Normal ranges determination with Quantile regression

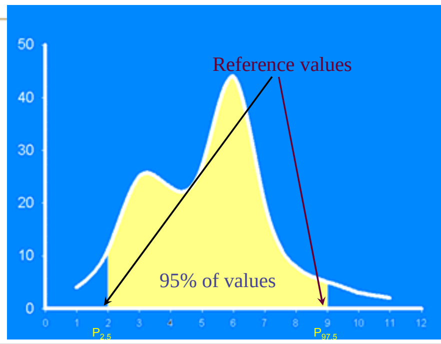
Luc Esserméant, Sept. 2012
NCSC 2012, Potsdam

Contents

Normal ranges or reference values definition

- *Statistical bases of reference values in laboratory medicine*, HARRIS E.K., BOYD J.C. , Dekker ed., 1995
  - « The interpretation of any measurement of an individual depends on the existence of a body of relevant information to which that measurement may be referred for comparison. »
  - « …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_τ is percentile τ%,
  - reference ranges are noted $P_{2.5}$ and $P_{97.5}$
  - Median is P_{50}

Example

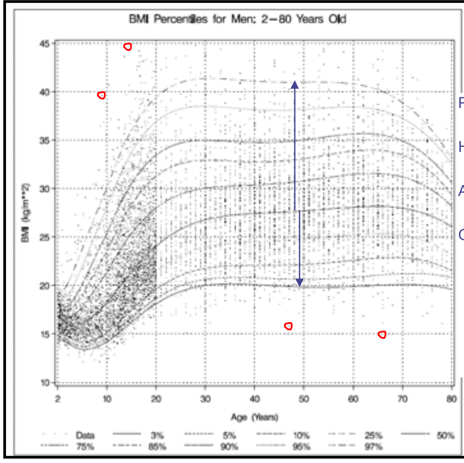


Calculation reference limits

- Parametric approach
  - if the distribution is known, directly or after transformation, percentiles are easily determined.
    - Example: For a normal distribution, P_τ = mean + $z_{1-\tau}$ * SD, where $z_{1-\tau}$ is the percentile $1-\tau$ 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. 5 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**



Possibly

Heteroscedasticity

Asymmetry

Outliers

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 $Y = y_1, y_2, \ldots, y_n$
- Mean of Y is the central point that minimizes the arithmetic mean of the quadratic deviations

$$\arg\min_b \sum_i (y_i - b)^2$$

Proof $$0 = \frac{\partial \sum_i (y_i - b)^2}{\partial b} = 2\sum_i (y_i - b) = 2\sum_i y_i - 2nb \Leftrightarrow b = \frac{\sum_i y_i}{n}$$

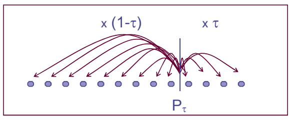
- The median is the central point that minimizes the arithmetic mean of the absolute deviations

$$\arg\min_b \sum_i |y_i - b|$$

Proof $$0 = \frac{\partial \sum_i |y_i - b|}{\partial b} = \sum_i 1(y_i - b > 0) - 1(y_i - b < 0) \Leftrightarrow b = \text{median}$$

Percentile on univariate sample by optimization

- The percentile $P\tau$ is the point that minimizes the weighted arithmetic mean of the absolute deviations



$$\arg\min_b \sum_i \rho_\tau (y_i - b) = \arg\min_b \sum_{i;y_i \ge b} \tau |y_i - b| + \sum_{i;y_i < b} (1-\tau)|y_i - b|$$

with "check function" $\rho_\tau(u) = u(\tau - 1(u < 0))$

Proof

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

$$Q_Y(\tau) = F^{-1}(\tau) = \inf\{ y : F(y) \ge \tau\} = \inf\{ y : P(Y<y) \ge \tau\}$$

- Check that the quantile QY(t) = F-1(t) is the solution of

$$\arg\min_b E(\rho_\tau(y-b)) = \arg\min_b (\tau - 1)\int_{-\infty}^{b} (y-b)dF(y) + \tau \int_{b}^{+\infty} (y-b)dF(y)$$

- Solution obtained setting the derivative to 0:

$$0 = (\tau - 1)\int_{-\infty}^{b} -dF(y) + \tau \int_{b}^{+\infty} -dF(y) = -\tau \underbrace{\left(\int_{-\infty}^{b} dF(y) + \int_{b}^{+\infty} dF(y) \right)}_{=1} + \int_{-\infty}^{b} dF(y)$$

$$\Leftrightarrow \tau = F(b)$$

Quod Erat Demonstrandum

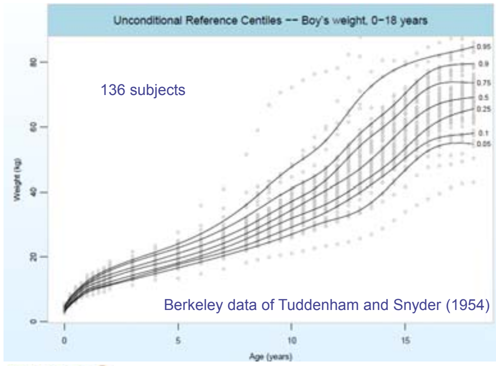
Linear regression quantile

- Consider a covariate $X = x_1, x_2, \ldots, x_n$
- OLS regression estimates the linear conditional mean function $E(Y|X=x) = x'\beta$ by solving $\arg\min_\beta \sum_i (y_i - x_i'\beta)^2$
- Likewise, quantile regression estimates the linear conditional quantile function $Q_{y|X}(\tau) = x'\beta_\tau$, by solving

$$\hat{\beta}_\tau(Y,X) = \arg\min_\beta \underbrace{\sum_i \rho_\tau(y_i - x_i'\beta)}_{\text{Objective function}}$$

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

Example: growth chart



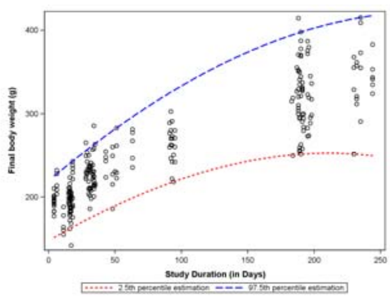
Conditional growth chart

- A simple AR(2) model

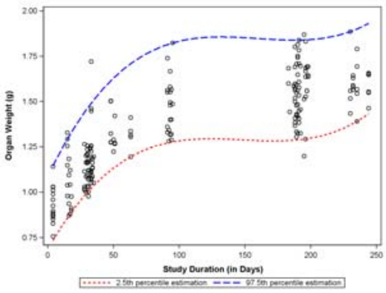
$Q_\tau(W_t) = \beta_{\tau 0} + \beta_{\tau 1} W_{t-1} + \beta_{\tau 2} W_{t-2} + \beta_{\tau 3} H_t,$

- 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



SANOFI

Goodness of fit criteria (not in QUANTREG SAS 9.2)

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - x_i'\hat{\beta})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{SSE_F}{SSE_R}$$

F denotes Full model and R reduced (empty) model

- A natural analog of R² is

$$R^1 = 1 - \frac{\sum_i \rho_\tau\left(y_i - x_i'\hat{\beta}_\tau\right)}{\sum_i \rho_\tau\left(y_i - x_i'\tilde{\beta}_\tau\right)}, \text{ with } \hat{\beta}_\tau \text{ and } \tilde{\beta}_\tau \text{ estimators from full and reduced models}$$

- Koenker (2005) suggested an adapted AIC (aAIC) where the likelihood is replaced by the empirical risk

$$AIC_\tau = -2Log\left(\frac{1}{n}\sum_i \rho_\tau\left(y_i - x_i'\hat{\beta}_\tau\right)\right) + 2p$$

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

SANOFI

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 bandwith.
- 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**

SANOFI

References (1/2)

- Univariate sample
  - **HARRIS E.K. and BOYD J.C., *Marcel Dekker*, *Statistical bases of reference values in laboratory medicine,* 1995.**
  - **SILVERMAN B.W., *Chapman & Hall, Density estimation for statistics and data analysis, 1996.***
  - **JONES M.C., MARRON J.S. and SHEATHER S.J., *JASA*, *A brief Survey of bandwith selection for density estimation*, Vol 91 N°433, 1996.**
  - **HARRELL F.E. and DAVIS C.E., *A new distribution-free quantile estimator*, *Biometrika*, Vol 69, N°3, 1982.**
  - **DERZKO G., *An intrinsic approach to nonparametric density estimation*, *C.R. Acad. Sci. Paris*, t. 327, Série 1, 1998.**
- SAS/STAT User’s guide
- With a covariate
  - **HEALY M.J.R., RASBASH J. and YANG M., Distribution-free estimation of age-related centiles, Annals of human biology, Vol 15 N°1 (1988), pp 17-22**
  - **PAN H.Q., GOLDSTEIN H. and YANG Q., Non-parametric estimation of age-related centiles over wide age ranges, Annals of human biology, Vol 17 N°6 (1990),pp 475-481**
  - **HAUSPIE R., Application of the method of Healy and Pan for estimating centile lines, Bull. et Mém. de la Soc. d’ Anthrop. de Paris, n.s., t. 3, n° 3-4, 1991, pp. 257-273.**
  - **GANNOUN A., GIRARD S., GUINOT C. and SARACCO J.,Trois méthodes non paramétriques pour l'estimation de courbes de référence- application à l'analyse de propriétés biophysiques de la peau, Revue de Statistique Appliquée, 50 no. 1 (2002), p. 65-89**

SANOFI

References (2/2)

- Quantile regression
  - KOENKER R. and BASSETT G., Regression quantiles, Econometrica, Vol 46 (1978), pp33-50
  - KOENKER R. and MACHADO A.F., Goodness of Fit and Related Inference Processes for Quantile Regression, Journal of the American Statistical Association, Vol. 94, No. 448. (Dec., 1999), pp. 1296-1310.
  - REDDEN D.T., FERNANDEZ J.R. and ALLISON D.B., A simple significance test for quantile regression, Statistics in Medicine, 2004; 23: 2587 – 2597
  - FRIEDRICH N., ALTE D., VOLZKE H., SPILCKE-LISS E., LUDERMANN J., LERCH M.M., KOHLMANN T., NAUCK M., WALLASCHOFSKI H., Reference ranges of serum IGF-1 and IGFBP-3 levels in a general adult population: Results of the Study of Health in Pomerania (SHIP), ScienceDirect, Growth Hormone & IGF Research 18 (2008) 228-237.
  - CADE B.S., NOON B.R. and FLATHER C.H., Quantile regression reveals hidden bias and uncertainty in habitat models, 2005, Ecology 86: 786-800.
  - CHEN Colin, Growth Charts of Body Mass Index (BMI) with Quantile Regression, SAS Institute Inc. Cary, NC, U.S.A.
  - WEI Ying, PERE Anneli, KOENKER Roger and HE Xuming, Quantile Regression Methods for Reference Growth Charts, Statistics in Medicine, 2006; 25: 1369-1382.

Back-up

Empirical estimation in SAS

- You want the percentile τ of an univariate ordered sample $X = x_1, x_2, .., x_n$.
- Note $n\tau = j + g$, 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 $x_{n\tau}$. $P_\tau = (1-g)x_j + gx_{j+1}$
     - If $n=100$, $P_{5\%}=5$ if $n=50$, $P_{5\%}=2.5$
  2. Observation numbered closest to $n\tau$. Let i=integer part of $n\tau+0.5$
     $P_\tau = x_i 1(g\neq 0.5) + x_j 1(g=0.5 \ \& \ j \text{ even}) + x_{j+1} 1(g=0.5 \ \& \ j \text{ odd})$
     - If $n=100$, $P_{5\%}=5$ if $n=50$, $P_{5\%}=2$
  3. Empirical distribution function $P_\tau = x_j 1(g=0) + x_{j+1} 1(g>0)$
     - If $n=100$, $P_{5\%}=5$ if $n=50$, $P_{5\%}=3$
  4. Weighted average at $x_{(n+1)\tau}$. 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
     $P_\tau = 0.5(x_j + x_{j+1}) 1(g=0) + x_{j+1} 1(g>0)$
     - If $n=100$, $P_{5\%}=5.5$ if $n=50$, $P_{5\%}=3$

Kernel estimation

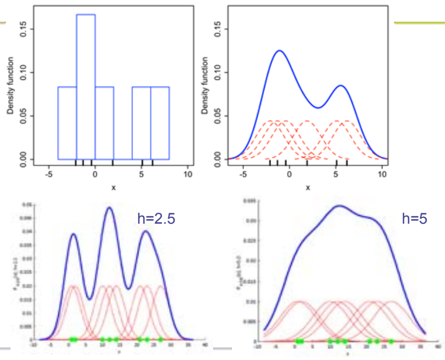
The kernel density estimator is defined as:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \text{with } \int_{\Re} K(x)dx = 1$$

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, Epanenchnikov,…) is not crucial compared to the choice of the bandwidth.
- Choice of the bandwith
  - Optimal bandwith can be estimated by several methods: Sheather-Jones Plug In (SJPI) is the SAS default.
- Easy to run with SAS proc KDE

Kernel estimation illustrated

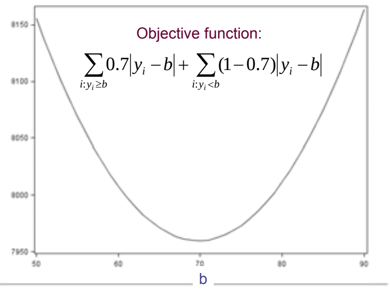


SANOFI

Example of objective function

N=99, Y = 1, 2, …, 98,10000

- Median = 50
- Mean = 150
- P70 = 70



SANOFI

Equivariance in Quantile regression

- Scale equivariance: for any a>0

  $\hat{\beta}_\tau(ay, X) = a\hat{\beta}_\tau(y, X)$

  $\hat{\beta}_\tau(-ay, X) = a\hat{\beta}_{1-\tau}(y, X)$
- Shift equivariance: for any γ

  $\hat{\beta}_\tau(y + X\gamma, X) = \hat{\beta}_\tau(y, X) + \gamma$
- Equivariance to reparameterization of design: for any nonsingular A

  $\hat{\beta}_\tau(y, XA) = A^{-1}\hat{\beta}_\tau(y, X)$
- Equivariance to monotonic transformations: for a nondecreasing function h, **$Q_{h(Y)|X}(\tau) = h(Q_{Y|X}(\tau))$**

  not true for the mean as E(h(Y)) ≠ h(E(Y))

SANOFI

Inference in Quantile regression

- Using asymptotic properties

  $\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \rightarrow N(0, \Lambda_\tau)$

  $\Lambda_\tau = \tau(1-\tau)(E[f_{u_\tau}(0|X)X_i'X_i])^{-1}E[X_i'X_i]E[f_{u_\tau}(0|X)X_i'X_i]^{-1}$
  - Allows to compute confidence interval, likelihood ratio and Wald tests **but type I error dramatically inflated for small sample size (n<500)**
  - Option CI=SPARSITY in PROC QUANTREG
- 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
  - The 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)

SANOFI

SAS code

```
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 βj=0 */
RUN;
```

Useful options: ALGORITHM=, CI=

Usefull MODEL option: QUANTILE=(list of quantiles|ALL)

SANOFI

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.

SANOFI

Fit criteria

- Fit Criteria

$$\begin{cases} MWAR_F(\tau) = \min\limits_{\text{full-model}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau[y_i - x_i\beta_\tau] \right\} \\ MWAR_R(\tau) = \min\limits_{\text{reduced-model}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau[y_i - x_i\beta_\tau] \right\} \end{cases}$$

$$R^1(\tau) = 1 - \frac{MWAR_F(\tau)}{MWAR_R(\tau)} \quad (\text{vs. } R^2)$$

$$AIC(\tau) = 2n\log(MWAR(\tau)) + 2p$$

$$SIC(\tau) = 2n\log(MWAR(\tau)) + p\log n$$

$$AICC(\tau) = 2n\log(MWAR(\tau)) + 2(p+1)\frac{n}{n-p-2}$$

$$\text{Sawa's } BIC(\tau) = 2n\log(MWAR(\tau)) + n\log\frac{n+p}{n-p-2}$$

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

SANOFI