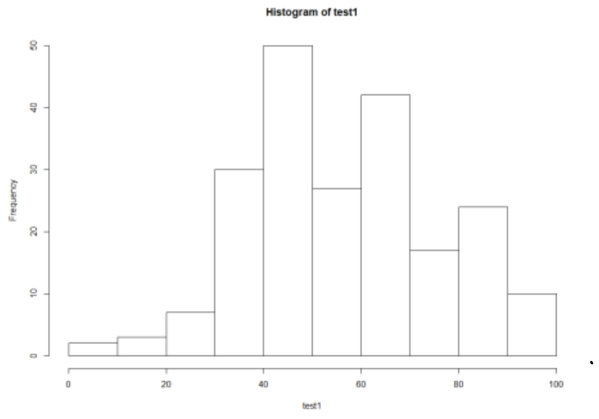
Lecture 6

May 9, 2018 6:26 PM



BSB ~~244~~ 249. or JHE.

Chapter 6: Descriptive Statistics

6.1 Numerical Summary.

Sample Mean.

n observations.

$x_1 \dots x_n$.

Sample $\Downarrow$ Statistic $\bar{x}$ $\Rightarrow$ parameter μ $\Uparrow$ Population

$$\bar{x} = \frac{x_1 + \dots x_n}{n} = \frac{\sum x_i}{n} \Rightarrow \mu.$$

Sample variance.

$$S^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \dots + (x_n-\bar{x})^2}{n-1}$$

$\Rightarrow \sigma^2$.

$\Rightarrow \sigma^2.$

Degree of freedom

$\underline{S \leftarrow n-1}$ degree of freedom.

$$(x_1-\bar{X})+(x_2-\bar{X})+\cdots+(x_n-\bar{X})=0$$

$$\bar{x}=\frac{x_1+\cdots x_n}{n}.$$

$$n\bar{X}=x_1+\cdots x_n$$

$$(x_1-\bar{X})+(x_2-\bar{X})\cdots\cdots(x_n-\bar{X})=0$$

Population mean and variance.

$$\mu=\frac{x_1+\cdots x_N}{N}$$

$$\sigma^2=\frac{(x_1-\mu)^2+\cdots+(x_N-\mu)^2}{N}$$

Sample Range.

$x_1 \cdots\cdots x_n.$ range:

$$r=\max(x_i)-\min(x_i).$$

Median:

a measure of central tendency that divide the data into equal parts, half below the median and half above median.

. . . 4 . 5 . 6

median.

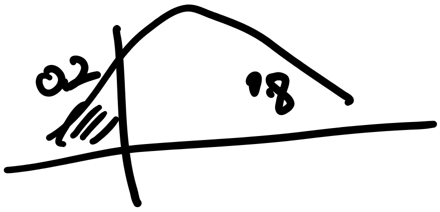
Even : 1, 2, $\underbrace{3, 4}$, 5, 6.

$$\text{median} = \frac{3+4}{2}$$

odd : 1, 2, 3, 4, 5

median = 3.

Quantile:



20th quantile

20% quantile.

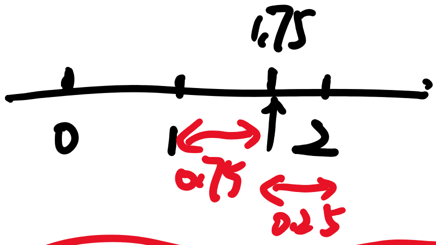
$$Q_1 = \frac{n+1}{4} (\text{th})$$

$$Q_3 = \frac{3(n+1)}{4} (\text{th.})$$

Even: 1, 2, 3, 4, 5, 6.

$n=6$ $\quad Q_1 = \frac{7}{4}$ th

$= 1.75$ (th)



1.75th

1 $\boxed{2}$ 3 $\boxed{4}$ 5 $\boxed{6}$ 7

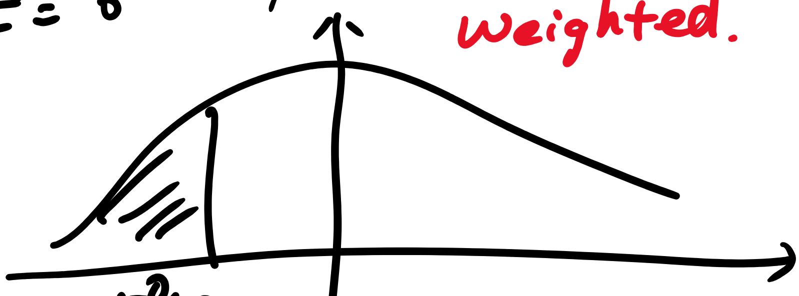
$$Q_1 = \frac{n+1}{4} = \frac{8}{4} = 2$$

$$Q_3 = \frac{3(n+1)}{4} = 6$$

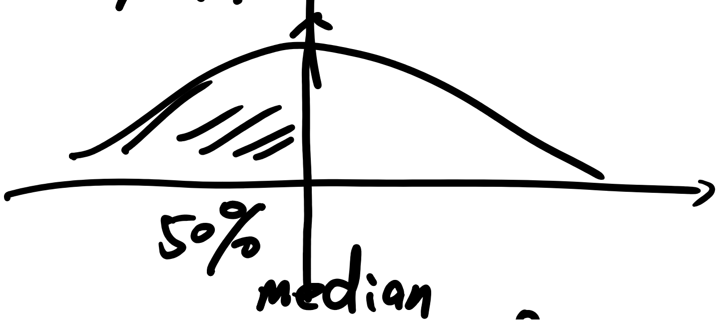
$$Q_1 = \underline{1 \cdot 0.25 + 2 \cdot 0.75}$$

weighted.

Q_1

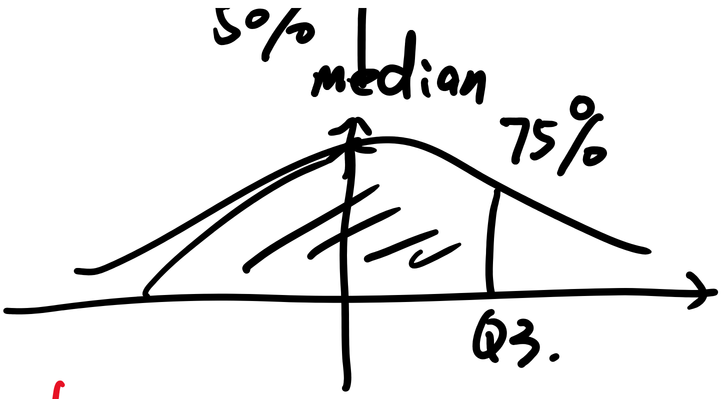


25% Q1.



50%

median



1 10 16 18 20 21.

$Q_1 = \frac{n+1}{4} = 1.75$.

1 0.75 0.25 10

$Q_1 = 1 \cdot 0.25 + 1.75 \cdot 10$

Mode:

1 2 2 2 3 3 3 3 4 5 6 7

mode = 3

A stem and leaf diagram

Stem	leaf

Example:

61, 63, 64, 65, 65, 66, 70, 71

71, 73, 75, 77, 78, 78, 79, 81,

71, 73, 75, 77, 78, 78, 79, 81,

83, 84, 84, 87, 88, 88, 92, 93, 95.

Stem	Leaf	Frequency
6	1 3 4 5 5 6	6
(9) 7	0 1 1 3 5 7 8 8 9	9
10 = (7+3) 8	1 3 4 4 7 8 8	7
3 9	2 3 5	3

← # in (the row contains median)

total sum of last two rows.

19th 20th.

N = 25

Median = 13th = 78

	Stem	Leaf	Fre
5	6		5
11 = 6+5	7		6
18 = 8 5+6+7	8		7
(8)	9	median	8
17 = 9+8	10		9
8			8

N = 43.

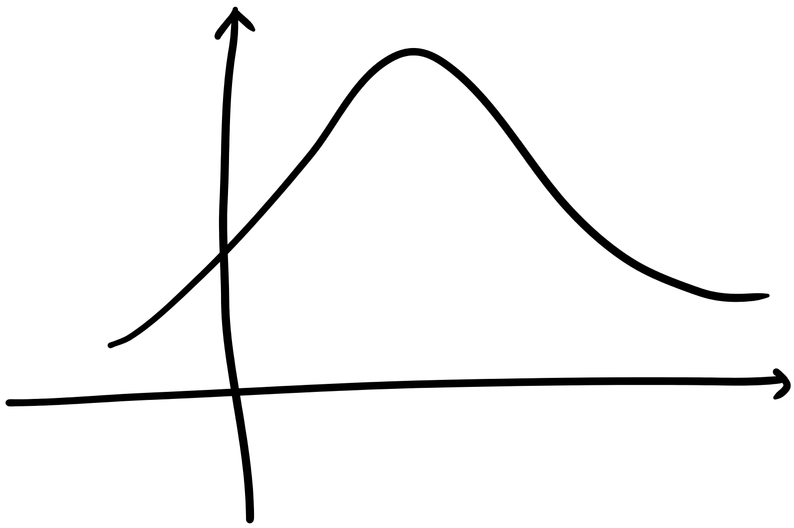
median = 22(th)

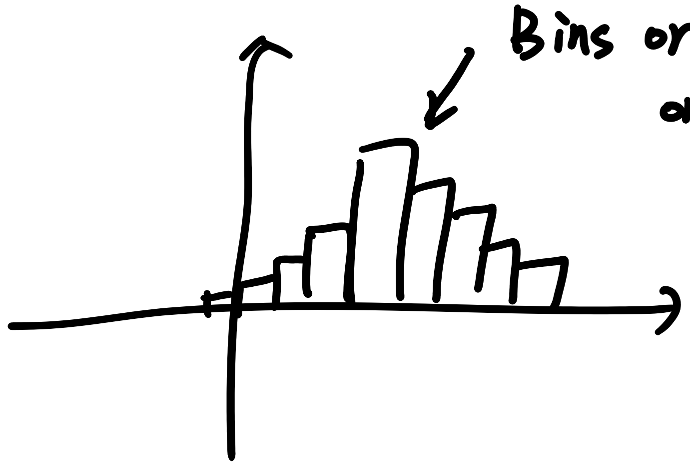
$Q_1 = \frac{n+1}{4}$ th = 6.5th. $\frac{66+70}{2} = 68$

$Q_3 = \frac{3(n+1)}{4}$ th = 19.5th. 84 + 87 ...

$Q_3 = \frac{3(n+1)}{4}$ th $=$ 19.5 th $\frac{84+87}{2} = 85.5$

Frequancy distribution and histogram.





Sample size n.

$x_1, x_2 \ldots\ldots x_n$.

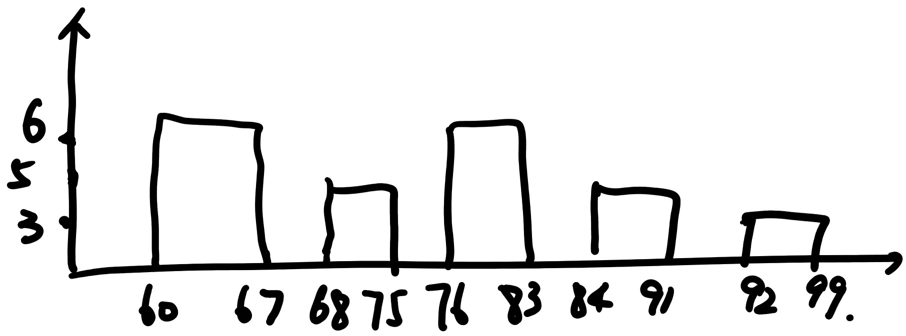
# of bins $= \sqrt{n}$.

$\underset{min}{61} - \underset{max}{95}$. $n = 25$

# of bins = 5.

	#	Relative Frequency
60 – 67	6	$\frac{6}{25}$
68 – 75	5	$\frac{5}{25}$
76 – 83	6	$\frac{6}{25}$
84 – 91	5	$\frac{5}{25}$
92 – 99	3	$\frac{3}{25}$.

8



Cumulative Frequency plot.

$6 = 6.$

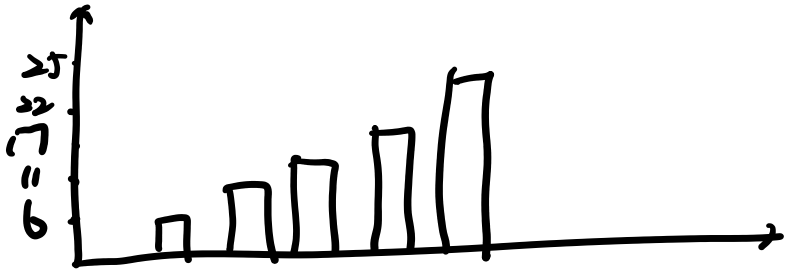
$6+5 = 11$

$6+5+6 = 17$

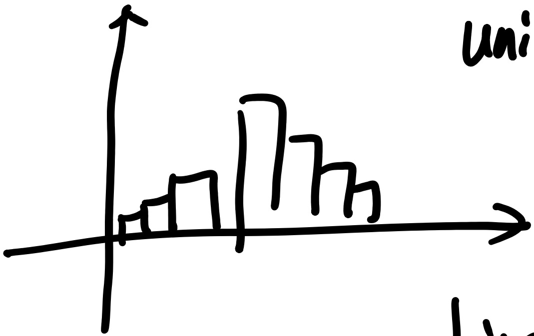
$6+5+6+5 = 22$

$6+5+6+5+3 = 25$

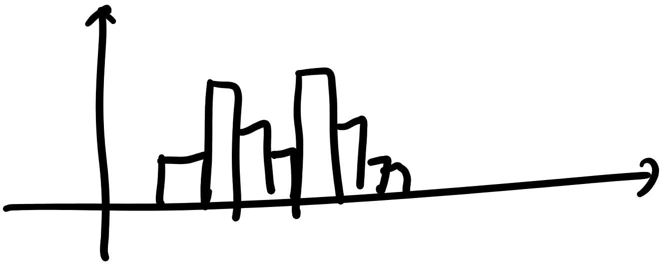
6 + 5 + 6 + 5 + 3 = 25



unimodal.



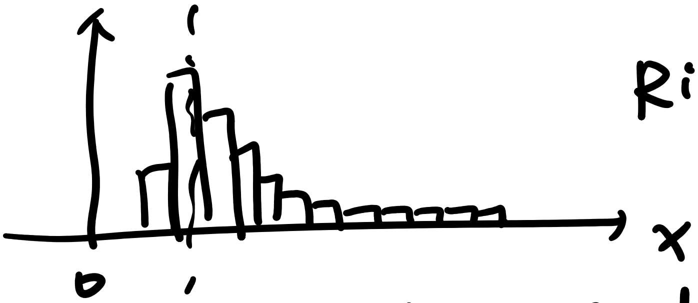
bimodal



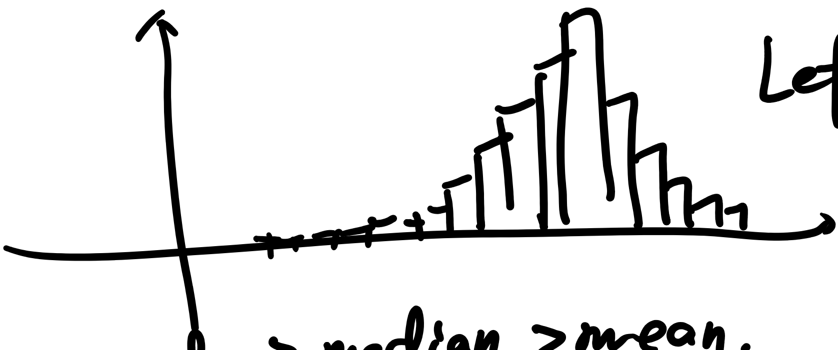
. . . . . .

Right-skewed

Income



mean > median > mode.



Left-skewed.

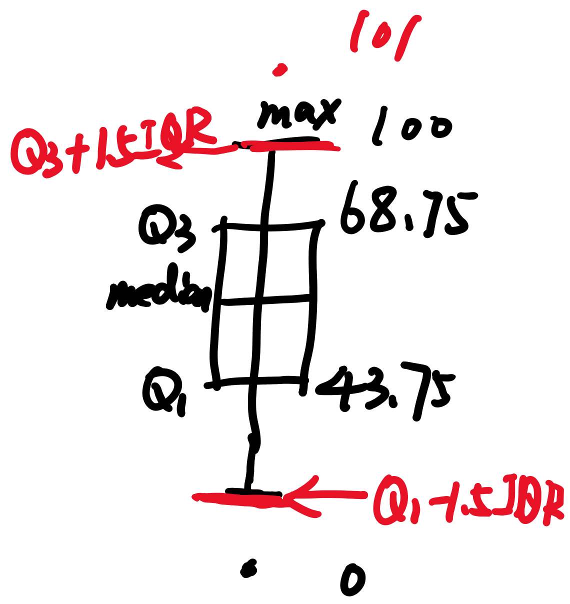
mode > median > mean.

Box plot: Box and wisker plot.

Box plot: Box and wisker plot.

$IQR = Q_3 - Q_1$

Interquantile Range.



If a point in the data set,

is greater than

$Q_3 + 1.5\ IQR$

or smaller than

$Q_1 - 1.5\ IQR.$

then it is an outlier.

Scater plot

plot y against x.

x: weight

y: height.

Scatter plot is designed to graphically display the potential relationship between two variables.

display ... between two variables.

Recall

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$= \frac{S_{xy}}{S_x \cdot S_y}$$

Population (σ_{xy})

$cov(x,y) = S_{xy}$ Sample

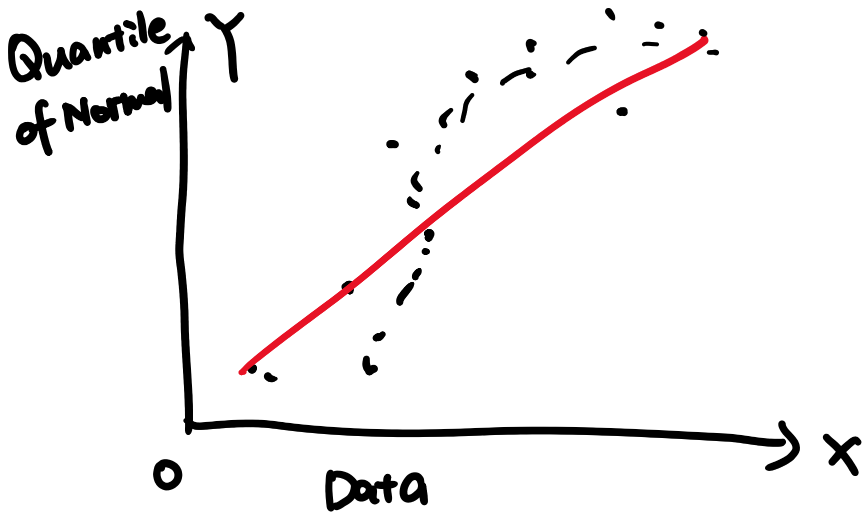
Correlation coefficient.

$$S_{xy} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}$$

$-1 \le r_{xy} \le 1$ linear association.

Normal probability Plot:

Check whether the sample follows Normal Distribution.

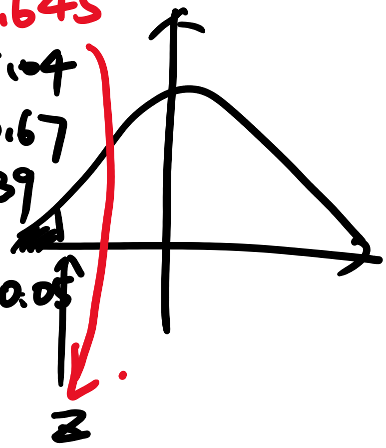


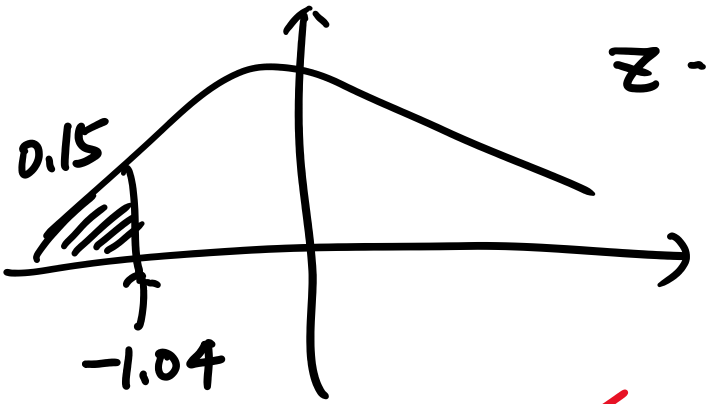
If the data set follows normal. The Normal probability plot will falls most on a straight line.

The Normal probability plot will
almost on a straight line.

n=10

j	x_j	$\frac{j-0.5}{n}$	z_j
1	176	0.05	-1.645
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220.	0.95.	1.645.

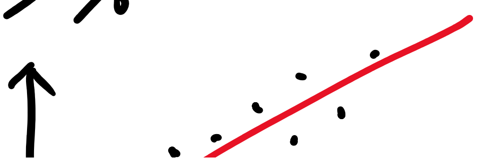




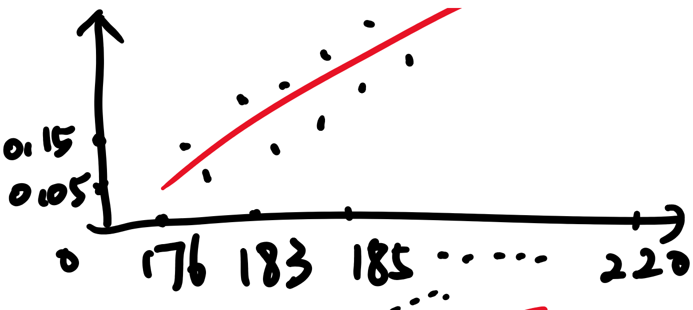
Z - standard normal.

5% value smaller than -1.645

5% value smaller than 5%

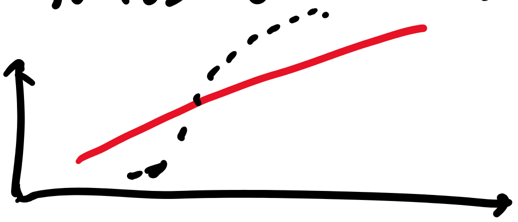


Normal



Normal

Data. x.



Not Normal.

Chapter 7:

Point Estimation of parameters and Sampling Distribution.

· $\bar{X}_1 \cdots \bar{X}_n$ Assume Poisson Distribution then whats the λ for it.

$$\hat{\lambda} = \frac{X_1 + \cdots X_n}{n} = \bar{X} \Longrightarrow \lambda.$$

statistic
a function of $x_1 \cdots x_n$

true λ.
or (true mean)
parameter

$\xRightarrow{\text{Inference}}$

Point Estimator:

A point estimator of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic.

value $\hat{\theta}$ of a statistic.

Then $\hat{\theta}$ is called point estimator.

$\hat{\mu} = \bar{x}$ sample mean

$\hat{\sigma} = S$ sample standard deviation.

$\hat{p} = \frac{x}{n}$ ← $\frac{\text{\# of 6 for a dice.}}{\text{\# of total trials.}}$

to test whether it's a fair dice.

Sampling Distribution:

The probability distribution for a $\underline{\text{statistic}}$ ($\bar{x}, S, \hat{p}$) is called sampling distribution.

Central Limit Theorem:

$x_1, x_2, \ldots x_n$ are i.i.d. sample of

⟺ Independent & identically distributed

⟺ Random Sample.

size n. with mean μ and variance σ^2.

Then.

The limiting distribution of $\underline{\bar{X}}$.

is $\underline{\text{Normal}(\mu, \frac{\sigma^2}{n})}$.

Then.

$Z = \frac{\bar{x} - \mu}{\ldots}$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

when $n \to \infty$, is a standard normal.

Examples in previous lecture note.

$\bar{x}_1$, $\bar{x}_2$ as the sample mean for two random sample.

Then. if they have σ_1^2, σ_2^2 are variance

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

as a standard normal approximately.

Example:

$x_1 \to \mu_1 = 5000 \quad \sigma_1 = 40 \quad n_1 = 16$

$x_2 \to \mu_2 = 5050 \quad \sigma_2 = 30 \quad n_2 = 25.$

Then assume x_1 and x_2 are independent

what's the probability that $\bar{x}_2 - \bar{x}_1$ is at least 25 ? $(\bar{x}_2 - \bar{x}_1)$.

$$Z = \frac{(25) - (5050 - 5000)}{\sqrt{\frac{40^2}{16} + \frac{30^2}{25}}}$$

$= -2.14$

$P(Z > -2.14) = 0.9838.$