VE401 Probabilistic Methods in Eng.
RC 4

CHEN Xiwen

UM-SJTU Joint Institute

March 27, 2020

Table of contents

Reliability

Failure Density, Reliability and Hazard Rate

Weibull Distribution

Basic Statistics

Samples and Data

Estimating Parameters

Estimating Intervals

Case Study

Reliability

Failure Density, Reliability and Hazard Rate

Weibull Distribution

Basic Statistics

Samples and Data

Estimating Parameters

Estimating Intervals

Case Study

Definitions

Suppose A is a black box unit.

- ***Failure density*** f_A: distribution of the time T that A fails.
- ***Reliability function*** R_A: the probability that A is working at time t, $R_A(t) = 1 - F_A(t)$.
- ***Hazard rate*** ρ_A:

$$\rho_A(t) := \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t | t \leq T]}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t]}{P[T \geq t] \cdot \Delta t} = \frac{f_A(t)}{R_A(t)},$$

$$R_A(t) = e^{-\int_0^t \rho_A(x) \mathrm{d}x}.$$

One often has information on ρ_A, but not F_A or R_A.

Series and Parallel Systems

- Series system with k components.

$$R_s(t) = \prod_{i=1}^{k} R_i(t),$$

where R_i is the reliability of the i-th component.

- Parallel system with k components.

$$R_p(t) = 1 - \prod_{i=1}^{k} (1 - R_i(t)).$$

Reliability
Failure Density, Reliability and Hazard Rate
Weibull Distribution

Basic Statistics
Samples and Data
Estimating Parameters
Estimating Intervals
Case Study

Weibull Distribution

- Density function. $\alpha, \beta > 0$ are parameters,

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

- Mean.

$$\mu = \alpha^{-1/\beta}\Gamma(1 + 1/\beta).$$

- Variance.

$$\sigma^2 = \alpha^{-2/\beta}\Gamma(1 + 2/\beta) - \mu^2.$$

- Reliability features.

$$\rho(t) = \alpha\beta t^{\beta-1},\ R(t) = e^{-\alpha t^{\beta}},\ f(t) = \rho(t)R(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}.$$

Reliability
  Failure Density, Reliability and Hazard Rate
  Weibull Distribution

Basic Statistics
  Samples and Data
  Estimating Parameters
  Estimating Intervals
  Case Study

Definitions

- ***Statistics*** aims to gain information about the parameters of a distribution by conducting experiments.
- ***Population***: a large collection of instances which we want to describe probability.
- ***Random sample of size n from distribution of X***: a collection of n independent random variables $X_1, \ldots, X_n$, each with the same distribution as X. ($\Leftrightarrow$ n i.i.d. random variables.)
- ***x-th percentiles***: d_x such that $x\%$ of values in sampled data are less than or equal to d_x. (***first, second, third quartile*** $\Rightarrow$ $x = 25, 50, 75$.)
- ***Interquartile range***: $\mathrm{IQR} = q_3 - q_1$, measures the dispersion of the data.
- ***Precision***: smallest decimal place of data $\{x_1, \ldots, x_n\}$.
- ***Sample range***: $\max\{x_i\} - \min\{x_i\}$.

Visualization — Histograms

Choose bin width / number of bins.

- Sturges's rule.

$$k = \lceil \log_2(n) \rceil + 1, \qquad h = \frac{\max\{x_i\} - \min\{x_i\}}{k},$$

  rounding ***up*** to the precision of the data.

- Freedman-Diaconis rule.

$$h = \frac{2 \cdot \mathrm{IQR}}{\sqrt[3]{n}}.$$

Sketch.

1. Choose bin width h.
2. Find minimum of data $\min\{x_i\}$, subtract $1/2$ of precision.
3. Successively add bin width and categorize all the data.

Visualization — Stem-and-Leaf Diagrams

1. Choose a convenient number of leading decimal digits to serve as stems.
2. Label the rows using the stems.
3. For each datum of the random sample, note down the digit following the stem in the corresponding row.
4. Turn the graph on its side to get an impression of its distribution.

Visualization — Stem-and-Leaf Diagrams

```
Stem | Leaves
-----+----------------------------------------------------------
   0 | 0000000111112222222222233334444445555566666777777888899999
   1 | 00011111223344444455555678899
   2 | 223669
   3 | 012456
   4 |
   5 | 2
   6 | 8
Stem units: 100
```

Visualization — Boxplots

1. Calculate q_1, q_2, q_3 and TQR.
2. Find ***inner fences*** and ***outer fences*** by

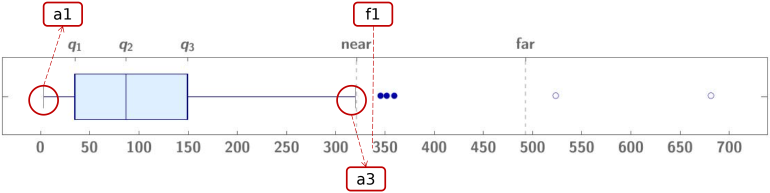
$$f_1 = q_1 - \frac{3}{2}\mathrm{IQR}, \qquad f_3 = q_3 + \frac{3}{2}\mathrm{IQR},$$
$$F_1 = q1 - 3\mathrm{IQR}, \qquad F_3 = q_3 + 3\mathrm{IQR},$$

   and find ***adjacent values***

$$a_1 = \min\left\{x_k : x_k \geq f_1\right\}, \qquad a_3 = \max\{x_k : x_k \leq f_3\}.$$

3. Identify ***near outliers*** and ***far outliers***.

Visualization — Boxplots



Reliability
- Failure Density, Reliability and Hazard Rate
- Weibull Distribution

Basic Statistics
- Samples and Data
- Estimating Parameters
- Estimating Intervals
- Case Study

Definitions

- ***Statistic***: a random variable that is derived from $X_1, \ldots, X_n$.
- ***Estimator***: a statistic that is used to estimate a population parameter.
- ***Point estimate***: a value of the estimator.
- ***Unbiased***: expectation of an estimator $\widehat{\theta}$ is equal to the true parameter.

$$\mathsf{E}[\widehat{\theta}] = \theta, \qquad \text{bias} = \theta - \mathsf{E}[\widehat{\theta}].$$

- ***Mean square error***:

$$\begin{aligned} \mathrm{MSE}(\widehat{\theta}) &= \mathsf{E}[(\widehat{\theta} - \theta)^2] \\ &= \mathsf{E}[(\widehat{\theta} - \mathsf{E}[\widehat{\theta}])^2] + (\theta - \mathsf{E}[\widehat{\theta}])^2 \\ &= \mathrm{Var}[\widehat{\theta}] + (\text{bias})^2. \end{aligned}$$

Estimating Parameters — The Method of Moments

Method of moments. Given a random sample $X_1, \ldots, X_n$ of a random variable X, for any integer $k \geq 1$,

$$\widehat{\mathrm{E}[X^k]} = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

is an unbiased estimator for the kth moment of X.

Proof. Denote $\mu_k = \mathrm{E}[X^k]$, then

$$\begin{aligned} \mathrm{E}\left[\widehat{\mu_k}\right] &= \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i^k\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[X_i^k] = \frac{1}{n} \cdot n\mu_k = \mu_k. \end{aligned}$$

Estimating Parameters — Method of Maximum Likelihood

Method of maximum likelihood. Given a random sample $X_1, \ldots, X_n$ of a random variable X with parameter θ and density f_X, the ***likelihood function*** is given by

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i).$$

The maximum likelihood estimator (MLE) of θ is given by

$$\widehat{\theta} = \arg\max_{\theta} \; L(\theta).$$

In most of the cases, we equivalently maximize the ***log-likelihood***

$$\ell(\theta) = \ln \; L(\theta), \qquad \widehat{\theta} = \arg\max_{\theta} \; \ell(\theta).$$

Estimating Mean

Method of moments.

- Estimating mean μ.

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- Biasness. As we have noted earlier,

$$\mathrm{E}\left[\widehat{\mu}\right] = \mu.$$

Estimating Mean

Maximum likelihood estimate. Suppose X follows a normal distribution with <u>unknown</u> mean μ and <u>known</u> variance σ^2, and we wish to estimate variance σ^2.

- <u>Estimating variance σ^2.</u>

$$L(\mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[\frac{1}{\sigma^2}\left(\sum_{i=1}^n X_i^2 - 2\mu\sum_{i=1}^n X_i + n\mu^2\right)\right].$$

$$\begin{aligned}\widehat{\mu} &= \arg\max_{\mu}\left\{-\frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\left(\sum_{i=1}^n X_i^2 - 2\mu\sum_{i=1}^n X_i + n\mu^2\right)\right\} \\ &= \frac{1}{n}\sum_{i=1}^n X_i.\end{aligned}$$

- <u>Biasness</u>. As seen earlier, the estimator is unbiased.

Estimating Variance

Method of moments.

- Estimating variance σ^2.

$$\widehat{\sigma^2} = \widehat{\mathrm{E}[X^2]} - \widehat{\mathrm{E}[X]}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

- Biasness. This estimator is not unbiased since

$$\mathrm{E}[X_i^2] = \mathrm{Var}[X_i] + \mathrm{E}[X_i]^2 = \sigma^2 + \mu^2,$$

$$\mathrm{E}[\overline{X}^2] = \mathrm{Var}[\overline{X}] + \mathrm{E}[\overline{X}]^2 = \frac{\sigma^2}{n} + \mu^2,$$

and thus

$$\mathrm{E}[\widehat{\sigma^2}] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Estimating Variance

Maximum likelihood estimate. Suppose X follows a Poisson distribution with parameter k, and we wish to estimate variance k (since both mean and variance of Poisson distribution are k).

- Estimating variance k. We know from lecture slides that

$$L(k) = e^{-nk} \frac{k^{\sum X_i}}{\prod X_i!},$$

$$\widehat{k} = \arg\max_k \left\{ -nk + \ln k \sum_{i=1}^{n} X_i - \ln \prod_{i=1}^{n} X_i \right\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- Biasness. Although both the MLE estimate for mean and variance are sample mean, the estimators are unbiased.

Summary

- Unbiased estimator for mean and variance.

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \widehat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

- Unbiased estimator for moments.

$$\widehat{\mathrm{E}[X^k]} = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

- MLE estimator for parameters.

$$\widehat{\theta} = \arg\max_{\theta}\ \ell(\theta) = \arg\max_{\theta}\ \sum_{i=1}^{n} \ln f_X(x_i).$$

Reliability
Failure Density, Reliability and Hazard Rate
Weibull Distribution

Basic Statistics
Samples and Data
Estimating Parameters
Estimating Intervals
Case Study

Confidence Intervals

Definition. Let $0 \leq \alpha \leq 1$. A $100(1-\alpha)\%$ ***(two-sided) confidence interval*** for a parameter θ is an interval $[L_1, L_2]$ such that

$$P[L_1 \leq \theta \leq L_2] = 1 - \alpha.$$

In most cases, we use ***centered confidence interval*** with

$$P[\theta < L_1] = P[\theta > L_2] = \frac{\alpha}{2}.$$

The $100(1-\alpha)\%$ ***upper confidence bound*** and ***lower confidence bound*** for θ are given by L_u, L_l such that

$$P[\theta \leq L_u] = 1 - \alpha, \qquad P[L_l \leq \theta] = 1 - \alpha.$$

Basic Distributions

Standard normal distribution.

- Density function.

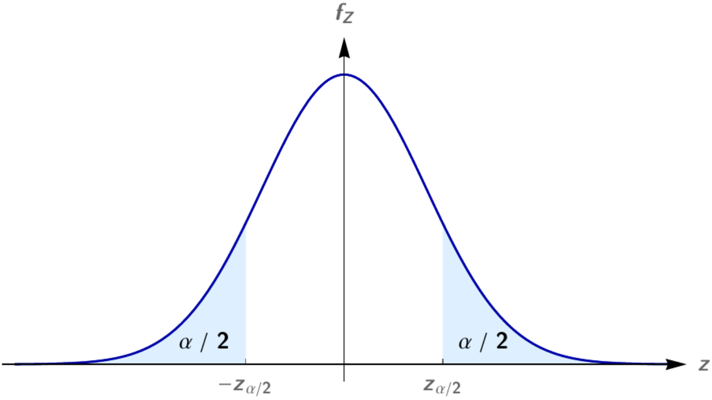
$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{z^2/2}, \qquad z \in \mathbb{R}.$$

- Statistical values. Command for x such that $P[X \geq x] = p$: `InverseCDF[NormalDistribution[0, 1], 1-p]`.

$$\alpha = 0.05 \quad \Rightarrow \quad z_\alpha = 1.64485, \quad z_{\alpha/2} = 1.95996.$$

Basic Distributions

Standard normal distribution.



Basic Distributions

Chi-squared distribution.

- Origin. $Z_1, \ldots, Z_n$ are i.i.d. random variables.

$$Z_i \sim \text{Normal}(0,1) \quad \Rightarrow \quad \chi_n^2 = \sum_{i=1}^{n} Z_i^2 \sim \text{ChiSquared}(n).$$

- Density function. $f_{\chi_n^2}(x) = 0$ for $x < 0$ and

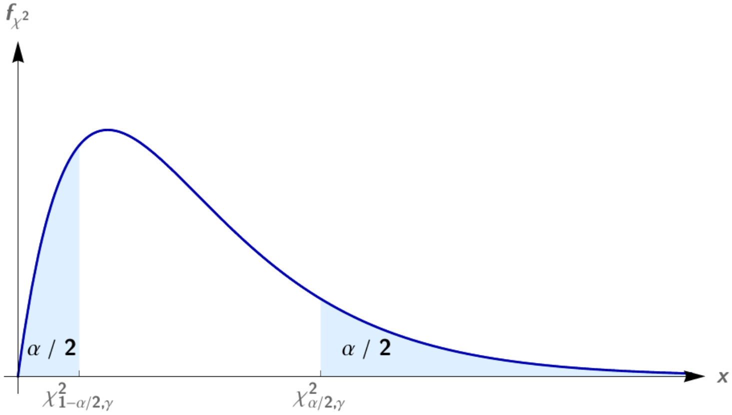
$$f_{\chi_n^2}(x) = \frac{1}{2^{\gamma/2}\Gamma(\gamma/2)} x^{\gamma/2-1} e^{-x/2}, \qquad x \geq 0,$$

  where γ is the degree of freedom.

- Statistical values. Command for x such that $P[X \geq x] = p$: `InverseCDF[ChiSquareDistribution[n], 1-p]`.

Basic Distributions

Chi-squared distribution.



Basic Distributions

Student T-distribution.

- Origin. Z, χ^2_γ are i.i.d. random variables such that

$$Z \sim \text{Normal}(0,1), \qquad \chi^2_\gamma \sim \text{ChiSquared}(\gamma),$$
$$\Rightarrow \quad T_\gamma = \frac{Z}{\sqrt{\chi^2_\gamma/\gamma}} \sim \text{StudentT}(\gamma).$$

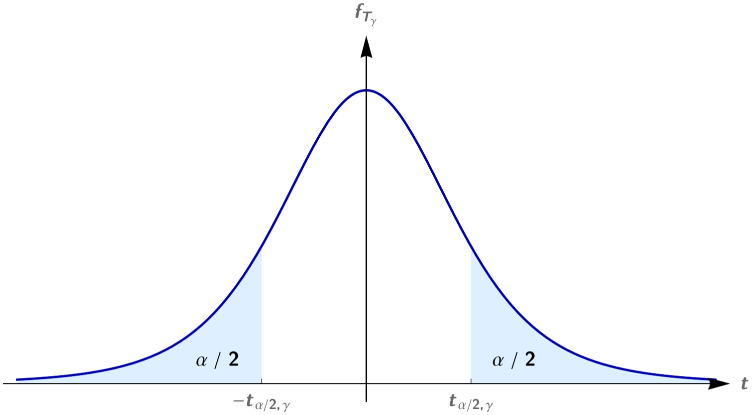
- Density function.

$$f_{T_\gamma}(t) = \frac{\Gamma((\gamma+1)/2)}{\Gamma(\gamma/2)\sqrt{\pi\gamma}} \left(1 + \frac{t^2}{\gamma}\right)^{-\frac{\gamma+1}{2}}, \qquad t \in \mathbb{R}.$$

- Statistical values. Command for x such that $P[X \geq x] = p$: `InverseCDF[StudentTDistribution[n], 1-p]`.

Basic Distributions

Student T-distribution.



Interval Estimation for Mean and Variance

Mean. Suppose we have a random sample of size n from a normal population with ***unknown*** mean μ and ***known*** variance σ^2.

- Statistic and distribution.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0, 1)\,.$$

- $100(1-\alpha)\%$ two-sided confidence interval for μ.

$$\overline{X} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}.$$

- $100(1-\alpha)\%$ one-sided interval for μ.

$$L_u = \overline{X} + \frac{z_\alpha \cdot \sigma}{\sqrt{n}}, \qquad L_l = \overline{X} - \frac{z_\alpha \cdot \sigma}{\sqrt{n}}.$$

Interval Estimation for Mean and Variance

Variance. Suppose we have a random sample of size n from a normal population with ***unknown*** mean μ and ***unknown*** variance σ^2.

- Statistic and distribution.

$$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2} \sim \text{ChiSquared}\,(n-1)\,.$$

- $100(1-\alpha)\%$ two-sided confidence interval for σ^2.

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right].$$

- $100(1-\alpha)\%$ one-sided interval for σ^2.

$$L_u = \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}, \qquad L_l = \frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}.$$

Interval Estimation for Mean and Variance

Mean. Suppose we have a random sample of size n from a normal population with ***unknown*** mean μ and ***unknown*** variance σ^2.

- Statistic and distribution.

$$T_{n-1} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim \mathrm{StudentT}\,(n-1)\,.$$

- $100(1-\alpha)\%$ two-sided confidence interval for μ.

$$\overline{X} \pm \frac{t_{\alpha/2,n-1} S}{\sqrt{n}}.$$

- $100(1-\alpha)\%$ one-sided interval for σ^2.

$$L_u = \overline{X} + \frac{t_{\alpha,n-1} S}{\sqrt{n}}, \qquad L_l = \overline{X} - \frac{t_{\alpha,n-1} S}{\sqrt{n}}.$$

Reliability
- Failure Density, Reliability and Hazard Rate
- Weibull Distribution

Basic Statistics
- Samples and Data
- Estimating Parameters
- Estimating Intervals
- Case Study

Case Study

Suppose we obtain $n = 70$ sample points from simulation.

```
In[ ]:= X = Round[RandomVariate[NormalDistribution[4.5, 2], 70], 0.01]

Out[ ]= {1.67, 3.6, 2.67, 11.3, 3.86, 2.67, 4.43, 5.86, 3.12, 2.86, 7.24, 3.31, 4.98, 6.68, 3.27, 6.32,
  3.94, 4.14, 4.9, 1.98, 7.27, 5.84, 1.33, 7.86, 4.12, 2.39, 9., 5.03, 6.03, 7.85, 1.94, 3.52, 5.49, 6.57,
  8.9, 7.73, 5.18, 4.3, 7.37, 5.02, 6.82, 1.24, 3.66, 0.94, 2.22, 5.37, 3.13, 2.44, 3.43, 3.89, 4.53, 1.37,
  4.88, 3.15, 1.63, 0.62, 3.49, 3.06, 2.76, 5.47, 3.26, 5.77, 6.64, 5.74, 2.19, 1.42, 3.82, 2.76, 2.29, 6.93}
```

We would like to:

1. visualize these data points,
2. obtain point estimates for mean and variance (suppose they are unknown), and
3. obtain interval estimates for
   3.1 mean when variance is known,
   3.2 mean and variance when variance is unknown.

Case Study

Histogram. Using Freedman-Diaconis Rule,

$$q_1 = 2.76, \quad q_3 = 5.84 \quad \Rightarrow \quad \mathrm{IQR} = q_3 - q_1 = 3.08,$$

and

$$h = \frac{2\mathrm{IQR}}{\sqrt{n}} = 0.736261 \approx 0.74 \quad \text{(rounding up)}.$$

Then the lower bound of the first bin is

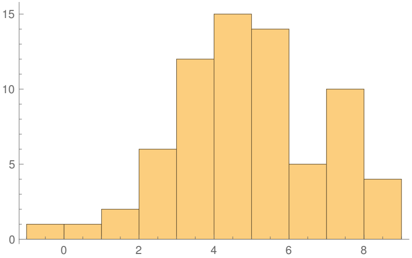
$$\min\{x_i\} - \text{pre.}/2 = 0.62 - 0.005 = 0.615.$$

Case Study

Histogram.

```
In[ ]:= {q1, q2, q3} = Quartiles[X]
        iqr = InterquartileRange[X]
        h = 2 iqr / Sqrt[70]
        Min[X] - 0.005
Out[ ]= {2.76, 3.915, 5.84}
Out[ ]= 3.08
Out[ ]= 0.736261
Out[ ]= 0.615
```

```
Histogram[X, "FreedmanDiaconis"]
Out[ ]=
```



Case Study

Stem-and-leaf diagram. We use stem units as 1.

In[•]:=
```
Needs["StatisticalPlots`"]
StemLeafPlot[Floor[X, 0.1], IncludeEmptyStems → True]
```

Out[•]=

Stem	Leaves
0	69
1	23346699
2	1223466778
3	0111223445668889
4	11345899
5	0013447788
6	0356689
7	223788
8	9
9	0
10	
11	3

Stem units: 1

Case Study

Boxplots. The inner fences and outer fences are determined as

$$f_1 = q_1 - \frac{3}{2}\text{IQR} = -1.86, \qquad f_3 = q_3 + \frac{3}{2}\text{IQR} = 10.46,$$
$$F_1 = q_1 - 3\text{IQR} = -6.48, \qquad F_3 = q_3 + 3\text{IQR} = 15.08,$$

and adjacent values

$$a_1 = \min\{x_k : x_k \geq f_1\}, \qquad a_3 = \max\{x_k : x_k \leq f_3\}.$$

Mathematica commands $\Rightarrow$

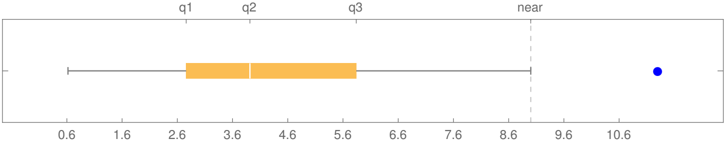
```
In[ ]:= f1 = q1 - 3/2*iqr
        f3 = q3 + 3/2*iqr
        F1 = q1 - 3 iqr
        F3 = q3 + 3 iqr
        a1 = Min[Select[X, # ≥ f1 &]]
        a3 = Max[Select[X, # ≤ f3 &]]
```

Case Study

Boxplots.

```
BoxWhiskerChart[
 X, {"Outliers", {"Outliers", Blue}, {"FarOutliers", Red}},
 AspectRatio → 1/7, BarOrigin → Left,
 GridLines → {{{a3, Dashed}, {F3, Dashed}}, None}, ImageSize → Large, FrameTicks → {
   {None, None},
   {Range[Min[Floor[X, 0.1]], Max[Ceiling[X, 0.1]]],
    {{q1, "q1"}, {q2, "q2"}, {q3, "q3"}, {a3, "near"}, {F3, "far"}}}}
]
```

Out[ ]=



Case Study

Point estimate for mean and variance. We use unbiased estimators for mean and variance.

- Mean.

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = 4.38.$$

- Variance.

$$\widehat{\sigma^2} = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = 4.90.$$

Case Study

Interval estimate for mean and variance.

- Mean. (Variance $\sigma^2 = 4$.) A 95% two-sided confidence interval for mean μ is given by

$$\mathrm{CI} = \left[\overline{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right] = [3.91, 4.85].$$

- Variance. (Variance unknown.) A 95% two-sided confidence interval for variance σ^2 is given by

$$\mathrm{CI} = \left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] = [3.60, 7.05].$$

- Mean. (Variance unknown.) A 95% two-sided confidence interval for mean μ is given by

$$\mathrm{CI} = \left[\overline{X} - \frac{t_{\alpha/2,n-1}S}{\sqrt{n}}, \overline{X} + \frac{t_{\alpha/2,n-1}S}{\sqrt{n}}\right] = [3.21, 5.55].$$

Thanks for your attention!