Historic Curiosity

- The article Toothpaste Detergents: A Potential Source of Oral Soft Tissue Damage" (Intl. J. of Dental Hygiene, 2008: 193198) contains the following statement:
- "Because the sample number for each experiment (replication) was limited to three wells per treatment type, the data were assumed to be normally distributed."
- As justification for this leap of faith, the authors wrote that "Descriptive statistics showed standard deviations that suggested a normal distribution to be highly likely. Note: This argument is not very persuasive.

Sample percentiles

- The basis for our construction is a comparison between percentiles of the sample data and the corresponding percentiles of the distribution under consideration.
- We know that the (100p)th percentile of a continuous distribution with cdf $F(\cdot)$ is the number $\eta(p)$ that satisfies $F(\eta(p)) = p$. That is, $\eta(p)$ is the number on the measurement scale such that the area under the density curve to the left of $\eta(p)$ is p.

Sample percentiles

- The 50th-sample percentile should separate the smallest 50% of the sample from the largest 50%, the 90th percentile should be such that 90% of the sample lies below that value and 10% lies above, and so on.
- Unfortunately, we run into problems when we actually try to compute the sample percentiles for a particular sample of n observations. If, for example, $n = 10$ we can split off 20% of these values or 30% of the data, but there is no value that will split off exactly 23% of these ten observations

Sample percentiles

- Order n observations in the ascending order. Then, the ith smallest observation is $[100(i - .5)/n]$ sample percentile
- Sample percentiles corresponding to intermediate percentages can be obtained by linear interpolation.
- For example, if n $=$ 10, the percentages corresponding to the ordered sample observations are $100(1.5)/10 = 5\%$, $100(2.5)/10 = 15\%$, 25%,, and $100(10.5)/10 = 95\%$. The 10th percentile is then halfway between the 5th percentile (smallest sample observation) and the 15th percentile (second-smallest observation).
- For our purposes, such interpolation is not necessary because a probability plot will be based only on the percentages $100(i.5)/n$ corresponding to the n sample observations

Example

- The value of a certain physical constant is known to an experimenter. The experimenter makes $n = 10$ independent measurements of this value using a particular measurement device and records the resulting measurement errors (error = observed value true value). These observations appear in the accompanying table

Figure :

Percentage	5	15	25	35	45
z percentile	-1.645	-1.037	-.675	-.385	-.126
Sample observation	-1.91	-1.25	-.75	-.53	.20
Percentage	55	65	75	85	95

Figure :

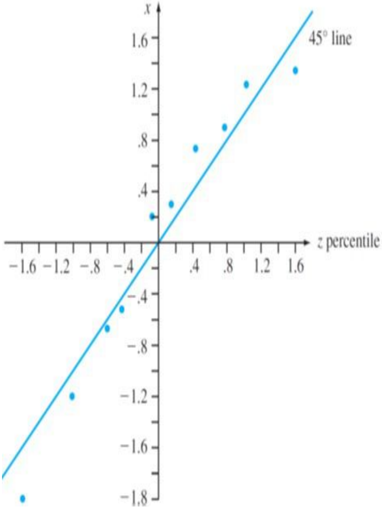
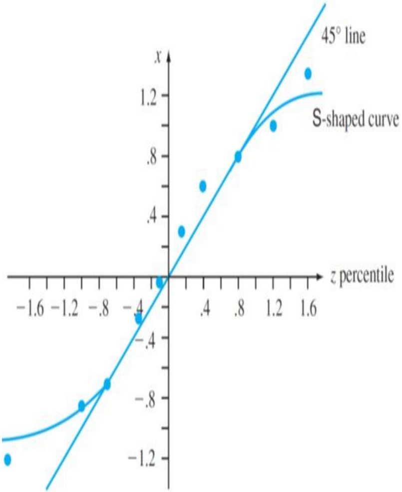


Figure :


45° line
x
1.2
.8
.4
S-shaped curve
z percentile
−1.6 −1.2 −.8 −.4
.4
.8
1.2
1.6
−.4
−.8
−1.2

Normal probability plot

- A plot of n pairs $([(100-0.5)/n]$th percentile,ith smallest observation) is called the normal probability plot
- If the sample observations $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ the plot should resemble a straight line with the intercept μ and slope σ