Introduction to the Theory of Statistics Part 2

PM522b

Meredith Franklin

Division of Biostatistics, University of Southern California

Slides 1, 2015

Course Details

- Book: Statistical Inference, 2nd Ed. Casella G and Berger RL. Wadsworth & Brooks, 2002
- Lecture slides will be posted on Blackboard
- Additional handouts will be posted as we go along
- Chapter 5 properties of random samples, order statistics Chapters 6-12
- More on the theory of regression than presented in CB
- We will use R for computation and visualization
- Grading: Homework (7 @5% each, 35%), Midterm Exam (25%), Final Exam (40%)

Course Details

- Software: we will use R
- Intro to R posted on Blackboard
  - functions for distributions
  - writing custom functions
  - sampling data
  - simulating data
- Homework will mostly be handwritten solutions, but some computation
- Exams all handwritten, in preparation for the screening exam

Topics Covered

1. Introduction to statistics and statistical inference
   - Review of cdf, pmf, pdf
   - Bridging from probability to inference
2. Review of random variables, random samples, functions of random variables (CB Ch 5)
   - Relating samples to populations
   - Empirical distribution functions
   - Order statistics
   - Graphical representations of statistics

CDF

- First half of the PM522 series focused on probability and the development of cumulative distribution functions (cdf), probability mass functions (pmf), and probability distribution functions (pdfs).
- Recall the cumulative distribution function (cdf) for a discrete random variable:

$$F(x) = P(X \leq x), \forall x$$

  which has three conditions:

  1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
  2. $F(x)$ is a non-decreasing function of x
  3. $F(x)$ is right continuous

- For continuous random variables, $F(x)$ is a continuous function of X
- We can say a random variable X is continuous if $F(x)$ is a continuous function of x. Similarly a random variable X is discrete if $F(x)$ is a step function of x.

PMF

A probability mass function (pmf) evaluated at a value corresponds to the probability that a random variable takes that value.

- The pmf of a discrete random variable X:

$$f(x) = P(X = x), \forall x$$

To be a valid pmf, the probability must satisfy:

1. $f(x) \geq 0 \forall x$
2. $\sum_x f(x) = 1$ (the sum is taken over all values of x)
3. $P(X \in A) = \sum_{x \in A} f(x)$

Example

X is the result of flipping a coin where X=0 is tails and X=1 is heads. If the coin is fair, $P(x) = (1/2)^x (1/2)^{1-x}$ for $x = 0, 1$
If we do not know whether the coin is fair or not, $P(x) = \theta^x \theta^{1-x}$ for $x = 0, 1$

PDF

A probability density function (pdf) is a function associated with a continuous random variable. Areas under pdfs correspond to probabilities for a random variable.

- The pdf of a continuous random variable X is the function that satisfies:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

  And hence,

$$\frac{dF(x)}{dx} = f(x)$$

- Using the fundamental theorem of calculus, the derivative of the cdf is the pdf (when f(x) is continuous).
- To be a valid pdf, the function f must satisfy
  1. $f(x) \geq 0 \forall x$
  2. The area under $f(x)$ is one

Distributions and parameters

- In PM522a you learned specific types of discrete (Discrete Uniform, Hypergeometric, Binomial, Poisson, Negative Binomial, Geometric) and continuous (Uniform, Gamma, Exponential, Normal, Beta, Lognormal) distribution functions.
- The parameters of these functions were assumed to be known.
- Using a pdf with known parameters, we can say something about a random variable X

Example

$X \sim f_X(x|\theta), x \in R$ and $\theta \in \Theta$ are parameters
If $f_X(x|\theta)$ is the binomial distribution then we know $X \sim \text{binomial}(n, p)$ where n and p are our parameters
$\theta = (n, p)$
Furthermore we can calculate $E(X) = np$ and $V(X) = np(1 - p)$

Distributions and random samples

A Numerical Example

What is the probability that a family of 3 children will have 2 girls given that the probability of having a girl is 1/2?
In R: choose(3, 2) * 0.5^2*0.5^1 OR dbinom(2,3,1/2) =0.375

A Less Obvious Example

Suppose we toss a coin 10 times and observe 8 heads. What is the probability of heads?
If the coin was perfectly fair, then we could assume $\theta = 1/2$. But a) we don't know anything about the coin, and b) having flipped 8/10 heads does not support that P(heads)=1/2.

Distributions and random samples

- The examples above illustrate a sequence of n Bernoulli trials
- The distribution is often denoted Bernoulli(p) meaning there is only one parameter in the distribution because we known n
- The normal distribution has two parameters

Example

$X \sim f_X(x|\theta), x \in R$ and $\theta \in \Theta$ are parameters
If $f_X(x|\theta)$ is the normal distribution then we know $X \sim N(\mu, \sigma^2)$ where μ and σ are our parameters
$\theta = (\mu, \sigma^2)$
Furthermore we can find properties of these parameters $E(X) = \mu$ and $V(X) = \sigma^2$

Statistical Inference

- We need to bridge from probability to (inferential) statistics
- Populations to samples: data
- Experiments are performed to collect information (data) from which we can (imperfectly) understand the population
- A random sample is drawn from our population and we need a suitable function to describe the population from the sample
- We want to make *inference* about a population based on information contained in this random sample
- Always remember: the sample is NOT the population

Random Samples

- In statistics and statistical inference, we have random samples of X
- We don't know the pdf of X but want to be able to say something about its distribution
- A random variable could be represented by any possible pdf, however one model will be more probable than the others
- $\mathbf{X} = (X_1, X_2, ..., X_n)$ is a set of iid random variables with an unknown distribution function
- $X \sim f_X(x|\theta), x \in R$ and $\theta \in \Theta$ and we further define $\Theta \in R^d$ as the parameter space
- We regard $f_X(x|\theta)$ as the parametric model function

Random Samples

- The objective of statistical inference is thus to assess aspects of our unknown parameters θ given random samples
- Notation: X and X_i represent random variables; x and x_i represent observed values of the random variable X
- Notation: boldface denotes multiple variates where $\mathbf{X}$ represents random variables $(X_1, X_2, ..., X_n)$, and $\mathbf{x}$ represents observations $(x_1, x_2, ..., x_n)$
- There are three major components to statistical inference: point estimation, confidence/interval estimation, and hypothesis testing
- Point estimation is a single value estimate of θ_i computed from the data x
- Confidence estimation provides a set of values having a probability of including the true (but unknown) value of θ_i
- Hypothesis testing involves setting up a hypothesis about θ_i and assessing the plausibility of the hypothesis using the data x
- We will also focus on the theory of linear regression and anova in the second half of the term

Frequentist vs Bayesian Inference

Two types of inference exist: Frequentist and Bayesian
In the context of understanding the unknown parameter θ given random samples, we can describe the two approaches. Suppose the unknown parameter of interest is the mean μ of a normal distribution and we have observations $x_1, x_2, ...x_n$:

- Frequentist approach:
  - We do not make any further probabilistic assumptions on the parameter
  - Treat μ as a fixed but unknown constant
  - Use data reduction techniques to summarize the information in the sample (i.e. sample mean). This summary is a function which is also known as a statistic.
  - The data are a repeatable random sample. That is, sampling is infinite.
  - Assessment of the suitability of the estimate for our unknown parameter is based in how it would perform if done repeatedly (frequency interpretation)
  - That is, uncertainty in the estimate for μ

Frequentist vs Bayesian Inference

Two types of inference exist: Frequentist and Bayesian
In the context of understanding the unknown parameter θ given random samples, we can describe the two approaches. Suppose the unknown parameter of interest is the mean μ of a normal distribution and we have observations $x_1, x_2, ...x_n$:

- Bayesian approach:
  - Treat μ as having a probability distribution, not fixed
  - The prior distribution on the unknown parameter is either known, assumed on some information, or drawn from thin air
  - The uncertainty in μ is taken into account with the prior, without using the observations
  - Use Bayes' theorem to modify the probability of our unknown parameter given the observations
  - The posterior distribution is the modified prior distribution of the unknown μ

Random Variables, Functions, and Samples

- The classical, frequentist approach is concerned with experiments that are replicated a fixed number of times
- Replication means that each repetition is performed under identical conditions and is mutually independent (iid)
- We use the sample to extract information used to draw inferences about the population

Empirical Distribution Function

- For discrete probability distributions we can define the empirical distribution function (edf)

Empirical Distribution Function (edf)

Let our sample $x_1, x_2, ...x_n$ be iid random variables with cdf F_n
The edf associated with the sample $\hat{F}_n$ is the discrete distribution function defined by assigning probability $1/n$ to each x_i
Example edf: A fair die is rolled $n = 20$ times resulting in the sample $x = 1, 2, 3, 6, 3, 4, 5, 2, 5, 1, 2, 4, 4, 2, 3, 5, 6, 1, 2, 6$ the edf $\hat{P}_{20}$ assigns the probabilities:

x_i	$\#\mathrm{x}_i$	$\hat{P}_{20}(x_i)$
1	3	0.15
2	5	0.25
3	3	0.15
4	3	0.15
5	3	0.15
6	3	0.15

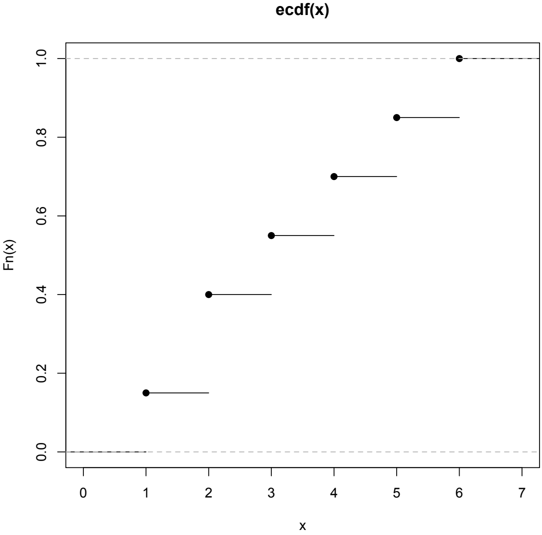
Empirical Distribution Function

- The true probabilities are 1/6 but the empirical probabilities range from 0.15 to 0.25
- The fact that the empirical probabilities $\hat{P}_n$ differ from P_n is sampling variation
- $\hat{P}_n(A) = \#\{x_i \in A\}\frac{1}{n}$
- The empirical cumulative distribution function associated with $\hat{P}_n$ is denoted $\hat{F}_n$

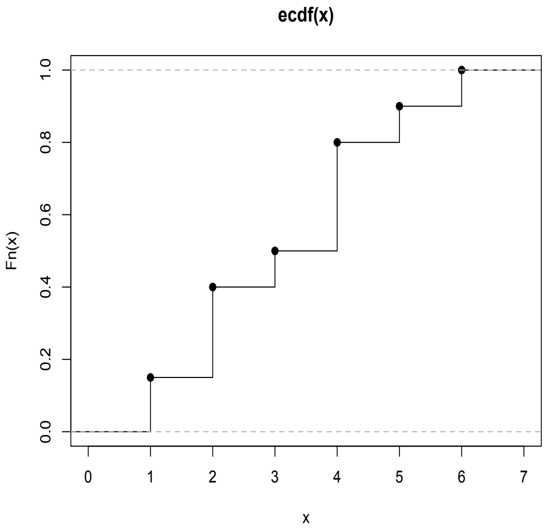
Definition: Empirical cdf

$$\hat{F}_n(a) = \hat{P}_n(X \leq a) = \frac{\#\{x_i \leq a\}}{n}$$

Empirical CDF



Empirical CDF



Relating samples to populations: Mean

- Expected values are another common estimate of the population from our random sample
- Let $E(X_i) = \mu$ denote the population mean
- We can use the plug-in principle to estimate the mean
- For our sample $x_1, x_2, ...x_n$, $\hat{\mu}_n = \sum_{i=1}^{n} \frac{x_i}{n}$

Example: mean of the empirical distribution

A fair die is rolled $n = 20$ times resulting in the sample
$x = \{1, 2, 3, 6, 3, 4, 5, 2, 5, 1, 2, 4, 4, 2, 3, 5, 6, 1, 2, 6\}$ the population mean is:
$\mu = E(X_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$
But the sample mean is 3.35
$\hat{\mu}_{20} \neq \mu$

Relating samples to populations: Variance

- Variance is another common estimate of the population from our random sample
- Let $V(X_i) = \sigma^2$ denote the population variance
- We can use the plug-in principle to estimate the variance of the empirical distribution
- For our sample $x_1, x_2, ...x_n$, $\hat{\sigma}_n^2 = \sum_{i=1}^{n} \frac{(x_i - \hat{\mu}_n)^2}{n}$

Example: variance of the empirical distribution

A fair die is rolled $n = 20$ times resulting in the sample
$x = \{1, 2, 3, 6, 3, 4, 5, 2, 5, 1, 2, 4, 4, 2, 3, 5, 6, 1, 2, 6\}$ the population variance is:
$\sigma^2 = E(X_i^2) - (E(X_i))^2 = \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} - 3.5^2 = 2.92$
But the sample variance is 1.73
$\hat{\sigma}_{20}^2 \neq \sigma^2$

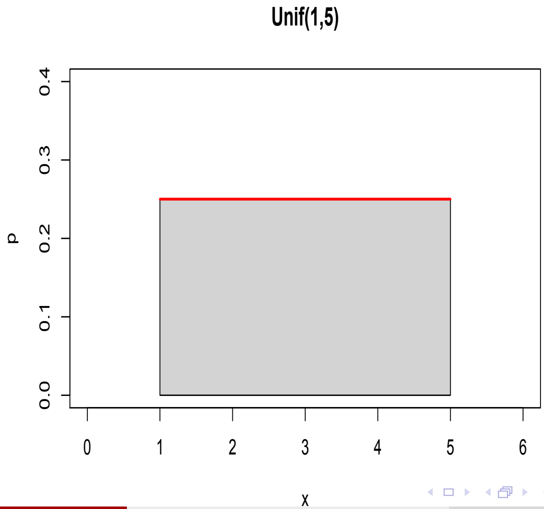
Relating samples to populations: Quantiles

- Quantiles are another common estimate of the population from our random sample
- The estimate of the population quantile is the corresponding quantile of the empirical distribution (e.g. median (2nd quantile or 50%) and interquartile range (3rd-1st quntile or 75%-25%))
- We can use the plug-in principle to estimate the quantiles of the empirical distribution

Example: quantiles of the empirical distribution

If we take $n = 20$ draws from a Uniform distribution $X \sim U(1,5)$ resulting in the sample x={4.92, 4.89, 1.93, 2.25, 3.08, 2.58, 3.91, 3.11, 2.56, 1.16, 3.55, 3.57, 1.16, 1.02, 2.20, 4.80, 4.94, 4.99, 2.68, 4.58} the population quantiles are:
$Pr[X \leq x] \geq q$ and $Pr[X \geq x] \geq 1 - q$ where q is the qth quantile, $0 < q < 1$
For a continuous r.v., $F(x) = q$, so for $X \sim U(1,5), F(x) = 1/2$ when $x = 3$

Relating samples to populations: Quantiles and the Uniform Distribution



Order Statistics

- Sample (empirical) quantiles are determined through order statistics.
- The order statistic of a random sample is denoted $X_{(1)}, X_{(2)}, ...X_{(n)}$ and satisfies $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ where $X_{(1)} = \min_{1 \leq i \leq n} X_i$
- For any number q between 0 and 1, the qth quantile is the observation that approximately nq of the observations are less than this observation and $n(1-q)$ are greater
- If nq is an integer, then the qth quantile is any real number such that $X_{(nq)} \leq X \leq X_{(nq+1)}$
- if nq is not an integer, then the qth quantile is $X_{\lceil nq \rceil}$ where $\lceil nq \rceil$ is the ceiling (smallest integer greater or equal to nq)
- The percentile is often used and is defined as the $100q$th sample percentile

Order Statistics

Example con't: quantiles of the empirical distribution

Recall our (ordered) random sample x={1.02, 1.16, 1.16, 1.93, 2.20, 2.25, 2.56, 2.58, 2.68, 3.08, 3.11, 3.55, 3.57, 3.91, 4.58, 4.80, 4.89, 4.92, 4.94, 4.99}

The median, $q = 0.5$ is any number between $x_{(10)} = 3.08$ and $x_{(11)} = 3.11$
The 25%ile, $q = 0.25$ is any number between $x_{(5)} = 2.20$ and $x_{(6)} = 2.25$
The 75%ile, $q = 0.75$ is any number between $x_{(15)} = 4.58$ and $x_{(16)} = 4.80$
The 99%ile $q = 0.99$ is $x_{(19.8)}$ which is $x_{(20)} = 4.99$ since $\lceil nq \rceil = \lceil 19.8 \rceil = 20$

Note: the population median (3) is not equal to the sample median $q = 0.5$ which is the mean of $x_{(10)} = 3.08$ and $x_{(11)} = 3.11$, $x = 3.095$

Order Statistics

- Note what we can have a non-unique median when nq is an integer
- This is commonly dealt with by the following:
- When n is odd then the empirical median is: $x_{\lceil n/2 \rceil}$
- When n is even then the empirical median is: $\frac{x_{(n/2)}+x_{n/2+1}}{2}$

Order Statistics: Discrete Distributions

- For a random sample $X_1, ..., X_n$ from a **discrete** distribution with pmf $f_X(x_i) = p_i$ and the possible values of X are in ascending order $x_1 < x_2 < ... < x_i$ then

$$\begin{aligned} P_0 &= 0 \\ P_1 &= p_1 \\ P_2 &= p_1 + p_1 \\ &\vdots \\ P_i &= p_1 + p_2 + ... + p_i \end{aligned}$$

- The order *statistics* from the sample are $X_{(1)}, X_{(2)}, ...X_{(n)}$, so:

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

and

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} [P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}$$

Order Statistics: Discrete Distributions

- To prove $P(X_{(j)} \leq x_i)$, fix i and define Y to be a random variable that is the count of the number of $X_1, ..., X_n$ that are less than or equal to x_i
- Thus, the event $\{X_{(j)} \leq x_i\}$ can be thought of as a success and $\{X_{(j)} > x_i\}$ can be thought of as a failure
- With these definitions of success and failures, Y is defined as the number of successes in n trials. In other words, Y $\sim$ Bin(n,P$_i$)
- Relating back to our X's, the event $\{X_{(j)} \leq x_i\}$ is equivalent to the event $\{Y \geq j\}$ and we express this with the Binomal probability
- $P(X_{(j)} \leq x_i) = P(Y \geq j)$ and following this, the equality $P(X_{(j)} = x_i) = P(X_{(j)} \leq x_i) - P(X_{(j)} \leq x_{i-1})$. Thus the two equations are established.

Order Statistics: Discrete Distributions

Example: Probability of a discrete order random variable

Suppose we roll a dice 15 times (independent rolls), $P(X_i = x) = 1/6$. What is the probability that the third largest roll is at least 5?

We have the ordered random variables $X_{(1)}, ..., X_{(15)}$ with the third largest being the 13th of the 15 rolls. Thus, we want to find $P(X_{(13)} \geq 5)$.
From the definition $P_i = p_1 + p_2 + ... + p_i$, We have $P_i = P(x < 5) = 4/6$

$$\begin{aligned} P(X_{(13)} \leq 5) &= \sum_{k=13}^{15} \binom{15}{k} (4/6)^k (1-4/6)^{15-k} \\ &= 105(2/3)^{13}(1/3)^2 + 15(2/3)^{14}(1/3) + (2/3)^{15} \\ &= 0.07936 \end{aligned}$$

Thus $P(X_{(13)} \geq 5) = 1 - P(X_{(13)} \leq 5) = 1 - 0.07936 = 0.92064$

Order Statistics: Continuous Distributions

- For a random sample with order statistics $X_{(1)}, X_{(2)}, ...X_{(n)}$ from a **continuous** distribution with cdf $F_X(x)$ and pdf $f_X(x)$.
  The CDFthe pdf of $X_{(j)}$ is:

$$f(X_{(j)}(x)) = \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1-F_X(x)]^{n-j}$$

- The proof of this lies in taking the derivative of the cdf of $X_{(j)}$ to obtain the pdf (see CB theorem 5.4.4)
- As in the discrete case, define Y to be a random variable that is the count of the number of $X_1, ..., X_n$ that are less than or equal to x
- Thus, the event $\{X_{(j)} \leq x\}$ can be thought of as a success
- With this definition of success, Y is defined as the number of successes in n trials. In other words, Y $\sim$ Bin(n,F$_x$(x))
- Although X is continuous, by this definition Y is a counting variable and is discrete

Order Statistics: Continuous Distributions

- From the pdf of $X_{(j)}$, $f(X_{(j)}(x))$ we can dissect it into three terms of interest:
  - $[F_X(x)]^{j-1}$ representing the j-1 sample items below x_i
  - $[1 - F_X(x)]^{n-j}$ representing the n-j sample items above x_i
  - f_X (x) representing the sample item near x_i

Example: Uniform Order Statistic

Suppose we have $X_{(1)}, X_{(2)}, ...X_{(5)}$ from a Uniform distribution on [0,1], what is the pdf of the the second order statistic? For Unif[0,1]:

$$f_X(x) = \begin{cases} 1, 0 \leq x \leq 1 \\ 0, \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0, x < 0 \\ x, 0 \leq x \leq 1 \\ 1, x > 1 \end{cases}$$

Order Statistics: Continuous Distributions

Example: Uniform Order Statistic, con't

$$f_{X_{(2)}}(x_2) = \frac{5!}{(2-1)!(5-2)!} f_X(x_2)[F_X(x_2)]^{2-1}[1 - F_X(x_2)]^{5-2}$$
$$= \begin{cases} 20x_2(1-x_2)^3, 0 \le x_2 \le 1 \\ 0, \text{otherwise} \end{cases}$$

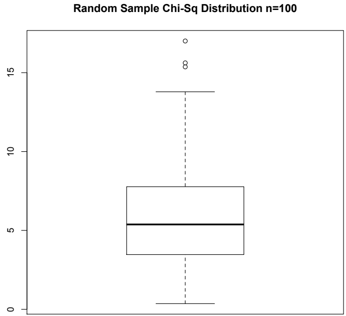
We also note that the jth order statistic from a uniform [0,1] has a beta(j, n-j+1) distribution

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j}$$
$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{(n-j+1)-1}$$

From which the expected value and variance for the uniform order statistics can be defined: $E(X_{(j)}) = \frac{j}{n+1}$ and $\text{Var}(X_{(j)}) = \frac{j(n-j+1)}{(n+1)^2(n+2)}$

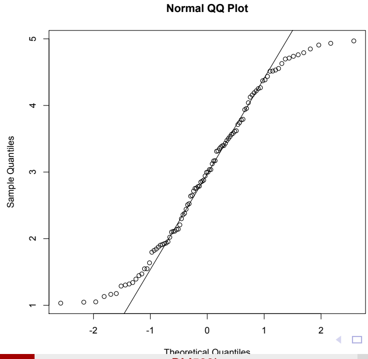
Graphical Representations

- Graphical uses of quantiles can be useful in determining aspects of the population from our random sample
- Box plots: gives an indication of symmetry of distribution
    - create a box around the 1st and 3rd quartile (25% and 75%)
    - add a line at the median (50%)
    - extend whiskers to extreme values (1.5 iqr or 5%-95%)
    - add outliers as points beyond the whiskers

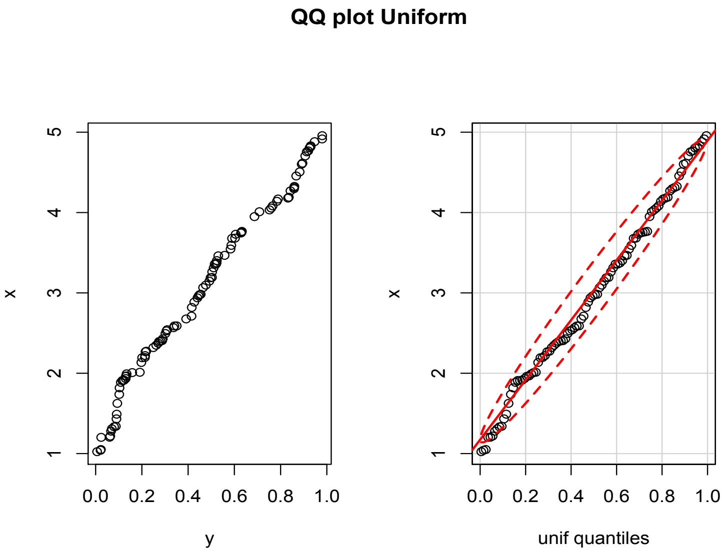


Graphical Representations

- Graphical uses of quantiles can be useful in determining aspects of the population from our random sample
- QQ plots: gives an indication of how close the ditribution of your random sample is to a theoretical distribution
    - called a normal QQ or normal probability plot when you compare to normal quantiles
    - QQ plot is similar to the EDF



Graphical Representations



Sampling from the Normal Distribution

Under the assumption of normality, there are a few properties of $\bar{X}$ and S^2 that are important. First, recall for our sample,

$$\bar{X} = \frac{1}{n}\sum_{i-1}^{n} X_i$$

$$S^2 = \frac{\sum_{i-1}^{n}(X_i - \bar{X})}{n-1}$$

For $X \sim (\mu, \sigma^2)$

1. $\bar{X}$ and S^2 are independent
2. $\bar{X} \sim N(\mu\sigma^2/n)$, namely $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2/n$
3. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

For 2, we recall that the sum of independently normally distributed random variables also has a normal distribution. Also, a linear transformation of a normally distributed variable is also normally distributed.

Sampling from the Normal Distribution

Proving 1, the independence between $\bar{X}$ and S^2, we look at n-1 deviations $(X_1 - \bar{X}, X_2 - \bar{X}, ..., X_{n-1} - \bar{X})$ and show that $\bar{X}$ is independent of $X_i - \bar{X}$ by showing $\text{Cov}(\bar{X}, X_i - \bar{X}) = 0$. Since S^2 is a function of $X_i - \bar{X}$ then it is independent of $\bar{X}$.