Introduction to Q-Q Plots

Purpose: compare data with given dist'n.

Common question: dist'n generates data?

Basis: for c.d.f. $F(x) = P\{X \le x\}$,

write $p = F(q)$ and $q = F^{-1}(p)$

where p = "probability"

and q = "quantile"

Q-Q plot: given p, plot "empirical q" vs. "theoretical q"

Show QQToyEg1.ps, data cdf in red, theoretical in blue, black dashed lines are p's, data q's thin red, theoretical q's thin blue. Bottom plot uses theoretical q's as x's, and data q's as y's. Compare to green 45 degree line.

Simulated examples

I. Weibull(1) (exp.) data:

Show EGQQWeibull1.ps, different panels are different theoretical dist's.

a. Theory = normal, log-norm: poor fit.

Upper and lower left, note Q-Q curves away from 45 degree line

a. Theory = Pareto, Weibull: good fit

Upper center, right

b. understand "sampling error" through blue overlays (100 sims, theor. dist.)

Inside blue envelope is "good fit".

c. log scale "highlights different regions of dist'n".

Bottom center, left

Simulated examples

II. Pareto(1.5) data:

Show EGQQPareto1.5.ps

a. Gaussian, Weibull: terrible fit

Top, right and left

b. Pareto: fits, but “wide range”

Top center (because of extreme “outliers”)

c. Log scale: more useful view of data

Bottom, center

d. Log-N nearly fits

Bottom left

e. Weibull misses on log scale, too

Bottom right

Model Fitting

i.e. Choice of “parameters”

Mode I. Given parameters (trial and error)

Mode II: Parameter Estimation

a. Gaussian: $\hat{\mu} = \overline{X}$, $\hat{\sigma} = s$

b. Quantile Matching: choose for equality (i.e. cross 45° line) at given quantiles, q_1 & q_2

Used in above examples

HTTP Response Size Data

$n = 734{,}815$ values over one hour on the UNC $\leftrightarrow$ MCNC link

Distributional Shape I: "Body"

- "Smooth Histogram"

Top part of movie, SmithData1p3.mpg

- SiZer $\Rightarrow$ Bumps are "really there"

Bottom part of movie, and SmithData1p2.ps

- Log scale important

SmithData1p1.ps – massive skewness obscures other features

HTTP Response Size Data

Distributional Shape II: "Tail" (Q-Q plots)

a. Full Data vs. Pareto: estimated Tail Par: $\hat{\alpha} \approx 1.25$
   SmithData1p11.ps

b. 1^{st} 50,000 vs. Pareto: estimated Tail Par: $\hat{\alpha} = 1.5$, "bad fit farther out"
   SmithData1p21.ps

c. 1^{st} 50,000 vs. Pareto: given α, σ
   SmithData1p12.ps, looks much better

d. 1^{st} 50,000 vs. Weibull: terrible fit
   SmithData1p21.ps

e. Conclusion: not exactly Pareto, but not a bad model (for tails of dist'n).