Serik Sagitov, Chalmers and GU, February 14, 2018

Solutions chapter 10

Problem 10.2

IID sample $(X_1, \ldots, X_n)$ from the uniform distribution $\mathrm{U}(0,1)$.

(a) For a fixed x, the empirical distribution function $F_n(x)$ is the sample proportion estimate of $p = F(x) = x$. The variance of $F_n(x)$ is

$$\sigma^2_{F_n(x)} = \frac{p(1-p)}{n} = \frac{x(1-x)}{n},$$

the that

$$\sigma_{F_n(x)} = \sqrt{\frac{x(1-x)}{n}}, \quad x \in [0,1].$$

(b) Generate 100 samples of size $n = 16$. For each sample plot $F_n(x) - F(x)$ and relate what you see to your answer to (a). Matlab code

```
x=rand(16,100);
y=sort(x)';
for k=1:100
plot(y(k,:),(1:16)/16-y(k,:),'.')
hold on
end
```

Problem 10.5

We have

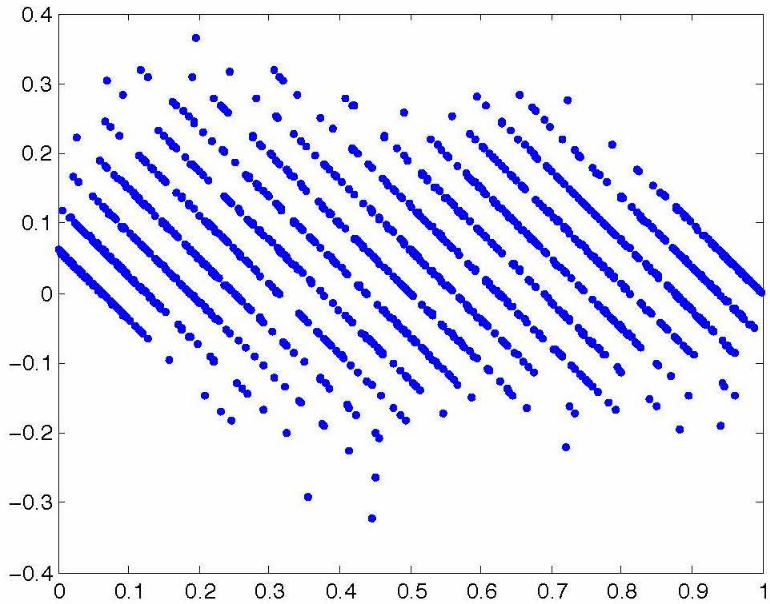
$$F_n(u) = \frac{1}{n}[1_{\{X_1 \le u\}} + \ldots + 1_{\{X_n \le u\}}], \quad \mathrm{E}(F_n(u)) = F(u),$$
$$F_n(v) = \frac{1}{n}[1_{\{X_1 \le v\}} + \ldots + 1_{\{X_n \le v\}}], \quad \mathrm{E}(F_n(v)) = F(v).$$

Assuming $u < v$, we get

$$\begin{aligned}
\mathrm{E}(F_n(u) \cdot F_n(v)) &= \frac{1}{n^2}[\sum_{i=1}^n \mathrm{E}(1_{\{X_i \le u\}} 1_{\{X_i \le v\}}) + \sum_{i=1}^n \sum_{j \ne i} \mathrm{E}(1_{\{X_i \le u\}} 1_{\{X_j \le v\}})] \\
&= \frac{1}{n^2}[\sum_{i=1}^n F(u) + \sum_{i=1}^n \sum_{j \ne i} F(u)F(v)] \\
&= \frac{1}{n}[F(u) + (n-1)F(u)F(v)].
\end{aligned}$$

Finish by using

$$\mathrm{Cov}(F_n(u), F_n(v)) = \mathrm{E}(F_n(u) \cdot F_n(v)) - \mathrm{E}(F_n(u)) \cdot \mathrm{E}(F_n(v)).$$



Problem 10.6

Ordered sample $n = 59$

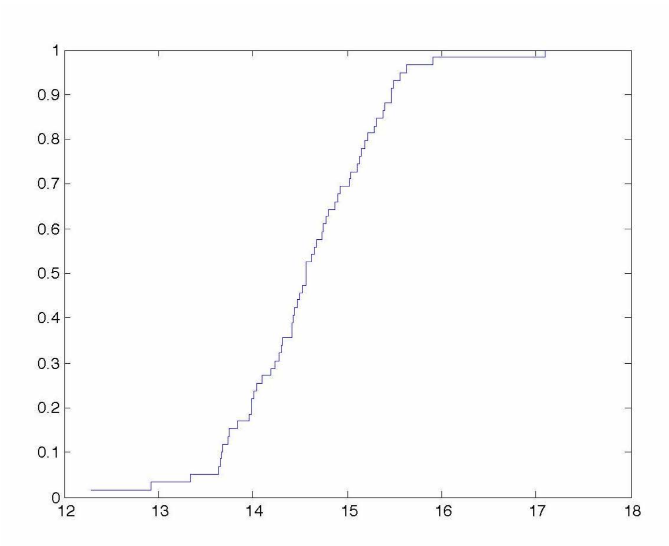
12.28 12.92 13.33 13.64 13.65 13.66 13.68	
13.73 13.75 13.83 13.96 13.98 13.98 14.01	
14.04	25% quantile
14.10 14.19 14.23 14.27 14.30 14.32 14.41	
14.41 14.43 14.44 14.47 14.49 14.52 14.56	
14.57	50% quantile
14.57 14.62 14.65 14.68 14.73 14.75 14.77	
14.80 14.87 14.90 14.92 15.02 15.03 15.10	
15.13	75% quantile
15.15 15.18 15.21 15.28 15.31 15.38 15.40	
15.47 15.47 15.49 15.56 15.63 15.91 17.09	

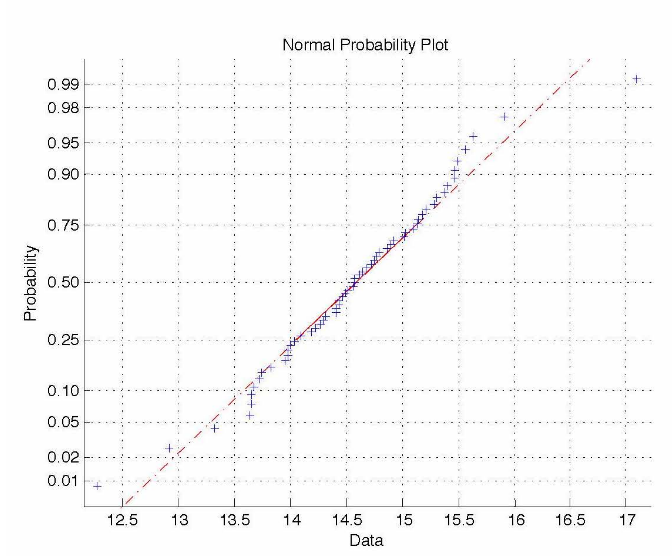
(a) Use Matlab commands

```
x=data vector;
stairs(sort(x),(1:length(x))/length(x))   % empirical cdf
hist(x)                                   % histogram, the same as hist(x,10)
normplot(x)                               % normal probability plot
prctile(x,90)                             % 0.90-quantile
```

The distribution appears to be rather close to normal. The 10% quantile

$$\frac{X_{(6)} + X_{(7)}}{2} = \frac{13.66 + 13.68}{2} = 13.67.$$





(b) Since $\bar{x} = 14.58$ and $s = 0.78$, the one-sided 99% of the population distribution for the natural wax is

$$(-\infty, 14.58 + 2.33 \cdot 0.78) = (-\infty, 16.40).$$

Expected means

1% dilution	$\mu_1 = 14.58 \cdot 0.99 + 85 \cdot 0.01 = 15.28$	can not be detected
3% dilution	$\mu_3 = 14.58 \cdot 0.97 + 85 \cdot 0.03 = 16.69$	can be detected
5% dilution	$\mu_5 = 14.58 \cdot 0.95 + 85 \cdot 0.05 = 18.10$	can be detected

Problems 10.11, 10.13, 10.14.

$$1 - F(t) = e^{-\alpha t^\beta}, \quad f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^\beta}, \quad h(t) = \alpha\beta t^{\beta-1}.$$

- If $\beta = 1$, then $h(t) = \alpha$ is constant and the distribution is memoryless.
- If $0 < \beta < 1$, then $h(t)$ decreases with t meaning that the longer you live the healthier you become.
- If $\beta > 1$, then $h(t)$ increases with t meaning that the older individuals die more often than the younger.

Problem 10.29

Stem and leaf display for $n = 26$ observations including $k = 5$ outliers:

133:7
134:134
135:002244
135:88

136:36
High: 141.2, 143.3, 146.5, 147.8, 148.8

Let N be the number of outliers in a non-parametric bootstrap sample.

(a) Due to sampling with replacement we have $N \sim$ Bin $(26, \frac{5}{26})$.

(b) Find P($N \geq 10$):

$$\mathrm{P}(N \leq 9) = \mathrm{binocdf}(9, 26, 5/26) = 0.9821,$$
$$\mathrm{P}(N \geq 10) = 1 - 0.9821 = 0.018.$$

(c) In $B = 1000$ bootstrap samples, we expect

$$B \cdot \mathrm{P}(N \geq 10) = 18$$

samples to contain 10 or more of outliers.

(d) The probability that a bootstrap sample is composed entirely of these outliers is

$$\mathrm{P}(N = 25) = (5/26)^{26} = 2.4 \cdot 10^{-19}.$$

Problem 10.37

Same data as in Problem 10.6.

(a) The Matlab commands

```
trimmean(x,10)
trimmean(x,20)
```

give $\bar{X}_{0.1} = 14.586$ and $\bar{X}_{0.2} = 14.605$.

> m = trimmean(X,percent) calculates the trimmed mean of the values in X. For a vector input, m is the mean of X, excluding the highest and lowest k data values, where k=n*(percent/100)/2 and where n is the number of values in X.

(b) An approximate 90% CI for the mean is

$$14.58 \pm 1.645 \cdot \frac{0.78}{\sqrt{59}} = 14.58 \pm 0.17 = (14.41; 14.75)$$

(c) Nonparametric 90% CI for the population median M is $(X_{(k)}, X_{(60-k)})$, where $\mathrm{P}(Y < k) = 0.05$ and $Y \sim$ Bin $(59, 0.5)$. Applying the normal approximation for Bin (n, p) with continuity correction

$$\mathrm{P}(Y < k) = \mathrm{P}(Y \leq k - 1) \approx \Phi\Big(\frac{k - 0.5 - np}{\sqrt{np(1-p)}}\Big),$$

we arrive at equation

$$\frac{k - 0.5 - \frac{59}{2}}{\sqrt{\frac{59}{4}}} = -1.645,$$

that gives $k = 24$. This yields

$$(X_{(k)}, X_{(60-k)}) = (X_{(24)}, X_{(36)}) = (14.43; 14.75).$$

(d) The Matlab commands for the non-parametric bootstrap

```
n=59; B=1000;
z=x(random('unid',n,n,B));  % ('unid',n) - uniform discrete [1, n], 1000 samples of size n
t1=trimmean(z,10);
t2=trimmean(z,20);
std(t1)
std(t2)
```

give the standard errors 0.1034 and 0.1004 for $\bar{X}_{0.1}$ and $\bar{X}_{0.2}$ respectively.

(e) Matlab commands

```
c11=prctile(t1,5)
c12=prctile(t1,95)
c21=prctile(t2,5)
c22=prctile(t2,95)
```

give 90% CIs

$$\text{for } \mu_{0.1}: \quad (2\bar{X}_{0.1} - c12; 2\bar{X}_{0.1} - c11) = (14.435; 14.765),$$
$$\text{for } \mu_{0.2}: \quad (2\bar{X}_{0.2} - c22; 2\bar{X}_{0.2} - c21) = (14.463; 14.784).$$

(f) Matlab commands

```
iqr(x)
median(abs(x-median(x)))
```

Warning: mad(x) in Matlab stands for the mean abs. dev.

(g) Matlab commands (vector z comes from the (d) part)

```
q=prctile(z,75);
hist(q)
std(q)
```

give the standard error 0.1332 of the upper quartile.

Problem 10.40

Matlab command (x = control and y = seeded data)

```
qqplot(x,y)
```

fits the line $y = 2.5x$ claiming 2.5 times more rainfall from seeded clouds.

Matlab command

```
qqplot(log(x),log(y))
```

fits the line

$$\log(y) = 2 + 0.8 \log(x)$$

meaning a decreasing slope in the relationship $y = 7.4x^{0.8}$.