1 - statistical inference

9.1. Estimate the unknown parameter θ from a sample

$$3, 3, 3, 3, 3, 7, 7, 7$$

drawn from a discrete distribution with the probability mass function

$$\begin{cases} P(3) &= \theta \\ P(7) &= 1-\theta \end{cases}.$$

discrete

methods of moment

maximum likelihood

Compute two estimators of θ: $m_1 = \text{mean} = \frac{3.5 + 7.3}{8} = 4.5$ $E(x) = \sum x P(x) = 3\theta + 7(1-\theta) = 7 - 4\theta$

(a) the method of moments estimator; $7 - 4\theta = 4.5$ $\hat{\theta} = \boxed{0.625}$

(b) the maximum likelihood estimator. $L = \theta.\theta.\theta.\theta.\theta\,(1-\theta)(1-\theta)(1-\theta)$ product of all prob. $L = \theta^5 (1-\theta)^3$

ln of both sides $\to$ $\ln L = 5\ln\theta + 3\ln(1-\theta)$

derivatives $\to$ $\frac{L'}{L} = \frac{5}{\theta} - \frac{3}{1-\theta} = 0$ $5 - 5\theta - 3\theta = 0$ $\hat{\theta} = \boxed{\frac{5}{8}}$

min or max? second deriv $-\frac{5}{\theta^2} - \frac{3}{(1-\theta)^2} < 0$ max if max

Also,

(c) Estimate the standard error of each estimator of θ.

(c) $7 - 4\theta = \bar{x}$, $\theta = \frac{\bar{x}}{4}$

standard error

$\sigma(\theta) = \frac{\sigma(\bar{x})}{4} = \frac{\sigma(x)}{4\sqrt{n}} = \frac{\sqrt{E(x^2) - E(x)^2}}{4\sqrt{n}} = \frac{\sqrt{24 - (4.5)^2}}{4\sqrt{8}} = \boxed{0.17116}$

$E(x^2) = 3^2\frac{5}{8} + 7^2\frac{3}{8} = \frac{45 + 147}{8} = 24$

9.2. The number of times a computer code is executed until it runs without errors has a Geometric distribution with unknown parameter p. For 5 independent computer projects, a student records the following numbers of runs:

$$3 \quad 7 \quad 5 \quad 3 \quad 2$$

Estimate p

(a) by the method of moments;

(b) by the method of maximum likelihood.

9.3. Use method of moments and method of maximum likelihood to estimate

(a) parameters a and b if a sample from Uniform(a, b) distribution is observed;

(b) parameter λ if a sample from Exponential(λ) distribution is observed;

(c) parameter μ if a sample from Normal(μ, σ) distribution is observed, and we already know σ;

(d) parameter σ if a sample from Normal(μ, σ) distribution is observed, and we already know μ;

(e) parameters μ and σ if a sample from Normal(μ, σ) distribution is observed, and both μ and σ are unknown.

9.4. A sample of 3 observations ($X_1 = 0.4,\ X_2 = 0.7,\ X_3 = 0.9$) is collected from a continuous distribution with density

$m_1 = \frac{0.4 + 0.7 + 0.9}{3} = \frac{2}{3}$

$$f(x) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$\mu_1 = \int_0^1 x\,\theta x^{\theta-1}$ to calculate expected value of cont. dis $= \theta \int_0^1 x^{\theta} dx = \theta \frac{x^{\theta+1}}{\theta+1}\Big|_0^1 = \frac{\theta}{\theta+1} = \frac{2}{3}$ method of moment $\boxed{\theta = 2}$

continuous dis.
methods of moment
max likelihood

Estimate θ by your favorite method.

max likelihood $L = \theta(0.4)^{\theta-1} \cdot \theta(0.7)^{\theta-1} \cdot \theta(0.9)^{\theta-1}$ $L = \theta^3 (0.252)^{\theta-1}$ $\ln L = 3\ln\theta + (\theta-1)\ln(0.252)$ $\frac{L'}{L} = \frac{3}{\theta} + \ln(0.252) = 0$ $\boxed{\theta = 2.17}$

9.5. A sample $(X_1, \ldots, X_{10})$ is drawn from a distribution with a probability density function

$$\frac{1}{2}\left(\frac{1}{\theta}e^{-x/\theta} + \frac{1}{10}e^{-x/10}\right), \quad 0 < x < \infty$$

The sum of all 10 observations equals 150. mean = 15

(a) Estimate θ by the method of moments. $\int_0^\infty x f(x) = \frac{1}{2}\frac{1}{\theta}\int_0^\infty x e^{-\frac{x}{\theta}} + \frac{1}{10}\int_0^\infty e^{-\frac{x}{10}} x$ $\frac{\theta}{2} + 5 = 15$ $\boxed{\theta = 20}$

(b) Estimate the standard error of your estimator in (a).

$\sigma(\hat{\theta})$ $\frac{\hat{\theta}}{2} + 5 = \bar{x}$ $\hat{\theta} = 2\bar{x} - 10$ $\sigma(\hat{\theta}) = 2\sigma(\bar{x})$ sample $\sigma(\bar{x}) = \frac{\sigma(x)}{\sqrt{n}}$ popul. $E(x^2) = \int_0^\infty x^2 f(x)dx = \theta^2 + 100$ $Var(x) = E(x^2) - E^2(x) = \theta^2 + 100 - (\frac{\theta}{2} + 5)^2 = 0.75\theta^2 - 5\theta + 75$ $\hat{\theta} = 20$ $n = 10$ $\sigma(\hat{\theta}) = 2\frac{\sigma(x)}{\sqrt{n}} = 2\sqrt{\frac{0.75(20)^2 - 5\cdot 20 + 75}{10}} = \boxed{5.24}$

9.6. Verify columns 3-5 in Table 9.1 on p. 273. Section 9.4.7 will help you with the last row of the table.

9.7. In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the average number of concurrent users at 100 randomly selected times is 37.7, with a standard deviation $\sigma = 9.2$.

(a) Construct a 90% confidence interval for the expectation of the number of concurrent users. $[36.1867,\ 39.2133]$ one_mean_conf(37.7, 100, 9.2, 1.64485); $z_{\alpha/2} = z_{0.05}$

(b) At the 1% significance level, do these data provide significant evidence that the mean number of concurrent users is greater than 35? $H_0: \mu = 35$ $H_1: \mu > 35$ test value $= +2.93478$ yes, $z_\alpha = z_{0.01} = 2.326$ one_mean(37.7, 35, 9.2, 100); reject H_0 H_1 yes it is greater

$H_0: \mu = a$ $H_a: \mu > a$ reject H_0 $z = t$ right tailed

$H_0: \mu = a$ $H_a: \mu < a$ reject H_0 $z = t$ left tailed

9.8. Installation of a certain hardware takes random time with a standard deviation of 5 minutes.

(a) A computer technician installs this hardware on 64 different computers, with the average installation time of 42 minutes. Compute a 95% confidence interval for the population mean installation time. $[40.775,\ 43.225]$

(b) Suppose that the population mean installation time is 40 minutes. A technician installs the hardware on your PC. What is the probability that the installation time will be within the interval computed in (a)?

9.9. Salaries of entry-level computer engineers have Normal distribution with unknown mean and variance. Three randomly selected computer engineers have salaries (in \$ 1000s):

$$30, 50, 70$$

$\bar{x} = 50$ $e = \frac{20}{\sqrt{3}}$ sample!! $t(\frac{\alpha}{2}, 2)$ two tailed $\to 2.9199$

unknown pop var with $n < 30$

(a) Construct a 90% confidence interval for the average salary of an entry-level computer engineer. $[16.2829,\ 83.1717]$

(b) Does this sample provide a significant evidence, at a 10% level of significance, that the average salary of all entry-level computer engineers is different from \$80,000? Explain. $H_0: \mu = 80000$ $H_a: \mu \neq 80000$ same level no, 80000 is in interval

(c) Looking at this sample, one may think that the starting salaries have a great deal of variability. Construct a 90% confidence interval for the standard deviation of entry-level salaries.

std dev $\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$ $n = 3$ $\alpha = 0.1$ $\sigma(\bar{x}) = 20$ p: 0.05, df: 2 $\chi^2 = 5.991$; p: 0.95, df: 2 $\chi^2 = 0.1$ $\sqrt{\frac{2\cdot 400}{5.991}} < \sigma < \sqrt{\frac{2\cdot 400}{0.1}} \to [11.55,\ 89.4]$

9.10. We have to accept or reject a large shipment of items. For quality control purposes, we collect a sample of 200 items and find 24 defective items in it. $z_{\alpha/2} = 2.053$

(a) Construct a 96% confidence interval for the proportion of defective items in the whole shipment. $[0.0728,\ 0.1671]$

(b) The manufacturer claims that at most one in 10 items in the shipment is defective. At the 4% level of significance, do we have sufficient evidence to disprove this claim? $\to$ no, 0.1 is in interval
Do we have it at the 15% level? $H_0: p \neq 0.1$ $H_a: p \not< 0.1$ $t = 0.842$ α $0.15 \to 1.43953 \to [0.869,\ 0.153] \to$ no 0.1 is in interval

9.11. Refer to Exercise 9.10. Having looked at the collected sample, we consider an alternative supplier. A sample of 150 items produced by the new supplier contains 13 defective items.
Is there significant evidence that the quality of items produced by the new supplier is higher than the quality of items in Exercise 9.10? What is the P-value? $n_1: 200$ $x_1 = 24$ $n_2 = 150$ $x_2 = 13$

$H_0: p_1 = p_2$ $H_a: p_1 < p_2$ $t = 1.0249$ pooled-$t = 1.00769$ $\alpha_{0.02} = 2.053$ left tailed inverse (I) $p = 0.1586 \to$ no high, not enough evidence

9.12. An electronic parts factory produces resistors. Statistical analysis of the output suggests that resistances follow an approximately Normal distribution with a standard deviation of 0.2 ohms. A sample of 52 resistors has the average resistance of 0.62 ohms.

(a) Based on these data, construct a 95% confidence interval for the population mean resistance.

(b) If the actual population mean resistance is exactly 0.6 ohms, what is the probability that an average of 52 resistances is 0.62 ohms or higher?

9.13. Compute a P-value for the right-tail test in Example 9.25 on p. 272 and state your conclusion about a significant increase in the number of concurrent users.

9.14. Is there significant difference in speed between the two servers in Example 9.21 on p. 263?

(a) Use the confidence interval in Example 9.21 to conduct a two-sided test at the 5% level of significance.

(b) Compute a P-value of the two-sided test in (a).

(c) Is server A really faster? How strong is the evidence? Formulate the suitable hypothesis and alternative and compute the corresponding P-value.

State your conclusions in (a), (b), and (c).

9.15. According to Example 9.17 on p. 257, there is no significant difference, at the 5% level, between towns A and B in their support for the candidate. However, the level $\alpha = 0.05$ was chosen rather arbitrarily, and the candidate still does not know if he can trust the results when planning his campaign. Can we compare the two towns at *all* reasonable levels of significance? Compute the P-value of this test and state conclusions.

9.16. A sample of 250 items from lot A contains 10 defective items, and a sample of 300 items from lot B is found to contain 18 defective items.

$z_{\alpha/2} = z_{0.01} = 2.3263$ $[-0.0629966,\ 0.0229966]$

(a) Construct a 98% confidence interval for the difference of proportions of defective items.

(b) At a significance level $\alpha = 0.02$, is there a significant difference between the quality of the two lots? no their level are same + two tailed if $p_1 = p_2$ $p_1 - p_2 = 0$ is in interval they can be same

9.17. A news agency publishes results of a recent poll. It reports that candidate A leads candidate B by 10% because 45% of the poll participants supported Ms. A whereas only 35% supported Mr. B. What margin of error should be reported for each of the listed estimates, 10%, 35%, and 45%? Notice that 900 people participated in the poll, and the reported margins of error typically correspond to 95% confidence intervals.

A: 0.45 B: 0.35 $n = 900$ $z_{\alpha/2} = 1.96$

0.45 one_mean_prop(405, 900, 1.96) $800(0.45)$ $0.0325 \to \%3.25$

0.35 one_mean_prop(315, 900, 1.96) $900(0.35)$ $0.031162 \to \%3.116$

two_prop_conf(900, 900, 405, 315, 1.96) $p_1 - p_2 = 0.15$ $0.045 \to \%4.5$

9.18. Consider the data about the number of blocked intrusions in Exercise 8.1, p. 233.

(a) Construct a 95% confidence interval for the difference between the average number of intrusion attempts per day before and after the change of firewall settings (assume equal variances).

(b) Can we claim a significant reduction in the rate of intrusion attempts? The number of intrusion attempts each day has approximately Normal distribution. Compute P-values and state your conclusions under the assumption of equal variances and without it. Does this assumption make a difference?

before: $n = 14$ $s: 7.616$ $\bar{x} = 50$

after: $n = 20$ $\bar{x} = 40.2$ $s = 7.957$

eq variances $\to$ $s_p = 7.82$ use cat $t(v, \frac{\alpha}{2}) = 2.0369$ $20 + 14 - 2 = 32$ $[4.2489,\ 15.35]$ two_mean_confident(50, 40.2, 7.82, 7.82, 14, 20, 2.0369)

equal var $\to$ $t = 3.596$ $v = 32$

unequal var $\to$ $t = 3.625$ $v = 30$

$H_0: x_1 \leq x_2$ $H_1: x_1 > x_2$

inverse t $t = 3.596$ $p = 0.0005$ $t = 3.625$ $p = 0.005$

$\leftarrow$ $P < 0.01 \to$ reject H_0

yes it reduced

Null hypothesis H_0	Parameter, estimator θ, $\hat{\theta}$	If H_0 is true: $\mathbf{E}(\hat{\theta})$	If H_0 is true: $\mathrm{Var}(\hat{\theta})$	Test statistic $Z = \dfrac{\hat{\theta} - \theta_0}{\sqrt{\mathrm{Var}(\hat{\theta})}}$
One-sample Z-tests for means and proportions, based on a sample of size n				
$\mu = \mu_0$	μ, $\bar{X}$	μ_0	$\dfrac{\sigma^2}{n}$	$\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
$p = p_0$	p, $\hat{p}$	p_0	$\dfrac{p_0(1-p_0)}{n}$	$\dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
Two-sample Z-tests comparing means and proportions of two populations, based on independent samples of size n and m				
$\mu_X - \mu_Y = D$	$\mu_X - \mu_Y$, $\bar{X} - \bar{Y}$	D	$\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}$	$\dfrac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$
$p_1 - p_2 = D$	$p_1 - p_2$, $\hat{p}_1 - \hat{p}_2$	D	$\dfrac{p_1(1-p_1)}{n} + \dfrac{p_2(1-p_2)}{m}$	$\dfrac{\hat{p}_1 - \hat{p}_2 - D}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}$
$p_1 = p_2$	$p_1 - p_2$, $\hat{p}_1 - \hat{p}_2$	0	$p(1-p)\left(\dfrac{1}{n} + \dfrac{1}{m}\right)$, where $p = p_1 = p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}$ where $\hat{p} = \dfrac{n\hat{p}_1 + m\hat{p}_2}{n+m}$

Hypothesis H_0	Conditions	Test statistic t	Degrees of freedom
$\mu = \mu_0$	Sample size n; unknown σ	$t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$	$n - 1$
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown but equal standard deviations, $\sigma_X = \sigma_Y$	$t = \dfrac{\bar{X} - \bar{Y} - D}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$	$n + m - 2$
$\mu_X - \mu_Y = D$	Sample sizes n, m; unknown, unequal standard deviations, $\sigma_X \neq \sigma_Y$	$t = \dfrac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$	Satterthwaite approximation, formula (9.12)

TABLE 9.2: *Summary of T-tests.*

9.1 parameter estimation

methods of moment = calculate mean from given data $= \sum x \cdot P(x)$ (discrete)

$= \int x f(x)$ (continuous)

max likelihood = L = product of all given values probabilities

→ write both side as $\ln$

→ take derivative $= 0$

→ find θ : check if it is max point by checking second derivative L is negative

standard error = find a equation like $\theta = a\bar{x} + b$ → does not matter

$\bar{x}$ → mean

$\sigma(\theta)$ = is error

to find it $\sigma(\theta) = a \cdot \sigma(\bar{x})$ (sample)

$= a \cdot \frac{\sigma(x)}{\sqrt{n}}$ (population)

$E(x) = \text{mean} = \sum x P(x) = \int x f(x)$

$E(x^2) = \sum x^2 P(x) = \int x^2 f(x)$

$= a \cdot \sqrt{\frac{E(x^2) - E(x)^2}{n}}$

CONFIDENCE INTERVALS

$\boxed{\text{one population}}$

1- mean (μ)

2- proportion (p)

3- variance (σ^2)

$\boxed{\text{1-mean}}$

ⓐ known variance

$\bar{x} - e < \mu < \bar{x} + e \qquad e = \left(z_{\frac{\alpha}{2}}\right) \cdot \frac{\sigma}{\sqrt{n}}$

ⓑ unknown variance with $n \geq 30$

$\bar{x} - e < \mu < \bar{x} + e \qquad e = \left(z_{\frac{\alpha}{2}}\right) \cdot \frac{s}{\sqrt{n}}$ $\qquad s$ = std dev of sample

ⓒ unknow variance with $n < 30$

$\bar{x} - e < \mu < \bar{x} + e \qquad e = t_{(\frac{\alpha}{2}, n-1)} \cdot \frac{s}{\sqrt{n}}$ $\qquad n-1$ = degrees of freedom

$\boxed{\text{2- proportion}}$

$$\hat{p} - e < p < \hat{p} + e \qquad e = \left(z_{\frac{\alpha}{2}}\right) \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\boxed{\text{3- variance}}$

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1).s^2}{\chi^2_{(1-\frac{\alpha}{2})}}$$

two populations

1- mean $(\mu_1 - \mu_2)$
2- proportion $(p_1 - p_2)$
3- variance

1- mean

ⓐ known populations variances

$$(\bar{x}_1 - \bar{x}_2) - e < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + e \qquad e = \left(z_{\frac{\alpha}{2}}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

ⓑ unknow populations variances with $\sigma_{samp} = \sigma_{pop}$

Soruda assuming equal population ver der

$$(\bar{x}_1 - \bar{x}_2) - e < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + e$$

$$e = t\left(\frac{\alpha}{2}, v\right) \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$v = n_1 + n_2 - 2 \qquad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

ⓒ unknow populations variances with $\sigma_{samp} \neq \sigma_{pop}$

we know populations

$$(\bar{x}_1 - \bar{x}_2) - e < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + e$$

$$e = t\left(\frac{\alpha}{2}, v\right) \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

closest int $\simeq v = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$ → Satterthwaite formula

$$\min(n_1, n_2) < v < (n_1 + n_2)$$

2-proportion

$$(\hat{p}_1 - \hat{p}_2) - e < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + e \qquad e = z_{\left(\frac{\alpha}{2}\right)} \cdot \sqrt{\frac{\hat{p}_1 \cdot (1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

HYPOTHESIS TESTING

One population

1- mean (μ)
2- proportion (p)
3- variance (σ^2)

1-mean

ⓐ known (pop) variance

test statistic = $Z = \dfrac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$

claimed value in the hypotesis

σ

• if $\sigma(x)$ given

$\sigma(\bar{x}) = \dfrac{\sigma(x)}{\sqrt{n}}$ in no form

• if $\sigma(\bar{x})$ is given

$z = \dfrac{\bar{x} - \mu_0}{\sigma(\bar{x})}$

critical value = z_α (one sided)

two sided: $\frac{\alpha}{2}$

ⓑ unknown (pop) variance with $n \geq 30$

test statistic = $Z = \dfrac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$

in hypothesis

sample std dev

critical value = z_α (one sided)

two sided: $\frac{\alpha}{2}$

ⓒ unknown (pop) variance with $n < 30$

test statistic = $t = \dfrac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$

in hypothesis

sample std dev

from calculator (two sided) or (one sided)

critical value = $t\,(\alpha, n-1)$

degree of freedom

2-proportion

test statistic = $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

in hypothesis

critical value = z_α (one sided)
($\frac{\alpha}{2} \rightarrow$ two sided)

two populations

1- mean ($\mu_1 - \mu_2$)
2- proportion ($p_1 - p_2$)
3- variance

1-mean

ⓐ known (pop) variance

test statistic = $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

critical value = z_α (one sided)
($\frac{\alpha}{2} \rightarrow$ two sided)

ⓑ unknown population variance with $var_1 = var_2$ $(s_1 = s_2)$

$$\text{test statistic} = t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

just hint

critical value $= t(\alpha, n+m-2)$

one / two tailed

degrees of freedom

s_p → standard dev of population $= \sigma_1 = \sigma_2$

paranteze alınmış hali

$$s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

ⓒ unknown population variance with $var_1 \neq var_2$

$$\text{test statistic} = t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

critical value $= t(\alpha, v)$

satterthwaite approx

$$v = \frac{\frac{s_x^2}{n} + \frac{s_y^2}{m}}{\frac{s_x^4}{n^2(n-1)} + \frac{s_y^4}{m^2(m-1)}}$$

2-proportion

$$\text{test statistic} = Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

critical value $= z_\alpha$

(pooled better approx)

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} \qquad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

p value

Testing H_0 with a P-value

For	$\alpha < P$,	accept H_0
For	$\alpha > P$,	reject H_0
Practically,		
If	$P < 0.01$,	reject H_0
If	$P > 0.1$,	accept H_0

this table

check if $p < 0.08$ or $p > 0.1$ ..

Hypothesis H_0	Alternative H_A	P-value	Computation
	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{Z \geq Z_{\text{obs}}\}$	$1 - \Phi(Z_{\text{obs}})$
$\theta = \theta_0$	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{Z \leq Z_{\text{obs}}\}$	$\Phi(Z_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\lvert Z\rvert \geq \lvert Z_{\text{obs}}\rvert\}$	$2(1 - \Phi(\lvert Z_{\text{obs}}\rvert))$

Hypothesis H_0	Alternative H_A	P-value	Computation
	right-tail $\theta > \theta_0$	$\boldsymbol{P}\{t \geq t_{\text{obs}}\}$	$1 - F_\nu(t_{\text{obs}})$
$\theta = \theta_0$	left-tail $\theta < \theta_0$	$\boldsymbol{P}\{t \leq t_{\text{obs}}\}$	$F_\nu(t_{\text{obs}})$
	two-sided $\theta \neq \theta_0$	$\boldsymbol{P}\{\lvert t\rvert \geq \lvert t_{\text{obs}}\rvert\}$	$2(1 - F_\nu(\lvert t_{\text{obs}}\rvert))$

TABLE 9.4: *P-values for T-tests (F_ν is the cdf of T-distribution with the suitable number ν of degrees of freedom).*

```
#include <iostream>
#include <math.h>

double df_satt_formula(double s1,double s2,double n1, double n2);//std1
std2 n1 n2 -- no mean values
//one population mean confidence interval
void one_mean_conf(double mean, double n, double std_dev, double
zt_val);
void one_prop_conf(double portion_in_n, double total_n, double z_val);
//two populations mean confidence interval
void two_mean_conf(double mean1, double mean2, double std_dev1,
double std_dev2, double n1, double n2, double zt_val);//known pop var +
is also for. unknow unqual variances
void two_mean_conf_unk_eq(double mean1, double mean2, double
std_dev1, double std_dev2, double n1, double n2, double t_val);//unknown
pop var, equal
void two_prop_conf(double total_n1, double total_n2, double
portion_in_n1, double portion_in_n2, double z_val);
//one population
double one_mean(double mean, double claimed_mean, double std_dev,
```

```
double n);
double one_prop(double claimed_p , double portion_in_n, double
total_n);
//two populations
double two_mean(double mean1, double mean2, double std_1, double
std_2, double n1, double n2);
double two_prop(double total_n1, double total_n2, double portion_in_n1,
double portion_in_n2);
//pooled is used mostly
double two_prop_pooled(double total_n1, double total_n2, double
portion_in_n1, double portion_in_n2);

double sp(double s1,double s2,double n1, double n2){
    return sqrt(((n1-1)*s1*s1+(n2-1)*s2*s2)/(n1+n2-2));
}




int main()
{
   one_prop_conf(55, 154, 1.43953);
    std::cout<<"---\n";
   // std::cout << two_mean(6.2, 5.8, 1.5, 1.1, 55, 55);
    std::cout<<"\n---\n";
   // two_prop_conf(400,400, 192, 148, 1.95996);

}


double df_satt_formula(double s1,double s2,double n1, double n2){
    return (s1*s1/n1 + s2*s2/n2)*(s1*s1/n1 + s2*s2/n2)/ (s1*s1*s1*s1/
(n1*n1*(n1-1)) + s2*s2*s2*s2/(n2*n2*(n2-1)));
}

void one_mean_conf(double mean, double n, double std_dev, double
zt_val){
    double e = zt_val * std_dev/sqrt(n);
    std::cout<<"mean: "<<mean<<"\nmargin:"<<e<<"\ninterval: [ "<<mean-
e<<" , "<<mean+e<<" ]\n";
}


void one_prop_conf(double portion_in_n, double total_n, double z_val){
```

```
    double p = portion_in_n/total_n;
    double e = z_val * sqrt(p*(1-p)/total_n);
    std::cout<<"propotion: "<<p<<"\nmargin:"<<e<<"\ninterval: [ "<<p-
e<<" , "<<p+e<<" ]\n";
}


void two_mean_conf(double mean1, double mean2, double std_dev1,
double std_dev2, double n1, double n2, double zt_val){
    double e = zt_val * sqrt(std_dev1*std_dev1/n1 + std_dev2*std_dev2/
n2);
    double dif = mean1-mean2;
    std::cout<<"mean differences: "<<dif<<"\nmargin:"<<e<<"\ninterval:
[ "<<dif-e<<" , "<<dif+e<<" ]\n";
}



void two_mean_conf_unk_eq(double mean1, double mean2, double
std_dev1, double std_dev2, double n1, double n2, double t_val){
    double sp = sqrt( ( (n1-1)*std_dev1*std_dev1 +
(n2-1)*std_dev2*std_dev2 ) / (n1+n2-2));
    double e = t_val * sp* sqrt(1/n1 + 1/n2);
    double dif = mean1-mean2;
    std::cout<<"mean differences: "<<dif<<"\nmargin:"<<e<<"\ninterval:
[ "<<dif-e<<" , "<<dif+e<<" ]\n";
}

void two_prop_conf(double total_n1, double total_n2, double
portion_in_n1, double portion_in_n2, double z_val){
    double p_1=portion_in_n1/total_n1;
    double p_2=portion_in_n2/total_n2;
    double dif = p_1-p_2;
    double e = z_val * sqrt(p_1*(1-p_1)/total_n1 + p_2*(1-p_2)/total_n2);
    std::cout<<"proportion differences:
"<<dif<<"\nmargin:"<<e<<"\ninterval: [ "<<dif-e<<" , "<<dif+e<<" ]\n";
}


//hypothesis
double one_prop(double new_p , double portion_in_n, double total_n){
    double old_p=portion_in_n/total_n;
    return (old_p-new_p)*sqrt(total_n / (new_p*(1-new_p)));
}
```

```
double one_mean(double mean, double claimed_mean, double std_dev,
double n){
    return (mean-claimed_mean)*sqrt(n)/std_dev;
}

double two_mean(double mean1, double mean2, double std_1, double
std_2, double n1, double n2){
    return (mean1-mean2)/sqrt((std_1*std_1)/n1 +(std_2*std_2)/n2);
}

double two_prop(double total_n1, double total_n2, double portion_in_n1,
double portion_in_n2){
    double p_1=portion_in_n1/total_n1;
    double p_2=portion_in_n2/total_n2;
    return (p_1-p_2)/sqrt(p_1*(1-p_1)/total_n1 + p_2*(1-p_2)/total_n2);
}

double two_prop_pooled(double total_n1, double total_n2, double
portion_in_n1, double portion_in_n2){
    double p_1=portion_in_n1/total_n1;
    double p_2=portion_in_n2/total_n2;
    double p=(portion_in_n1+portion_in_n2)/(total_n1 + total_n2);
    return (p_1-p_2)/sqrt(p*(1-p)*(1/total_n1 +1/total_n2));
}
```