Biostats VHM 801 Course
Fall 2020, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment 1

The assignment is based on a subset of the data collected for an MSc project in shellfish aquaculture at AVC (Aaron Ramsay, Vase Tunicate, *Ciona intestinalis*: Ecology and Mitigation Strategies, UPEI 2008). The socking experiment was specifically discussed in a published article, Ramsay et al. (2008), *Aquaculture* **275**, 194–200. The three socking conditions were different densities.

The solution presents one way in which the descriptive analysis could be done, but other ways are possible as well. It is more detailed than required for a 100% mark, among other things by including all the variables for the descriptive analysis when only two selected variables were required for the assignment. All analyses shown used Minitab 19, but other Minitab versions or other programs, such as Stata, would give similar figures and results.

1. Study and variable types

The study is mostly *experimental*, because of the socking conditions imposed at the time of socking. This represents a clear intervention by the experimenter, whereas the time of socking in either the fall or spring could be viewed also as normal production practices that were not specifically put in place for the study (although the setup with mussels from both socking periods represented on the same longlines was a specific study feature). Once the mussels were socked, no further interventions were undertaken up till the time of partial harvest, so this part was observational.

The variables Sock, Time and Condition are nominal categorical, whereas the remaining seven variables are all quantitative (and with ratio scale, if the further subdivision between ratio and interval scale from the Stephens textbook is adopted). The two abundance variables are counts and therefore discrete, but their distribution may be approximated by continuous distributions because the counts span very wide ranges. The other variables are measured on a continuous scale.

In experimental design terminology we would label Time and Condition as factors, and the 15 socks per factor combination would be considered as replicates (or replications), having exactly the same experimental conditions. Socking time could be considered as a blocking factor because it seems a bit artificial to view the socks as being allocated to socking times, and because the socking conditions would have been randomized within the two socking times separately. This is similar to studies with a factor such as gender, where gender is an inherent property of the individuals not an assigned condition. There are no practical implications of whether the socking time is considered as a treatment factor or a blocking factor. The randomization described in Question 6 corresponds to considering socking time as a treatment factor at a different level than the socking conditions. Depending on the choice of terminology for the socking time, one could either say the study has 3 treatments and 2 blocks, or that is has 6 treatments formed by the combinations of the two factors.

2. Descriptive analysis

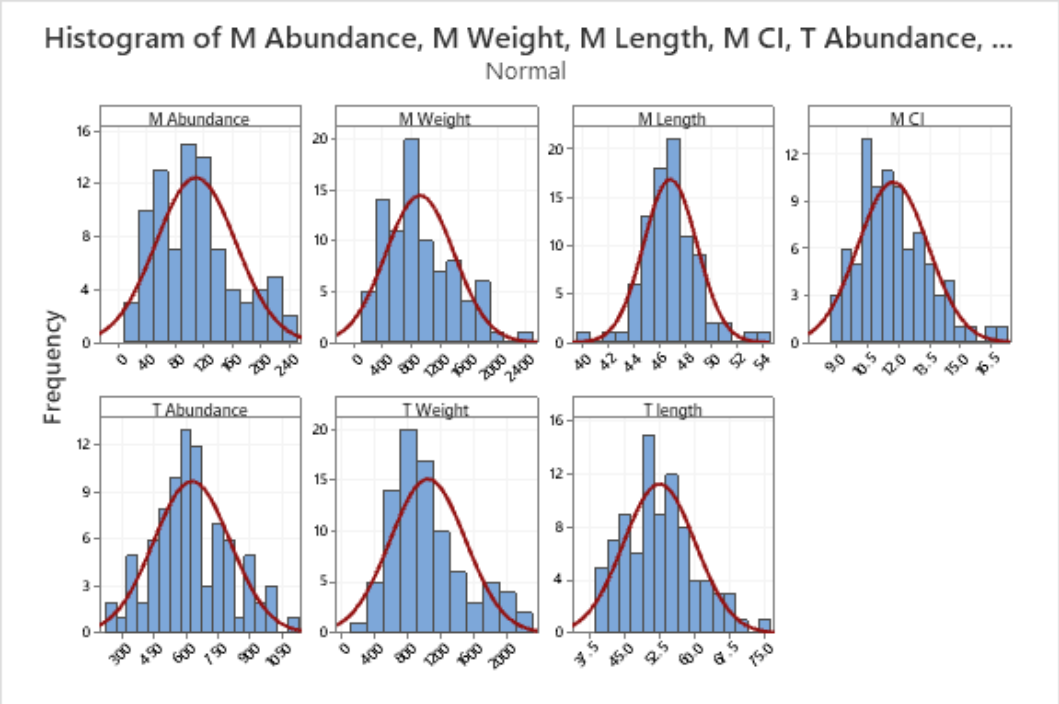
According to the text, the data should be considered as a simple random sample (or i.i.d., independent and identically distributed, observations) from the population of socks, although it really is not. The two tables below gives the most useful descriptive statistics for continuous variables; the list of descriptive statistics should as a minimum include the mean, median and standard deviation. All variables have 87 non-missing values (no values were recorded for 3 of the 90 socks), and different displays can show the distribution well: a histogram with overlaid normal curve (preferably with 9–12 bins), a stem-plot and also a dotplot. The boxplot is just a graphical representation of the descriptive statistics.

Descriptive statistics for mussel *variables:*

Statistic	*abundance*	*weight (g)*	*length (mm)*	*cond. index (%)*
mean	109.4	918.8	46.75	11.75
minimum	22.0	129.3	39.75	8.79
1st quartile	63.0	544.9	45.48	10.49
median	101.0	860.0	46.82	11.54
3rd quartile	136.0	1216.0	47.79	12.88
maximum	232.0	2434.5	53.62	16.77
standard deviation	55.6	479.4	2.06	1.69
inter-quartile range	73.0	671.1	2.31	2.39
skewness	0.58	0.74	0.24	0.63

Descriptive statistics for tunicate *variables:*

Statistic	*abundance*	*weight (g)*	*length (mm)*
mean	628.8	1037.8	52.43
minimum	264.0	295.9	39.46
1st quartile	512.0	714.5	47.20
median	624.0	948.2	51.92
3rd quartile	736.0	1272.0	56.71
maximum	1088.0	2172.1	75.12
standard deviation	178.3	457.1	7.68
inter-quartile range	224.0	557.5	9.51
skewness	0.31	0.84	0.53



Finally, brief summaries of the distributions based on the computed statistics and histograms with overlaid normal curve:

- *mussel abundance*: unimodal; centered around 100 mussels; quite large standard deviation relative to the mean (cv $= s/\bar{x} = 0.51$); somewhat right-skewed without suspected outliers,
- *mussel weight*: unimodal; centered around 900 g; quite large standard deviation relative to the mean (cv $= 0.54$); right-skewed with a long right tail; one potential outlier on the boxplot (not shown) in the right tail but does not look like a real outlier,
- *mussel length*: unimodal; centered around 46.5 mm; close to symmetrical but has too heavy tails compared to a bell-shaped curve, due to three potential outliers in the right tail and one in the left tail, a biological assessment of the potential outliers should be undertaken,
- *mussel CI*: unimodal; centered around 11.5 %; somewhat right-skewed with a long right tail; one potential outlier in the right tail but does not look like a real outlier,
- *tunicate abundance*: unimodal; centered around 625 tunicates; quite symmetrical and bell-shaped; one potential outlier in the right tail but does not look like a real outlier,
- *tunicate weight*: unimodal; centered around 1000 g; quite large standard deviation relative to the mean (cv $= 0.44$); clearly right-skewed with long right tail; one potential outlier in the right tail but hardly a real outlier,
- *tunicate length*: unimodal; centered around 52 mm; somewhat right-skewed with a too short left tail; two potential outliers in the right tail but do not look like real outliers.

3. Descriptive analysis to compare socking times and condition

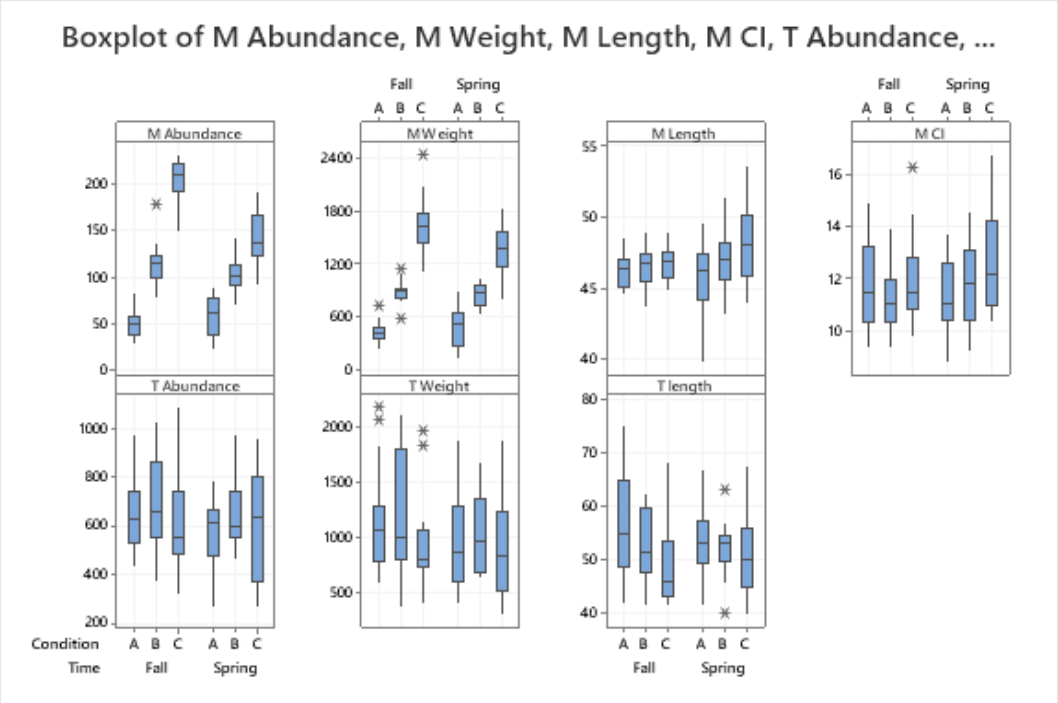
The focus in this part is on comparative statistics and graphics, rather than repeating everything for the six different treatments (the 3 socking conditions times the 2 socking times). Among the descriptive statistics it is primarily of interest to compare the center and spread of the different distributions. The distribution shape is poorly determined because of the small number ($\leq$15) of observations per group. None of the variables have strongly skewed distributions so we choose the most standard measures of center and spread, the mean and standard deviation. In view of the small sample size per group, the best comparitive graphical display is by boxplots.

Mean and standard deviation per treatment for mussel *variables:*

Socking		Mussel (mean ± standard deviation)			
time	condition	*abundance*	*weight (g)*	*length (mm)*	*cond. index* (%)
Fall	A	51.5±13.9	434±120	46.3±1.1	11.8±1.7
	B	113.3±24.3	884±134	46.6±1.4	11.2±1.3
	C	204.5±25.6	1643±339	46.9±1.2	11.9±1.8
Spring	A	55.7±21.2	491±223	45.8±2.6	11.2±1.5
	B	103.7±17.8	839±129	46.9±2.0	11.8±1.7
	C	142.2±27.0	1346±310	48.2±2.8	12.7±2.0

Mean and standard deviation per treatment for tunicate *variables:*

Socking time	condition	Tunicate (mean ± standard deviation) *abundance*	*weight (g)*	*length (mm)*
Fall	A	643±152	1154±497	56.0±10.0
	B	693±184	1212±569	52.9± 6.9
	C	614±206	965±454	48.6± 7.7
Spring	A	567±147	961±393	53.8± 6.8
	B	651±147	1017±326	52.2± 5.2
	C	601±230	899±461	50.5± 7.8



Again, brief summaries of the distributions based on the computed statistics and graphs:

- *mussel abundance*: very clear differences in location for socking conditions (increasing from A to C); socking times approximately equal for conditions A − B, but condition C is higher with fall than spring socking; variable spread, possibly increasing with the mean,
- *mussel weight*: very clear differences in location for socking conditions (increasing from A to C); socking times approximately equal for conditions A − B, and condition C is somewhat higher with fall than spring socking; very variable spread, and largest with highest means; one possible real outlier for fall socking with condition C,
- *mussel length*: fairly similar in location across all six groups; some indication of larger spread for spring socking; no indications of outliers (all the previously noted suspect observations now seem fine),
- *mussel CI*: fairly similar in both location and spread across all six groups, despite some variation with no obvious pattern; one potential outlier in the right tail but not the same as previously noted and probably no real outlier,
- *tunicate abundance*: fairly similar in both location and spread across all six groups, despite some variation with no obvious pattern; no potential outliers indicated,

- *tunicate weight*: fairly similar in both location and spread across the six treatments; most distributions appear right-skewed, and there are four potential outliers in the right tail,
- *tunicate length*: some indication of differences in center between socking condition, in particular for fall socking (declining from A to C); also some differences in spread but the spread does not seem to follow the mean; two potential outliers indicated for spring B but these do not seem particularly extreme when comparing to the other distributions,

4. Final descriptive analysis to describe the distributions, accounting for socking times and condition

For each of the variables, we here need to look at six distributions instead of the single distribution in Q2. In order to avoid making this solution excessively long, only normal probability plots will be computed, and all of these have been collected in an appendix. Note that histograms should be avoided with the small sample sizes per group: they can be quite misleading. To aid in the interpretation of the plots and the indicated P-values for the AD-normality test it is also helpful to look at the symmetry and suspected outliers in the boxplots and to compute some descriptive statistics, in particular the skewness statistic (which will be referred to in the summaries without including a complete list). Note also that when doing six normality tests per variable, we would typically want to apply a somewhat stricter significance level, maybe set at $\alpha = 0.01$, in order to reduce the chance of getting of false significances (this issue will be discussed more later in the course).

- *mussel abundance*: no clear indications of non-normality: only group with $P < 0.05$ is (Fall,C) with some left-skewness; none of the features from Q2 seen here,
- *mussel weight*: no clear indications of non-normality by probability plots or normality tests, and even the visually clear potential outlier in (Fall,C) does not fall outside the confidence limits; no consistent or strong right-skewness within groups,
- *mussel length*: all distributions look nicely normal,
- *mussel CI*: a few of the groups show some curvature in the normality plot ($\sim$ right-skewness), but the normality tests are all clearly non-significant (except for (Fall,C) with $P = 0.08$),
- *tunicate abundance*: also here a couple of groups with curvature in the normality plot ($\sim$ right-skewness), but the normality tests are all clearly non-significant (except for (Spring,B) with $P = 0.07$),
- *tunicate weight*: the right-skewness was already noted above, and in the two distributions with suspected outliers it is strong enough to give formal evidence against normality (Fall,A & C),
- *tunicate length*: most distributions look fairly normal, but the right-skewness of (Fall,C) is strong enough for evidence against normality.

5. Allocation of mussel seed to the socks

The main consideration is to distribute the mussel seed equally among the socks with the three different socking conditions (both for fall and spring socking). The mussel seed used for fall and spring socking is hardly the same, although an effort can be made to use seed from the same source and to avoid any other obvious differences. At both of the socking events (fall/spring), we would ideally want every mussel seed to have the same probability to enter into each of the socks, but that may be difficult to achieve strictly (due to the large number of seed and their small size). It would obviously be good to mix the seed well before allocating it to the socks. It could also be suggested to fill up the socks partially in a rotating scheme (e.g. each sock filled up one third, then in the second round every sock is filled up to two-thirds, and in the final round every sock is completed). Another variant of this idea, although less satisfactory because it will increase variability *between*

the socks, is to fill up the socks for conditions A – C in a rotating (or blocked) scheme, such as (A,B,C,A,B,C,. . . ,A,B,C). If such a scheme was used and the position in the sequence was thought to be a source of variation, the blocks should be noted and used in the subsequent analysis. In practice, it may be sufficient for the experimenter to be aware of the risk of inducing bias between the socking conditions from the selection of mussel seed and to take sensible steps in the socking procedure to make the seed distributed as equally as possible onto the socks.

6. Experimental design and randomization

A simple sketch of the experimental layout, with its 3 longlines, each split into two half longlines and again each split into 3 blocks, is shown below:

```
longline 1
           half longline                          half longline
|-----------------------------------||-----------------------------------|
|           |           |           ||           |           |           |
    block       block       block       block       block       block

longline 2
           half longline                          half longline
|-----------------------------------||-----------------------------------|
|           |           |           ||           |           |           |
    block       block       block       block       block       block

longline 3
           half longline                          half longline
|-----------------------------------||-----------------------------------|
|           |           |           ||           |           |           |
    block       block       block       block       block       block
```

The randomization of socking times and conditions onto the design of three longlines is most conveniently done in two steps. This is a result of the two factors, socking times and conditions, being applied to different experimental units. All socks (socking groups) on a halfline were socked at the same time (either fall or spring), therefore socking times are applied to halflines. When a direction of each longline has been decided upon, two halflines are created. Then the two socking times are allocated randomly to halflines (within each line); this involves three independent random choices, which could be done by simple procedures (cards from a pile, numbers from a hat). The following Minitab commands generate a valid randomization.

```
Name c1 "line"
Set 'line'
  1( 1 : 3 / 1 )2
  End.
Name c2 "socktime"
TSet 'socktime'
  3(  "S" "F" )1
  End.
Base 201012.
Name c3 "randomhline"
Random 6 'randomhline';
Uniform 0.0 1.0.
Sort 'line'-'randomhline' 'line'-'randomhline';
By 'line' 'randomhline'.
Print 'line'-'randomhline'.
```

Data

Row	line	socktime	randomhline
1	1	S	0.723831
2	1	F	0.985629
3	2	F	0.046795
4	2	S	0.195992
5	3	S	0.495046
6	3	F	0.655727

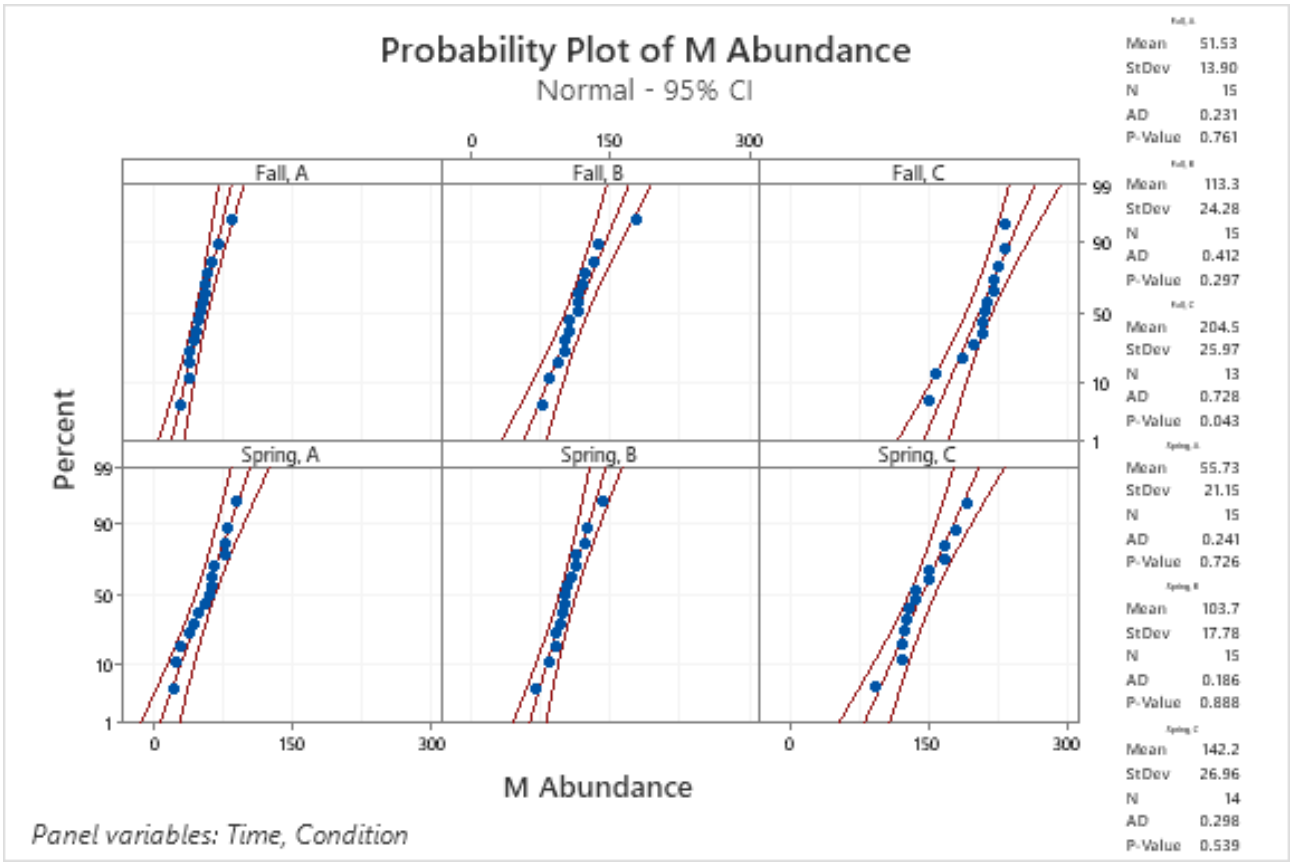
The socking conditions were applied to individual socks located at the sock locations along the line. On each halfline there were 3 blocks, this equates to six blocks per longline, and 18 blocks in total. The randomization requires the three socking conditions to be randomly allocated to the three sock locations with each block. The following Minitab commands generate a valid randomization. For convenience, the socking times are entered into the design as well in the order determined by the first randomization.

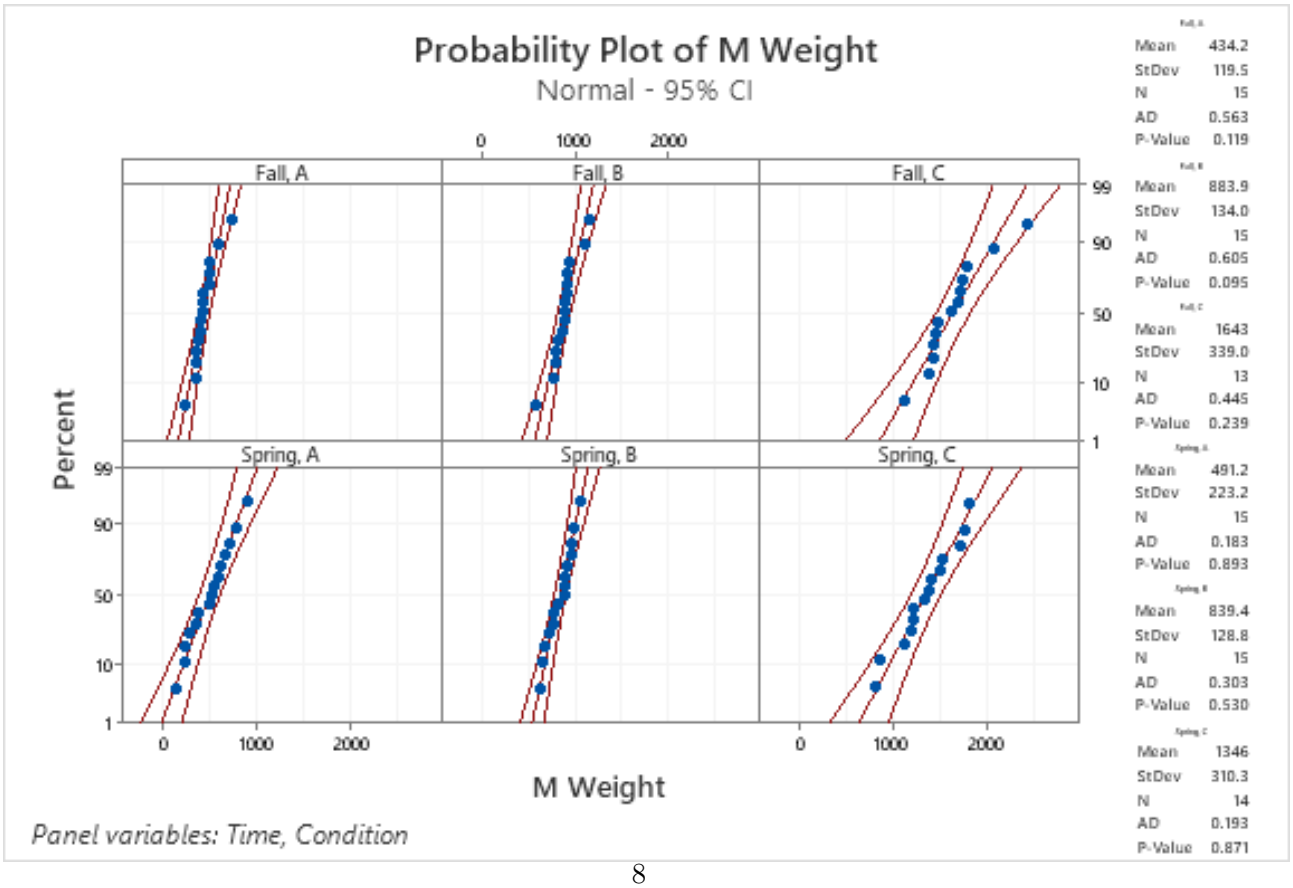
```
Name c5 "line2"
Set 'line2'
  1( 1 : 3 / 1 )18
  End.
Name c6 "socktime2"
TSet 'socktime2'
  1(  "S" "F" "F" "S" "S" "F" )9
  End.
Name c7 "block"
Set 'block'
  3( 1 : 6 / 1 )3
  End.
Name c8 "sockcond"
TSet 'sockcond'
  18(  "A" "B" "C" )1
  End.
Name c9 "randomloc"
Random 54 'randomloc';
Uniform 0.0 1.0.
Sort  'line2'-'randomloc'  'line2'-'randomloc';
By 'line2' 'block' 'randomloc'.
Print 'line2'-'randomloc'.
```

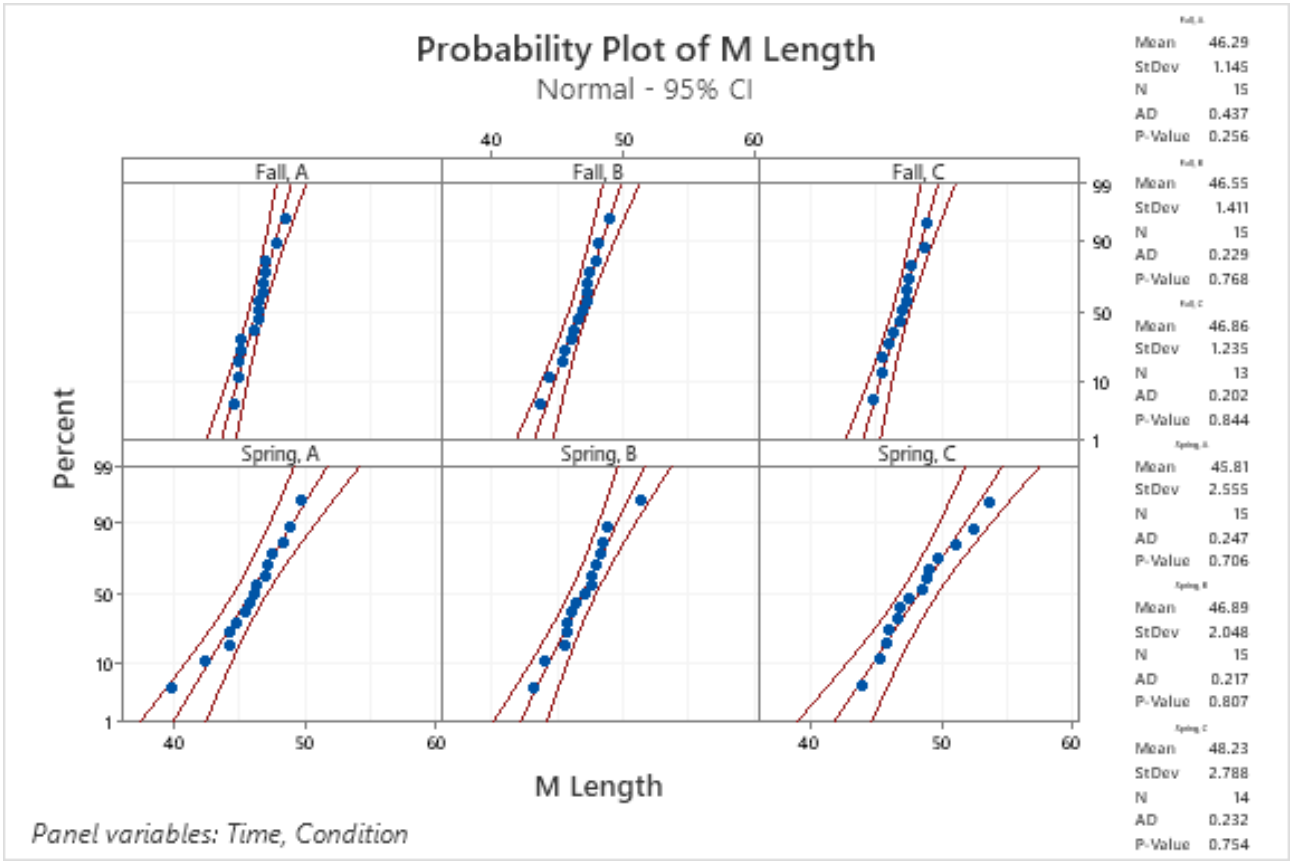
Data

Row	line2	socktime2	block	sockcond	randomloc
1	1	S	1	A	0.361621
2	1	S	1	B	0.660354
3	1	S	1	C	0.885711
4	1	S	2	C	0.229786
5	1	S	2	B	0.290462
6	1	S	2	A	0.914751
7	1	S	3	C	0.279621
8	1	S	3	B	0.625032
9	1	S	3	A	0.983714
10	1	F	4	C	0.238113
11	1	F	4	A	0.751107
12	1	F	4	B	0.827579
13	1	F	5	B	0.191902
14	1	F	5	C	0.360328
15	1	F	5	A	0.959049
16	1	F	6	A	0.002679
17	1	F	6	C	0.331861
18	1	F	6	B	0.835684
19	2	F	1	C	0.174599
20	2	F	1	A	0.272474
21	2	F	1	B	0.964983
22	2	F	2	A	0.338544
23	2	F	2	C	0.575941
24	2	F	2	B	0.887351
25	2	F	3	A	0.324410
26	2	F	3	C	0.791394
27	2	F	3	B	0.978708
28	2	S	4	C	0.050422
29	2	S	4	B	0.656690
30	2	S	4	A	0.745895
31	2	S	5	B	0.042195
32	2	S	5	A	0.302664
33	2	S	5	C	0.794033
34	2	S	6	A	0.592697
35	2	S	6	C	0.624632
36	2	S	6	B	0.767540
37	3	S	1	B	0.203994
38	3	S	1	A	0.331159
39	3	S	1	C	0.758898
40	3	S	2	B	0.590168
41	3	S	2	A	0.913685
42	3	S	2	C	0.924525
43	3	S	3	B	0.066147
44	3	S	3	C	0.237661
45	3	S	3	A	0.927422
46	3	F	4	A	0.246606
47	3	F	4	B	0.271469
48	3	F	4	C	0.978334
49	3	F	5	B	0.214365
50	3	F	5	C	0.753494
51	3	F	5	A	0.962484
52	3	F	6	A	0.297505
53	3	F	6	C	0.419698
54	3	F	6	B	0.936501

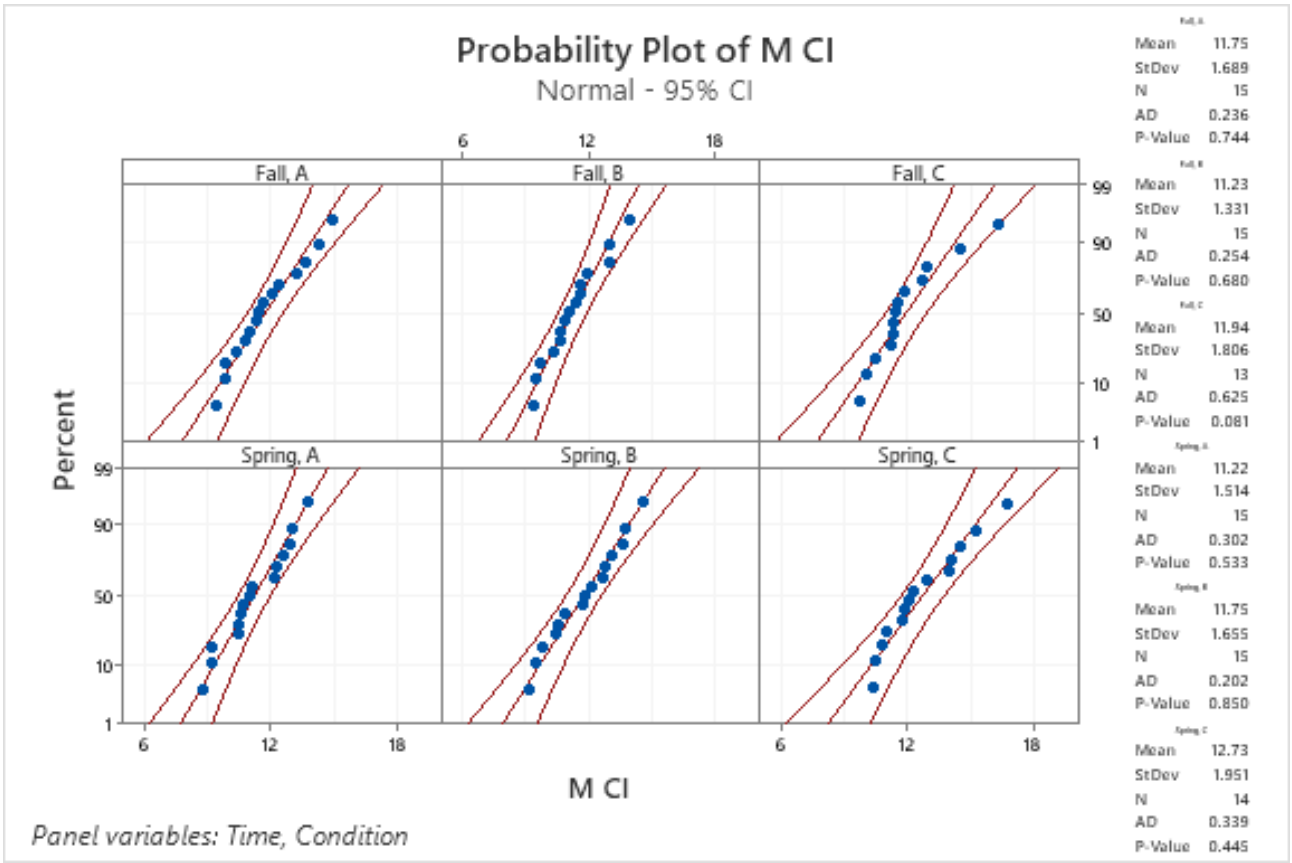
Appendix: Normal probability plots for each variable (Q4)



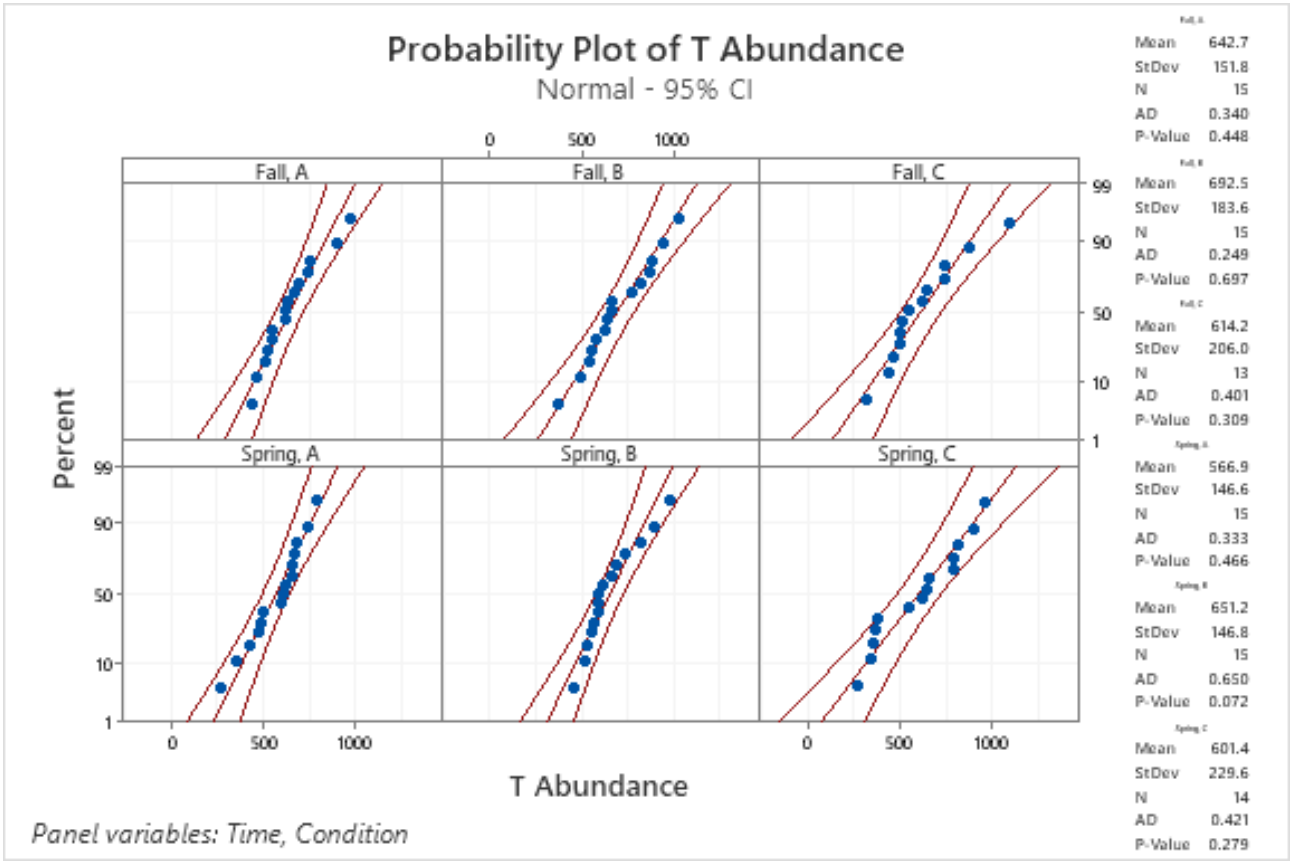


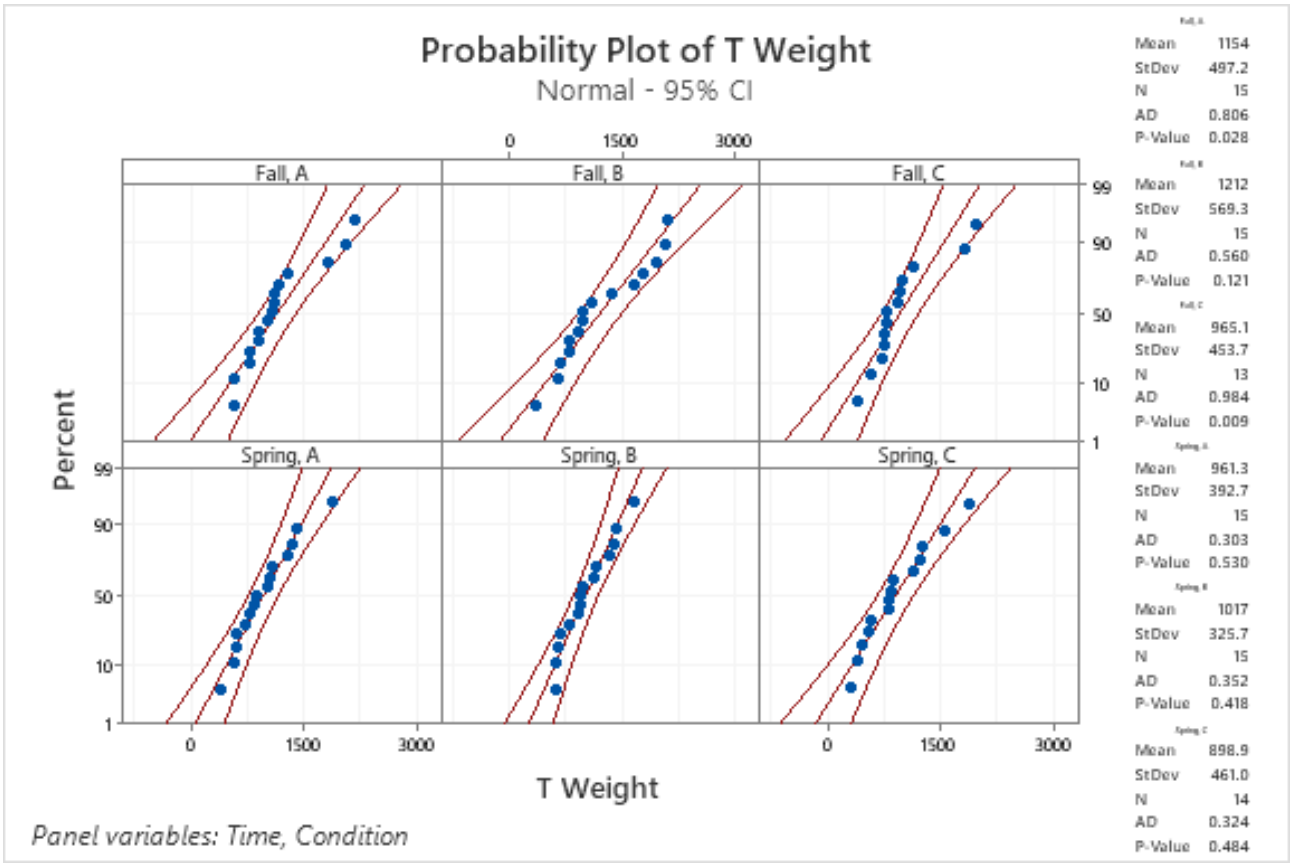


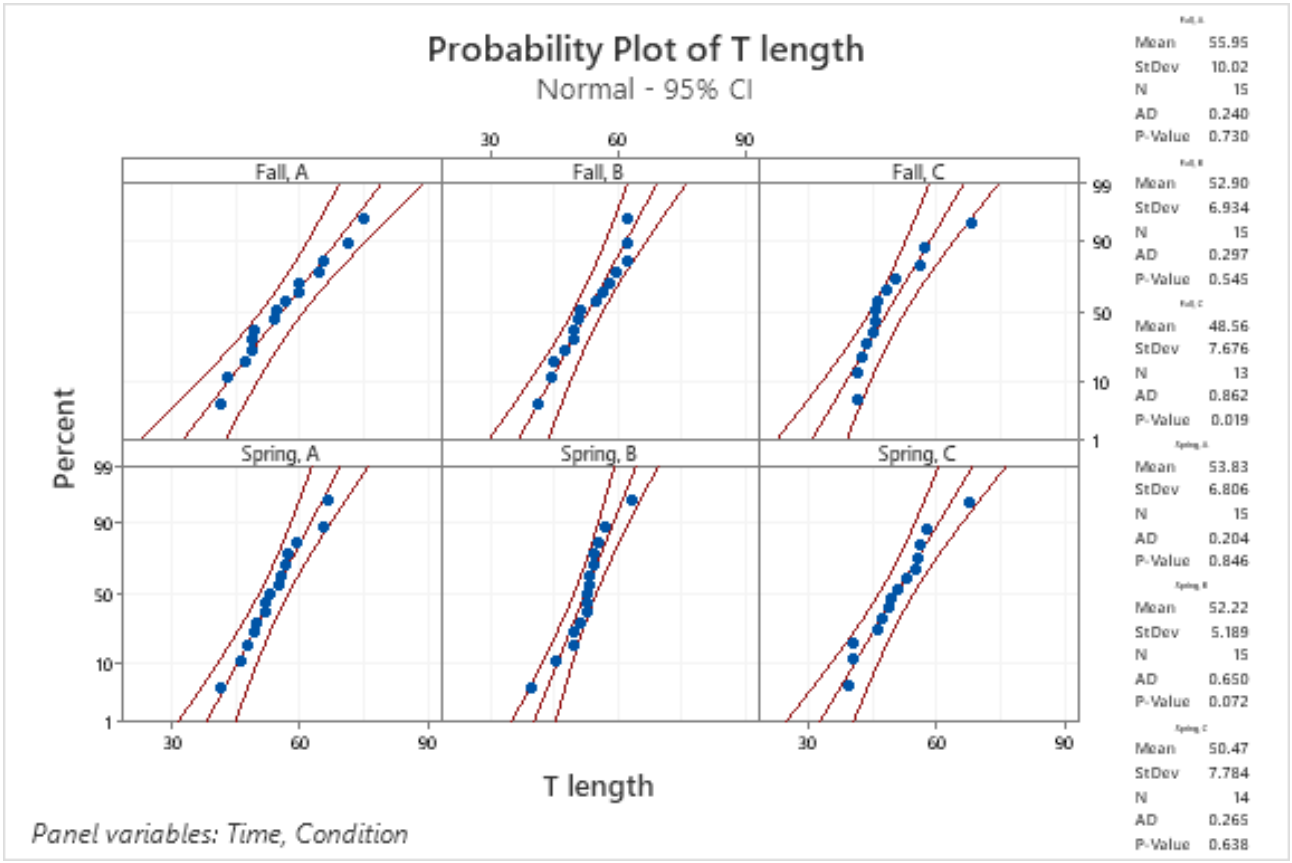
Panel variables: Time, Condition



Panel variables: Time, Condition


Probability Plot of T Abundance
Normal - 95% CI
Fall, A
Fall, B
Fall, C
Spring, A
Spring, B
Spring, C
Percent
T Abundance
Panel variables: Time, Condition
Mean 642.7
StDev 151.8
N 15
AD 0.340
P-Value 0.448
Mean 692.5
StDev 183.6
N 15
AD 0.249
P-Value 0.697
Mean 614.2
StDev 206.0
N 13
AD 0.401
P-Value 0.309
Mean 566.9
StDev 146.6
N 15
AD 0.333
P-Value 0.466
Mean 651.2
StDev 146.8
N 15
AD 0.650
P-Value 0.072
Mean 601.4
StDev 229.6
N 14
AD 0.421
P-Value 0.279


Probability Plot of T Weight
Normal - 95% CI
Fall, A
Fall, B
Fall, C
Spring, A
Spring, B
Spring, C
Percent
T Weight
Panel variables: Time, Condition
Mean 1154
StDev 497.2
N 15
AD 0.806
P-Value 0.028
Mean 1212
StDev 569.3
N 15
AD 0.560
P-Value 0.121
Mean 965.1
StDev 453.7
N 13
AD 0.984
P-Value 0.009
Mean 961.3
StDev 392.7
N 15
AD 0.303
P-Value 0.530
Mean 1017
StDev 325.7
N 15
AD 0.352
P-Value 0.418
Mean 898.9
StDev 461.0
N 14
AD 0.324
P-Value 0.484


Probability Plot of T length
Normal - 95% CI
Fall, A
Fall, B
Fall, C
Spring, A
Spring, B
Spring, C
Percent
T length
Panel variables: Time, Condition
Mean 55.95
StDev 10.02
N 15
AD 0.240
P-Value 0.730
Mean 52.90
StDev 6.934
N 15
AD 0.297
P-Value 0.545
Mean 48.56
StDev 7.676
N 13
AD 0.862
P-Value 0.019
Mean 53.83
StDev 6.806
N 15
AD 0.204
P-Value 0.846
Mean 52.22
StDev 5.189
N 15
AD 0.650
P-Value 0.072
Mean 50.47
StDev 7.784
N 14
AD 0.265
P-Value 0.638