Statistics

Statistics - Basic

- *mode* (眾數): the value with highest frequency
- *median* (中位數): the value in the middle
- *mean* (平均數): the average value ($\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$)

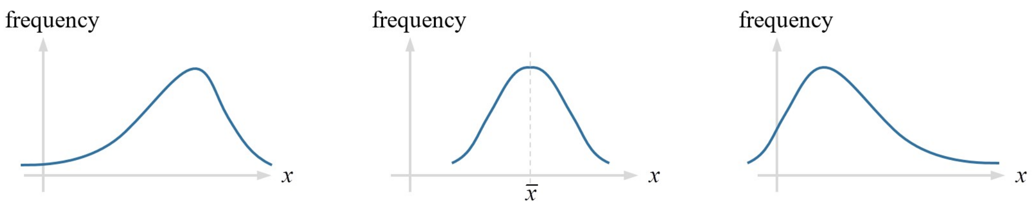
Statistics - Basic

Example. 23, 29, 20, 32, 23, 33, 25, 21

X	20	21	23	25	29	32	33
frequency	1	1	2	1	1	1	1

- mode(23, 29, 20, 32, 23, 33, 25, 21) = 23
- median(23, 29, 20, 32, 23, 33, 25, 21) = median(20, 21, 23, 23, 25, 29, 32, 33) = $\frac{23+25}{2} = 24$
- mean(20, 21, 23, 23, 25, 29, 32, 33) = $\frac{23+29+20+32+23+33+25+21}{8} = 25.75$

Skewness



(a) Negative skew: $x_{\text{median}} < x_{\text{mode}}$

(b) Symmetric: $\bar{x} = x_{\text{median}} = x_{\text{mode}}$

(c) Positive skew: $x_{\text{mode}} < x_{\text{median}}$

Statistical Graphs

Statistical Graphs (a.k.a. statistical diagrams, 統計圖表) = {bar chart (長條圖), histogram (直方圖), scatter plot (散佈圖), line chart (折線圖), boxplot (箱形圖), pie chart (圓形圖), …}

Expectation

X	x_1	x_2	…	x_m
frequency	f_1	f_2	…	f_m

$$p_i = \frac{f_i}{\sum^m f_i} \Rightarrow$$

X	x_1	x_2	…	x_m
P	p_1	p_2	…	p_m

Expectation (期望值, a.k.a. expected value),

$$E(X) = \frac{\sum_{i=1}^{m} x_i f_i}{\sum_{i=1}^{m} f_i} = \sum_{i=1}^{m} x_i P(X = x_i) = \sum_{i=1}^{m} x_i p_i \text{ where } p_i \equiv p(x_i) \equiv P(X = x_i) \equiv \frac{f_i}{\sum_{i=1}^{m} f_i}$$

$$\Rightarrow \boxed{E(X) = \sum_{i=1}^{m} x_i p_i} \text{ where } \boxed{\sum_{i=1}^{m} p_i = 1}$$

Expectation

Properties:

(1) $\boxed{E(aX+b)=aE(X)+b}$

Proof. $E(aX+b)=\sum_{i=1}^{m}(ax_i+b)P(x_i)=a\sum_{i=1}^{m}x_i p(x_i)+b\sum_{i=1}^{m}p(x_i)=aE(X)+b$ □

Bias of Estimates (估計之偏差)

Let $\hat{\theta}$: estimate of θ $(\hat{\theta}\in\Theta)$

where $\hat{\theta}$ is called the *estimate* (估計值), and

Θ is called the *estimator* (估計量),

the *bias of estimate* (偏誤估計) is defined as

$$\boxed{bias(\hat{\theta})\equiv E(\hat{\theta})-\theta}\ .$$

$\hat{\theta}$ is called

(i) *unbiased*, if $bias(\hat{\theta})=0$;

(ii) *biased*, if $bias(\hat{\theta})\neq 0$;

(iii) *asymptotically unbiased*, if $\lim_{m\to+\infty} bias(\hat{\theta})=0$ where m : number of samples.

Deviation

- centered data, $x_{ic} = x_i - \overline{x}$
- average absolute devitation $= \dfrac{1}{m}\sum_{i=1}^{m}\left|x_i - \overline{x}\right|$
- variance (方差) $= \dfrac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x})^2$
- standard deviation $= \sqrt{\dfrac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x})^2}$

Variance

The *variance*,

$$
\begin{aligned}
Var(X) &\doteq \frac{1}{m}\sum_{i=1}^{m}(X_i - E(X))^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\left\{X_i^2 - 2E(X)X_i + [E(X)]^2\right\} \\
&= \frac{1}{m}\left\{\sum_{i=1}^{m}X_i^2 - 2E(X)\sum_{i=1}^{m}X_i + m[E(X)]^2\right\} \\
&= \frac{1}{m}\left\{\sum_{i=1}^{m}X_i^2 - 2m[E(X)]^2 + m[E(X)]^2\right\} \qquad \left[\because E(X) = \frac{1}{m}\sum_{i=1}^{m}X_i\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}X_i^2 - [E(X)]^2 \\
&= E(X^2) - [E(X)]^2
\end{aligned}
$$

$$\Rightarrow \; Var(X) = E(X^2) - [E(X)]^2 \quad \text{or} \quad \boxed{Var(X) = \langle X^2 \rangle - \langle X \rangle^2}$$

Variance

Properties:

(1) $\boxed{E(aX+b)=aE(X)+b}$

Proof. $E(aX+b)=\sum_{i=1}^{m}(ax_i+b)P(x_i)=a\sum_{i=1}^{m}x_i p(x_i)+b\sum_{i=1}^{m}p(x_i)=aE(X)+b$ □

(2) $\boxed{Var(aX+b)=a^2Var(X)}$

Proof.
$$\begin{aligned}Var(aX+b)&=E\left((aX+b)^2\right)-[E(aX+b)]^2\\&=E(a^2X^2+2abX+b^2)-[aE(X)+b]^2\\&=a^2E(X^2)+2abE(X)+b^2-a^2[E(X)]^2-2abE(X)-b^2\\&=a^2\left[E(X^2)-[E(X)]^2\right]\\&=a^2Var(X)\end{aligned}$$
□

Sampling

Let μ : population mean
σ : population standard deviation

$\bar{x}$: sample mean
s : sample standard deviation

Suppose X_i ($i=1, 2, ..., m$) is a random sample from a population with mean μ, i.e. $E(X_i)=\mu$.

Then $E(\bar{X})=E\left(\frac{1}{m}\sum_{i=1}^{m}X_i\right)=\frac{1}{m}\sum_{i=1}^{m}E(X_i)=\frac{1}{m}(m\mu)=\mu$.

Hence the *expectation of the sample mean* is $\boxed{E(\bar{X})=\mu}$.

Sampling

$$
\begin{aligned}
Var(\bar{X}) &= Var\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right) \\
&= \frac{1}{m^2}\sum_{i=1}^{m} Var(X_i) \\
&= \frac{1}{m^2}(m\sigma^2) \\
&= \frac{\sigma^2}{m}
\end{aligned}
$$

Hence the *population variance of the sample mean* is $\boxed{Var(\bar{X}) = \frac{\sigma^2}{m}}$.

Sampling

The *population variance*,

$$
\begin{aligned}
\sigma^2 &= \frac{1}{m}\sum_{i=1}^{m}(X_i - \mu)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}(X_i^2 - 2\mu X_i + \mu^2) \\
&= \frac{1}{m}\left(\sum_{i=1}^{m} X_i^2 - 2\mu\sum_{i=1}^{m} X_i + m\mu^2\right) \\
&= \frac{1}{m}\left(\sum_{i=1}^{m} X_i^2 - 2m\mu^2 + m\mu^2\right) \\
&= \frac{1}{m}\sum_{i=1}^{m} X_i^2 - \mu^2
\end{aligned}
$$

$$\Rightarrow \quad \sigma^2 = E(X^2) - \mu^2$$

Sampling

The (biased) sample variance is

$$S^2 = \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}(X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$

$$= \frac{1}{m}\left(\sum_{i=1}^{m}X_i^2 - 2\bar{X}\sum_{i=1}^{m}X_i + m\bar{X}^2\right)$$

$$= \frac{1}{m}\left(\sum_{i=1}^{m}X_i^2 - 2m\bar{X}^2 + m\bar{X}^2\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}X_i^2 - \bar{X}^2$$

$$\Rightarrow \; S^2 = E(X^2) - \bar{X}^2$$

Sampling

Taking the expectation on $S^2 = E(X^2) - \bar{X}^2$, the estimate (biased) sample variance is

$$E(S^2) = E\left(E(X^2) - \bar{X}^2\right)$$

$$= E(X^2) - E(\bar{X}^2)$$

$$= E(X^2) - \left[E(\bar{X})\right]^2 - \left\{E(\bar{X}^2) - \left[E(\bar{X})\right]^2\right\}$$

$$= E(X^2) - \mu^2 - Var(\bar{X})$$

$$= \sigma^2 - \frac{\sigma^2}{m}$$

$$= \frac{(m-1)\sigma^2}{m}$$

Sampling

Let s be the unbiased sample variance, s.t. $E(s^2) = s^2 = \sigma^2$.

Hence $s^2 = \frac{m}{m-1}\sigma^2$

$$\Rightarrow \boxed{s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \mu)^2}$$