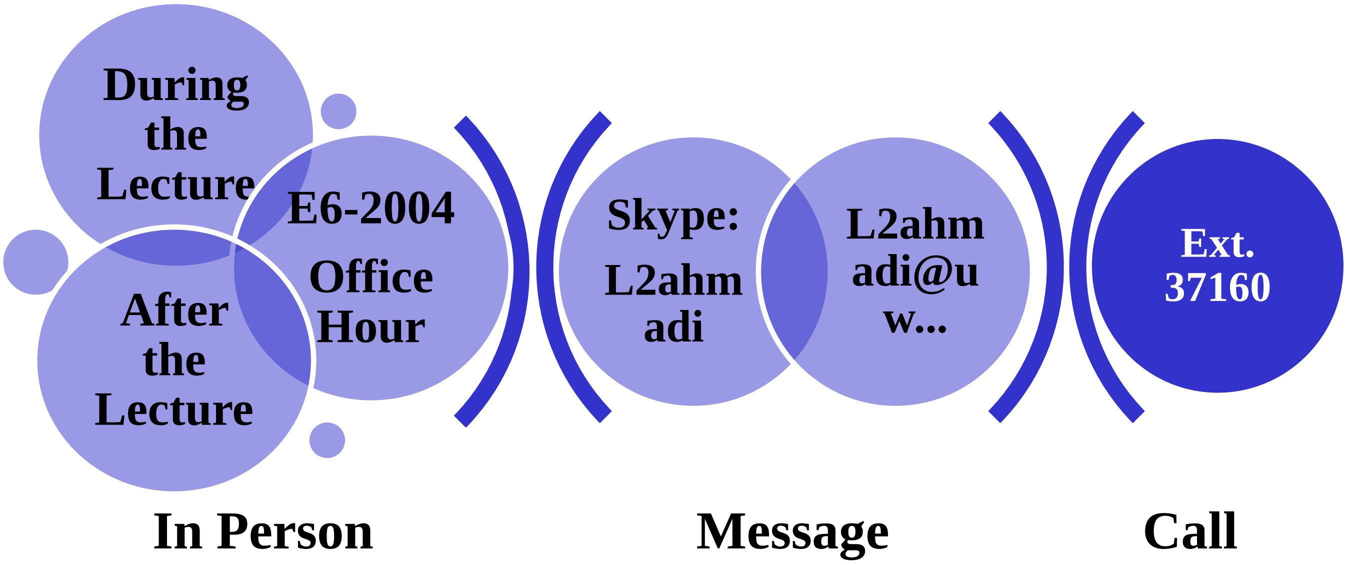
CHAPTER 1
STATISTICAL BACKGROUND

Instructor: Lena Ahmadi



During the Lecture

E6-2004 Office Hour

After the Lecture

Skype: L2ahmadi

L2ahmadi@uw...

Ext. 37160

In Person

Message

Call

Tel: (519) 888-4567 x36634, E-mail: penlidis@uwaterloo.ca

Major Topics

- **Chapter 1:**
  Statistical Background (Mb: CHAPTER 2)

- **Chapter 2:**
  Regression Analysis

- **Chapter 3:**
  Statistical Design of Exp

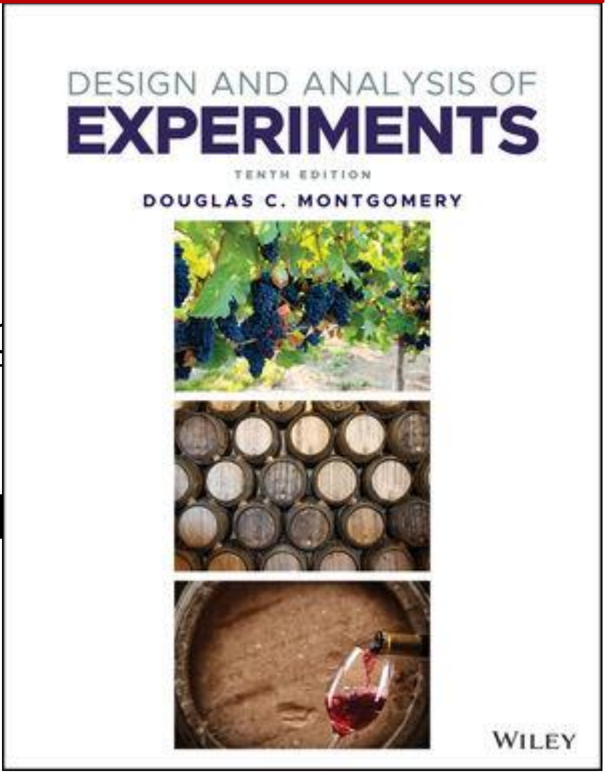
- **Chapter 4:**
  Design/Analysis of Singl
  Experiments

- **Chapter 5:**
  Blocking

- **Chapter 6:**
  Multifactor Experiments

- **Chapter 7:**
  Multifactor Experiments

- **Chapter 8:**
  Concluding Remarks



Course Targets

- Collect data (process or product information)
- Information (info) on **error** (allows one to make comparisons!)
- Develop mathematical **models** (and check for validity of models)
- Why are the above steps useful?

Course Targets (cont'd)

- Process **understanding**! (only way to design/simulate)
- Process **optimization**!
- Product **improvement** (if one understands property characteristics!)
- Identification and control
- **Model-based** approaches!
- … (sky is the limit!)

Review basic statistical concepts including:

- Random variables
- Experimental error and measurement
- Mean and variance
- Random sampling and sample statistics
- Normal distribution
- Central limit theorem
- Confidence intervals
- t-distribution
- Hypothesis testing
- Variance ratio tests; X^2 and F-distributions

WHY STATISTICS?

- What role does statistics play in scientific endeavour?
- Historical origin: measurements related to the “state”.
- Statistics deals with the gathering and the analysis of data.

- *What quantities that you deal with have variability associated with them?*
- Statistics helps us with decision-making in the face of uncertainty!

DECISION-MAKING

During the process of decision-making we are interested in statistical inference which is:

The process of stating something about a population based on a sample (drawn from the population)

Tools for the inference process:

Estimation / Confidence Intervals

Hypothesis Testing

RANDOM VARIABLE

- A random variable is a *chance* quantity, i.e., a mathematical variable whose value is determined by *chance.*
- When we perform an experiment there will be some unique value of the random variable which will correspond to this outcome.
- The value of the random variable cannot be predicted with certainty, because the outcome of the experiment cannot be predicted with certainty.
- However, we can usually express some knowledge about the possible values of the random variable.
- A statement of the probabilities of all possible outcomes of a random variable is called the probability distribution (or probability density function).
- We differentiate between a continuous and a discrete random variable.

EXPERIMENTAL ERROR

- When an experiment is carried out under what are, as nearly as possible, the same conditions, the observed results are never quite identical. The fluctuations which occur are referred to as:
  - noise
  - experimental error or variation
  - "error' (bias, stochastic part, …)
  - uncertainty, etc.
- In statistics, "error" is a technical, emotionally neutral, term. It refers to often uncontrollable variation and *does not* associate blame!
- Sources can include: measurement, analytical, sampling, ambient conditions, skill or alertness of personnel, age and purity of reagents, etc.
- An awareness of the potential sources and effects of experimental error is essential in the design and analysis of experiments
- **Experimental error is a random variable!**

EXPERIMENTAL DATA

- An experimental result or datum in scientific work is often a numerical measurement which describes the outcome of an experimental run (trial, treatment); e.g., yield for a chemical experiment.
- One very useful way of dealing with measurements is to hypothesize the following (simplest) model:

$$y_i = \mu + \varepsilon_i$$

$$\begin{aligned} y_i &= \text{random variable representing the} \\ &\quad \text{experimental data for the i}^{\text{th}} \text{ experiment.} \\ \mu &= \text{"true" value} \\ \varepsilon_{\text{i}} &= \text{random variable representing the} \\ &\quad \text{experimental error for the i}^{\text{th}} \text{ experiment.} \end{aligned}$$

- The error ε cannot be predicted exactly, but can be dealt with using probability theory.

MEAN AND VARIANCE

- These are quantities used to summarize information about random variables and their probability distributions (probability density functions).

- The ***mean (μ)*** is a measure of location. As such it represents a "typical" value or the long-term average value of the random variable. It is also referred to as the *arithmetic mean or <u>expectation</u>*:

$$E(X) = \mu = \sum_x xp(x) \quad \text{discrete r.v.}$$

$$E(X) = \mu = \int_x xf(x)dx \quad \text{continuous r.v.}$$

- The ***variance (σ²)*** is a measure of dispersion or scale. Having established a typical value of a random variable, it is also of interest to have some idea of the "spread":

$$\begin{aligned} V(X) = \sigma^2 &= E\left\{[X - E(X)]^2\right\} \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

Basic properties of E ( ) operator

- E (c) = c   c is a constant
- E (aX) = a E (X)   linearity of operator

Therefore:

- E [aX ± bY] = E [aX] ± E [bY] = a E (X) ± b E (Y)

Basic properties of V ( ) operator

- V (X) or Var (X)
- **Var [aX ± bY] =**

  a^2 Var (X) + b^2 Var (Y)

  ± 2 a b Cov (X, Y)
- Never forget the above formula! We will use throughout the course!

- Cov (X, Y) ? (covariance)

Cov ( ) operator
(more in Ch 2, notes)

- Cov (as the word signifies) relates two random variables!

- Cov (X, Y) =

  E { [X-E (X)] [Y-E (Y)] }

- If X,Y independent, then

  Cov (X, Y) = 0
- Converse statement? Be careful!

A few more remarks…

- $E(XY) \neq E(X)\,E(Y)$, in general (equality holds only if X,Y are independent)
- Square root of Var gives the standard deviation (st dev or st error or se or s.e.)
- So, Var (X) = σ^2 (for X)
- σ^2 vs σ

POPULATION

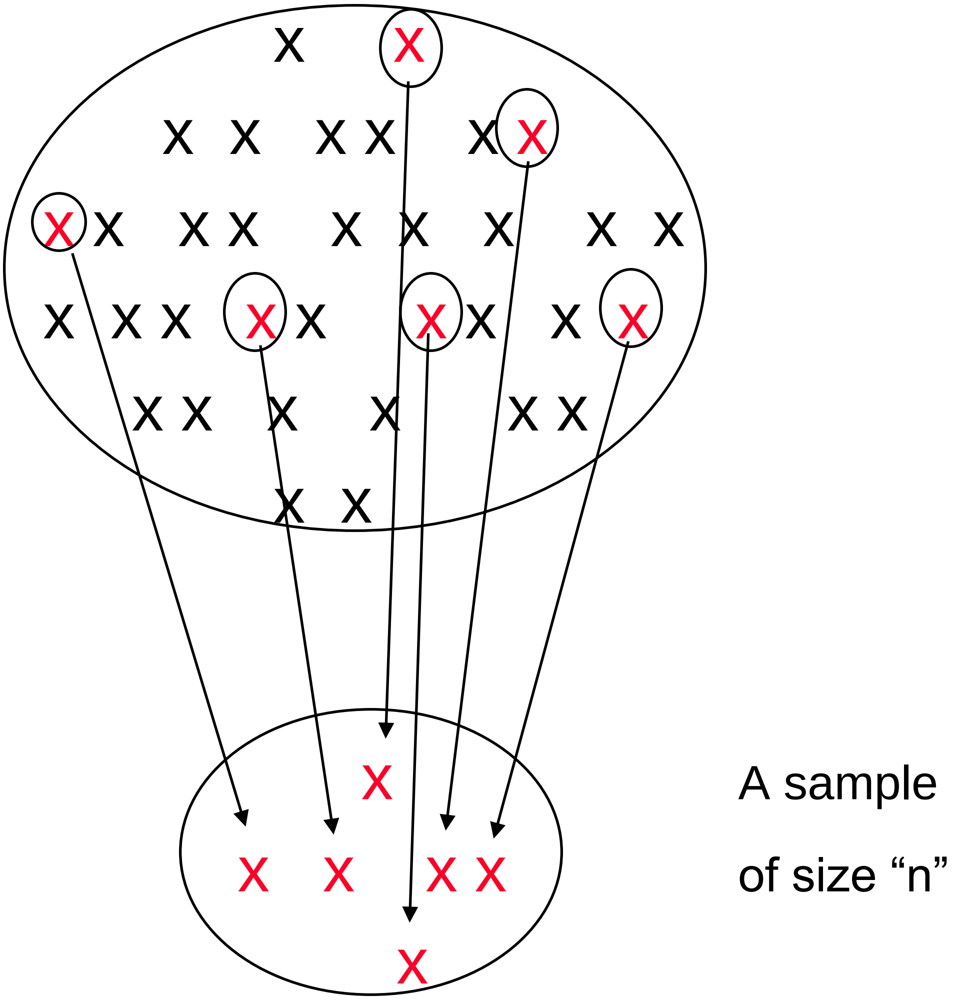
- A *population* is the whole set from which we are collecting information.
- A *population (reference set)* is very large and can be a hypothetical concept including past, present and future production

Example: Suppose we are interested in measuring the lifespan of a particular type of battery.

- Often, as in the example above, it is impossible to make observations on 100% of the population, so we draw a random sample from it.

RANDOM SAMPLE (fair representation)

- In a random sample, each member of the population has an equal chance of being included in the sample.



A sample

of size “n”

RANDOM SAMPLE

- To make inferences we must describe or summarize the data in our sample
- We make use of descriptors (indicators, metrics) called “sample statistics”. The most commonly used are:
  - Sample Mean:
    - *Average of all data values; gives a sense of location.*
  - Sample Variance:
    - *Average of the squared deviations from the mean; gives an indication of the scatter, spread or range of the data.*

SAMPLE STATISTICS

	Sample	Population
Mean	$\bar{X}$	μ
Variance	S^2	σ^2
Std. Deviation	S	σ

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right\}$$

$$S = \sqrt{S^2}$$

A few remarks …

- Note that

$$E\,(s^2) = \sigma^2$$

- If Var (X) = σ^2, then

$$\text{Var (mean)} = \sigma^2 / n$$

More remarks…

- And do not forget to check the CoV or CV or coefficient of variation; (Why? Free of location!)

- CV = (s / mean)
- ‘% error’ = 100 CV

SAMPLE STATISTICS EXAMPLE: YIELD DATA

58.0	66.4	61.8	61.9	60.8	62.7
63.4	63.0	64.1	63.9	65.8	60.1
65.7	62.4	57.9	59.1	64.9	62.7
62.5	63.3	63.0	63.8	60.0	65.3
65.8	62.5	62.1	61.6	62.0	62.4
61.4	65.1	62.6	62.8	62.9	63.1
64.0	64.4	62.6	63.4	62.6	64.3
	63.4	60.3			

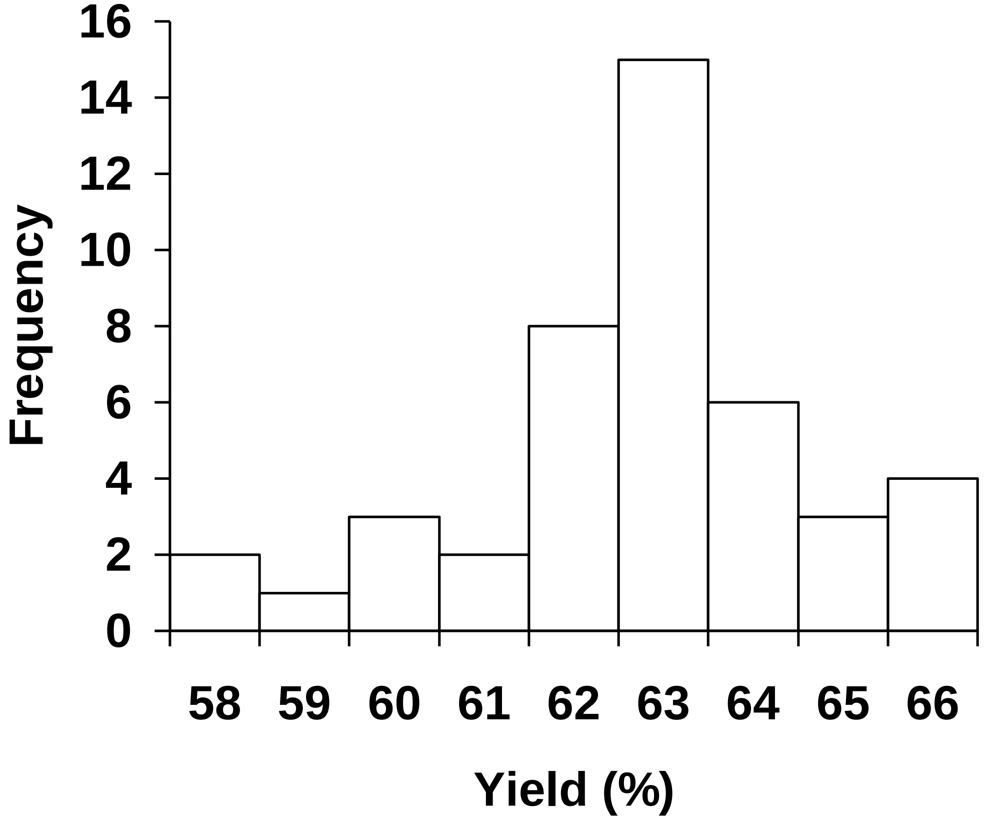
$$\bar{X} = ?$$

$$S^2 = ?$$

$$S = ?$$

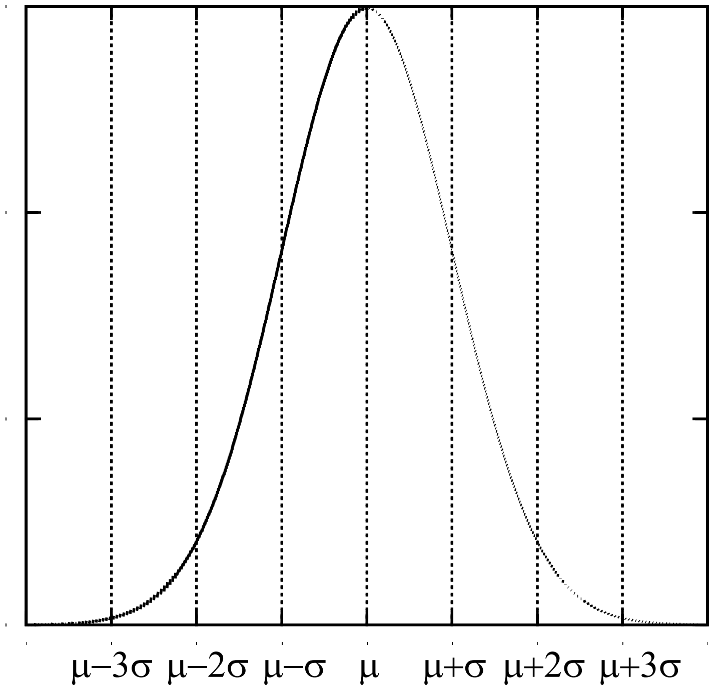
DESCRIBING A SAMPLE

- Pictures are useful to help describe a sample.
- A histogram (or frequency diagram), for example, indicates the shape of the data distribution; frequencies become probabilities in the long term).



NORMAL DISTRIBUTION

One of the most important distributions used for statistical inference (Gaussian, ‘bell-shaped’ curve).



$$P[-\sigma < X - \mu < \sigma] \cong 0.68$$

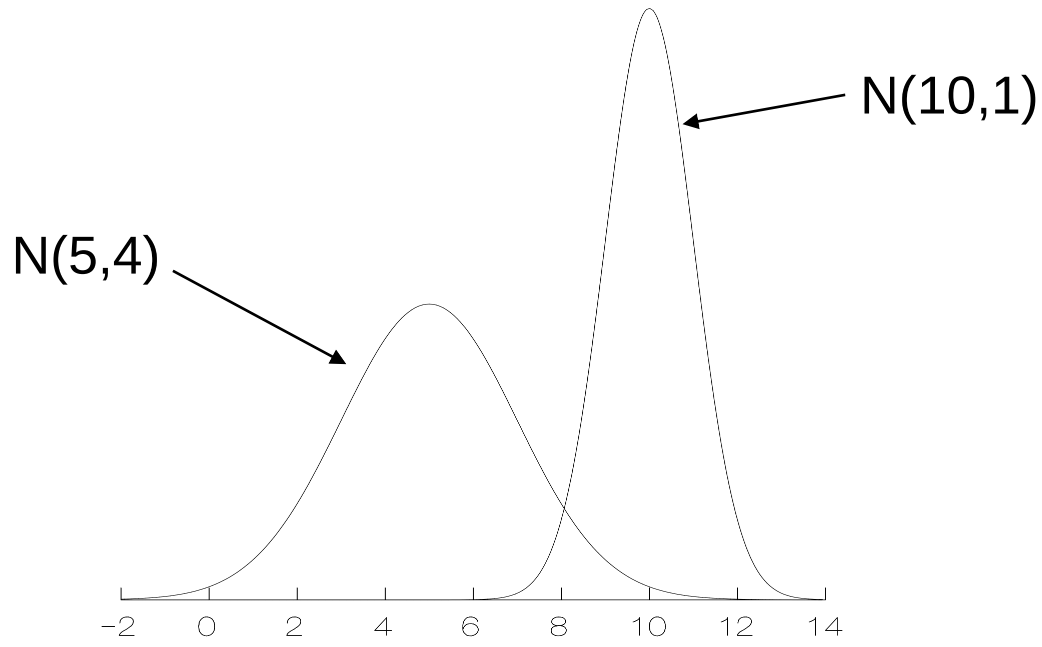
$$P[-2\sigma < X - \mu < 2\sigma] \cong 0.95$$

$$P[-3\sigma < X - \mu < 3\sigma] \cong 0.99$$

NORMAL DISTRIBUTION

The normal distribution requires two parameters, the mean and variance (μ, σ^2), to describe its shape (many examples)!

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}$$



STANDARD NORMAL DISTRIBUTION

- Consider the random variable Z defined by:

$$Z = \frac{X - \mu}{\sigma}, \text{ where } X : N(\mu, \sigma^2).$$

$$E(Z) = \mu_Z = 0;\ V(Z) = \sigma_Z^2 = 1$$

- Z is called the “standard” or “unit” normal random variable (unit normal deviate); if X ~ N(μ, σ²), then Z ~ N(0,1)
- To look up probabilities associated with any normal random variable X, first transform to Z and then look up the probability in a table; i.e., values of Z are tabulated.
- Example:

$$\text{Given} : X : N(8{,}4).\ \text{Find } P(X \leq 10)?$$

$$Z = \frac{10 - 8}{2} = 1.00$$

More remarks…

- E (Z) = ?
- Var (Z) = ?
- Refresh your memory on use of z-tables!

STANDARD NORMAL DISTRIBUTION

- Example:

Given: $X : N(8,4)$, find the value of x for which $P(X \le x) = 0.75$.

$$P(X \le x) = P(Z \le z) = 0.75$$

From the Table $z = 0.6745$

$$x = z\sigma + \mu = 0.6745 * 2 + 8 = 9.349$$
$$P(X \le 9.349) = 0.75$$

- Normal probability rules (basis of SQC/SPC,..):

$$P[-\sigma < X - \mu < \sigma] \cong 0.68$$

$$P[-2\sigma < X - \mu < 2\sigma] \cong 0.95$$

$$P[-3\sigma < X - \mu < 3\sigma] \cong 0.99$$

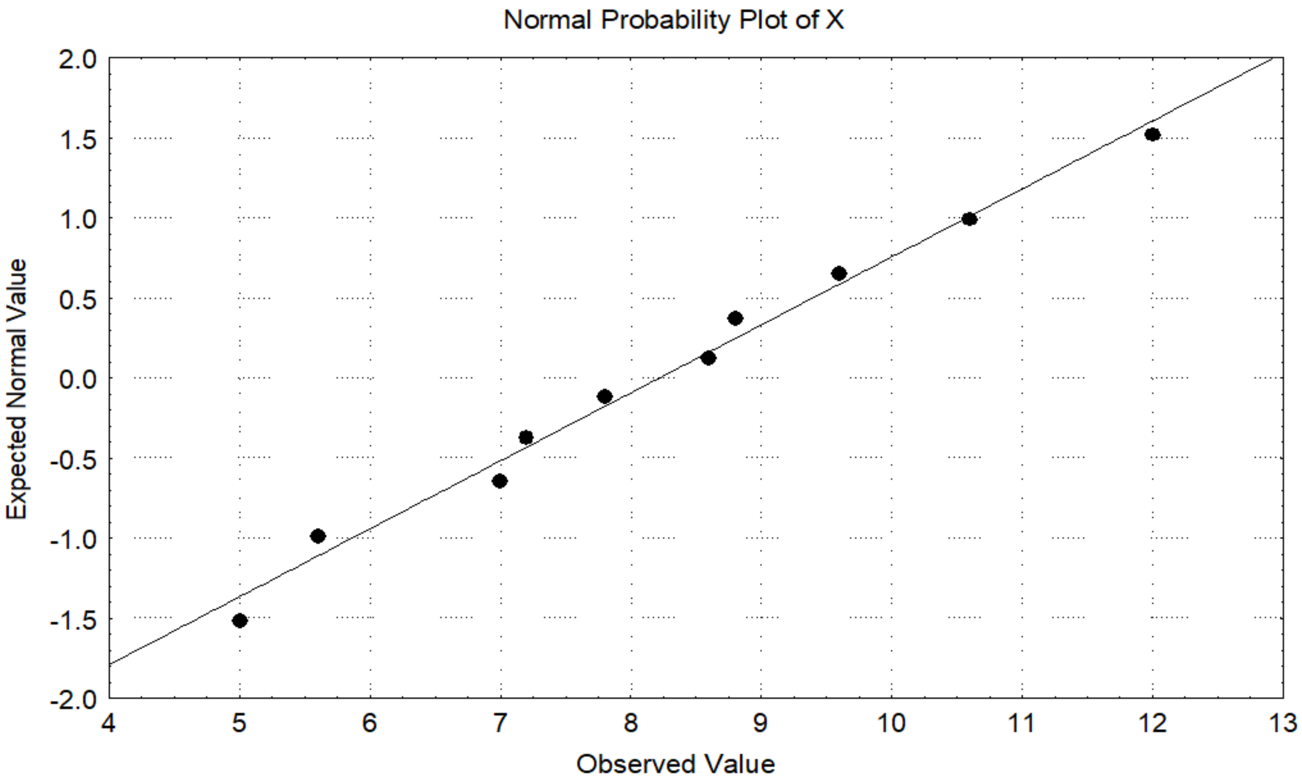
Exercise

- Involves the Normal distn; typical process specs example
- In an industrial process, the (property) of (a specimen) is considered important. The buyer sets specifications on the (property) to be 3 ± 0.01 (appropriate units). The implication is that no part outside of this range will be accepted. It is known that in the process in question, the (property) follows a normal distn with $\mu = 3$ and $\sigma = 0.005$ (units?). On the average, what is the % of product to be scrapped, given the production specs?
- (Hint: Start with a probability statement, as per 2nd yr Stats!)

NORMAL PROBABILITY PLOTS

A normal probability plot (NPP) is a tool for determining if a sample comes from a normal distribution. If it does, then the data should follow a straight line pattern on this plot.

Rank	X_i	P(rank)	Expected Value
1	5	0.05	-1.645
2	5.6	0.15	-1.035
3	7	0.25	-0.675
4	7.2	0.35	-0.385
5	7.8	0.45	-0.125
6	8.6	0.55	0.125
7	8.8	0.65	0.385
8	9.6	0.75	0.675
9	10.6	0.85	1.035
10	12.0	0.95	1.645



NPP remarks…

- P (rank) = (i – 0.5) / n

- n = sample size (# of data)
- i = index (rank x_i (observed values) **from low to high**)
- Expected Normal Value on y-axis is the z-value from the tables corresponding to P(rank)
- P(rank) = cumulative probability value
- NPPs (very useful!) will be revisited in later chapters!
- Regular NPPs vs Half-Normal Plots

SHAPE OF DATA AND THE CENTRAL LIMIT THEOREM

- *Do all raw data we collect follow the normal distribution?* No, but **averages do!**
- Central Limit Theorem (CLT):

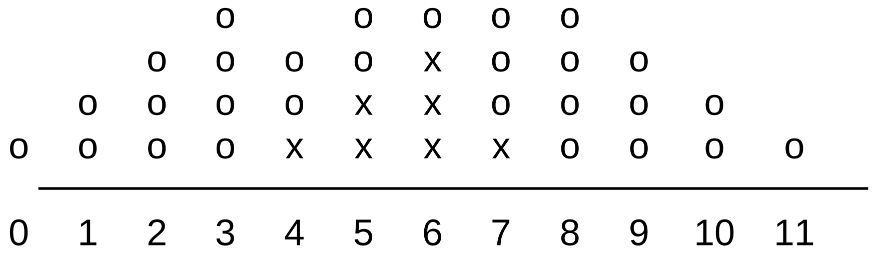
  For a random variable X with mean μ and variance σ^2, the distribution of the sample mean $\bar{\mathrm{X}}$ from a sample of size n on X follows the normal distribution with mean μ and variance σ^2/n. This is true regardless of the shape of the distribution of X.

- This, at least in part, explains why many measurements do follow the normal distribution.

DISTRIBUTIONS

Lab tests of notched Izod bars: 5 bars/sample

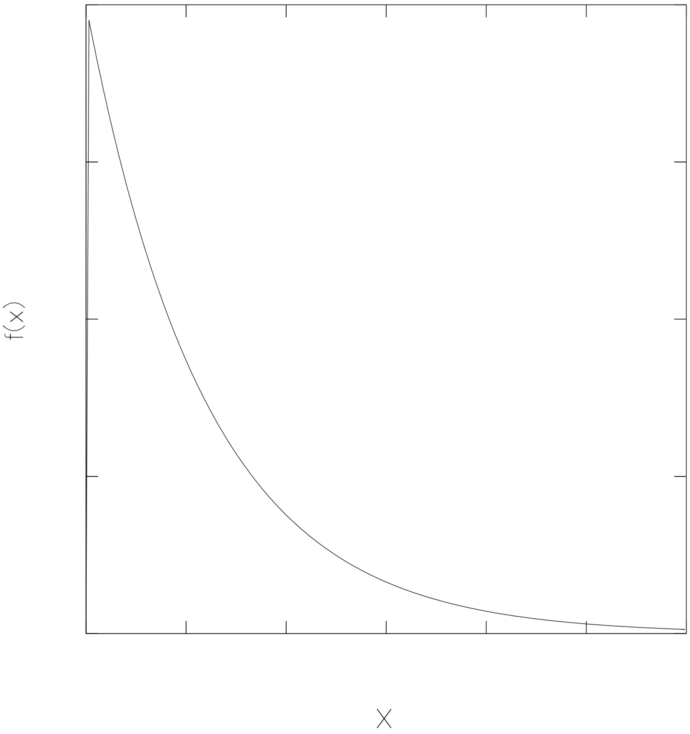
	0	1	2	3	4	5	6	7	8	9	10	11
Sample 1		o		o		o		o		o		
Mean 1						x						
Sample 2			o		o				o	o	o	
Mean 2								x				
Sample 3		o			o			o	o		o	
Mean 3							x					
Sample 4				o	o		o	o		o		
Mean 4							x					
Sample 5			o			o	o		o	o		
Mean 5							x					
Sample 6		o		o		o	o	o				
Mean 6					x							
Sample 7			o	o		o	o					o
Mean 7						x						



```
            o       o   o   o   o
        o   o   o   o   x   o   o   o
    o   o   o   o   x   x   o   o   o   o
o   o   o   o   x   x   x   x   o   o   o   o
  ----------------------------------------------
0   1   2   3   4   5   6   7   8   9   10  11
```

OTHER DISTRIBUTIONS

- An extremely non-normal distribution:



Summary

- Statistics is the science of decision-making in the face of uncertainty.
- Sample statistics estimate population parameters (**fixed, yet unknown**).
- Put another way, we analyze the sample(s) and we make inferences about the population.
- Individual observations may or may not be normally distributed.
- **Averages tend to be normally distributed** (Central Limit Theorem).

CONFIDENCE INTERVALS (CIs)

- The sample mean is a point estimate. As a point estimate, it has the undesirable property that we do not know how far away it might be from the true population mean (it is unlikely to exactly equal μ).

- We would like to have an estimate/indicator which combines information from variability and sample size.

- A confidence interval (CI) is such an estimate.

IS THE MEAN ENOUGH?

Two companies, A and B, are producing pellets to be injection molded. The plasticizing time target is 1.8 seconds. You have your choice of which company to buy from.

A	B
1.8 s	1.9 s

CONFIDENCE INTERVAL FOR THE MEAN

We can construct a confidence interval about the sample mean, $\overline{X}$, such that we are 95.44% confident that the interval contains the true mean, μ. The 95.44% confidence interval would be:

$$\overline{X} \pm 2\frac{\sigma}{\sqrt{n}}$$

A 99.74% confidence interval would be:

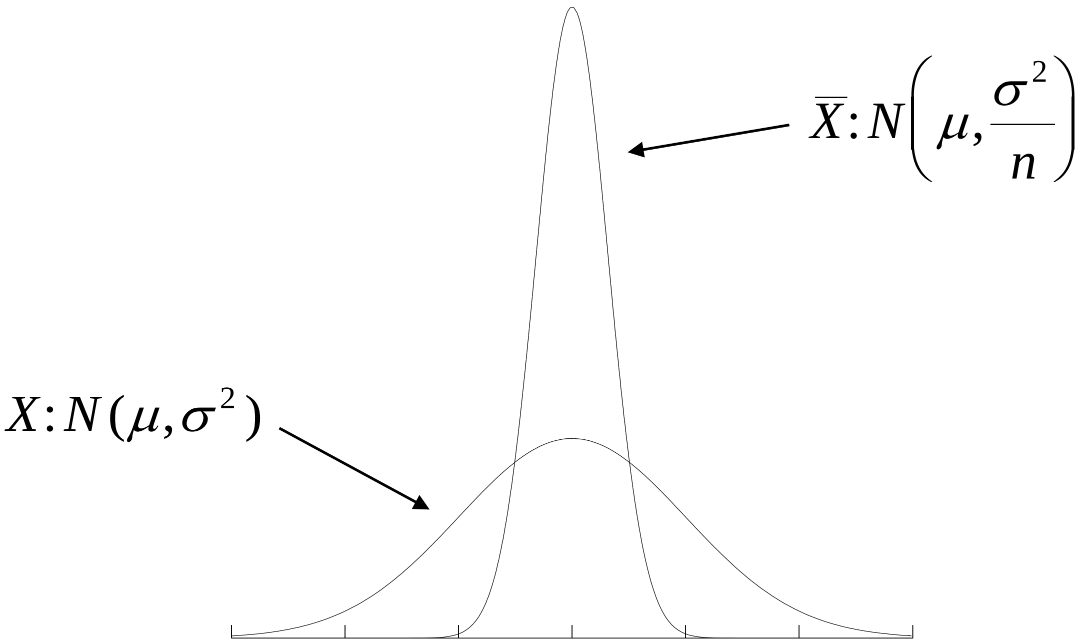
$$\overline{X} \pm 3\frac{\sigma}{\sqrt{n}}$$

Example: For the two companies A and B from the previous page assume the standard deviations indicated are values of σ:

Company A: $1.8 \pm 2\frac{1.41}{\sqrt{2}} = (-0.19, 3.79)$

Company B: $1.9 \pm 2\frac{0.11}{\sqrt{20}} = (1.85, 1.95)$

X AND $\bar{\text{X}}$ DISTRIBUTION



CONFIDENCE INTERVALS

- [An estimate of a statistic ± (a multiplier) · (the st dev of the estimate of the statistic)]

- General guiding formula for a CI

- CI on a population characteristic (population parameter)!

- The multiplier takes into account some level of significance (α, alpha)

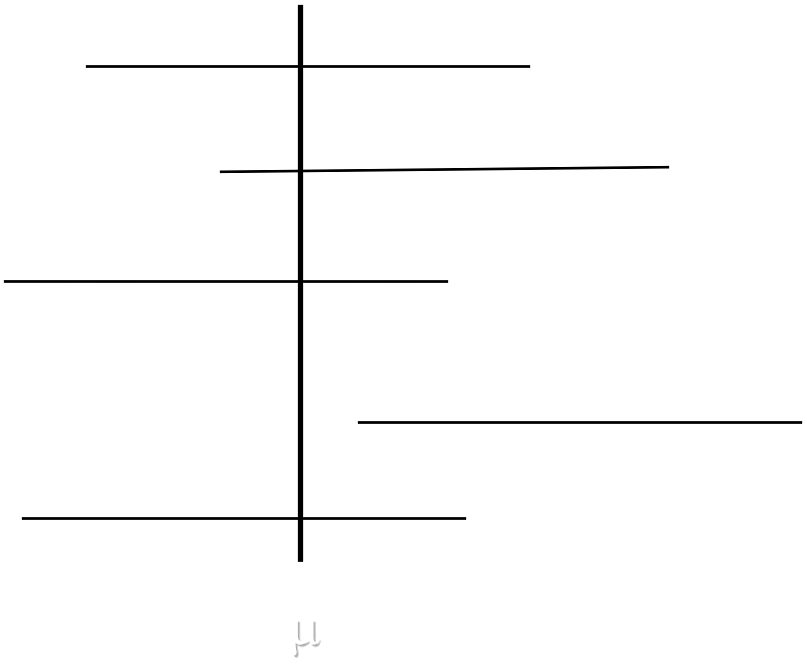
CONFIDENCE INTERVALS

- What are the multipliers for exactly 90, 95 and 99% confidence intervals? Check z-tables…

Confidence Level	Factor
90%	1.645
95%	1.96
99%	2.58

CONFIDENCE INTERVALS

- What do we mean by the word “confidence”:



- For a 95% confidence interval: On repeated sampling from the population, 95% of the numerical intervals generated are expected to contain μ; by chance, 5% (1 in 20) will not.

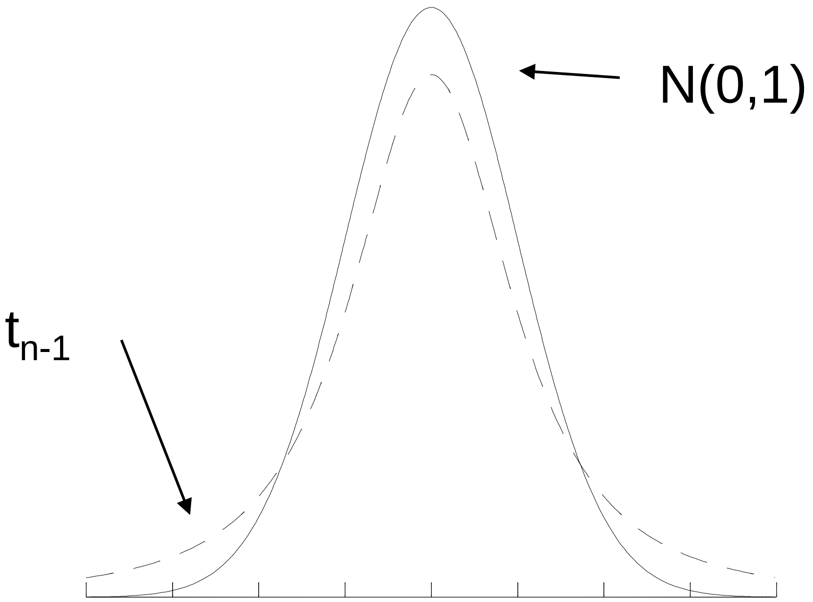
REMARKS ABOUT CONFIDENCE INTERVALS

- The approach assumes **no systematic error** is present.
- 2-sided vs 1-sided
- Confidence intervals could be made **narrower** by decreasing the confidence level (increasing the significance level) or increasing n.
- The population mean, μ, is constant (fixed); the interval changes from sample to sample.
- **Probability is based on the distribution of** $\bar{X}$; not on the individual observations on X.
- The distribution of $\bar{X}$ approaches the normal distribution according to the Central Limit Theorem, as n increases.
- The choice of the factor (1.96 at 95% confidence, for example) yields the shortest interval.

The t-DISTRIBUTION

- When the sample size is small and the variance is unknown, it must be estimated by S^2. The sample mean $\bar{X}$ follows a “t” distn (‘Student t’ distn) rather than a normal distn.
- Normal: $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
- t-distribution: $\bar{X} \pm t \frac{S}{\sqrt{n}}$
- “t”-factor is the tabulated value of the t distribution with n-1 degrees of freedom (df) and a significance level α.
- Confidence level = 100% (1-α)
- Default: 95% for CI; 5% for α

THE t-DISTRIBUTION



N(0,1)

t_{n-1}

The t-distribution has a different shape for different sample sizes (elongated tails; as n becomes large, then ?)

CONFIDENCE INTERVALS

- Tutorial, Ch 1_ex 1 and Ch1_ex 2, on CIs

- Supposed to be review material from 2nd yr Stats

- **Next?** Hypothesis Testing (HT)!

- HT and CIs complementary; same math background, but from a slightly different angle (slightly different starting point)

HYPOTHESIS TESTING (HT)

- Suppose a reaction currently has a desirable yield of 70%
- Researchers have come up with a cheaper catalyst, but they are not sure it will give the same yield (claim? An alternative!)

- The "null hypothesis" ('devil's advocate', 'status quo') to disprove is:

$$H_0 : \mu_0 = 70\%$$

- To test this, we could collect a random sample of yield data using the new catalyst and determine if the sample mean is different from 70%.
- But how different is different?

HYPOTHESIS TESTING

- We could reject the null hypothesis if, for example, a 95% confidence interval calculated for the new sample mean does not include 70%.

$$\bar{X}_{\text{sample}} - t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \le \mu_0 \le \bar{X}_{\text{sample}} + t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$-t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \le \mu_0 - \bar{X}_{\text{sample}} \le t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$-t_{n-1,\frac{\alpha}{2}} \le \frac{\mu_0 - \bar{X}_{\text{sample}}}{S/\sqrt{n}} \le t_{n-1,\frac{\alpha}{2}}$$

$$\left| \frac{\bar{X}_{sample} - \mu_0}{S/\sqrt{n}} \right| \le t_{n-1,\frac{\alpha}{2}}$$

- The last line is true if 70% is a reasonable value for the mean yield in light of the data.

HYPOTHESIS TESTING

- Therefore, if $$\left| \frac{\bar{X}_{sample} - \mu_0}{S / \sqrt{n}} \right| \geq t_{n-1,\frac{\alpha}{2}}$$

  we would reject the null hypothesis H_0: μ_0 = 70%, and conclude that the new catalyst has indeed changed the yield.

- The above procedure has the important property that if we carry out a hypothesis test at a 95% confidence level, there is a 5% chance that we will wrongly reject the null hypothesis; i.e., claim that the yield is different from 70% when it is not.

HYPOTHESIS TESTING

- Accompanying a null hypothesis H_o is an alternative hypothesis H_1 which we “accept” if we reject H_0. In its most general form:

$$H_1 : \mu \neq 70\% \qquad \text{(two - sided)}$$

$$\text{or}$$

$$H_1 : \mu > 70\% \qquad \text{(one - sided)}$$

$$\text{or}$$

$$H_1 : \mu < 70\% \qquad \text{(one - sided)}$$

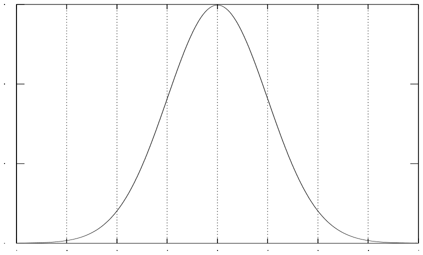
- **The strength of a hypothesis test lies in rejecting H_0.** Therefore, whatever is to be claimed should be stated as the alternative or “research” hypothesis H_1.

- When do we use a one-sided vs. a two-sided test (analogous to 1-sided vs. 2-sided CIs)?

HYPOTHESIS TESTING

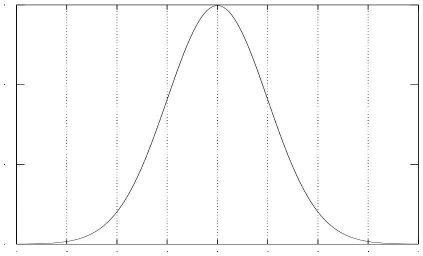
Alternative hypothesis:

Two-sided, ≠



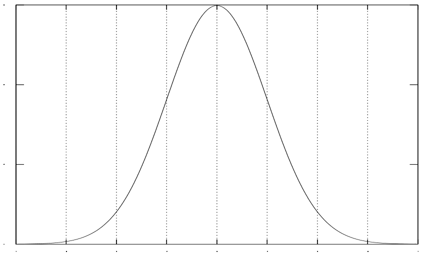
μ_0

One-sided, >



μ_0

One-sided, <



μ_0

HYPOTHESIS TESTING

- Example: A soft-drink beverage company purchases 300 ml bottles from a glass company. It is important to ensure that the bottles meet the minimum internal pressure or bursting strength of 200 psi.

Either H_0: $\mu_0 = 200$ psi H_1: $\mu > 200$ psi (1)
or H_0: $\mu_0 = 200$ psi H_1: $\mu < 200$ psi (2)

(1): If H_0 is rejected, bottles ok.
⇨ If H_0 not rejected, bottles should not be used.
Forces bottle manufacturer to “demonstrate” that bottles are ok.

(2): Bottles are ok, unless H_0 is rejected.

- (1) is appropriate if we have had trouble with this supplier; there is some probability that the null hypothesis is not rejected even though true mean pressure is greater than 200 psi.
- (2) is appropriate if we are generally happy with bottle supplier and a small deviation below 200 psi is not serious.

SUMMARY OF HYPOTHESIS TESTING

1. State (null) H_0 and H_1 — H_0: $\mu_0 = 70\%$
   H_1: $\mu \neq 70\%$

2. Set the significance level of the test (confidence level = 1 - α) — $\alpha = 0.05$

3. Choose the test statistic

4. Determine the distribution — $\bar{X} : N(\mu_0, \sigma^2/n)$

5. Collect a random sample, calculate the test statistic and make a decision on H_0 :

If $\left| \dfrac{\bar{X}_{sample} - \mu_0}{S/\sqrt{n}} \right| \geq t_{n-1,\frac{\alpha}{2}}$ then we reject H_0

If $\left| \dfrac{\bar{X}_{\text{sample}} - \mu_0}{S/\sqrt{n}} \right| < t_{n-1,\frac{\alpha}{2}}$ then we fail to reject H_0

P-VALUES

- Introduce concept schematically
- Commonly the value of α is set at 0.01 or 0.05. The philosophy behind setting this value is to "control" the probability of rejecting a correct (true) null hypothesis.
- Dilemma: What if the test statistic value is close to the critical value?
- An alternative to pre-specifying the value of α is to calculate the probability of exceeding the value you observed given the H_0 - the p-value.
- The amount of evidence for **rejecting the null hypothesis** is then given according to the following guidelines:

$P<0.01$	**very strong evidence**
$0.01<P<0.025$	**strong evidence**
$0.025<P<0.05$	**moderate**
$0.05<P<0.1$	**weak**

HYPOTHESIS TESTING

- Tutorial, Ch 1_ex 3 and Ch1_ex 4, on HT

- Supposed to be review material from 2nd yr Stats

- Can also be done via CIs

- Try to calculate P-values as well

- Essentially, with P-value you calculate your problem's α (alpha)

ESTIMATING THE DIFFERENCE BETWEEN TWO MEANS (Case 1 and Case 2)

- Sometimes we are interested in examining the difference between two means ($\mu_1 - \mu_2$). Suppose we have made n_1 measurements on X_1 and n_2 measurements on X_2.

$$\bar{X}_1 : N(\mu_1, \frac{\sigma_1^2}{n_1}) \qquad \bar{X}_2 : N(\mu_2, \frac{\sigma_2^2}{n_2})$$

$$\bar{X}_1 - \bar{X}_2 : N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

- We distinguish between two cases, which depend upon how much is known about the variances:

Case 1: σ_1^2 and σ_2^2 known (almost never used):

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Case 2:

Case 2a: **σ_1^2 and σ_2^2 “equal” and unknown** (widely used):

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\nu}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_p^2 = \text{"pooled" estimate}$$

$$\nu = n_1 + n_2 - 2$$

Case 2b: **σ_1^2 and σ_2^2 unequal and unknown**:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1 - 1}\right] + \left[\left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2 - 1}\right]}$$

(rounded up (by convention) to the nearest integer)

ESTIMATING DIFFERENCES

- Tutorial, Ch 1_ex 5, on comparing two means
- Supposed to be review material from 2nd yr Stats
- Many alternatives in approaching these types of problems (CIs vs HT)

ESTIMATING DIFFERENCES

- Essentially, in Case 2a, we can make the assumption that population variances 1 and 2 are unknown to us but sufficiently close (‘equal’), and proceed with the use of the ‘pooled’ variance.

- Is that a good assumption?

- We can always make the assumption and then check back (and hence, justify it).

HYPOTHESIS TESTS FOR VARIANCES (Case 3)

- Two companies A and B produce polycarbonate pellets. Here are the summary statistics for two pellet samples:

Company	*N*	$\bar{X}$	*Variance*	*Std. Dev.*
A	100	0.024 g/pellet	$2.25\text{x}10^{-6}$ $(\text{g/pellet})^2$	0.0015 g/pellet
B	100	0.025 g/pellet	$1.64\text{x}10^{-5}$ $(\text{g/pellet})^2$	0.0041 g/pellet

- The pellets seem to have the same average weights, but there is a difference in the pellet weight variability.

F-TEST

- We can examine the difference by using an F-test which involves the ratio of the two variances

$$S_1^2 \big/ S_2^2$$

- When testing for differences, the smaller observed variance is always in the denominator.

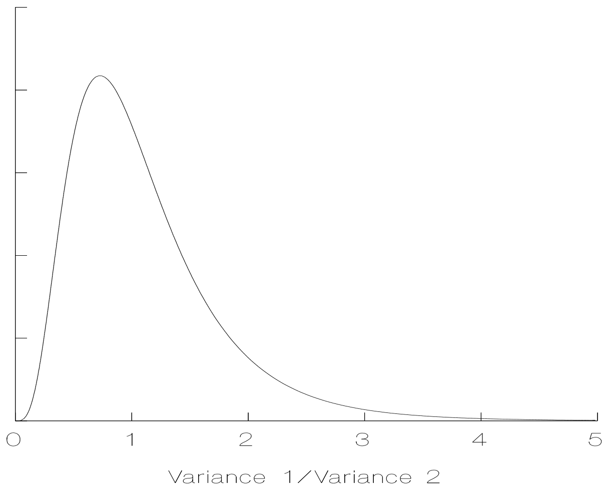
- The test involves the following hypotheses:

$$H_0 : \sigma_1^2 \big/ \sigma_2^2 = 1 \qquad H_1 : \sigma_1^2 \big/ \sigma_2^2 > 1$$

- The null above says that the two variances are equal.

F-DISTRIBUTION

- The ratio is compared to an F-distribution to see if the variances are equal.



- The F-distribution requires two numbers to define its shape: F(df_{num}, df_{den}) where:

df_{num} = degrees of freedom in numerator variance (n_1-1)

df_{den} = degrees of freedom in denominator variance (n_2-1)

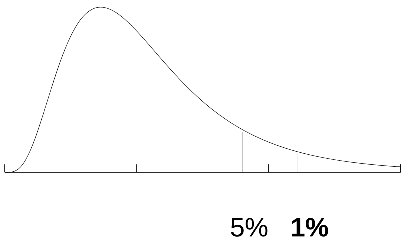
F-TEST

- This probability can be found in an F-table:

		degrees of freedom in numerator			
		1	2	3	4
degrees of freedom in denominator	1	161	200	216	225
		4052	**4999**	**5403**	**5625**
	2	18.51	19.00	19.16	19.25
		98.49	**99.01**	**99.17**	**99.25**
	3	10.13	9.55	9.28	9.12
		34.12	**30.81**	**29.46**	**28.71**

regular type = 5%

bold type = 1%



- The F-table above is for 5% and 1% significance levels. It provides "critical" F-values that are cut-off values for those probabilities.

F-TABLE

- To use the F-table:

1. Choose your level of significance (e.g., 1% or 5%).

2. Determine df_{num} and df_{den} from the respective sample sizes.

3. Look up critical F-value from table:

If calculated ratio > critical F-value,
reject H_0, hence the
variances in the ratio seem
significantly different.

F-TEST

- Back to the pellet weight example:

$$S_B^2 = 1.64\text{x}10^{-5} \qquad S_A^2 = 2.25\text{x}10^{-6}$$

$$df_B = 100 - 1 = 99 \qquad df_A = 100 - 1 = 99$$

$$F_{observed} = \frac{S_B^2}{S_A^2} = \frac{0.0000164}{0.00000225} = 7.29$$

- Compare to the appropriate F-distn (F-tables):

Since F_{crit} is 1.4 → $F_{obs.} > F_{table}$ (99,99) at $\alpha = 0.05$,

Therefore, the variances are significantly different from each other (i.e., reject H_0)

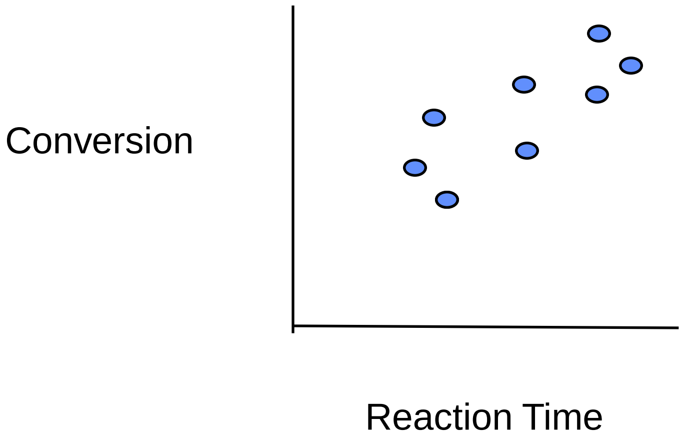
Remarks…

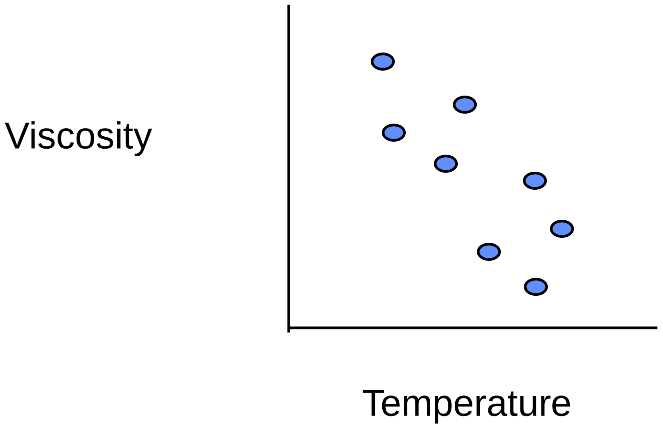
- Tutorial, Ch 1_ex 6, on obtaining a CI on a variance

- Supposed to be review material from 2nd yr Stats

CORRELATION

- Sometimes, two variables are related to each other in a linear fashion:





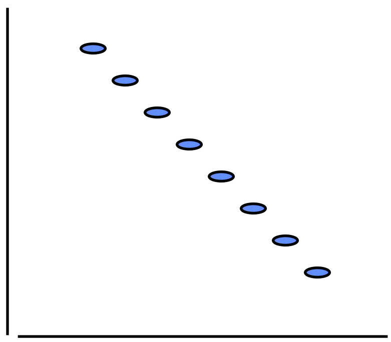
CORRELATION

- A measure of the strength of the **linear** relationship between two variables X and Y, is given by the sample correlation coefficient:

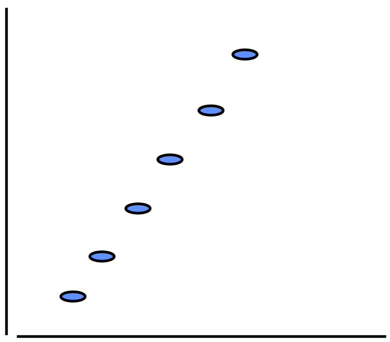
$$r = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum\left(X - \overline{X}\right)^2 \sum\left(Y - \overline{Y}\right)^2}}$$

- The correlation coefficient
  - can range from -1 to +1
  - represents stronger positive (or negative) correlation as it gets closer to positive (or negative) 1
  - is an indicator of **linear association** only!

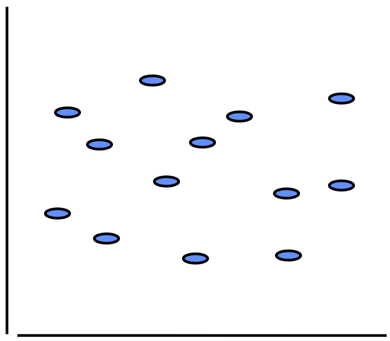
CORRELATION



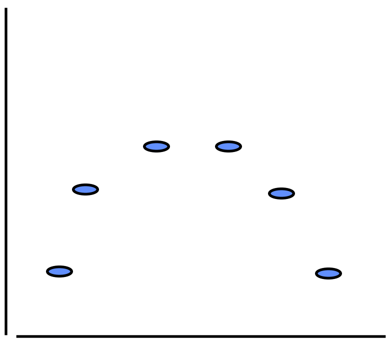
r =



r =



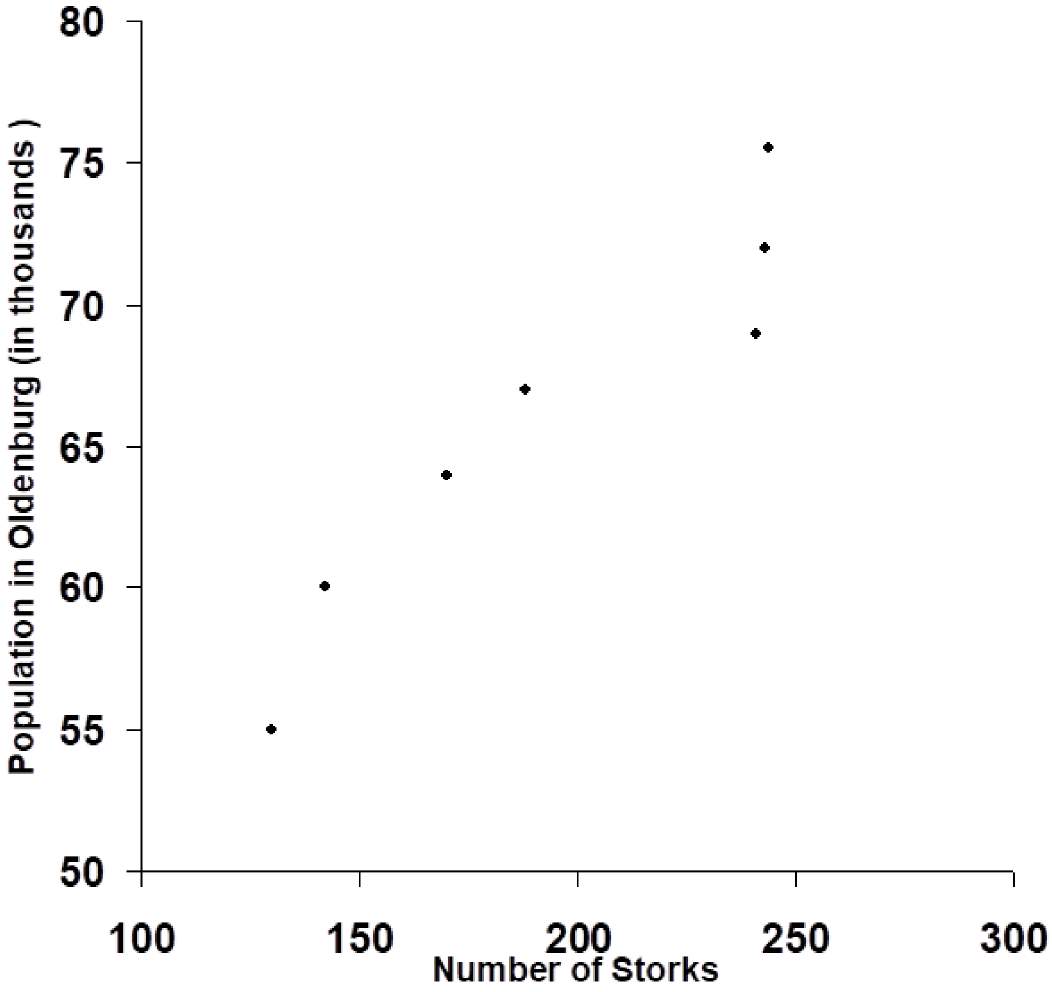
r =



r =

CORRELATION

***Remember:** Correlation does not necessarily imply **<u>causation</u>**!*

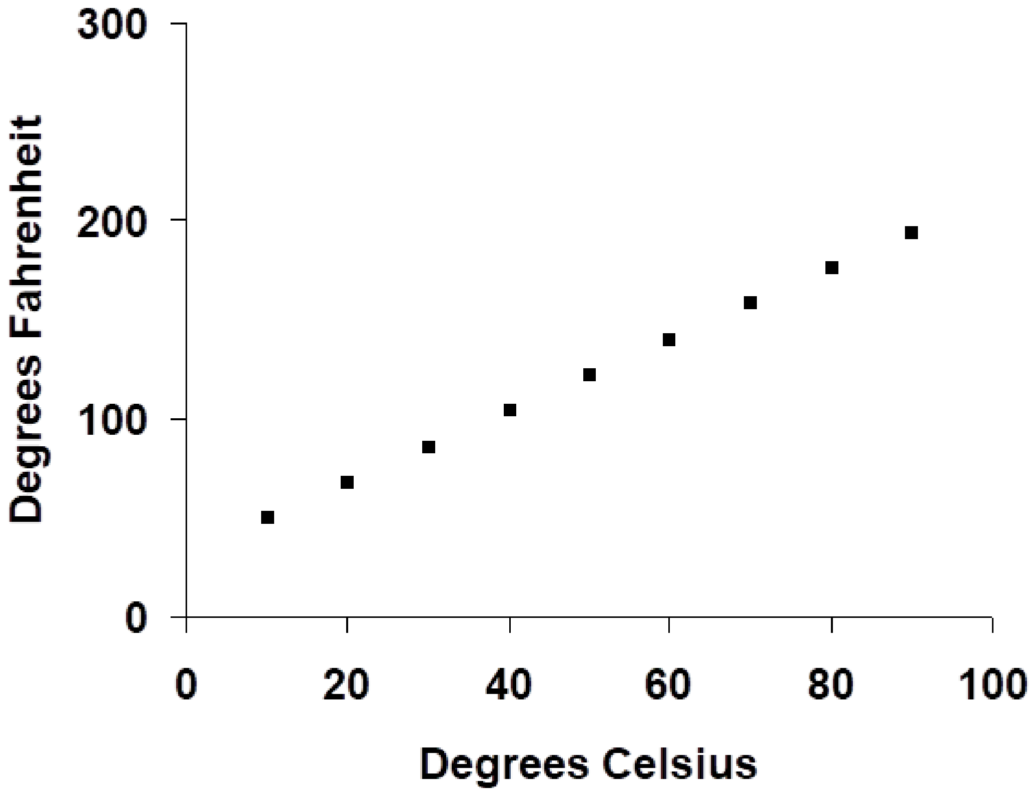


A plot of the population of Oldenburg at the end of each year against the number of storks observed in that year, 1930-1936; many other examples over the years!

SIMPLE LINEAR REGRESSION

When two variables are related linearly, a line can be used to represent and model the relationship between them.

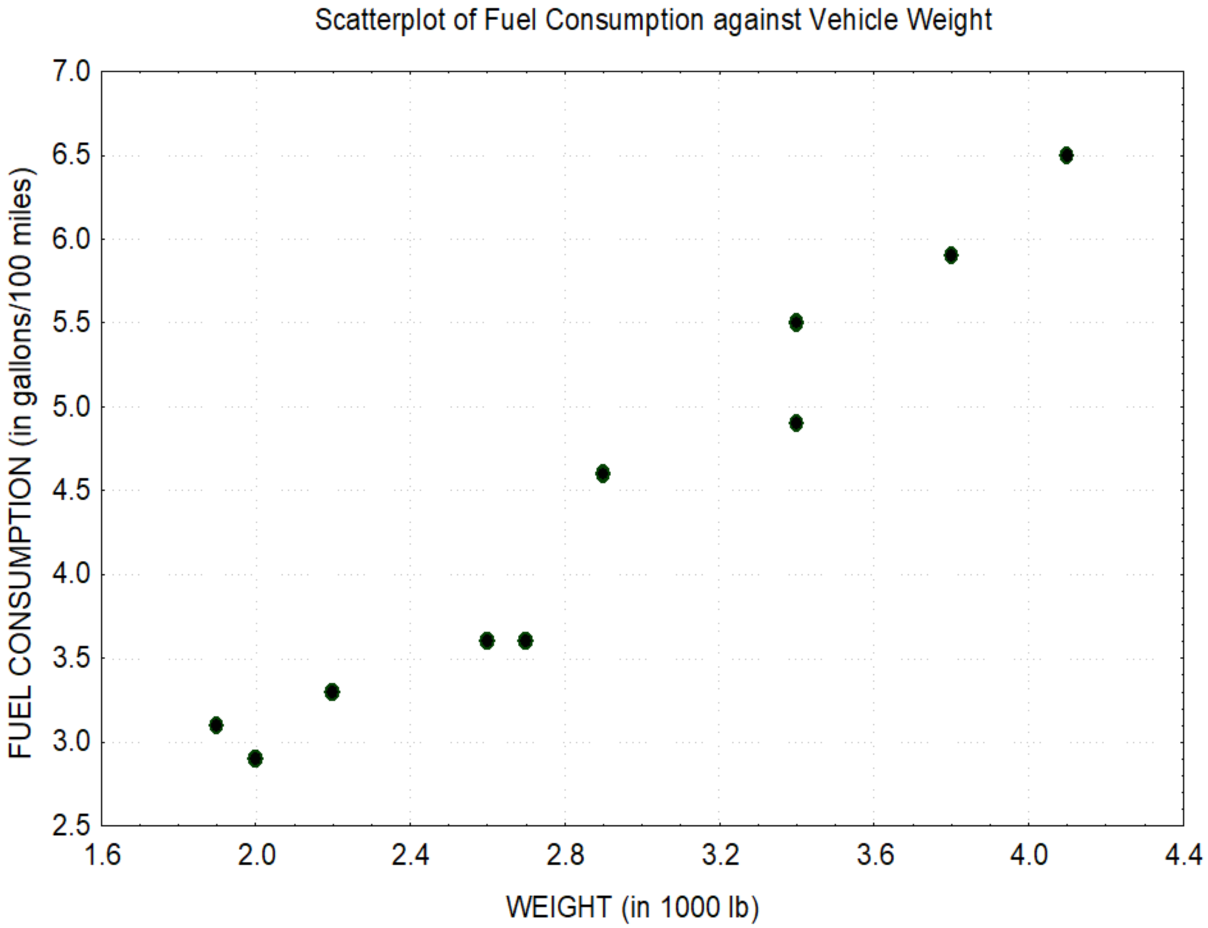
$$Y = aX + b$$



$T_F = T_C * 9/5 + 32$

SIMPLE LINEAR REGRESSION

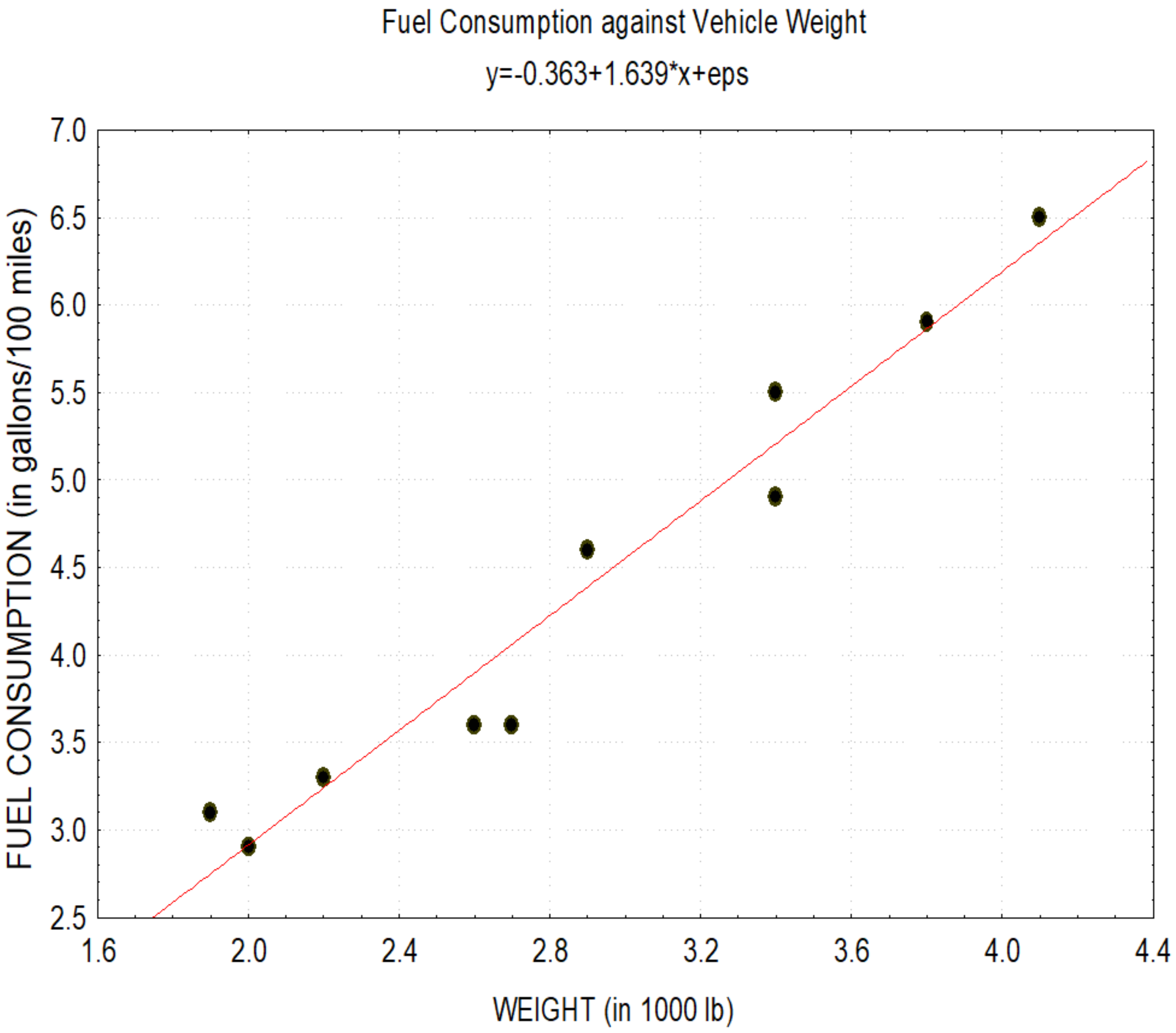
- Some relationships are close to linear, but are not exact due to experimental error.



- These data could be modeled by a straight line. But how do we find the “best” line?

SIMPLE LINEAR REGRESSION

- We choose an equation (slope and intercept) that minimizes the vertical distance between the data and the assumed model.



SIMPLE LINEAR REGRESSION

- Mathematically, we can accomplish this by minimizing the sum of squared **residuals** (errors):

$$SSE = \sum\left(Y - \hat{Y}\right)^2 = \sum e^2$$

where $\hat{Y}$ = the estimate of Y provided by the chosen line.

- The slope and intercept, respectively (see below), that minimize the SSE are given by:

$$\hat{a} = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sum\left(X - \overline{X}\right)^2}$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}$$

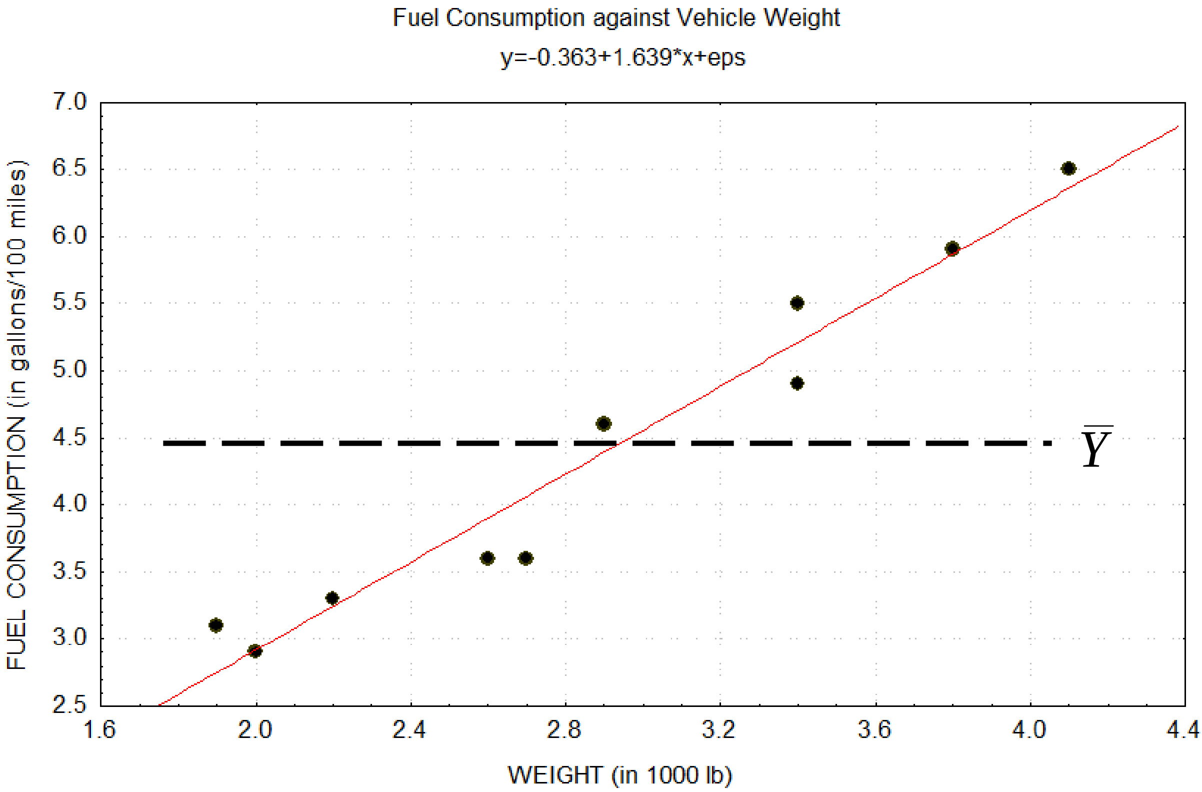
SIMPLE LINEAR REGRESSION

- We can explain the regression in terms of variability in the data accounted for by the model:

$$Y - \overline{Y} = \hat{Y} - \overline{Y} + Y - \hat{Y}$$

distance accounted for by line

distance unaccounted for by line



SIMPLE LINEAR REGRESSION

- With some algebra we can decompose the **total variation** in the data into the amount **explained by the regression (model)** and the amount **unexplained by the regression**:

$$\sum\left(Y-\overline{Y}\right)^2=\sum\left(Y-\hat{Y}\right)^2+\sum\left(\hat{Y}-\overline{Y}\right)^2$$

$$\begin{aligned}\text{where } \sum\left(Y-\overline{Y}\right)^2 &= \text{ total sum of squares } = \text{ SST}\\ \sum\left(Y-\hat{Y}\right)^2 &= \text{ S. of S. unexplained } = \text{ SSE}\\ \sum\left(\hat{Y}-\overline{Y}\right)^2 &= \text{ S. of S. explained } = \text{ SSR}\\ &\quad\ \text{by regression}\end{aligned}$$

$$\therefore\ \text{SST} = \text{SSR} + \text{SSE}$$

ANALYSIS OF VARIANCE

- This breakdown of the variability can be summarized in an Analysis of Variance (ANOVA) table ('book-keeping'):

Source	SS	df	MS	F
Regression	$\sum\left(\hat{Y}-\overline{Y}\right)^2$	p-1	SSR/(p-1)	MSR/MSE
Error	$\sum\left(Y-\hat{Y}\right)^2$	n-p	SSE/(n-p)	
Total	$\sum\left(Y-\overline{Y}\right)^2$	n-1		

where df = number of independent observations to compile each sum-of-squares

p = number of parameters in the model (here 2)

n = number of observations

MS = mean square (variance)

So, what is MS essentially? MSE or MS_E? F (F probe) above?

SIMPLE LINEAR REGRESSION

- The last column of the ANOVA table refers to the observed F-value for a hypothesis test which is looking at the overall significance of the regression.
- For a simple straight line this is equivalent to testing the slope against zero:

$$H_0 : a = 0 \qquad H_1 : a \neq 0$$

- In a simple linear regression case, this is the same as testing the hypothesis that the variation explained by the model (i.e., via regression) is significantly larger than the error term:

$$H_0 : \frac{MSR}{MSE} = 1 \qquad H_1 : \frac{MSR}{MSE} > 1$$

Where MSR = SSR/df_{SSR} = variance explained by model
MSE = SSE/df_{SSE} = variance due to errors

- What type of test is this?

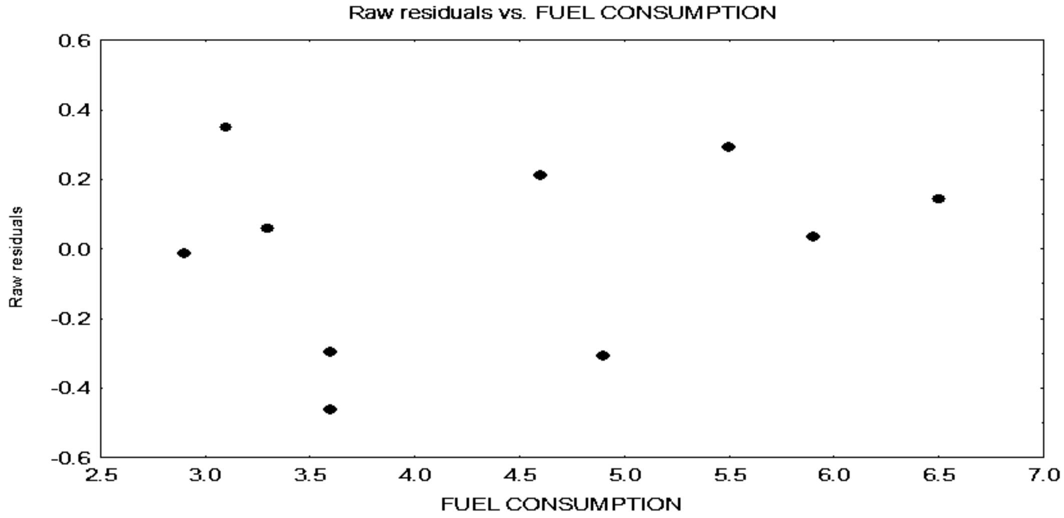
SIMPLE LINEAR REGRESSION

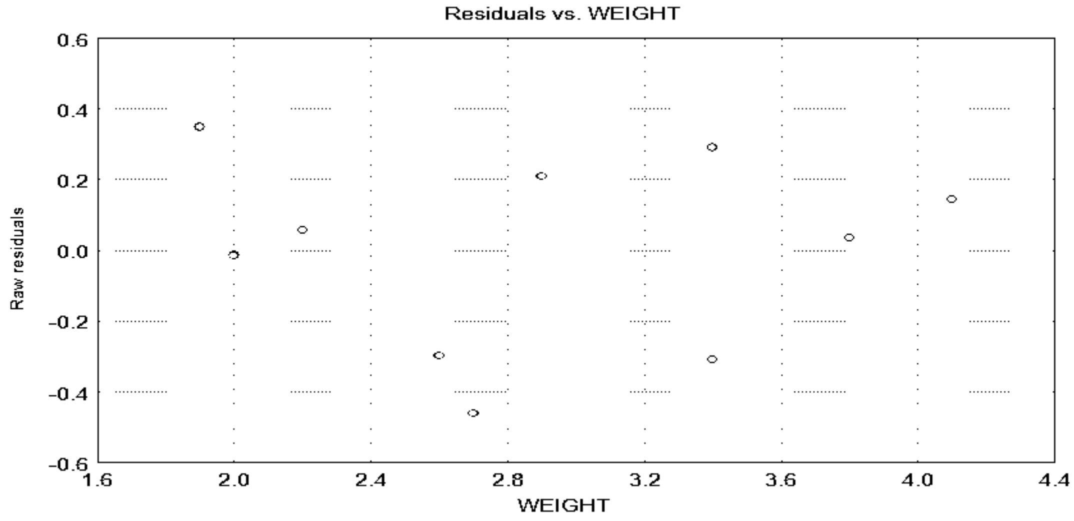
- To test this hypothesis, we compare the ratio of MSR/MSE to the F-distribution. If MSR/MSE > F(p-1,n-p), then we are in the rejection region of the null hypothesis H_0, hence it seems that the slope (and therefore the model) is significant.

- If the model is significant, the model explains a major amount of the variability in the data.

REGRESSION DIAGNOSTICS (more in Ch 2, notes)

- The most valuable regression diagnostic tool is the residual plot (e vs something).
- In examining residual plots we are looking for trends or patterns.





REGRESSION DIAGNOSTICS

- **Provided the residuals are “well-behaved”,** another useful quantity to examine is R^2, the “Coefficient of Determination”:

$$R^2 = \frac{SSR}{SST}$$

- This ratio (R^2 above or r^2 from correlation coefficient slide) indicates what percentage of the variation present in the data is accounted for by the model. R^2 ranges from 0 to 1 and, in general, the closer it is to 1, the better the model explains the data.

- **It is unwise** to place too much emphasis on R^2 alone, or to consider it without examining residual plots. It is possible to have a low R^2 value with a model that fits the data reasonably well. Under what circumstances might this happen?

Example: ANSCOMBE’S DATA (R^2 = 0.67)

Anscombe's Data

Observa-tion	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

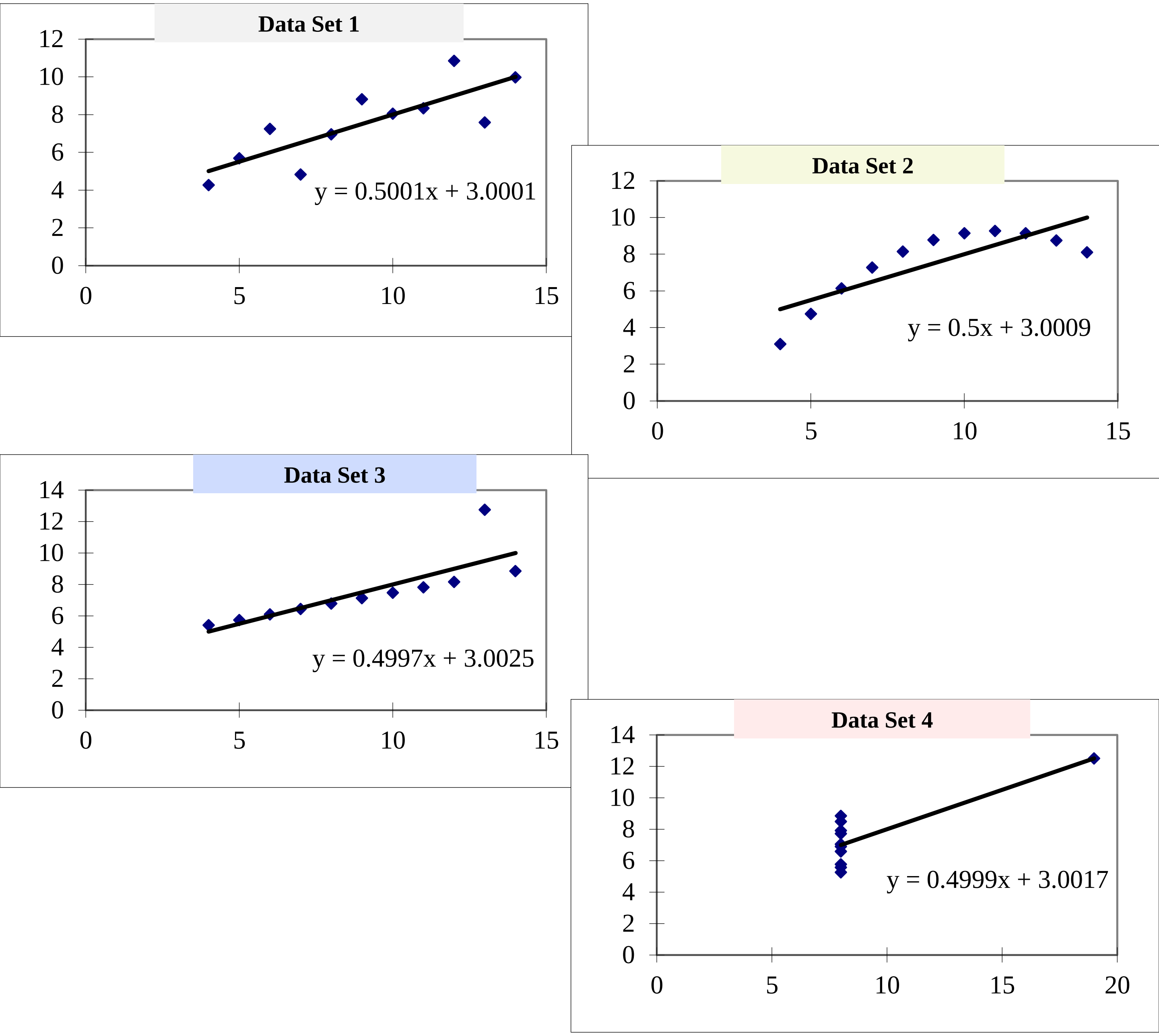
Summary Statistics

N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.501	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

Sources:
Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut: Graphics Press, 1983), pp. 14-15.

F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, vol. 27 (Feb 1973), pp. 17-21.

ANSCOMBE’S DATA
(R^2 = 0.67)



MULTIPLE LINEAR REGRESSION
(more on this in Ch 2, notes)

- Often we have more than one X (independent, regressor, input, setting) variable. We then carry out Multiple Linear Regression instead of Simple Linear Regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots. + \beta_p X_{pi} + \varepsilon_i \quad i = 1,..,n$$

- To test the **overall significance of the model**, we test the hypothesis that all the β_i's (slopes) are equal to zero:

$$H_0 : \beta_i = 0 \qquad H_1 : \text{some of } \beta_i \neq 0$$

- Again, this involves an F-test in which the ratio MSR/MSE is compared against F(p-1,n-p) at a preselected confidence level.

SUMMARY

- In this section we have summarized the statistical methods necessary for

  ESTIMATION

  HYPOTHESIS TESTING and

  CONFIDENCE INTERVALS

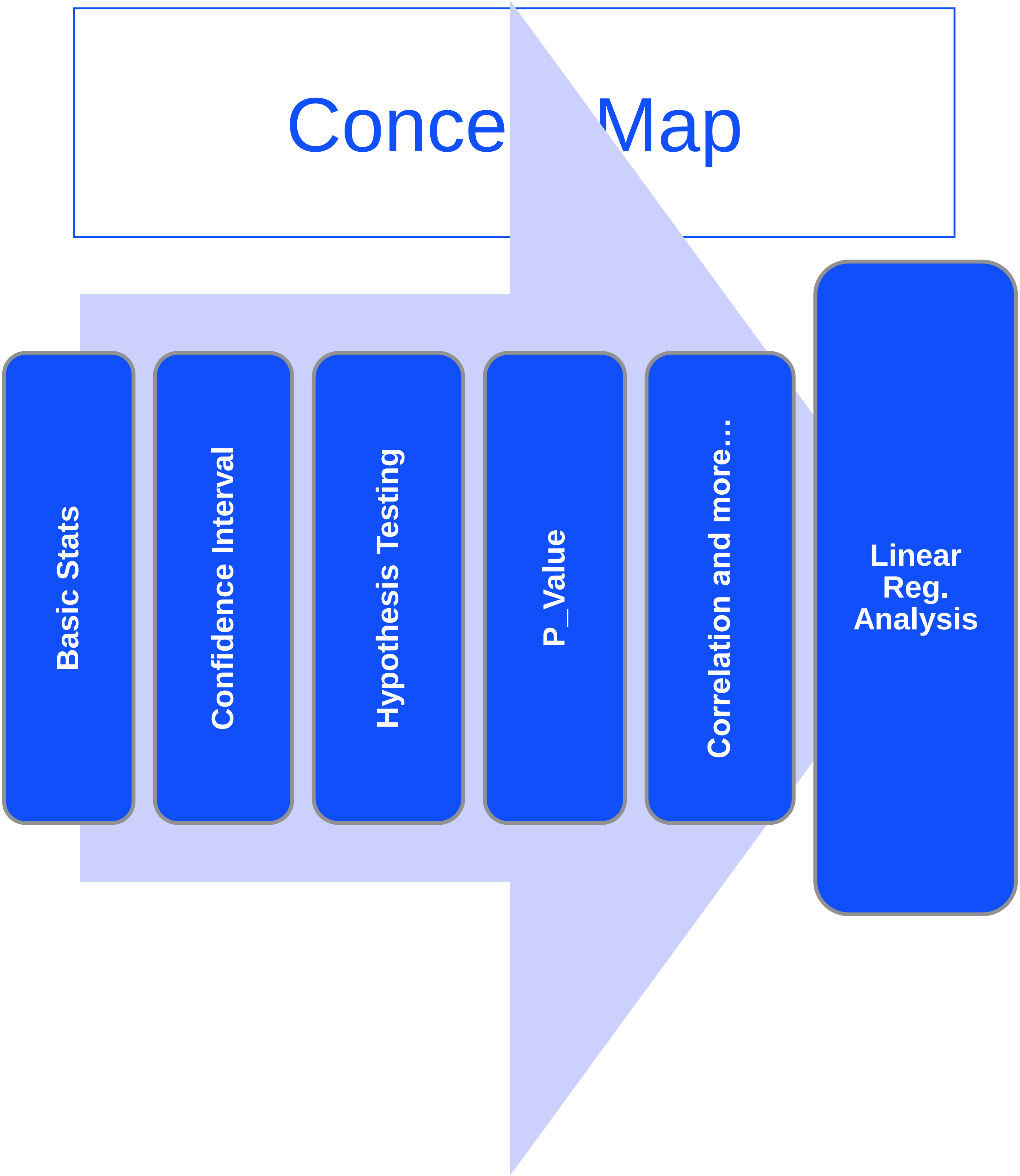
  REGRESSION MODELLING (linear)

  ANOVA

  (Normal Probability Plot, Residual Plots)

- These techniques are an essential part of analyzing and therefore understanding experimental data (and a quick overview of a 2nd-yr Eng or Sci Applied Statistics course).

Conce Map


Basic Stats
Confidence Interval
Hypothesis Testing
P_Value
Correlation and more…
Linear Reg. Analysis

Helpful Review

Paired comparison test (Paired *t*-test):

"Special case of case II"

It is a special case of the two sample t-tests (Variance Unknown) occurs when the observations on the two populations of interest are collected in pairs. Each pair of observations, say (X_{1j},X_{2j}), is taken under homogeneous conditions, but these conditions may change from one pair to another. The test procedure consists of analyzing the differences between readings (d). The Hypothesis test would be $H_0 : \mu_d = 0$ $H_1 : \mu_d \neq 0$

and the test statistics would be $t_o = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$.

Where $\bar{d}$ is the average of all d (differences between readings), and S_d is the standard deviations of all d (differences between readings).

(See section 2.5, pg 48 of Montgomery book, 10th ed.; or pg 53 of 8th ed.).