Statistical Testing

Statistical Testing

- There are many types of decisions that people would like to make using data.
  - Is a new additive better than the old one?
  - Does the new test machine give the same results as the old one?
  - Did the changes we made upstream affect the product quality?
- We call this type of decision making using data, ***statistical testing***.

Steps in a Statistical Test

1. State the null hypothesis (***Ho***), and alternative hypothesis (***Ha***).
2. Collect a representative sample from the population of interest.
3. Compute a test statistic from sample values, <u>assuming ***Ho*** is true.</u>

Steps in a Statistical Test

4. Determine the probability of getting a more extreme value of the test statistic. (The ***p-value***.)
5. If the ***p-value*** is <u>small</u> (e.g. **p < 0.05**), the result is unlikely if ***Ho*** is true, so ***we reject Ho***. Otherwise, we <u>fail to reject</u> ***Ho***.

Steps in a Statistical Test

- The hypothesis in Step 1 and the formula for the test statistic in Step 3 change depending on the specific situation.
  - We will discuss later.
- ***Minitab*** will calculate the statistics, but it is a good idea to understand what hypothesis is being tested.

Statistical Testing

- When using data to make a statistical decision, there are **two** types of ***errors*** that can occur.
- **TYPE 1 ERROR** – You decide that something has happened, when in fact, the result is only due to random variation.
  - For example, rejecting good product is known as Producer's risk and is a Type 1 error .

Statistical Testing

- **TYPE 2 ERROR** – You fail to detect a real change has happened.
  - For example, *accepting* bad product is called Consumer's risk and is a Type 2 error.

Statistical Testing

- In statistical testing we want the probability (risk) of making a **Type 1** error small, usually no greater than 10%.
- But, the small **Type 1** error will increase the probability of a **Type 2** error.

Statistical Testing

- ***Alpha (α)*** risk ≡ probability of **Type 1** error α, is known as the ***p-value*** and should be small (we said $p < 0.05$).
- ***Beta (β)*** risk ≡ probability of **type 2** error β, is dependent on α and the amount of data (***n***).

Statistical Testing

- The following table may help to clarify the two decision error types discussed.

	True Situation	
Your Decision	No Difference	Difference
No Difference	No error	***Type II Error Control with β (consumer's risk)***
Difference	***Type I Error Control with α (producer's risk)***	No error ***(Control with Power)***

Statistical Testing

- To control for Type 1 error, the α risk level can be adjusted (e.g. from .05 to .01).
  - One problem with this is that it may increase the probability of a Type 2 error.
- There are several choices that can help to control the Type 2 error rate, given the Type 1 error level is chosen at α.

Statistical Testing

- Ways to decrease the Type 2 error (with a given α risk):
  - The researcher may **increase** the difference that may be detected.
  - Run a **pre-test** to get a better estimate of variability.
  - **Increase** the sample size, my favorite.

Statistical Testing

- A way to assess these choices is to examine the ***power*** of the statistical test.
- ***Power analysis*** a way to analyze, assess and adjust your **sample size** and **risk**, given the **difference** you desire to detect and the **variability** in the data.

Power of a Statistical Test

- ***Power*** is defined as the probability of correctly rejecting the null hypothesis; that is, rejecting Ho when you should be rejecting Ho.
  - The probability of detecting a “true” effect when it exists.

Power of a Statistical Test

- You should design an experiment with a <u>sample size</u> (***n***) that achieves the highest possible ***power*** given the α risk (Type 1), minimum detectable difference (***d***) and estimates of variability (σ).

- We will discuss this in detail later.

One Sample *t* Test

Comparing a population **average** (or mean) to a **standard**.

Hypothesis:

$$H_0 : \mu = \text{constant}$$

$$H_a : \mu \neq \text{constant (2 sided test)}$$

or

$$H_a : \mu > \text{constant (1 sided test)}$$

or

$$H_a : \mu < \text{constant (1 sided test)}$$

One Sample *t* Test

Comparing a population **average** to a **standard**.

Test Statistic:

$$t = \frac{\left(\bar{x} - \text{constant}\right)}{\left(s/\sqrt{n}\right)}$$

s is the standard deviation

n is the sample size

One Sample *t* Test

- Suppose we want to know if the **average** of a set of moisture measurements are less than 8%.
- The one sample ***t*** test would be the appropriate statistical test.
  - **WHY**?

One Sample *t* Test

1. The ***hypothesis*** we wish to test is:
   - Ho: $\mu = 8$
   - Ha: $\mu < 8$
2. There was a ***representative sample*** of 50 sheets collected at rollup of 50 jumbos.
3. Use Minitab to ***calculate*** the ***t*** statistic.

One Sample *t* Test

- Open the '***Minitab Dataset 6.xls***' worksheet.
- There are 3 measures across the sheet for moisture (in percentage).
  - These are repeated measures, although they do represent CD variation too.
- Average the ***Moisture*** by the ***Sample ID*** and create a new summary worksheet.

One Sample *t* Test

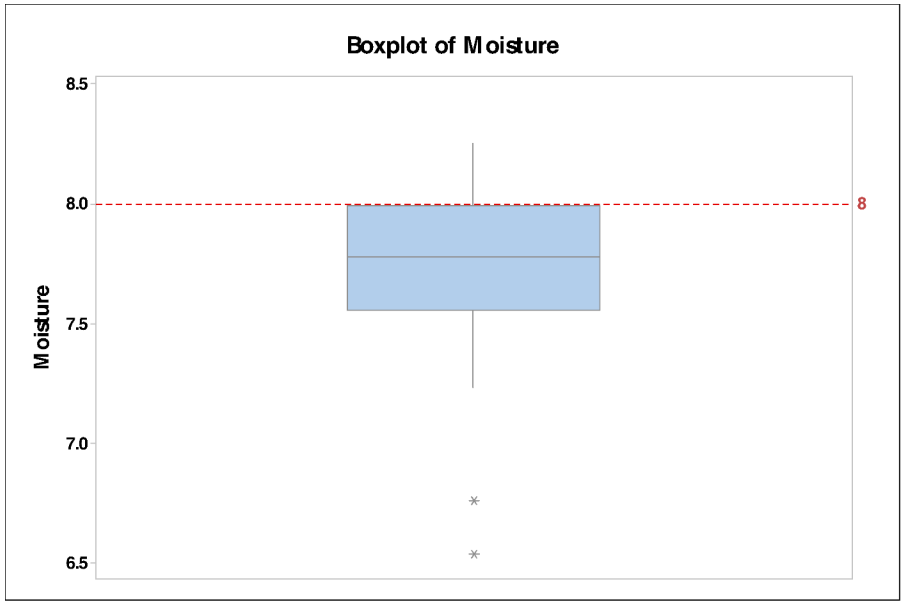
- Average the ***Moisture*** by the ***Sample ID*** and create a new summary worksheet.
- The worksheet should resemble the following:

Worksheet 2 ***

↓	C1	C2
	Sample ID	Moisture
1	1	8.02000
2	2	7.65667
3	3	7.29000
4	4	8.16000

One Sample *t* Test

- Create a **Simple Boxplot** of ***Moisture***.
- Add a **Reference Line** at Y = 8.

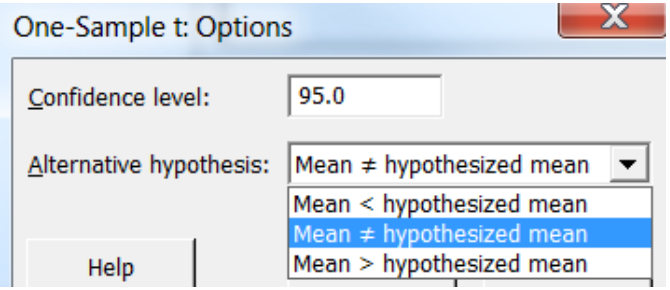


One Sample *t* Test

- Select **Stat** → **Basic Statistics**
  → **1-Sample t…**
- Select ***Moisture*** in the **One or more samples, each in a column** box.
- Click the check box next to **Perform hypothesis test**.
- Enter ***8*** in the **Hypothesized mean** box.

One Sample *t* Test

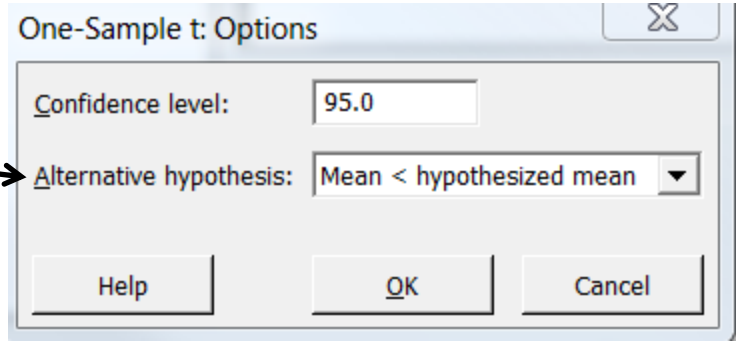
- Click on the **Options…** button → click on the down arrow to bring up the menu.



- All the different hypothesis tests that were discussed on slide 10-16 are listed.

One Sample *t* Test

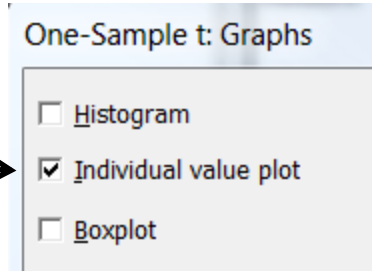
- Now select Mean < hypothesized mean from the pull down menu as the **Alternative hypothesis** option.



- Click **OK** once.

One Sample *t* Test

- Click the **Graphs…** button → select only the **Individual value plot**.



- Click **OK** twice.
- Go to the **Sessions Folder**

One Sample *t* Test

```
Test of μ = 8 vs < 8

Variable   N     Mean    StDev  SE Mean    95% Upper Bound      T      P
Moisture  50  7.73573  0.34778  0.04918            7.81819  -5.37  0.000
```

Mean Moisture

***t* statistic**

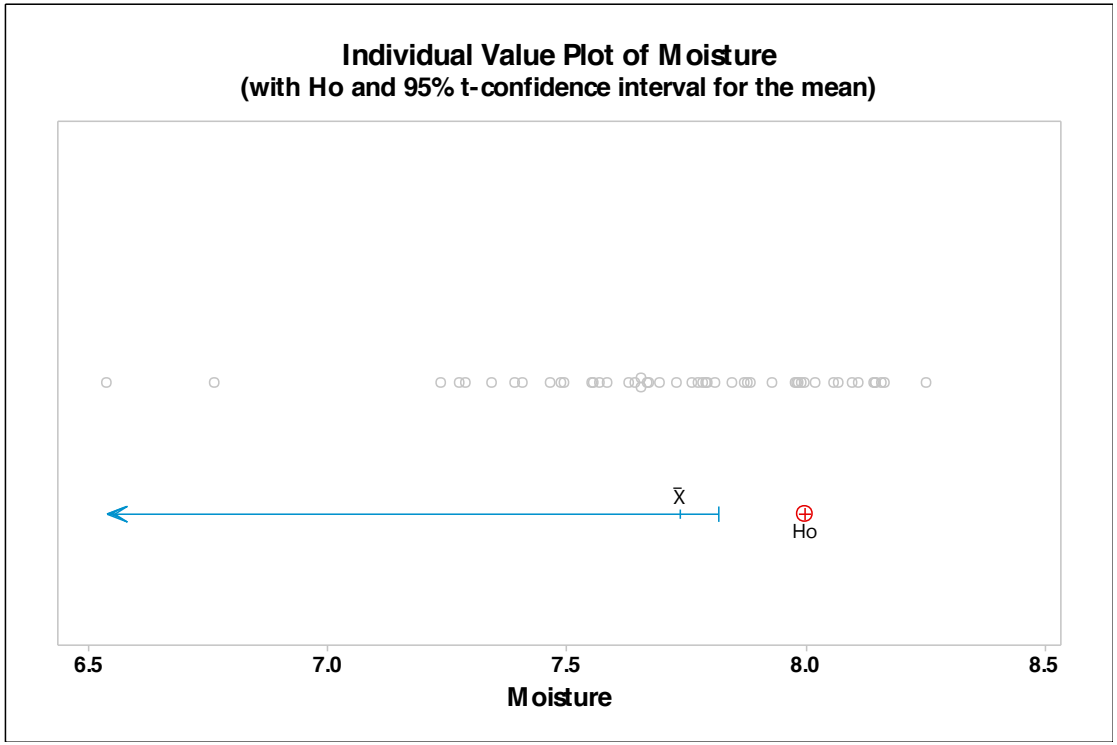
p value

<u>Interpretation</u>:

3. The ***t*** statistic is **-5.37**
4. The p value = 0.000, which implies $p < 0.05$
5. Therefore, we reject the null hypothesis (Ho) and conclude the average moisture is less than 8%

One Sample *t* Test

Plot shows the average ***Moisture*** $(\bar{X})$ is **statistically significantly lower** than 8% (***Ho***)



Two Sample *t* Test

Comparing two (2) population means from two different and independent sample groups.

Hypothesis:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2 \text{ (2 sided test)}$$

or

$$H_a : \mu_1 > \mu_2 \text{ (1 sided test)}$$

or

$$H_a : \mu_1 < \mu_2 \text{ (1 sided test)}$$

Two Sample *t* Test

Comparing two (2) population means from two different and independent sample groups.

Test Statistic:

(Assumes unequal variances)

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

s_1^2 is the variance for group 1

s_2^2 is the variance for group 2

n_1 is the sample size for group 1

n_2 is the sample size for group 2

Two Sample *t* Test

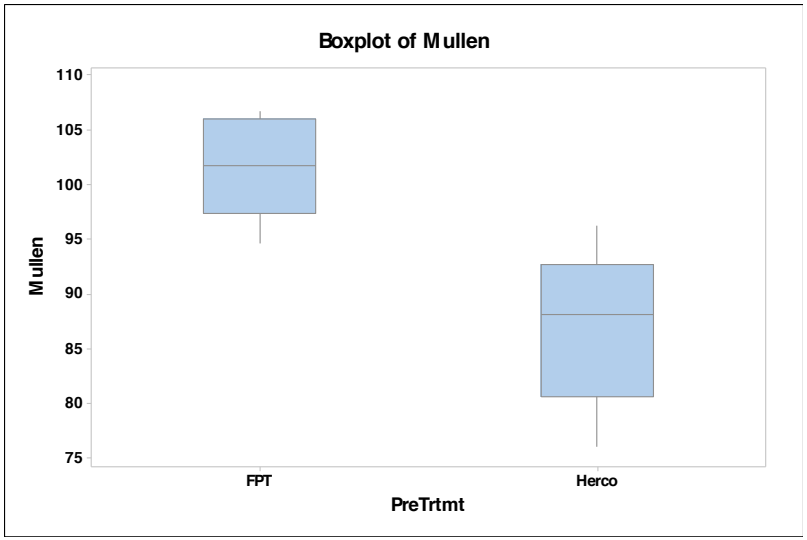
- Suppose we wish to test whether the average mullen for two fiber pre-treatments are statistically different.
- The two sample ***t*** test with independent samples would be the appropriate statistical test.
  - **WHY**?

Two Sample *t* Test

1. The hypothesis we wish to test is:
   - Ho: $\mu_{fpt} = \mu_{herco}$
   - Ha: $\mu_{fpt} \neq \mu_{herco}$
2. A representative sample of 10 averages for both **FPT** and **Herco** were collected.
3. Use Minitab to calculate the ***t*** statistic.

Two Sample *t* Test

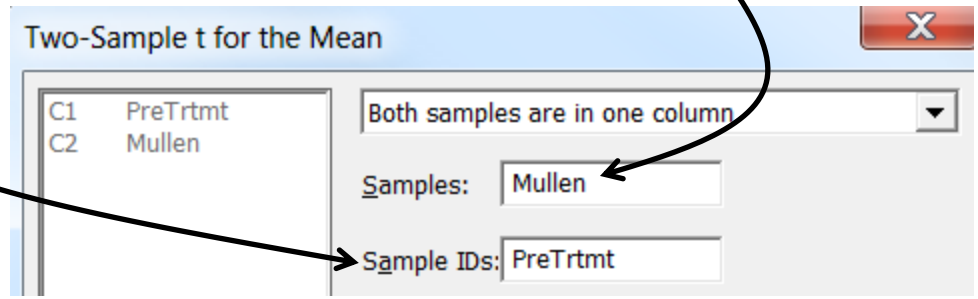
- Open the Excel worksheet ***Minitab Dataset 7.xls***
- Create a **Boxplot With Groups** of **Mullen** with **PreTrtmt** as the **Categorical Variable**.



Two Sample *t* Test

- Select **Stat** → **Basic Statistics**
  → **2- Sample t…**
- The default data format is **Both samples are in one column**, which is ok for us.

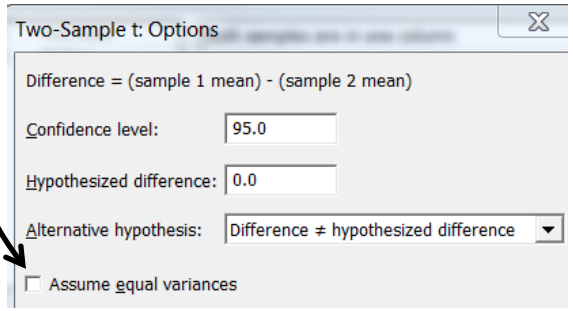
Two Sample *t* Test

- Select ***Mullen*** for **Samples** and ***PreTrtmt*** for **Sample IDs**.
- 

- Click the **Options…** button

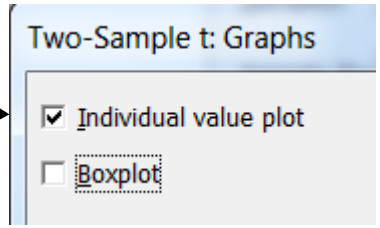
Two Sample *t* Test

- Keep the default **Options**.
- Keeping **Assume equal variance** unchecked is okay.
  - This allows for the variances of the two samples not to be equal. If the variances are equal it does not hurt, but if the variance are unequal it could lead to a decision error.



Two Sample *t* Test

- Click the **Graphs…** button.
- Select the **Individual value plot** graph by clicking in the check box.



- Click **OK** twice.
- Go to the **Session Folder**.

Two Sample *t* Test

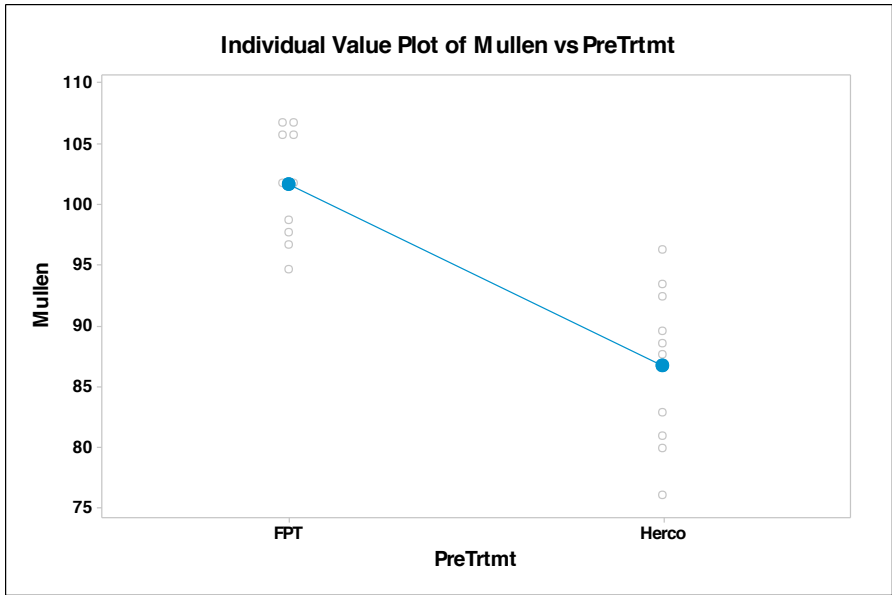
```
PreTrtmt    N     Mean  StDev  SE Mean
FPT        10   101.63   4.52      1.4
Herco      10    86.79   6.59      2.1

Difference = μ (FPT) - μ (Herco)
Estimate for difference:  14.8405
95% CI for difference:  (9.4572, 20.2238)
T-Test of difference = 0 (vs not =): T-Value = 5.88  P-Value = 0.000  DF = 15
```

<u>Interpretation</u>

3. The ***t*** statistic is **5.88**
4. The P-Value=0.000, which is less than **0.05**
5. Therefore, we **reject *Ho*** and conclude that there is a **significant difference** between FPT and Herco.

Two Sample *t* Test

The plot shows that the **FPT** pre-treatment has statistically significant higher average **Mullen** than the **Herco**.



Paired t-test

Comparing two different measurements on the same sample (paired).

Hypothesis:

$$H_0 : \mu_D = 0$$

$$H_a : \mu_D \neq 0 \text{ (2 sided test)}$$

or

$$H_a : \mu_D > 0 \text{ (1 sided test)}$$

or

$$H_a : \mu_D < 0 \text{ (1 sided test)}$$

μ_D is the average difference between the two means.

Paired t-test

Comparing two different measurements on the <u>same</u> sample.

<u>Test Statistic:</u>

$$t = \frac{\left(\bar{x}_D\right)}{\left(s_D / \sqrt{n}\right)}$$

$\bar{x}_D$ is the average of the differences

s_D is the standard deviation of the differences

Paired t-test

- The Lab manager was interested in testing if there was a difference in viscosity test results between **two** Technicians. The techs each measured the same samples.
- The Paired ***t*** test would be the appropriate statistical test.
  - **WHY**?

Paired t-test

1. The hypothesis we wish to test is:

   Ho: $\mu_{difference} = 0$

   Ha: $\mu_{difference} \neq 0$

2. A representative sample of 25 viscosity tests on the same sample, for both Technician #1and Tech #2, were collected.

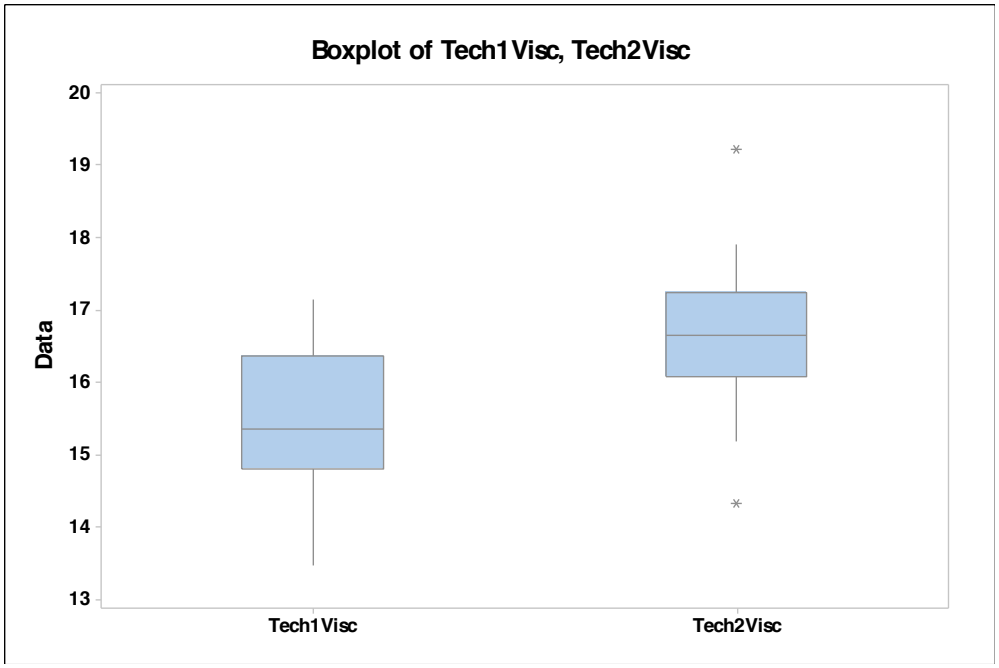
3. Use Minitab to calculate the ***t*** statistic.

Paired t-test

- Open the Excel worksheet ***Minitab Dataset 16.xls***
- The dataset has **Viscosity** test results for Tech #1 (**Tech1Visc**) and Tech #2 (**Tech2Visc**).

Paired t-test

- Create a **Boxplot of Multiple Y's Simple** with **Tech1Visc** and **Tech2Visc**.



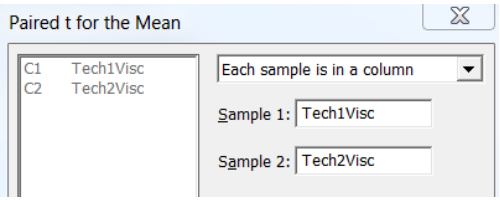
Paired t-test

- Select **Stat** → **Basic Statistics**
  → **Paired t…**
- The data format is **Each sample is in one column**.

Paired t-test

- Select **Tech1Visc** as **Sample 1**

  → select **Tech2Visc** as **Sample 2**

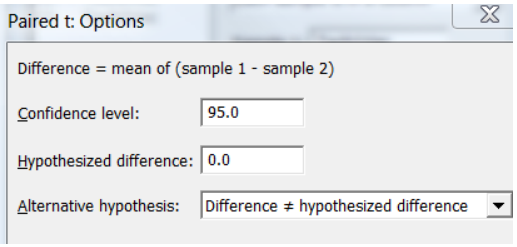


- Click on the **Graphs…** button.

  → select **Individual value plot of differences** → click **OK** once

Paired t-test

- Click on the **Options…** button.



- The default options are ok.
  - You can change ***Ha*** as previously discussed.
- Click **OK** twice.
- Go to the **Session Folder**.

Paired t-test

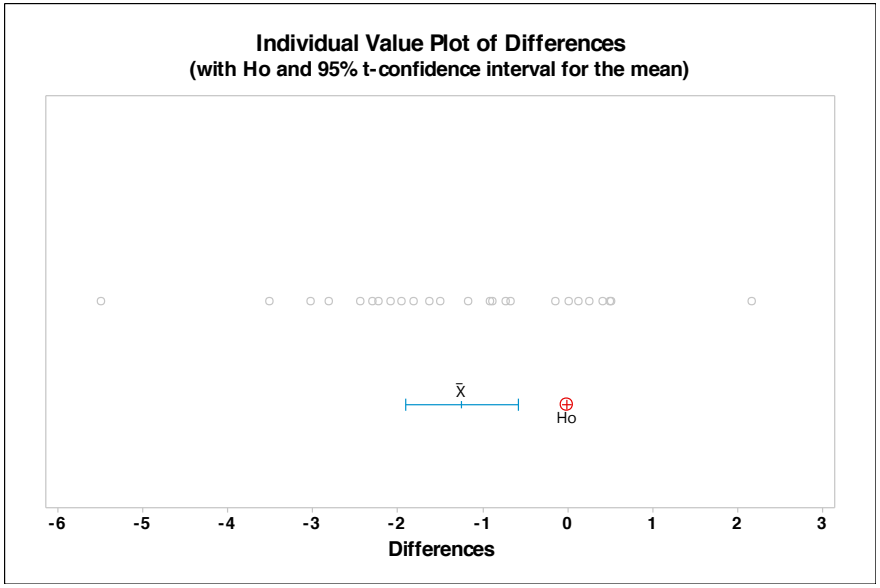
```
Paired T for Tech1Visc - Tech2Visc

              N    Mean  StDev  SE Mean
Tech1Visc    25  15.449  1.028    0.206
Tech2Visc    25  16.695  0.976    0.195
Difference   25  -1.246  1.616    0.323


95% CI for mean difference: (-1.913, -0.579)
T-Test of mean difference = 0 (vs ≠ 0): T-Value = -3.86  P-Value = 0.001
```

<u>Interpretation</u>

3. The ***t*** statistic is **-3.86**
4. The P-Value=0.001, which is less than **0.05**
5. Therefore, we **reject Ho** and conclude that there is a significant difference between Tech #1 and Tech #2.

Paired t-test

The plot shows that the **difference** is average viscosity between operator is **significant**. Tech #2 is significantly higher.



EXERCISE

- R&D was interested in testing if a new additive would enhance formation. No additive and .4 lbs/T were run in a trial.
- The dataset '***Exercise 3.xls***' contains the results.
- Perform the appropriate ***t*** test and report your findings.

Power of a Statistical Test

- We now return to discussing of ***Power***.
- Remember ***Power*** is defined as the probability of detecting a “true” effect when it exists.

Power of a *t* Test

- When calculating ***Power***, the "***effect size***" is an important concept to learn.
- The <u>effect size</u> (Δ) tells you how much of an "***effect***" you can detect (in standard deviation units, $\boldsymbol{\sigma}$) given the <u>minimum detectable difference</u> (***d***).

$$\Delta = \frac{d}{\sigma}$$

Power of a *t* Test

- ***d***, the minimum detectable difference.
  - the smallest difference between two means that has engineering or quality significance.
- Suppose a trial is run to assess two different additives effect on mullen.
  - Define the mean mullen for additive A and B as μ_A and μ_B, respectively.
- If a difference of 50 or more in mullen is important then,

$$d = \min(\mu_A - \mu_B) = 50.$$

Power of a *t* Test

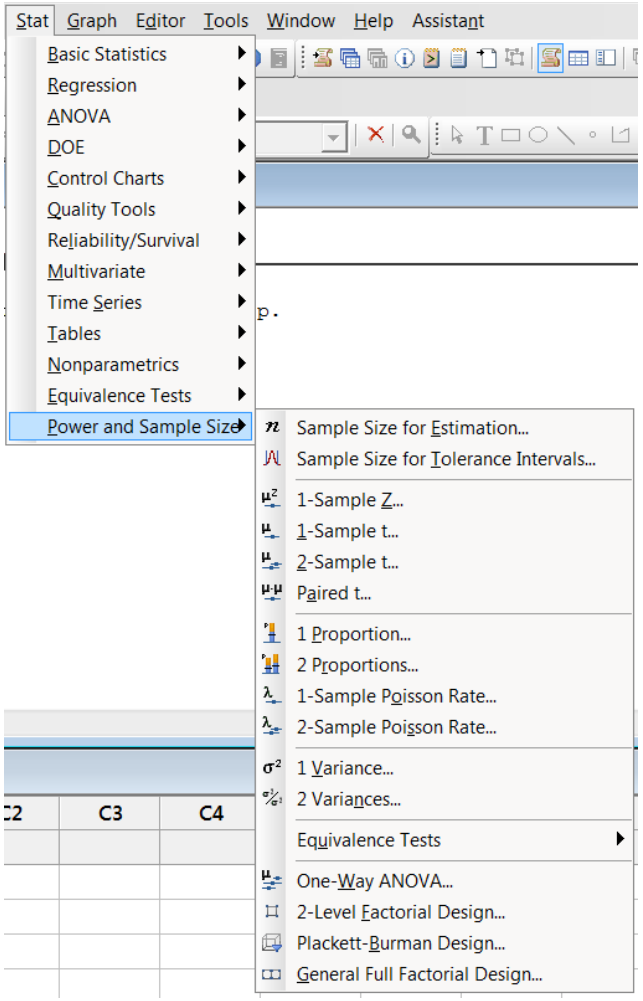
- The ***effect size*** (Δ) can be <u>indexed</u> as follows:
  - small effects: $\Delta = .2 \rightarrow .6$
  - medium effects: $\Delta = .7 \rightarrow 1$
  - large effects: $\Delta = 1.1 \rightarrow 1.5$
    - Effect sizes are in terms of standard deviation, so a <u>large effect</u> is one that is detectable around <u>one</u> standard deviation or more.
- The smaller the effect size, the larger the **sample size**.

Power of a *t* Test

- ***How Does Power Increase?***
- When sample size (***n***) increases, **power** increases.
- As variation (σ) decreases, **power** increases.
- As α risk gets larger, **power** increases.
- When ***d*** increases, **power** increases.

Power of a *t* Test

- Minitab can calculate power for many statistical tests.
- In the following example, we are interested in the power of a two sample *t* test with independent samples.



Power of a *t* Test

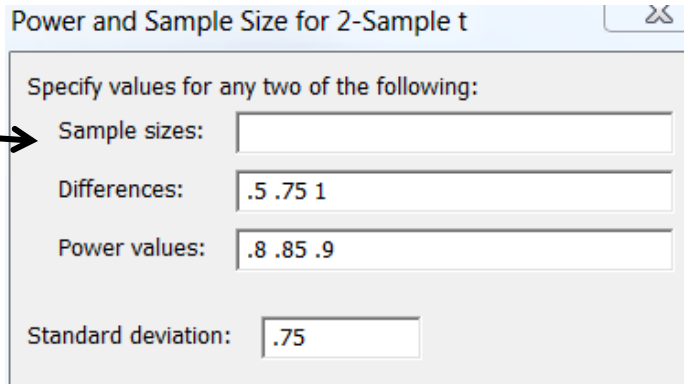
- Select **Stat** → **Power and Sample Size** → **2 Sample t…**
  - We want the sample sizes for desired differences, with several power values and an estimate of the standard deviation.
- For example, you want to detect a sheet caliper difference of **0.5** mil to **1** mil. You believe the standard deviation is **~0.75** mil.

Power of a *t* Test

- Set the **Standard Deviation** box to .75
- Type in .5, .75 and 1 in the **Differences** entry, using a space as the delimiter.
- ***Good*** statistical ***power*** is in the range of 80% to 90%.
- Set the **Power values** to .8, .85, .9 using a space as the delimiter.

Power of a *t* Test

Your **Power and Sample Size** window should look like the following.



Leave the defaults for the **Options…** and **Graphs…** as is.

Click **OK.**
Got to the **Session folder**.

Power of a *t* Test

The table shows the **sample sizes** and power occur for each difference.

2-Sample t Test

Testing mean 1 = mean 2 (versus ≠)
Calculating power for mean 1 = mean 2 + difference
α = 0.05 Assumed standard deviation = 0.75

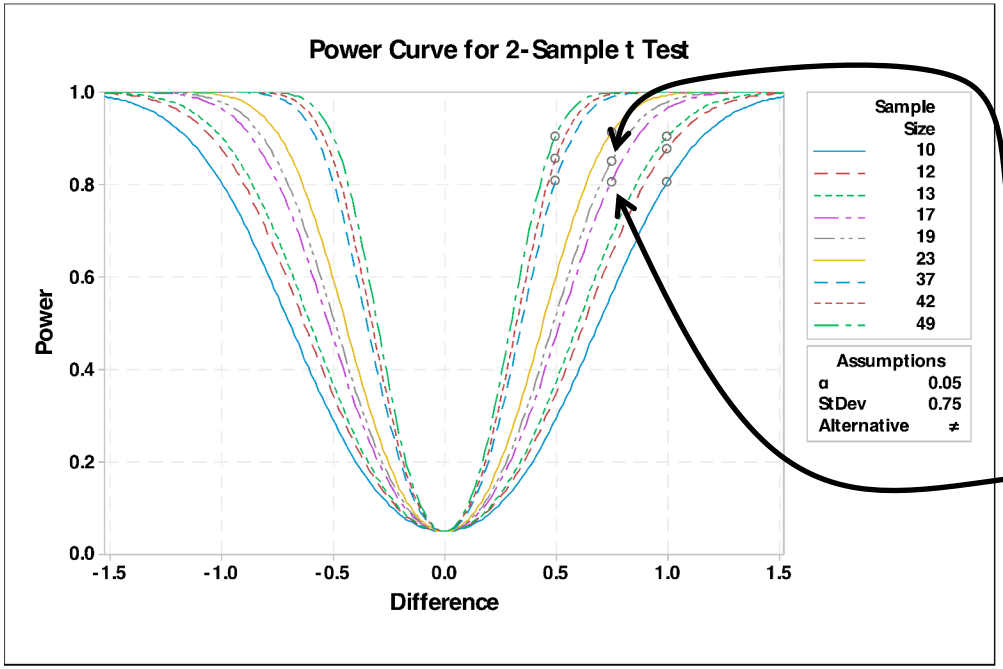
Difference	Sample Size	Target Power	Actual Power
0.50	37	0.80	0.807587
0.50	42	0.85	0.855227
0.50	49	0.90	0.904339
0.75	17	0.80	0.807037
0.75	19	0.85	0.850612
0.75	23	0.90	0.912498
1.00	10	0.80	0.804913
1.00	12	0.85	0.877082
1.00	13	0.90	0.903290

The sample size is for each group.

Sample sizes are for **<u>EACH</u>** group, so the total samples required for a difference=1.0 would be 10+10, 12+12, 13+13.

Power of a *t* Test

<u>Power curves</u>: Gives the tabular output in a plot.



If you want a power > .8 to detect a difference < 1 with a sample size < 20, choose either ***n***=17 or ***n***=19.

EXERCISE

- Suppose the Fiberline process owner would like to understand the effect of an increase in temperature on Kappa.
- There is the normal (control) temp and one temp increase, so this is a 2-sample t test.
- Find the sample size for a power > 0.80, with a detectable difference of 0.5, 1.0 and 1.5, and with a standard dev = 1.1

Replication of Samples

- We will now discuss *replication* vs. *repeated measures*, an important consideration.
  - The discussion centers on whether there was independent *replication* or dependent *repeated measures*.
- The specific situation depends on how the experiment was **executed,** what was **measured** and how.

Replication of Samples

- A *replication* is an independent observation of the **treatment** combinations applied to the **experimental unit**.
  - If we are examining the effect of Additive A, two replications of Additive A would mean collecting sheets from two separate production runs using Additive A.
    - Additive is the treatment and the sheet is the **experimental unit**.

Replication of Samples

- Repeated measures, or sub-sampling, is *splitting or dividing up* the same sample to generate two or more observations.
  - Example: making one sheet with Additive A, cutting the sheet into 10 pieces and making 10 measurements does NOT constitute 10 replications, but 10 sub-samples.
  - 10 ***replications*** requires 10 ***independent*** sheets to be made and sampled.

Replication of Samples

- A good designed experiment (DOE) requires replication.
- An estimate of the *experimental error* cannot be obtained unless the treatments are replicated.
  - This means you must make true independent replications to get a valid estimate of the error variance.
- A repeated measure only estimates how much that particular experimental unit varies, but is <u>not</u> an estimate of the experimental error.