Interval Estimation

Example : Election 2000 voting poll from 2386 interviews

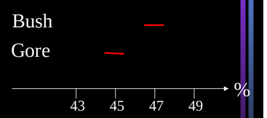
Result of Nationwide Gallup poll November 5, 2000	
George W. Bush	47%
Al Gore	45%
Ralph Nader	4%
Buchanan	1%
Not yet decided	3%

Margin of error : 0.05%

47% > 45%

Can Bush get more votes than Gore?

Interval estimates



"Bush will win the race"

Interval Estimation

Example : VSRS $\{X_1, X_2, \dots, X_n\}$ from $N(\mu, \sigma^2)$

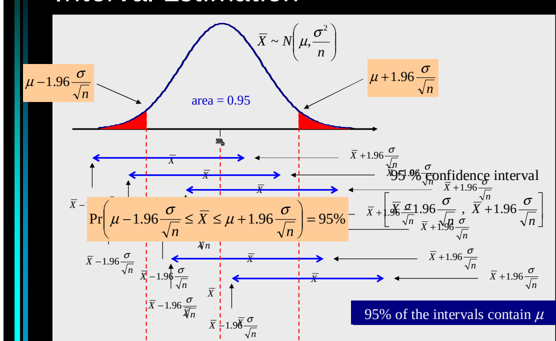
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\Pr\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = \Phi\left(\frac{\mu + 1.96\sigma/\sqrt{n} - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\mu - 1.96\sigma/\sqrt{n} - \mu}{\sigma/\sqrt{n}}\right)$$

$$= 0.975 - (1 - 0.975)$$

$$= 0.95$$

Interval Estimation



Confidence Interval

L, U ——— Statistics calculated from the sample

The interval [L, U] is called a 100(1-α)% confidence interval (C.I.) for the parameter θ if

$$\Pr(L \le \theta \le U) = 1 - \alpha \quad \text{for any possible } \theta$$

Confidence level

α	Confidence level
0.01	99%
0.05	95%
0.1	90%

Confidence Interval

Normal population $N(\mu, \sigma^2)$ μ unknown σ^2 known

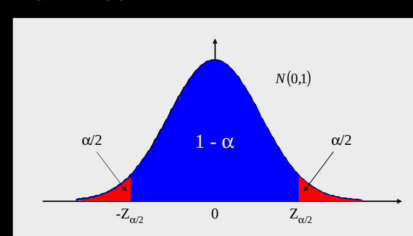
VSRS $\{X_1, X_2, \dots, X_n\}$

$$\Pr\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \; 99 \; 0.95$$

100(1-α)% C.I. for μ : $\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$

point estimate margin of error

Tail Area



$Z_{0.05} = 1.645$

$Z_{0.025} = 1.96$

$Z_{0.15} = 1.036$

$Z_{0.9} = -1.282$

α	0.400	0.300	0.200	0.100	0.050	0.025	0.010	0.005	0.001
z_α	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090
$z_{\alpha/2}$	0.842	1.036	1.282	1.645	1.960	2.240	2.576	2.807	3.291

Confidence Interval

Example : X — alcohol concentration of French wine

$X \sim N(\mu, 1.2^2)$

VSRS $\{X_1, X_2, ..., X_{60}\}$ ⟶ $\overline{X} = 9.3$

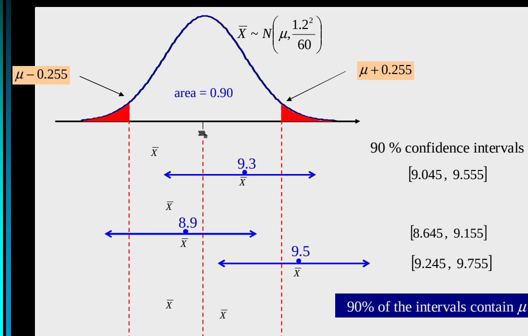
A 90% C.I. for μ : $\overline{X} \pm Z_{0.05}\frac{\sigma}{\sqrt{n}} = 9.3 \pm (1.645)\frac{1.2}{\sqrt{60}}$

$= 9.3 \pm 0.255$

$= [9.045, 9.555]$

? ~~$\Pr(9.045 \le \mu \le 9.555) = 0.9$~~ ?

Confidence Interval



Normal Model, Unknown Variance

Normal population $N(\mu, \sigma^2)$ μ unknown σ^2 unknown

VSRS $\{X_1, X_2, ..., X_n\}$

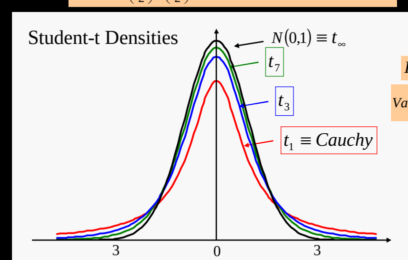
$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$ $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

$$\Pr\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

100(1-α)% C.I. for μ : $\overline{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$

Student-t Distribution

$$f(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{1}{2}\right)} r^{-\frac{1}{2}}\left(1 + \frac{x^2}{r}\right)^{-\frac{r+1}{2}}, \quad -\infty < x < \infty$$

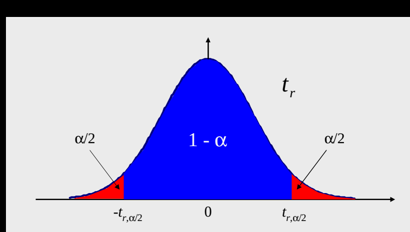


$X \sim t_r$

$E(X) = 0$ for $r > 1$

$Var(X) = \frac{r}{r-2}$ for $r > 2$

Student-t Distribution



Student-t Distribution Table

	$P(T \le t)$						
	0.60	0.75	0.90	0.95	0.975	0.99	0.995
r	$t_{0.40}(r)$	$t_{0.25}(r)$	$t_{0.10}(r)$	$t_{0.05}(r)$	$t_{0.025}(r)$	$t_{0.01}(r)$	$t_{0.005}(r)$
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012
14	0.258	0.692	1.345	1.761	2.145	2.624	2.997
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947

$X \sim t_r$

$t_{8,0.05} = 1.860$ $t_{11,0.025} = 2.201$

Normal Model, Unknown Variance

Example : X — Frequency of elephant call

$X \sim N(\mu, \sigma^2)$

VSRS $\{X_1, X_2, ..., X_{12}\}$ ⟶ $\bar{X} = 22.33$ $S^2 = 56.424$

A 95% C.I. for μ : $\bar{X} \pm t_{11,0.025} \frac{S}{\sqrt{n}} = 22.33 \pm (2.201)\sqrt{\frac{56.424}{12}}$

$= 22.33 \pm 4.773$

$= [17.557, 27.103]$

~~$\Pr(17.557 \le \mu \le 27.103) = 0.95$~~

Two Samples Problem

Population 1: μ_x, σ_x^2

Population 2: μ_y, σ_y^2

VSRS

$\{X_1, X_2, ..., X_m\}$ $\{Y_1, Y_2, ..., Y_n\}$

estimate

Compare μ_x and μ_y by : $\mu_x - \mu_y$ ⟵ unknown

Normal Models, Known Variances

Normal populations $N(\mu_x, \sigma_x^2)$ $N(\mu_y, \sigma_y^2)$ μ_x, μ_y unknown, $\sigma_x^2 \cdot \sigma_y^2$ known

Independent VSRSs $\{X_1, X_2, ..., X_m\}$ $\{Y_1, Y_2, ..., Y_n\}$

Point estimator for μ_x - μ_y : $\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right)$

$$\Pr\left((\bar{X} - \bar{Y}) - Z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \le \mu \le (\bar{X} - \bar{Y}) + Z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}\right) = 1 - \alpha$$

100(1-α)% C.I. for μ : $(\bar{X} - \bar{Y}) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}$

point estimate — margin of error

Normal Models, Unknown Variances

Normal populations $N(\mu_x, \sigma_x^2)$ $N(\mu_y, \sigma_y^2)$ μ_x, μ_y unknown, $\sigma_x^2 \cdot \sigma_y^2$ unknown

Independent VSRSs $\{X_1, X_2, ..., X_m\}$ $\{Y_1, Y_2, ..., Y_n\}$

Assumption : Equal variances $\sigma_x^2 = \sigma_y^2 = \sigma^2$

Pooled sample variance $S_{pool}^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}$

100(1-α)% C.I. for μ_x - μ_y : $(\bar{X} - \bar{Y}) \pm t_{m+n-2,\alpha/2} S_{pool}\sqrt{\frac{1}{m} + \frac{1}{n}}$

Normal Models, Unknown Variances

Example : growth of tomatoes

Fertilizer A (X)	12, 11, 7, 13, 8, 9, 10, 13
Fertilizer B (Y)	13, 11, 10, 6, 7, 4, 10

Assumptions : (i) Normal populations

(ii) Independent samples (iii) Equal variance

$m = 8,\ n = 7,\ \bar{X} = 10.371,\ \bar{Y} = 8.714,\ S_x^2 = 5.125,\ S_y^2 = 9.905$

Pooled sample variance : $S_{pool}^2 = \frac{(8-1)(5.125) + (7-1)(9.905)}{8+7-2} = 7.331$

95% C.I. for μ_x - μ_y : $1.661 \pm 3.027 = [-1.366, 4.688]$ $\sqrt{\frac{1}{8} + \frac{1}{7}}$

Large Sample Approximation

Arbitrary Population: μ, σ^2

VSRS $\{X_1, X_2, ..., X_n\}$

For large n

$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot\sim N(0,1)$ ✓

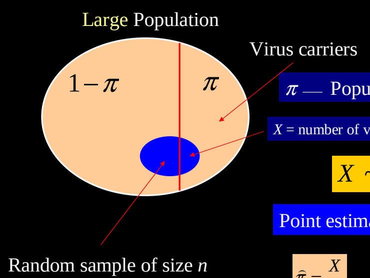
$\Pr\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$

Approximate 100(1-α)% C.I. for μ :

$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

IF σ is unknown, replace it by S .

Population Proportion



π — Population proportion

X = number of virus carriers in the sample

$X \sim b(n, \pi)$

Point estimator for π :

$\hat{\pi} = \frac{X}{n}$ — Sample proportion

Population Proportion

For large n $\quad \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0,1)$

$$\Pr\left(-Z_{\alpha/2} \le \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \le Z_{\alpha/2} \right) \approx 1-\alpha$$

Approximate 100(1-α)% C.I. for π :

$$\frac{2\hat{\pi} + Z_{\alpha/2}^2 \pm \sqrt{\left(2\hat{\pi} + Z_{\alpha/2}^2/n\right)^2 - 4\hat{\pi}^2\left(1 + Z_{\alpha/2}^2/n\right)}}{2\left(1 + Z_{\alpha/2}^2/n\right)}$$

$$\approx \hat{\pi} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Population Proportion

Example : 41 people out of 500 persons were unemployed

$$\hat{\pi} = \frac{41}{500} = 0.082$$

Approximate 95% C.I. for π :

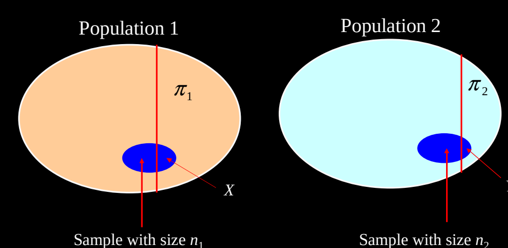
$$\hat{\pi} \pm Z_{0.025}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.082 \pm (1.96)\sqrt{\frac{(0.082)(0.918)}{500}}$$

$$= 0.082 \pm 0.024$$

$$= [5.8\% \ , \ 10.6\%]$$

Estimated unemployment rate is 8.2% with margin of error 2.4% .

Two Samples Problem



Compare π_1 and π_2 by : $\pi_1 - \pi_2$ ← unknown

Two Samples Problem

Point estimators $\quad \hat{\pi}_1 = \frac{X}{n_1} \quad \hat{\pi}_2 = \frac{Y}{n_2}$

For large n_1, n_2 $\quad \frac{X - n_1\pi_1}{\sqrt{n_1\pi_1(1-\pi_1)}} \sim N(0,1) \quad \frac{Y - n_2\pi_2}{\sqrt{n_2\pi_2(1-\pi_2)}} \sim N(0,1)$

Approximate 100(1-α)% C.I. for π_1 - π_2 :

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Two Samples Problem

Example : ability of detergents to remove stains

Detergent	Successes	Failures	Total
1	63	28	91
2	42	37	79

$\hat{\pi}_1 = \frac{63}{91} = \frac{9}{13} = 69.23\%$

$\hat{\pi}_2 = \frac{42}{79} = 53.16\%$

Approximate 90% C.I. for π_1 - π_2 :

$\pi \, \pi_1 > \pi_2 \, 0$

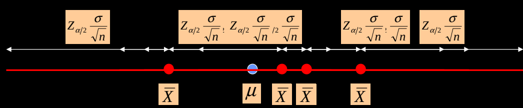
$$(\hat{\pi}_1 - \hat{\pi}_2) \pm Z_{0.05}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

$$= \left(\frac{9}{13} - \frac{42}{79}\right) \pm (1.645)\sqrt{\frac{(9/13)(4/13)}{91} + \frac{(42/79)(37/79)}{79}}$$

$$= 0.1607 \pm 0.1219 = [3.88\% \ , \ 28.26\%]$$

Sample Size Determination

100(1-α)% C.I. for μ : $\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$



D : precision requirement

$$D = \frac{Z_{\alpha/2}\sigma}{\sqrt{n}} \Leftrightarrow n = \frac{Z_{\alpha/2}^2 \sigma^2}{D^2}$$

With probability 1-α, estimation error is at most D.

$$D = Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \Leftrightarrow n = \frac{Z_{\alpha/2}^2 \pi(1-\pi)}{D^2}$$

Sample Size Determination

Example : survey on entertainment expenses

$\sigma = \$400$ ——— from past surveys

Precision requirement :

95% confidence that estimation error is at most $120

$$n = \frac{Z_{0.025}^2 \sigma^2}{D^2} = \frac{(1.96)^2 (400)^2}{(120)^2} = 42.68 \approx 43$$