Section 4.4

Day 8 (2/7)

1 Quantiles (Percentiles)

Definition: Let X be a continuous random variable with cumulative distribution function $F(\cdot)$. For $0 < p < 1$, the pth quantile of X is defined to be

$$\xi_p = F^{-1}(p)$$

i.e. $F(\xi_p) = p$.

Special cases:

- $\xi_{.25} = Q_1$ is called the .25th quantile, or the 25th percentile or the 1st quartile
- $\xi_{.50} = Q_1$ is called the .50th quantile or the 50th percentile, or the 2nd quartile
- $\xi_{.75} = Q_1$ is called the .75th quantile, or the 75th percentile or the 3rd quartile

Theorem: Let $(Y_1, Y_2, \ldots, Y_n)$ be order statistics from a population with continuous cdf $F(\cdot)$ on the domain $\mathcal{D}$. For $k = 1, 2, \ldots, n$,

$$E[F(Y_k)] = \frac{k}{n+1}$$

Proof.

$$\begin{aligned} E[F(Y_k)] &= \int_{\mathcal{D}} F(y_k) g(y_k) dy_k \\ &= \int_{\mathcal{D}} F(y_k) \left[F(y_k)\right]^{k-1} f(y_k) \left[1 - F(y_k)\right]^{n-k} \frac{n!}{(k-1)!(n-k)!} dy_k \\ &= \frac{n!}{(k-1)!(n-k)!} \int_0^1 z^k (1-z)^{n-k} dz \end{aligned}$$

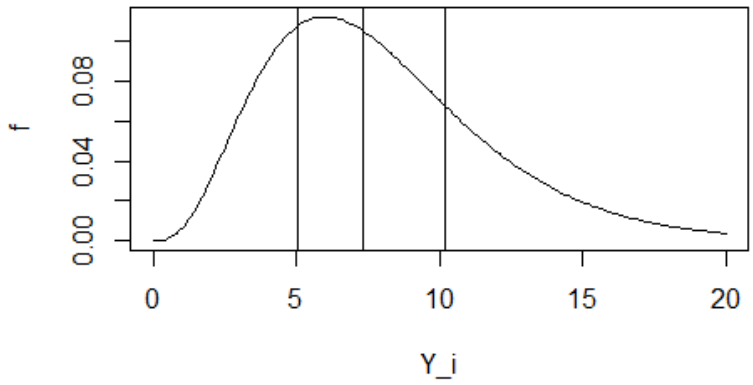
by letting $z = F(y_k)$, so that $dz = f(y_k)dy_k$. Comparing to the integral of a Beta pdf,

$$\begin{aligned} E[F(Y_k)] &= \frac{n!}{(k-1)!(n-k)!} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} \\ &= \frac{n!}{(k-1)!(n-k)!} \frac{k!(n-k)!}{(n+1)!} \\ &= \frac{k}{n+1} \end{aligned}$$

□

This makes intuitive sense. Given a random sample of size $n = 3$ from F, then the order statistics are expected to divide the area into quarters

j	$E[F(Y_j)]$
1	1/4
2	1/2
3	3/4



(Go to R simulation in the Appendix)

Since the theorem says $E[F(Y_k)] = k/(n+1)$, it follows that

$$E(Y_k) \doteq F^{-1}\left(\frac{k}{n+1}\right)$$

Therefore Y_k is an estimator of the quantile ξ_p where $p \doteq k/(n+1)$, or equivalently, $k = (n+1)p$.

1.1 Sample quantiles

Definition: Given order statistics $Y_1, \ldots, Y_n$ the pth sample quantile (also called the $100p$th sample percentile) is

$$Y_{[(n+1)p]}$$

where $[\,\cdot\,]$ is the largest integer smaller than the argument.

Comments:

1. The pth sample quantile is also called the $100p$th sample percentile.
2. In R, the sample quantile above is calculated using the function *quantile(x,p,type=1)*.
3. There are many *interpolation methods* for defining sample quantiles. The *quantile(x,p,type=6)* function in R calculates the pth quantile as follows. Let $r = (n+1)p - [(n+1)p]$. Then the pth sample quantile is
$$Y_{[(n+1)p]} + r\left(Y_{[(n+1)p]+1} - Y_{[(n+1)p]}\right)$$

Example 4.4.4 Suppose a random sample of size $n = 15$ yields the following ordered values. Calculate the .40th sample quantile.

56	70	89	94	96	101	102	102
102	105	106	108	110	113	116	

Solution: Using truncation, the .60th sample quantile is

$$Y_{[(15+1)(.40)]} = Y_{[6.4]} = Y_6 = 101$$

The interpolation method gives

$$Y_6 + .4\,(Y_7 - Y_6) = 101 + .4(102 - 101) = 101.4$$

Calculation in R:

```
> a<-c(56,70,89,94,96,101,102,102,102,105,106,108,110,113,116)
> quantile(a,.4, type=1)
40%
101
> quantile(a,.4, type=6)
  40%
101.4
```

1.2 Five-number summary

Definition: Given order statistics $Y_1, \ldots, Y_n$ the *five-number summary* is

$$(Y_1, Q_1, Q_2, Q_3, Y_n)$$

i.e. the minimum, the three quartiles, and the maximum.

Example 4.4.4 (con't.): Calculate the five-number summary.

The minimum is $Y_1 = 56$. Since $n + 1 = 16$, the first quartile is $Q_1 = Y_{[16(.25)]} = Y_4 = 94$. Simlarly $Q_2 = Y_{[16(.50)]} = Y_8 = 102$, and $Q_3 = Y_{[16(.75)]} = Y_{12} = 108$. Finally, the maximum is $Y_{15} = 116$, so that

$$(Y_1, Q_1, Q_2, Q_3, Y_n) = (56, 94, 102, 108, 116)$$

Based on the five-number summary, the data range from 56 to 116, the middle 50% of observations range from 94 to 108, and the middle of the data is 102.

1.3 The Boxplot

The *box-and-whisker plot*, or *boxplot*, is a quick plot of the data that makes use of the five-number summary. It has the following features:

- The box goes from Q_1 to Q_3 and encloses the middle 50% of the data. The length of this box is IQR=$Q_3 - Q_1$ and is called the interquartile range.
- A vertical line is then drawn at the median.
- Whiskers are drawn from the edge of the box to the farthest observations *within the fences*. The lower and upper fences are

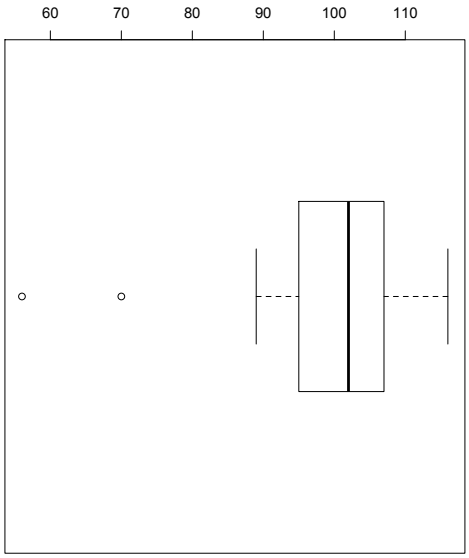
$$\text{LF} = Q_1 - 1.5(\text{IQR}) \text{ and } \text{UF} = Q_3 + 1.5(\text{IQR})$$

- Any observations outside the fences are potential outliers and denoted by a symbol 0.

Example 4.4.4 con't. Drawing a boxplot in R:

```
> a<-c(56,70,89,94,96,101,102,102,102,105,106,108,110,113,116)
> boxplot(a)
```

Figure 1: Boxplot of data in Example 4.4.4



1.4 Q-Q plot

Recall that if $Y_1, \ldots, Y_n$ are order statistics from a distribution with cdf $F(\cdot)$, then $E[F(Y_k)] = k/(n+1)$. Collectively, we expect

$$F(Y_1) \doteq \frac{1}{n+1}, F(Y_2) \doteq \frac{2}{n+1}, \ldots, F(Y_n) \doteq \frac{n}{n+1}$$

Suppose we suspect that the underlying distribution is $N(\mu, \sigma^2)$. Then $F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$ so that

$$\Phi(\frac{Y_1-\mu}{\sigma}) \doteq \frac{1}{n+1}, \Phi(\frac{Y_2-\mu}{\sigma}) \doteq \frac{2}{n+1}, \ldots, \Phi(\frac{Y_n-\mu}{\sigma}) \doteq \frac{n}{n+1}$$

and hence

$$\Phi^{-1}\left(\frac{k}{n+1}\right) \doteq \frac{Y_k-\mu}{\sigma}$$

for $k = 1, \ldots, n$. Graphically, the $X - Y$ plot of

$$\left[\Phi^{-1}\left(\frac{k}{n+1}\right), \frac{Y_k-\mu}{\sigma}\right]$$

should fall along the 45-degree line. More simply, the plot of

$$\left[\Phi^{-1}\left(\frac{k}{n+1}\right), Y_k\right]$$

should fall along a straight line. In other words, if the sample came from a distribution with cdf of the form $F\left(\frac{x-a}{b}\right)$ then the quantiles of F should be linearly related to the sample quantiles. This provides a way to check for whether the sample came from a presumed underlying distribution F.

Example 4.4.4 (con't.) To check whether the data came from a normal distribution, we plot the following points and compare to a straight line:

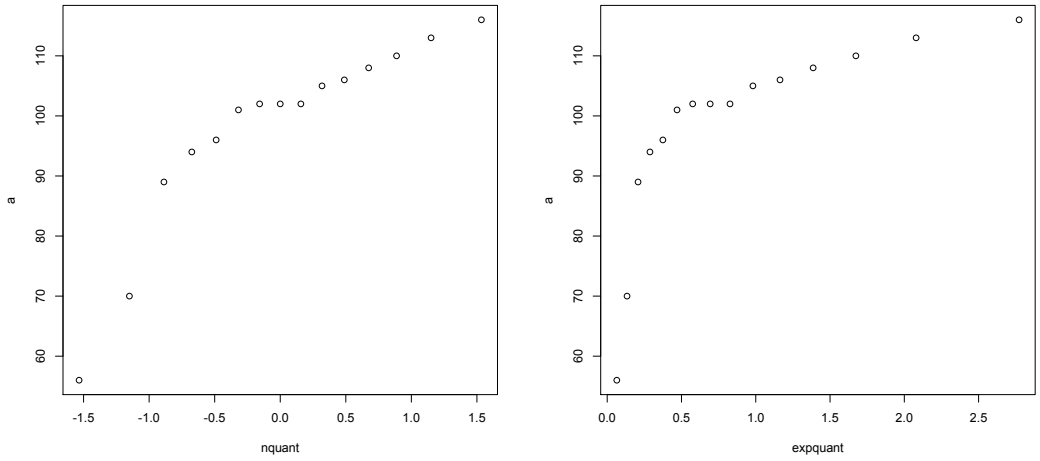
$$\left[\Phi^{-1}\left(\frac{1}{16}\right), 56\right], \left[\Phi^{-1}\left(\frac{2}{16}\right), 70\right], \ldots, \left[\Phi^{-1}\left(\frac{15}{16}\right), 116\right]$$

To check whether the data came from, say, an exponential distribution, replace Φ^{-1} above with F^{-1}, where F is cdf of exponential. Since $F(t) = 1 - e^{-t}$, then $F^{-1}(p) = -\ln(1-p)$. Or use the built-in F^{-1} function in R.

Example 4.4.4 con't. Draw q-q plots to check whether the data comes from Normal or Exponential

```
> a<-c(56,70,89,94,96,101,102,102,102,105,106,108,110,113,116)
>
> b<-(1:15)/16       #Calculate k/(n+1)
> b
 [1] 0.0625 0.1250 0.1875 0.2500 0.3125 0.3750 0.4375 0.5000 0.5625 0.6250 0.6875
[12] 0.7500 0.8125 0.8750 0.9375
> nquant<-qnorm(b)  #Calculate Normal quantiles
> plot(nquant, a)   #Plot Normal quantiles against sample quantiles
>
> expquant<-qexp(b)    #Calculate Exponential quantiles
> plot(expquant, a)    #Plot Exponential quantiles against sample quantiles
```

Figure 2: Normal and Exponential q-q plots



1.5 Confidence interval for quantiles

Recall that the pth quantile of X is $\xi_p = F^{-1}(p)$ i.e. $F(\xi_p) = p$. For a sample of size n from F, let $Y_1, Y_2, \ldots, Y_n$ be the order statistics. A point estimate of ξ_p is the pth sample quantile $Y_{[(n+1)p]}$. We now discuss a way to construct a confidence interval for ξ_p based on the order statistics.

Example 4.4.7 (Example 4.4.4 con't.) Recall that we have a sample of size $n = 15$ and the order statistics are given below. Construct a confidence interval for the population median $\xi_{.50}$.

$$\begin{array}{cccccccc} 56 & 70 & 89 & 94 & 96 & 101 & 102 & 102 \\ 102 & 105 & 106 & 108 & 110 & 113 & 116 & \end{array}$$

Solution. Note that

$$\begin{aligned} P\left(Y_5 < \xi_{.50} < Y_{11}\right) &= P\left(5 \text{ or more } X \text{ values are less than } \xi_{.50} \textit{ and } 10 \text{ or fewer are less than } \xi_{.50}\right) \\ &= \sum_{j=5}^{10} P(B = j), \text{ where } B \sim \text{Bin(n=15, p=.50)} \\ &= .88 \end{aligned}$$

so that (Y_5, Y_{11}) is an 88% confidence interval for the median $\xi_{.50}$. Similarly, we calculate coverage probabilities of different intervals

Confidence Interval	Confidence level
(Y_5, Y_{11})	.88
(Y_4, Y_{11})	.92
(Y_5, Y_{12})	.92
(Y_4, Y_{12})	.96

■

Theorem: Let $(Y_1, Y_2, \ldots, Y_n)$ be order statistics from a population with continuous cdf F. Then for $0 < p < 1$,

$$P\left(Y_i < \xi_p < Y_j\right) = \sum_{w=i}^{j-1} P(B = j), \text{ where } B \sim \text{Bin(n, p)}$$
$$= \sum_{w=i}^{j-1} \binom{n}{w} p^w (1-p)^{n-w}$$

Homework 8:

1. Exercise 4.4.25
2. Exercise 4.4.28

R simulation of $E[F(Y_k)]$

```
> # Draw n=3 from N(50, 10)
> set.seed(4321)
> sort(rnorm(3,mean=50,sd=10))
[1] 45.73243 47.76388 57.17607
> # Calculate the area to the left
> pnorm(c(45.73, 47.76, 57.17), mean=50,sd=10)
[1] 0.3346897 0.4113787 0.7633130
> # The theorem says on the average, these will be (.25,.50,.75)
>
> # Now repeat 40 times and take average
> set.seed(4321)
> m2<-matrix(rnorm(3*40,mean=50,sd=10), byrow=TRUE, ncol=3)
> m3<-apply(m2,1,sort)   #Sort within rows
> m3
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
[1,] 45.73243 48.71643 47.02833 42.81302 37.39015 38.24885 34.11738 49.91070
[2,] 47.76388 58.41446 51.96005 49.32764 37.78218 50.73478 42.52619 49.96974
[3,] 57.17607 66.09347 62.40746 53.44367 61.39464 65.73316 54.83522 55.93358
         [,9]    [,10]    [,11]    [,12]    [,13]    [,14]    [,15]    [,16]
[1,] 47.61966 36.26030 51.40279 42.48481 48.24334 48.53204 42.28962 51.00105
[2,] 49.00798 41.66190 56.62126 45.24888 52.37211 61.83290 54.19156 58.42398
[3,] 50.47783 52.96513 61.31040 58.52414 57.51964 62.67496 57.14528 60.37409
        [,17]    [,18]    [,19]    [,20]    [,21]    [,22]    [,23]    [,24]
[1,] 50.19726 52.09720 51.48414 38.18396 44.70630 45.80346 42.20837 36.91364
[2,] 50.90360 64.74833 60.39212 50.46477 48.75776 49.05093 50.48157 51.82745
[3,] 52.30993 70.80248 67.18528 60.28431 49.80894 58.23488 52.14547 57.29065
        [,25]    [,26]    [,27]    [,28]    [,29]    [,30]    [,31]    [,32]
[1,] 44.28851 44.48092 25.49984 50.31526 25.60680 39.28311 39.60584 48.40305
[2,] 57.64100 52.02619 40.72632 55.64471 55.85280 46.57207 44.12216 63.08704
[3,] 59.26927 54.32404 51.86917 68.67515 62.45953 50.42843 51.55513 68.56493
        [,33]    [,34]    [,35]    [,36]    [,37]    [,38]    [,39]    [,40]
[1,] 33.79804 55.43005 41.63647 39.13821 42.71048 37.75895 36.64560 40.47263
[2,] 49.58546 58.04652 47.47122 52.43426 57.15838 40.34103 44.90040 50.16356
[3,] 66.89409 60.56656 64.03281 64.13896 65.50334 43.61227 61.46552 51.75032

> m4<-pnorm(m3,mean=50,sd=10)   # Apply CDF
> apply(m4,1,mean)
[1] 0.2833679 0.5415392 0.7710309
```

R simulation of Example 4.4.7 (confidence interval for median)

```
> # Draw n=15 from N(50, 10) [1000 times]
> set.seed(4620)
> m5<-matrix(rnorm(15*1000,mean=50,sd=10), byrow=TRUE, ncol=15)
> for (i in (1:1000)){m5[i,]<-sort(m5[i,])}   #Sort within rows
> head(m5, n=20)
          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
 [1,] 17.88581 36.78729 40.54091 41.68576 44.52682 45.28793 46.17200 47.63912
 [2,] 41.89124 45.54545 47.06323 48.91430 49.02108 49.43477 51.69051 51.69914
 [3,] 34.81414 39.34964 44.19606 44.23190 46.46846 47.08980 47.53320 47.83171
 [4,] 30.47548 34.38069 35.10455 39.01921 42.33776 43.17207 44.41712 45.74908
 [5,] 28.89853 30.86142 41.62095 46.78307 46.95241 48.30782 48.62136 48.74436
 [6,] 37.10378 40.34033 41.32040 43.66377 48.33588 48.39084 48.41661 48.67023
 [7,] 33.08612 38.36518 39.59152 43.08303 43.72503 43.82536 44.30809 45.84291
 [8,] 36.24410 41.03497 42.38523 42.76117 43.60378 44.16188 46.91753 48.41747
 [9,] 38.04349 39.63772 39.99042 40.38919 41.65635 44.16263 45.55332 46.27501
[10,] 37.36846 39.44652 39.93094 42.22524 42.29618 43.83049 46.88907 48.55857
[11,] 30.75324 40.62151 45.56832 46.00827 46.54933 46.74879 49.06279 49.88751
[12,] 36.56361 37.09939 40.74262 41.46406 46.86732 46.98230 48.09429 49.23607
[13,] 33.48457 33.93648 35.28079 38.71828 41.49487 43.18269 43.52061 45.05162
[14,] 29.63935 33.78501 38.67712 40.02562 46.84669 48.10387 48.57025 52.62359
[15,] 35.47360 39.90698 41.50730 45.83174 46.84642 49.51874 49.55556 50.65318
          [,9]    [,10]    [,11]    [,12]    [,13]    [,14]    [,15]
 [1,] 56.27999 56.48514 57.48322 58.36972 60.07842 60.62316 66.04873
 [2,] 52.84193 55.13958 55.45283 62.14647 62.31566 66.11343 70.84980
 [3,] 50.73802 52.78208 52.95212 54.53134 56.47979 58.49865 60.37340
 [4,] 45.87925 47.44299 50.17336 51.26176 55.89564 62.91314 64.95676
 [5,] 49.79084 49.86732 51.87419 53.93631 57.74183 59.35399 64.57303
 [6,] 49.02708 49.26062 49.62747 55.52441 56.91864 57.67922 68.98423
 [7,] 47.87785 49.04429 50.57622 51.18651 52.85999 55.20088 69.80480
 [8,] 52.15543 52.24337 52.32120 54.93222 55.95468 64.05503 76.81176
 [9,] 46.34682 49.17773 49.86216 51.62961 53.50981 58.11889 63.40196
[10,] 50.67895 53.95362 54.63043 59.25396 63.72291 65.34428 67.94885
[11,] 51.39008 51.60333 52.10218 53.90596 54.51686 57.38464 58.09550
[12,] 49.74052 49.97341 50.16547 50.62983 54.34116 60.32591 66.32068
[13,] 45.46668 46.13692 54.01575 54.17334 55.35044 68.01376 71.00739
[14,] 53.05237 55.23646 55.49834 57.20082 58.76321 59.72078 63.19818
[15,] 54.22943 54.97939 55.43109 56.20493 60.79737 67.40060 71.37653

> Y5<-m5[,5]
> Y11<-m5[,11]
> head(Y5<50 & Y11>50, n=15)
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
[14]  TRUE  TRUE
> mean(Y5<50 & Y11>50)
[1] 0.884
```