

Biol 531 – Advanced topics in biological data analysis

Linear Models

- Course content
- Review of hypothesis testing, t-tests

Statistical survival skills for biologists

- This class will focus on analytical skills that are needed by professional biologists
- To function as a biological scientist, you need to be able to:
 - Design studies
 - Analyze data
 - Present results
 - Interpret the literature

Study design

- Focus on the logical construction of studies
- Why do we use:
 - Randomization?
 - Control groups?
- How do we get reliable experimental results in the face of:
 - Confounded variables?
 - Random variation in outcomes?
- How do fields that are unable to routinely use manipulative experiments make progress?
 - How can we use our understanding of experimental design to improve non-manipulative studies?

Analyze data

- Graphical, descriptive analysis – looking at your data
 - What are the major patterns?
 - How are the data distributed? Are there outliers?
- Evaluating hypotheses about the structure in your data
 - How do you match an analytical technique to the question you are asking?
 - How do you reach scientific conclusions based on statistical analytical results?

Interpreting the literature

- Become a critical consumer of the primary literature
 - What statistical techniques are being used, and why?
 - How to interpret presented results? Is the p-value enough?
 - Which are alternative approaches to the same analysis, and which are addressing different questions?
- You should know what authors assume you will know, and thus don't bother explaining to you
- Some examples...

Cell/molecular bio

The human splicing code reveals new insights into the genetic determinants of disease

Hui Y. Xiong,^{1,2,3*} Babak Alipanahi,^{1,2,3*} Leo J. Lee,^{1,2,3*} Hannes Bretschneider,^{1,3,4} Daniele Merico,^{5,6,7} Ryan K. C. Yuen,^{5,6,7} Yimin Hua,⁸ Serge Gueroussov,^{2,7} Hamed S. Najafabadi,^{1,2,3} Timothy R. Hughes,^{2,3,7} Quaid Morris,^{1,2,3,7} Yoseph Barash,^{1,2,9} Adrian R. Krainer,⁸ Nebojsa Jojic,¹⁰ Stephen W. Scherer,^{3,5,6,7} Benjamin J. Blencowe,^{2,5,7} Brendan J. Frey^{1,2,3,4,5,7,10†}

In white blood cells, 101 pairs of individuals that have differing SNPs (13). When we examined 99 exons that exhibited a significant difference in RNA-seq-assessed Ψ between pairs of individuals and whose predicted difference in Ψ was above a noise threshold, we found that our technique correctly predicted the direction of change in 73% of cases ($P = 3.5 \times 10^{-6}$, binomial test).

We found that intronic disease SNVs that are more than 30 nt from any splice site are 9.0 times as likely to disrupt splicing regulation relative to common SNPs in the same region ($P = 5.1 \times 10^{-68}$, two-sample t test, $n = 1639$ and $n = 24,535$). Within exons, synonymous disease SNVs are on average 9.3 times as likely as synonymous SNPs to disrupt splicing regulation ($P = 8.0 \times 10^{-116}$, two-sample t test, $n = 2652$ and $n = 4510$).

Missense SNVs have previously been examined mainly in the context of how they alter protein function (7). Our method enables the exploration of their effects on splicing regulation. We found that missense disease SNVs are not more likely to disrupt splicing than missense SNPs ($P = 0.22$, two-sample t test, $n = 58,918$ and $n = 2981$), which contradicts previously published evidence that they do ($P \approx 0.05$) (9). However, when we examined 789 and 1757 missense disease SNVs that minimally and maximally alter protein function as indicated by Condell (21) analysis, we found that SNVs that minimally alter protein function are on average 5.6 times as likely to disrupt splicing regulation ($P = 4.5 \times 10^{-14}$, two-sample t test), elucidating a “disease by misregulation” mechanism (13).

We found that within introns, the regulatory scores of 457 SNPs that were implicated in genome-wide association studies (GWAS) and that map to regulatory regions (22) are quite similar to non-GWAS SNPs ($P = 0.27$, KS test, $n = 262,804$), whereas the scores of disease SNVs are significantly higher ($P < 1 \times 10^{-320}$, KS test, 71.2%, $n = 280,638$). Fewer than 5% of GWAS SNPs are esti-

If they don't explain, you're expected to know already

Ecology

Eisenia density and macroalgal community structure

Over the course of this study (2004–2005), *Eisenia* densities at the study site were significantly higher in summer (up to 17 individuals/m²) compared to winter seasons (up to 11 individuals/m²) [two-way, Model III ANOVA, $F_{1,1} = 191.500$, $P = 0.046$; Fig. 2A)]. This pattern did not vary across years (ANOVA, $F_{1,186} = 1.684$, $P = 0.196$) or between seasons and years (ANOVA, $F_{1,186} = 0.015$, $P = 0.903$).

Foliose algal community structure differed with respect to canopy, season, and year (Fig. 2B, C); see Appendix B for complete ANOVA results for each macroalgal group. Foliose red algae was highest in percent cover in the canopy (12–16%) compared to the canopy-free (2–7%) zone for all of 2004 and summer 2005, but had a similar percent cover between the two

habitats in winter 2005 (5–7%) [three-way Model III ANOVA, canopy \times year; $P < 0.001$; Fig. 2B)]. In contrast, brown algal cover was up to six times higher in the canopy-free compared to the canopy zone across all seasons and years sampled (canopy; $P = 0.019$; Fig. 2C). Articulated coralline cover was overall higher in winter in 2005 (38–40%) than in summer (21–25%) [season \times year; $P = 0.03$]]. However, in 2004, articulated coralline cover was higher in the canopy zone each season (canopy \times year; $P = 0.028$; Fig. 2D). Crustose corallines were nearly 2.5 \times greater in abundance in the canopy-free compared to the canopy zone in winter of 2004, but showed similar abundances in both habitats in winter 2005. In the summer seasons, however, percent cover was higher under the *Eisenia* canopy (39–40%) compared to the canopy-free zone (25–26%) in both 2004 and 2005 (canopy \times season \times year; $P = 0.004$; Fig. 2E).

These results indicate that although both habitats have high abundances of articulated and crustose coralline algae, they differ in foliose algal species. The understory assemblage is dominated by foliose red algae with extremely low abundances of brown algae. In contrast, the assemblage in the canopy-free zone is

ppendix A;

Kelp canopy facilitates understory algal assemblage via competitive release during early stages of secondary succession

KYLLA M. BENES¹ AND ROBERT C. CARPENTER

Department of Biology, California State University, 18111 Nordhoff Street, Northridge, California 91330-8303 USA

Physiology

Effect of Nutritional Status on the Osmoregulation of Green Sturgeon (*Acipenser medirostris*)

Liran Y. Haller¹
Silas S. O. Hung¹
Seunghyung Lee¹
James G. Fadel¹
Jun-Ho Lee²
Maryann McEnroe³
Nann A. Fangue^{4,*}

¹Department of Animal Science, University of California, Davis, California 95616; ²Department of Marine Biomaterials and Aquaculture, Feeds and Foods Nutrition Research Center, Pukyong National University, Busan, Korea; ³School of Natural and Social Sciences, Purchase College, State University of New York, Purchase, New York 10577; ⁴Department of Wildlife, Fish, and Conservation Biology, University of California, Davis, California 95616

Accepted 11/1/2014; Electronically Published 12/18/2014

K⁺ ATPase), and m
The largest disturba
treatments across al
action between feed
the highest salinity i
during the first 72
interactions of thes
plications on green
form restoration and
estuarine environme

Introduction

Green sturgeon (*Ac*

Table 1: Growth performance, body proximate composition, and plasma metabolites in juvenile green sturgeon following a 4-wk feed restriction trial

Response variable	OFR group			
	12.5%	25%	50%	100%
Growth performance: ^a				
Final body weight (g)	180.0 ± 1.9 ^D	202.8 ± 2.1 ^C	248.1 ± 3.1 ^B	331.5 ± 2.7 ^A
SGR (% BW d ⁻¹)	-.40 ± .0 ^D	.02 ± .0 ^C	.78 ± .0 ^B	1.89 ± .0 ^A
Feed efficiency (%)	-206.2 ± 10.1 ^C	4.5 ± 7.4 ^B	96.9 ± 2.8 ^A	116.4 ± 2.2 ^A
Condition factor	.31 ± .0 ^C	.32 ± .0 ^{BC}	.37 ± .0 ^{AB}	.38 ± .0 ^A
Body proximate composition: ^b				
Moisture (%)	86.8 ± .3 ^A	85.3 ± .3 ^{AB}	84.3 ± .2 ^B	81.9 ± .6 ^C
Crude protein (%)	9.5 ± .1 ^B	10.9 ± .1 ^A	11.2 ± .1 ^A	12.1 ± .4 ^A
Crude lipids (%)	1.4 ± .1 ^B	1.3 ± .1 ^B	1.6 ± .2 ^B	3.6 ± .4 ^A
Body energy (kJ g ⁻¹) ^c	2.8 ± .1 ^C	3.1 ± .1 ^{BC}	3.4 ± .1 ^B	4.3 ± .2 ^A
Plasma metabolites: ^d				
Glucose (mg dL ⁻¹)	66.3 ± 1.8 ^B	56.7 ± .3 ^B	96.7 ± 1.8 ^{AB}	116.3 ± 9.2 ^A
Lactate (mg dL ⁻¹)	8.3 ± 2.7 ^A	18.1 ± .6 ^A	21.8 ± 6.1 ^A	33.3 ± 9.6 ^A
Triglycerides (mg dL ⁻¹)	11.3 ± 1.8 ^B	25.1 ± 5.0 ^B	171.0 ± 21.6 ^A	232.9 ± 28.8 ^A
Total protein (g L ⁻¹)	8.8 ± .1 ^C	10.3 ± .2 ^{BC}	12.3 ± .2 ^B	16.1 ± 1.2 ^A

Note. Values are means ± SE. Means with different superscript capital letters in each row are significantly different by the Tukey HSD test based on a one-way ANOVA ($P < 0.05$). BW = body weight; OFR = optimal feeding rate; SGR = specific growth rate.

^aN = 3 (three replicate tanks per feeding treatment). Average initial body weight was 202 ± 1.5 g.

^bN = 3 (three fish per replicate tank were pooled, and the three replicate tanks were averaged). Initial body proximate composition was 83.7% ± 0.1% moisture, 11.6% ± 0.2% crude protein, and 2.0% ± 0.3% crude lipids.

^cBody energy was calculated as 4.18 × [5.65 × (% crude protein × 100⁻¹) + 9.4 × (% crude lipids × 100⁻¹) + 4.23 × (% nitrogen-free extract × 100⁻¹)].

^dN = 3 (means of three fish in each replicate tank were averaged).

Some exotic, new developments...

How do animals optimize the size–number trade-off when aging? Insights from reproductive senescence patterns in marmots

VÉRANE BERGER, JEAN-FRANÇOIS LEMAITRE, JEAN-MICHEL GAILLARD, AND AURÉLIE COHAS¹

Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622, Villeurbanne, France

Trait	Model	k	AIC	Δ AIC	AICw
Offspring mass ($N = 549$)	Base	25	6058.02	4.19	0.06
	Base + Age	26	6055.47	1.65	0.22
	Base + Age ²	27	6057.26	3.43	0.09
	Base + T(10)	27	6058.18	4.35	0.06
	Base + F(age)	37	6053.83	0.00	0.49
	Base + S(age)	26	6057.47	3.65	0.08

Why statistical analysis?

- Did you really get into Biology to study statistics?
- It's everywhere, can't avoid it, assumed you know it...why, exactly?

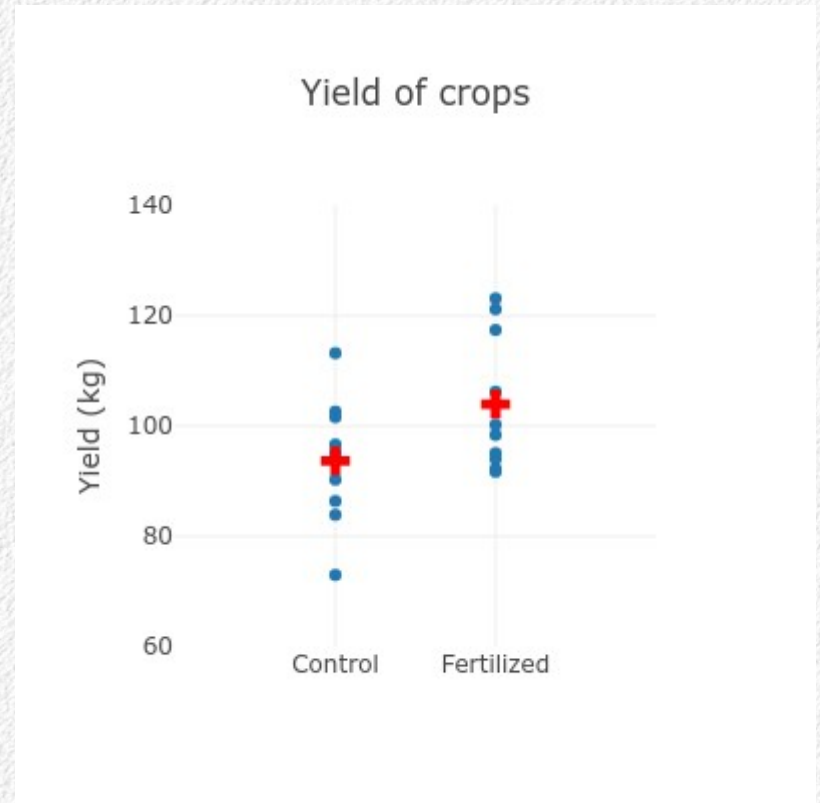
How do we know what we know?

- In science, we base what we know on evidence
- Experiments are the workhorse, the source of much of our evidence
- But, we work with biological material, which is variable
- This causes experimental uncertainty
- Example – conduct an experiment testing the effectiveness of a fertilizer for growing crops
 - Two groups (fertilizer treated and control)
 - Which gives more yield?
- Some hypothetical data (quick demo...)



What we learned from the demo...

- Even if there is no effect of treatment, two samples can have means that look quite different
- Identical sample means almost never happen, even if there is no difference at the population level
- If there is a difference at the population level, the samples of data will start looking different consistently (more convincingly when the difference is large)



Random differences can look real

- Simulated data sets were randomly selected from a single normal distribution with a mean of 100 kg
 - Randomly selected ten data points and called them “treatment”
 - Randomly selected ten other data points and called them “control”
- We only know this because I made up the data
- Normally, we don't know whether the experimental outcome is a real biological effect, or just random sampling
- So, what do we do to avoid being fooled by chance differences?

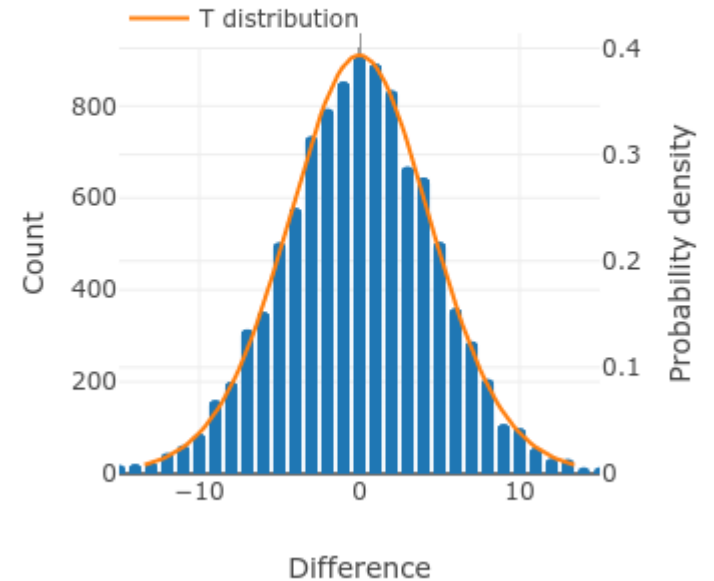
Drawing conclusions in the face of uncertainty

- We can't know for sure if experimental differences are real or just random sampling variation
- But, we can estimate the sizes of differences between groups to expect from random sampling
- We can then ask whether our observed differences are big compared to what we expect to get from random sampling
- How much variation to expect from random sampling? A simulation...

What we learned from the simulation

- Randomly sampling two groups from the same population many times results in:
 - Bell-shaped distribution of differences between means
 - Many small, near zero differences
 - The bigger the difference the less likely it is to occur by chance
 - The larger the sample size the less variability in differences occurs
 - Any given difference is likely to be different from zero, but on average the differences are zero
 - The Student's t-distribution is a good mathematical model for this **sampling distribution** of differences

Distribution of differences between means
mean of diffs. = -0.05, s of diffs. = 4.53



Standard error = measure of spread in the sampling distribution

- We measured standard error in the simulation by generating thousands of random differences and calculating their standard deviation
- This isn't possible in the real world – we do experiments once, not thousands of times
- Fortunately, there is a simple relationship between the standard deviation of the data and the standard error of the sampling distribution – $s_{\bar{x}} = s/\sqrt{n}$
- Both the standard deviation of the data (s) and the sample size (n) are known for a single sample, so we can estimate the standard error from a single sample of data

Standard error of a difference

- But, we have difference between two means, not a single mean
- The standard error of a difference between means is calculated with:

$$s_{diff} = \sqrt{\frac{s_{control}^2}{n_{control}} + \frac{s_{treatment}^2}{n_{treatment}}}$$

- This is a measure of how much random variation we expect, if there is no difference in means for the control and treatment populations
- We can compare the amount of difference between our treatment and control group sample means
 - If the amount of difference between the groups is big compared to this se, we have confidence that the treatment had an effect (the difference is probably not random)
 - If the amount of difference is not big compared to this se, it could easily just be random variation, and not a real treatment effect

For example...

- Control:

- $\bar{X} = 101.2$

- $s = 10.5$

- $n = 10$

$$s_{diff} = \sqrt{\frac{10.5^2}{10} + \frac{9.3^2}{10}} = 4.2$$

- Treatment:

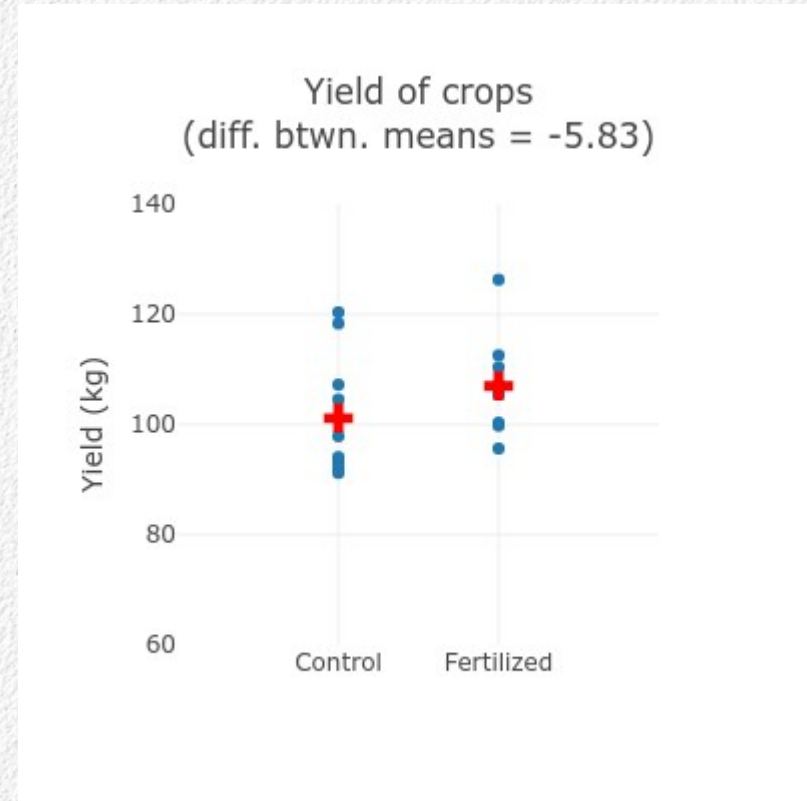
- $\bar{X} = 107.0$

- $s = 9.3$

- $n = 10$

Difference between means is 5.8, which is $5.8/4.2 = 1.38$ standard errors

Is this a lot?

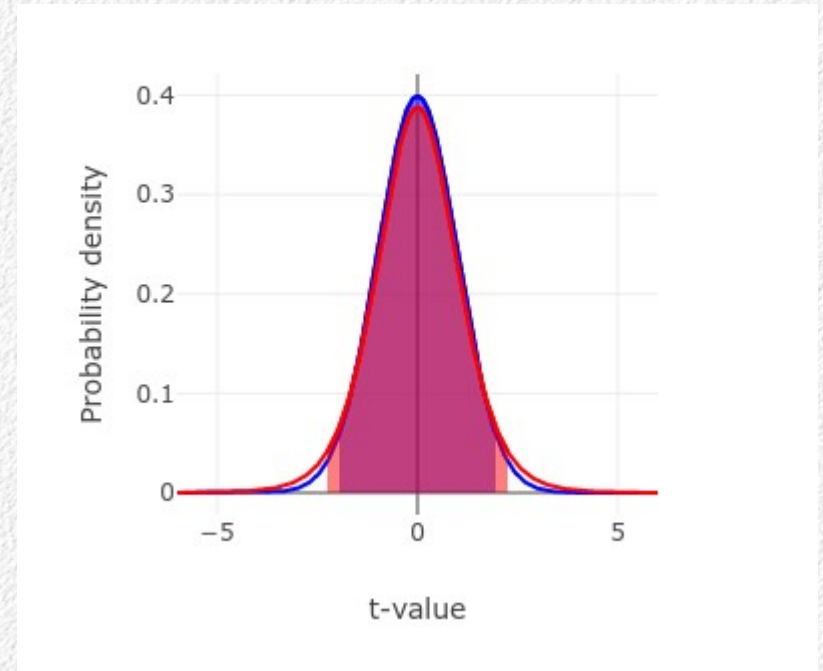


Using the t-distribution as a model of random sampling

- We can use the t-distribution to tell us if 1.38 standard errors between the means would happen commonly, or uncommonly, due to random sampling
- Back to the web pages...

What we learned from the t-distribution app

- With our $n = 10$, $df = 8$, 95% of random differences fall within 2.26 standard errors of zero
- If we take 2.26 as a benchmark of typical sizes for random differences, our difference of 1.38 s.e. is not very big – we can't be confident it's a real difference

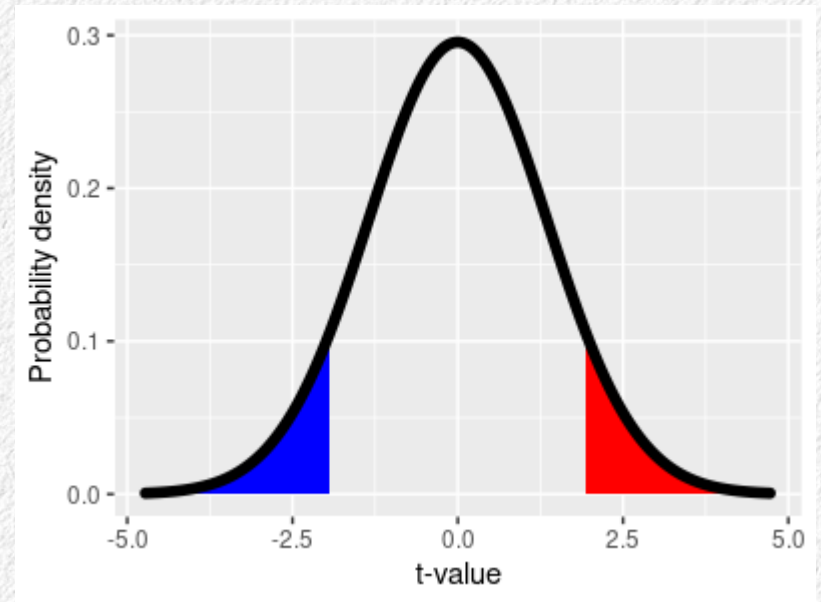


Hypothesis testing

- This general approach is often formalized into a null hypothesis significance test
- Start by hypothesizing no difference between population means (i.e. no effect of treatment, the **null hypothesis**)
 - Symbolically, $H_0: \mu_{\text{control}} = \mu_{\text{treatment}}$
- Calculate the amount of difference between the groups observed in the sample means, as a **test statistic**
 - $t_{\text{observed}} = \text{difference} / (\text{standard error of difference})$
- Use the t-distribution as a sampling distribution to calculate the probability of obtaining the difference observed by chance, if the null is true (the p-value)
- If the p-value is < 0.05 , reject the null – conclude that there probably is a difference between population means
- If the p-value is > 0.05 , retain the null – conclude that the population means are the same (or, rather, that your data do not give you enough evidence to conclude otherwise)

Our example as a t-test

- Our observed t-value of 1.38 above or below the mean difference of 0 encompasses 15% of the random difference expected
- Or, $p = 0.15$
- Since $p > 0.05$, we retain the null
- This is a two-tailed test – random differences as big as we observed, but in either direction, are included



Errors in hypothesis testing

- The benchmark we compare p against is our **alpha** level
 - Usually set to 0.05
 - Since $p < 0.05$ causes us to reject, this is the probability of a random outcome large enough to cause us to reject the null
 - It is thus an error rate – if we get a random difference that we conclude is non-random, this is a Type I error = a false positive
- We can also get differences that are real, but are so small they could easily have been random
 - When this happens, $p > 0.05$, even though the difference is real
 - The probability of this is **beta**, and it is the probability of a **Type II error** = a false negative
 - The probability of detecting a difference when there is one there to detect is $1-\beta$, and is called **statistical power**

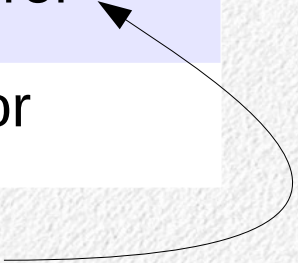
Possible outcomes of our test – types of errors

	Populations are different	Populations are not different
Reject the null	No error	Type I error
Retain the null	Type II error	No error

Determined by size of difference, sample size, sampling variation (β errors)



Set by you (α -level)



Effects of sample size on error

- Error can never be eliminated, but we do have some control over the chances of an error
 - Type I error (α) is set by you when you decide the p-value that will indicate a positive result (i.e. a rejected null)
 - Type II error (β) isn't set, but can be minimized with large sample size and good experimental design
- For a particular sample size, they trade off
 - Lower chance of Type I error \rightarrow higher chance of Type II
 - Lower chance of Type II \rightarrow higher chance of Type I
- Only way to reduce Type II without increasing Type I is to increase sample size (or collect data more carefully)
- Another online illustration...

Types of t-tests – matching analysis to design

Design	Example	t-test type
Two groups compared	Treatment vs. control	2-sample t-test
One group compared to a hypothetical mean	Mean body temperature of the class vs. 98.6	1-sample t-test
Measurements of paired samples	Right bicep circumference vs. left bicep circumference	Paired t-test

One-sample designs

- A single set of measured data is collected, and a sample mean is calculated
- This mean is compared against a number, representing the hypothetical mean for the population
 - Simple example would be something like normal body temp, 98.6° (the example you will use for the review exercise)
 - Null hypothesis is $H_0: \mu = 98.6$, if you reject the null conclude population body temp is not 98.6
- Not very common, because we don't usually know what the population parameter should be, and need to estimate it from data

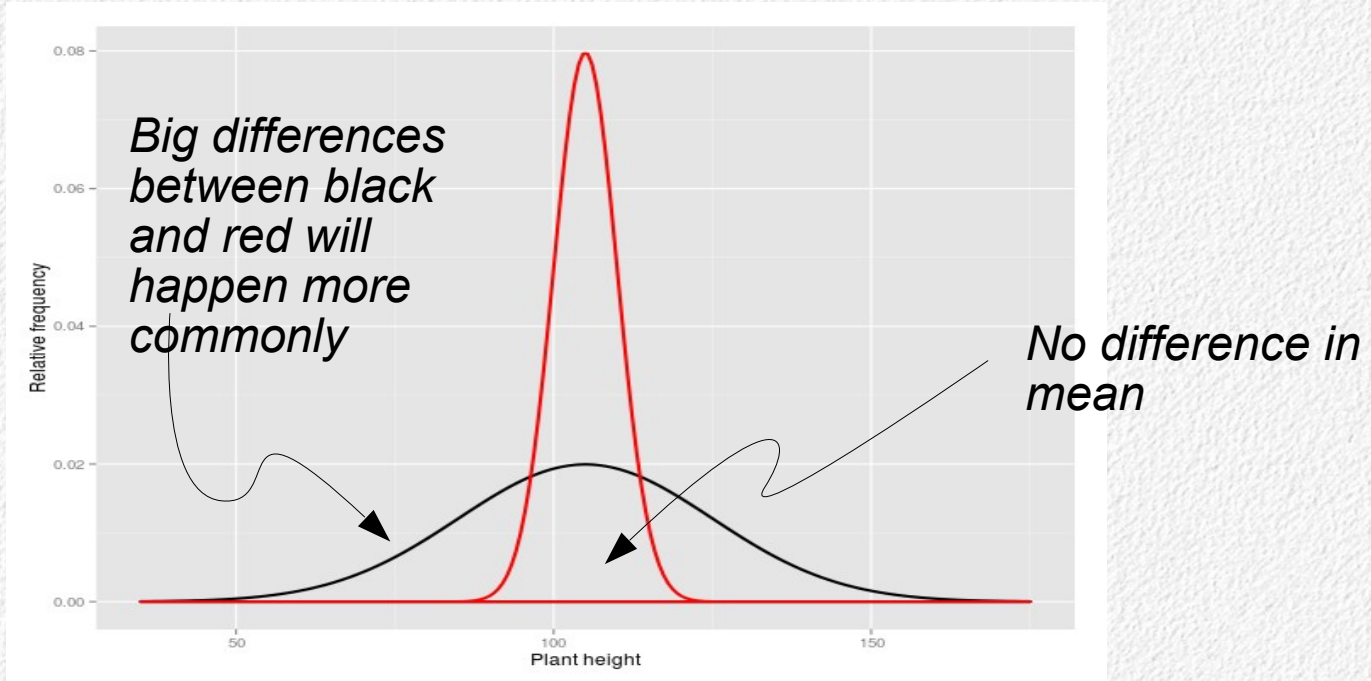
Paired designs

- Two-sample design, but the data points are not independent between the groups
 - Usually because they are two different measures of the same subjects (before/after treatment, right/left side of body, etc.)
- Problem is that variation between the experimental subjects can obscure small, consistent differences
- Solution is to use the difference between the two sets of measurements
 - Example: measure uptake of CO₂ in a set of plants grown in chambers with low atmospheric CO₂, and then in chambers with high atmospheric CO₂ (the example you will use for the review assignment)
 - Compare mean of differences against 0 (Ho: $\mu_{\text{diff}} = 0$) with a one-sample t-test
 - If the null is rejected conclude that the differences are not 0 at the population level

Assumptions of tests

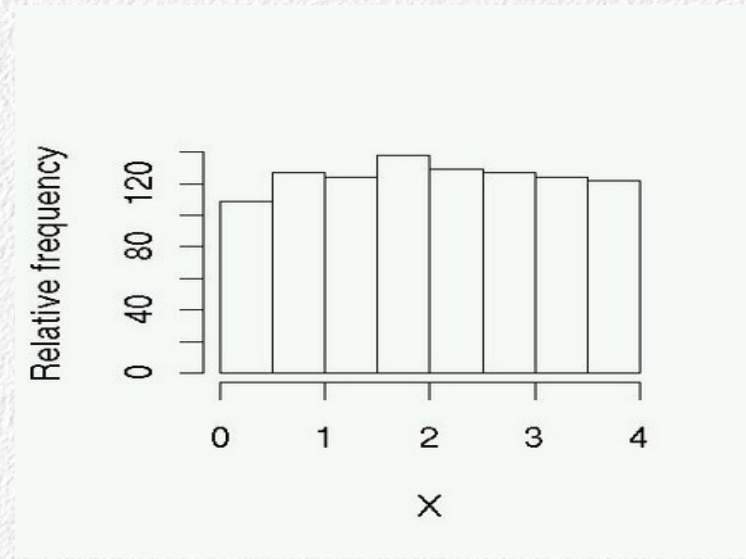
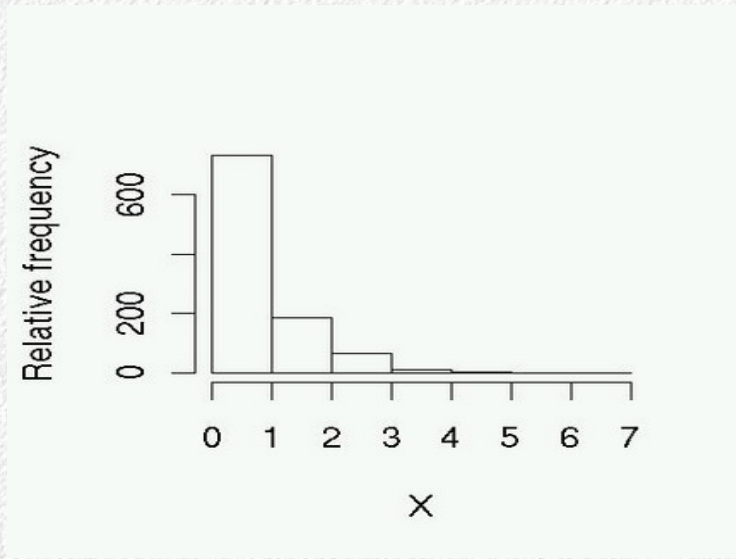
- Statistical assumptions are the conditions that need to be true for the test to work as expected
- General statistical assumptions
 - Observations are randomly sampled
 - Observations are independent
- “Parametric” assumptions (needed for the t-distribution to be an accurate model of random sampling)
 - Variances between the groups are equal (two-sample t-tests only...why?)
 - A.k.a. homogeneity of variances, homoscedasticity
 - The populations are normally distributed (differences are normal in a paired t)

Violating the HOV assumption



Note: there are ways to treat lack of homogeneity of variances... more later

Violating the normal population assumption



We can test for this, and if the data are non-normal we may treat the data (transformations), or use a different type of test (nonparametrics) – more later

Take-home: the purpose of hypothesis testing

- Random sampling can produce results that look like real experimental effects
- You can never be completely certain that your results are not just random sampling variation
- But, you can calculate the probability that your results are random sampling variation, and draw conclusions in light of this uncertainty