Assignment 1

Q1 : (a)(i) : WTS: $\sum_{i=1}^{n}(y_i-\bar{y})=0$.

$$\sum_{i=1}^{n}(y_i-\bar{y}) = \sum_{i=1}^{n} y_i - n\cdot\bar{y}$$
$$= \sum_{i=1}^{n} y_i - n\cdot\left(\frac{1}{n}\cdot\sum_{i=1}^{n} y_i\right)$$
$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i$$
$$= 0 \text{ as required.}$$

$\therefore$ Q.E.D.

(a) ii WTS : $S_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\cdot\bar{y}^2$.

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})$$
$$= \sum_{i=1}^{n} y_i\cdot(y_i-\bar{y}) - \bar{y}\cdot\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})}_{\Downarrow \; 0 \text{ as proved previously.}}$$
$$= \sum_{i=1}^{n} y_i^2 - \bar{y}\sum_{i=1}^{n} y_i$$

$$\bar{y} = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i$$
$$\therefore \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\cdot\sum_{i=1}^{n} y_i\cdot\sum_{i=1}^{n} y_i$$
$$= \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\sum_{i=1}^{n} y_i^2$$
$$= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \text{ as required.}$$

$\therefore$ Q.E.D

(b) $n=2$ $M \in \mathbb{N}$

(i) $\bar{y}=0$ $S_{yy}=M$.

$\bar{y}=\frac{y_1+y_2}{2}$ $S_{yy}=(y_1-\bar{y})^2+(y_2-\bar{y})^2$.

$\bar{y}=0 \Rightarrow y_1+y_2=0$.

$S_{yy}=M \in \mathbb{N} \Rightarrow y_1, y_2$ are integer or square root.

$\therefore$ $y_1=M=-y_2$ is one set : e.g $\{y_1, y_2\}=\{2,-2\}$

(ii) $\bar{y}=M$ $S_{yy}=0$.

$S_{yy}=(y_1-\bar{y})^2+(y_2-\bar{y})^2=0$

$y_1=\bar{y}=y_2 \Rightarrow y_1=y_2=M$ is the whole set

e.g : for one specific set : $\{y_1, y_2\}=\{3,3\}$

(C) (i) $z_i=\frac{1}{0.454} y_i$

$y_i=0.454 z_i$

(ii). $\bar{y}=32$. $\bar{z}=\frac{1}{0.454} \cdot \bar{y}$

$=70.48$

≈ 70.5

(iii) $S_{zz}=20$.

$S_{yy}=\sum_{i=1}^{2}(y_i-\bar{y})^2$

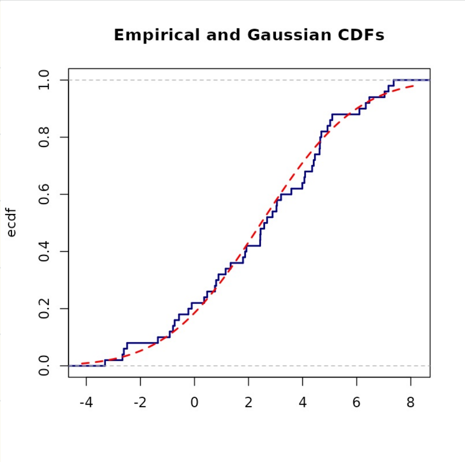
$=\sum_{i=1}^{2}(0.454 z_i-0.454\bar{z})^2$

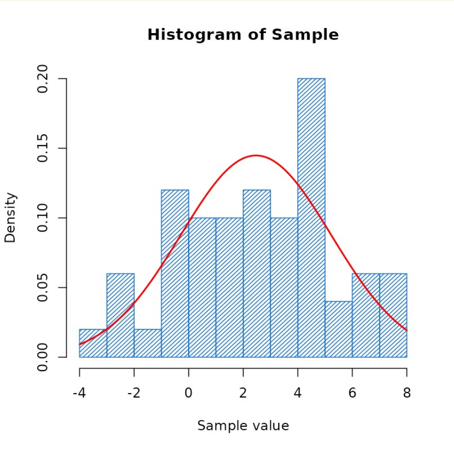
$=0.454^2 \cdot S_{zz}$

$=4.12$

Q2.

(a)





mean: 2.468

variance: 7.59

skewness: −0.226

kurtoisis: 2.284

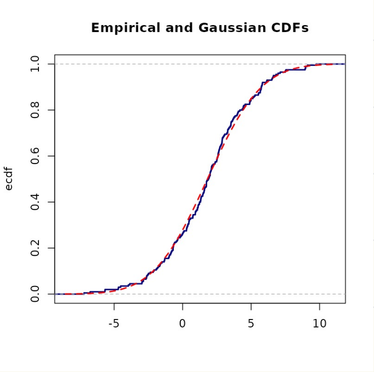
mean: 1.81

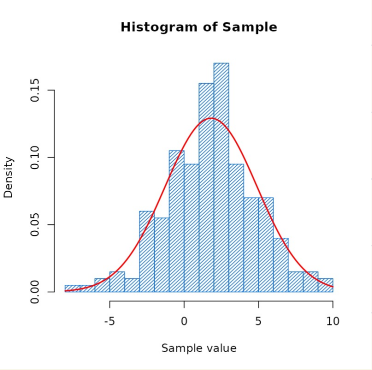
variance: 9.541

skewness: −0.109

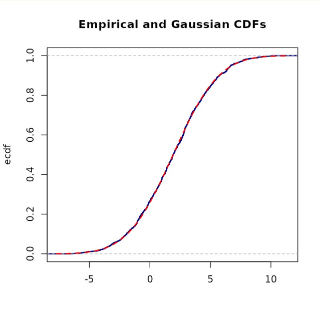
kurtoisis: 3.227

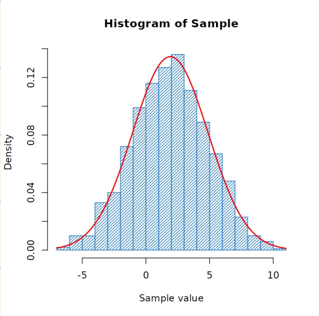
b





c.





mean: 1.917

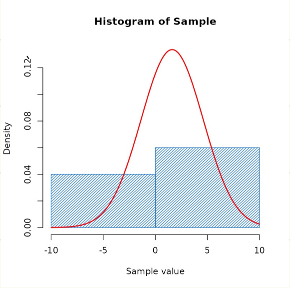
variance: 8.797

skewness: −0.013

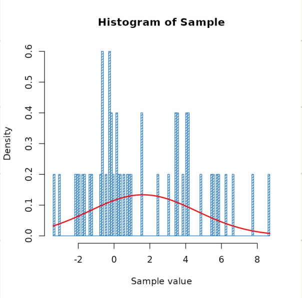
kurtoisis: 2.784

d. for the same value of $\mu=2$ and $\sigma=3$, for different sample, there exist different value of mean and variance but the value of kurtosis is always between 2 to 4 and with the increase in sample size, the value for kurtoisis become closer to 3 and the skewness becomes closer to the normal distribution with skewness 0. In this case, we may conclude that with large amount of data, the accuracy increases and trend become more obvious.

e. the lower number of bins will provide less amount of information than larger ones.



number of bin = 1



number of bin = 100

Q3:

(a) The potential population in this given set should be all the players of the fictional mobile game. This includes both active and not-active players.

(b) This should be an observational study. The data is collected by asking each players individually or collected from central game center. There has variate but there is no specific change in variate to see other cases so it is not experimental study or sample survey.

c. ① skill_grade:

type: ordinal variate.

Reason: The skill_grade is a rank system from lowest of F to highest of A $\Rightarrow A > B > C > D > E > F$

Thus, the order makes it an ordinal variable.

② time_overworld:

type: continuous variate.

Reason: The time_overworld is a collection of data where the time can be any number of a continuous variable so it is continuous variate.

③ device_age

type: Discrete variate.

Reason: The age of device can only be measured in whole years so it is a discrete variate.

4.

(a). minimum: 0.

25^{th} quantile: 2.94

median: 7.09

75^{th} quantile: 14.44

maximum: 65.65

$Range = y_n - y_0$

$= 65.65 - 0$

$= 65.65$.

$IQR = q(0.75) - q(0.25)$

$= 14.44 - 2.94$

$= 11.5$

(b). 1. The mean $= 10.121 > 7.090$

$\therefore$ The sample mean is higher than median.

2. The mean $>$ median $\Rightarrow$ the distribution should be positive skewness which is right tail.

3. skewness $= 1.685$

$$\text{skewness}(x) = \frac{E(x^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

$x \sim Exp(\lambda)$ $\quad \mu = E(x) = \frac{1}{\lambda}$. $\quad \sigma = \sqrt{Var(x)} = \frac{1}{\lambda}$.

$E(x^3) = M_x'''(0)$. $\quad M_x'''(t) = \frac{6\lambda}{(\lambda-t)^4}$.

$M_x(t) = \frac{\lambda}{\lambda - t}$ $\quad M_x'''(0) = \frac{6\lambda}{\lambda^4} = \frac{6}{\lambda^3}$

$\text{skew}(x) = \lambda^3 \cdot \frac{6}{\lambda^3} - 4 = 2$ close to 1.685

The skewness of time_combat is close to exponential distribution so shape should be relatively the same. but different at extreme cases.

C. 1. Shown in part B. The skewness of time-combat is very close to skewness of the exponential distribution. so the approximate curve should be very similar.

2. Both curve are positive skewed with right tail shows that most players spend a small amount of time in combat. For $X \sim Exp(\lambda)$ it is also obvious to see that when $P(X=x)$ and x is small, $P(X=x)$ holds alot.

3. The rate of decrease for both combat and exp are relatively the same.

4. Range: The range for combat is really similar to the exponential distribution

combat : 65.65

Exp : 63.4

Q5:

(a). y: 39 41 42 46 46 46 48 49 52 53, 59 11

y min = 39.

$q(0.25) = (11+1) \times 0.25$

$= 3^{rd}$ position $= 42$.

$q(0.5) = 46$.

$q(0.75) = 3 \times 3 = 9^{th}$ position $= 52$

$y_n = 59$.

(b) $\text{skew}_{\text{bowley}} = \dfrac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$

$$= \frac{52 + 42 - 2 \times 46}{52 - 42} = \frac{1}{5} = 0.2$$

(c) $\bar{y} = \dfrac{41+46+46+52+39+48+42+53+49+46+59}{11}$

$= 47.36$

$\sum_{i=1}^{11} (y_i - \bar{y})^2 = 336.5$.

$\frac{1}{n} \cdot \Downarrow = 336.5 \div 11 = 30.6$

$\sum_{i=1}^{11} (y_i - \bar{y})^3 = 858.51$.

$\frac{1}{n} \cdot \Downarrow = 78.05$

$\therefore g_1 = \dfrac{78.05}{30.6^{3/2}}$

$= 0.461$

d. skewness = 0.461, it is slightly positive with right tail.