Using Appropriate Statistics – Statistics for Artificial Intelligence

Lecturer: Steffen Christensen
Authors: Steffen Christensen,
Mark Wineberg

Top 5 Experimental Analysis Myths in AI

i. Results from 1 run shows you anything other than proof-of-concept
   - It doesn't
ii. The single best value achieved in a set of runs tells you anything significant about the population distribution
   - No
iii. Using the same seed for your random number generator in both treatments controls for *anything*
   - It doesn't
iv. The mean performance of your *entire* population is worth doing statistics on
   - You normally want best-of-run
v. One system is obviously better than the other when looking at the data or graph - there is no need for a statistical analysis
   - If it is so obvious, then will be easy to show statistically
     - might as well do the stats
     - shows that you are objectively confident in your conclusion

Top 12 Statistics Myths in AI

1. My mean result being better than yours means my technique is superior to yours
   - In the best case you need to perform a T test to assert this claim
2. Reporting the mean value of a statistic is good enough
   - You need some representative range
3. Reporting the mean and standard deviation of a statistic is good enough
   - Need number of runs
4. Your data are normally distributed
   - Not usually

Top 12 Statistics Myths in AI

5. The mean performance of the best-of-run individuals of your system is what matters
   - It's usually the median you want
6. 10 runs is enough to show significant differences between groups
   - It can be, but the statistics required to show this are hairy
7. 95% confidence levels are generally sufficient
   - Try 99.9%
8. Drawing 95% confidence intervals around each sample mean on a graph implies that it's a rare event if any of the true means fall outside the CIs
   - Nope; need Bonferroni correction

Top 12 Statistics Myths in AI

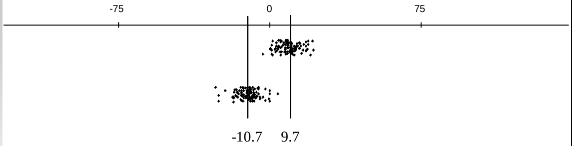
9. Reporting the results of several comparisons where each is made at a 95% confidence level means that all conclusions are valid simultaneously
   - Nope; need Bonferroni correction for that too
10. 95% confidence intervals can be computed using the sample mean ± 1.96 standard deviations of the mean
    - Nope; need the Student's T score given your degrees of freedom
11. An experimental setup where more than one parameter is varied can be treated like one where exactly one parameter varies
    - Need ANOVA, MANOVA or regression
12. One can infer trends from observed data beyond the data you've generated
    - Generally, this would be a consequence of some model, and you probably haven't supported said model with enough experimental data

Myth 1: Averages are Everything

- We might get unlucky with our data distribution – a simple comparison between two averages might not give the same result as the comparison between two distributions
- Consider the following samples of two distributions (blue and green), which are normally distributed and have the following exact parameters:

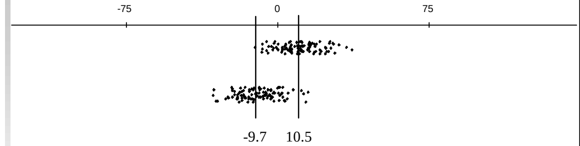
Mean	StdDev	Reps
+10	s = 5, 10, 50	N = 100, 5
-10	s = 5, 10, 50	N = 100, 5

Sampling From Two Normal Distributions



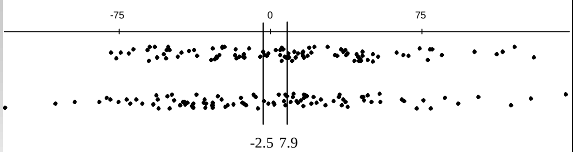
Mean	StdDev	Reps
+10	5	100
-10	5	100

Sampling From Two Normal Distributions



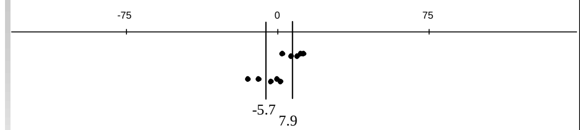
Mean	StdDev	Reps
+10	10	100
-10	10	100

Sampling From Two Normal Distributions



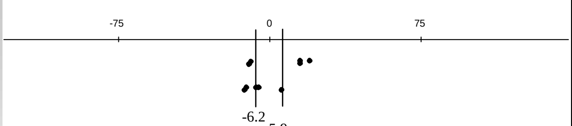
Mean	StdDev	Reps
+10	50	100
-10	50	100

Sampling From Two Normal Distributions



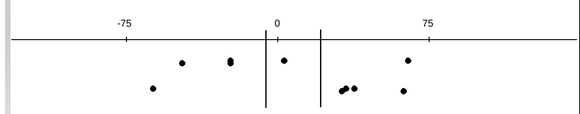
Mean	StdDev	Reps
+10	5	5
-10	5	5

Sampling From Two Normal Distributions



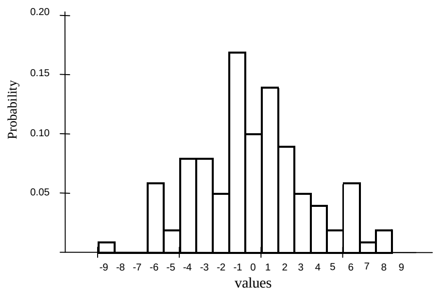
Mean	StdDev	Reps
+10	10	5
-10	10	5

Sampling From Two Normal Distributions



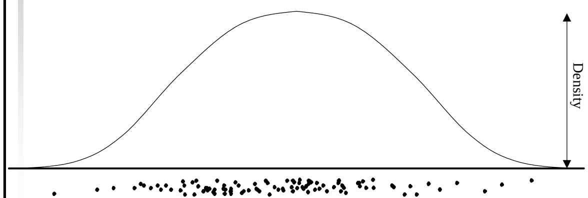
Mean	StdDev	Reps
+10	50	5
-10	50	5

Probability Function



Probability Density Functions

- The greater the density of observed outcomes, the larger the density function's value
- Probability = integration (sum) of the density function between a range of values



What Are We Interested In?

- For most statistical analysis for AI the question is
  - Is my new way better than the old way?
  - Statistically this translates into a statement about the difference between means: “Is the difference between ‘my mean’ and ‘the old mean’ greater than zero?”
- However, to answer this question you must first be able to estimate the true mean of both distributions
  - Of course the true mean will not be where the sample average is
  - So what does the sample average tell us?

The Sample Average as a Sum of *n* Random Variables

- Let us consider the sample average of 10 points

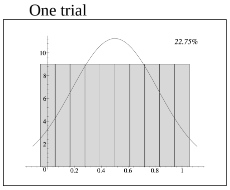
$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$$

- Another way of writing this is $\bar{x} = \frac{1}{10}(x_1 + x_2 + ... + x_{10})$
- We can say that x_1 is drawn from X_1, x_2 is drawn from X_2, etc.
  - Where X_1, X_2, … are themselves separate identically-distributed random variables
  - So what we really need to know is the behavior of the sum of many identically distributed random variables
- This has been studied, producing a useful result known as the *Central Limit Theorem*: the sum of many identically distributed random variables tends to a Gaussian

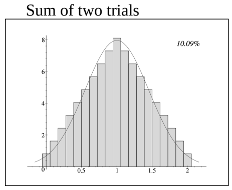
Central Limit Theorem

- The sum of many *independent, identically distributed (IID)* random variables approaches a Gaussian normal curve
- E.g. Uniform distribution on [0, 1]:

One trial



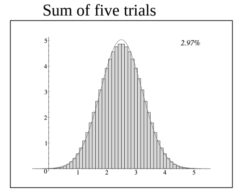
Sum of two trials



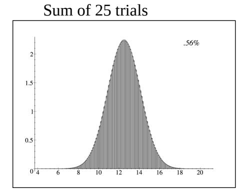
Central Limit Theorem

- E.g. Uniform distribution (continued):

Sum of five trials

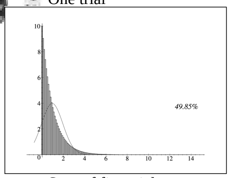


Sum of 25 trials

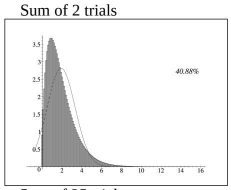


Exponential Distribution

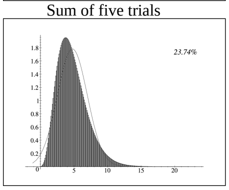
One trial



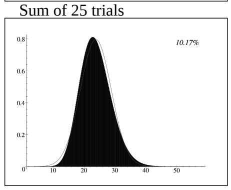
Sum of 2 trials



Sum of five trials

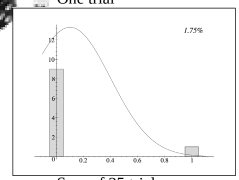


Sum of 25 trials

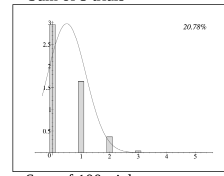


Binomial Distribution (p = 0.9)

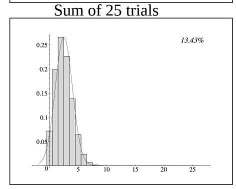
One trial



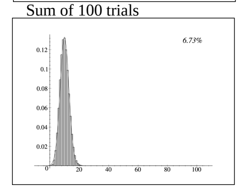
Sum of 5 trials



Sum of 25 trials



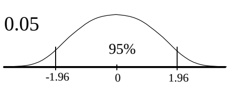
Sum of 100 trials



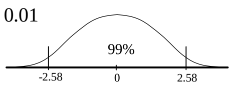
Estimating the Mean: Confidence Intervals Around the Average

The wider the cutoff values, the more likely that the true mean will fall within that range

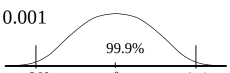
$a = 0.05$



$a = 0.01$



$a = 0.001$



- Cutoff values are called *z-scores*
- α is the probability of obtaining values outside the cutoffs
- Since cutoffs are on both sides, probability of each side is α /2
- Confidence Level is 1 – α
- Use "– NORMSINV(α / 2)" to compute *z-scores* in Excel
- Use *T-scores* with small sample sizes: "TINV(α, N - 1)"

Estimating the Mean: Confidence Intervals Around the Average

- These cutoff values form a confidence interval so named because you are confident that the mean lies within this interval (for a given confidence level)
- So for 95% coverage, the confidence interval for the sample mean $\overline{X}$ would be

$$\overline{X} = m_{\overline{X}} \pm 1.96\, s_{\overline{X}}$$

Using $s_{\overline{X}} \cong \frac{s_X}{\sqrt{n}}$ and $m_{\overline{X}} = m_X$, we get:

$$\overline{X} = m_X \pm 1.96 \frac{s_X}{\sqrt{n}}$$

Estimating the Mean: Confidence Intervals Around the Average

- Since we want to estimate m_X, the true mean of the original distribution, we must rearrange the equation and solve for m_X

Error Bounds

$$m_X = \overline{X} \pm 1.96 \frac{s_X}{\sqrt{n}}$$

Confidence Intervals

$$\overline{X} - 1.96 \frac{s_X}{\sqrt{n}} \le m_X \le \overline{X} + 1.96 \frac{s_X}{\sqrt{n}}$$

$$m_X \in \left[\overline{X} - 1.96 \frac{s_X}{\sqrt{n}}, \overline{X} + 1.96 \frac{s_X}{\sqrt{n}} \right]$$

All formulations are equivalent

There is a 95% chance that the true mean of *X* will lie within the confidence interval

In other words there is only a 5% chance that the true mean lies outside the confidence interval formed around the sampled mean

T-Test Step by Step

1. Compute the 2 averages X_1 and X_2
2. Compute standard deviations s_1 and s_2
3. Compute degrees of freedom: $N_1 + N_2 - 2$
4. Calculate T
5. Computes the area under the T distribution greater than the T value
   - *This is the p-value*
   - Use TDIST(T, N - 1, 2) in Excel
     - The final "2" means "two-sided", which is the usual case

$$T = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

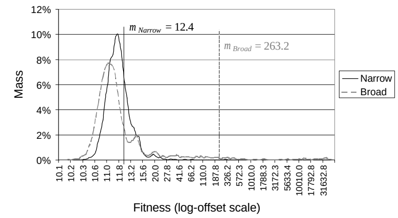
When The CLT Fails You

- Everything we have done so far depends on the Central Limit Theorem holding
  - But this is not always true
  - ***In fact for AI it rarely holds***
- Problems occur when
  - …you have a non-zero probability of obtaining infinity
    - Mean and standard deviation are infinite!
  - …the sample average depends highly on a few scores
    - When the mean of your distribution is not measuring what you want, consider using the median instead (rank-based statistics)
- AI alert!
  - Many data in artificial intelligence are often highly skewed because some local optima in the search space are very unfit
  - Example follows

When The CLT Fails You

From a node layout problem where fitness is absolute error (minimization)

- Here are the PDFs of 2 AI parameter settings, named Broad and Narrow for convenience



- Here Broad's mean is much *worse* than Narrow's because of its extended tail, even though Broad often beats Narrow in practice!
  - We don't really care about the 8% of trials where Broad performs badly

So what should we do?

- There are tests that use Ranks instead of actual values
  - These are called **Non-Parametric** Tests
  - They measure how interspersed the samples from the two treatments are
    - If the result is "alternating" it is assumed that there is no effective difference

Non-Parametric Tests

- Ranks are uniformly distributed (think of percentiles – uniform on [0%, 100%])
- The sum of ranks and average of ranks will be approximately normally distributed because of the Central Limit Theorem, as long as we have 5 or more samples
  - This result is independent of the particular distributions of the 2 treatments
  - So we can perform a T-Test on the ranks
- 2 other techniques with similar results are commonly seen
  - Wilcoxon's Rank-Sum test
  - Mann-Whitney U test
- All are nearly equivalent, and the test is often called the "Mann-Whitney-Wilcoxon test" by statisticians

How To Rank the Data

- Augment each data point with a treatment identifier and an additional slot for its rank
- Sort the data sets together by value
  - record the ranks of all values in their rank slot
    - assign the average rank of tied values to each tied value
- Resort by the original order thus splitting the data sets back out
  - keep the combined ranking with each data point
- Apply your T test on the ranked values

A	0.03
A	0.91
A	0.64
A	0.99
A	0.64
A	0.16
A	0.16
A	0.91
A	0.16
A	0.27

Two sets of Data

B	0.64
B	0.08
B	0.16
B	0.27
B	0.02
B	0.01
B	0.16
B	0.03
B	0.03
B	0.64

Ranked Example

Ranked Example

A	0.03
A	0.91
A	0.64
A	0.99
A	0.64
A	0.16
A	0.16
A	0.91
A	0.16
A	0.27
B	0.64
B	0.08
B	0.16
B	0.27
B	0.02
B	0.01
B	0.16
B	0.03
B	0.03
B	0.64

Combine the data into a single array

Ranked Example

A	0.99
A	0.91
A	0.91
A	0.64
A	0.64
B	0.64
B	0.64
A	0.27
B	0.27
A	0.16
A	0.16
A	0.16
B	0.16
B	0.16
B	0.08
A	0.03
B	0.03
B	0.03
B	0.02
B	0.01

Sort the combined data

Ranked Example

		ranks
A	0.99	1
A	0.91	2
A	0.91	3
A	0.64	4
A	0.64	5
B	0.64	6
B	0.64	7
A	0.27	8
B	0.27	9
A	0.16	10
A	0.16	11
A	0.16	12
B	0.16	13
B	0.16	14
B	0.08	15
A	0.03	16
B	0.03	17
B	0.03	18
B	0.02	19
B	0.01	20

Give each data element its corresponding rank

Ranked Example

		ranks	
A	0.99	1	
A	0.91	2	t1
A	0.91	3	t1
A	0.64	4	t2
A	0.64	5	t2
B	0.64	6	t2
B	0.64	7	t2
A	0.27	8	t3
B	0.27	9	t3
A	0.16	10	t4
A	0.16	11	t4
A	0.16	12	t4
B	0.16	13	t4
B	0.16	14	t4
B	0.08	15	
A	0.03	16	t5
B	0.03	17	t5
B	0.03	18	t5
B	0.02	19	
B	0.01	20	

Average tied ranks together

t1	2.5
t2	5.5
t3	8.5
t4	12
t5	17

Identify ties

Ranked Example

		ranks	
A	0.99	1	
A	0.91	2	t1
A	0.91	3	t1
A	0.64	4	t2
A	0.64	5	t2
B	0.64	6	t2
B	0.64	7	t2
A	0.27	8	t3
B	0.27	9	t3
A	0.16	10	t4
A	0.16	11	t4
A	0.16	12	t4
B	0.16	13	t4
B	0.16	14	t4
B	0.08	15	
A	0.03	16	t5
B	0.03	17	t5
B	0.03	18	t5
B	0.02	19	
B	0.01	20	

Average tied ranks together

t1	2.5
t2	5.5
t3	8.5
t4	12
t5	17

Replace tied ranks with average tied ranks

Ranked Example

		ranks	
A	0.99	1	
A	0.91	2.5	t1
A	0.91	2.5	t1
A	0.64	5.5	t2
A	0.64	5.5	t2
B	0.64	5.5	t2
B	0.64	5.5	t2
A	0.27	8.5	t3
B	0.27	8.5	t3
A	0.16	12	t4
A	0.16	12	t4
A	0.16	12	t4
B	0.16	12	t4
B	0.16	12	t4
B	0.08	15	
A	0.03	17	t5
B	0.03	17	t5
B	0.03	17	t5
B	0.02	19	
B	0.01	20	

Average tied ranks together

t1	2.5
t2	5.5
t3	8.5
t4	12
t5	17

Replace tied ranks with average tied ranks

		ranks
A	0.99	1
A	0.91	2.5
A	0.91	2.5
A	0.64	5.5
A	0.64	5.5
A	0.27	8.5
A	0.16	12
A	0.16	12
A	0.16	12
A	0.03	17
B	0.64	5.5
B	0.64	5.5
B	0.27	8.5
B	0.16	12
B	0.16	12
B	0.08	15
B	0.03	17
B	0.03	17
B	0.02	19
B	0.01	20

Resort by treatment

Ranked Example

		ranks
A	0.99	1
A	0.91	2.5
A	0.91	2.5
A	0.64	5.5
A	0.64	5.5
A	0.27	8.5
A	0.16	12
A	0.16	12
A	0.16	12
A	0.03	17
B	0.64	5.5
B	0.64	5.5
B	0.27	8.5
B	0.16	12
B	0.16	12
B	0.08	15
B	0.03	17
B	0.03	17
B	0.02	19
B	0.01	20

Perform T Test on Ranks

	A_{rank}	B_{rank}
avg	7.85	13.15
stdDev	5.28	5.33

	Non Parm. T Test	
$s = s_A^2 + s_B^2$	7.50	
$s_T = s / \sqrt{n}$	2.37	n=10
$avg_A - avg_B / s_T$	2.23	*T-score*
p-value	0.038	

Ranked Example

A Non-Parametric 'Mean': The Median

- Average of a data set that is not normally distributed produces a value that behaves non-intuitively
  - Especially if the probability distribution is skewed
    - Large values in 'tail' can dominate
    - Average tends to reflect the typical value of the "worst" data not the typical value of the data in general
- Instead use the Median
  - 50th percentile
  - Counting from 1, it is the value in the $\frac{n+1}{2}$ position
    - If n is even, $(n+1)/2$ will be between 2 positions, average the values at that position

A Confidence Interval Around the Median: Thompson-Savur

- Find the b the binomial value that has a cumulative upper tail probability of $\alpha/2$
  - The binomial distribution is used instead of the normal distribution because median based on a two sided sign test
- The lower percentile $l = \frac{b}{n-1}$
- The upper percentile $u = 1 - l$
- Confidence Interval is $[value_l, value_u]$
  - i.e. $value_l \le median \le value_u$
  - With a confidence level of $1 - a$

A Confidence Interval Around the Median: Thompson-Savur

- In Excel:
  - To calculate b use `CRITBINOM(`n`,`1/2`,`$\alpha/2$`)`
  - to compute the $value_u$ use the function `PERCENTILE(`dataArray, u`)`
  - to compute the $value_l$ use the function `PERCENTILE(`dataArray, l`)`

A Confidence Interval Alternative to the Ranked T Test

- Find the median confidence interval for the two data sets
- If the confidence intervals do not overlap
  - Data sets from different distributions
  - With a confidence level of 1 - α where α is the upper tail probability used in computing b
  - Advantages:
    - Gives better understanding of system
      - see median values with error bounds
      - easy to draw on a graph
  - Disadvantage:
    - Not as sensitive as the ranked T test

Repetitions

- What is the number of repetitions needed to see if there is a difference between two means or between two medians?
  - Depends on the underlying distributions
    - But underlying distributions are unknown
- Rule of thumb
  - Perform a minimum of 30 repetitions for each system
  - Performing 50 to 100 repetitions is usually better

More Than 2 Treatments

- Preceeding stats to be used for simple experiment designs
- More sophisticated stats needs to be done if:
  - Comparing multiple systems instead of just 2 treatments
    - E.g. comparing the effect on a Genetic Algorithm of using no mutation, low, medium and high levels of mutation
      - We say there are 4 *levels* of the mutation variable
      - Need $\binom{4}{2} = 6$ possible comparisons to test all pairs of treatments
  - Called a 'multi-level' analysis

Multiple Levels: Post-hoc Analysis

- For 4 levels of mutation there are 6 comparisons possible
  - *Each one* of the comparison holds at a 95% C.L. independent of the other comparisons
  - If *all* comparisons are to hold at once the odds are $0.95 \times 0.95 \times 0.95 \times \ldots \times 0.95 = (0.95)^6 = 0.735$
  - So in practice we only have 73.5% C.L
    - Wrong 1/4 of the time
- For 7 levels of mutation there are 21 comparisons possible
  - C.L. = $(0.95)^{21} = 0.341$
    - Chances are better than half that at least one of the decisions is wrong!

The Bonferroni Corrections for Tests

- To correct, choose a smaller α

$$a' = \frac{a}{m}$$

  - Where *m* is the number of comparisons
  - So for 95% CL use $\alpha = 0.025/6 = 0.004167$
  - For a Z test the critical value changes from 1.96 to 2.64
- Called a Bonferroni post-hoc correction
  - There are other post-hoc techniques such as Tukey and Scheffé that can be more powerful than Bonferroni
- You should apply the Bonferroni correction:
  - To T-tests (T tests and ranked T Tests)
  - To Confidence Intervals and Error Bounds
  - Whenever you mean "all the significant results we obtained hold at once"

The Bonferroni Corrections for Experiments

- The Bonferroni Correction is more widely applicable than just for multi-level comparisons
- We really need to control for the dilution of the confidence levels throughout the study, whether or not the CLs are applied to analyses of independent 'phenomena'
- We must *divide* the α used for each CL test by the total number of CL tests in the study
- To apply the Bonferroni correction to p-values *multiply* the p-values by the number of CL tests performed
  - "Probabilities" bigger than 1 means "not significant"

The Bonferroni Correction for Experiments

- Example:
  - A robot dog has been created
    - Genetic Programming is used to control the ear wiggles of the robot
    - a Genetic Algorithm is used to optimize its tail wagging ability
  - A study is being done to improve both the ears and the tail independently, and we want to be 95% confident in our over all tests
    - For the ears the GP is tested with 3 different sets of terminal nodes
    - For the tail the GA is tested with 4 different fitness functions
    - There are $\binom{3}{2} + \binom{4}{2} = 3 + 6 = 9$ total CL inferences used in the study
    - Consequently the a used for any CL should be $a = 0.025 / 9 = 0.0028$

Multiple Factors

- Most of the time, there are many different properties we are interested in studying
  - e.g. We may be trying out various kinds of crossovers, with and without mutation, under different selection pressures
  - Each of the above parameters has multiple levels
  - This is called a multiple factor analysis
    - with each factor having multiple levels
  - Use Analysis of Variance or General Linear Models to analyze
    - See text books on ANOVA and GLMs

Multiple Factors: Factorial Design

- When dealing with multiple factors with multiple levels
  - Important that all combinations of factor levels are tried
  - A given combination of factor levels is called a treatment
  - If you want accurate information about each possible interaction, each treatment should be repeated at least 30 times
    - If you interested largely in main effects, 10 repetitions is often fine, if you have enough levels

Multiple Factors: Factorial Design

E.g. if we have 2 EC systems, new and standard (New and Std) and we want to see their behavior under

- crossover and no crossover (x and ~~x~~)
- 3 different selection pressures (p1, p2 and p3)

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
S	New	New	New	New	New	New	Std	Std	Std	Std	Std	Std
X	x	x	x	~~x~~	~~x~~	~~x~~	x	x	x	~~x~~	~~x~~	~~x~~
P	p1	p2	p3	p1	p2	p3	p1	p2	p3	p1	p2	p3

Multiple Factors: Factorial Design

- If we are performing 50 reps per treatment
  - In previous example we have $S \times X \times P \times 50 = 2 \times 2 \times 3 \times 50 = 12 \times 50 = 600$ experiments to perform
- The number of experiments goes up as the product of the number of levels in each factor
  - This is exponential in the number of factors
  - Consequently, carefully choose the factors and factor levels that you study in your experiments
  - Minimize what factors you vary (focus your experiments on the relevant factors)