- Points outside upper and lower tails – outliers

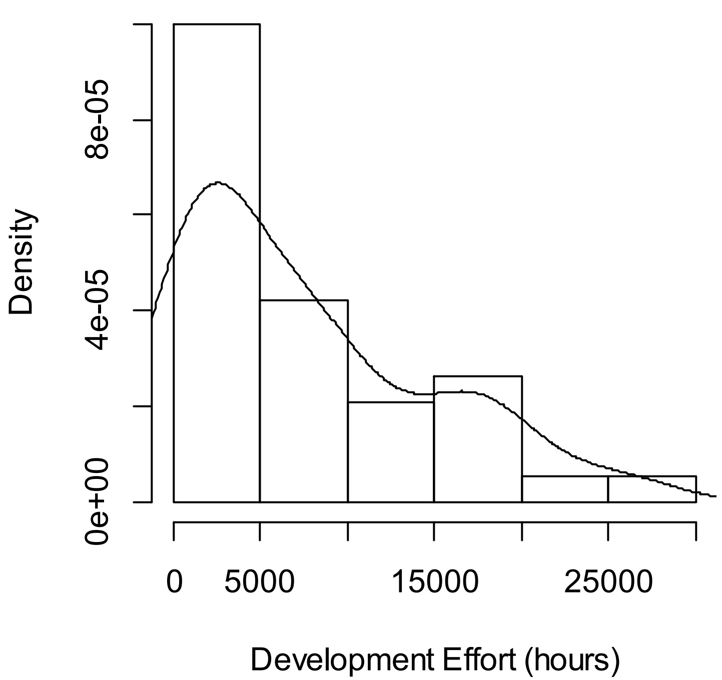
# Kernel Density

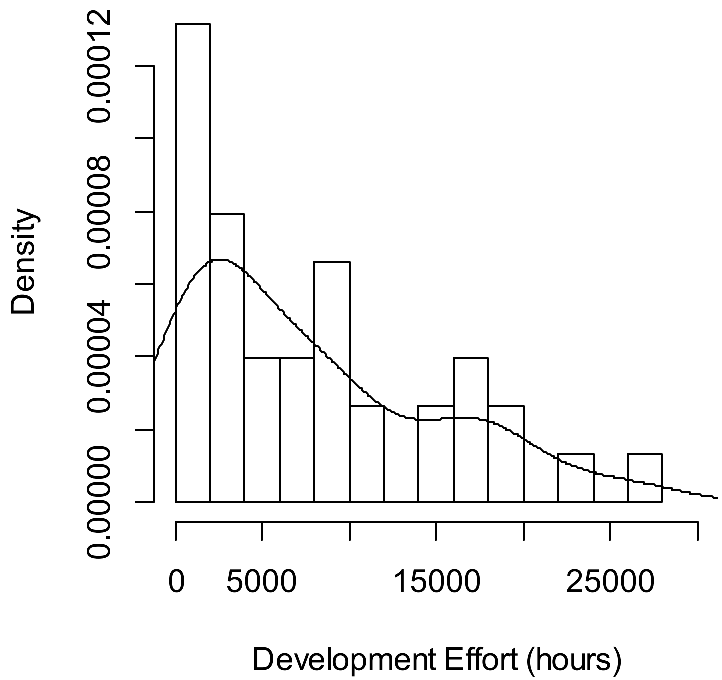
- Let ($x_1$,$x_2$,…,$x_3$) be iid (independently and identically distributed) with unknown density f, it kernel density is

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

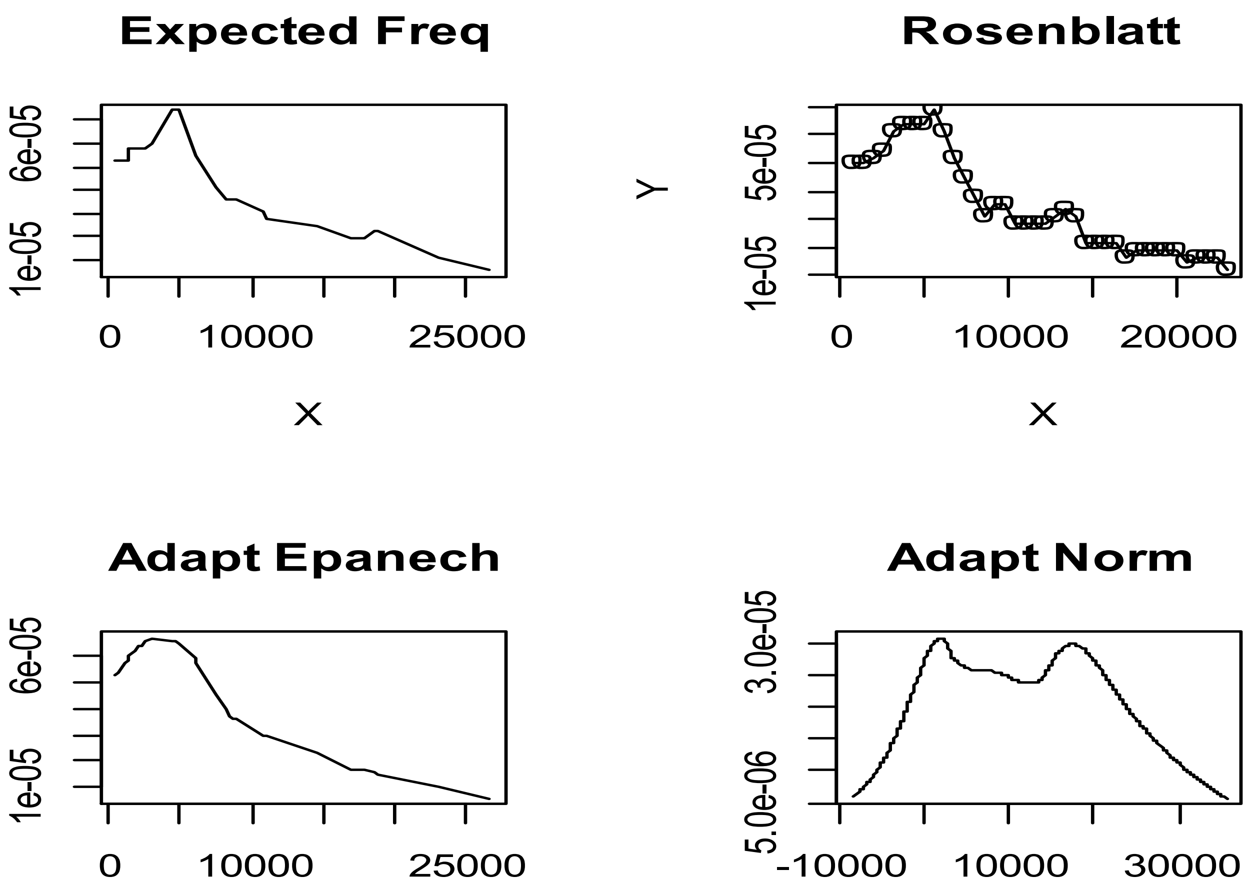
  - h>0 is a smoothing parameter called the bandwidth (which should be a small as the data allow)
- There are many kernel functions
  - Uniform, biweight, Epanechnikov, Normal

# Example Kernel Density



# Other Kernels



Keele University

# Kernel density estimators

- Provide "smoothed"  depiction of histogram
- Doesn't depend on bin end
- Does depend on "bandwidth"
  - Equivalent to histograms bin width
- Provide a non-parametric estimate of the unknown probability density function
- Importance of Kernel Density Estimators
  - Parameters and their standard errors can be estimated from empirical function rather than data
    - Using numerical methods for integration and differentiation
  - Idea generalises to multivariate datasets
  - Mostly used for regression problems

# Central Tendency/Location

- Mean (average) = $(\Sigma x_i)/N$
  - Very vulnerable to large values
- Median (50 percentile)
  - Very stable
  - Based putting measurement into ascending order
    - If N is even, median = $(x_{(N/2)}+x_{(N/2)+1})/2$
    - If N is odd, median =$x_{(N+1)\backslash 2}$
- Trimmed mean , based on mean of values after
  - M% of upper & lower values removed
- Winsorized mean , based on mean of values after
  - M% of upper and lower values replaced with upper & lower values respectively
- Geometric mean for proportions $m_g = \sqrt[n]{\Pi_{\mathrm{i}=1}^{\mathrm{n}}(p_i)}$

# Trimmed Means

- A robust measure of central tendency
- Remove X% smallest & largest values
  - Usually X=20
- Meant to be a compromise between
  - Mean include all values
  - Median excluding all but one or two
    - i.e. maximally trimmed estimate of central tendency
- Windorized means
  - Find X%ile and 100-X%ile values
    - Usually 20 percentile and 80 percentile
    - Replace lowest 20% with 20 percentile value
    - Replace largest 20% with 80 percentile value
  - Take average of amended data set

# Location Metrics for Data set

| Metric | Value |
|---|---|
| Mean | 7678.289 |
| Median | 5430 |
| 20% Trimmed Mean | 6123.458 |
| 20% Windsorized mean | 6796.026 |
| Geometric mean | 4431.826 |

- Geometric mean=$e^m$
  - where m=Mean(LN($X_i$))
- Mean of LN transformed observations

# Measure of Scale/Dispersion

- Variance (Squared standard deviation)
  - Average Squared difference between measure and its mean $=\Sigma(x_i-m)^2/(N-1)$
    - Very vulnerable to large values
  - Also versions for trimmed & Winsorized samples
    - Less vulnerable to large values
- Median absolute deviation (MAD) $= \Sigma|x_i-M|/N$
  - Normal distribution MAD$\sim z_{0.75}\sigma$
- Interquantile range = 75 percentile-25 percentile
- Variance for trimmed/Winsorized means

# Scale Metrics

| Metric | Value |
|---|---|
| Sample Variance | 50912220 |
| Sample Standard deviation | 7135.28 |
| Standard error of mean | 1157.495 |
| 20% Trimmed Mean SE | 1414.929 |
| 20% Windsorized Mean SE | 1365.714 |
| Interquartile Range | (1750,11023) |
| Median Absolute Deviation (for Normal data MAD=0.6745σ) | 4037.494 |

# Standard error of Median? -1

- McKean Schrader

$$Let\ k = \frac{n+1}{2} - z_{0.995}\sqrt{\frac{n}{4}} \qquad z_{0.995} = 2.5758$$

  - Round k down to nearest integer
  - Put observations in order $X_{(1)}$, $X_{(2)}$, … X(n)
  - Estimate of SE of Median is

$$s_M^2 = \left(\frac{X_{(n-k+1)} - X_k}{5.1517}\right)^2$$

  - 1300.552
- Generally recommended but
  - Has problem if there are tied values in data set
  - There are two ties values in data set

# Standard error of Median? -2

- Maritz-Jarrett Estimate
  - Based on a beta function
  - 1841.72
- Kernel density method
  - 1269.011 (Rosenblatt's shifted histogram)
  - 1093.338 (expected frequency curve)
  - 1242.662 (adaptive kernel)
- Bootstrap
  - 1790 (1000 bootstrap samples)