VE401-Mid Review

Zhang Hexin

2020.3.22

Introduction

- Zhang Hexin (张何欣)
- 1234567809zhx@sjtu.edu.cn
- Junior Student majored in ECE
- Interests: data science
- Love sports, movies……

minion

扫一扫二维码，加我QQ。

Statistics-Visualization

- **Stem-and-Leaf diagram**

```
Needs["StatisticalPlots`"]

StemLeafPlot[Floor[Data, 10], IncludeEmptyStems → True]
```

Stem	Leaves
0	000000011111222222222223333444445555566666777777888899999
1	00011111223344444455555678899
2	223669
3	012456
4	
5	2
6	8

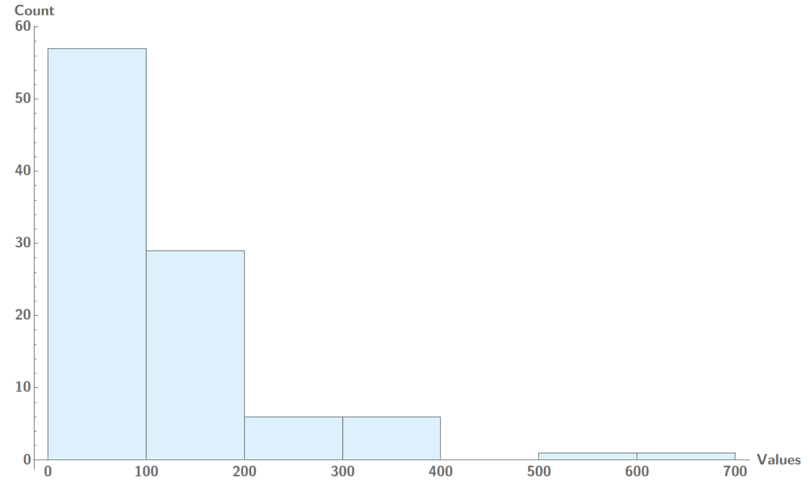
Stem units: 100

Statistics-Visualization

- **Histogram-method1**
- Sturges rule: $k = \lceil \log_2(n) \rceil + 1,$

The data range is $682 - 3 = 679$ and Sturges's rule (based on 100 data) gives $k = 7$. We calculate $679/7 = 97$, which should be rounded up by one to $h = 98$.

- Bin width: $h = \dfrac{\max\{x_i\} - \min\{x_i\}}{k},$



Mathematica: `Histogram[Data, "Sturges"]`

- Finally, 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.

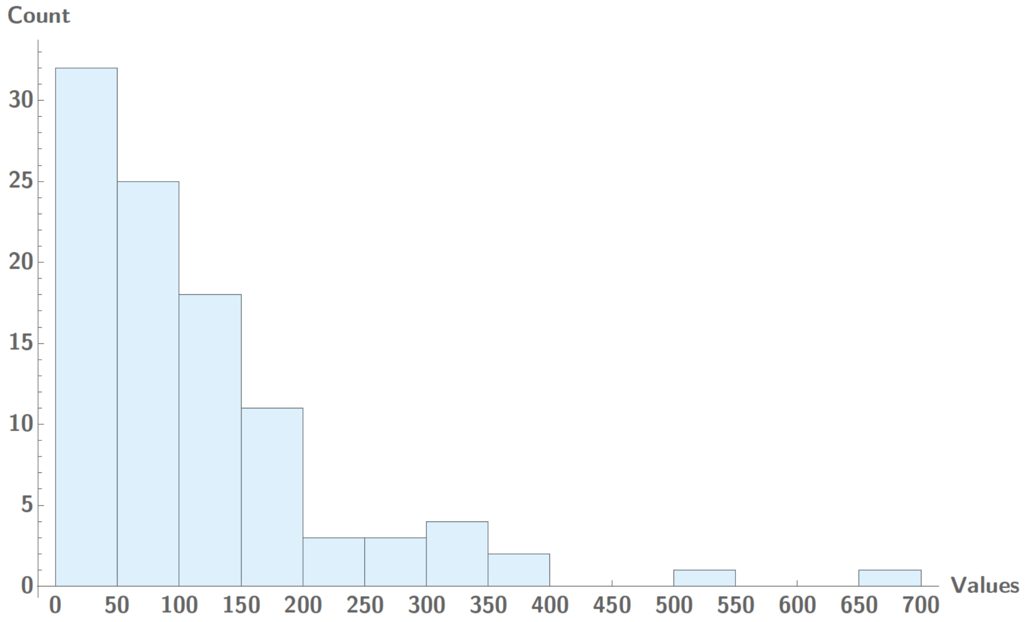
Statistics-Visualization

- **Histogram-method2**
- The Freedman-Diaconis Rule: Bin width:

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

- Similarly, 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 our example, we have $\frac{2 \cdot \mathrm{IQR}}{\sqrt[3]{n}} = 49.34$, which we round up to 50.



Mathematica: `Histogram[Data, "FreedmanDiaconis"]`

Statistics-Visualization

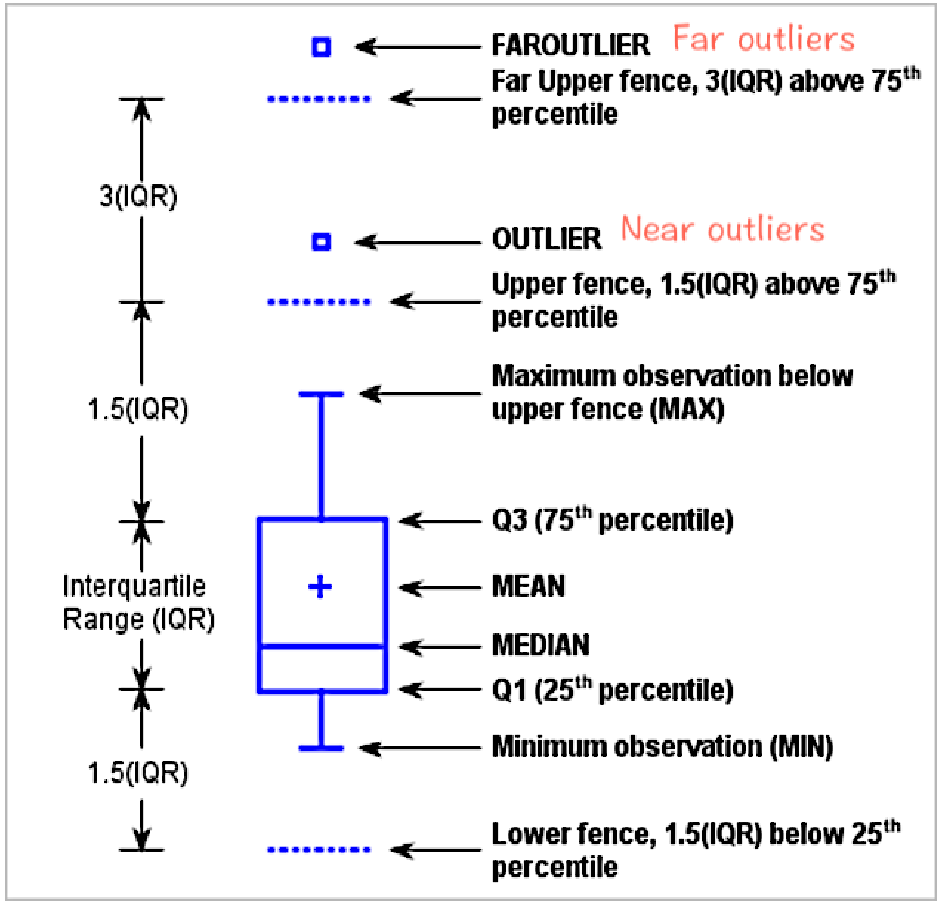
- **Box plot**
- q_1, q_2 , q_3 : 25%, 50%, 75% of the data are no greater than the first/second/third quartile
- Interquatile range: $q_3 - q_1$
- Inner/Outer fences:

$$f_1 = q_1 - \frac{3}{2} IQR. \qquad f_3 = q_3 + \frac{3}{2} IQR$$

$$F_1 = q_1 - 3IQR. \qquad F_3 = q_3 + 3IQR$$

- Adjacent values

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



Statistics-Estimation

- Estimator vs. point estimate
- Unbiased vs. biased estimator $\mathsf{E}[\widehat{\theta}] = \theta,$
- Mean square error (MSE) $$\begin{aligned}\mathsf{MSE}(\widehat{\theta}) &= \mathsf{E}[(\widehat{\theta} - \mathsf{E}[\widehat{\theta}])^2] + (\theta - \mathsf{E}(\widehat{\theta}))^2 \\ &= \mathsf{Var}\,\widehat{\theta} + (\mathsf{bias})^2.\end{aligned}$$
- Sample mean and sample variance

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

Statistics-MoM

- $\widehat{\mathrm{E}[X^k]} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ is an unbiased estimator for the kth moment of X.
- 2.2.6. Example. Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with parameters α and β. We know that

$$\mathrm{E}[X] = \alpha\beta, \qquad \operatorname{Var} X = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \alpha\beta^2.$$

Replacing the moments with M_1 and M_2, we obtain

$$M_1 = \hat{\alpha}\hat{\beta}, \qquad M_2 - M_1^2 = \hat{\alpha}\hat{\beta}^2.$$

This gives first $M_2 - M_1^2 = M_1\hat{\beta}$ and then

$$\hat{\beta} = \frac{M_2 - M_1^2}{M_1}, \qquad \hat{\alpha} = \frac{M_1}{\hat{\beta}} = \frac{M_1^2}{M_2 - M_1^2}.$$

Statistics-ML

- Maximum the likelihood $L(\theta) = \prod_{i=1}^{n} f_{X_\theta}(x_i).$

- Then, the location of the maximum is then chosen to be the estimator.

Procedure:

1. Take $In()$ on both sides.
2. Differentiate with respect to θ
3. Let the equation equals 0, then solve the equation.

Statistics-MoM & ML

Example: estimating the maximum number of a consecutive discrete series

Suppose the discrete series is $\{1, 2, 3, \ldots n\}$ and each appears with equal probability. Now a given sample is $\{1, 2, 96\}$.

- Using the method of moments,

$$E[X] = \frac{1+n}{2}, \text{ an estimation is given by } n = 2\overline{X} - 1$$

  In our case, $\hat{n} = 2 \times 33 - 1 = 65$. However this is ridiculous since there is already 96 in the sample.

- Using the method of maximum likelihood, we could first write $f(x)$ as

$$f(x) = \begin{cases} \frac{1}{n} & x \leq n \\ 0 & x > n \end{cases}$$

  $1/n$ when $x \leq n$ comes from the fact that each element from the series appears with equal probability.

  Therefore, to give maximum $L(n)$, we want our $\prod_i f(x_i)$ to be as large as possible. Since the sample size is fixed, this can be achieved by making n as close to $\max x$ as possible, in our case, $\overline{n} = 96$. However, intuitively this is also not a proper solution.

Statistics-Interval Estimate

- A 100(1-α)% two-sided confidence interval for $\boldsymbol{\theta}$ is an interval s.t.

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

- A 100(1-α)% (one-sided confidence interval) upper confidence bound is an interval s.t. $P[\theta \le L] = 1 - \alpha.$ And lower confidence bound is an interval s.t $P[L \le \theta] = 1 - \alpha.$
- Interval estimation for Mean with variance known

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \overline{X} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}.$$

- Interval estimation for Mean with variance unknown

$$T_{n-1} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \qquad \overline{X} \pm t_{\alpha/2,n-1} S/\sqrt{n}$$

Joint Distri. of Sample Mean and Variance

13.11. Theorem. Let $X_1, \ldots, X_n$, $n \geq 2$, be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then

(i) The sample mean $\overline{X}$ is independent of the sample variance S^2,

(ii) $\overline{X}$ is normally distributed with mean μ and variance σ^2/n,

(iii) $(n-1)S^2/\sigma^2$ is chi-squared distributed with $n-1$ degrees of freedom.

Statistics-Interval Estimate

- Interval Estimation of Variability

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

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

- One sided confidence interval

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}} \qquad \frac{(n-1)S^2}{\chi^2_{\alpha,n-1}} \le \sigma^2.$$

Bivariate Random Variable

Definition:

- *Discrete bivariate random variable* is a map $(X, Y): S \longrightarrow \Omega$, together with a function $f_{XY}: \Omega \longrightarrow R$.
- *Continuous bivariate random variable* is a map $(X, Y): S \longrightarrow R^2$ together with a function $f_{XY}: R^2 \longrightarrow R$. For $\Omega \subset R^2$

$$P[(X, Y) \in \Omega] = \iint f_{XY}(x, y) d(x, y)$$

- If X is continuous but Y is discrete, then

$$F_{XY}(x, y) = P[X \le x, Y \le y] = \sum_{v \le y} \int_{-\infty}^{x} f_{XY}(u, v) du$$

Marginal & Conditional Density

Definition:

In the context of two random variables X and Y, the distribution of X is known as the *marginal density* of X, which can be characterised by

$$f_{XY} = \sum_y f_{XY}(x,y) \qquad f_{XY} = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

Definition:

Conditional density is defined by

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Conditional Expectation

Discrete case:

$$E[Y|x] := \sum_{y} y f_{Y|x}(y) \qquad E[X|y] := \sum_{x} x f_{X|y}(x)$$

Continuous case:

$$E[Y|x] := \int y f_{Y|x(y)dy} \qquad E[Y|x] := \int x\, f_{X|y(x)dx}$$

Covariance

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - E[X]E[Y]$$

- If X and Y are independent, $Cov(X,Y) = 0$.
- If $Cov(X,Y) = 0$, X and Y are not necessarily independent.

Definition:

Pearson Correlation Coefficient are defined as:

$$\rho_{XY} := \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}$$

Variable Transformation

Definition:

Continuous bivariate random variable $((X, Y), f_{XY})$ and $\varphi: R^2 \to R^2$ a differentiable bijection map with inverse φ^{-1}. Then $(U, V) = \varphi o(X, Y)$ is a continuous bivariate random variable

$$f_{UV}(u, v) = f_{XY}(x, y) o \varphi^{-1}(u, v) |\det D\varphi^{-1}(u, v)|$$

Where $D\varphi^{-1}$ is the Jacobian of φ^{-1}

- Transform the map (X, Y) to $(Z, *)$
- Find f_{Z*} from f_{XY}
- Find the marginal density f_Z

Exercise 1

- Let X and Y be independent random variables following continuous **uniform** distributions on the interval $[0,1]$ and let $Z = X + Y$. Find the density f_Z of Z.

- Solution: Consider $H: (X, Y) \to (X + Y, Y) =: (U, V)$, so $D\varphi = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $D\varphi^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$, so $\det D\varphi^{-1} = 1$, so

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_{XY}(u - v, v)\,dv$$

Exercise 1

- Let X and Y be independent random variables following continuous **uniform** distributions on the interval $[0,1]$ and let $Z = X + Y$. Find the density f_Z of Z.

- Solution: $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-w) f_Y(w) dw = \int_0^1 f_X(z-w)\, dw = -\int_z^{z-1} f_X(u) du = \int_{z-1}^{z} f_X(u) du$
- The integral vanishes if $z < 0$ or $z > 2$. If $z \in [0,1]$, we got $\int_0^z 1 du = \mathrm{z}$ and If $z \in [1,2]$, we got $\int_{z-1}^{1} 1 du = 2 - \mathrm{z}$

Exercise 1*

- Let X and Y be independent random variables following continuous **exponential** distributions and let $Z = X + Y$. Verify that f_Z follows gamma distribution.

- Solution: $f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(u-v) f_Y(v) dv = \int_0^u f_X(u-v) f_Y(v) dv = \left(\frac{1}{\beta}\right)^2 \int_0^u e^{-u/\beta} dv = \left(\frac{1}{\beta}\right)^2 e^{-u/\beta} u$

- Notice that the MGF of the sum of the two i.i.d. r.v. is the product of their MGF.

Exercise 1**

- Now Let us consider the **discrete** case. You toss two dices, if the sum you get is not {7, 8, 9}, you pass; if you get 7, you must drink the wine but as much as you want; if you get 8, you must drink half the bottle; if you get 9, you must drink all. Therefore, we want to know the probability of get {7, 8, 9}.
- For $Y = 7$, the possibilities are $(X_1, X_2) = (1,6)(2,5)(3,4)(4,3)(5,2)(6,1)$ and X_1, X_2 are independent of each other, so $P[X_1, X_2] = P[X_1]P[X_2]$

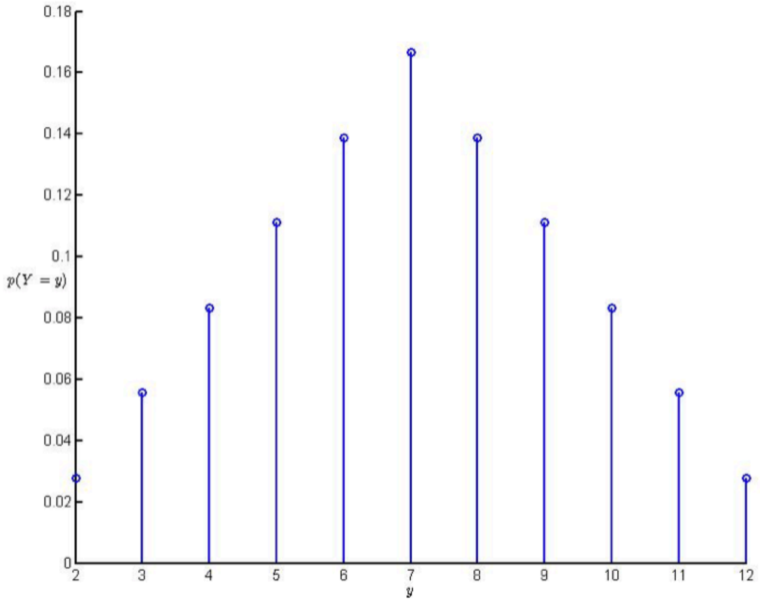
$$P[Y = 7] = \sum_{i=1}^{6} p(X_1 = i)p(X_1 = 7 - i)$$

- Discussion: the relationship with convolution?

Exercise 1**

$$P[Y=7]=\sum_{i=1}^{6} p(X_1=i)p(X_1=7-i)$$

- The formula is not only the discrete case of $f_{X+Y}(u)=\int_{-\infty}^{\infty} f_{XY}(u-v,v)dv$ but also a form of convolution $f(y)=f(x)*g(x)=\int f(w)g(y-w)dw$.



- For more information: https://www.cnblogs.com/yymn/p/4493165.html

Thank you!

Q & A