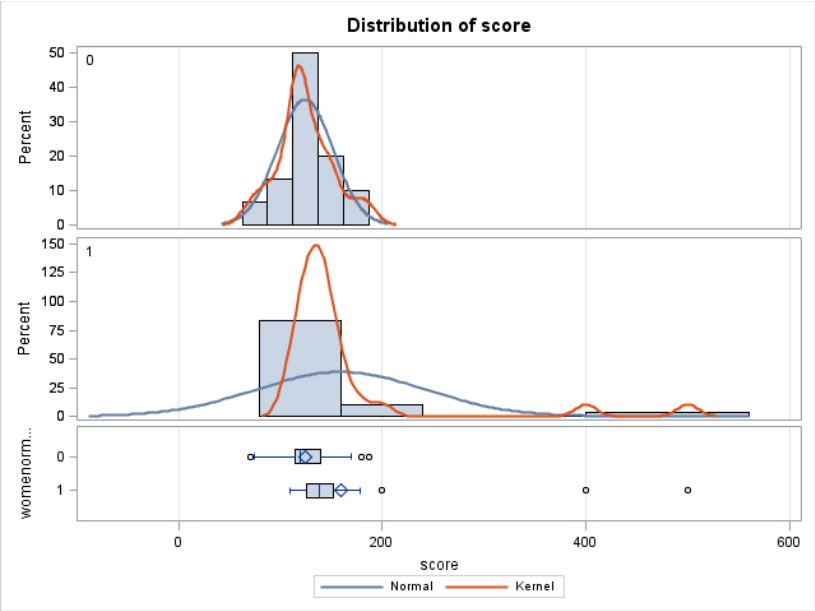
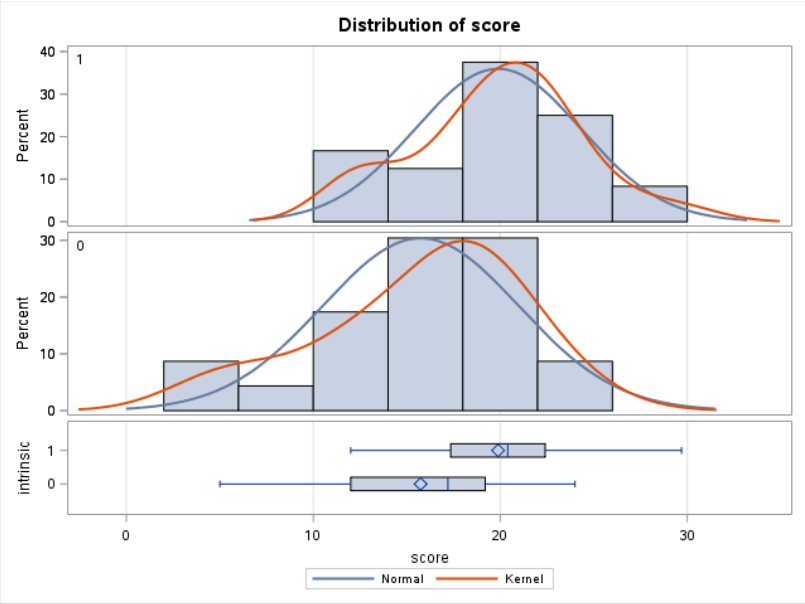
A Closer Look at Assumptions

ROBUSTNESS OF T-TOOLS

Assumptions for T-Tools: Equal Variances

Looking at Histograms & Test for Equal Variances

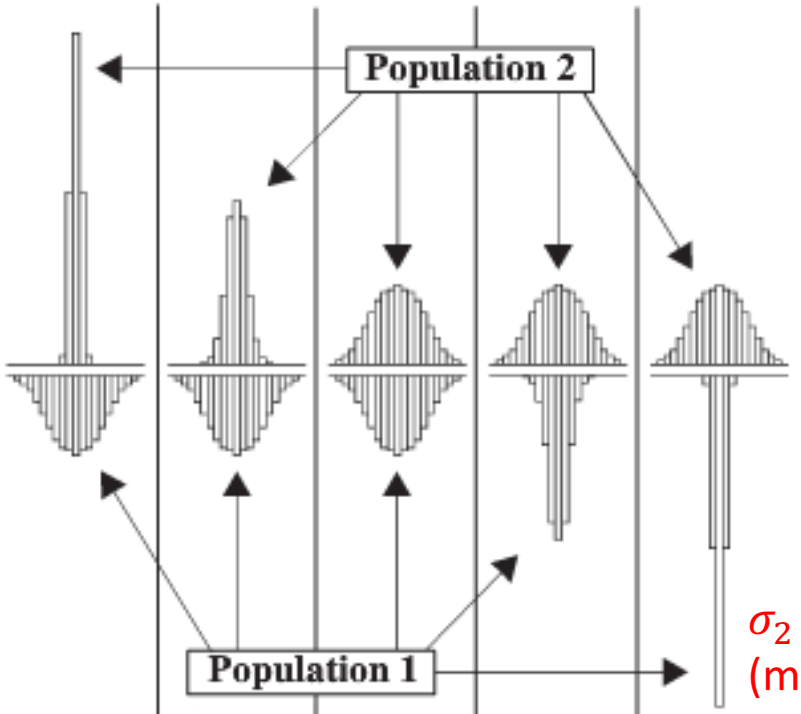




Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	29	9.08	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	23	1.40	0.4289

DISPLAY 3.5 Percentage of successful 95% confidence intervals when the two populations have different standard deviations (but are normal) with possibly different sample sizes (each percentage is based on 1,000 computer simulations)



$n_1 = n_2$
(equal coverage)

$\sigma_2 < \sigma_1$ and $n_1 < n_2$
(less coverage)

$\sigma_2 > \sigma_1$ and $n_1 < n_2$
(more coverage)

n_1	n_2	$\sigma_2/\sigma_1 = 1/4$	$\sigma_2/\sigma_1 = 1/2$	$\sigma_2/\sigma_1 = 1$	$\sigma_2/\sigma_1 = 2$	$\sigma_2/\sigma_1 = 4$
10	10	95.2	94.2	94.7	95.2	94.5
10	20	83.0	89.3	94.4	98.7	99.1
10	40	71.0	82.6	95.2	99.5	99.9
100	100	94.8	96.2	95.4	95.3	95.1
100	200	86.5	88.3	94.8	98.8	99.4
100	400	71.6	81.5	95.0	99.5	99.9

Success rates for 95% intervals

Assumptions of Two sample T-Tools

1. Samples are drawn from a Normally distributed population.
2. If it is a two sample test, both populations are assumed to have the same standard deviation (same shape).
3. The observations in the sample are independent of one another.

Assumptions for T-Tools: Independence

Independence

A major assumption underlying the Central Limit Theorem and/or the T-tools is **INDEPENDENCE**

INDEPENDENCE: "Whenever knowledge about one observation gives information about another observation"

- Example: If I measure your parents' height, that would give me information about your height without even measuring you.

Two common types of independence

- CLUSTER EFFECTS: Observations are naturally in subgroups
- SERIAL EFFECTS: Observations are collected over time/space.

Independence: Cluster Effects

Two common types of independence

- CLUSTER EFFECTS: Observations are naturally in subgroups
- SERIAL EFFECTS: Observations are collected over time/space.

Examples:

- Genetic Similarity: If we sample from all cattle at a ranch, some of the cattle could be dependent due to genetics
- Sociodemographic: People tend to spend their days with people of the same race/income/…

Independence: Serial Effects

Two common types of independence

- CLUSTER EFFECTS: Observations are naturally in subgroups
- SERIAL EFFECTS: Observations are collected over time/space.

Examples:

- Suppose we are testing the effect of new software on productivity. For all the subjects, we have them perform a task using old software and then the same task with new software
- We grow 10 different bacterial colonies on a single culture medium. The size of bacterial colonies near each other with be **negatively** associated due to resource competition

Independence: Serial Effects

Two common types of independence

- CLUSTER EFFECTS: Observations are naturally in subgroups
- SERIAL EFFECTS: Observations are collected over time/space.

How to diagnose?

- There are statistical techniques for estimating/testing/accounting for these effects.
- Some of these methods will be discussed later in the class/the next class.

For now, just be aware of these types of assumption violations

Outliers and Resistance

Outliers

A common issue in data analysis is the presence of **OUTLIERS**

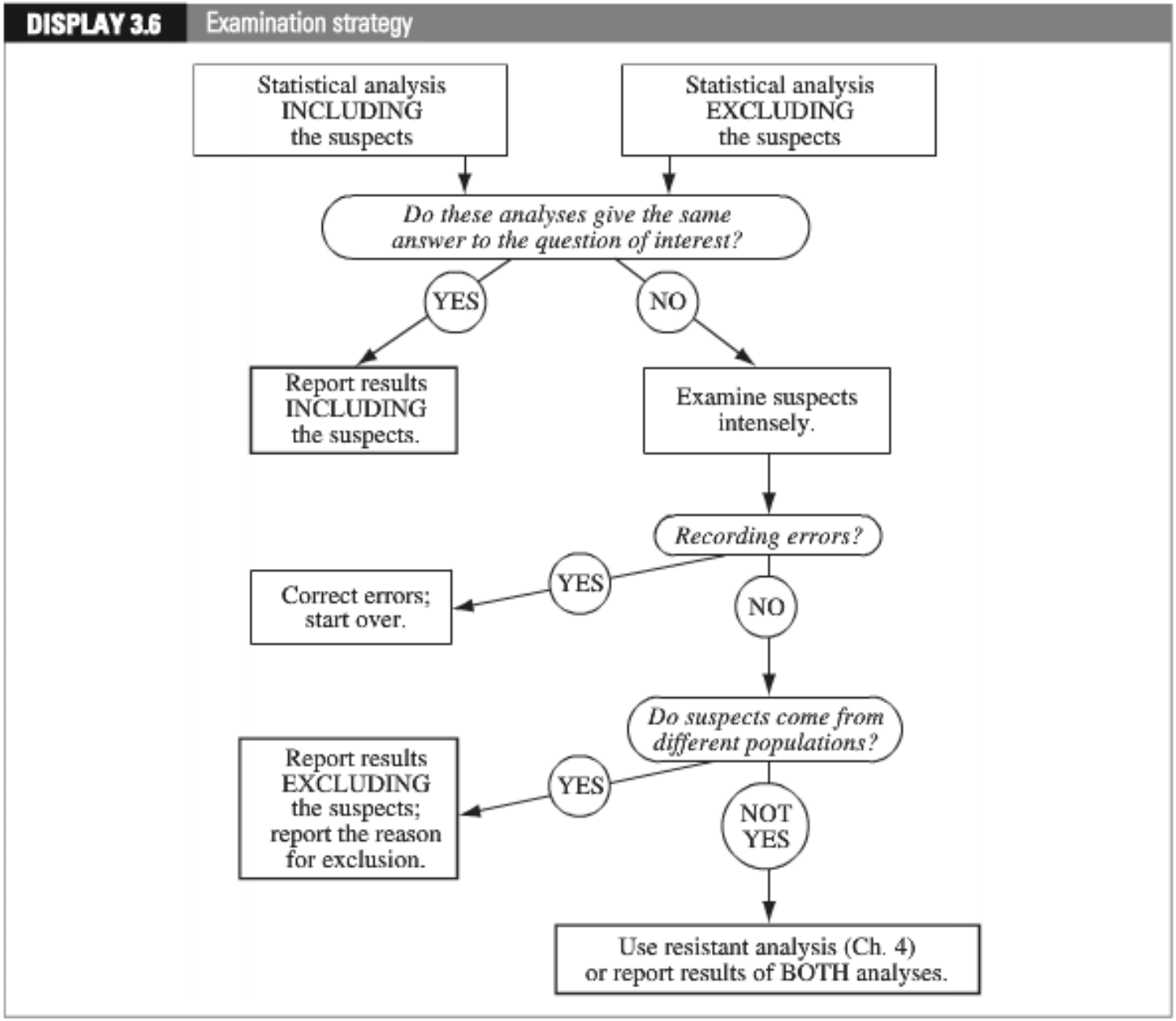
OUTLIERS: Observations that are judged to be “far” from “typical”

What is far? What is typical?

Sample averages are not resistant to outliers → neither are t-tools

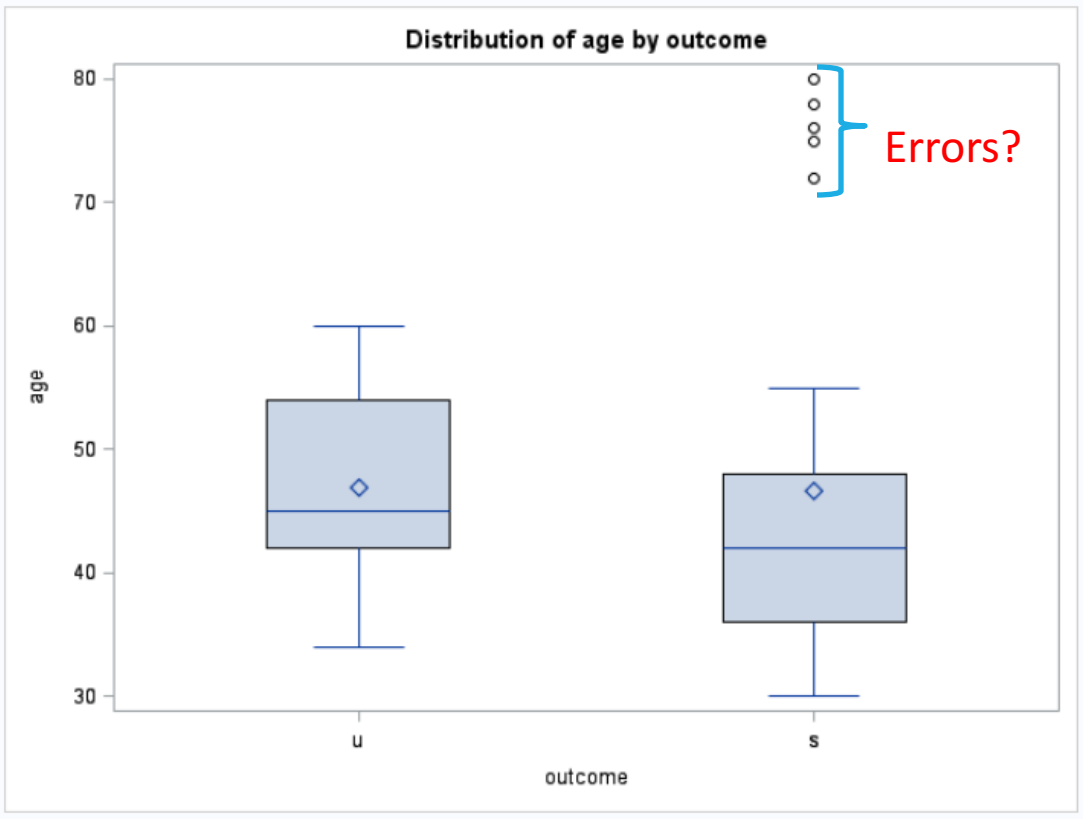
(Caveat: outlier is a bit pejorative/misleading. An alternative term would be “extreme” or “influential” observation)

Strategy For Data Sets with Outliers



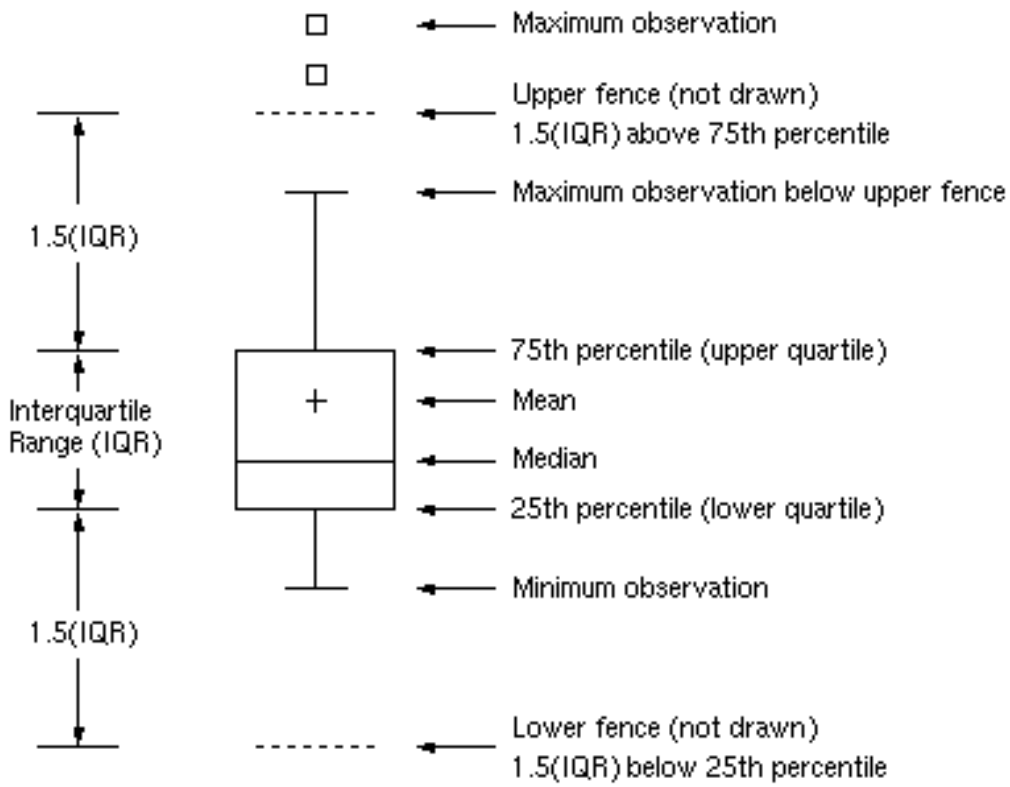
Example: Age by Promotion Status

```
proc boxplot data = promotion;
plot age*outcome / boxstyle = schematic;
run;
```

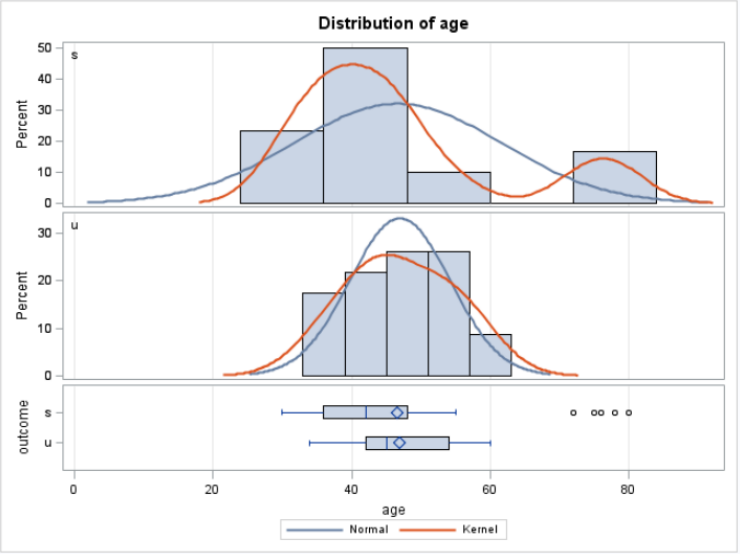


SAS procedure:
The “schematic” option

Figure 28.8: Schematic Box-and-Whiskers Plot



Re-check Assumptions



Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	22	4.27	0.0008

Results with Outliers

```
proc ttest data = promotion sides = l alpha = .05;
class outcome;
var age;
run;
```

$H_0: \mu_s = \mu_u$

$H_1: \mu_s < \mu_u$

Variable: age

outcome	N	Mean	Std Dev	Std Err	Minimum	Maximum
s	30	46.6333	14.9193	2.7239	30.0000	80.0000
u	23	46.9565	7.2204	1.5056	34.0000	60.0000
Diff (1-2)		-0.3232	12.2089	3.3837		

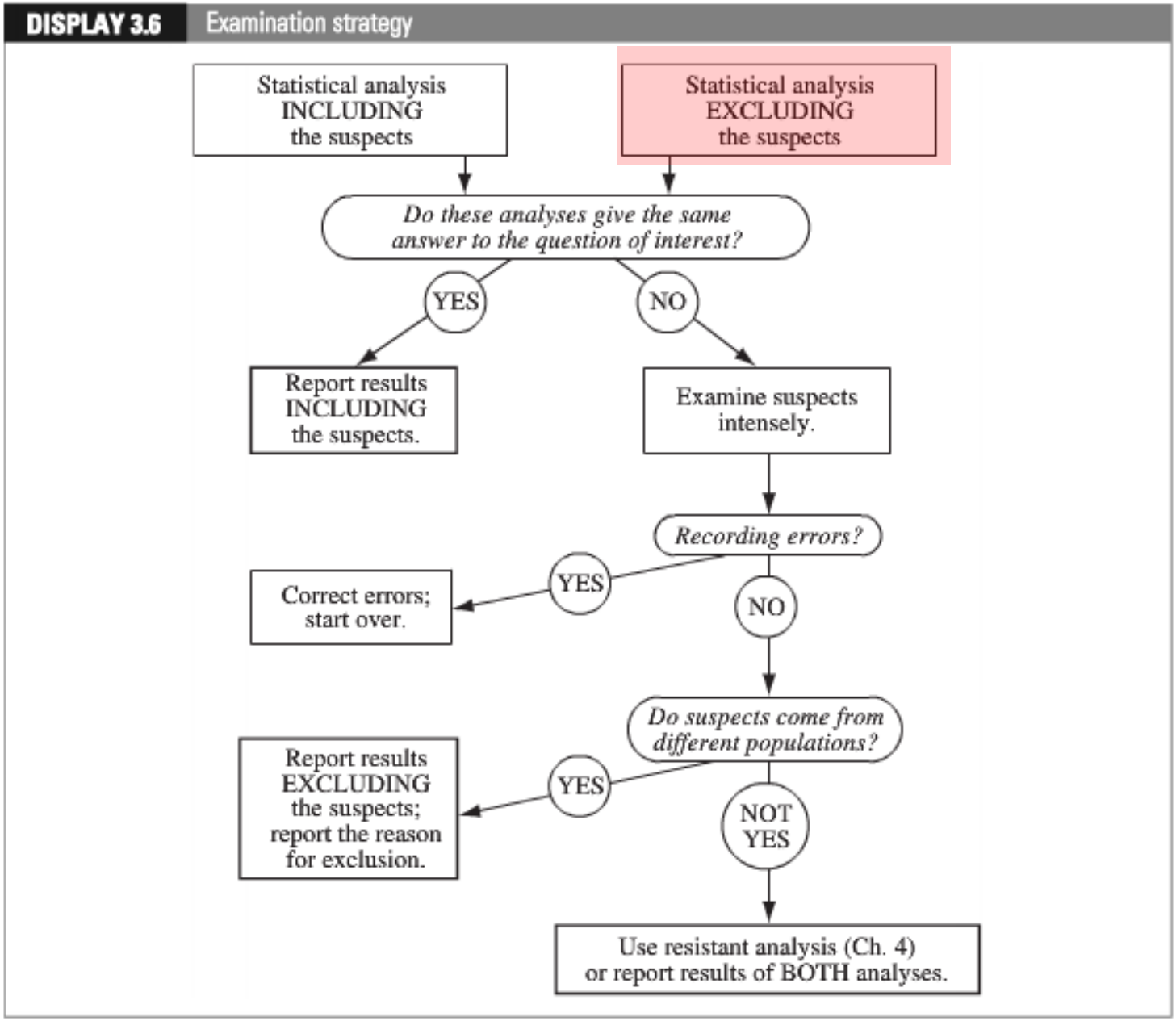
outcome	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
s		46.6333	41.0624	52.2043	14.9193	11.8818	20.0562
u		46.9565	43.8342	50.0789	7.2204	5.5842	10.2194
Diff (1-2)	Pooled	-0.3232	-Infty	5.3454	12.2089	10.2316	15.1406
Diff (1-2)	Satterthwaite	-0.3232	-Infty	4.9061			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	51	-0.10	0.4621
Satterthwaite	Unequal	44.011	-0.10	0.4589

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	22	4.27	0.0008

Statistical Conclusion:
The data provide no evidence against the null hypothesis that the mean age of the "successful" group is lower than the mean age of the "unsuccessful" group (one sided, two sample pooled t-test p-value = 0.4621). The assumptions about normality and equal variances seem suspect, however.

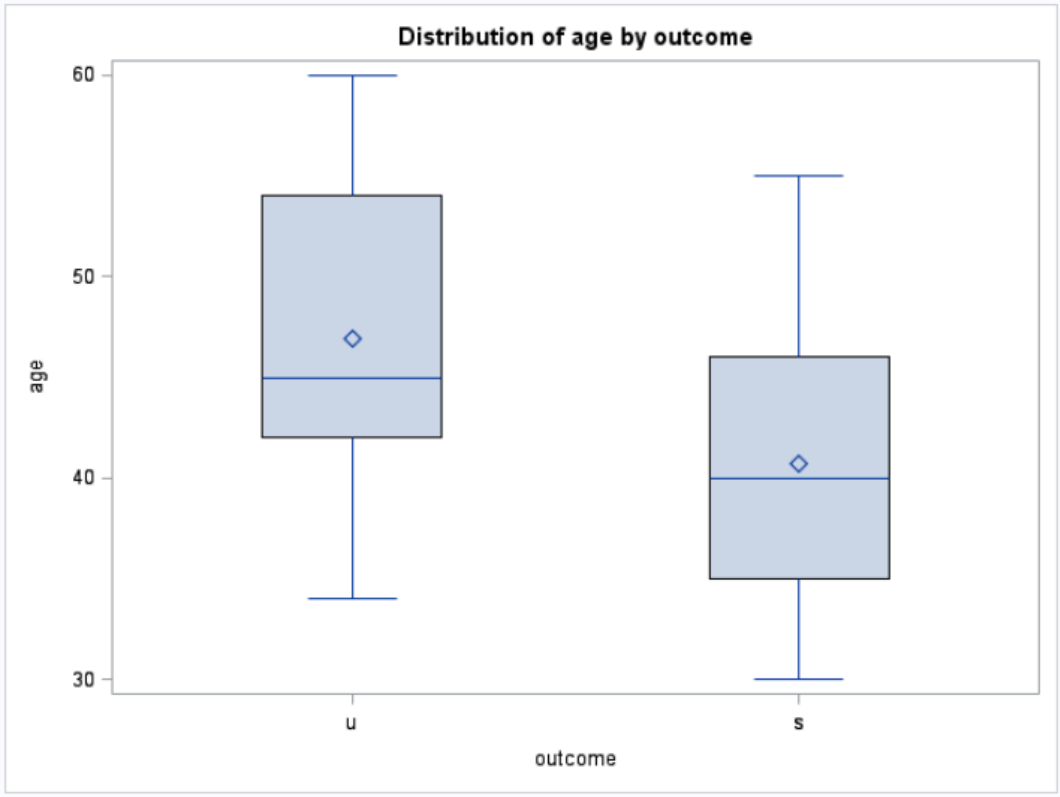
Strategy For Data Sets with Outliers



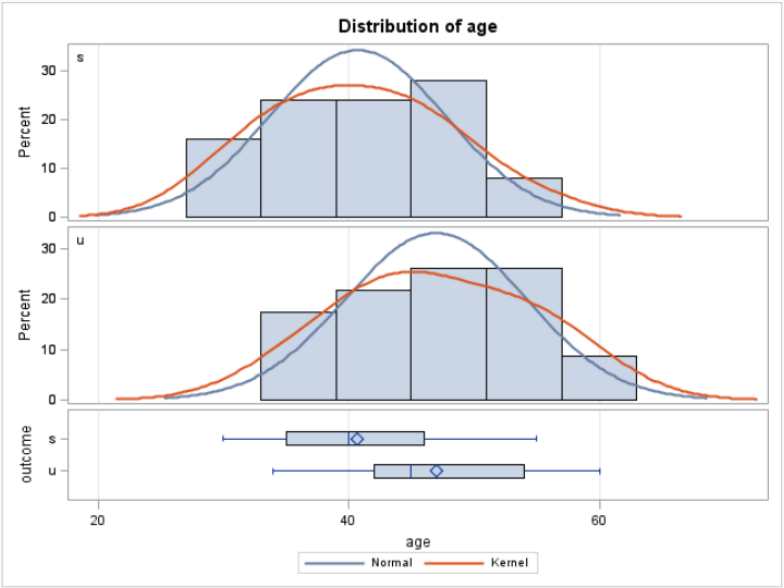
New Box Plot Without Outliers

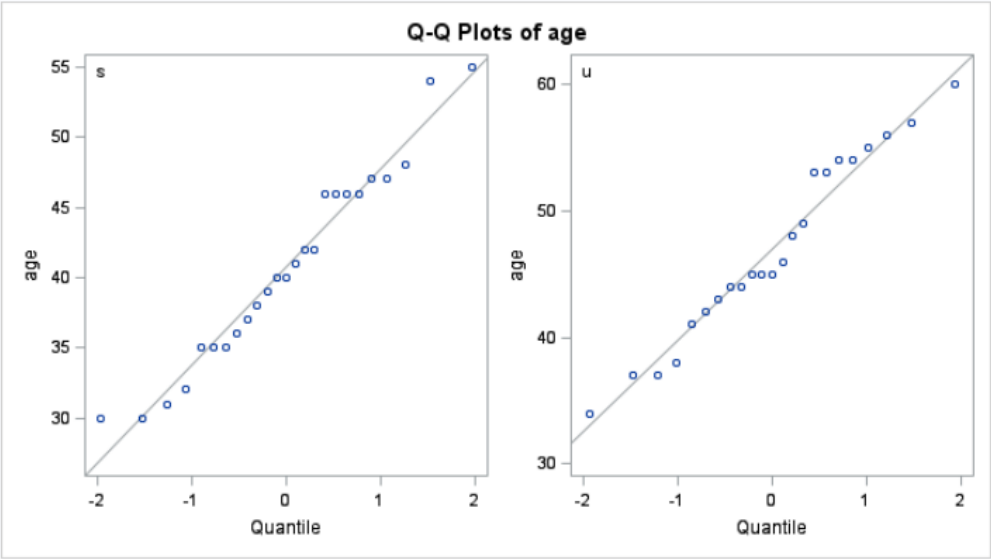
```
data promotion_no_outlier;
set promotion;
if outcome eq "s" and age > 66 then delete;
run;
```

```
proc boxplot data = promotion_no_outlier;
plot age*outcome / boxstyle = schematic;
run;
```



Re-check Assumptions





Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	24	1.07	0.8736

Results Excluding the Suspects

```
proc ttest data = promotion_no_outlier sides = l;
class outcome;
var age;
run;
```

$H_0: \mu_s = \mu_u$

$H_1: \mu_s < \mu_u$

Variable: age

outcome	N	Mean	Std Dev	Std Err	Minimum	Maximum
s	25	40.7200	6.9912	1.3982	30.0000	55.0000
u	23	46.9565	7.2204	1.5056	34.0000	60.0000
Diff (1-2)		-6.2365	7.1017	2.0519		

outcome	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
s		40.7200	37.8342	43.6058	6.9912	5.4589	9.7258
u		46.9565	43.8342	50.0789	7.2204	5.5842	10.2194
Diff (1-2)	Pooled	-6.2365	-Infty	-2.7921	7.1017	5.9014	8.9197
Diff (1-2)	Satterthwaite	-6.2365	-Infty	-2.7864			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	46	-3.04	0.0020
Satterthwaite	Unequal	45.375	-3.04	0.0020

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	24	1.07	0.8736

Statistical Conclusion:

There is strong evidence against the null hypothesis that the mean age of the "successful" group is lower than the mean age of the "unsuccessful" group (one sided, two sample pooled t-test p-value =0.002). We estimate that the mean difference is -6.2365 years, with up to -2.7921 years a plausible value (95% one-tailed pooled t-dist. confidence interval).

Included v. Excluded

Variable: age

outcome	N	Mean	Std Dev	Std Err	Minimum	Maximum
s	30	46.6333	14.9193	2.7239	30.0000	80.0000
u	23	46.9565	7.2204	1.5056	34.0000	60.0000
Diff (1-2)		-0.3232	12.2089	3.3837		

outcome	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
s		46.6333	41.0624	52.2043	14.9193	11.8818	20.0562
u		46.9565	43.8342	50.0789	7.2204	5.5842	10.2194
Diff (1-2)	Pooled	-0.3232	-Infty	5.3454	12.2089	10.2316	15.1406
Diff (1-2)	Satterthwaite	-0.3232	-Infty	4.9061			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	51	-0.10	0.4621
Satterthwaite	Unequal	44.011	-0.10	0.4589

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	22	4.27	0.0008

Included

Variable: age

outcome	N	Mean	Std Dev	Std Err	Minimum	Maximum
s	25	40.7200	6.9912	1.3982	30.0000	55.0000
u	23	46.9565	7.2204	1.5056	34.0000	60.0000
Diff (1-2)		-6.2365	7.1017	2.0519		

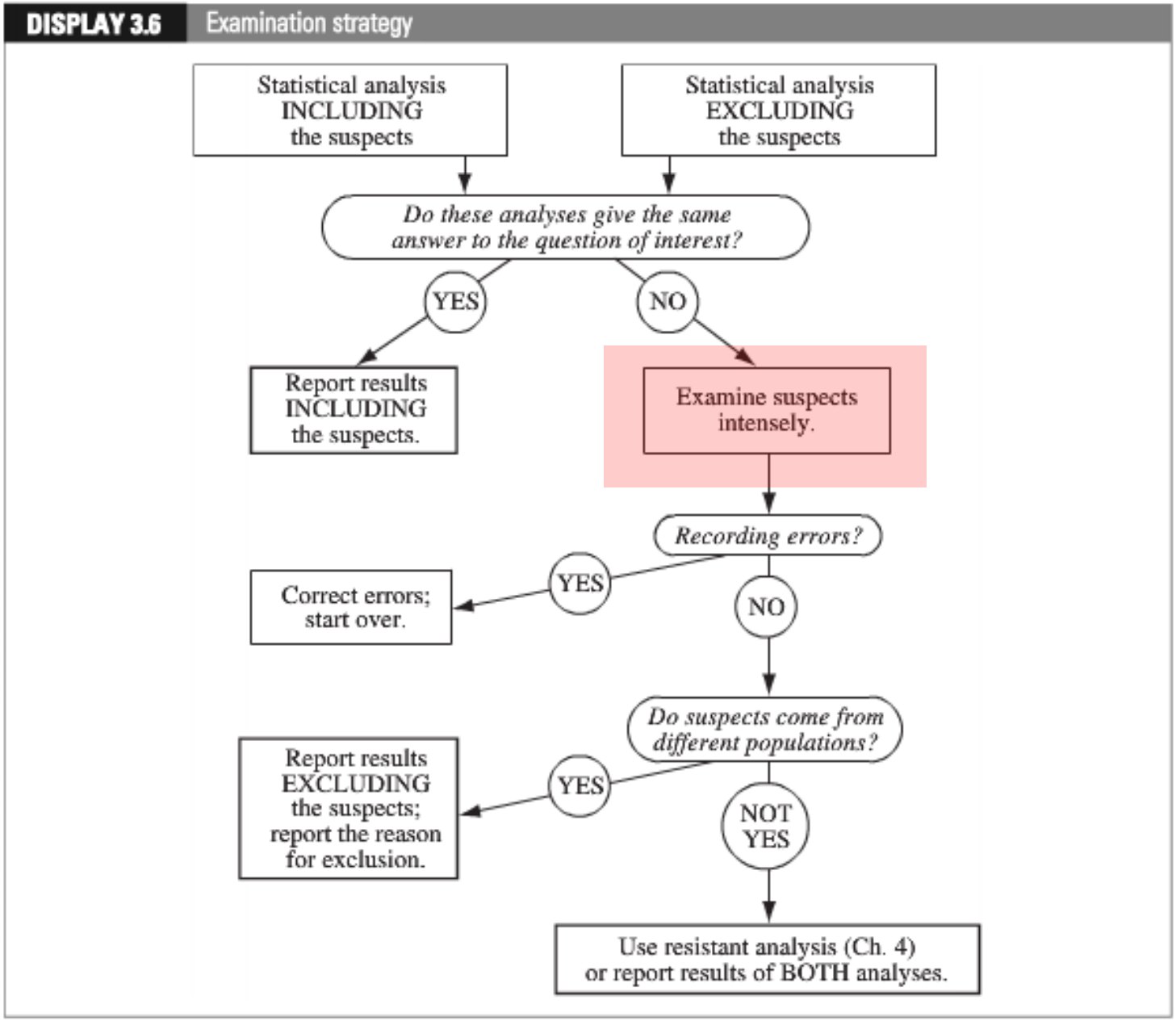
outcome	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
s		40.7200	37.8342	43.6058	6.9912	5.4589	9.7258
u		46.9565	43.8342	50.0789	7.2204	5.5842	10.2194
Diff (1-2)	Pooled	-6.2365	-Infty	-2.7921	7.1017	5.9014	8.9197
Diff (1-2)	Satterthwaite	-6.2365	-Infty	-2.7864			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	46	-3.04	0.0020
Satterthwaite	Unequal	45.375	-3.04	0.0020

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	24	1.07	0.8736

Excluded

Strategy For Data Sets with Outliers



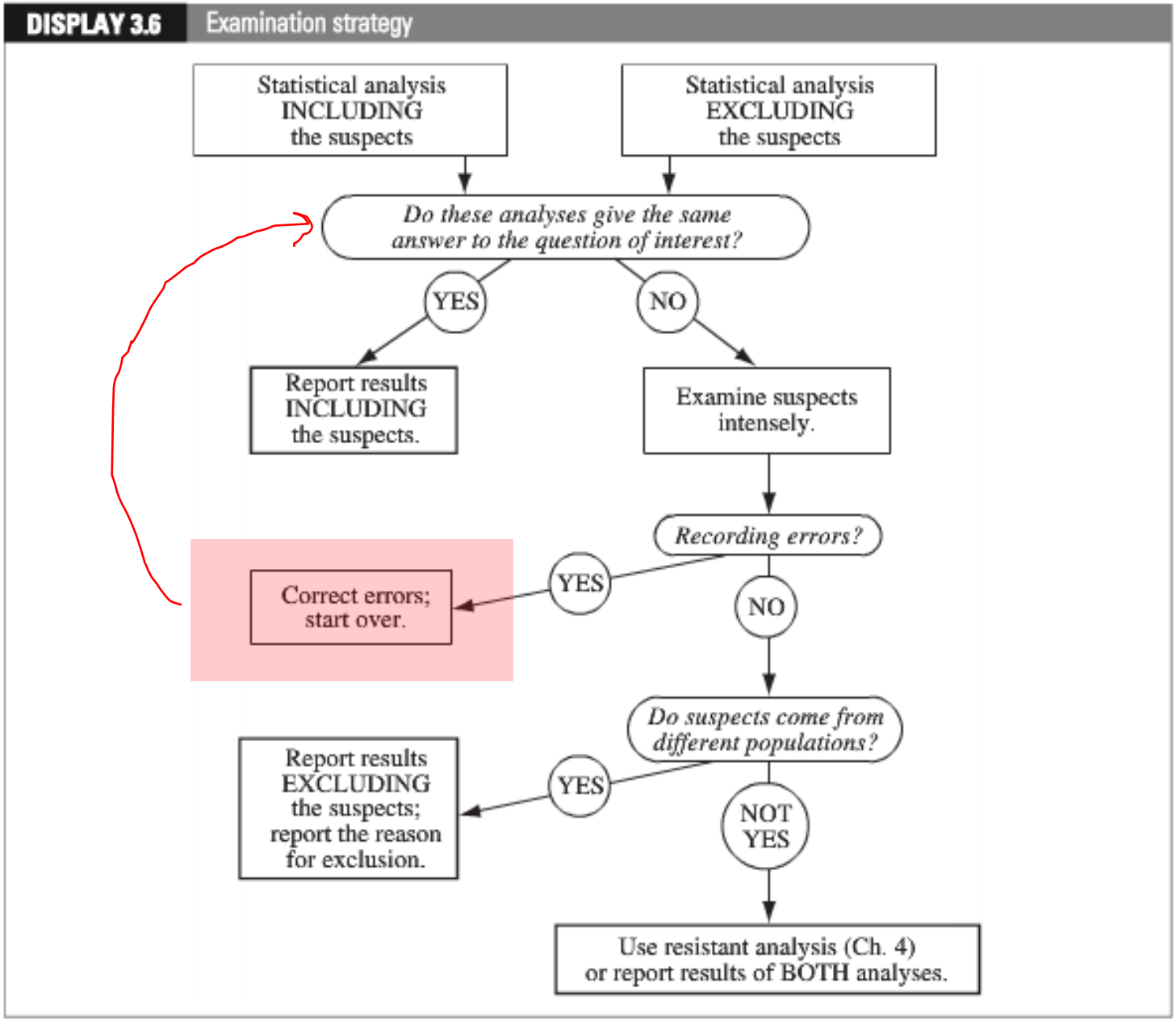
Examination of Outliers

Careful review of the outlying observations was conducted and it was found that the data point that was recorded as “80 years” was actually supposed to be recorded as a “30 years”!

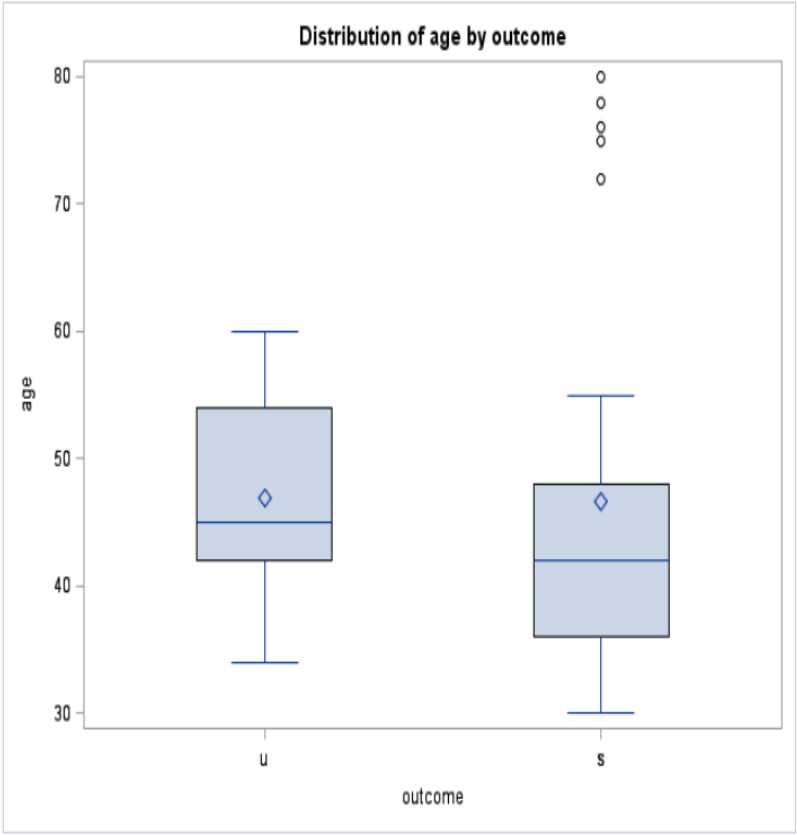
The other outliers were found to be recorded correctly.

```
data promotion_correction;
set promotion;
if age = 80 then age = 30;
run;
```

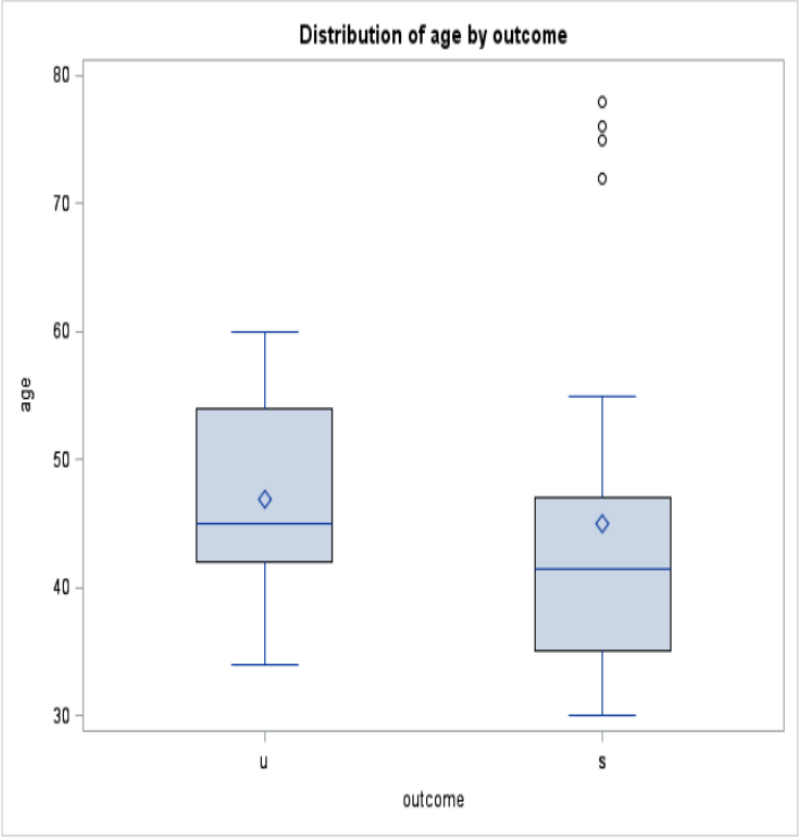
Strategy For Data Sets with Outliers



New Box Plot After Correction

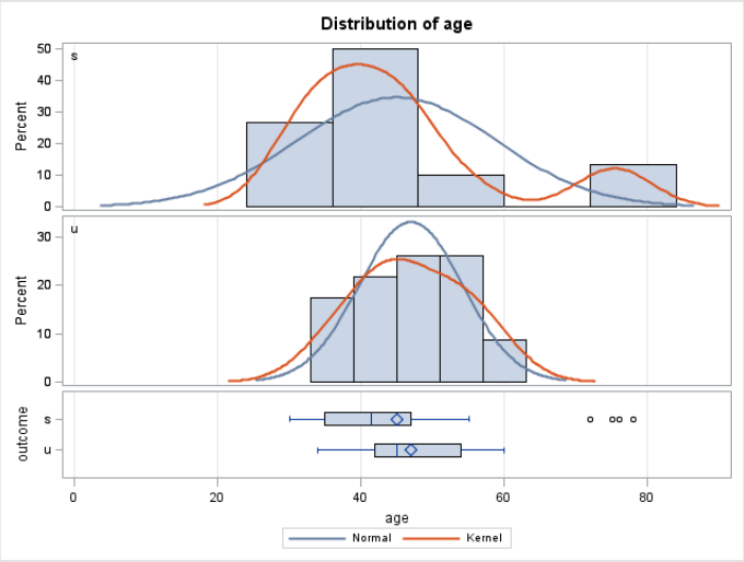


Original



After Correction

Another Look at Assumptions



Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	22	3.66	0.0026

Another Look At the Promotion Data

Variable: age

outcome	N	Mean	Std Dev	Std Err	Minimum	Maximum
s	30	44.9667	13.8152	2.5223	30.0000	78.0000
u	23	46.9565	7.2204	1.5056	34.0000	60.0000
Diff (1-2)		-1.9899	11.4463	3.1723		

outcome	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
s		44.9667	39.8080	50.1254	13.8152	11.0026	18.5720
u		46.9565	43.8342	50.0789	7.2204	5.5842	10.2194
Diff (1-2)	Pooled	-1.9899	-Infty	3.3247	11.4463	9.5926	14.1949
Diff (1-2)	Satterthwaite	-1.9899	-Infty	2.9418			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	51	-0.63	0.2666
Satterthwaite	Unequal	45.699	-0.68	0.2508

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	29	22	3.66	0.0026

After Correction

Variable: age

outcome	N	Mean	Std Dev	Std Err	Minimum	Maximum
s	25	40.7200	6.9912	1.3982	30.0000	55.0000
u	23	46.9565	7.2204	1.5056	34.0000	60.0000
Diff (1-2)		-6.2365	7.1017	2.0519		

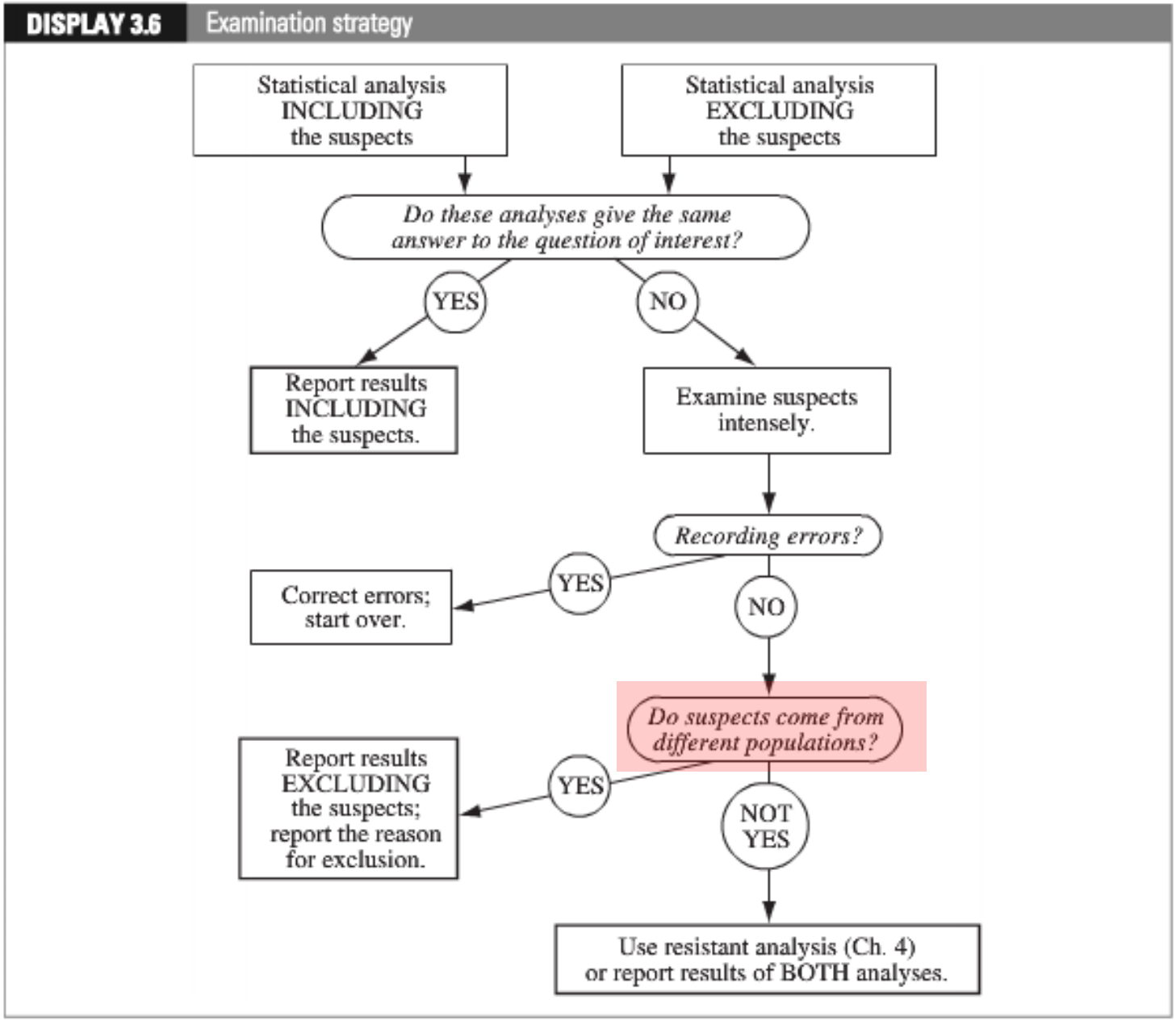
outcome	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
s		40.7200	37.8342	43.6058	6.9912	5.4589	9.7258
u		46.9565	43.8342	50.0789	7.2204	5.5842	10.2194
Diff (1-2)	Pooled	-6.2365	-Infty	-2.7921	7.1017	5.9014	8.9197
Diff (1-2)	Satterthwaite	-6.2365	-Infty	-2.7864			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	46	-3.04	0.0020
Satterthwaite	Unequal	45.375	-3.04	0.0020

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	24	1.07	0.8736

Excluding all Outliers (Suspects)

Strategy For Data Sets with Outliers



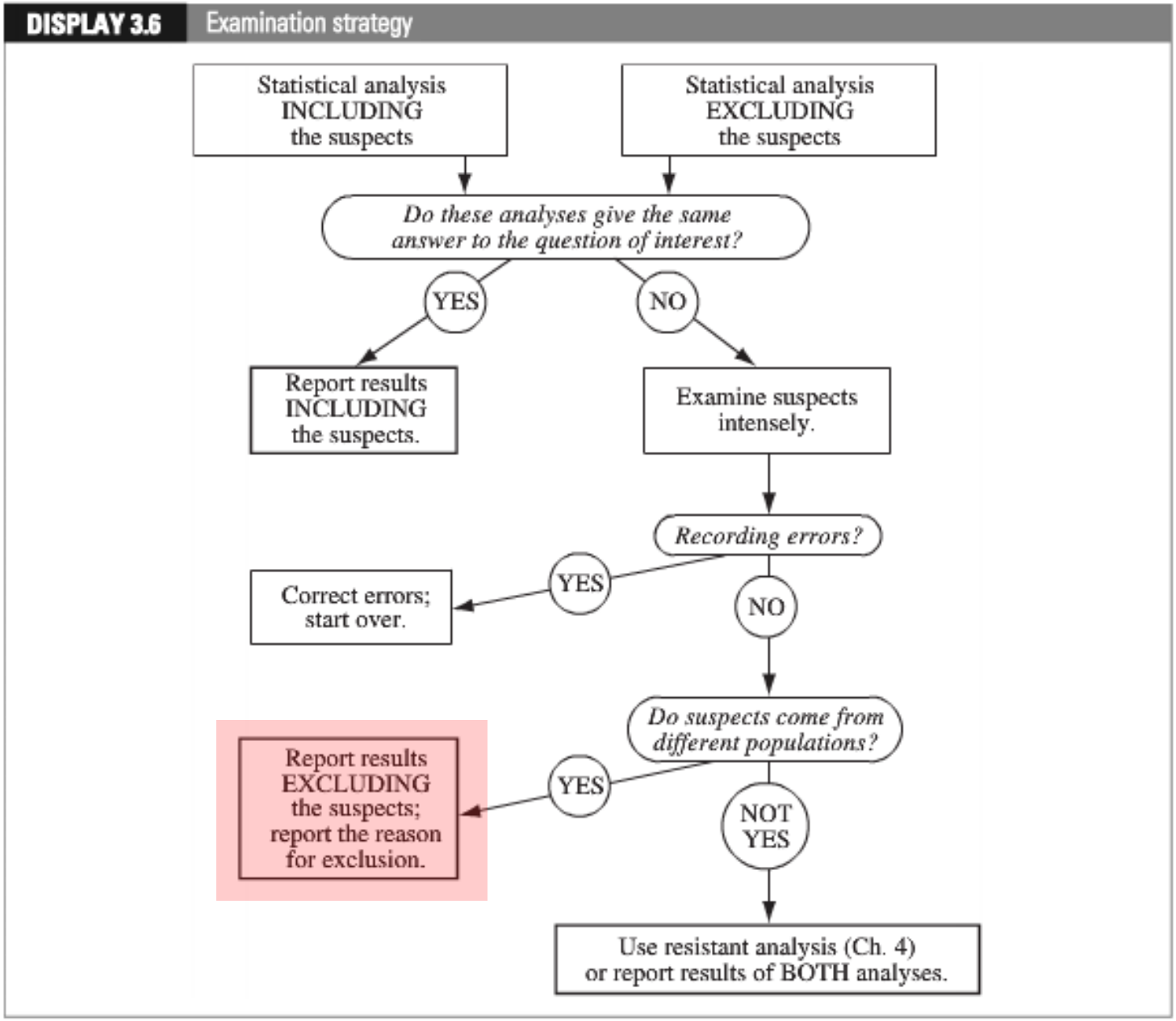
Examination of Outliers

- We carefully review the remaining outliers
- The >70 years are the heads of the company, and had given themselves a promotion.
- We are interested in age discrimination. Management is not going to discriminate against itself….

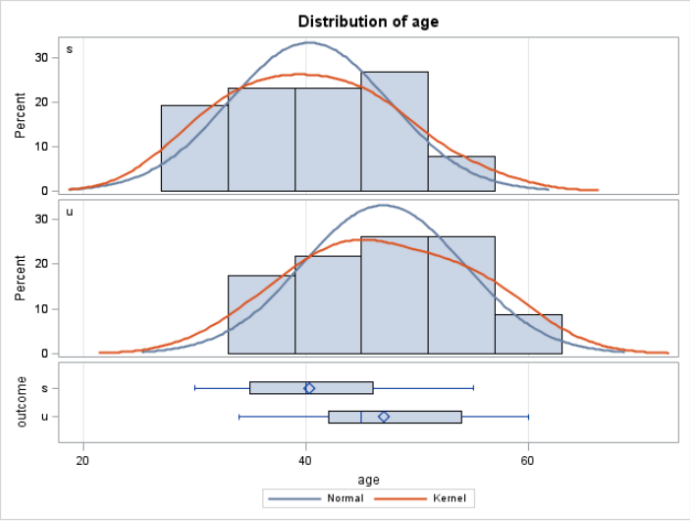
  → they are not a part of the population we are interested in.
- Therefore, we will exclude these observations from the study.

```
data promotion_no_outlier_correction;
set promotion_correction;
if outcome eq "s" and age > 66 then delete;
run;
```

Strategy For Data Sets with Outliers



With Correction with no outliers.



Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	25	1.02	0.9637

Report Results Excluding the Suspects

Variable: age

outcome	N	Mean	Std Dev	Std Err	Minimum	Maximum
s	26	40.3077	7.1653	1.4052	30.0000	55.0000
u	23	46.9565	7.2204	1.5056	34.0000	60.0000
Diff (1-2)		-6.6488	7.1912	2.0585		

outcome	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
s		40.3077	37.4136	43.2018	7.1653	5.6194	9.8910
u		46.9565	43.8342	50.0789	7.2204	5.5842	10.2194
Diff (1-2)	Pooled	-6.6488	-Infty	-3.1949	7.1912	5.9864	9.0075
Diff (1-2)	Satterthwaite	-6.6488	-Infty	-3.1920			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	47	-3.23	0.0011
Satterthwaite	Unequal	46.184	-3.23	0.0011

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	22	25	1.02	0.9637

$H_0: \mu_s = \mu_u$

$H_1: \mu_s < \mu_u$

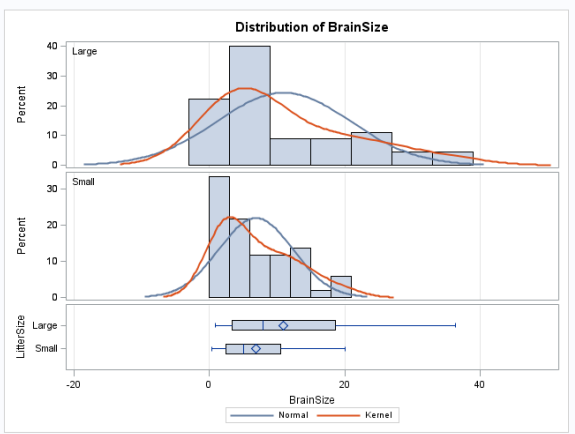
Statistical Conclusion:

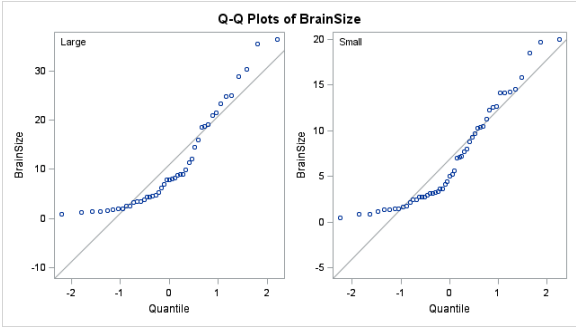
There is strong evidence against the null hypothesis that the mean age of the "successful" group is lower than the mean age of the "unsuccessful" group (one sided, two sample pooled t-test p-value = 0.0011). We estimate that the mean difference is -6.6488 years, with up to -3.1949 years a plausible value (95% one-tailed pooled t-dist. confidence interval).

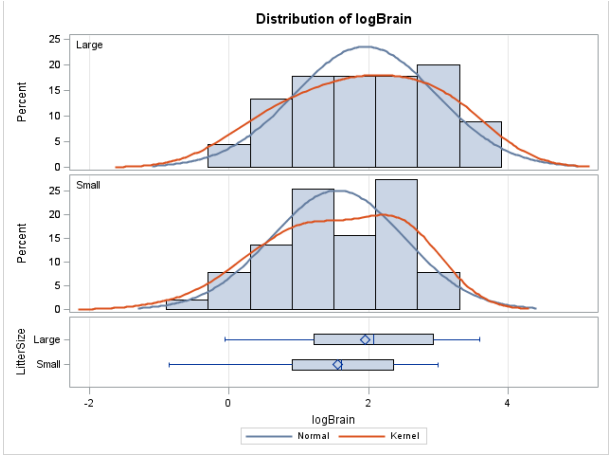
We excluded workers with age > 66 years due to being be management.

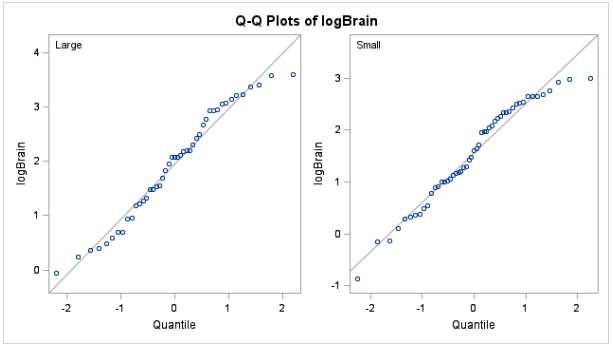
Transformations

Checking Assumptions…









Checking Assumptions…

LitterSize	N	Mean	Std Dev	Std Err	Minimum	Maximum
Large	45	10.9684	9.8369	1.4664	0.9400	36.3500
Small	51	6.8859	5.4603	0.7646	0.4200	20.0000
Diff (1-2)		4.0826	7.8201	1.5994		

LitterSize	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Large		10.9684	8.0131	13.9236	9.8369	8.1435	12.4260
Small		6.8859	5.3501	8.4216	5.4603	4.5687	6.7876
Diff (1-2)	Pooled	4.0826	0.9064	7.2588	7.8201	6.8442	9.1230
Diff (1-2)	Satterthwaite	4.0826	0.7815	7.3836			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	94	2.55	0.0123
Satterthwaite	Unequal	66.83	2.47	0.0161

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	44	50	3.25	<.0001

Original Data

LitterSize	N	Mean	Std Dev	Std Err	Minimum	Maximum
Large	45	1.9494	1.0163	0.1515	-0.0619	3.5932
Small	51	1.5525	0.9522	0.1333	-0.8675	2.9957
Diff (1-2)		0.3970	0.9827	0.2010		

LitterSize	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Large		1.9494	1.6441	2.2548	1.0163	0.8413	1.2838
Small		1.5525	1.2846	1.8203	0.9522	0.7967	1.1837
Diff (1-2)	Pooled	0.3970	-0.00211	0.7960	0.9827	0.8601	1.1465
Diff (1-2)	Satterthwaite	0.3970	-0.00394	0.7979			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	94	1.98	0.0512
Satterthwaite	Unequal	90.684	1.97	0.0522

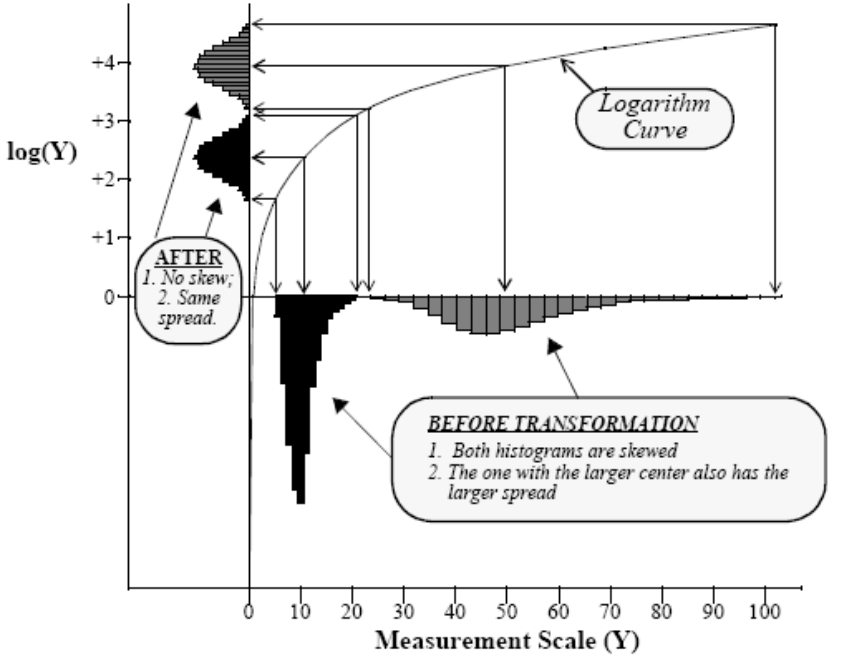
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	44	50	1.14	0.6529

Log Transformed Data

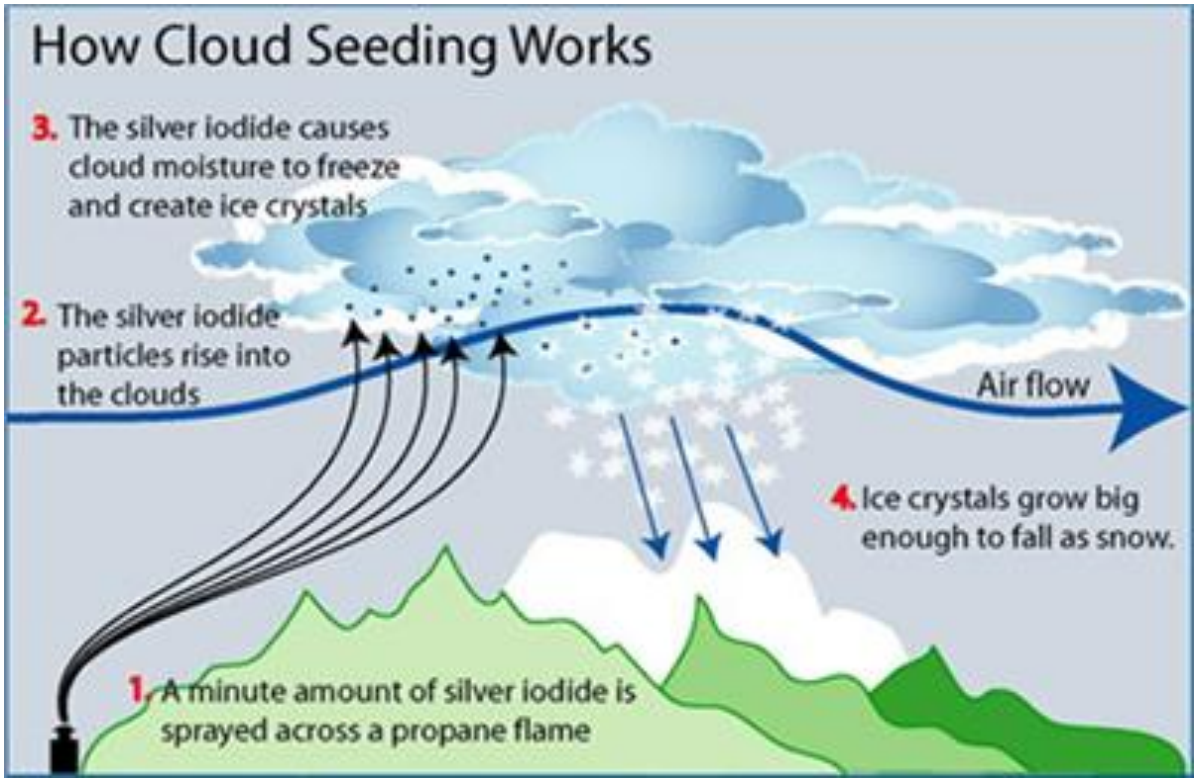
Log Transformation

Display 3.8 **p. 69**

The logarithmic transformation used to arrive at favorable conditions for the two-sample t-analysis



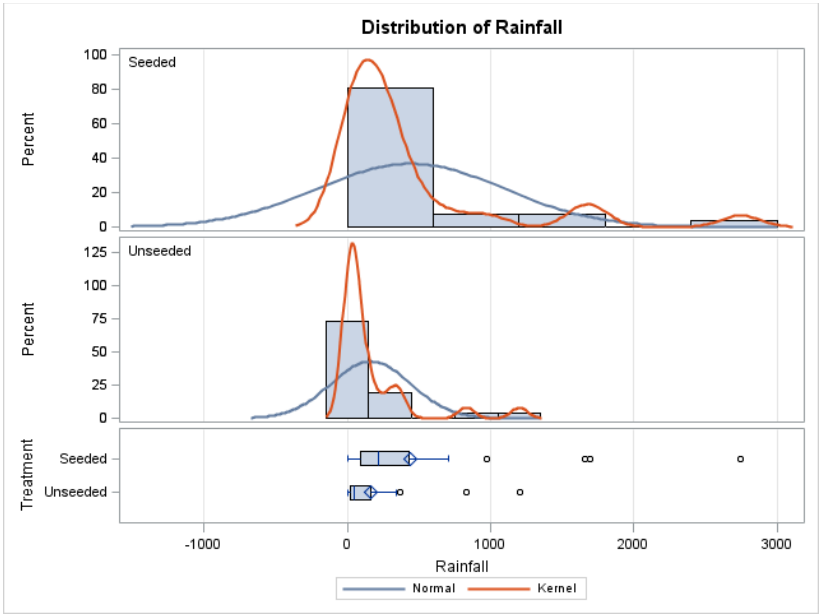
Example: Cloud Seeding

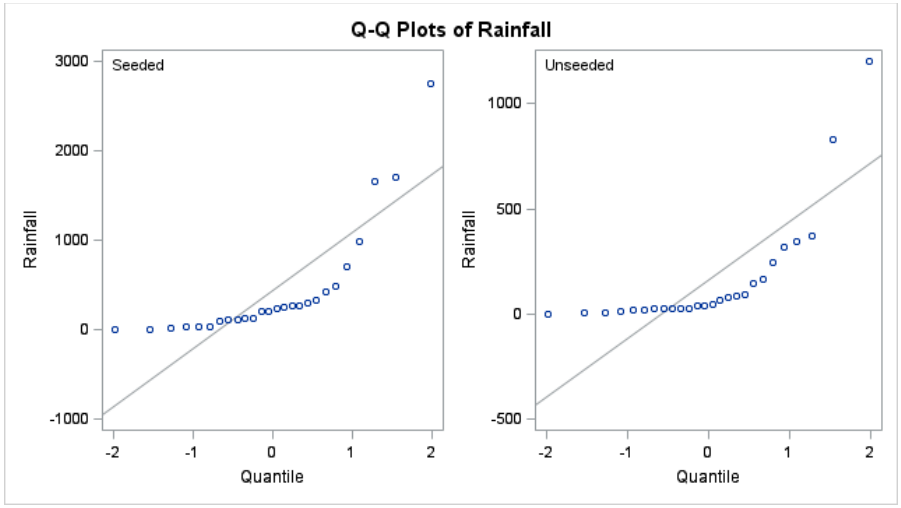


Does Cloud Seeding Work?

- On days that were deemed suitable for cloud seeding, a random mechanism was used to decide whether to seed the target cloud on that day or to leave it unseeded as a control
- Precipitation was measured as the total rain volume falling from the cloud base following the seeding run (as measured by radar)
- We would like to test whether cloud seeding is effective in increasing precipitation.

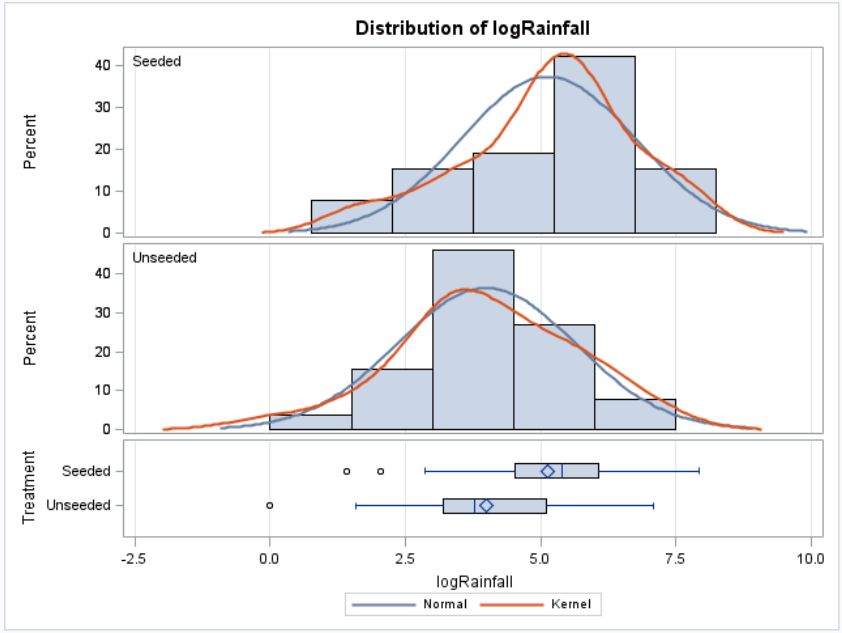
Cloud Seeding: Original Data

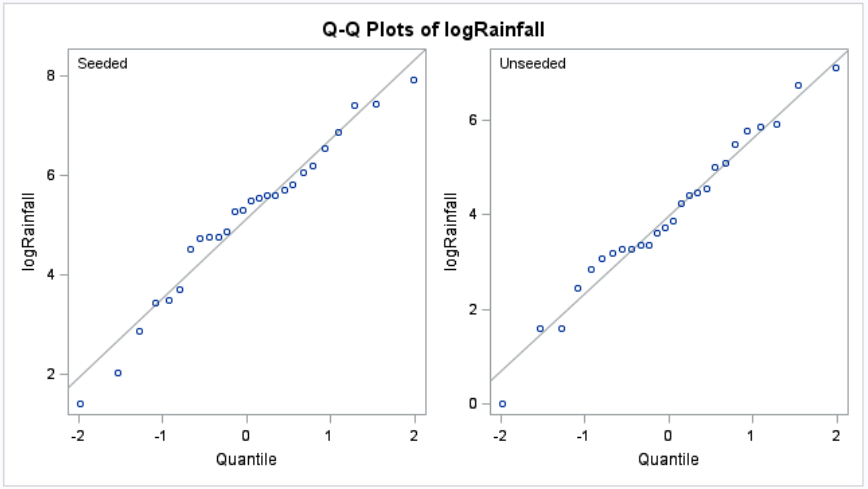





```
proc ttest data = cloud sides = u alpha = .05;
class treatment;
var rainfall;
run;
```

After Log Transformation





Create lograin

Analyze lograin

```
data cloud;
set cloud;
lograin = log(rainfall);
run;

proc ttest data = cloud sides = 2 alpha = .05;
class treatment;
var lograin;
run;
```

Hypothesis Test

Variable: lograin

Treatment	N	Mean	Std Dev	Std Err	Minimum	Maximum
Seeded	26	5.1342	1.5995	0.3137	1.4110	7.9178
Unseeded	26	3.9904	1.6418	0.3220	0	7.0922
Diff (1-2)		1.1438	1.6208	0.4495		

Treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Seeded		5.1342	4.4881	5.7802	1.5995	1.2544	2.2080
Unseeded		3.9904	3.3272	4.6536	1.6418	1.2876	2.2664
Diff (1-2)	Pooled	1.1438	0.3904	Infty	1.6208	1.3562	2.0148
Diff (1-2)	Satterthwaite	1.1438	0.3904	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	50	2.54	0.0070
Satterthwaite	Unequal	49.966	2.54	0.0070

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	25	25	1.05	0.8971

H_0: Cloud Seeding does not work.
H_1: Cloud Seeding does work.

$H_0: \mu_{seeded} = \mu_{unseeded}$
$H_1: \mu_{seeded} > \mu_{unseeded}$

$H_0: \mu_{log(seeded)} = \mu_{log(unseeded)}$
$H_1: \mu_{log(seeded)} > \mu_{log(unseeded)}$

Statistical Conclusion:
There is strong evidence against the null hypothesis that the mean log(precip.) of the "seeded" group is greater than the mean log(precip.) of the "unseeded" group (one sided, two sample pooled t-test p-value = 0.007)…

H_0: *Cloud Seeding does not work.*
H_1: *Cloud Seeding does work.*

$H_0: \mu_{seeded} = \mu_{unseeded}$
$H_1: \mu_{seeded} > \mu_{unseeded}$ → $H_0: \mu_{log(seeded)} = \mu_{log(unseeded)}$
$H_1: \mu_{log(seeded)} > \mu_{log(unseeded)}$

Hypothesis Test

Variable: lograin

Treatment	N	Mean	Std Dev	Std Err	Minimum	Maximum
Seeded	26	5.1342	1.5995	0.3137	1.4110	7.9178
Unseeded	26	3.9904	1.6418	0.3220	0	7.0922
Diff (1-2)		1.1438	1.6208	0.4495		

Treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Seeded		5.1342	4.4881	5.7802	1.5995	1.2544	2.2080
Unseeded		3.9904	3.3272	4.6536	1.6418	1.2876	2.2664
Diff (1-2)	Pooled	1.1438	0.3904	Infty	1.6208	1.3562	2.0148
Diff (1-2)	Satterthwaite	1.1438	0.3904	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	50	2.54	0.0070
Satterthwaite	Unequal	49.966	2.54	0.0070

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	25	25	1.05	0.8971

Statistical Conclusion (continued):
… We estimate that the median precip. for the "seeded" group is $e^{1.1438} = 3.1387$ times the median precip. for the "unseeded" group, with at least a multiplicative effect of $e^{0.3904} = 1.4776$ (95% one-tailed pooled t-dist. confidence interval).

Scope of Inference:
As this is a randomized study, we can infer that seeding mechanism caused this difference. We can only extend these results to this location and to "suitable" days.