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

1. **True or False (and state why).**

   (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.

   Solution: T, T, F, T, T

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 $H_0 : p = p_0 = 0.5$.

   (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., $P(\chi^2_3 \leq q.90 = 6.25) = .90$)

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 $Z \sim N(0,1) : \Phi(z) = P(Z \leq z)$.)

Solution:

(a). The Pearson chi-square statistic for a multinomial distribution with $(m = 2)$ cells is

$$\mathcal{X}^2 = \sum_{j=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

where the observed counts are

$O_1 = x = 38$ and $O_2 = n - x = 62$,

and the expected counts under the null hypothesis are

$E_1 = n \times p_0 = n \times 1/2 = 50$ and $E_2 = (n - x) \times (1 - p_0) = (n - x) \times (1 - 1/2) = 50$

Plugging these in gives

$$\begin{aligned} \mathcal{X}^2 &= \sum_{j=1}^{m} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(38-50)^2}{50} + \frac{(62-50)^2}{50} \\ &= \frac{144}{50} + \frac{144}{50} = \frac{288}{50} = 5.76 \end{aligned}$$

(b). The approximate distribution of $\mathcal{X}^2$ is chi-squared with degrees of freedom $q = dim(\{p, 0 \leq p \leq 1\}) - dim(\{p : p = 1/2\}) = (m - 1) - 0 = 1$.

(c). The P-value of the Pearson chi-square statistic is the probability that a chi-square random variable with $q = 1$ degrees of freedom exceeds the 5.76, the observed value of the statistic. Since 5.76 is greater than $q.975 = 5.02$ and less than $q.99 = 6.63$, (the percentiles of the chi-square distribution with $q = 1$ degrees of freedom) we know that the P-value is smaller than $(1 - .975) = .025$ but larger than $(1 - .99) = .01$.

(d). Substituing $O_1 = x$ and $O_2 = (n - x)$

and $E_1 = n \times p_0 = n/2$ and $E_2 = n \times (1 - p_0) = n/2$

in the formula from (a) we get

$$\begin{aligned} \mathcal{X}^2 &= \sum_{j=1}^{m} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(x - n/2)^2}{n/2} + \frac{((n - x) - n/2)^2}{n/2} \\ &= \frac{(x - n/2)^2}{n/2} + \frac{((n/2 - x))^2}{n/2} \\ &= 2 \times \tfrac{(x-n/2)^2}{n/2} \\ &= \tfrac{4}{n} \times |x - n/2|^2 \end{aligned}$$

(e). Since X is the sum of n independent $Bernoulli(p)$ random variables,

$$E[X] = np \text{ and } Var(X) = np(1-p)$$

so by the CLT

$$X \sim N(np, np(1-p)) \text{ (approximately)}$$

which is $N(\frac{n}{2}, \frac{n}{4})$ when the null hypothesis ($p = 0.5$) is true.

The significance level of the test is the probability of rejecting the null hypothesis when it is true which is given by:

$$\begin{aligned}
\alpha = P(Reject\ H_0 \mid H_0) &= P(|X - n/2| > k \mid H_0) \\
&= P(|\tfrac{X-n/2}{\sqrt{n/4}}| > \tfrac{k}{\sqrt{n/4}} \mid H_0) \\
&\approx P(|N(0,1)| > \tfrac{k}{\sqrt{n/4}}) \\
&= 2 \times [1 - \Phi(\tfrac{k}{\sqrt{n/4}})]
\end{aligned}$$

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.

Solution: The random variable $S \sim Binomial(n, \theta)$. If $\theta \sim beta(a = 1, b = 1)$, then because the beta distribution is a conjugate prior for the binomial distribution, the posterior distribution of θ given S is

$$beta(a* = a + S, b* = b + (n - s))$$

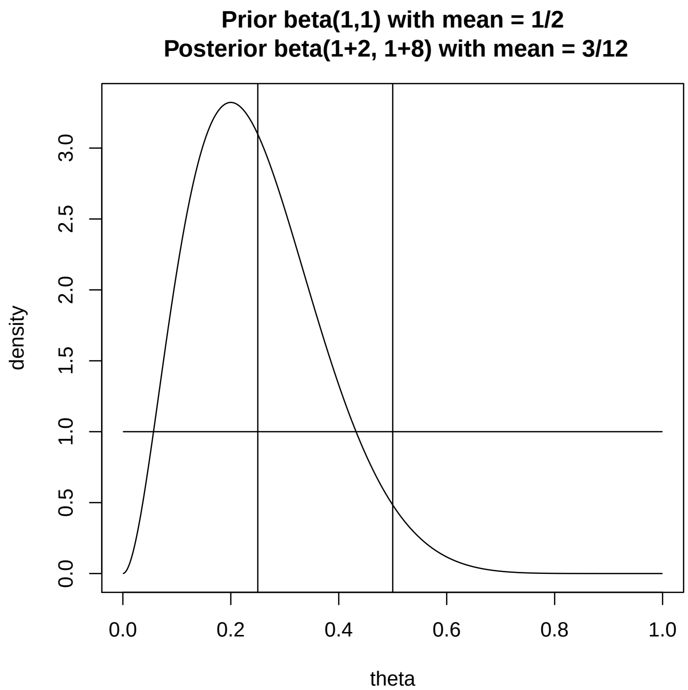
For $S = 2$, the posterior distribution of θ is thus $beta(a = 3, b = 9)$

Since the mean of a $beta(a, b)$ distribution is $a/(a + b)$, the prior mean is $1/2 = 1/(1 + 1)$, and the posterior mean is $3/12 = (a + s)/(a + b + n)$

These densities are graphed below

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

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



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

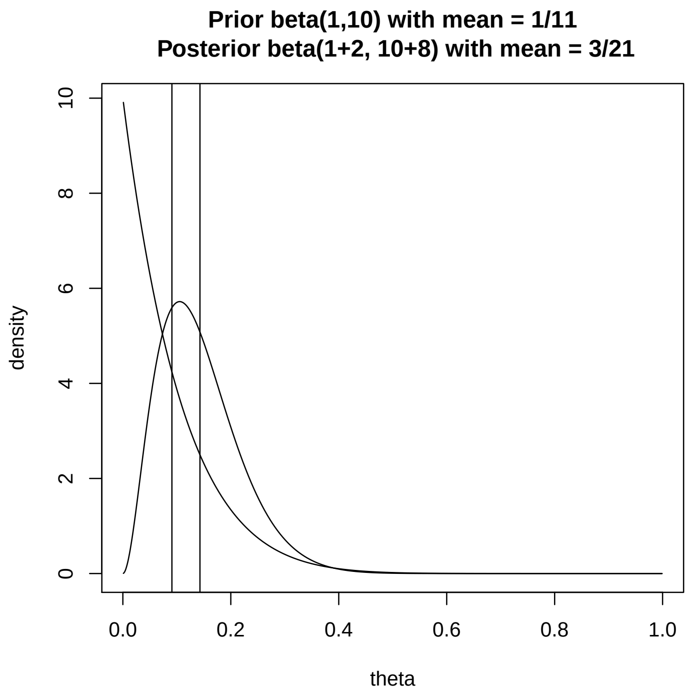
Solution: The random variable $S \sim Binomial(n, \theta)$. If $\theta \sim beta(a = 1, b = 10)$, then because the beta distribution is a conjugate prior for the binomial distribution, the posterior distribution of θ given S is

$$beta(a* = a + S, b* = b + (n - s))$$

For $S = 2$, the posterior distribution of θ is thus $beta(a = 3, b = 18)$

Since the mean of a $beta(a, b)$ distribution is $a/(a + b)$, the prior mean is $1/11 = 1/(10 + 1)$, and the posterior mean is $3/21 = (a + s)/(a + b + n)$

These densities are graphed below



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

Solution: The prior in (a) is a uniform distribution on the interval $0 < \theta < 1$. It is a flat prior and represents ignorance about θ such that any two intervals of θ have the same probability if they have the same width.

The prior in (b) gives higher density to values of θ closer to zero. The mean value of the prior in (b) is $1/11$ which is much smaller than the mean value of the uniform prior in (a) which is $1/2$.

(d). Suppose that X $S = 1$ 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.) Solution: The joint distribution of X and θ has pdf/cdf:

$$f(x, \theta) = f(x \mid \theta)\pi(\theta)$$

where $f(x \mid \theta)$ is the pmf of a $Bernoulli(\theta)$ random variable and $\pi(\theta)$ is the pdf of a $beta(a, b)$ distribution.

The marginal distribution of X has pdf

$$
\begin{aligned}
f(x) &= \int_0^1 f(x,\theta)d\theta \\
&= \int_0^1 \theta^x(1-\theta)^{1-x}\pi(\theta)d\theta \\
&= \int_0^1 \theta\pi(\theta)d\theta, \qquad for \;\; x=1 \\
and \qquad &= 1-\int_0^1 \theta\pi(\theta)d\theta \quad for \;\; x=0
\end{aligned}
$$

That is, X is $Bernoulli(p)$ with $p = \int_0^1 \theta\pi(\theta)d\theta = E[\theta \mid prior] = a/(a+b)$.

(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.)

Solution: The marginal distribution of X afer the sample is computed using the same argument as (c), replacing the prior distribution with the posterior distribution for θ given $S = s$.

$$X \text{ is } Bernoulli(p)$$

with

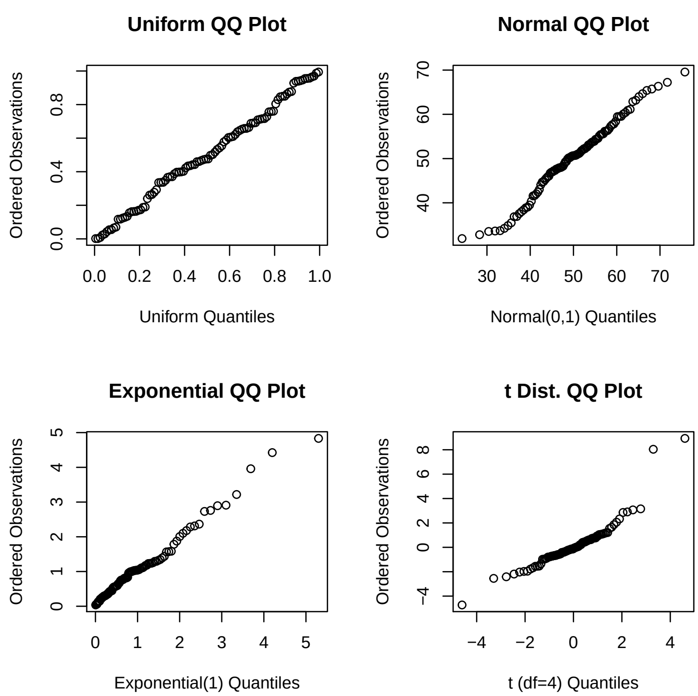
$$p = \int_0^1 \theta\pi(\theta \mid S)d\theta = E[\theta \mid S] = (a+s)/(a+b+n).$$

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:



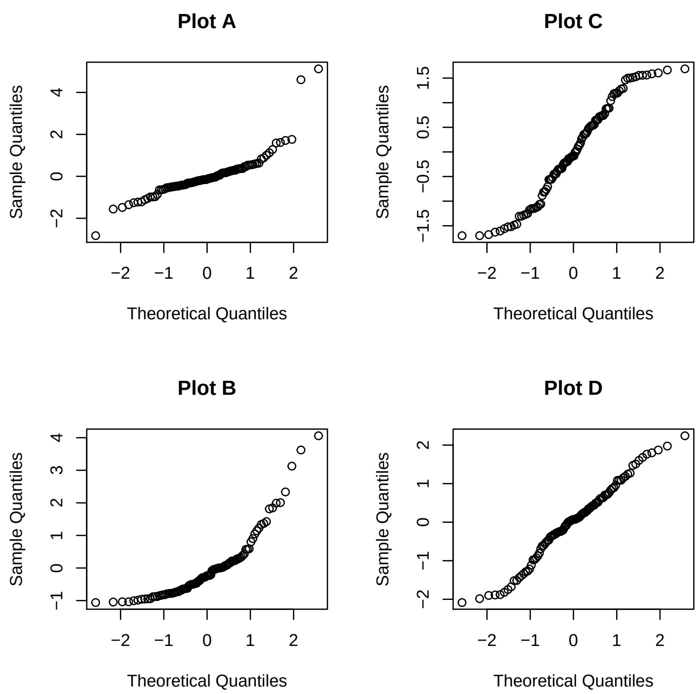
For each sample, the values were re-scaled to have sample mean zero and sample standard deviation 1

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

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 __



Solution:

The Student's t sample has two extreme high values and one extreme low value which are evident in Plot A, so

Plot A = t distribution

Plot B is the only plot that has a bow shape which indicates larger observations are higher than would be expected for a normal sample and smaller observations are less small than would be expected for a normal sample. This is true for the Exponential distribution which is asymmetric with a right-tail that is heavier than a normal distribution.

Plot B = Exponential.

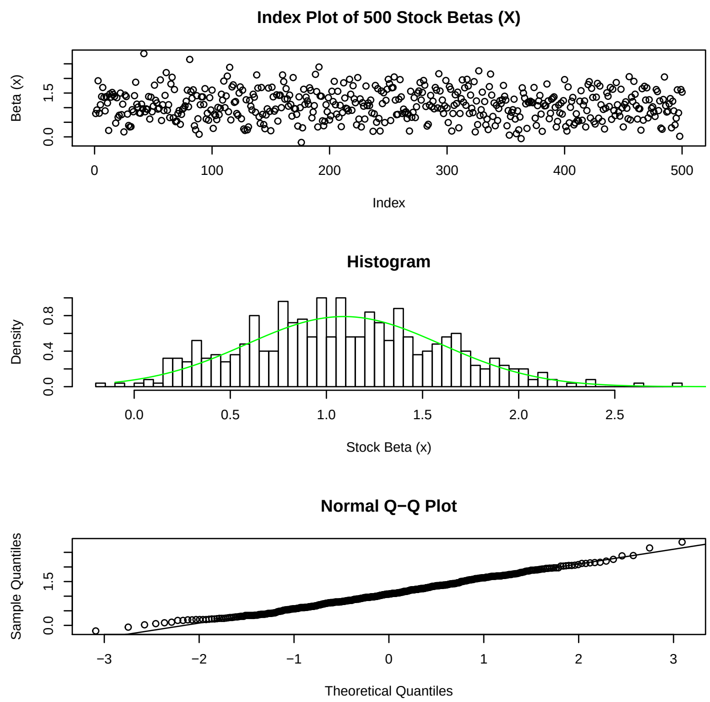
The $Uniform(0, 1)$ sample has true mean 0.5 and true variance equal to $E[X^2] - (E[X])^2 = 1/3 - (1/2)^2 = 1/12$. For a typical sample, the standardized sample values will be bounded (using the true mean and standard deviation to standardize, the values would no larger than $+(1 - .5)/\sqrt{1/12} = 1.73$). For Plot C the range of the standardized values is smallest, consistent with what would be expected for a sample from a uniform distribution.

Plot C = Uniform distribution.

The QQ Plot for the normal distribution is unchanged and follows a straight-line pattern indicating consistency of the ordered observations with the theoretical quantiles – distribution

Plot D = Normal

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?

Solution: Yes, the values are consistent with being a random sample from a Normal distribution. The normal QQ-plot is quite straight.

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

Solution: Consider the extremes of the distribution. The high positive points appear a bit higher than would be expected for a normal sample suggesting there are some outlier stocks with higher betas than would be expected under a normal model. The lowest values near zero appear a bit above the straight line through most of the ordered points, suggesting that the stocks with lowest beta values aren't as low as might be expected under a normal model.

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 $X = 1.6$.

Solution: This is the case of a normal conjugate prior distribution for the normal sample observation. The posterior distribution of θ is given by

$$[\theta \mid X = x] \sim N(\mu_*, \sigma_*^2)$$

where

$$\tfrac{1}{\sigma_*^2} = \tfrac{1}{\sigma_0^2} + \tfrac{1}{\sigma_{prior}^2}$$

and

$$\mu_* = \frac{(\frac{1}{\sigma_0^2})x + (\frac{1}{\sigma_{prior}^2})\mu_{prior}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_{prior}^2}}$$

Plugging in values we get

$$\tau_*^2 = (0.186)^2$$

$$\mu_* = \frac{(\frac{1}{.2^2})1.6 + (\frac{1}{.5053^2})1.0902}{(\frac{1}{.2^2}) + (\frac{1}{.5053^2})} = 1.531$$

(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?

Solution:

(d). Yes, the posterior mean is a weighted average of X and μ_{prior} which will always be between the two values.

(e). The variance of the posterior distribution $\tau_*^2 = (0.186)^2$ is less than $(.5053)^2 = \sigma_{prior}^2$. From part (c), the posterior variance does not vary with the outcome $X = x$.