

# FE Review - Statistics

---

DR. GERRY KNAPP, P.E. (LOUISIANA)

A solid green horizontal bar spanning the width of the slide, located at the bottom.

# Statistics Topics on FE

---

Number of questions:

- **Chemical:** 4–6
- **Civil:** 2-3?  
(part of math)
- **Electrical:** 4–6
- **Environmental:** 4–6
- **Industrial:** 10–15
- **Mechanical:** 4–6
- **Other:** 6–9

<b>Measures of central tendencies and dispersions (e.g., mean, mode, standard deviation), measurement uncertainty, confidence intervals. IE, ChE: control limits</b>	All
<b>Basic laws of probability &amp; probability distributions (e.g., discrete, continuous, normal, binomial, empirical). EE &amp; IE: Conditional probability; IE: permutations &amp; combinations, sets</b>	All
<b>Expected value (weighted average)</b>	All
<b>Regression &amp; goodness of fit (linear, multiple, curve fitting, correlation coefficient, R<sup>2</sup>, least squares). IE: Residual analysis</b>	All
<b>Sampling &amp; hypothesis testing (central limit, sampling distributions, standard error, normal, t, chi-square, types of error, sample size, outlier testing, significance)</b>	ChE, EE, IE, Other
<b>Analysis of Variance (ANOVA). IE: Factorial Design</b>	Che, IE, EE

# Definitions

---

## Population

- Collection of all measurements of interest to the statistical analysis.
- May be of infinite size.

## Sample

- Subset of data extracted from population (usually randomly)

## Parameter

- Characteristic of *population*

## Statistic

- Characteristic of *sample*

**Variation:** changes, differences, uncertainty, randomness in a population

**Probability:** two (equivalent) perspectives:

- Percentage of a population meeting some criteria
- Chance of any one observation from the population meeting some criteria

# Measures of central tendencies and dispersions

---



# Common Statistics

STATISTIC	SYMBOL	FORMULA	Population Symbol
Mean	$\bar{X}$	$\frac{\sum X_i}{n}$	$\mu$
Median	m	Order data smallest to largest If n is odd, middle value; if n is even, midpoint between middle 2 values	-
Mode	-	Value which repeats most often (its possible to have zero or multiple modes)	-
Range	R	$\text{Max}(X_i) - \text{Min}(X_i)$	-
Variance	$s^2$	$\frac{\sum(X_i - \bar{X})^2}{n-1}$ for population $\sigma^2$ , divide by n	$\sigma^2$
Standard Deviation	s	Sqrt( $s^2$ )	$\sigma$

**Pg. 63, Reference book ver. 10**

Others: Coefficient of variation (CV), geometric mean, root mean square (RMS)

# Examples

---

What is the sample variance of the following numbers?

2,4,6,8,10,12,14

- a) 4.32
- b) 5.29
- c) 8.00
- d) 18.70

SOLN:

First find mean:  $\bar{X} = \frac{\sum X_i}{n} = \frac{2+4+6+8+10+12+14}{7} = 8$

Then  $S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{(2-8)^2 + (4-8)^2 + (6-8)^2 + (8-8)^2 + (10-8)^2 + (12-8)^2 + (14-8)^2}{7-1} = 18.67$  (d)

Double checks: should be positive! Asking for S or S<sup>2</sup>? Or  $\sigma$  or  $\sigma^2$ ?

---

What is the median of the following data?

10, 7, 12, 4, 6, 8

- a) 4
- b) 7.5
- c) 8
- d) 10

SOLN:

Put in ascending (sort) order: 4,6,7,8,10,12

If n odd, middle item; if even, average of middle two items

$$m = (7+8) / 2 = 7.5 \text{ (b)}$$

Follow-up: What is the mode of this data?



# Confidence Intervals

---

Provide a bracket on the location of the true population mean given a sample  $\bar{X}$  and S. Based on Student-T distribution

- If standard deviation  $\sigma$  is unknown:  $\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2, n-1} \left( \frac{S}{\sqrt{n}} \right)$  ← Standard error
- If standard deviation  $\sigma$  is known:  $\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

May also be stated as:

$$\bar{X} \pm \left( t_{\alpha/2} \frac{S}{\sqrt{n}} \right) \leftarrow \text{Half width}$$

$\alpha$ =significance level (type 1 error) = 1 – confidence level

Z values: use table on page 44, t-values: use student-t table pg. 46

IEs: also refresh on formulas for CI on difference means and on variance



# Examples

You collect 10 observations from an experiment. The sample average is 14.0, and the standard deviation is 5.8. The 90% confidence interval on the mean is:

- a)  $11.57 < \mu < 16.43$
- b)  $10.68 < \mu < 17.32$
- c)  $8.2 < \mu < 19.8$
- d)  $8.78 < \mu < 19.22$

SOLN:

Identify what you have:  $n=10$ ,  $\bar{X} = 14.0$ ,  $S=5.8$ , confidence=0.9 and population variance  $\sigma$  unknown

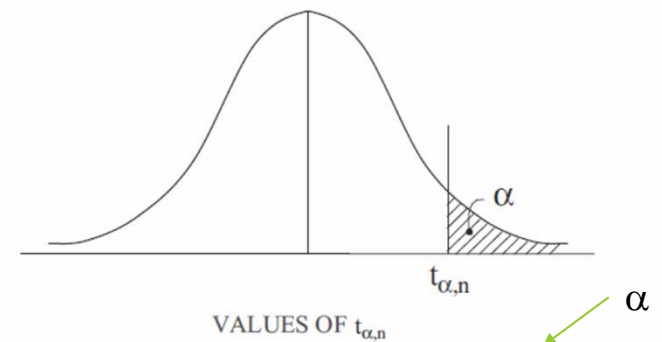
Use t, not Z.  $\alpha = 1 - 0.9 = 0.1$ .

Find  $t_{0.1/2=0.05, n=10} = 1.812$

$$\text{Then: } 14.0 - 1.812 \frac{5.8}{\sqrt{10}} < \mu < 14.0 + 1.812 \frac{5.8}{\sqrt{10}}$$

$$\Rightarrow 10.68 < \mu < 17.32 \text{ (b)}$$

STUDENT'S *t*-DISTRIBUTION



n	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	n
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	1
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	2
3	0.765	0.978	1.350	1.638	2.353	3.182	4.541	5.841	3
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	4
5	<b>0.727</b>	<b>0.920</b>	<b>1.156</b>	<b>1.476</b>	<b>2.015</b>	<b>2.571</b>	<b>3.365</b>	<b>4.032</b>	<b>5</b>
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	6
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	7
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	8
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	9
10	<b>0.700</b>	<b>0.879</b>	<b>1.093</b>	<b>1.372</b>	<b>1.812</b>	<b>2.228</b>	<b>2.764</b>	<b>3.169</b>	<b>10</b>

---

A random sample is selected from a normal distribution with variance 46.12336. If the width of a 95% confidence interval about the sample mean is 4.5, what is the size of the sample?

SOLN:

Variance is known, so will use Z rather than t:  $Z_{0.05/2} = 1.96$

If width of CI is 4.5, half width is  $4.5/2 = 2.25$

CI halfwidth =  $Z_{\alpha/2} * \sigma / \sqrt{n}$ , so have  $2.25 = 1.96 * \sqrt{46.12336} / \sqrt{n}$

Solve for n:  $\sqrt{n} = 1.96 * \sqrt{46.12336} / 2.25$

$$n = (1.96^2) * 46.12336 / (2.25^2) = \mathbf{35}$$

	<i>Values of <math>Z_{\alpha/2}</math></i>
CI	$Z_{\alpha/2}$
80%	1.2816
90%	1.6449
95%	1.9600
96%	2.0537
98%	2.3263
99%	2.5758

# Control Limits/Charts

## Average and Range Charts

$n$	$A_2$	$D_3$	$D_4$
2	1.880	0	3.268
3	1.023	0	2.574
4	0.729	0	2.282
5	0.577	0	2.114
6	0.483	0	2.004
7	0.419	0.076	1.924
8	0.373	0.136	1.864
9	0.337	0.184	1.816
10	0.308	0.223	1.777

$X_i$  = an individual observation

$n$  = the sample size of a group

$k$  = the number of groups

$R$  = (range) the difference between the largest and smallest observations in a sample of size  $n$ .

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \dots + \bar{X}_k}{k}$$

$$\bar{R} = \frac{R_1 + R_2 + \dots + R_k}{k}$$

The  $R$  Chart formulas are:

$$CL_R = \bar{R}$$

$$UCL_R = D_4 \bar{R}$$

$$LCL_R = D_3 \bar{R}$$

The  $\bar{X}$  Chart formulas are:

$$CL_{\bar{X}} = \bar{\bar{X}}$$

$$UCL_{\bar{X}} = \bar{\bar{X}} + A_2 \bar{R}$$

$$LCL_{\bar{X}} = \bar{\bar{X}} - A_2 \bar{R}$$

## Standard Deviation Charts

$n$	$A_3$	$B_3$	$B_4$
2	2.659	0	3.267
3	1.954	0	2.568
4	1.628	0	2.266
5	1.427	0	2.089
6	1.287	0.030	1.970
7	1.182	0.119	1.882
8	1.099	0.185	1.815
9	1.032	0.239	1.761
10	0.975	0.284	1.716

$$UCL_X = \bar{\bar{X}} + A_3 \bar{S}$$

$$CL_X = \bar{\bar{X}}$$

$$LCL_X = \bar{\bar{X}} - A_3 \bar{S}$$

$$UCL_S = B_4 \bar{S}$$

$$CL_S = \bar{S}$$

$$LCL_S = B_3 \bar{S}$$

**Pg 82-83 in Ref.  
handbook ver. 10**

# Examples

A process measurement historically has  $\mu=10.0$  and  $\sigma=4.2$ . Sample size is 8. What are the 3 sigma control limits for this process?

SOLN:

- $UCL = 10 + 3*4.2/\text{sqrt}(8) = 14.45$
- $CL = 10.0$
- $LCL = 10 - 3*4.2/\text{sqrt}(8) = 5.55$

Data is initially collected to create a control chart:  $\bar{X}=14.0$ ,  $\bar{R}= 3.5$ . Sample size  $n$  is 7. What are the  $\bar{X}$  chart control limits for this process?

SOLN:

- Given:  $\bar{X}=14.0$ ,  $\bar{R}= 3.5$ ,  $n=7$
- $A_2$  for  $n=7$  is 0.419
- $CL = \bar{X} = 14.0$
- $UCL = \bar{X} + A_2\bar{R} = 14.0 + 0.419*3.5 = 15.47$
- $LCL = \bar{X} - A_2\bar{R} = 14.0 - 0.419*3.5 = 12.53$

Average and Range Charts

$n$	$A_2$	$D_3$	$D_4$
2	1.880	0	3.268
3	1.023	0	2.574
4	0.729	0	2.282
5	0.577	0	2.114
6	0.483	0	2.004
7	0.419	0.076	1.924
8	0.373	0.136	1.864
9	0.337	0.184	1.816
10	0.308	0.223	1.777

$X_i$  = an individual observation

$n$  = the sample size of a group

$k$  = the number of groups

$R$  = (range) the difference between the largest and smallest observations in a sample of size  $n$ .

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \dots + \bar{X}_k}{k}$$

$$\bar{R} = \frac{R_1 + R_2 + \dots + R_k}{k}$$

The  $R$  Chart formulas are:

$$CL_R = \bar{R}$$

$$UCL_R = D_4\bar{R}$$

$$LCL_R = D_3\bar{R}$$

The  $\bar{X}$  Chart formulas are:

$$CL_X = \bar{X}$$

$$UCL_X = \bar{X} + A_2\bar{R}$$

$$LCL_X = \bar{X} - A_2\bar{R}$$

# Measurement Uncertainty

---

See also expectation later this review and "Combinations of Random Variables" in the FE Reference Handbook for formulas for mean and variance of linear combinations of variables

Given a desired state or measurement  $y$ , which is a function of different measured or available states  $x_i$ :

$$y = f(x_1, x_2, \dots, x_n)$$

Given the individual states  $x_i$  and their standard deviations  $\sigma_{x_i}$ , and assuming that the different  $x_i$  are uncorrelated, the Kline-McClintock equation can be used to compute the expected standard uncertainty of  $y$  ( $\sigma_y$ ) is:

$$\sigma_y = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \sigma_{x_2}^2 + \dots + \left(\frac{\partial f}{\partial x_n}\right)^2 \sigma_{x_n}^2}$$

# Example

---

Given function  $y = x_1^2 - x_2^2 x_1$  where measured values  $x_1 = 4 \pm 0.1$  and  $x_2 = 16 \pm 0.5$  ( $\pm$  values indicating standard deviations of each), what is  $y$  and the uncertainty of  $y$ ,  $\sigma_y$ ?

SOLN:

$$y = 4^2 - 16^2 * 4 = -1,008$$

$$\sigma_y = \sqrt{\left(\frac{dy}{dx_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{dy}{dx_2}\right)^2 \sigma_{x_2}^2}$$

$$= \sqrt{(2x_1 - x_2^2)^2 * 0.1^2 + (0 - 2x_2 * x_1)^2 * 0.5^2}$$

$$= \sqrt{(2 * 4 - 16^2)^2 * 0.1^2 + (0 - 2 * 16 * 4)^2 * 0.5^2}$$

$$= 68.6$$

$$y = -1,008 \pm 68.6$$

# Basic laws of probability & probability distributions

---

# Equally Likely Outcomes

---

If an experiment can result in  $N$  equally likely outcomes, and  $n$  of those outcomes make up event  $A$ , the probability of event  $A$  is


$$P(A) = \frac{n}{N}$$

Example: Single die toss; let  $A = \{2, 4, 6\}$ , so

- $P(A) = 3/6 = 1/2 = 0.5$

Example: Single coin toss, let  $A = \{H\}$

- $P(A) = 1/2 = 0.5$





# Probability characteristics

---

Probability of an event is a number between zero and one,  $0 \leq P(A) \leq 1$ .

$P(A) = 1$ , means event  $A$  occurs with certainty.

$P(A) = 0$ , means event  $A$  is impossible.

$P(A') = 1 - P(A)$ , that is, the probability of not  $A$  is simply 1 minus the chance of  $A$  happening.

## Property 2. Law of Total Probability

$P(A + B) = P(A) + P(B) - P(A, B)$ , where

$P(A + B)$  = the probability that either  $A$  or  $B$  occur alone or that both occur together,

$P(A)$  = the probability that  $A$  occurs,

$P(B)$  = the probability that  $B$  occurs, and

$P(A, B)$  = the probability that both  $A$  and  $B$  occur simultaneously.

If the events are mutually exclusive ( $A_1, A_2, \dots$ )

$$P(A_1 + A_2 + \dots) = P(A_1) + P(A_2) + \dots$$

If the events  $A_1..A^n$  are a partition of  $S$ ,

$$\begin{aligned} P(A_1 + A_2 + \dots + A_n) \\ = P(A_1) + P(A_2) + \dots + P(A_n) = 1 \end{aligned}$$

If  $A$  and  $A'$  are complementary,  $P(A) + P(A') = 1$

# Mutually Exclusive vs. Independent

---

## MUTUALLY EXCLUSIVE

Events cannot occur at the same time

$$P(A, B) = 0$$

$$P(A + B) = P(A) + P(B)$$

## INDEPENDENT

One event occurring does not affect the likelihood of the other event

Events could occur simultaneously, but they have no effect on the likelihood of the other occurring

$$P(A, B) = P(A) * P(B)$$

$$P(A + B) = P(A) + P(B) - P(A \cap B)$$

# Examples

---

Two students are working independently on a problem. Their respective probabilities of solving the problem are  $1/3$  and  $3/4$ . What is the probability that at least one of them will solve the problem?

- a)  $1/2$
- b)  $5/8$
- c)  $2/3$
- d)  $5/6$

SOLN:

Given  $P(A)=1/3$ ,  $P(B)=3/4$ . Looking for  $P(A+B)$ .

$$= P(A) + P(B) - P(A,B) = P(A) + P(B) - P(A)*P(B) \quad (\text{since A and B independent})$$

$$= 1/3 + 3/4 - (1/3)(3/4) = 5/6 \quad (\text{d})$$

---


A coin is flipped and a 6-sided die thrown. What is the probability of getting heads and a 5?

- a)  $1/2$
- b)  $1/3$
- c)  $1/6$
- d)  $1/12$

SOLN:

12 possible outcomes ( $2*6$ ), all equally likely. Only one produces the outcome of interest.

Pr =  $1/12$  (d)



---

The only possible outcomes of an experiment are A,B,C,or D, all mutually exclusive.  $P(A)=0.3$ ,  $P(B)=0.2$ ,  $P(C)=0.4$ .  $P(D)$ ?

- a) 0.3
- b) 0.5
- c) 0.1
- d) 0.7

SOLN: Since a partition, sum must add to 1. So  $P(D) = 1 - 0.3 - 0.2 - 0.4 = 0.1$  (c)

Same question, but what is  $P(A,C)$ ?

SOLN: 0

$P(A+C)$ ?

SOLN:  $0.3+0.4 = 0.7$



# Conditional Probability & Bayes Theorem

---

Probability that event B occurs given that event A has already occurred.

- The occurrence of A shrinks the sample space from S to A.

$$\text{Formula: } P(B|A) = \frac{P(B,A)}{P(A)}$$

$$\text{Bayes Theorem: } \Pr(A|B) = \frac{P(B|A)*P(A)}{P(B)}$$

$$\text{If have partition } A_1, A_2, \dots, A_j \text{ then: } \Pr(A_j|B) = \frac{P(B|A_j)*P(A_j)}{\sum_i P(B|A_i)P(A_i)}$$

---

An assembly is composed of two plastic parts A and B, which are created at the same time in a plastic injection mold. Given part A is defective, there is an 80% chance part B is also defective. If part A is defective 10% of the time, what is P(A defective and B defective)?

SOLN:

Given:  $P(AD)=0.1$ ,  $P(BD|AD)=0.8$

From Conditional Probability:  $P(BD|AD) = P(AD,BD) / P(AD)$

Rewriting:  $P(AD,BD) = P(AD)*P(BD|AD) = 0.1*0.8 = 0.08$

---

A rare disease exists in which only 1 in 500 are affected. A test for the disease has the following accuracy:

- 95% correct positive result (patient actually has the disease)
- 1% false positive (patient does not have the disease but tests positive)

If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

SOLN:

Define events  $D$ =has disease and  $POS$ =positive result.

Given:  $P(D) = 1/500 = 0.002$ ,  $P(D') = 1 - 0.002 = 0.998$ ,  $P(POS|D) = 0.95$ ,  $P(POS|D') = 0.01$

$$P(D|POS)? = \frac{P(POS|D)*P(D)}{P(POS|D)*P(D)+P(POS|D')*P(D')} = \frac{0.95*0.002}{0.95*0.002+0.01*0.998} = 0.1599$$



# Probability Distributions

A **random variable (rv)** is a variable whose observed value is determined by chance and is therefore unknown before the outcome of an experiment

- Can characterize the likelihood of values based on the rv's probability distribution (more on this momentarily)
- **Discrete random variable:** An rv whose set of possible outcomes (values) is countable
- **Continuous random variable:** An rv that takes on any value over a continuous scale

The pattern of variation in a rv is called its **distribution**

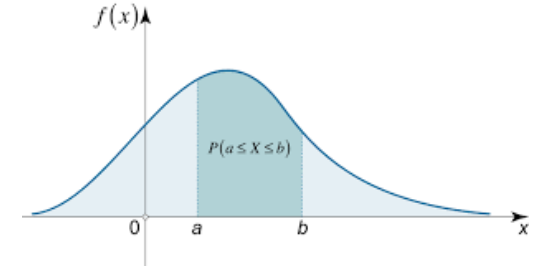
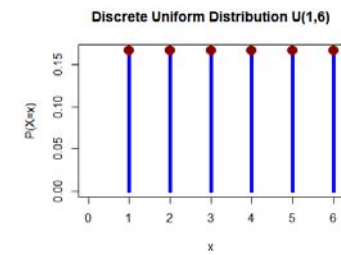
- A distribution defines how likely each possible value of the rv is to occur

Reference guide has:

- PDF formula and CDF table for **Binomial** distribution [a discrete rv]
- CDF **Normal** distribution table [continuous rv]. Requires unit normal rv:  $Z = (X - \mu) / \sigma$ 
  - note unit normal table says x in left column but really means z values!
  - Note: Z is the number of standard deviations of X above or below the mean  $\mu$
- Upper tail (1-CDF) tables for **Student-t**, **F**, **Chi-Square** distributions [continuous rv]

Some others are easy to manually work out so may appear:

- **Empirical** (typically discrete)
- **Uniform** distribution (either discrete or continuous rv)
- **Triangle** distribution (continuous rv)



# Distribution Functions

	Discrete	Continuous
<b>f(x)</b> : Probability distribution function ( <b>PDF</b> ) (aka probability mass function ( <b>PMF</b> ) for discrete only)	$=P(X = x)$	$=\frac{\delta F(x)}{\delta x}$  (rate, not a probability)
<b>F(x)</b> : Cumulative distribution function ( <b>CDF</b> )	$=P(X \leq x)$  $=\sum_{X < x} f(X)$	$=P(X \leq x)$  $=\int_{X < x} f(X) dX$

# Binomial table

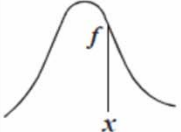
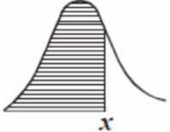
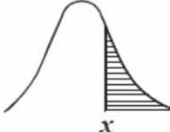
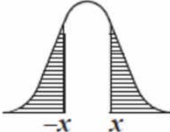
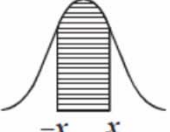
---

**Cumulative Binomial Probabilities  $P(X \leq x)$**

		<i>P</i>										
<i>n</i>	<i>x</i>	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
1	0	0.9000	0.8000	0.7000	0.6000	0.5000	0.4000	0.3000	0.2000	0.1000	0.0500	0.0100
2	0	0.8100	0.6400	0.4900	0.3600	0.2500	0.1600	0.0900	0.0400	0.0100	0.0025	0.0001
	1	0.9900	0.9600	0.9100	0.8400	0.7500	0.6400	0.5100	0.3600	0.1900	0.0975	0.0199
3	0	0.7290	0.5120	0.3430	0.2160	0.1250	0.0640	0.0270	0.0080	0.0010	0.0001	0.0000
	1	0.9720	0.8960	0.7840	0.6480	0.5000	0.3520	0.2160	0.1040	0.0280	0.0073	0.0003
	2	0.9990	0.9920	0.9730	0.9360	0.8750	0.7840	0.6570	0.4880	0.2710	0.1426	0.0297
4	0	0.6561	0.4096	0.2401	0.1296	0.0625	0.0256	0.0081	0.0016	0.0001	0.0000	0.0000
	1	0.9477	0.8192	0.6517	0.4752	0.3125	0.1792	0.0837	0.0272	0.0037	0.0005	0.0000
	2	0.9963	0.9728	0.9163	0.8208	0.6875	0.5248	0.3483	0.1808	0.0523	0.0140	0.0006
	3	0.9999	0.9984	0.9919	0.9744	0.9375	0.8704	0.7599	0.5904	0.3439	0.1855	0.0394
5	0	0.5905	0.3277	0.1681	0.0778	0.0312	0.0102	0.0024	0.0003	0.0000	0.0000	0.0000

# Normal Table

Unit Normal Distribution ( $\mu = 0, \sigma = 1$ )

					
$x$	$f(x)$	$F(x)$	$R(x)$	$2R(x)$	$W(x)$
0.0	0.3989	0.5000	0.5000	1.0000	0.0000
0.1	0.3970	0.5398	0.4602	0.9203	0.0797
0.2	0.3910	0.5793	0.4207	0.8415	0.1585
0.3	0.3814	0.6179	0.3821	0.7642	0.2358
0.4	0.3683	0.6554	0.3446	0.6892	0.3108
0.5	0.3521	0.6915	0.3085	0.6171	0.3829
0.6	0.3332	0.7257	0.2743	0.5485	0.4515
0.7	0.3123	0.7580	0.2420	0.4839	0.5161
0.8	0.2897	0.7881	0.2119	0.4237	0.5763
0.9	0.2661	0.8159	0.1841	0.3681	0.6319

$f(x)$ : PDF  
 $F(x)$ : CDF,  $P(X < x)$   
 $R(x)$ : 1-CDF,  $P(X > x)$   
 (aka, reliability function)  
 $W(x)$ :  $F(x) - F(-x)$

What if you have a negative  $x$ ?  
 Normal distribution is symmetric, so  
 $F(x) = R(-x) = 1 - F(-x)$

# Examples

---

A rocket launch for a particular model of rocket has a 5% chance of failure on each launch. What is the probability of at least one failure in the next 10 launches?

SOLN:

Binomial distribution.  $n=10$ ,  $p=0.05$ . Note: only have equation for  $f(x)$ , not  $F(x)$  (pg. 42)

$$P(X \geq 1) = 1 - P(X=0) = 1 - P_{10}(0) = 1 - \frac{10!}{0!(10-0)!} * 0.05^0 * (1 - 0.05)^{10-0} = 0.40$$

If  $p$  had been large, you could also have used the Binomial CDF table for  $1 - P(X \leq 0)$  (but the smallest  $p$  in the table is 0.1)

---

The water content of soil from a borrow site is normally distributed with a mean of 14.2% and standard deviation of 2.2%. What is the probability that a sample taken from the site will have a water content above 16.4% or below 12%?

- a) 0.13
- b) 0.25
- c) 0.32
- d) 0.42

SOLN:

Normally distributed – use standard (unit) normal table (pg. 76).

Looking for  $P(X > 16) + P(X < 12)$ .

Convert X to Z:  $P(Z > (16.5-14.2)/2.2) + P(Z < (12-14.2)/2.2) = P(Z > 1) + P(Z < -1) = 1 - W(1) = 0.32$  (c)

---

Continuous random variable X varies between 0 and 10 with constant  $f(x)=0.1$ . What is  $P(X>3)$ ?

SOLN:

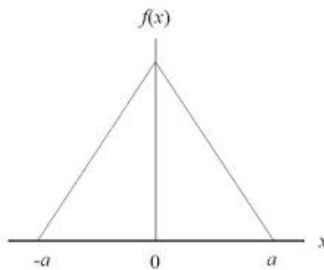
This is continuous uniform distribution:  $f(x)$  is basically a rectangle. Probabilities (area under  $f(x)$  curve) can be determined by proportion

$$P(X>3) = (10-3)/(10-0) = 0.7$$

---

For the probability density function shown, where  $a=0.5$ , what is the probability of the random variable  $x$  being greater than 0.25?

- a) 0.125
- b) 0.22
- c) 0.25
- d) 0.33



SOLN:

This is a triangular distribution.  $P(x > 0.25)$  is equal to the area under the curve to the right of 0.25.

The area under the  $f(x)$  curve must equal 1. Therefore at its peak the height of  $f(x)$  must be 2.

The equation for the curve would then be  $2-4x$

Integrating from 0.25 to 0.5:

$$P(x > 0.25) = \int_{0.25}^{0.5} (2 - 4x) dx = 2 \left[ x - x^2 \right]_{0.25}^{0.5} = 0.5 - 0.375 = 0.125 \text{ (a)}$$



# Counting Outcomes – Permutations, Combinations

---

A *permutation* is a particular sequence of a given set of objects. A *combination* is the set itself without reference to order.

1. The number of different *permutations* of  $n$  distinct objects taken  $r$  at a time is

$$P(n, r) = \frac{n!}{(n - r)!}$$

$nPr$  is an alternative notation for  $P(n, r)$

2. The number of different *combinations* of  $n$  distinct objects taken  $r$  at a time is

$$C(n, r) = \frac{P(n, r)}{r!} = \frac{n!}{[r!(n - r)!]}$$

$nCr$  and  $\binom{n}{r}$  are alternative notations for  $C(n, r)$

3. The number of different *permutations* of  $n$  objects taken  $n$  at a time, given that  $n_i$  are of type  $i$ , where  $i = 1, 2, \dots, k$  and  $\sum n_i = n$ , is

$$P(n; n_1, n_2, \dots, n_k) = \frac{n!}{n_1! n_2! \dots n_k!}$$

**Multiplication Rule:** Applies if events A and B are independent  
#outcomes = #outcomes(A)\*#outcomes(B)

# Examples

---

10 buildings to be inspected today and there are 10 inspectors. How many different ways could the buildings and inspectors be paired up?

SOLN:

- Multiplication rule:  $10 \times 10 = 100$

A sound wall is being built next to a freeway. 7 different colored bricks are being used, and the design stipulates that every six bricks (across) must be of a different color. How many ways (orders) could the six brick colors be placed?

SOLN:

- Sequence (ordering) indicates permutations
- ${}_7P_6 = 7! / (7-6)! = 5,040$

# Sets

---

## De Morgan's Law

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$

## Associative Law

$$A \cup (B \cup C) = (A \cup B) \cup C$$

$$A \cap (B \cap C) = (A \cap B) \cap C$$

## Distributive Law

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

# Examples

---

$S = \{1..10\}$ ,  $A = \{1,3\}$ ,  $B = \{1,4,5\}$ .

What is  $(A,B)$ ?

SOLN: intersection,  $= \{1\}$

What is  $(A+B)$ ?

SOLN: union,  $= \{1,3,4,5\}$

What is the set  $(A,B)'$ ?

SOLN:

- Apply De Morgans:  $(A,B)' = (A'+B') = \{2,4,5,6,7,8,9,10\} + \{2,3,6,7,8,9,10\} = \{2,3,4,5,6,7,8,9,10\}$

Expected value  
(Expectation)

---

# Expected value

---

Let  $X$  be a rv with probability distribution  $f(x)$

- If  $f(x)$  is discrete:
  - mean or expected value of  $X$  is:  $\mu = E[X] = \sum_x x * f(x)$
  - $E[g(X)] = \sum_x g(x) * f(x)$
- If  $f(x)$  is continuous:
  - $\mu = E[X] = \int_{-\infty}^{\infty} x * f(x) dx$
  - $E[g(X)] = \int_{-\infty}^{\infty} g(x) * f(x) dx$
- $E[X]$  is the centroid or mean of the pdf (central moment: center of gravity)
- $E[X]$  is in same units as  $X$ . It is NOT a probability.
- See also formulas for variance of  $X$ ,  $g(X)$ , and for mean and variance of sums of independent random variables on **page 65-66 of reference manual**

Simplifications:

For constant  $a$ ,

- $E[a] = a$
- $\sigma_a^2 = 0$

Rules:

$$E[aX + b] = aE[X] + b$$

$$\sigma_{aX+b}^2 = a^2 \sigma_x^2$$

$$E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)]$$

# Example

---

Let  $X$  be number of successful wells out of 5 drilled. The probability distribution for  $X$  follows.  $\$(3+4X)$  million in profit is earned. What is the expected profit for the 5 wells?

x:	0	1	2	3	4	5
f(x):	0.1	0.4	0.3	0.1	0.05	0.05

SOLN:

Simplifying:  $E[3+4x] = 3 + 4E[x]$

First:  $E[x] = 0*0.1+1*0.4+2*0.3+3*0.1+4*0.05+5*0.05 = \$1.75$  million

Then:  $E[3+4x] = 3 + 4E[x] = 3+4*1.75 = \$10$  million

# Regression & goodness of fit

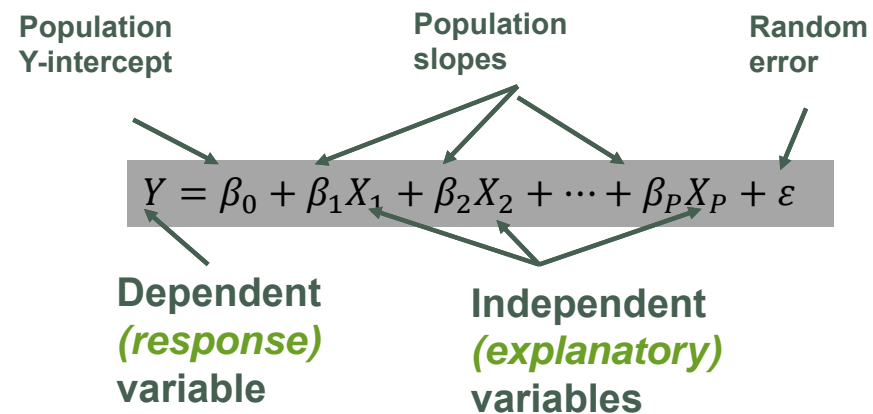
---



# Linear Regression Model

---

Relationship between one dependent & one or more independent variables is a linear function



# Fitting the Linear Regression Model (Least Squares) – one variable

$y = \hat{a} + \hat{b}x$ , where

$y$ -intercept:  $\hat{a} = \bar{y} - \hat{b}\bar{x}$ ,

and slope:  $\hat{b} = S_{xy}/S_{xx}$ ,

$$S_{xy} = \sum_{i=1}^n x_i y_i - (1/n) \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right),$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - (1/n) \left( \sum_{i=1}^n x_i \right)^2,$$

$n$  = sample size,

$$\bar{y} = (1/n) \left( \sum_{i=1}^n y_i \right), \text{ and}$$

$$\bar{x} = (1/n) \left( \sum_{i=1}^n x_i \right).$$

**Pg 69-70 ref  
handbook**

**Standard Error of Estimate**

$$S_e^2 = \frac{S_{xx} S_{yy} - S_{xy}^2}{S_{xx}(n-2)} = MSE, \text{ where}$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - (1/n) \left( \sum_{i=1}^n y_i \right)^2$$

**Confidence Interval for  $a$**

$$\hat{a} \pm t_{\alpha/2, n-2} \sqrt{\left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) MSE}$$

If CI contains 0,  $a$  is not considered significant

**Confidence Interval for  $b$**

$$\hat{b} \pm t_{\alpha/2, n-2} \sqrt{\frac{MSE}{S_{xx}}}$$

If CI contains 0,  $x$  is not considered significant

**Sample Correlation Coefficient**

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$R$  in -1 to 1

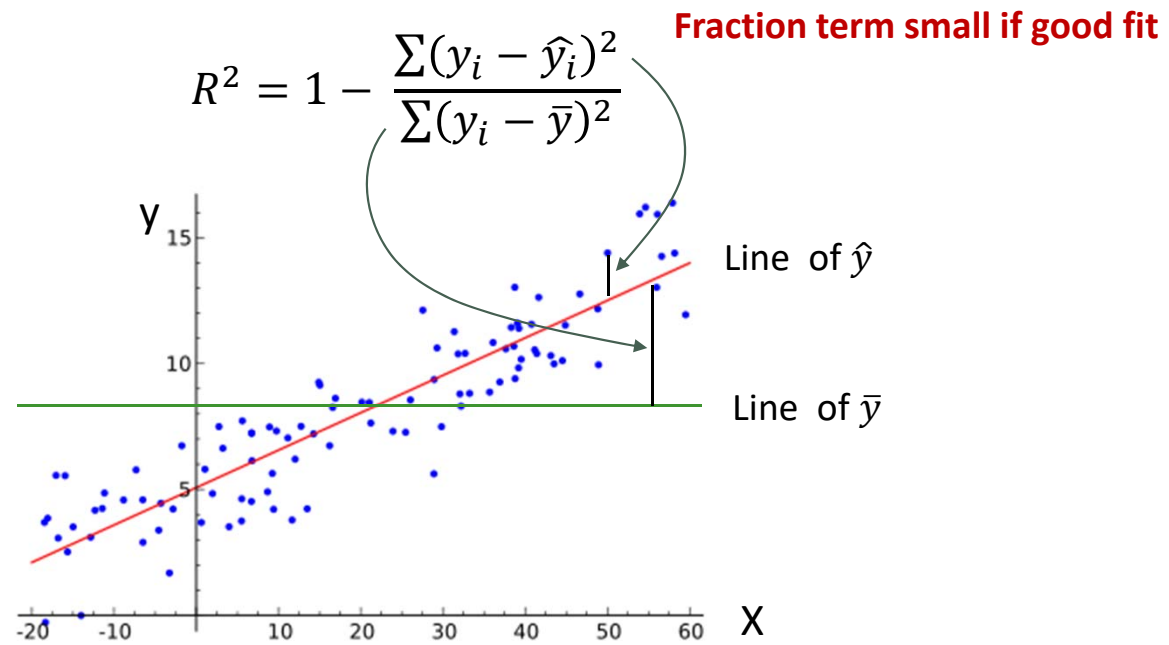
$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

$R^2$  in 0 to 1

May also be given p-values for  $a$  and  $b$ . If  $p < \alpha$ , significant; else assume 0

# R<sup>2</sup> Values & Fit

---



# Examples

---

When conducting a t-test for a parameter estimated by simple linear regression using a formula which consists of the parameter estimate divided by its standard deviation, you are testing whether that parameter is equal to:

- a) Its actual value
- b) Anything but the value you estimated
- c) Zero
- d) One

SOLN: (c)



---

An analysis is being conducted comparing the height of a building (X, in meters) of a building and the age of the building (Y, in years). Given the data provided below, what is the least squares estimate of the a (intercept) coefficient?

$$\sum_i X_i = 167$$

$$\sum_i Y_i = 49$$

$$\sum_i X_i Y_i = 366$$

$$\sum_i X_i^2 = 425$$

$$\text{no. observations} = 62$$

SOLN:

- Note: you may also see the intercept denoted as  $B_0$  or  $\beta_0$
- $a = \bar{Y} - b \cdot \bar{X}$
- $\bar{X} = \sum_i X_i / n = 167 / 62 = 2.69$
- $\bar{Y} = \sum_i Y_i / n = 49 / 62 = 0.79$
- $b = S_{xy} / S_{xx}$ 
  - $S_{xx} = \sum_i X_i^2 - (1/n)(\sum_i X_i)^2 = 425 - (1/62)(167)^2 = -24.82$
  - $S_{xy} = \sum_i X_i Y_i - (1/n)(\sum_i X_i)(\sum_i Y_i) = 366 - (1/62) \cdot 167 \cdot 49 = 234.02$
- $b = 234.02 / -24.82 = -9.43$
- Finally:  $a = 0.79 - (-9.43)(2.69) = 26.16$

Check: Make sure whether the question is asking for the slope or intercept. Chances are both will be in the answer list!

---

The results of a certain regression analysis (comparing the hardness of soil with its clay content) for 27 observations give a  $\beta_1$  estimate of -2.1. Given this information, what is the largest value of (the estimate of) the standard deviation of the estimate of the  $\beta_1$  parameter such that you could say, with 95% confidence, that the true value of the coefficient is not equal to zero?

SOLN:

- Note:  $\beta_1$  is another notation for **b (the slope)**. Statistics packages often use this notation and it is also common in multiple variable regression.
- We would say coefficient is 0 if the 95% CI contains 0. In this case, that means the CI half width must be  $< 2.1$
- Half-width =  $t_{\alpha/2, n-2} \sqrt{MSE/S_{xx}}$ , where  $\sqrt{MSE/S_{xx}}$  is the standard deviation of the estimate so can just say =  $t_{\alpha/2, n-2} * \sigma_b$
- $\alpha=1-\text{confidence} = 1-0.95 = 0.05$ . Half this is 0.025. Also,  $n=27$ , so  $n-2 = 25$
- From student t table on pg. 77,  $t_{0.025, 25} = 2.06$
- Solve for  $\sigma_b$ :  $2.1 = 2.06 * \sigma_b \rightarrow \sigma_b = 2.1/2.06 = 1.02$

---

A firm has developed a regression analysis linking the average rainfall (X, in cm) in an area and how often, on average, roofs have to be replaced (Y, in years). The firm wishes to rate the performance of the analysis. Given the data below, which has been tabulated from the collected data, what is the value of the  $R^2$  measure from this analysis?

$$\begin{array}{llll} \sum_i X_i = 254,064 & \sum_i Y_i = 80,936 & \sum_i X_i Y_i = 7,870,000 & \\ \sum_i X_i^2 = 24,056,521 & \sum_i Y_i^2 = 2,608,322 & & \text{no. observations} = 3,098 \end{array}$$

SOLN:

$$R^2 = S_{xy}^2 / (S_{xx} * S_{yy})$$

- $S_{xx} = \sum_i X_i^2 - (1/n)(\sum_i X_i)^2 = 24056521 - (1/3098)*(254064)^2 = 3,220,976.7$
- $S_{yy} = \sum_i Y_i^2 - (1/n)(\sum_i Y_i)^2 = 2608322 - (1/3098)*(80936)^2 = 493,849.4$
- $S_{xy} = \sum_i X_i Y_i - (1/n)(\sum_i X_i)(\sum_i Y_i) = 7870000 - (1/3098)*254064*80936 = 1,232,516.5$

$$R^2 = 1232516.5^2 / (3220976.7 * 493849.4) = 0.955$$

---


While fitting a regression line to X (independent) and Y (dependent) variables, you find  $R = -0.98$ . Which of the following is the correct interpretation:

- a) X and Y are not linearly correlated
- b) X and Y are positively correlated
- c) X and Y are strongly negatively correlated
- d) X and Y are uncorrelated

SOLN: (c)

R tells direction as well as magnitude of correlation. A negative correlation indicate X and Y are negatively correlated (tend to move in opposite directions)

Both R and  $R^2$  only tell about LINEAR correlation. They can be 0 and X and Y may still be nonlinearly correlated.





# Hypothesis Testing

---



# Hypothesis Testing

---

Formalized method of determining if your hypothesis (claim) is true, with a certain level of confidence that the result is correct

- Rejecting a hypothesis: sample evidence does not support the hypothesis on the population
- Reject if there is only a small probability of having obtained the sample when the assumed population parameter(s) are true.

Null hypothesis:  $H_0$

- Aka, “default hypothesis”; usually the status quo
- Assumed true, but really is what you are trying to prove is not true

Alternative hypothesis:  $H_1$

- What you are trying to prove is true

Test statistic:

- Depends on what testing (Z, t, Chi-square, F)

Criterion for acceptance/rejection:

- **Significance level  $\alpha$** . Typically 0.05 (sometimes 0.01, 0.10)
- Pre-select this level before testing the statistics

See tables on page 73 and 74 of the reference handbook:

- **Test on single mean**: variance known (Table A - Normal) or unknown (Table B – student-t dist.)
- **Tests on difference in means**: variance known, unknown but equal, unknown and unequal (Table B)
- **Tests on variance of a single population** (Table C – Chi-Square dist.)
- **Tests on variances between two populations** (Table C – F dist.)

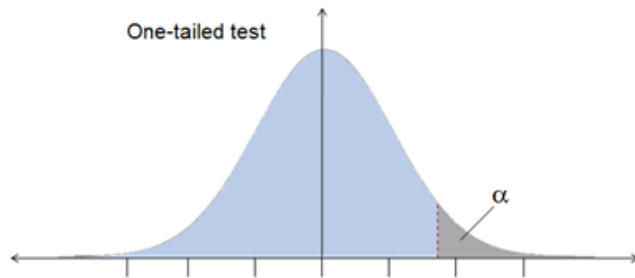
# Stating the hypothesis: 1 vs. 2 tails

---

## 1-TAILED TEST

$$H_0: \mu = x$$

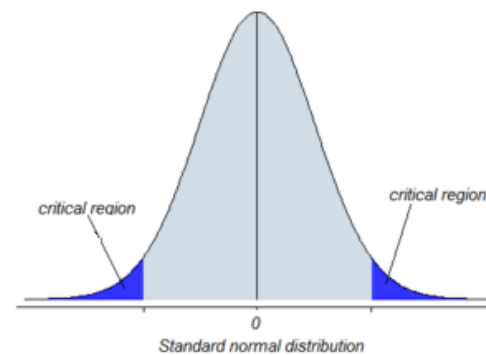
$$H_1: \mu < x \quad (\text{or } \mu > x)$$



## 2-TAILED TEST

$$H_0: \mu = x$$

$$H_1: \mu \neq x$$

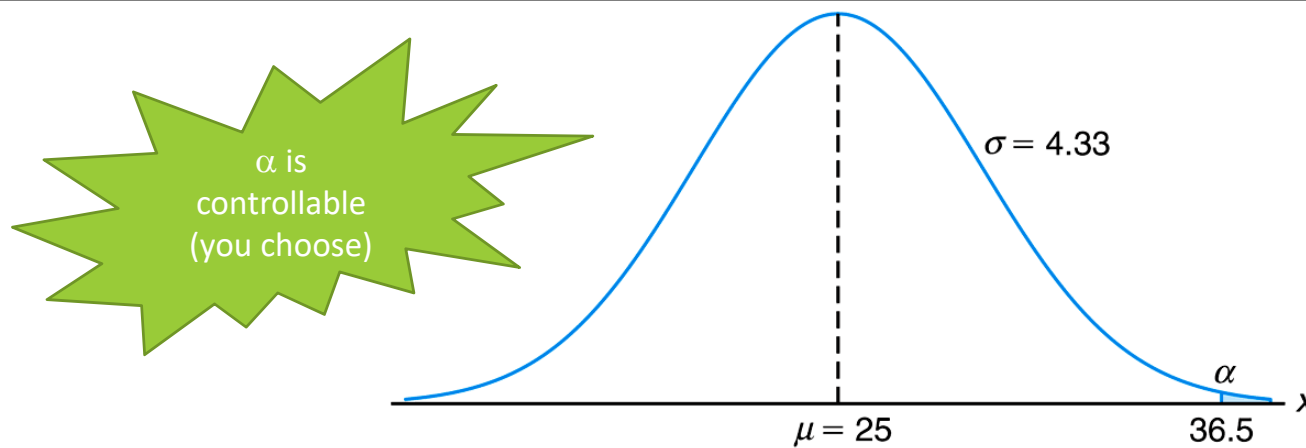


# Sources of error

---

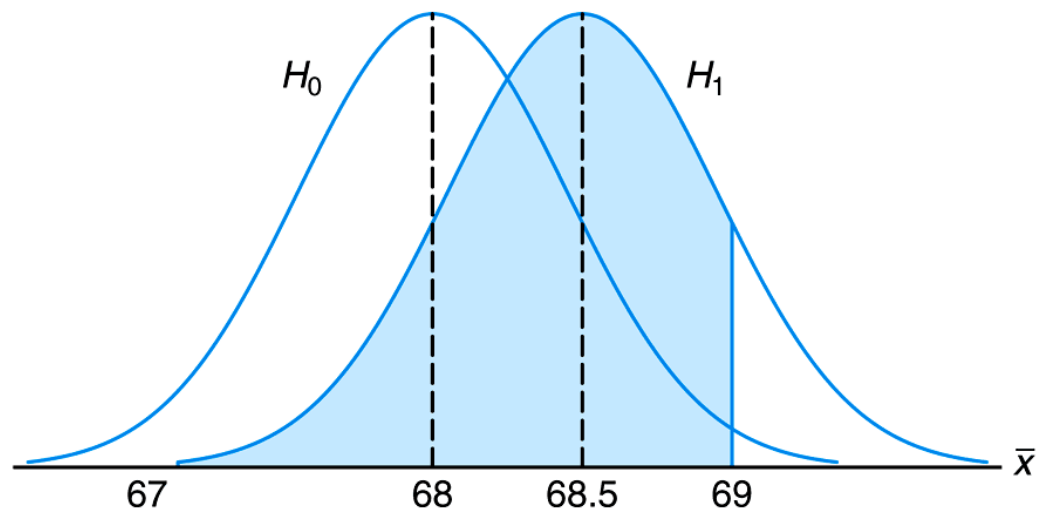
		Reality (truth)	
		<b>H<sub>0</sub> is true</b>	<b>H<sub>0</sub> is false</b>
Your conclusion	<b>Do not reject H<sub>0</sub></b>	Correct ( <i>confidence</i> )	Type II error $\beta$
	<b>Reject H<sub>0</sub></b>	Type I error $\alpha$	Correct ( <i>power</i> )

# Confidence and Type I ( $\alpha$ ) error



		Reality (truth)	
		$H_0$ is true	$H_0$ is false
Your conclusion	Do not reject $H_0$	Correct (confidence= $1-\alpha$ )	Type II error $\beta$
	Reject $H_0$	Type I error $\alpha$	Correct ( <i>power</i> )

## Type II ( $\beta$ ) error and Power



		Reality (truth)	
		$H_0$ is true	$H_0$ is false
Your conclusion	Do not reject $H_0$	Correct (confidence)	Type II error $\beta$
	Reject $H_0$	Type I error $\alpha$	Correct ( <i>power</i> )= $1-\beta$

# p-value

---

Probability of outcomes at least as far from the mean as your sample outcome if in fact the null hypothesis is true.

Physical interpretation: The lowest level of significance at which the test would reject  $H_0$ .

- Reject  $H_0$  if  $\alpha \geq p$

# Examples

---

You conduct a hypothesis test and end up rejecting the null hypothesis. Later, you find out that  $H_0$  was actually true. What type of error did you commit? Typographic, Human, Type I, or Type II?

SOLN: Type 1



---

A sample of 23 numbers is drawn from an unknown distribution. Significance level is 0.05. The sample mean is 57.4 and the sample standard deviation is 1.3. If  $H_0$  is  $\mu=55$  and  $H_1$  is  $\mu>55$ , what is the test critical value for the hypothesis test?

SOLN:

Test on single mean.  $\sigma$  not given, so unknown.

Significance = Type I error =  $\alpha$

From Table B, for  $H_1 \mu > \mu_0$ , critical value is  $t_{\alpha, n-1}$ . From student-t table,  $t_{0.05, 22} = 1.717$

Checks: Selected correct test? Correct  $H_1$ ? Correct  $\alpha$ ?



---

You take a sample of 15 from a population and calculate a sample standard deviation of 12.2. Historically,  $\sigma=8.1$  for this process. You are concerned with whether variance has increased. What test statistic value should be used?

SOLN:

- This is a test of variance for a single population. Use Chi-Square (top of Table C)
- Test statistic is chi-square statistic:  $= (n-1)*s^2 / \sigma^2 = 14*12.2^2 / 8.1^2 = 31.76$

Same question. What is the p-value?

- $p = P(\chi^2 > 31.76)$  for  $dof=n-1=14$ . From Chi-Square table on pg 79, for  $n=14$ ,  $p$  falls between 0.005 and 0.01. Interpolate or knock out possible answers based on this.

Checks: Selected correct test? Correct  $H_1$ ? Correct  $\alpha$ ? Given  $s$ ,  $s^2$ ,  $\sigma$ , or  $\sigma^2$ ?



# Analysis of Variance (ANOVA) & Factorial Design

---

# Analysis of Variance (ANOVA) & Factorial Designs

---

## One-Way ANOVA:

- Used with 3 or more **treatment groups** for MEAN DIFFERENCES
  - Is variability between groups just randomness or due to treatment differences?
- Assumptions:
  - The response variable is normally distributed for all treatment groups
  - The population standard deviations are equal for all treatment groups
  - Randomization, such that samples from the  $g$  populations can be treated as *independent* random samples

## Randomized Complete Block Design:

- Similar use and assumptions to One-Way ANOVA
- There may be factors other than treatment that impact response variable, such as perhaps gender, age, ethnicity, etc.
- So assign units based on both treatment level AND blocks

## Two-Factor (or Two-Way) Factorial Design:

- Similar use and assumptions to One-Way ANOVA
- Rather than treatments and blocks, have factor variables, each with multiple possible discrete values (levels)
- Formulas are given for the two-factor design: For  $a$  levels of Factor A,  $b$  levels of Factor B, and  $n$  repetitions (units or observations) per cell

Formulas are quite computationally intensive for time limits of exam. Questions are most likely to be conceptual understanding, or they will provide many of the sums already calculated.

# One-way ANOVA

Let a "dot" subscript indicate summation over the subscript. Thus:

$$y_{i\cdot} = \sum_{j=1}^n y_{ij} \quad \text{and} \quad y_{\cdot\cdot} = \sum_{i=1}^k \sum_{j=1}^n y_{ij}$$

Given independent random samples of size  $n_i$  from  $k$  populations, then

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^k n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

$$SS_{\text{total}} = SS_{\text{treatments}} + SS_{\text{error}}$$

If  $N =$  total number observations

$$N = \sum_{i=1}^k n_i, \text{ then}$$

$$SS_{\text{total}} = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N}$$

$$SS_{\text{treatments}} = \sum_{i=1}^k \frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{N}$$

$$SS_{\text{error}} = SS_{\text{total}} - SS_{\text{treatments}}$$

Test Statistic for significance of effect. Test is:  
 $H_0$ : no difference in treatment means  
 $H_1$ : difference in treatment means  
 Use F critical value for given DOF of treatments (numerator) and total (denominator) and  $\alpha$ .  
 You may also be given p-value(s)

One-Way ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Between Treatments	$k - 1$	$SS_{\text{treatments}}$	$MST = \frac{SS_{\text{treatments}}}{k - 1}$	$\frac{MST}{MSE}$
Error	$N - k$	$SS_{\text{error}}$	$MSE = \frac{SS_{\text{error}}}{N - k}$	
Total	$N - 1$	$SS_{\text{total}}$		

# Randomized Complete Block Design

For  $k$  treatments and  $b$  blocks

$$\sum_{i=1}^k \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2 = b \sum_{i=1}^k (\bar{y}_{i.} - \bar{y}_{..})^2 + k \sum_{j=1}^b (\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^b (\bar{y}_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y}_{..})^2$$

$$SS_{\text{total}} = SS_{\text{treatments}} + SS_{\text{blocks}} + SS_{\text{error}}$$

$$SS_{\text{total}} = \sum_{i=1}^k \sum_{j=1}^b y_{ij}^2 - \frac{y_{..}^2}{kb}$$

$$SS_{\text{treatments}} = \frac{1}{b} \sum_{i=1}^k y_{i.}^2 - \frac{y_{..}^2}{bk}$$

$$SS_{\text{blocks}} = \frac{1}{k} \sum_{j=1}^b y_{.j}^2 - \frac{y_{..}^2}{bk}$$

$$SS_{\text{error}} = SS_{\text{total}} - SS_{\text{treatments}} - SS_{\text{blocks}}$$

Randomized Complete Block ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Between Treatments	$k - 1$	$SS_{\text{treatments}}$	$MST = \frac{SS_{\text{treatments}}}{k - 1}$	$\frac{MST}{MSE}$
Between Blocks	$n - 1$	$SS_{\text{blocks}}$	$MSB = \frac{SS_{\text{blocks}}}{n - 1}$	$\frac{MSB}{MSE}$
Error	$(k - 1)(n - 1)$	$SS_{\text{error}}$	$MSE = \frac{SS_{\text{error}}}{(k - 1)(n - 1)}$	
Total	$N - 1$	$SS_{\text{total}}$		

# Two-Way Factorial Design

For  $a$  levels of Factor A,  $b$  levels of Factor B, and  $n$  repetitions per cell:

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{...})^2 = bn \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 + an \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2 + n \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij.})^2$$

$$SS_{\text{total}} = SS_A + SS_B + SS_{AB} + SS_{\text{error}}$$

$$SS_{\text{total}} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n y_{ijk}^2 - \frac{y_{...}^2}{abn}$$

$$SS_A = \sum_{i=1}^a \frac{y_{i..}^2}{bn} - \frac{y_{...}^2}{abn}$$

$$SS_B = \sum_{j=1}^b \frac{y_{.j.}^2}{an} - \frac{y_{...}^2}{abn}$$

$$SS_{AB} = \sum_{i=1}^a \sum_{j=1}^b \frac{y_{ij.}^2}{n} - \frac{y_{...}^2}{abn} - SS_A - SS_B$$

$$SS_{\text{error}} = SS_T - SS_A - SS_B - SS_{AB}$$

Two-Way Factorial ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
A Treatments	$a - 1$	$SS_A$	$MSE_A = \frac{SS_A}{a - 1}$	$\frac{MSE_A}{MSE}$
B Treatments	$b - 1$	$SS_B$	$MSE_B = \frac{SS_B}{b - 1}$	$\frac{MSE_B}{MSE}$
AB Interaction	$(a - 1)(b - 1)$	$SS_{AB}$	$MSE_{AB} = \frac{SS_{AB}}{(a - 1)(b - 1)}$	$\frac{MSE_{AB}}{MSE}$
Error	$ab(n - 1)$	$SS_{\text{error}}$	$MSE = \frac{SS_{\text{error}}}{ab(n - 1)}$	
Total	$abn - 1$	$SS_{\text{total}}$		

# Examples

---

A hardness testing machine operates by pressing a tip into a metal test "coupon". Four tip types are each tested on four different coupons (the coupons may vary slightly in hardness). We are interested in whether there is a difference between the tip types. What kind of test should be used?

- a) Randomized Complete Block Design
- b) One-Way ANOVA
- c) Two-Way Factorial Design
- d) T-test hypothesis test

SOLN: (a). Here tip type is the treatment of interest, and coupon is a "nuisance" factor (block) we need to control for.



---

Same as last problem. Given the following summary and  $\alpha=0.05$ , what is the F test statistic for tip type, and is it significant?

Source	DF	SS	MS
Coupon	3	0.82500	0.27500
Tip	3	0.38500	0.12833
Error	9	0.08000	0.00889
Total	15	1.29000	

SOLN:

F test statistic:

- $MST=0.12833$  (given),  $MSE= 0.00889$  (given),  $F = MST/MSE = .12833/0.00889= 14.44$
- Critical value  $F_{0.05,4-1=3,9-1=8} = 3.86$
- Since  $14.44 > 3.86$ , reject  $H_0$ : treatment effect is significant at  $\alpha=0.05$

---

Same as last problem. Given the following summary and  $\alpha=0.05$ , which of the following is correct?

Source	DF	SS	MS	F	P
Coupon	3	0.82500	0.27500	30.94	0.000
Tip	3	0.38500	0.12833	14.44	0.001
Error	9	0.08000	0.00889		
Total	15	1.29000			

- a) Neither the treatment or blocking factor were significant
- b) The treatment factor was significant but not the blocking factor
- c) The blocking factor was significant but not the treatment factor
- d) Both the treatment and blocking factors were significant

SOLN: (d)

Since both the treatment (tip) and blocking (coupon) p-values are  $< \alpha$ , both are significant

