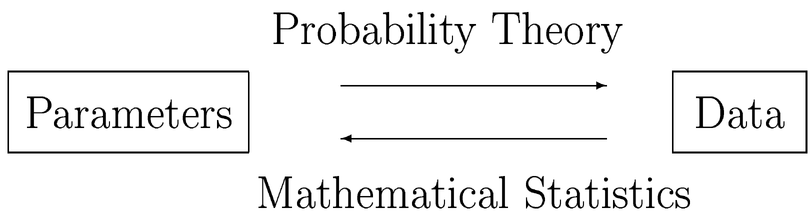
Serik Sagitov, Chalmers Tekniska Högskola, August 16, 2004

4. Parameter estimation



4.1 Random sampling

Def 1: population distribution

Population is a set of elements $\{1, 2, ..., N\}$
labeled by values $\{x_1, x_2, ..., x_N\}$, population size N
PD = population distribution of x-values
find PD: random sampling versus enumeration

> Randomisation in sampling is a guard against investigator's biases even unconscious

Two kinds of sampling errors
systematic (accuracy) and random (precision)

Ex 1: sampling design errors

Systematic errors caused by sampling designs
selection bias: Roosevelt unpredicted victory in 1936
non-response bias: questionnaire vs interview
response bias: potentionally embarrassing information

Ex 2: color preference

histogram: students choice of green/yellow/red T-shirt
PD = (p_1, p_2, p_3)

Def 2: iid sample

$(X_1, \ldots, X_n)$ with observations X_i being
Independent and Identically Distributed

4.2 Population parameters and estimates

Examples of population parameters

population mean μ and standard deviation σ
PD $= (p_1, \ldots, p_r)$, population proportion p_i
PD $= \mathrm{U}(0, \theta)$, interval length θ
PD $= \mathrm{Exp}(\lambda)$, population distribution rate λ

Def 3: point estimate

a function $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ of the data
representing the unknown population parameter θ

Sampling distribution = distribution of $\hat{\theta}$

different samples give different values of $\hat{\theta}$

Point estimate $\hat{\theta}$ is a certain number after sampling but $\hat{\theta}$ is a random variable before sampling

Sample mean and variance

Common estimates of μ and σ^2

sample mean $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$
sample variance $s^2 = \frac{1}{n-1} \Sigma_{i=1}^{n} (X_i - \bar{X})^2$

Approximate sampling distribution $\bar{X} \approx \mathrm{N}(\mu, \frac{\sigma^2}{n})$

$$\boxed{s^2 = \frac{1}{n-1}(X_1^2 + \ldots + X_n^2) - \frac{n}{n-1}\bar{X}^2}$$

Sample proportion

PD = Bernoulli(p)

$X = 1$ with probability p and $X = 0$ with q

sample count $X_1 + \ldots + X_n \sim \text{Bin}(n, p)$

sample proportion $\hat{p} = \bar{X}$

Approximate sampling distribution $\hat{p} \approx \text{N}(p, \frac{pq}{n})$

4.3 Unbiased estimates

Def 4: unbiased estimate

$\hat{\theta}$ is an unbiased estimate of θ, if $\text{E}(\hat{\theta}) = \theta$

no systematic error

Sample mean, variance, and proportion

all three are unbiased estimates $\text{E}(\bar{X}) = \mu$, $\text{E}(\hat{p}) = p$

$\text{E}(s^2) = \sigma^2$ explains the factor $\frac{1}{n-1}$ instead of $\frac{1}{n}$ in s^2

Ex 3: recombination fraction

Hemophilia and color blindness:

two recessive traits carried on the X chromosome

Family data: color blind mormor, homophiliac morfar

both mother $(\frac{ch^+}{c^+h})$ and father $(\frac{c^+h^+}{Y})$ are normal

4 daughters, 6 sons = $1(\frac{ch}{Y})+2(\frac{c^+h}{Y})+2(\frac{ch^+}{Y})+1(\frac{c^+h^+}{Y})$

Point estimate of the recombination fraction p

$\hat{p} = \frac{\text{number of recombinations}}{\text{number of sons}} = \frac{2}{6} = 0.33$

4.4 Estimated standard error

Def 5: standard error

The standard error of $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}}$

estimated standard error $s_{\hat{\theta}}$ = an estimate of $\sigma_{\hat{\theta}}$

Def 6: consistent estimate

a point estimate becoming accurate and precise for sufficiently large sample size n

$$\hat{\theta} \text{ is consistent if } \mathrm{E}((\hat{\theta} - \theta)^2) \to 0 \text{ as } n \to \infty$$

Sample mean and proportion

Two unbiased and consistent estimates:

$\bar{X}$ for μ with $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

$\hat{p}$ for p with $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

$$\text{Estimated standard errors } s_{\bar{X}} = \frac{s}{\sqrt{n}},\ s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$$

Report point estimates in the form: $\bar{X}(s_{\bar{X}})$ or $\hat{p}(s_{\hat{p}})$

Ex 3: recombination fraction

$\hat{p} = 0.33,\ s_{\hat{p}} = \sqrt{\frac{0.33 \cdot 0.67}{5}} = 0.21$

Result report

recombination fraction is estimated as 0.33 (0.21)

Ex 4: cuckoos' eggs

Length and breadth of 243 eggs in mm with frequencies

19	19.5	20	20.5	21	21.5	22	22.5	23	23.5	24	24.5	25
1	1	7	3	29	13	54	38	47	22	21	5	2

14	14.5	15	15.5	16	16.5	17	17.5	18	18.5	19
1	1	5	9	73	51	80	15	7	0	1

length $\bar{X} = 22.41$, $s = 1.08$, $s_{\bar{X}} = 0.069$

breadth $\bar{X} = 16.54$, $s = 0.66$, $s_{\bar{X}} = 0.042$

4.5 Confidence intervals

Def 7: CI for a population parameter

CI for θ of confidence level $x\%$ = an interval estimate

that covers θ with frequency $\frac{x}{100}$

when computed for many independent samples

Approximate CI

approximate $100(1-\alpha)\%$ CI for μ: $\bar{X} \pm z_{\alpha/2} \cdot s_{\bar{X}}$

approximate $100(1-\alpha)\%$ CI for p: $\hat{p} \pm z_{\alpha/2} \cdot s_{\hat{p}}$

Normal distribution table: $\Phi(z_\alpha) = 1 - \alpha$

$100(1-\alpha)$	68%	80%	90%	95%	99%	99.9%
z_α	0.47	0.84	1.28	1.64	2.33	3.09
$z_{\alpha/2}$	1.00	1.28	1.64	1.96	2.58	3.30

Ex 4: cuckoos' eggs

68% CI $\mu_L = 22.41 \pm 0.069$, $\mu_B = 16.54 \pm 0.042$

95% CI $\mu_L = 22.41 \pm 0.135$, $\mu_B = 16.54 \pm 0.083$

99% CI $\mu_L = 22.41 \pm 0.178$, $\mu_B = 16.54 \pm 0.109$

> The higher is confidence level the wider is the CI
> the larger is sample the narrower is the CI

Exact CI for the mean

Exact $100(1-\alpha)\%$ CI for μ: $\bar{X} \pm t_{\alpha/2,n-1} \cdot s_{\bar{X}}$

assuming that PD = $N(\mu, \sigma^2)$ with unknown μ and σ

Coefficient $t_{\alpha/2,n-1}$ comes from the table of

t-distribution with $(n-1)$ degrees of freedom

> Exact CI for μ is larger than approximate
> the difference is greater for small samples

Compare $t_{.025,k}$ with $z_{.025} = 1.96$

k=3	4	5	6	7	8	9	15	24	120
3.18	2.78	2.57	2.45	2.37	2.31	2.26	2.13	2.06	1.98

Ex 5: comparison of two measurements

Two methods of measuring the fat content % of meat

are compared on 16 hotdogs

16 differences of measurements: $\bar{X} = 0.53$, $s = 1.06$

Exact and approximate 95% CI for the mean difference

exact CI $= 0.53 \pm 2.13 \cdot \frac{1.06}{\sqrt{16}}$ or $(-0.03, 1.08)$

approximate CI $= 0.53 \pm 1.96 \cdot \frac{1.06}{\sqrt{16}}$ or $(0.01, 1.05)$

Ex 6: weight gain in rats

Four diets with different amount and source of protein

beef low	90	76	90	64	86	51	72	90	95	78
beef high	73	102	118	104	81	107	100	87	117	111
cereal low	95	107	97	80	98	74	74	67	89	58
cereal high	98	74	56	111	95	88	82	77	86	92

Average weight gain and estimated s.e. $\bar{X}$ $(s_{\bar{X}})$

beef low 79.2 (4.39), beef high 100 (4.79)

cereal low 83.9 (4.97), cereal high 85.9 (4.75)

build approximate and exact 99% CIs

4.6 Prediction interval

Assuming normal PD

predict a new observation X_{n+1} from n earlier obs

Approximate $100(1-\alpha)\%$ PI of X_{n+1}: $\bar{X} \pm z_{\alpha/2} \cdot s$

exact $100(1-\alpha)\%$ PI is $\bar{X} \pm t_{\alpha/2,n-1} \cdot \sqrt{s^2 + \frac{s^2}{n}}$

Two variance components: $\mathrm{Var}(X_{n+1} - \bar{X}) = \sigma^2 + \frac{\sigma^2}{n}$

population variance plus the sampling error in $\bar{X}$

Ex 7: fat content of hot dogs

Fat content (%) of $n = 10$ hot dogs: ordered sample

16.0, 17.0, 19.5, 20.9, 21.0, 21.3, 22.8, 25.2, 25.5, 29.8

Compare the exact 95% CI and exact 95% PI

CI for average fat content $21.9 \pm 2.26 \cdot \frac{4.13}{\sqrt{10}}$ or 21.9 ± 2.96

PI for the fat content of your hot dog 21.9 ± 9.81

4.7 Two methods of finding point estimates

Method of Moments Estimate

substitute population moments with sample moments

If $\mathrm{E}(X) = f_1(\theta_1, \theta_2)$ and $\mathrm{E}(X^2) = f_2(\theta_1, \theta_2)$
solve $\bar{X} = f_1(\tilde{\theta}_1, \tilde{\theta}_2)$ and $\overline{X^2} = f_2(\tilde{\theta}_1, \tilde{\theta}_2)$

a simple method, gives a first approximation for a MLE

Second sample moment $\overline{X^2} = \frac{1}{n}(X_1^2 + \ldots + X_n^2)$

Maximum Likelihood Estimate

find a parameter value that best supports the data

Def 8: likelihood function

$L(\theta) = f(x_1, \ldots x_n|\theta)$

is the joint pmf/pdf of the data set $(X_1, \ldots, X_n)$

with fixed observations $(x_1, \ldots, x_n)$ and variable θ

The MLE $\hat{\theta}$ is the value of θ that maximizes $L(\theta)$

Large sample properties of MLE

If sample is iid, then $L(\theta) = f(x_1|\theta) \ldots f(x_n|\theta)$

MLE is asymptotically unbiased, consistent, and

asymptotically efficient = minimal standard error

PD	MME = MLE	Corrected MLE
$\mathrm{N}(\mu, \sigma^2)$	$\hat{\mu} = \bar{X}$	$\bar{X}$
	$\hat{\sigma}^2 = \frac{1}{n}\Sigma(X_i - \bar{X})^2$	s^2
$\mathrm{Bin}(1, p)$	$\hat{p}$	$\hat{p}$
$\mathrm{Pois}(\mu)$	$\hat{\mu} = \bar{X}$	$\bar{X}$
$\mathrm{Exp}(\lambda)$	$\hat{\lambda} = 1/\bar{X}$	no formula

Ex 8: bus waiting time

Waiting times for a bus in minutes: 2, 7, 4, 15, 11

$X \sim \mathrm{U}(0, \theta)$, θ = fixed time between two busses

$\mathrm{E}(X) = \frac{\theta}{2}$, MME: $\tilde{\theta} = 2\bar{X} = 15.6$ min

$L(\theta) = \Pi_i f(x_i|\theta) = \Pi_i \frac{1}{\theta} 1_{\{x_i \le \theta\}} = (\frac{1}{\theta})^5 1_{\{\theta \ge 15\}}$

MLE $\hat{\theta} = 15$ min

General MLE formula $\hat{\theta} = \max(X_1, \ldots, X_n)$

$\mathrm{E}(\hat{\theta}) = \frac{n}{n+1}\theta$

corrected MLE $= \frac{n+1}{n}\hat{\theta} = 18$ min

Ex 9: DNA sequences

Given a DNA sequence

GTG CAC TGG ACT GCT GAG GAG AAG

estimate nucleotide frequences (p_A, p_C, p_G, p_T)

sample size $n = 24$, PD $= (p_A, p_C, p_G, p_T)$

Nucleotide counts distribution

$(Y_A, Y_C, Y_G, Y_T) \sim \mathrm{Mn}(24; p_A, p_C, p_G, p_T)$

$L(p_A, p_C, p_G, p_T) = \binom{24}{6,4,10,4} p_A^6 p_C^4 p_G^{10} p_T^4$

obtain MLE $\hat{p}_A = \frac{6}{24}$, $\hat{p}_C = \frac{4}{24}$, $\hat{p}_G = \frac{10}{24}$, $\hat{p}_T = \frac{4}{24}$

Ex 10: capture/recapture method

For estimating the unknown population size N

step 1: 100 animals have been tagged and released

step 2: 50 animals are captured with 20 tagged

Sampling without replacement (dependent observations)

number of tagged animals $X \sim \text{Hg}(N, 50, \frac{100}{N})$

Likelihood function

$L(N) = \text{P}(X = 20) = \frac{\binom{100}{20}\binom{N-100}{30}}{\binom{N}{50}}$

$\frac{L(N)}{L(N-1)} = 1 - \frac{20}{N} \cdot \frac{N-250}{N-130}$ larger than 1 if $N > 250$

Maximum likelihood estimate

$\hat{N} = 250$ equates two proportions $\frac{100}{\hat{N}} = \frac{20}{50}$

Ex 11: randomized response method

prison population size $N = 500$ inmates

with Np heroin users, Nq non-users

Bill rolls a die in private and responds to the statement

"I use heroin" with probability $\frac{5}{6}$ or

"I do not use heroin" with probability $\frac{1}{6}$

Observed number of "yes" answers $Y = 125$

$Y = Y_p + Y_q$, where $Y_p \sim \text{Bin}(Np, \frac{5}{6})$, $Y_q \sim \text{Bin}(Nq, \frac{1}{6})$

Observed proportion of "yes" answers $\pi = \frac{Y}{N} = 0.25$

$\text{E}(\pi) = \frac{1+4p}{6}$, $\text{Var}(\pi) = \frac{1}{N} \cdot \frac{5}{6} \cdot \frac{1}{6}$, $\sigma_\pi = \frac{1}{60}$

Method of moment estimate

solve the equation $\frac{1+4\hat{p}}{6} = \pi$ to find $\hat{p} = 0.125$

$\sigma_{\hat{p}} = \frac{6}{4} \cdot \frac{1}{60} = 0.025$, 95% CI for p is 0.125 ± 0.049

Posterior probabilities if $p = 0.125$

P(Bill uses heroin | Bill said "yes")=0.417

P(Bill uses heroin | Bill said "no")=0.028