Biasedness in Pg. 31, 32
Page 39 - Plug-in estimators
Page 44 - MLE

11/2/21

TD 2223 - DATA ANALYSIS

Plug-in, MoM, MLE

Probability - its the study of quantification of uncertainity
The real world is not just uncertain, but also variable

Data analysis is the set of tools and techniques that helps us extract useful information from real-world variability

Questions in Adv. probability theory are very well defined, whereas question in data analysis isn't

Components —
- collecting / generating useful data in correct fashion
- extracting information from data by summarising it
- correctly interpreting statistical analysis done by others.

Lecture 1.2
Data analysis is also an art -
- Different data can be collected for same question of interest
- Measurement resolutions, methods, visualizations could be different
- Data set will inevitably have errors.

Role of computers : collection, storage & retreival of data (complex)
processing large & complex data sets.

Data processing - large & cheap memory is available
processor speed and capacity has grown substantially
new paradigms & techniques have emerged

Data analysis packages
- Important to understand the methods used by the packages

A good method needs to be selected depending upon the question and data

Output from packages needs to be interpreted.

Statistics : Theory underlying data analysis. Two types -
Descriptive - summarizing and visualising a given data
Inferential - estimating properties of the population using sampled data i.e. going beyon collected data set

(2) 22/2

Lecture 2

Describing and Visualising data

Level of measurement

All data available is a result of measurement. This decides what kind of operations and process can be carried out on the data.

Three kinds of LoM –

1. Nominal – most basic type where data collected consists of words or numeric codes

Eg: Aadhar card & list of cities of students of the class.

This produces qualitative data.

i.e. it wouldn't make sense to add the data.

But you can group or classify the data – its a feature/property of this level of measurement.

Many surveys collect through nominal LoM.

2. Ordinal LoM

Its a type of measurement where we try to see what are the relative rankings of data collected

Eg: Height of students – its possible to rank the ~~values~~ students without values

This LoM produces Ranked data.

It still lacks certain features – we know the tallest and shortest by not know by how much.

A feature of the ranked data is order – order is important

3. Interval / Ratio LoM

It is the richest level of measurement

Eg: Measuring heights of students in cm.

From this, we can always go back to ranked data

This LoM produces quantitative data

The property of this kind of data is –

- Equal intervals – can be used to compare differences in a meaningful way
- Absolute zero – On any scale, to compare differences & ratios, the scale needs to have absolute 0.

Lecture 3 - Frequency plots. (Histogram)

Frequency distribution is similar to probability distribution but here the data available is finite

Consider a r.v $X \sim p(x)$ where $p(x)$ is probability distribution. This says that the probability that X takes value in $[x, x+dx]$ is $p(x).dx$

Freq. distribution plot lets us take the finite data and represent it theoretically as $p(x)$.

Histograms were invented by Carl Pearson.
All kinds of data can be represented through this

Example - Plotting using python
Generate random numbers - np.random.normal() = x
Plot it in a histogram - plt.hist(x)
It takes a default set of bins and groups them into them (binning)
Default no. of bins = 10 # Too less bins - hides the shape of graph
Using too many bins (200 for 200 data points) will show a lot of gaps - which are not real i.e if we'd taken 2000 data points, they'd have been filled
Using too many bins will make the noise in the data more prominent
So using too less or too many bins will not give accurate representation.

→ So how many bins to use?
It depends on the no. of data points we have
There is no set formula, its subjective.
* There is a thumb rule called the Square root rule
i.e "for n data points, use $\sqrt{n}$ bins"
* Sturge's rule - For n data points, use $\log_2 n + 1$ rule
* Freedman-Diaconis rule - just use : bins = 'fd'

4 → The probability distribution / pmf should be normalised
i.e. $\int P(x).dx = 1$.

We can plot relative frequency. To do this use the argument - density = True. Very useful.

→ Some graphs although technically correct can be misleading
One way that is done is by manipulating the Y-axis limits

Eg: Yes / No - equal proportion but visually different.

13/s

Lecture 04

Error propogation

- All data have uncertainity because no measurement is perfect.
- Uncertainity in data ⇒ uncertainity in prediction and that should be quantified

Example —
Toxic chemical A with safer chemical B
Specific heat capacity (C_p) decides toxicity.
$C_p(A) = 8.3$ $C_p(B) = 8.9$

Uncertainities A - [7.8, 8.8]   B - [8.4, 9.4]

Since intervals overlap, we can't say for sure that $C_p(A)$ is different than $C_p(B)$

→ Best estimate & uncertainties of directly measurable quantities.

- Temp, length, voltage — measured directly
  If value is closer to a particular marking, thats taken as best estimate.

- Sum of uncertainties on either side is chosen
  ⇒ sum is distance b/w consecutive markings.

Uncertainties in measurement of complex quantities
Some quantities can't be directly measured —

$KE = \frac{1}{2} mv^2 \qquad v = \frac{l}{t}$

So here, error in measurement of l and t move into calculation of v and further propogate to E.
So, uncertainties may propogate in complex fashion

⟹ Reporting uncertainties
Best estimate ± Uncertainty i.e. $x_0 \pm \delta_x$

Usually, uncertainties should be rounded to 1 SF
Reporting uncertainty with many decimal places is strange ⟹ you can't be so sure of your ut!

Best estimate should be the same order of magnitude as uncertainty

1363.25 m ± 40 m (incorrect) → (1360 ± 40) m

Fractional uncertainty (Relative) : $\frac{\delta_x}{|x_0|}$ could also be expressed in percentage
Useful to calculate error propogation.

Error Propogation 2

⟹ Uncertainties in sum and difference

$w_1 = x_b \pm \delta_x \qquad w_2 = y_b \pm \delta_y$

Best estimate of $(w_1 + w_2) = x_b + y_b$

$\max(w_1 + w_2) = (x_b + y_b) + (\delta_x + \delta_y)$

$\min(w_1 + w_2) = (x_b + y_b) - (\delta_x + \delta_y)$

Uncertainty in sum : sum of uncertainties. $= \boxed{\delta_x + \delta_y}$

⑥ Similarly in $\max(w_1 - w_2)$ and $\min(w_1 - w_2)$ the uncertainty in difference of w_1 and w_2 is still sum of uncertainties : $\delta_x + \delta_y$

⇒ Uncertainty in product

$$\max(w_1 w_2) = (x_b + \delta_x)(y + \delta_y) \approx x_b y_b + (x_b \delta_y + y_b \delta_x)$$

$$\min(w_1 w_2) \approx x_b y_b - (x_b \delta_y + y_b \delta_x)$$

there, $\delta_x \delta_y$ is negligible

- Uncertainty – $(x_b \delta_y + y_b \delta_x)$
- Fractional uncertainty – $\dfrac{\delta x_1}{x_b} + \dfrac{\delta_y}{y_b}$

⇒ We can add fractional uncertainties !

⇒ Uncertainty is quotient –

$$\max\left(\frac{w_1}{w_2}\right) = \frac{x_b + \delta_x}{y_b - \delta_y} = \frac{x_b}{y_b}\left(\frac{1 + \frac{\delta_x}{x_b}}{1 - \frac{\delta_y}{y_b}}\right)$$

$$\approx \frac{x_b}{y_b}\left(1 + \frac{\delta_x}{x_b} + \frac{\delta_y}{y_b}\right) \quad \because \frac{1}{1-a} \approx 1+a$$

$|||^{ly}$, $$\min\left(\frac{w_1}{w_2}\right) = \frac{x_b}{y_b}\left[1 - \left(\frac{\delta_x}{x_b} + \frac{\delta_y}{y_b}\right)\right]$$

- Fractional uncertainty – $\boxed{\dfrac{\delta_y}{|y_b|} + \dfrac{\delta_x}{|x_b|}}$

Say, $$q = \frac{x_1 \times x_2 \cdots x_m}{y_1 \times y_2 \cdots y_n}$$

Fractional uncertainty – $\dfrac{\delta x_1}{|x_b|} + \dots \dfrac{\delta_{x_m}}{|x_{m_b}|} + \dots \dfrac{\delta y_n}{|y_{n_b}|}$

→ Special cases : Uncertainty in $\boxed{\begin{array}{l} Bx \text{ — } B\delta x \\ x^n \text{ — } n\dfrac{\delta x}{|x_b|} \end{array}}$

Function of single variable, say $q(x)$

$$q(x_b + \delta_x) \simeq q(x_b) + \underbrace{\boxed{\left(\frac{dq}{dx}\right)_{x_b} \delta_x}}_{\text{Uncertainty}}$$

This is calculated using taylor series.

Ex: derive the special cases.

Example — measuring g

$$g = \frac{4\pi^2 L}{T^2}$$

Unc in
L — δL
T — δT
T^2 — $2 T_b \delta T$

$$\frac{\delta g}{|g|} = 4\pi^2 \left(\frac{\delta L}{|L|} + \frac{2 \cdot T \, \delta T}{T^2} \right) = 4\pi^2 \left(\frac{\delta L}{|L|} + \frac{2\,\delta T}{|T|} \right)$$

Uncertainty in T contributes more to δg than δL

Error Propogation 3

Repeatable measurements.

Measuring time using stopwatch has manual errors. To get around this, we make repeated measurements — $t_1, t_2 \ldots t_n$

Best estimate : $\frac{1}{n} \sum_i t_i$ — min & max errors cancel out to give best estimate

Uncertainty interval : $[t_{min}, t_{max}]$

Pairing problem

Let $q = q(x, y)$. Suppose we measure both x & y several times, best estimate should be?

$$q(\langle x \rangle, \langle y \rangle) \quad \text{or} \quad \langle q(x_i, y_i) \rangle \quad ?$$

Area — $A = xy$

We can calculate $A_1, A_2 \dots A_n$ for (x_1, x_n) & (y_1, y_n).
But there's no reason x_1 and y_1 only should be paired. Since they're independent quantities, x_1 could be paired with any y_i.
So, its better to calculate $\langle x \rangle$, $\langle y \rangle$ and then plug it into $A(\langle x \rangle, \langle y \rangle)$.
This also works well when no. of measurements of x & y are different.

This depends on the context. Consider, $g = \frac{4\pi^2 L}{T^2}$
(L_1, L_n) and (T_1, T_n).
If we decide to use many pendulums with different lengths — and time period depends on this. So, pairings (L_i, T_i) are important.
$\langle L \rangle$ would be incorrect!
$\therefore$ Best estimate — $\langle g(L_i, T_i) \rangle$

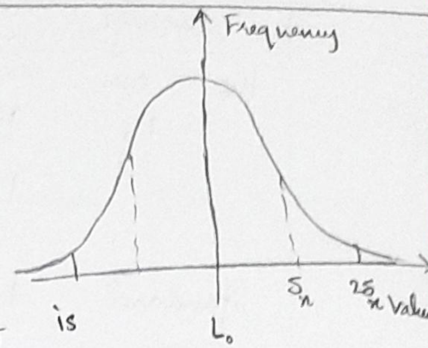
Random errors
Caused by inherent errors in act of measurement.
They're both +ve & -ve
$\Rightarrow$ Their avg is 0 when enough measurements are carried out

Systematic errors
They affect all measurements equally — cannot be reduced by increasing the no. of measurements.
Eg: difference in scale

Assumption of normality in repeated measurement.
Values obtained by repeatedly measuring a value can be thought of as having a normal distribution provided the measurements are independent.
Normal distribution —

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ : true value
σ : uncertainty.



Histogram of measured values essentially forms a gaussian curve

If an instrument is bad, it'll give a flatter, broader curve

But the real question – uncertainty in the average – how far it is from the true value

⇒ Estimating uncertainties for (any) finite data – $\hat{\mu}$ → function of measurements

$$\boxed{x_b = \hat{\mu}_x = \frac{1}{n} \sum_i x_i}$$

Spread in x : std. deviation $\boxed{\hat{\sigma}_x = \sqrt{\frac{1}{n-1} \sum_i (x_i - x_b)^2}}$ sample standard distribution

$(n-1)$ gives a better estimate. This is because once we've fixed x_b, we only have $(n-1)$ degrees of freedom

Uncertainty in value of x_b depends on $\sqrt{n}$ and the standard distribution of sample –

$$\boxed{\delta_x = \frac{\hat{\sigma}_x}{\sqrt{n}}}$$

68% confidence interval – $x_b \pm \delta_x$

95% confidence interval – $x_b \pm 2\delta_x$

Addition in quadrature

Consider two rv – $X_1 \sim N(\mu_1, \sigma_1^2)$ $X_2 \sim N(\mu_2, \sigma_2^2)$

Let $Y = X_1$ ~~and~~ $+ X_2$

It can be shown that, $\boxed{Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)}$

Var $(Y) = \sigma_1^2 + \sigma_2^2$ ⇒ std dev$(Y) = \sqrt{\sigma_1^2 + \sigma_2^2}$

Uncertainty in Y – $\delta_Y = \sqrt{\delta_{x_1}^2 + \delta_{x_2}^2}$

Works when you can justify normal distribution for random errors

* Addition in quadrature results in lesser uncertainty *

Uncertainty in counting

A process that produces something with a fixed average rate can be modelled as Poisson process.

We can count the no. of occurences and its governed by Poisson distribution —

$$p(k) = \frac{\lambda e^{-\lambda}}{k!} \quad \text{where } \lambda : \text{avg rate.}$$

Standard deviation : $\sigma = \sqrt{\lambda}$

Square root rule : Uncertainty in best estimate,

$\lambda \pm \sqrt{\lambda}$. Fractional uncertainty : $\frac{1}{\sqrt{\lambda}}$

14/3

Lecture 5

Statistical correlation 1

There could be positive or negative correlation.

- Given r.vs X and Y, we can talk about statistical relationship in addition to pdfs. This allows us to predict things.
- If P(y) is known, best possible guess of Y is $\langle Y \rangle$ or median (Y)
- Relationship can be thought of as information that allows more accurate guessing of value of Y by knowing that of X than without it.

Height and weight of children from Hong Kong

mean (w) = 57.7 kg   mean (h) = 172.6 cm

If asked to guess the height of a new child, then (in the absence of any other info) its best to guess the mean. i.e

$$y_{pred} = \langle Y \rangle$$

Best implies that avg. squared error is minimised —

$$E^2 = \frac{1}{n-1} \sum_i (y_i - y_{pred})^2$$

Can we predict Y better when value of X is known?

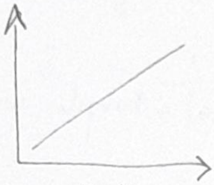
Scatterplot — Height vs weight

If we look at the plot, there is some correlation. Those with higher height have higher weight.

So if we know the height, we can predict the weight better.

Types of correlation —

- Positive : On an average, low values are paired with low values and high values with high
- Negative : Low values are paired with high values on average
- No correlation — no preference while pairing.



Positive

Negative

None

We can also have linear or non-linear correlation

linear

Non-linear

Statistical correlation 2

Measuring the strength of linear correlation

More closely the scatter of points resembles a straight line, stronger the correlation.

Pearson's correlation coefficient of sample

$$\boxed{r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}}$$

$$\bar{x} = \frac{1}{n} \sum_i x_i \qquad \bar{y} = \frac{1}{n} \sum_i y_i$$

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad S_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

S - sample std. deviation

When $x_i < \bar{x}$, if $y_i < \bar{y}$ also & similarly when $x_i > \bar{x}$, ~~dif~~ $y_i > \bar{y}$ — the low values are paired with low — then r is +ve

If $x_i < \bar{x}$ when $y_i > \bar{y}$, then r is -ve

Pearson r is independent of units of measurement.

Sample covariance and r

Given (x_i, y_i) for $i = 1, 2, \ldots n$, sample covariance is defined as —

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad \Rightarrow \quad r = \frac{S_{xy}}{S_x \cdot S_y}$$

Features —

- r lies between -1 and +1
- Sign of r indicates the ~~strength~~ nature of linear relationship
- Absolute value of r indicates the strength of the relationship.
- r is not applicable for non-linear relationship.

Linear Regression

Assume that an observed character (X, Y) is produced by $Y = a + bX + \varepsilon$ where —
we assume that Y is related to X by a linear, deterministic relationship.

ε : Noise that is normally distributed

Homoscedasticity : uncertainty σ in each y_i is same

Errors in measurement of X can be neglected. (because uncertainty in Y is much greater)

How to estimate a, b given the data? Best fit?

Ans: Minimize the squared error (aka predictive or residual error)

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \frac{1}{n-1}\ ?$$

Given (x_i, y_i) for $i = 1, 2 \ldots n$

$$\begin{cases} \hat{a} = \dfrac{1}{n(n-1)} \dfrac{\sum x_i^2 \sum y - \sum x \sum xy}{S_x^2} \\[2ex] \hat{b} = \dfrac{1}{(n-1)} \dfrac{\sum xy - \frac{1}{n}\sum x \sum y}{S_x^2} \end{cases}$$

26/3

Statistical Correlation 3

Deriving the expression of a & b

Given (x_i, y_i) for $i = 1, 2 \ldots n$

We're just estimating a & b — to the best of our ability. True value could be different

Minimizing square error, $E = \sum_{i=1}^{n} \Big(y_i - \underbrace{(a + bx_i)}_{Y_{predicted}} \Big)^2$

So we can see that E is a function of a & b, not x & y.

i.e. we've to minimise E wrt a & b by differentiating E ~~to 0~~ & equating to 0.

★ $$\frac{\partial E}{\partial a} = \sum_{i=1}^{n} 2\left(y_i - (a+bx_i)\right)(-1)$$

$$\frac{\partial E}{\partial a} = 0 \Rightarrow \sum_{i=1}^{n} y_i - a(n) - b\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow n\bar{y} - a(n) - bn\bar{x} = 0$$

$$\boxed{a = \bar{y} - b\bar{x}}$$

★ $$\frac{\partial E}{\partial b} = \sum_{i=1}^{n} 2\left(y_i - (a+x_ib)\right)(-x_i) = 0$$

$$\sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$$

$$\sum x_i y_i - (\bar{y} - b\bar{x})\sum x_i - b\sum x_i^2 = 0$$

$$\sum x_i y_i - \bar{y}\sum x_i + b\left(\bar{x}\sum x_i - \sum x_i^2\right) = 0$$

$$b = \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \bar{x}\sum x_i} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

Sample std. dev: $s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{n-1}\sum\left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right)$

$$\cancel{S_x = \frac{1}{n-1}\left[\sum x_i^2 - 2\bar{x}^2 + n\bar{x}^2\right]}$$

$$S_x^2 = \frac{1}{n-1}\left[\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2\right] = \frac{1}{n-1}\left[\sum x_i^2 - \frac{2}{n}\left(\sum x_i\right)^2 + \frac{n}{n^2}\left(\sum x_i\right)^2\right]$$

$$S_x^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]$$ — subsituting this in denominator in b.

$$\therefore \left\{ \hat{b} = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{(n-1)\, S_x^{\,2}} \right\}$$

$\hat{b}$ is an approximation based on our data.
Using the value of $\hat{b}$, we can calculate the value of $\hat{a}$ -

$$\hat{a} = \frac{1}{n(n-1)} \; \frac{\Sigma x_i^2 \, \Sigma y_i - \Sigma x_i \, \Sigma x_i y_i}{S_x^{\,2}}$$

Statistical Correlation 4

Relationship between r and regression line

Pearson's r and the slope of of the regression line are related by -

$$b = r \frac{S_{\cancel{X}}}{S_{\cancel{X}}} \qquad \frac{S_Y}{S_X} \qquad b = r\frac{S_Y}{S_X}$$

The y-intercept of regression line is given by,

$$\left\{ a = \bar{y} - r\bar{x}\frac{S_x}{S_Y} \right\} \quad \text{Y-intercept}$$

We saw that $b = \dfrac{\Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i}{(n-1)\, S_x^2}$

$$r = \frac{1}{(n-1)} \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{S_x S_Y} = \frac{1}{n-1} \cdot \frac{\Sigma\left[x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x}\bar{y}\right]}{S_x S_Y}$$

$$r = \frac{1}{(n-1)\, S_x S_Y}\left[\Sigma x_i y_i - \frac{1}{n}\Sigma y_i \Sigma x_i \cancel{- \frac{1}{n}\Sigma x_i \Sigma y_i} + \cancel{n\frac{\Sigma x_i \Sigma y_i}{n^2}}\right]$$

$$r = \frac{1}{(n-1)} \frac{\left[\Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i\right]}{S_x S_Y} = b \cdot \frac{S_x}{S_Y}$$

$$\therefore b = r\frac{S_Y}{S_x}$$

Previously, $E_{tot} = \Sigma (y_i - \bar{y})^2$ – error obtained by predicting the mean value each time

But why is r a measure of strength of linear relationship?

Residual error : $\boxed{E_{res} = \Sigma \left(y_i - \hat{a} - \hat{b}x_i\right)^2}$

these estimates are predicted through $\hat{a}$ & $\hat{b}$.

Residual – difference b/w what's observed and what's predicted

If there is a strong correlation b/w x and y, we should be able to predict it better

i.e E_{res} should be smaller.

Coefficient of determination,

$$R^2 = \frac{E_{tot} - E_{res}}{E_{tot}} = \boxed{1 - \frac{E_{res}}{E_{tot}} = r^2}$$

It tells us the reduction in error by assuming that there's a correlation between x and y. If the reduction isn't much, ($E_{res} \sim E_{tot}$) then there is no significant correlation.

If there's a strong correlation, then we can predict y better which means error is reduced $E_{res} < E_{tot}$

27/3

Statistical correlation 5

$$E_{res} = \Sigma (y_i - \hat{a} - \hat{b}x_i) = \Sigma \left[\left(y_i - \{\bar{y} - b\bar{x}\} - \hat{b}x_i\right)^2\right]$$

$$E_{res} = \Sigma \left((y_i - \bar{y}) - b(x_i - \bar{x}) \right)^2$$

$$E_{res} = \Sigma (y_i - \bar{y})^2 - 2b\, \Sigma (y_i - \bar{y})(x_i - \bar{x}) + b^2\, \Sigma (x_i - \bar{x})^2$$

From eq^n for r

$$E_{res} = (n-1)\, S_Y^2 - 2b \cdot (n-1)\, S_x \cdot S_Y\, r + b^2 \cdot (n-1)\, S_x^2$$

$$E_{res} = (n-1)\, S_Y^2 - 2(n-1) \cdot r \frac{S_Y}{\cancel{S_x}}\, \cancel{S_x} \cdot S_Y\, r + (n-1) \frac{\cancel{S_x^2} \cdot r^2\, S_Y^2}{\cancel{S_x^2}}$$

$E_{res} = (n-1)\, S_Y^2 - r^2 (n-1) . S_Y^2$

$\therefore\ E_{res} = (n-1)\, S_Y^2 \left[1 - r^2\right]$ → reduces with increased correlation ∴ y can be predicted better.

Simplified expression.

→ $E_{tot} = \Sigma (y_i - \bar{y})^2 = (n-1)\, S_Y^2$

$R^2 = 1 - \frac{E_{res}}{E_{tot}} = 1 - (1 - r^2) = r^2$

This is for linear model : $y = a + bx$

If R^2 is big, then model is a good predictor of the data in comparison to the model that just predicts the mean each time.

R^2 can also be negative!

Using r

- Check assumption of linearity (at least visually).
- Check for ~~no~~ homoscedasticity – spread in values of y remains the same for any x

greater range of y for higher x

Heteroscedasticity

→ Correlation is not causation

Strong correlation doesn't imply that one causes the other

High correlation means: $P(Y|X) \neq P(Y)$

X allows for better prediction because there's association

Cause : undefined notion in traditional statistics

High correlation could be a result of direct causation, existence of confounders or collider bias

↳ common cause of X & Y

Age → Shoe size, reading ability

3/4

Review of Random variables

Book: All of statistics

Sample space : Set of all possible outcomes

Random variable is a map from sample space to real numbers

Coin toss : $\{1, 0\}$

Deck of cards : $\{1, ---, 52\}$

Discrete rv

Prob. mass function : pmf $\mathbb{P}(x)$: Prob that X takes value x

Continuous rv

Prob. density function : pdf $p(x).dx$: P that X takes value in range $[x, x+dx]$

Cumulative distribution function

$$F(x) = \mathbb{P}(X \leq x) = \int_0^x p(x).dx$$

$$X \sim F$$

For independent rv : $P(X=x, Y=y) = P(X=x).P(Y=y)$

⟹ Inverse CDF (Quantile)

$$F^{-1}(q) = \inf\{x : F(x) > q\} \quad \text{for } q \in [0,1]$$

q is the probability. $F^{-1}(q)$: The infimum value of x for which $F(x)$ attains q

First quartile : $F^{-}(1/4)$

Median : $F^{-}(1/2)$

Interquartile range: $\left[F^{-}\left(\frac{1}{4}\right), F^{-}\left(\frac{3}{4}\right)\right]$

↳ Middle 50% of values lie in this range

Expectation

- Discrete : $E[X] = \sum x \, \mathbb{P}(x)$

Continuous : $E[X] = \int x \, p(x) \, dx$

$$E[X] = \int x \, dF(x)$$

- k^{th} moment : $E[X^k]$

- $E\left[\sum_i a_i X_i\right] = \sum_i a_i E[X_i]$

If X_i are independent, $E\left[\prod_i X_i\right] = \prod_i E[X_i]$

Variance

$$V(X) = E\left[(X - E[X])^2\right]$$

If X_i are independent,

$$V\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \, V(X_i)$$

Convergence of RV

Just as a series can converge, RV can also converge.

Say, $X_n \sim N\left(\frac{1}{n}, \sigma^2\right)$ where $n \in \mathbb{N}$

We could sort of say that this converges to $N(0, \sigma^2)$

But to be more clear,

1. Convergence in probability

If for every $\epsilon > 0$, $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$, then X_n is said to converge in probability.

Written as, $X_n \xrightarrow{P} X$

(20) 2. Convergence in distribution

X_n converges to X in distribution if

$$\lim_{n \to \infty} F_n(t) = F(t)$$

at all t for which F is continuous

Written as, $X_n \rightsquigarrow X$

* Convergence in probability implies the convergence in distribution, but the ~~convergence~~ converse is not true. *

Introduction to Inferential Statistics

This part allows us to generalise from given data and not just summarising it.

Population : Complete set of observations of interest

Sample : Subset of observations drawn from the population

Hypothetical population: imagined population from which observed data is thought to be drawn

Inferential statistics : estimating properties of underlying population by studying a sample

The question is asked about the population and studied in a sample to answer it.

Two methods to estimate properties —

1. Surveys
   - Sampling the population so that resulting sample is representative of population.
   - Uniform random sample best represents the underlying population

2. Experiments
   - Finding the effect of interventions
   - Control group vs Treatment group
   Random assignment to 2 groups
   - Is the difference b/w two groups attributable to intervention?
   - Can work even when sample is not completely random (convinience sampling).

Types of Inferential statistics –

- Parametric inference
  Statistically well defined.
  Assumes that data comes from a population that can be adequately modelled by a prob. distribution that has a fixed set of parameters.
  Eg: Estimate variance given 100 nos. drawn from Gaussian distribution with mean 0

- Non-parametric (Distribution free) inference
  No distribution is assumed.
  Eg: estimating the mean income of people in a city by randomly choosing 1000 people.

Statistical Inf. in the language of RV. 5/4

Fundamental problem:
Given a IID sample of $X_1, X_2 \dots X_n$, then how to infer their CDF 'F'?

- Statistical model – restricting to st. line
  Set $\mathcal{F}$ of distributions or regression functions
- Parametric model
  Set $\mathcal{F}$ that can be parametrized – with finite parameters

$$\mathcal{F} = \{ N(x; \mu, \sigma) : \mu, \sigma \in \mathbb{R}, \sigma > 0 \}$$

Set of all possible normal distribution

- Non parametric model
  Set $\mathcal{F}$ that cannot be parametrized by finite no. of parameters.

22 * 1D parametric estimation

Let $X_1, \dots X_n$ be IID Bernoulli (p) observations.

How to estimate p?

* 2D parametric estimation

Say $X_1 \dots X_n \sim N(\mu, \sigma)$. How to estimate μ, σ?

Non-parametric estimation —

- CDF : Let $X_1 \dots X_n \sim F$. How to estimate F assuming $F \in \{\text{all CDFs}\}$

Can't be done with finite parameters

∴ Distribution free analysis

- Statistical functionals: $X_1 \dots X_n \sim F$. How to estimate $\mu = E[X_i]$ assuming that μ exists.

Functional : Takes a function as an input and produces a real no.

Eg: Mean of distribution

Three types of inferences —

Point estimation : to find one value

Confidence sets ; sets that contain the value of interests with some prob.

Hypothesis testing: To test if something is true or not

5/4

Point estimation 1

- Single "best guess" of a quantity of interest
- Point estimate of θ is denoted by $\hat{\theta}$

Most common estimate is the mean of all values.

- for IID $X_1, X_2 \dots X_n$,

$$\hat{\theta} = g(X_1, X_2 \dots X_n)$$

The function g is called the $\boxed{\text{estimator}}$ of θ

Eg: German Tank Problem – highest value, $2 \times$ mean value

Eg. 2 : For IID normal r.v $X_1 \dots X_n$

$$\hat{\mu} = \frac{1}{n}\sum_i X_i$$

Note : $\hat{\theta}$ is also a r.v. because it's a function of data

5/5

Point estimation 2

Consider IID Normal r.v.s $X_1, X_2 \dots X_n$ with unknown μ and σ^2

$$\boxed{\mu = \int x\, P(x) \,.\, dx}$$ where $P(x)$ is the pdf.

This is the defn of mean of a distribution

So, $$\boxed{\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

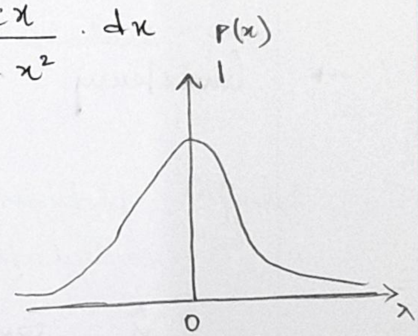
In ~~ther~~ terms of ~~$\hat{\mu}$ as~~ stat. inference,

$\hat{\mu}_1$ and μ are not the same.

$\hat{\mu}$ is an estimator of mean and can change with data given

Consider Cauchy distribution whose pdf –

$$P(x) = \frac{c}{1+x^2} \qquad \mu = \int_{-\infty}^{\infty} \frac{cx}{1+x^2} \,.\, dx$$



p(x) diverges at a point to the value of μ is infinite

i.e. undefined

Say we draw a sample from the distribution and calculate $\hat{\mu}$, the mean won't be close to 0 – the estimator will give farther and farther values, even if sample size is huge

p(x) is symmetric around 0.

(24)

Say. $\hat{\mu}_2 = x_1$ — guess based on first value of distribution

$\hat{\mu}_3 = \frac{1}{n'} \sum_{x_i > c} x_i$ where c is any arbitrary number

Bias and Consistency

→ Bias : $\boxed{\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta}$ (Expected value; True value)

$\hat{\theta}$ is unbiased when $E[\hat{\theta}] = \theta$

★ For $\hat{\mu}_1$ — $E[\hat{\mu}_1] - \mu = \frac{1}{n} \sum_i E[x_i] - \mu$

For normal distribution $E[x_i] = \mu$

$\Rightarrow \frac{1}{n} \cdot n\mu - \mu = 0$

$\hat{\mu}_1$ estimator is unbiased

★ $E[\hat{\mu}_2] - \mu = E[x_1] - \mu = \mu - \mu = 0$

★ $E[\hat{\mu}_3] - \mu$ Here, $E[\hat{\mu}_3] > c$

$\Rightarrow c - \mu \neq 0$

So this estimator is biased.

→ Consistency — $\hat{\theta}$ is consisten if — Convergence In probability for large enough sample size

$$\boxed{\hat{\theta}_n \xrightarrow{P} \theta}$$

$P(|\hat{\theta} - \theta| > \epsilon) \to 0$

$\hat{\mu}_2$ can be arbitrarily small or large despite the sample size So, its not consistent

$\hat{\mu}_1$ is consistent

$\hat{\mu}_3$ is also <u>not</u> consistent.

An unbiased, consistent estimator is the best

Sampling Distribution

Say $X_1, X_2, \ldots X_n \sim \text{Exp}(\beta)$ $\quad p(x) = \frac{1}{\beta} e^{-x/\beta}$ for $x > 0$

then, $\langle x \rangle = \beta$

Say $\hat{\beta} = \frac{1}{n} \sum_i X_i$

If we repeat the experiment many times, we'll get different values of $\hat{\beta}$.

The distribution of $\hat{\beta}$ is known as sampling distribution.

~~Sample~~ Standard error

If the distribution of sampling distribution is very wide, then the estimate of $\hat{\beta}$ is not good.

Standard error is the std dev of sampling distribution.

$$\boxed{se = \sqrt{var(\hat{\theta})}}$$

$\hat{se}$ is the estimator of standard error.

→ Simulation

The sampling distribution is completely different from the underlying distribution

Mean-squared-error [MSE] of an estimator

$$MSE = E\left[(\hat{\theta} - \theta)^2\right]$$

Theorem : $MSE = bias^2(\hat{\theta}) + var(\hat{\theta})$

Theorem : If $bias \to 0$ and $se \to 0$ as $n \to \infty$ then $\hat{\theta}_n$ is consistent and unbiased (given).

Asymptotic normality

An estimator is <u>asymptotically</u> <u>normal</u> if –

$$\boxed{\frac{\hat{\theta}_n - \theta}{se} \rightsquigarrow N(0,1)}$$

↳ std. normal distribution

Z-score : $\boxed{\frac{x - \mu}{\sigma}}$: If x is a normal distribution this will normalise it st. $\mu = 0$ and $\sigma^2 = 1$

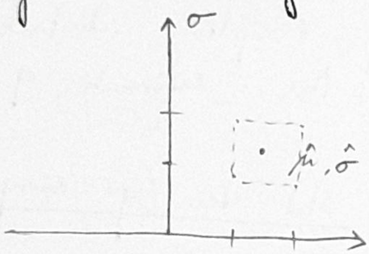
5/5

Confidence Sets & Hypothesis testing

↳ a region in parameter space that contains the quantity of interest.

Confidence set is assigned a level of confidence

Eg: Estimating μ and σ of normal distribution



Say, you're 95% confidence that μ and σ fall in that region – 2D confidence set.

Error bars are 1D confidence set.

$1 - \alpha$ confidence interval for parameter θ is an open interval $C_n = (a, b)$ where –

$a = a(x_1, x_2 \dots x_n)$

$b = b(x_1, x_2 \dots x_n)$

Such that $\boxed{P(\theta \in C_n) \geq 1 - \alpha}$

Usually people choose ~~as~~ $\alpha = 0.05$

$\Rightarrow$ (a, b) traps θ with ~~a~~ probability $1 - \alpha$, its called the coverage of confidence interval.

Hypothesis testing

Principled way of deciding whether observed data is sufficient to reject the default position.

Default position - null hypothesis (H_0)

Complementary position - alternative hypothesis (H_1)

Eg: Deciding whether a coin is fair or not - [doesn't assume anything]

$H_0 : p = \frac{1}{2}$

$H_1 : p \neq \frac{1}{2}$

If $p = 0.9$ then its fair to say its loaded.

Non-parametric Estimators 1

Population mean : $\mu = \int x \cdot p(x) \cdot dx$ - independent of distribution

also

Population variance : $\sigma^2 = \int (x-\mu)^2 p(x) \cdot dx$

Population correlation coefficient : $r = \frac{1}{\sigma_x \cdot \sigma_y} \iint (x-\mu_x)(y-\mu_y) \cdot P(x,y) \cdot dx dy$

Any estimator of distribution-free quantities is known as non-parametric estimator

for $X_1, X_2 .. X_n$ IID

Sample mean : $\hat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

Sample variance : $\hat{\sigma}^2 = S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$

Sample correlation coefficient : $\hat{r} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{S_x S_y}$

These estimators are dependent on the sample values.

28

NPE 2

Sample mean – most important NPE

We assume that $X_1, X_2 \dots X_n$ are IID and have finite mean and finite variance

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

By the Central Limit Theorem, distribution of $\bar{X}_n$ converges to distribution of normal r.v.

$$\bar{X}_n \rightsquigarrow X$$

1. What is the mean of sampling distribution of $\bar{X}_n$? What is the se?

WKT, Sampling distribution of $\bar{X}_n$ is Gaussian.

At very large n, the distribution → normal

Mean : $E[\bar{X}_n] = E\left[\frac{1}{n} \sum_i X_i\right] = \frac{1}{n} \sum_i E[X_i]$

$$\Rightarrow \frac{1}{n} \cdot n\mu = \mu$$

The mean of sampling distribution is the same as population mean (regardless of distribution).

2. Std error.

variance of sampling distribution :

$$V(\bar{X}_n) = V\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\sigma^2$$

$$\therefore se = \sqrt{V(\bar{X}_n)} = \frac{\sigma}{\sqrt{n}}$$

Std dev of sampling distribution (se) is reduced by factor of $1/\sqrt{n}$ from the σ of population (og distribution).
std dev

So, to increase the accuracy by a factor of 10, the sample should be increased by a factor of 100. *

$\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$: This rv is same as standard normal rv i.e $Z \sim N(0,1)$

Z - score : $\boxed{z = \dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}$

8|5

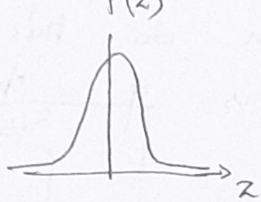
NPE 3

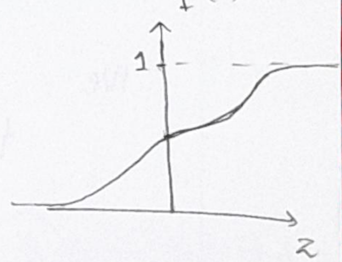
$$\text{Bias}(\bar{X}_n) = \mathbb{E}[\vec{X}_n] - \mu = \mu - \mu$$
$$= 0$$

Consistency : $\hat{\theta}_n \xrightarrow{P} \theta$ i.e. $P(|\hat{\theta}_n - \theta| > \epsilon) \to 0$

Std normal rv $Z \sim N(0,1)$

CDF : $\Phi(z) = \mathbb{P}(Z \leq z)$

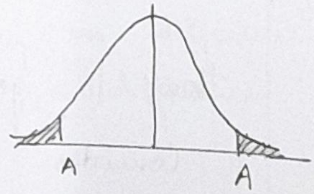




$$P(|\bar{X}_n - \mu| > \epsilon) = P\left(\frac{|X_n - \mu|}{\sigma/\sqrt{n}} > \frac{\sqrt{n}\epsilon}{\sigma}\right)$$

$$= P\left(|Z| > \frac{\sqrt{n}\epsilon}{\sigma}\right)$$

This is possible when $Z > \dfrac{\sqrt{n}\epsilon}{\sigma}$ or $Z < -\dfrac{\sqrt{n}\epsilon}{\sigma}$



$$= 2P\left(Z > \frac{\sqrt{n}\epsilon}{\sigma}\right)$$

$$= 2\left(1 - P\left(Z \leq \frac{\sqrt{n}\epsilon}{\sigma}\right)\right)$$

$$= 2\left(1 - \Phi\left(\frac{\sqrt{n}\epsilon}{\sigma}\right)\right)$$

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 2\left(1 - \lim_{n\to\infty}\Phi\left(\frac{\sqrt{n}\epsilon}{\sigma}\right)\right)$$

$$= 2(1-1) = 0$$

So estimator is also consistent

(30)

8/5

NPE 4

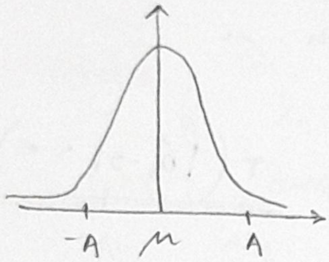
WKT, $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$\bar{X}_n$ is an estimator of μ.

But we should also know the range of values in which μ lies with a certain confidence level.

i.e the $1-\alpha$ confidence interval

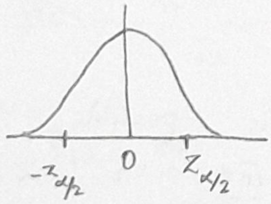
α : level of significance



Constructing an interval $(-A, A)$ whose midpt is μ and the probability that $\bar{X}_n$ lies in this interval is $(1-\alpha)$

$$\int_{\mu-A}^{\mu+A} p(x) \cdot dx = 1-\alpha$$

We can do this much more easily in terms of std normal distribution: $z \sim N(0,1)$.



$$\boxed{\int_{-z_{\alpha/2}}^{z_{\alpha/2}} \phi(z) \cdot dz = 1-\alpha}$$

For $\alpha = 0.05$, $1-\alpha = 0.95$

$\Rightarrow \boxed{z_{\alpha/2} \simeq 1.96}$

How to generalise

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

$$\frac{X-\mu}{\sigma} \sim N(0,1) = z$$

To find $x_{\alpha/2}$, we can use this,

$$\frac{\mu \pm x_{\alpha/2} - \mu}{\sigma} = \pm z_{\alpha/2}$$

$$\therefore \boxed{\pm x_{\alpha/2} = \pm z_{\alpha/2} \cdot \sigma}$$

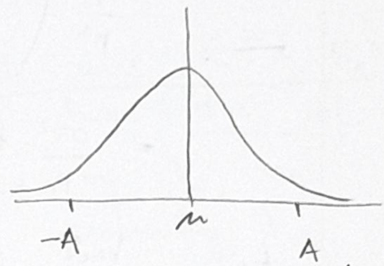
$E[XY] = E[X] \cdot E[Y]$ only if X, Y are independent

So, for $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, we can say that the interval for α is given by,

$$\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \ \mu + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Note that μ is a fixed quantity and $\bar{X}_n$ is a rv. So its correct to say that $\bar{X}_n$ occurs within the given interval 95% ($\alpha = 0.05$) of the time and not the other way round

$$|\bar{X}_n - \mu| < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{95\% of time}$$



The interval $(\bar{X}_n - A, \ \bar{X}_n + A)$ may or may not contain μ.

25/5

NPE 5

Sample variance

$$\hat{\sigma_1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \quad \text{for } X_1, X_2 \dots X_n \text{ i.i.d rvs}$$

$$\hat{\sigma_2}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

When n is large, there's virtually no difference

But when n is small, there's a difference

Recall : Bias $= E[\hat{\theta}] - \theta$

$$(n-1) E[\hat{\sigma_1}^2] = E\left[\sum_{i=1}^{n}(x_i^2 - 2\bar{x}_n x_i + \bar{x}_n^2)\right]$$

$$= \sum_{i=1}^{n}\left(E(x_i^2) - 2\bar{x}_n E(x_i) + E(\bar{x}_n^2)\right) \quad - (1)$$

$\star$ $E[X_i^2] = \sigma^2 + \mu^2$

$V(X_i) = E[X_i^2] - (E[X_i])^2 \Rightarrow \sigma^2 = E[X_i^2] - \mu^2$

$\star$ $E[X_i \bar{X}_n] = E\left[X_i \frac{1}{n} \sum_{j=1}^{n} X_j\right] = \frac{1}{n} \sum_{i,j=1}^{n} E[X_i X_j]$

$= \frac{1}{n} \left(E[X_j^2] + \sum_{j \neq i} E[X_i] \cdot E[X_j] \right)$

$= \frac{1}{n} \left(\sigma^2 + \mu^2 + (n-1)\mu^2 \right) = \frac{\sigma^2}{n} + \mu^2$

$\star$ $E[\bar{X}_n^2] = E\left[\frac{1}{n^2} \left(\sum_{i=1}^{n} X_i \right)^2 \right] = \frac{1}{n^2} E\left[\sum_i \sum_j X_i X_j \right]$

$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} E[X_i X_j]$ # for n terms out of n^2, independence doesn't hold where $i=j$

$= \frac{1}{n^2} \left(n(\sigma^2 + \mu^2) + (n^2 - n)\mu^2 \right)$

$= \frac{\sigma^2}{n} + \cancel{\frac{\mu^2}{n}} - \cancel{\frac{\mu^2}{n}} + \mu^2 = \frac{\sigma^2}{n} + \mu^2$

Going back to Eq^n 1

$(n-1) E[\hat{\sigma}_1^2] = \sum_{i=1}^{n} \left(\sigma^2 + \cancel{\mu^2} - 2\left(\frac{\sigma^2}{n} + \cancel{\mu^2} \right) + \frac{\sigma^2}{n} + \cancel{\mu^2} \right)$

$= \sum_{i=1}^{n} \left(\sigma^2 \left(1 - \frac{1}{n} \right) \right) = n \left(\frac{n-1}{n} \right) \sigma^2$

$= (n-1)\sigma^2$

$\therefore E[\hat{\sigma}_1^2] = \sigma^2$

$\therefore$ This estimator for sample variance is unbiased.

$n \cdot E[\hat{\sigma}_2^2] = (n-1)\sigma^2 \Rightarrow E[\hat{\sigma}_2^2] = \left(1 - \frac{1}{n} \right) \sigma^2$

So for smaller values, the second estimator undervalues the sample variance

Although asymptotically, the estimator is unbiased.

For constructing confidence interval for $(1-\alpha)$,

$*\left(\bar{X}_n - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \; \bar{X}_n + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right) *$ since we're using an estimator, $z_{\alpha/2}$ is not the right value for $(1-\alpha)$ CI

Allows us to construct confidence interval without using (knowing) μ or σ.

NPE6

Degrees of freedom

$X_1, X_2 \ldots X_n$ represents n independent numbers.

$\bar{X}_n = \frac{1}{n} \sum_i x_i$: We can say $\bar{X}_n$ has n degrees of freedom

Sample variance : $\hat{\sigma}^2$ / S_x^2 $= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$

Here, we don't have n independent values — once $\bar{X}_n$ is fixed, we have only $(n-1)$ free variables are present

If we have a formula which uses $\bar{X}_n$ & S_x^2, the computed quantity will have $(n-2)$ DoF.

So in $\hat{\sigma}_2^2$ where we divided by $1/n$, we were underestimating S_x^2 by assuming it has n DoF

When calculating CI, we assumed that if n is sufficiently large, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightarrow Z(0,1)$ [∵ CLT]

CI for μ : $\left(\bar{X}_n - Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \; , \; \bar{X}_n + Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right)$

By going from $\sigma \rightarrow \hat{\sigma}$, the coverge shrinks, its not $(1-\alpha)$ anymore. This needs to be rectified

NPE 7

$t = \frac{\bar{X}_n - \mu}{\hat{\sigma}_n / \sqrt{n}}$ As $n \rightarrow \infty$, $\hat{\sigma}_n \xrightarrow{P} \sigma$

When $n \rightarrow \infty$, ~~whe~~ we can justify normality of t and $\hat{\sigma}_n \rightarrow \sigma$.

But when n is small, t need not be normally distributed unless x_i itself is normally distributed. And even then, it might not be so distributed $\because \hat{\sigma}_n \neq \sigma$.

(34) We'll assume that $X_1, X_2 \dots X_n$ are normally i.i.d distributed. The distribution of t was discovered by William Gosset – pseudoname Student. Hence Student's t-distribution

$$\left\{ \psi(t) = \underbrace{\frac{1}{\sqrt{\pi k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(k/2\right)}}_{\text{Normalisation factor}} \frac{1}{\left(1 + \frac{t^2}{k}\right)^{k+1/2}} \right\}$$

$$\Rightarrow \int_{-\infty}^{\infty} \psi(t) = 1$$

Here, k: degrees of freedom $\boxed{k = n-1}$

If n is small, then t distribution is given by $\psi(t)$ for k degrees of freedom

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x}\, dx \qquad \Gamma(z) = (z-1)!$$

$z \in \mathbb{R}$
$z > 0$

For $z \in$ Integer

Computational 'proof': we'll plot t distribution and then plot $\psi(t)$ on top to make sure they match.

To do this, we'll take $n = 3$ of X_i with $\bar{X}_n$, $\hat{\sigma}_n$. We'll take values of $t \left(= \frac{\bar{X}_n - \mu}{\hat{\sigma}_n / \sqrt{n}} \right)$ 10,000 times and plot histogram.

Student's t - distribution has thicker tails in the plot

$\Rightarrow$ Its confidence interval for $1-\alpha$ will not fit what we'd calculated using std. normal distribution

We would be underestimating the CI.

To construct a good confidence interval, we need a value $(t_{\alpha/2})$ similar to $z_{\alpha/2}$.

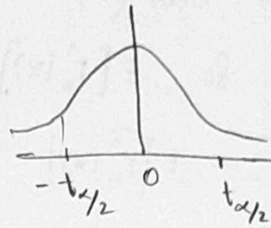
$$* \left(\bar{X}_n - t_{\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}} ,\ \bar{X}_n + t_{\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}} \right) *$$

Histogram is sensitive to sample size ∴ of noise

$\hat{F}_n$ puts the mass $\frac{1}{n}$ at each x_i

$t_{\alpha/2}$ is defined so that,

$$\int_{-t_{\alpha/2}}^{t_{\alpha/2}} \Psi(t)\, dt = 1-\alpha$$



in Student's t-distribution

Say, $\alpha = 0.05$. We have a table which gives values of t for different α and k.

We can use scipy.stats we can import module 't'

$$\gg t.\text{ppf}\,(q = 1 - \underbrace{0.05}_{\alpha}/2\,,\ \underbrace{df}_{k} = 2)$$

As $n \to \infty$ $\quad t \xrightarrow{P}$ std. normal

3|6

EDF 1 At any pt x, think of $\hat{F}_n$ as a point estimation of F_n.

Not its pdf

Can we estimate the distribution itself from a sample – not μ or σ^2? We'll restrict to estimating the distribution ~~for~~ through its CDF

$$F(x) = P(X \leq x)$$

When it comes to estimating the shape of distribution, the CDF is more robust ∵ less fluctuations in the no. of points in a range

Also, CDF is same fn for continuous & discrete rv

The main goal is to find out how good our estimates are from the sample.

Estimator of CDF – Empirical Distribution function (EDF)

Consider $X_1, \dots X_n$ iid rv

EDF : $\boxed{\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)}$ $\quad$ I : indicator fn

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > 0 \end{cases}$$

(36)

Bias and consistency of EDF

Here, we fix value of x.

EDF $\hat{F}_n$ is just a sample mean of Bernoulli RV with success prob. $F(x) = p$

$\text{Bias}(\theta) = \mathbb{E}[\hat{\theta}] - \theta$

So, $E[\hat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n} E[I(x_i \le x)]$

What's p that I takes value 1?

$E[\hat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n} F(x) = \frac{1}{n} \cdot n F(x)$

Think of it like Bernoulli rv

$E[W] = 1 \times p + 0(1-p)$

$E[\hat{F}_n(x)] = F(x)$

It takes value 1 with prob. F_x

$\therefore$ Estimator is unbiased

It's also consistent if

$\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$ (for fixed value x)

(?) Recall theorem: mean-squared error $\to 0 \Rightarrow$ estimator is consistent

MSE = bias^2 + Variance

WKT, bias = 0

page 25

Recall

$V(aX + bY) = a^2 V(X) + b^2 V(Y)$

$V(\hat{F}_n(x)) = \sum_{i=1}^{n} \frac{1}{n^2} V(I(x_i \le x))$

$= \sum_{i=1}^{n} \frac{1}{n^2} F(x)(1-F(x))$

Again $I \sim \text{Bern}(F(x))$

$= \frac{1}{n^{\cancel{2}}} \cdot \cancel{n} F(x)(1-F(x))$

$$\therefore \boxed{V(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}}$$

This tends to 0 as $n \to \infty$

$\therefore$ MSE $\to 0 \Rightarrow$ Estimator is consistent.

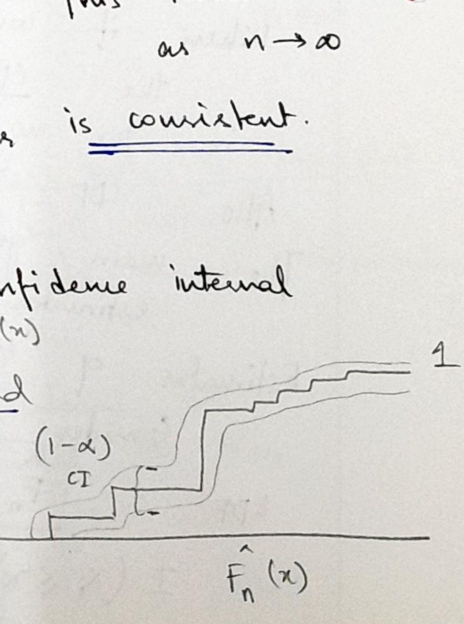
EDF 2

Next we need to construct confidence interval

CI forms a band around $\hat{F}_n(x)$

So, it's called a confidence band

We don't know $F(x)$, so in the variance formula, we replace it by $\hat{F}_n(x)$ for large n



$$\Rightarrow \left\{ \sqrt{V(\hat{F}_n(x))} = \left[\frac{1}{n} \hat{F}_n(x) \left(1 - \hat{F}_n(x)\right) \right]^{1/2} = \hat{se} \right\}$$

Confidence Interval : $\hat{F}_n(x) \pm z_{\alpha/2} \, \hat{se}$: Hard to guarantee that this is $(1-\alpha)$ CI

So, we'll do something akin to Student's t-distribution.

DKW inequality : for any $\varepsilon > 0$, for a fixed value of x,

$$P\left(\sup_x \left| F(x) - \hat{F}_n(x) \right| > \varepsilon \right) \leq 2e^{-2n\varepsilon^2}$$

Through this we can compute $(1-\alpha)$ CI. Equate RHS $2e^{-2n\varepsilon_n^2} = \alpha$ P that the two values differ should be less than α.

Critical epsilon : $\boxed{\varepsilon_n = \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}}$

$\hat{F}_n(x) + \varepsilon_n \not> 1$ $\quad \hat{F}_n(x) - \varepsilon_n \not< 0$: cannot be

$$\left\{ \begin{aligned} L(x) &= \max\{\hat{F}_n(x) - \varepsilon_n, 0\} \\ U(x) &= \min\{\hat{F}_n(x) + \varepsilon_n, 1\} \end{aligned} \right\}$$

These functions give us the confidence band

$$\boxed{P\left(L(x) \leq F(x) \leq U(x) \right) \geq 1-\alpha}$$

P that this band traps the true $F(x)$ for any value of x is $> 1-\alpha$.

EDF 3

Plug-in estimators

Statistical functionals : It's a fn T of the CDF which produces a real no.

$$T(F) \rightarrow R$$

Say, $\mu = T(F)$ We have $X_1, X_2 \ldots X_n$ iid rv

Plug-in principle : Once we have something like $\mu = T(F)$, we can also write, $\hat{\mu} = T(\hat{F}_n)$

A Statistical functional is linear if —

$\boxed{T(F) = \int \eta(x)\, p(x)\, dx}$ i.e. if there exists a fn $\eta(x)$ such that the integral can be written.

$p(x) \downarrow$ pdf

If we take 2 CDFs — $aF + bG$ and we apply statistical functional —

$$T(aF + bG) = \int \eta(x) \left(a p(x) + b q(x)\right) . dx$$

$$= a \int \eta(x) . p(x) . dx + b \int \eta(x) . q(x) . dx$$

$$= a\,T(F) + b\,T(G) \; *$$

So its linear $\because$ T of linear comb. is a linear comb. of Ts.

$\mu = \int x\, p(x) . dx$ here $\eta(x) = x$

$\sigma^2 = \int (x-\mu)^2\, p(x) . dx$ $\quad \eta(x) = (x-\mu)^2$

Say, X is a discrete, uniform rv which has sample size $\because \{8, 9, 10\}$

Say we draw 4 times & get : $\{9, 8, 10, 8\}$

We can construct a $\hat{F}_n(x)$.

Say we then construct a rv Y with CDF $\hat{F}_n(x) \Rightarrow P(Y=8) = 1/2$, $P(Y=9) = 1/4$, or $P(Y=10)$ "

Even if X is cont, Y is always discrete — at most it can take n values.

pmf of Y : $\boxed{P_Y(y) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = y)}$

We can find expectation value based on $\hat{F}_n$

$$\left\{ T(\hat{F}_n(x)) = \frac{1}{n} \sum_{i=1}^{n} \eta(x_i) \right\}$$

$\therefore$ We have our plug-in estimator $\hat{F}_n(x)$ that can be used

To replace F by $\hat{F}_n$ in T(F), instead of pdf, we use a r.v. Y whose CDF is exactly $\hat{F}_n$. Y is always discrete

$T(F) = \mu = \int x \, p(x) \cdot dx$ $\quad \eta(x) = x$. Here, plug-in estimator is,

$$T(\hat{F_n}) = \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n \quad \text{sample mean}$$

Through this we can construct plug-in estimator for various quantities. Say, variance

$$\sigma^2 = \int x^2 p(x) \cdot dx - \left(\int x p(x) \cdot dx \right)^2$$

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^2 = \frac{1}{n} \sum X_i^2 - \frac{1}{n^2} \sum_i \sum_j X_i X_j$$

$$= \frac{1}{n} \sum X_i^2 - \frac{1}{n} \sum_i X_i \left(\frac{1}{n} \sum_j X_j \right)$$

$$= \frac{1}{n} \sum X_i^2 - \frac{1}{n} \sum_i X_i \bar{X}_n = \frac{1}{n} \sum_i \left(X_i^2 - X_i \bar{X}_n \right)$$

$$\hat{\sigma^2} = \frac{1}{n} \sum_i \left(X_i^2 - 2 X_i \bar{X}_n + X_i \bar{X}_n \right) = \frac{1}{n} \sum_i \left(X_i^2 - 2 X_i \bar{X}_n \right) + \bar{X}_n \frac{1}{n} \sum_i X_i$$

$$= \frac{1}{n} \sum \left(X_i^2 - 2 X_i \bar{X}_n \right) + \frac{n \bar{X}_n^2}{n} = \frac{1}{n} \sum_i \left(X_i - \bar{X}_n \right)^2$$

$1/n$ common, n vanishes inside Σ

$$\therefore \hat{\sigma}^2 = \frac{1}{n} \sum_i (X_i - \bar{X}_n)^2$$

$\therefore$ slightly biased, but bias vanishes when $n \to \infty$

Bootstrap 1

Sample : $X_1, X_2 \dots X_n \sim F$ iid rv

Statistical functional : $\theta = T(F)$

Estimator : $\hat{\theta}_n = g(X_1, X_2 \dots X_n)$

Sampling distribution is the distribution of $\hat{\theta}_n$

Standard error se_n is the std dev of this sampling distributⁿ

$1 - \alpha$ CI for θ : $\hat{\theta}_n \pm z_{\alpha/2} \hat{se}_n$ (Assuming $\hat{\theta}_n$ is normally distributed)

40

Estimation of std error

For $\bar{X}_n$, $se_n = \frac{\sigma}{\sqrt{n}}$ $\Rightarrow$ $\hat{se}_n = \frac{\hat{\sigma}_n}{\sqrt{n}}$

Estimation becomes difficult when
- $T(F)$ is complicated
- $T(F)$ is not linear

→ $T(F)$ is complicated due to skewness - a measure of its asymmetry

$$K_3 = \frac{1}{\sigma^3} \int (x-\mu)^3 \, p(x) \cdot dx$$

Plug in estimator: $\hat{K}_3 = \frac{1}{n\sigma^3} \sum_{i=1}^{n} (X_i - \bar{X}_n)^3$

Computing the variance of $\hat{K}_3$ is difficult.

→ Median

It's the value x such that $F(x) = 0.5$

Let F and G be two CDFs : $H(x) = aF(x) + (1-a)G(x)$

If $T(F)$ denotes the median, then T is not linear

Consider the estimator:

$$\hat{M} = \begin{cases} Y_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1}\right) & \text{otherwise} \end{cases}$$

where Y_j represents the sample X_i sorted in ascending order.

Computing variance of $\hat{M}$ (ie constructing a CI) is difficult

Bootstrap conf. interval : $\hat{\theta}_n \pm z_{\alpha/2} \, \hat{se}_{boot}$

Bootstrap 2

When its hard to estimate variance of sampling distribut

bootstrap can be used. Its one of the 'resampling methods' that samples from $\hat{F}_n$.

The method estimates the sampling distribution (sam d) of $\hat{\theta}_n$ and its sam variance $V_{boot}(\hat{\theta}_n)$ is then taken as estimate of variance of sampling dist.

$$\boxed{\hat{se}_n = \sqrt{V_{boot}(\hat{\theta}_n)}}$$

can be estimated with large accuracy ∵ N can be huge but step 1 depends on n

Step 1 : Replace $V_F(\hat{\theta}_n)$ by $V_{\hat{F}_n}(\hat{\theta}_n)$ – assuming underlying dist. is EDF, not CDF

Step 2 : Estimate $V_{\hat{F}_n}(\hat{\theta}_n)$ as follows –

a) Draw N samples, each of size n from $\hat{F}_n$

b) Bootstrap sample can be drawn from $\hat{F}_n$ by * drawing n values from original sample with replacement. *

c) Compute $\hat{\theta}_n^*$ of $\hat{\theta}_n$ for each of these samples

$$\hat{\theta}^*_{n,j} = g(x^*_{1,j}, \ldots x^*_{n,j})$$

d) $$\hat{se}_{boot} = \sqrt{V_{boot}(\hat{\theta}_n)} = \sqrt{\frac{1}{N}\sum_{j=1}^{n}\left(\hat{\theta}^*_{n,j} - \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}^*_{n,i}\right)^2}$$

Bootstrap sometimes fails

An estimated sam d should converge to true sampling distribution as $n \to \infty$

Let $x_1, \ldots x_n \sim U(0, \theta)$ $\quad \hat{\theta}_n = \max\{x_1, x_2 \ldots x_n\}$

$$\lim_{n\to\infty} P(\hat{\theta}^*_n = \hat{\theta}_n) = 1 - \frac{1}{e} \approx 0.632$$

Refer Pg. 3 of notes

The bootstrap maximum will be same as sample maximum ~~ab~~ around 60% of the time

Since the sampling dis is continuous, the P should have been 0.

So, here bootstrap fails

4/6

Parametric Inference 1

Here, CDF is known to us – the shape is known but parameters that control the shape of F are unknown.

Eg: estimating mean, variance of underlying Gaussian from the data

Method of moments (MoM) data

If we have ~~moments~~, we can calculate the sample moments & if we know the form, we can estimate sample moments.

Suppose F contains k different parameters –

$\alpha_j(\theta) = E[X^j]$ for $j = 1, 2, \dots k$

MoM estimator is defined as –

$$\boxed{\alpha_j(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^j} \Big\} \text{ Sample moment}$$

There are k equations in k unknowns, so we can simultaneously solve for estimators of all k parameters.

4 min

Two types of parameters – Parameters of interest – unknown – k

Nuisance parameters – known / don't care

$\int x^j p(x; \theta).dx = \alpha_j(\theta)$ x is integrated out, so LHS is a fn of θ

Say we have $X_1, X_2 \dots X_n \sim N(\mu, \sigma^2)$ & we've to construct estimators for μ, σ^2

$\mu = \int x . p(x)^{\theta} dx$: First moment

$\mu^2 + \sigma^2 = \int x^2 \, p(x, \theta) \, dx$: Second moment

$$V(x) = E[X^2] - (E[X])^2$$

$$\sigma^2 = E[X^2] - \mu^2 \Rightarrow \therefore E[X^2] = \sigma^2 + \mu^2$$

From this, we can get MoM estimators -

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\hat{\mu}_n^2 + \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$ i.e. $\bar{X}_n = \hat{\mu}_n$ Refer Plug-in estimators part Pg. 39

$$\therefore \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2$$

Biased estimator - negligible for $n \to \infty$

These parameters are not-distribution-free - they're defined according to shape of distribution

In this case it was general.

Properties : MoM estimators are consistent : $\hat{\theta}_n \xrightarrow{P} \theta$

They are usually unbiased when $n \to \infty$

They are asymptotically normal : $\frac{\hat{\theta}_{n,j} - \theta_j}{\sigma_j} \rightsquigarrow N(0,1)$

4/6

Parametric Inference 2

Maximum Likelihood

You know the form of distribution & interested in a particular parameter - most likely value of θ is the one that maximizes the probability of generating the observed sample

44

Likelihood fn : $\boxed{\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p(x_i, \theta)}$ $= p(x_1, x_2 \dots x_n | \theta)$ $\because x_i$ are i.i.d

pdf - we're maximising the pdf instead of P

log-likelihood fn : $l_n(\theta) := \log \mathcal{L}_n(\theta)$

So we maximise $l_n(\theta)$ wrt θ

$$\boxed{l_n(\theta) = \sum_{i=1}^{n} \log(p(x_i, \theta))}$$

$l_n(\theta)$ and $\mathcal{L}_n(\theta)$ have the same global maxima because logarithms is a monotonically increasing function. Also, using l_n makes it easier when dealing with exponential, gamma or guassian distribution. Also using l_n keeps the numbers in a manage-able range

Consider $X_1, X_2 \dots X_n \sim F \quad N(\mu, \sigma^2)$

$$\mathcal{L}_n(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

$$l_n(\mu, \sigma^2) = -\log(2\pi)^{n/2} - n\log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Here, the constants don't play a role in maximisation they become 0 when differentiated

★ $$\frac{d\, l_n(\mu, \sigma^2)}{d\mu} = \frac{-1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \hat{\mu})(-1) = 0$$ (To maximise) $\sigma \neq 0$

$$\Rightarrow \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$$ Same estimator

$$\frac{d\, l_n(\mu, \sigma^2)}{d\sigma} = \frac{-n}{\hat{\sigma}} + \frac{\cancel{2}}{\cancel{2}\hat{\sigma}^3} \sum_{i=1}^{n} (x_i - \hat{\mu}_n)^2 = 0$$

$$\Rightarrow \hat{\sigma}_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_n)^2$$

★ Same for MoM estimator. ★

Not true ∀ estimations

Properties of Max. likelihood estimators (MLEs) –

- They are consistent : $\hat{\theta}_n \xrightarrow{P} \theta$
- They're also asymptotically normal : $\frac{\hat{\theta}_n - \theta}{se} \rightsquigarrow N(0,1)$

They're unbiased when $n \to \infty$

We can use parametric bootstrap to construct CI

4/6

Hypothesis testing 1

It's the 3rd type of problem.

Hypothesis : a statement that claims something is true.

Applicable when we need to take a decision based on data.

Null hypothesis (H_0) : default position – assumes nothing special is happening

Alternative hypothesis (H_1) : Opposite of H_0 – claims that something special is happening

If data contains sufficient evidence to support H_1, we reject H_0.

We ask how rare is the observed data if we assume that H_0 is true

Important Points –

- Randomness forbids a deterministic decision.
- Multiple observations are required
- Sample must be representative of the population.

Hypothesis Testing 2

H_0 : Person cannot predict coin tosses better than random guesses / $\mu_{chennai} = \mu_{india}$

H_1 : Person can predict. / $\mu_{chennai} \neq \mu_{india}$

~ Two-tailed test ~

If H_1 was : $\mu_{chennai} > \mu_{india}$

H_0 : $\mu_{chennai} \leq \mu_{india}$

One-tailed test

H_0 & H_1 should be opposites

Also, this is a Right-tailed test ∵ H_1 μ_{ind} μ_c

Hypothesis Testing 3

Rejecting H_0

The margin at which it becomes really improbable that guesses are random, then we can reject H_0.

Say n coin tosses – k or more tosses should be guessed correctly. $p = 0.5$

$$p(k) = \sum_{i=k}^{n} {}^nC_i \, p^i (1-p)^{n-i}$$: Prob. of guessing $> k$ tosses correctly (Randomly).

We should keep this p small, say $p(k) < 10^{-5}$ (rare event)

So we can be fairly sure that the events didn't occur by chance

$p(k) < 10^{-5}$ is the level of significance (similar to concept of α).

For this, $k = 72$ for $n = 100$

⇒ If a person predicts 72 or more coin tosses, then we reject H_0 & accept H_1 that the person has supernatural abilities

If $p(k) < \alpha$ then the result is statistically significant. ⇒ also called test of significance

There are many hypothesis tests – we'll focus of μ of the sample. i.e. avg of the underlying distribution.

Hypothesis testing 4

Z- test : $z \sim N(0,1)$

H_0 : $\mu \geq 200$ g

H_1 : $\mu < 200$ g $\quad \bar{x}_n = \frac{1}{n} \sum_i x_i$

We take a representative sample

Say we get $\bar{x}_n = 198.02$ g

If we take multiple samples, we'll get a gaussian sampling distribution from which.

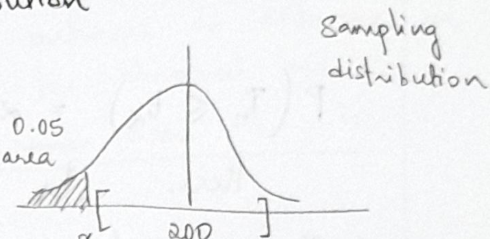
$$\hat{se}_n = \frac{\hat{\sigma}_n}{\sqrt{n}} \quad \text{where} \quad \hat{\sigma}_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

delta degree of freedom ddof = 1 for unbiased estimator

We get $\hat{\sigma}_n = 4.7$ $\quad \hat{se}_n = 0.706$

$\hat{se}_n$: std dev of sample distribution

If we can say that our value falls outside of this interval then its significant i.e. $\alpha = 0.05$



Interval that contains 95% of data

But ours is a left-hand test

$$\boxed{P\left(\bar{x}_n \leq x_\alpha\right) = 0.05}$$

$$P\left(\frac{\bar{x}_n - 200}{\hat{se}_n} \leq \frac{x_\alpha - 200}{\hat{se}_n}\right) = 0.05$$

$$P\left(z \leq z_\alpha\right) = 0.05$$

So $Z_\alpha = -1.643$ for our data

$$\frac{\bar{x}_n - 200}{\hat{se}_n} = -2.80 < -1.643 = Z_\alpha$$

If the true mean was 200, there's less than 5% chance that we get a $\bar{x}_n$ less than 198.357 by chance. Our $\bar{x}_n = 198.02$. So we can reject H_0 and accept H_1.

(48)

For 2-tailed test, we consider $z_{\alpha/2}$

where $P(\bar{X}_n \leq -z_{\alpha/2}) + P(\bar{X}_n \geq z_{\alpha/2}) = 0.05$

For gaussian, $z_{\alpha/2} = -z_{\alpha/2}$

Hypothesis testing 5

Mileage — 40 km/L

$H_0 : \mu \geq 40$

$H_1 : \mu < 40$

Say, $\bar{X}_n = 38.63$ $n = 5$

$\dfrac{\bar{X}_n - \hat{\mu}}{se}$: won't have normal distribution

$T_n = \dfrac{\bar{X}_n - \mu}{\hat{\sigma}_n / \sqrt{n}}$: Student's t-distribution. From this, we can proceed with z-test.

$P(T_n \leq t_\alpha) = \alpha = 0.05$

Here, $t_\alpha = -2.132$

If $t < t_\alpha$, then we should reject H_0

Here, $T_n = \dfrac{38.63 - 40}{3.383/\sqrt{5}} = -0.906$

$\Rightarrow T_n > t_\alpha$

Hence, we cannot reject H_0

HT 6

Two imp. points —

- Rejecting H_0 is a stronger decision than retaining it
- Rejecting H_0 refers to population while statistically significant result refers to the sample

Types of errors –

Type I : Rejecting H_0 when H_0 is true (more serious than II)

Type II : Retaining H_0 when H_1 is true

Size: Prob. of making I error $= \alpha$

II error $= \beta$

P. of not making type II error : $(1-\beta)$ is the power of the test.

P-values

An alternative to rejecting H_0, no level of significance is specified

los should be specified b4 collecting data

p-value : P of observing an outcome as extreme or more extreme as the observed one

Here, it only reports how ~~rat~~ rare H_1 is.

$$\alpha_j(\theta) = \int x^j \, p(x_i, \theta) \, dx$$

$$\alpha_j(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} x_i^{j}$$

$$l_n(\theta) = \sum_{i=1}^{n} \log\left[p(x_i, \theta)\right]$$ — minimise this