# Mathematical Foundations of Data Science

Session 5 – Inferential Statistics 3

Nandan Sudarsanam,
Department of Data Science and AI,
Wadhwani School of Data Science and AI,
Indian Institute of Technology Madras

# Rejecting and failing to reject the null hypothesis

- Acceptance Matrix for hypothesis tests

| | | Decision | |
|---|---|---|---|
| | | Reject The Null Hypothesis | Fail to Reject the Null hypothesis |
| Actual | Null hypothesis is true | Type 1 Error (or Producer Risk, False Positive, alpha-risk) | Correct Decision (1-alpha) |
| | Alternate Hypothesis is true | Correct Decision (Power = 1-Beta) | Type 2 Error (Consumer risk, False Negative, Beta risk) |

# Type I and Type II Errors

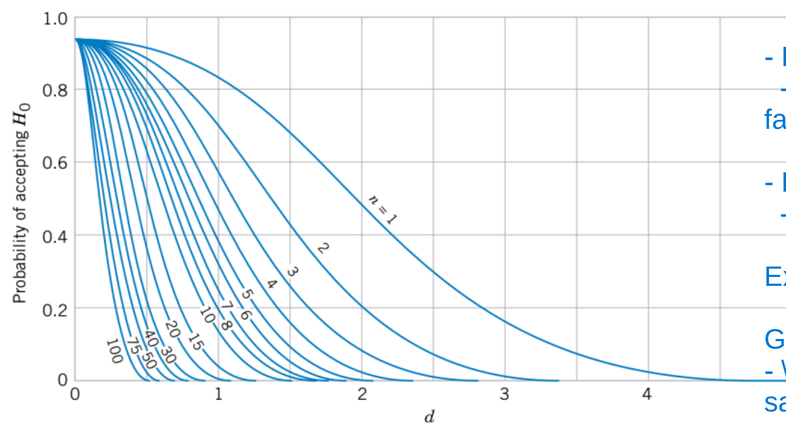
- Prior to any data collection
  could be as high as alpha (
  analysis it is exactly equal t

- Type II error ($\beta$) is more co

  - It is a function of Delta: $\delta = |\mu - \mu_0|$
  - It is a function of standard deviation: $\sigma$
  - Often we focus on $d = {}^{\delta}/_{\sigma}$
  - It is a function of Sample Size: *n*
  - It is a function of the Type I error: $\alpha$

# OC curves

- A graph of $\beta$ versus d for a given sample size ($n$) is known as an OC (Operational Characteristic) curve:




(a) O.C. curves for different values of $n$ for the two-sided normal test for a level of significance $\alpha = 0.05$.

OC Curve Interpretation:

- **X-axis:** Effect size (\(d\)) or \(|\mu - \mu_0|\)
- **Y-axis:** Probability of Type II error (\(\beta\))

- For small effect sizes:
  - \(\beta\) is large, indicating a higher chance of failing to reject a false null hypothesis.

- For large effect sizes:
  - \(\beta\) decreases, increasing the test's power.

Example:

Given:
- \(\alpha = 0.05\), \(\sigma = 10\), and various sample sizes, calculate \(\beta\) for different values of effect size (\(d\)) and plot \(\beta\) vs. \(d\).

Steps to Plot OC Curve:

1. **Choose Effect Sizes (d):**
   Pick a range of effect sizes, e.g., \(d = 0.1, 0.2, \dots, 1.0\).

2. **Compute \(\beta\) for Each Effect Size:**
   Use the z-test formula for power and Type II

# Introduction

- So far statistical inference was confined to input variables that could take up two possible values (two sample tests), or there was no notion of an input variable (single sample tests).
- ANOVA
  - When there are three or more states of a single variable we can use ANOVA
- Chi-Square Test of Independence
  - Can be used when we want to compare multiple proportions

# BASICS of ANOVA

- Tests the hypothesis that: $\mu_A = \mu_B = \mu_C = \mu_D$
- Why not multiple pairwise comparisons using t-tests?
- What do you do after a test? Tukey, Bonferroni, Scheffe tests
- Take the table:

| | 1 | 2 | n |
|---|---|---|---|
| A | $y_{1,1}$ | $y_{1,2}$ | $y_{1,n}$ |
| B | $y_{2,1}$ | … | … |
| C | … | … | … |
| D | … | … | $y_{4,n}$ |

# ANOVA OUTPUT

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F-Stat |
|---|---|---|---|---|
| Between Treatments | $n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{\bar{y}}_{..})^2$ or SSB | a-1 | MSB = SSB/DoF | F = MSB/MSE |
| Error within treatments | $\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j} - \bar{y}_{i.})^2$ or SSE | N-a | MSE = SSE/DoF | |
| Total | $\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j} - \bar{\bar{y}}_{..})^2$ or SST | N-1 | MST = SST/DoF | |

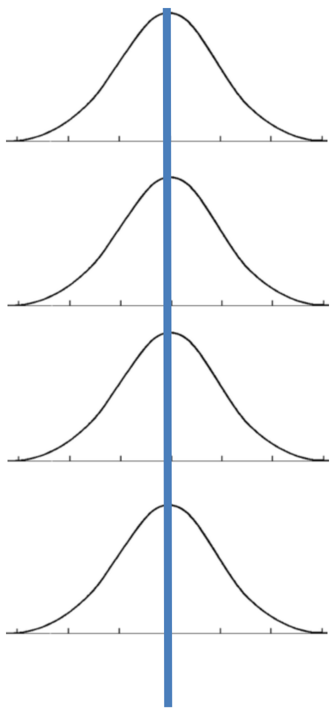
Compare F calculated against the F-distribution with a-1,N-a degrees of freedom and get a p-value

# Why F for difference in means??

- The F is the ratio of two variances (where the samples come from a normal distribution and the null hypothesis is that the variances are equal)
- MSB is a way of calculating total variance
- MSE is a way of calculating total variance
- MSB, MSE and MST will be equal if the null hypothesis is true
- However is the null hypothesis is not, then MSB>MST>MSE

# MSB and MSE

**MSE**



Total distribution of Y

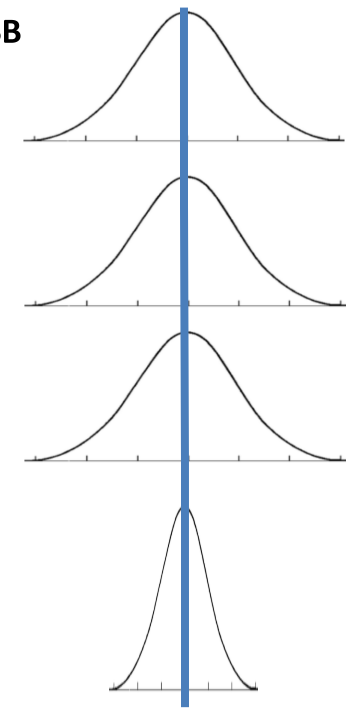
Distribution of treatment A

Distribution of treatment B

Distribution of treatment C

.
.
.

**MSB**



Total distribution of Y

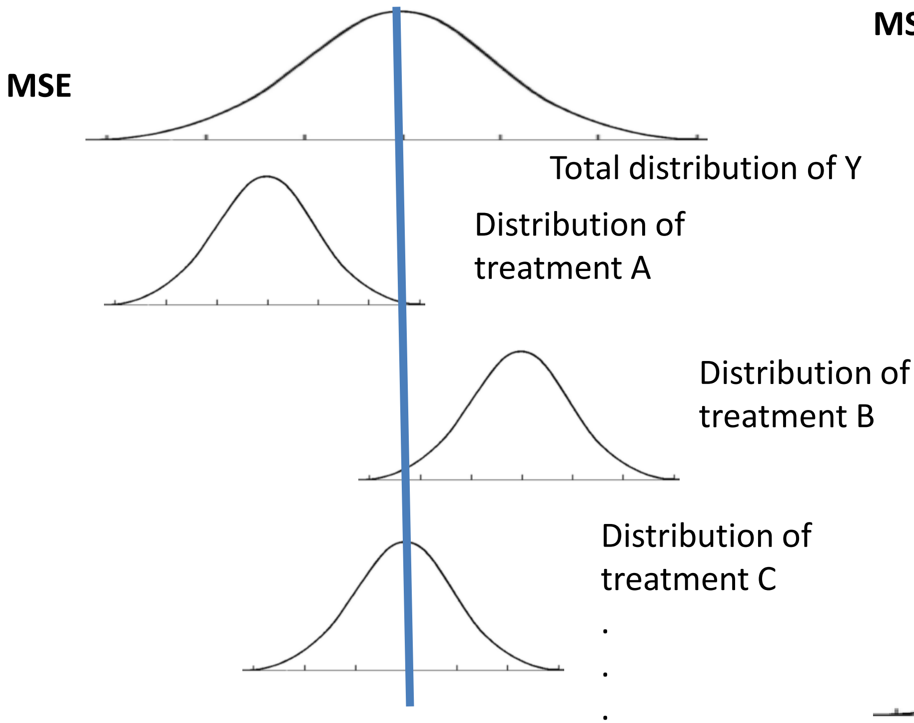
Distribution of treatment A

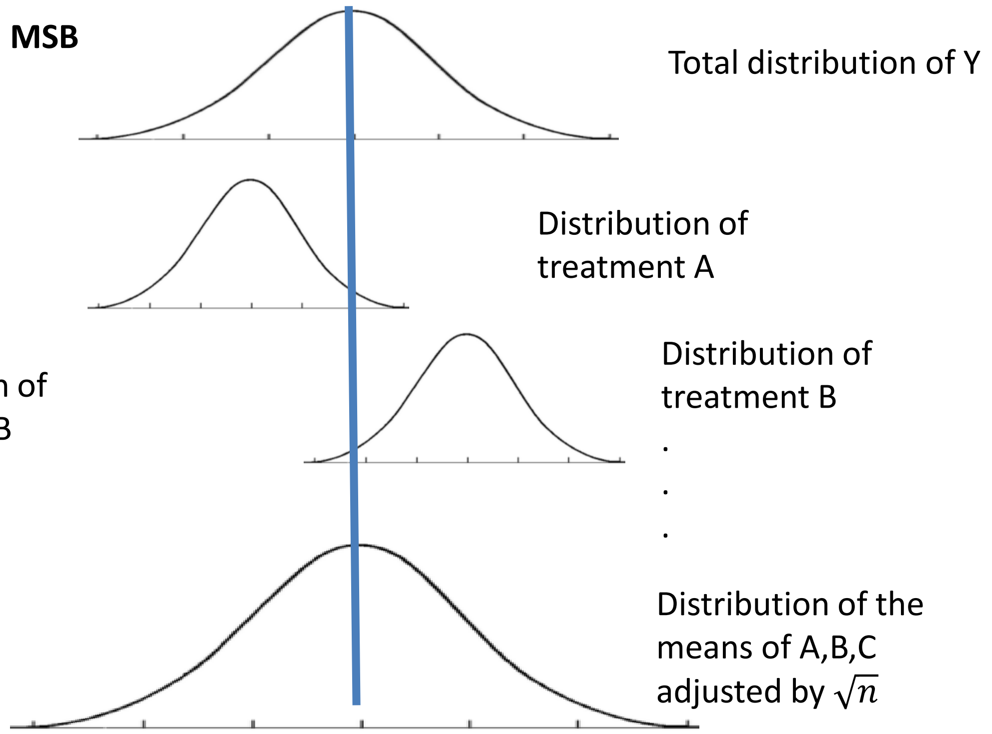
Distribution of treatment B

.
.
.

Distribution of the means of A,B,C

# MSB and MSE


MSE
Total distribution of Y
Distribution of treatment A
Distribution of treatment B
Distribution of treatment C
.
.
.


MSB
Total distribution of Y
Distribution of treatment A
Distribution of treatment B
.
.
.
Distribution of the means of A,B,C adjusted by $\sqrt{n}$

# Chi-Square TOI

- When using categorical variables
- Use this to test:
  - Does the input categorical variable effect the output categorical variable (works 2 or more states of the input or output variable)
  - Independence between two variables
  - Construct a contingency table:

| Exercise \ Smoking habit | Heavy | Regular | Occasional | Never |
|---|---|---|---|---|
| Frequent | 7 | 9 | 12 | 87 |
| Some | 3 | 7 | 4 | 84 |
| None | 1 | 1 | 3 | 18 |

# Chi Square TOI continued

- Create Theo values for this table in accordance to the assumption of independence
- It can be done row wise or column wise, but each cell gets an expected value
- Then if the null hypothesis is true then the test statistic is:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

With (r-1)*(c-1) degrees of freedom (or rc-c-r+1)