Saanich & Statistics

MICB425 Project 1
2018 Term 2

Kim Dill-McFarland kadm@mail.ubc.ca
Steven Hallam shallam@mail.ubc.ca

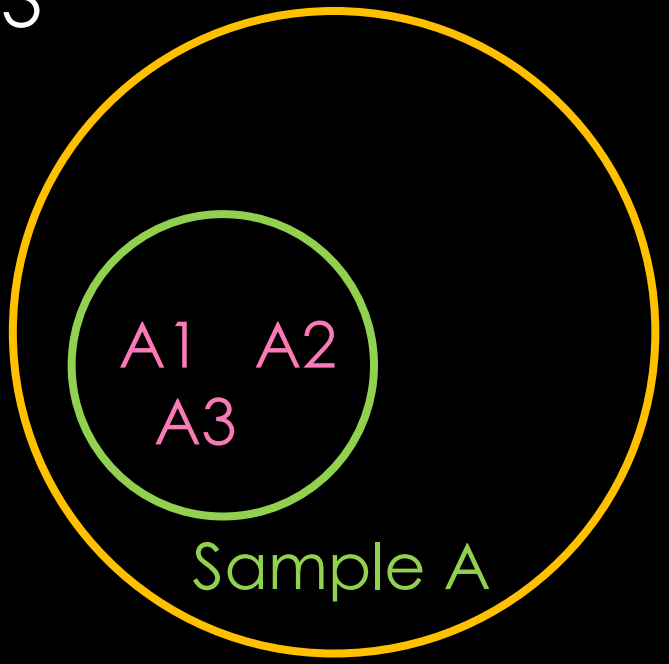
Learning objectives

Students will be able to:

- Distinguish between a population and a sample, and between parameters and statistics
- Apply t-tests, ANOVA, or linear models to appropriate data sets
- Run t-tests, ANOVA, and linear models in RStudio and interpret the output

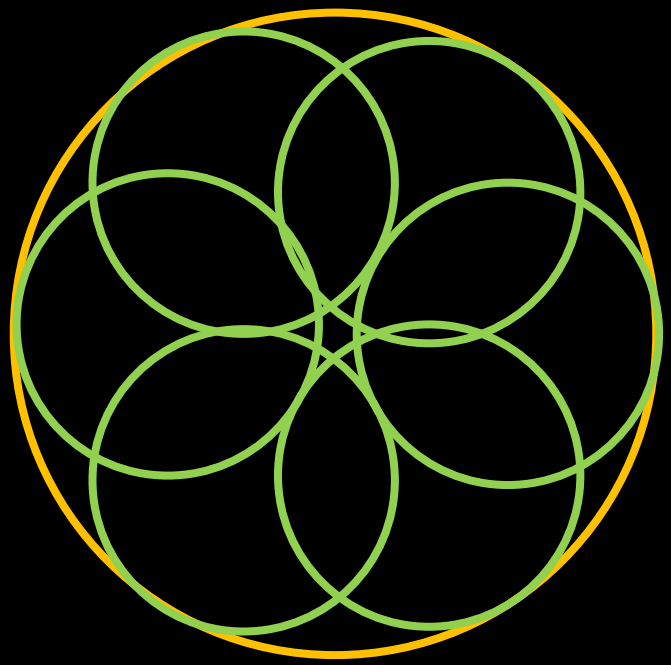
Observation vs. Sample vs. Population

- Observations A1, A2, A3
  - Small bags taken home
- Sample A
  - All small bags taken by your group
- Population A
  - Original large Ziploc of candies



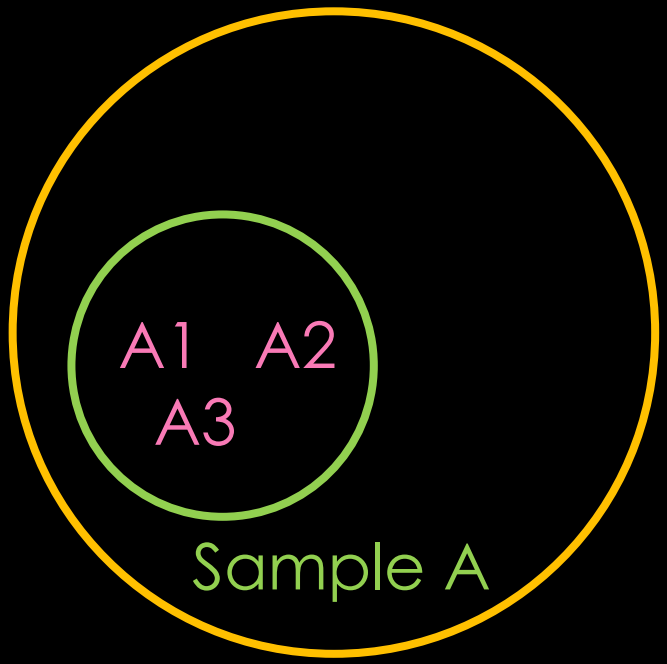
Many samples

- **Many** samples of **enough** observations accurately represents the original population



All candies in Population A

Statistic vs. parameter



- Use a sample to calculate ***statistics***
- Test hypotheses using ***statistics*** to draw conclusions about the true ***parameters*** of the overall population

Project 1

- Population of microbes at a given depth in Saanich Inlet: unknown/ unmeasurable
- A sample taken from each given depth with 1 observation for each depth
- No true replication at any given depth
- Pseudo-replication across the water column

Biological hypothesis

Alpha-diversity differs with oxygen levels in Saanich Inlet.

t-test

- Compare 2 *sample means* in order to make a statistical inference about the underlying *population means*

- Null hypothesis H_0: the *population* means are not different
- Alternate hypothesis H_1: the *population* means are different

- Application: Does alpha-diversity differ between oxic and anoxic environments (pseudo-replicates)?

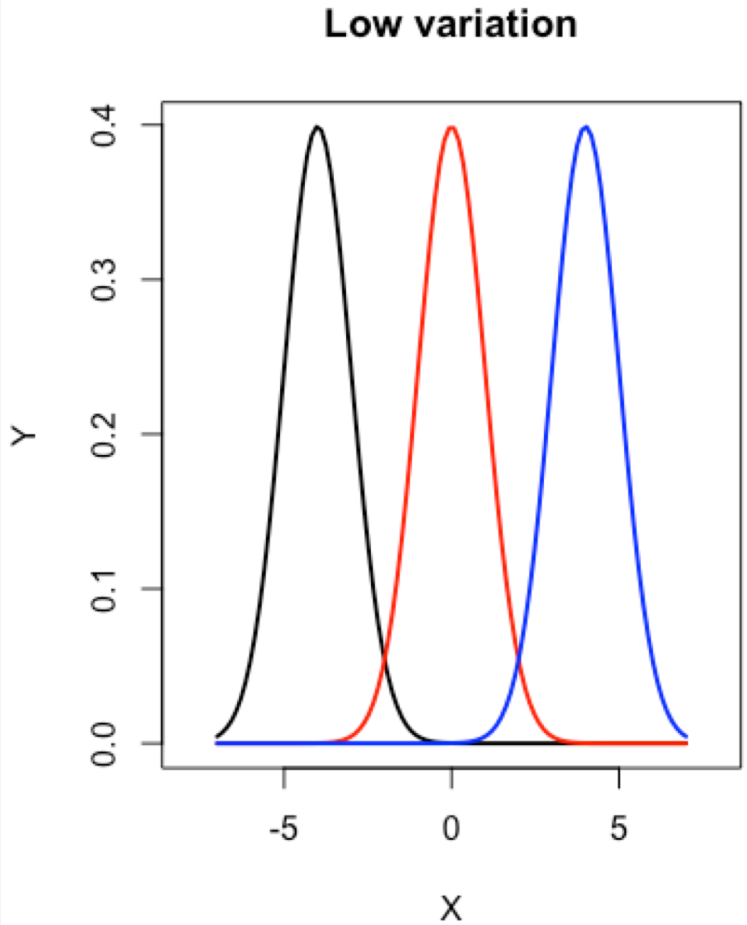
t-test assumptions

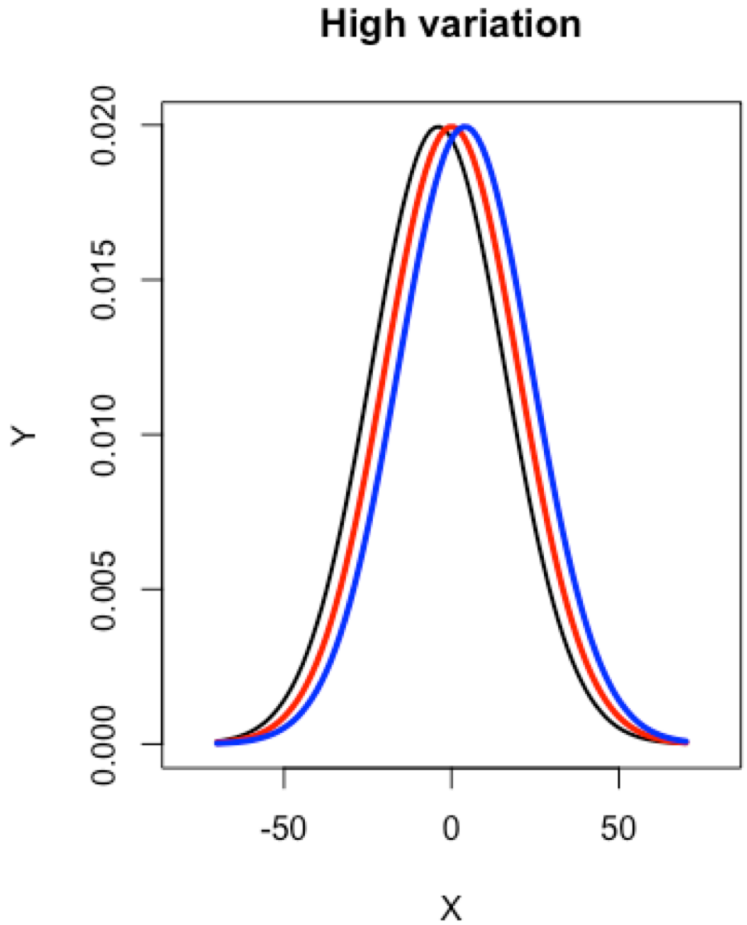
1. Simple random sample
2. Approximately normal distributions of sample means
    a. Central Limit Theorem
3. Reasonably large sample size
4. Equal sample sizes*
5. Equal population variances*

* Different test statistics depending on if these are true or not

Why do we care about *variance* when we are testing the equality of population *means*?

Mean *and* variance





test statistic unequal variance

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

Sample means

Sample variance

Sample size

Sample	Shannon	Depth (m)	Oxygen group
Saanich_010	3.945966	10	oxic
Saanich_100	4.273747	100	oxic
Saanich_120	3.937085	120	oxic
Saanich_135	3.203605	135	oxic
Saanich_150	2.349901	150	anoxic
Saanich_165	2.350053	165	anoxic
Saanich_200	2.465456	200	anoxic

t-test in R

```
t.test(Shannon ~ o2_group, data = m.meta.alpha, var.equal = TRUE)

Welch Two Sample t-test
data:  Shannon by o2_group

t = -5.3858, df = 5, p-value = 0.002976

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: -2.1444774  -0.7587843
sample estimates: mean in group anoxic  mean in group oxic
                         2.388470                3.840101
```

Analysis of variance (ANOVA)

- Compare 2+ *sample means* in order to make a statistical inference about the underlying *population means*

- t-test is a special case of ANOVA where you have only 2 samples

- Null hypothesis H_0: all of the *population* means are equal
- Alternate hypothesis H_1: all of the *population* means are not equal
  - At least 1 is different but ANOVA does not tell you which one

Sample	Shannon	Depth (m)	Oxygen group
Saanich_010	3.945966	10	oxic
Saanich_100	4.273747	100	oxic
Saanich_120	3.937085	120	oxic
Saanich_135	3.203605	135	oxic
Saanich_150	2.349901	150	anoxic
Saanich_165	2.350053	165	anoxic
Saanich_200	2.465456	200	anoxic

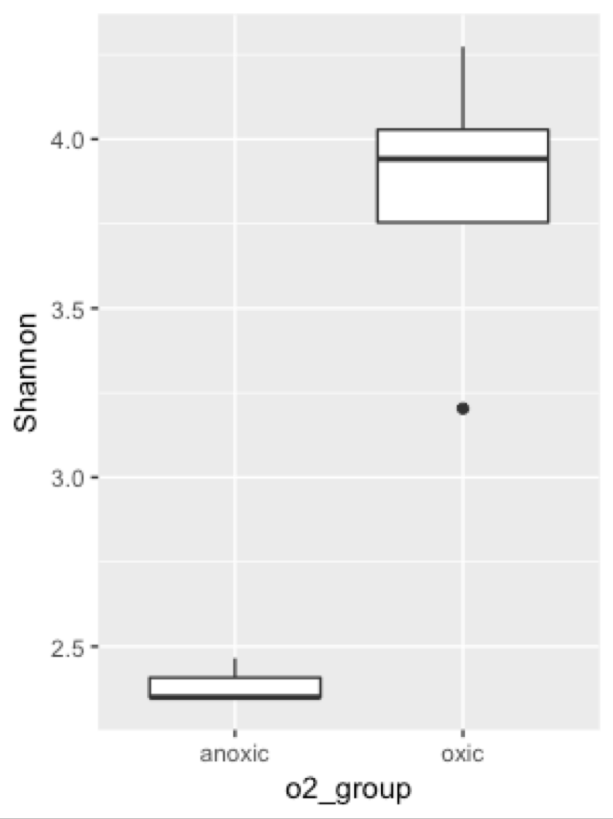
ANOVA in R

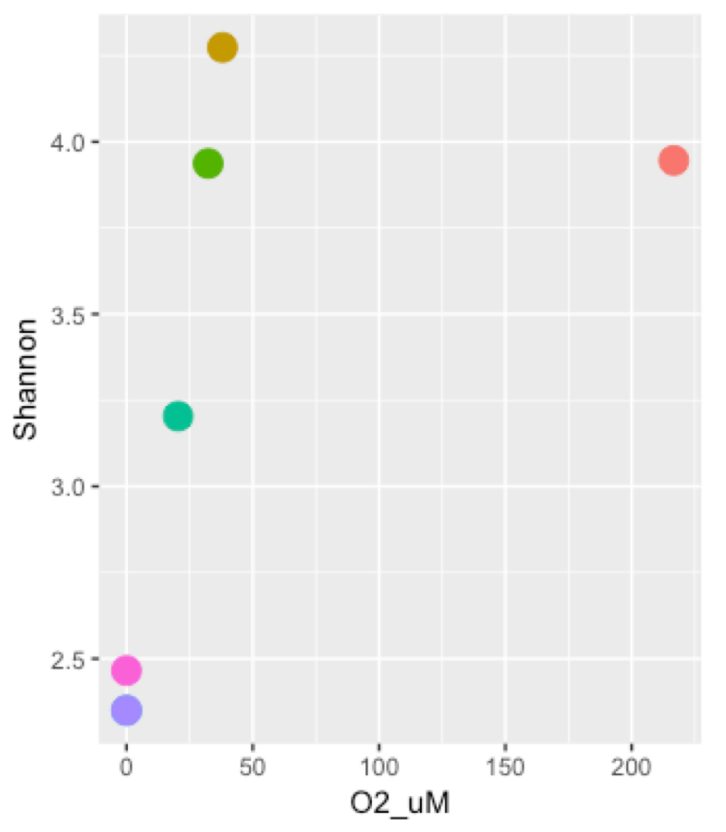
```
summary(aov(Shannon ~ o2_group, data = m.meta.alpha))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
o2_group	1	3.612	3.612	29.01	0.00298
Residuals	5	0.623	0.125		

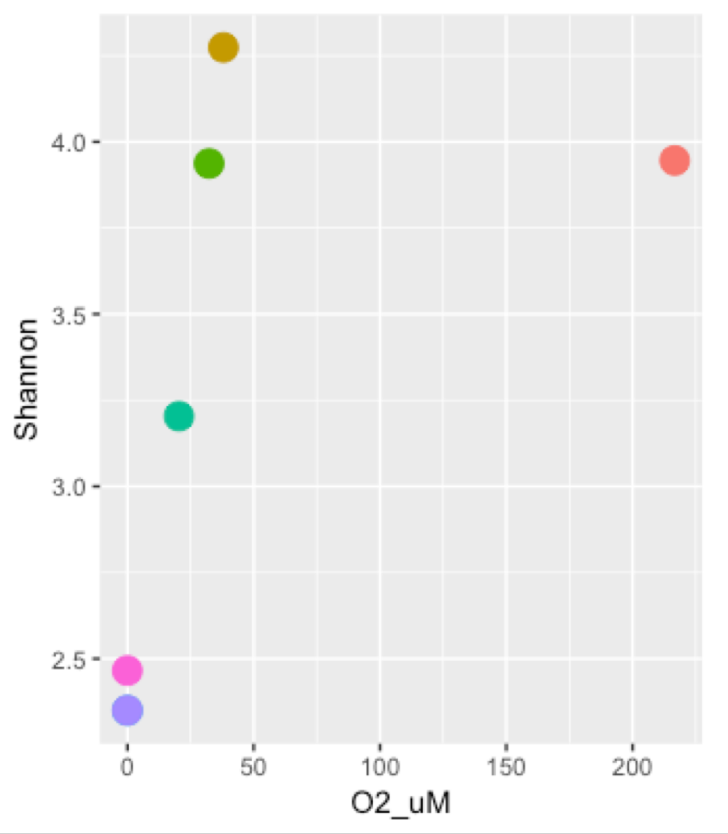
Welch Two Sample t-test t = -5.3858, df = 5, p-value = 0.002976

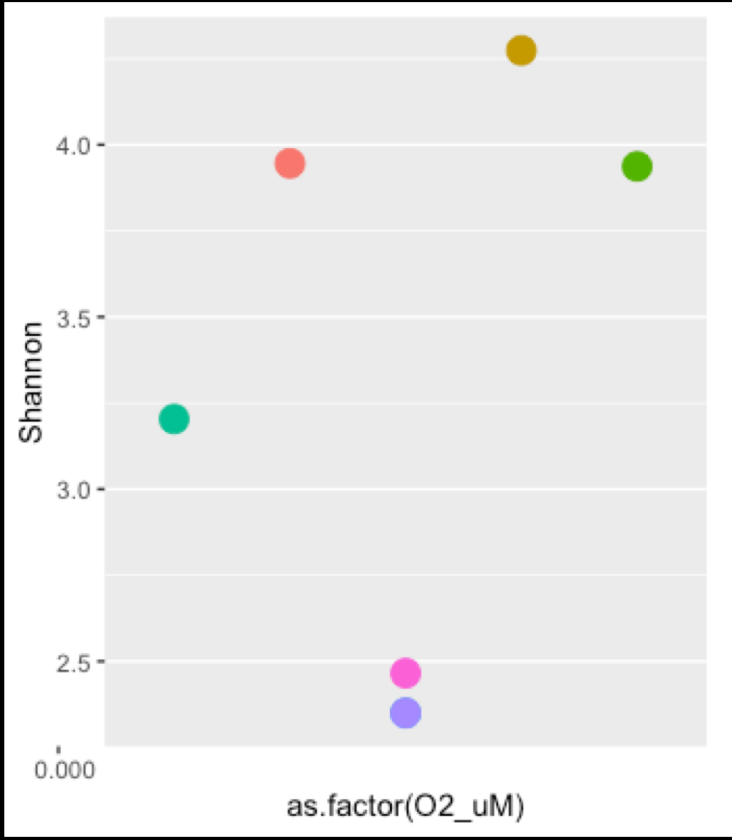
But oxygen is not binary





And oxygen is actually continuous





Oxygen is a continuous variable with order

Sample	Shannon	Depth (m)	Oxygen group	Oxygen (uM)
Saanich_010	3.945966	10	oxic	216.667
Saanich_100	4.273747	100	oxic	38.012
Saanich_120	3.937085	120	oxic	32.354
Saanich_135	3.203605	135	oxic	20.446
Saanich_150	2.349901	150	anoxic	0
Saanich_165	2.350053	165	anoxic	0
Saanich_200	2.465456	200	anoxic	0

ANOVA vs. LM

- Are they just groups? Or do you care about order?
- Are the groups discrete? Or taken from a continuous spectrum?
- ANOVA = just groups
- Linear model (LM) = groups or continuous

Linear model

- Compare 2 *samples* in order to make a statistical inference about the relationship (A) between the 2 *populations*
  - Continuous variable: $y = Ax + B$

- ANOVA is a special case of LM

- Null hypothesis H_0: slope (A) is equal to zero
- Alternate hypothesis H_1: slope (A) is not equal to zero

Linear model of discrete variable

Sample	Shannon	Depth (m)	Oxygen group
Saanich_010	3.945966	10	oxic
Saanich_100	4.273747	100	oxic
Saanich_120	3.937085	120	oxic
Saanich_135	3.203605	135	oxic
Saanich_150	2.349901	150	anoxic
Saanich_165	2.350053	165	anoxic
Saanich_200	2.465456	200	anoxic

Linear model of discrete variable

summary(lm(Shannon ~ o2_group, data = m.meta.alpha))

Residuals:

1	2	3	4	5	6	7
0.10587	0.43365	0.09698	-0.63650	-0.03857	-0.03842	0.07699

Coefficients:

	Estimate	Std. Error	t value.	Pr(>|t|)
(Intercept)	2.3885	0.2037	11.723	7.94e-05
o2_group.oxic	1.4516	0.2695	5.386.	0.00298

Residual standard error: 0.3529 on 5 degrees of freedom
Multiple R-squared: 0.853, Adjusted R-squared: 0.8236
F-statistic: 29.01 on 1 and 5 DF, p-value: 0.002976

t-test vs. ANOVA vs. LM

O2 GROUPS

- P-value for t-test = ANOVA = LM = 0.00298

Oxygen is a continuous variable with order

Sample	Shannon	Depth (m)	Oxygen group	Oxygen (uM)
Saanich_010	3.945966	10	oxic	216.667
Saanich_100	4.273747	100	oxic	38.012
Saanich_120	3.937085	120	oxic	32.354
Saanich_135	3.203605	135	oxic	20.446
Saanich_150	2.349901	150	anoxic	0
Saanich_165	2.350053	165	anoxic	0
Saanich_200	2.465456	200	anoxic	0

Linear model of O2 in uM

summary(lm(Shannon ~ O2_uM, data = m.meta.alpha))

Residuals:

1	2	3	4	5	6	7
-0.3215	1.0917	0.7894	0.1283	-0.6012	-0.6010	-0.4856

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.951099	0.336727	8.764	0.000321
o2_group	0.006076	0.003990	1.523	0.188331

Residual standard error: 0.7607 on 5 degrees of freedom
Multiple R-squared: 0.3168, Adjusted R-squared: 0.1802
F-statistic: 2.319 on 1 and 5 DF, p-value: 0.1883

Two different questions