NAME: ________________

18.443 Exam 2 Spring 2015
Statistics for Applications
4/9/2015

1. **True or False:**

   (a). The significance level of a statistical test is not equal to the probability that the null hypothesis is true.

   (b). If a 99% confidence interval for a distribution parameter θ does not include θ_0, the value under the null hypothesis, then the corresponding test with significance level 1% would reject the null hypothesis.

   (c). Increasing the size of the rejection region will lower the power of a test.

   (d). The likelihood ratio of a simple null hypothesis to a simple alternate hypothesis is a statistic which is higher the stronger the evidence of the data in favor of the null hypothesis.

   (e). If the p-value is 0.02, then the corresponding test will reject the null at the 0.05 level.

2. **Testing Goodness of Fit**.

   Let X be a binomial random variable with n trials and probability p of success.

   (a). Suppose $n = 100$ and $X = 38$. Compute the Pearson chi-square statistic for testing the goodness of fit to the multinomial distribution with two cells with $H_0 : p = 0.5$.

   (b). What is the approximate distribution of the test statistic in (a), under the null Hypothesis H_0.

   (c). What can you say about the P-value of the Pearson chi-square statistic in (a) using the following table of percentiles for chi-square random variables ? (i.e., $P(\chi_3^2 \leq q.90 = 6.25) = .90$)

df	q.90	q.95	q.975	q.99	q.995
1	2.71	3.84	5.02	6.63	9.14
2	4.61	5.99	7.38	9.21	11.98
3	6.25	7.81	9.35	11.34	14.32
4	7.78	9.49	11.14	13.28	16.42

(d). Consider the general case of the Pearson chi-square statistic in (a), where the outcome $X = x$ is kept as a variable (yet to be observed). Show that the Pearson chi-square statistic is an increasing function of $|x - n/2|$.

(e). Suppose the rejection region of a test of H_0 is $\{X : |X - n/2| > k\}$ for some fixed known number k. Using the central limit theorem (CLT) as an approximation to the distribution of X, write an expression that approximates the significance level of the test for given k. (Your answer can use the cdf of $Z \sim N(0,1) : \Phi(z) = P(Z \leq z)$.)

3. **Reliability Analysis**

   Suppose that $n = 10$ items are sampled from a manufacturing process and S items are found to be defective. A $beta(a, b)$ prior [1] is used for the unknown proportion θ of defective items, where $a > 0$, and $b > 0$ are known.

   (a). Consider the case of a beta prior with $a = 1$ and $b = 1$. Sketch a plot of the prior density of θ and of the posterior density of θ given $S = 2$. For each density, what is the distribution's mean/expected value and identify it on your plot.

   (b). Repeat (a) for the case of a $beta(a = 1, b = 10)$ prior for θ.

   (c). What prior beliefs are implied by each prior in (a) and (b); explain how they differ?

   (d). Suppose that $X = 1$ or 0 according to whether an item is defective (X=1). For the general case of a prior $beta(a, b)$ distribution with fixed a and b, what is the marginal distribution of X before the $n = 10$ sample is taken and S is observed? (Hint: specify the joint distribution of X and θ first.)

   (e). What is the marginal distribution of X after the sample is taken? (Hint: specify the joint distribution of X and θ using the posterior distribution of θ.)

[1] A $beta(a, b)$ distribution has density $f_\Theta(\theta) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, 0 < \theta < 1.$

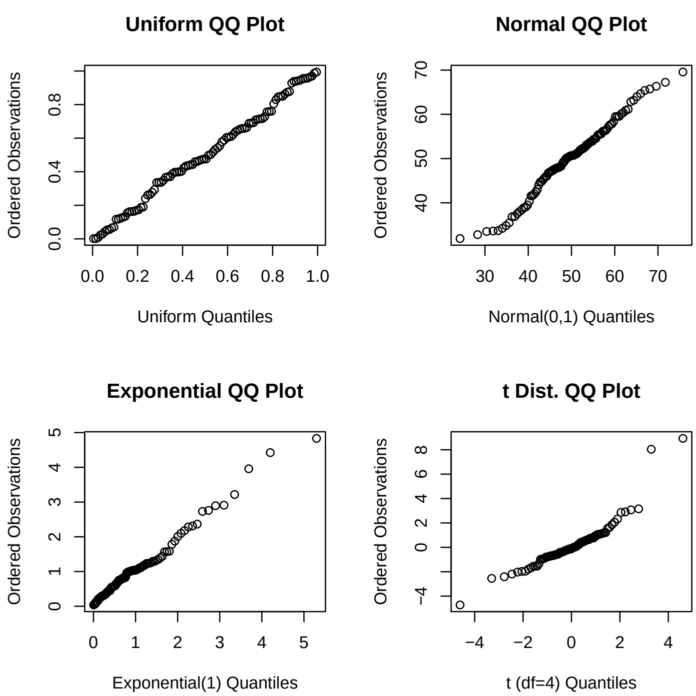
Recall that for a $beta(a, b)$ distribution, the expected value is $a/(a + b)$, the variance is $\dfrac{ab}{(a+b)^2(a+b+1)}$. Also, when both $a > 1$ and $b > 1$, the mode of the probability density is at $(a - 1)/(a + b - 2)$,

4. **Probability Plots**

Random samples of size $n = 100$ were simulated from four distributions:

- $Uniform(0, 1)$
- $Exponential(1)$
- $Normal(50, 10)$
- Student's t (4 degrees of freedom).

The quantile-quantile plots are plotted for each of these 4 samples:



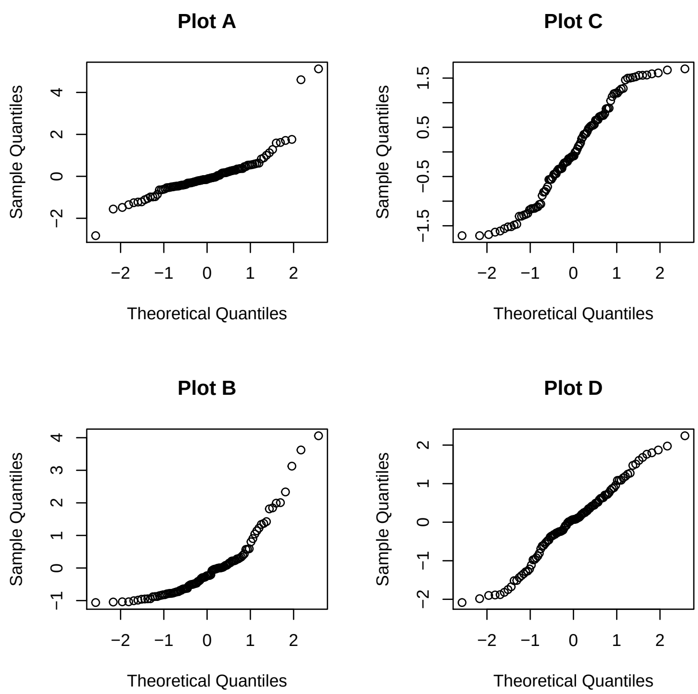
For each sample, the values were re-scaled to have sample mean zero and sample standard deviation 1

$$\{x_i, i = 1, \ldots, 100\} \Longrightarrow \{Z_i = \tfrac{x_i - \overline{x}}{s_x}, i = 1, \ldots, 100\}$$

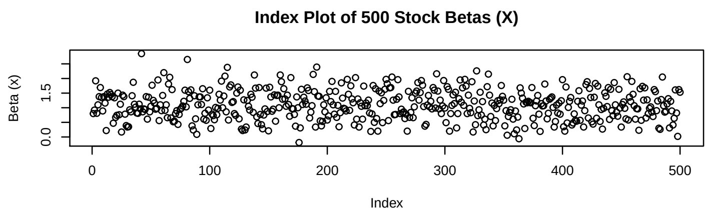
where $\overline{x} = \frac{1}{n}\sum_1^n x_i$ and $s_x^2 = \frac{1}{n}\sum_1^n (x_i - \overline{x})^2$

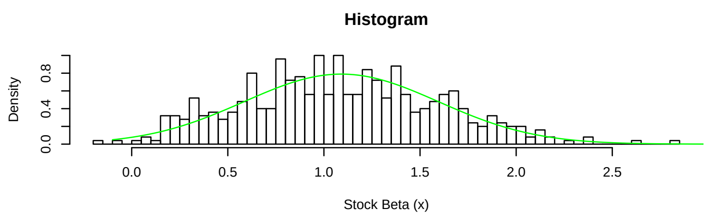
The Normal QQ plot for each set of standardized sample values is given in the next display but they are in a random order. For each distribution, identify the corresponding Normal QQ plot, and explain your reasoning.

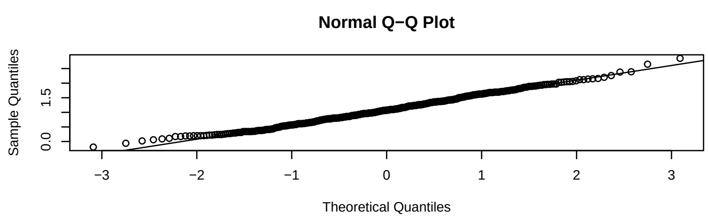
- $Uniform(0, 1)$ = Plot __
- $Exponential(1)$ = Plot __
- $Normal(50, 10)$ = Plot __
- Student's t (4 degrees of freedom) = Plot __



5. **Betas for Stocks in S&P 500 Index**. In financial modeling of stock returns, the Capital Asset Pricing Model associates a "Beta" for any stock which measures how risky that stock is compared to the "market portfolio". (Note: this name has nothing to do with the beta(a,b) distribution!) Using monthly data, the Beta for each stock in the S&P 500 Index was computed. The following display gives an index plot, histogram, Normal QQ plot for these Beta values.







For the sample of 500 $Beta$ values, $\overline{x}$ =1.0902 and s_x =0.5053.

(a). On the basis of the histogram and the Normal QQ plot, are the values consistent with being a random sample from a Normal distribution?

(b). Refine your answer to (a) focusing separately on the extreme low values (smallest quantiles) and on the extreme large values (highest quantiles).

Bayesian Analysis of a Normal Distribution. For a stock that is similar to those that are constituents of the S&P 500 index above, let $X = 1.6$ be an estimate of the Beta coefficient θ.

Suppose that the following assumptions are reasonable:

- The conditional distribution X given θ is Normal with known variance:
  $$X \mid \theta \sim Normal(\theta, \sigma_0^2), \text{ where } \sigma_0^2 = (0.2)^2.$$
- As a prior for θ, assume that θ is Normal with mean and variance equal to those in the sample
  $$\theta \sim Normal(\mu_{prior}, \sigma_{prior}^2)$$
  where $\mu_{prior} = 1.0902$ anad $\sigma_{prior} = 0.5053$

(c). Determine the posterior distribution of θ given $X = 1.6$.

(d). Is the posterior mean between X and μ_{prior}? Would this always be the case if a different value of X had been observed?

(e). Is the variance of the posterior distribution for θ given X greater or less than the variance of the prior distribution for θ? Does your answer depend on the value of X?