VE401, Probabilistic Methods in Eng.

Recitation Class - Week 5

Zhanpeng Zhou

UMJI-SJTU Joint Institute

March 29, 2021

Mail: zzp1012@sjtu.edu.cn / WeChat: zzp01012

Table of contents

1. Reliability
   - Failure Density, Reliability and Hazard Rate
   - Common Distributions for Reliability Studies
2. Basic Statistics
3. Supplementary Materials
4. Exercises

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:

$$\begin{aligned}\rho_A(t) &:= \lim_{\Delta t \to 0} \frac{P[t \le T \le t + \Delta t | t \le T]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P[t \le T \le t + \Delta t]}{P[T \ge t] \cdot \Delta t} = \frac{f_A(t)}{R_A(t)}, \\ R_A(t) &= e^{-\int_0^t \rho_A(x) \mathrm{d}x}.\end{aligned}$$

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.

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

Exponential Distribution

- Density function. $\beta > 0$ is a parameter,

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

- Mean.

$$\mu = \frac{1}{\beta}.$$

- Variance.

$$\sigma^2 = \frac{1}{\beta^2}.$$

- Reliability features.

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

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.

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

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

Steps.

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	0000000111112222222222233334444455555666667777778888999999
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 IQR.
2. Find ***inner fences*** and ***outer fences*** by

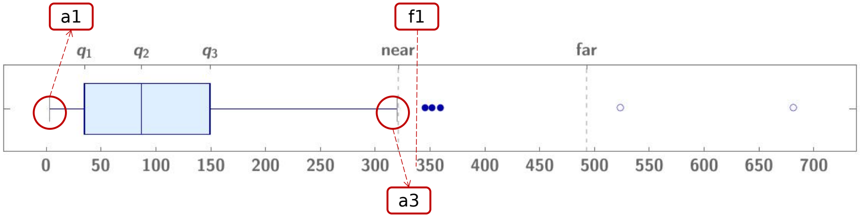
$$f_1 = q_1 - \frac{3}{2}\text{IQR}, \qquad f_3 = q_3 + \frac{3}{2}\text{IQR},$$
$$F_1 = q1 - 3\text{IQR}, \qquad F_3 = q_3 + 3\text{IQR},$$

   and find ***adjacent values***

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

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

Visualization — Boxplots



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.

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

- ***Mean square error***:

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

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 ***likeliho-od 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 distribut-ion with unknown mean μ and known variance σ^2, and we wish to estimate mean μ.

- Estimating mean μ.

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

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

- Biasness. 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

$$\begin{aligned} L(k) &= e^{-nk} \frac{k^{\sum X_i}}{\prod X_i!}, \\ \widehat{k} &= \arg\max_k \left\{ -nk + \ln k \sum_{i=1}^n X_i - \ln \prod_{i=1}^n X_i \right\} \\ &= \frac{1}{n} \sum_{i=1}^n X_i. \end{aligned}$$

- 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.

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

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.

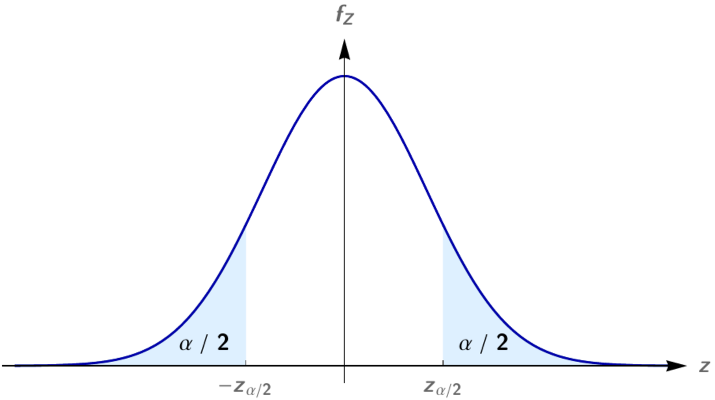
$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{z^2/2}, \qquad z \in \mathbb{R}.$$

- 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

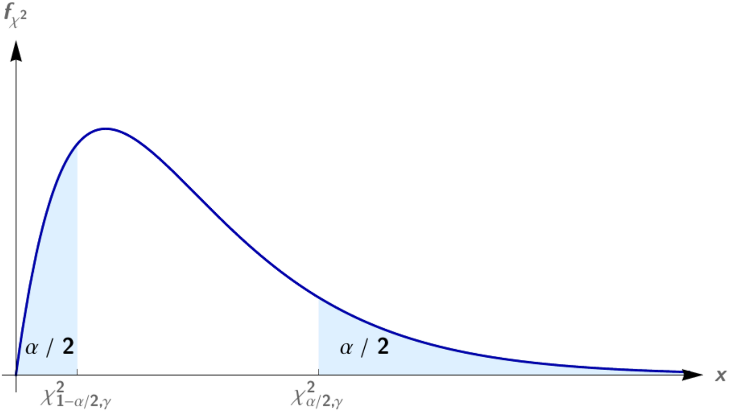
$$f_{\chi_n^2}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \qquad x \geq 0,$$

where n 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

Chi distribution.

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

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

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

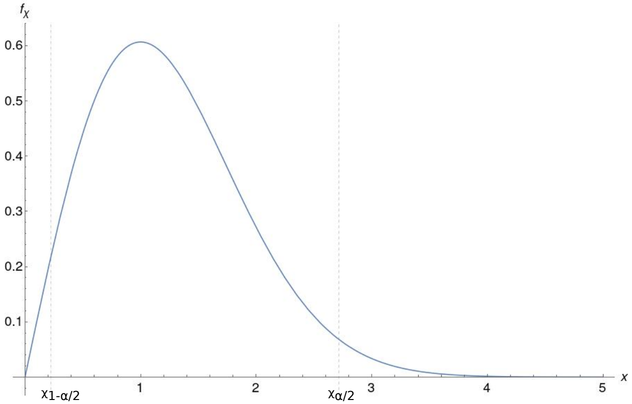
$$f_{\chi_n}(x) = \frac{2}{2^{n/2}\Gamma(n/2)} x^{n-1} e^{-x^2/2}, \qquad x \geq 0,$$

  where n is the degree of freedom.

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

Basic Distributions

Chi distribution.



Basic Distributions

Student T-distribution.

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

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

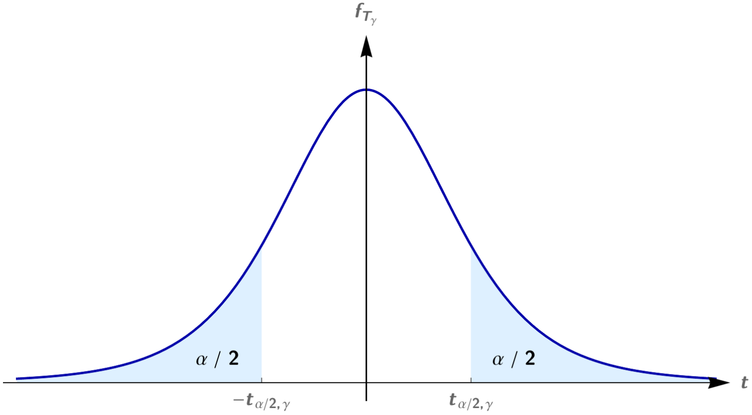
- Density function.

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

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

Basic Distributions

Student T-distribution.



Summary

Suppose $X_1, \ldots, X_n$ are samples from a population X, where X follows normal distribution with mean μ and variance σ^2.

- Normal distribution.

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

- Chi-squared distribution.

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

- Chi distribution.

$$\chi_{n-1} = \sqrt{\frac{(n-1)S^2}{\sigma^2}} \sim \text{Chi}\,(n-1)\,.$$

- Student T-distribution.

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

Interval Estimation for Mean (Variance Known)

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 (Variance Unknown)

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.

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

- $100(1-\alpha)\%$ 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.

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

Interval Estimation for 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 Standard Deviation

Std. Deviation. 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.

$$\chi_{n-1} = \sqrt{\frac{(n-1)S^2}{\sigma^2}} \sim \mathrm{Chi}\,(n-1)\,.$$

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

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

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

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

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
   1. mean when variance is known,
   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

$$h = \frac{2\mathrm{IQR}}{\sqrt[3]{n}} = 1.49468 \approx 1.50 \quad \text{(rounding up)}.$$

Then the lower bound of the first bin is

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

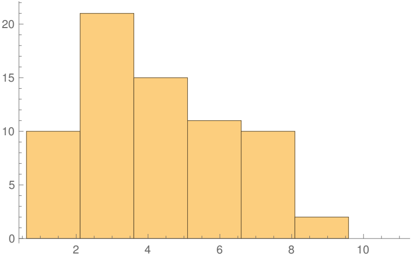
Case Study

Histogram.

```
In[ ]:= {q1, q2, q3} = Quartiles[X]
        iqr = InterquartileRange[X]
        h = 2 iqr / ∛70
        Min[X] - 0.005
Out[ ]= {2.76, 3.915, 5.84}
Out[ ]= 3.08
Out[ ]= 1.49468
Out[ ]= 0.615
```

```
In[ ]:= Histogram[X, {Min[X] - 0.005, Max[X], h}]
```

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

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

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 $\Rightarrow$

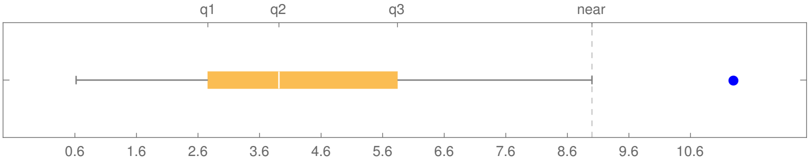
```
In[ ]:= f1 = q1 - 3/2*iqr
        f3 = q3 + 3/2*iqr
        F1 = q1 - 3 iqr
        F3 = q3 + 3 iqr
        a1 = Min[Select[X, # ≥ f1 &]]
        a3 = Max[Select[X, # ≤ f3 &]]
```

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

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

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

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

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

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

German Tank Problem

German Tank Problem. Suppose there exists an unknown number of tanks which are sequentially numbered from 1 to N. A random sample of these tanks is taken and their sequence numbers observed. Try to estimate N from these observed numbers, by using:

- the method of moments
- the method of maximum likelihood

What is the good method?

SP20 Assignment 3.4

A mathematics textbook has 200 pages on which typographical errors in the equations could occur. Suppose there are in fact five errors randomly dispersed among these 200 pages.

1. What is the probability that a random sample of 50 pages will contain at least one error?
2. How large must the random sample be to assure that at least three errors will be found with 90% probability? (You may use a normal approximation to the binomial distribution.)

SP20 Assignment 3.4 Sol. I

1.
The problem is to randomly place the five errors in 200 pages, and each error has the same probability of being placed among the sampled pages.

$$\begin{aligned} P[\text{at least 1 error in 50 pages}] &= 1 - P[\text{0 error in 50 pages}] \\ &= 1 - \left(\frac{200-50}{200}\right)^5 \\ &= 76.27\%. \end{aligned}$$

2. Let the sample size be k. The number of selected errors follows a binomial distribution with

$$p = \frac{k}{200}, \qquad n = 5,$$

SP20 Assignment 3.4 Sol. II

and thus the mean and standard deviation are given by

$$\mu = 5p = \frac{k}{40}, \qquad \sigma = \sqrt{5p(1-p)} = \sqrt{\frac{k}{40}\left(1 - \frac{k}{200}\right)}.$$

Let X be the number of errors in the sample. Then

$$P[X \geq 3] \geq 90\% \quad \Rightarrow \quad P[Y \geq 2.5] \geq 90\%,$$

where Y follows normal distribution. Transforming to standard normal variable Z, we have

$$P\left[Z \geq \frac{2.5-\mu}{\sigma}\right] \geq 0.9 \quad \Rightarrow \quad F\left[\frac{2.5-\mu}{\sigma}\right] \leq 0.1 \quad \Rightarrow \quad \frac{2.5-\mu}{\sigma} \leq -1.28,$$

which gives $k \geq 150$.

SP20 Assignment 3.4 Sol. III

Note. Some of you may have noticed that the requirements for "good approximation" specified in lecture slides are not satisfied. However, if we calculate using $p = 0.75$ and $n = 5$ for binomial distribution,

$$P[X \geq 3] = 1 - \texttt{CDF}[\texttt{BinomialDistribution}[5, 0.75], 2] = 0.896484,$$

which is quite close to 90%. This posterior validation shows the approximation is reasonable.

SP20 Assignment 3.11

A system consists of two independent components connected in series. The life span (in hours) of the component follows a Weibull distribution with $\alpha = 0.006$ and $\beta = 0.5$; the second has a lifespan in hours follows the exponential distribution with $\beta = 1/25000$.

1. Find the reliability of the system at 2500 hours.
2. Find the probability that the system will fail before 2000 hours.
3. If the two components are connected in parallel, what is the system reliability at 2500 hours?

SP20 Assignment 3.11 I

1.

$$R_s(t) = R_1(t) \cdot R_2(t)$$

Due to R_1 follows Weibull Distribution with $\alpha = 0.006$ and $\beta = 0.5$, then:

$$R_1(t) = e^{-\alpha t^{\beta}} = e^{-0.006 t^{0.5}}$$

Also, we have already knew that $R_2(t)$ follows exponential distribution with $\beta = 1/25000$, then:

$$R_2(t) = 1 - \int_0^t \frac{1}{25000} e^{-x/25000} dx = e^{-t/25000}$$

Thus:

$$R_s(2500) = e^{-0.006 \times 2500^{0.5}} \cdot e^{-2500/25000} \approx 0.6703$$

2.

SP20 Assignment 3.11 II

$$P[X < 2000] = 1 - R_s(2000) = 1 - e^{-0.006 \times 2000^{0.5}} \times e^{-2000/25000} \approx 0.2941$$

3.

$$\begin{aligned} R_p(2500) &= 1 - (1 - R_1(2500))(1 - R_2(2500)) \\ &= 1 - (1 - e^{-0.006 \times 2500^{0.5}})(e^{-2500/25000}) \\ &\approx 0.9753 \end{aligned}$$

SP20 Assignment 4.2 I

Let $X_1, \ldots, X_n$ be a random sample of size n from a random variable with variance σ^2. We have seen that the sample variance

$$S_{n-1}^2 := \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2$$

is an unbiased estimator for σ^2. It can be shown that

$$\operatorname{Var}(S_{n-1}^2) = \operatorname{MSE}(S_{n-1}^2) = \frac{1}{n}\left(\operatorname{E}[(X - \overline{X})^4] - \frac{n-3}{n-1}\sigma^4\right) = \frac{1}{n}\left(\gamma_2 + \frac{2n}{n-1}\right) \tag{1}$$

where $\gamma_2 := \operatorname{E}[(X - \mu)^4]/\sigma^4 - 3$ is called the *excess kurtosis* of a distribution.

SP20 Assignment 4.2 II

1. Show that if X follows a normal distribution with mean μ and variance σ^2,

$$\mathrm{MSE}(S_{n-1}^2) = \frac{2}{n-1}\sigma^4.$$

2. For $a > 0$ set

$$S_a^2 := \frac{n-1}{a} S_{n-1}^2.$$

Find $\mathrm{MSE}(S_a^2)$ and show that the mean square error is minimized for

$$a = n + 1 + \frac{n-1}{n}\gamma_2.$$

In the case of a normal distribution with mean μ and variance σ^2, show that this reduces to $a = n + 1$.

SP20 Assignment 4.2 Sol. I

1.

Recall MGF of normal distribution

$$m_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

Therefore, for a standard normal distribution

$$m_Z(t) = e^{\frac{t^2}{2}}$$

Thus

$$E[Z^4] = \frac{d^4 e^{\frac{t^2}{2}}}{dt^4} = (3e^{\frac{t^2}{2}} + (t^4 + 5t^2)e^{\frac{t^2}{2}})|_{t=0} = 3$$

Define a random variable X follows normal distribution with mean μ and variance σ^2

$$Z = \frac{X - \mu}{\sigma}$$

SP20 Assignment 4.2 Sol. II

Thus

$$
\begin{aligned}
E[(\frac{X-\mu}{\sigma})^4] &= 3 \\
E[(X-\mu)^4] &= 3\sigma^4 \\
\frac{E[(X-\mu)^4]}{\sigma^4} - 3 &= 0 \\
\gamma_2 &= 0 \\
\frac{1}{n}(\gamma_2 + \frac{2n}{n-1})\sigma^4 &= \frac{1}{n}(\frac{2n}{n-1})\sigma^4 \\
\mathrm{MSE}(S_{n-1}^2) &= \frac{2}{n-1}\sigma^4
\end{aligned}
$$

Thus, the statement is proved.

SP20 Assignment 4.2 Sol. III

2.

$$
\begin{aligned}
&\mathrm{MSE}(S_a^2)\\
=&\mathrm{MSE}(\frac{n-1}{a}S_{n-1}^2)\\
=&E[(\frac{n-1}{a}S_{n-1}^2-\sigma^2)^2]\\
=&E[(\frac{n-1}{a}S_{n-1}^2-\frac{n-1}{a}\sigma^2+\frac{n-1-a}{a}\sigma^2)^2]\\
=&E[(\frac{n-1}{a})^2(S_{n-1}^2-\sigma^2)^2+2(\frac{n-1}{a})(S_{n-1}^2-\sigma^2)(\frac{n-1-a}{a}\sigma^2)+(\frac{n-1-}{a}\\
=&(\frac{n-1}{a})^2\mathrm{MSE}(S_{n-1}^2)+(\frac{n-1-a}{a}\sigma^2)^2\\
=&(\frac{n-1}{a})^2\frac{1}{n}(\gamma_2+\frac{2n}{n-1})\sigma^4+(\frac{n-1-a}{a})^2\sigma^4\\
=&(\frac{(n-1)^2}{n}\gamma_2+2(n-1)+(n-1)^2)\frac{1}{a^2}-2(n-1)\frac{1}{a}+1)\sigma^4
\end{aligned}
$$

SP20 Assignment 4.2 Sol. IV

Thus, to optimize the MSE, we should take a as

$$\frac{1}{a} = -\frac{-2(n-1)}{2(\frac{(n-1)^2}{n}\gamma_2 + 2(n-1) + (n-1)^2)}$$

$$a = n - 1 + 2 + \frac{n-1}{n}\gamma_2$$

$$a = n + 1 + \frac{n-1}{n}\gamma_2$$