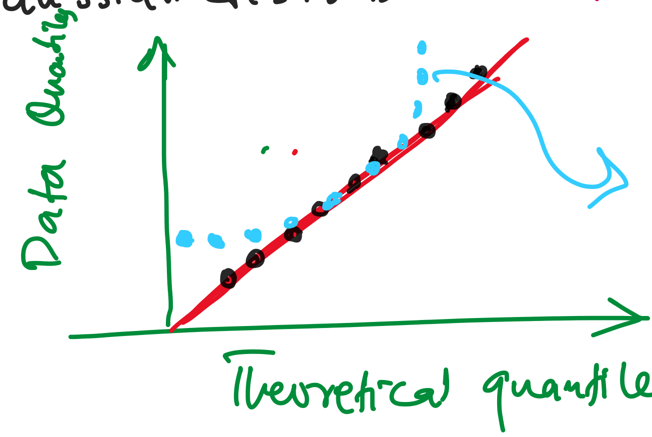
Monday, 7 February 2022 11:49

Sample: $\{x_1, x_2 \ldots\ldots x_n\}$

- underlying distribution ?
- My Guess is that this sample comes from the Gaussian distribution. ⇒ Hypothesis.



Quantile-Quantile plots.

Then the underlying dist. func is not Gaussian.

Data-Based Estimates of Descriptive Statistics

if you $P(x)$ probability dist.

Mean: $E(x) = \sum x_i P(X = x_i)$ & $E(x) = \int_{-\infty}^{+\infty} x P(x) dx$.

Var: $Var(x) = \sum_{r=i} E[(x - x_i)]$ & $E(x^2) = \int_{-\infty}^{+\infty} x^2 P(x) dx$.

$$Var(x) = E(x^2) - (E(x))^2 = E(x^2) - \mu^2.$$

Suppose Now you don't know the underlying distribution function.

$$\{x_1, x_2 \ldots x_n\}$$

Estimators → Statistics.

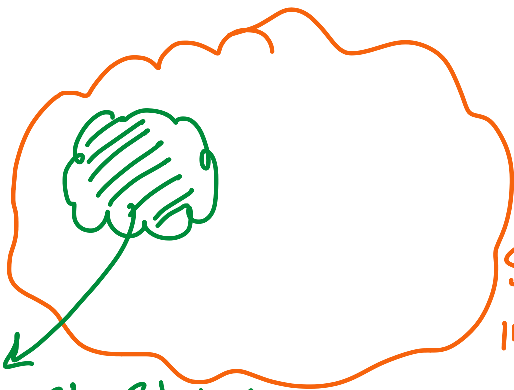
biased

Not biased.

Estimator should be unbiased

→ - Sample Statistics

- Population Statistics. → You know the underlying statistics → Mean & variance will be the "true" mean & vari.



"Sample space?
"population sample" Ω

Sample Statistics.

N Measurements. x_i. $i = 1, 2, \dots N$ $\{x_i\}$

Sample Mean: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

Sample Standard deviation: $\bar{s} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$

$N-1$

The reason for the N-1 term istead of the naively expected N is due to the fact that $\bar{x}$ is also determined from the data. → Bessel's Correction.

True standard deviation is σ^2

Sample standard deviation is s^2.

for a Gaussian dist. the underestimation varies from 20% for N=2 to 3% for N=10.

less than 1% for N > 30.

For the large sample size we reduce the "underestimation" to less than 1%.

→ What do you mean by large Number?

→ It depends on your particular case and level of accuracy.

but generally the transition N=10 to N = 100.

"Massive" data set the transition may occur at N of the order of Million or even billion,

$\underbrace{(\bar{X}, s)}_{\text{estimators from the sample.}}$ $\underbrace{(\mu, \sigma)}_{\text{Truth from the population.}}$

These estimators have a variance and a bias

Mean Squared errors (MSE)

$MSE = V + bias^2$.

$MSE = V + bias$

Bias: expectation of the difference b/w the estimator and the truth value.

$N \to \infty$, $V \to 0$, bias $\to 0$

Consistent estimators

- Central limit theorem
- law of large Numbers.

(1) Law of large Numbers:

let $X_1, X_2, \ldots$ be a sequence of independent random variables with a common distribution and $E(|X_1|) < \infty$ then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu = E[X_1]$$

as $n \to \infty$; for all $\varepsilon > 0$

$$P(|\bar{X}_n - \mu| > \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

Sample mean $\bar{X}_n$ gives a good approximation of the population mean $\mu = E(X_1)$ when N is large.

Sample Mean → Random Quantity.
Population Mean → fixed Number - (Par).

(2) Central limit Theorem :

let $X_1\, X_2 \dots$ be a sequence of i.i.d. random variables with mean $\mu = E(X_1)$ and finite variance $\sigma^2 = E((X_1 - \mu)^2) > 0$ Then

$$P\left(\sqrt{n}\,(\bar{X}_n - \mu) < x\,\sigma\right) \rightarrow \Phi(x)$$

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt \quad \text{and} \quad \phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}.$$

Distribution approaches the Gaussian dist for large Sample Size.

Next lecture:

- Examples
- Discuss some discrete &

Continuous ~~d~~istribution functions.