VE401 RC Week5

Wang Yangyang

UM-SJTU JI

2022 Spring

Outline

Outline

Reliability

Definition

Suppose a unit A fails ***randomly***, and we describe the time it fails by the continuous random variable T_A.
The density of T_A is called the ***failure density*** f_A. The cumulative distribution function of T_A is denoted by F_A. We note that

$$\begin{aligned} f_A(t) &= \lim_{\Delta t \to 0} \frac{P[t \leqslant T \leqslant t + \Delta t]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{F_A(t + \Delta t) - F_A(t)}{\Delta t} \end{aligned}$$

The ***reliability function*** R_A gives the probability that A is working at time $t \geqslant 0$

$$\begin{aligned} R_A(0) &= 1 \\ R_A(t) &= 1 - P[\text{ component A fails before time } t] \\ &= 1 - F_A(t) \end{aligned}$$

Hazard

Definition

Hazard rate ρ_A defined by

$$\begin{aligned}\rho_A(t) &= \lim_{\Delta t \to 0} \frac{P[t \leqslant T \leqslant t + \Delta t \mid t \leqslant T]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P[t \leqslant T \leqslant t + \Delta t]}{P[T \geqslant t] \cdot \Delta t} = \frac{f_A(t)}{R_A(t)}\end{aligned}$$

Interpretation

- If ρ is decreasing, then as time goes by a failure is more likely to occur earlier in the time interval.
- If ρ is steady, a failure tends to occur during this period due mainly to random factors.
- If ρ is increasing, then as time goes by a failure is more likely to occur.

Hazard

Theorem

Let X be a random variable with ***failure density*** f, ***reliability function*** R, and ***hazard rate*** ρ. Then

$$R(t) = e^{-\int_0^t \rho(x)\,dx}$$

Proof

Proof. Since $R(x) = 1 - F(x)$ we have $R'(x) = -F'(x)$. Therefore,

$$\rho(x) = \frac{f(x)}{R(x)} = \frac{F'(x)}{R(x)} = -\frac{R'(x)}{R(x)}$$

$$R'(x) = -\rho(x)R(x)$$

Solving this equation with $R(0) = 1$ (because A is always working at the beginning), we obtain the result.

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.

$$\rho(t) = \alpha\beta t^{\beta-1}$$

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

Outline

System

R_i is the reliability of the ith component, then

1. reliability of a series system with k components

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

2. reliability of a parallel system with k components

$$R_p(t) = 1 - P[\text{all components fail before t}] = 1 - \prod_{i=1}^{k} (1 - R_i(t))$$

Outline

Sample and Percentile

Definition

A ***random sample*** of size n from the distribution of X is a collection of n independent random variables $X_1, \dots, X_n$, each with the same distribution as X.

- i.i.d random variables
- Sample size n should neither be too small or large, usually smaller than 5% of the population.

Percentiles: The x th percentile is defined as the value d_x of the data such that $x\%$ of the values of the data are less than or equal to d_x.

Quantile

Definition

Quartiles:(special cases of percentiles)

- 25% of the data are no greater than the first quartile q_1.
- 50% are no greater than the second quartile q_2 (median).
- 75% are no greater than the third quartile q_3.

Definition

Interquartile Range: $\mathrm{IQR} = q_3 - q_1$.

- Median describes location of data.
- IQR describes dispersion of data.

Calculating Quartiles

Suppose that our list of n data has been ordered from smallest to largest, so that

$$x_1 \leqslant x_2 \leqslant x_3 \leqslant \cdots \leqslant x_n$$

Then the median q_2 :

$$q_2 = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even} \end{cases}$$

The first quartile q_1 :

- the median of the smallest $n/2$ elements if n is even.
- the average of the median of the smallest $(n-1)/2$ elements and the median of the smallest $(n+1)/2$ elements of the list if n is odd.

The third quartile q_3 : replace "smallest" with "largest" in the above definition.

Outline

1. Reliability
   - Reliability and Hazard
   - System
2. **Basic Statistics**
   - Samples and Data
   - **Data Visualization**
3. Estimator
   - Parameter Estimation
   - Interval Estimation
4. Supplementary Materials
   - Exercise and Discussion

Histogram

1. Determine ***bin width***
   - Sturges's Rule

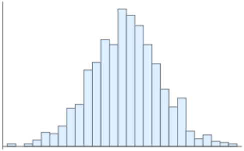
$$k = \lceil \log_2(n) \rceil + 1, \quad h = \frac{\max\{x_i\} - \min\{x_i\}}{k}$$

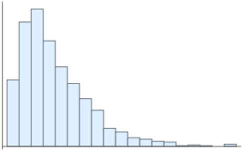
   - Freedman-Diaconis Rule

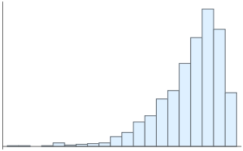
$$h = \frac{2 \cdot \mathrm{IQR}}{\sqrt[3]{n}}$$

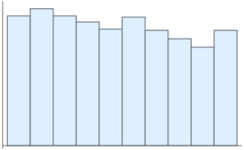
   which should be rounded up to the precision of the data.
2. Determine ***the lower boundary***:
   Ideally, take the smallest datum, subtract one-half of the smallest decimal of the data and then successively add the bin width to obtain the bins.
   In practice, just choosing the lower boundary where ***no datum can lie on the boundary*** is fine.

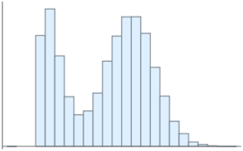
Histogram

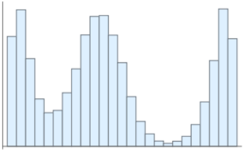

Symmetric,
unimodal


Positive skew,
unimodal

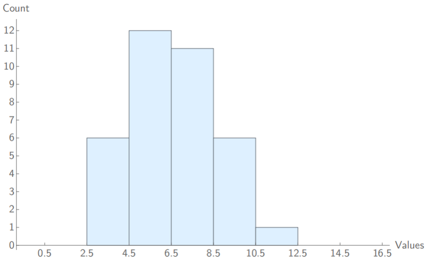

Negative skew,
unimodal


Symmetric,
no prominent mode


Bimodal


Multimodal

Details and Interpretation



(1/2) for labelling the axes
(1/2) for starting point that prevents data from falling on the boundary.
(1) is for the general shape and correctness of the histogram.
This histogram has a ***unimodal*** shape (1/2 Mark) which is consistent with a normal distribution.
It is not significantly ***skewed***, (1/2 Mark) again consistent with a normal distribution.
Therefore, there is no evidence that the data does not come from a normal distribution. (1 Mark)

Stem-and-Leaf Diagram

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.

Stem	Leaves
0	00000001111122222222222333344444555556666677777788899
1	00011111223344444455555678899
2	223669
3	012456
4	
5	2
6	8

Stem Units: 100 (Important!)

Box-and-Whisker Plot

We define the ***inner fences***

$$f_1 = q_1 - \frac{3}{2}\mathrm{IQR}, \quad f_3 = q_3 + \frac{3}{2}\mathrm{IQR}$$

The "whiskers" (lines extending to the left and right of the box) end at the adjacent values

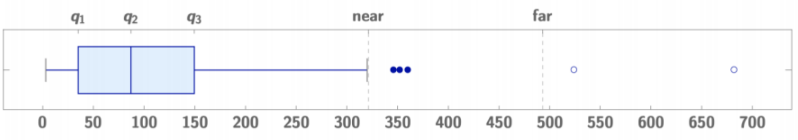
$$a_1 = \min\{x_k : x_k \geqslant f_1\}, \quad a_3 = \max\{x_k : x_k \leqslant f_3\}$$

We define the ***outer fences***

$$F_1 = q_1 - 3\mathrm{IQR}, \quad F_3 = q_3 + 3\mathrm{IQR}$$

Measurements x_k that lie outside the inner fences but inside the outer fences are called ***near outliers***.
Those outside the outer fences are known as ***far outliers***.

Box-and-Whisker Plot

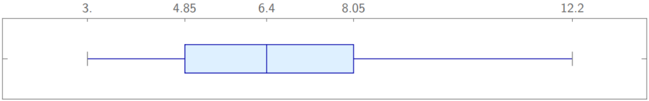


Interpretation

Interpretation: If data is obtained from a normal distribution, one would expect to see

- a symmetric median line in the middle of the box;
- equally long whiskers;
- very few near outliers and no far outliers.

Details and Interpretation



(1/2) for the general shape of the boxplot,
(1/2) for labelling the ordinate
(1) for the correct drawing indicating the whisker values, q_1, q_2, q_3 and for correctly identified outlier(s), if any.
The whiskers are moderately ***asymmetric*** (1/2)
but the ***median line*** is not too far from the center of the box (1/2).
There is no ***outlier***, (1/2)
and in summary no strong evidence that the data does not come from a normal distribution. (1/2)

Outline

Estimation

- *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*.

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

Sample Mean and Sample Variance

Theorm

Let $X_1, \dots, X_n$ be a random sample of size n from a distribution with mean μ. The *sample mean* $\bar{X}$ is an unbiased estimator μ.
Let $\bar{X}$ be the sample mean of a random sample of size n from a distribution with mean μ and variance σ^2. Then

$$\operatorname{Var}\bar{X} = \mathrm{E}\left[(\bar{X}-\mu)^2\right] = \frac{\sigma^2}{n}$$

- $\operatorname{MSE}\bar{X} = \operatorname{Var}\bar{X}$
- We can make MSE $\bar{X}$ small by taking n large enough.

The unbiased *sample variance*

$$S^2 := \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2$$

Method of Moments

Given a random sample $X_1, \ldots, X_n$ of a random variable X, for any integer $k \geqslant 1$,

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

is an unbiased estimator for the k th moment of X.

Proof

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

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

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

$$\ell(\theta) = \ln L(\theta), \quad \widehat{\theta} = \arg\max_{\theta} \ell(\theta)$$

Estimating Mean - MOM

- Estimating mean μ.

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

- Biasness. As we have noted earlier,

$$\mathrm{E}[\widehat{\mu}] = \mu$$

Estimating Mean - MLE

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.

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

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

- Biasness. As seen earlier, the estimator is unbiased.

Estimating Variance - MOM

- Estimating variance σ^2

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

- Biasness. This estimator is not unbiased since

$$\mathrm{E}\left[X_i^2\right] = \mathsf{Var}\left[X_i\right] + \mathrm{E}\left[X_i\right]^2 = \sigma^2 + \mu^2$$
$$\mathrm{E}\left[\bar{X}^2\right] = \mathsf{Var}[\bar{X}] + \mathrm{E}[\bar{X}]^2 = \frac{\sigma^2}{n} + \mu^2$$

and thus

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

Estimating Variance - MLE

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.

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{X}\right)^2$$

- Unbiased estimator for moments.

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

- MLE estimator for parameters.

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

Outline

Summary

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

- Normal distribution.

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

- Student T-distribution.

$$T_{n-1} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \text{ Student } T(n-1)$$

- Chi-squared distribution.

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

- Chi distribution.

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

Estimation for Mean (Variance Known)

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.

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

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

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

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

$$L_u = \bar{X} + \frac{z_\alpha \cdot \sigma}{\sqrt{n}}, \quad L_l = \bar{X} - \frac{z_\alpha \cdot \sigma}{\sqrt{n}}$$

Estimation for Mean (Variance Unknown)

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.

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

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

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

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

$$L_u = \bar{X} + \frac{t_{\alpha,n-1} S}{\sqrt{n}}, \quad L_l = \bar{X} - \frac{t_{\alpha,n-1} S}{\sqrt{n}}$$

Estimation for 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.

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

Estimation for 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.

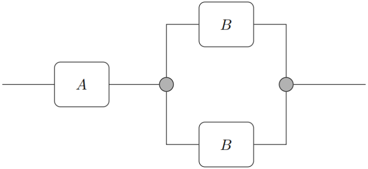
$$L_u = \frac{\sqrt{(n-1)S^2}}{\chi_{1-\alpha,n-1}}, \quad L_l = \frac{\sqrt{(n-1)S^2}}{\chi_{\alpha,n-1}}.$$

Outline

1. System Fail Time

Exercise

Consider the following system of components:

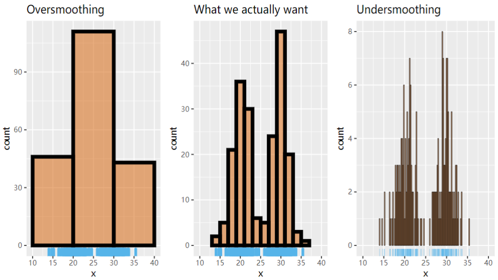


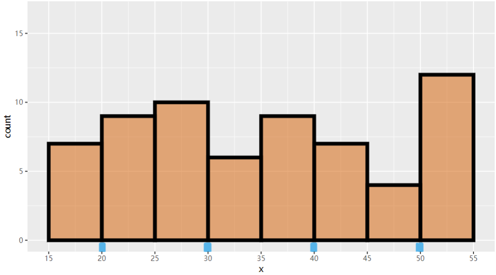
The system will fail if either component A or both components marked B fail. The components A and B have failure densities

$$f_A(t) = \frac{1}{100}e^{-t/100}, f_B(t) = \frac{1}{50}e^{-t/50}, \quad t \geqslant 0$$

respectively. What is the expected time of failure of the system?

2. Misleading Bining





https://aakinshin.net/posts/misleading-histograms/

3. Uniform Distribution Estimation

Exercise

Estimator $\widehat{\Theta}$ is called an unbiased estimator for Θ if $E(\widehat{\Theta}) = \Theta$ (notice that $\widehat{\Theta}$ is indeed a random variable!). Consider a Uniform distribution on the interval $(0,\ \mathrm{A})$.

- Is the maximum likelihood estimator for A unbiased?
- Is $\widehat{A}_1 = 2\bar{X}_n$ an estimator for A ? Is it a reasonable estimator for A ? Is the above defined $\widehat{A}_1$ an unbiased estimator for A?
- Is $\widehat{A}_2 = 2$ an estimator for A ? Is it a reasonable estimator for A ? Is the above defined $\widehat{A}_2$ an unbiased estimator for A ?

End

Credit to Zhanpeng Zhou (TA of SP21)
Credit to Fan Zhang (TA of SU21)
Credit to Liying Han (TA of SP21)
Credit to Zhenghao Gu (TA of SP20)