Normal ranges determination with Quantile regression
Luc Esserméant, Sept. 2012
NCSC 2012, Potsdam

Contents

- Normal ranges or reference values definition and calculation
- Covariate-dependent reference limits

Normal ranges or reference values definition

- Statistical bases of reference values in laboratory medicine, HARRIS E.K., BOYD J.C., Dekker ed., 1995
- "...the conventional « reference range » is defined by a pair of numbers (the reference limits) that bound the central 95% of a collection of values(....). The word central means that 2.5% of the values lie above the upper limit and 2.5% below the lower limit."

- Notations: P_{\tau} is percentile \( \tau \% \)
  - reference ranges are noted \( P_{2.5} \) and \( P_{97.5} \)
  - Median is \( P_{50} \)

Example

- Reference values
- 95% of values
Calculation reference limits

- **Parametric approach**
  - If the distribution is known, directly or after transformation, percentiles are easily determined.
    - Example: For a normal distribution, $P = \text{mean} + z_{1-\alpha} \times SD$, where $z_{1-\alpha}$ is the percentile 1- of the normal distribution $N(0,1)$.
  - In case of deviation to Normality due to outliers, robust approaches are available, replacing mean and SD by robust estimators like resp. median and MAD or Sn or Qn or using M-estimators.

- **Non parametric approaches**
  - **Empirical**: i.e., 0.025(n+1)th and 0.975(n+1)th ordered values with interpolation. 3 methods in SAS.
  - **Spline**: determined from a smoothed representation of the cumulative distribution, obtained by joining a series of cubic polynomial segments.
  - **Kernel**: weighted average quantile.

Contents

- Normal ranges or reference values definition and calculation
- Covariate-dependent reference limits

Calculation covariate-dependent reference values

- **Parametric method**: polynomial regression with individual prediction intervals (See Royston, 1991)

- Non parametric methods of Healy (1988) and its modification by Pan (1990) allow to compute normal ranges with a continuous covariate using the following process:
  1. Cut the covariate in overlapping intervals
  2. Compute the normal ranges per interval
  3. Build a polynomial regression of normal ranges vs covariate

- Kernel approaches are also available (bivariate distribution, double kernel.)

- Quantile regression
Introduction to Quantile regression

Advantages
- Model directly the percentiles of interest
- No distributional assumptions
- Robust to outliers and heterogeneity
- Provide a complete picture of the distribution
- Easy to run in R and SAS PROC QUANTREG
- Easy to understand and explain!

Mean and median of univariate sample by optimization

Consider an univariate sample \( Y = y_1,y_2,..,y_n \)

Mean of \( Y \) is the central point that minimizes the arithmetic mean of the quadratic deviations

\[
\text{arg min}_b \sum (y_i - b)^2
\]

Proof
\[
0 = \frac{-\sum (y_i - b)^2}{\partial b} = \sum (y_i - b) = 2 \sum y_i - 2nb \implies b = \frac{\sum y_i}{n}
\]

Median is the central point that minimizes the arithmetic mean of the absolute deviations

\[
\text{arg min}_b \sum |y_i - b|
\]

Proof
\[
0 = \frac{\sum |y_i - b|}{\partial b} \implies \sum (y_i - b_+ - b_-) \implies b = \text{median}
\]

Percentile on univariate sample by optimization

The percentile \( P_t \) is the point that minimizes the weighted arithmetic mean of the absolute deviations

\[
\text{arg min}_b \sum \rho_t(y_i - b) \implies \text{arg min}_b \sum |y_i - b| - \sum (1 - \tau) |y_i - b|
\]

with "check function" \( \rho_t(u) = u(t - |u| < 0) \)

Proof
- If random variable \( Y \) has a distribution function \( F \), then the quantile \( Q_Y(t) \) is

\[
Q_Y(t) = F^{-1}(t) = \inf \{ y : F(y) \geq t \} = \inf \{ y : P(Y < y) \geq t \}
\]

- Check that the quantile \( Q_Y(t) = F^{-1}(t) \) is the solution of

\[
\text{arg min}_b \int (y - b)^2 f(y) dy + \int (y - b) - b f(y) dy
\]

- Solution obtained setting the derivative to 0:

\[
0 = (t - b) \frac{1}{2} \hat{f}(y) + \int (y - b) f(y) = \frac{1}{2} \hat{f}(y) + \int \hat{f}(y) + \int f(y) \implies \frac{1}{2} \hat{f}(y)
\]

\( \tau = F(b) \)

Quod Erat Demonstrandum
Linear regression quantile

- Consider a covariate $X = x_1, x_2, \ldots, x_n$.
- OLS regression estimates the linear conditional mean function $E(Y|X = x)$ by solving $\arg\min_x \sum (y_i - x_i' \beta)^2$.

Likewise, quantile regression estimates the linear conditional quantile function $Q(Q, X(\tau)) = x' \beta\tau$ by solving $\hat{\beta}(Y, X) = \arg\min_x \sum \rho(\tau - (y_i - x_i' \beta))$.

Unfortunately, it’s not everywhere differentiable, so standard numerical algorithms do not work and linear programming must be used. Simplex ($n \leq 5000$ & $p \leq 100$), interior point ($n > 5000$ & $p < 100$) and smoothing algorithms ($p > 100$) are proposed in SAS PROC QUANTREG (ALGORITHM option).

Example: growth chart

- A simple AR(2) model $Q_2(W_t) = \beta_0 + \beta_1 W_{t-1} + \beta_2 W_{t-2} + \beta_3 H_t$.
  - where $W_t$ is the current weight at time $t$.
  - $W_{t-1}$ and $W_{t-2}$ are two prior weights at time $t-1$ and $t-2$, respectively.
  - $H_t$ is the current height at time $t$.

Example: final body weight in females rats
Example: epididymides weight in males rats

![Graph showing epididymides weight in males rats]

Goodness of fit criteria (not in QUANTREG SAS 9.2)

\[ R^2_F = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} \]
\[ R^2_R = 1 - \frac{\sum (Y_i - \hat{Y}_i')^2}{\sum (Y_i - \bar{Y})^2} \]
\[ F \text{ denotes Full model and R reduced (empty) model} \]

- A natural analog of \( R^2 \) is
  \[ R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \hat{Y}_i')^2} \] with \( \hat{Y}_i \) and \( \hat{Y}_i' \) estimators from full and reduced models.

- Koenker (2005) suggested an adapted AIC (aAIC) where the likelihood is replaced by the empirical risk:
  \[ AIC_c = -2\text{Log} \left( \frac{1}{n} \sum (Y_i - \hat{Y}_i'')^2 \right) + 2p \]

other "adaptations" are proposed in the litterature, as for AICc, BIC.

Conclusion and perspectives

- Quantile regression allows to calculate Normale ranges with (and without) covariates, without hypotheses of distribution or even independence, without a priori choice of a kernel or a bandwidth.

- Quantile regression is easy to run in SAS 9.2, and future version (upcoming SAS/STAT 12.1) will propose wonderfull new tools:
  - The new QUANTSELECT procedure for quantile regression model selection works similarly to the GLMSELECT procedure. Selection methods include forward, backward, stepwise, and LASSO. PROC QUANTSELECT uses variable selection criteria such as AIC, SBC, and AICC.
  - The new QUANTLIFE procedure performs quantile regression for censored data.

References (1/2)

- Univariate sample

- With a covariate
Quantile regression

- Chen Colin, Growth Charts of Body Mass Index (BMI) with Quantile Regression, SAS Institute Inc. Cary, NC, U.S.A.
- Wei Ying, Perre Anneli, Koenker Roger and He Xuming, Quantile Regression Methods for Reference Growth Charts, Statistics in Medicine, 2006; 25: 1369-1382.

References (2/2)

Back-up

Empirical estimation in SAS

- You want the percentile \( \tau \) of an univariate ordered sample \( X = x_1, x_2, ..., x_n \).
- Note \( n = j + g \), with \( j \) integer and \( g \) fractional. For example,
  - if \( n=100 \) and \( \tau = 5\% \), \( j=5 \) and \( g=0 \)
  - if \( n=50 \) and \( \tau = 5\% \), \( j=2 \) and \( g=0.5 \)
- SAS proc univariate proposes the following formulae (option PCTLDEF=):
  1. Weighted average at \( x_j ) \): \( P_\tau = (1-g)x_j + gx_{j+1} \)
     - if \( n=100 \), \( P_{5\%} = 5 \)
     - if \( n=50 \), \( P_{5\%} = 2.5 \)
  2. Observation numbered closest to \( n\tau \). Let integer part of \( n\tau = j \). Let \( g \) fractional. For example,
     - if \( n=100 \) and \( \tau = 5\% \), \( j=5 \) and \( g=0 \)
     - if \( n=50 \) and \( \tau = 5\% \), \( j=2 \) and \( g=0.5 \)
  3. Empirical distribution function \( P_\tau = x_j + (g \neq 0) \)
     - if \( n=100 \), \( P_{5\%} = 5 \)
     - if \( n=50 \), \( P_{5\%} = 2.5 \)
  4. Weighted average at \( x_j ) \): Replace \( n \) by \( n+1 \) in formula 1
     - if \( n=100 \), \( P_{5\%} = 5.05 \)
     - if \( n=50 \), \( P_{5\%} = 2.55 \)
  5. Empirical distribution function with averaging \( P_\tau = 0.5(x_j + x_{j+1}) + (g > 0) \)
     - if \( n=100 \), \( P_{5\%} = 5.5 \)
     - if \( n=50 \), \( P_{5\%} = 3 \)

Kernel estimation

The kernel density estimator is defined as:

\[
\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x - X_i}{h} \right)
\]

With \( h \) called the bandwidth and \( K \) the kernel.

- Choice of the Kernel
  - The Kernel is a distribution, usually chosen unimodal and symmetric. Its choice (Gaussian, rectangular, triangular, Epanechnikov, ...) is not crucial compared to the choice of the bandwidth.
- Choice of the bandwidth
  - Optimal bandwidths can be estimated by several methods: Sheather-Jones Plug In (SJPI) is the SAS default.
- Easy to run with SAS proc KDE

SANOFI [Image]
Kernel estimation illustrated

Example of objective function

Equivariance in Quantile regression

- Scale equivariance: for any $a > 0$
  \[ \hat{\beta}_i(aY, X) = a\hat{\beta}_i(Y, X) \]
  \[ \hat{\beta}_i(-aY, X) = a\hat{\beta}_i(Y, X) \]

- Shift equivariance: for any $\gamma$
  \[ \hat{\beta}_i(Y + X\gamma, X) = \hat{\beta}_i(Y, X) + \gamma \]

- Equivariance to reparameterization of design: for any nonsingular $A$
  \[ \hat{\beta}_i(Y, AX) = A^{-1}\hat{\beta}_i(Y, X) \]

- Equivariance to monotonic transformations: for a nondecreasing function $h$,
  \[ Q_{\text{osp}}(t) = hQ_{\text{osp}}(t) \]
  not true for the mean as $E(h(Y)) \neq h(E(Y))$

Inference in Quantile regression

- Using asymptotic properties
  \[ \sqrt{n}(\hat{\beta} - \beta) \sim N(0, \Lambda_n) \]
  \[ \Lambda_n = \tau(1 - \tau)(E[Q_{\text{osp}}(0X)|X]\{X\})^{-1}E[X\{X\}]\{E[Q_{\text{osp}}(0X)|X]\{X\}]^{-1} \]
  allows to compute confidence interval, likelihood ratio and Wald tests
  but type I error dramatically inflated for small sample size ($n<500$)

- Using bootstrap
  allows to compute confidence interval and tests but is unstable for small
data sets ($n<100$) and computationally intense for huge data sets.
  Option CI=RESAMPLING in PROC QUANTREG (default if $n>5000$ or $p>20$)

- Using rank test statistic
  unlike Wald tests or likelihood ratio tests, requires
  no estimation of the nuisance parameter under i.i.d. error models.
  Option CI=RANK in PROC QUANTREG (default if $n<5000$ and $p<20$)
PROC QUANTREG DATA = sas-data-set <options>;
BY variables;
CLASS variables; /* generate indicator variables*/
<label: MODEL response = <effects> < / options > ;
<label: TEST effects < / WALD | LR > ; /* test \beta_j=0 */
RUN;

Useful options: ALGORITHM=n, CI=
Usefull MODEL option: QUANTILE=(list of quantiles|ALL)

---

**SAS code**

**SAS output**

- **Model information**
  - report the name of the data set and the response variable, the number of covariates, the number of observations, algorithm of optimization and the method for confidence intervals.
- **Summary statistics**
  - report the sample mean and standard deviation, sample median, MAD and interquartile range for each variable included in the MODEL statement.
- **Quantile objective function**
  - report the quantile level to be estimated, the optimized objective function and the predictive value at covariate mean
- **Parameter Estimates**
  - report the estimated coefficients and their 95% confidence intervals.

---

**Fit criteria**

- **Fit Criteria**
  - $R^2 = \frac{\text{MPE}_{\text{LR}}(\tau)}{\text{MPE}_{\text{LR}}(\tau)}$ (vs. $R^2$)
  - $AIC(\tau) = 2\log(\text{MPE}_{\text{LR}}(\tau)) + 2p$
  - $SK(\tau) = 2\log(\text{MPE}_{\text{LR}}(\tau)) + p \log(n)$
  - $AICC(\tau) = 2\log(\text{MPE}_{\text{LR}}(\tau)) + 2(p+1) \frac{n-p-2}{n-p}$
  - Sawat's $BIC(\tau) = 2\log(\text{MPE}_{\text{LR}}(\tau)) + p \log(n) \frac{n+p}{n-p-2}$

Yonggang Yao, Nonparametric Statistics — 2010 JSM, Vancouver, Canada