COVARIANCE CONTINUATION:

covar(X, Y) is not independent of the units in which X and Y are measured.

eg: Suppose X & Y are rvs measured in cm such that cov(X, Y) = 0.15

If we change the unit to mm, then $X_1 = 10X$, $Y_1 = 10Y$ are new rvs covar(X_1, Y_1) = covar(10X, 10Y) = 15 $(k^2 var(X, Y))$

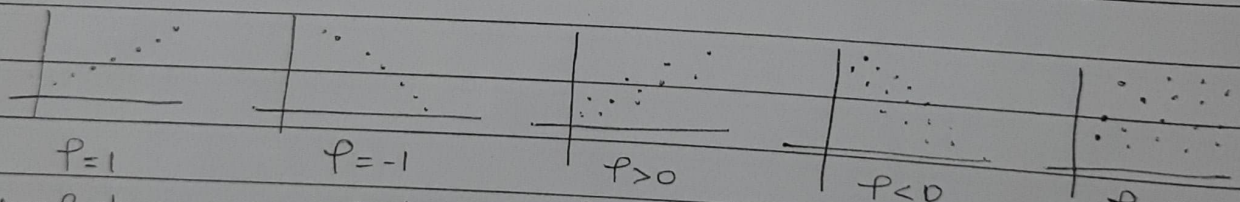
Solution So we normalize covariance to measure the correlation in absolute scale

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}$$

$\sigma_X = SD(X)$, $\sigma_Y = SD(Y)$

ρ_{XY} always lies between [-1, 1]

This is called coefficient of correlation btw X and Y.



perfect +ve relationship | perfect -ve relationship | linear +ve relationship | linear -ve relationship

"Measures only linear relationship btw 2 rvs"

RESULT $var(ax+bY) = a^2 var(X) + b^2 var(Y) + 2ab\,\sigma_X \sigma_Y \rho_{XY}$ (correlation)

$$var(ax+bY) = a^2 var(X) + b^2 var(Y) + 2ab\,cov(X,Y)$$

Proof:

$$var(ax+bY) = E[(ax+bY) - (a\bar{x} + b\bar{y})]^2$$
$$= E[a(X - M_x) + b(Y - M_y)]^2 \quad (a+b)^2$$
$$= E[a^2(X-M_x)^2 + b^2(Y-M_y)^2 + 2ab(X-M_x)(Y-M_y)]$$
$$= a^2 E(X-M_x)^2 + b^2 E(Y-M_y)^2 + 2ab\,E[(X-M_x)(Y-M_y)]$$
$$= a^2 var(X) + b^2 var(Y) + 2ab\,cov(X,Y)$$

$a = 1/\sigma_x \quad b = \frac{1}{\sigma_y}$

* $\text{var}\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_x}\right) = \frac{1}{\sigma_x^2}\text{var}(X) + \frac{1}{\sigma_y^2}\text{var}(Y) + 2ab\sigma_x\sigma_y r_{xy}$

$= 2 + 2r_{xy}$

* $\text{var}\left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}\right) = 2 - 2r_{xy}$

PROPERTIES:

① $-1 \le r_{xy} \le 1$

※ Variance of any rv is non-negative

$\text{var}\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y}\right) \ge 0$	$\text{var}\left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_x}\right) \ge 0$
$2 + 2r_{xy} \ge 0$	$2 - 2r_{xy} \ge 0$
$r_{xy} \ge -1$	$r_{xy} \le 1$

② $r_{xy} = 1$ iff $Y = aX + b$ for constants a, b & $a > 0$

$r_{xy} = -1$ iff $Y = aX + b$ for constants a, b & $a < 0$

Proof: * Suppose $r_{xy} = 1$

$\because 2 - 2r_{xy} \Rightarrow \text{var}\left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}\right) = 0$ if $\frac{X}{\sigma_x} - \frac{Y}{\sigma_y} = c$, constant

$\frac{Y}{\sigma_y} = \frac{X}{\sigma_x} - c \qquad Y = \underbrace{\frac{\sigma_y}{\sigma_x}}_{a} X - \underbrace{c\sigma_y}_{b}$

* let $Y = aX + b$

$$r_{xy} = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$$

eg If x and y are rvs with joint pdf
$f(x,y) = x+y \quad 0<x,y<1$.
Show that x and y are not linearly rel

Ans:

First we find marginal densities of x &

$$f_x(x) = \int_0^1 x+y\,dy = xy + \frac{y^2}{2}\Big|_0^1$$

$f_x(x) = x + 1/2 \quad 0<x<1$.

III^{rly} $f_y(y) = y + 1/2 \quad 0<y<1$

$$E(x) = \int x f_x(x)\,dx$$
$$= \int_0^1 x^2 + x/2\,dx = \frac{x^3}{3} + \frac{x^2}{4}\Big|_0^1$$

$E(x) = 1/3 + 1/4 = 7/12$

III^{rly} $E(y) = 7/12$

$$E(xy) = \iint xy\,f(x,y)\,dy\,dx$$
$$= \int_0^1\int_0^1 xy(x+y)\,dy\,dx$$
$$= \int_0^1 \frac{x^2y^2}{2} + \frac{xy^3}{3}\Big|_0^1\,dx$$
$$= \int_0^1 \frac{x^2}{2} + \frac{x}{3}\,dx = \frac{x^3}{6} + \frac{x^2}{6}\Big|_0^1$$

$E(xy) = 1/6 + 1/6 = 2/6 = 1/3$

$$E(x^2) = \int x^2 f_x(x)\,dx$$
$$= \int_0^1 x^2(x+1/2)\,dx$$
$$= x^4/4 + 1/2\,x^3/3\,y\Big|_0^1 = 1/4 + 1/6 = \frac{10}{24} = \frac{5}{12}$$

III^{rly} $E(y^2) = 5/12$

$$Var(x) = E(x^2) - E(x)^2 = \frac{5}{12} - \frac{49}{144} = \frac{60-49}{144} = \frac{11}{144}$$

III^{rly} $E(y^2) - E(y)^2 = var(y) = 11/144$.

Camlin | Page
Date / /

correlation $= \frac{\text{covar}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$

$= \frac{1/3 - \frac{49}{144}}{\frac{\sqrt{11}}{12}\frac{\sqrt{11}}{12}} = \frac{\frac{-1}{144}}{11/144} = \frac{-1}{144} \times \frac{144}{11} = \frac{-1}{11} \neq 1$

∴ They are not linearly related.

Mean & variance of sum of rvs

For any list of rvs $X_1, X_2 \ldots X_n$
Let $W_n = X_1 + X_2 + \ldots X_n$

$$E(W_n) = \sum_i E(X_i)$$

$$\boxed{\text{Var}(W_n) = \sum_i \text{var}(X_i) + 2 \sum_{i<j} \text{covar}(X_i, X_j)}$$

If $X_1, X_2 \ldots X_n$ are pairwise independant, or pairwise correlated then

$$\text{var}(W_n) = \sum_i \text{var}(X_i)$$

(Widely)

RESULT: Sum of iid rvs
If $X_1 = X_2 = \ldots X_n$ are iid rvs each with mean "μ" and SD "σ" then

$$E(W_n) = n\mu$$
$$\text{var}(W_n) = n\sigma^2$$

eg: $X_1, X_2 \ldots$ be a sequence of rvs with $E(X_i) = 0$ and $\text{covar}(X_i, X_j) = 0.8^{|i-j|}$. Find the mean & variance of $Y_i = X_i + X_{i-1} + X_{i-2}$

Soln:

$\text{covar}(X,Y) = E(XY) - E(X)E(Y)$
$\text{covar}(X,X) = E(X^2) - E(X)E(X)$

$E(Y_i) = 0$

$\text{Var}(Y_i) = \text{var}(X_i + X_{i-1} + X_{i-2})$
$= \text{var}(X_i) + \text{var}(X_{i-1}) + \text{var}(X_{i-2}) +$
$2\,\text{covar}(X_i, X_{i-1}) + 2\,\text{covar}(X_{i-1}, X_{i-2}) +$
$2\,\text{covar}(X_i, X_{i-2})$ . . . . . .

$= 1 + 1 + 1 + 2(0.8) + 2(0.8) + 2(0.64)$

$\text{covar}(Y_i) = 7.48$

eg: At a party of $n \geq 2$ people, each person puts his/her hat in a common box. They are shuffled and each person draws a hat, without replacement. If a person draws his/her hat it is success. What are mean & variance of the no. of successes?

Let $X_i = \begin{cases} 1 & \text{if person } i \text{ draws his/her own hat} \\ 0 & \text{o/w} \end{cases}$

$P(X_i = 1) = 1/n \qquad P(X_i = 0) = 1 - 1/n$

No. of successes $S_n = X_1 + X_2 + \ldots X_n$

Binomial

$E(X_i) = 1/n \qquad \text{var}(X_i) = 1/n(1 - 1/n)$

pq

$E(S_n) = n \times 1/n = 1$

$\text{covar}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$

To find $E(X_i X_j)$

$X_i X_j = \begin{cases} 0 & \text{o/w} \\ 1 & \text{iff } X_i = X_j = 1 \end{cases}$

$E(X_i X_j) = 1 \, P_{X_i X_j}(1,1) \qquad f_{(X|Y)}(x|y) = \dfrac{f(x,y)}{f_Y(y)}$

$= 1 \, P_{X_i|X_j}(1|1) \, P_{X_j}(1)$

Given $X_j = 1$, the j^{th} person draws his hat then $X_i = 1$ iff i^{th} person draws his hat from $n-1$ other hats $P_{X_i|X_j}(1|1) = \dfrac{1}{n-1}$

$E(X_i X_j) = \dfrac{1}{n-1} \dfrac{1}{n}$

$$\text{covar}(X_i, X_j) = \frac{1}{n(n-1)} - \frac{1}{n^2}$$

$$\text{var}(S_n) = \text{var}(X_1) + \text{var}(X_2) + \ldots \text{var}(X_n) + {}^{n}C_2 \, \text{cov}(X_i, X_j)$$

$$= n\left(\frac{1}{n} - \frac{1}{n^2}\right) + 2\,\frac{n(n-1)}{2}\left[\frac{1}{n(n-1)} - \frac{1}{n^2}\right]$$

$$\text{var}(S_n) = \left(1 - \frac{1}{n}\right) + \left(1 - \frac{n-1}{n}\right) = 1$$

Note: Suppose each person immediately returns the hat into Iron. What are mean & variance?

$E(X_i) = E(S_n) = 1 \qquad \text{var}(S_n) = n\left(\frac{1}{n} - \frac{1}{n^2}\right) = 1 - 1/n$

LIMIT THEOREMS

fundamental issues related to the asymptotic behaviour of sequence of rvs

1. Markov inequality
2. Chebyshev's inequality
3.
4.

① Markov Inequality

If X is a rv that takes non-negative value then for any $a > 0$

$$P[X \geq a] \leq \frac{E(X)}{a}$$

upper bound for $P[X \geq a]$

Proof:

$$E(x) = \int_0^\infty x f(x)\,dx \quad \rightarrow \text{removing this will become} \leq$$

$$= \int_0^a x f(x)\,dx + \int_a^\infty x f(x)\,dx$$

$$\leq \int_a^\infty x f(x)\,dx \quad \text{taking min value for } x \text{ in interval } a \text{ to } \infty$$

$$\leq \int_a^\infty f(x)\,dx$$

$$\leq a \int_a^\infty f(x)\,dx$$

$$\boxed{E(x) \leq a\, P[X \geq a]}$$

② Chebyshev's inequality.

If X is a rv with μ and variance σ^2, then for any values of $K > 0$

$$P[|X-\mu| \geq K] \leq \frac{\sigma^2}{K^2}$$

Proof: $(X-\mu)^2$ is a non-negative rv

By Markov's inequality,

$$P[(X-\mu)^2 \geq K^2] \leq \frac{E(X-\mu)^2}{K^2} = \frac{\sigma^2}{K^2}$$

$$P[|X-\mu| \geq K] \leq \frac{\sigma^2}{K^2}$$

If a rv has small var then the probability that it takes values far away from mean is small.

Note: If $K = n\sigma$, Chebyshev's inequality between

$$P[|X-\mu| \geq n\sigma] \leq \frac{1}{n^2}$$

The probability that X deviates from mean by atleast n SD is $\leq 1/n^2$.

eg Suppose that no. of items produced in a factory in a week is a rv with a mean of 500

i. What is it can be said about the prob that this week is production will be atleast 1000?

ii. If the variance is known to be 100, what can be said about the probability that the week is production will be btw 400 & 600?

Soln:

i. X no. of items produced in a week

$X \geq 0$ & $E(X) = 500$

By Markov's inequality,

$$P[X \geq 1000] \leq \frac{500}{1000} = 0.5$$

ii. Given $\sigma^2 = 100$

By Chebyshev's inequality,

$$P[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}$$

$k = 100$
$400 - 500 = -100$
$600 - 500 = 100$

$$P[|X - 500| \geq 100] \leq \frac{100}{100^2} = 0.01$$

$$P[|X - 500| < 100] \geq 0.99$$

$$P[400 < X < 600] \geq 0.99$$

eg A person post office handles an average of 10000 letters per day. What can be said about the prob that it will handle
i. atleast 15000 letters tomorrow?
ii. lesser than 15000 letters tomorrow?
If the variance is 2000, what can be said about the probability that it will handle btw 8000 and 12000 letters tomorrow?

Soln: X: no. of letters handled

$E(X) = 10000$ $\geq E(X)/a$ Markov's inequality

i. $P[X \geq 15000] \leq \frac{10000}{15000} = \frac{2}{3}$

ii. $P[X < 15000] \geq 1/3$

$\sigma^2 = 2000$ Chebyshev's inequality

$P[8000 < X < 12000] = P[|X - \mu| < 2000] = ?$

$$P[|X - \mu| \geq 2000] \leq \frac{2000}{2000^2} = \frac{1}{2000}$$

$$\therefore P[8000 < X < 12000] = 1 - 1/2000 = 1999/2000$$

Note: The bounds obtained for the probabilities in Markov's and Chebyshev's inequalities are not close to actual probabilities.

eg: roll a dice X - o/c of the die

$E(X) = 1/6(1+2+3+4+5+6) = 7/2 \quad (21/6)$

$E(X^2) = 1/6(1^2+2^2+3^2+4^2+5^2+6^2) = 91/6$

$Var(X) = E(X^2) - E(X)^2 = 91/6 - \left(\frac{21}{6}\right)^2 = 35/12$

By Markov's inequality

$P[X \geq 6] \leq \frac{21/6}{6} = \frac{21}{36} = 0.583$

By Chebyshev's inequality

$P[|X - 21/6| \geq 3/2] \leq \frac{35/12}{9/4} = 1.296$

But the trivial bound $P[|X - 21/6| \geq 3/2] \leq 1$

Exact probabilities.

$P[X \geq 6] = 1/6 = 0.167 \not\leq 0.583$ (too far)

$P[|X - 21/6| \geq 3/2]$

$|X - 21/6| \leq 3/2$

$-3/2 \leq X - 21/6 \leq 3/2$

$21/6 - 3/2 \leq X \leq 3/2 + 21/6$

$12/6 \leq X \leq 30/6$

$2 \leq X \leq 5$

$P[X \leq 2 \text{ or } X \geq 5]$ $P[2 \leq X \leq 5] = 4/6 = 2/3$

$= 0.667 \leq 1$

(too far)

Note: If $Var(X) = 0$, then X is constant with probability 1.

We can prove this ↑ using Chebyshev's inequality.

Laws of large numbers

weak law	strong law
describes how a seq of prob converges	describes how a seq of rv behave in the limit

* Weak law of large numbers (WLL)

Let $X_1, X_2, \ldots X_n$ be a sequence of iid rvs each having limit mean $E(X_i) = \mu$ and $var(X_i) = \sigma^2$

Then for any $\epsilon > 0$

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| > \epsilon\right\} \to 0 \text{ as } n \to \infty$$

$$\lim_{n\to\infty} P\left\{\left|\frac{X_1 + X_2 + \ldots X_n}{n} - \mu\right| > \epsilon\right\} = 0$$

(or)

$$P\left\{\left|\frac{X_1 + X_2 + \ldots X_n}{n} - \mu\right| \leq \epsilon\right\} \to 1 \text{ as } n \to \infty$$

$\frac{X_1 + X_2 + \ldots X_n}{n}$ ↓ sample mean; μ ↓ true mean

Proof:

Let $M_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$

$$E(M_n) = \frac{1}{n} E(X_1) + E(X_2) + \ldots E(X_n) = \frac{1}{n} n\mu = \mu$$

$$var(M_n) = \frac{1}{n^2}\left[var(X_1) + var(X_2) + \ldots + var(X_n)\right] = \frac{1}{n^2} n\sigma^2$$

$$var(M_n) = \frac{\sigma^2}{n}$$

$var(ax) = a^2 var(x)$

$var(ax+Y) = var(x) + var(Y)$

By Chebyshev's inequality,

$$P\{|M_n - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2} \quad \text{for any } \epsilon > 0$$

for any fixed $\epsilon > 0$

$$P\left\{\left|\frac{X_1 + X_2 + \ldots X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0 \text{ as } n \to \infty$$

WLL and relative frequency interpretation of probability:

In a repeatable experiment with sample space S, Let A be an event. Suppose the experiment is repeated independantly.

For $i = 1$ to n,

Let $X_i = \begin{cases} 1 & \text{if } A \text{ occurs in } i^{th} \text{ trial} \\ 0 & \text{if } A \text{ does not occur} \end{cases}$

$E(X_i) = 1P(A) + 0\,P(\bar{A}) = P(A)$

and $S_n = X_1 + X_2 + \dots X_n = \#(A)$ { no. of times A occurs in n trials

Then by WLL

$$\lim_{n\to\infty} P\left\{\left|\frac{X_1+X_2+\dots X_n}{n} - E(X_i)\right| \geq \epsilon\right\} = 0$$

$$\lim_{n\to\infty} P\left\{\left|\frac{\#(A)}{n} - P(A)\right| \geq \epsilon\right\} = 0$$

When n is large, no. of times A occurs out of n times is close to P(A).

* Strong law of large numbers (SLL)

$$P\left\{\lim_{n\to\infty} \frac{X_1+X_2+\dots X_n}{n} = \mu\right\} = 1$$

With probability 1, the sequence of sample means converges to the true mean.

Central limit theorem (CLT)

Sum of large number of iid rvs have a distribution that is approximately normal.

Let $X_1, X_2 \ldots$ be a sequence of iid rvs each with mean $E(X_i) = \mu$ and $var(X_i) = \sigma^2$. Then distribution of $\frac{X_1 + X_2 + \ldots X_n - n\mu}{\sigma\sqrt{n}}$ tends to follow SND as $n \to \infty$

ie) $\frac{\frac{X_1 + X_2 + \ldots X_n}{n} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is close to SND as $n \to \infty$

$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ $\quad z = \frac{X - \mu}{\sigma}$

eg The lifetime in hrs, of a component is a rv with mean 500 and SD 35. What is the prob that a rv random sample of 49 bulbs will have mean life btw 488 and 505 hrs?

Soln:

$\bar{x}$ - mean life of $n = 49$ bulbs

By CLT $\mu = 500$, $\sigma_{\bar{x}} = 35/\sqrt{49}$

$\bar{X} = N(500, 35/\sqrt{49}) \sim N(500, 5)$ $\quad z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 5}{5}$

$P[488 < \bar{X} < 505] = P[-2.4 < z < 1]$

$= \Phi(1) - \Phi(-2.4)$

$= 0.8413 - 0.0082$

$= 0.8331$

eg: The distribution of annual earnings of all bank managers has a negatively skewed distribution with mean \$19000 and \$2000. In a random sample of 30 managers, what is the probability that mean annual earnings is more than \$19750?

Soln: $\bar{x}$ - mean annual earnings of 30 managers

$\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

By CLT $\mu_{\bar{x}} = 19000$ $\sigma_{\bar{x}} = \frac{2000}{\sqrt{30}}$ =

$\bar{X} \sim N\left(19000, \frac{2000}{\sqrt{30}}\right)$ $\therefore z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$P[\bar{x} > 19750] = P\left[z > \frac{19750 - 19000}{2000/\sqrt{30}}\right]$

$= P[z > 2.054]$

$= 1 - \varphi(2.054)$

$= 1 - 0.9798$

$= 0.0202$

eg: The time it takes for a student to finish an aptitude test has a pdf $f(x) = \begin{cases} 6(x-1)(2-x) & 1 < x < 2 \\ 0 & dw \end{cases}$

Approximate the prob that the average length of time it takes for a random sample of 15 students to complete the test is less than 1 hr 25 mins?

Soln:

For $i = 1$ to 15, let x_i be the time for i^{th} student to finish the test

$\bar{x} = \frac{x_1 + x_2 + \dots x_{15}}{15}$

$\mu = E(x_i) = \int_1^2 x\, 6(x-1)(2-x)\, dx = \frac{1+2}{2} = \frac{3}{2} = 1.5$

$E(x_i^2) = \int_1^2 x^2 f(x)\, dx = \frac{23}{10} = 2.3$

$\sigma^2 = Var(x) = E(x_i^2) - E(x_i)^2 = 2.3 - 2.25 = 0.05$

$\bar{X} \sim N\left(1.5, \frac{\sqrt{0.05}}{\sqrt{15}}\right)$ $P(z < -1.44)$

$P\left(\bar{x} < \frac{85}{60}\right) = 0.0749$

If 20 random numbers are selected from (0,1) approximate the prob that sum of them is atleast 8.

Ans:

For i=1 to 20, let X_i be the i^{th} no. selected

$$P\left[\sum_{i=1}^{20} X_i \geq 8\right] = P\left[\bar{X} \geq 8/20\right]$$

$X \sim$ uniform (0,1) * $\mu = \frac{a+b}{2} = \frac{0+1}{2} = 0.5$

* $\sigma^2 = \frac{(a-b)^2}{12} = \frac{(0-1)^2}{12} = \frac{1}{12}$

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}) = N(0.5, 1/\sqrt{240})$$

$$P\left[\bar{X} \geq 8/20\right] = P\left[z \geq \frac{8/20 - 0.5}{1/\sqrt{240}}\right] \rightarrow 0.0645$$

$$= P\left[z \geq -15.504\right] = 1 - \varphi(-1.55)$$

$$= 1 - 0.0606 = 0.9394$$

STATISTICS METHODS

- descriptive
- inferential
  - estimation
  - hypothesis

$$CI = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$ — Confidence interval

$n \geq 30$ large (becomes normal dist)

$n < 30$ (follows t-distribution)

degrees of freedom $\rightarrow n-1$

$$CI = \bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

eg: $n = 70 \geq 30$, σ unknown $\quad \bar{X} \sim$ normal$(0,1)$ $\quad \mu$ σ SND

$\bar{x} = 1759 \quad s = 380$

99% $\rightarrow$ 2.58 $\quad$ 95% $\rightarrow$ 1.96 $\quad$ 90% $\rightarrow$ 1.645 ※

$$CI = 1759 \pm 2.58 \frac{380}{\sqrt{70}}$$

$$= 1759 \pm 117.180$$

$$CI = 1641.82 \,/\, 1876.18$$

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}$$ — population proportion p

$$CI = \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

eg: $\hat{p} = \frac{25}{100} = 0.25 \quad$ 95%

$$CI = 0.25 \pm 1.96 \sqrt{\frac{0.25(0.75)}{100}}$$

$$= 0.25 \pm 0.00735 \quad 0.08487$$

$$CI = 0.25735 \,/\, 0.24265 \quad 0.3348 \,/\, 0.1651$$

Sampling error

parameter = statistics + error

$\bar{x} \pm (Z\sigma/\sqrt{n})$ → margin of error

$$e = Z\sigma/\sqrt{n} \qquad n = \frac{Z^2\sigma^2}{e^2} = \left(\frac{Z\sigma}{e}\right)^2$$

"Population mean" check

eg: $\sigma = 45$ $e = \pm 5$ 90%. n (sample size) = ?

$$n = \left(\frac{z\sigma}{e}\right)^2 = \left(\frac{1.645 \times 45}{5}\right)^2 = 219.188$$

≈ 220 (always round up)

eg: 95%. $e = \pm 5$ $\sigma = 15$

$$n = \left(\frac{z\sigma}{e}\right)^2 = \left(\frac{1.96 \times 15}{5}\right)^2 = 34.5744$$

≈ 35

Population proportion

$$e = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad n = \frac{z^2 \hat{p}(1-\hat{p})}{e^2}$$

eg: $\hat{p} = 0.12$, $e = 3$, 95%.

$$n = \frac{(1.96)^2 \, 0.12 \, (0.88)}{9/100^2} = 0.0452 \times 100^2 = 450$$

Characteristics for selecting estimators (unbiased)

* unbiased (how much calc time value differs from true value is called bias)
* efficient (σ^2 small more efficient)
* sufficient
* consistent

* unbiased

$E(\hat{\theta}) - \theta = 0$ bias $= 0$ [$\hat{\theta}$ is an unbiased estimation of θ]

sample mean & sample median → unbiased

Hypothesis:

is a statement or claim about the parameter of one or more populations

Date

Two sided hypothesis

$H_0 : \mu = 8.0$ (null hypothesis)

$H_1 : \mu \neq 8.0$ (opp to null hypothesis)

→ alternate hypothesis / research

One sided hypothesis

$H_0 : \mu \leq 8.0$	$\mu \geq 8.0$
$H_1 : \mu > 8.0$	$\mu < 8.0$

"Null hypothesis is a hypothesis of no diff".
It always contain equality sign. It may / may not be rejected.

Hypothesis testing begins with the assumption that null hyp is true. Either one hypothesis will be the result / true / claim.

CRITICAL VALUE

if $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$

test statistics $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim SND$

$$\boxed{z = \frac{t - E(t)}{SE}}$$

t - statistics
SE - standard error

* p-value method: compare p value & α
if p-value $< \alpha$, reject H_0

* critical value method: compare $|z|$ & $|z_\alpha|$

$|z| > |z_\alpha|$ reject
$|z| < |z_\alpha|$ donot reject

σ known case

$H_0: \mu \leq 12 \quad H_1: \mu > 12 \rightarrow$ right-tailed

$n = 40 \quad \bar{x} = 13.25 \quad \sigma = 3.2 \quad \alpha = 0.05 \quad z_\alpha = 1.96$

test statistics: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{13.25 - 12}{3.2/\sqrt{40}} = 0.39 \times \sqrt{40} = 2.466$

By critical value approach,

$|z| > z_\alpha \qquad 2.466 > 1.96$

reject H_0

∴ Service goal of 12 mins or less is not achieved

* Left tailed $P[Z \leq z]$
* Right tailed $1 - P[Z \leq z]$
* Two-tailed $2[P[Z \leq z]] \;/\; 2[1 - P[Z \leq z]]$

σ unknown case

use t-distribution

rejection rule for critical value approach:

$H_0: \mu \geq \mu_0$ rej H_0 if $t \leq -t_\alpha$ left

$H_0: \mu \leq \mu_0$ rej H_0 if $t \geq t_\alpha$ right

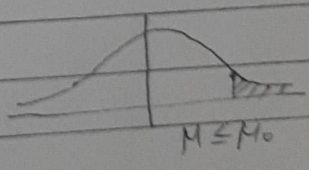
$H_0: \mu = \mu_0$ rej H_0 if $|t| > |t_\alpha|$ 2 tailed

eg: (shown)

$H_0: \mu \leq 130$

$H_1: \mu > 130$

$n = 9 \quad \bar{x} = 131.08 \quad s = 1.5 \quad \alpha = 0.01$



test statistics: $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{131.08 - 130}{1.5/\sqrt{9}} = 2.16 = z$

critical value approach

For $\alpha = 0.01$, $z_\alpha = 2.58$

$P[Z \leq z] = 0.01$

$z_\alpha = -2.33 \qquad |z| < |z_\alpha|$ ∴ donot reject H_0

p-value approach:

$1 - P[Z \leq 2.16] = 1 - 0.9846 = 0.0154$

$0.0154 > 0.01$

p-value $> \alpha$ Hence do not reject H_0

∴ Manufacturer's claim is correct

eg: $H_0 : \mu = 1000$

$H_1 : \mu > 1000$

$n = 64 \quad \bar{X} = 1,038 \quad s = 146 \quad \alpha = 0.05$

test statistics : $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1,038 - 1000}{146/\sqrt{64}} = 2.0821$

critical value approach

For $\alpha = 0.05 \quad z_\alpha = 1.645$

$z > z_\alpha$ reject H_0

σ - unknown

eg: $H_0 : \mu \geq 8$

$H_1 : \mu < 8$

$n = 25 \quad \bar{X} = 7.73 \quad s = 0.77 \quad \alpha = 0.05$

test statistics $= \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ with dof $n-1 = 24$

$= \frac{7.73 - 8}{0.77/\sqrt{25}}$

$\boxed{t = -1.75}$

Critical value approach:

$t = \bar{x} \mu_0$

refer table (t-dist)

for $\alpha = 0.05$, one tailed dof $= 24 \quad t_\alpha = 1.71$

$1.75 > 1.71 \quad |t| > |t_\alpha| \quad$ ∴ reject H_0

p-value approach:

$P[Z \leq -1.75] = P[Z \geq 1.75]$ (∵ they are symmetri

eg:

$H_0: \mu \geq 5$

$H_1: \mu < 5$

$n = 20 \quad \bar{x} = 4.6 \quad S = 2.2 \quad \alpha = 0.05$

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{4.6 - 5}{2.2/\sqrt{20}} = -0.813$$

Critical value approach, $\alpha = 0.05 \quad df = 19 \quad t_\alpha = 1.73$

~~$|t| \leq |t_\alpha|$~~ $t \leq -t_\alpha$ is false

∴ Donot reject H_0

95% confidence interval for μ:

$$CI = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

For LOC 95%, df = 19,

$t_{\alpha/2} = 2.09$ (always 2 tail)

$$= 4.6 \pm (2.09) \frac{2.2}{\sqrt{20}} = 4.6 \pm 1.028$$

$$= 5.628 \,/\, 3.572$$

eg: For the sample, $\bar{x} = 14.4$, $S = 0.158$, $n = 5$

$H_0 = \mu = 14.0$

$H_1 = \mu \neq 14.0$

$\alpha = 0.05$

test statistics: $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{14.4 - 14}{0.158/\sqrt{5}} = 5.66$

$t_\alpha = 2.13$ p-value = 0.004

∴ Claim is wrong.

Camlin
Date

independant samples

eg: $n_1 = 20$ $n_2 = 25$ σ_1, σ_2 known → ND
$\bar{X}_1 = 29.8$ $\bar{X}_2 = 34.7$ unknown → t. dis
$\sigma_1 = 4.0$ $\sigma_2 = 5.0$

$H_0 = \mu_1 = \mu_2$ $\alpha = 0.01$
$H_1 : \mu_1 \neq \mu_2$

test statistics $Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \dfrac{29.8 - 34.7}{\sqrt{\frac{16}{20} + \frac{25}{25}}}$

$= \dfrac{-4.9}{\sqrt{0.04 + 0.04}} = -3.6522$

independant small samples

assumption : $\sigma_1 = \sigma_2 = \sigma$ (var are equal but unknown)

$T = \dfrac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$ $S_p \rightarrow$ pooled variance

eg: $\bar{X}_1 = 85$ $S_1 = 4$ $\bar{X}_2 = 81$ $S_2 = 5$ $\alpha = 0.05$
$n_1 = 12$ $n_2 = 10$

$H_0 : \mu_1 - \mu_2 \leq 2$ $d_0 = \mu_1 - \mu_2 = 2$
$H_1 : \mu_1 - \mu_2 > 2$

test statistics $t = \dfrac{(\bar{X}_1 - \bar{X}_2) - d_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ with $df = n_1 + n_2 - 2$
$= 20$

$S_p^2 = \dfrac{11(16) + 9(25)}{20} = \sqrt{20.05} = 4.478$

$t = \dfrac{(85 - 81) - 2}{4.478 \sqrt{\frac{1}{12} + \frac{1}{10}}} = \dfrac{2}{4.478 \sqrt{0.183}} = 1.044$

Critical value approach:

at LOS = 5%, one tailed, df = 20, $t_\alpha = 1.72$

$|t| < t_\alpha$ ∴ donot reject H_0

Matched pair comparisons

eg: x_i = before

y_i = after

$d_i = x_i - y_i$, $\bar{d}$ = mean(d_i)

$$T = \frac{\bar{d} - 0}{s_d/\sqrt{n}} \qquad df = n-1$$

eg: d_i : 9 13 2 5 -2 6 6 5 2 6

$\bar{d} = 5.7$ $\quad s_d = 4.08$ $\quad H_0: \mu_d = 0$

$\alpha = 0.05$ $\quad H_1: \mu_d < 0$

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{5.2}{4.08/\sqrt{10}} = 4.0303$$

$d_f = 9$ $\quad \alpha = 0.05$ $\quad t_\alpha = 1.83$

$t > t_\alpha$