3Y03 - 3J04 PROBABILITY & STATISTICS FOR (CIVIL) ENGINEERING

C01 - Lecture 20

Last time: PROBABILITY PLOTS

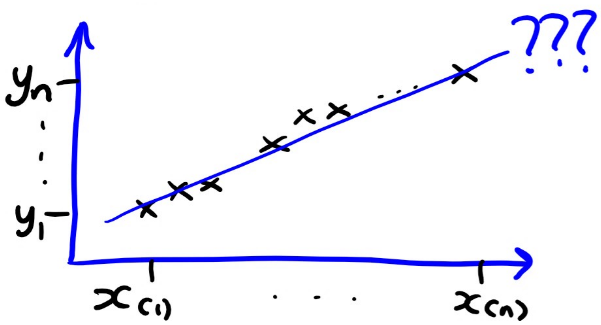
Given sample $\{x_1, \dots, x_n\}$ ↳ use this to check a hypothesis about underlying probability distribution

Our hypothesis: X has pdf $f(x)$

— Reorder sample: $\{x_{(1)}, \dots, x_{(n)}\}$

— Find $\{y_1, \dots, y_n\}$ with

$$F(y_j) = P(X \le y_j) = \frac{j-0.5}{n} \quad \text{for each } j:$$

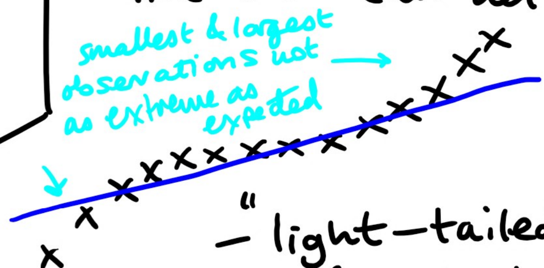


Textbook talks only about normal distribution where $F(x) = \Phi(x)$ & in that case y_j-values are "z_j-values" from normal table.

There is also "normal probability paper" where we plot $100\left(\frac{j-0.5}{n}\right)$ against $x_{(j)}$ but y-axis is distorted so that good hypothesis → straight line

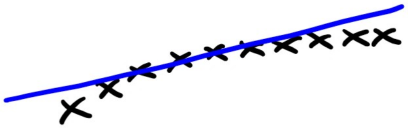
No need to find y_j (or z_j) values.

Even if we don't get a straight line we can detect features of distribution relative to hypothesis



— "light-tailed" e.g. uniform instead of normal

— "heavy-tailed"



– "right-skewed" → both smallest & largest observations are bigger than expected.

(Below line : bigger than expected;
Above line : smaller " ")

---

7. Point Estimation of Parameters & Sampling Distributions (7.1, 7.2)

Statistical Inference → Parameter Estimation
→ Hypothesis Testing

↓

Drawing conclusions about underlying population based on sample

Parameter : any function of underlying distribution
e.g. population mean μ,
" variance σ^2

↑ Usually by default called θ

→ Want to estimate using sample $\{x_1, \dots, x_n\}$

→ Remember : observations are considered to be independent random variable $X_1, \dots, X_n$ with same distribution (population)

(& sample $\{x_1, \dots, x_n\}$ represents one set of possible values of $\{X_1, \dots, X_n\}$, with some samples more likely than others, depending on underlying distribution)

(↖ a "random sample")

A <u>statistic</u> is a function of the sample. $h(X_1, \dots, X_n)$

Any function of sample is also as a random variable

e.g. $\bar{X}$ = sample mean

S^2 = sample variance

Each statistic has its own prob. distribution, called a <u>sampling distribution</u>.

Given a parameter θ an <u>estimator</u> $\hat{\Theta} = h(X_1, \dots, X_n)$ is a statistic used to "guess" (<u>estimate</u>) the parameter θ

some function

e.g. sample mean $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ as estimator for μ

sample variance $S^2 = \frac{1}{n-1}\left(X_1^2 + \dots + X_n^2 - n\bar{X}^2\right)$ as estimator for σ^2

After we select the sample $\{x_1, \dots, x_n\}$ & calculate a value of the estimator $\hat{\Theta}$, that value is referred to as $\hat{\theta}$, the <u>point estimate</u>.

e.g. $\bar{x}$ as point estimate for ($\bar{X}$ i.e. for) μ

s^2 " " " " (S^2 " ") σ^2

Other interesting parameters:

→ proportion p of population in some class of interest

Estimate: $\hat{p} = \dfrac{\text{\# items in sample in class}}{n}$

→ difference in means $\mu_1 - \mu_2$ of 2 populations

Estimate: $\bar{x}_1 - \bar{x}_2$

(μ_1: mean of pop^n 1
μ_2: mean of pop^n 2)

→ difference in proportions $p_1 - p_2$ of 2 populations

Estimate: $\hat{p}_1 - \hat{p}_2$

(p_1: proportion of pop^n 1
p_2: proportion of pop^n 2)

Not always one choice for an estimator e.g.
sample mean & median used to estimate pop. mean.

Textbook often assumes that pop. r.vs X_is are normally distributed (hmm)

But notice we already saw if $X_i \sim N(\mu, \sigma^2)$
then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

In general: This says: this is a safe assumption about $\bar{X}$ in general, even when X_is NOT normally distributed.

Central Limit Theorem If $\{X_1, \ldots, X_n\}$ is random sample & underlying pop. mean is μ & pop. variance is σ^2

(in particular σ^2 finite) then as n gets big ($n \to \infty$)

the limiting form of $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0,1)$

i.e. $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$ as $n \longrightarrow \infty$.