STAT 231

October 5, 2016

Syllabus $\leq$ Friday, Oct 7.

Tutorial : 3-30 → Surya

STP 105.

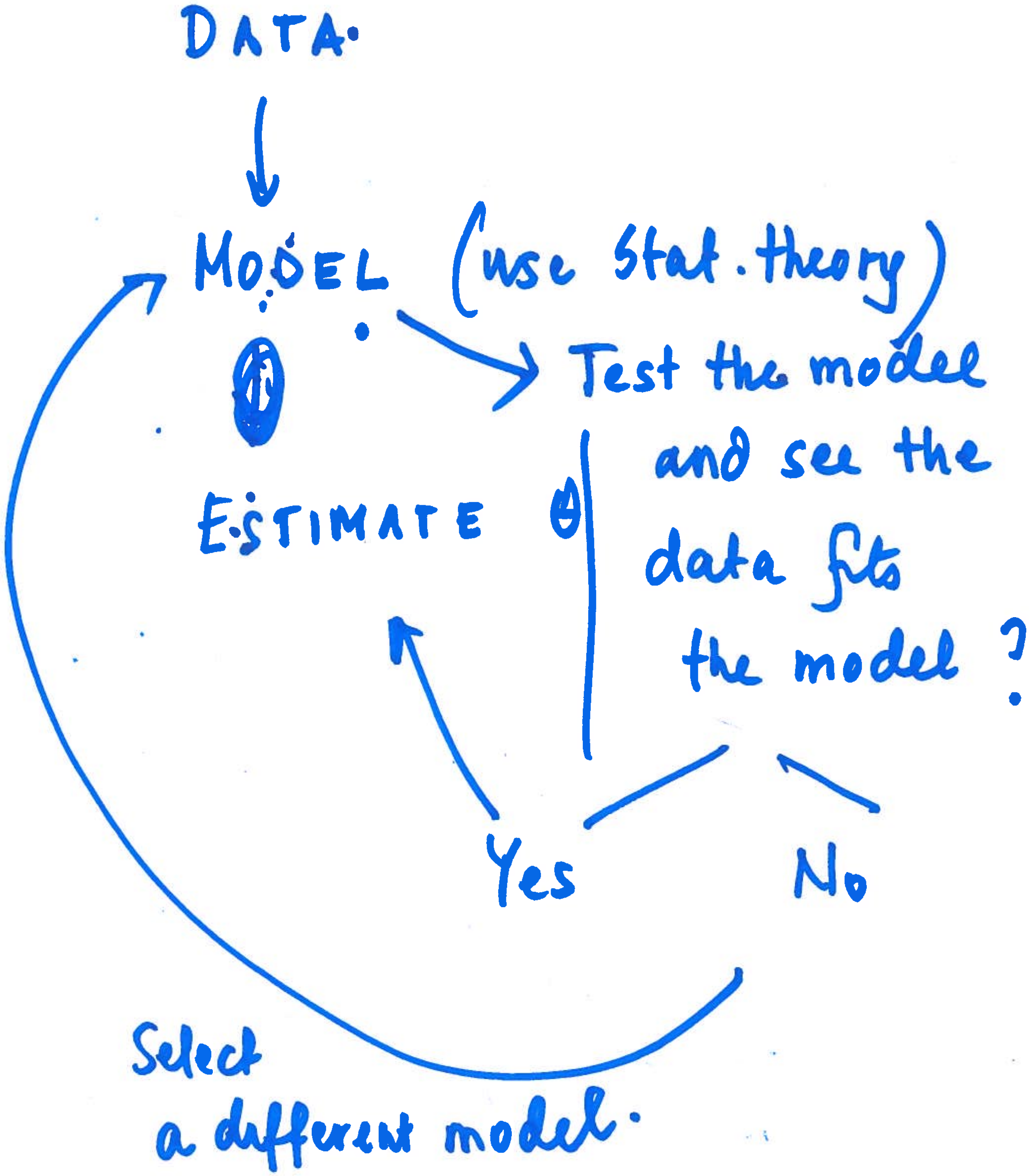
6-00 → Cyntha.

DC

Roadmap

MODEL SELECTION METHODS

- Graphical methods
  - Relative Frequency
  - cdf
  - Q-Q-plot
  - Run charts.
- Numerical methods: (Ch 7)
  - Observed versus Expected frequence's


DATA.
MODEL (use stat. theory)
Test the model and see the data fits the model?
ESTIMATE θ
Yes
No
Select a different model.

How to select the "right" model ?

Graphical ways.

- Relative frequency histogram method.

Example: Suppose the model chosen is Gaussian (μ, σ).

Draw the density histogram

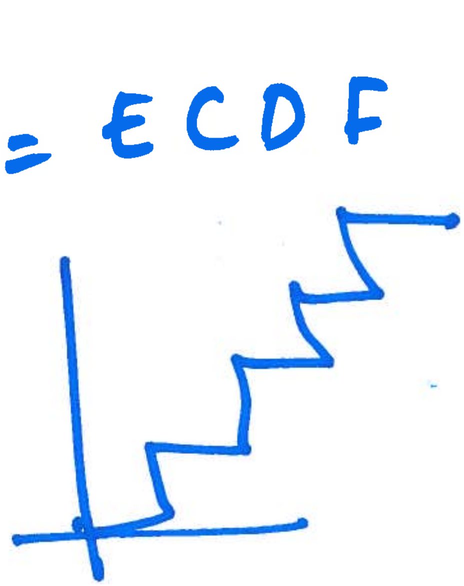
D.

The parameters of the superimposed distribution is chosen by the method of max. Likelihood.

· Compar· the empirical cdf with the theoretical cdf.

$$\{y_1, \ldots y_n\}$$

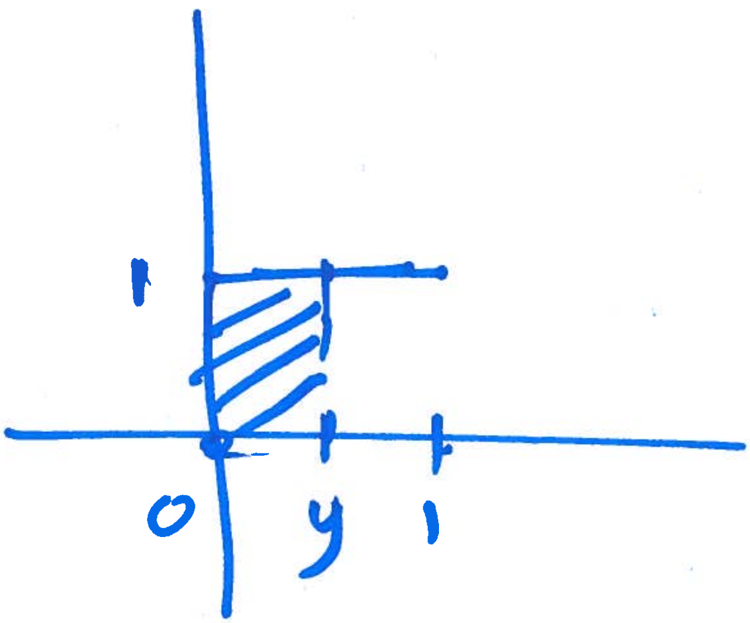
$$\hat{F}(y) = \frac{\#\text{ of obs} \le y}{n} = ECDF$$



$$F(y) = P(Y \le y)$$

Example $\{y_1, \ldots y_n\}$ DATA

Model: $Y_i \sim \text{Uni}(0,1)$ independent.

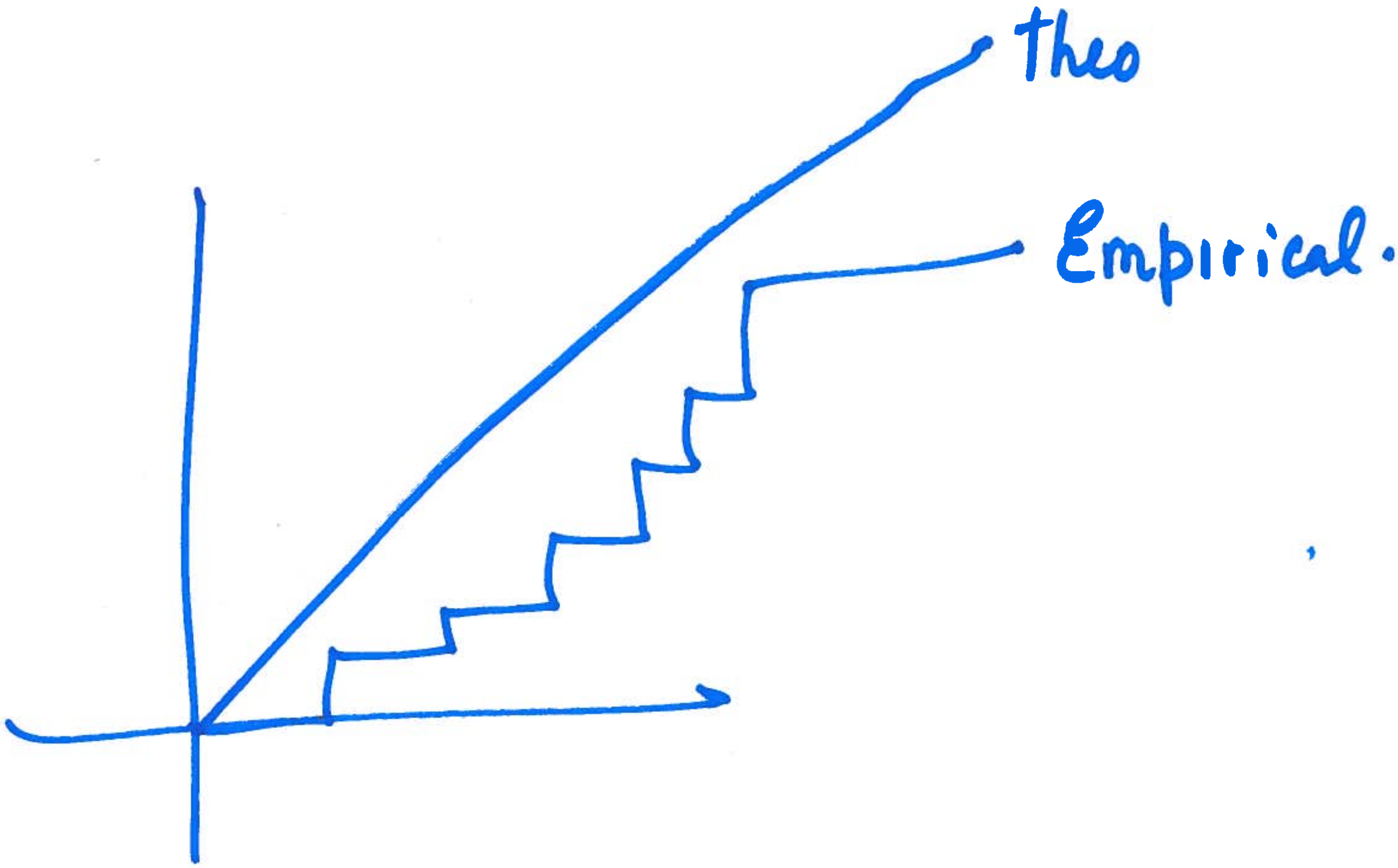


$F(y)$

$P(Y \leq y) = y$

cdf of $U(0,1) = F(y) = y$

$= 45^\circ$ line



Superimpose the theoretical cdf on the empirical cdf and compare their shapes.

The Q-Q-plot

Typically, the Q-Q plot is used to check for Gaussian.

Q

$$\{y_1, \dots y_n\}$$

arranged.

$$\Downarrow$$

$$\{y_{(1)}, y_{(2)} \dots y_{(n)}\}$$

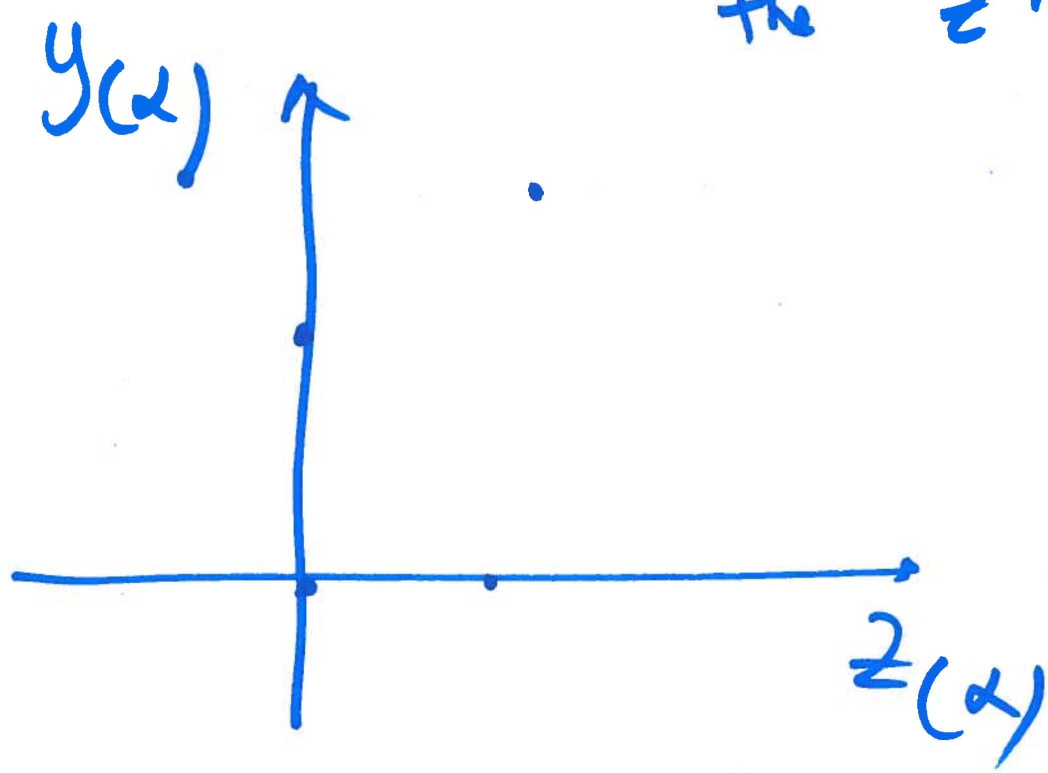
The Q-Q plot: plots the sample quantiles against the quantiles of the standard normal distribution.

$$Z \sim \mathcal{G}(0, 1)$$

$QQ \to (y_{(\alpha)}, z_{(\alpha)})$

$y_{(\alpha)} = \alpha^{th}$ quantile of the data set

$z_{(\alpha)} = \alpha^{th}$ quantile of the $Z \sim G(0,1)$



$y_1, \ldots \ldots y_{101}$

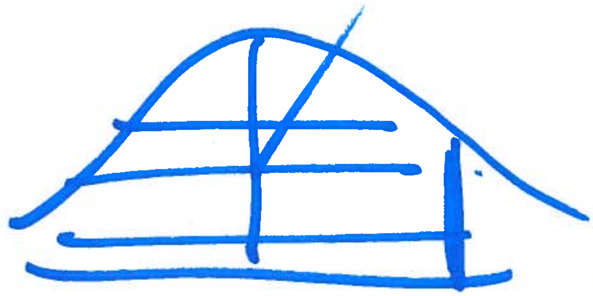
Median $\longrightarrow$ y-value $\rightarrow$ median of the data

z-value $\longrightarrow$ 0

95^{th} percentile $\rightarrow$ y-value $\rightarrow$ 95^{th} %ile of the data

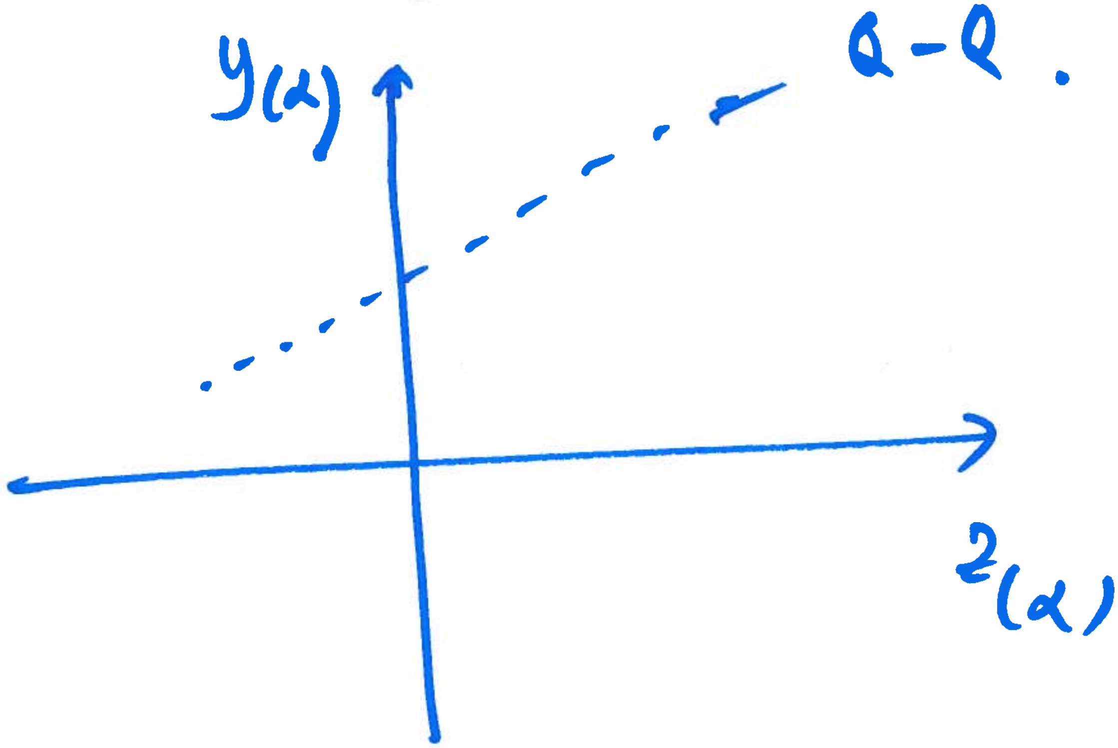
0.95

z-value : 95^{th} percentile of z



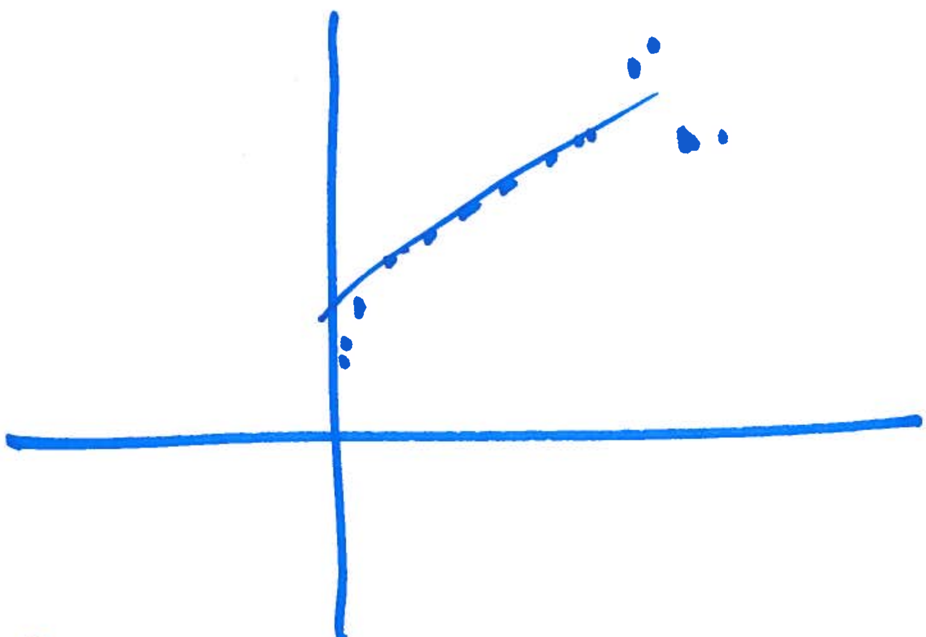
$a = 1.64$

If the Q - Q plot is a straight line, that is strong evidence that the data is Gaussian.



Notes about the Q-Q plot

- Even if the data is actually Gaussian the extreme points are typically off the line



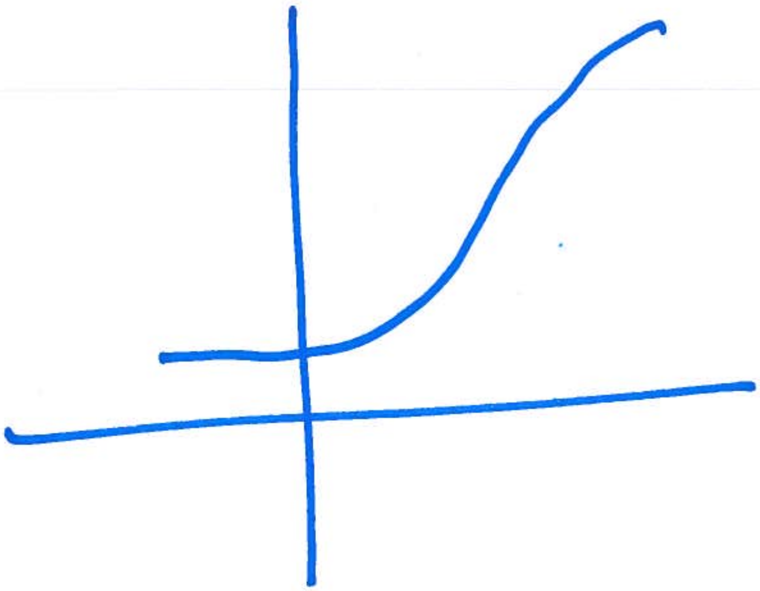
(Why?)

- Apply the Q-Q plot to other distributions. (Exponential).

- Suppose the distribution is asymmetric.

skewed distribution

(Try drawing Q-Q plot of Exponential against Normal)



Typically, with a higher kurtosis, the Q-Q plot looks like a $\int$.

Numerical methods of model checking

$\overset{Y}{\cancel{X}}_i$ = # of accidents on a highway in a 1 hour period during rush-hour.

$\{y_1, \dots, y_n\}$

- $Y_i \sim \text{Poi}(\mu) \rightarrow$ SUGGESTED MODEL.

DATA

# of acci	frequency	Expected frequencies
0	10	
1	20	
2	40	
3	20	
4	10	
5	0	
6	0	
⋮	⋮	

Calculate $\hat{p}$

$$\hat{P}(Y=0) = \frac{e^{-\hat{\mu}}\,\hat{\mu}^{0}}{0!} = 0\%$$

$$\cancel{\times n}$$

Expected frequency of 0 $\rightarrow$

$$\frac{e^{-\hat{\mu}}\,\hat{\mu}^{0}}{0!} \times n$$

n = sample size

We compare the observed with the expected frequencies and see whether they are "close enough"

<u>Continuous</u>

$Y_i \sim Exp(\mu)$

$y_1, \ldots, y_n$.

Data

	Freq
[0, 100]	20
[100, 200]	30
[200, 300)	20
≥ 300	10