NOTE: The notes is a summary for materials discussed in the class and is not supposed to substitute the textbook. Materials that appear in the textbook but do not appear in the lecture notes might still be tested. Please send me an email if you find typos.
1 Ordered statistics (deterministic)

Let $x_1, \ldots, x_n$ be unsorted real numbers (not necessarily distinct). Recall that the ordered statistics are

\begin{align*}
y_1 &:= \text{smallest of } x_1, x_2, \ldots, x_n; \\
y_2 &:= \text{second smallest of } x_1, x_2, \ldots, x_n; \\
& \vdots \\
y_n &:= \text{largest of } x_1, x_2, \ldots, x_n.
\end{align*}
2 Ordered statistics (random)

Let $X_1, \ldots, X_n$ be random numbers now. The ordered statistics are

\[
Y_1 := \text{smallest of } X_1, X_2, \ldots, X_n;
\]
\[
Y_2 := \text{second smallest of } X_1, X_2, \ldots, X_n;
\]
\[
\vdots
\]
\[
Y_n := \text{largest of } X_1, X_2, \ldots, X_n.
\]

We call $Y_k$ the \textbf{k-th order statistic} of $X_1, \ldots, X_n$.

Note that $Y_k$ are all random numbers since $X_1, \ldots, X_n$ are random numbers.
3 Ordered statistics: Example

Let $X_1$ and $X_2$ be two independent Bernoulli random variables with success probability $\frac{1}{2}$,

$$P[X_i = 0] = P[X_i = 1] = \frac{1}{2}.\]

There are four possibilities, each with equal probability:

$X_1 = X_2 = 0$, which implies $Y_1 = Y_2 = 0$;

$X_1 = 0, X_2 = 1$, which implies $Y_1 = 0, Y_2 = 1$;

$X_1 = 1, X_2 = 0$, which implies $Y_1 = 0, Y_2 = 1$;

$X_1 = X_2 = 1$, which implies $Y_1 = Y_2 = 1$;
In conclusion, we $Y_1$ and $Y_2$ have distribution

\[ P[Y_1 = 0] = \frac{3}{4}; \quad P[Y_1 = 1] = \frac{1}{4}; \]
\[ P[Y_2 = 0] = \frac{1}{4}; \quad P[Y_2 = 1] = \frac{3}{4}. \]

Note that $Y_1$ and $Y_2$ are NOT independent.
4 Cumulative distributive function

Recall that the cumulative distributive function (cdf) of $X$ is the function $F_X : \mathbb{R} \to [0, 1]^1$

$$F_X(x) := P[X \leq x] \quad \text{for all real number } x.$$

We now compute the cdf of $Y_1, \ldots, Y_n$.

---

1This means that $F_X$ is a function that maps real numbers to another real number that is between 0 and 1.
5 Cdf of $Y_n$

Lemma 1. Let $X_1, \ldots, X_n$ be independent, identical random variables. Then, for any real number $y$,

$$P[Y_n \leq y] = (P[X_1 \leq y])^n.$$ 

Proof. We have (BT)

$$P[Y_n \leq y] = P[X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y]$$

$$= P[X_1 \leq y] P[X_2 \leq y] \cdots P[X_n \leq y]$$

$$= P[X_1 \leq y] P[X_1 \leq y] \cdots P[X_1 \leq y]$$

$$= (P[X_1 \leq y])^n,$$

as desired. \qed
6 Cdf of $Y_1$

**Lemma 2.** Let $X_1, \ldots, X_n$ be independent, identical random variables. Then, for any real number $y$,

$$P[Y_1 \leq y] = 1 - (P[X_1 > y])^n.$$

**Proof.** The proof is (almost) identical to the one in Lemma 1 and thus is left as an exercise\(^2\). \qed

---

\(^2\)When I was a student I had a (misguided) view that it is unforgivable for the instructor to make notes with proofs left as exercises. It was only after I become older that I realize it is important for you to do these calculations yourself as part of the learning experience, and there are no shortcuts for this.
7 Probability density function

Let $X$ be a continuous random variable. The **probability density function (pdf)** of $X$ is the derivative of cdf,

$$f_X(x) := \frac{\partial}{\partial x} F_X(x) \quad \text{for all real number } x.$$

Note that the pdf and cdf satisfies

$$F_X(x) := \int_{-\infty}^{x} f_X(t) \, dt \quad \text{for all real number } x.$$
8 Pdf of ordered statistics: Example

Let $X_1, X_2, X_3$ be independent, uniform random variables on $[0, 1]$, (BT)

$$f_{X_i}(x) = \begin{cases} 
1 & \text{if } 0 \leq x \leq 1; \\
0 & \text{otherwise.}
\end{cases}$$

We now compute the pdf and cdf of $Y_1, \ldots, Y_3$.

First note that

$$P[Y_i \leq y] = \begin{cases} 
0 & \text{if } y \leq 0; \\
1 & \text{if } y \geq 1;
\end{cases}$$

$$f_{Y_i}(y) = 0 \quad \text{if } y \leq 0 \text{ or } y \geq 1.$$ 

So we are left with the case $0 \leq y \leq 1$. 

10
The cdf of $Y_3$ is equal to, for any $y \in [0, 1]$, (BT)

$$F_{Y_3}(y) = (P[X_1 \leq y])^3 \quad (\text{by Lemma 1})$$

$$= \left( \int_0^y 1 \, dx \right)^3 = y^3.$$

The pdf of $Y_3$ is then equal to, for any $y \in [0, 1]$,

$$f_{Y_3}(y) = \frac{\partial}{\partial y} F_{Y_3}(y) = \frac{\partial}{\partial y} y^3 = 3y^2.$$
The cdf of $Y_1$ is equal to, for any $y \in [0, 1]$, (BT)

$$F_{Y_1}(y) = 1 - \left( P[X_1 > y] \right)^3$$

(by Lemma 2)

$$= 1 - \left( \int_y^1 1 \, dx \right)^3$$

$$= 1 - (1 - y)^3 = 3y - 3y^2 + y^3.$$

The pdf of $Y_1$ is then equal to, for any $y \in [0, 1]$,

$$f_{Y_1}(y) = 3 - 6y + 3y^2.$$
We now compute the pdf and cdf of $Y_2$ for $y \in [0, 1]$. For $Y_2$ (which is the second smallest of $X_1, X_2, X_3$) to be smaller than $y$, at least two of $X_1, X_2, X_3$ are smaller than $y$. There are three possibilities

- $X_1$ and $X_2$ are smaller than $y$; or
- $X_1$ and $X_3$ are smaller than $y$; or
- $X_2$ and $X_3$ are smaller than $y$;

These three events are overlapping when all $X_1, X_2, X_3$ are smaller than $y$. 
This implies that (BT)

\[ F_{Y_2}(y) = P[Y_2 \leq y] \]

\[ = P[X_1 \leq y, X_2 \leq y] + P[X_1 \leq y, X_3 \leq y] + P[X_2 \leq y, X_3 \leq y] - 2P[X_1 \leq y, X_2 \leq y, X_3 \leq y] \]

\[ = P[X_1 \leq y]P[X_2 \leq y] + P[X_1 \leq y]P[X_3 \leq y] + P[X_2 \leq y]P[X_3 \leq y] - 2P[X_1 \leq y]P[X_2 \leq y]P[X_3 \leq y] \]

\[ = y \times y + y \times y + y \times y - 2y \times y \times y \]

\[ = 3y^2 - 2y^3. \]

The pdf of \( Y_2 \) is thus equal to

\[ f_{Y_2}(y) = 6y - 6y^2. \]
9 Pdf of ordered statistics

Theorem 3. Let \( X_1, \ldots, X_n \) be independent, identical continuous random variables. Then the pdf of \( Y_k \) (for \( 1 \leq k \leq n \)) is equal to

\[
f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!} \left[ F_{X_1}(y) \right]^{k-1} \left[ 1 - F_{X_1}(y) \right]^{n-k} f_{X_1}(y),
\]

where \( n! := 1 \times 2 \times 3 \times \ldots \times n \) is the product of the first \( n \) positive integers.
Chapter 10 Checking for normal random variables

Suppose for an unknown random variable $X$, how to check if $X$ is a normal random variable?

- If $n$ is large, you can draw histogram or stem-and-leaves diagram, and check if you see a bell curve;

- If $n$ is small, you can draw the q-q plot and check if you see a straight line.
We now learn how to draw the q-q plot.

Suppose that we have the following samples:

\[
\begin{array}{cccccccccc}
1.24 & 1.36 & 1.28 & 1.31 & 1.35 & 1.20 & 1.39 & 1.35 & 1.41 & 1.31 \\
1.28 & 1.26 & 1.37 & 1.49 & 1.32 & 1.40 & 1.33 & 1.28 & 1.25 & 1.39 \\
1.38 & 1.34 & 1.40 & 1.27 & 1.33 & 1.36 & 1.43 & 1.33 & 1.29 & 1.34 \\
\end{array}
\]

Figure 1: Sample of size \( n = 30 \) of a certain unknown random variable \( X \), taken from the textbook.

(1 minute pause, open textbook page 262)
<table>
<thead>
<tr>
<th>k</th>
<th>Diameters in mm (x)</th>
<th>$p = k/31$</th>
<th>$z_{1-p}$</th>
<th>k</th>
<th>Diameters in mm (x)</th>
<th>$p = k/31$</th>
<th>$z_{1-p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.20</td>
<td>0.0323</td>
<td>-1.85</td>
<td>16</td>
<td>1.34</td>
<td>0.5161</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>0.0645</td>
<td>-1.52</td>
<td>17</td>
<td>1.34</td>
<td>0.5484</td>
<td>0.12</td>
</tr>
<tr>
<td>3</td>
<td>1.25</td>
<td>0.0968</td>
<td>-1.30</td>
<td>18</td>
<td>1.35</td>
<td>0.5806</td>
<td>0.20</td>
</tr>
<tr>
<td>4</td>
<td>1.26</td>
<td>0.1290</td>
<td>-1.13</td>
<td>19</td>
<td>1.35</td>
<td>0.6129</td>
<td>0.29</td>
</tr>
<tr>
<td>5</td>
<td>1.27</td>
<td>0.1613</td>
<td>-0.99</td>
<td>20</td>
<td>1.36</td>
<td>0.6452</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>1.28</td>
<td>0.1935</td>
<td>-0.86</td>
<td>21</td>
<td>1.36</td>
<td>0.6774</td>
<td>0.46</td>
</tr>
<tr>
<td>7</td>
<td>1.28</td>
<td>0.2258</td>
<td>-0.75</td>
<td>22</td>
<td>1.37</td>
<td>0.7097</td>
<td>0.55</td>
</tr>
<tr>
<td>8</td>
<td>1.28</td>
<td>0.2581</td>
<td>-0.65</td>
<td>23</td>
<td>1.38</td>
<td>0.7419</td>
<td>0.65</td>
</tr>
<tr>
<td>9</td>
<td>1.29</td>
<td>0.2903</td>
<td>-0.55</td>
<td>24</td>
<td>1.39</td>
<td>0.7742</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>1.31</td>
<td>0.3226</td>
<td>-0.46</td>
<td>25</td>
<td>1.39</td>
<td>0.8065</td>
<td>0.86</td>
</tr>
<tr>
<td>11</td>
<td>1.31</td>
<td>0.3548</td>
<td>-0.37</td>
<td>26</td>
<td>1.40</td>
<td>0.8387</td>
<td>0.99</td>
</tr>
<tr>
<td>12</td>
<td>1.32</td>
<td>0.3871</td>
<td>-0.29</td>
<td>27</td>
<td>1.40</td>
<td>0.8710</td>
<td>1.13</td>
</tr>
<tr>
<td>13</td>
<td>1.33</td>
<td>0.4194</td>
<td>-0.20</td>
<td>28</td>
<td>1.41</td>
<td>0.9032</td>
<td>1.30</td>
</tr>
<tr>
<td>14</td>
<td>1.33</td>
<td>0.4516</td>
<td>-0.12</td>
<td>29</td>
<td>1.43</td>
<td>0.9355</td>
<td>1.52</td>
</tr>
<tr>
<td>15</td>
<td>1.33</td>
<td>0.4839</td>
<td>-0.04</td>
<td>30</td>
<td>1.49</td>
<td>0.9677</td>
<td>1.85</td>
</tr>
</tbody>
</table>

Figure 2: Table of samples, ordered from smallest to largest, with corresponding values of $p$ and $z_{1-p}$.
Figure 3: The q-q plot for the example. The x-axes are indexed by values of $y_k$, and the y-axes are indexed by values of $z_{1-p_k}$. The 30 black dots are the plot of $(y_k, z_{1-p_k})$, and the straight line is a line that “best fit” those 30 dots.
Here is how to draw the q-q plot:

1. Compute the order statistics $y_1 \leq y_2 \leq \ldots \leq y_n$.
   That is, order the samples from smallest to largest.

2. Write $p_1, p_2, \ldots p_n$, where $p_k := \frac{k}{n+1}$.

3. Compute $z_{1-p_1}, z_{1-p_2}, \ldots, z_{1-p_n}$, which can be found
   from Table Va and Vb of the textbook Appendix C.
   This number $z_\alpha$ is the real number such that

   $P(Z > z_\alpha) = \alpha,$

   where $Z$ is the standard normal random variable
   (mean 0 and variance 1).
4. Plot $n$ dots in the plane, $(y_1, z_1-p_1)$, $(y_2, z_1-p_2)$, \ldots, $(y_n, z_1-p_n)$.

5. Squint your eyes, and try to see if these points form a straight line. Later in Section 6.5, we will learn linear regression to draw this line.