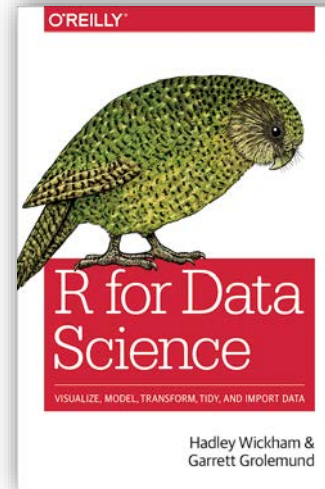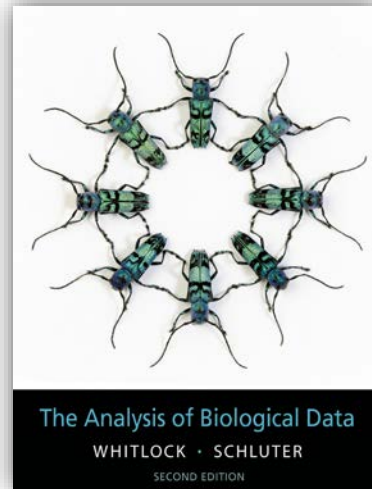# Data Science in Bioinformatics

Palle Villesen & Thomas Bataillon

Outline for week 05

- Any questions from last week ?

- Chapter 6
  - Open discussion & Exercises

- Thursday
  - Who will be responsible

Hypothesis testing

4 basic steps

- State null hypothesis
- Compute test statistic
- Determine p value
- Conclude

Why statistical tests?

- One use of hypothesis testing is deciding whether experimental results contain enough information to **cast doubt on conventional wisdom**
  - That means reject the null hypothesis

Null hypothesis

A good null hypothesis is a statement that would be interesting to reject

Examples

- H0: Tomato plants do not exhibit a higher rate of growth when planted in compost rather than soil.

- H0: World is Flat

- H0: The Geocentric Model: Earth is the centre of the Universe and it is Spherical

- H0: The Heliocentric Model: Sun is the centre of the universe
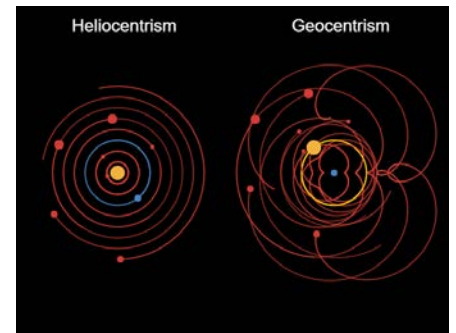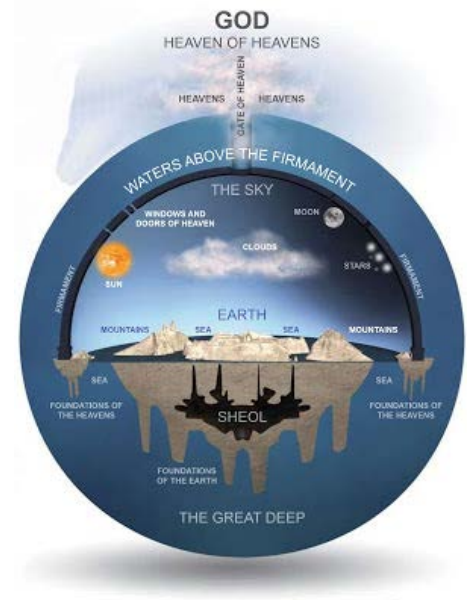
Discuss

- How do you formulate a null-hypothesis, if you have absolutely no idea about the expected parameter value, and how do you judge if a null-hypothesis is "good" or not, if there's no data - or theory based assumptions of the true parameter value.

# Polio vaccine – what is the H0?

- 401974 students in two groups
- 0.016% got polio with vaccine
  - 33/200987
- 0.057% got polio without vaccine
  - 115/200987

Null distribution

The null distribution is the sampling distribution of outcomes for a test statistic under the assumption that the null hypothesis is true

H0 and its null distribution

- Distribution arising under the null hypothesis

- A null hypothesis (H0):

a (precise) statement about a property of a population

- Test of an hypothesis: How unlikely is my observation under H0 ?

  If unlikely enough -> reject H0

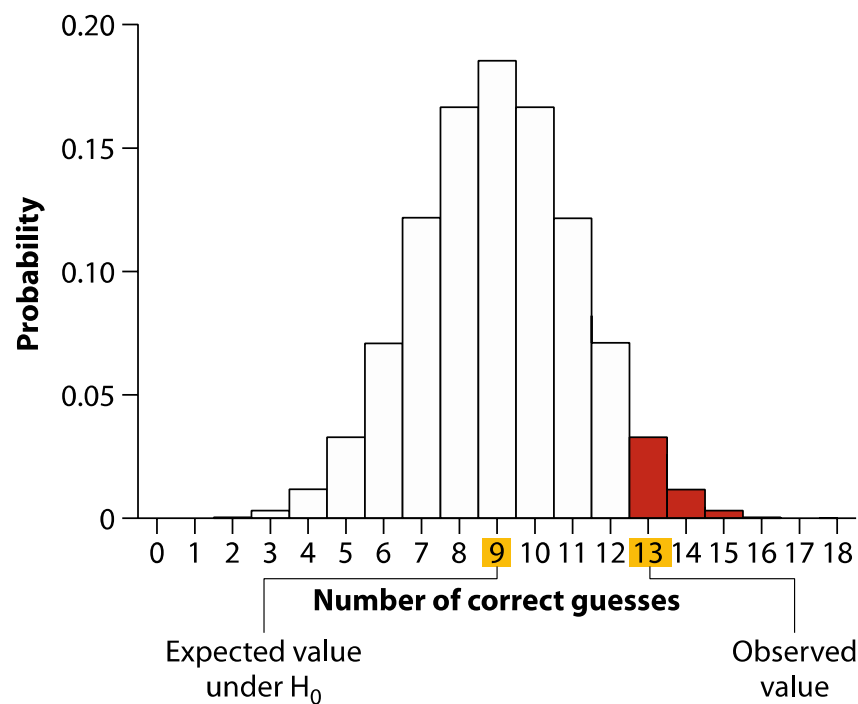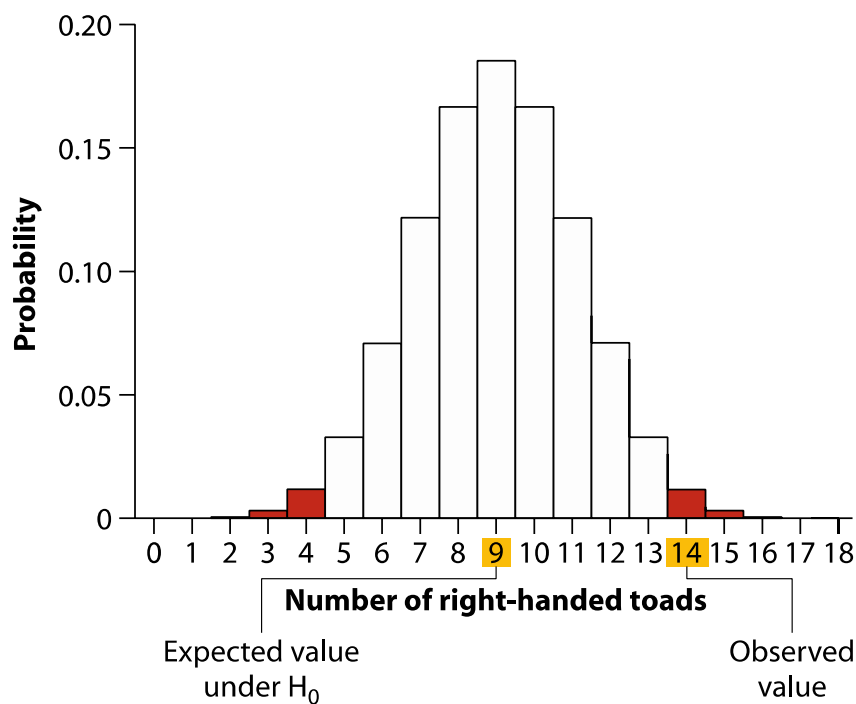 If not -> do not reject H0... does it mean accepting it ?
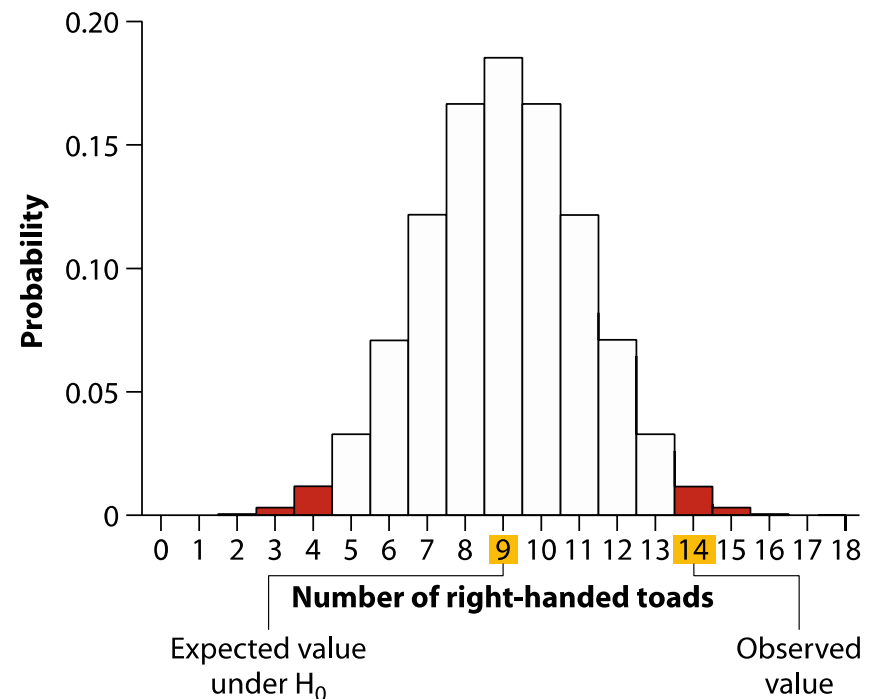
P values

Compare the observed test stastitic
to the null distribution

What is a P-value?
The probability of obtaining outcomes as extremes or more extreme than the observed data at hand under the hypothesis H0
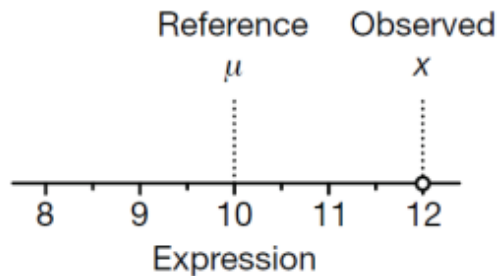**How do we define "extreme"?**
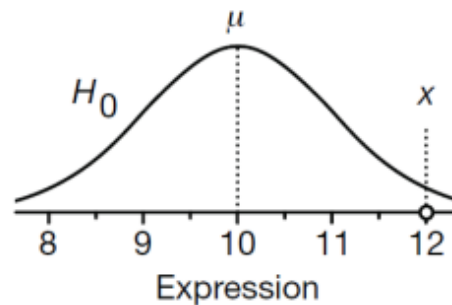

"Two sided" or "one sided"?
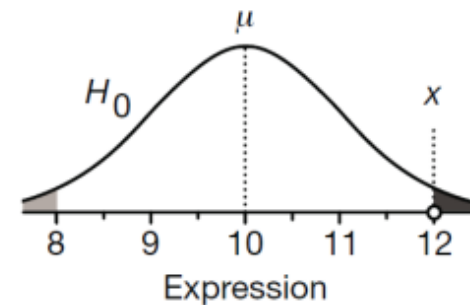
How to obtain p

a Experimental observation

Reference μ Observed x

8   9   10   11   12
Expression

b Distribution of reference expression values

μ

H_0 x

8   9   10   11   12
Expression

c Probability of observing a more extreme value

μ

H_0 x

8   9   10   11   12
Expression

P value

- A *p* value is not a measure of how right you are, or how significant the difference is; it's a measure of ***how surprised you should be*** if there is no actual difference between the groups, but you got data suggesting there is. A bigger difference, or one backed up by more data, suggests more surprise and a smaller *p* value.

But there are limitations

- The *p* value is a measure of surprise, not a measure of the size of the effect. I can get a tiny *p* value by either measuring a huge effect – "this medicine makes people live four times longer" – or by measuring a tiny effect with great certainty.

- **Statistical significance does not mean your result has any *practical* significance.**

http://www.statisticsdonewrong.com/data-analysis.html#the-power-of-p-values

Significance

- In statistics, a result is called statistically significant if it is **unlikely to have occurred by chance alone**, according to a pre-determined threshold probability, the significance level.

The big issue

- There's no mathematical tool to tell you if your hypothesis is true; you can only see whether it is consistent with the data, and if the data is sparse or unclear, your conclusions are uncertain.

- But we can't let that stop us.

- www.statisticsdonewrong.com

Type I and Type II errors

	H0 is TRUE	H0 is False
DECISION (based on $\mathbb{D}$)		
Reject H0	Type I error (α)	Correct
Do not reject H0	Correct	Type II error (β)

Choosing a significance level (α) for a P-value sets the type I error rate at α.
Varying α will have consequences for the rate of type II error and statistical power
Power = $1 - \beta$
IF your test is well behaved/ calibrated and p-value well calculated THEN you will reject
Wrongly H0 a proportion α of the time when making a test

What about H0?

- When do we know that H0 is true?

- H0: HPV vaccine is not harmfull
- Ha: HPV vaccine causes serious side effects

Statistics in three different ways

- Using known test with known distribution
  - Compare test statistic to known distribution of H0
  - E.g. t-test
- **Simulating** test statistic "under H0"
  - Compare test statistic to simulated distribution of H0
- **Permute** data → test statistic "under H0"
  - Compare test statistic to permuted distribution of H0

Do the poor guys on the left have higher blood pressure?
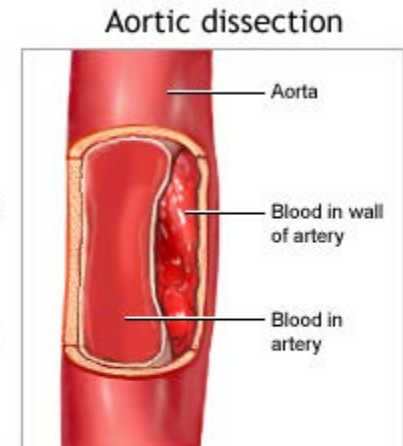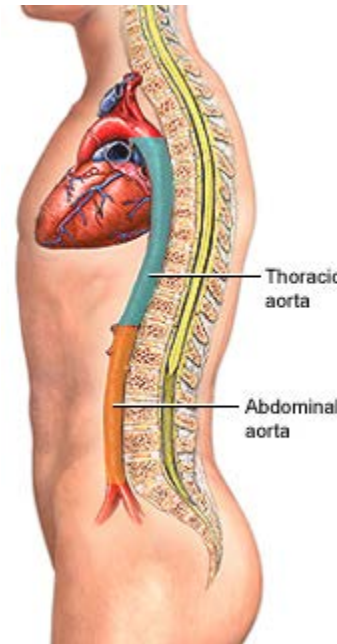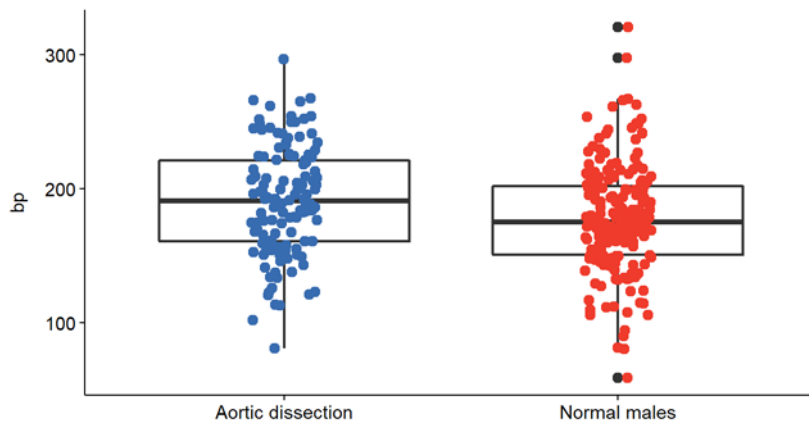
My own test statistic

ts = abs(mean(g1)-mean(g2))

Measures difference between two groups
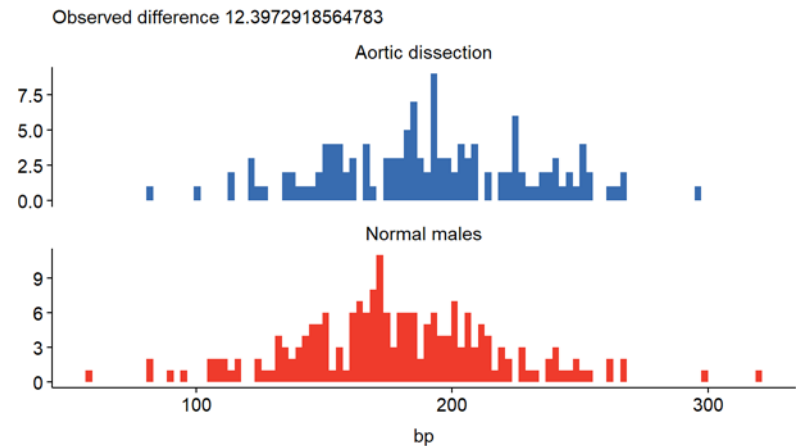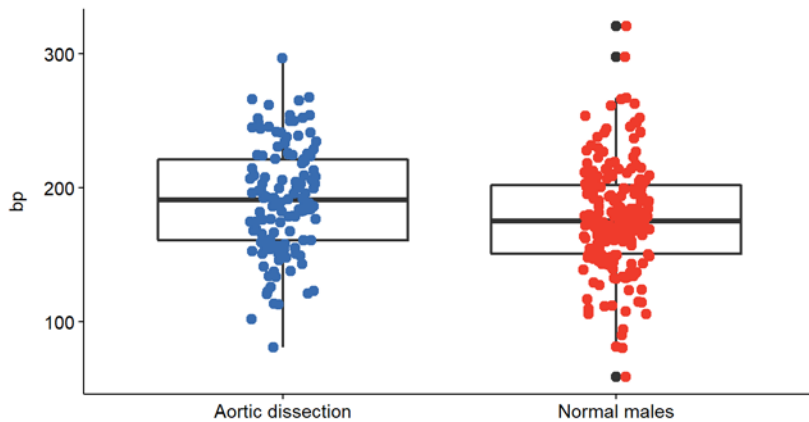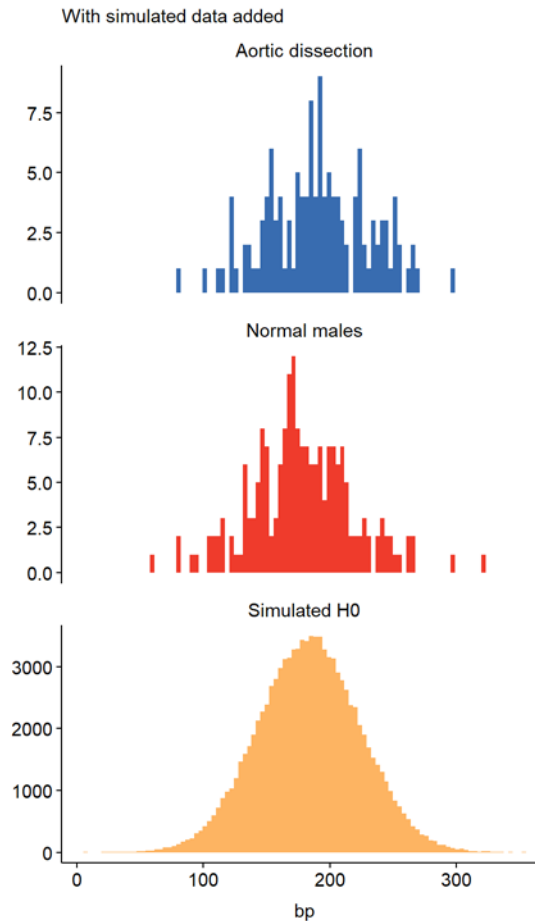
Is the observed difference significant?

Normally we use a "distribution of ts under H0" – but what does that even mean?

Statistical test: did sick and normal come from the same population?

H0 = Sick comes from the blue "population"

H0 = Normal comes from the blue population

H0 = Sick and Normal are the same
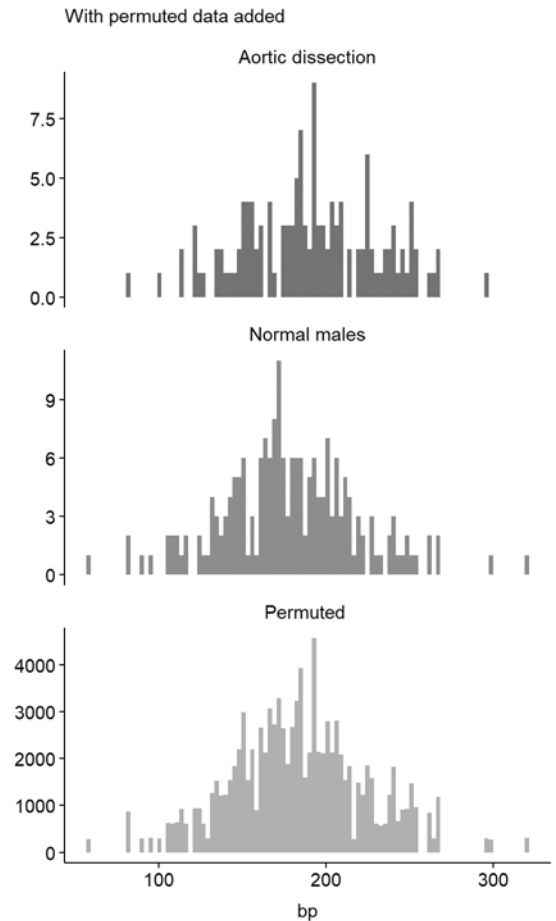
Simulating the null distribution

- H0: Only one **population = one mean**
- Draw two **samples** from that H0 **population** (H0: same mean)
- Calculate the test-statistic between the two **samples**
- Repeat 100000 times

Simulating data to get H0 distribution

```
d = read.table("Aortic_dissection.tsv")
g1 = d$bp[d$group != "Normal males"]
g2 = d$bp[d$group == "Normal males"]

mean_h0 = mean(d$bp)
sd_h0   = sd(d$bp)

ts = rep(0,100000)
for (i in 1:length(ts)) {
  ng1   = rnorm(length(g1), mean=mean_h0, sd=sd_h0)
  ng2   = rnorm(length(g1), mean=mean_h0, sd=sd_h0)
  ts[i] = abs(mean(ng1)-mean(ng2))
}
```
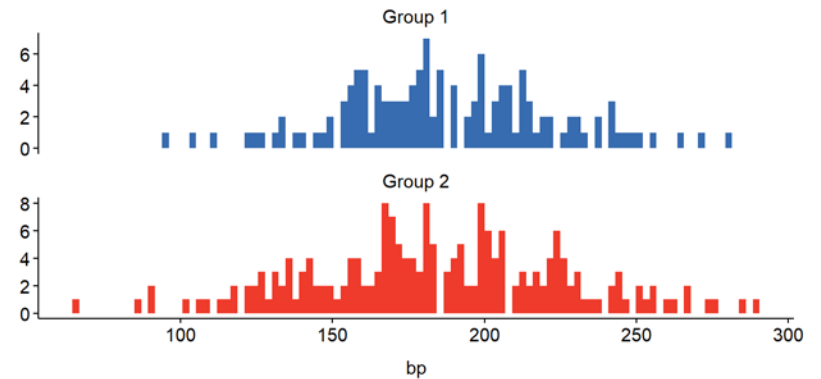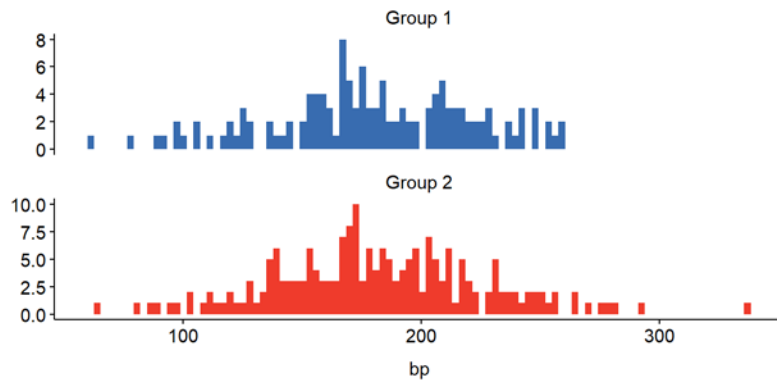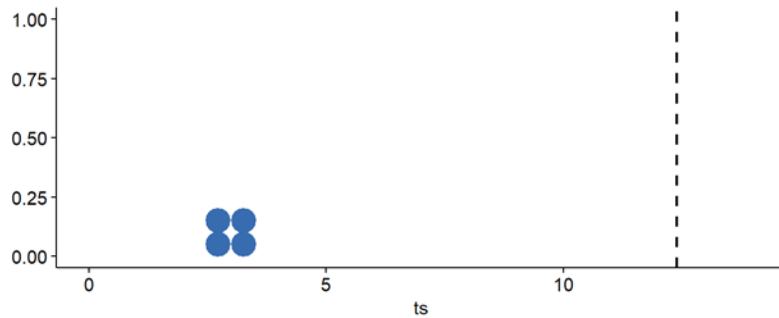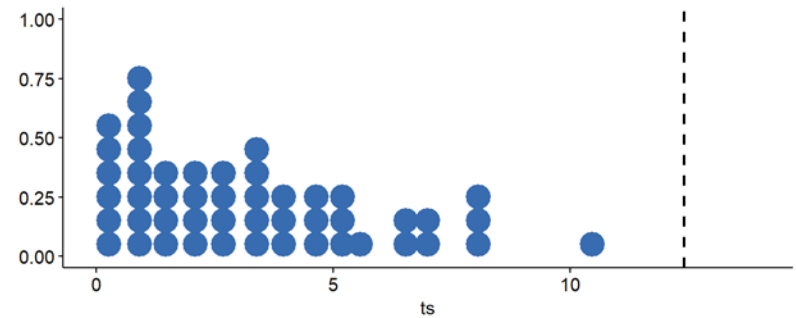
Simulations
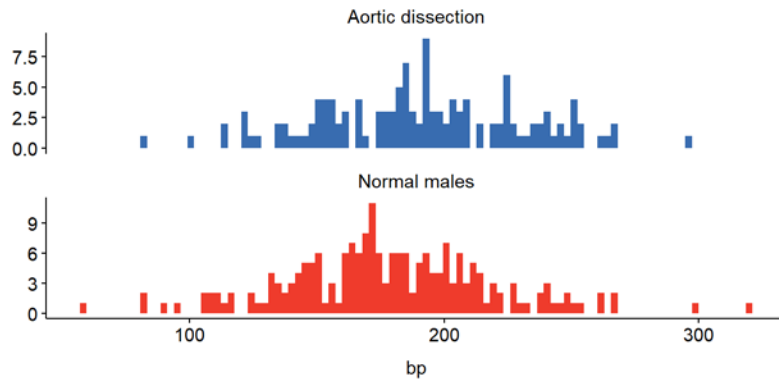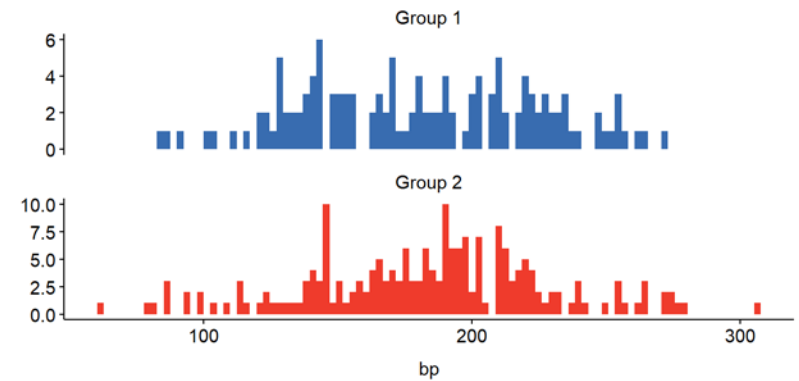
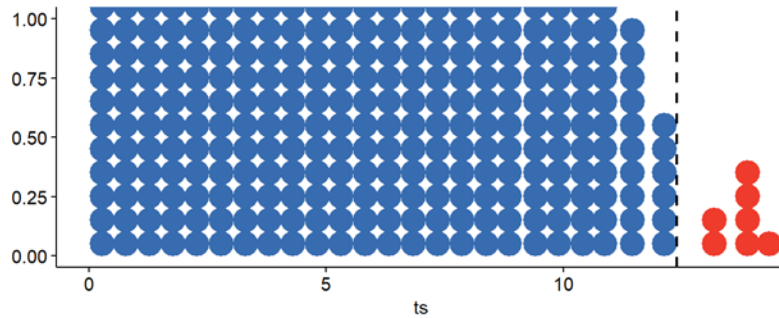$$p = \frac{n_{\text{more extreme}} + 1}{n_{\text{simulations}} + 1}$$

The first 4 test values
Number of observations > 12.3972918564783 = 0
Estimated p < 0.2
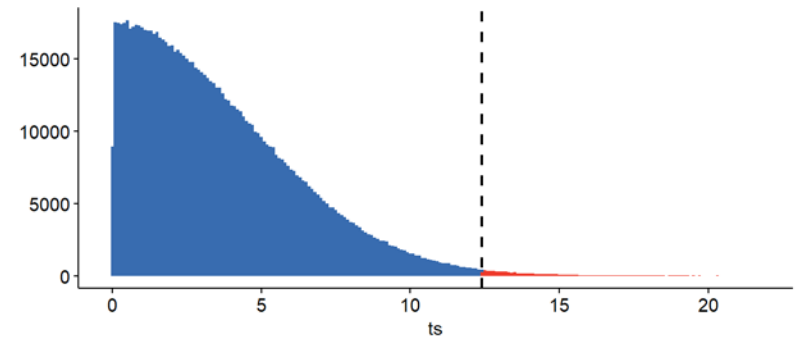
The first 49 test values
Number of observations > 12.3972918564783 = 0
Estimated p < 0.02

The first 1999 test values
Number of observations > 12.3972918564783 = 9
Estimated p < 0.005

Number of extreme values: 6348 of 1000000
Simulated H0 p < 0.00634899365100635

If you can't get the probability distribution (like the chi-square) and can't simulate it – because your data doesn't "behave nicely"

PERMUTATION TESTS (RANDOMIZATION)

Basic idea of permutations

- We use the data to generate **permuted datasets – that are expected to follow the H0**
- For patients/control – we **randomly switch patients and controls**
- Before:
  - Calculate test statistic of observed data
- Loop:
  - Permute data
  - Calculate test statistic of permuted data
- P-value:
  - How many permutations were more extreme than the observed

Statistical test: did sick and normal come from the same population?

H0 = Sick comes from the blue "population"

H0 = Normal comes from the blue population

H0 = Sick and Normal are the same
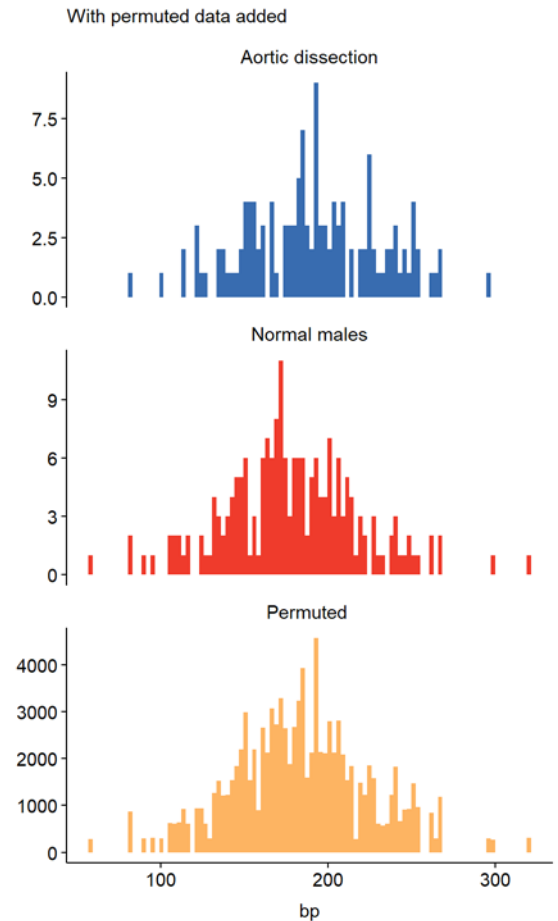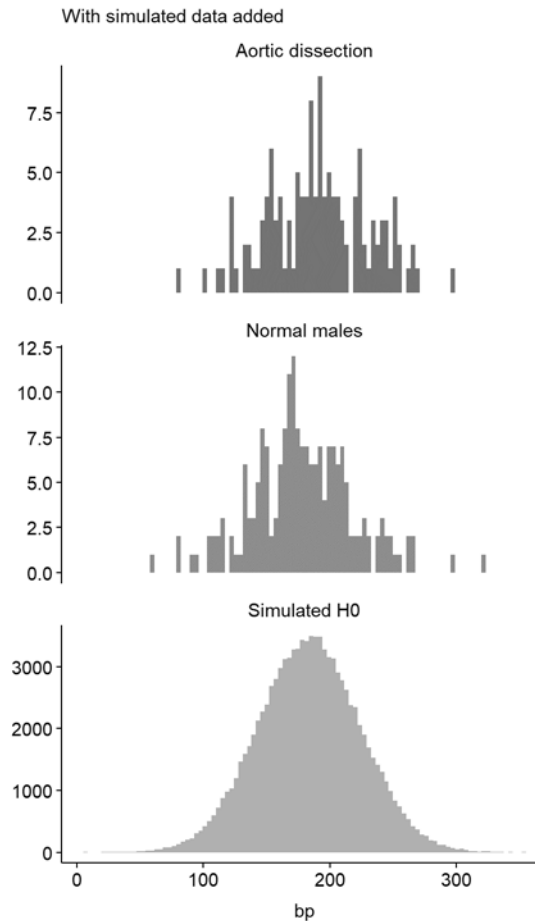
Permuting data to get H0

```
set.seed(1)
permutations  = 100000
ts = rep(NA, permutations)

for (i in 1:length(ts)) {
  pmg = sample(d$group)                    # Mix samples between groups
  ng1 = d$bp[pmg != "Normal males"]        # The values from the sick group
  ng2 = d$bp[pmg == "Normal males"]        # The values from the normal group
  ts[i]    = abs(mean(ng1) - mean(ng2))    # Test statistic (H0 is true)
  if(i %% 1000==0) {
    cat(i,"\n")
    flush.console()
  }
}

extreme_values      = sum(ts >= observed_difference)
pval_estimated      = extreme_values/length(ts)
```
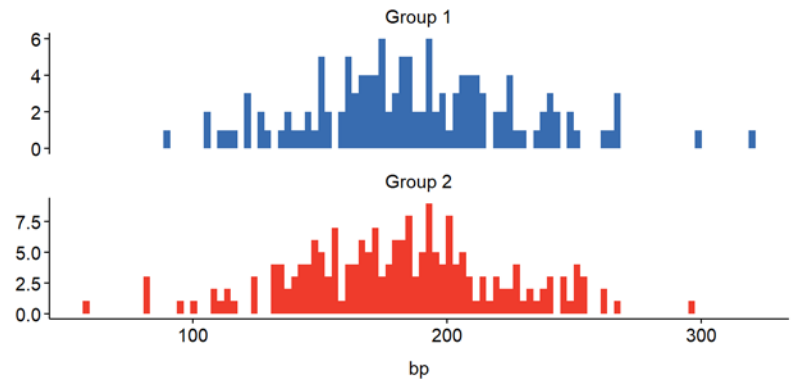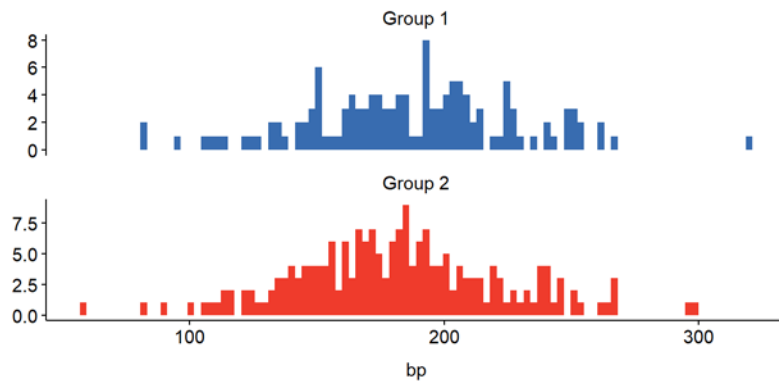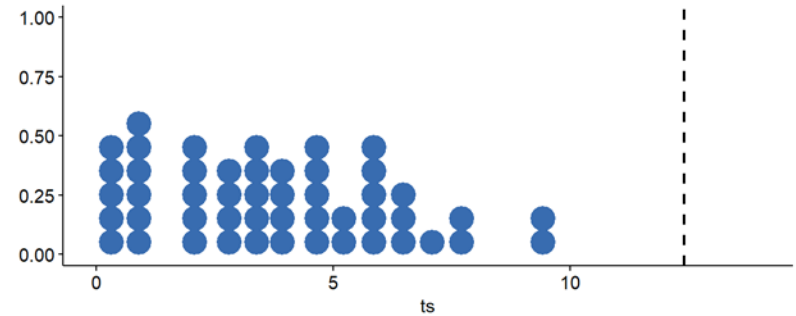
Permutations
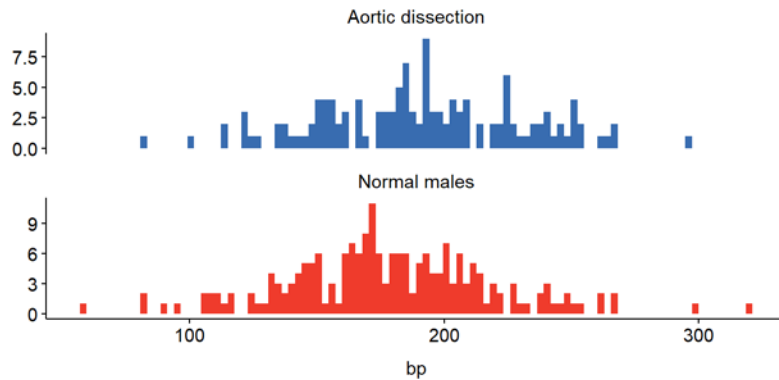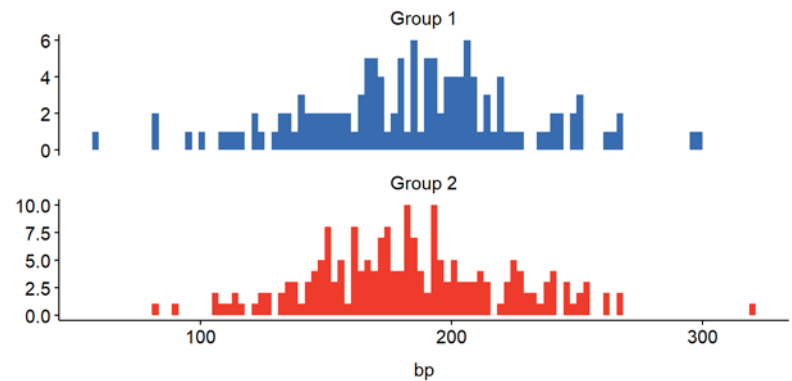
$$p = \frac{n_{more\ extreme} + 1}{n_{simulations} + 1}$$

The first 4 test values
Number of observations > 12.3972918564783 = 0
Estimated p < 0.2
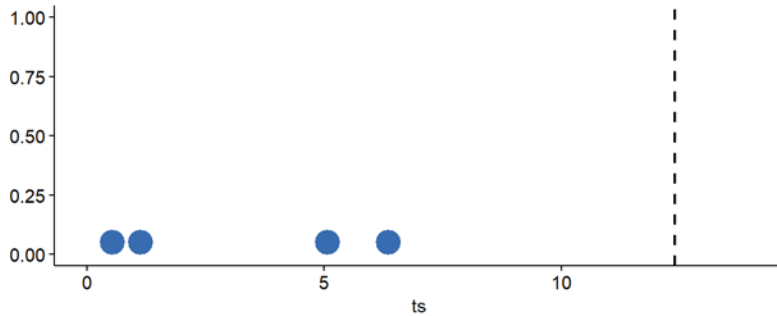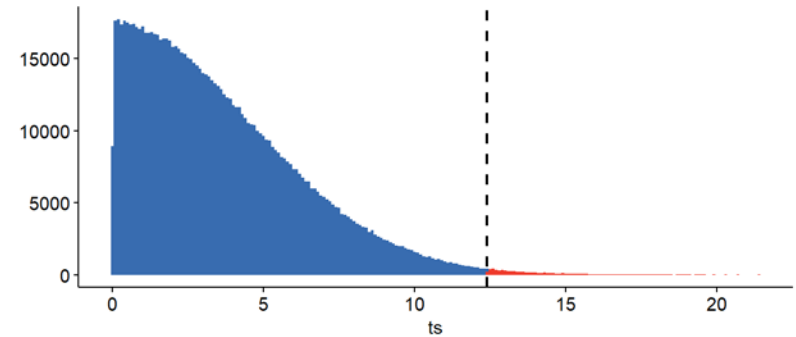
The first 49 test values
Number of observations > 12.3972918564783 = 0
Estimated p < 0.02

The first 1999 test values
Number of observations > 12.3972918564783 = 12
Estimated p < 0.0065
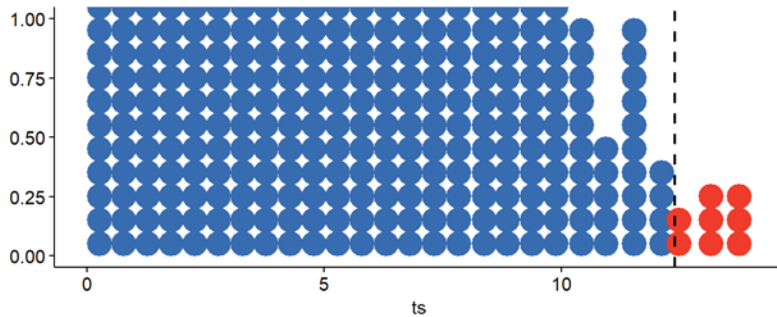
Number of extreme values: 6182 of 1000000
Permuted H0 p < 0.00618299381700618

t.test(g1,g2, var.equal=T)

```
    Welch Two Sample t-test

data:   g1 and g2
t = 2.7555, df = 281.6, p-value = 0.006242
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
  3.541172 21.253411
sample estimates:
mean of x mean of y
 190.8181   178.4208
```

Statistics in three different ways

- T.test
  - P = 0.006242

- Simulating test statistic "under H0"
  - P = 0.00635

- Permute data → test statistic "under H0"
  - P = 0.00619

What was wrong?

- Is my test statistic good or bad?
- ts = abs(mean(g1) – mean(g2))

- Number of observations?
  - "Sample size"
- Variation of sample?
  - "Variance" or "standard deviation"

T test (wikipedia)

Independent two-sample *t*-test [edit]

Equal sample sizes, equal variance [edit]

This test is only used when both:

- The two sample sizes (that is, the number, *n*, of participants of each group) are equal;
- It can be assumed that the two distributions have the same variance.

Violations of these assumptions are discussed below.

The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 X_2} \cdot \sqrt{\frac{1}{n}}}$$

where

$$s_{X_1 X_2} = \sqrt{s_{X_1}^2 + s_{X_2}^2}$$

Here $s_{X_1 X_2}$ is the grand standard deviation (or pooled standard deviation), 1 = group one, 2 = group two. $s_{X_1}^2$ and $s_{X_2}^2$ are the unbiased estimators of the variances of the two samples. The denominator of *t* is the standard error of the difference between two means.
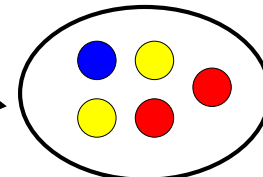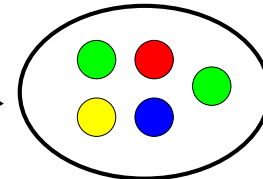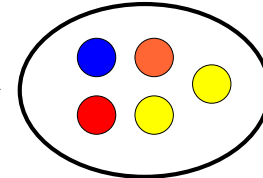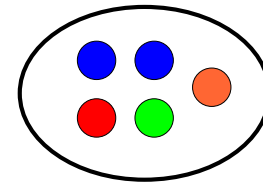
For significance testing, the degrees of freedom for this test is $2n - 2$ where *n* is the number of participants in each group.

Population   VS   sample

UNKNOWN

KNOWN
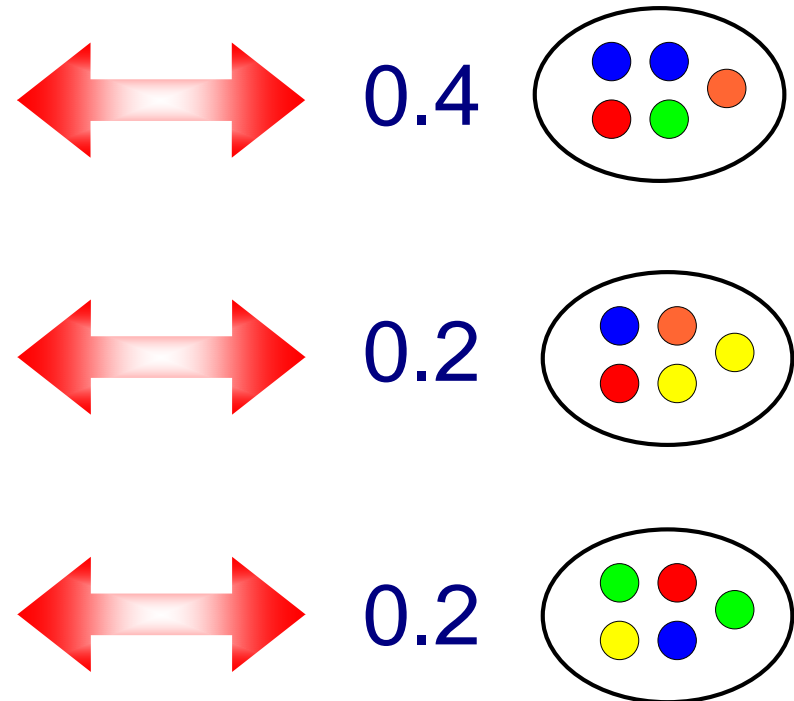
- Virtually infinite

- Parameters

- Probability distribution

- n obs: $x_1, x_2, \ldots x_n$

- Parameter estimates

- Sampling distribution

Estimating with uncertainty

- Estimates of parameters are based on a finite sample from the population.

- **Sampling error** is chance-induced variation between the estimate and the true parameter.

0.4

0.2

0.2

Hypothesis testing

- From sample to general knowledge about population
- **Sampling error** is chance-induced variation between the estimate and the true parameter
- Is the sample more blue than expected?
- H0: it is a random sample, frequency(blue) = 0.23
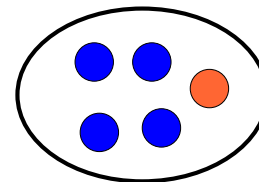
Think binomial

- What is the chance of "blue" (under H0)
- What is the chance of "not blue" (under H0)

- What is the chance of BBBBN ? BBNBB? NBBBB
- What is more extreme than the observed?
- What is the chance of 5 blues (under H0)?

- What is the total chance of same or more extreme than the observed?

Hypothesis testing

- H0: our sample is from the pool, frequency(blue) = 0.23
- HA: ?
- Formulate and discuss the test of this sample
- What is the observed? What is more extreme?
- What is the frequency of more extreme = pvalue?
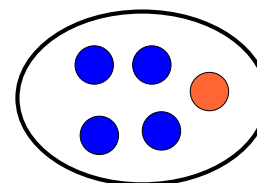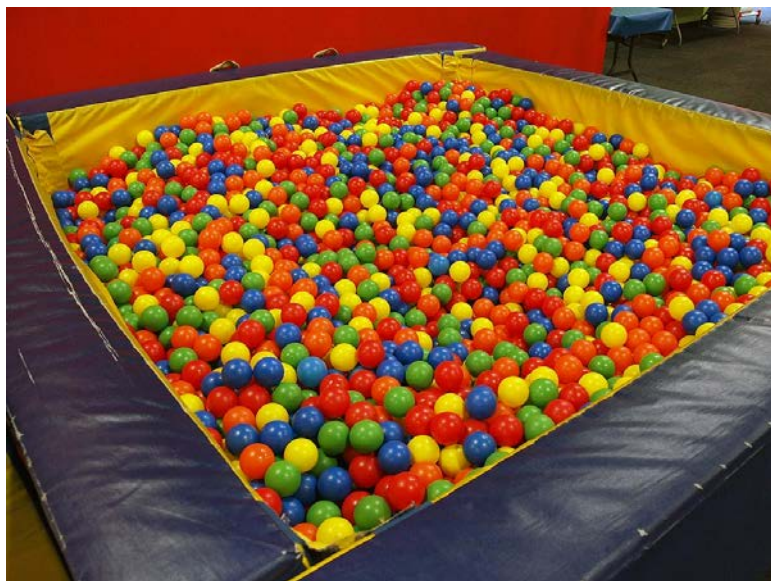
Questions to think about

Imagine that you make 1000 test using the threshold α for significance of each test,

Define a random variable X as the number of false positive (type I errors)?

How is X distributed?

How many type I errors do you expect?

4 basic steps

- State hypothesis
- Compute test statistic
- Determine p value
  - Compare TS with null
- Conclude
  - $P < 0.05$? 0.001? 0.00001?
  - Number of tests

Presenting thursday