STATISTICS
TEXAS A&M UNIVERSITY

Handout 01:

Review of Basic Statistical Concepts

Samples and Populations

- A population is the set of all possible outcomes of an experiment or process.

- A sample is a subset of the population which is the representative of the population.
  - Researchers deal with samples rather than populations.
  - With designed experiments, we create two or more populations. Each test or run is the sample.

Samples and Populations

A beverage company wanted to see if people in the United States liked their new logo. Which choice best represents a population?

A. selection of logo artists.

B. Every person in the United States.

C. A selection of shoppers from different states.

D. 3,800 children age 5 - 15

A musician wanted to see what people who bought his last album thought about the songs. Which choice best represents a sample?

A. Every person who bought the album.

B. selection of people who didn't want to buy the album.

C. 250 girls who bought the album.

D. A selection of 3,294 people who bought the album.

A gaming website wanted to find out which console its visitors owned. Which choice best represents a population?

A. Visitors to the 3DS section.

B. All of the website visitors.

C. Visitors to the PS4 section.

D. Visitors who are on the website for more than 5 minutes.

Samples and Populations

Before a nation wide election, a polling place was trying to see who would win. Which choice best represents a sample?

A. A selection of voters over age 50.
B. A selection of male voters.
C. A selection of voters of different ages.
D. All voters.

A toy store owner tracking how much kids spend each month on toys. Which choice best represents a population?

A. All of the kids who buy toys.
B. 227 rich kids.
C. 228 boys age 7 - 15
D. 235 kids from age 10 to 15.

A mayor wanted to see if the people in his town thought he was doing a good job. Which choice best represents a sample?

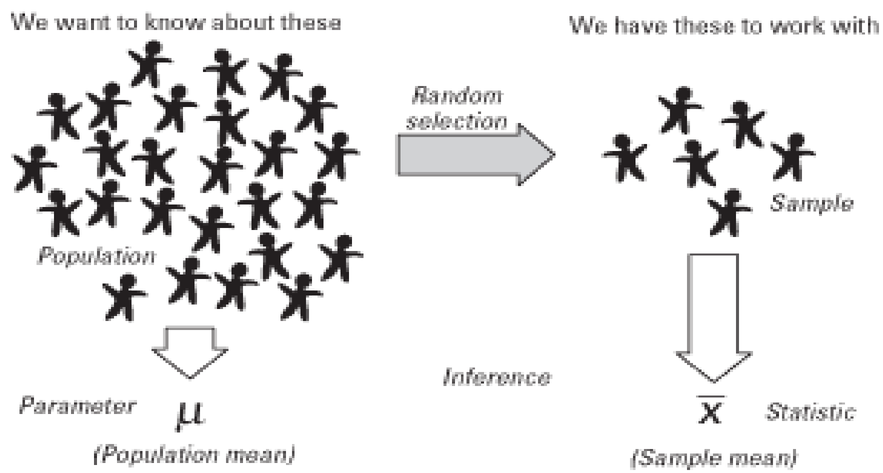
A. 1,000 unemployed voters.
B. The mayor's family.
C. The residents of the town.
D. 242 voters.

Any others from your field of study?

Statistics and Parameters

Statistics are measurements or attributes about a sample.

Parameters are measurements or attributes about a population. Parameters are usually unknown and are estimated by statistics.



With a designed experiment, we determine if changing a variable manifests itself in the parameters that we estimate.

Statistics and Parameters

Identify the following as statistics or parameters.

$\bar{y}$

μ

σ

β

s

$\hat{\pi}$

π

τ

λ

STATISTICS
TEXAS A&M UNIVERSITY

Variables

A **variable** is any characteristics, number, or quantity that can be measured or counted.

Discrete

Continuous

Categorical

Ordinal

Binary

Dependent

Independent

Examples for each variable type

Hypotheses

- "Hypotheses are single tentative guesses, good hunches – assumed for use in devising theory or planning experiments intended to be given a direct experimental test when possible". (Eric Rogers, 1966)
- "A hypothesis a conjectural statement of the relation between two or more variables". (Kerlinger, 1956)
- "Hypothesis a formal statement that presents the expected relationship between an independent and dependent variable."(Creswell, 1994)
- "An hypothesis is a statement or explanation that is suggested by knowledge or observation but has not, yet, been proved or disproved."(MacleodClark J and Hockey L, 1981)

Types of Hypotheses

- Null hypothesis (H_0)

The null hypothesis represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved.

Has serious outcome if incorrect decision is made!

- Alternative hypothesis (H_a or H_1)

The alternative hypothesis is a statement of what a hypothesis test is set up to establish.

Opposite of Null Hypothesis.

Only reached if H_0 is rejected.

Frequently "alternative" is actual desired conclusion of the researcher!

Types of Hypotheses

- In a clinical trial of a new drug

  H_0: there is no difference between the two drugs on average.

  H_a: the two drugs have different effects, on average.

OR

  H_a: the new drug is better than the current drug, on average

- A nutritionist, working in a zoo, will develop a menu plan for the group of monkeys. In order to get all the vitamins they need, the monkeys have to be given fresh leaves as part of their diet. Choices include leaves of the following species: (a) A (b) B (c) C (d) D and (e) E. If you know that in the wild the monkeys eat mainly B leaves, but you suspect that this could be because they are safe whilst feeding in B trees, whereas eating any of the other species would make them vulnerable to predation. You design an experiment to find out which type of leaf the monkeys actually like best: You offer the monkeys all five types of leaves in equal quantities, and observe what they eat.

correctly specified H_a: When offered all five types of leaves, the monkeys will preferentially feed on B leaves.

incorrectly specified H_a: When offered all five types of leaves, the monkeys will preferentially eat the type they like best.

Type of Tests

- Right tailed test

  H_0: parameter $\leq$ constant value

  H_a: parameter > constant value

- Left tailed test

  H_0: parameter $\geq$ constant value

  H_a: parameter > constant value

- Two tailed test

  H_0: parameter = constant value

  H_a: parameter $\neq$ constant value

STATISTICS
TEXAS A&M UNIVERSITY

Hypothesis Testing and Significance

A hypothesis test is used to assess the evidence provided by the data in favor of some claim about the population.

There are four steps to conduct a hypothesis test.

1. State the null (H_0) and alternative (H_a) hypotheses.
2. State alpha (significance level).
3. Collect data, compute sample statistics, and compute the *p*-value under H_0.
4. Make a decision:

   If *p*-value$\leq\alpha$, there is sufficient evidence to reject H_0.

   If *p*-value$>\alpha$, there is not sufficient evidence to reject H_0.

You can never prove the null hypothesis. You can only state that you have evidence to reject that hypothesis.

Hypothesis Testing and Significance

How about the confidence interval (C.I) instead of pvalue in decision making ?

3. Collect data, compute sample statistics, and compute C.I for the parameter of interest.

Advantage ?

4. Make a decision:

If the hypothesized value under H_0

falls between the limits of the C.I, fail to reject H_0.

does not fall between the limits of the C.I., reject H_0.

How about if you have a one tailed test but you have calculated the two tailed C.I.?

Types of Errors and Power

You perform a hypothesis test and make a decision. Was the decision correct? Any kind of error?

	REALITY	
DECISION	**H_0 True**	**H_0 False**
Fail to reject H_0	correct	Type II error
Reject H_0	Type I error	correct

The probability of Type I error is denoted by α.

The probability of Type II error is denoted by β.

Power of the test $=1-\beta$.

Types of Errors and Power

In the court, we assume innocence until proven guilty, so in a court case innocence is the Null hypothesis.
Type 1 error is the error of convicting an innocent person.
Type 2 error is the error of letting a guilty person go free.

<u>Identify the following as type I error and type II error</u>

1. Shepherd and wolf

H_0:no wolf present

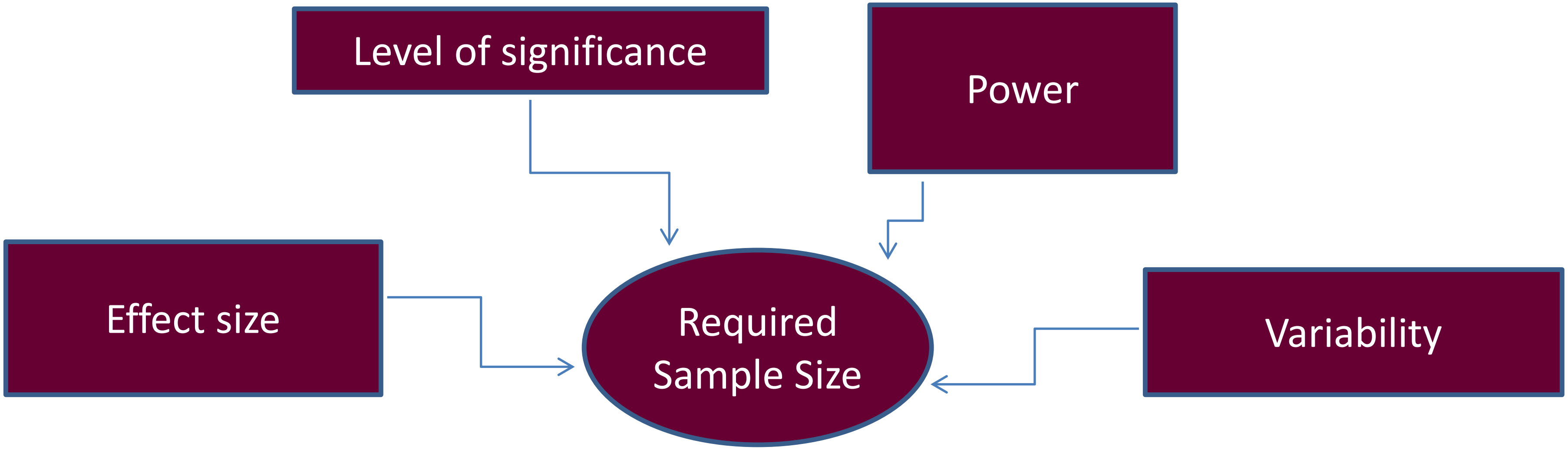
A. the shepherd wrongly indicated there was no wolf present and continued to play Candy Crush on his iPhone.
B. the shepherd wrongly indicated there was a wolf present by calling "Wolf! Wolf!"

2. Mileage versus fuel additive

H_0: fuel additive does not increase gas mileage

A. Falsely declared that the additive makes a difference
B. Falsely declared that the additive does not make a difference

How Many Observations?



<u>Given that you hold all other terms constant, the required sample size increases if:</u>

1. You want to detect a smaller effect size (the difference that is practically important to you).

2. You decrease the level of significance level (alpha) .

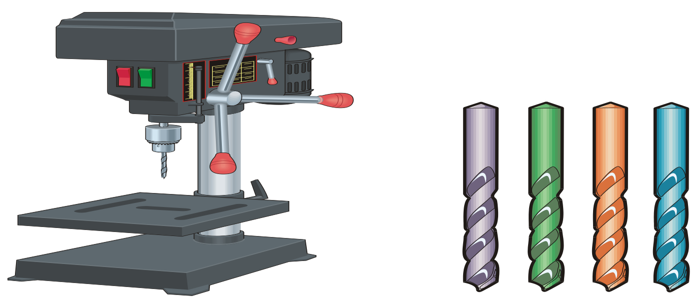
3. You require a higher power.

4. The variability in the data increases.

Hardness Measurement

The hardness of a metal sheet is determined by using a drill press to place an indentation on the metal sheet.

A measurement instrument is then used to measure this indentation, and a measurement of hardness is obtained.



A manufacturer of metal sheets specifies that the hardness of its metal sheets is 9.5, based on its measurement system. To test this claim, a customer requests that a significance test be conducted.

STATISTICS
TEXAS A&M UNIVERSITY

Demonstration Information

- Data Model and Assumptions

  The model represents what you think is true about your population. You assume that there is a single population with values that are normally distributed with mean, μ and standard deviation, σ.

- Preliminary sample information indicates that the hardness difference you need to detect is about 0.6 and the standard deviation is about 0.5.

- The company would like the power of the test to be at least 0.95 and to test at a level of significance of 0.05.

- This information will be used to determine the required sample size.

Determining Power and Sample Size

JMP:

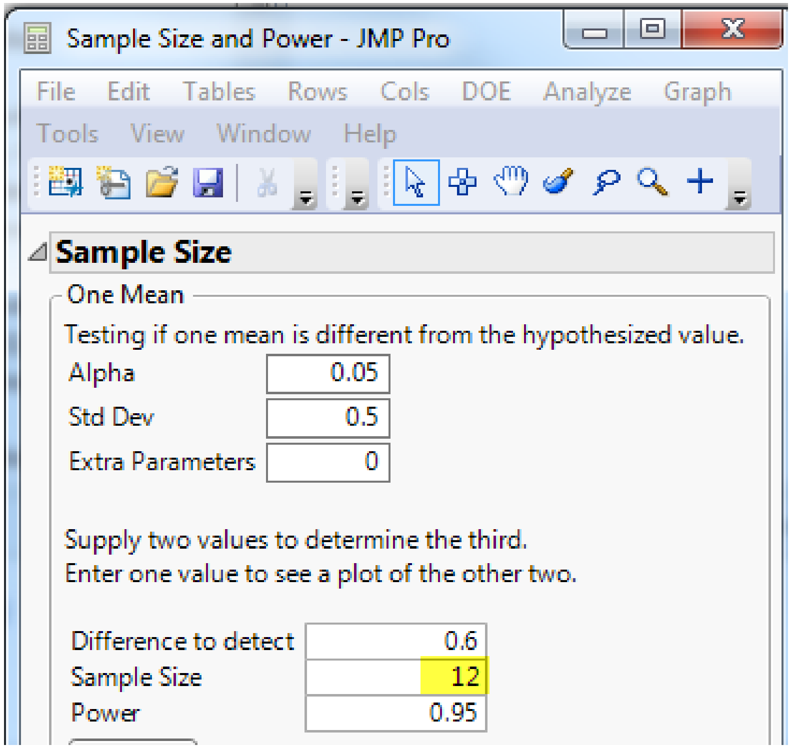
Select DOE ⇒ Sample Size and Power

Select One Sample Mean

Type 0.5 for the error standard deviation

Type 0.6 for the difference to detect

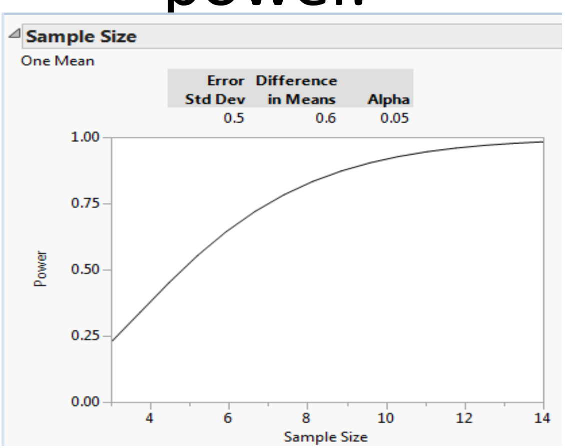
Type 0.95 for the Power

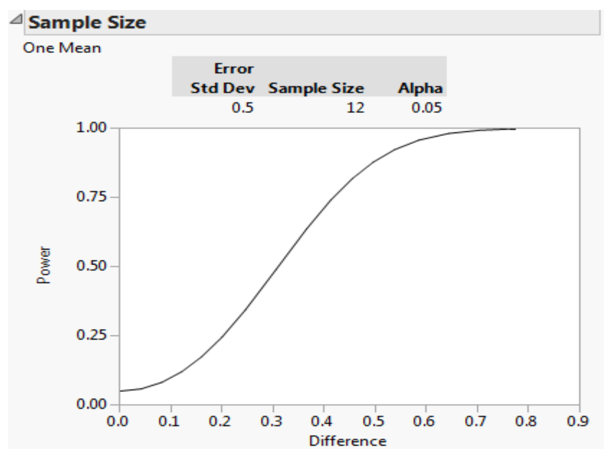


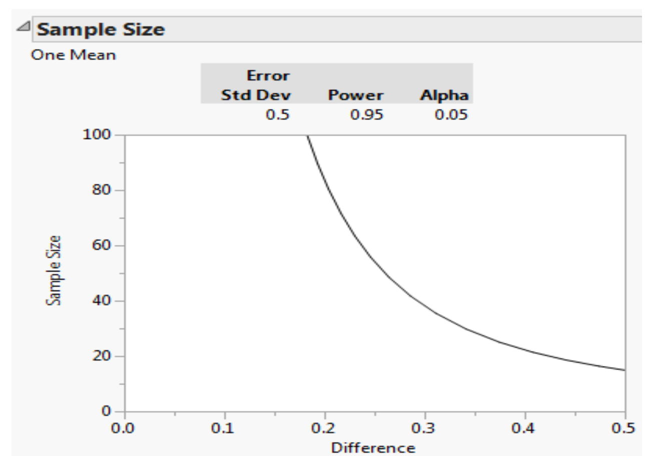
Determining Power and Sample Size

If we had two unspecified fields in the platform, we could see the plot showing the relationship between the following two values

1. Power as a function of sample size, given the specific effect size.
2. Power as a function of the effect size, given sample size
3. Effect size as a function of sample size, for a given power.







Determining Power and Sample Size

```
proc power;
  onesamplemeans test=t
  nullmean = 9.5
  mean  = 10.1
  stddev = 0.5
  power = .95
  ntotal = . ;
run;
```

The POWER Procedure
One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Null Mean	9.5
Mean	10.1
Standard Deviation	0.5
Nominal Power	0.95
Number of Sides	2
Alpha	0.05

Computed N Total	
Actual Power	N Total
0.965	12

```
proc power;
  onesamplemeans test=t
  nullmean = 9.5
  mean  = 8.9
  stddev = 0.5
  power = .95
  ntotal = . ;
run;
```

The POWER Procedure
One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Null Mean	9.5
Mean	8.9
Standard Deviation	0.5
Nominal Power	0.95
Number of Sides	2
Alpha	0.05

Computed N Total	
Actual Power	N Total
0.965	12

Determining Power and Sample Size

```
proc power;
  onesamplemeans test=t
  nullmean = 9.5
  mean  = 10.1
  stddev = 0.5
  power = .
  ntotal = 12 ;
run;

proc power;
  onesamplemeans test=t
  nullmean = 9.5
  mean  = 8.9
  stddev = 0.5
  power = .
  ntotal = 12 ;
run;
```

The SAS System

The POWER Procedure
One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Null Mean	9.5
Mean	8.9
Standard Deviation	0.5
Total Sample Size	12
Number of Sides	2
Alpha	0.05

Computed Power
Power
0.965

The SAS System

The POWER Procedure
One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Null Mean	9.5
Mean	10.1
Standard Deviation	0.5
Total Sample Size	12
Number of Sides	2
Alpha	0.05

Computed Power
Power
0.965

Analyzing Data from a Completely Randomized Design:

Download the data from ecampus-Lecture Datasets. THEN,

In JMP:

Select **File** $\Rightarrow$ **Open** $\Rightarrow$ **hardness.JMP** $\Rightarrow$ **Open**

Select **Analyze** $\Rightarrow$ **Distribution**

Select **Hardness** $\Rightarrow$ **Y,columns**

Select **OK**

Click the red triangle next to **Hardness** and then select **Normal Quantile Plot**

Click the red triangle next to **Hardness** and then select **Continuous Fit** $\Rightarrow$ **Normal**

Click the red triangle next to **Fitted Normal** and then select **Goodness of Fit**

Click the red triangle next to **Hardness** and then select **Test Mean**

Type **9.5** for the hypothesized mean

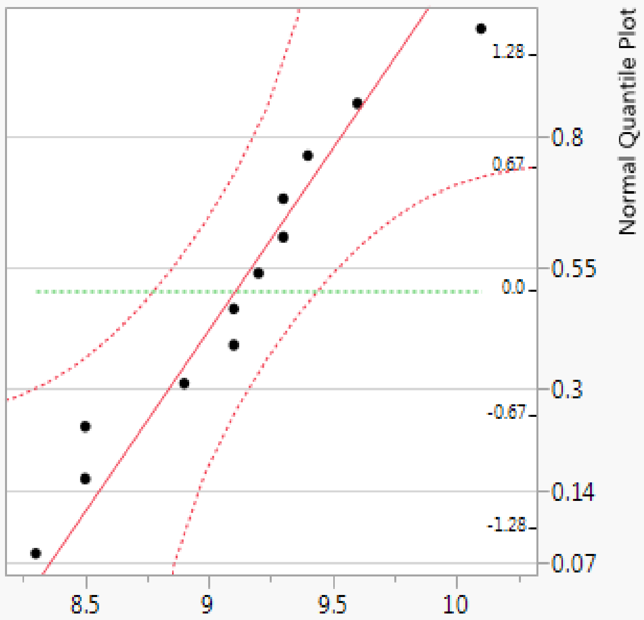
Select **OK**

Answer: Is the data normally distributed?

Is there sufficient evidence to reject H0?

Hardness Data

Hardness



Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	9.1083333	8.7861082	9.4305585
Dispersion	σ	0.5071459	0.3592597	0.8610719

-2log(Likelihood) = 16.7595679739322

Goodness-of-Fit Test

Shapiro-Wilk W Test

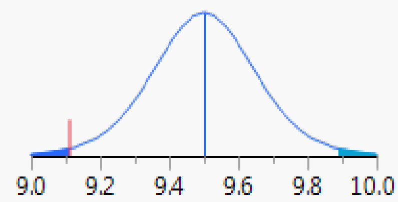
W	Prob<W
0.959229	0.7728

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Test Mean

Hypothesized Value	9.5
Actual Estimate	9.10833
DF	11
Std Dev	0.50715

	t Test
Test Statistic	-2.6753
Prob > |t|	0.0216*
Prob > t	0.9892
Prob < t	0.0108*



Quantiles

100.0%	maximum	10.1
99.5%		10.1
97.5%		10.1
90.0%		9.95
75.0%	quartile	9.375
50.0%	median	9.15
25.0%	quartile	8.6
10.0%		8.36
2.5%		8.3
0.5%		8.3
0.0%	minimum	8.3

Summary Statistics

Mean	9.1083333
Std Dev	0.5071459
Std Err Mean	0.1464004
Upper 95% Mean	9.4305585
Lower 95% Mean	8.7861082
N	12
Variance	0.257197
Interquartile Range	0.775

$$H_0: \mu = 9.5$$

$$H_a: \mu \neq 9.5$$

Test Mean

Hypothesized Value	9.5
Actual Estimate	9.10833
DF	11
Std Dev	0.50715
Sigma given	0.5

	z Test
Test Statistic	-2.7135
Prob > |z|	0.0067*
Prob > z	0.9967
Prob < z	0.0033*

Hardness Data

```
PROC UNIVARIATE DATA=Hardness mu0=9.5 normal;
  VAR hardness;
        QQplot hardness /Normal(mu=est sigma=est);
RUN;
```

The UNIVARIATE Procedure
Variable: hardness

Moments			
N	12	Sum Weights	12
Mean	9.10833333	Sum Observations	109.3
Std Deviation	0.50714591	Variance	0.25719697
Skewness	0.12232829	Kurtosis	0.10731991
Uncorrected SS	998.37	Corrected SS	2.82916667
Coeff Variation	5.56793309	Std Error Mean	0.14640041

Basic Statistical Measures			
Location		Variability	
Mean	9.108333	Std Deviation	0.50715
Median	9.150000	Variance	0.25720
Mode	8.500000	Range	1.80000
		Interquartile Range	0.65000

Note: The mode displayed is the smallest of 3 modes with a count of 2.

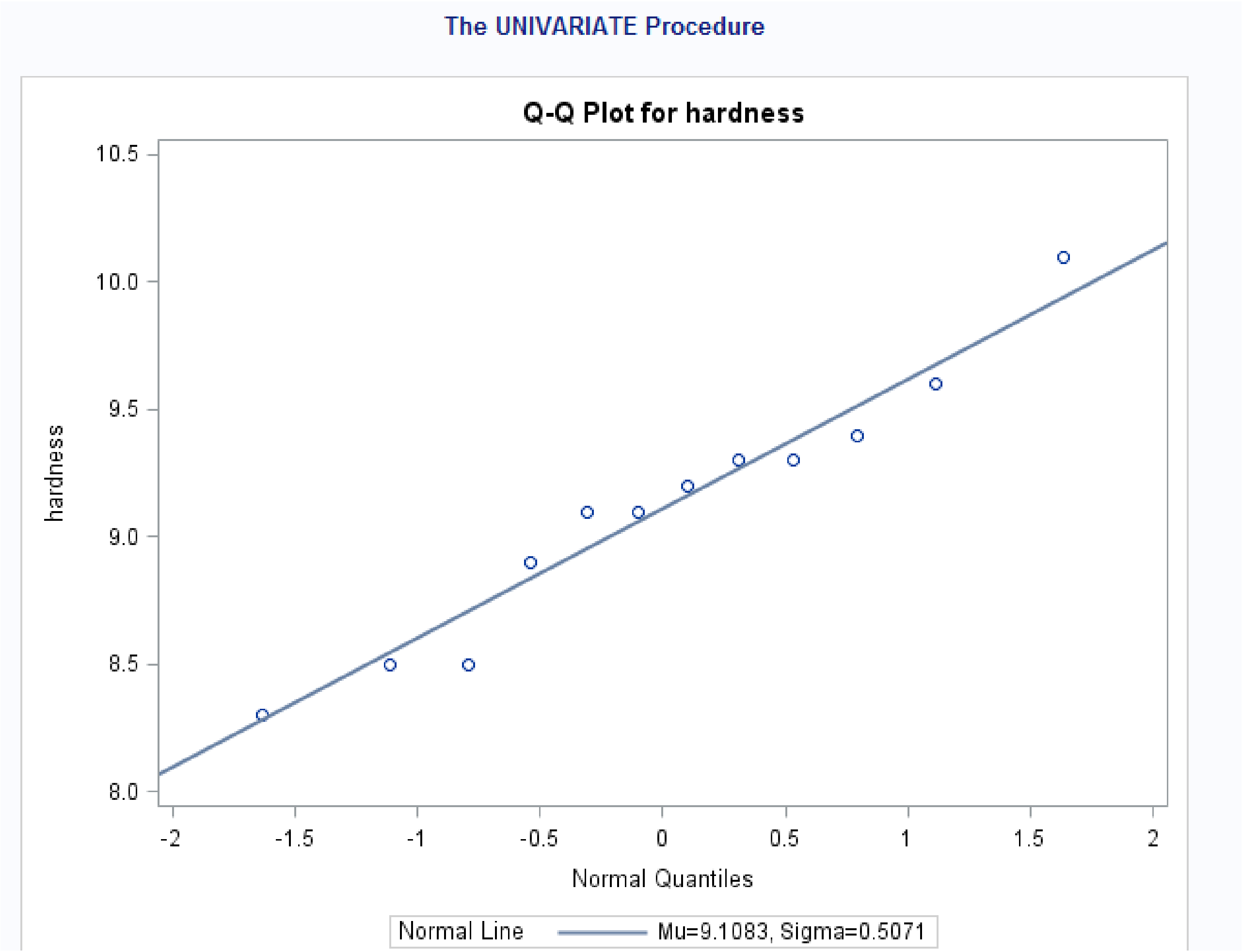
Tests for Location: Mu0=9.5				
Test	Statistic		p Value	
Student's t	t	-2.67531	Pr > |t|	0.0216
Sign	M	-4	Pr >= |M|	0.0386
Signed Rank	S	-29	Pr >= |S|	0.0215

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.959229	Pr < W	0.7728
Kolmogorov-Smirnov	D	0.160112	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.046178	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.288644	Pr > A-Sq	>0.2500

Quantiles (Definition 5)	
Level	Quantile
100% Max	10.10
99%	10.10
95%	10.10
90%	9.60
75% Q3	9.35
50% Median	9.15
25% Q1	8.70
10%	8.50
5%	8.30
1%	8.30
0% Min	8.30

Hardness Data

```
PROC UNIVARIATE DATA=Hardness mu0=9.5 normal;
   VAR hardness;
            QQplot hardness /Normal(mu=est sigma=est);
RUN;
```



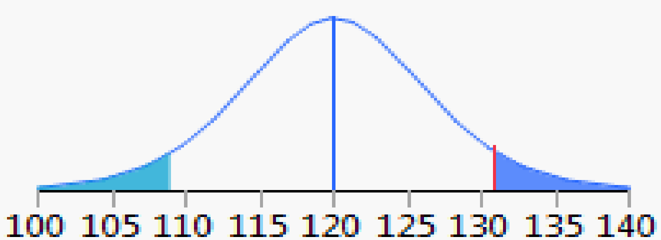
Days Data

A company wants to determine whether there is evidence that the mean shelf life of its carbonated beverage exceeds 120 days. Ten bottles are randomly selected and tested then the results are stored in Days.JMP.

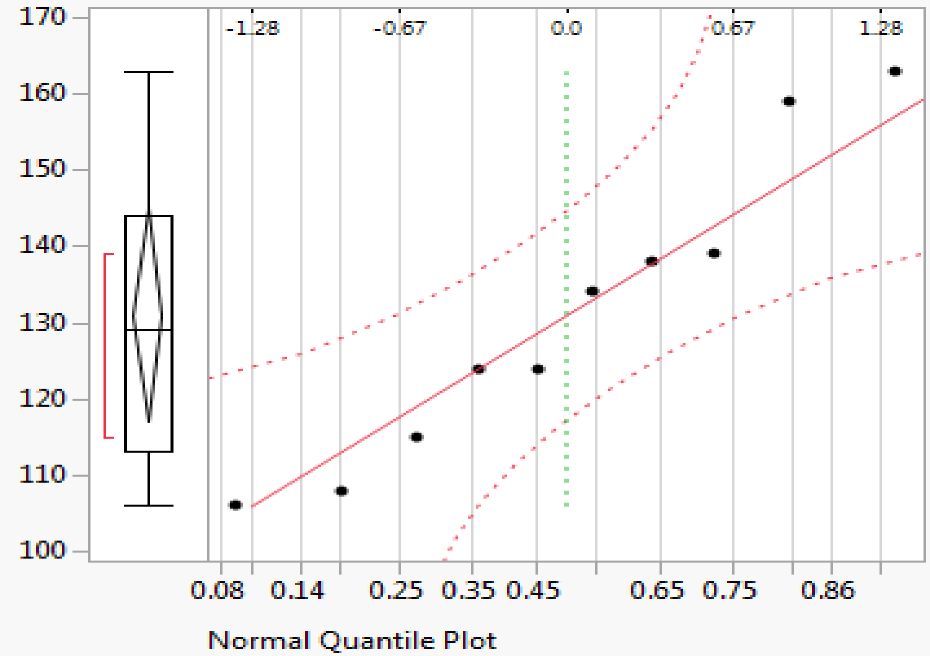
Test Mean

Hypothesized Value	120
Actual Estimate	131
DF	9
Std Dev	19.5448

	t Test
Test Statistic	1.7798
Prob > |t|	0.1088
Prob > t	0.0544
Prob < t	0.9456



Days



Normal(131,19.5448)

Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	131	117.01848	144.98152
Dispersion	σ	19.54482	13.443616	35.681223

-2log(Likelihood) = 86.8329767501592

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.936541	0.5152

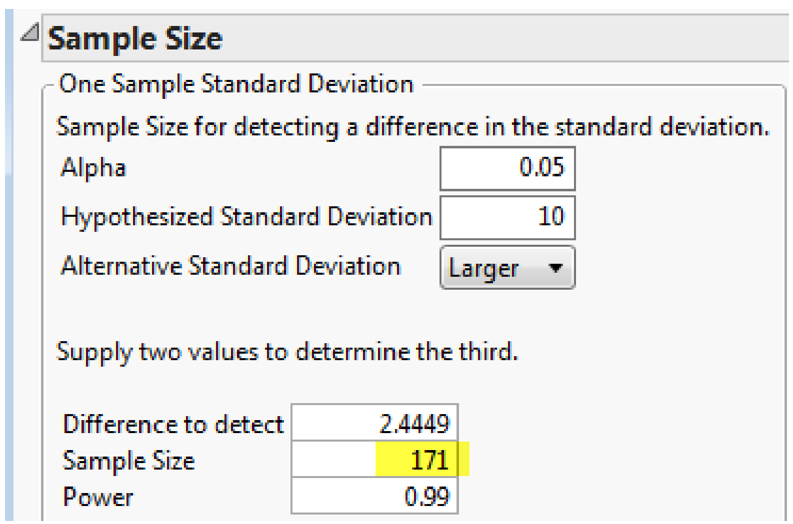
Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Engineering Example

The variance of resistivity measurements on a lot of silicon wafers is claimed to be 100 ohm-cm.

The buyer is unwilling to accept a shipment if the variance is greater than 155 ohm-cm for a particular lot (this measure is 55 ohm-com above the baseline of 100 ohm-cm) so the difference to detect with the standard deviation is $\sqrt{155} - \sqrt{100}$=2.4499.

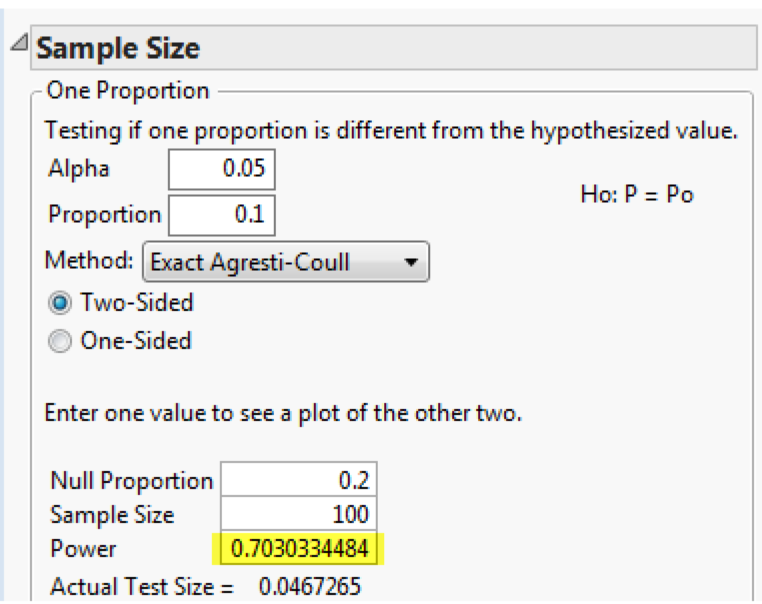
To detect an increase in the standard deviation of 2.4449 for a standard deviation of 10 with an alpha of 0.05 and power of 0.99:



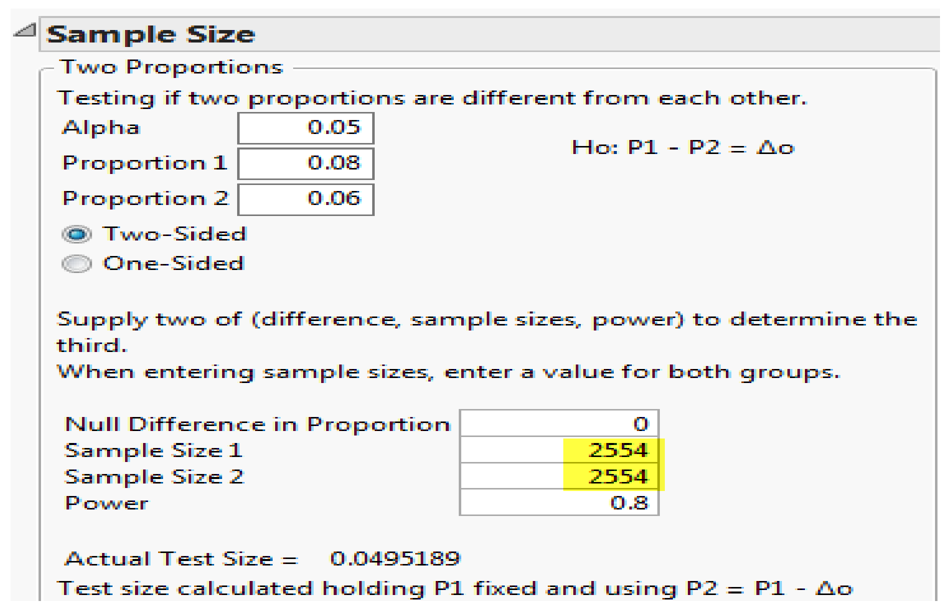
STATISTICS TEXAS A&M UNIVERSITY

Engineering Example

Suppose that an assembly line has a historical proportion of defects equal to 0.1 and you want to know the power to detect the proportion is different from 0.2 given an alpha level of 0.05 and a sample size of 100.



Suppose you are responsible for two silicon wafer assembly lines. Based on the knowledge from many runs, one of the assembly lines has a defect rate of 8%; the other line has a defect rate of 6%. You want to know the sample size necessary to have 80% power to reject the null hypothesis of equal proportions of defects for each line.



Introduction to Design of Experiments

What is Design:

Experimental design is the planning phase of data collection. It defines the structure of the experiment to ensure the efficient use of collected data.

Why do you experiment:

- To discover the sources of variation in a measured response
- To collect evidence to support or rebut a theory
- To determine a consistent result of a system, product design, or process
- To find conditions that yield a maximum or minimum response in a specified range
- To compare values of the response at different settings of the controllable variables
- To build a predictive model

Introduction to Design of Experiments

What is Experimental Error:

Experimental error is the variation among identically and independently treated experimental units. The various origins of experimental error include:

- The natural variation of experimental units
- The variability in measurement of the response
- The inability to reproduce the treatment conditions exactly from one unit to another
- The interactions of treatments and experimental units
- Any other extraneous factors that influence the measured the characteristics

Ad-Hoc Analysis versus Experimental Design

Ad-Hoc Regression	Experimental Design
Process outcomes are measured	Process outcomes are measured
Objective of the process is to make all units identical	Objective of the experimental design is to discern which input variables influence a response
The ranges of input values are limited	The ranges of the input variables are purposefully manipulated
The values of the input variables might be correlated	There is zero, or near zero, correlation between input variables

Design and Analysis

Design and analysis go hand-in-hand.
You analyze data to answer your questions.
You design an experiment so that the analysis is simple and correct.

The process for experimenting:

1. Define the purpose of the experiment.
2. Document the specific questions to be answered.
3. Define the population of interest.
4. Determine the need for sampling.
5. Define the data collection protocol :
   Describe the process.
   Identify sources of variability in the process.
   Determine the "best" design for the experiment.
   Delineate the experimental procedure.
6. Collect the data.
7. Analyze the data.

Basic Terms

Factor is an independent or predictor variable that is a possible source of variation.

Factor Level is a particular value of a factor. In other words, it is the specific types or amounts of a factor used in the experiment.

Treatment or design point is a combination of factor levels used in the experiment. In single factor studies, a treatment is the same as factor level.

Response is the variable that measures the outcome of interest.

Effect is a change in the response due to a change in the factor level.

Experimental unit is what receives the treatment.

Run is a single observation for a treatment. In other words, a run is the combination of factor levels and the value of the response variable.

Replication occurs when you assign the same treatment again to a new experimental unit.

Match the term on the left with the appropriate experimental component on the right.

Suppose an experiment is to be conducted to test the effect of a particular drug on the blood pressure of women. The three dosages are 3, 5, and 7 mg.

1. Factor
2. Factor Levels
3. Response
4. Run
5. Experimental Unit

A. 3 mg, 5 mg, and 7 mg
B. Blood pressure reading
C. Drug
D. Dosage and corresponding blood pressure reading
E. A woman

Three Basic Principles of Designs

- *Randomization* of runs prevents systematic biases from being introduced into the experiment. Randomization refers not only to performing the runs in a random order, but also to resetting the conditions after each run.

- *Blocking* is a design technique used to reduce or control variability from nuisance factors.

- *Replication* enables the experimenter to obtain an estimate of experimental error.