ABI Summer 2021 Introduction to Statistics

Guest Session 2

Javier Cabrera*, Volha Tryputsen** & Davit Sargsyan**

July 7, 2021

(*)Rutgers, the State University of New Jersey
(**)Janssen Pharmaceutical Research and Development (J&J)

Know Your Tools

- What are the research questions you are trying to answer?
- What data do you need to answer them?
- Can you design an experiment that will provide you with the data? Think about the experimental design long and hard!
- What are the right tools, i.e., appropriate statistical techniques that can help you answer the questions?
- Do you understand the statistical methods well enough to be confident with your inference? Things to consider:
  - Distributions
  - Missing data
  - Outliers

Research Questions

Ultimately, we are developing a drug that can cure or manage a disease

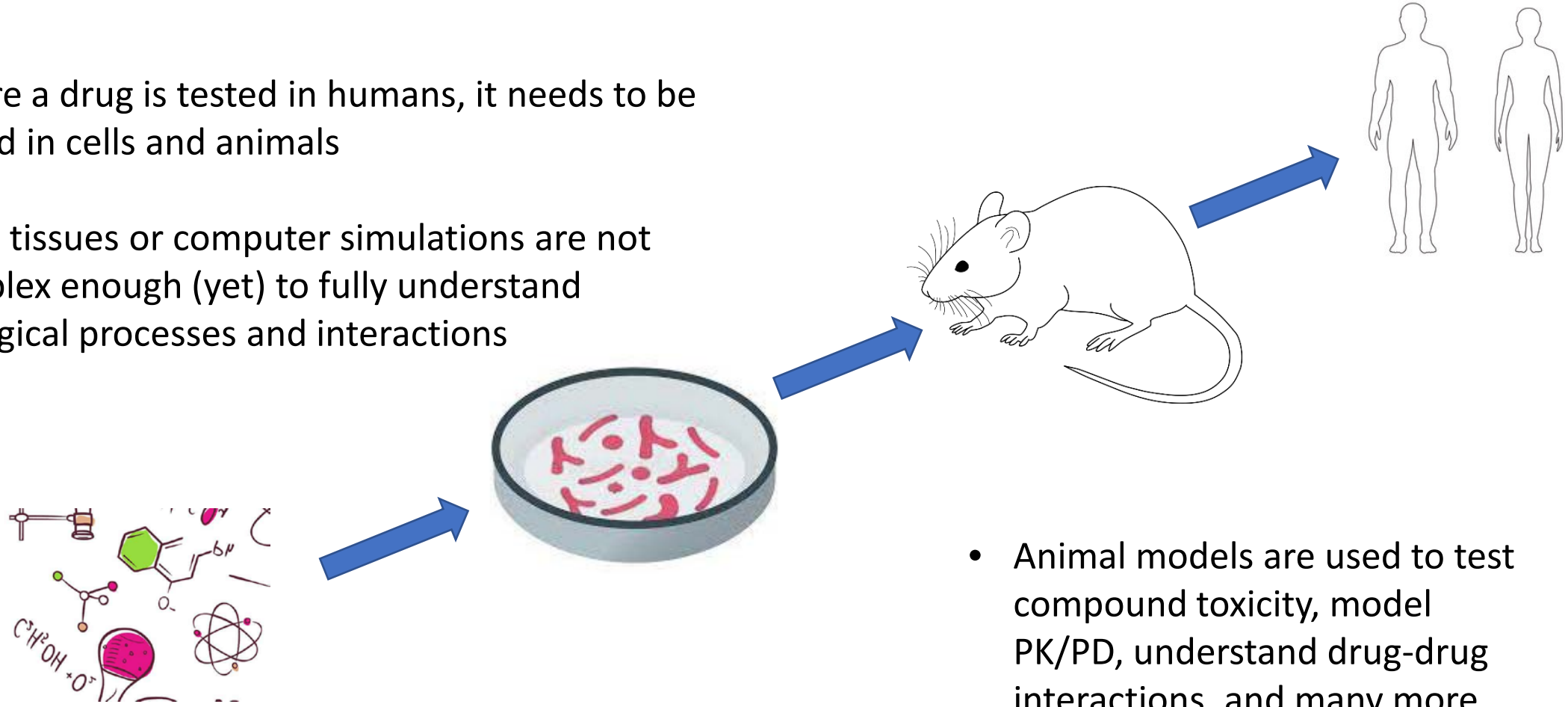
Is my drug better than the standard treatment or placebo?

Depending on a therapeutic area, these can be:

- Can the drug reduce or eliminate cancer (oncology)
- Can the drug reduce inflammation (immunology)
- Can the drug help reduce weight, correct blood pressure, help damaged heart, regulate insulin (cardio-vascular and metabolism)
- Can the drug eliminate harmful bacteria and viruses and prevent new infections (infectious diseases and vaccines)

Animal Models in Drug Development

- Before a drug is tested in humans, it needs to be tested in cells and animals
- Cells, tissues or computer simulations are not complex enough (yet) to fully understand biological processes and interactions

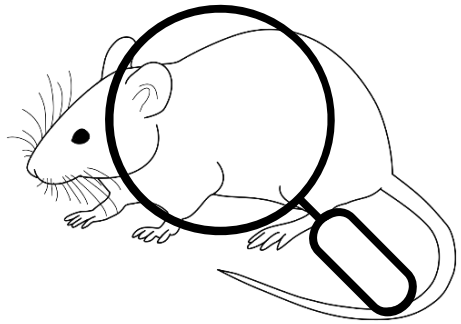


- Animal models are used to test compound toxicity, model PK/PD, understand drug-drug interactions, and many more

Know Your Tools

- What are the research questions you are trying to answer?
- What data do you need to answer them?
- Can you design an experiment that will provide you with the data? Think about the experimental design long and hard!
- What are the right tools, i.e., appropriate statistical techniques that can help you answer the questions?
- Do you understand the statistical methods well enough to be confident with your inference? Things to consider:
  - Distributions
  - Missing data
  - Outliers

Study Endpoints

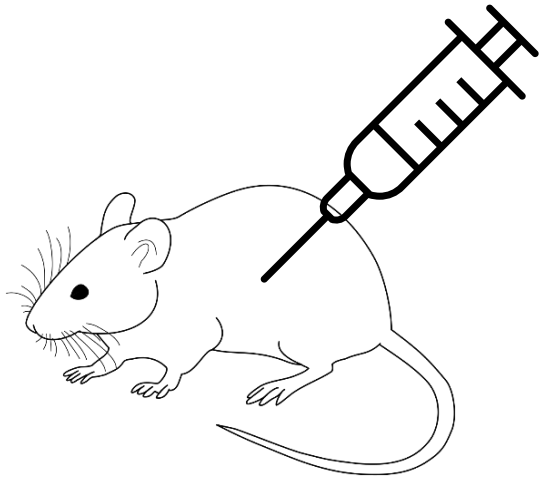


Endpoints that can be observed directly:

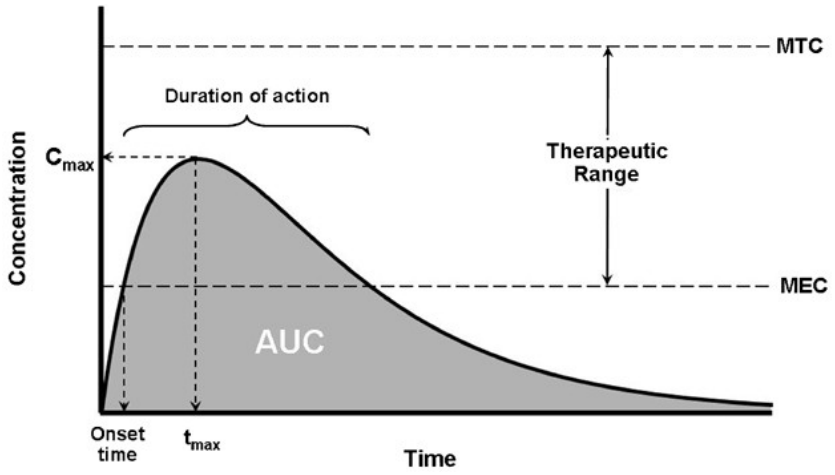
- Survival rate and time
- Tumor volume
- Body weight change
- Skin inflammation
- Bacterial count

Biomarkers from blood or tissue

- Protein concentration (Western blot, proteomics)
- Gene expression (microarrays, qPCR, RNA- and DNA-seq, DNA methylation and acetylation)
- Cell counting (flow cytometry)

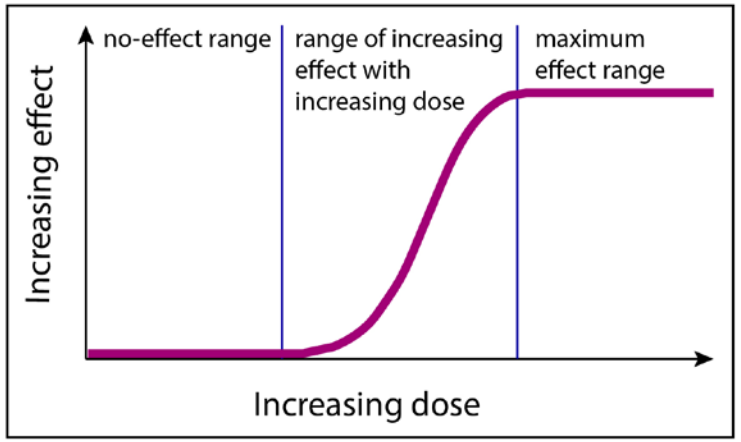


Pharmacokinetic/Pharmacodynamics (PK/PD)



Pharmacokinetic: what a body does to a drug

- Drug concentration in blood or urine over time
- Estimate large number of parameters: total drug exposure (AUC), distribution volume, drug elimination half-life, maximum concentration, renal clearance, etc.



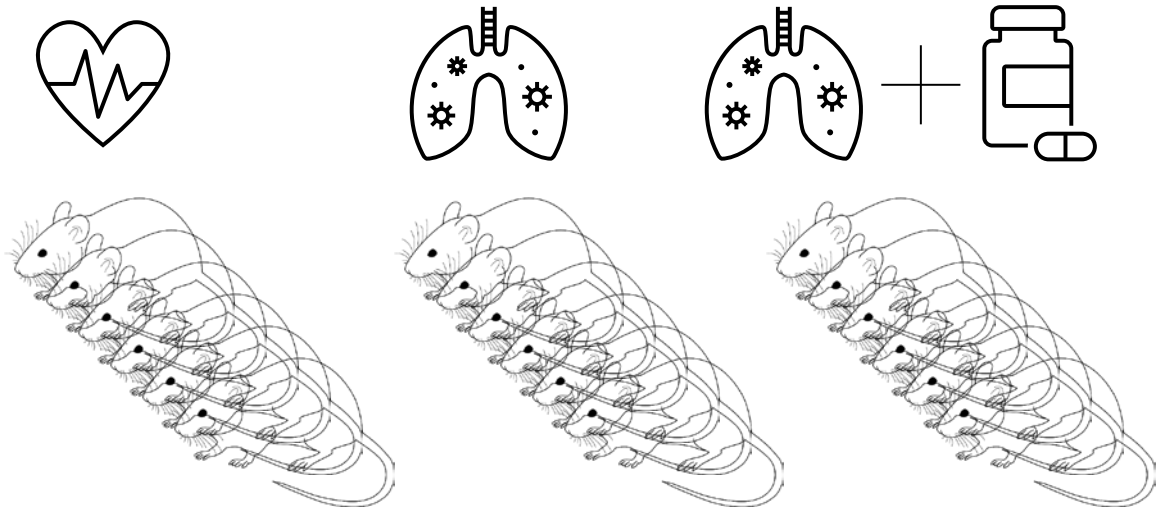
Pharmacodynamics: what a drug does to a body

- Estimate dose-response or exposure-response relationship
- Calculate parameters: top and bottom asymptotes, slope, EC50/IC50, etc.

Know Your Tools

- What are the research questions you are trying to answer?
- What data do you need to answer them?
- Can you design an experiment that will provide you with the data? Think about the experimental design long and hard!
- What are the right tools, i.e., appropriate statistical techniques that can help you answer the questions?
- Do you understand the statistical methods well enough to be confident with your inference? Things to consider:
  - Distributions
  - Missing data
  - Outliers

Basics of Animal Study Design



Group A Group B Group C

Is my drug working?

- What is the range of values for my endpoints in healthy animals? Use naïve animals as controls (**Group A**)
- Are they sufficiently different in sick animals? Use sick untreated animals as negative controls (**Group B**).
- Can the drug prevent or reverse the effect of the disease? Use the drug in sick animals (treatment group, **Group C**)

Study Design Considerations

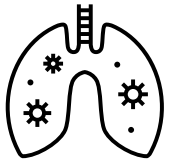
- What questions do we need to answer with the study? The research hypotheses must be negatable!
- What are the primary, i.e., the most important endpoints of the study?
- What is the criteria for success? Minimum difference in primary endpoints that is **biologically meaningful**.
- How much variability we expect? Use published data and previous experiments.
- How many control and treatment groups do we need?
- How many animals do we need that, given all of the above, we have a good chance to show **statistically significant** results?
- What variables and experimental conditions can we control for?

Know Your Tools

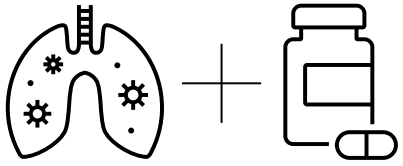
- What are the research questions you are trying to answer?
- What data do you need to answer them?
- Can you design an experiment that will provide you with the data? Think about the experimental design long and hard!
- What are the right tools, i.e., appropriate statistical techniques that can help you answer the questions?
- Do you understand the statistical methods well enough to be confident with your inference? Things to consider:
  - Distributions
  - Missing data
  - Outliers

Comparing groups

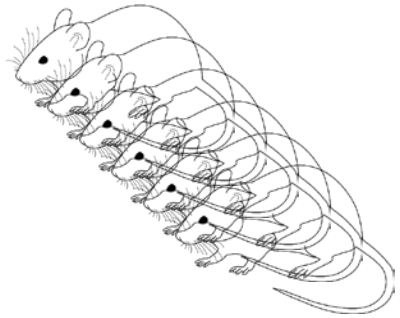
Group A: healthy

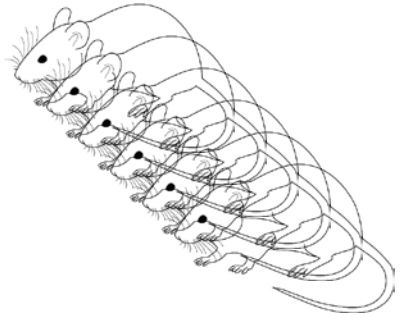


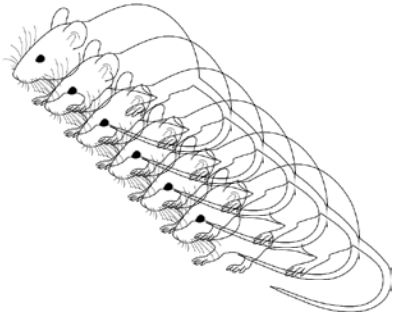
Group B: sick



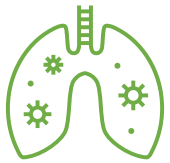
Group C: treated sick



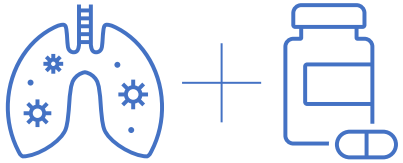




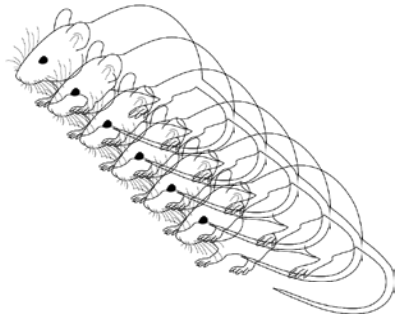
1. Comparing 2 groups

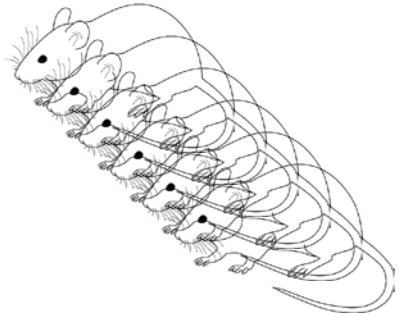


Group B: sick

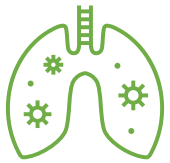


Group C: treated sick

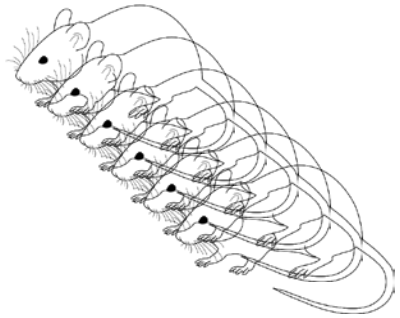


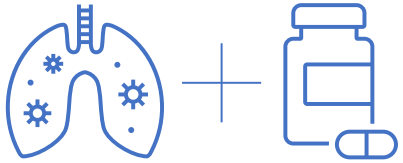


Mouse weight

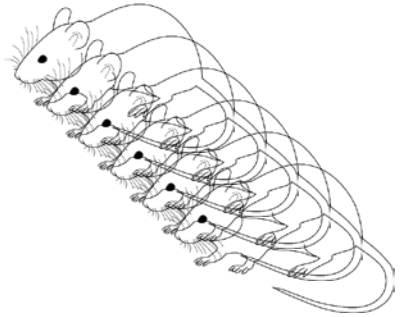


Group B: sick

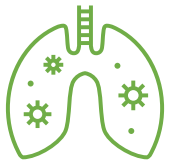




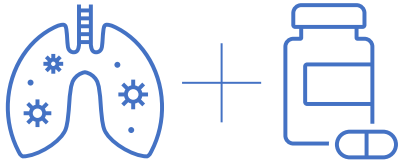
Group C: treated sick



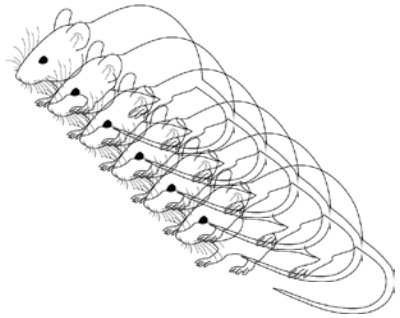
Weight distributions

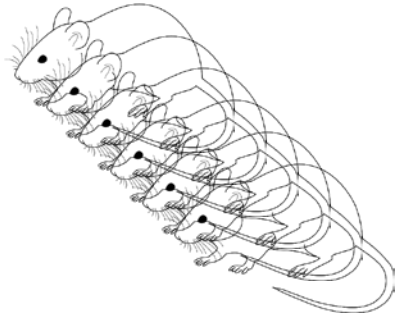


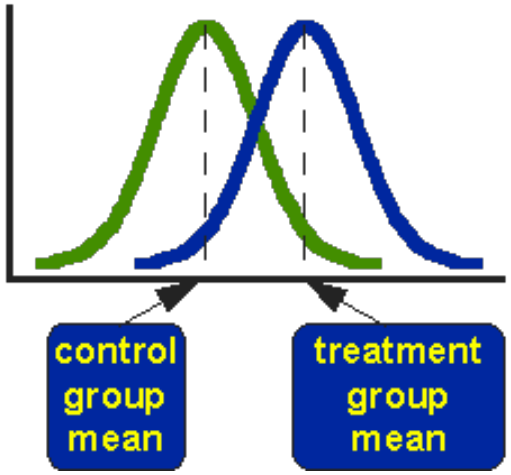
Group B: sick



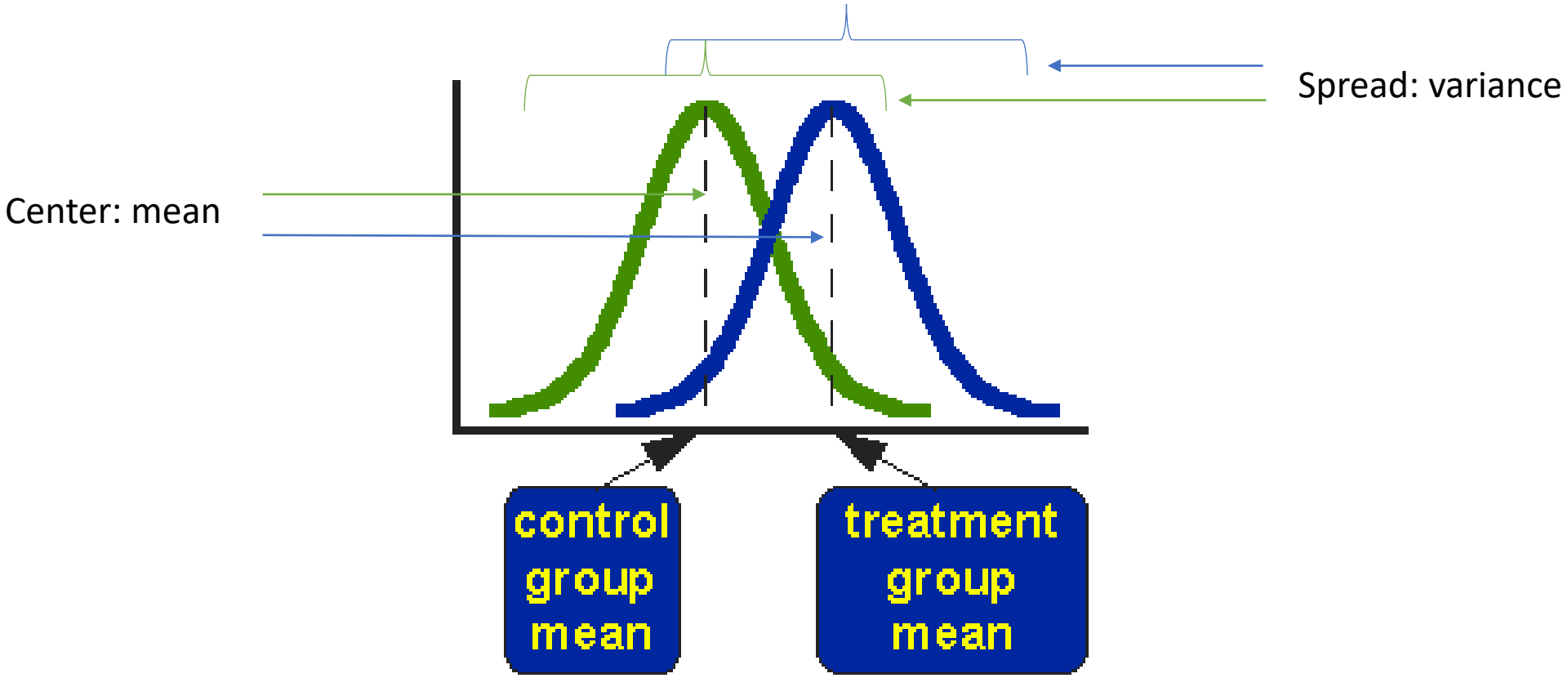
Group C: treated sick



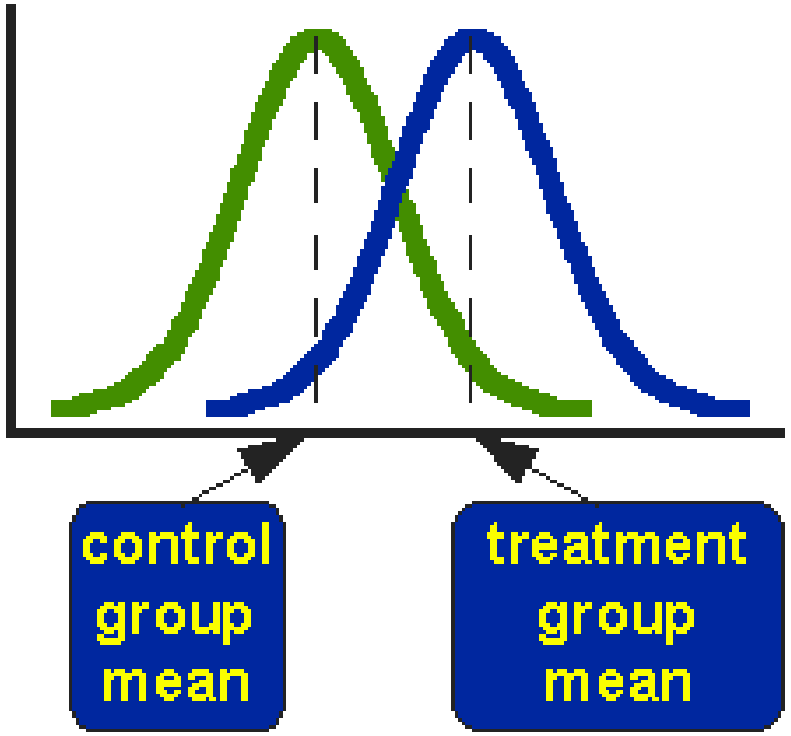




Weight distribution`s center and spread

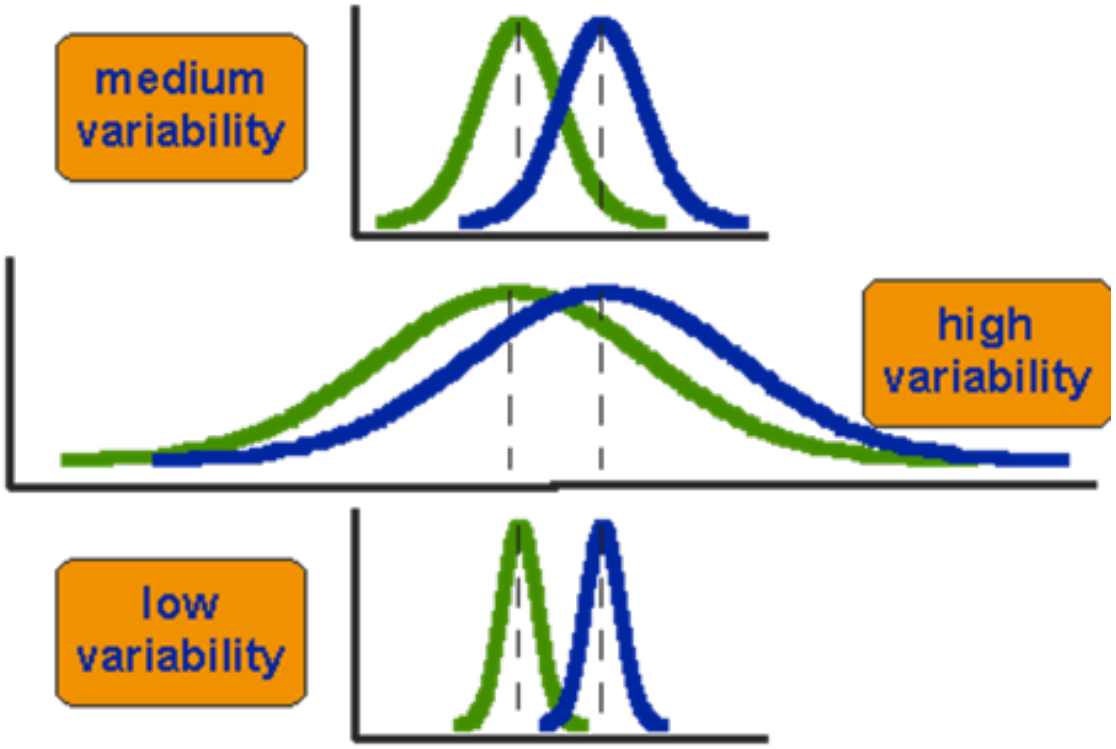


Comparing 2 means



Question: Are mean weights in control are treated groups statistically different?

Difference between means - 3 scenarios

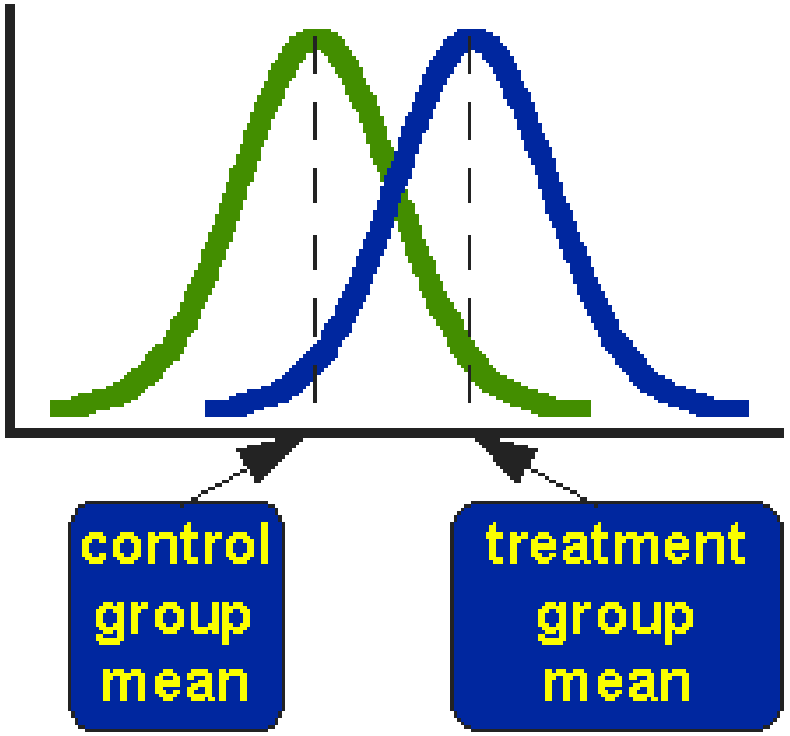


1. The difference between the means is the same in all three.

2. The two groups appear most different or distinct in the bottom or low-variability case.

3. When we are looking at the differences between weights for 2 groups, we must judge the difference between their means relative to variability!

Comparing 2 Groups' Means: t-Test

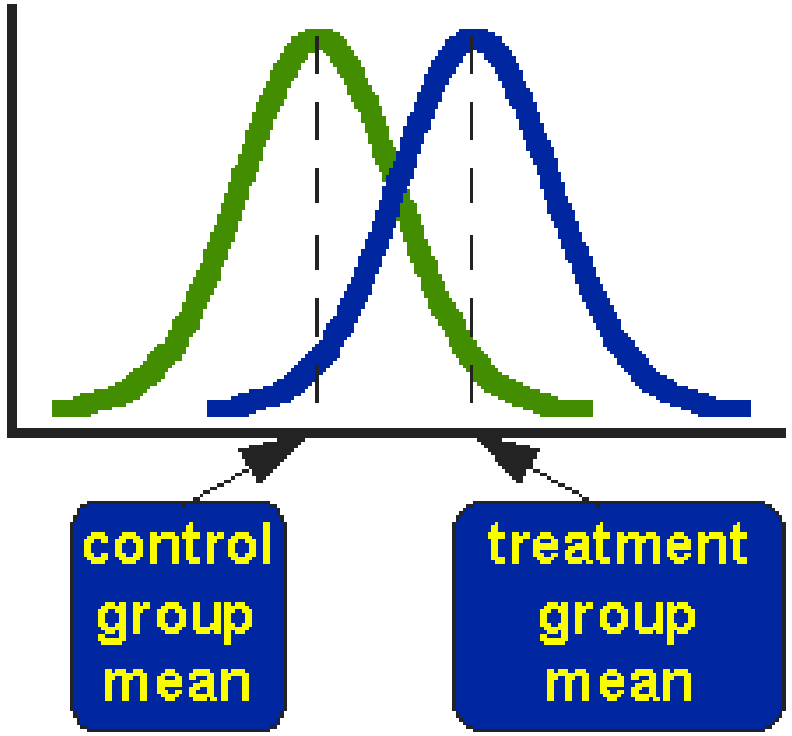


$$\frac{signal}{noise} = \frac{difference\ between\ two\ means}{groups\ variance} =$$

$$\frac{\overline{X_T} - \overline{X_C}}{SE(\overline{X_T} - \overline{X_C})} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}} = t - value$$

$$\text{SE}(\bar{X}_T - \bar{X}_C) = \sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}$$

Comparing 2 Groups' Means: t-Test



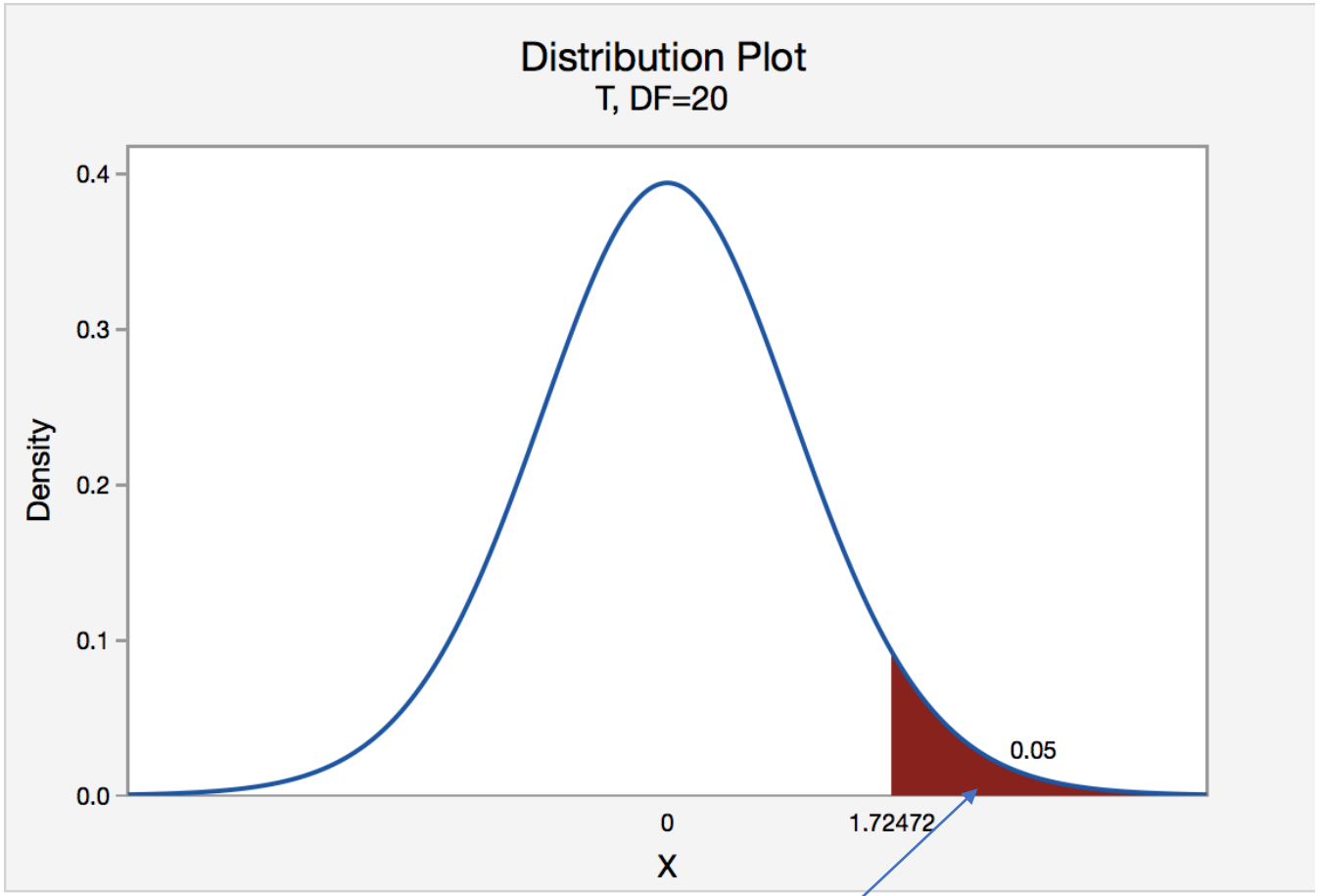
$$\frac{signal}{noise} = \frac{difference\ between\ two\ means}{groups\ variance} =$$

$$\frac{\overline{X_T} - \overline{X_C}}{SE(\overline{X_T} - \overline{X_C})} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\mathrm{var}_T}{n_T} + \frac{\mathrm{var}_C}{n_C}}} = t - value$$

number of animals in treated and control groups

Variance in treated and control groups

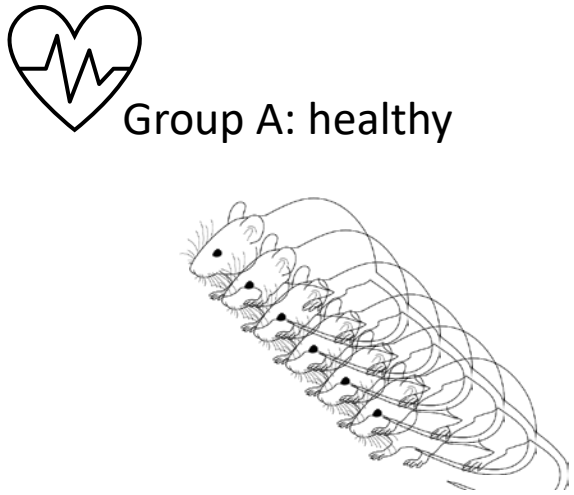
Comparing 2 Groups' Means: t-Test

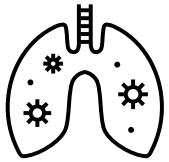


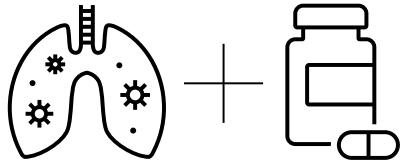
- Compute the t-value.
- To test the significance, you need to set a significant level alpha (usually alpha = 0.05). This means that 5% of time you would find a statistically significant difference between the means even if there was none (i.e., by "chance").
- You also need to determine the degrees of freedom (df) for the test = the sum of the mice in both groups minus 2.
- Given the alpha level, the df, and the t-value, you can determine whether the t-value is large enough to be significant.

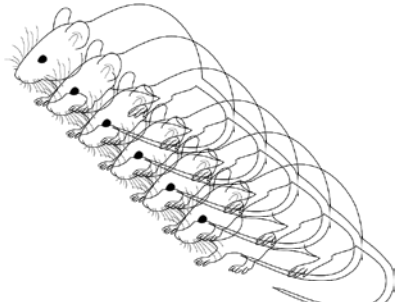
If it is, you can conclude that the difference between the means for the two groups is different.

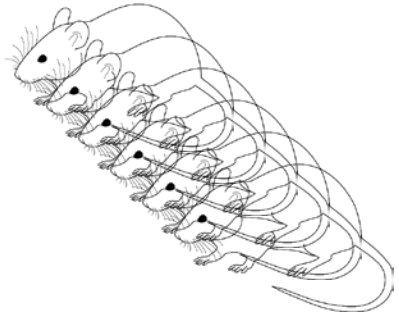
Comparing More Than 2 Groups' Means with ANOVA

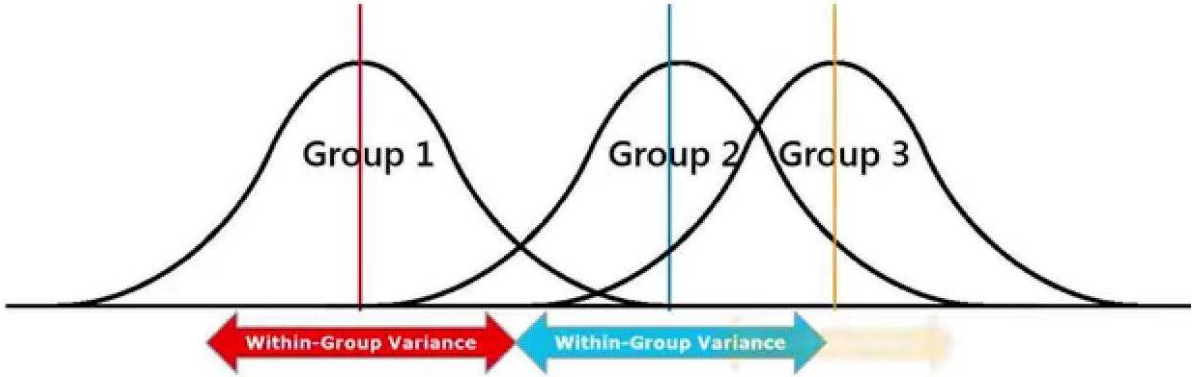

Group A: healthy


Group B: sick


Group C: treated sick





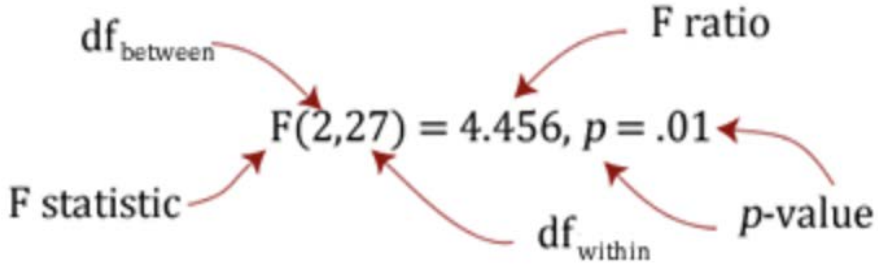


- Null: All group means are equal.
- Alternative: Not all group means are equal.

Comparing More Than 2 Groups' Means with ANOVA

Source	*SS*	*df*	*MS*	*F*	*Sig.*
Between	SS_b	k-1	MS_b	MS_b/MS_w	*p* value
Within	SS_w	N-k	MS_w		
Total	$SS_b + SS_w$	N-1			

Source	*SS*	*df*	*MS*	*F*	*Sig.*
Between	91.476	2	45.733	4.467	.021
Within	276.400	27	10.237		
Total	367.867	29			



Comparing More Than 2 Groups' Means with ANOVA and post hoc

The ANOVA test tells you whether you have an overall difference between your groups, but it does not tell you which specific groups differed – post hoc tests do.

Because post hoc tests are run to confirm where the differences occurred between groups, they should only be run when you have a shown an overall statistically significant difference in group means.

Post-hoc tests:
- **Tukey's honestly significant difference (HSD)** post hoc test - for all pair-wise comparison -
- **Dunnett`s test** – for all treatment to control comparisons

Group Assignment

Group 1	Group 2	Group 3	Group 4	Group 5
Gisane (R)	Arabo (R)	Susie (R)	Tamara (R)	Mher (R)
Arvin	Nerses	Ashkhen	Vardan	Vika
Satenik	Liana	Anush (R)	Hripsime	Nelli

The Royal Guinea Pig Problem

- *Once upon a time there were King and Queen*
- *They had a favorite pet – Peppa the Guinea Pig*
- *And life was great – until one day…*

Source: https://www.123rf.com/photo_42732599_a-cartoon-queen-with-a-confused-expression-.html

The Royal Guinea Pig Problem

- *Once upon a time there were King and Queen*
- *They had a favorite pet – Peppa the Guinea Pig*
- *And life was great – until one day… Peppa became very ill*

The Royal Guinea Pig Problem

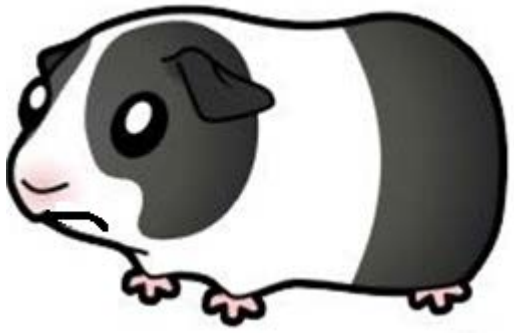
- The Guinea pig lost appetite and began losing weight . It even stopped eating its favorite treat - carrot

Available resources:

- Castle basement full of mice and rats
- 2 types of diet: regular grass or carrot-rich
- 3 natural compounds to use as treatments

- You are commissioned to find a cure for Peppa – and get half of the Kingdom
- Or else…

Peppa's Information
Age: 5 years 2 months; Sex: Female, Skin Color: Pink;
Favorite snack: carrots



	Last Week	This Week
Diet	Grass and Carrots	Grass
Weight	550g	423g
Food Intake	128g/day	105g/day

Now, Lets Design the Experiment!

Research Question(s):

1. Diagnose Peppa first! You noticed inflammation in the guts.
2. Does the treatment make a difference?
3. Is diet affecting weight?

Endpoints:

1. Body weight
2. Food intake
3. Activity

Define success:

1. Significant difference between sick and treated mice body weight
2. No significant body weight change from the baseline weight

Study design:

1. 5 Treatment groups: Naïve, negative control (harsh chemical to make mice sick), 3 treatments (all sick animals). We will not consider combinations
2. Diet: grass only vs crass + carrots

Data analysis and visualization:

1. t-Test: Naïve vs. Negative control to see if the inflammation made the animals sick
2. ANOVA + pairwise comparison: what treatment9s) work?
3. Dot plots, histograms, boxplots

Reporting :

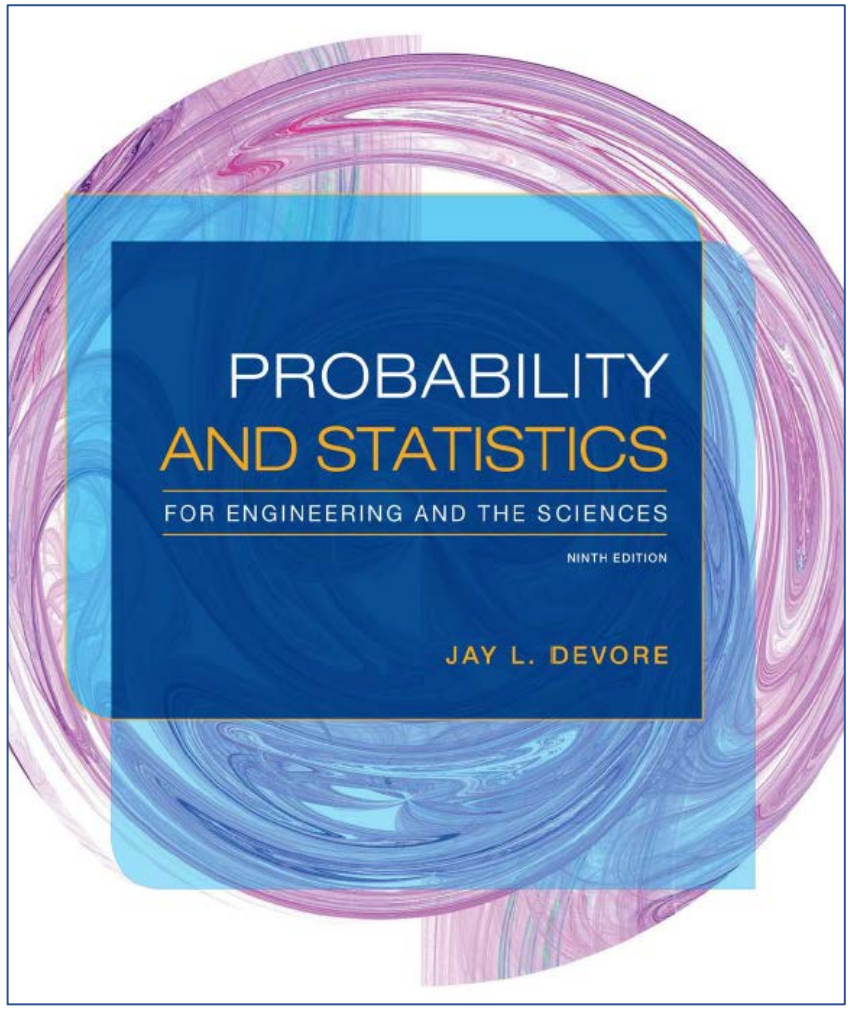
Explain your results clearly to a non-scientists (aka King and Queen)

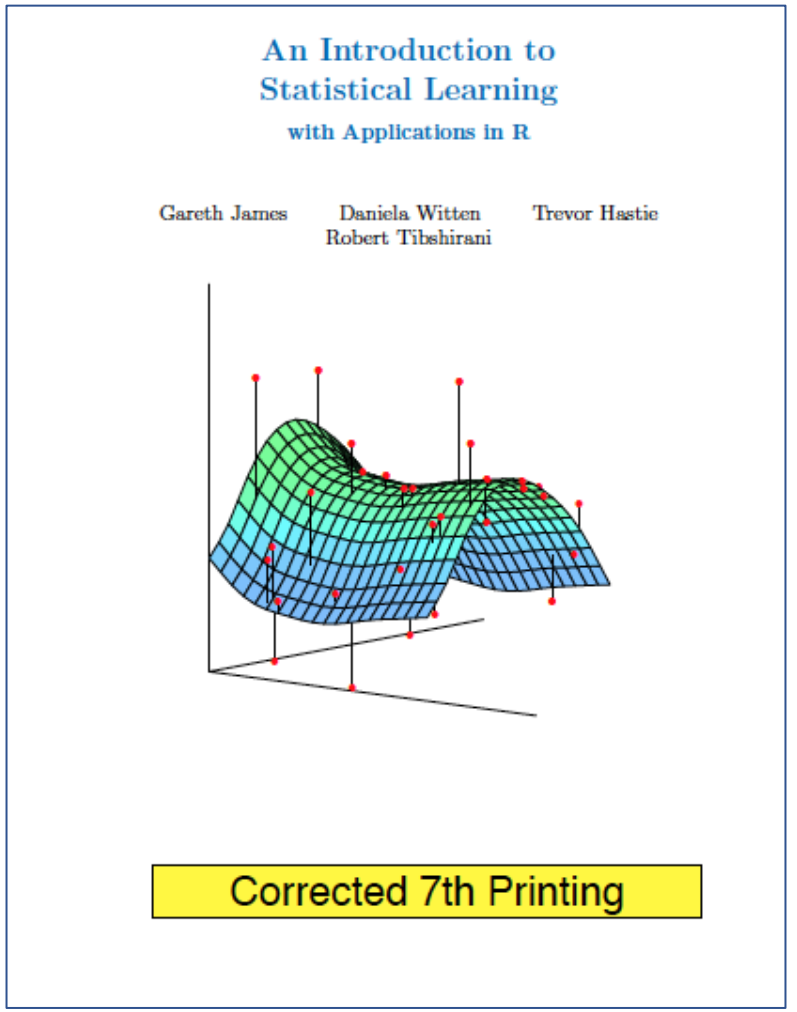
Homework (Due Next Tuesday)

- Clone our GitHub https://github.com/sargdavid/abi_statistics
- Find the data file in the “data” folder
- Visualize the data (dot plot, histogram, boxplot)
- Data analysis (t-Test, ANOVA with pair-wise comparisons)
- Report the results: what is the best treatment available, and how confident you are that it will work. Hint: take into account the patient’s preferences
- Report limitations of your study

Questions

References





https://www.statlearning.com/

Topics in Statistics Level I

Topics in Statistics Level II

10 The Analysis of Variance

11 Multifactor Analysis of Variance

12 Simple Linear Regression and Correlation

13 Nonlinear and Multiple Regression

15 Distribution-Free Procedures

Topics in Statistics Level III

Probability: Central Limit Theorem and Bayes' Theorem

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, $\overline{X}$ has approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma^2_{\overline{X}} = \sigma^2/n$, and T_o also has approximately a normal distribution with $\mu_{T_o} = n\mu$, $\sigma^2_{T_o} = n\sigma^2$. The larger the value of n, the better the approximation.

If a statistic T is some linear combination of RV's Then for large samples T / SE(T) is approximately normally distributed regardless of the population of the data. This is the fundamental theorem of all statistics.

Bayes' Theorem

Let $A_1, A_2, \ldots, A_k$ be a collection of k mutually exclusive and exhaustive events with *prior* probabilities $P(A_i)$ $(i = 1, \ldots, k)$. Then for any other event B for which $P(B) > 0$, the *posterior* probability of A_j given that B has occurred is

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)} \quad j = 1, \ldots, k \qquad (2.6)$$

Suppose that we assume a model for the Data based on some parameter-s θ, P(Data $|\theta$)
And suppose we give a prior distribution to θ, P(θ).
P(θ) represents our prior believe before collecting the data and then we collect the data Data.

Then Bayes theorem gives us the posterior distribution of θ given the observe Data
P(θ | Data) incorporating the observed data to the distribution of θ.

Probability homework

Q 1:

A is a disease : Prostate Cancer {A1, A2}

A1: Healthy A2: Prostate cancer

B is a genotype variable, it has 3 possible values {B1, B2, B3}

The following tables give the conditional probabilities of A and of B|A

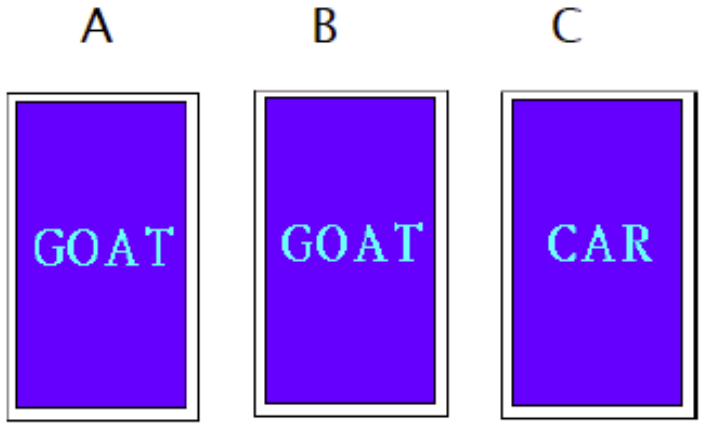
	Prob(Bj|Ai)	
Genotype	A1: Healthy	A2: Prostate C.
B1	0.2	0.8
B2	0.4	0.1
B3	0.4	0.1
Total	1	1
Disease	A1: Healthy	A2: Prostate C.
Prob	0.8	0.2

The Data: For each patient we measure the genotype either B1, B2 or B3.

Calculate the posterior distribution of A (prostate cancer) given a genotype Bj.

Q 2:

Goat game (a TV Game): Behind three closed doors, one has a car, and two have a goat.



a. Player chooses one door, say door C

b. Game host shows a goat behind a door, say door A

c. There are two closed doors left B, C. The player is asked if he would like to switch.

Question : Is it better to switch (S) or not to switch (S^c)?

The Two-Sample *t* Test and t-Confidence Interval

Pooled t-Interval Procedure

Purpose To find a confidence interval for the difference between two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1 **For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n_1 + n_2 - 2$.**

Step 2 **The endpoints of the confidence interval for $\mu_1 - \mu_2$ are**

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

where s_p is the pooled sample standard deviation.

Step 3 **Interpret the confidence interval.**

Note: The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

Pooled t-Test

Purpose To perform a hypothesis test to compare two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1 **The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is**

H_a: $\mu_1 \neq \mu_2$ (Two tailed) or H_a: $\mu_1 < \mu_2$ (Left tailed) or H_a: $\mu_1 > \mu_2$ (Right tailed)

Step 2 **Decide on the significance level, α.**

Step 3 **Compute the value of the test statistic**

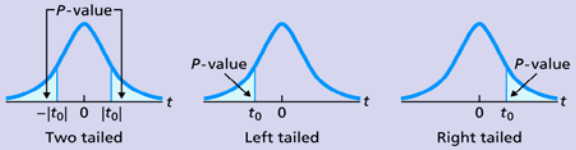
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{(1/n_1) + (1/n_2)}},$$

where

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

Denote the value of the test statistic t_0.

Step 4 **The t-statistic has $\text{df} = n_1 + n_2 - 2$. Use Table IV to estimate the P-value, or obtain it exactly by using technology.**



Step 5 **If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.**

Single-Factor ANOVA

- The relevant hypotheses are
- H_0: $\mu_1 = \mu_2 = \cdots = \mu_I$ *(POPULATION MEANS ARE EQUAL)*

- versus
- H_a: at least two the of the μ_i's are different

- If $I = 4$, H_0 is true only if all four μ_i's are identical. H_a would be true, for example, if

$$\mu_1 = \mu_2 \neq \mu_3 = \mu_4, \text{ if } \mu_1 = \mu_3 = \mu_4 \neq \mu_2,$$

  or if all four μ_i's differ from one another.
- Assumptions
  - Samples are independent
  - Populations are Normal
  - POPULATION VARIANCES ARE EQUAL

The Test Statistic

Mean square for treatments is given by

$$\text{MSTr} = \frac{J}{I-1}[(\overline{X}_{1\cdot} - \overline{X}_{\cdot\cdot})^2 + (\overline{X}_{2\cdot} - \overline{X}_{\cdot\cdot})^2 + \cdots + (\overline{X}_{I\cdot} - \overline{X}_{\cdot\cdot})^2]$$
$$= \frac{J}{I-1}\sum_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

and **mean square for error** is

$$\text{MSE} = \frac{S_1^2 + S_2^2 + \cdots + S_I^2}{I}$$

The test statistic for single-factor ANOVA is $F = \text{MSTr/MSE}$.

Theorem

Let $F = \text{MSTr/MSE}$ be the test statistic in a single-factor ANOVA problem involving I populations or treatments with a random sample of J observations from each one. When H_0 is true and the basic assumptions of this section are satisfied, F has an F distribution with $\nu_1 = I - 1$ and $\nu_2 = I(J-1)$. Because a larger f is more contradictory to H_0 than a smaller f, the test is upper-tailed:

$$P\text{-value} = P(F \geq f \text{ when } H_0 \text{ is true})$$
$$= \text{area under the } F_{I-1,I(J-1)} \text{ curve to the right of } f$$

Statistical software will provide an exact P-value. Refer to Section 9.5 for a description of how our book's table of F critical values, Table A.9, can be used to obtain an upper or lower bound (or both) on the P-value.

Multiple Comparisons in ANOVA

- Several of the most frequently used procedures are based on the following central idea.
- First calculate a confidence interval for each pairwise difference $\mu_i - \mu_j$ with $i < j$. Thus if $I = 4$. the six required CIs would be for $\mu_1 - \mu_2$ (but not also for $\mu_2 - \mu_1$), $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_2 - \mu_3$, $\mu_2 - \mu_4$, and $\mu_3 - \mu_4$.
- Then if the interval for $\mu_1 - \mu_2$ does not include 0, conclude that μ_1 and μ_2 *differ significantly* from one another; if the interval does include 0, the two μ's are judged not significantly different.

Tukey's Procedure (the *T* Method)

With probability $1 - \alpha$,

$$\bar{X}_{i\cdot} - \bar{X}_{j\cdot} - Q_{\alpha,I,IJ-1)}\sqrt{\text{MSE}/J} \le \mu_i - \mu_j \le \bar{X}_{i\cdot} - \bar{X}_{j\cdot} + Q_{\alpha,I,IJ-1)}\sqrt{\text{MSE}/J} \quad (10.4)$$

for *every* i and j $(i = 1,\ldots, I$ and $j = 1,\ldots, I)$ with $i < j$.

2^p Factorial Experiments

- If an experimenter wishes to study simultaneously the effect of p different factors on a response variable and the factors have $I_1, I_2, \ldots, I_p$ levels, respectively, then a complete experiment requires at least $I_1 \cdot I_2, \cdot \ldots \cdot I_p$ observations.

  In such situations, the experimenter can often perform a "screening experiment" with each factor at only two levels to obtain preliminary information about factor effects.

  An experiment in which there are p factors, each at two levels, is referred to as a **2^p factorial experiment.**

2^3 Experiments

As in Section 11.3, we let X_{ijkl} and x_{ijkl} refer to the observation from the lth replication, with factors A, B, and C at levels i, j, and k, respectively. The model for this situation is

$$X_{ijkl} = \mu + \alpha_i + \beta_j + \delta_k + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} + \gamma_{ijk} + \epsilon_{ijkl}$$

for $i = 1, 2$; $j = 1, 2$; $k = 1, 2$; $l = 1, \ldots, n$. The $\in_{ijkl}$'s are assumed independent, normally distributed, with mean 0 and variance σ^2.

Because there are only two levels of each factor, the side conditions on the parameters of (11.14) that uniquely specify the model are simply stated:

$$\alpha_1 + \alpha_2 = 0, \ldots, \gamma_{11}^{AB} + \gamma_{21}^{AB} = 0, \gamma_{12}^{AB} + \gamma_{22}^{AB} = 0,$$

$$\gamma_{11}^{AB} + \gamma_{12}^{AB} = 0, \gamma_{21}^{AB} + \gamma_{22}^{AB} = 0,$$

and the like.
These conditions imply that there is only one functionally independent parameter of each type (for each main effect and interaction).