QQ图回归模拟

宋歌 2015080086 数52

2018 年 4 月 10 日

1 实验的目的

对不同的分布、同一分布不同的参数和不同的样本量，观察它们的QQ图的特点。

2 选择的因素

样本量$n = 30, 100, 1000$；不同的模型：正态分布、Cauchy分布、t分布、χ^2分布、Poisson分布、二项分布；各个模型中的不同参数。

3 假设的基本理由

- 随着试验次数的增大，某一事件发生的频率逐渐趋于该事件发生的概率。
- 可以将分位数作为分布的特征，通过分位数基本确定两个分布之间的关系。
- 若$Z \sim N(0,1)$，则有$X = \sigma Z + \mu \sim N(\mu, \sigma^2)$
  若$Z \sim Cauchy(0,1)$，则有$X = \sigma Z + \mu X \sim Cauchy(\mu, \sigma)$
- 随着自由度逐渐增大，t分布逐渐接近标准正态分布。
- Poisson定理：在n重贝努力试验中，事件A在每次试验中发生的概率为p，出现A的总次数K服从二项分布$B(n,p)$，当n很大p很小，$\lambda = np$大小适中时，二项分布可用参数为$\lambda = np$的Poisson分布来近似。

4 详细的研究方法

4.1 观测值与理论分布

- 产生观测值：对于每一个模型，产生n个服从该模型的随机数x_i，并从小到大排序，作为QQ图中的纵坐标值。
- 选取分位数：产生与样本量相应数目的该模型分布的分位数，作为QQ图中的横坐标值。
- 作图：作QQ图，并在其上用黑色画出散点的最小二乘拟合直线，用红色画出理论直线。
  横纵坐标的分位数来自相同参数的相同分布时，该理论直线为过原点的斜率为1的直线;
  对于正态分布和Cauchy分布，其横坐标的分布为标准的$N(0,1), Cauchy(0,1)$分布，而纵坐标可能来自$N(\mu,\sigma^2), Cauchy(\mu,\sigma)$，此时其理论直线有如下形式。

4.1.1 正态分布

设观测值x来自$N(\mu,\sigma^2)$的分布总体，由于QQ图中横坐标是来自标准正态总体z的分位数，故理论上应有

$$x = \sigma z + \mu$$

对每一个模型，在QQ图中用黑色画出散点的最小二乘拟合直线，用红色画出上述理论直线，并进行比较。

4.1.2 Cauchy分布

设观测值x来自$Cauchy(\mu,\sigma)$的分布总体，由于QQ图中横坐标是来自$Cauchy(0,1)$总体z的分位数，故理论上应有

$$x = \sigma z + \mu$$

对每一个模型，在QQ图中用黑色画出散点的最小二乘拟合直线，用红色画出上述理论直线，并进行比较。

4.2 不同分布之间的关系

4.2.1 t分布与正态分布

已知随着t分布自由度的增大，t分布逐渐接近于标准正态分布。故可以在QQ图中比较自由度较大的t分布与标准正态分布。

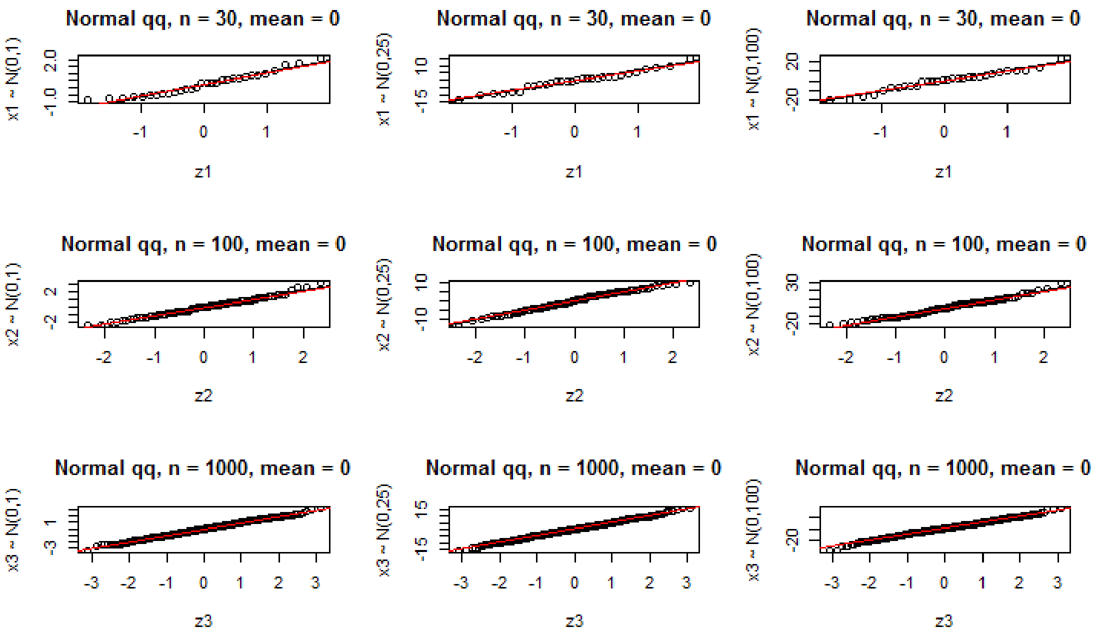
4.2.2 二项分布与Poisson分布

已知在二项分布中，当实验次数n很大，成功概率p很小时，二项分布可以用参数为$\lambda = np$的Poisson分布来近似。故可以在QQ图中比较n很大p很小的二项分布与$\lambda = np$的Poisson分布。

5 结果及讨论

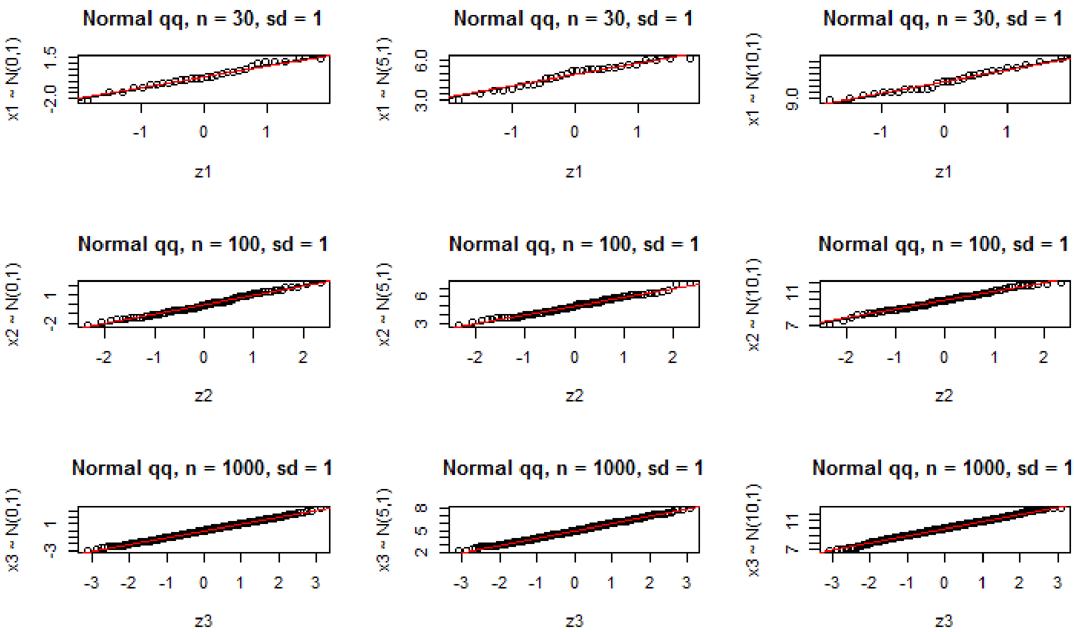
5.1 正态分布

- 固定均值为0，取标准差$sd = 1, 5, 10$时
  所作的红色理论直线以样本均值为截距、样本标准差为斜率。此时截距近似为0.
  对于不同的样本量$n = 30, 100, 1000$结果如下



  可见拟合直线和理论直线的截距近似为0；
  可见随着标准差的增大，拟合直线的斜率增大；
  随着样本量的增大，拟合直线与理论直线越接近。

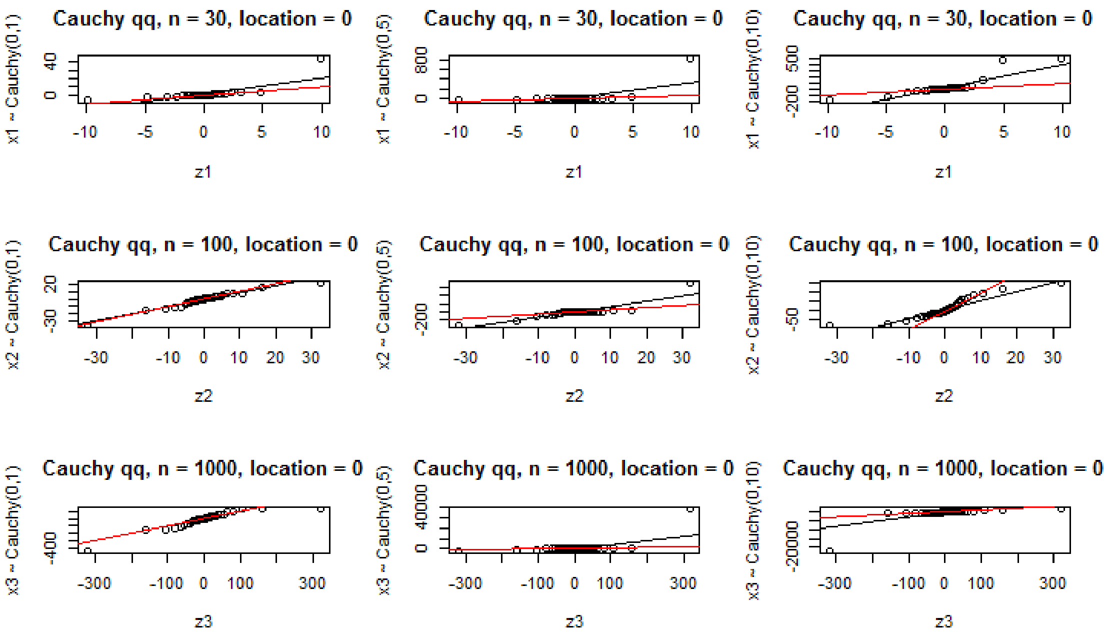
- 固定方差为1，取均值$mean = 0, 5, 10$时
  所作的红色理论直线以样本均值为截距、样本标准差为斜率。此时斜率近似为1.
  对于不同的样本量$n = 30, 100, 1000$结果如下



可见拟合直线和理论直线的斜率近似为1；
可见随着均值的增大，拟合直线的截距增大；
随着样本量的增大，拟合直线与理论直线越接近。

5.2 Cauchy分布

- 固定location为0，取$scale = 1, 5, 10$时
  所作的红色理论直线分别以0为截距、$1, 5, 10$为斜率。
  对于不同的样本量$n = 30, 100, 1000$结果如下



可见拟合直线和理论直线的截距近似为0；

可见随着$scale$的增大，拟合直线的斜率增大；

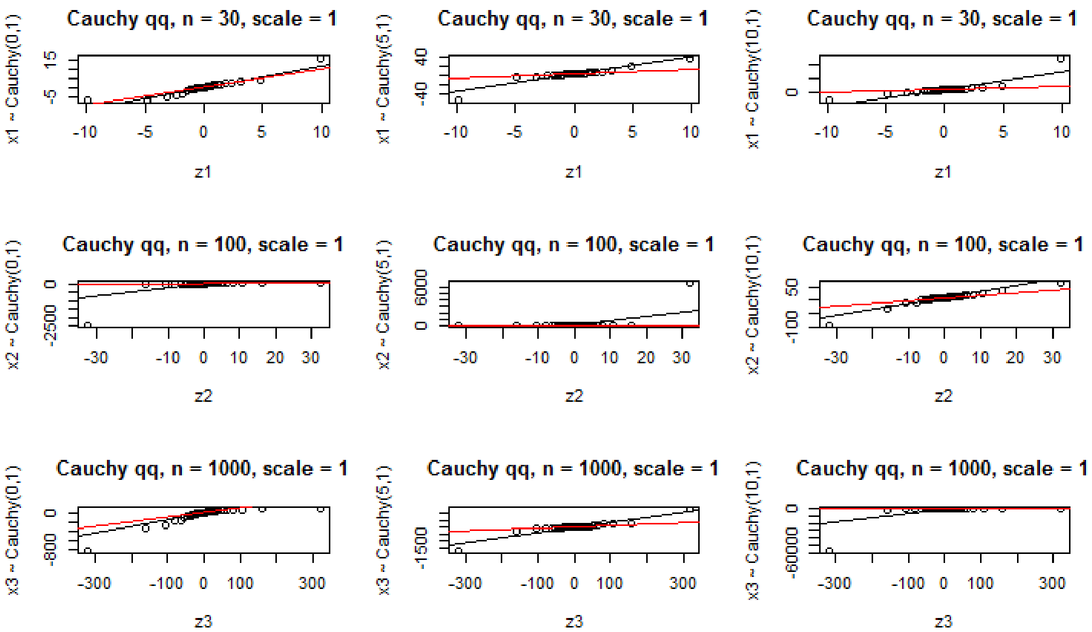
注意到此时对于Cauchy分布，由于样本方差太大，图中的最小二乘拟合直线对于数据的拟合并不好。但对于第一列$Cauchy(0,1)$分布拟合得还不错。

但此时单从图中来看，散点的分布与红色理论直线拟合得较好，但拟合程度与样本量的大小没有显著的关系。

- 固定scale为1，取$location = 0, 5, 10$时

  所作的红色理论直线分别以1为斜率、$0, 5, 10$为截距。

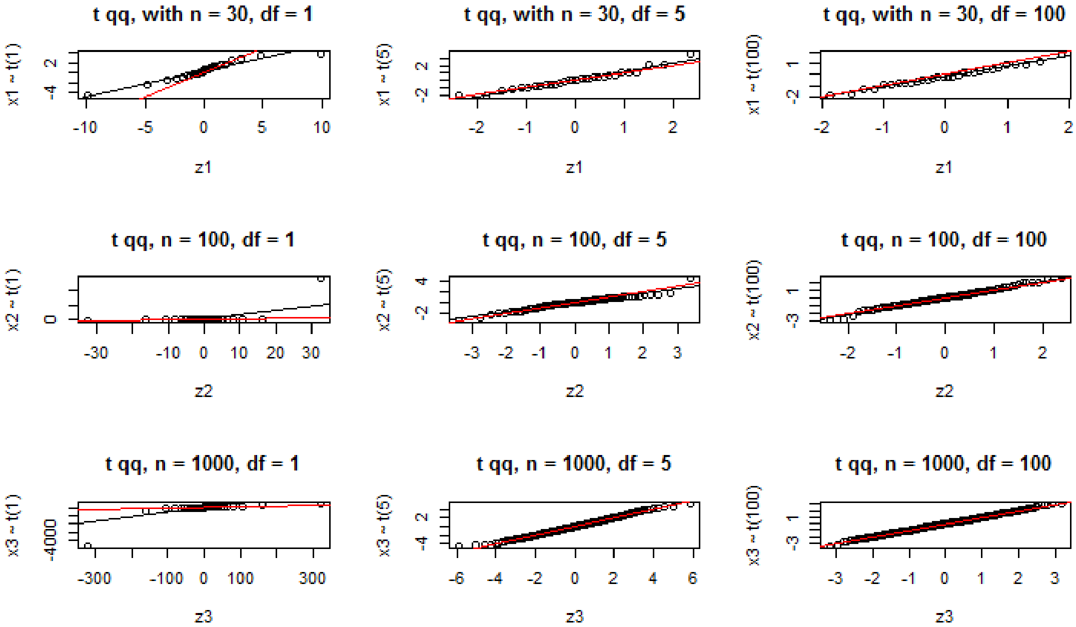
  对于不同的样本量$n = 30, 100, 1000$结果如下



可见随着$location$的增大，拟合直线的截距增大；
注意到此时对于Cauchy分布，由于样本方差太大，图中的最小二乘拟合直线对于数据的拟合并不好。但对于第一列$Cauchy(0,1)$分布拟合得还不错。
但此时单从图中来看，散点的分布与红色理论直线拟合得较好，但拟合程度与样本量的大小没有显著的关系。

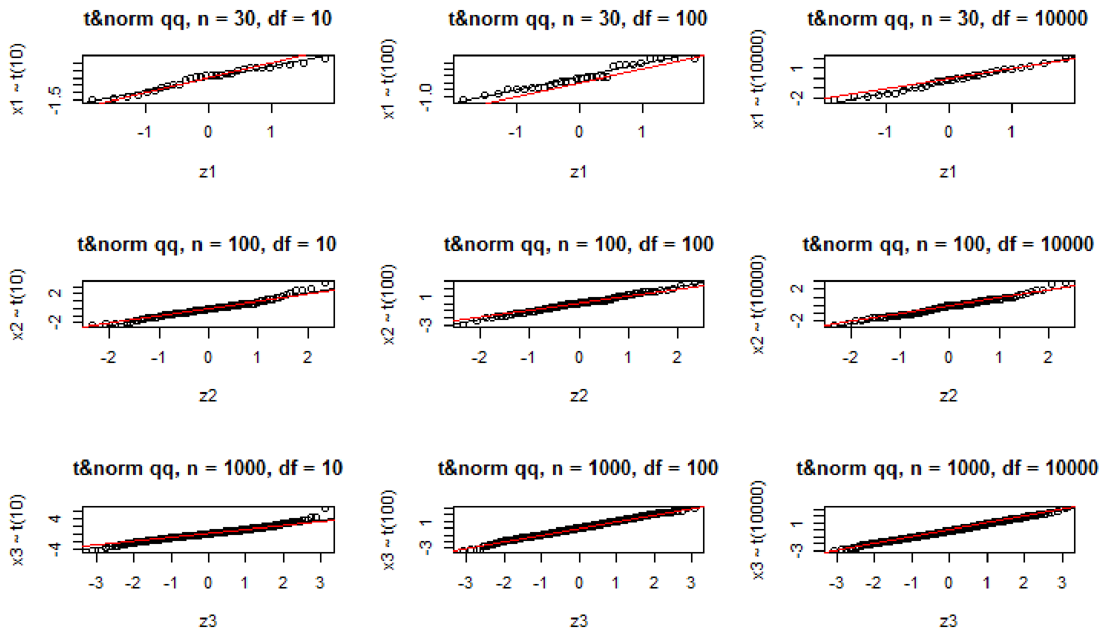
5.3 t分布

- t分布与t分布：
  选取自由度$df = 1, 5, 100$
  所作的红色理论直线为过原点斜率为1的直线。
  对于不同的样本量$n = 30, 100, 1000$结果如下：



可见随着自由度的增大，拟合程度越来越好；
随着样本量的增大，红色理论直线与黑色拟合直线越来越接近。

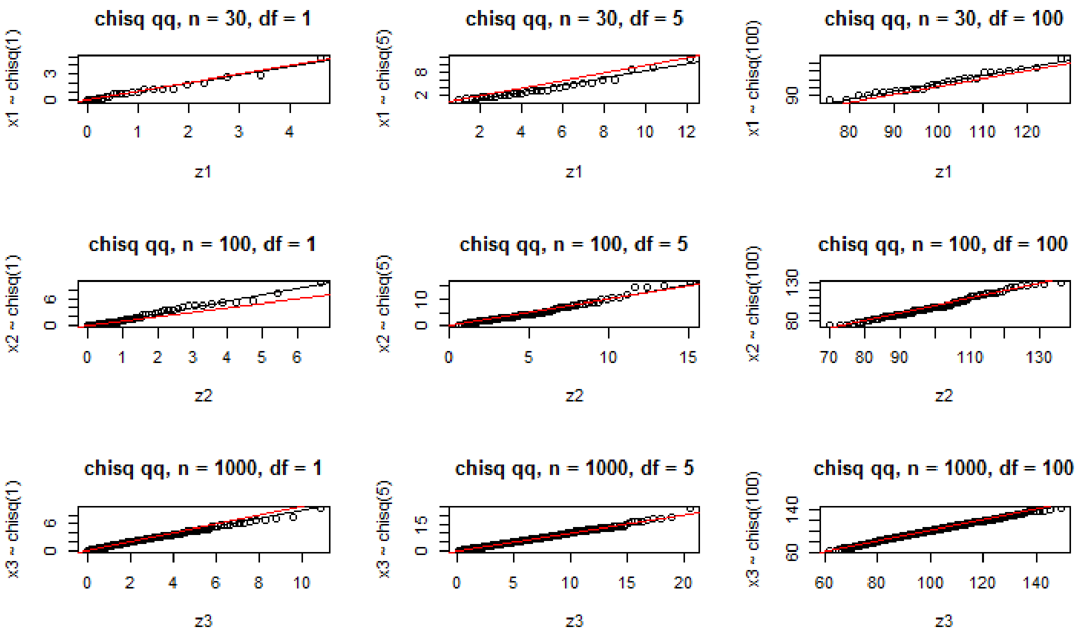
- t分布与正态分布：
  选取自由度$df = 10, 100, 10000$
  所作的红色理论直线为过原点斜率为1的直线。
  对于不同的样本量$n = 30, 100, 1000$，t分布与正态分布的关系如下：



可见随着自由度增大，t分布越来越接近正态分布；
随着样本量的增大，红色理论直线与黑色拟合直线越来越接近。

5.4 卡方分布

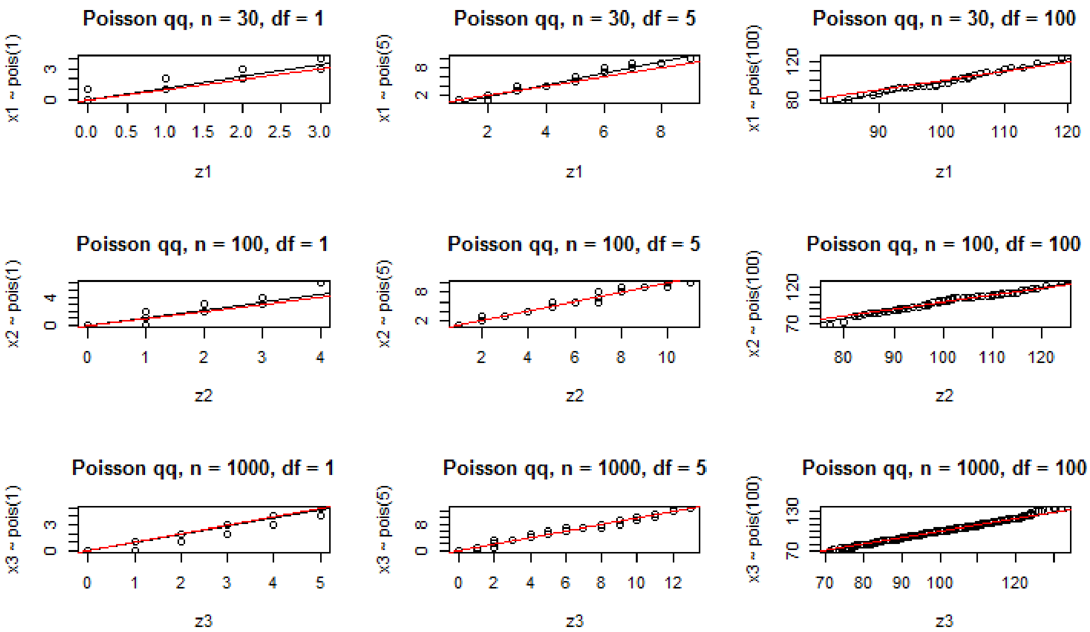
- 选取自由度$df = 1, 5, 100$
  所作的红色理论直线为过原点斜率为1的直线。
  对于不同的样本量$n = 30, 100, 1000$ 结果如下：



可见随着自由度增大，红色理论直线与黑色拟合直线越来越接近；
随着样本量的增大，红色理论直线与黑色拟合直线越来越接近。

5.5 Poisson分布

- 选取参数$\lambda = 1, 5, 100$
  所作的红色理论直线为过原点斜率为1的直线。
  对于不同的样本量$n = 30, 100, 1000$结果如下：

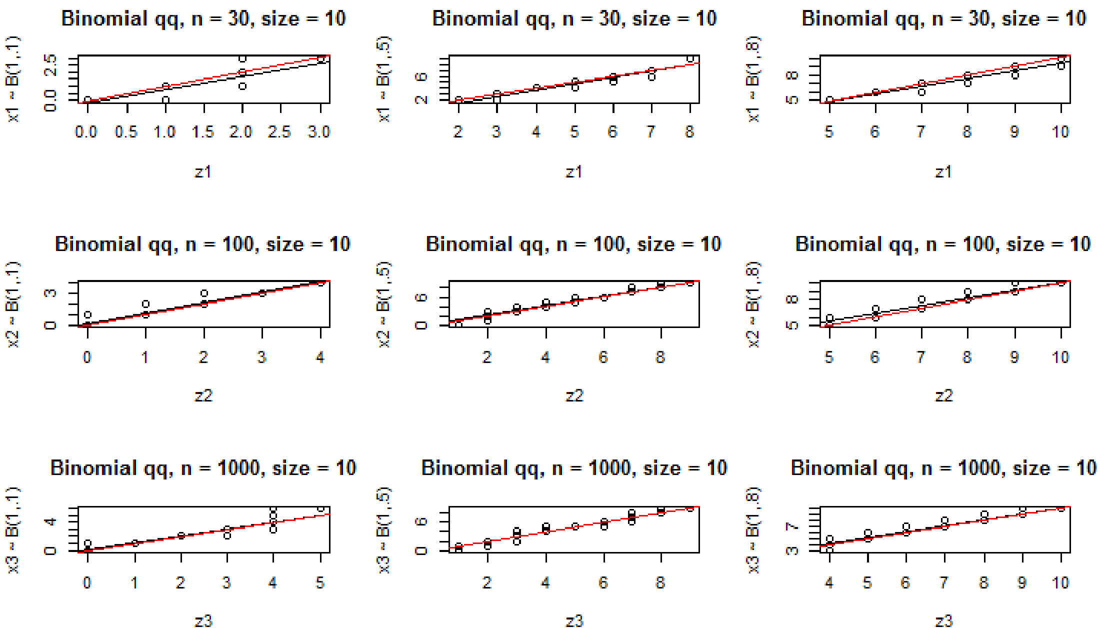


可见随着样本量的增大，红色理论直线与黑色拟合直线越来越接近；
拟合程度与参数没有显著关系。

5.6 二项分布

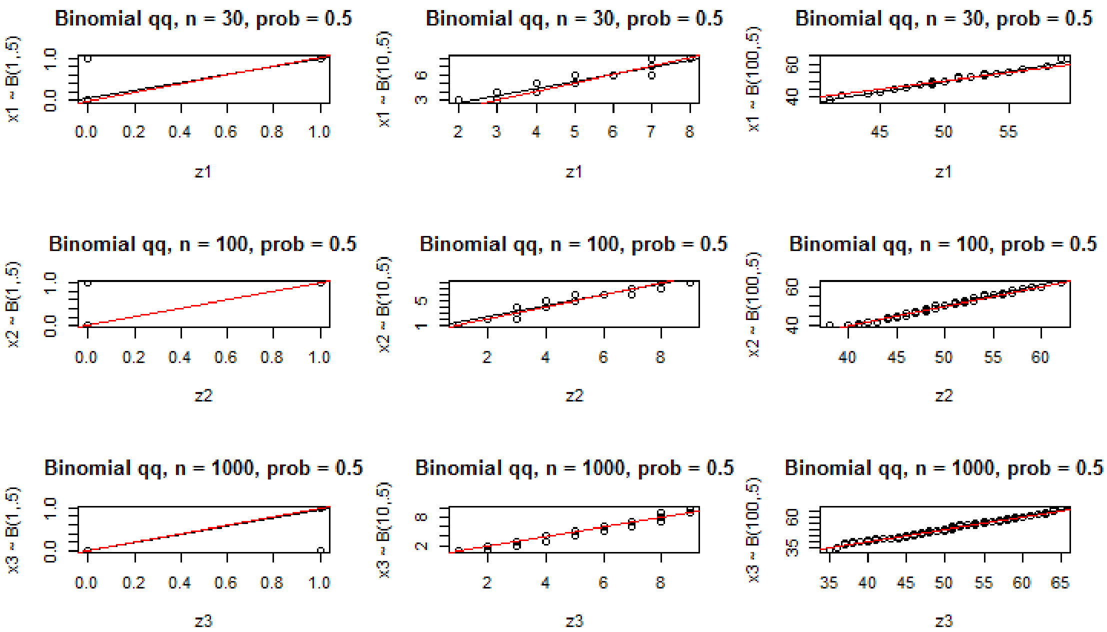
5.6.1 二项分布与二项分布

- 固定size为10，取$prob = 0.1, 0.5, 0.8$时
  所作的红色理论直线为过原点斜率为1的直线。
  对于不同的样本量$n = 30, 100, 1000$结果如下



均拟合得很好，拟合程度与参数没有显著关系。

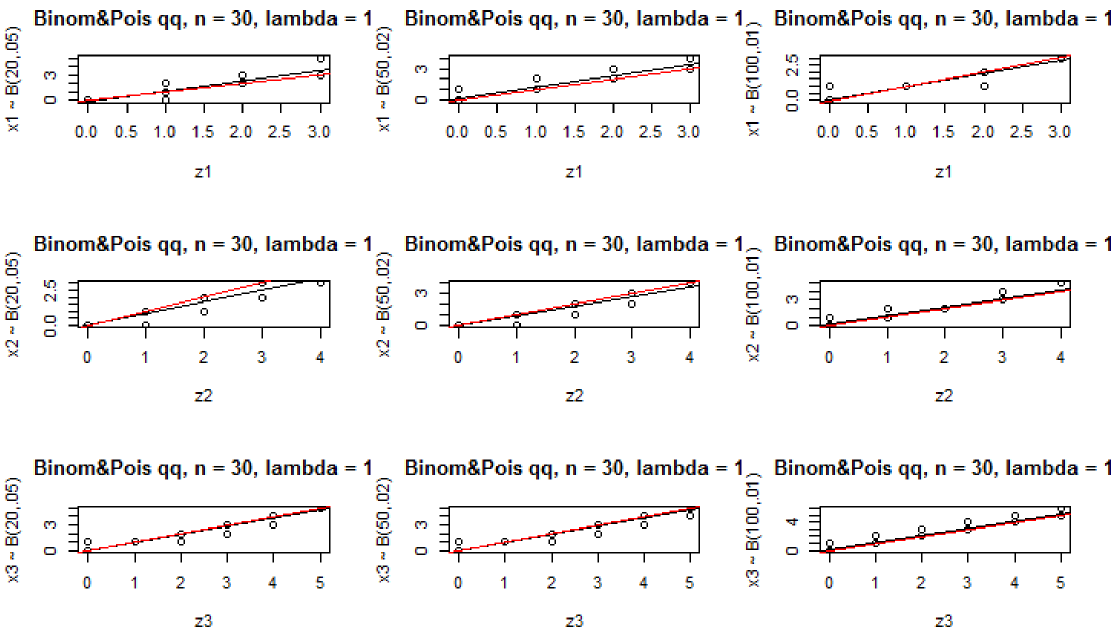
- 固定$prob = 0.5$，取$size = 1, 10, 100$时
  所作的红色理论直线为过原点斜率为1的直线。
  对于不同的样本量$n = 30, 100, 1000$结果如下



均拟合得很好，拟合程度与参数没有显著关系。

5.6.2 二项分布与Poisson分布

- 固定横坐标的分布为$P(\lambda=1)$，选取三组二项分布的参数值$(20, 0.05), (50, 0.02), (100, 0, 01)$所作的红色理论直线为过原点斜率为1的直线。
  对于不同的样本量$n = 30, 100, 1000$，二项分布与Poisson分布的关系：



可见随着样本量的增大，红色理论直线与黑色拟合直线越来越接近；
随着n越大、p越小，二项分布越来越接近Poisson分布。