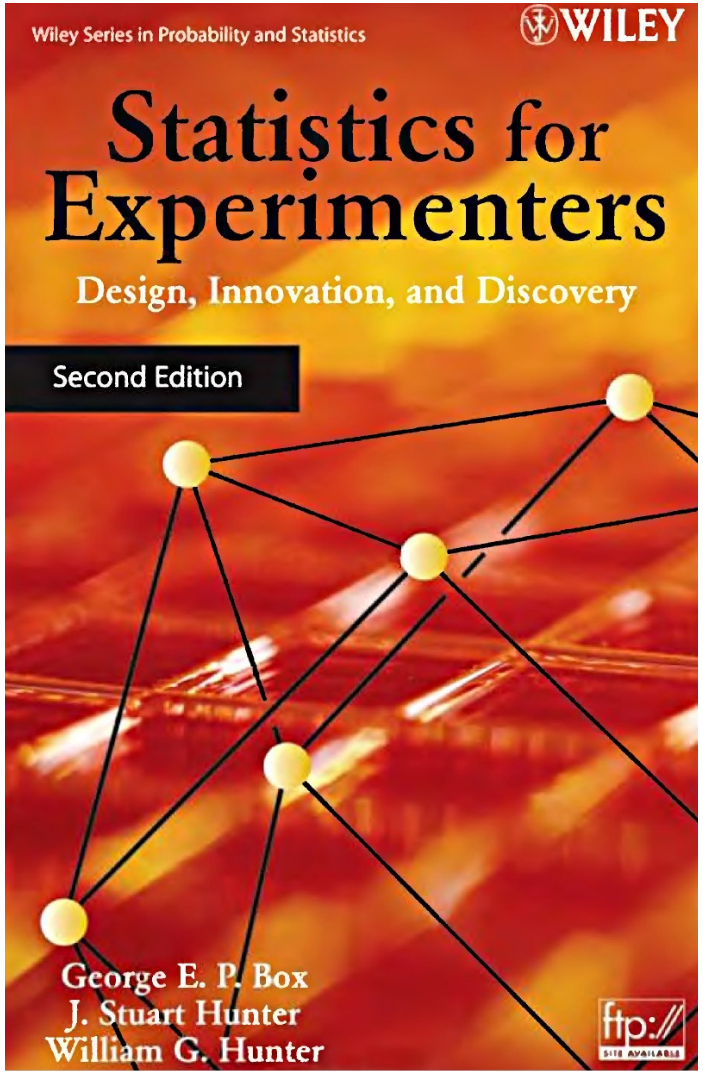
Powertrain Calibration Optimisation

Introduction to Statistics

Overview

- Basic concepts
- Continuous distributions
- Estimation
- Significance tests
- Regression
- ANOVA

Powertrain Calibration Optimisation



Wiley Series in Probability and Statistics

WILEY

Statistics for Experimenters

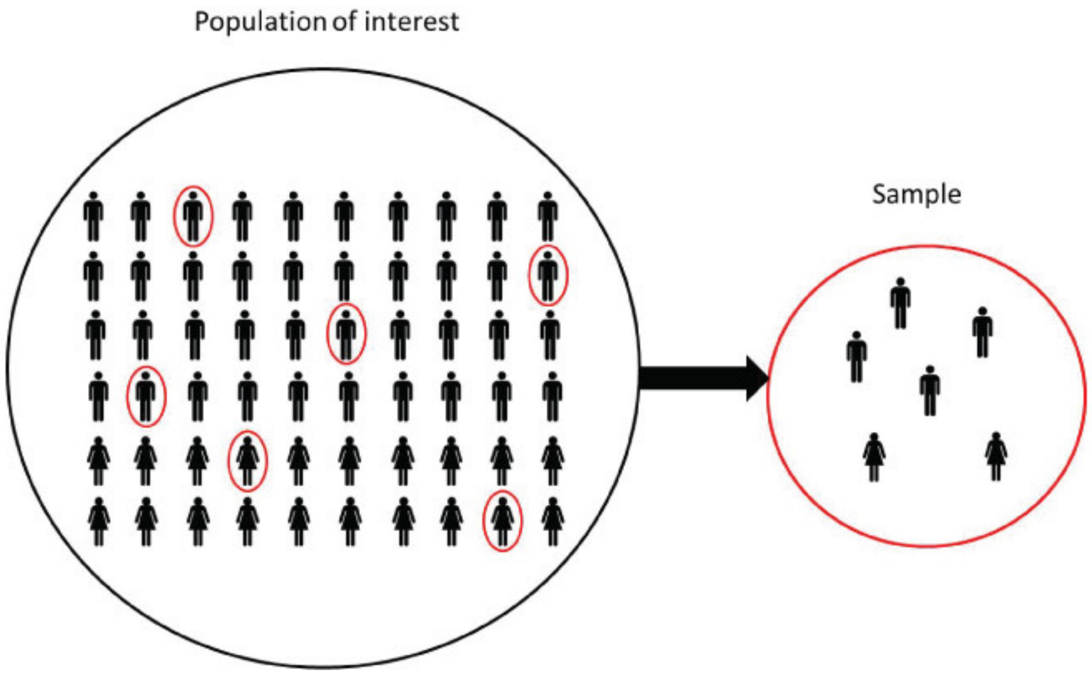
Design, Innovation, and Discovery

Second Edition

George E. P. Box
J. Stuart Hunter
William G. Hunter

ftp://
SITE AVAILABLE

Population vs sample



From a set of data, $\bar{y}$ and $s = \sum_N \frac{(x_i - \eta)^2}{N-1}$ form **estimates** of the *population* mean, μ and standard deviation, σ

Definition - Degrees of Freedom

- How many choices?
- Degrees of freedom relate to the number of ‘observations’ that are free to vary when estimating statistical parameters

$$mean = \frac{x_1 + x_2 + \cdots x_n}{n}$$

In calculating the mean only $n - 1$ observations are ‘free to vary’

Making measurements – location and spread

Mean, $\bar{x} = \frac{x_1+x_2+\cdots+x_N}{N} = \sum_N \frac{x_i}{N}$

Also known as the expectation of x i.e. $E(x)$.

Standard deviation, s, variance, s^2

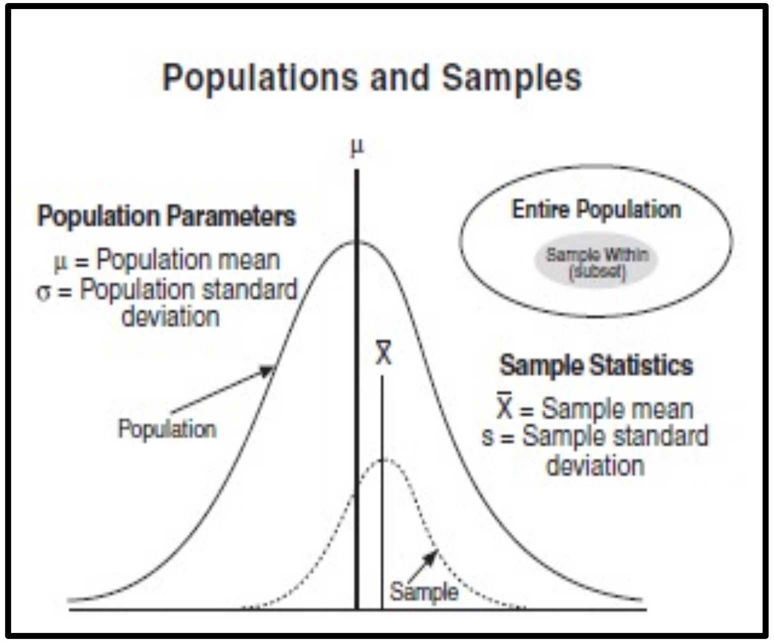
$$s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})^2}{N-1} = \sum_N \frac{(x_i-\bar{x})^2}{N-1}$$

sample variance

For ease of calculation,

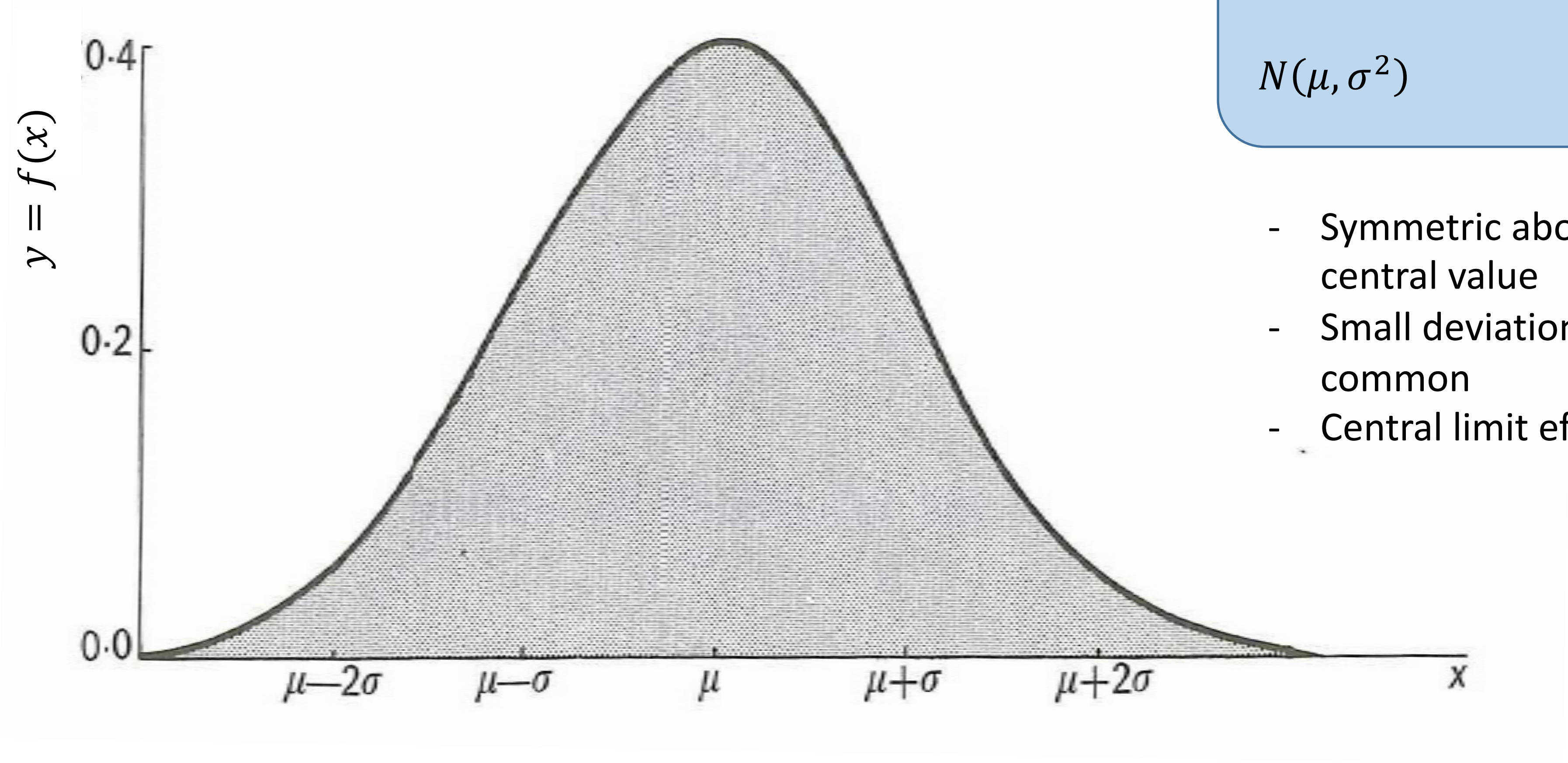
$$s^2 = \frac{\sum_N x_i^2 - N\bar{x}^2}{N-1}$$

Why $N-1$?



Probability distributions

Normal Probability Distribution

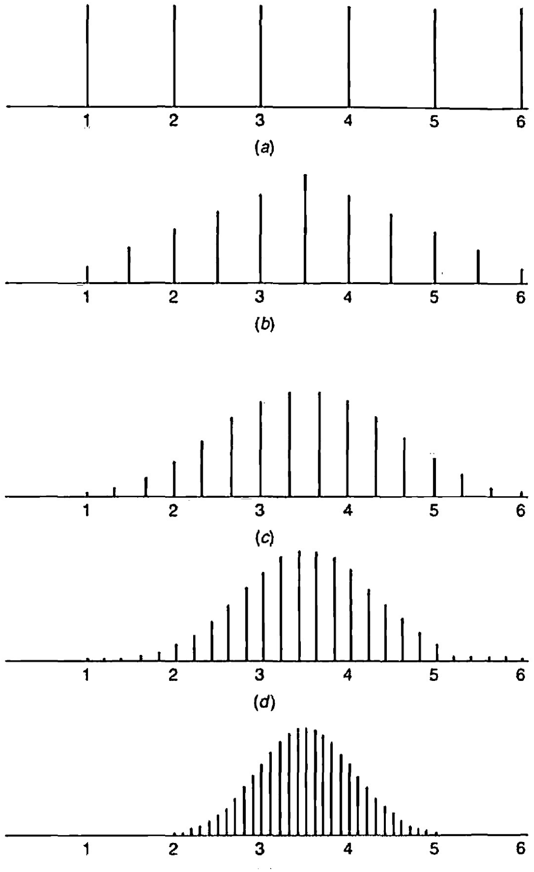


$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2 / 2\sigma^2}$$

$$N(\mu, \sigma^2)$$

- Symmetric about some central value
- Small deviations more common
- Central limit effect

Central limit effect



Average scores of (100 rolls)

One die

Two die

Three die

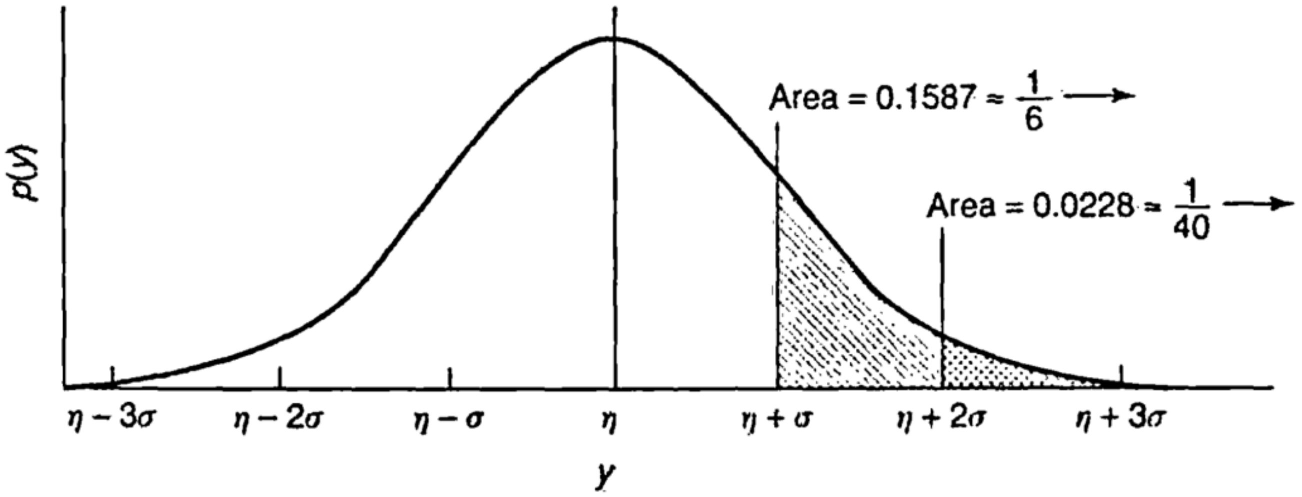
Five die

Ten die

In many experiments the error is an aggregate of a number of component errors and the distribution will tend to be “normal” in distribution

Probability

η and σ^2 fully characterise a distribution, $N(\eta, \sigma^2)$



<u>Probability density</u> is given by a point on the line p(y)

$p(y > \eta + \sigma) = \frac{1}{6}$ i.e. the area under the curve.

Often it is easier to express probability in terms of the standard deviate;

$$z = \frac{y - \eta}{\sigma}$$

$$z(0, 1)$$

i.e. z has a mean of 0 and variance, $s^2 = 1$. So that;

$$= p(y > \eta + \sigma)$$

$$= p(y - \eta > \sigma)$$

$$= p\left(\frac{y-\eta}{\sigma} > 1\right)$$

$$= p(z > 1)$$

(which can be found from tables)

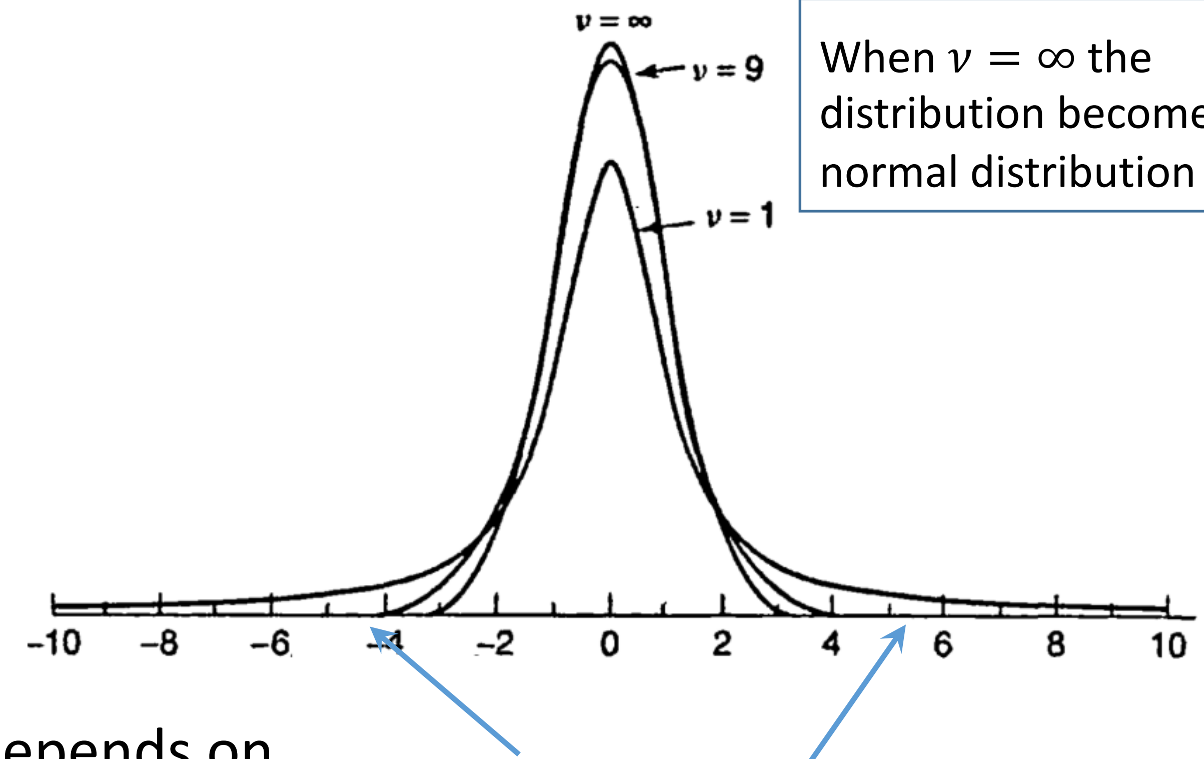
If σ is unknown (which is normally the case)

A substitution can be made for σ *using* s, the sample standard deviation;

$$z = \frac{y - \eta}{\sigma}$$

i.e.

$$t = \frac{y - \eta}{s}$$

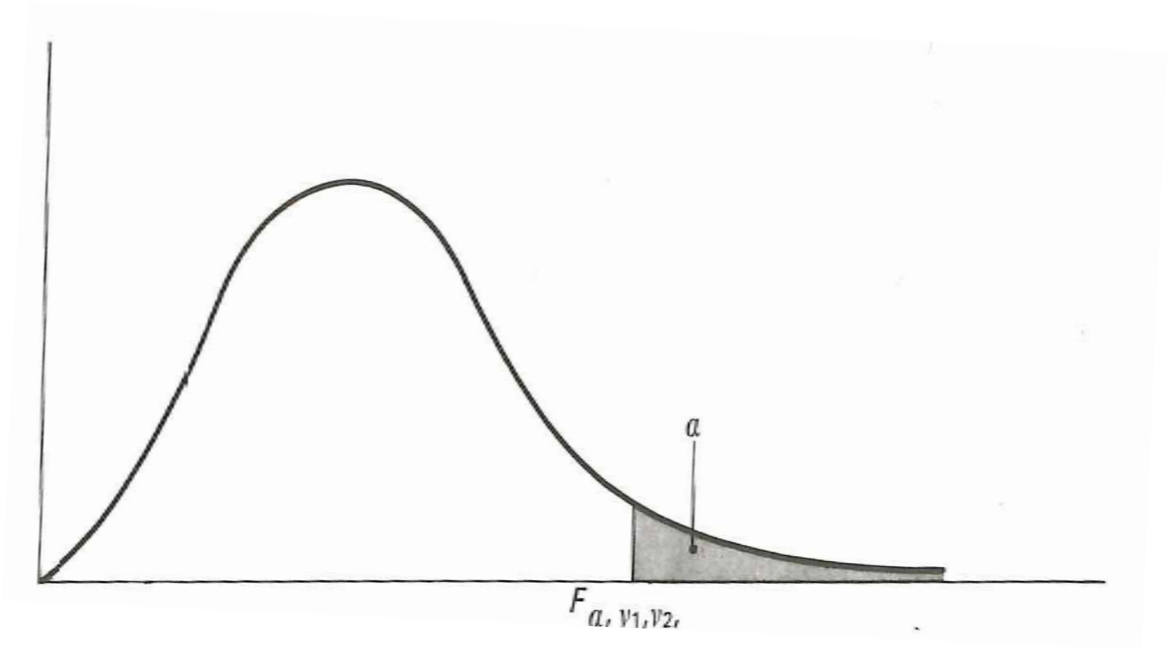


When $\nu = \infty$ the distribution becomes the normal distribution

Fatter tail compared with normal distribution

the ‘student’ or ‘t’ distribution depends on degrees of freedom available for estimation of s.

F-distribution



Can obtain ratio of two sample variances;

F statistic is ${s_1}^2/{s_2}^2$

F depends on the estimates and the dof of the variance estimates

Degrees of freedom of population variances;

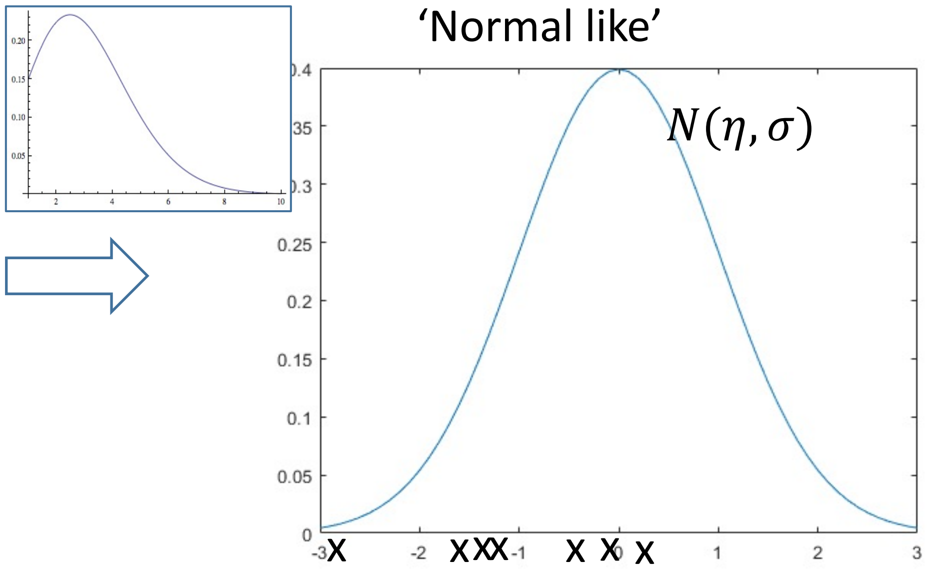
$\nu_1 = n_1 - 1$
$\nu_2 = n_2 - 1$

So F test statistic is designated

F_{ν_1,ν_2}

Standard Error of the Mean

- Take n random samples from a normal distribution with mean, η and standard deviation, σ. Calculate the sample η and σ. Repeat.
- The sample means will form a distribution with the same mean, η but a **smaller standard deviation** $\sigma/\sqrt{n}$ (the **standard error of the sample mean**).



For a sample of size n the sample mean is $\bar{x}$

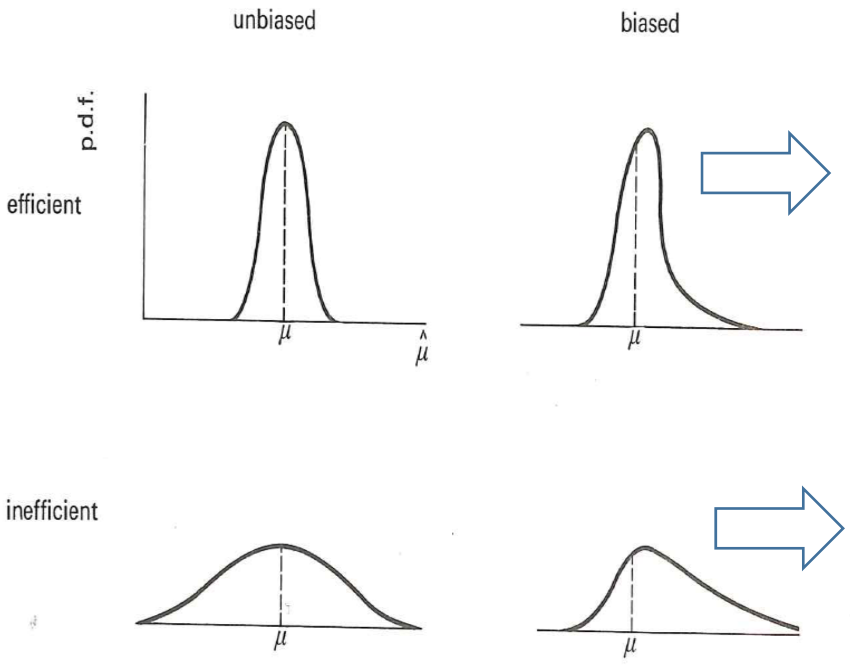
The standard error is an estimate of the standard deviation of the sample means for sample size n

$$\boxed{SE_m = \frac{\sigma}{\sqrt{n}}}$$

Intuitively it is a measure of how sample size affects the likely accuracy of the sample means relative to the population mean.

Bias and efficiency

- **Bias** – an estimator is said to be biased if the mean of its sampling distribution is not equal to the value it is estimating.
- **Efficiency** – an efficient unbiased estimator is the minimum variance unbiased estimator (MVUE).



$$bias(\hat{\mu}) = \mathrm{E}(\mu) - \mu$$

$$var(\hat{\mu}_1) < var(\hat{\mu}_2)$$

more efficient estimator $\hat{\mu}_1$

Significance testing

Testing a theory about the population

Null hypothesis	H_0	What we are testing ….
Alternative hypothesis	H_1	An alternative ….

- Test statistic
- Level of significance
- One tailed and two tailed tests

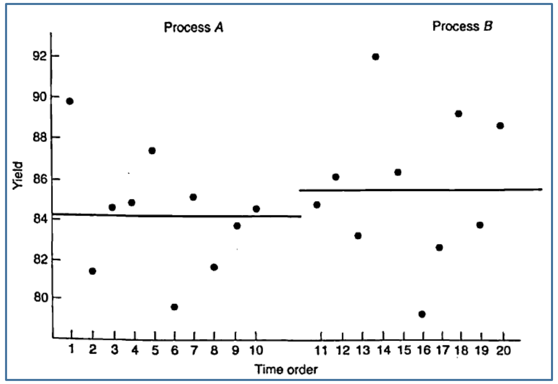
How to know if a treatment is significant?

Batch Process A

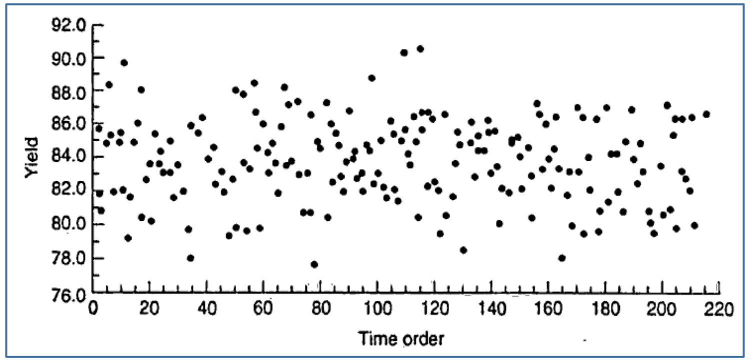
Batch Process B

$\bar{y}_a = 84.24$

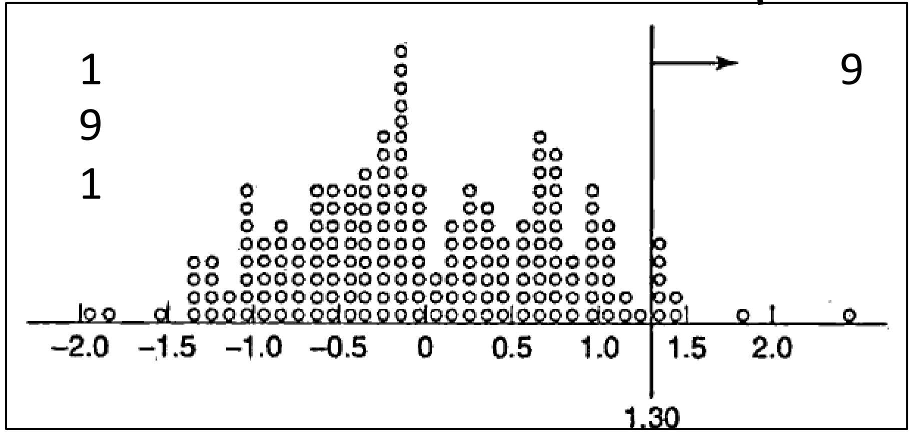
$\bar{y}_b = 85.54$



Previous yield data



$$\frac{9}{191} = 0.047$$



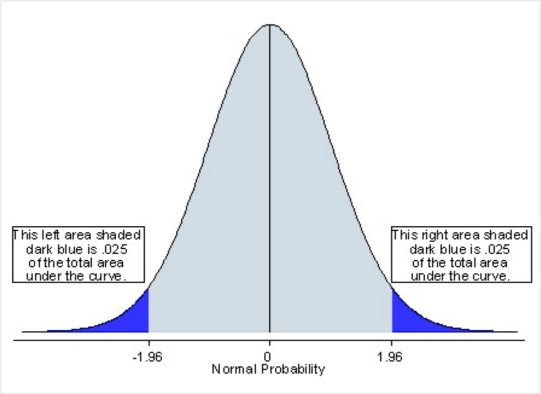
Difference in means of sequential batches

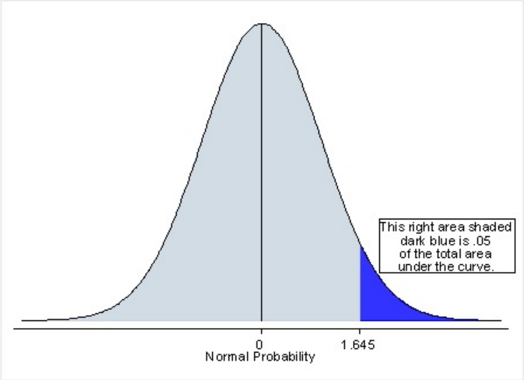
An example - composition of a chemical compound

- The iron content of a compound should be 12.1%. Tests on nine different samples are being used to examine this assumption.

- Null hypothesis i.e. there is no difference in the sample $(n = 9)$ mean
  - $H_0 : \mu = 12.1\%$

- Alternative hypothesis
  - $H_1 : \mu \neq 12.1\%$





Example (continued)

The analysis of nine samples gave the following values for % content of iron.

11.7	12.2	10.9	11.4	11.3	12.0	11.1	10.7	11.6

$$\bar{y} = 11.43$$
$$s^2 = 0.24$$
$$s = 0.49$$

$$t = \frac{(\bar{y} - \eta)}{s/\sqrt{n}}$$
$$= \frac{(11.43 - 12.1)}{0.49/\sqrt{9}} = -4.1$$

Standard deviation of the mean (estimate)

Degrees of freedom: eight because nine samples and one DoF used for population mean, $\bar{x}$

Example (continued)

1. Two tailed test
2. 5% level of significance
3. Eight degrees of freedom (from tables)

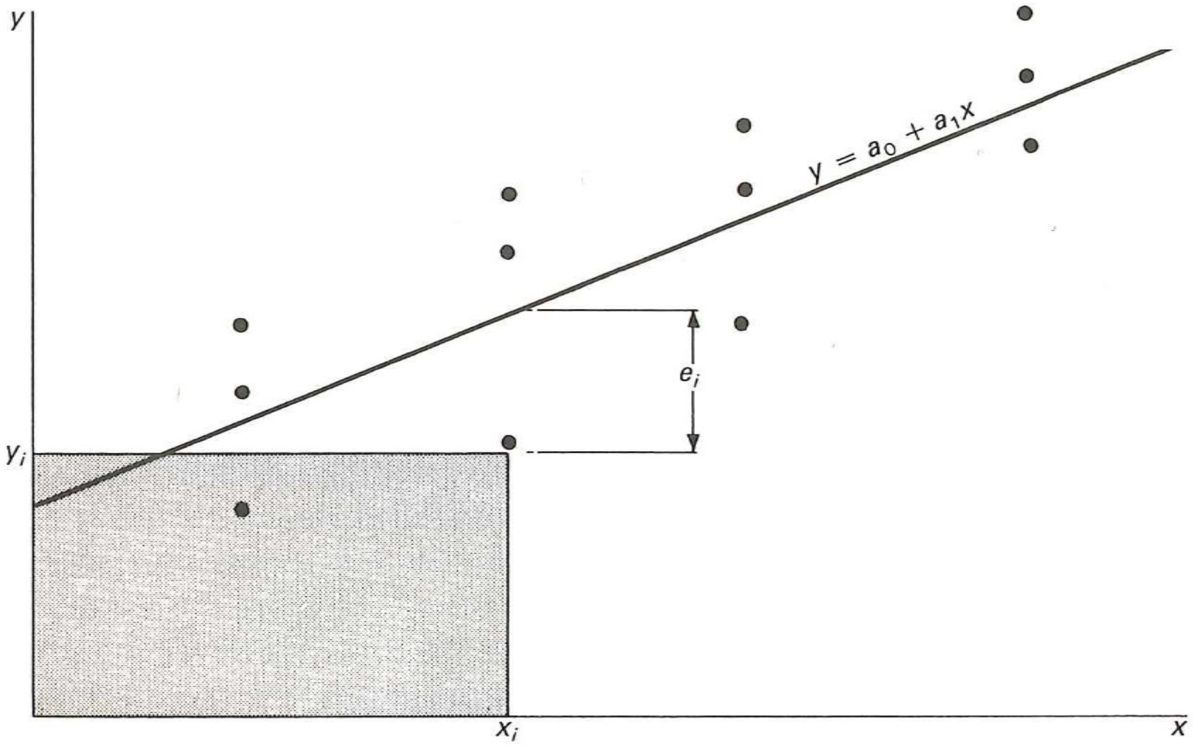
Test statistic is designated: $t_{0.025,8} = 2.31$ (from tables)

In fact, $t_{0.005,8} = 3.36$ (from tables)

So even at the 1% level, the result is significant.

Regression

Fitting a line or curve to the data in order to predict the mean value of the dependent variable for a given value of the controlled variable



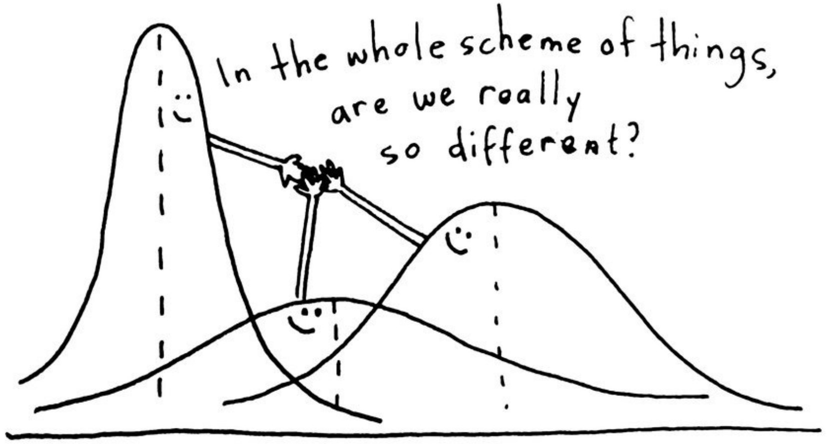
Model = $f(x, \beta)$

$$\min\left(\sum y_i - f(x_i, \beta)\right)$$

Analysing variance (ANOVA)

For comparing more than two entities

	A	B	C	D
	62	63	68	56
	60	67	66	62
	63	71	71	60
	59	64	67	61
	63	65	68	63
	59	66	68	64
Treatment avg	**61**	**66**	**68**	**61**
Overall avg	**64**	**64**	**64**	**64**



Deviation from overall average

Deviations **between** treatments

deviations **within** treatments

y_{ti}

A	B	C	D	$\boldsymbol{y_{ti} - \bar{y}}$	$\boldsymbol{\bar{y}_t - \bar{y}}$	$\boldsymbol{y_{ti} - \bar{y}_t}$
62	63	68	56	-2 -1 4 -8	-3 2 4 -3	1 -3 0 -5
60	67	66	62	-4 3 2 -2	-3 2 4 -3	-1 1 -2 1
63	71	71	60	-1 7 7 -4	-3 2 4 -3	2 5 3 -1
59	64	67	61	-5 0 3 -3	-3 2 4 -3	-2 -2 -1 0
63	65	68	63	-1 1 4 -1	-3 2 4 -3	2 -1 0 2
59	66	68	64	5 2 4 0	-3 2 4 -3	-2 0 0 3
			Sum of squares	**340**	**228**	**112**
			Degrees of freedom	**23**	**3**	**20**

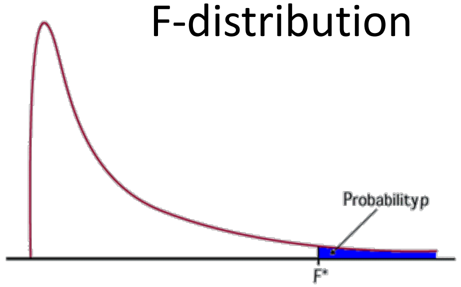
y_{ti} individual coagulation results

$\bar{y}_t$ treatment average

$\bar{y}$ overall average

ANOVA Table

Source of variation	Sum of squares	d.f.	$\boldsymbol{\frac{\chi^2}{\nu}}$
Between treatments	$\sum(\bar{y}_{\boldsymbol{t}} - \bar{\boldsymbol{y}})^2 = 228$	$n-1=3$	$\frac{\sum(\bar{y}_{\boldsymbol{t}} - \bar{\boldsymbol{y}})^2}{n-1} = 76$
Within treatments	$\sum(\boldsymbol{y}_{\boldsymbol{ti}} - \bar{\boldsymbol{y}}_{\boldsymbol{t}})^{\boldsymbol{2}} = 112$	$n-1=20$	$\frac{\sum(y_{ti} - \bar{y}_t)^2}{n-1} = 5.6$
Total about the overall average	**340**	**23**	



F-distribution

$$s_1^2/\sigma_1^2 \Big/ s_1^2/\sigma_1^2 = \frac{\sum(\bar{y}_t - \bar{y})^2}{n-1} \Big/ \frac{\sum(y_{ti} - \bar{y}_t)^2}{n-1} \sim F_{\nu_1,\nu_2}$$

$$F_{3,20} = 13.6$$

Significant at 0.001 i.e. we can be confident that treatments do result in different means, we can reject H_0