Chapter - 5

Univariate Statistical Analysis

5.1 → Data Mining Tasks in Discovering knowledge in Data

It involves six task (Already cover in Chapter-1)

→ Description (understanding existing data)
Estimation (from sample, compute value & use they value to estimate true value)
Prediction (Predict the future)
Classification (classify the data object into predefined class)
Clustering (grouping)
Association ("go together")

5.2 → Statistical Approaches to Estimation and prediction

	Prediction	Estimation
Definition	→ predicts future or unknown values based on pattern in existing data.	→ Assigns a current unknown numerical value to a target variable.
Time oriented	Future-oriented	→ present-oriented.
Type of target	→ can be numerical or categorical	→ Always numerical (continuous)
Technique used	→ Regression, DT, NN classification model	→ Regression, KNN, Bayesian model.
O/P Nature	→ Future number or class label	→ Produces numeric Estimate.

5.3 Statistical Inference

→ It is a process of drawing conclusions about a population based on sample data;

→ In data mining, it is used to understand patterns, relationship and predictions using statistical techniques.

→ Statistical inference consists of methods like estimation and tests of hypothesis about population characteristics based on the information contained in the sample.

→ Population is the collection of all elements (persons, item or data) of interest on ~~the~~ a particular study.

→ Sample statistics to estimate unknown population.

Sample statistics	Estim	Population parameters (unknown)
Mean → $\bar{X}$	→	μ
SD → S	→	σ
Proportion → P	→	π

5.4 How confident are we in our estimates

Estimation

→ we can't check the churn behaviour of all feature cell phone customers.

→ So we take a sample, compute a value & use it to estimate the true value.

Ex If your churn sample, if 600 out of 3333 customers churned

estimate of churn rate $= \frac{600}{3333} = 0.18 = 18\%$.

Point Estimation (provides a single value not a range)

→ It is a single ~~population that~~ number calculated from the sample that is used to guess the population parameter.

Ex)

→ To find confident calculate accuracy of the estimate.

→ Accuracy of the estimate can be calculate through sampling error?

→ why we need sampling error? because a sample never perfectly represents the population.
(subset of the population)

So sampling Error = | Sample Estimate − Population Parameter |
(unknown)

i.e the population parameter is usually unknown. So we can't compute sampling error exactly. that's why we estimate it using standard error (SE)

→ Ex IF the population parameter is known (rare case)

then sampling error (SE) = | sample estimate − population parameter |

Ex Population mean = 50

sample mean = 47

$SE = |47 - 50| = |-3|$

→ Ex IF the population parameter is unknown (common case)

then we estimate the sampling error using standard error (SE)

ie standard error = estimated sampling error.

$\Rightarrow SE = \frac{s}{\sqrt{n}}$ where s = sample standard deviation

n = sample size.

Ex SD = 6 , sample size = 100

$SE = \frac{6}{\sqrt{100}} = \frac{6}{10} = 0.6$.

so sampling error is ± 0.6

5.5 confidence interval estimation of the mean

confidence interval (CI)

It gives a range of values within which the true population mean is likely to lie.

It is based on

→ sample mean
→ sample standard deviation
→ sample size
→ level of confidence (90%, 95%, 99%)

→ **confidence interval estimate**

It contais.

(I) point estimate → A single number from the sample
ex sample mean ($\bar{x}$)

(II) A margin of error ÷ how far the estimate might be from the true value. (It measures the precision of the interval estimate)
ex Smaller margin of error = more precision

(III) confidence level (90%, 95%, 99%)
↳ this tells how confident we are that the interval contain the true paramater.

Why Standard error.

→ Ex: If the population

→ General form of a confidence interval

$\boxed{CI = \text{point Estimate} \pm \text{margin Error}}$

It creates two bounds.

Lower limit = point estimate − ME

upper limit = " + ME

→ The t-interval for estimating a population mean when;

(I) the SD is unknown (usually true)

(II) either the population is normal OR sample size is large ($n \geq 30$)

then use t-distribution.

So, using t-interval

$$CI = \bar{x} \pm t_{\frac{\alpha}{2}, n-1} \times \frac{s}{\sqrt{n}}$$

where

$\bar{x}$ = sample mean (point estimate)

$t_{\frac{\alpha}{2}, n-1}$ = (t critical value at chosen confidence level)

s = sample standard deviation

n = sample size

$\frac{s}{\sqrt{n}}$ = Standard Error

$t \times SE$ = margin of error

α = Significance level

(It represents the probability of making a mistake by rejecting a true statement about the population)

Confiden level	α
90%	0.10
95%	0.05
99%	0.01

* why use t- instead of z.

→ z is used when population SD is known (rare)

→ t " " " SD must be estimated from the sample (common case)

∴ sample size increases, → t - distribution ≈ normal distribution.

* 95% confidence interval means :- If we take many samples & build intervals, 95% of those intervals will contains the true population mean.

Q From a customer service calls statistics as given sample mean $\bar{x} = 1.563$, sample SD $= s = 1.315$ sample size $n = 3333$ & confidence level $= 95\%$. use t interval ($t_{\alpha/2, n-1} \approx 1.96$) t-critical value

S-1 compute standard error $SE = \frac{s}{\sqrt{n}} = \frac{1.315}{\sqrt{3333}}$

$= \frac{1.315}{57.732} \simeq 0.02278$

S-2 compute margin of error $ME = t \times SE$

$= 1.96 \times 0.02278$

≈ 0.045

S-3 Form the confidence interval

Lower bound $= \bar{x} - ME = 1.563 - 0.045 = 1.518$

upper bound $= \bar{x} + ME = 1.563 + 0.045 = 1.608$

Q $\bar{x} = 1.607$, $n = 28$ for 95% t confidence interval estimate

$df = n-1 = 27$

$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = 1.607 \pm 2.052\,(1.8 \ldots)$ $\alpha/2 = 0.025$ $t_{0.025,\,27}$ (df 27, table 0.025)

Interpretation

We are 95% confident that the true mean number of customer service calls for the population lies between 1.518 & 1.608.

Q The contents of seven similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2 & 9.6 liters.

90%	t critical value	≈ 1.645
95%	"	≈ 1.96
99%	"	≈ 2.57

Find 95% CI for the mean contents of all such containers, assuming an approximately normal distribution, $\bar{x} = 10.0$ & $s = 0.283$ & $t_{0.025} = 2.447$

Hence CI $= 10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + 2.447\left(\frac{0.283}{\sqrt{7}}\right)$

$\Rightarrow 9.74 < \mu < 10.26$

Q The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per millimeter. Find the 95% & 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per millimeter.

Soln

Sample mean, $\bar{x} = 2.6$ g/mL

Population standard deviation $\sigma = 0.3$ g/mL (known)

Sample size $n = 36$

So we z-based interval because σ is known.

S-1 Compute Standard Error (SE) $= \frac{\sigma}{\sqrt{n}} = \frac{0.3}{6} = 0.05$

S-2 Z-critical values

→ For 95% confidence $= 1.96$

99% " $= 2.576$

-3 Margin of error (ME)

$95\% \rightarrow ME_{95} = Z_{0.025} \times SE = 1.96 \times 0.05 = 0.098$

$99\% \rightarrow ME_{99} = Z_{0.005} \times SE = 2.576 \times 0.05 = 0.1288$

-4 confidence intervals

95% CI

lower $= \bar{x} - ME = 2.6 - 0.098 = 2.502$

upper $= \bar{x} + ME = 2.6 + 0.098 = 2.698$

for 95% CI the true mean zinc concⁿ lies betⁿ 2.502 & 2.698 g/ml

99% CI

Lower $= 2.6 - 0.1288 = 2.4712$

upper $= 2.6 + 0.1288 = 2.7288$

For 99% CI the true mean zinc concentration lies between 2.4712 & 2.7288 g/ml

situation	use
σ known	Z-interval
σ unknown, $n < 30$	t-interval
σ unknown, $n \geq 30$	t-interval (preffered) but Z-interval can be used
Population normal + σ unknown	t-interval but $CI = \bar{x} \pm Z\left(\frac{s}{\sqrt{n}}\right)$

5.6 How to reduce the margin of Error

The margin of error (ME) for a 95% confidence interval for the population mean μ is

$$ME = \boxed{t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)}$$

≈ 0.045

So now smaller is the ME, the more precise our estimation.

So how to reduce

ME contains 3 quantities.

(1) $t_{\alpha/2}$, which depends upon confidence level & sample size.

(2) Sample standard deviation (s), which is the characteristics of the data.

may not be change

n, the sample size, we can reduce as follows

(i) If confidence level is reduced then $t_{\alpha/2}$ then reduce ME, (Not recommended)

(ii) By increasing the sample size, we can reduce ME while maintain a constant level of confidence.

eg sample 5000 customers.

with same $s = 1.315$. the ME.

$$ME = t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = 1.96\left(\frac{1.315}{\sqrt{5000}}\right)$$

$$= 0.036$$

with increase in sample size reduce the ME.